Econometrics for Honors Students

Table of Contents

Chapter 1. Background issues in statistics

1.1. Statistical terminology[*]

Important definitions in statistics

It is not unusual for students to forget important concepts learned in an earlier course. This set of definitions is intended to stir memories of those wonderful times when you were learning statistics and econometrics. It is not intended to replace a statistics course but to provide you with a handy guide to the denfinition of some important terms in the statistical tools used by economists.

Random variables

Random experiment

A random experiment is an experiment whose outcome is uncertain.

Outcome space

The outcome space (also sometimes referred to as the sample space) is the list of all possible outcomes of a random experiment.

Example 1.1. Single toss of a coin.

Consider the toss of a coin. Since the outcome is uncertain, tossing the coin is an example of a random experiment. The outcome space consists of a heads and a tails. If we let *X* be 0 if the outcome is a heads and let *X* equal 1 if the outcome is a tails, then *X* is a random variable. Since *X* only can take on integer values (0 or 1), it is a discrete random variable.

Random variable

A random variable is a number that can be assigned to an outcome of a random experiment. A discrete random variable has a finite number of possible values while a continuous random variable has an infinite number of potential values.

Non-stochastic variable

A non-stochastic variable is any variable that is not a random variable; i.e., does not represent the outcome of a random experiment.

Example 1.2. Multiple tosses of a coin.

Let *x* equal the number of heads that occur when a coin is tossed n times. The tossing of the coin *n* times is a random experiment. The outcome space of this random experiment is an integar between 0 and *n*. Since the value *x* is equal represents the outcome of a random experiment, it is a random variable.

Random sample

A random sample of size *n* out of a population of size *N* has the characteristic that every member of the population is equally likely to be chosen.

Example 1.3. Height of college age women.

Consider a random sample of the population of college age women. The height, *x*, of any woman chosen from this population is a random variable with a value somewhere in the outcome space, where the outcome space is a number between (say) 24 and 96 inches. Since in theory we can have as accurate a measurement as we might like, *x* can be thought of as being a continuous random variable.

Probability

General terms

Probability distribution for a discrete random variable.

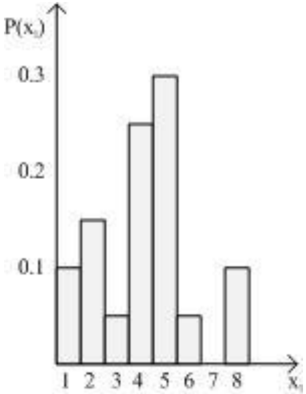
Consider a discrete random variable *x* $_i$ that represents an outcome of the *n* potential outcomes of a random experiment—that is, the set of potential outcomes is represented by $\mathbf{X}=(x_1, \ldots ,x_n)$. $\Pr(x_i)$ **Any function is a probability if and only if (1)** $\Pr(x_i) \geq 0 \text{ for all } i=1, \ldots ,n,$ **(2)** $\Pr\left(x_i \cup x_j\right)=\Pr(x_i)+\Pr\left(x_j\right)$ **for all i and j, and (3)** $\sum_{i=1}^{n} \Pr(x_i)=1.$ **An example of a discrete distribution is in Example 4.**

Example 1.4.

Discrete distribution.

Figure 1 illustrates a discrete probability distribution where *x* $_i$ goes from 1 to 8. The areas in the shaded rectangles sum to 1.

Figure 1.1. A discrete probability function



The areas of the rectangles sum to 1.

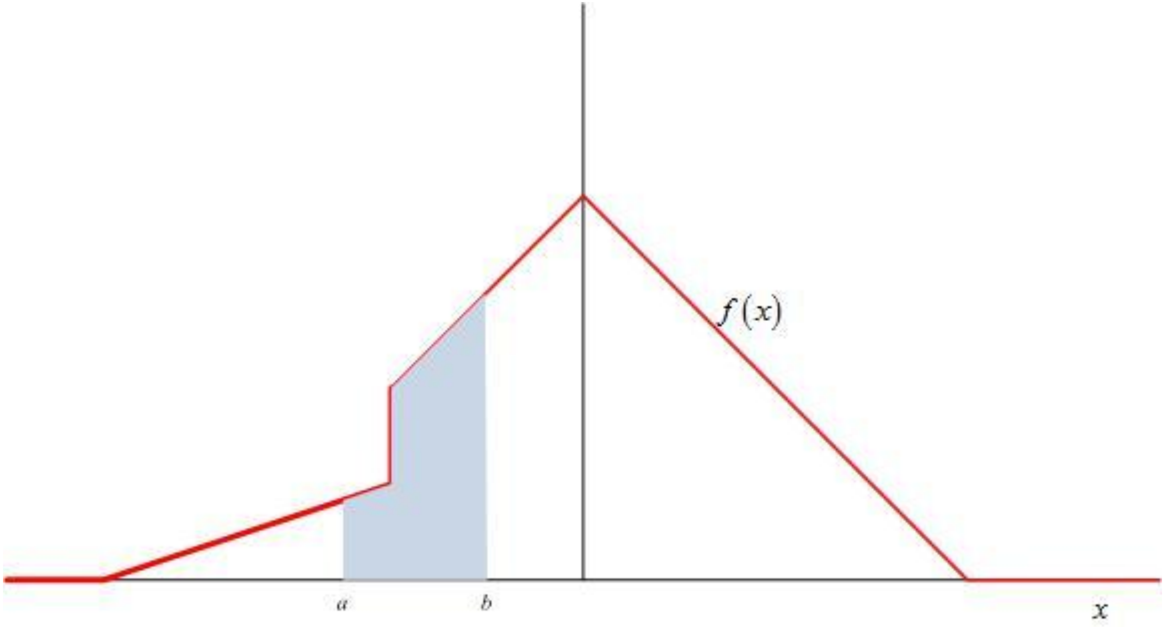
Probability density function.

If *x* $_i$ is a continuous random variable, the concept of a probability distribution is replaced by a probility density function (pdf). A function, *f*(*x*), is a pdf for the continuous random

variable x if and only if (1) *f*(*x*) ≥ 0 for – ∞ < *x* < ∞; (2) $\int_{-\infty}^{\infty} f(x)dx = 1;$ **and (3) *f*(*x*) has a finite number of discontinuities. By definition** $\Pr(a \leq x \leq b) = \int_{a}^{b} f(x)dx.$ **Example 5 offers an example of a pfd.**

Example 1.5. Probability distribution function for a continuous random variable.

Figure 1.2.



The red line is the pdf for the random variable *x*. The shaded in area under the pdf is equal to the probability that *x* falls between *a* and *b*. The total area under the pdf is equal to 1.

Cumulative distribution function (cdf).

The cumulative distribution function is given by *F*(*x*) = Pr(*X* ≤ *x*). For a discrete variable the cdf is $F(x_w) = \sum_{i=1}^{w} f(x_i).$ **For a continuous distribution, the cdf is** $F(x) = \int_{-\infty}^{x} f(w)dw.$ **Example 6 illustrates the calculation of the cumulative distribution function for a continuous random variable.**

Example 1.6. The cumulative distributon function.

Let *f*(*x*) = *x* ^{2} be the pdf for the random variable x defined between 0 and 1. The cumulative distribution function for any *a* is $F(a) = \int_0^a x^2\,dx = \frac{1}{3}x^3\Big|_0^a = \frac{a^3}{3}.$

Mathematical expectation

Mathematical expectation for a function.

The mathematical expectation of the function *g*(*x*) is $E(g(x)) = \int_x g(x) f(x) dx$ **where *x* is a random variable. Example 7 shows the calculation of the expected value of a function.**

Example 1.7. Expected value calculation.

Let *f*(*x*) = *x* ^{a} be a pdf for 0 ≤ *x* ≤ 1 and *a* > 0. Let *g*(*x*) = *x* ^{3} . We can calculate

$$E[g(x)] = \int_0^1 \left(x^3\right) x^a \, dx = \int_0^1 x^{a+3} dx = \frac{1}{a+4} x^{a+4} \Big|_0^1 = \frac{1}{a+4}.$$

The mean of a distribution.

The population mean, *μ*, of a random variable, x, with a pdf of *f*(*x*) is defined to be the expected value of x: $\mu = E(x) = \int x f(x) dx.$ **Example 8 illustrates the calculation of the population mean.**

Example 1.8. Calculation of the population mean.

Assume we have the same pdf used in Example 7. The population mean for this distribution is $\mu = E[x] = \int_0^1 (x)x^a\,dx = \int_0^1 x^{a+1}\,dx = \frac{1}{a+2}x^{a+2}\Big|_0^1 = \frac{1}{a+2}.$

The variance of a distribution.

The population variance, *σ* ², of a distribution is $\sigma^2 = E\left[(x-\mu)^2\right]$**. Example 9 shows a shortcut way to calculate the population variance.**

Example 1.9. Calculation of the population variance using the expected value operator.

Define the variance operator, *V*, to be:

$$V(x) = E\left[(x-\mu)^2\right].$$

Then,

$$E\left[(x-\mu)^2\right] = \int (x-\mu)^2 f(x)dx.$$

Squaring the term in the integral gives: $\int (x^2 - 2\mu x + \mu^2) f(x)dx = E(x^2 - 2\mu x + \mu^2).$

Expand of the left-hand-side of this equality:

$$\int x^2 f(x)dx - \int 2\mu x f(x)dx + \int \mu^2 f(x)dx = E(x^2) - E(2\mu x) + E(\mu^2).$$

Thus, we have established that:

$$E[(x - \mu)^2] = E(x^2) - E(2\mu x) + E(\mu^2).$$

Evaluating the last two terms gives

$$E(2\mu x) = \int 2\mu x f(x)dx = 2\mu \int x dx = 2\mu^2$$

and

$$E(\mu^2) = \int \mu^2 f(x)dx$$

or, since $\int f(x)dx = 1,$ **that** $E(\mu^2) = \mu^2.$ **Thus,** $E[(x - \mu)^2] = E(x^2) - 2\mu^2 + \mu^2$ **or**

$$E[(x - \mu)^2] = E(x^2) - \mu^2.$$

For example, in Example 8 we found that $\mu = \frac{1}{a+2}.$ **The expected value of *x*** ^{2} **is**

$$E\left[x^2\right] = \int_0^1 \left(x^2\right) x^a \, dx = \int_0^1 x^{a+2} \, dx = \frac{1}{a+3} x^{a+3} \Big|_0^1 = \frac{1}{a+3}.$$

Thus, the variance of the distribution is

$$V(x) = \frac{1}{a+3} - \left(\frac{1}{a+2}\right)^2$$

or

$$V(x) = \frac{(a+2)^2 - (a+3)}{(a+3)(a+2)^2} = \frac{a^2 + 3a + 1}{(a+3)(a+2)^2}.$$

Expected value operation rules.

As shown in Example 9, the expected value operation allows several linear operations. Let *a* and *b* be a non-stochastic variables and *x* be a random variable. Then we have

1. ***E*(*a*) = *a*,**

2. $E\left(ax^2 + bx + c\right) = aE\left(x^2\right) + b\mu + c.$

3. $E(\, a\, x + b\,) = a\, \mu + b,$

These rules work both for discrete and continuous random variables.

Joint distributions

The joint pdf for two random variables.

Any function, *f*(*x*,*y*), that has the characteristics

1. *f*(*x*,*y*) ≥ 0 for all *x* and *y* and

2. $$\int_y \int_x f(x, y) dx dy = 1$$

is a joint pdf. This definition can be extended easily to include more than two random variables.

Covariance between two random variables.

If x and y are random var een the two variables, *C o v*(*x,y*) or *σ* $_{xy}$, is defined to be $Cov(x,y)=E[(x-\mu_x)(y-\mu_y)]$. **Expansion gives the alternative definition that *σ* $_{xy}$ = *E*(xy) − *μ* $_x$ *μ* $_y$.**

Stochastic independence.

The random variables *x* and *y* are stochastically independent if and only if *σ* $_{xy}$ = 0. An equivalent definition of independence is that *x* and *y* are stochastically independent if and only if *f*(*x,y*) = *f*(*x*)*f*(*y*), or, in words, if the joint pdf of the two random variables is equal to the product of the pdf of each random variable. From the definition of covariance it is easy to see that if two random variables are stochastically independent then *E*(xy) = *μ* $_x$ *μ* $_y$.

Correlation coefficient.

The correlation coefficient, *ρ*, is defined to be $\rho_{xy}=\dfrac{\sigma_{xy}}{\sigma_x\sigma_y}.$ **The correlation coefficient is a unitless number that varies between -1 and +1. Clearly, two random variables are stochastically independent if and only if *ρ* $_{xy}$ = 0.**

Discrete distributions

Binomial distribution.

The discrete random variable *x* has a binomial distribution if $f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0, 1, \ldots, n \\ 0 & \text{elsewhere} \end{cases}$ **where** $\binom{n}{x} = \frac{n!}{x!(n-x)!}$. **For the binomial distribution, $\mu = n\,p$ and $\sigma^2 = n\,p(1-p)$.**

Uniform distribution.

The discrete random variable *x* has a uniform distribution if $f(x) = \left\{\begin{matrix} \frac{1}{b-a} \text{ if } a \le x \le b \\ 0 \ \text{ elsewhere} \end{matrix}\right\}$.

The mean and variance of the uniform distribution are $\mu = \frac{a+b}{2}$ **and** $\sigma^2 = \frac{(b-a)^2}{12}$.

Poisson distribution.

$$f(x)=\begin{cases}\dfrac{m^{x}e^{-m}}{x\,!}, & x=0,1,\ \ldots\\ 0 & \text{elsewhere}\end{cases}$$

has a Poisson distribution if

For the Poisson distribution $\mu = \sigma^{2} = m$. The Poisson distribution is used quite often in queuing theory to, among other things, describe the arrival of customers at a cashier's station.

Continuous distributions

Expotential distribution.

The continuous random variable x has an exponential distribution if

$$f(x)=\left.\begin{cases}\lambda e^{-\lambda x}, & \text{for } x\geq 0\\ 0 & \text{for } x<0\end{cases}\right\}.$$

The cumulative exponential distribution is given by $F(x) = 1 - e^{-\lambda x}$, for $x \geq 0$. The exponential distribution describes the times between events that occur continuously and independently at a constant rate (as in a Poisson process). The mean and variance of an exponential distribution are $\mu = \lambda^{-1}$ and $\sigma^{2} = \lambda^{-2}$.

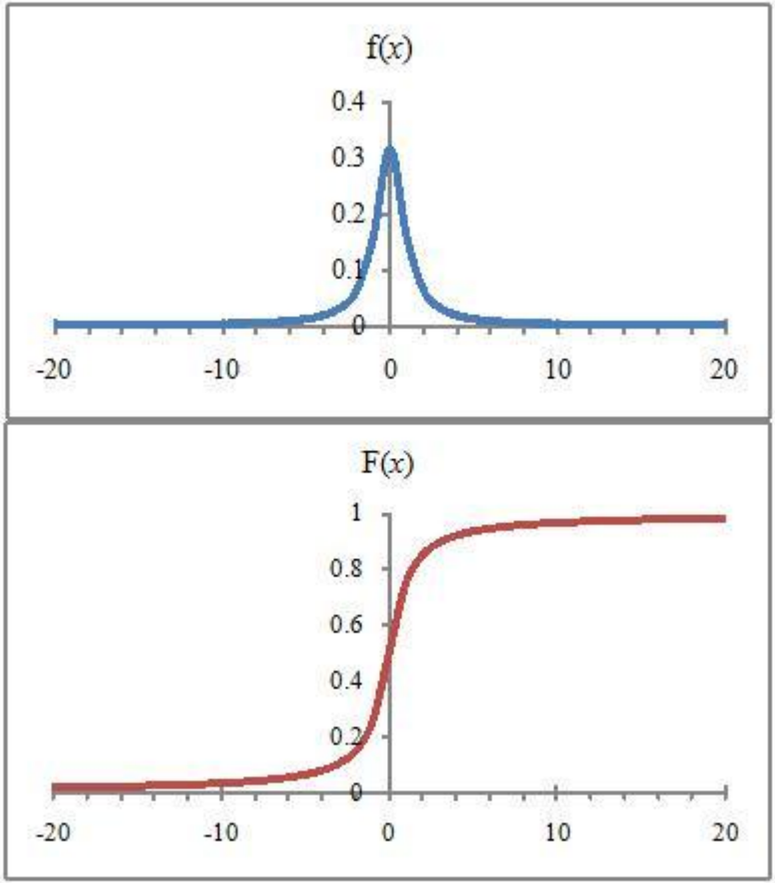
Cauchy distribution.

A randon < ∞, has a Cauchy (or Cauchy-Lorentz) distribution if its pdf is

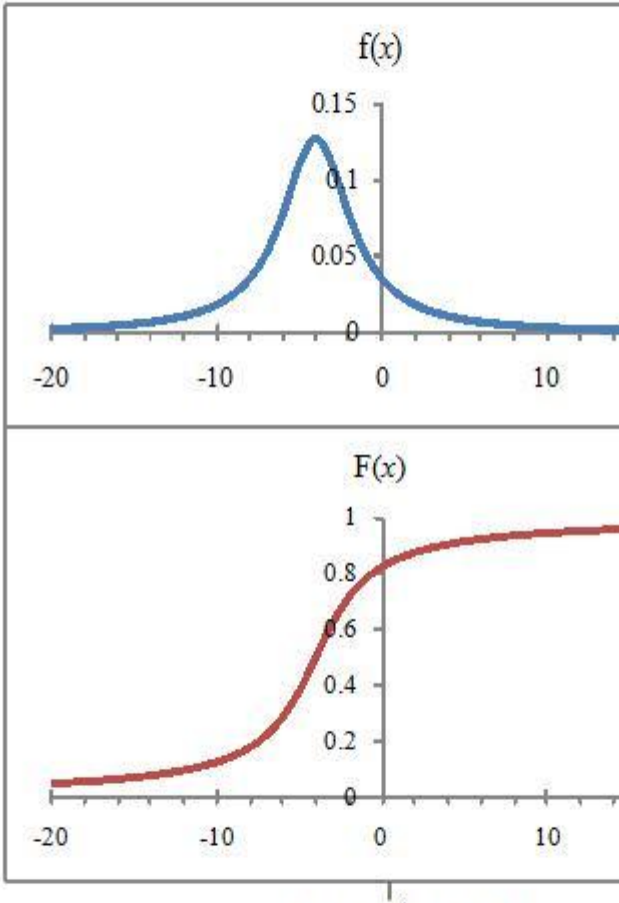
$$f(x)=\frac{1}{\pi}\left[\frac{\gamma}{(x-x_0)^2+\gamma^2}\right].$$

The parameter *x* $_0$ locates the peak of the pdf while *γ* specifies the half-width of the pdf at the half maximum. Figure 3 shows the pdf and cumulative function for two values of these two parameters.

Figure 1.3. The Cauchy distribution.



$\gamma = 1$ and $x_0 = 0$



$\gamma = 2.5$ and $x_0 = -4$

The two panels represent the Cauchy distribution for two sets of values of x_0 and γ.

Normal distribution.

The continuous random variable x has a normal distribution with a mean of μ and a variance of σ^2 if its pdf is $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ for $-\infty \leq x \leq \infty$. The distribution is symmetric around the mean.

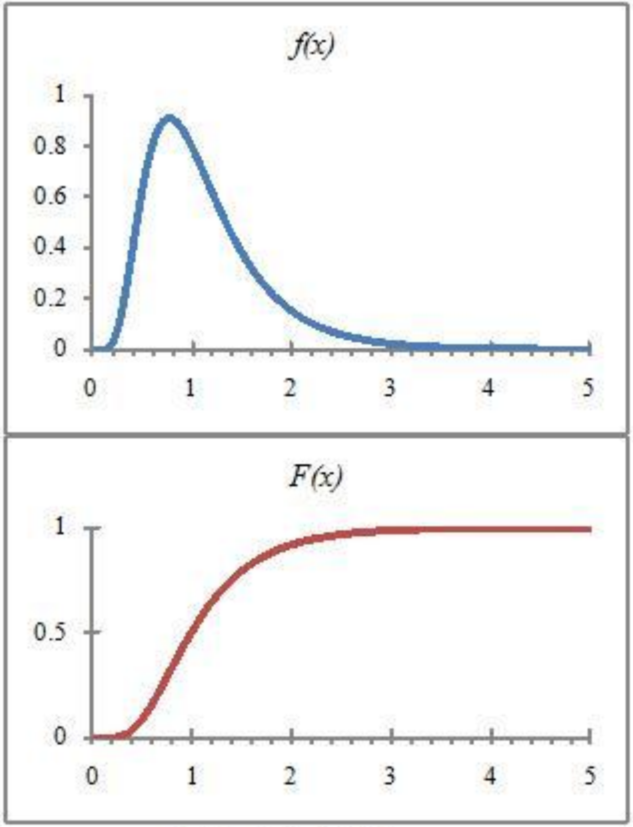
Log normal distribution.

The continuous random variable x has log normal distribution if y has a normal distribution and $x = e^y$. Thus, if $y \sim N(\mu, \sigma^2)$, then the pdf of a log normal distribution is

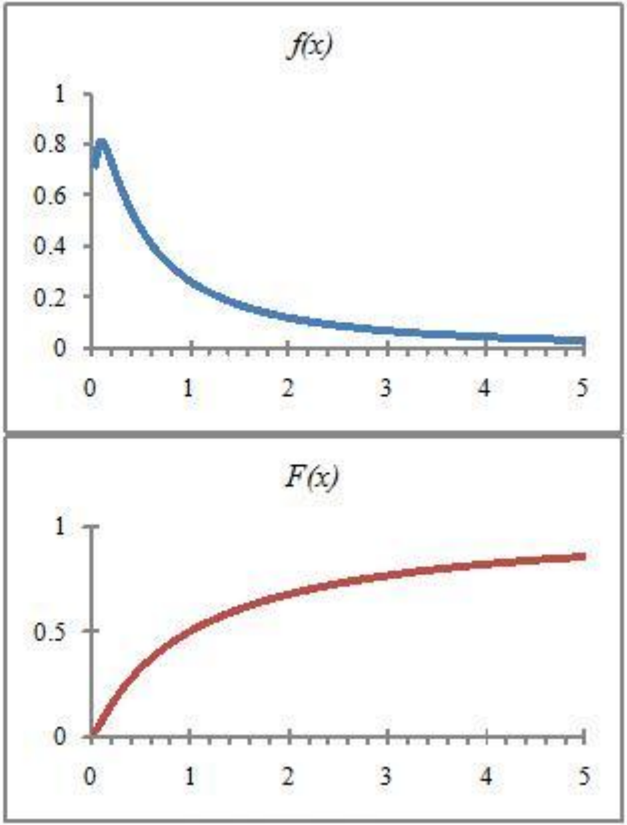
$$f(x)=\left\{\begin{matrix}\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}, & \text{for } x>0\\ 0 \quad \text{otherwise} & \end{matrix}\right\}.$$

The mean and variance of *x* are $\mu_x = e^{\mu+\frac{\sigma^2}{2}}$ **and** $\sigma_x^2=\left(e^{\sigma^2}-1\right)e^{2\mu+\sigma^2}.$ **Because the distribution is skewed downward for variances over 1, the log normal distribution is sometimes used to describe income distributions (where there are relatively few very wealthy people and incomes generally are positive. Figure 4 shows the graphs of the pdf and cumulative functions for the log normal distributions for two values of *σ*.**

Figure 1.4. The log-normal distribution.



$\mu = 0$ and σ=0.5



$\mu = 0$ and σ=1.5

The two panels illustrate the log-normal distribution for two values of *σ*..

Gamma distribution.

A positive random variable *x* has a gamma distribution if its pdf is

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha - 1} e^{-\frac{x}{\beta}}$$

for *x* > 0 and 0 elsewhere. *Γ*(*α*) is known as the gamma function and is defined to be

$$\Gamma(\alpha) = \int_0^{\infty} y^{\alpha - 1} e^{-y} dy = (\alpha - 1)\,!.$$

The gamma function is often used to model waiting times like waiting for death. Its mean and variance are given by *μ* = *α β* and *σ* ^{2} = *α β* ^{2} .

Chi-square distribution.

A chi-square distribution (*χ* ^{2} (*k*)) is the sum of *k* independent standard normal random variables and is a special case of the gamma distribution (with $\alpha = \frac{k}{2}$ and *β* = 2). The pdf

of a chi-square distribution with k degrees of freedom is $f(x) = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$ where $x > 0$. Its mean and variance are $\mu = k$ and $\sigma^2 = 2k$. If $y = \sum_{i=1}^{k} x_i^2$ where the x_i 's are independently drawn from the standard normal distribution (N(1, 0)), then $y_i \sim \chi^2(k)$.

Student's t-distribution.

Consider two random variables, x and v. Assume that $x \sim N(0,1)$ and $v \sim \chi^2(r)$ and are stochastically independent. Then the random variable $t = \frac{w}{\sqrt{\frac{v}{r}}}$ has the t-distribution with r *degrees of freedom*. The pdf and cumulative function of t are $f(t) = \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{r\pi}\,\Gamma\left(\frac{r}{2}\right)}\left(1+\frac{t^2}{r}\right)^{-\left(\frac{r+1}{2}\right)}$ and $F(t) = \frac{1}{2} + t\Gamma\left(\frac{t}{2}\right)$. The mean and variance of the distribution are 0 for $r > 1$ and $\frac{r}{r-2}$ for $t > 2$, respectively.[1] The t-distribution plays a

prominent role in hypothesis testing that is well-known to all undergraduate economics majors.

F distribution.

Consider two stochastically independent chi-square random variable such that

$u \sim \chi^2(r_1)$ **and** $v \sim \chi^2(r_2)$ **and *u,v* > 0. The new random variable** $f = \frac{\frac{u}{r_1}}{\frac{v}{r_2}}$ **has a F-distribution with *r* ₁ and *r* ₂ degrees of freedom. The pdf for the F-distribution is**

$$g(f)=\frac{\Gamma\left(\frac{r_1+r_2}{2}\right)\left(\frac{r_1}{r_2}\right)}{\Gamma\left(\frac{r_1}{2}\right)\Gamma\left(\frac{r_2}{2}\right)}\frac{f^{\frac{r_1}{2}-1}}{\left(1+\frac{r_1 f}{r_2}\right)^{\frac{r_1+r_2}{2}}}.$$

The F-distribution is used in testing if population variances are equal and in performing likelihood ratio tests.

Multinomial distribution.

Consider the *n* rando $x_i \sim N\left(\mu_i, \sigma_i^2\right)$ **,*x* $_2$,⋯,*x* $_n$ where each variable has a normal**

$\sigma_{ij} = E\left[(x_i - \mu_i)\left(x_j - \mu_{j.}\right)\right]$ **and the covariance between of the variables is**

ange the variances and covariances into a *n*-by-*n*

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix}$$

matrix where **that is known as the variance-covariance matrix.**

$$(\mathbf{x} - \boldsymbol{\mu}) = \begin{pmatrix} x_1 - \mu_1 \\ \vdots \\ x_n - \mu_n \end{pmatrix} \qquad (\mathbf{x} - \boldsymbol{\mu})'$$

as its transpose. Then,

$$(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma} (\mathbf{x} - \boldsymbol{\mu}) = \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - \mu_i)\left(x_j - \mu_j\right)\sigma_{ij},$$

where $\sigma_{ii} = \sigma_i^2$. If $|\boldsymbol{\Sigma}|$ **is the determinant of the variance-covariance matrix, then the pdf for the joint distribution of**

these random variables is

$$f(x_1, x_2, \ldots, x_n) = \frac{1}{(2\pi)^{n/2} |\mathbf{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})' \mathbf{\Sigma}(\mathbf{x}-\boldsymbol{\mu})}.$$

If the random variables are stochastically independent the covariances are equal to 0 and the pdf becomes

$$f(x_1, x_2, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}\left(\prod_{i=1}^{n} \sigma_1^2\right)^{\frac{1}{2}}} e^{-\frac{1}{2}\sum_{i=1}^{n} \frac{(x_i-\mu_i)^2}{\sigma_i^2}}.$$

If the *n* random variables are all drawn from the same normal distribution with a mean of μ and a variance of *σ* ^{2} , then the pdf simplifies to

$$f(x_1, x_2, \ldots, x_n) = \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} (x_i-\mu)^2}.$$

Characteristics of an estimator of a population parameter θ

Finite estimators

Bias.

The bias of an estimator is defined to be $B\left(\hat{\theta}\right)=E\left(\hat{\theta}\right)-\theta.$ **An estimator is unbiased if and only if** $B\left(\hat{\theta}\right)=0.$

Mean square error.

The mean square error (MSE) of an estimator is defined to be $MSE\left(\hat{\theta}\right)=E\left[\left(\hat{\theta}-\theta\right)^2\right].$ **It is relatively easy to show that** $MSE\left(\hat{\theta}\right)=V\left(\hat{\theta}\right)+\left(B\left(\hat{\theta}\right)\right)^2.$ **Often a biased estimator with a smaller MSE may be preferred to an unbiased estimator with a relatively larger MSE.**

Efficiency.

An estimator $\hat{\theta}$ is relatively more efficient than $\tilde{\theta}$ if and only if $V\left(\hat{\theta}\right) < V\left(\tilde{\theta}\right).$ Generally, we would prefer to use the most efficient estimator available (if it is unbiased).

Asymtoptic estimators

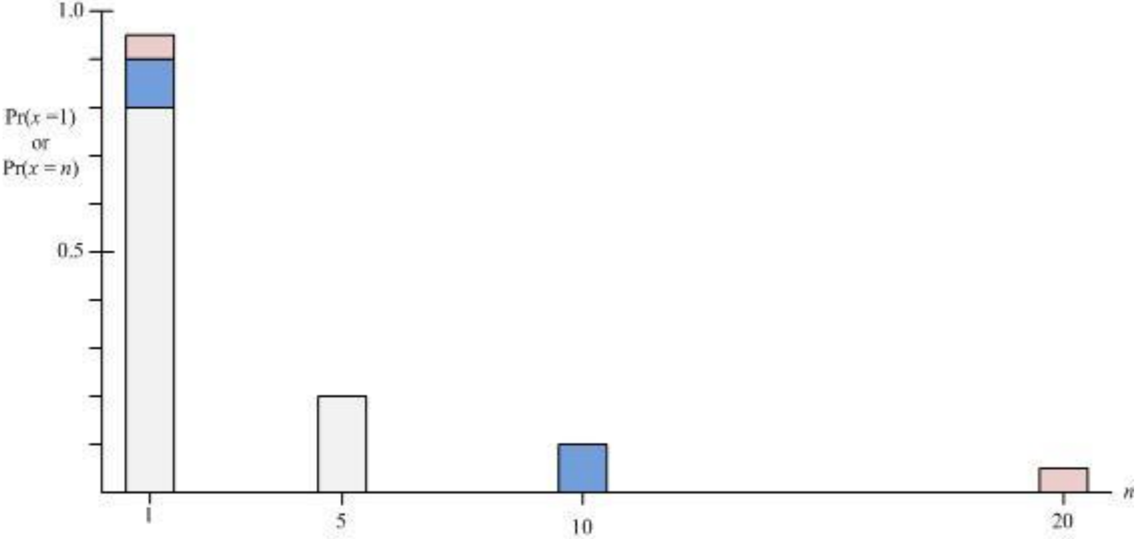
plim.

x_n converges to a constant, *c*, if $\lim_{n \to \infty} \Pr(|x_n - c| > \varepsilon) = 0$ for any positive *ε*. We can write this relationship as *plim*x_n = *c*.

Example 1.10.

Greene[2] offers this example of plim: Suppose x_n equals 0 with probability $1 - \left(\frac{1}{n}\right)$ and *n* with probability $\left(\frac{1}{n}\right).$ As *n* increases, the second point becomes more remote from the first point. However, at the same time the probability of observing the second point becomes more and more unlikely. This effect is shown in Figure 5 where as *n* increases the probability distribution concentrates more and more on 1.

Figure 1.5. Example of plim.



The probability x = 1 is the area of the gray box centered on 1 for n = 5; the gray area plus the blue area for n = 10; and the sum of the gray, blue, and red areas for n = 20; the probability x = n is the area of the box centered on n.

Consistency.

The estimator $\hat{\theta}$ is a consistent estimator of ϑ if and only if $\operatorname{plim}\hat{\theta} = \theta.$

Asymmtotically unbiased.

An estimator $\hat{\theta}$ is an asymtotically unbiased estimator of ϑ if $\lim_{n \to \infty} E\left[\hat{\theta}\right] = \theta.$

1.2. The maximum likelihood estimation method[*]

The Maximum Likelihood Method

Introduction

The maximum likelihood (ML) method is an alternative to ordinary least squares (OLS) and offers a more general approach to the problem of finding estimators of unknown

population parameters. In these notes we present an intuitive introduction to the ML technique. We begin our discussion with a description of continuous random variables.

Continuous random variables

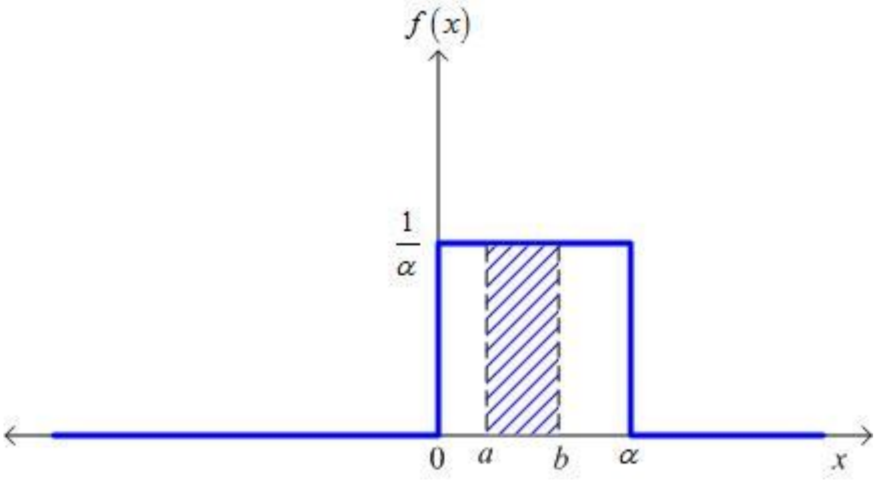
Assume that *x* is a continuous random variable over the interval – ∞ ≤ *x* ≤ ∞. Because of the assumption of continuity we need some special definitions.

***Probability density function*. Any function *f*(*x*) that has the following characteristics is a probability density function (pdf): (1) *f*(*x*) > 0 and (2) $\int_{-\infty}^{\infty} f(x)dx = 1.$ The probability that *x* has a value between *a* and *b* is given by $\Pr(a \le x \le b) = \int_{a}^{b} f(x)dx.$ Here are two examples of the probability density functions (pdf) of continuous random variables.**

Example 1.11. Uniform distribution

Let $f(x) = \frac{1}{\alpha}$ for $0 \leq x \leq \alpha$ and 0 elsewhere, where $\alpha > 0$. A graph of the pdf for this distribution is shown in Figure 1.

Figure 1.6. Probability distribution function of a uniform distribution.



The probability *x* falls between *a* and *b* is given by the colored in area.

It is easy to see from the graph that $f(x) = \frac{1}{\alpha} > 0$ **and**

$$\Pr(a \le x \le b) = \int_{-\infty}^{\infty} f(x)dx = \int_{0}^{\alpha} \frac{1}{\alpha} dx = 1.$$

Moreover, as shown in Figure 1, the area under the pdf curve between *a* and *b* is equal to the probability that *x* lies between *a* and *b*; that is,

$$\Pr(a \le x \le b) = \int_{a}^{b} \left(\frac{1}{\alpha}\right) dx = \frac{x}{\alpha}\bigg|_{a}^{b} = \frac{b-a}{\alpha}.$$

The calculation of the mean and variance of this distribution is relatively simple. The population mean is given by $\mu_x = E(x) = \int_{0}^{\alpha} x f(x) dx$ **or** $\mu_x = \int_{0}^{\alpha} x\left(\frac{1}{\alpha}\right) dx = \frac{x^2}{2\alpha}\bigg|_{0}^{\alpha} = \frac{\alpha}{2}.$

The population variance[3] is given by $V(x)=E\left[(x-\mu_x)^2.\right]$ Thus,

$$V(x)=\int_0^\alpha \left(x-\frac{\alpha}{2}\right)^2\left(\frac{1}{\alpha}\right)dx=\int_0^\alpha \left(x^2-\alpha x+\frac{\alpha^2}{4}\right)\left(\frac{1}{\alpha}\right)dx$$

or

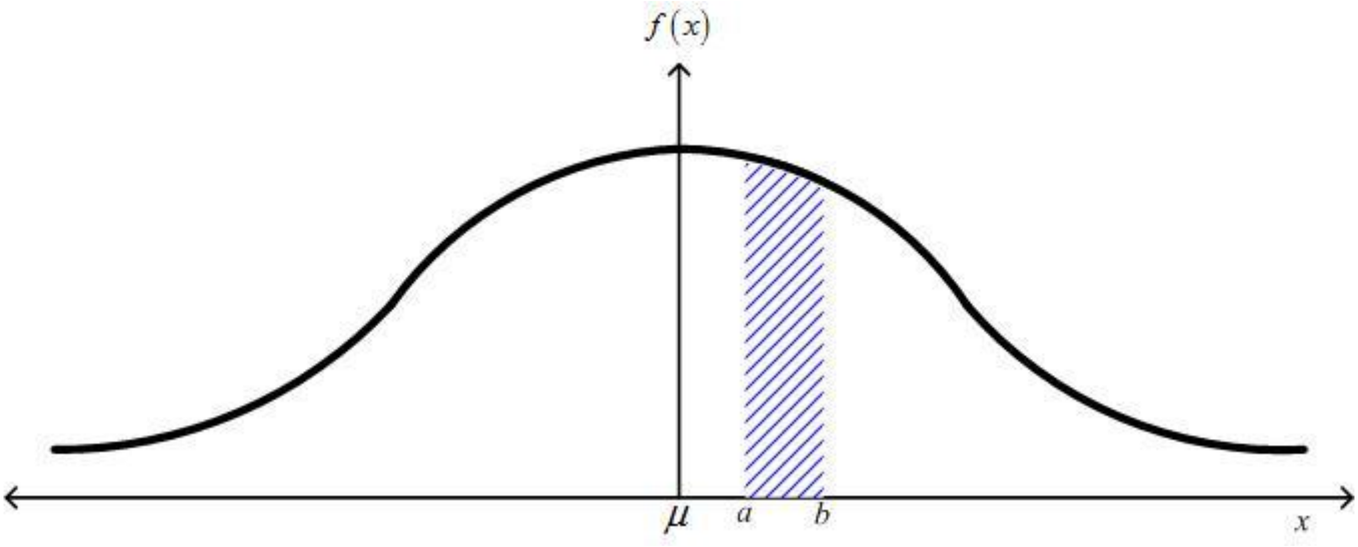
$$V(x)=\frac{x^3}{3\alpha}-\frac{x^2}{2}+\frac{\alpha}{4}x\Big|_0^\alpha=\frac{\alpha^2}{3}-\frac{\alpha^2}{2}+\frac{\alpha^2}{4}=\frac{\alpha^2}{12}.$$

Because of the simple mathematical form of the uniform pdf, the calculations in Example 1 are relatively straight forward. While the calculations for random variables with a pdf that has a more complicated form are generally more difficult (if algebraically possible), the basic methodology remains the same. Example 2 considers the case of a more complicated pdf.

Example 1.12. The Normal distribution.

A random variable with a mean of *μ* and a variance of *σ* ^{2} that has a *normal distribution*—that is, $x \sim N\left(\mu, \sigma^2\right)$— **has the pdf** $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. **A typical graph of this pdf is given in Figure 2. The area under the curve between values of *x* of *a* and *b* is equal to the probability that *x* falls between *a* and *b*.**

Figure 1.7. Probability distribution function of a Normal distribution.



The probability *x* falls between *a* and *b* is given by the shaded area.

Joint distributions of samples and the ML method.

Most of the statistical work that economists use involves the use of a sample of observations. It is usual to assume that the members of the sample are drawn independently of each other. The implication of this assumption is that *the pdf of the joint distribution is equal to the product of the pfd of each observation*; i.e.,

(1.1)

$$f(x_1, x_2, \ldots, x_n) = f(x_1) f(x_2) \cdots f(x_n).$$

The pdf of the joint distribution shown in (1) is known as the *likelihood function*. If the sample were not independently drawn, the pdf of joint distribution could not be written in such a simple form because of the covariance among the members of the sample would not be equal to zero. The logarithm of this function (or as it is referred to, the log of the likelihood function) is given by the sum

$$L(x_1, x_2, \ldots, x_n) = \ln f(x_1) + \ln f(x_2) + \cdots + \ln f(x_n) = \sum_{i=1}^{n} \ln f(x_i).$$

The maximum likelihood method involves choosing as estimators of the unknown parameters of the distribution the values that maximize the likelihood function. However, because the logarithm is a monotonically increasing function[4], maximizing the log of the likelihood function is equivalent to maximizing the likelihood function. The following example of this procedure illustrates how to derive ML estimators.

Example 1.13. The ML estimator of the population mean and population variance.

Assume that $x \sim N\left(\mu, \sigma^2\right)$. **Consider a sample of size n drawn independently from this distribution. The likelihood function is the product of the pdf of each observation or:**

(1.2)

$$f(x_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \Rightarrow L(x_1, x_2, \ldots, x_n) = \frac{1}{\sigma^n (2\pi)^{\frac{n}{2}}} e^{-\frac{\sum_{i=1}^{n} (x_i-\mu)^2}{2\sigma^2}}.$$

Thus, the log of the likelihood function of this sample is

$$L(x_1, x_2, \ldots, x_n) = -\frac{n\ln 2\pi}{2} - n\ln\sigma - \frac{\sum_{i=1}^{n} (x_i-\mu)^2}{2\sigma^2}.$$

In the ML method we want to find the estimators of the mean and variance, $\widehat{\mu}$ **and** $\widehat{\sigma}$ **, that maximize the log of the likelihood function. Substituting in the parameter estimates into the log of the likelihood function gives our problem as:**

(1.3)

$$\underset{\widehat{\mu}\ ,\ \widehat{\sigma}}{Max}\ L(x_1, x_2,\ \ldots\ , x_n) = \underset{\widehat{\mu}\ ,\ \widehat{\sigma}}{Max}\left[-\frac{n\ln 2\pi}{2} - n\ln\ \widehat{\sigma}\ - \frac{\sum (x_i - \ \widehat{\mu}\ )^2}{2\ \widehat{\sigma}\ ^2}\right].$$

Setting the derivatives of the log of the likelihood function with respect to $\widehat{\mu}$ and $\widehat{\sigma}$ equal to 0 gives:

(1.4)

$$\frac{\partial L(x_1, x_2,\ \ldots\ , x_n)}{\partial\ \widehat{\mu}} = \frac{\sum (x_i - \ \widehat{\mu}\ )}{\widehat{\sigma}\ ^2} = 0 \ \text{ and}$$

(1.5)

$$\frac{\partial L(x_1, x_2,\ \ldots\ , x_n)}{\partial\ \widehat{\sigma}} = -\frac{n}{\widehat{\sigma}} + \frac{\sum (x_i - \ \widehat{\mu}\ )^2}{\widehat{\sigma}\ ^3} = 0.$$

Solving these two equations simultaneously gives:

(1.6)

$$\widehat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x} \text{ and } \widehat{\sigma}^2 = \frac{\sum (x_i - \widehat{\mu})^2}{n}.$$

Notice the fact that the estimator of the population mean is equal to the sample mean, a result that is the same as the one you found in your introductory statistics course. However, the *unbiased* estimator of the population variance used in that course is

$$s^2 = \frac{\sum (x_i - \widehat{\mu})^2}{n-1}.$$

Thus, one of the common "problems" with using a ML estimator is that quite often they are *biased estimators* of a population parameter. On the other hand, under very general conditions ML estimators are *consistent*, are *asymptotically efficient*, and have an *asymptotically normal distribution* (these are desirable large sample size characteristics of potential estimators and are discussed in advanced statistics courses).[5]

Application of the ML method to regressions

The discussion above illustrates the basics of the ML method—you form the log of the likelihood function and then find the values of the parameter estimates that maximize this function. In most cases the maximization will not yield answers in closed form—that is, you cannot find a neat algebraic formula as we did for the population mean. However, you can use computer programs to search for the values of the parameter estimates that maximize this function. Thus, in most cases in advanced regression models you often will treat the ML method as a "black box" and not concern yourself with the estimation details. However, I illustrate one more example of the ML technique.

Example 1.14. The ML estimators for a simple regression.

Assume that we want to estimate the population parameters for the regression model $y_i = \beta x_i + \varepsilon_i$, where we assume that

1. $\varepsilon_i \sim N(0, \sigma^2)$,
2. $E(\varepsilon_i \varepsilon_j) = 0$ **for $i \neq j$,**
3. $y_i = Y_i - \overline{Y}$ **and** $x_i = X_i - \overline{X}$ **(this assumption allows us to ignore the estimation of the intercept term), and**

4. *x*$_i$ is a non-stochastic variable.

The assumption of a normally distributed error term implies that $\varepsilon_i = y_i - \beta x_i \sim N(0, \sigma^2)$.

Thus, the pdf of the error term is $f(\varepsilon_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i - \beta x_i)^2}{2\sigma^2}}$ **and, thus, the likelihood function[6] is:**

(1.7)

$$\prod_{i=1}^{n} f(\varepsilon_i) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i - \beta x_i)^2}{2\sigma^2}} = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \prod_{i=1}^{n} e^{-\frac{(y_i - \beta x_i)^2}{2\sigma^2}}$$

and the log of the likelihood function is

$$L(\varepsilon_1, \varepsilon_2, \ \ldots, \varepsilon_n) = -n\ln\sqrt{2\pi} - n\ln \widehat{\sigma} - \frac{\sum_{i=1}^{n}\left(y_i - \widehat{\beta}\ x_i\right)^2}{2\ \widehat{\sigma}^{\ 2}}.$$

We find the estimators $\widehat{\beta}$ and $\widehat{\sigma}$ in the same manner as we did for the sample mean and variance. Differentiating the log of the likelihood function and setting these first derivatives equal to 0 gives the following two first-order conditions:

(1.8)

$$\frac{\partial L(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)}{\partial \widehat{\beta}} = \frac{2\sum_{i=1}^{n} \left(y_i - \widehat{\beta}\, x_i\right) x_i}{2\, \widehat{\sigma}^{\,2}} = 0$$

and

(1.9)

$$\frac{\partial L(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)}{\partial \widehat{\sigma}} = -\frac{n}{\widehat{\sigma}} + \frac{\sum_{i=1}^{n} \left(y_i - \widehat{\beta}\, x_i\right)^2}{\widehat{\sigma}^{\,3}} = 0.$$

Thus, the ML estimators are:

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} \text{ and } \widehat{\sigma}^{\,2} = \frac{\sum_{i=1}^{n} \left(y_i - \widehat{\beta}\, x_i\right)^2}{n}.$$

Notice that in this simple case the ML estimator of β is the same as the OLS estimator of β . Also, notice that the ML estimator of σ^2 is biased—the (unbiased) OLS estimator of σ^2 is

$$s^2 = \frac{\sum_{i=1}^{n} \left(y_i - \widehat{\beta}\, x_i\right)^2}{n-2}.$$

You can use the examples in this module as the basis of your understanding of the ML method. When you see that the ML method is used in a computer program, you can be fairly certain that the program uses one of the many optimizing subroutines to find the maximum of the log of the likelihood program. You can consult the help files with the computer program to see what underlying distribution is used to set up the log of the likelihood function. A concept related to the maximum likelihood estimation method

worth exploring is the likelihood ratio test (see the module by Don Johnson entitled The Likelihood Ratio Test for an introduction to this key statistical test.)

Exercises

Exercise 1.2.1.

Consider the following functions. For each of them, (1) prove that the function is a pdf; (2) calculate the mean and variance of each distribution, and (3) find the maximum likelihood estimator of the parameter ϑ. Sketch a graph of each of the distributions for a representative value of ϑ.

1. $f(x;\vartheta) = (\vartheta + 1)x^{\vartheta}$ where $0 \le x \le 1$ and $\vartheta > 0$.
2. $f(x;\vartheta) = \vartheta e^{-\vartheta x}$ where $0 \le x < \infty$ and $\vartheta > 0$.

[1] **The mean of the t-distribution is undefined for *t* ≤ 1. The variance of the distribution is ∞ for 1 < *r* ≤ 2 and undefined for *r* ≤ 1.**

[2] **Greene, William H. (1990). *Econometric Analysis* (New York: Macmillan Publishing Company): 103.**

[3] **Quite often, as in the exercises at the end of this module, it is easier to calculate the variance of a distribution using the alternative formula for the variance:** $\sigma_x^2 = V(x) = E(x-\mu)^2 = E\left(x^2\right) - \mu^2,$ **where** $E\left(x^2\right) = \int x^2 f(x)dx.$

[4] **The function *g*(*y*) is monotonically increasing for y if *g* ′ (*y*) > 0. Because** $\frac{d}{dx}\ln x = \frac{1}{x} > 0 \text{for} x > 0,$ **the logarithm function is monotonically increasing for positive values of *x*.**

[5] **Intuitively, what these concepts mean is that as the sample size increases the estimator becomes more precise (the variance becomes smaller and an bias disappears) and the distribution of the estimator approaches the normal distribution. The formal definitions of these terms involve advanced statistical concepts that are reported here only in the**

interest of completeness. An estimator $\left(\hat{\theta}\right)$ of the parameter ϑ is consistent if and only if $\operatorname{plim}\hat{\theta} = \theta.$ This estimator has an asymptotically normal distribution if $\hat{\theta} \xrightarrow{a} N\left(\theta, \{\mathbf{I}(\theta)\}^{-1}\right).$ An unbiased estimator is more efficient that another unbiased estimator if it has a smaller variance than the alternative estimator. An asymptotically efficient is an estimator whose mean square error tends to zero as the sample size increases. The mean square error (MSE) is defined to be

$$MSE\left(\hat{\theta}\right) = E\left[\left(\hat{\theta} - \theta\right)^2\right] = V\left(\hat{\theta}\right) + \left(Bias\left[\hat{\theta}\right]\right)^2.$$

An estimator is asymptotically efficient if

$$\lim_{n \to \infty} MSE\left(\hat{\theta}\right) = 0.$$

See any advanced statistics text or Statistical terminology for further information on these concepts.

[6] **The symbol $\prod_{i=1}^{n} x_1$ is equivalent to the product $x_1 x_2 \cdots x_n$.**

Chapter 1. Background issues in statistics

1.1. Statistical terminology[*]

Important definitions in statistics

It is not unusual for students to forget important concepts learned in an earlier course. This set of definitions is intended to stir memories of those wonderful times when you were learning statistics and econometrics. It is not intended to replace a statistics course but to provide you with a handy guide to the denfinition of some important terms in the statistical tools used by economists.

Random variables

Random experiment

A random experiment is an experiment whose outcome is uncertain.

Outcome space

The outcome space (also sometimes referred to as the sample space) is the list of all possible outcomes of a random experiment.

Example 1.1. Single toss of a coin.

Consider the toss of a coin. Since the outcome is uncertain, tossing the coin is an example of a random experiment. The outcome space consists of a heads and a tails. If we let *X* be 0 if the outcome is a heads and let *X* equal 1 if the outcome is a tails, then *X* is a random variable. Since *X* only can take on integer values (0 or 1), it is a discrete random variable.

Random variable

A random variable is a number that can be assigned to an outcome of a random experiment. A discrete random variable has a finite number of possible values while a continuous random variable has an infinite number of potential values.

Non-stochastic variable

A non-stochastic variable is any variable that is not a random variable; i.e., does not represent the outcome of a random experiment.

Example 1.2. Multiple tosses of a coin.

Let *x* equal the number of heads that occur when a coin is tossed n times. The tossing of the coin *n* times is a random experiment. The outcome space of this random experiment is an integar between 0 and *n*. Since the value *x* is equal represents the outcome of a random experiment, it is a random variable.

Random sample

A random sample of size *n* out of a population of size *N* has the characteristic that every member of the population is equally likely to be chosen.

Example 1.3. Height of college age women.

Consider a random sample of the population of college age women. The height, *x*, of any woman chosen from this population is a random variable with a value somewhere in the outcome space, where the outcome space is a number between (say) 24 and 96 inches. Since in theory we can have as accurate a measurement as we might like, *x* can be thought of as being a continuous random variable.

Probability

General terms

Probability distribution for a discrete random variable.

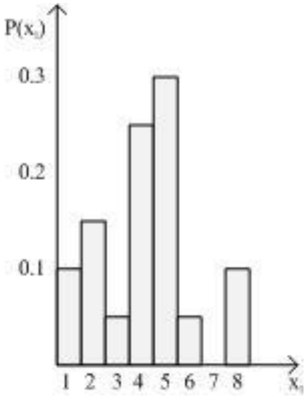
Consider a discrete random variable *x* $_i$ that repr $\mathbf{X}=(x_1, \ldots, x_n)$. $\Pr(x_i)$ *n* potential outcomes of a random experiment—that is, the set of potential outcomes is represented by Any functi $\sum_{i=1}^{n} \Pr(x_i)=1.$ ity if and only if (1) $\Pr(x_i)\geq 0 \text{ for all } i=1, \ldots, n,$ (2) $\Pr(x_i \cup x_j)=\Pr(x_i)+\Pr(x_j)$ for all i and j, and (3) An example of a discrete distribution is in Example 4.

Example 1.4.

Discrete distribution.

Figure 1 illustrates a discrete probability distribution where *x* $_i$ goes from 1 to 8. The areas in the shaded rectangles sum to 1.

Figure 1.1. A discrete probability function



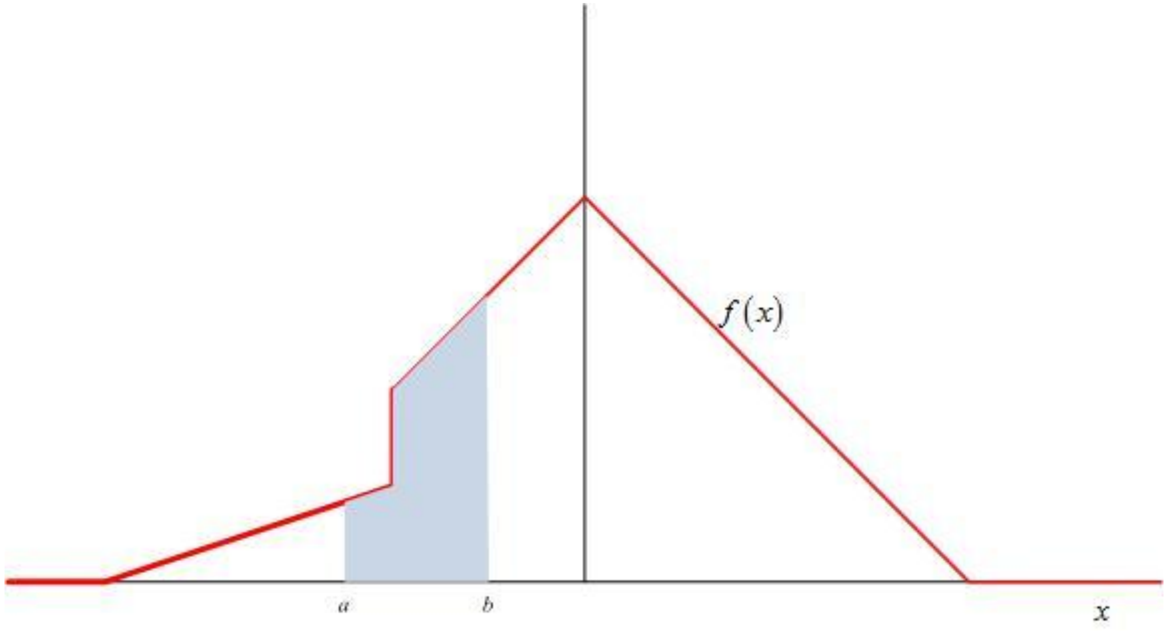
The areas of the rectangles sum to 1.

Probability density function.

If *x* $_i$ is a continuous random variable, the concept of a probability distribution is replaced by a probility de (pdf). A function, ***f***(*x*), is a pdf for the continuous random va (*x*) ≥ 0 for – ∞ < *x* < ∞; (2) $\int_{-\infty}^{\infty} f(x)dx = 1;$ and (3) ***f***(*x*) has a finite number of discontinuities. By definition $\Pr(a \le x \le b) = \int_{a}^{b} f(x)dx.$ Example 5 offers an example of a pfd.

Example 1.5. Probability distribution function for a continuous random variable.

Figure 1.2.



The red line is the pdf for the random variable *x*. The shaded in area under the pdf is equal to the probability that *x* falls between *a* and *b*. The total area under the pdf is equal to 1.

Cumulative distribution function (cdf).

The cumulative distribution function is given by $F(x) = \Pr(X \le x)$. For a discrete variable the cdf is $F(x_w) = \sum_{i=1}^{w} f(x_i).$ For a continuous distribution, the cdf is $F(x) = \int_{-\infty}^{x} f(w)dw.$ Example 6 illustrates the calculation of the cumulative distribution function for a continuous random variable.

Example 1.6. The cumulative distributon function.

Let $f(x) = x^2$ be the pdf for the random variable x defined between 0 and 1. The cumulative distribution function for any a is

$$F(a) = \int_0^a x^2\,dx = \tfrac{1}{3}x^3\Big|_0^a = \frac{a^3}{3}.$$

Mathematical expectation

Mathematical expectation for a function.

The mathematical expectation of the function $g(x)$ is $E(g(x)) = \int_x g(x)f(x)dx$ where x is a random variable. Example 7 shows the calculation of the expected value of a function.

Example 1.7. Expected value calculation.

Let *f*(*x*) = *x* ^{a} be a pdf for 0 ≤ *x* ≤ 1 and *a* > 0. Let *g*(*x*) = *x* ^{3} . We can calculate

$$E[g(x)] = \int_0^1 (x^3) x^a \, dx = \int_0^1 x^{a+3} \, dx = \frac{1}{a+4} x^{a+4} \Big|_0^1 = \frac{1}{a+4}.$$

The mean of a distribution.

The population mean, *μ*, of a random variable, x, with a pdf of *f*(*x*) is defined to be the expected value of x: $\mu = E(x) = \int x f(x) dx.$ **Example 8 illustrates the calculation of the population mean.**

Example 1.8. Calculation of the population mean.

Assume we have the same pdf used in Example 7. The population mean for this distribution is

$$\mu = E[x] = \int_0^1 (x) x^a \, dx = \int_0^1 x^{a+1} \, dx = \frac{1}{a+2} x^{a+2} \Big|_0^1 = \frac{1}{a+2}.$$

The variance of a distribution.

The population variance, *σ* ^{2} , of a distribution is $\sigma^2 = E\left[(x-\mu)^2\right]$ **Example 9 shows a shortcut way to calculate the population variance.**

Example 1.9. Calculation of the population variance using the expected value operator.

Define the variance operator, *V*, to be:

$$V(x) = E\left[(x-\mu)^2\right]$$

Then,

$$E\left[(x-\mu)^2\right] = \int (x-\mu)^2 f(x)dx.$$

Squaring the term in the integral gives: $\int\left(x^2 - 2\mu x + \mu^2\right)f(x)dx = E\left(x^2 - 2\mu x + \mu^2\right).$

Expand of the left-hand-side of this equality:

$$\int x^2 f(x)dx - \int 2\mu x f(x)dx + \int \mu^2 f(x)dx = E\left(x^2\right) - E(2\mu x) + E\left(\mu^2\right).$$

Thus, we have established that:

$$E\left[(x-\mu)^2\right] = E\left(x^2\right) - E(2\mu x) + E\left(\mu^2\right).$$

Evaluating the last two terms gives

$$E(2\mu x) = \int 2\mu x f(x)dx = 2\mu \int x dx = 2\mu^2$$

and

$$E\left(\mu^2\right) = \int \mu^2 f(x)dx$$

or, since $\int f(x)dx = 1,$ **that** $E\left(\mu^2\right) = \mu^2.$ **Thus,** $E\left[(x-\mu)^2\right] = E\left(x^2\right) - 2\mu^2 + \mu^2$ **or**

$$E\left[(x-\mu)^2\right] = E\left(x^2\right) - \mu^2.$$

For example, in Example 8 we found that $\mu = \frac{1}{a+2}.$ **The expected value of *x*** ^{2} **is**

$$E\left[x^2\right] = \int_0^1 \left(x^2\right)x^a\,dx = \int_0^1 x^{a+2}\,dx = \frac{1}{a+3}x^{a+3}\Big|_0^1 = \frac{1}{a+3}.$$

Thus, the variance of the distribution is

$$V(x) = \frac{1}{a+3} - \left(\frac{1}{a+2}\right)^2$$

or $$V(x) = \frac{(a+2)^2 - (a+3)}{(a+3)(a+2)^2} = \frac{a^2+3a+1}{(a+3)(a+2)^2}.$$

Expected value operation rules.

As shown in Example 9, the expected value operation allows several linear operations. Let *a* and *b* be a non-stochastic variables and *x* be a random variable. Then we have

1. ***E*(*a*) = *a*,**
2. $E\left(ax^2 + bx + c\right) = aE\left(x^2\right) + b\mu + c.$
3. ***E*(*a x* + *b*) = *a μ* + *b*,**

These rules work both for discrete and continuous random variables.

Joint distributions

The joint pdf for two random variables.

Any function, *f*(*x,y*), that has the characteristics

1. ***f*(*x,y*) ≥ 0 for all *x* and *y* and**
2. $\int_y \int_x f(x,y)dxdy = 1$

is a joint pdf. This definition can be extended easily to include more than two random variables.

Covariance between two random variables.

If x and y are random variables, then the covariance between the two variables, *C o v*(*x,y*) or σ_{xy}, is defined to be $Cov(x,y) = E[(x-\mu_x)(y-\mu_y)]$. **Expansion gives the alternative definition that $\sigma_{xy} = E(xy) - \mu_x\mu_y$.**

Stochastic independence.

The random variables *x* and *y* are stochastically independent if and only if $\sigma_{xy} = 0$. An equivalent definition of independence is that *x* and *y* are stochastically independent if and only if $f(x,y) = f(x)f(y)$, or, in words, if the joint pdf of the two random variables is equal to the product of the pdf of each random variable. From the definition of covariance it is easy to see that if two random variables are stochastically independent then $E(xy) = \mu_x\mu_y$.

Correlation coefficient.

The correlation coefficient, *ρ*, is defined to be $\rho_{xy} = \dfrac{\sigma_{xy}}{\sigma_x\sigma_y}$. **The correlation coefficient is a unitless number that varies between -1 and +1. Clearly, two random variables are stochastically independent if and only if $\rho_{xy} = 0$.**

Discrete distributions

Binomial distribution.

The discrete random variable *x* has a binomial distribution if $f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0, 1, \ldots, n \\ 0 & \text{elsewhere} \end{cases}$ **where** $\binom{n}{x} = \dfrac{n!}{x!(n-x)}$. **For the binomial distribution, $\mu = n\,p$ and $\sigma^2 = n\,p(1-p)$.**

Uniform distribution.

The discrete random variable *x* has a uniform distribution if $f(x)=\begin{cases}\frac{1}{b-a} & \text{if } a\le x\le b\\ 0 & \text{elsewhere}\end{cases}$**. The mean and variance of the uniform distribution are** $\mu=\frac{a+b}{2}$ **and** $\sigma^2=\frac{(b-a)^2}{12}$.

Poisson distribution.

The discrete random variable *x* has a Poisson distribution if $f(x)=\begin{cases}\frac{m^x e^{-m}}{x!}, & x=0,1,\ \ldots\\ 0 & \text{elsewhere}\end{cases}$ **For the Poisson distribution $\mu = \sigma^2 = m$. The Poisson distribution is used quite often in queuing theory to, among other things, describe the arrival of customers at a cashier's station.**

Continuous distributions

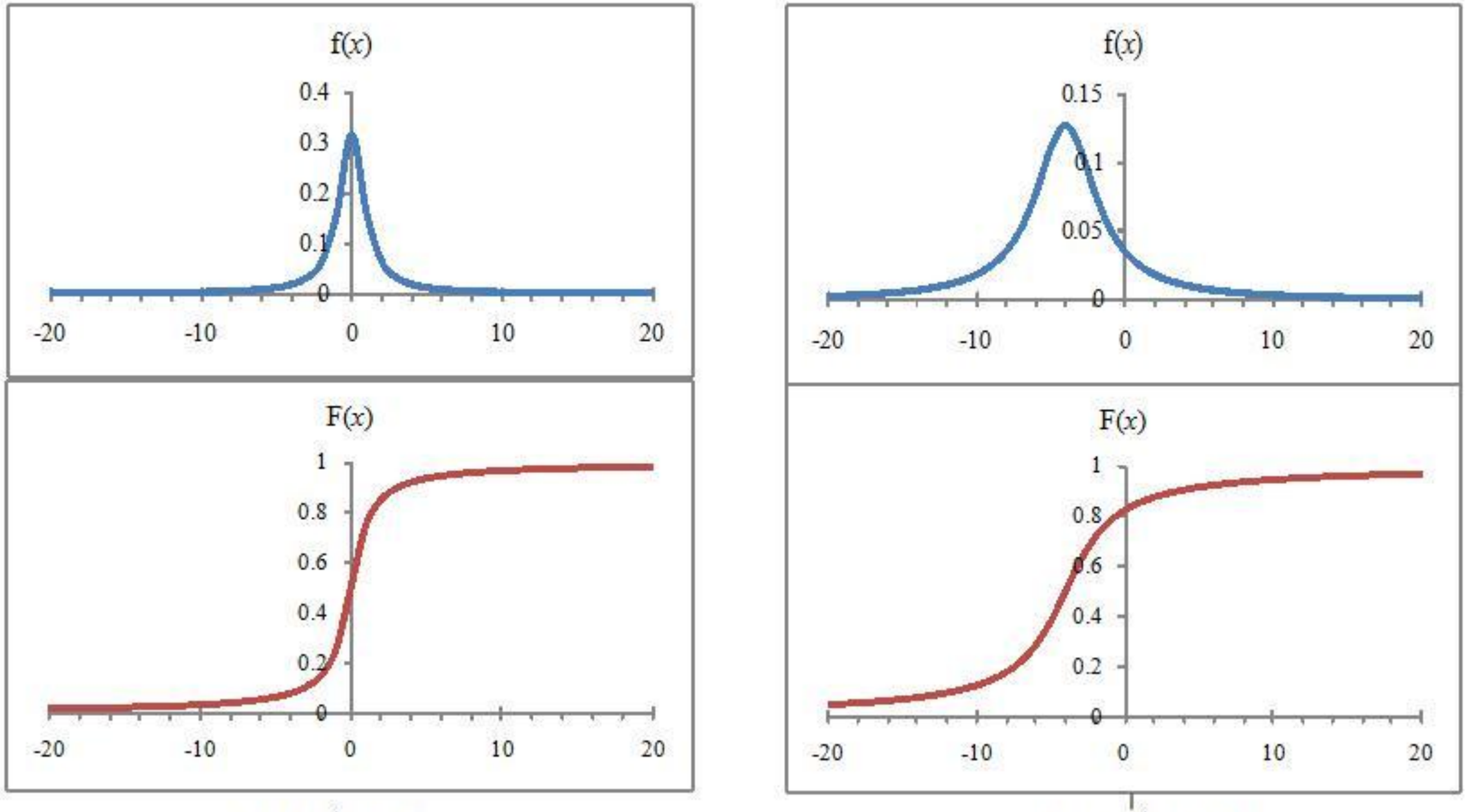
Expotential distribution.

The continuous random variable *x* has an exponential distribution if $f(x)=\begin{cases}\lambda e^{-\lambda x}, & \text{for } x\ge 0\\ 0 & \text{for } x<0\end{cases}$**. The cumulative exponential distribution is given by $F(x) = 1 - e^{-\lambda x}$, for $x \ge 0$. The exponential distribution describes the times between events that occur continuously and independently at a constant rate (as in a Poisson process). The mean and variance of an exponential distribution are $\mu = \lambda^{-1}$ and $\sigma^2 = \lambda^{-2}$.**

Cauchy distribution.

A random variable *x*, where $-\infty < x < \infty$, has a Cauchy (or Cauchy-Lorentz) distribution if its pdf is $f(x)=\frac{1}{\pi}\left[\frac{\gamma}{(x-x_0)^2+\gamma^2}\right]$**. The parameter x_0 locates the peak of the pdf while γ specifies the half-width of the pdf at the half maximum. Figure 3 shows the pdf and cumulative function for two values of these two parameters.**

Figure 1.3. The Cauchy distribution.



$\gamma = 1$ and $x_0 = 0$

$\gamma = 2.5$ and $x_0 = -4$

The two panels represent the Cauchy distribution for two sets of values of x_0 and γ.

Normal distribution.

variable *x* has a normal distribution with a mean of *μ* and a variance of *σ* ^{2} if its pdf is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

for − ∞ ≤ *x* ≤ ∞. The distribution is symmetric around the mean.

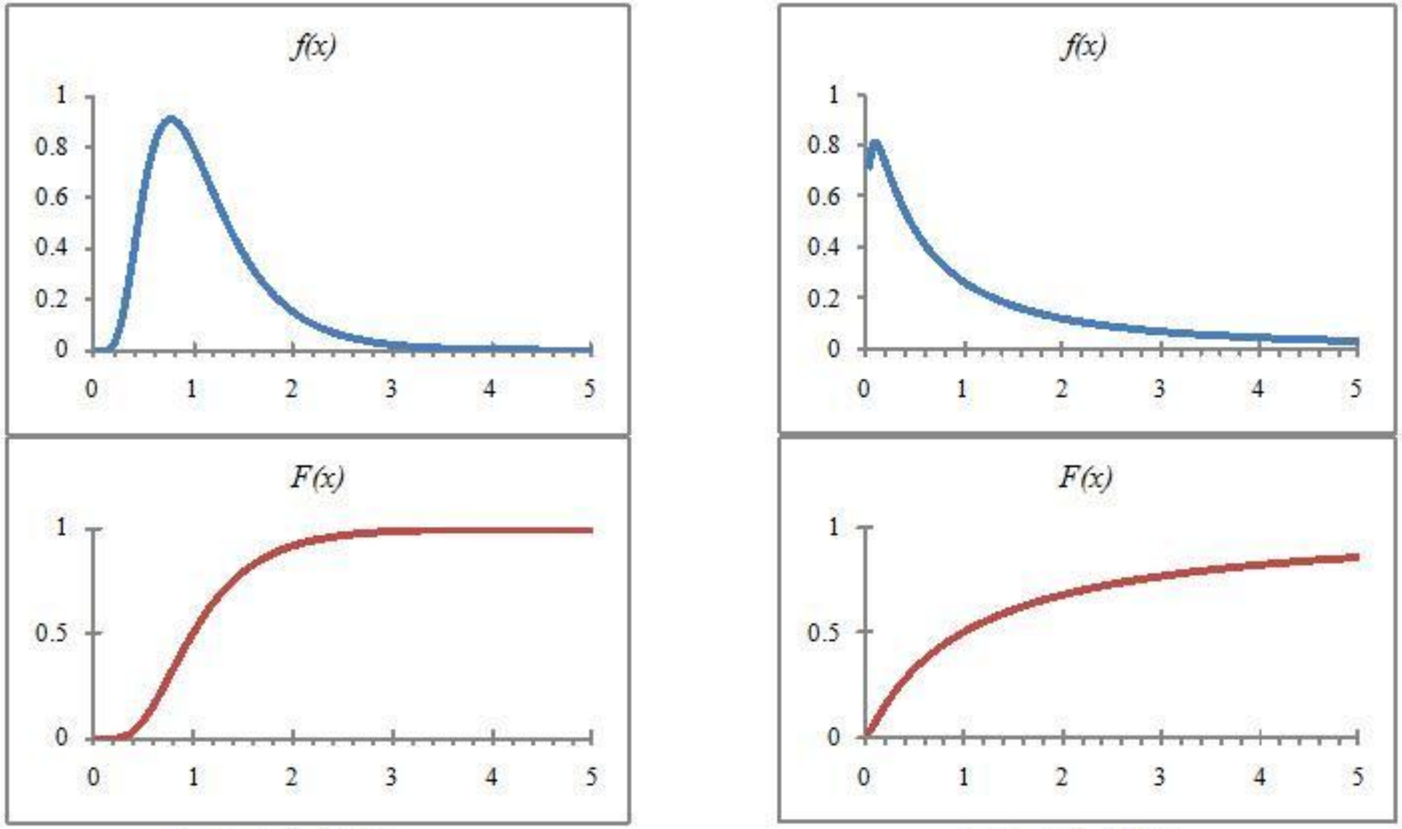
Log normal distribution.

The continuous random variable *x* has log normal distribution if *y* has a normal distribution and *x* = *e* ^{y}. Thus, if $y \sim N(\mu, \sigma^2)$, **then the pdf of a log normal distribution is**

$$f(x) = \begin{Bmatrix} \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}, & \text{for } x > 0 \\ 0 & \text{otherwise} \end{Bmatrix}.$$

The mean and variance of *x* are $\mu_x = e^{\mu + \frac{\sigma^2}{2}}$ **and** $\sigma_x^2 = \left(e^{\sigma^2} - 1\right)e^{2\mu + \sigma^2}$. **Because the distribution is skewed downward for variances over 1, the log normal distribution is sometimes used to describe income distributions (where there are relatively few very wealthy people and incomes generally are positive. Figure 4 shows the graphs of the pdf and cumulative functions for the log normal distributions for two values of *σ*.**

Figure 1.4. The log-normal distribution.



$\mu = 0$ and σ=0.5

$\mu = 0$ and σ=1.5

The two panels illustrate the log-normal distribution for two values of *σ*..

Gamma distribution.

A positive random variable *x* has a gamma d[illegible] $f(x)=\frac{1}{\Gamma(\alpha)\beta^{\alpha}}x^{\alpha-1}e^{-\frac{x}{\beta}}$ for *x* > 0 and 0 elsewhere. *Γ*(*α*) is known as the gamma function and is defined to be $\Gamma(\alpha)=\int_0^{\infty}y^{\alpha-1}e^{-y}dy=(\alpha-1)\,!.$ The gamma function is often used to model waiting times like waiting for death. Its mean and variance are given by *μ* = *α β* and $\sigma^2 = \alpha\beta^2$.

Chi-square distribution.

A chi-square distribution (χ^2 (*k*)) is the sum of *k* independent standard normal random variables and is a special case of the gamma distribution (with $\alpha=\frac{k}{2}$ and *β* = 2). The pdf of a chi-square distribution with *k* degrees of freedom is $f(x)=\frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)}x^{\frac{k}{2}-1}e^{-\frac{x}{2}}$ where *x* > 0. Its mean and variance are *μ* = *k* and σ^2 = 2*k*. If $y=\sum_{i=1}^{k}x_i^2$ where the x_i 's are independently drawn from the standard normal distribution (N(1, 0)), then $y_i \sim \chi^2 (k)$.

Student's t-distribution.

Consider two random variables, *x* and *v*. Assume that *x* ~ *N*(0,1) and $v \sim \chi^2 (r)$ and are stochastically independent. Then the random variable $t=\frac{w}{\sqrt{\frac{v}{r}}}$ has the t-distribution with *r degrees of freedom*. The pdf and cumulative function of *t* are $f(t)=\frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{r\pi}\Gamma\left(\frac{r}{2}\right)}\left(1+\frac{t^2}{r}\right)^{-\left(\frac{r+1}{2}\right)}$ and $F(t)=\frac{1}{2}+t\Gamma\left(\frac{t}{2}\right).$ The mean and variance of the distribution are 0 for *r* > 1 and $\frac{r}{r-2}$ for *t* > 2, respectively.[1] The t-distribution plays a prominent role in hypothesis testing that is well-known to all undergraduate economics majors.

F distribution.

Consider two sto lly independent chi-square random variable such that $u \sim \chi^2(r_1)$ and $v \sim \chi^2(r_2)$ and *u*,*v* > 0. The new

$$f = \frac{\frac{u}{r_1}}{\frac{v}{r_2}}$$

ution with *r*$_1$ and *r*$_2$ degrees of freedom. The pdf for the F-distribution is

$$g(f) = \frac{\Gamma\left(\frac{r_1 + r_2}{2}\right)\left(\frac{r_1}{r_2}\right)}{\Gamma\left(\frac{r_1}{2}\right)\Gamma\left(\frac{r_2}{2}\right)} \frac{f^{\frac{r_1}{2} - 1}}{\left(1 + \frac{r_1 f}{r_2}\right)^{\frac{r_1 + r_2}{2}}}.$$

The F-distribution is used in testing if population variances are equal and in performing likelihood ratio tests.

Multinomial distribution.

Consider the *n* random variables *x*$_1$,*x*$_2$,⋯,*x*$_n$ where each variable has a normal distribution—that is, $x_i \sim N\left(\mu_i, \sigma_i^2\right)$ and the covariance between of the variables is $\sigma_{ij} = E\left[(x_i - \mu_i)\left(x_j - \mu_{j.}\right)\right]$ We can arrange the variances and covariances into a *n*-by-*n* matrix where $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix}$ that is known as the variance-covariance matrix. Define the vector $(\mathbf{x} - \boldsymbol{\mu}) = \begin{pmatrix} x_1 - \mu_1 \\ \vdots \\ x_n - \mu_n \end{pmatrix}$ and $(\mathbf{x} - \boldsymbol{\mu})'$ as its transpose. Then,

$$(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma} (\mathbf{x} - \boldsymbol{\mu}) = \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - \mu_i)\left(x_j - \mu_j\right)\sigma_{ij},$$

where *σ*$_{ii}$ = *σ*$_i{}^2$. If $|\boldsymbol{\Sigma}|$ is the determinant of the variance-covariance matrix, then the pdf for the joint distribution of these random variables is

$$f(x_1, x_2, \ldots, x_n) = \frac{1}{(2\pi)^{n/2} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma} (\mathbf{x} - \boldsymbol{\mu})}.$$

If the random variables are stochastically independent the covariances are

equal to 0 and the pdf becomes

$$f(x_1, x_2, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}\left(\prod_{i=1}^{n} \sigma_1^2\right)^{\frac{1}{2}}} e^{-\frac{1}{2}\sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2}}.$$

If the *n* random variables are all drawn from the same normal distribution with a mean of μ and a variance of σ^2, then the pdf simplifies to

$$f(x_1, x_2, \ldots, x_n) = \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} (x_i - \mu)^2}.$$

Characteristics of an estimator of a population parameter θ

Finite estimators

Bias.

The bias of an estimator is defined to be $B\left(\hat{\theta}\right) = E\left(\hat{\theta}\right) - \theta.$ **An estimator is unbiased if and only if** $B\left(\hat{\theta}\right) = 0.$

Mean square error.

The mean square error (MSE) of an estimator is defined to be $MSE\left(\hat{\theta}\right) = E\left[\left(\hat{\theta} - \theta\right)^2\right].$ **It is relatively easy to show that** $MSE\left(\hat{\theta}\right) = V\left(\hat{\theta}\right) + \left(B\left(\hat{\theta}\right)\right)^2.$ **Often a biased estimator with a smaller MSE may be preferred to an unbiased estimator with a relatively larger MSE.**

Efficiency.

An estimator $\hat{\theta}$ is relatively more efficient than $\tilde{\theta}$ if and only if $V\left(\hat{\theta}\right) < V\left(\tilde{\theta}\right).$ Generally, we would prefer to use the most efficient estimator available (if it is unbiased).

Asymtoptic estimators

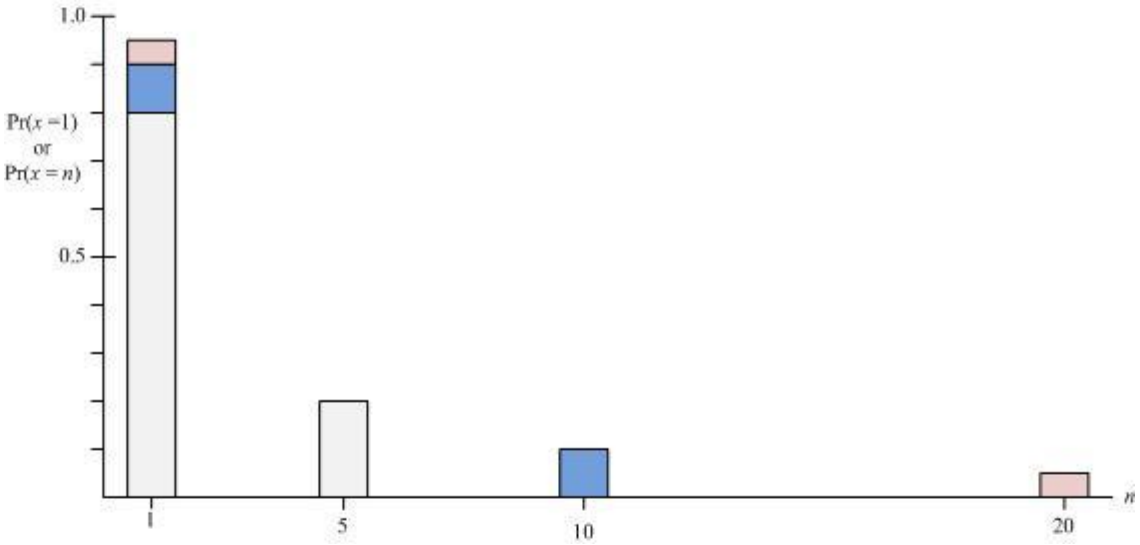
plim.

x $_n$ **converges to a constant, *c*, if $\lim_{n \to \infty} \Pr(|x_n - c| > \varepsilon) = 0$ for any positive *ε*. We can write this relationship as *plimx*** $_n$ **= *c*.**

Example 1.10.

Greene[2] offers this example of plim: Suppose *x* $_n$ **equals 0 with probability $1 - \left(\frac{1}{n}\right)$ and *n* with probability $\left(\frac{1}{n}\right).$ As *n* increases, the second point becomes more remote from the first point. However, at the same time the probability of observing the second point becomes more and more unlikely. This effect is shown in Figure 5 where as *n* increases the probability distribution concentrates more and more on 1.**

Figure 1.5. Example of plim.



The probability x = 1 is the area of the gray box centered on 1 for n = 5; the gray area plus the blue area for n = 10; and the sum of the gray, blue, and red areas for n = 20; the probability x = n is the area of the box centered on n.

Consistency.

The estimator $\hat{\theta}$ **is a consistent estimator of *ϑ* if and only if** $\text{plim}\,\hat{\theta} = \theta.$

Asymmtotically unbiased.

An estimator $\hat{\theta}$ **is an asymtotically unbiased estimator of *ϑ* if** $\lim_{n \to \infty} E\left[\hat{\theta}\right] = \theta.$

1.2. The maximum likelihood estimation method[*]

The Maximum Likelihood Method

Introduction

The maximum likelihood (ML) method is an alternative to ordinary least squares (OLS) and offers a more general approach to the problem of finding estimators of unknown population parameters. In these notes we present an intuitive introduction to the ML technique. We begin our discussion with a description of continuous random variables.

Continuous random variables

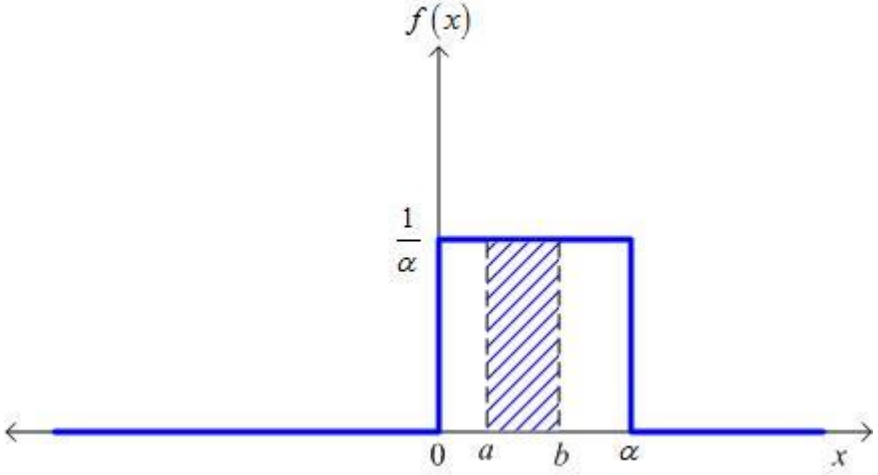
Assume that x is a continuous random variable over the interval $-\infty \le x \le \infty$. Because of the assumption of continuity we need some special definitions.

Probability density function. Any function $f(x)$ that has the following characteristics is a probability density function (pdf): (1) $f(x) > 0$ and (2) $\int_{-\infty}^{\infty} f(x)dx = 1.$ The probability that x has a value between a and b is given by $\Pr(a \le x \le b) = \int_{a}^{b} f(x)dx.$ Here are two examples of the probability density functions (pdf) of continuous random variables.

Example 1.11. Uniform distribution

Let $f(x) = \frac{1}{\alpha}$ for $0 \le x \le \alpha$ and 0 elsewhere, where $\alpha > 0$. A graph of the pdf for this distribution is shown in Figure 1.

Figure 1.6. Probability distribution function of a uniform distribution.



The probability *x* falls between *a* and *b* is given by the colored in area.

It is easy to see from the graph that $f(x)=\frac{1}{\alpha}>0$ **and** $\Pr(a\le x\le b)=\int\limits_{-\infty}^{\infty} f(x)dx=\int\limits_{0}^{\alpha}\frac{1}{\alpha}dx=1.$ **Moreover, as shown in Figure 1, the area under the pdf curve between *a* and *b* is equal to the probability that *x* lies between *a* and *b*; that is,**

$$\Pr(a\le x\le b)=\int\limits_{a}^{b}\left(\frac{1}{\alpha}\right)dx=\frac{x}{\alpha}\Big|_{a}^{b}=\frac{b-a}{\alpha}.$$

The calculation of the mean and variance of this distribution is relatively simple. The population mean is given by

$\mu_x=E(x)=\int\limits_{0}^{\alpha}xf(x)dx$ **or** $\mu_x=\int\limits_{0}^{\alpha}x\left(\frac{1}{\alpha}\right)dx=\frac{x^2}{2\alpha}\Big|_{0}^{\alpha}=\frac{\alpha}{2}.$

The population variance[3] is given by $V(x)=E\left[(x-\mu_x)^2.\right]$Thus, $V(x)=\int_0^{\alpha}\left(x-\frac{\alpha}{2}\right)^2\left(\frac{1}{\alpha}\right)dx=\int_0^{\alpha}\left(x^2-\alpha x+\frac{\alpha^2}{4}\right)\left(\frac{1}{\alpha}\right)dx$ or $V(x)=\frac{x^3}{3\alpha}-\frac{x^2}{2}+\frac{\alpha}{4}x\Big|_0^{\alpha}=\frac{\alpha^2}{3}-\frac{\alpha^2}{2}+\frac{\alpha^2}{4}=\frac{\alpha^2}{12}.$

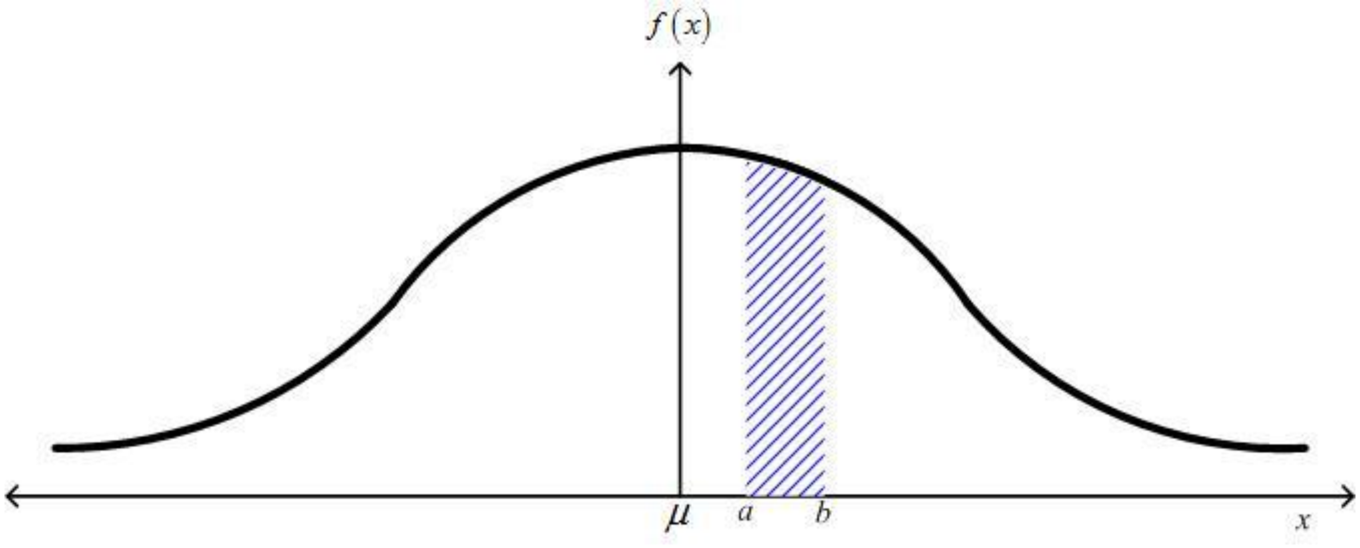
Because of the simple mathematical form of the uniform pdf, the calculations in Example 1 are relatively straight forward. While the calculations for random variables with a pdf that has a more complicated form are generally more difficult (if algebraically possible), the basic methodology remains the same. Example 2 considers the case of a more complicated pdf.

Example 1.12. The Normal distribution.

A random variable with a mean of *μ* and a variance of *σ* ^{2} that has a *normal distribution*—that is, $x\sim N\left(\mu,\sigma^2\right)$— **has the pdf**

$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ **. A typical graph of this pdf is given in Figure 2. The area under the curve between values of *x* of *a* and *b* is equal to the probability that *x* falls between *a* and *b*.**

Figure 1.7. Probability distribution function of a Normal distribution.



The probability *x* falls between *a* and *b* is given by the shaded area.

Joint distributions of samples and the ML method.

Most of the statistical work that economists use involves the use of a sample of observations. It is usual to assume that the members of the sample are drawn independently of each other. The implication of this assumption is that *the pdf of the joint distribution is equal to the product of the pfd of each observation*; i.e.,

(1.1)

$$f(x_1, x_2, \ \ldots, x_n) = f(x_1)f(x_2) \cdots f(x_n).$$

The pdf of the joint distribution shown in (1) is known as the *likelihood function*. If the sample were not independently drawn, the pdf of joint distribution could not be written in such a simple form because of the covariance among the members of the sample would not be equal to zero. The logarithm of this function (or as it is referred to, the log of the likelihood function) is given by the sum $L(x_1, x_2, \ldots, x_n) = \ln f(x_1) + \ln f(x_2) + \cdots + \ln f(x_n) = \sum_{i=1}^{n} \ln f(x_i).$ **The maximum likelihood method involves choosing as estimators of the unknown parameters of the distribution the values that maximize the likelihood function. However, because the logarithm is a monotonically increasing function[4], maximizing the log of the likelihood function is equivalent to maximizing the likelihood function. The following example of this procedure illustrates how to derive ML estimators.**

Example 1.13. The ML estimator of the population mean and population variance.

Assume that $x \sim N\left(\mu, \sigma^2\right).$ **Consider a sample of size n drawn independently from this distribution. The likelihood function is the product of the pdf of each observation or:**

(1.2)

$$f(x_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \Rightarrow L(x_1, x_2, \ldots, x_n) = \frac{1}{\sigma^n (2\pi)^{\frac{n}{2}}} e^{-\frac{\sum_{i=1}^{n} (x_i-\mu)^2}{2\sigma^2}}.$$

Thus, the log of the likelihood function of this sample is $L(x_1, x_2, \ldots, x_n) = -\frac{n\ln 2\pi}{2} - n\ln\sigma - \frac{\sum_{i=1}^{n} (x_i-\mu)^2}{2\sigma^2}.$ **In the ML method we want to find the estimators of the mean and variance,** $\widehat{\mu}$ **and** $\widehat{\sigma}$ **, that maximize the log of the likelihood function. Substituting in the parameter estimates into the log of the likelihood function gives our problem as:**

(1.3)

$$\underset{\widehat{\mu}\,,\,\widehat{\sigma}}{Max}\; L(x_1, x_2, \ldots, x_n) = \underset{\widehat{\mu}\,,\,\widehat{\sigma}}{Max}\left[-\frac{n\ln 2\pi}{2} - n\ln\widehat{\sigma} - \frac{\sum (x_i - \widehat{\mu})^2}{2\widehat{\sigma}^2}\right].$$

Setting the derivatives of the log of the likelihood function with respect to $\widehat{\mu}$ and $\widehat{\sigma}$ equal to 0 gives:

(1.4)

$$\frac{\partial L(x_1, x_2, \ldots, x_n)}{\partial \widehat{\mu}} = \frac{\sum (x_i - \widehat{\mu})}{\widehat{\sigma}^{2}} = 0 \text{ and}$$

(1.5)

$$\frac{\partial L(x_1, x_2, \ldots, x_n)}{\partial \widehat{\sigma}} = -\frac{n}{\widehat{\sigma}} + \frac{\sum (x_i - \widehat{\mu})^2}{\widehat{\sigma}^{3}} = 0.$$

Solving these two equations simultaneously gives:

(1.6)

$$\widehat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x} \text{ and } \widehat{\sigma}^{2} = \frac{\sum (x_i - \widehat{\mu})^2}{n}.$$

Notice the fact that the estimator of the population mean is equal to the sample mean, a result that is the same as the one you found in your introductory statistics course. However, the *unbiased* estimator of the population variance used in that course is $s^2 = \frac{\sum (x_i - \widehat{\mu})^2}{n-1}$.

Thus, one of the common "problems" with using a ML estimator is that quite often they are *biased estimators* of a population parameter. On the other hand, under very general conditions ML estimators are *consistent*, are *asymptotically efficient*, and have an *asymptotically normal distribution* (these are desirable large sample size characteristics of potential estimators and are discussed in advanced statistics courses).[5]

Application of the ML method to regressions

The discussion above illustrates the basics of the ML method—you form the log of the likelihood function and then find the values of the parameter estimates that maximize this function. In most cases the maximization will not yield answers in closed form—that is, you cannot find a neat algebraic formula as we did for the population mean. However, you can use computer programs to search for the values of the parameter estimates that maximize this function. Thus, in most cases in advanced regression models you often will treat the ML method as a "black box" and not concern yourself with the estimation details. However, I illustrate one more example of the ML technique.

Example 1.14. The ML estimators for a simple regression.

Assume that we want to estimate the population parameters for the regression model $y_i = \beta x_i + \varepsilon_i$, where we assume that

1. $\varepsilon_i \sim N(0, \sigma^2)$,
2. $E(\varepsilon_i \varepsilon_j) = 0$ **for $i \neq j$,**
3. $y_i = Y_i - \overline{Y}$ **and** $x_i = X_i - \overline{X}$ **(this assumption allows us to ignore the estimation of the intercept term), and**
4. **x_i is a non-stochastic variable.**

The assumption of a normally distributed error term implies that $\varepsilon_i = y_i - \beta x_i \sim N(0, \sigma^2)$. **Thus, the pdf of the error term is**

$$f(\varepsilon_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i - \beta x_i)^2}{2\sigma^2}}$$

and, thus, the likelihood function[6] is:

(1.7)

$$\prod_{i=1}^{n} f(\varepsilon_i) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i - \beta x_i)^2}{2\sigma^2}} = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \prod_{i=1}^{n} e^{-\frac{(y_i - \beta x_i)^2}{2\sigma^2}}$$

and the log of the likelihood function is $L(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n) = -n\ln\sqrt{2\pi} - n\ln\widehat{\sigma} - \frac{\sum_{i=1}^{n}\left(y_i - \widehat{\beta}\, x_i\right)^2}{2\widehat{\sigma}^{\,2}}.$

We find the estimators $\widehat{\beta}$ and $\widehat{\sigma}$ in the same manner as we did for the sample mean and variance. Differentiating the log of the likelihood function and setting these first derivatives equal to 0 gives the following two first-order conditions:

(1.8)

$$\frac{\partial L(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)}{\partial\, \widehat{\beta}} = \frac{2\sum_{i=1}^{n}\left(y_i - \widehat{\beta}\, x_i\right)x_i}{2\widehat{\sigma}^{\,2}} = 0$$

and

(1.9)

$$\frac{\partial L(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)}{\partial\, \widehat{\sigma}} = -\frac{n}{\widehat{\sigma}} + \frac{\sum_{i=1}^{n}\left(y_i - \widehat{\beta}\, x_i\right)^2}{\widehat{\sigma}^{\,3}} = 0.$$

Thus, the ML estimators are:

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} \text{ and } \widehat{\sigma}^{\,2} = \frac{\sum_{i=1}^{n}\left(y_i - \widehat{\beta}\, x_i\right)^2}{n}.$$

Notice that in this simple case the ML estimator of *β* is the same as the OLS estimator of *β* . Also, notice that the ML estimator of σ^2 is biased—the (unbiased) OLS estimator of σ^2 is $s^2 = \frac{\sum_{i=1}^{n} \left(y_i - \widehat{\beta}\ x_i\right)^2}{n-2}$.

You can use the examples in this module as the basis of your understanding of the ML method. When you see that the ML method is used in a computer program, you can be fairly certain that the program uses one of the many optimizing subroutines to find the maximum of the log of the likelihood program. You can consult the help files with the computer program to see what underlying distribution is used to set up the log of the likelihood function. A concept related to the maximum likelihood estimation method worth exploring is the likelihood ratio test (see the module by Don Johnson entitled The Likelihood Ratio Test for an introduction to this key statistical test.)

Exercises

Exercise 1.2.1.

Consider the following functions. For each of them, (1) prove that the function is a pdf; (2) calculate the mean and variance of each distribution, and (3) find the maximum likelihood estimator of the parameter ϑ. Sketch a graph of each of the distributions for a representative value of ϑ.

1. $f(x;\vartheta) = (\vartheta + 1)x^{\vartheta}$ where $0 \le x \le 1$ and $\vartheta > 0$.
2. $f(x;\vartheta) = \vartheta e^{-\vartheta x}$ where $0 \le x < \infty$ and $\vartheta > 0$.

[1] The mean of the t-distribution is undefined for $t \leq 1$. The variance of the distribution is ∞ for $1 < r \leq 2$ and undefined for $r \leq 1$.

[2] Greene, William H. (1990). *Econometric Analysis* (New York: Macmillan Publishing Company): 103.

[3] Quite often, as in the exercises at the end of this module, it is easier to calculate the variance of a distribution using the alternative formula for the variance: $\sigma_x^2 = V(x) = E(x-\mu)^2 = E\left(x^2\right) - \mu^2,$ where $E\left(x^2\right) = \int x^2 f(x)dx.$

[4] The function $g(y)$ is monotonically increasing for y if $g'(y) > 0$. Because $\frac{d}{dx}\ln x = \frac{1}{x} > 0 \text{for} x > 0,$ the logarithm function is monotonically increasing for positive values of *x*.

[5] Intuitively, what these concepts mean is that as the sample size increases the estimator becomes more precise (the variance becomes smaller and an bias disappears) and the distribution of the estimator approaches the normal distribution. The formal definitions of these terms involve advanced statistical concepts that are reported here only in the interest of completeness. An estimator $\left(\hat{\theta}\right)$ of the parameter ϑ is consistent if and only if $\text{plim}\,\hat{\theta} = \theta.$ This estimator has an asymptotically normal distribution if $\hat{\theta} \overset{a}{\rightarrow} N\left(\theta, \{\mathbf{I}(\theta)\}^{-1}\right)$ An unbiased estimator is more efficient that another unbiased estimator if it has a smaller variance than the alternative estimator. An asymptotically efficient is an estimator whose mean square error tends to zero as the sample size increases. The mean square error (MSE) is defined to be $MSE\left(\hat{\theta}\right) = E\left[\left(\hat{\theta}-\theta\right)^2\right] = V\left(\hat{\theta}\right) + \left(Bias\left[\hat{\theta}\right]\right)^2.$ An estimator is asymptotically efficient if $\lim_{n \to \infty} MSE\left(\hat{\theta}\right) = 0.$ See any advanced statistics text or Statistical terminology for further information on these concepts.

[6] The symbol $\prod_{i=1}^{n} x_1$ is equivalent to the product $x_1 x_2 \cdots x_n$.

Chapter 2. Advanced topics in econometrics

2.1. Logit and Probit Regressions[*]

Logit and Probit models

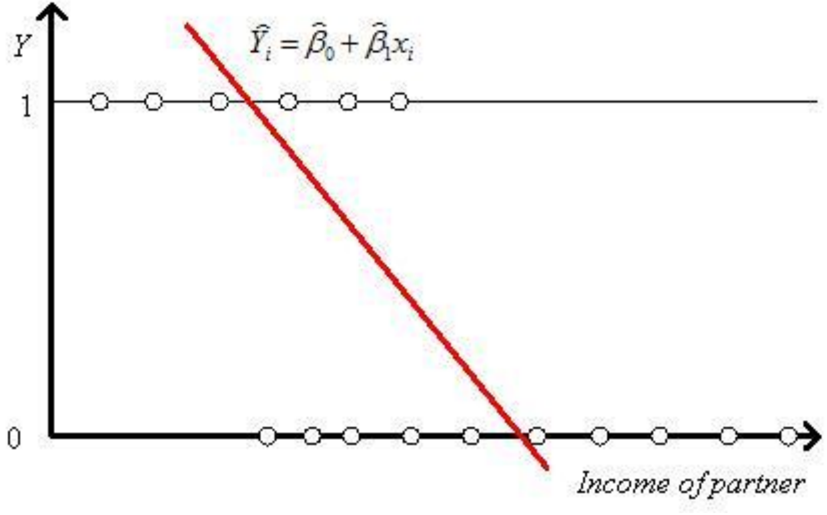
Introduction

Consider a model that "explains" whether a wife enters the work force. It is straight forward to think of potential explanatory variables—her potential wage rate, the income of her partner, the number of children under the age of 6 in the household, and the number of children in the household between the ages of 6 and 18 are candidates to be independent variables used to explain the wife's decision to enter the labor force. The dependent variable, *Y*, however, is a dummy variable because the wife chooses either to enter the labor force ($Y = 1$) or not to enter the labor force ($Y = 0$). An OLS model of the form:

$$(2.1)\quad Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

does not make sense. Figure 1 shows what the data of this model might look like when graphed against one of the explanatory variables. Figure 1 also includes the regression line that an OLS estimation of (1) will yield. It is easy to see one problem with this approach—the predicted values of *Y* that can be greater than 1 and less than 0. In addition, special properties must be attributed to the error term and it is the simple properties ascribed to the error term that make the OLS model so attractive.[7]

Figure 2.1. Linear regression line for a discrete dependent variable

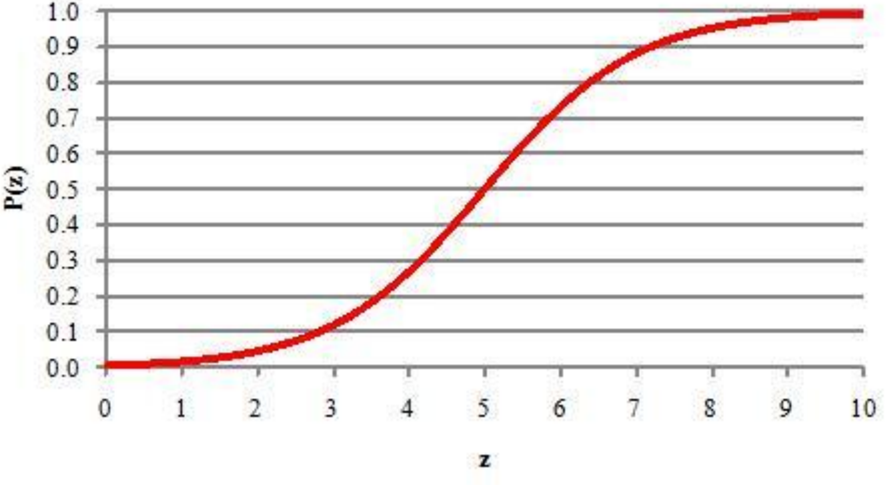


The linear regression line can be a poor representation of a discrete dependent variable.

The logit model

There does exist another approach to the modeling problem—assume that the dependent variable is *the probability that the wife is in the labor force*. For instance we might assume that we have a linear probability model of the form $\Pr(x_i) = \beta_0 + \beta_1 x_i + \varepsilon_i.$ This model can be estimated reasonably successfully if the observed frequencies are well away from their bounds of 0 and 1.[8] However, is more appealing to assume that the probability varies monotonically with *x* and remains within the bounds of [0,1], as shown in Figure 2. This S-shaped curve is known as the *sigmoid curve* and can be represented algebraically for some variable *z* by: $\Pr(z) = \frac{e^z}{1 + e^z}.$

Figure 2.2. The signoid function.



The signoid function forces the dependent variable to be between 0 and 1.

We can simplify our analysis by using a bit of algebra. First, the inverse probability is $1 - \Pr(z) = 1 - \frac{e^{z}}{1+e^{z}} = \frac{1}{1+e^{z}}$. **Thus,**

(2.2)

$$\frac{\Pr(z)}{1-\Pr(z)}=\frac{\frac{e^{z}}{1+e^{z}}}{\frac{1}{1+e^{z}}}=e^{z}.$$

Taking the natural logarithm of (2) gives $\ln\left(\frac{\Pr(z)}{1-\Pr(z)}\right)=z.$ Assuming that *z* is a linear function of *x* (and, more generally, of other variables) gives the *logit* model:

(2.3)

$$\ln\left(\frac{\Pr(x_i)}{1-\Pr(x_i)}\right)=\beta_0+\beta_1 x_i+\varepsilon_i.$$

We can estimate the parameters of this model using maximum likelihood methods. In the *probit* model the error term is assumed to be normally distributed with a mean of zero and a unit variance.[9] In the logit model the error term is assumed to have a *standardized logistic distribution*. This distribution has a mean of 0 and a variance of 1 and is very similar to a normal distribution with the same mean and variance.[10] While the choice of which model to use generally is personal, it should be noted that the ratio of the parameter of a logit model to the parameter of a probit model (using the same data set) usually varies between 1.6 and 2.0. We focus on the logit model in the balance of this discussion.

Interpretation of the logit model parameters

The interpretation of the economic meaning of the parameter values in a logit model is not very obvious.[11] One simple, but not often used, interpretation comes from taking the first-derivative of (3) with respect to *x*:

(2.4)

$$\ln(\text{odds}\,Y=1)=\beta_0+\beta_1 x+\varepsilon \Rightarrow \frac{\partial \ln(\text{odds}\,Y=1)}{\partial x}=\beta_1.$$

Thus, in the labor force participation model one interpretation is that β_1 is equal to the change in the natural logarithm of the odds that the wife is in the labor force due to a one unit change in the independent variable x. This interpretation is both awkward and not really economically informative.

Stata offers two command for estimating a logit regression—logit and logistic. The logit command returns the parameter estimates as shown in (3). The logistic command returns the odds ratio rather than the parameter estimates. The odds ratio is equal to $e^{\beta}{}_{1}$. Thus,

one can go from the odds ratio reported by the logistic command to the parameter estimates merely by taking the natural logarithm of the odds ratio. The interpretation of the odds ratio is straightforward. For example, assume that *y* = 1 means that the birth weight of an individual is less than 2,500 grams and *y* = 0 means that the birth weight is greater than 2,500 grams. A logit parameter estimate of -0.27 is equivalent to an odds ratio of 0.97 (i.e., $e^{-0.27} = 0.97$). An odds ratio of 0.97 means that odds of a baby being underweight are 0.97 times those of the odds of a baby being of normal weight. To see what is being said re-write (2.3) as:

$$\frac{\Pr(x)}{1-\Pr(x)} = e^{\beta_0 + \beta_1 x + \varepsilon}.$$

A one unit change in *x* implies that:

$$\frac{\Pr(x+1)}{1-\Pr(x+1)} = e^{\beta_0 + \beta_1 (x+1) + \varepsilon}$$

or

$$\frac{\Pr(x+1)}{1-\Pr(x+1)} = e^{\beta_0 + \beta_1 x + \varepsilon} e^{\beta_1}$$

or

$$\frac{\Pr(x+1)}{1-\Pr(x+1)} = e^{\beta_1}\left(\frac{\Pr(x)}{1-\Pr(x)}\right).$$

Thus, $e^{\widehat{\beta}_1}$ is equal to the percent change in the odds that *y* equals 1 (a baby is born underweight) due to a one unit change in *x*. The logistic command reports $e^{\widehat{\beta}_1}$ while the logit command reports $\widehat{\beta}_1$. Because of the ease of interpretation of the odds ratio, *Stata* argues that the logistic command is the proper one to use.

Elasticities

Another route to follow is to try to find something that can be interpreted as an elasticity. Elasticities are important enough topic in economics for us to discuss them here in some detail. The reason they are so attractive to economists is that they have no units and, thus, can be compared across different commodities. For instance, it is quite reasonable to compare the demand elasticity for apples with the demand elasticity for pearl

necklaces in spite of the fact that the units of measuring apples and necklaces are different. There are a few important ways that elasticities appear in regressions.

Linear regression elasticities

In a linear regression of the form (ignoring the subscripts and the error term)

$$Y = \beta_0 + \beta_1 x,$$

we would calculate the elasticity of *Y* with respect to *x* to be

$$\eta_{Yx} = \frac{x}{Y}\frac{\partial Y}{\partial x} = \beta_1 \frac{x}{Y}.$$

Clearly, researchers need to choose the levels of *Y* and *x* at which to report this elasticity; it is traditional to calculate the elasticity at the means. Thus, economists typically report

$$\eta_{Yx} = \beta_1 \frac{\bar{x}}{\bar{Y}}.$$

Constant elasticities

Consider the following demand equation:

$$(2.5)\ q = \alpha\, p^{-\beta} e^{\varepsilon},$$

where q is the quantity demanded, p is the price the good is sold at, $\alpha,\beta > 0$, and ε is an error term. The price elasticity of demand is given by

$$\eta_{qp} = \frac{p}{q}\frac{\partial q}{\partial p} = \frac{p}{\alpha p^{-\beta} e^{\varepsilon}}\left(-\beta\alpha p^{-\beta-1} e^{\varepsilon}\right) = -\beta.$$

In other words, this demand curve has a *constant price elasticity of demand* equal to $-\beta$. Moreover, we can convert the estimation of this equation into a linear regression by taking the natural logarithm of both sides of (5) to get $\ln q = \ln\alpha - \beta\ln p + \varepsilon$.

The logit equation and the quasi-elasticity

It is not appropriate to use the normal formula for an elasticity with (3) because the dependent variable is itself a number without units between 0 and 1. As an alternative it makes more sense to calculate the *quasi-elasticity*, which is defined as:

(2.6)

$$\eta(x) = x\frac{\partial \Pr(x)}{\partial x}.$$

Since

$$\ln\left(\frac{\Pr(x_i)}{1-\Pr(x_i)}\right) = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

we can calculate this elasticity as follows:

$$\frac{\partial\left(\ln\left(\frac{\Pr(x_i)}{1-\Pr(x_i)}\right)\right)}{\partial x} = \beta_1.$$

Focusing on the left-hand-side, we get:

$$\frac{1-\Pr(x_i)}{\Pr(x_i)}\frac{(1-\Pr(x_i))\frac{\partial \Pr(x_i)}{\partial x} + \Pr(x_i)\frac{\partial \Pr(x_i)}{\partial x}}{(1-\Pr(x_i))^2} = \beta_1$$

or

$$\frac{1}{\Pr(x_i)(1-\Pr(x_i))}\frac{\partial \Pr(x_i)}{\partial x}=\beta_1$$

or

(2.7)

$$\frac{\partial \Pr(x_i)}{\partial x}=\beta_1 \Pr(x_i)(1-\Pr(x_i)).$$

Thus, we see from (6) that the quasi-elasticity is given by:

(2.8)

$$\eta(x_i)=\beta_1 x_i \Pr(x_i)(1-\Pr(x_i)).$$

The quasi-elasticity measures the percentage point change in the probability due to a 1 percent increase of *x*. Notice that it is dependent on what value of *x* it is evaluated at. It is usual to evaluate (8) at the mean of *x*. Thus, the quasi-elasticity at the mean of *x* is:

$$\eta(\bar{x})=\beta_1 \bar{x}\Pr(\bar{x})(1-\Pr(\bar{x})),$$

where

$$\Pr(\overline{x}) = \frac{e^{\beta_0 + \beta_1 \overline{x}}}{1 + e^{\beta_0 + \beta_1 \overline{x}}}.$$

Hypothesis testing

The researcher using the logit model (and any regression estimated by ML) has three choices when constructing tests of hypotheses about the unknown parameter estimates—(1) the Wald test statistic, (2) the likelihood ratio test, or (3) the Lagrange Multiplier test. We consider them in turn.

The Wald test

The Wald test is the most commonly used test in econometric models. Indeed, it is the one that most statistics students learn in their introductory courses. Consider the following hypothesis test:

(2.9)

$$\begin{aligned} &\mathrm{H}_0 : \beta_1 = \beta \\ &\mathrm{H}_\mathrm{A} : \beta_1 \neq \beta. \end{aligned}$$

Quite often in these test researchers are interested in the case when *β* = 0 —i.e., in testing if the independent variable's estimated parameter is statistically different from zero. However, *β* can be any value. Moreover, this test can be used to test multiple restrictions on the slope parameters for multiple independent variables. In the case of a hypothesis test on a single parameter, the t-ratio is the appropriate test statistic. The t-statistic is given by

$$t = \frac{\widehat{\beta}_i - \beta}{\mathrm{s.e}\left(\widehat{\beta}_i\right)} \sim t_{n-k-1},$$

where *k* is the number of parameters in the mode that are estimated. The F-statistic is the appropriate test statistic when the null hypothesis has restrictions on multiple parameters. See Cameron and Trivedi (2005: 224-231) for more detail on this test. According to Hauck and Donner (1977) the Wald test may exhibit perverse behavior when the sample size is small. For this reason this test must be used with some care.

The likelihood ratio test

The likelihood ratio test is based on a comparison of the maximum log of likelihood function for the unrestricted model with the maximum log of likelihood function for the

model with the restrictions implied by the null hypothesis. Consider the null hypothesis given in (9). Let *L*(*β*) be the value of the likelihood function when *β* $_1$ be the value of the likelihood function when is restricted to being equal to *β* and $L\left(\widehat{\beta}_1\right)$ be the value of the likelihood function when there is no restriction on the value of *β*. Then the appropriate test statistic is

$$LR = -2\left[\ln L(\beta) - \ln L\left(\widehat{\beta}_1\right)\right].$$

The likelihood ratio statistic has the Chi-square distribution χ^2 (*r*), where *r* is the number of restrictions. Thus, using a likelihood ratio test involves two estimations—one with no restrictions on the model and one with the restrictions implied by null hypothesis. Since the likelihood ratio test does not appear to exhibit perverse behavior with small sample sizes, it is an attractive test. Thus, we will run through an example of how to execute the test using *Stata*. The example we are using is from the *Stata* manual, volume 2, pp. 353-355.

Example 2.1. Underweight births.

In this model we estimate a model that explains the likelihood that a child will be born with a weight under 2,500 grams (low). The eight explanatory variables used in the model are listed in Table 1. The model to be estimated is:

(2.10)

$$\ln\left(\frac{\Pr(Low)}{1-\Pr(Low)}\right)=\beta_1 Age+\beta_2 Lwt+\beta_3 RaceB+\beta_4 RaceO$$

$$+\beta_5 Smoke+\beta_6 Ptl+\beta_7 Ht+\beta_8 Ui+\varepsilon.$$

Also, we want to test the null hypothesis that the coefficients on Age, Lwt, Ptl, and Ht are all zero. The first step is to estimate the unrestricted regression using the command:

```
. logistic low age lwt raceb raceo smoke ptl ht ui
```

Variable name	Definition
Age	Age of mother
Lwt	Weight at last menstrual period

RaceB	Dummy variable =1 if mother is black; 0 otherwise
RaceO	Dummy variable = 1 if mother in neither white or black; 0 otherwise
Smoke	Dummy variable = 1 if mother smoked during pregnancy; 0 otherwise
Ptl	Number of times mother had premature labor
Ht	Dummy variable = 1 if mother has a history of hypertension; 0 otherwise
Ui	Dummy variable = 1 there is presence in mother of uterine irritability; 0 otherwise
Ftv	Number of visits to physician during first trimester

Table 2.1. Definition of the explanatory variables.

The results of this estimation are shown in column 2 of Table 2. Next we save the results of this regression with the command:

```
. estimates store full
```

where “full” is the name that we will refer to when we want to recall the estimation results from this regression. Now we estimate the logistic regression with the omitting the variables whose parameters are to be restricted to being equal to zero:

```
. logistic low raceb raceo smoke ui
```

The results of this estimation are reported in column 3 of Table 2. Finally we run the likelihood ratio test with the command:

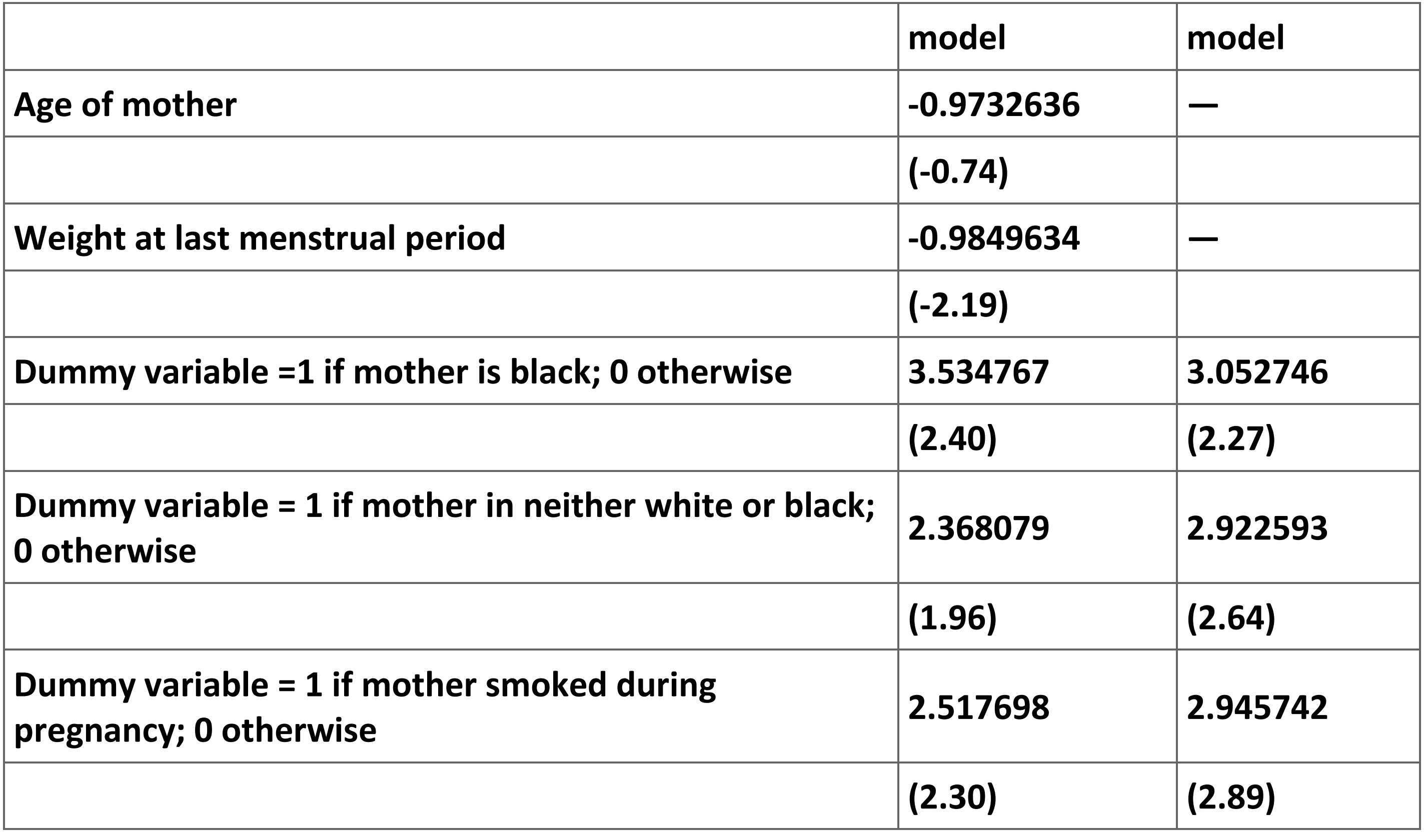
```
. lrtest full .
```

Notice that we refer to the first regression with the word “full” and to the second regression with the second period. The results of this command are as follows:

```
Likelihood-ratio test LR chi2(4) = 14.42

(Assumption: . nested in full) Prob > chi2 = 0.0061
```

The interpretation of these results is that the omitted variables are statistically significant at the 0.6 percent level.[12]

Explanatory variable	**Unrestricted**	**Restricted**



	model	model
Age of mother	-0.9732636	—
	(-0.74)	
Weight at last menstrual period	-0.9849634	—
	(-2.19)	
Dummy variable =1 if mother is black; 0 otherwise	3.534767	3.052746
	(2.40)	(2.27)
Dummy variable = 1 if mother in neither white or black; 0 otherwise	2.368079	2.922593
	(1.96)	(2.64)
Dummy variable = 1 if mother smoked during pregnancy; 0 otherwise	2.517698	2.945742
	(2.30)	(2.89)

Number of times mother had premature labor	1.719161	—
	(1.56)	
Dummy variable = 1 if mother has a history of hypertension; 0 otherwise	6.249602	—
	(2.64)	
Dummy variable = 1 if there is presence in mother of uterine irritability; 0 otherwise	2.1351	2.419131
	(1.65)	(2.04)
Log likelihood	-100.724	-107.93404
Number of observations	189	189
pseudo-R^2	0.1416	0.0801

Table 2.2. Estimation results for (2.10).

Note: Parameter estimates are odds ratios; z statistics are shown in parentheses.

The Lagrange multiplier test

The intuition behind the Lagrange multiplier (LM) test (or score test) is that the gradient of the log of the likelihood function is equal to zero at the maximum of the likelihood function.[13] If the null hypothesis in (2.9) is correct, then maximizing the log of the likelihood function for the restricted model is equivalent to maximizing the log of the likelihood function with the constraint specified by the null hypothesis. The LM test measures how close the Lagrangian multipliers of this constrained maximization problem are to zero—the closer they are to zero, the more likely that the null hypothesis can be rejected.

Economists generally do not make use of the LM test because the test is complicated to compute and the LR test is a reasonable alternative. Thus, as a practical matter the Wald test and the LR test are reasonable alternative test statistics to use to test most linear restrictions on the parameters. Moreover, since the calculations are relatively easy, it may make sense to calculate both test statistics to be sure they produce consistent conclusions. However, when the sample size is small, the LM test probably is preferred.

Goodness-of-fit measures

The standard measure of goodness-of-fit in the linear OLS regression model is R^2. No such measure exists for non-linear models like the logit model. Several potential alternatives have been developed in the literature and are known collectively as pseudo-R^2. Many of these measures are discussed in McFadden (1974), Amemiya (1981), and Maddala (1983). In case any reader really cares about the pseudo-- R^2, a practical approach is to report the value that the computer program reports.

One addition measure of goodness-of-fit is a measure called percentage correctly predicted. This variable is computed in one of several ways. One way is to use the observed values of the independent variable to forecast the probability the dependent variable equal one. Then, if the predicted probability is above some critical value, you assume that the predicted value of the dependent value is one. If it is below this value, you assume the predicted value of the dependent variable is zero. Then you construct a table that compares the predicted values of the dependent variable with the actual value of the dependent as shown in Table 3.

	Predicted

Actual	$\widehat{Y}=0$	$\widehat{Y}=1$
Y = 0	n_{00}	n_{01}
Y = 1	n_{10}	n_{11}

Table 2.3. Percent correctly predicted.

The percentage correctly predicted is equal to the sum of the diagonal elements, that is, $n_{00} + n_{11}$, over the sample size. The main problem with this measure is that the choice of the cutoff point is arbitrary. Traditionally, a cutoff point used has been 0.5. However, there is no reason why this cutoff is the appropriate one. Cramer (2003, 67) suggests that a more appropriate cutoff point is the sample frequency—that is, $\frac{n_{10}+n_{11}}{n_{00}+n_{01}+n_{10}+n_{11}}$. The bottom line is that the uncertainty about the proper choice of cutoff point is a major problem with using the percentage correctly predicted as a measure of goodness-of-fit.

Additional notes on binary variable models

One of the key choices in the various binary variable models involves the cumulative distribution function. The Table 4 shows the four commonly used binary outcome models along with the cumulative distribution functions:

Model	**Probability density function**	**Cumulative distribution function**	**Marginal effects,** $\frac{\partial p}{\partial x_j}$
Logit	**Logistic**	$\Lambda(\mathbf{x}'\boldsymbol{\beta}) = \frac{e^{\mathbf{x}'\boldsymbol{\beta}}}{1+e^{\mathbf{x}'\boldsymbol{\beta}}}$	$\Lambda(\mathbf{x}'\boldsymbol{\beta})[1-\Lambda(\mathbf{x}'\boldsymbol{\beta})]\beta_j$
Probit	**Normal***	$\Phi(\mathbf{x}'\boldsymbol{\beta}) = \int_{-\infty}^{\mathbf{x}'\boldsymbol{\beta}} \phi(\mathbf{x}'\boldsymbol{\beta})dx$	$\phi(\mathbf{x}'\boldsymbol{\beta})\beta_j$
Linear probability		$F(\mathbf{x}'\boldsymbol{\beta}) = \mathbf{x}'\boldsymbol{\beta}$	$\boldsymbol{\beta}_j$
Complementary log-log		$C(\mathbf{x}'\boldsymbol{\beta}) = 1 - e^{-e^{\mathbf{x}'\boldsymbol{\beta}}}$	$e^{-e^{\mathbf{x}'\boldsymbol{\beta}}} e^{\mathbf{x}'\boldsymbol{\beta}} \beta_j$

Table 2.4. Commonly used binary outcome models.

*** *φ*(·) is the probability density function (pdf) of the normal distribution.**

The logit, probit, and complementary log-log models are symmetric around zero and restrict 0 ≤ *p* ≤ 1. The linear does not impose either of these restrictions. Use of the complementary log-log regression sometimes is recommended when the sample is skewed such that there is a high proportion of ones and zeros. In general, economists use either the logit or probit models a majority of the time. Interestingly, there is no need to use robust estimation techniques for the logit and probit models if they are correctly specified. If use of the vce(robust) option produces substantially different parameter estimates than the estimates without the robust option, then it is likely that the models are misspecified. The linear model is inherently heteroskedastistic, implying that the vce(robust) option should be used.

The parameter estimates are comparable across the first three models in Table 4. In particular,

1. $\widehat{\beta}_{\text{Logit}} \approx 4\,\widehat{\beta}_{\text{Linear}}$

2. $\widehat{\beta}_{\text{Probit}} \approx 2.5\,\widehat{\beta}_{\text{Linear}}$, and

3. $\widehat{\beta}_{\text{Logit}} \approx 1.6\,\widehat{\beta}_{\text{Logit}}$.

Example 2.2. Supplementary health insurance coverage.

These data come from wave 5 (2002) of the Health and Retirement Study (HRS), a panel survey sponsored by the National Institute of Aging. The sample is restricted to Medicare beneficiaries; there are 3,206 observations. The elderly can obtain supplementary insurance coverage either by purchasing it themselves or by joining employer-sponsored plans. The data is in the file Example.xls. The variables included are listed in Table ?.

Variable	**Definition**
Binary variables	
(ins	**= 1 if individual has purchased supplementary insurance from any source**

retire	= 1 if individual is retired
hstatusg	= 1 if individual assess his/her health status either as good, very good, or excellent
married	= 1 if married
hisp	= 1 if hispanic
female	= 1 if female
white	= 1 if white
sretire	= 1 if a retired spouse is present in household
Continuous variables	
age	Age of individual in years
hhincome	Household income
educyear	Years of education
chronic	Total number of chronic conditions

adl	Number of limitations on daily activity (up to 5)

Table 2.5. Definition of the variables used in Example 2.

***Stata* commands**

Place the data into the editor and then create a list of the independent variables. Now create a new variable equal to the log of income:

```
.generate linc = ln(hhinc)
```

[notice that 9 observations are eliminated.]

Create list of "extra" variables in order to shorten future commands:

```
. global extralist linc female white chronic adl sretire
```

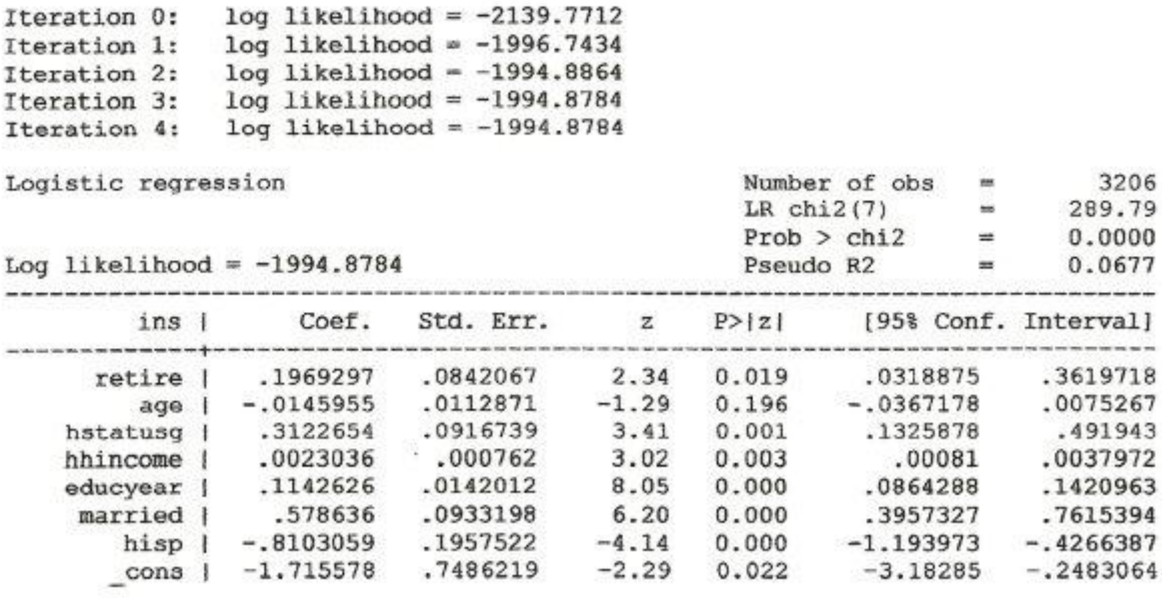
Summarize the variables in order to check for obvious typos (output is suppressed):

```
.summarize ins retire $xlist $extralist
```

Estimate logit regression (output is shown in Figure 3):

```
.logit ins retire $xlist
```

Figure 2.3. Stata regression output.



```
Iteration 0:   log likelihood = -2139.7712
Iteration 1:   log likelihood = -1996.7434
Iteration 2:   log likelihood = -1994.8864
Iteration 3:   log likelihood = -1994.8784
Iteration 4:   log likelihood = -1994.8784

Logistic regression                               Number of obs   =       3206
                                                  LR chi2(7)      =     289.79
                                                  Prob > chi2     =     0.0000
Log likelihood = -1994.8784                       Pseudo R2       =     0.0677

------------------------------------------------------------------------------
         ins |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      retire |   .1969297   .0842067     2.34   0.019     .0318875    .3619718
         age |  -.0145955   .0112871    -1.29   0.196    -.0367178    .0075267
    hstatusg |   .3122654   .0916739     3.41   0.001     .1325878     .491943
    hhincome |   .0023036    .000762     3.02   0.003       .00081    .0037972
    educyear |   .1142626   .0142012     8.05   0.000     .0864288    .1420963
     married |    .578636   .0933198     6.20   0.000     .3957327    .7615394
        hisp |  -.8103059   .1957522    -4.14   0.000    -1.193973   -.4266387
       _cons |  -1.715578   .7486219    -2.29   0.022     -3.18285   -.2483064
```

Estimate and save results from several models (the Stata command "quietly" suppresses the output from the command):

```
. estimates store blogit
.quietly probit ins retire $xlist
.estimates store bprobit
.quietly regress ins retire $xlist
.estimates store bols
.quietly logit ins retire $list, vce(robust)
. estimates store blogitr
.quietly probit ins retire $xlist, vce(robust)
.estimates store bprobitr
.quietly regress ins retire $xlist, vce(robust)
.estimates store bolsr
```

We can create table for comparing the models (output is suppressed):

```
.estimates table blogit blogitr bprobit bprobitr bols bolsr, t stats(N ll) b(%8.4f)
stfmt(%8.2f)
```

We now test for the presence of interaction variables:

```
.generate age2 = age*age

.generate agefem = age*fem

.generate agewhite = age*white

.generate agechronic = age*chronic

.global intlist age2 agefem agewhite agechronic

.quietly logit ins retire $xlist $intlist

.test $intlist

( 1) [ins]age2 = 0

( 2) [ins]agefem = 0

( 3) [ins]agewhite = 0

( 4) [ins]agechronic = 0

chi2( 4) = 7.45
```

```
Prob > chi2 = 0.1141
```

Likelihood ratio test

```
.quietly logit ins retire $xlist $intlist
.estimates store B
.quietly logit ins retire $xlist
.lrtest B
Likelihood-ratio test LR chi2(4) = 7.57
(Assumption: . nested in B) Prob > chi2 = 0.1088
```

Comparison with using the logistic command:

```
. logistic ins retire $xlist
```

The marginal effects at the mean will yield more useful results when the model is non-linear:

```
.quietly logit ins retire $xlist
.mfx
```

Let’s put the table comparing parameter estimates into a cleaned up table:

	Logit	Robust Logit	Probit	Robust Probit	OLS	Robust OLS
Individual retired	0.1969	0.1969	0.1184	0.1184	0.0409	0.0409
	(2.34)	(2.32)	(2.31)	(2.30)	(2.24)	(2.24)
Age of individual	-0.0146	-0.0146	-0.0089	-0.0089	-0.0029	-0.0029
	(-1.29)	(-1.29)	(-1.29)	(-1.32)	(-1.20)	(-1.25)
Health status	0.3123	0.3123	0.1977	0.1977	0.0656	0.0656
	(3.41)	(3.40)	(3.56)	(3.57)	(3.37)	(3.45)
Household income	0.0023	0.0023	0.0012	0.0012	0.0005	0.0005
	(3.02)	(2.01)	(3.19)	(2.21)	(3.58)	(2.63)
Years of education	0.1143	0.1143	0.0707	0.0707	0.0234	0.0234
	(8.05)	(7.96)	(8.34)	(8.33)	(8.15)	(8.63)
Individual married	0.5786	0.5786	0.3623	0.3623	0.1235	0.1235

	(6.20)	(6.15)	(6.47)	(6.16)	(6.38)	(6.62)
Individual is an Hispanic	-0.8103	-0.8103	-0.4731	-0.4731	-0.1210	-0.1210
	(-4.14)	(-4.18)	(-4.28)	(-4.36)	(-3.59)	(-4.49)
Intercept	-1.7156	-1.7156	-1.0693	-1.0693	0.1271	0.1271
	(-2.29)	(-2.36)	(-2.33)	(-2.40)	(0.79)	(0.83)
Sample size	3,206	3,206	3,206	3,206	3,206	3,206
Log of the likelihood function	-1994.88	-1994.88	-1993.62	-1993.62	-2104.75	-2104.75

Table 2.6. Comparison of Logit, Probit and OLS regressions with Insurance as the dependent variable.

(t-ratio or z-values in parentheses.)

As a last exercise use the following commands to generate a graph of the predicted values:

```
. quietly logit ins hhincome
. predict plogit, pr
. quietly probit ins hhincome
. predict pprobit, pr
. quietly regress ins hhincome
. predict pols, xb
. summarize ins plogit pprobit pols
. sort hhincome
.twoway (scatter ins hhincome, msize(vsmall)) (line plogit hhincome, lcolor(blue) lpattern
> (solid)) (line pprobit hhincome, lcolor(red) lpattern(tight_dot)) (line pols hhincome,
> lcolor(green) lpattern(longdash_shortdash)), ytitle(Predicted Probability)
xtitle(Household income)
```

Note: save file as a .tif file if you want to insert the graph directly into a word file.

Exercises

Exercise 2.1.1.

The determinants of physician advice. Physicians are expected to give lifestyle advice as a part of their normal interaction with their patients. Sometimes doctors choose not to comment on a patient’s lifestyle because they do not have time for personal comments, they feel the advice will be unwelcome, they feel that lifestyle choices are not any business of the physician, they find the discussion of lifestyle issues to be embarrassing, or they are not aware of the patient’s actual lifestyle choices. In this project we are interested in understanding when physicians choose to give advice concerning the consumption of alcohol.

The MS Excel file ktdata contains the responses to the 1990 National Health Interview Survey core questionnaire and special supplements from 2,467 males who were current drinkers in 1990. Individuals who are lifetime abstainers or who are former drinkers who have not consumed any alcohol in the past year are excluded from the sample. Table 7 contains the names and definitions of the variables collected in the survey.

Variable	Definition
Drinks	Total number of drinks taken in the past two weeks
Advice	Did your physician give you advice about alcohol consumption? Yes = 1, No = 0
Income	Monthly income in \$1,000 (there are 5 missing values denoted by a “.”)
Age30	Dummy variable equal to 1 if 30 < Age ≤ 40and 0 otherwise
Age40	Dummy variable equal to 1 if 40 < Age ≤ 50 and 0 otherwise
Age50	Dummy variable equal to 1 if 50 < Age ≤ 60 and 0 otherwise
Age60	Dummy variable equal to 1 if 60 < Age ≤ 70 and 0 otherwise
AgeGT70	Dummy variable equal to 1 if individual’s age is greater than 70 and 0 otherwise
Educ	Number of years of schooling (0 to 18)
Black	Dummy variable equal to 1 if the individual is a black and 0 otherwise
Other	Dummy variable equal to 1 if the individual is non-white and non-black and 0

	otherwise
Married	Dummy variable equal to 1 if the individual is married and 0 otherwise
Widow	Dummy variable equal to 1 if the individual is a widow and 0 otherwise
DivSep	Dummy variable equal to 1 if the individual is either divorce or separated and 0 otherwise
Employed	Dummy variable equal to 1 if the individual is currently employed and 0 otherwise
Unemploy	Dummy variable equal to 1 if the individual is currently unemployed and 0 otherwise
NE	Dummy variable equal to 1 if the individual lives in the Northeast US and 0 otherwise
MW	Dummy variable equal to 1 if the individual lives in the Midwest US and 0 otherwise
South	Dummy variable equal to 1 if the individual lives in the South and 0 otherwise
Medicare	Dummy variable equal to 1 if the individual receives Medicare and 0

	otherwise
Medicaid	**Dummy variable equal to 1 if the individual receives Medicaid and 0 otherwise**
Champus	**Dummy variable equal to 1 if the individual has military insurance and 0 otherwise**
HlthIns	**Dummy variable equal to 1 if the individual has health insurance and 0 otherwise**
RegMed	**Dummy variable equal to 1 if the individual has a regular source of medical care and 0 otherwise**
DRI	**Dummy variable equal to 1 if the individual sees the same doctor and 0 otherwise**
MajorLim	**Dummy variable equal to 1 if the individual has limits on major daily activity and 0 otherwise**
SomeLim	**Dummy variable equal to 1 if the individual has limits on some daily activity and 0 otherwise**
Diabetes	**Dummy variable equal to 1 if the individual has diabetes and 0 otherwise**

Heart	**Dummy variable equal to 1 if the individual has a heart condition and 0 otherwise**
Stroke	**Dummy variable equal to 1 if the individual has had a stroke and 0 otherwise**

Table 2.7. Definition of the variables in the *Excel* worksheet ktdata.

You are to estimate a logit regression of the form: $\ln\left(\frac{p}{1-p}\right)=\beta_0+\sum_i \beta_i x_i+\varepsilon,$ **where *p* is the probability that a patient received advice about his level of consumption of alcohol and x_i are the explanatory variables.**

Provide the following information:

1. **Make a table of the means of all of the variables.**

2. **Offer an economic justification for the inclusion of each explanatory variable you use in your regression (including a prediction of its expected sign).**

3. Make a table reporting the results of the estimation of (1) an OLS linear estimation, (2) a probit estimation, and (3) a logit estimation. Also include a column with the ratio of each of the logit parameters to the probit parameter. Do not use the abbreviated name of the explanatory variables in the table.

4. Present a table of results of a logit model with all of the variables and with whatever other models you feel are suggested by your empirical results. Discuss the results of the estimation and what the estimation tells you about how physicians decide whether to give advice on alcohol consumption to their male patients.

Exercise 2.1.2.

The Supply of Married Women in the Workforce. We are interested in understanding the decision of married women to enter the labor force. We have available two data sets, one using data from the United States and the other using data from Portugal. You are to estimate a logit regression for married women for each of the two data sets.

Variable	Definition

Working	**dummy variable = 1 if a married woman works during the year**
Fulltime	**dummy variable = 1 if a married woman works more than 1000 hours in a year**
Other	**the other household income in $100 (not in $1000)**
Age	**age of the wife**
Educ	**education years of the wife**
C0005	**number of children for ages 0 to 5**
C0613	**number of children for ages 6 to 13**
C1417	**number of children for ages 14 to 17**
NW	**1 if non-white, and 0 otherwise.**
HOwn	**1 if the home is owned by the household, and 0 otherwise**
HMort	**1 if the home is on mortgage, and 0 otherwise**
Prof	**1 if the husband is manager or professional, and 0 otherwise**
Sales	**1 if the husband is sales worker or clerical or craftsman, and 0 otherwise**

Farm	1 if the husband is farm-related worker
Unem	local unemployment rate in %

Table 2.8. US Data on Married Women.

Data Set 1: The data for this project are in the MS Excel file FLABOR. These data are observations on married females drawn from the 1987 wave of Michigan Panel Study of Income Dynamics (PSID). The data set has observations for 3,382 individuals.

Data Set 2: These data are from Portugal. The data set is a sample from Portuguese Employment Survey, from the interview year 1991, and has been provided by the Portuguese National Institute of Statistics (INE). The data are in the Excel file Martins. This file is organized into seven columns, corresponding to seven variables, with 2,339 observations.

Variable	Definition
Works	Dummy variable equal to 1 if the woman works, 0 otherwise

Child18	**The number of children younger than 18 living in the family**
Child03	**The number of children younger than 3 living in the family**
Age	**The woman’s age**
LogWomanWageRate	**The log of women's hourly wage rate (measured in escudos)**
Education	**The women's educational level, measured in years of schooling**
LogHusbandMonthlyWages	**The log of the husband's monthly wage (measured in escudos)**

Table 2.9. The Portuguese data set.

Answer the following questions:

1. **What factors other than wage levels determine the number of hours that a wife will spend in the work force? Remember to use economic theory in answering this question.**

2. **Clearly, one of the major factors in determining if a wife will enter the labor force is the wage level she can earn. The US data set does not include the wife's wage level. Is there any other variable in the data set that economic theory suggests will be a good proxy for wage levels?**

3. **The variable Age is a proxy for the work (or life) experience of a woman. We would expect that its effect on the probability that a woman will enter the labor force will be non-linear—that is, its marginal impact will be positive and decreasing. This reasoning suggests that you should use Age and Age^2 as explanatory variables. Can the same reasoning be used with the variable Education? What are your expectations about the signs of the parameters of these two explanatory variables? The same reasoning can be used about the number of years of education.**

4. **Estimate and report in a table the following two logit regressions: (1) US women enter the labor force at all and (2) US women enter the labor force for at least 1,000 hours if they enter the labor force,. In each of these cases, compare your results to a linear model.**

5. **The Portuguese data set has a different problem. We have reported the wage rate of women who are working, but no wage level for women who are not working. We will get around this problem by first using the data for women who actually work to**

estimate the relationship between wage rates and the age and education of the women. We will then use this relationship to predict the wage rate for both women who do work and women who do not work. We will then use this predicted wage rate data series as an independent variable in a logit model explaining the probability that a married woman will enter the labor force. When completing the logit regression be sure that you separate all of the children in a family into those 3 and under and those between 4 and 18. Also, include the years of education in this regression to see if a Portuguese married woman's taste for participation in the labor force increases or decreases with the level of her education.

6. Is it reasonable to compare your results for the two countries?

References

Amemiya, T. (1981). Quantitative Response Models: A Survey. *Journal of Economic Literature* 19: 1483-1536.

Cramer, J. S. (2003). *Logit Models from Economics and Other Fields* (Cambridge: Cambridge University Press).

Cameron, A. Colin and Pravin K. Trivedi (2005). *Microeconometrics: Methods and Applications* (Cambridge: Cambridge University Press).

Ladd, G. W. (1966). Linear Probability Functions and Discriminant Functions. *Econometrica* 34: 873-888.

Maddala, G. S. (1983). *Limited-Dependent and Qualitative Variables in Economics* (Cambridge: Cambridge University Press).

McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior. In P. Zarembka (ed.) *Frontiers in Econometrics* (New York: Academic Press): 105-142.

Wald, A. (1943). Test of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large. *Transactions of the American Mathematical Society* 54: 426-482.

2.2. Analysis of time series[*]

Analysis of Time-Series

Introduction

This module offers a brief introduction of some of the issues that arise in the analysis of time-series. Most of the topics covered are those that we attacked first by statisticians and economists. As such they do not demand the more sophisticated tools used by the more modern approaches to time-series. In spite of these shortcomings, they should give you some understanding of the issues that arise with the use of times-series in econometric analyses. One final note of explanation is necessary. These notes are designed to give you a brief introduction to how *Stata* handles time-series data. These notes are not a substitute for reading the *Stata* manual, completing a forecasting course, or reading standard texts on the rather complicated field.

Time-series analysis in *Stata*

Throughout this module we work with US macroeconomic data included in the MS Excel file Macro data.xls. The variables are real level of investments (RINV), real gross national product (RGNP), and real interest rate (RINTRATE). The real interest rate is approximated by the difference between the nominal interest rate and the rate of change of the price index from the previous year. The data are for the years 1963 to 1982. You can replicate the analysis done here by copying this data set into a *Stata* file.

The first step after entering the data set into *Stata*, is to declare that the data set is a time-series. The command to do this is:

```
. tsset year
```

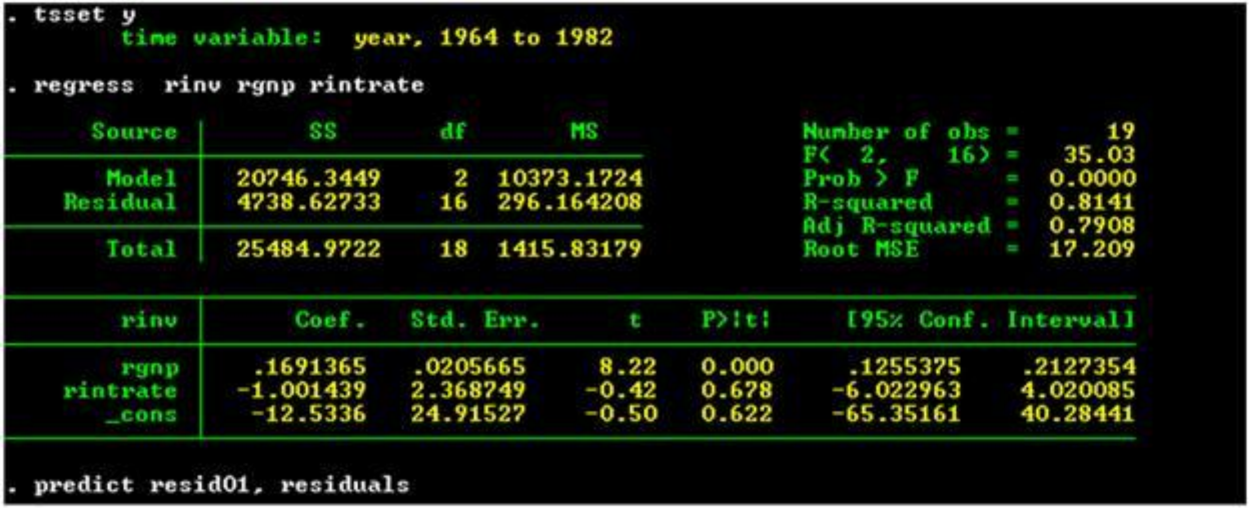
The data set can be broken into any number of time periods including daily, weekly, monthly, quarterly, halfyearly, yearly and generic.[14]

Assume that we want to estimate the following regression:

$$(2.11)\ RINV_t = \beta_0 + \beta_1 RGNP_t + \beta_2 RINTRATE_t + \varepsilon_t$$

using the data set in the appendix. Figure 1 shows this regression command and the resultant output.

Figure 2.4.

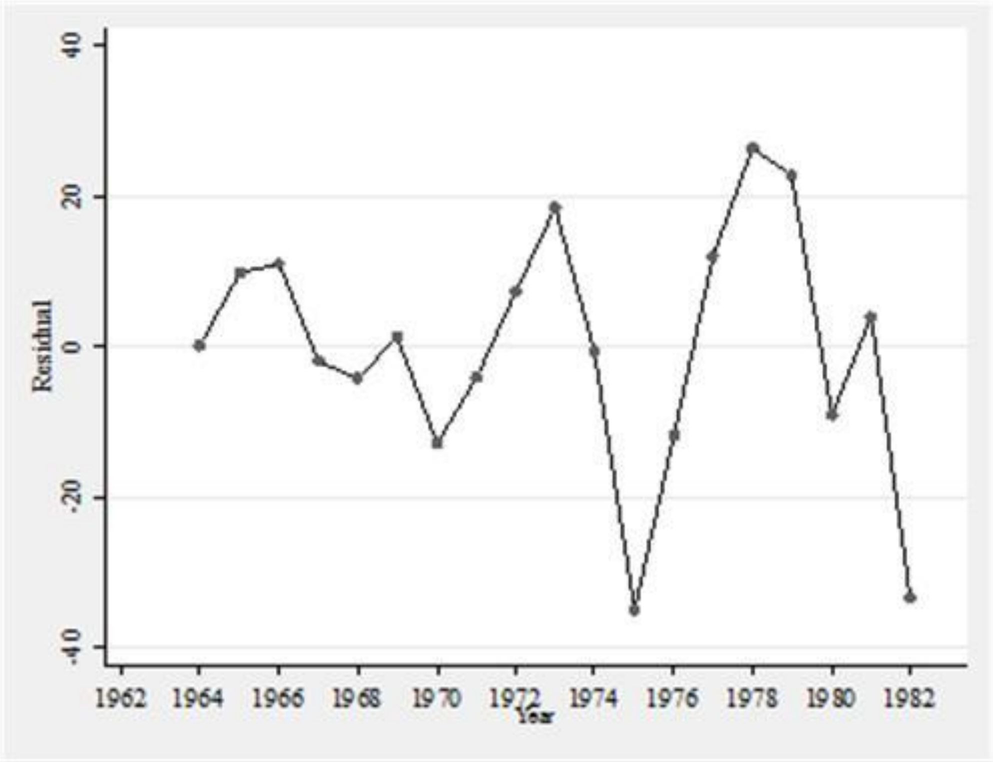


```
. tsset y
        time variable:  year, 1964 to 1982

. regress  rinv rgnp rintrate

      Source |       SS       df       MS              Number of obs =      19
-------------+------------------------------           F(  2,    16) =   35.03
       Model |  20746.3449     2  10373.1724           Prob > F      =  0.0000
    Residual |  4738.62733    16  296.164208           R-squared     =  0.8141
-------------+------------------------------           Adj R-squared =  0.7908
       Total |  25484.9722    18  1415.83179           Root MSE      =  17.209

------------------------------------------------------------------------------
        rinv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        rgnp |   .1691365   .0205665     8.22   0.000     .1255375    .2127354
    rintrate |  -1.001439   2.368749    -0.42   0.678    -6.022963    4.020085
       _cons |   -12.5336   24.91527    -0.50   0.622    -65.35161    40.28441
------------------------------------------------------------------------------

. predict resid01, residuals
```

OLS estimates for Equation (1).

On the surface the estimates seem “reasonable” because the signs on the two explanatory variables are what theory predicts they should be and the parameter for real GNP is statistically different from zero. However, an examination of the residuals shown in Figure 2 suggest that the error terms might exhibit autocorrelation.

Figure 2.5.



The residuals appear to be autocorrelated.

There are several issues that arise here. First, what sort of models can we use to account for autocorrelation? Second, what sorts of tests exist for detecting the existence of autocorrelation? We begin with the first of these questions by introducing the concept of first-order autocorrelation. Consider the following model:

$$(2.12)\ y_t = \beta_0 + \beta_1 x_t + \varepsilon_t .$$

We say that this model exhibits first-order autocorrelation if the error terms can be written as:

$$(2.13)\ \varepsilon_t = \rho\,\varepsilon_{t-1} + \mu_t ,$$

where $\mu_t \sim N\left(0, \sigma^2\right)$. Equation (3) implies that the error terms in (2) are correlated with each other. It is rather easy to show that, while the estimates of the unknown parameters are unbiased, the estimates of the standard errors are biased—downward if $1 > \rho > 0$ and upward if $-1 < \rho < 0$. This conclusion holds as long as the source of the autocorrelation is due to (3). If, on the other hand, the source of autocorrelation among the error terms in (2) is due to omitted explanatory variables (whose effects are absorbed in the error term), we have a potentially more serious problem. In particular, if the omitted

explanatory variables are correlated with the included explanatory variables (as is often true in time-series), then the estimates of the unknown slope parameters are also biased.

For the moment we will assume that Equations (2) and (3) are true representations of the world. What then can we do to estimate (2)? What we need to do is find a way to transform (2) so that the error term of whatever regression we estimate does not exhibit autocorrelation. In time period *t* − 1 we have:

(2.14) $y_{t-1} = \beta_0 + \beta_1 x_{t-1} + \varepsilon_{t-1}$.

Multiply (4) by *ρ* to get:

(2.15) $\rho y_{t-1} = \rho \beta_0 + \rho \beta_1 x_{t-1} + \rho \varepsilon_{t-1}$.

Now subtracting (5) from (4) gives:

$$y_t - \rho y_{t-1} = \beta_0 + \beta_1 x_t + \varepsilon_t - (\rho\beta_0 + \rho\beta_1 x_{t-1} + \rho\varepsilon_{t-1})$$

or, equivalently,

$$(y_t - \rho y_{t-1}) = \beta_0(1 - \rho) + \beta_1 (x_t - \rho x_{t-1}) + (\varepsilon_t - \rho\varepsilon_{t-1}).$$

Let

$y_t{}^* = y_{t-1} - \rho\, y_{t-1}$,

$\beta_0{}^* = \beta_0\,(1 - \rho)$,

and

$x_t{}^* = x_{t-1} - \rho\, x_{t-1}$.

Remember that (3) implies that $\mu_t = \varepsilon_t - \rho\, \varepsilon_{t-1}$. Thus, we have:

$$(2.16)\quad y_t{}^* = \beta_0{}^* + \beta_1 x_t{}^* + \mu_t,$$

where $\mu_t \sim N(0, \sigma^2)$. Thus, we have a regression for which the OLS estimates will be BLUE (Best Linear Unbiased Estimator) if we only knew the true value of ρ.

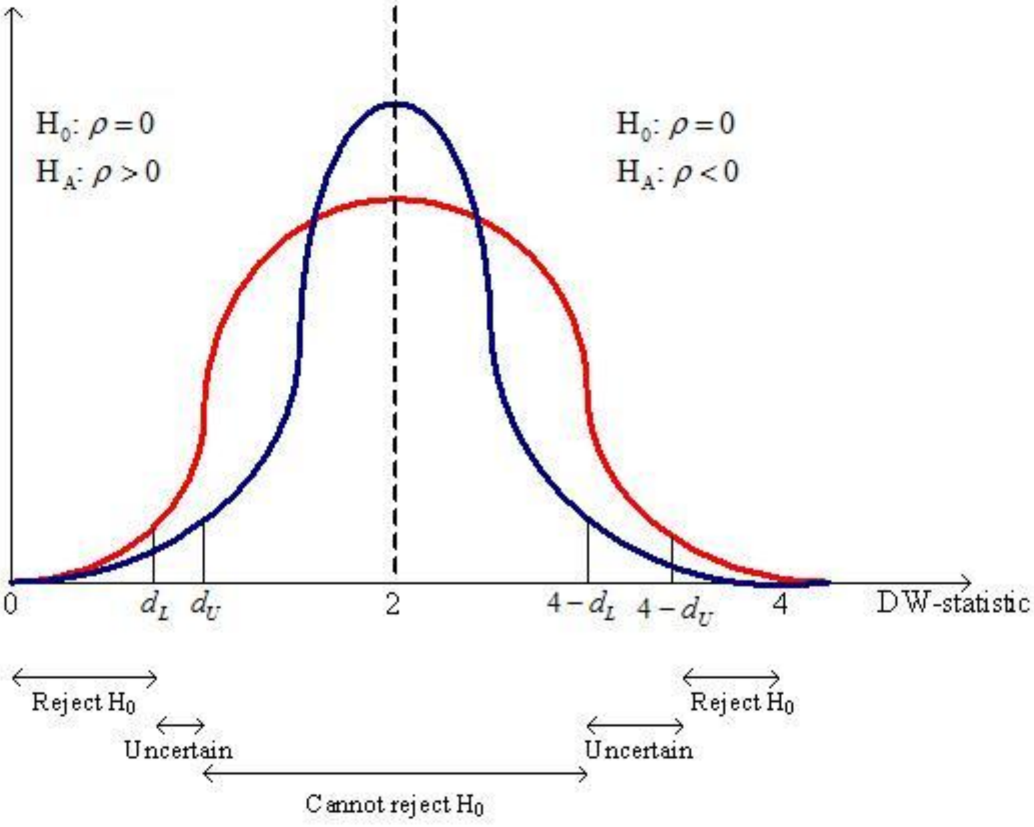
Cochran and Orcutt [1949] use this algebra to suggest one way to estimate (6). The estimation entails several steps. First, you use OLS to estimate (2). Second, you estimate (3) using the residuals from the first stage to approximate ε_t. This regression gives an estimate of ρ. In the third step, you use the estimate of ρ to construct estimates of $y_t{}^*$ and $x_t{}^*$. In the fourth step, you use the estimates of $y_t{}^*$ and $x_t{}^*$ to estimate (6); this will yield new estimates of β_0 and β_1 . You then repeat step (2) using these new estimates of β_0 and β_1 to calculate the residuals and then repeat with steps (3) and (4). You continue the process until the estimate of ρ does not change anymore (i.e., until the change in the estimate of ρ is less than some value chosen by the researcher). There are a multitude of alternative ways of estimating ρ. [See Greene (1990): Chapter 15 for a full discussion of these methods.] Once you have an estimator for ρ, there exist two major ways of completing the estimation—the Cochran-Orcutt procedure described above and the Prais-Winsten (1954) estimator. The latter estimation procedure does not involve dropping the first observation (as does the Cochran-Orcutt) estimator. In large samples these two estimation techniques are likely to be very similar. In small samples the two techniques may produce estimates that are substantially different.

We now turn to the issue of detecting the existence of autocorrelation. In what follows we focus mainly on the detection of first-order autocorrelation as shown in Equation (3). We can use the Durbin-Watson test to see if our suspicions are correct. The Durbin-Watson statistic tests the hypothesis:

(2.17)

$$\begin{aligned} &\mathrm{H}_0 : \rho = 0 \\ &\mathrm{H}_1 : \rho \neq 0 \end{aligned}$$

Figure 2.6.



H_0: $\rho = 0$
H_A: $\rho > 0$
H_0: $\rho = 0$
H_A: $\rho < 0$
0
d_L
d_U
2
$4 - d_L$
$4 - d_U$
4
DW-statistic
Reject H_0
Uncertain
Cannot reject H_0
Uncertain
Reject H_0

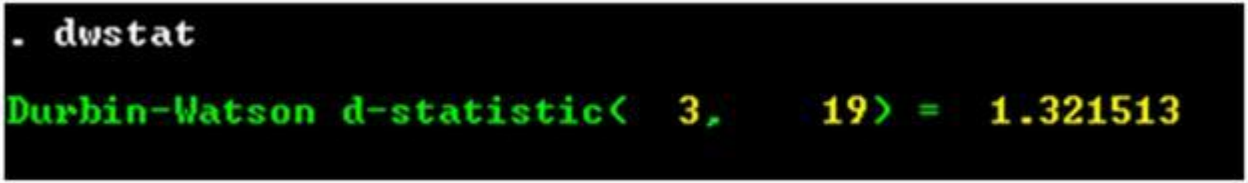
Limiting distributions for the Durbin-Watson statistic.

The details of the test statistic can be found in any econometrics textbook and need not detain us here. What you need to know about the DW-statistic are (1) it has a mean value of 2; (2) because its distribution lies between two limiting distributions, we need to look at two critical values. For this reason there are two critical values—one for each of the limiting distributions. Figure 3 illustrates the probability distribution function (pdf) for the Durbin-Watson statistic. The true pdf lies somewhere between the blue pdf and the red pdf. What is shown in the figure is the point below which, say, 5 percent of the distribution lies for each distribution. The true critical point lies somewhere between d_L and d_UThese values are relevant to testing the null hypothesis of no autocorrelation against the alternative hypothesis of positive autocorrelation (i. e., $\rho > 0$).

If $d < d_L$, we can reject the null hypothesis of no autocorrelation; if $d_U < d < 4 - d_U$, we cannot reject the null hypothesis; and if $d_L < d < d_U$, the results of the test are uncertain. Moreover, since the distributions are symmetric around 2 and between 0 and 4, the critical values for the alternative hypothesis of negative autocorrelation (i. e., $\rho > 0$) are 4 minus either the upper or lower critical values, as shown in Figure 3. Critical values for

the Durbin-Watson statistic can be found in the appendices of most econometric textbooks.

Figure 2.7.



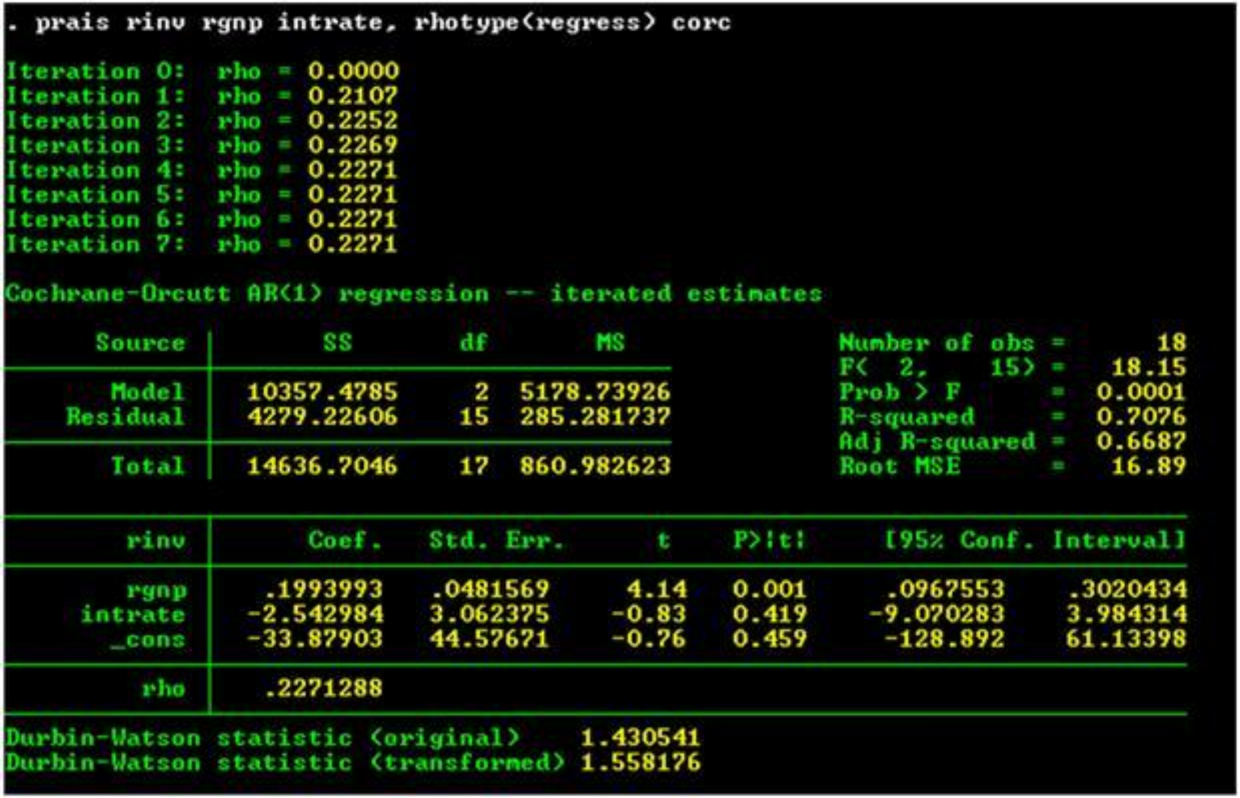
```
. dwstat

Durbin-Watson d-statistic(  3,    19) =  1.321513
```

Command for calculating the Durbin-Watson statistic in Stata.

The command for the test and the resultant DW-statistics for the estimate of Equation (2) are shown in Figure 4. The 5 percent level critical values for the Durbin-Watson statistic for a sample size of 19 with two parameters (less the intercept) estimated are 1.074 and 1.536—if the observed value of the DW-statistic is between 1.536 and 2.464, we can accept the null hypothesis that the residuals do not exhibit autocorrelation. Our value of 1.32 falls in the uncertain region where we are not sure if we can or cannot reject the null hypothesis.

At this point we can try the Cochran-Orcutt estimate. Figure 5 reports the results of using the Cochran-Orcutt estimation procedure. Notice that it took 7 iterations for the estimate of *ρ* to converge. If we use the Prais-Winsten estimation technique, we get the results shown in Figure 6. It is reassuring to see that the two estimation techniques do not yield estimates of the standard errors that are substantially different from each other.

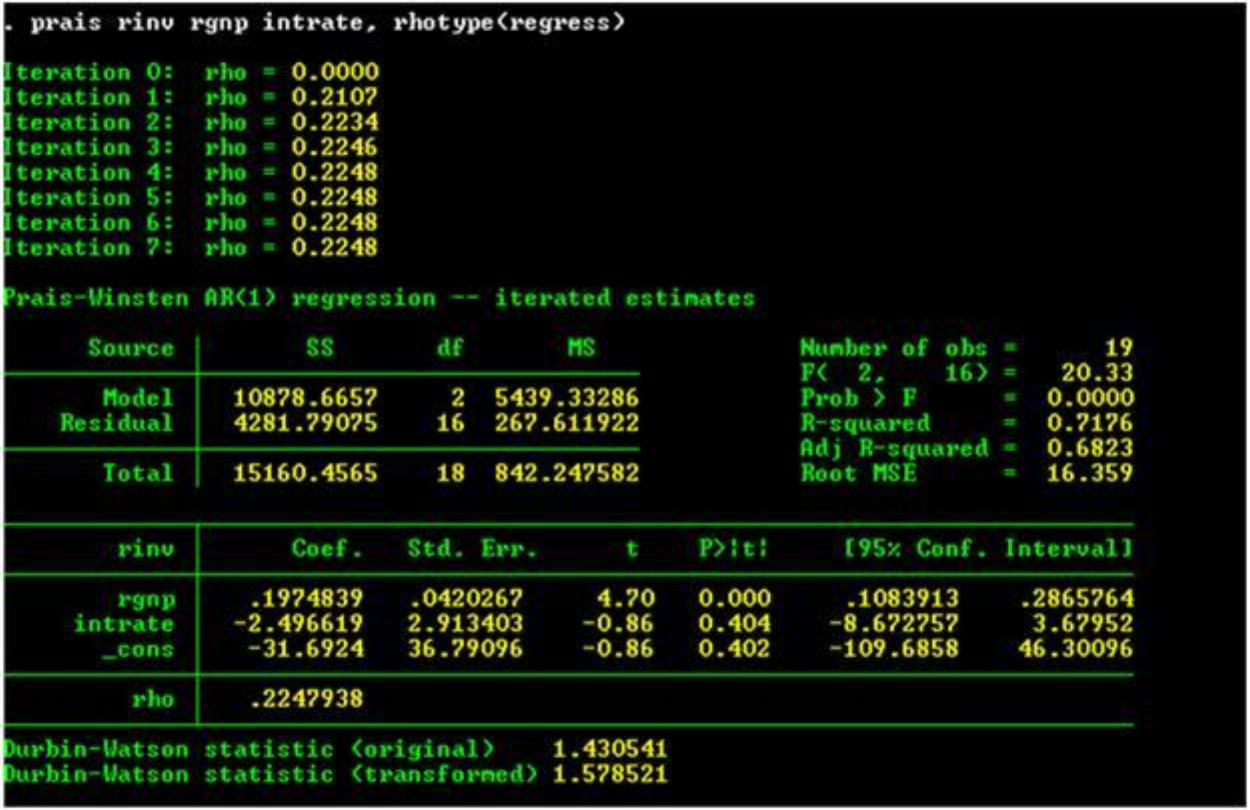
Figure 2.8.



```
. prais rinv rgnp intrate, rhotype(regress) corc

Iteration 0:  rho = 0.0000
Iteration 1:  rho = 0.2107
Iteration 2:  rho = 0.2252
Iteration 3:  rho = 0.2269
Iteration 4:  rho = 0.2271
Iteration 5:  rho = 0.2271
Iteration 6:  rho = 0.2271
Iteration 7:  rho = 0.2271

Cochrane-Orcutt AR(1) regression -- iterated estimates

      Source |       SS       df       MS              Number of obs =      18
-------------+------------------------------           F(  2,    15) =   18.15
       Model |  10357.4785     2  5178.73926           Prob > F      =  0.0001
    Residual |  4279.22606    15  285.281737           R-squared     =  0.7076
-------------+------------------------------           Adj R-squared =  0.6687
       Total |  14636.7046    17  860.982623           Root MSE      =   16.89

------------------------------------------------------------------------------
        rinv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        rgnp |   .1993993   .0481569     4.14   0.001     .0967553    .3020434
     intrate |  -2.542984   3.062375    -0.83   0.419    -9.070283    3.984314
       _cons |  -33.87903   44.57671    -0.76   0.459     -128.892    61.13398
-------------+----------------------------------------------------------------
         rho |   .2271288
------------------------------------------------------------------------------
Durbin-Watson statistic (original)    1.430541
Durbin-Watson statistic (transformed) 1.558176
```

Estimation of Equation (1) using the Cochran-Orcutt method.

Figure 2.9.



```
. prais rinv rgnp intrate, rhotype(regress)

Iteration 0:  rho = 0.0000
Iteration 1:  rho = 0.2107
Iteration 2:  rho = 0.2234
Iteration 3:  rho = 0.2246
Iteration 4:  rho = 0.2248
Iteration 5:  rho = 0.2248
Iteration 6:  rho = 0.2248
Iteration 7:  rho = 0.2248

Prais-Winsten AR(1) regression -- iterated estimates

      Source |       SS       df       MS              Number of obs =      19
-------------+------------------------------           F(  2,    16) =   20.33
       Model |  10878.6657     2  5439.33286           Prob > F      =  0.0000
    Residual |  4281.79075    16  267.611922           R-squared     =  0.7176
-------------+------------------------------           Adj R-squared =  0.6823
       Total |  15160.4565    18  842.247582           Root MSE      =  16.359

------------------------------------------------------------------------------
        rinv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        rgnp |   .1974839   .0420267     4.70   0.000     .1083913    .2865764
     intrate |  -2.496619   2.913403    -0.86   0.404    -8.672757     3.67952
       _cons |   -31.6924   36.79096    -0.86   0.402    -109.6858    46.30096
-------------+----------------------------------------------------------------
         rho |   .2247938
------------------------------------------------------------------------------
Durbin-Watson statistic (original)    1.430541
Durbin-Watson statistic (transformed) 1.578521
```

Estimation of Equation (1) using the Prais-Winsten estimator.

Using either the Cochran-Orcutt or the Prais-Winstn estimator is dependent on the assumption that the error terms exhibit first-order autocorrelation. Unfortunately, there is no particular reason (from a theoretical viewpoint) to believe in this assumption. Why, for instance, couldn't the error terms of Equation (2) exhibit second-order autocorrelation of the form:

$$(2.18)\ \varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + \mu_t ?$$

There is a more troubling possible explanation for the low Durbin-Watson statistic: the model may be misspecified. In particular, there may be important explanatory variables omitted from the regression. These omitted explanatory variables may exhibit autocorrelation and, thus, may be the source of autocorrelation in the error term. If the omitted explanatory variables are correlated with the included explanatory variables, then the parameter estimates are biased. The large difference in the estimate of parameter for real interest rates for the OLS regression and the Cochran-Orcutt estimate is suggestive of model misspecification.[15]

More modern time-series models

ARMA models

The model we described above is assumed to have first-order autoregressive error disturbances. Such a process is referred to as AR(1). The error structure in (8) is AR(2). If we apply this concept to a data series, we would call the following an AR(*p*) process:

(2.19)

$$y_t = \alpha_0 + \sum_{i=1}^{p} \beta_i y_{t-i}.$$

Another approach available to us is to think of a data as a weighted average of some error terms that are assumed to have a mean of zero, have a fixed variance, and be uncorrelated over time[16]:

(2.20)

$$y_t = \sum_{i=0}^{q} \beta_i \varepsilon_{t-i}.$$

A data series exhibiting this pattern is called a moving average process or MA(q). The error tern is known in the literature as white noise. A data series that has both autoregressive and moving average characteristics is call an autoregressive moving average (ARMA) series; an ARMA(p, q) is:

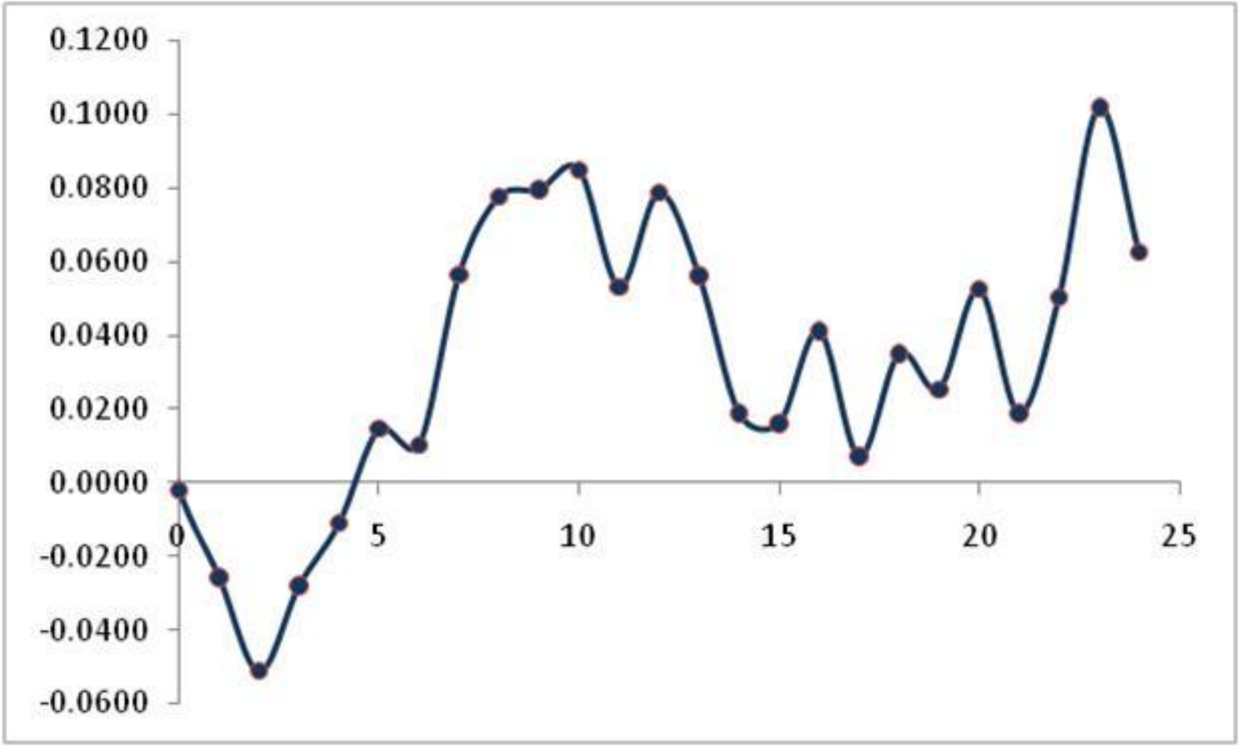
(2.21)

$$y_t = \alpha_0 + \sum_{i=1}^{p} \beta_i y_{t-i} + \sum_{i=0}^{q} \beta_i \varepsilon_{t-i}.$$

It may help to show two series constructed to have different ARMA patterns. Figure 7 shows one of the potential time series generated by the ARMA(2,1) process:

(2.22) $y_t = 0.67 y_{t-1} + 0.33 y_{t-2} + 0.1 \varepsilon_t + 0.05 \varepsilon_{t-1}$.

Figure 2.10.

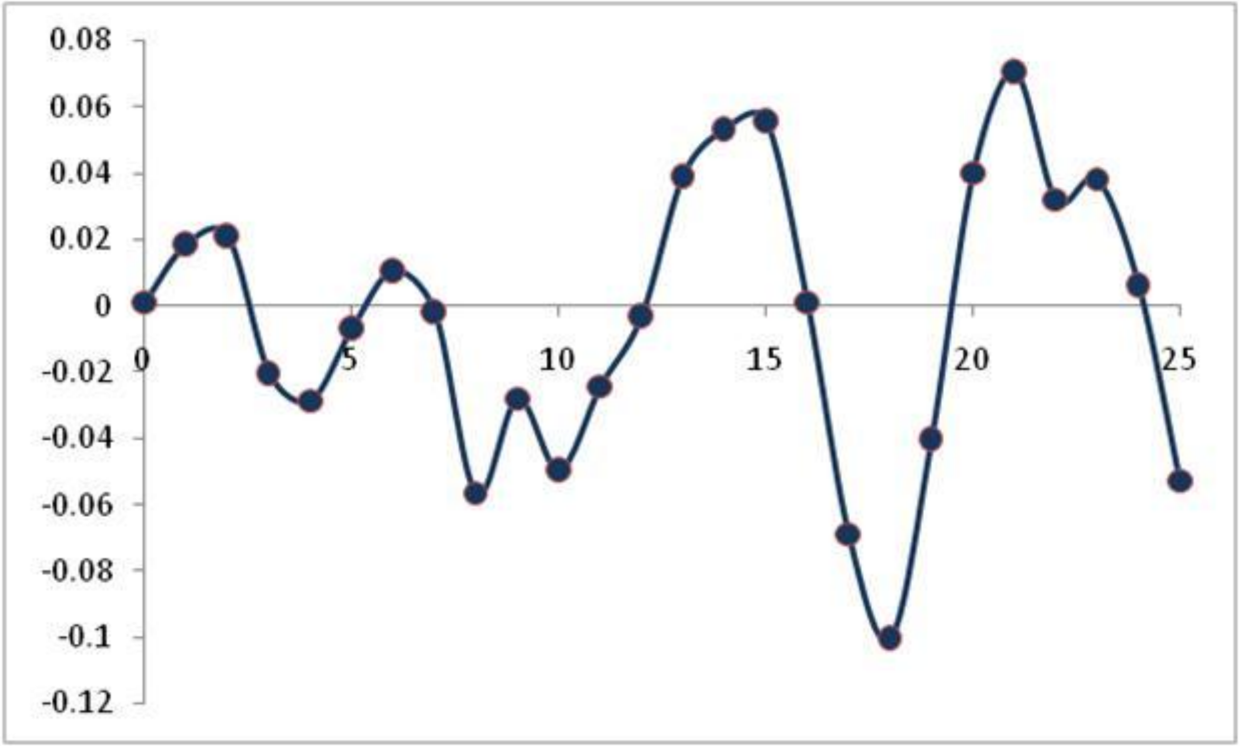


Graph of a ARMA(2,1) process.

Figure 8 shows one potential time series generated by the ARMA(1,1) process:

$$\text{(2.23)}\quad y_t = 0.67y_{t-1} + 0.1\varepsilon_t + 0.05\varepsilon_{t-1}.$$

Figure 2.11.



Graph of a ARMA(1,1) process.

Stationarity

Consider the time-series y_t. We define this stochastic process as *covariance stationary* if

(2.24)

$$E(y_t) = E(y_{t-s}) = \mu,$$

(2.25)

$$E\left[(y_t - \mu)^2\right] = E\left[(y_{t-s} - \mu)^2\right] = \sigma^2, \text{and}$$

(2.26)

$$Cov(y_t, y_{t-s}) = E[(y_t - \mu)(y_{t-s} - \mu)] = E\left[\left(y_{t-j} - \mu\right)\left(y_{t-j-s} - \mu\right)\right] = \gamma_s.$$

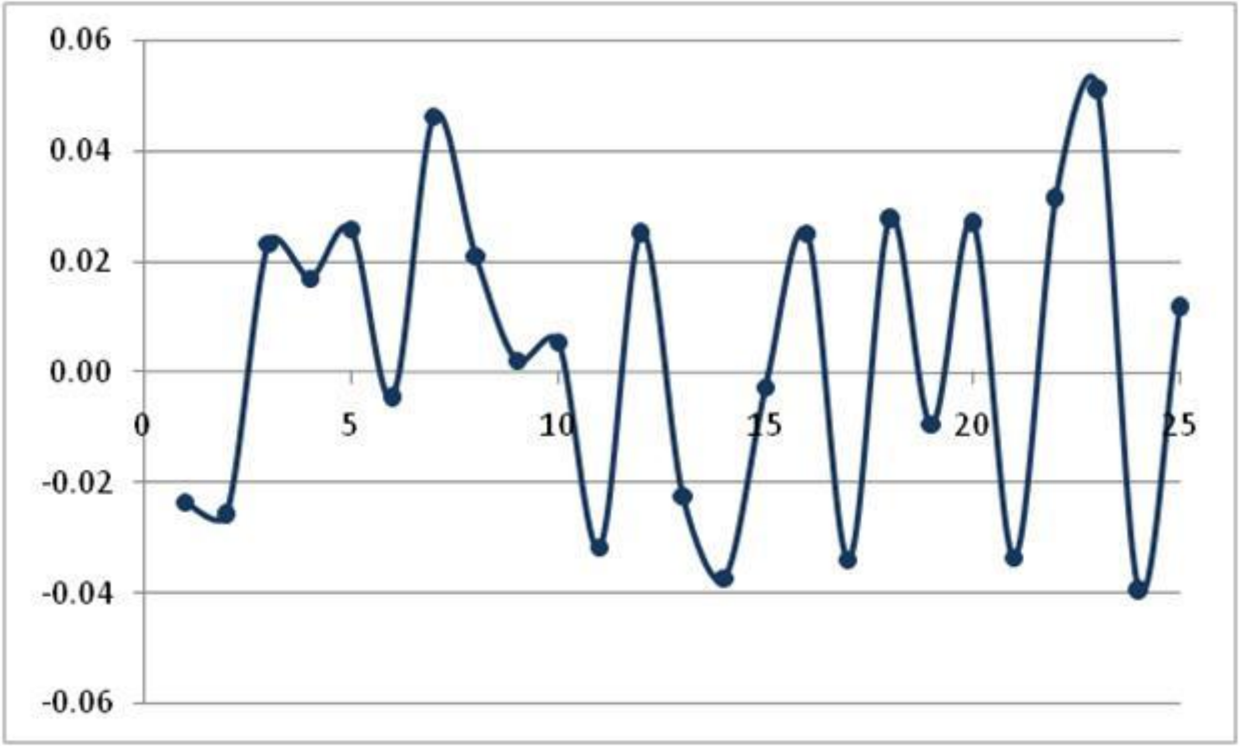
The last term, γ_s, is known as the autocovariance. A time-series is defined to be covariance stationary if its mean and all its autocovariances are unaffected by a change of time origin. We define the autocorrelation between y_t and y_{t-s} as:

(2.27)

$$\rho_s := \frac{\gamma_s}{\gamma_0}.$$

Quite often you can create a stationary time-series from a non-stationary time-series by taking the first-differences of the non-stationary series. If the first difference does not produce a stationary series, then one continues to take first differences until you find a stationary series. For instance, the time-series shown in Figure 7 appears to be non-stationary. The first differences of this series is shown in Figure 9. Using the imperfect eye, it would appear that the first differences of (13) is stationary. However, we really cannot tell anything for sure from the graph of a data set. We need to use the restrictions of the parameters derived in advanced texts to determine if a data set is stationary.[17]

Figure 2.12.



First-differences of the time-series of the ARMA(2,1) data.

The autocorrelation function

One of the major ways to identify the structure of a time series is to look at the autocorrelation function. The autocorrelation function, ρ_s, is the correlation between y_t and y_{t-s}. Stata uses the following formula to estimate it [StataCorp: p. 60] for a time-series:

The researcher then has to compare the actual autocorrelation function with the theoretical autocorrelation for comparable data series. To see to use the autocorrelation function consider the following five time series[18]:

(2.28)

$$\widehat{\rho}_s = \frac{\sum_{i=1}^{n-s} (y_t - \bar{y})(y_{t-s} - \bar{y})}{\sum_{i=1}^{n} (y_t - \bar{y})^2}.$$

(2.29) AR(1): $y_t = 0.7y_{t-1} + \varepsilon_t$,

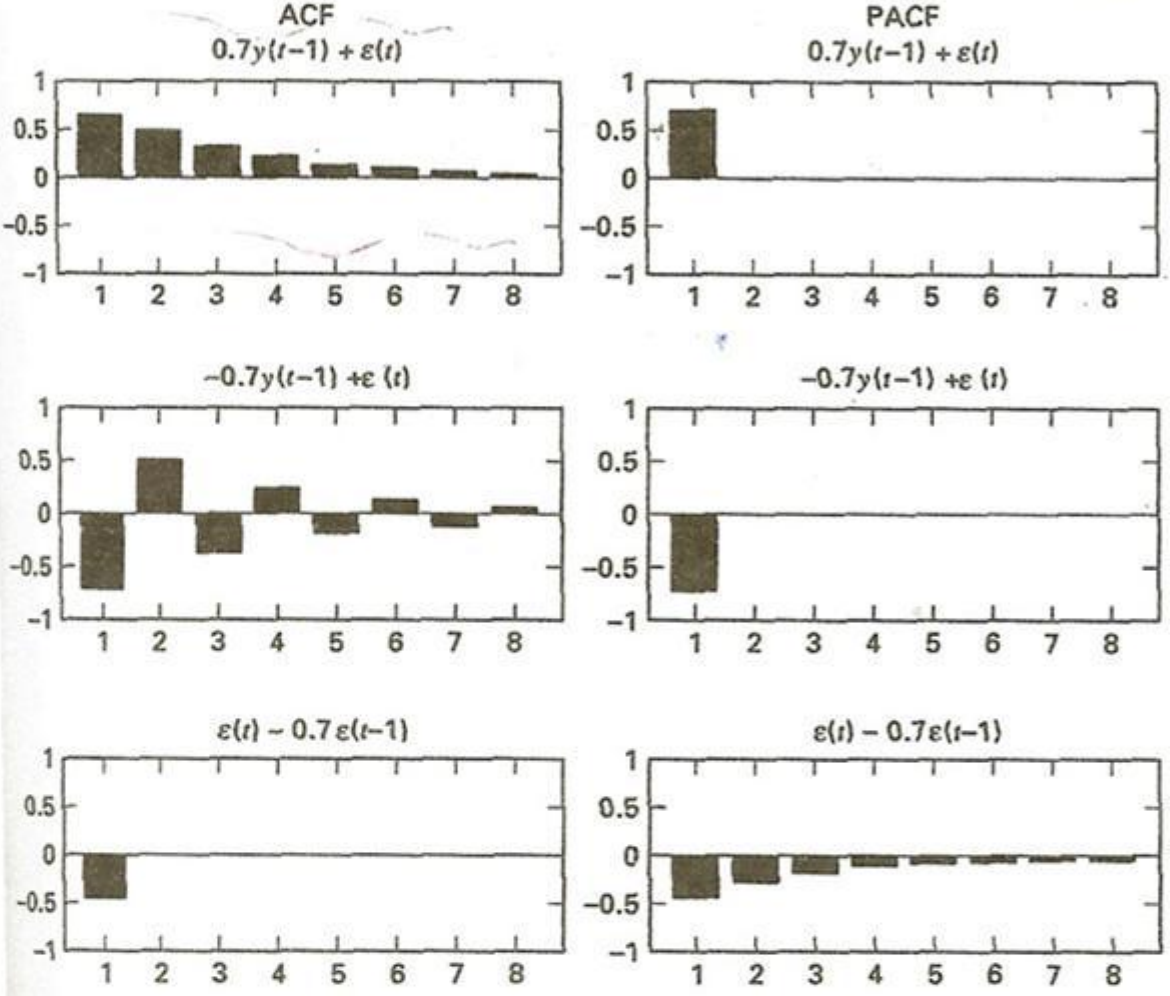
(2.30) AR(1): $y_t = -0.7y_{t-1} + \varepsilon_t$, (2.31) MA(1): $y_t = \varepsilon_t - 0.7\varepsilon_{t-1}$,

(2.32) ARMA(2, 1): $y_t = 0.7y_{t-1} - 0.49y_{t-2} + \varepsilon_t$, and

(2.33) ARMA(1, 2): $y_t = -0.7y_{t-1} + \varepsilon_t - 0.7\varepsilon_{t-1}$.

Each of these functions has a theoretical autocorrelation function; graphs of these autocorrelation functions are shown in the left column of Figure 10.[19]

Figure 2.13.



ACF

$0.7y(t-1) + \varepsilon(t)$

PACF

$0.7y(t-1) + \varepsilon(t)$

$-0.7y(t-1) + \varepsilon(t)$

$-0.7y(t-1) + \varepsilon(t)$

$\varepsilon(t) - 0.7\varepsilon(t-1)$

$\varepsilon(t) - 0.7\varepsilon(t-1)$

Examples of autocorrelation and partial autocorrelation functions.

There is additional function we can use to help identify the nature of a time-series. Consider the following regressions:

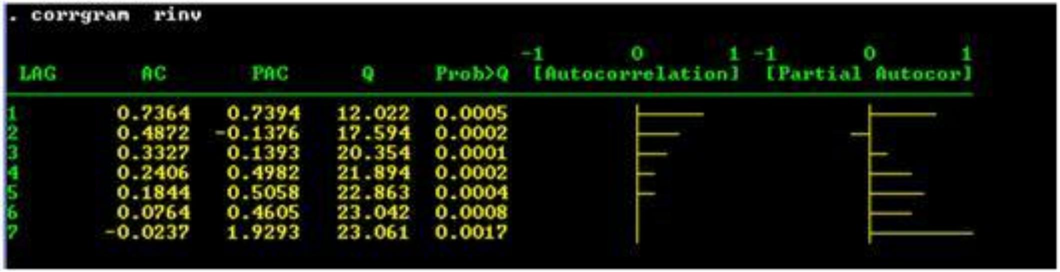
(2.34) $y_t^* = \varphi_{11} y^*_{t-1} + e_t$, $y_t^* = \varphi_{21} y^*_{t-1} + \varphi_{22} y^*_{t-2} + e_t$, **etc.,**

where $y_t^* = y_t - \bar{y}$.

Our interpretation of the φ_{ii} parameters is that they are the correlation between y_t and y_{t-i} controlling for all of the y_j where $j = 2,\ldots,(i - 1)$. Because these correlation coefficients control for values of y's observed between y_t and y_{t-i}, they are known as the partial autocorrelations. The theoretical partial autocorrelations are shown in the right column of Figure 10. Stata uses the command .corrgram varname to calculate the autocorrelations and partial autocorrelations for the time-series varname. Figure 11 shows the output when using this command on the real levels of investment. The autocorrelation function for this data set looks like the theoretical one for an AR(1) process. However, the partial autocorrelation function does not look like any of the

partial autocorrelation functions shown in Figure 11. Thus, it would not be safe to assume that real investment follows an AR(1) process.

Figure 2.14.

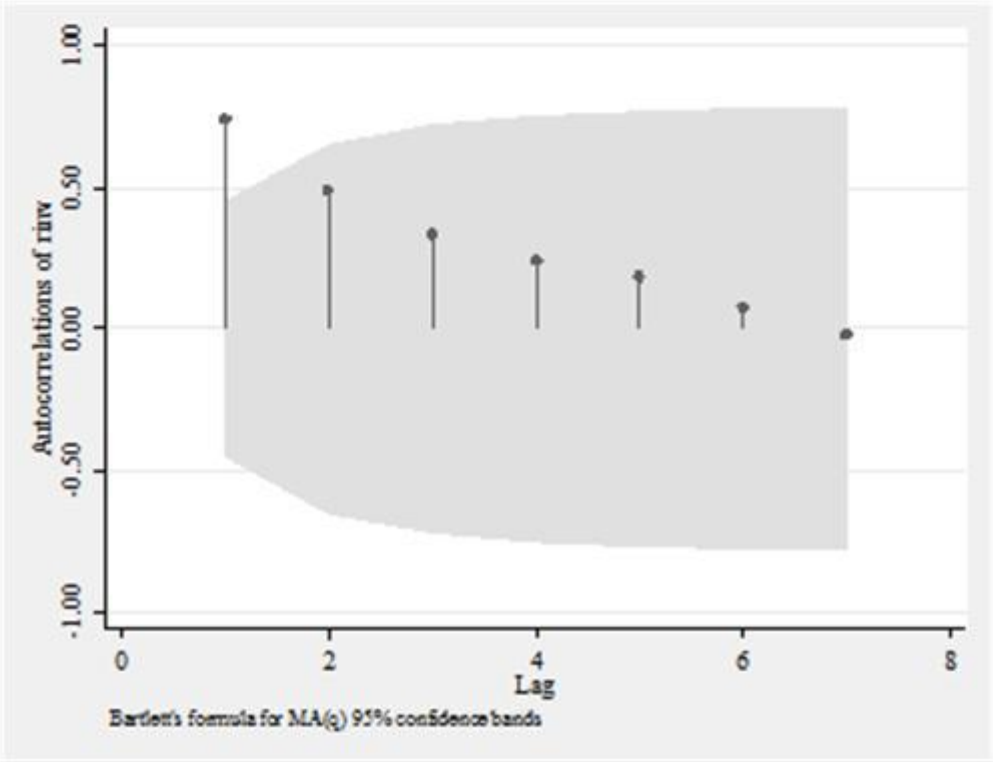


```
. corrgram  rinv

                                                   -1       0       1 -1       0       1
LAG       AC       PAC       Q      Prob>Q  [Autocorrelation]  [Partial Autocor]
-----------------------------------------------------------------------------------------
1        0.7364   0.7394   12.022  0.0005
2        0.4872  -0.1376   17.594  0.0002
3        0.3327   0.1393   20.354  0.0001
4        0.2406   0.4982   21.894  0.0002
5        0.1844   0.5058   22.863  0.0004
6        0.0764   0.4605   23.042  0.0008
7       -0.0237   1.9293   23.061  0.0017
```

Autocorrelation and partial autocorrelation functions for real investment.

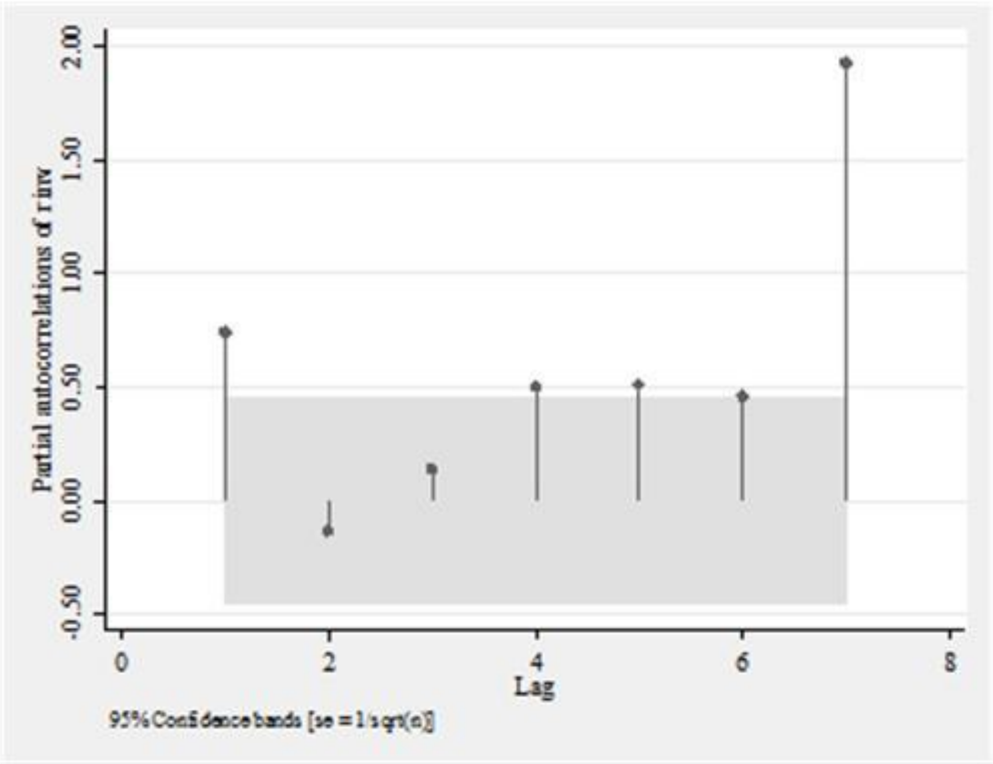
You can generate prettier graphs of the autocorrelation functions using the .ac varname command. For instance, the command .ac rinv generates the graph shown in Figure 12. The .pac varname generates a graph for the partial autocorrelations as is shown in Figure 13.

Figure 2.15.



Another graph of the autocorrelation function for real investment.

Figure 2.16.



Partial autocorrelations for real investments.

There are several generalizations one can use to help identify the process underlying a data series. Table 1 [Enders (2005): p. 85] offers a brief summary of these properties of the autocorrelation and partial autocorrelation functions.

Process	**Autocorrelation function**	**Partial autocorrelation function**
White-noise	**All $\rho_s = 0$**	**All $\varphi_{ss} = 0$**
AR(1): $\alpha_1 > 0$	**Direct exponential decay**	**$\varphi_{11} = \rho_1$; $\varphi_{ss} = 0$ for $s \geq 2$**
AR(1): $\alpha_1 > 0$	**Decays toward zero. Coefficients may oscillate**	**$\varphi_{11} = \rho_1$; $\varphi_{ss} = 0$ for $s \geq 2$**
AR(p)	**Decays toward zero; Coefficients may oscillate**	**Spikes through lag p. All $\varphi_{ss} = 0$ for $s > p$**
MA(1): $\beta > 0$	**Negative spike at lag 1. $\rho_s = 0$ for $s \geq 2$**	**Oscillating decay: $\varphi_{11} < 0$**
MA(1): $\beta < 0$	**Positive spike at lag 1. $\rho_s = 0$ for s**	**Decay: $\varphi_{11} > 0$**

	≥ 2	
ARMA(1, 1): $\alpha_1 > 0$	Exponential decay beginning at lag 1. Sign ρ_1 = sign $(\alpha_1 + \beta)$	Oscillating decay beginning at lag 1. $\varphi_{11} = \rho_1$
ARMA(1, 1): $\alpha_1 < 0$	Oscillating decay beginning at lag 1. Sign ρ_1 = sign $(\alpha_1 + \beta)$	Exponential decay beginning at lag 1. $\varphi_{11} = \rho_1$ and sign (ϕ_{ss}) = sign (ϕ_{11})
ARMA(p, q)	Decay (either direct or oscillatory) beginning at lag q	Decay (either direct or oscillatory) beginning at lag p

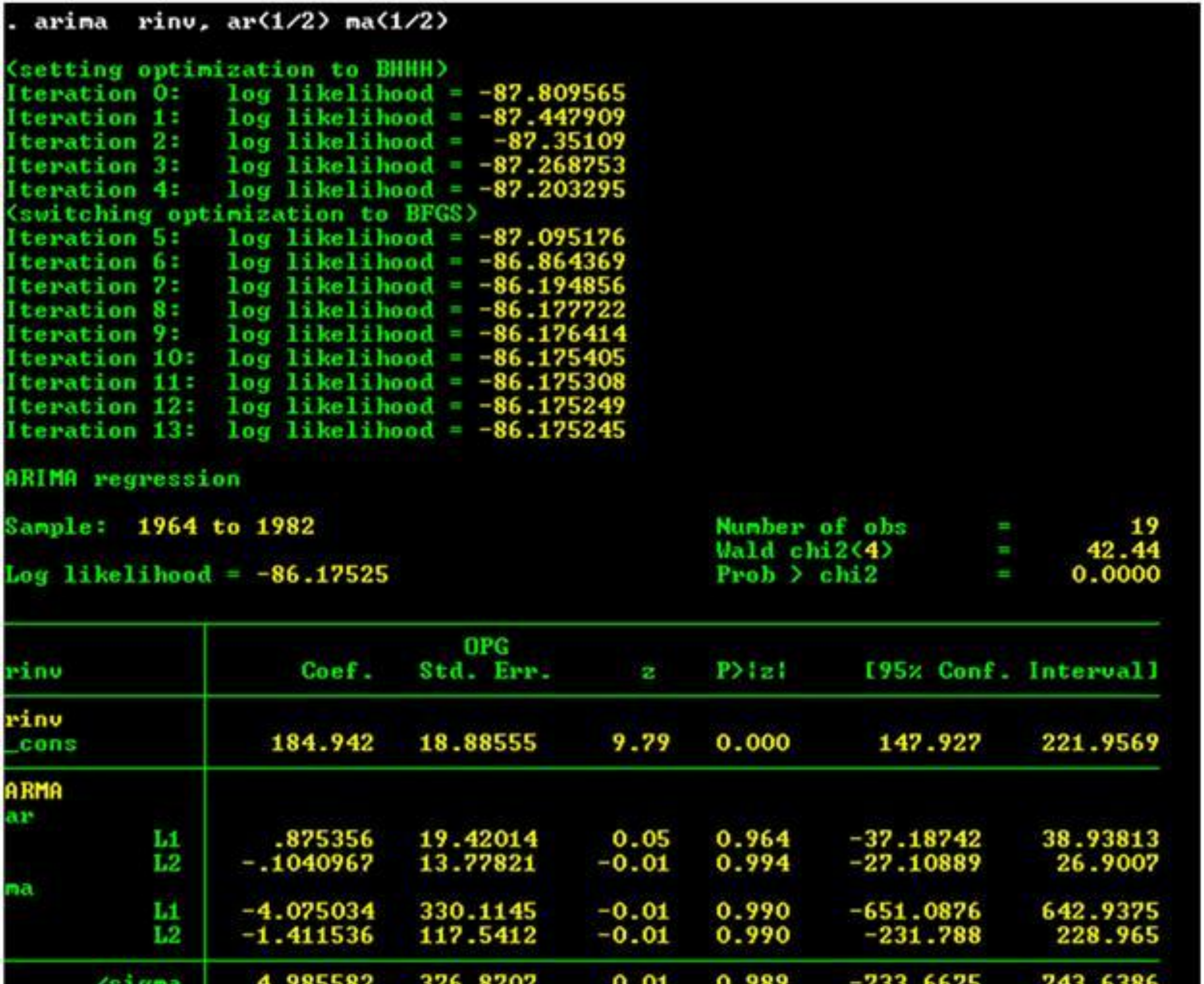
Table 2.10. Properties of the autocorrelation and partial functions.

Estimation of ARMA models

The estimation of ARMA models are relatively easy in *Stata*. The basic command to estimate an ARMA model is: .arima depvar [varlist], ar(*numlist*) ma(*numlist*).[20] The first thing to notice in the command that this command can apply to either to a single variable or to an equation. If [varlist] is omitted, *Stata* will produce an estimate of the

ARMA model for that variable; if the list is included, it will estimate the model with the disturbances allowed to have the ARMA structure specified in the command. Figure 14 reports the estimation of an ARMA model for real investment levels. Notice that we write AR(1/2) so that *Stata* knows to include both the first and second autoregressive term. A command of AR(2) would include only the second autoregressive term. In Figure 15 we report the ARMA (2, 1) estimation of (1).

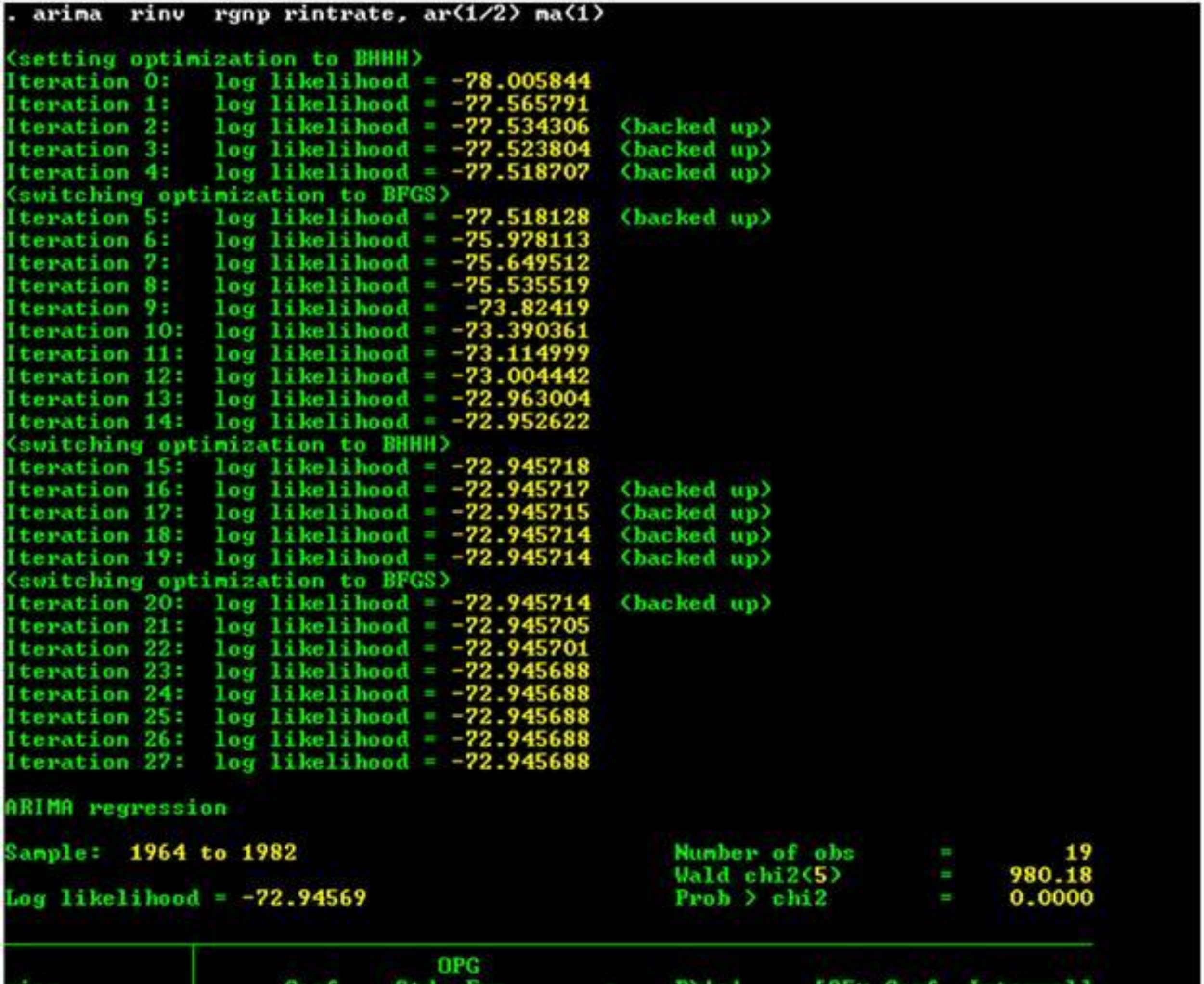
Figure 2.17.



```
. arima  rinv, ar(1/2) ma(1/2)

(setting optimization to BHHH)
Iteration 0:   log likelihood = -87.809565
Iteration 1:   log likelihood = -87.447909
Iteration 2:   log likelihood =  -87.35109
Iteration 3:   log likelihood = -87.268753
Iteration 4:   log likelihood = -87.203295
(switching optimization to BFGS)
Iteration 5:   log likelihood = -87.095176
Iteration 6:   log likelihood = -86.864369
Iteration 7:   log likelihood = -86.194856
Iteration 8:   log likelihood = -86.177722
Iteration 9:   log likelihood = -86.176414
Iteration 10:  log likelihood = -86.175405
Iteration 11:  log likelihood = -86.175308
Iteration 12:  log likelihood = -86.175249
Iteration 13:  log likelihood = -86.175245

ARIMA regression

Sample:  1964 to 1982                           Number of obs      =        19
                                                Wald chi2(4)       =     42.44
Log likelihood = -86.17525                      Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |                 OPG
        rinv |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
rinv         |
       _cons |    184.942   18.88555     9.79   0.000      147.927    221.9569
-------------+----------------------------------------------------------------
ARMA         |
ar           |
          L1 |    .875356   19.42014     0.05   0.964    -37.18742    38.93813
          L2 |  -.1040967   13.77821    -0.01   0.994    -27.10889     26.9007
ma           |
          L1 |  -4.075034   330.1145    -0.01   0.990    -651.0876    642.9375
          L2 |  -1.411536   117.5412    -0.01   0.990     -231.788     228.965
-------------+----------------------------------------------------------------
      /sigma |   4.985582   376.8707     0.01   0.989    -733.6675    743.6386
```

Estimation of an ARMA(2, 2) model of real investment.

Figure 2.18.



```
. arima  rinv  rgnp rintrate, ar(1/2) ma(1)

(setting optimization to BHHH)
Iteration 0:   log likelihood = -78.005844
Iteration 1:   log likelihood = -77.565791
Iteration 2:   log likelihood = -77.534306  (backed up)
Iteration 3:   log likelihood = -77.523804  (backed up)
Iteration 4:   log likelihood = -77.518707  (backed up)
(switching optimization to BFGS)
Iteration 5:   log likelihood = -77.518128  (backed up)
Iteration 6:   log likelihood = -75.978113
Iteration 7:   log likelihood = -75.649512
Iteration 8:   log likelihood = -75.535519
Iteration 9:   log likelihood =  -73.82419
Iteration 10:  log likelihood = -73.390361
Iteration 11:  log likelihood = -73.114999
Iteration 12:  log likelihood = -73.004442
Iteration 13:  log likelihood = -72.963004
Iteration 14:  log likelihood = -72.952622
(switching optimization to BHHH)
Iteration 15:  log likelihood = -72.945718
Iteration 16:  log likelihood = -72.945717  (backed up)
Iteration 17:  log likelihood = -72.945715  (backed up)
Iteration 18:  log likelihood = -72.945714  (backed up)
Iteration 19:  log likelihood = -72.945714  (backed up)
(switching optimization to BFGS)
Iteration 20:  log likelihood = -72.945714  (backed up)
Iteration 21:  log likelihood = -72.945705
Iteration 22:  log likelihood = -72.945701
Iteration 23:  log likelihood = -72.945688
Iteration 24:  log likelihood = -72.945688
Iteration 25:  log likelihood = -72.945688
Iteration 26:  log likelihood = -72.945688
Iteration 27:  log likelihood = -72.945688

ARIMA regression

Sample:  1964 to 1982                           Number of obs      =        19
                                                Wald chi2(5)       =    980.18
Log likelihood = -72.94569                      Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |                 OPG
```

Estimation of Equation (1) using an ARMA(2, 1) model.

	ARMA(1, 1)	ARMA(2, 1)	AR(1)	AR(2)	MA(1)
Intercept	185.307	185.6556	184.8208	185.2092	189.373
	(10.06)	(10.83)	(9.27)	(10.25)	(18.09)
AR (L1)	0.70936	1.76342	0.80307	0.95257	—
	(3.12)	(5.27)	(5.51)	(4.47)	—
AR (L2)	—	-0.81715	—	-0.18963	—
		(-3.21)		(-0.91)	
MA (L1)	0.26236	-0.99998	—	—	0.87262
	(0.90)	(-0.00)			(2.97)
Log likelihood	-86.1791	-85.8702	-86.47780	-86.21224	-88.48713

Wald χ2	26.96	422.60	30.36	31.65	8.81
Probability > χ2	0.0000	0.0000	0.0000	0.0000	0.0000
Sample size	19	19	19	19	19
(14,1)	1964-1982	1964-1982	1964-1982	1964-1982	1964-1982

Table 2.11. Estimation of various ARMA models of real investment.

The interpretation of these results is not obvious. We check the sensitivity of these results by estimation some other models. The results of these estimations are reported in Table 2 and Table 3. Based purely on ML tests, it would appear that AR(1) model in Table 2 is as good as any of the models describing the ARMA structure of real investments. On the other hand, the results reported in Table 3 suggests that the ARMA(2, 1) appears to be the best model to assume for the disturbance term in the estimates of Equation (1).

	AR(1)	ARMA(1, 1)	ARMA(2, 1)

Intercept	-14.49489	-13.37455	-16.89182
	(-0.26)	(-0.23)	(-1.68)
Real GNP	0.17006	0.16912	0.17253
	(3.96)	(3.78)	(20.18)
Real interest rate	-0.82517	-0.92007	-0.33692
	(-0.46)	(-0.33)	(-0.25)
AR (L1)	0.27953	-0.02028	0.85619
	(0.60)	(-0.02)	(1.46)
AR (L2)	—	—	-0.70702
			(-2.64)
MA (L1)	—	0.41151	-1.00000
		(0.42)	(-2.98)
Log likelihood	-78.7868	-78.4279	-72.94569

Wald χ^2	26.30	31.86	980.18
Probability > χ^2	0.0000	0.0000	0.0000
Sample size	19	19	19
Sample period	1964-1982	1964-1982	1964-1982

Table 2.12. Various ARMA estimates of Equation (1).

Other time-series concepts

There are a large number of additional time-series methods and issues that are not discussed in this module. These topics include, among others, ARCH and GARCH estimators, unit roots, the Dickey-Fuller test, and vector autoregression (VAR) models. There is no way to do justice to these topics in notes as short as these are. Moreover, it is necessary to discuss difference equations (the discrete version of differential equations) if one wants to understand many of these topic at anything more than an intuitive level. Those interested in these topics should enroll in the forecasting course (Economics 422)

or, if they cannot, plan to read several textbooks on whatever econometric tool they need to understand.

Exercise

Exercise 2.2.1.

This exercise is designed to be sure you know how to use *Stata* in analyzing time-series data sets; there is no economic content in the exercise. The MS Excel file Rabun County Temperature Data reports the morning temperature (MornTemp) observed in Rabun County, Georgia for every day between March 15, 2005 to November 2, 2008. The data set includes a variable “edate” that is the daily date in *Stata* notation. The data set also includes dummy variables for the season, the month, and the year of each observation (with the Winter, the December, and the 2008 dummy variables omitted).

a. Create a graph of (a) the data set morntemp, (2) the autocorrelations of morntemp, and (3) the partial autocorrelations of morntemp (you will have to set the matrix size to some number greater than 43 using the command .set matsize #).

b. Estimate the following models:

1. **ARMA(2,2) for morntemp.**
2. **ARMA(2,2) for morntemp as a function of the season dummy variables.**
3. **ARMA(2,2) for morntemp as a function of the monthly dummy variables.**
4. **ARMA(2,2) for morntemp as a function of the monthly dummy variables and the annual dummy variables.**
5. **ARMA(1,2) for morntemp as a function of the monthly dummy variables and the annual dummy variables.**
6. **ARMA(1,1) for morntemp as a function of the monthly dummy variables and the annual dummy variables.**

c. Arrange the parameter estimates in a table and comment on them. Include the results of estimating (6) using OLS; what is the DW-statistic for this regression?

References

Cochran, D. and G. Orcutt (1949). Application of Least Squares Regression to Relationships Containing Autocorrelated Error Terms. *Journal of the American Statistical Association* 44: 32-61.

Enders, Walter (1995). *Applied Econometric Time Series* (New York: John Wiley & Sons, Inc.).

Greene, William H. (1990). *Econometric Analysis* (New York: Macmillan Publishing Company).

StataCorp (2003). *Stata Statistical Software: Release 8.0: Stata Time-Series Reference Manual* (College Station, TX: Stat Corporation).

2.3. Panel Data Models[*]

Equation Chapter 1 Section 1Notes on Panel Data Models

Introduction

Panel data methods are appropriate when the researcher has available observations that are both cross-sectional and time series. For example, one could form a panel data set with observations on the per capita consumption of tobacco for a set of OECD countries over the period 1960 to 2005. Usually the data is "stacked"—that is, all of the observations for country A is listed together in order of year before the data for country B, etc. It is also possible to stack the data by year—countries A to Z for 1960, countries A to Z for 1961, and so on through 2005.

Let y_{it} be the per capita consumption of tobacco for country *i* in year *t*. We wish to model the per capita consumption of tobacco as a function of a set of observable independent variables like the price of tobacco, income, restrictions on tobacco advertising, and restrictions on tobacco consumption. Of course there are several sources of unobserved heterogeneity in that data set. In particular, we might expect that systematic differences in consumption patterns would exist due to differences in the customs and mores of the various countries in the sample. It also would be reasonable to assume that these country-level differences are be relatively stable over time. Additionally, we might expect that there would be differences the per capita consumption of tobacco over time due to changes in our understanding of the long run health effects of tobacco consumption. These changes might affect both (1) the level of consumption and (2) the responsiveness of the consumption of tobacco to changes in the explanatory variables.

In these notes we describe some of the ways of modeling panel data sets and discuss some of the issues associated with the estimation of these models. We also discuss how to use *Stata* to analyze panel data sets. We begin by considering some of the types of panel data model specifications.

Model specification

There are four general specifications of the panel data model available. The differences in these models reflect differing assumptions one might make and are listed below.

1. Slope coefficients are constant and the intercept varies over the individuals:

(2.35)

$$y_{it} = \alpha_i + \sum_{j=1}^{k} \beta_j x_{jit} + \varepsilon_{it}, i = 1, \ \ldots \ , N, i = 1, \ \ldots \ , N, \mathrm{and} t = 1, \ \ldots \ , T.$$

2. Slope coefficients are constant and the intercept varies over the individuals and over time:

(2.36)

$$y_{it} = \alpha_{it} + \sum_{j=1}^{k} \beta_j x_{jit} + \varepsilon_{it}, i = 1, \ldots, N, \text{and} t = 1, \ldots, T.$$

3. All coefficients vary over individuals:

$$y_{it} = \alpha_i + \sum_{j=1}^{k} \beta_{ji} x_{jit} + \varepsilon_{it}, i = 1, \ldots, N, \text{and} t = 1, \ldots, T. \qquad (2.37)$$

4. All coefficients vary over time and individuals:

$$y_{it} = \alpha_{it} + \sum_{j=1}^{k} \beta_{jit} x_{jit} + \varepsilon_{it}, i = 1, \ldots, N, \text{and} t = 1, \ldots, T. \qquad (2.38)$$

These four models can be classified further, depending on whether the researcher assumes that the coefficients of the model are fixed or random. However, most research in economics is restricted to estimation of (1) and (2) because they strike a reasonable

balance between being general enough without introducing unnecessary assumptions that can render estimation extremely difficult.

Estimation issues

Hsiao (2003: 27-30) discusses a convenient example of a panel data model that illustrates many of the important issues that arise with panel data. We make use of this example in what follows. Assume that we want to estimate a production function for farm production in order to determine if the farm industry exhibits increasing returns to scale. Assume the sample consists of observations for *N* farms over *T* years, giving a total sample size of *N T* . For simplicity, we assume that the Cobb-Douglas production is an adequate description of the production process. The general form of the Cobb-Douglas production function is:

$$(2.39)\quad q = \alpha_0 I_1^{\beta_1} \cdots I_k^{\beta_k},$$

where q is output and I_j is the quantity of the j-th input (for example, land, machinery, labor, feed, and fertilizer). The parameter, β_j, is the output elasticity of the j-th input; the farms exhibit constant returns to scale if the output elasticities sum to one and either increasing or decreasing returns to scale if they sum to a value greater than or less than one, respectively. is the quantity of the *j*-th input (for example, land, machinery, labor, feed, and fertilizer). The parameter, is the output elasticity of the *j*-th input; the farms exhibit constant returns to scale if the output elasticities sum to one and either increasing or decreasing returns to scale if they sum to a value greater than or less than one, respectively.

Taking the natural logarithm of (5) gives $\ln q = \ln \alpha_0 + \beta_1 \ln I_1 + \cdots + \beta_k \ln I_k$. We can rewrite this equation (adding an error term, as well as farm and year subscripts) giving:

$$(2.40)\quad y_{it} = \beta_0 + \beta_1 x_{1it} + \cdots + \beta_k x_{kit} + \varepsilon_{it},$$

where $y_{it} = \ln q_{it}$, , $\beta_0 = \ln \alpha_0$, $x_{jit} = \ln I_{jit}$, for $j = 1,\ldots,k$ and ε_{it} is an error term. One way to account for year and time effects is to assume:

$$(2.41)\quad \varepsilon_{it} = \lambda F_i + \eta P_t + \upsilon_{it},$$

where F_i is a measure of the unobserved farm specific effects on productivity and P_t is a measure of the unobserved changes in productivity that are the same for all farms but vary annually. Substitution of (7) into (6) gives:

$$y_{it} = (\beta_0 + \lambda F_i + \eta P_t) + \sum_{j=1}^{k} \beta_j x_{jit} + \upsilon_{it}$$

or

(2.42)

$$y_{it} = \alpha_{it} + \sum_{j=1}^{k} \beta_j x_{jit} + \upsilon_{it},$$

where $\alpha_{it} = \beta_0 + \lambda F_i + \eta P_t$. Thus, (8) is equivalent to (2). Moreover, if we assume that $\eta = 0$, we get

(2.43)

$$y_{it} = \alpha_i + \sum_{j=1}^{k} \beta_j x_{jit} + \upsilon_{it},$$

where $\alpha_i = \beta_0 + \lambda F_i$. Thus, (9) is equivalent to (1).

Fixed-effects models

A natural way to make (9) operational is to introduce a dummy variable, D_i , for each farm so that the intercept term becomes:

(2.44)

$$\alpha_i = \alpha_1 + \alpha_2 D_2 + \cdots + \alpha_m D_m = \alpha_1 + \sum_{j=2}^{m} \alpha_j D_j,$$

where *D* $_j$ = 1 if *j* = *i* and 0 otherwise. This substitution is equivalent to replacing the intercept term with a dummy variable for each farm and letting the farm dummy variable "sweep out" the farm-specific effects. In this specification the slope terms are the same for every farm while the intercept term is given for farm *j* by *α* $_1$ + *α* $_j$. Clearly, the intercept term for the first farm is equal to just *α* $_1$. This specification is known as the *fixed effect model* and is estimated using ordinary least squared (OLS). We can extend the fixed-effects model to fit (8) by including a dummy variable for each time period except one.

In sum, *fixed-effects models* assume either (or both) that the omitted effects that are specific to cross-sectional units are constant over time or that the effects specific to time are constant over the cross-sectional units. This method is equivalent to including a dummy variable for all but one of the cross-sectional units and/or a dummy variable for all but one of the time periods.

Random-effects models

An alternative approach to treating the α_i in (1) as fixed constants over time is to treat it as a random variable. intercepts vary due to individual level differences, we have

$$y_{it} = \alpha_i + \sum_{j=1}^{k} \beta_k x_{kit} + \varepsilon_{it}.$$

Treating α_i as a random variable is equivalent to setting the model up as:

(2.45)

$$y_{it} = \alpha + \sum_{j=1}^{k} \beta_j x_{jit} + (\alpha_i + \lambda_t + \varepsilon_{it}).$$

For simplicity we consider only the case when $\lambda_t = 0$. Thus, the error term for (11) is $(\alpha_i + \varepsilon_{it})$. We assume that

(2.46)

$$
\begin{aligned}
&E(\alpha_i) = E(\varepsilon_{it}) = 0,\\
&E(\alpha_i \varepsilon_{it}) = 0,\\
&E(\alpha_i \alpha_j) = \begin{cases} \sigma_\alpha^2 & \text{if}\, i = j \\ 0 & \text{if}\, i \neq j \end{cases}, \quad \text{and}\\
&E(\varepsilon_{it} \varepsilon_{js}) = \begin{cases} \sigma_\varepsilon^2 & \text{if}\, i = j,\ t = s \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

We also assume that all of the elements of the error term are uncorrelated with the explanatory variables, x_j.

The key econometric issue is that the presence of α_i in the error term means that the correlation among the residual of the same cross-sectional unit is not zero; the error terms for one farm, for instance, are correlated with each other. Therefore, the error terms exhibit heteroskedasticity. The appropriate estimation technique is generalized-least-squares, a technique that attempts to adjust the parameter estimates (and their standard error estimates) for heteroskedasticity and autocorrelation. Alternatively one can assume that α_i and ε_{it} are normally distributed and use a ML estimator. Hsiao [2003: 35-41] and Cameron and Trivedi [2005: 699-716] offer greater detail on the estimation of the parameters of both the fixed-effects and the random-effects models. It is enough for

our purposes to accept that the econometricians have found a number of ways to estimate these parameters.

Random-effects or fixed effect model?

Economists generally prefer to use fixed-effects models. The decision to use fixed-effects or random-effects does not matter when *T* is large because the two methods will yield the same estimates of the parameters. When the number of individual categories (*N*) is large and the number of time periods (*T*) is small, the choice of which model to use becomes unclear. Hsiao summarized this somewhat arcane issue with the following observations:

> ***If the effects of omitted variables can be appropriately summarized by a random variable and the individual (or time) effects represent the ignorance of the investigator, it does not see reasonable to treat one source of ignorance () as fixed and the other source of ignorance () as random. It appears that one way to unify the fixed-effects and random-effects models is to assume from***
>
> ***the outset that the effects are random. The fixed-effects model is viewed as one in which investigators make inferences conditional on the effects that are in the sample. The random-effects model is viewed as one in which investigators make***

> ***unconditional or marginal inferences with respect to the population of all effects. There is really no distinction in the "nature (of the effect)." It is up to the investigator to decide whether to make inference with respect to population characteristics or only with respect to the effects that are in the sample. Hsiao [2003: 43]***

Needless to say, Hsiao's advice may well leave many researchers without any idea of whether to use a random-effects or a fixed-effects model. *In your own research I suggest that you consult an econometrician for advice*.

There is one problem that arises when using a fixed-effects model. Assume that you have a sample of observations for a large number of individuals over a period of years. If you use a fixed-effects model, you will not be able to find parameter estimates for any variable like race or sex that do not change over the time period of the sample. The reason for this limitation is that the time-constant variables are perfectly correlated with the dummy variables used for the fixed-effects. A similar problem arises if the fixed-effects are for years (rather than individuals). You cannot include a variable is constant for all individuals in any given year. Quite often the individual-constant (or time-constant) variable is not of interest and nothing is lost by not having the parameter estimate. On the other hand, the random-effects model does not have this problem because the estimation makes use of differences amongst the individuals to estimate a parameter for

the individual-constant variable.[21] We discuss in the next section an example in which this "problem" arises.

What would be nice is if there were a statistical test that allows us to decide if the random-effects model is the appropriate model? The *Hausman test* offers such a statistical test. The Hausman (specification) test exploits the fact that the parameters for the random-effects model should be not be statistically different from those found using a fixed-effects specification. If one observes a chi-squared value greater than the critical value you can conclude that the parameter estimates for the random-effects model are statistically different from the parameter estimates for a model using an assumption of fixed-effects, then you can conclude that the random-effects model is misspecified. Unfortunately, the misspecification could be due to the fact that the fixed-effects model is appropriate or it could be due to the unobserved error terms being correlated with the included explanatory variables. If the latter is the case, then one might consider augmenting the model with an appropriate measure of the part of the unobserved effect that is correlated with the error term. What we are describing is that same thing that happens when omitted variables are correlated with the error term—the parameter estimates are biased. We include an example of how to use *Stata* to perform the Housman specification test.

Estimation of panel data models in Stata

General comments

There are three commands that matter in setting up the panel data. The first two commands precede the regression command because they establish which variable denotes the time period and which variable denotes the cross-sectional unit. These commands are:

```
.iis [variable name]
```

```
.tis [variable name]
```

The command for estimating the fixed-effects model is:

```
. xtreg depvar [varlist], fe
```

The command for estimating the random-effects model is:

```
. xtreg depvar [varlist], re
```

If the part of the command with the comma and either re or fe is omitted, *Stata* will assume that you want to estimate the random-effects model.

Understanding Stata output

To understand the *Stata* output we need to return to the algebra of the model. Assume that we are fitting a model of the following form:

(2.47)

$$y_{it} = \alpha + \sum_{j=1}^{k} \beta_j x_{jit} + \nu_i + \varepsilon_{it}, i = 1, \ldots, N, \text{and} t = 1, \ldots, T.$$

We can sum (13) over *t* (holding the individual unit constant) and divide by *T* to get:

(2.48)

$$\bar{y}_i = \alpha + \sum_{j=1}^{k} \beta_j \bar{x}_{ji} + \nu_i + \bar{\varepsilon}_i,$$

where $\bar{y}_i = \frac{\sum_{t=1}^{T} y_{it}}{T}$, $\bar{x}_{ji} = \frac{\sum_{t=1}^{T} x_{it}}{T}$, and $\bar{\varepsilon}_i = \frac{\sum_{t=1}^{T} \varepsilon_{it}}{T}$. Thus, (14) uses the mean values for each cross-sectional unit. We can subtract (14) from (13) to get:

(2.49)

$$(y_{it} - \bar{y}_i) = \sum_{j=1}^{k} \beta_j (x_{jit} - \bar{x}_{ji}) + (\varepsilon_{it} - \bar{\varepsilon}_i).$$

Equations (13), (14), and (15) are the basis of *Stats's* estimates of the parameters of the model. In particular, the command xtreg, fe uses OLS to estimate (15); this is known as the *fixed-effects* estimator (or the *within* estimator). The command xtreg, be uses OLS to estimate (14) and is known as the *between* estimator. The command xtreg, re—the random-effects estimator—is a weighted average of the between and within estimators, where the weight is a function of the variances of and (and respectively).[22]

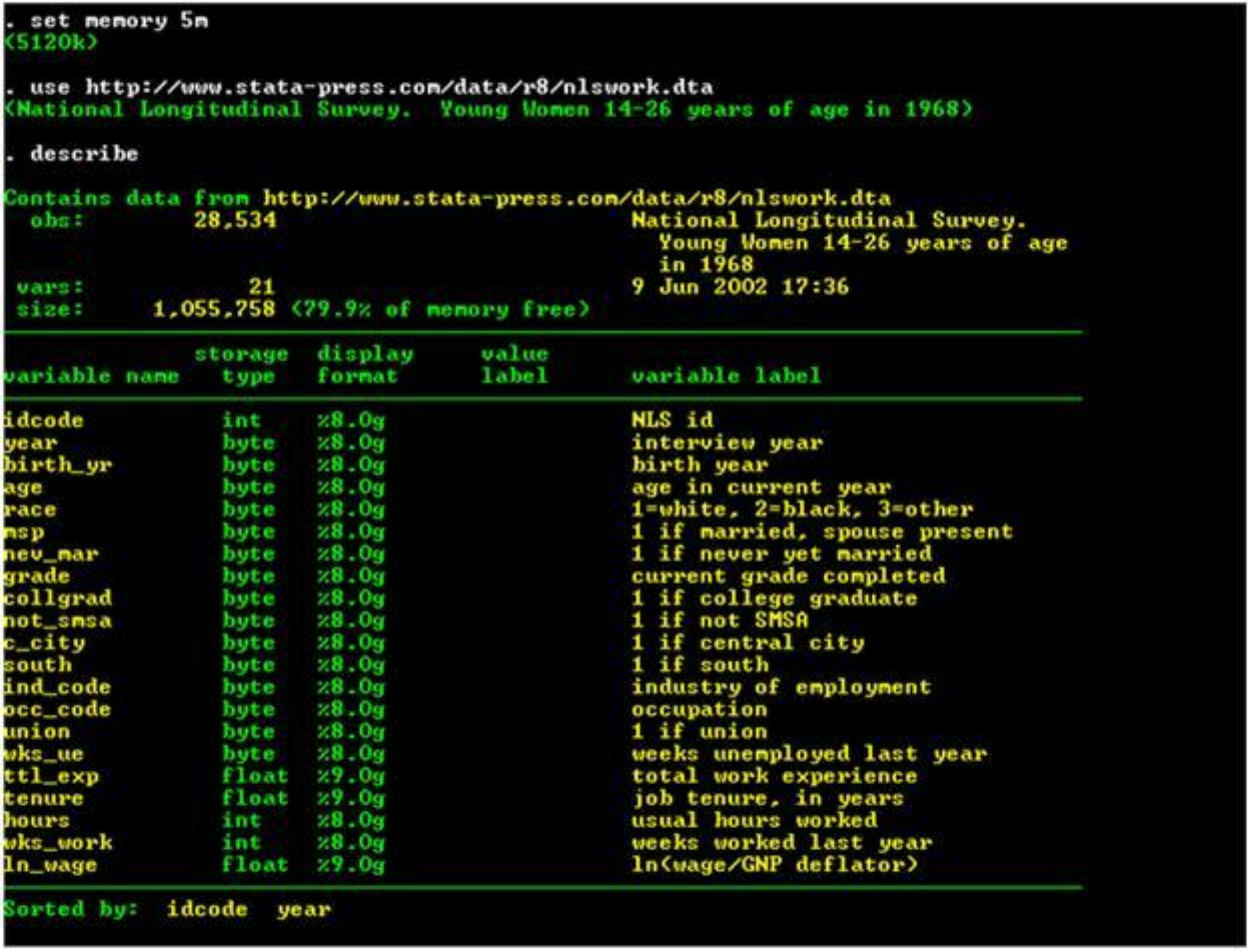
In general, you will not make use of the between estimator. However, these three equations do lie at the basis of the goodness-of-fit measures that *Stata* reports. In particular, *Stata* output reports three "R-squareds"[23]—the *overall-R^2* the *between-R^2* and

the *within-R*^{2} These three R-squareds are derived using one of the three equations. In particular, the overall-*R*^{2} uses (13); the between-*R*^{2} uses (14); and the within-*R*^{2} uses (15).

Example 2.3. A panel data analysis using *Stata*

In this example we follow the example offered in the *Stata* manual and use a large data set from the National Longitudinal Survey of wage data on 28,534 women who were between 14 and 26 years of age in 1968. The women were surveyed in each of the 21 years between 1968 and 1988 except for the six years 1974, 1976, 1979, 1981, 1984, and 1986. The study is focused on the determinants of wage levels, as measured by the natural logarithm of real wages.

Figure 2.19.

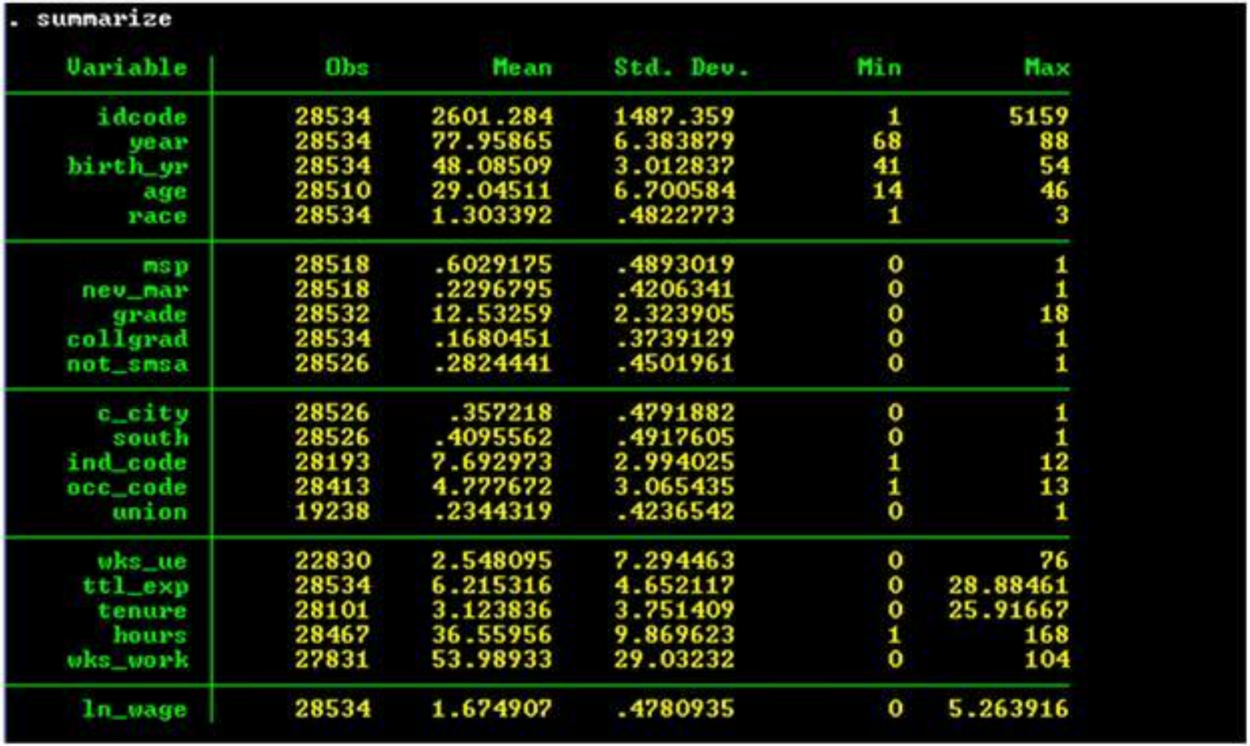


```
. set memory 5m
(5120k)

. use http://www.stata-press.com/data/r8/nlswork.dta
(National Longitudinal Survey.  Young Women 14-26 years of age in 1968)

. describe

Contains data from http://www.stata-press.com/data/r8/nlswork.dta
  obs:        28,534                          National Longitudinal Survey.
                                                Young Women 14-26 years of age
                                                in 1968
 vars:            21                          9 Jun 2002 17:36
 size:     1,055,758 (79.9% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
idcode          int    %8.0g                  NLS id
year            byte   %8.0g                  interview year
birth_yr        byte   %8.0g                  birth year
age             byte   %8.0g                  age in current year
race            byte   %8.0g                  1=white, 2=black, 3=other
msp             byte   %8.0g                  1 if married, spouse present
nev_mar         byte   %8.0g                  1 if never yet married
grade           byte   %8.0g                  current grade completed
collgrad        byte   %8.0g                  1 if college graduate
not_smsa        byte   %8.0g                  1 if not SMSA
c_city          byte   %8.0g                  1 if central city
south           byte   %8.0g                  1 if south
ind_code        byte   %8.0g                  industry of employment
occ_code        byte   %8.0g                  occupation
union           byte   %8.0g                  1 if union
wks_ue          byte   %8.0g                  weeks unemployed last year
ttl_exp         float  %9.0g                  total work experience
tenure          float  %9.0g                  job tenure, in years
hours           int    %8.0g                  usual hours worked
wks_work        int    %8.0g                  weeks worked last year
ln_wage         float  %9.0g                  ln(wage/GNP deflator)
-------------------------------------------------------------------------------
Sorted by:  idcode  year
```

Loading in the data set into *Stata* with a description of the data.

Figure 1 shows the commands used to put the data into *Stata*. The first command (set memory 5m) increases the size of the memory that the program uses; I did this because of the large sample size. The use command accesses that data from the *Stata* web site. The describe command calls up a description of the variables. Figure 2 presents a summary of the data using the command summerize.

Figure 2.20.



```
. summarize

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      idcode |     28534    2601.284    1487.359          1       5159
        year |     28534    77.95865    6.383879         68         88
    birth_yr |     28534    48.08509    3.012837         41         54
         age |     28510    29.04511    6.700584         14         46
        race |     28534    1.303392    .4822773          1          3
-------------+--------------------------------------------------------
         msp |     28518    .6029175    .4893019          0          1
     nev_mar |     28518    .2296795    .4206341          0          1
       grade |     28532    12.53259    2.323905          0         18
    collgrad |     28534    .1680451    .3739129          0          1
    not_smsa |     28526    .2824441    .4501961          0          1
-------------+--------------------------------------------------------
      c_city |     28526     .357218    .4791882          0          1
       south |     28526    .4095562    .4917605          0          1
    ind_code |     28193    7.692973    2.994025          1         12
    occ_code |     28413    4.777672    3.065435          1         13
       union |     19238    .2344319    .4236542          0          1
-------------+--------------------------------------------------------
      wks_ue |     22830    2.548095    7.294463          0         76
     ttl_exp |     28534    6.215316    4.652117          0   28.88461
      tenure |     28101    3.123836    3.751409          0   25.91667
       hours |     28467    36.55956    9.869623          1        168
    wks_work |     27831    53.98933    29.03232          0        104
-------------+--------------------------------------------------------
     ln_wage |     28534    1.674907    .4780935          0   5.263916
```

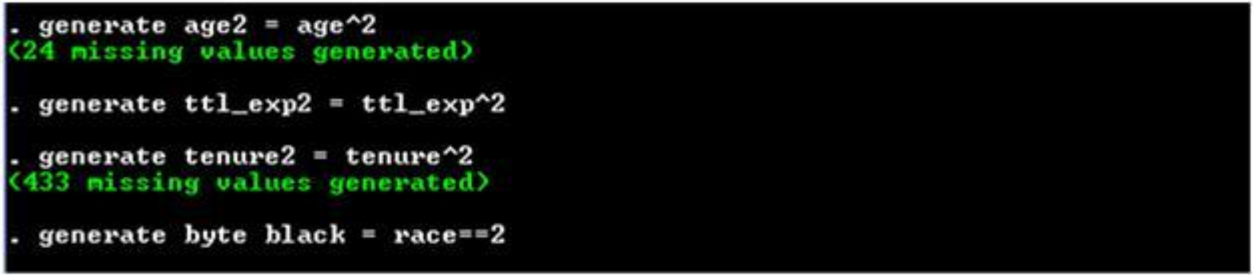
Summary of the data.

There are several transformations of the variables that we will need. In particular, we want to include the squares of several of the variables in our regression—age (age), work experience (ttl_exp), and job tenure (tenure). The reason we want to use the square of these variables is that we have reason to believe that wages have a non-linear relationship with these variables. For instance, consider the number of years a worker has been on the job, Tenure. Theory suggests that wages increase over a worker's work-life at a decreasing rate. Thus, if the equation we are estimating is $y = \ln w = \beta_0 + \beta_1 Tenure + \beta_2 Tenure^2 + \cdots$, what we expect is that: $\frac{\partial y}{\partial Tenure} = \beta_1 + 2\beta_2 Tenure > 0$ and $\frac{\partial^2 y}{\partial Tenure^2} = 2\beta_2 < 0.$ The only way that this last equation can be true is if $\beta_2 < 0$. Moreover, if this is true, the first-derivative implies that $\beta_1 > -2\beta_2 Tenure > 0$. Also, notice that we can determine the number of years in a job when wages reach a peak; y reaches a maximum at the age where $\frac{\partial y}{\partial Tenure} = \beta_1 + 2\beta_2 Tenure = 0$. or when

$Tenure = -\frac{\beta_1}{2\beta_2}$. The fact that $\frac{\partial^2 y}{\partial Tenure^2} = 2\beta_2 < 0$ guarantees that this point is indeed a maximum.

Additionally, because race is a categorical variable that has three potential values—1 if white, 2 if black, and 3 otherwise—we have to create a dummy variable in order to use this variable. The transformations we use are shown in Figure 3.

Figure 2.21.

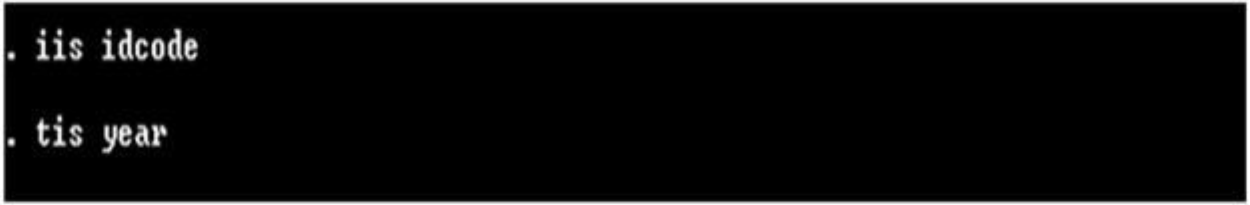


```
. generate age2 = age^2
(24 missing values generated)

. generate ttl_exp2 = ttl_exp^2

. generate tenure2 = tenure^2
(433 missing values generated)

. generate byte black = race==2
```

Transformations of the variables to create new variables.

The last step before estimating the regressions is to identify the data set as a panel data. shows the two commands that must be entered in order for *Stata* to know that idcode is the individual category and that year is the time series variable. Figure 4 shows these two commands.

Figure 2.22.



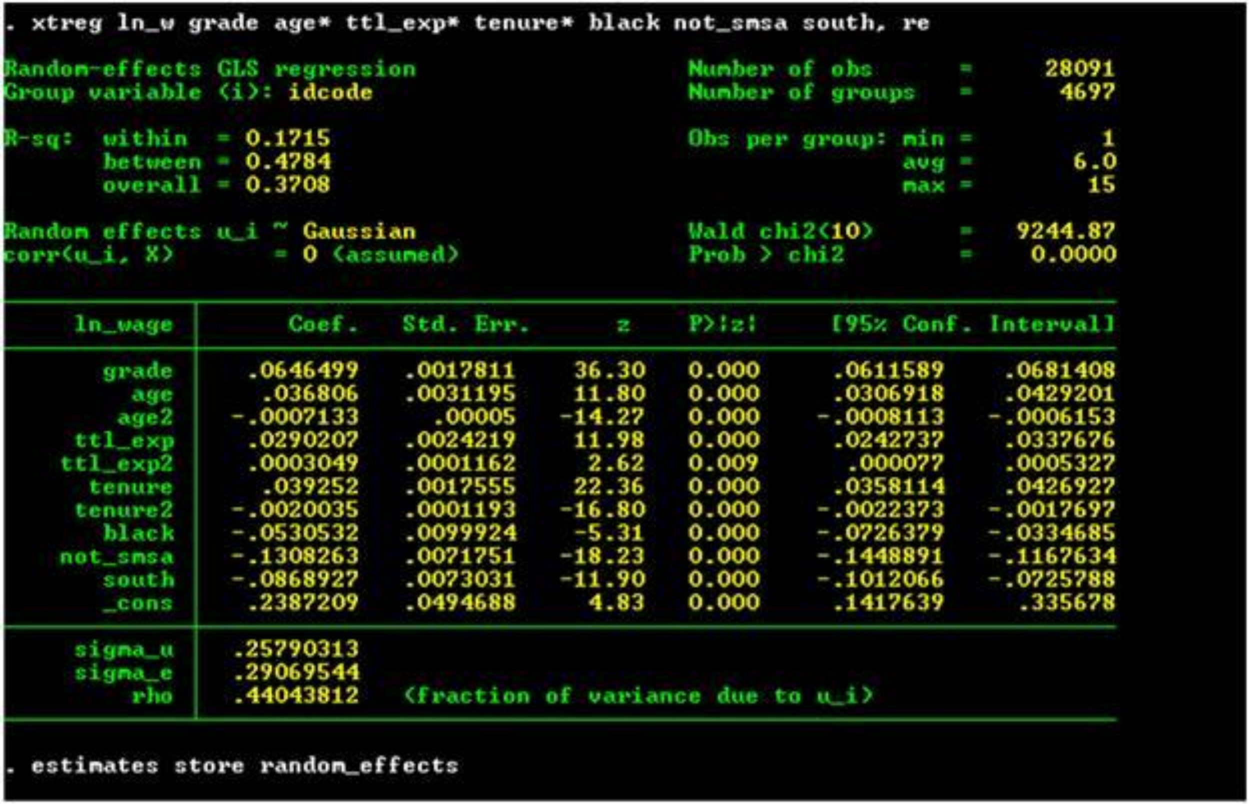
```
. iis idcode

. tis year
```

Declaring the category and time identifiers.

We are now ready to estimate the model (the natural logarithm of wages as a function of various variables). We begin with the random-effects model. Figure 5 shows the command and the results of the estimation of the random-effects model. There are several things to note here. First, in the command we are able to refer to all variables that have age in them by using age*, the * tells *Stata* to use and variable that begins with the

letters age. Second, we will need to use the estimation results in the Hausman test. Thus, we have stored these results in “random_effects” using the command estimates store random_effects.

Figure 2.23.



```
. xtreg ln_w grade age* ttl_exp* tenure* black not_smsa south, re

Random-effects GLS regression                   Number of obs      =     28091
Group variable (i): idcode                      Number of groups   =      4697

R-sq:  within  = 0.1715                         Obs per group: min =         1
       between = 0.4784                                        avg =       6.0
       overall = 0.3708                                        max =        15

Random effects u_i ~ Gaussian                   Wald chi2(10)      =   9244.87
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
     ln_wage |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       grade |   .0646499   .0017811    36.30   0.000     .0611589    .0681408
         age |    .036806   .0031195    11.80   0.000     .0306918    .0429201
        age2 |  -.0007133     .00005   -14.27   0.000    -.0008113   -.0006153
     ttl_exp |   .0290207   .0024219    11.98   0.000     .0242737    .0337676
    ttl_exp2 |   .0003049   .0001162     2.62   0.009      .000077    .0005327
      tenure |    .039252   .0017555    22.36   0.000     .0358114    .0426927
     tenure2 |  -.0020035   .0001193   -16.80   0.000    -.0022373   -.0017697
       black |  -.0530532   .0099924    -5.31   0.000    -.0726379   -.0334685
    not_smsa |  -.1308263   .0071751   -18.23   0.000    -.1448891   -.1167634
       south |  -.0868927   .0073031   -11.90   0.000    -.1012066   -.0725788
       _cons |   .2387209   .0494688     4.83   0.000     .1417639     .335678
-------------+----------------------------------------------------------------
     sigma_u |  .25790313
     sigma_e |  .29069544
         rho |  .44043812   (fraction of variance due to u_i)
------------------------------------------------------------------------------

. estimates store random_effects
```

The random-effects estimation.

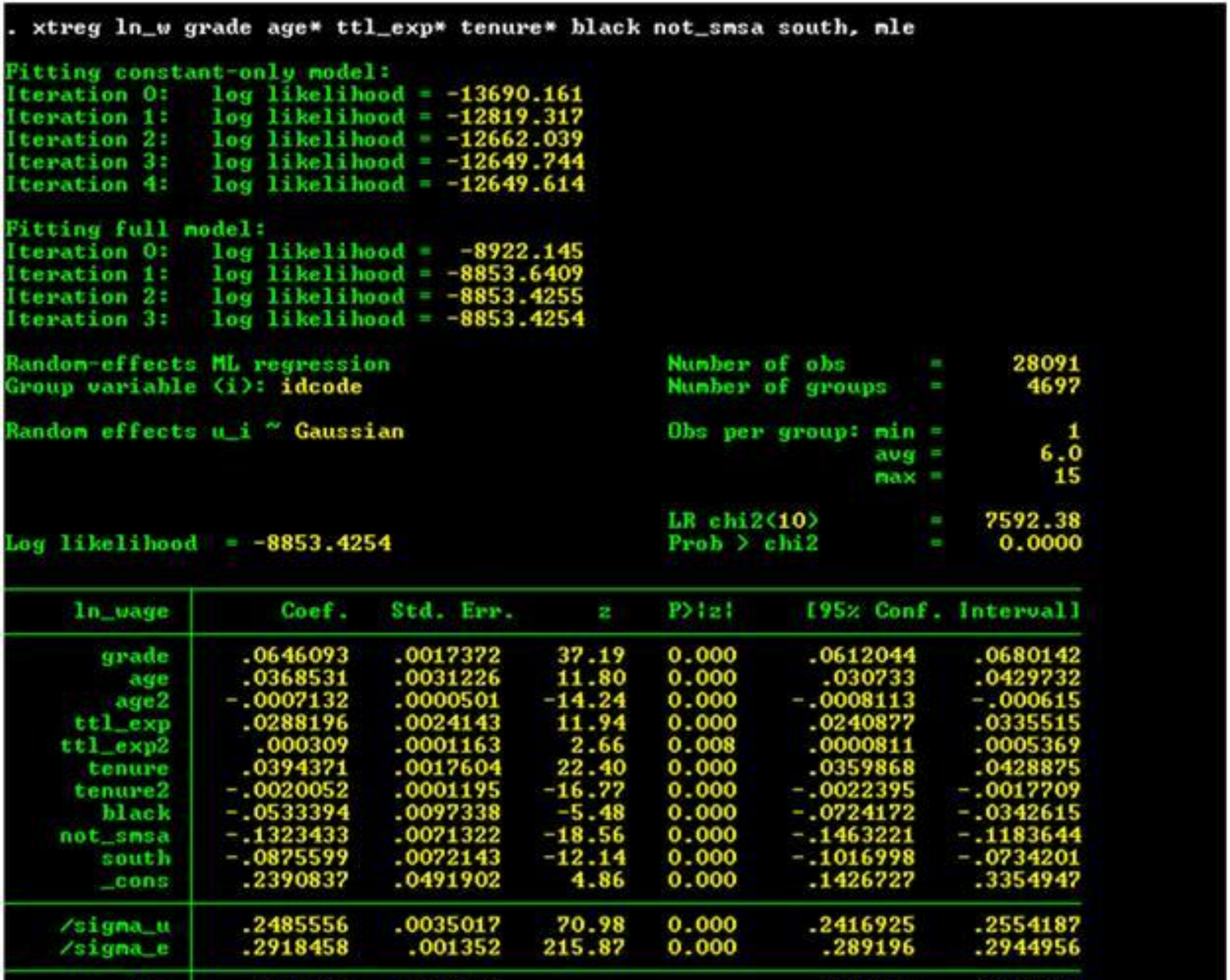
Notice that three R-squared values are reported in Figure 5. Also, wages reach a peak when the woman is $-\frac{0.036806}{2(-0.0007133)}=25.7998$ years old and after 9.795857 years on the job. The interpretation of the other variables demands a bit of algebra. For instance, the fact that black is a dummy variable affects our interpretation; when an individual is a black, her wage level is: $\ln w_{B} = \beta_{0} + \beta_{1} + \cdots$. When she is nonblack, her wage level is $\ln w_{NB} = \beta_{0} + \cdots$. Thus, we have: $\ln w_{B} - \ln w_{NB} = \beta_{1}$ or $\frac{w_B}{w_{NB}}=e^{\beta_1}=e^{-0.0530532}=0.94833.$ Thus, the wage level of a black is, everything else held constant, 94.8 percent of the wage level of a nonblack.

If we assume that grade is a continuous variable (it really is not), we have the following interpretation of the parameter: $\ln w = \beta_{0} + \beta_{1}\, grade + \cdots$ implies that $\frac{1}{w}\frac{\partial w}{\partial grade}=\beta_1$. Thus, in our case a increase of 1 year of schooling causes wages to increase by 6.46 percent.

We can compare the results of using the re option with using the mle option (which directs *Stata* to use maximum likelihood techniques to estimate the parameters of the

system. The mle parameter estimates, shown in Figure 6, are the same as those generated using the re command. However, the estimates of the standard errors (and, thus, the z-values) are different.

Figure 2.24.

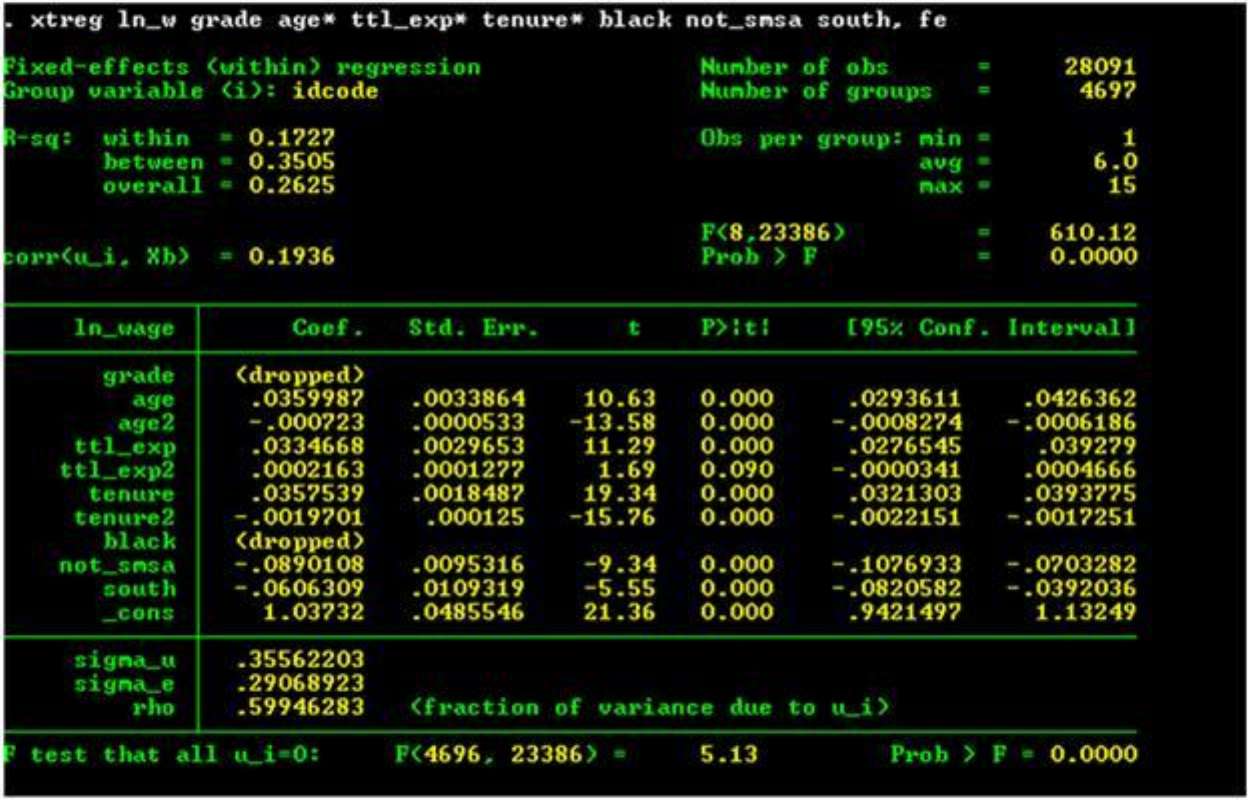


```
. xtreg ln_w grade age* ttl_exp* tenure* black not_smsa south, mle

Fitting constant-only model:
Iteration 0:   log likelihood = -13690.161
Iteration 1:   log likelihood = -12819.317
Iteration 2:   log likelihood = -12662.039
Iteration 3:   log likelihood = -12649.744
Iteration 4:   log likelihood = -12649.614

Fitting full model:
Iteration 0:   log likelihood =  -8922.145
Iteration 1:   log likelihood = -8853.6409
Iteration 2:   log likelihood = -8853.4255
Iteration 3:   log likelihood = -8853.4254

Random-effects ML regression                    Number of obs      =     28091
Group variable (i): idcode                      Number of groups   =      4697

Random effects u_i ~ Gaussian                   Obs per group: min =         1
                                                               avg =       6.0
                                                               max =        15

                                                LR chi2(10)        =   7592.38
Log likelihood  = -8853.4254                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
     ln_wage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       grade |   .0646093   .0017372    37.19   0.000     .0612044    .0680142
         age |   .0368531   .0031226    11.80   0.000      .030733    .0429732
        age2 |  -.0007132   .0000501   -14.24   0.000    -.0008113    -.000615
     ttl_exp |   .0288196   .0024143    11.94   0.000     .0240877    .0335515
    ttl_exp2 |    .000309   .0001163     2.66   0.008     .0000811    .0005369
      tenure |   .0394371   .0017604    22.40   0.000     .0359868    .0428875
     tenure2 |  -.0020052   .0001195   -16.77   0.000    -.0022395   -.0017709
       black |  -.0533394   .0097338    -5.48   0.000    -.0724172   -.0342615
    not_smsa |  -.1323433   .0071322   -18.56   0.000    -.1463221   -.1183644
       south |  -.0875599   .0072143   -12.14   0.000    -.1016998   -.0734201
       _cons |   .2390837   .0491902     4.86   0.000     .1426727    .3354947
-------------+----------------------------------------------------------------
    /sigma_u |   .2485556   .0035017    70.98   0.000     .2416925    .2554187
    /sigma_e |   .2918458    .001352   215.87   0.000      .289196    .2944956
-------------+----------------------------------------------------------------
```

The maximum likelihood estimation.

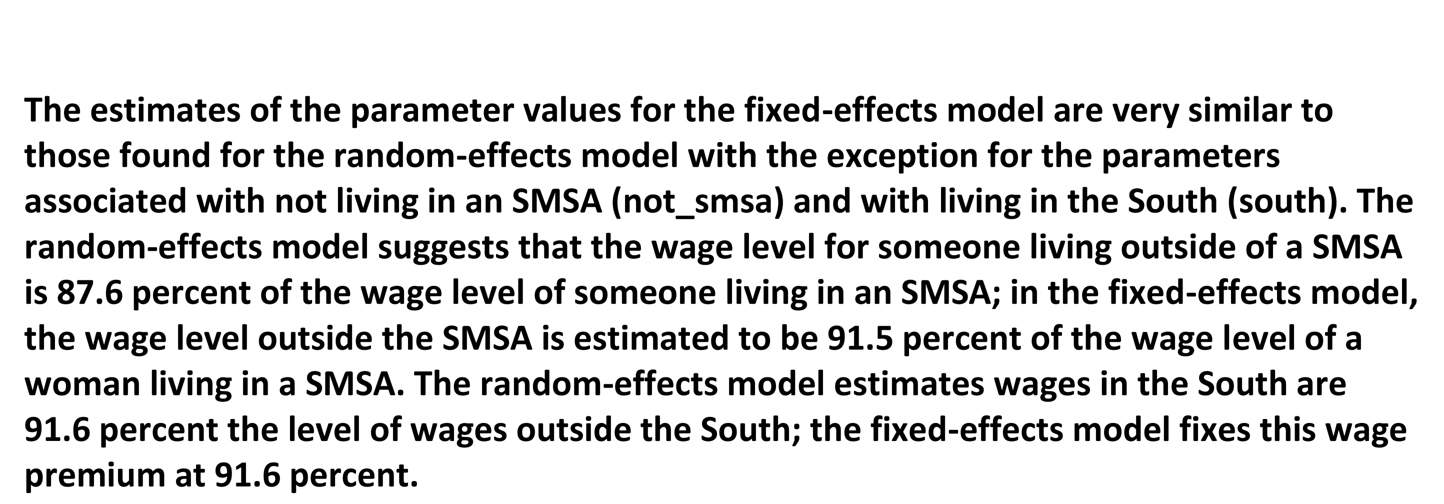
The estimation of the fixed-effects model is straightforward and is shown in Figure 7. The command is the same as in the random-effects model but with the re replaced by fe. Notice from the results that the variables grade and black are dropped from the estimation results. They are dropped because the amount of schooling and race of an individual is fixed over all observations. These two variables, thus, are perfectly correlated with the dummy variables that hold constant the individual level characteristics. The effects of education and race differences are absorbed into the residual.

Figure 2.25.



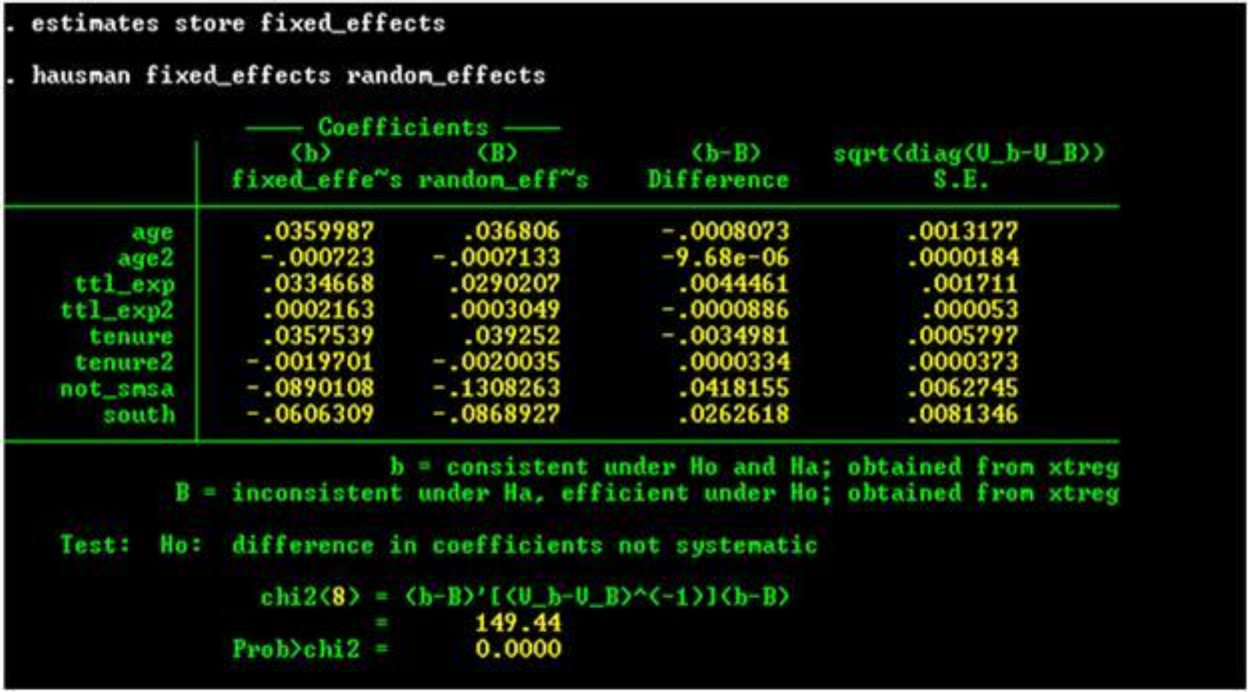
```
. xtreg ln_w grade age* ttl_exp* tenure* black not_smsa south, fe

Fixed-effects (within) regression               Number of obs      =     28091
Group variable (i): idcode                      Number of groups   =      4697

R-sq:  within  = 0.1727                         Obs per group: min =         1
       between = 0.3505                                        avg =       6.0
       overall = 0.2625                                        max =        15

                                                F(8,23386)         =    610.12
corr(u_i, Xb)  = 0.1936                         Prob > F           =    0.0000

------------------------------------------------------------------------------
     ln_wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       grade |  (dropped)
         age |   .0359987   .0033864    10.63   0.000     .0293611    .0426362
        age2 |   -.000723   .0000533   -13.58   0.000    -.0008274   -.0006186
     ttl_exp |   .0334668   .0029653    11.29   0.000     .0276545     .039279
    ttl_exp2 |   .0002163   .0001277     1.69   0.090    -.0000341    .0004666
      tenure |   .0357539   .0018487    19.34   0.000     .0321303    .0393775
     tenure2 |  -.0019701    .000125   -15.76   0.000    -.0022151   -.0017251
       black |  (dropped)
    not_smsa |  -.0890108   .0095316    -9.34   0.000    -.1076933   -.0703282
       south |  -.0606309   .0109319    -5.55   0.000    -.0820582   -.0392036
       _cons |    1.03732   .0485546    21.36   0.000     .9421497     1.13249
-------------+----------------------------------------------------------------
     sigma_u |  .35562203
     sigma_e |  .29068923
         rho |  .59946283   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(4696, 23386) =     5.13         Prob > F = 0.0000
```

The fixed-effects estimation.



The estimates of the parameter values for the fixed-effects model are very similar to those found for the random-effects model with the exception for the parameters associated with not living in an SMSA (not_smsa) and with living in the South (south). The random-effects model suggests that the wage level for someone living outside of a SMSA is 87.6 percent of the wage level of someone living in an SMSA; in the fixed-effects model, the wage level outside the SMSA is estimated to be 91.5 percent of the wage level of a woman living in a SMSA. The random-effects model estimates wages in the South are 91.6 percent the level of wages outside the South; the fixed-effects model fixes this wage premium at 91.6 percent.

Figure 2.26.



```
. estimates store fixed_effects

. hausman fixed_effects random_effects

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             | fixed_effe~s random_eff~s    Difference          S.E.
-------------+----------------------------------------------------------------
         age |    .0359987      .036806       -.0008073        .0013177
        age2 |    -.000723    -.0007133       -9.68e-06        .0000184
     ttl_exp |    .0334668     .0290207        .0044461         .001711
    ttl_exp2 |    .0002163     .0003049       -.0000886         .000053
      tenure |    .0357539      .039252       -.0034981        .0005797
     tenure2 |   -.0019701    -.0020035        .0000334        .0000373
    not_smsa |   -.0890108    -.1308263        .0418155        .0062745
       south |   -.0606309    -.0868927        .0262618        .0081346
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(8) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =      149.44
                Prob>chi2 =      0.0000
```

The Hausman test results.

The final issue we discuss in this example is the Hausman specification test. If the model is correctly specified and if *v_i* is uncorrelated with the explanatory variables, then the parameter estimates in the two models should not be statistically different. As shown in Figure 8, we first must same the results of the fixed-effects estimation using the command estimates store fixed_effects. The null hypothesis is that the the difference in that parameter estimates is not systematic. The appropriate test statistic is the $\chi^2(8)$, where the degrees of freedom are equal to the number of parameters in the model (8). The chi-squared statistic of 149.44 is greater than the critical value and we must reject the null hypothesis. The *Stata* offers this interpretation of this result:

> ***What does this mean? We have an unpleasant choice: we can admit that our model is misspecified—that we have not parameterized it correctly—or we can hold to our specification***
>
> ***being correct, in which case the observed differences must be due to the zero-correlation of and the assumption. [StataCorp: 202]***

Exercises

Exercise 2.3.1.

Estimation of a Labor Supply Function. An important issue in labor economics is the responsiveness of the number of hours worked to wages. Because labor supply curves can, in theory, be backward-bending, the sign and size of the impact of wages on the amount of labor supplied is an empirical issue. In this project you are to estimate the demand for labor curve for a cross-section of adult males.

(2.50)

The model to be estimated is:

$$y_{it} = \beta_0 + \beta_1 h_{it} + \beta_2 Age_{it} + \beta_3 Age^2_{it} + \beta_4 NC_{it} + \beta_5 HI_{it} + \varepsilon_{it}$$

where:

y_{it} = natural logarithm of individual *i*'s wage rate in year *t*,

h_{it} = natural logarithm of total number of hours worked by individual *i* in year *t*,

Age_{it} = age of individual *i* in year *t*,

NC_{it} = number of children of individual *i* in year *t*, and

HI_{it} = an dummy variable equal to 1 if individual *i* in year *t* has bad health and 0 otherwise.

The data are from Ziliak, James P. (1997) "Efficient Estimation with Panel Data When Instruments Are Predetermined: An Empirical Comparison of Moment-Condition Estimators," *Journal of Business & Economic Statistics* 15(4): 419-431. Ziliak (p. 423) describes his data as follows:

> ***The data used to estimate the life-cycle labor-supply parameters come from Waves XII-XXI (calendar years 1978-1987) of the PSID. The sample is selected on many dimensions and is similar to other research studying life-cycle models of labor supply. The sample is restricted to continuously married, continuously working, prime-age men aged 22-51 in 1978 from the Survey Research Center random subsample of the PSID. In addition the individual must either be paid an hourly wage rate or must be salaried, and he cannot be a piece-rate worker or self-employed. This selection process resulted in a balanced panel of 532 men over 10 years or 5,320 observations. The real wage rate, wit,. is the hourly wage reported by the panel participant rather than the average wage (annual earnings over annual hours) to minimize division bias (Borjas 1981).***

The data are available in the any of the three files , , and .

1. **Provide scatter plots among the dependent variable (Natural logarithm of hours) against each of the explanatory variables Natural logarithm of real wages, Age, Number of children, and Health. (Label these Figures 1 to 4.)**

2. **Present a table of the summary statistics for all of the variables in this data set (except *ID* and *Year*).**

3. **Provide a histogram of each of the following variables: Natural logarithm of hours, Natural logarithm of real wages, Age, and Number of children. (Label these Figures 5 to 8).**

4. **Estimate Equation (1) using (1) OLS (sometimes called a "pooled model"), (2) a "between" model (where the observations in the regression are the averages over the 10 years of each variable for each individual, (3) a fixed effects model, (4) a MLE random effects model and (5) a GLS random effects model. Present the results of your estimations in a single table and offer an interpretation for each parameter you estimate. Use Table 1 as shown below as a template for the table to present your results.**

	(1) Pooled	(2) Between	(3) Fixed Effects	(4) MLE Random Effects	(5) GLS Random Effects
Natural logarithm of real wages					
	( )	( )	( )	( )	( )
Age					
	( )	( )	( )	( )	( )
Age^2					
	( )	( )	( )	( )	( )
Number of children					
	( )	( )	( )	( )	( )
Health indicator					
	( )	( )	( )	( )	( )

Intercept					
	( )	( )	( )	( )	( )
R^2				—	—
σ_μ	—	—			
σ_ε	—	—			
Sample size					

Table 2.13. Hours and wages: Summary of linear panel model estimations (Dependent variable is the natural logarithm of total hours worked in a year; the observations consist of 532 adult males over the 10 year period 1978-1987).

Exercise 2.3.2.

The Effectiveness of Advertising Bans on Smoking. Anti-smoking activists often push for a total ban on cigarette advertisements. Indeed, one of the basic assumptions of the groups pushing the 1996 proposed settlement with the tobacco companies is that the amount of tobacco consumed is positively affected by the amount of tobacco advertising. There are

two mechanisms that might underlie such a relationship. The first mechanism suggests that the advertising increases the amount of cigarettes smoked by *current* smokers. Many economists doubt that the tobacco advertising increases the consumption of current smokers, arguing that the total consumption of cigarettes is unresponsive to advertisement. Instead, they argue that advertising is an effort by cigarette companies to affect the brand of cigarettes that current smokers consume. The second mechanism suggests that advertising is an effort by cigarette companies to induce non-smokers (especially children) to try cigarettes. The main reason that cigarette companies want non-smokers to try smoking, so the argument goes, is that some percentage of non-smokers who try cigarettes will become addicted and will form the future demand for cigarettes.

The effect of a total ban on advertising would be completely different if cigarette companies advertise with the hope of increasing the number of people addicted to cigarettes. In particular, the ban should have a small or negligible effect on current cigarette demand. Instead, the cigarette companies would face a steadily decreasing demand for their product. Such a decrease in demand would reduce future profits for these companies. If future profits fell enough, some of the companies might be forced out

of business. Clearly, it is this result that anti-smoking activists have in mind with their proposals to ban cigarette advertisements.

Finally, if advertising only induces current smokers to increase the number of cigarettes they consume, then the total ban on advertising should cause a one-time reduction in cigarette consumption that will reduce the profits of cigarette companies. However, which of these three mechanisms (if any) is correct is an empirical question.

Six European countries adopted a complete ban on cigarette advertising in the period after 1970. It this project we use annual data on smoking consumption in 22 developed countries for the 27 years between 1964 and 1990 to test the effect of a complete smoking ban on cigarette demand (giving us 594 observations). Moreover, since we have no *a priori* reason to choose one model specification over another, we check the stability of the estimated impact of an advertising ban on cigarette demand under several alternative model specifications.

We estimate three types of specifications of the model — the linear model, the log-linear model, and the log-log model. In general whether one uses a variable or the logarithm of the variable is the main difference in these three specifications. The linear model does not transform either the dependent or the independent variables. A variation on the linear models allows the use of the square and product of some of the independent

variables in order to take care of any non-linearity in the data. The log-linear model takes the same form as the linear model except that the dependent variable is the logarithm of variable under study. Finally, in the log-log model both the dependent and independent variables are, if possible, in logarithm form.

For example, for this problem the dependent variable in any of these specifications is either the per capita consumption of tobacco or the logarithm of the per capita consumption of tobacco. The dependent variables might include (1) the real price of tobacco in each country for each year, (2) a measure of the per capita income level of the country for each year, (3) the unemployment rate of the country for each year, (4) a measure of the age distribution of the population to measure smoking intensity by age, (5) a trend variable to account for the rising awareness of the health costs of smoking, (6) a dummy variable equal to one for years that a country has a complete ban on cigarette advertising, and (7) a set of 21 dummy variables identifying the country. Let *T it* be the measure of per capita cigarette consumption in country *i* for year *t*; *P it*, the price of tobacco; *I it*, the measure of per capita income level; *U it*, country *i*'s unemployment rate in year *t*; *A it*, country *i*'s age distribution in year *t*; *Year*, a trend variable; *B it*, the dummy variable for the ban; and *C i*, the dummy variable for country *i*.

Examples of the three models are:

1. **Linear:** $T_{it} = \beta_0 + \beta_1 P_{it} + \beta_2 I_{it} + \beta U_{it} + \beta_4 A_{it} + \beta_5 Year_t + \beta_6 B_{it} + \varepsilon_{it}$
2. **Log-Linear:** $\ln(T_{it}) = \beta_0 + \beta_1 P_{it} + \beta_2 I_{it} + \beta U_{it} + \beta_4 A_{it} + \beta_5 Year_t + \beta_6 B_{it} + \varepsilon_{it}$
3. **Log-Log:** $\ln(T_{it}) = \beta_0 + \beta_1 \ln(P_{it}) + \beta_2 \ln(I_{it}) + \beta U_{it} + \beta_4 A_{it} + \beta_5 Year_t + \beta_6 B_{it} + \varepsilon_{it}$

In models (1) and (2) it is possible to include additional explanatory variables that are the square of some of the currently included explanatory variables. In all three models it is possible to include as explanatory variables the product of the ban dummy and any of the currently included explanatory variables. Finally, in equation (2) we cannot take the logarithm of the unemployment rate because the data we have report zero levels of unemployment.

The data you will use in this project are in the *MS Excel* file Smkdata.xls. The variables included in the file are as follows:

Column	**Variable**	**Definition**
A	**Country**	**Name of country**

B	Country ID	Integar from 1 to 22, each designating a country
C	Year	Year of observation (1964, …, 1990)
D	Tobacco	Total grams of tobacco sold per individual 15 years or older
E	Price	Real price of 20 grams of tobacco in 1990 US cents (= Nominal price per E 20 grams of tobacco divided by the Gross Domestic Price deflator)
F	Consump	Per capita private final consumption expenditures in 1990 US dollars
G	Unemp	Number of unemployed persons per 1000 members of the workforce
H	AgeDist	Age distribution. This variable attempts to measure the differences in intensity of smoking as a function of age. It is equal to the relative consumption rate of tobacco in the UK observed between 1966 and 1981 by age group times the percentage of the population in the country in that age group.
I	Ban	Dummy variable equal to 1 if the country has a complete ban on tobacco advertising. The six countries in the sample with a complete

		ban and the first year of the ban are: Iceland (1972), Norway (1976), Finland (1979), Portugal (1984), Italy (1984), and Canada (1989).
J	BanTime	The number of years since the ban was put in place (if ban went into effect in 1972, then years 1964-1972 are equal to 0, year 1973 equals 1, year 1974 equals 2, etc.)

Table 2.14. Definition of the cigarette consumption data set.

(a) How do these variables match the ones suggested in the discussion of equations (1), (2), and (3)?

(b) Estimate the fixed effects models of the following versions of equations (1), (2), and (3):

1. **Equations (1), (2), and (3) as specified above.**
2. **Equations (1) and (2) with squared terms for the price, income, unemployment rate, and the age distribution included. This regression is designed to test for non-linearity.**

3. **Equations (1) and (2) with the squared terms mentioned in 2 that are statistically significant plus the following new variables: Ban*Time, Ban*Price, and Ban*Consump. (You must create these variables) This regression allows for an effect of the Ban on the slopes of the other explanatory variables.**

4. **Equation (3) with the following new variables: Ban*Log(Time), Ban*Log(Price), and Ban*Log(Consump).**

5. **Equations (1), (2), and (3) as estimated in 3 and 4 with a variable that counts the number of years that a total ban has been in effect (BanTime) and its square (BanTime^2). This regression allows for a changing impact of a ban the longer it is in effect.**

Report the results of your regressions in a table that allows you to comment on the stability of your estimation results over specifications.

(c) Do these results support any of the theories suggested above?

(d) What, if any, policy conclusions would you make given your estimations?

(e) Assume for the moment that you "believe" your results you got in (5). Sketch out a strategy you would follow to forecast the impact of a ban in a country that does not currently have a ban.

Note: The data in this problem are from Stewart, Michael J. (1993) "The Effect on Tobacco Consumption of Advertising Bans in OECD Countries," *International Journal of Advertising* 12(2): 155-180. The data set can be downloaded from the author's website.

Bibliography

Cameron, A. Colin and Pravin K. Trivedi (2005). *Microeconometrics: Methods and Applications* (New York: Cambridge University Press).

Greene, W. H. (2003). *Econometric Analysis*, 5th edition (Upper Saddle River, NJ: Prentice-Hall).

Hsiao, Cheng (2003). *Analysis of Panel Data*, 2nd Edition (New York: Cambridge University Press).

StataCorp (2003). *Stata Statistical Software: Release8.0* (College Station, TX: Stata Corporation).

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data* (Cambridge, MA: MIT Press).

2.4. Sample selectivity bias[*]

Sample Selection Bias

Introduction

These notes discuss how to handle one of the more common problems that arise in economic analyses—sample selection bias. Essentially, sample selection bias can arise whenever some potential observations cannot be observed. For instance, the students enrolled in an intermediate microeconomics course are not a random sample of all undergraduates. Students self-select when they enroll in any class or choose a major. While we do not know all of the reasons for this self-selection, we suspect that students choosing to take advanced economics courses have more quantitative skills than students choosing courses in the humanities. Since we do not observe the grades that students who did not enroll in the intermediate microeconomics class would have made had they enrolled, we can never observe the grades that they would have made. Under certain circumstances the omission of potential members of a sample will cause ordinary least squares (OLS) to give biased estimates of the parameters of a model.

In the 1970s James Heckman developed techniques that will correct the bias introduced by sample selection bias. Since then, most econometric computer programs include a command that automatically used Heckman’s method. However, blind use of these commands can lead to errors that would be avoided by a better understanding of his correction technique. This module is intended to provide this understanding.

In the first section I discuss the sources of sample selection bias by examining the basic economic model used to understand the problem. In the second section I present the estimation strategy first developed by Heckman. In the third section I discuss how to estimate the Heckman model in *Stata*. In the final section I examine an extended example of the technique. An exercise is included at the end of the discussion.

The model

Assume that there is an unobserved latent variable, $y_i{}^*$, and an unobserved latent index, $d_i{}^*$, such that:

(2.51)

$$y_i^* \ = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i \text{where} i = 1, \ \ldots \ , N;$$

(2.52)

$$d_i^* \ = \mathbf{z}'_i \boldsymbol{\gamma} + \nu_i \text{where} i = 1, \ \ldots \ , N;$$

(2.53)

$$d_i = \begin{cases} 1 \text{ if } d_i^* \ > 0 \\ 0 \text{ if } d_i^* \ \le 0 \end{cases}; \text{ and}$$

(2.54) ***y** $_i$ **= y** $_i$ *** d** $_i$ **.***

The matrix notation above means (1) that

1.

$$\mathbf{x}'_i \boldsymbol{\beta} = \begin{bmatrix} 1 \\ x_{1i} \\ \vdots \\ x_{Ki} \end{bmatrix}' \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{bmatrix} = [1 \quad x_{1i} \quad \cdots \quad x_{Ki}] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{bmatrix} = \beta_0 + \beta_1 x_{1i} + \ \cdots \ + \beta_K x_{Ki}$$

2. $$\mathbf{z}'_i\boldsymbol{\gamma} = \begin{bmatrix} 1 & z_{1i} & \cdots & z_{Li} \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_L \end{bmatrix} = \gamma_0 + \sum_{j=1}^{L} \gamma_j z_{ji}.$$

Substituting (1), (2) and (3) into (4) gives:

(2.55)

$$y_i = \begin{cases} \mathbf{x}'_i\boldsymbol{\beta} + \varepsilon_i & \text{if } \mathbf{z}'_i\boldsymbol{\gamma} + \nu_i > 0 \\ 0 & \text{if } \mathbf{z}'_i\boldsymbol{\gamma} + \nu_i \le 0 \end{cases}.$$

Note that *N* is the total sample size and *n* is the number of observations for which $d_i = 1$.

Since $y_i{}^*$ is not observed for (*N* − *n*), the question becomes why are these observations missing. A concrete example of such a model is a model of female wage determination. Equation (1) would model the wage rate earned by women in the labor force and Equation (2) would model the decision by a female to enter the labor force. In this case, y_i, the wage rate woman *i* receives, is a function of the variables in $\mathbf{x}_i;$ however, women not in the labor force are not included in the sample. If these missing observations are drawn randomly from the population, there is no need for concern. Selectivity bias arises

if the ($N - n$) omitted observations have unobserved characteristics that affect the likelihood that $d_i = 1$ and are correlated with the wage the woman would receive had she entered the labor force. For instance, a mentally unstable female is likely to earn relatively low wages and might be more unlikely to enter the labor force. In this case, the error terms, ε_i and v_i would be independent and identically distributed $N(0, \Sigma)$, where

(2.56)

$$\Sigma = \begin{bmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon \upsilon} \\ \sigma_{\upsilon \varepsilon} & \sigma_\upsilon^2 \end{bmatrix}$$

and (ε_i, ν_i) are independent of z_i. The selectivity bias arises because $\sigma_{\varepsilon v} \neq 0$. In effect the residual ε_i includes the same unobserved characteristics as does the residual v_i causing the two error terms to be correlated. OLS estimation of equation (1) would have a missing variable—the bias created by the missing observations (due to wage data not being available for women not in the work force). As in other cases of omitted variables, the estimates of the parameters of the model, $\hat{\beta}$, would be biased. Heckman (1979) notes in his seminal article on selectivity bias:

> ***One can also show that the least squares estimator of the population variance is downward biased. Second, a symptom of selection bias is that variables that do not belong in the true structural equation (variables in not in may appear to be statistically significant determinants of when regressions are fit on selected samples. Third, the model just outlined contains a variety of previous models as special cases. ...For a more complete development of the relationship between the model developed here and previous models for limited dependent variables, censored samples and truncated samples, see Heckman (1976). Fourth, multivariate extensions of the preceding analysis, while mathematically straightforward, are of consider-able substantive interest. One example is offered. Consider migrants choosing among K possible regions of residence. If the self selection rule is to choose to migrate to that region with the highest income, both the self selection rule and the subsample regression functions can be simply characterized by a direct extension of the previous analysis. (Notation has been altered to match the notation used in this module, see Heckman, 1979: 155)***

Estimation Strategy

Heckman (1979) suggests a two-step estimation strategy. In the first step a probit estimate of equation (2) is used to construct a variable that measures the bias. This variable is known as the “inverse Mills ratio.” Heckman and others demonstrate that

(2.57)

$$E[\varepsilon_i | \mathbf{z}_i, d_i = 1] = \frac{\sigma_{\varepsilon\nu}}{\sigma_\nu^2} \left\{ \frac{\phi(\mathbf{z}_i{}'\boldsymbol{\gamma})}{\Phi(\mathbf{z}_i{}'\boldsymbol{\gamma})} \right\},$$

where $\phi(\mathbf{z}_i{}'\boldsymbol{\gamma})$ and $\Phi(\mathbf{z}_i{}'\boldsymbol{\gamma})$ are the probability density function and the cumulative distribution functions, respectively, evaluated at $\mathbf{z}_i{}'\boldsymbol{\gamma}$. [24] The ratio in the brackets in equation (7) is known as the *inverse Mills ratio*. We will use an estimate of the inverse Mills ratio in the estimation of equation (5) to measure the sample selectivity bias.

The Heckman two-step estimator is relatively easy to implement. In the first step you use a maximum likelihood probit regression on the whole sample to calculate $\hat{\boldsymbol{\gamma}}$ from equation (2). You then use $\hat{\boldsymbol{\gamma}}$ to estimate the inverse Mills ratio:

(2.58)

$$\hat{\lambda}_i = \frac{\phi\left(\mathbf{z}_i{}' \hat{\boldsymbol{\nu}}\right)}{\Phi\left(\mathbf{z}_i{}' \hat{\boldsymbol{\nu}}\right)}.$$

In the second step, we estimate:

(2.59)

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \mu \hat{\lambda} + \eta_i$$

using OLS and where $E\left(\hat{\mu}\right) = \frac{\sigma_{\varepsilon\nu}}{\sigma_{\nu}^2}.$ Thus, a t-ratio test of the null hypothesis $H_0 : \mu = 0$ is equivalent to testing the null hypothesis $H_0 : \sigma_{\varepsilon\nu} = 0$ and is a test of existence of the sample selectivity bias.

An alternative approach to the sample selectivity problem is to use a maximum likelihood estimator. Heckman (1974) originally suggested estimating the parameters of the model by maximizing the average log likelihood function:

(2.60)

$$L=\frac{1}{N}\sum_{i=1}^{N}\left\{d_i\ln\left[\int_{-(\mathbf{z}'\boldsymbol{\gamma})}^{\infty}\phi_{\varepsilon\nu}(y_i-\mathbf{x}_i{}'\boldsymbol{\beta})d\nu\right]+(1-d_i)\left[\ln\int_{-(\mathbf{z}'\boldsymbol{\gamma})}^{\infty}\int_{-\infty}^{\infty}\phi_{\varepsilon\nu}(\varepsilon,\nu)d\varepsilon d\nu\right]\right]\right\},$$

where $\varphi_{\varepsilon\nu}$ is the probability density function for the bivariate normal distribution. Fortunately, Stata offers a single command for calculating either the two-step or the maximum likelihood estimators.

Estimation in *Stata*

Estimation of the two versions of the Heckman sample selectivity bias models is straightforward in *Stata*. The command is:

```
.heckman depvar [varlist], select(varlist_s) [twostep]
```

or

```
.heckman depvar [varlist], select(depvar_s = varlist_s) [twostep]
```

The syntax for maximum-likelihood estimates is:

```
.heckman depvar [varlist] [weight] [if exp] [in range], select([depvar_s =] varlist_s [,
offset(varname) noconstant]) [ robust cluster(varname) score(newvarlist|stub*)
nshazard(newvarname) mills(newvarname) offset(varname) noconstant
constraints(numlist) first noskip level(#) iterate(0) nolog maximize_options ]
```

The predict command has these options, among others:

xb, the default, calculates the linear predictions from the underlying regression equation.

ycond calculates the expected value of the dependent variable conditional on the dependent variable being observed/selected; E(y | y observed).

yexpected calculates the expected value of the dependent variable (y*), where that value is taken to be 0 when it is expected to be unobserved; y* = P(y observed) * E(y | y observed). The assumption of 0 is valid for many cases where nonselection implies non-participation (e.g., unobserved wage levels, insurance claims from those who are uninsured, etc.) but may be inappropriate for some problems (e.g., unobserved disease incidence).

Examples of these two commands are:

```
. heckman wage educ age, select(married children educ age)
. predict yhat
```

These two command would use the maximum likelihood estimate of the equations (1) wage as a function of education and age using a selection equation that used marital status, number of children, education level, and age to explain which individuals are participating in the labor force. The help file in *Stata* provides additional information on the structure of the Heckman command and is well worth printing out if you are dealing with a sample selectivity bias problem.

Example 2.4. Example from *Stata*

We will illustrate various issues of selection bias using the data set available from the *Stata* site. Retrieve the data set by entering:

```
. use http://www.stata-press.com/data/imeus/womenwk, clear
```

This data set has 2,000 observations of 15 variables. We can use the describe command (.describe) to get a brief description of the data set:

obs: 2,000				

vars: 15	9 Nov 2004 20:23			
size: 142,000	(86.5% of memory free)			
Variable Name	Storage Type	Display Format	Value Label	Variable Label
c1	double	%10.0g		
c2	double	%10.0g		
u	double	%10.0g		
v	(7,2)	%10.0g		
country	float	%9.0g		
age	int	%8.0g		
education	int	%8.0g		
married	byte	%8.0g		
children	int	%8.0g		
select	float	%9.0g		

wageful	float	%9.0g		
wage	float	%9.0g		
lw	float	%9.0g		
work	float	%9.0g		
lwf	float	%9.0g		

Table 2.15. Description of variables included in the data set from http://www.stata-press.com/data/imeus/womenwk.

We are interested in only a subset of these data. Table 2 reports the definitions of variables that are relevant for our analysis. We can get further insight into the data set using the summarize command. Table 3 reports the summary statistics for the data set.

Variable name	Definition

country	County of residence (categorical variable equal to 0, 1, ..., 9)
age	Age of the woman
education	Number of years of education of the woman
married	Dummy variable equal to 1 if the woman is married and 0 otherwise
children	Number of children that the woman has in their household
wage	Hourly wage rate of the woman
lw	Natural logarithm of hourly wage rate
work	Dummy variable equal to 1 if the individual is in the workforce and 0 otherwise

Table 2.16. Definition of the relevant variables in the data set.

Variable	Obs	Mean	Std. Dev	Min	Max
Age	2000	36.208	8.28656	20	59
education	2000	13.084	3.045912	10	20

married	2000	.6705	.4701492	0	1
children	2000	1.6445	1.398963	0	5
wage	1343	23.69217	6.305374	5.88497	45.80979
lw	1343	3.126703	.2865111	1.772402	3.824498
work	2000	.6715	.4697852	0	1

Table 2.17. Summary statistics of the relevant variables in the data set (using the command: .summarize age education married children wage lw work).

We are interested in modeling two things: (1) the decision of the woman to enter the labor force and (2) determinants of the female wage rate. It might be reasonable to assume that the decision to enter the labor force by a woman is a function of age, marital status, the number of children, and her level of education. Also, the wage rate a woman earns should be a function of her age and education.

The decision to enter the labor force

We can use a probit regression to model the decision of a woman to enter the labor force. The results of this estimation are reported in Table 4. However, we can use the

predict command to produce some results that we can use to be sure that we understand what the regression results mean. In particular, type in the following two commands:

```
.predict zbhat, xb
```

```
.predict phat, p
```

These two commands will predict (1) the linear prediction (zbhat) and (2) the predicted probability that the woman will be in the workforce (phat). Table 5 reports the values of these two variables for observations 1 through 10.

. probit work age education married children	
Iteration 0: log likelihood = -1266.2225	
Iteration 4: log likelihood = -1027.0616	
Probit estimates Number of obs = 2000	

LR chi2(4) = 478.32						
Prob > chi2 = 0.0000						
Log likelihood = -1027.0616 Pseudo R2 = 0.1889						
work	Coef.	Std. Err.	z	P>z	[95% Conf. Interval]	
age	.0347211	.0042293	8.21	0.000	.0264318	.0430105
education	.0583645	.0109742	5.32	0.000	.0368555	.0798735
married	.4308575	.074208	5.81	0.000	.2854125	.5763025
children	.4473249	.0287417	15.56	0.000	.3909922	.5036576
_cons	-2.467365	.1925635	-12.81	0.000	-2.844782	-2.089948

Table 2.18. Probit estimation of the decision to enter the labor force.

Observation	zbhat	phat
1	-0.68900	0.24541
2	-0.20290	0.41961
3	-0.48067	0.31538
4	-0.16818	0.43322
5	0.34859	0.63630
6	0.58758	0.72159
7	0.97357	0.83486
8	0.45978	0.67716
9	0.01799	0.50718
10	0.32628	0.62790

Table 2.19. Predicted values of zbhat and phat for observations 1 through 10.

The interpretation of the numbers in Table 5 is straightforward. Consider individual 1. The z-value predicted for this individual is -0.68. Using the standard normal tables reported in Table 11 it is easy to see:

(2.61) $\Phi(z \le -0.69)$ = Pr(Individual 1 is in the labor force)

(2.62)

$$\begin{aligned}\Phi(z \le -0.69) &= 0.5 - \Phi(0 \le z \le 0.69)\\ &\approx 0.5 - 0.2549\\ &\approx 0.2451.\end{aligned}$$

The difference between this number and the value reported for *phat* in Table 5 is due to rounding error.

A little later we will want to calculate the inverse Mills ratio. As noted in (8), the formula for the inverse Mills ratio is:

(2.63)

$$\hat{\lambda}_i = \frac{\phi\left(\mathbf{z}_i{}' \hat{\boldsymbol{\nu}}\right)}{\Phi\left(\mathbf{z}_i{}' \hat{\boldsymbol{\nu}}\right)}.$$

The variable phat is equal to $\Phi\left(\mathbf{z}_i{}'\,\hat{\boldsymbol{\nu}}\right)$. Stata offers an easy way to calculate $\phi\left(\mathbf{z}_i{}'\,\hat{\boldsymbol{\nu}}\right)$ with the function “normden(zbhat)” as follows:

.generate imratio = normden(zbhat)/phat

Table 6 repeats Table 5 with the estimate of the inverse Mills ratio for the first 10 observations.

Observation	zbhat	phat	Inverse Mills Ratio
1	-0.6889973	0.2454125	1.2821240
2	-0.2029016	0.4196060	0.9313837
3	-0.4806706	0.3153753	1.1269680
4	-0.1681804	0.4332207	0.9079438
5	0.3485867	0.6363002	0.5900134
6	0.5875849	0.7215945	0.4652062

7	0.9735670	0.8348642	0.2974918
8	0.4597758	0.6771615	0.5300468
9	0.0179909	0.5071769	0.7864666
10	0.3262833	0.6278950	0.6024283

Table 2.20. Calculation of the inverse Mills ratio for the first 10 observations.

The two Heckman estimates

One of the great advantages of using an econometrics program like *Stata* is that the authors quite often have created a command that does all of the work for the user. In our case, the commands we need to run to generate the maximum likelihood estimate of the Heckman model are:

```
. global wage_eqn wage educ age
. global seleqn married children age education
```

```
. heckman $wage_eqn, select($seleqn)
```

Notice that we have used the global command to create a shortcut for referring to each of the two equations in the estimation. The command for the Heckman two-stage estimate is:

```
.heckman $wage_eqn, select($seleqn) twostage
```

```
.predict mymills, mills
```

(1) Explanatory variable	(2) Maximum likelihood estimate	(3) Heckman two-step	(4) Probit estimate of the selection equation
Wage Equation			
Education	0.9899537	0.9825259	—
	(18.59)	(18.23)	
Age	0.2131294	0.2118695	—

	(10.34)	**(9.61)**	
Intercept	**0.4857752**	**0.7340391**	**—**
	(0.45)	**(0.59)**	
Selection equation			
Married	**0.4451721**	**0.4308575**	**0.4308575**
	(6.61)	**(5.81)**	**(5.81)**
Children	**0.4387068**	**0.4473249**	**0.4473249**
	(15.79)	**(15.56)**	**(15.56)**
Age	**0.0365098**	**0.0347211**	**0.0347211**
	(8.79)	**(8.21)**	**(8.21)**
Education	**0.0557318**	**0.0583645**	**0.0583645**
	(5.19)	**(5.32)**	**(5.32)**
Intercept	**-2.491015**	**-2.467365**	**-2.467365**

	(-13.16)	(-12.81)	(-12.81)
σ	0.7035061	0.67284	—
λ	6.004797	5.9473529	—
(Mills)λ	4.224412	4.001615	—
		(6.60)	
Observations	2000	2000	2000
Number of women not working	657	657	657
Number of women working	1343	1343	1343
Log likelihood	-5178.304	—	-1027.0616
Wald χ^2 (2)	508.44	—	—
Probability > χ^2	0.0000	—	—
Wald χ^2 (4)	—	551.37	—

Probability > χ^2	—	0.0000	—
LR test of independent equations (ρ = 0)			
$\chi^2(1)$	61.20	—	478.32
Probability > χ^2	0.0000	—	0.0000

Table 2.21. Comparison of Heckman Maximum-Likelihood and the Heckman Two-Step Estimates with the Probit Estimates of the Selection Equation.

The second command reports the estimates of the inverse Mills ratio; we have retrieved these values in order to check our earlier calculations. Table 7 reports the results of these two estimations. Column 2 reports the maximum-likelihood estimates; Column 3 reports the Heckman two-step estimates; and Column 3 reports the probit estimate of selection equation as reported in Table 4. The estimates for the two methods are very similar. Of course, the probit estimates in Column 4 exactly match the results reported for the selection equation in Column 3. As a final check, Table 8 reports the values of the inverse Mills ratio reported in Table 6 with the values of the inverse Mills ratio calculated in the

Heckman two-step method. The two estimates are identical except for some rounding errors.

Observation	As calculated from probit estimate	As reported by the Heckman two-step
1	1.2821240	1.2821240
2	0.9313837	0.9313837
3	1.1269680	1.1269680
4	0.9079438	0.9079438
5	0.5900134	0.5900134
6	0.4652062	0.4652061
7	0.2974918	0.2974918
8	0.5300468	0.5300469
9	0.7864666	0.7864666
10	0.6024283	0.6024283

Table 2.22. Inverse Mills Ratio Comparison.

Exercise

Exercise 2.4.1. The supply of married women in the workforce.

We are interested in understanding the decision of married Portugese women to enter the labor force. We have available data from Portugal. The data set is a sample from Portuguese Employment Survey, from the interview year 1991, and has been provided by the Portuguese National Institute of Statistics (INE). The data are in the Excel file Martins. This file is organized in the following way. There are seven columns, corresponding to seven variables, and 2,339 observations.

a) Estimate the following equation using OLS: $Wages = f\left(age, age^2, education\right)$ **using the observations for women actually working.**

b) What is the potential source of selection bias?

c) Estimate a wage equation for the Portuguese data three ways: (1) using OLS, (2) using the Heckman two-step method, and (3) using the ML method. Report all three estimates in a single table. For consistency, we will assume that the appropriate explanatory variables for wages are (1) age, (2) the square of age, and (3) the years of education. Further, assume that women do not enter the labor force because (1) presence of children under the age of 3, (2) presence of children between 3 and 18, (3) husband's wage level, (4) the level of education of the woman, and (5) the age of the woman.

Appendix A.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517

0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545

1.7	**0.4554**	**0.4564**	**0.4573**	**0.4582**	**0.4591**	**0.4599**	**0.4608**	**0.4616**	**0.4625**	**0.4633**
1.8	**0.4641**	**0.4649**	**0.4656**	**0.4664**	**0.4671**	**0.4678**	**0.4686**	**0.4693**	**0.4699**	**0.4706**
1.9	**0.4713**	**0.4719**	**0.4726**	**0.4732**	**0.4738**	**0.4744**	**0.4750**	**0.4756**	**0.4761**	**0.4767**
2.0	**0.4772**	**0.4778**	**0.4783**	**0.4788**	**0.4793**	**0.4798**	**0.4803**	**0.4808**	**0.4812**	**0.4817**
2.1	**0.4821**	**0.4826**	**0.4830**	**0.4834**	**0.4838**	**0.4842**	**0.4846**	**0.4850**	**0.4854**	**0.4857**
2.2	**0.4861**	**0.4864**	**0.4868**	**0.4871**	**0.4875**	**0.4878**	**0.4881**	**0.4884**	**0.4887**	**0.4890**
2.3	**0.4893**	**0.4896**	**0.4898**	**0.4901**	**0.4904**	**0.4906**	**0.4909**	**0.4911**	**0.4913**	**0.4916**
2.4	**0.4918**	**0.4920**	**0.4922**	**0.4925**	**0.4927**	**0.4929**	**0.4931**	**0.4932**	**0.4934**	**0.4936**
2.5	**0.4938**	**0.4940**	**0.4941**	**0.4943**	**0.4945**	**0.4946**	**0.4948**	**0.4949**	**0.4951**	**0.4952**
2.6	**0.4953**	**0.4955**	**0.4956**	**0.4957**	**0.4959**	**0.4960**	**0.4961**	**0.4962**	**0.4963**	**0.4964**
2.7	**0.4965**	**0.4966**	**0.4967**	**0.4968**	**0.4969**	**0.4970**	**0.4971**	**0.4972**	**0.4973**	**0.4974**
2.8	**0.4974**	**0.4975**	**0.4976**	**0.4977**	**0.4977**	**0.4978**	**0.4979**	**0.4979**	**0.4980**	**0.4981**
2.9	**0.4981**	**0.4982**	**0.4982**	**0.4983**	**0.4984**	**0.4984**	**0.4985**	**0.4985**	**0.4986**	**0.4986**

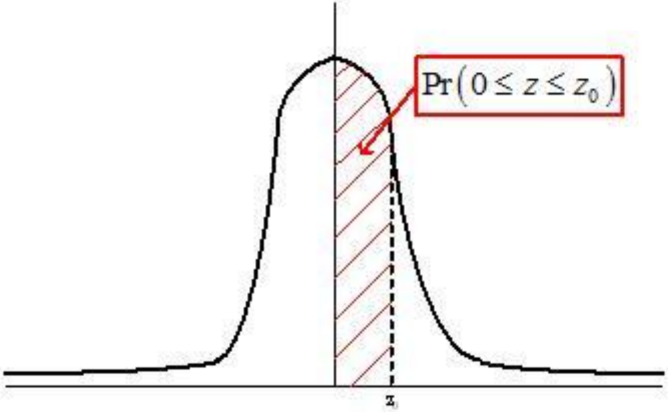
3.0	**0.4987**	**0.4987**	**0.4987**	**0.4988**	**0.4988**	**0.4989**	**0.4989**	**0.4989**	**0.4990**	**0.4990**

Table 2.23. Normal Distribution.

z~N(0, 1).

$$\Phi(z_0)=\int_{-\infty}^{z_0}\phi(z)dz=0.5+\Pr(0\leq z\leq z_0)$$

Figure 2.27. The Normal Distribution



References

Bourguignon, François, Martin Fournier, and Marc Gurgand (2007). Selection Bias Corrections Based on the Multinomial Logit Model: Monte Carlo Comparisons. *Journal of Economic Surveys* 21(1): 174-205.

Chiburis, Richard and Michael Lokshin (2007). Maximum Likelihood and Two-Step Estimation of an Ordered-Probit Selection Model. *The Stata Journal* 7(2): 167-182.

Dahl, G. B. (2002). Mobility and the Returns to Education: Testing a Roy Model with Multiple Markets. *Econometrica* 70(6): 2367-2420.

Dubin, Jeffrey A. and Douglas Rivers (1989). Selection Bias in Linear Regression, Logit and Probit Models. *Sociological Methods and Research* 18(2 & 3): 360-390.

Heckman, James (1974). Shadow Prices, Market Wages and Labor Supply. *Econometrica* 42(4):679-694.

Heckman, James (1976) “The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models,” *The Annals of Economic and Social Measurement* 5: 475-492.

Heckman, James (1979). Sample Selection Bias as a Specification Error. *Econometrica* 47(1): 153-161.

Jimenez, Emanuel and Bernardo Kugler (1987). The Earnings Impact of Training Duration in a Developing Country: An Ordered Probit Model of Colombia's *Servicio Nacional de Aprendizaje* (SENA). *Journal of Human Resources* 22(2): 230-233.

Lee, Lung-Fei (1983). Generalized Econometric Models with Selectivity. *Econometrica* 51(2): 507-512.

McFadden, Daniel L. (1973). Conditional Logit Analysis of Qualitative Choice Behavior. In P. Zarembka *Frontiers in Econometrics* (New York: Academic Press).

Newey, W. K. and Daniel L. McFadden (1994). Large Sample Estimation and Hypothesis Testing. In R. F. Engle and D. L. McFadden (eds.) *Handbook of Econometrics* (Amsterdam: North Holland).

Schmertmann, Carl P. (1994). Selectivity Bias Correction Methods in Polychotomous Sample Selection Models. *Journal of Econometrics* 60(1): 101-132.

Vella, Francis (1998). Estimating Models with Sample Selection Bias: A Survey. *The Journal of Human Resources* 33(1):127-169.

2.5. Endogenous explanatory variables[*]

Endogenous Explanatory Variables

Introduction

One of the most common problems complicating the research of an economist is created by the inclusion of endogenous variables as an explanatory variable. The variable on the left-hand-side of a regression is an endogenous variable; its level is determined by the levels of the explanatory variables—that is, the variables on the right-hand-side of the equation. In OLS we assume that the explanatory variables are independent of the error term. However, if the level of one of these explanatory variables is determined by the levels of the other variables in the model, that explanatory variable actually is an endogenous variable. In a nutshell the problem with having endogenous explanatory

variables is that these endogenous variables cause the error term in the model to be correlated with the explanatory variables thus causing the OLS estimator to be biased. This problem is also known as *simultaneous equation bias* and it is a problem that is subtly different from sample selection bias. See "What is the difference between 'endogeneity' and 'sample selection bias"'?" for an excellent discussion of the difference between these two econometric problems.

In this module we explore both the statistical and algebraic issues raised by the inclusion of endogenous explanatory variables in a model. This introduction is too sketchy to give you a thorough understanding of the many problems raised by simultaneous equation bias. Hopefully, by the time you finish the module along with the problem set, you will have an least an intuitive understanding of the problem and will be able to recognize it when you come across the problem in your own research. If you think the model you are estimating may have simultaneous equation bias, you should seek the advice of an econometrician.

The Statistical Problem

Imagine we know with certainty that the following model fully describes the true state of the supply and demand for wheat. First, the demand for wheat in any year, q_t, is a function of the price of wheat, p_t^w, the income of the average individual, I_t, and the price of corn, p_t^c. Second, in any year the price of wheat is a function of the amount of wheat brought to market, q_t, and a weather index, W_t, that is positively related to the amount of wheat that is harvested. Third, the error terms in the supply and demand functions are due purely to measurement errors—that is, there are no omitted variables in the model. Thus, we have the following two equation model:

(2.64)

Demand:

$$q_t = \alpha_0 + \alpha_1 p_t^w + \alpha_2 I_t + \alpha_3 p_t^c + \varepsilon_t$$

and

Supply:

$$p_t^w = \beta_0 + \beta_1 q_t + \beta_2 W_t + \eta_t.$$

We assume that the error terms each are normally distributed with a mean of zero and a constant variance. Moreover, we assume that the two error terms are independent of each other—that is, we are assuming that:

$$
\begin{aligned}
&\varepsilon_t \sim N\left(0, \sigma_\varepsilon^2\right), \\
&\eta_t \sim N\left(0, \sigma_\eta^2\right), \text{ and} \\
&E(\varepsilon_t \eta_t) = 0.
\end{aligned}
\tag{2.65}
$$

Finally, we assume that income, the price of corn, and the weather index are non-stochastic variables—i.e., these variables are independent of the two error terms. Clearly, the price of wheat and the quantity of wheat are stochastic variables.[25]

What we have here is an ideal model in the sense that we know and can measure all of the variables in the model. The model as written has two *endogenous* variables—q_t and p_t^w —and three exogenous variables— I_t, p_t^c, and W_t. Equations (1) and (2) are known as *structural equations*. What makes this model useful for our purposes is that there is an endogenous explanatory variable in each of the two structural equations.

What we ultimately want to know is if we can use ordinary least squares (OLS) to obtain unbiased estimates of the parameters in Equations (1) and (2). One of the assumptions of OLS is that each of the explanatory variables are independent of the error term, ε_t; if this assumption is violated, OLS will produce biased estimates of the slope parameters. Thus, what we need to do is see if the error term in each equation is independent of the endogenous variable on the right-hand-side of that equation. That is, we want to see if $E(\varepsilon_t p_t^w) = 0$ and $E(\eta_t q_t) = 0.$

It is convenient in answering our question to use the two structural equations to find what are known as the *reduced form equations*—that is, one equation for each endogenous variable in which the endogenous variable is written as a function solely of exogenous variables and error terms. We can find the reduce form equations by solving the structural equations simultaneously for the endogenous variables. Substituting (2) into (1), we get:

$$q_t = \alpha_0 + \alpha_1 (\beta_0 + \beta_1 q_t + \beta_2 W_t + \eta_t) + \alpha_2 I_t + \alpha_3 p_t^c + \varepsilon_t$$

$$q_t = \alpha_0 + \alpha_1 \beta_0 + \alpha_1 \beta_1 q_t + \alpha_1 \beta_2 W_t + \alpha_1 \eta_t + \alpha_2 I_t + \alpha_3 p_t^c + \varepsilon_t$$

$$q_t - \alpha_1 \beta_1 q_t = (\alpha_0 + \alpha_1 \beta_0) + \alpha_1 \beta_2 W_t + \alpha_2 I_t + \alpha_3 p_t^c + (\varepsilon_t + \alpha_1 \eta_t)$$

or

(2.66)

$$q_t = \frac{\alpha_0 + \alpha_1 \beta_0}{1 - \alpha_1 \beta_1} + \frac{\alpha_1 \beta_2}{1 - \alpha_1 \beta_1} W_t + \frac{\alpha_2}{1 - \alpha_1 \beta_1} I_t + \frac{\alpha_3}{1 - \alpha_1 \beta_1} p_t^c + \frac{\varepsilon_t + \alpha_1 \eta_t}{1 - \alpha_1 \beta_1}.$$

Substituting (1) into (2) yields:

$$p_t^w = \beta_0 + \beta_1 (\alpha_0 + \alpha_1 p_t^w + \alpha_2 I_t + \alpha_3 p_t^c + \varepsilon_t) + \beta_2 W_t + \eta_t$$

$$\boldsymbol{p_t^{w} = \beta_0 + \beta_1 \alpha_0 + \alpha_1 \beta_1 p_t^{w} + \alpha_2 \beta_1 I_t + \alpha_3 \beta_1 p_t^{c} + \beta_1 \varepsilon_t + \beta_2 W_t + \eta_t}$$

or

(2.67)

$$p_t^w = \frac{\beta_0 + \beta_1 \alpha_0}{1 - \alpha_1 \beta_1} + \frac{\alpha_2 \beta_1}{1 - \alpha_1 \beta_1} I_t + \frac{\alpha_3 \beta_1}{1 - \alpha_1 \beta_1} p_t^c + \frac{\beta_2}{1 - \alpha_1 \beta_1} W_t + \frac{\beta_1 \varepsilon_t + \eta_t}{1 - \alpha_1 \beta_1}.$$

Equations (4) and (5) are the reduced form equations for this model. We can use them to calculate $E(\varepsilon_t p_t^w) = 0$ **and** $E(\eta_t q_t) = 0.$ **In particular,**

$$E(\varepsilon_t p_t^w) = E\left[\varepsilon_t \left(\frac{\beta_0 + \beta_1 \alpha_0}{1 - \alpha_1 \beta_1} + \frac{\alpha_2 \beta_1}{1 - \alpha_1 \beta_1} I_t + \frac{\alpha_3 \beta_1}{1 - \alpha_1 \beta_1} p_t^c + \frac{\beta_2}{1 - \alpha_1 \beta_1} W_t + \frac{\beta_1 \varepsilon_t + \eta_t}{1 - \alpha_1 \beta_1}\right)\right]$$

$$E(\varepsilon_t p_t^w)=E\left[\varepsilon_t\left(\frac{\beta_0+\beta_1\alpha_0}{1-\alpha_1\beta_1}+\frac{\alpha_2\beta_1}{1-\alpha_1\beta_1}I_t+\frac{\alpha_3\beta_1}{1-\alpha_1\beta_1}p_t^c+\frac{\beta_2}{1-\alpha_1\beta_1}W_t\right)+\varepsilon_t\left(\frac{\beta_1\varepsilon_t+\eta_t}{1-\alpha_1\beta_1}\right)\right]$$

or

(2.68)

$$E(\varepsilon_t p_t^w)=E\left[\varepsilon_t\left(\frac{\beta_0+\beta_1\alpha_0}{1-\alpha_1\beta_1}+\frac{\alpha_2\beta_1}{1-\alpha_1\beta_1}I_t+\frac{\alpha_3\beta_1}{1-\alpha_1\beta_1}p_t^c+\frac{\beta_2}{1-\alpha_1\beta_1}W_t\right)\right]+E\left(\frac{\beta_1\varepsilon_t^2+\eta_t\varepsilon_t}{1-\alpha_1\beta_1}\right).$$

Factoring out the non-stochastic terms from the expected value operators gives:

$$E(\varepsilon_t p_t^w)=\left(\frac{\beta_0+\beta_1\alpha_0}{1-\alpha_1\beta_1}+\frac{\alpha_2\beta_1}{1-\alpha_1\beta_1}I_t+\frac{\alpha_3\beta_1}{1-\alpha_1\beta_1}p_t^c+\frac{\beta_2}{1-\alpha_1\beta_1}W_t\right)E[\varepsilon_t]+\frac{\beta_1E(\varepsilon_t^2)}{1-\alpha_1\beta_1}+\frac{E(\eta_t\varepsilon}{1-\alpha_1}$$

Moreover, by assumption $E(\varepsilon_t)=0,\ E(\eta_t\varepsilon_t)=0,$ **and** $E(\varepsilon_t^2)=\sigma_\varepsilon^2.$ **Thus, we get:**

(2.69)

$$E(\varepsilon_t p_t^w)=\frac{\beta_1\sigma_\varepsilon^2}{1-\alpha_1\beta_1}\neq 0.$$

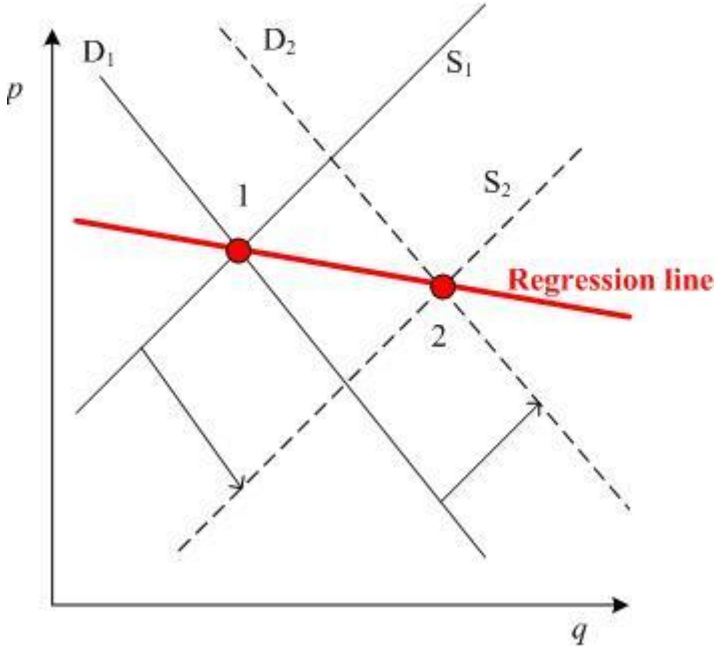
A similar analysis yields:

(2.70)

$$E(\eta_t q_t) = \frac{\alpha_1 \sigma_\eta^2}{1 - \alpha_1 \beta_1} \neq 0.$$

Equations (6) and (7) are what create the endogeneity problem (or *simultaneous equation bias*)—using OLS to estimate the parameters of equations that have an endogenous variable as an explanatory variable yields biased estimates of the unknown parameters. Figure 1 illustrates the endogeneity problem. In this figure we have demand and supply equations that have both risen due to changes in exogenous variables. What the researcher observes are two (red) points: (1) the intersection of the old demand and supply curves and (2) the intersection of the new demand and supply curves.

Figure 2.28.



The simultaneous equation problem.

The thick red line shows the regression that would result from using OLS to estimate either of the two structural equations. As illustrated, an OLS estimate of the slope estimate will be biased. We need to use some other estimation technique than OLS.

Estimation

As noted earlier, the basic problem created by the endogeneity problem is that the endogenous explanatory variable is correlated with the error term. The most logical approach would be to replace this variable with one that is not correlated with the error term but highly correlated with the endogenous variable. Consider the value of the price predicted by the *reduced form* equation (5):

(2.71)

$$\widehat{p}_t^w = \widehat{\gamma}_0 + \widehat{\gamma}_1 I_t + \widehat{\gamma}_2 p_t^c + \widehat{\gamma}_3 W_t$$

where $\widehat{\gamma}_i$ is the OLS estimate of $\gamma_0 = \frac{\beta_0 + \beta_1 \alpha_0}{1 - \alpha_1 \beta_1}$, $\gamma_1 = \frac{\alpha_2 \beta_1}{1 - \alpha_1 \beta_1}$, $\gamma_2 = \frac{\alpha_3 \beta_1}{1 - \alpha_1 \beta_1}$, and $\gamma_3 = \frac{\beta_2}{1 - \alpha_1 \beta_1}$.

Clearly, $\widehat{p}_t^w$ is correlated with p_t^w. It also is true that the covariance between $\widehat{p}_t^w$ and ε_t goes to zero as the sample size increasing. Thus, we can use (8) to construct a variable that will produce a consistent estimator of α_1. It is this conclusion that underlies the strategy of both two-stage least squares (TSLQ) and instrumental variable (IV) estimators.

Two-stages least squares

The easiest way to understand two-stage least squares is to think of the estimation process as being in the following two steps (although the computer programs calculate the estimators in one step):

Stage 1: obtain a OLS predictions for any endogenous variable on the right-hand side of the equation to be estimated using as the explanatory variables all of the exogenous variables in the system.

Stage 2: estimate the parameters of the equation using OLS and replacing the endogenous variable on the right-hand side of the equation by the its predictions as obtained in step 1.

For obvious reasons he TSLS method works best when the full model is specified or when you know and can measure all of the exogenous variables in the system.

Instrumental variables (IV)

While the use of instrumental variable (IV) estimators is appropriate in a large number of situations, the two situations where they are most commonly used are (1) in the presence of endogenous explanatory variables and (2) in cases when errors arise in the measurement of an explanatory variable (or the *errors-in-variables* problem). Since I have already described the endogeneity problem, I now turn to a brief discussion of errors-in-variables.

Consider the following simple model:

$$\text{(2.72)}\ y_i = \beta_1 x_i^* + \varepsilon_i \text{ and } x_i = x_i^* + \mu_i\,.$$

In this model the researcher observes x_i but not the desired x_i^* because of some random measurement error. Using OLS to estimate (9) using the observable x_i instead of the correct x_i^* is equivalent to estimating:

(2.73)

$$y_i = \beta_1 x_i + (\varepsilon_i - \beta_1 \mu_i).$$

The important thing to note in estimating (10) using OLS is that the explanatory variable, x_i, is correlated with the error term, $(\varepsilon_i - \beta_1 \mu_i).$ As was the case with the endogeneity problem, the OLS estimate of β_1 is biased. Murray (2006) summarizes the situation as follows:

> ***In both examples, ordinary least squares estimation is biased because an explanatory variable in the regression is correlated with the error term in the regression. Such a correlation can result from an endogenous explanator, a mismeasured explanator, an omitted explanator, or a lagged dependent variable among the explanators. I call all such explanators "troublesome." Instrumental variable estimation can consistently estimate coefficients when ordinary least squares cannot—that is, the instrumental variable estimate of the coefficient will almost certainly be very close to the coefficient's true value if the sample is sufficiently large—despite troublesome explanators. [Murray (2006a): 112]***

Consider a regression that includes a "troublesome explanator," like x_i^* in (9). Assume that there exists a variable z_i (or set of variables) that (1) is correlated with the "troublesome explanator," (2) is uncorrelated with the error term—like ε_i in (9), and (3) is

not one of the explanatory variables in the equation to be estimated. Greene (1990: 300) offers the following example of such a variable. Self-reported income tends to be a very "noisy" variable because sometimes people forget to report minor sources of income and sometimes they deliberately misreport their income. If the regression you are estimating uses income as explanatory variable of consumption, OLS will yield biased estimates. On the other hand, the number of checks written in a month by the household head might serve as an instrumental variable. Clearly, the number of checks written might well be positively correlated with income and there is no reason to assume that it is correlated with the error term in the consumption equation.[26]

It is usually fairly easy to identify instances when IV estimation methods are appropriate. This is especially true when one of the explanatory variables is possibly an endogenous variable. The real problem arises in finding an instrumental variable or a set of instrumental variables. However, assuming you have one or more instrumental variables, the IV method follows the same steps as described above for TSLS. In the first stage you estimate a regression of the "troublesome variable" as a function of the instruments and the exogenous variables in the equation—i.e., you estimate the reduced form equation. In the second stage you use OLS to estimate the original equation with the value of the "troublesome variable" predicted by the first stage regression substituted for the actual values of the "troublesome variable."

In a sense TSLS is a IV estimation. The exogenous variables not in a particular regression play the role of the instruments. Thus, in the IV estimation of (1), the weather index is the instrument. In the estimation of (2) the price of corn and the income level are the IVs. Thus, in a fully specified model, the exogenous variables excluded from the regression play the role of instrumental variables. In other situations the choice of an appropriate instrument can be very difficult. The selection process demands creativity both in finding the instrument and in defending the choice.

The use either of IV or TSLS comes at a cost. First, the OLS estimators are more precise (i.e., have a smaller standard error) than the TSLS or IV estimators. Second, selecting invalid or weak instruments can create results that are not meaningful. So how does one know if they have chosen a good set of instruments? There is no easy answer to this question. Murray (2006a: 116-117) discusses some possible tests of the validity of an instrumental variable. In the end, however, the "success" of your instrument may depend more on how convincing your justifications are than any statistical test. Some economists, like Steven Levitt, make a living coming up with and justifying the use of some very creative instrumental variables. Murray (2006a) offers a detailed discussion of IV and should be read by any student planning to make use either of TSLS or IV regression estimators.

The identification problem

There is an additional issue that arises with estimating systems of equations—identification. Essentially, identification is an *algebraic* problem. Consider the reduced form equations given earlier in (4) and (5):

$$q_t = \frac{\alpha_0 + \alpha_1 \beta_0}{1 - \alpha_1 \beta_1} + \frac{\alpha_1 \beta_2}{1 - \alpha_1 \beta_1} W_t + \frac{\alpha_2}{1 - \alpha_1 \beta_1} I_t + \frac{\alpha_3}{1 - \alpha_1 \beta_1} p_t^c + \frac{\varepsilon_t + \alpha_1 \eta_t}{1 - \alpha_1 \beta_1}$$

and

$$p_t^w = \frac{\beta_0 + \beta_1 \alpha_0}{1 - \alpha_1 \beta_1} + \frac{\alpha_2 \beta_1}{1 - \alpha_1 \beta_1} I_t + \frac{\alpha_3 \beta_1}{1 - \alpha_1 \beta_1} p_t^c + \frac{\beta_2}{1 - \alpha_1 \beta_1} W_t + \frac{\beta_1 \varepsilon_t + \eta_t}{1 - \alpha_1 \beta_1}.$$

OLS estimation of both of these equations yields unbiased estimates of the parameters in the reduced form equations. Identification asks if we can retrieve the parameters of the structural equations from the reduced form equations. Say, for instance, that we re-write the reduced form equations as:

$$(2.74)\quad q_t = \delta_{10} + \delta_{11} W_t + \delta_{12} I_t + \delta_{13} p_t^{c} + \gamma_1$$

and

$$p_t^{w} = \delta_{20} + \delta_{21} I_t + \delta_{22} p_t^{c} + \delta_{23} W_t + \delta_2 . \tag{2.75}$$

Table 1 shows each of the parameters in (11) and (12) in terms of the parameters of the two reduced form equations. We can recover the parameters of the structural equations by algebraic manipulation of the relationships in Table 1. (This method of estimation—that is, estimating the reduced form equations of a model using OLS and then solving algebraically for the parameters of the structural equations is referred to in the literature as *indirect least squares*.) For instance,

$$\frac{\delta_{21}}{\delta_{12}} = \frac{\left(\dfrac{\alpha_2 \beta_1}{1 - \alpha_1 \beta_1}\right)}{\left(\dfrac{\alpha_2}{1 - \alpha_1 \beta_1}\right)} = \beta_1$$

and

$$\frac{\delta_{11}}{\delta_{23}} = \frac{\left(\dfrac{\alpha_1 \beta_2}{1 - \alpha_1 \beta_1}\right)}{\left(\dfrac{\beta_2}{1 - \alpha_1 \beta_1}\right)} = \alpha_1 .$$

Explanatory variable	Equation (11)	Equation (12)
Intercept	$\delta_{10}=\frac{\alpha_0+\alpha_1\beta_0}{1-\alpha_1\beta_1}$	$\delta_{20}=\frac{\beta_0+\beta_1\alpha_0}{1-\alpha_1\beta_1}$
I_t	$\delta_{11}=\frac{\alpha_1\beta_2}{1-\alpha_1\beta_1}$	$\delta_{21}=\frac{\alpha_2\beta_1}{1-\alpha_1\beta_1}$
p_t^c	$\delta_{12}=\frac{\alpha_2}{1-\alpha_1\beta_1}$	$\delta_{22}=\frac{\alpha_3\beta_1}{1-\alpha_1\beta_1}$
W_t	$\delta_{13}=\frac{\alpha_3}{1-\alpha_1\beta_1}$	$\delta_{23}=\frac{\beta_2}{1-\alpha_1\beta_1}$
Error term	$\gamma_1=\frac{\varepsilon_t+\alpha_1\eta_t}{1-\alpha_1\beta_1}$	$\delta_2=\frac{\beta_1\varepsilon_t+\eta_t}{1-\alpha_1\beta_1}$

Table 2.24. Parameters of the structural and reduced form equations.

One can continue in a likewise manner to find formulae for other of the structural $\beta_1 = \frac{\delta_{22}}{\delta_{13}}$. 's. However, an interesting problem does arrive $\frac{\delta_{22}}{\delta_{13}} = \frac{\delta_{21}}{\delta_{12}}$, also true that Since there is no *a priori* reason to believe that we have two estimates of ***β*** $_1$. This result illustrates the point that there are three possibilities when calculating the structural parameters from the reduced form equations—first, there may be more than one formula for a structural parameter; second, there may be only one formula for a structural parameter; or third, there may be no formula for a structural parameter. We say in the first case that the equation is over-identified; is exactly identified in the second case; and is under-identified in the third case. It turns out that in the case of an over-identified equation we can to use TSLS to estimate the structural parameters. However, in the case of an exactly identified equation, the TSLS estimators are equal to the indirect-least-squares estimators that can be calculated using estimates of the reduced form equations. Finally, an under-identified equation cannot be estimated by any technique.

Clearly, we need to know how to identify if an equation is either over-identified, exactly identified, or under-identified. A necessary rule is that the number of exogenous variables in a system of equation that are not included in a particular regression must be greater than or equal to the number of endogenous variables on the right-hand-side of the equation for the equation to be either exactly or over identified. Consider the following three-equation model, where the endogenous variables are y_1, y_2, and y_3 and the exogenous variables are represented by x_1 with $i = 1,\ldots,5$:

$$(2.76)\ y_1 = \beta_{10} + \beta_{12} y_2 + \beta_{13} y_3 + \alpha_{11} x_1 + \alpha_{12} x_2 + \alpha_{15} x_5 ,$$

$$(2.77)\ y_2 = \beta_{20} + \beta_{21} y_1 + \alpha_{23} x_3 , \text{ and}$$

$$(2.78)\ y_3 = \beta_{30} + \beta_{31} y_1 + \alpha_{31} x_1 + \alpha_{32} x_2 + \alpha_{33} x_3 + \alpha_{34} x_4 + \alpha_{35} x_5 .$$

The error terms in these three equations are omitted because they are irrelevant to determining if an equation is identified—remember, identification is an algebraic problem, not a statistical issue. There are 3 endogenous variables in the system and 3 equations in the system. Also, there are 5 exogenous variables in the system of equations. Equation (13) is exactly identified; Equation (14) is over-identified; and Equation (15) is under-identified. What this means is (1) Equation (13) can be estimated directly from the reduced form equation (using indirect-least-squares) or using TSLS; (2) Equation (14) must be estimated using TSLS; and Equation (15) cannot be estimated. Table 2 summarizes how to determine if an equation is or is not identified. Basically, if the number in column 2 equals the number in column 3, the equation is exactly identified. If the number in column 2 is less than the number in column 3, the equation is

over-identified. Finally, if the number in column 2 is greater than the number in column 3, the equation is under-identified.[27]

Equation	**Number of endogenous variables on right-hand-side**	**Number of exogenous variables excluded from the equation**	**Identification**
$y_1 = \beta_{10} + \beta_{12} y_2 + \beta_{13} y_3 + \alpha_{11} x_1 + \alpha_{12} x_2 + \alpha_{15} x_5$	**2**	**2**	**Exactly**
$y_2 = \beta_{20} + \beta_{21} y_1 + \alpha_{23} x_3$	**1**	**4**	**Over**
$y_1 = \beta_{10} + \beta_{12} y_2 + \beta_{13} y_3 + \alpha_{12} x_2 + \alpha_{13} x_3 + \alpha_{15} x_5$	**1**	**0**	**Under**

Table 2.25. Identification of the equations in the example model.

One other thing to notice is the similarity of TSLS to IV estimation. The exogenous variables play the role of instruments in TSLS estimation. By implication, the instruments in an IV estimation must not include any of the exogenous variables in the equation.[28] Similarly, one of the

ways to isolate potential instruments in a regression is to think of what system of equation the equation is and then ask what exogenous variables in that system are not included in the equation. These excluded exogenous variables are potential instruments.

TSLS and IV in Stata

The command for estimating an equation in *Stata* using two-stages least squares (TSLS) is a bit tricky. Assume that you want to estimate equations (13) and (14) in the model discussed above.[29] For simplicity assume that each variable assumes the name for it in Table 2. Thus, in our *Stata* commands Y1 refers to variable Thus, in our Stata commands Y1 refers to variable y_1 and so on. The command to estimate either a TSLS or an IV regression is the same.[30] The command, ivreg, consists of three major parts—(1) the name of the dependent variable is followed by (2) a list of the names of the exogenous variables that are being used as explanatory variables and then followed in parentheses by (3) the information needed to estimate the first stage (the list of the endogenous variables that are explanatory variables along with the names of the exogenous variables in the system that are excluded from the equation or, in the case of IV, a list of the instruments).[31]

Equation to be estimated	*Stata* command
$y_1 = \beta_{10} + \beta_{12} y_2 + \beta_{13} y_3 + \alpha_{12} x_2 + \alpha_{13} x_3 + \alpha_{15} x_5$	.ivreg y1 x2 x3 x5 (y2 y3 = x1 x4)
$y_2 = \beta_{20} + \beta_{21} y_1 + \alpha_{23} x_3$	.ivreg y2 x3 (y1 = x1 x2 x4 x5)

Table 2.26. *Stata* command for estimating TSLS and IV regressions.

Example 2.5.

An example from *Stata.* The *Stata* manual offers the following example analysis. Assume that you want to use state level data from the 1980 census to estimate the following system of equations:

$$hsngval = \alpha_0 + \alpha_1 faminc + \alpha_2 reg2 + \alpha_3 reg3 + \alpha_4 reg4 + \varepsilon \quad (2.79)$$

and

$$(2.80)\ rent = \beta_0 + \beta_1 hsngval + \beta_2 pcturban + v,$$

where *hsngval* is the median dollar value of owner-occupied housing; *rent* is the median monthly gross rent; *fainc* is family income; *pcturban* is the percent of the state population living in an urban area; and *reg2*, *reg3*, and *reg4* are dummy variables that designate the region of the country where the state is located. In this example we focus on estimating (17).

We begin by loading the data set and describing the data.

```
. use http://www.stata-press.com/data/r8/hsng2

(1980 Census housing data)

.describe
```

Contains data from http://www.stata-press.com/data/r8/hsng2.dta	
obs: 50	1980 Census housing data

vars: 16	3 Sep 2002 12:25

size: 3,600 (99.7% of memory free)	

variable name	storage type	display format	value label	variable	label
state	str14	%	14s		State
division	int	%	8.0g	division	Census division
region	int	%	8.0g	region	Region
pop	long	%	10.0g		Population in 1980
popgrow	float	%	6.1f		Pop. growth 1970-80
popden	int	%	6.1f		Pop/sq. mile
pcturban	float	%	8.1f		Percent urban
faminc	long	%	8.2f		Median family inc.,

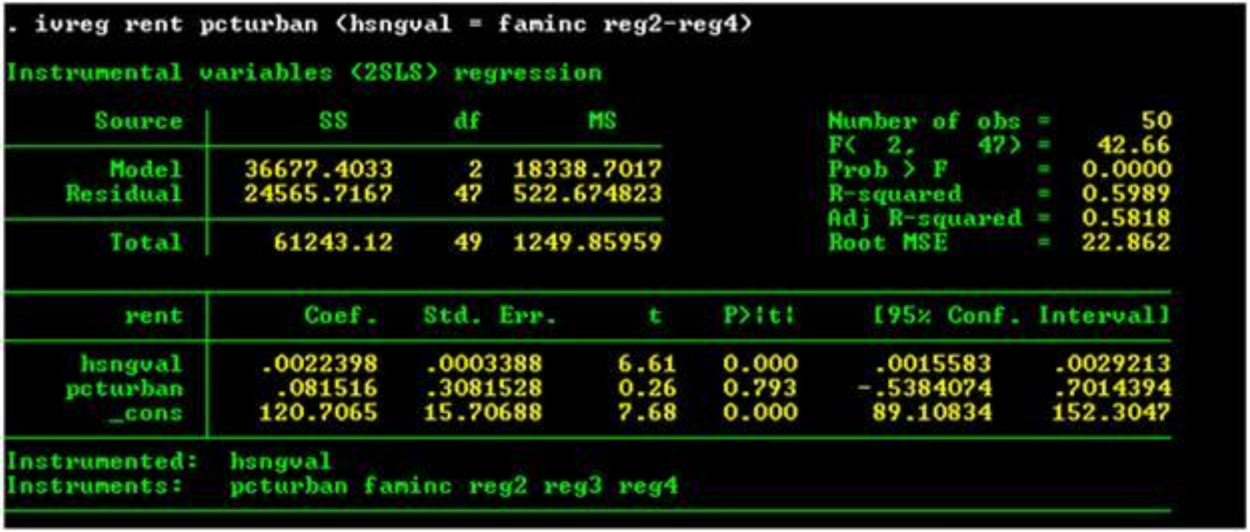
					1979
hsng	long	%	10.0g		Hsng units 1980
hsnggrow	float	%	8.1f		% housing growth
hsngval	long	%	9.2f		Median hsng value
rent	long	%	6.2f		Median gross rent
reg1	float	%	9.0g		
reg2	float	%	9.0g		
reg3	float	%	9.0g		
reg4	float	%	9.0g		

Sorted by: state

Table 2.27. Description of the *Stata* data set used in the example.

Now we estimate equation (17) using TSLS as shown in Figure 2.

Figure 2.29. Two-stages least square estimate of the example.



```
. ivreg rent pcturban (hsngval = faminc reg2-reg4)

Instrumental variables (2SLS) regression

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  2,    47) =   42.66
       Model |  36677.4033     2  18338.7017           Prob > F      =  0.0000
    Residual |  24565.7167    47  522.674823           R-squared     =  0.5989
-------------+------------------------------           Adj R-squared =  0.5818
       Total |    61243.12    49  1249.85959           Root MSE      =  22.862

------------------------------------------------------------------------------
        rent |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     hsngval |   .0022398   .0003388     6.61   0.000     .0015583    .0029213
    pcturban |    .081516   .3081528     0.26   0.793    -.5384074    .7014394
       _cons |   120.7065   15.70688     7.68   0.000     89.10834    152.3047
------------------------------------------------------------------------------
Instrumented:  hsngval
Instruments:   pcturban faminc reg2 reg3 reg4
------------------------------------------------------------------------------
```

The manual continues the example to include some testing of the model including the Hausman test. Students using TSLS and IV should read the discussion in the *Stata* manual thoroughly.

Exercises

Exercise 2.5.1.

Cigarette advertising and sales. A great deal of controversy exists over the issue of whether advertising expenditures affect sales. This controversy is particularly sharp when it affects policy decisions. An example of this phenomenon is the controversy over the impact of cigarette advertising on advertising sales. While many public policy experts advocate bans on cigarette advertising, a majority of economists caution against bans on cigarette advertising. The economists point out that there is little theoretical reasons to believe that cigarette advertising affects total demand for cigarettes. Instead, economists argue that cigarette advertising only affects brand choice and not the number of cigarettes that people smoke. Moreover, these economists point out that there is also little empirical evidence that supports the argument that cigarette advertising affects the demand for cigarettes. Given the negative impact advertising bans have on freedom of speech, most economists conclude that the negative effects of cigarette advertising bans outweigh the benefits of the bans.

In this exercise we address this issue by using data used originally by Richard Schmalensee (1972) in his Ph.D. dissertation. You will use these data to estimate a simple two-equation model of the cigarette advertising industry.

We use annual data for the period 1955 to 1967 to estimate the impact of cigarette advertising on aggregate demand for cigarettes and the impact of cigarette consumption on cigarette advertising. We begin with a model of the demand for cigarettes. We assume that the demand for cigarettes is given by:

(2.81)

$$q_t = f(pc_t, y_t, A_t, D64),$$

where

q_t = cigarettes consumed per person over age 15,

pc_t = retail price of cigarettes,

y_t = real disposable personal income per capita (1958 dollars),

A_t = real advertising expenditures per individual over age 15 (1960 dollars), and

D64 = a dummy variable equal to 1 for the years 1964 through 1967 and zero otherwise.

We include the dummy variable for years after 1964 to pick up the negative impact on cigarette sales of the 1964 report of the US Surgeon General's Advisory Committee (1964) announcing that the government believed that there was enough evidence available to conclude that cigarette smoking causes cancer. We expect the signs of the parameters with the price of cigarettes and the dummy variable to be negative. We expect that the sign of the parameters with income and advertising to be positive.

Next we turn to a model of the supply of advertising. We assume:

$$A_t = g(q_t, pa_t, m_t), \tag{2.82}$$

where:

pa_t = advertising price index, and

m_t = gross profits as a percentage of gross sales.

The last variable needs a bit of explaining. The amount of advertising in the industry should be a function of degree of competition in the industry. If the market were perfectly competitive, there would be no reason for any firm to advertise. If the firm were a monopoly, there also would be no reason to advertise. However, if the market is an oligopoly, then a firm would advertise in an effort to gain market share by differentiating its product from the product of its competitors.

The traditional measure of the degree of monopoly power that a firm has is the ratio of its marginal profits to its marginal cost:

(2.83)

$$m = \frac{p - mc}{mc},$$

where *p* is output price, *mc* is marginal cost, and *m* is the measure of monopoly power. Since we cannot observe the firms' marginal costs, we approximate *m* by the ratio of gross profits to gross sales. We expect the impact of the degree of monopoly to have a non-linear impact on advertising expenditures.

The data used to estimate our two equations are listed in Table 5 and are available in the MS Excel file Cigarette sales and advertising data.xls. These data are with the exception of

disposable personal income from Schmalensee (1972: 273-290). The disposable personal income data are from the Department of Commerce (1975: Table F26, page 225).

Specification of the Model. Equations (18) and (19) are, as written, very general and need further specification before they can be estimated. We will assume that the two equations take a log-log form. In particular, we assume that we want to estimate:

$$\ln(q_t) = \alpha_0 + \alpha_1 \ln(pc_t) + \alpha_2 \ln(y_t) + \alpha_3 \ln(A_t) + \alpha_4 D64_t \tag{2.84}$$

and

$$\ln(A_t) = \beta_0 + \beta_1 \ln(q_t) + \beta_2 \ln(pa_t) + \beta_3 m_t + \beta_4 m_t^2. \tag{2.85}$$

Year	Cigarettes Sold per Person Over Age 15	Retail Price of Cigarettes	Real Advertising per Person Over Age 15	Advertising Price Index	Degree of Monopoly	Disposable Personal Income in 1958 dollars

1955	3163.090	93.9693	0.96100	95.4775	18.595	1659
1956	3230.517	94.7049	1.09969	94.3800	19.207	1673
1957	3313.033	94.2535	1.22180	96.2125	20.165	1683
1958	3479.063	94.7712	1.40471	97.8300	21.736	1666
1959	3584.930	98.1779	1.45816	98.2800	22.042	1735
1960	3676.912	100.0000	1.37863	100.0000	22.04	1749
1961	3743.354	99.8677	1.31871	102.0400	22.465	1756
1962	3733.504	99.6761	1.35467	102.9725	22.226	1814
1963	3775.886	101.3630	1.51345	103.9525	22.848	1867
1964	3648.211	102.3110	1.73665	103.4775	23.168	1948
1965	3710.075	105.7510	1.59761	103.7225	23.598	2047
1966	3689.386	108.0450	1.71062	104.2200	25.085	2127
1967	3652.016	109.2490	1.71444	104.6125	26.310	2164

Table 2.28. Cigarette Industry Data, 1955-1967.

Answer the following six questions:

a) Which variables in the model are exogenous and which are endogenous?

b) Check and see if equations (18) and (19) are underidentified, exactly identified, or overidentified.

c) Estimate equations (21) and (22) using ordinary least squares.

d) Estimate equations (21) and (22) using two-stage least squares. Present the results in a table that for comparison reasons includes the results from the OLS estimation. Be sure to include the R^2 and the Durbin-Watson statistic.

e) Which side of the advertising-sales controversy do your results appear to support?

f) How well-specified does your model appear to be? Why?

Exercise 2.5.2.

Exercise 2. Demand and supply of commercial loans. We are interested in estimating the demand for commercial loans by business firms and the supply of commercial loans by banks. We have available in Table 6 monthly data from the U. S. commercial loan market for the period from January, 1979 through December, 1984 and available in the MS Excel file Exercise 2.xls.[32] Define:

Q_t = total commercial loans (billions of dollars)

R_t = average prime rate charged by banks

RS_t = 3-month Treasury bill rate (represents an alternative rate of return for banks)

RD_t = Aaa corporate bond rate (represents the price of alternative financing to firms)

X_t = industrial production index (represents firms' expectation about future economic activity)

y_t = total bank deposits (billions of dollars) (represents a scale variable).

The demand and supply equations to be estimated, respectively, are as follows:

$$(2.86)\ Q_t = \beta_0 + \beta_1 R_t + \beta_2 RD_t + \beta_3 X_t + \mu_t$$

and

$$(2.87)\ Q_t = \alpha_0 + \alpha_1 R_t + \alpha_2 RS_t + \alpha_3 y_t + \varepsilon_t .$$

Questions

a) What are the endogenous and exogenous variables in this model?

b) Solve for the two "reduced form" equations of this model. Estimate these two equations using the data in Table 6.

c) Check the "order" condition for identification of each equation of the model.

d) Estimate equations (23) and (24) using ordinary least squares using the data in Table 6.

e) Estimate equations (23) and (24) using two-stage least squares. Report the results of the estimations for part 4 and 5 in a single table. Be sure to include the t-ratios, R^2's, and Durbin-Watson statistics for each of the equations estimated.

f) Perform the Hausman Specification Test on both equations.[33]

g) When presenting this model, Maddala notes "[T]he model postulated here is not necessarily the right model for the problem of analyzing the commercial loan market." Is there anything in the results reported above that suggests that the model may be mis-specified?

N	Date	*Q*	*R*	*RD*	*X*	*RS*	*y*
1	January-79	251.8	11.75	9.25	150.8	9.35	994.3
2	February-79	255.6	11.75	9.26	151.5	9.32	1002.5
3	March-79	259.8	11.75	9.37	152.0	9.48	994.0
4	April-79	264.7	11.75	9.38	153.0	9.46	997.4
5	May-79	268.8	11.75	9.50	150.8	9.61	1013.2
6	June-79	274.6	11.65	9.29	152.4	9.06	1015.6
7	July-79	276.9	11.54	9.20	152.6	9.24	1012.3
8	August-79	280.5	11.91	9.23	152.8	9.52	1020.9

9	September-79	288.1	12.90	9.44	151.6	10.26	1043.6
10	October-79	288.3	14.39	10.13	152.4	11.70	1062.6
11	November-79	287.9	15.55	10.76	152.4	11.79	1058.5
12	December-79	295.0	15.30	11.31	152.1	12.64	1076.3
13	January-80	295.1	15.25	11.86	152.2	13.50	1063.1
14	February-80	298.5	15.63	12.36	152.7	14.35	1070.0
15	March-80	301.7	18.31	12.96	152.6	15.20	1073.5
16	April-80	302.0	19.77	12.04	152.1	13.20	1101.1
17	May-80	298.1	16.57	10.99	148.3	8.58	1097.1
18	June-80	297.8	12.63	10.58	144.0	7.07	1088.7
19	July-80	301.2	11.48	11.07	141.5	8.06	1099.9
20	August-80	304.7	11.12	11.64	140.4	9.13	1111.1
21	September-80	308.1	12.23	12.02	141.8	10.27	1122.2

22	**October-80**	**315.6**	**13.79**	**12.31**	**144.1**	**11.62**	**1161.4**
23	**November-80**	**323.1**	**16.06**	**11.94**	**146.9**	**13.73**	**1200.6**
24	**December-80**	**330.6**	**20.35**	**13.21**	**149.4**	**15.49**	**1239.9**
25	**January-81**	**330.9**	**20.16**	**12.81**	**151.0**	**15.02**	**1223.5**
26	**February-81**	**331.3**	**19.43**	**13.35**	**151.7**	**14.79**	**1207.1**
27	**March-81**	**331.6**	**18.04**	**13.33**	**151.5**	**13.36**	**1190.6**
28	**April-81**	**336.2**	**17.15**	**13.88**	**152.1**	**13.69**	**1206.0**
29	**May-81**	**340.9**	**19.61**	**14.32**	**151.9**	**16.30**	**1221.4**
30	**June-81**	**345.5**	**20.03**	**13.75**	**152.7**	**14.73**	**1236.7**
31	**July-81**	**350.3**	**20.39**	**14.38**	**152.9**	**14.95**	**1221.5**
32	**August-81**	**354.2**	**20.50**	**14.89**	**153.9**	**15.51**	**1250.3**
33	**September-81**	**366.3**	**20.08**	**15.49**	**153.6**	**14.70**	**1293.7**
34	**October-81**	**361.7**	**18.45**	**15.40**	**151.6**	**13.54**	**1224.6**

35	November-81	365.5	16.84	14.22	149.1	10.86	1254.1
36	December-81	361.4	15.75	14.23	146.3	10.85	1288.7
37	January-82	359.8	15.75	15.18	143.4	12.28	1251.5
38	February-82	364.6	16.56	15.27	140.7	13.48	1258.3
39	March-82	372.4	16.50	14.58	142.7	12.68	1295.0
40	April-82	374.7	16.50	14.46	141.5	12.70	1272.1
41	May-82	379.3	16.50	14.26	140.2	12.09	1286.1
42	June-82	386.7	16.50	14.81	139.2	12.47	1325.8
43	July-82	384.4	16.26	14.61	138.7	11.35	1307.3
44	August-82	384.5	14.39	13.71	138.8	8.68	1321.7
45	September-82	395.0	13.50	12.94	138.4	7.92	1335.5
46	October-82	393.7	12.52	12.12	137.3	7.71	1345.2
47	November-82	398.9	11.85	11.68	135.7	8.07	1358.1

48	December-82	395.3	11.50	11.83	134.9	7.94	1409.7
49	January-83	392.4	11.16	11.79	135.2	7.86	1385.4
50	February-83	392.3	10.98	12.01	137.4	8.11	1412.6
51	March-83	395.9	10.50	11.73	138.1	8.35	1419.5
52	April-83	393.5	10.50	11.51	140.0	8.21	1411.0
53	May-83	391.7	10.50	11.46	142.6	8.19	1413.1
54	June-83	395.3	10.50	11.74	144.4	8.79	1443.8
55	July-83	397.7	10.50	12.15	146.4	9.08	1438.1
56	August-83	400.6	10.89	12.51	149.7	9.34	1461.4
57	September-83	402.7	11.00	12.37	151.8	9.00	1448.9
58	October-83	405.3	11.00	12.25	153.8	8.64	1459.0
59	November-83	412.0	11.00	12.41	155.0	8.76	1499.4
60	December-83	420.1	11.00	12.57	155.3	9.00	1508.9

61	January-84	424.4	11.00	12.20	156.2	8.90	1504.1
62	February-84	428.8	11.00	12.08	158.5	9.09	1499.3
63	March-84	433.1	11.21	12.57	160.0	9.52	1494.5
64	April-84	439.7	11.93	12.81	160.8	9.69	1501.5
65	May-84	447.3	12.39	13.28	162.1	9.83	1541.3
66	June-84	452.9	12.60	13.55	162.8	9.87	1532.9
67	July-84	454.4	13.00	13.44	164.4	10.12	1535.5
68	August-84	455.2	13.00	12.87	165.9	10.47	1539.0
69	September-84	459.9	12.97	12.66	166.0	10.37	1549.9
70	October-84	467.7	12.58	12.63	165.0	9.74	1578.9
71	November-84	468.7	11.77	12.29	164.4	8.61	1578.2
72	December-84	476.8	11.06	12.13	164.8	8.06	1631.2

Table 2.29. Monthly Data for the U.S. Commercial Loan Market, January 1979 to December 1984.

References

Angrist, Joshua D. and Alan B. Krueger (2001). Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic Perspectives* 15(4): 69–85.

Berndt, Ernst R. (1991). *The Practice of Econometrics* (Reading, MA: Addison-Wesley Publishing Company).

Greene, William H. (1990). *Econometric Analysis* (New York: Macmillan Publishing Company).

Murray, Michael P. (2006a). Avoiding Invalid Instruments and Coping with Weak Instruments. *Journal of Economic Perspectives* 20(4): 111-132.

Murray, Michael P. (2006b). *Econometrics: A Modern Introduction.* (Boston: Addison-Wesley): Chapter 13.

Schmalensee, Richard (1972). *The Economics of Advertising* (Amsterdam: North-Holland Publishing Company).

StataCorp (2003). *Stata Statistical Software: Release 8* (College Station, TX: Stata Corporation): Volume 2: Reference G-M, pages 186-194.

Stock, James H, and Mark W. Watson (2003). *Introduction to Econometrics* (Boston, MA: Addison-Wesley): Chapter 10.

US Department of Commerce (1975). *Historical Statistics of the United States: Colonial Times to 1970* (Washington: Government Printing Office).

US Surgeon General's Advisory Committee (1964). *Smoking and Health* (Washington: Government Printing Office).

2.6. Replication of econometric studies[*]

Replication

Introduction

One of the most important first steps in a science experiment is to replicate the results of earlier research. For a variety of reasons (most of them practical and not theoretically sound) economists generally do not undertake this step; what they tend to do is report the results of earlier papers and then compare their results with the earlier results without asking the question of whether these earlier results were reported accurately. Omitting this step in a world of honest careful researchers might seem to be a minor problem. However, there is enough casual evidence to suggest that a large portion of the econometric results reported in the journals cannot be replicated because the original researcher (1) does not have the data set used in the research because it has been lost for a variety of reasons, (2) cannot share the data set because it is proprietory, (3) is unwilling to share the data set because there are other issues they wish to investigate using the data set, or (4) just are unwilling to share the data set. For this reason much of the published econometrics research has never been replicated. In recognization of this problem several journals like the *Journal of Applied Econometrics* now require that authors submit the data set they used to the journal to be posted on the web for use by any other researcher. Whether this effort has been successful will not be clear unless someone undertakes to replicate the work in this journal to see if all of the data necessary to replicate an article have been posted and if the regressions included in the

article actually can be replicated. It is very unlikely anyone would undertake such an effort given the fact that no journal will publish results that are merely a replication of previously published articles.

In this module we explore some of the difficulties that exist in replicating existing research by undertaking to replicate some of the results reported in the Butler, Finegan, and Siegfried (1998) (BFS, hereafter) article analyzing the effect of a student's calculus background on the grade he or she earns in intermediate microeconomics or in intermediate macroeconomics.[34] The goal of this module is to (1) help students to learn how to read in detail an article that appears in a typical economics trade journal, (2) introduce them to ordered probit, an advanced econometrics tool, and (3) teach them how to present and discuss the results of an estimation of a model in an economics paper. While most of the discussion in this module focuses on using *Stata* in this replication, one can use most any econometrics program they are comfortable with to replicate some of the results reported in the BFS article.

Butler, Finegan, and Siefried (1998).

The obvious first step is to find and print a copy of the article by Butler, Finegan, and Siefried. In fact, do not proceed any further in reading this module until you have read

the article. We will discuss in class what the authors do in the paper and how clearly they present their conclusions. In this first pass at the article you are to pay attention to how convincing you find their arguments to be. Since everyone in the class has completed an intermediate microeconomics course, your discussion of their conclusions should reflect your own experiences. Also, you need to be able to discuss in class the estimation strategy they use in the paper. In particular, you will need to be able to identify what the source of the data is and what equations did they estimate. Also, try to determine how the estimations in the "first" stage are used in the estimations of the "second" stage. Why did the authors use a two-stage estimation strategy?

Also, what do you think the authors mean in their description of their estimation strategy by their statement about the estimation methods they use:

> ***Estimation Methods and Expectations***
>
> ***To cope with the selection bias problem, we use a two-stage estimation procedure. The first stage employs an ordered probit model to predict the highest level of calculus attained by each student prior to taking each intermediate economic theory course.... In the second stage, the student's grade in MICRO-2 ... (the `outcome') is regressed on the actual level of calculus attained, the grade earned in that calculus course, the predicted residual in the grade equation that we would expect on the***

> *basis of the actual level of calculus attained, and a roster of control variables reflecting ability and motivation. Individuals are the unit of observation. Ordinary least squares estimation is used because there are twelve categories of grades which are commonly interpreted as cardinal measures of performance (as is implied by the calculation of `grade point averages'). (Butler, Finegan, and Siegfried, 1998: 188)*

The ordered-probit model

In what follows you are to “replicate” the equations the authors estimate in the paper for the intermediate microeconomics course. In order to complete this assignment you will need to figure out several things including (1) what an ordered-probit model is and (2) how to use *Stata* to estimate an ordered-probit model. In this section of the module we introduce the ordered-probit model. I strongly encourage you to consult Greene (1990: 703-706) for an excellent and clear discussion of the ordered-probit model. The discussion here follows Greene closely.

It is common for surveys to have questions that require the responder to choose one of several categories that have an innate order to them. For instance, most course evaluations ask the respondent to choose an answer to a question that reflects their agreement with a statement about the course. For instance, the question might read, "The Professor was interested in the material taught in the class" where the student completing the evaluation would choose a number from 1 to 9 where a 1 indicates complete disagreement with the statement and a 9 reflects complete agreement with the statement. Thus, there is an order to the potential answers. Using a logit, probit, or multilogit model would completely ignore this order. A linear regression is inappropriate because OLS treats the difference between answers of 1 and 2 as being the same as the difference between a 7 and and 8, when in fact the numbers only provide a ranking.

Consider a latent variable, *y**, that is not observed but where $y=\boldsymbol{\beta}'\mathbf{x}+\varepsilon.$ We want to estimate the $\boldsymbol{\beta}_k$' s in the vector $\boldsymbol{\beta}=(\beta_0 \;\; \beta_1 \;\; \cdots \;\; \beta_K).$ [35] We may not observe *y** but we do observe:

The μ_i's in (1) are parameters that must be estimated along with $\boldsymbol{\beta}.$ As usual, we assume that the error term ε is normally distributed (with a normalized mean and variance arbitrarily set to 0 and 1, respectively). It is trivial to estimate the model with the error terms having a logistic distribution, but this chance in assumptions appears to make virtually no difference in practice).[36] With the normal distribution, we have:

(2.88)

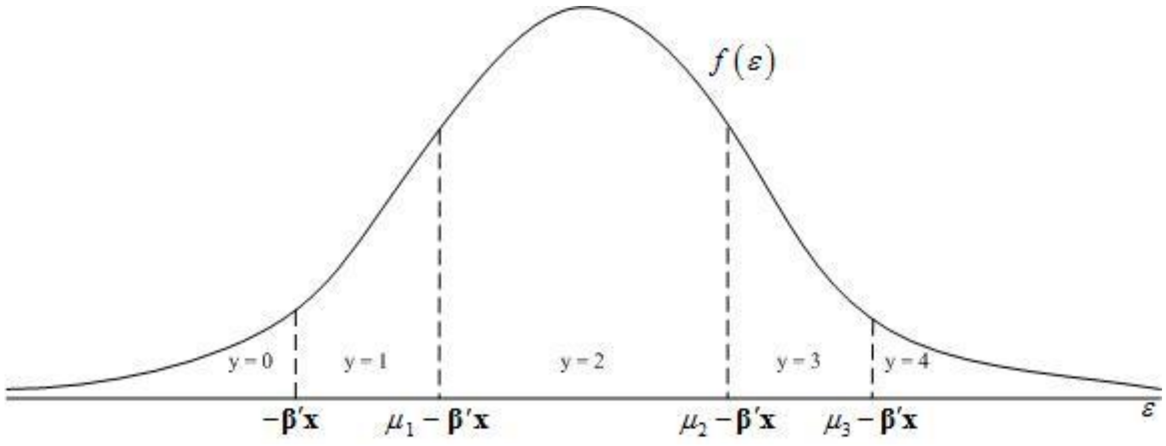
$$y = \begin{cases} 0 \text{if} y^* < 0, \\ 1 \text{if} 0 \leq y^* < \mu_1, \\ 2 \text{ if} \mu_1 \leq y^* < \mu_2, \\ \vdots \\ J \text{if} \mu_{J-1} \leq y^*. \end{cases}$$

(2.89)

$$\begin{aligned} &\Pr(y = 0) = \Phi(-\boldsymbol{\beta}'\mathbf{x}), \\ &\Pr(y = 1) = \Phi(\mu_1 - \boldsymbol{\beta}'\mathbf{x}) - \Phi(-\boldsymbol{\beta}'\mathbf{x}), \\ &\Pr(y = 2) = \Phi(\mu_2 - \boldsymbol{\beta}'\mathbf{x}) - \Phi(\mu_1 - \boldsymbol{\beta}'\mathbf{x}), \\ &\vdots \\ &\Pr(y = J) = 1 - \Phi(\mu_{J-1} - \boldsymbol{\beta}'\mathbf{x}), \end{aligned}$$

where $\boldsymbol{\Phi}(\cdot)$ is the cumulative normal function. In order for all of the probabilities to be positive, we need $\mu_1 < \mu_2 < \cdots < \mu_{J-1}$, as shown in Figure 1. One thing to note in Figure 1 is that the cutoff locations change when the values of the explanatory variables change.

Figure 2.30.



Distribution of the error term in the ordered-probit model.

The estimation strategy from here follows the usual maximum likelihood method. The computer program forms the likelihood function and then chooses the values of the parameters (including the cutoffs) that maximize this likelihood function.

The estimated coefficients are not equal to the marginal effects of a change in one of the explanatory variables (as is also true with the logit and probit models). Consider the

simple example Greene (1990, 704) describes. Assume that there are three categories. Then (2) becomes:

(2.90)

$$\Pr(y=0)=1-\Phi(\boldsymbol{\beta}'\mathbf{x}),$$
$$\Pr(y=1)=\Phi(\mu-\boldsymbol{\beta}'\mathbf{x})-\Phi(-\boldsymbol{\beta}'\mathbf{x}),$$
$$\Pr(y=2)=1-\Phi(\mu-\boldsymbol{\beta}'\mathbf{x}).$$

Figure 2 shows this situation. The solid curve shows the distribution of *y* and *y**.

Increasing one of the x's while holding the β constant (that is, changing $\hat{\boldsymbol{\beta}}'\mathbf{x}_0$ to $\hat{\boldsymbol{\beta}}'\mathbf{x}_1)$ is the same as shifting the entire distribution of *y* and *y** to the right with $\hat{\mu}$ remaining constant. As a result the probabilities that y takes on the values of 0, 1, and 2 change. Clearly, as shown in Figure 2, Pr(*y* = 0) decreases and Pr(*y* = 2) increases. The Pr(*y* = 1), on the other hand, may increase or decrease and, thus, the effect of an increase in one of the explanatory variables is ambiguous. It is easy to show this result algebraically. The marginal effects for the 3 probabilities in (3) are, assuming $\boldsymbol{\beta}>0:$

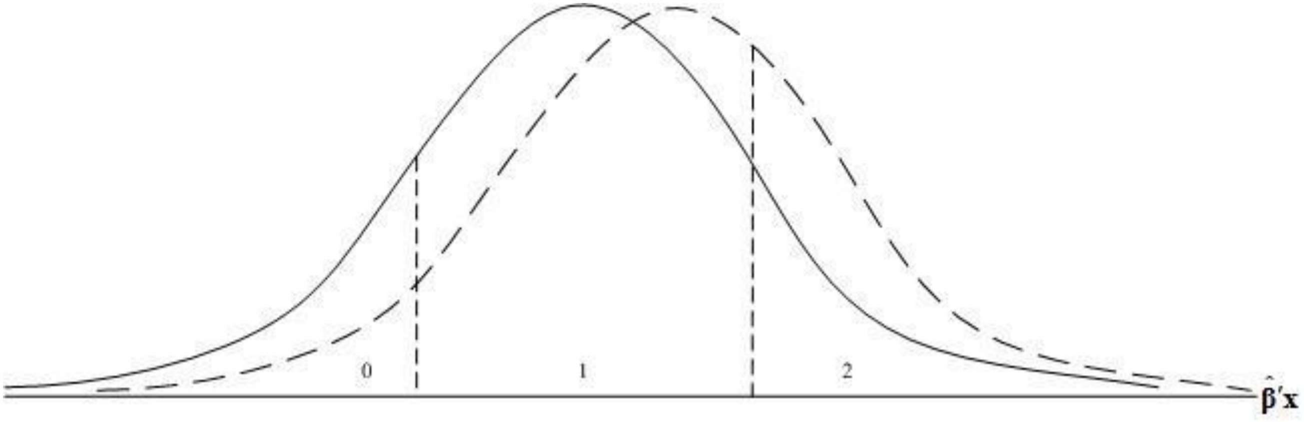
(2.91)

$$\frac{\partial \Pr(y=0)}{\partial \mathbf{x}} = -\phi(\boldsymbol{\beta}'\mathbf{x})\boldsymbol{\beta} < 0,$$

$$\frac{\partial \Pr(y=1)}{\partial \mathbf{x}} = \phi(\mu - \boldsymbol{\beta}'\mathbf{x})\boldsymbol{\beta} - \phi(\boldsymbol{\beta}'\mathbf{x})\boldsymbol{\beta},$$

$$\frac{\partial \Pr(y=2)}{\partial \mathbf{x}} = \phi(\mu - \boldsymbol{\beta}'\mathbf{x})\boldsymbol{\beta} > 0.$$

Figure 2.31.



A rise in one of the explanatory variables whose parameter is positive will shift the probability distribution of the outcome to the right (from the solid line to the dashed line).

In general, only the sign's of the change Pr(*y* = 0) and Pr(*y* = *J*) are unambiguous. Greene (1990, 705) cautions that "“[w]e must be very careful in interpreting the coefficients in this model.... Indeed, without a fair amount o extra calculation, it is quite unclear how the coefficients in the ordered-probit model should be interpreted.”"

The BFS Dataset

The data used by BFS are available at the *Journal of Applied Econometrics* data website or in the *MS Excel* file *Vanderbilt data set.xls* . Table 1 identifies the variables in the dataset.

Column	Code	Variable definition
A	Obs	Observation number

B	SID	Student ID
C	Grade	Grade earned in Economics 231, A = 4, A- = 3.7, etc.
D	SelCorr	Variable correcting for selection bias
E	Soph	Dummy variable = 1 if student is a sophomore
F	Senior	Dummy variable = 1 if student is a senior
G	Same	Dummy variable = 1 if student took both intermediate classes the same year
H	Skip	Dummy variable = 1 if student took the intermediate classes at least one semester apart
I	HighestMath	Highest level of math attained (the dependent variable, 0-6 corresponding to Math 170, 171a, 172a, 171b, 172b, 221a, 221b)
J	M170	Dummy variable = 1 if student's highest level of math was Math 170
K	M171a	Dummy variable = 1 if student's highest level of math was Math 171A

L	M172a	Dummy variable = 1 if student's highest level of math was Math 172a
M	M171b	Dummy variable = 1 if student's highest level of math was Math 171b
N	M172b	Dummy variable = 1 if student's highest level of math was Math 172b
O	M221a	Dummy variable = 1 if student's highest level of math was Math 221a
P	M221b	Dummy variable = 1 if student's highest level of math was Math 221b
Q	GE100	Grade in Economics 100
R	GDE100	Individual instructor grade deflator in Economics 100
S	GE101	Grade in Economics 101
T	GDE101	Individual instructor grade deflator in Economics 101
U	GDE231	Individual instructor grade deflator in Economics 231

V	Size	Class size
W	FGPA	Freshman GPA
X	Female	Dummy variable =1 if student is a female
Y	MSAT	Score on Math section of the SAT
Z	VSAT	Score on Verbal section of the SAT
AA	TE231	Teacher of Economics 231 (numerical code)
AB	SE231	Section of Economics 231 (numerical code)
AC	GM170	Grade in highest math class: Math 170
AD	GM171a	Grade in highest math class: Math 171a
AE	GM172a	Grade in highest math class: Math 172a
AF	GM171b	Grade in highest math class: Math 171b
AG	GM172b	Grade in highest math class: Math 172b
AH	GM221a	Grade in highest math class: Math 221a

AI	GM221b	Grade in highest math class: Math 221b
AJ	GHM	Grade in highest math class
AK	Foreign	Dummy variable = 1 if student passed foreign language proficiency test
AL	EMEcon	Dummy variable = 1 if expected major is economics
AM	EMOSS	Dummy variable = 1 if expected major is another social science
AN	EMNS	Dummy variable = 1 if expected major is a natural science
AO	EMH	Dummy variable = 1 if expected major is in the humanities
AP	AM1	Dummy variable = 1 if student completed 1 year of advanced math in high school
AQ	AM2	Dummy variable = 1 if student completed 2 years of advanced math in high school
AR	AM3	Dummy variable = 1 if student completed 3 years of advanced math in high school
AS	Phy1	Dummy variable = 1 if student completed 1 course in physics in

		high school
AT	Phy2	Dummy variable = 1 if student completed 2 courses in physics in high school
AU	Chem1	Dummy variable = 1 if student completed 1 course in chemistry in high school
AV	Chem2	Dummy variable = 1 if student completed 2 courses in chemistry in high school

Table 2.30. Definition of the variables included in the Vanderbilt data set.

Replication of the Ordered Probit Regression

At this point we are ready to begin the replication. Since it is easy to get lost in the process, I have created a list of steps that include both instructions on what to do and questions you need to answer. As part of this exercise you will be asked to complete several tables of results. In order to make this effort easier, I have provided a MS Word file, Tables for ordered probit discussion.doc, with the tables to be completed in it.

1. Load the data in *Stata* from *Excel*.

2. Convert MSAT and VSAT to MSAT/100 and VSAT/100, respectively, using the commands:

```
.replace msat = msat/100

.replace vsat = vsat/100
```

3. Common sense dictates that we should calculate the means and standard deviations of the variables to be sure that there are no entry errors. We need to construct a table that compares the means and standard deviations reported in BFS with those in our dataset. Table 2, which has the means and standard deviations reported by BFS, gives a place to put the means and standard deviations for the variables in our dataset. Fill in the information missing from Table 2.

	Our data		**Butler, et al.**	
Variable	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**
msat			**6.25**	**0.60**
foreign			**0.11**	**0.32**

female			0.39	0.49
emecon			0.34	0.48
emoss			0.17	0.38
emns			0.21	0.41
emh			0.07	0.25
am1			0.49	0.50
am2			0.45	0.50
am3			0.01	0.11
phy1			0.67	0.47
Phy2			0.02	0.14
chem1			0.82	0.39
chem2			0.12	0.32

Table 2.31. Means and standard deviations of the data.

4. Estimate the ordered probit regression using (in *Stata*) the commands:

```
.global indvar msat foreign female emecon emoss emns emh am1 am2 am3 phy1 phy2 chem1 chem2
```

```
.oprobit highestmath $indvar
```

5. Use the result of this estimation to complete Table 3.[37]

highestmath	Coef.	Std. Err.	z	P>z	[95% Conf. Interval]	
msat1						
foreign						
female						
emecon						

emoss						
emns						
emh						
am1						
am2						
am3						
phy1						
Phy2						
chem1						
chem2						
_cut1						
_cut2						

_cut3						
_cut4						
_cut5						
_cut6						
Observations						
Log likelihood						
LR χ^2(14)						
Prob > χ^2						
Pueudo-R^2						

Table 2.32. Results of *Stata* ordered-probit regression.

6. Compare your results with the table reported in the article. The table in the article is Table II on page 193 and is reproduced in Figure 3. What we are interested in is

comparing column 4 in Figure 3 with columns 2 and 4 in Table 3. Table 4 below offers a model for this comparison.

Figure 2.32.

Table II. Ordered probit estimates of level of calculus attained[a]

Variable[b]	Expected sign	Students taking MICRO-2		Students taking MACRO-2	
		Mean (SD)	Coefficient (t-value)	Mean (SD)	Coefficient (t-value)
Constant		—	−3·09 (5·48)	—	−2·62 (3·95)
SAT-math × 10^{-2}	+	6·25 (0·60)	0·50[d] (6·12)	6·25 (0·60)	0·48[d] (5·23)
Foreign lang. proficiency [1,0]	−	0·11 (0·32)	0·02 (0·14)	0·09 (0·29)	0·23 (1·22)
Sex (female = 1; male = 0)	?	0·39 (0·49)	0·25[d] (2·59)	0·36 (0·48)	0·22[e] (1·96)
Expected major:					
Economics	?	0·34 (0·48)	−0·11 (0·86)	0·36 (0·48)	−0·18 (1·31)
Other social science	?	0·17 (0·38)	−0·29[e] (1·99)	0·15 (0·36)	−0·27 (1·59)
Natural science	+	0·21 (0·41)	0·43[d] (3·10)	0·20 (0·40)	0·32[e] (2·05)
Humanities	−	0·07 (0·25)	−0·37[e] (1·78)	0·07 (0·26)	−0·39[e] (1·80)
Years of HS Advanced Math (Y_m)					
$1 \leqslant Y_m < 2$	+	0·49 (0·50)	0·24 (1·07)	0·49 (0·50)	−0·00 (0·02)
$Y_m = 2$	+	0·45 (0·50)	0·93[d] (4·04)	0·45 (0·50)	0·67[d] (2·83)
$Y_m > 2$	+	0·01 (0·11)	0·77[e] (1·70)	0·01 (0·11)	0·28 (0·55)
Years of HS physics (Y_p)					
$1 \leqslant Y_p < 2$	+	0·67 (0·47)	0·26[d] (2·71)	0·67 (0·47)	0·27[d] (2·50)
$Y_p \geqslant 2$	+	0·02 (0·14)	0·38 (1·07)	0·01 (0·11)	−0·11 (0·20)
Years of HS chemistry (Y_c)					
$1 \leqslant Y_c < 2$	+	0·82 (0·39)	−0·12 (0·69)	0·82 (0·39)	−0·18 (0·75)
$Y_c \geqslant 2$	+	0·12 (0·32)	0·17 (0·75)	0·13 (0·34)	0·20 (0·75)
TRUNCATION POINTS[c]					
(1)	+		0·27[d]		0·21[d]

Results of ordered probit regression as reported in Butler, et al.

Table 4. Comparison of ordered probit estimations.

	Our estimates		Butler, et al. estimates	
	Estimate	z	Estimate	t-value
msat1			0.05	6.12
foreign			0.02	0.14
female			0.25	2.59
emecon			-0.11	0.86
emoss			-0.29	1.99
emns			0.43	3.10
emh			-0.37	1.78

am1			0.24	1.07
am2			0.93	4.04
am3			0.77	1.70
phy1			0.26	2.71
Phy2			0.38	1.07
chem1			-0.12	0.69
chem2			0.17	0.75
Intercept			-3.09	5.48
_cut1			0.27	7.29
_cut2			0.33	8.16
_cut3			1.52	20.32
_cut4			1.79	23.07
_cut5			2.04	23.72

_cut6				

Table 2.33. Comparison of ordered-probit estimations.

7. It is easy to see from Table 4 is that almost without exception the estimates of the parameters and their t-ratios are very similar. The exception arises with the estimates of the truncation points (_cut# in the *Stata* results). We will have to figure out what these are estimates of in order to make sense of them. Figure 1 shows the "cutoffs" that are being estimated. Footnote c in the BFS Table II on page 193 (shown in Figure 3) offers a useful observation:

> *In an ordered probit, an underlying, normally distributed, latent variable has a mean which is a function of observable variables. The latent variable gives rise to a set of observed dummy variables for ordered categories based on ranges between unobserved but estimable truncation points which correspond to levels of effort, ability, or other factors reflected in the explanatory variables. If L categories are observed, there are L − 1 truncation points, of which the first is normalized to be zero, so that L − 2 truncation points are estimated and reported in the table. The values correspond to standard deviations of the latent normally distributed variable.*

The key idea is that the values of cutoffs are relative and can be normalized around any value. Notice that the *Stata* results do not report an intercept term but do report six cutoff values. Moreover, the difference between the estimate by *Stata* for the first cutoff (3.08402) and the estimate for the second cutoff (3.356916) is equal to 0.272896, which is itself equal to the first truncation point reported by BFS (1998: 193). Use Table 5 to report the difference between the first cutoff value and each of the cutoff points reported by *Stata*.

Cutoff	Estimate	Estimate - _cut1	BFS Truncation Points
_cut1	3.0840		
_cut2	3.3569		0.27
_cut3	3.4146		0.33
_cut4	4.6013		1.52
_cut5	4.8774		1.79
_cut6	5.1202		2.04

Table 2.34. Reconciling *Stata* estimates of cutoff points with Butler, et al.'s truncation points.

The second part of the reconciliation of the two sets of results is to compute the t-ratios. To do this we need to compute the standard deviation of the estimates of the cutoff points reported by *Stata*. To do this we need to retrieve the variance-covariance matrix from the regression. First, let's see what we are interested in computing. Let $\hat{\beta}_i$ be the estimate of the *i* th cutoff point. In column 3 of Table 5 you computed $\hat{\alpha}_i = \hat{\beta}_i - \hat{\beta}_1$ for *i* = 2,…,6 . The variance of the new variable is:

(2.92)

$$V\left(\hat{\alpha}_i\right) = V\left(\hat{\beta}_i\right) - 2Cov\left(\hat{\beta}_i \hat{\beta}_1\right) + V\left(\hat{\beta}_1\right) = \sigma_i^2 - 2\sigma_{i1} + \sigma_1^2$$

The variance-covariance matrix will give us estimates of these variances and covariances. When there are *j* parameters in a regression equation, this matrix is defined to be:

$$\hat{\Sigma} = \begin{bmatrix} \hat{\sigma}^2_{\beta_1} & \hat{\sigma}_{\beta_1 \beta_2} & \cdots & \sigma_{\beta_1 \beta_k} \\ \hat{\sigma}_{\beta_2 \beta_1} & \hat{\sigma}^2_{\beta_2} & \cdots & \hat{\sigma}_{\beta_2 \beta_k} \\ \vdots & \vdots & \ddots & \\ \sigma_{\beta_k \beta_1} & \sigma_{\beta_k \beta_2} & \cdots & \hat{\sigma}^2_{\beta_k} \end{bmatrix}.$$

If you type the command .vce, *Stata* will report $\hat{\Sigma}$ as shown in Figure 4. We need the section of this matrix shown in Part A of Table 6. Use equation (5) to estimate the standard errors of the estimates of the cutoff points and complete Part B of Table 6 and compares the t-ratios with the values reported by Butler, et al. (and shown in the last column 4 of Table 6). Are you satisfied that we have been able to come reasonably close to the results reported in the article?

Figure 2.33.

	msat	foreign	female	emecon	emoss	emns	emh	am1	am2	am3	phy1	phy2	chem1	chem2	_cut1	_cut2	_cut3
msat	0.007																
foreign	-0.001	0.020															
female	0.001	-0.002	0.009														
emecon	0.000	0.000	0.001	0.015													
emoss	-0.001	0.000	-0.001	0.009	0.021												
emns	0.000	-0.001	0.000	0.009	0.009	0.019											
emh	0.000	-0.002	0.000	0.009	0.009	0.009	0.040										
am1	0.000	-0.002	-0.001	0.000	0.001	0.002	0.002	0.047									
am2	-0.001	-0.001	-0.001	-0.001	0.001	0.001	0.002	0.043	0.048								
am3	-0.004	0.002	0.000	-0.003	0.000	-0.006	-0.007	0.042	0.044	0.178							
phy1	-0.001	0.001	0.001	0.001	0.000	0.000	0.003	0.000	0.000	-0.002	0.010						
phy2	0.000	0.001	0.000	0.001	0.000	-0.001	0.000	0.001	0.001	-0.006	0.007	0.091					
chem1	0.000	0.001	0.000	0.001	0.000	0.000	0.001	0.000	0.000	0.004	0.000	0.004	0.033				
chem2	-0.001	0.002	0.000	0.000	0.000	-0.002	0.001	0.000	-0.002	0.006	0.000	0.005	0.030	0.047			
_cut1	0.040	-0.006	0.012	0.010	0.006	0.008	0.012	0.043	0.033	0.018	0.002	0.009	0.029	0.025	0.329		
_cut2	0.041	-0.006	0.012	0.010	0.006	0.008	0.012	0.043	0.034	0.018	0.002	0.009	0.029	0.026	0.329	0.330	
_cut3	0.041	-0.006	0.012	0.010	0.006	0.008	0.012	0.043	0.034	0.018	0.002	0.009	0.029	0.026	0.329	0.330	0.331
_cut4	0.041	-0.006	0.012	0.010	0.005	0.009	0.011	0.043	0.035	0.020	0.003	0.010	0.029	0.026	0.332	0.333	0.334
_cut5	0.041	-0.006	0.012	0.010	0.005	0.009	0.011	0.043	0.035	0.020	0.003	0.010	0.029	0.026	0.333	0.334	0.334
_cut6	0.041	-0.006	0.013	0.010	0.005	0.009	0.011	0.043	0.035	0.020	0.003	0.010	0.029	0.026	0.333	0.334	0.335

Stata estimate of the variance-covariance matrix.

<table>
<tr><td colspan="7"><b>Part A. Relevant portion of the variance-covariance matrix.</b></td></tr>
<tr><td></td><td>_cut1</td><td>_cut2</td><td>_cut3</td><td>_cut4</td><td>_cut5</td><td>_cut6</td></tr>
<tr><td>_cut1</td><td>0.329</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>_cut2</td><td>0.329</td><td>0.330</td><td></td><td></td><td></td><td></td></tr>
<tr><td>_cut3</td><td>0.329</td><td>0.330</td><td>0.331</td><td></td><td></td><td></td></tr>
<tr><td>_cut4</td><td>0.332</td><td>0.333</td><td>0.334</td><td>0.341</td><td></td><td></td></tr>
<tr><td>_cut5</td><td>0.333</td><td>0.334</td><td>0.334</td><td>0.341</td><td>0.343</td><td></td></tr>
<tr><td>_cut6</td><td>0.333</td><td>0.334</td><td>0.335</td><td>0.342</td><td>0.343</td><td>0.345</td></tr>
<tr><td colspan="7"><b>Part B. Calculation of the t-ratios (with comparison of values reported in BFS)</b></td></tr>
</table>

	V($\hat{\beta}$)	St. Dev.($\hat{\beta}$	t-ratio	BFS t-ratio	
_cut2				7.29	
_cut3				8.16	
_cut4				20.32	
_cut5				23.07	
_cut6				23.72	

Table 2.35. Calculation of the t-ratios for the cutoff estimates.

8. The next step in the process is to generate the term we will use in the estimation of the grade regression to account for the potential sample selection bias. To do this we will need to find a reference in the literature that offers a clear description of what we need to do. As it turns out, a reasonable explanation of the appropriate estimation technique is available in Jimenez and Kugler (1987). Since much of what follows comes directly from this article, I highly recommend you read it yourself.

The gist of the method suggests that the potential sample bias is accounted for by an inverse Mills ratio for each of the categories. What we need to do is calculate:

(2.93)

$$\hat{\lambda}_i = \frac{\phi\left(\hat{\mu}_j - \hat{z}_i^*\right) - \phi\left(\hat{\mu}_{j+1} - \hat{z}_i^*\right)}{\Phi\left(\hat{\mu}_{j+1} - \hat{z}_i^*\right) - \Phi\left(\hat{\mu}_j - \hat{z}_i^*\right)}$$

for the category that the individual actually is in. What we will do is calculate (6) for all of the categories and then sum the product of this number and a dummy variable indicating if a course is the highest math class completed by an individual. Since the dummy variables will equal 0 for math categories an individual is not in, the resulting sum will preserve the value of (6) that is associated with the category the individual does belong to.

It is clear from (6) that we will need to retain the 6 cutoffs. We can do this with the commands:

```
. generate cutoff1 = _b[_cut1]

. generate cutoff2 = _b[_cut2]
. generate cutoff3 = _b[_cut3]
```

```
. generate cutoff4 = _b[_cut4]
```

```
. generate cutoff5 = _b[_cut5]
```

```
. generate cutoff6 = _b[_cut6]
```

Technically, this step is not necessary since the parameter estimates are preserved until the next regression is estimated; I suggest doing this purely as a precaution.

9. Preserve the predicted values of the ordered-probit using the command:

```
. predict zhat, xb
```

```
. predict phat1 phat2 phat3 phat4 phat5 phat6 phat7, p
```

These two commands will generate for each observation the predicted mean category of math classes and the probability that this individual will fall in each category. To see what is going on we will retrieve some representative values of these variables and then graph them for one individual. Table 7 reports these values for 10 individuals in the sample. Now consider individual 2. Fitting a normal distribution with a mean of 4.25 and using the critical values from our estimation yields the probabilities that the individual is in each of the categories. For example, the probability that individual 1 will have completed no math classes is equal to 0.1223. Figure 5 illustrates the results for individual 1. The dashed vertical lines are the six cutoff values that are the same for each individual. The solid vertical line is the zhat for individual 1. The heavy blue line represents the normal probability density function for this individual. While, there is, of course, a different probability distribution for each individual, the cutoff values are the same for all members of the sample.

Observation	Highest Math Class	zhat	Pr(0)	Pr(1)	Pr(2)	Pr(3)	Pr(4)	Pr(5)	Pr(6)
1	3	3.9657	0.1890	0.0824	0.0194	0.4467	0.0816	0.0568	0.1241
2	0	4.2507	0.1217	0.0640	0.0158	0.4355	0.0975	0.0731	0.1923
165	0	3.5982	0.3036	0.1011	0.0225	0.4149	0.0575	0.0364	0.0640
166	6	4.6914	0.0540	0.0370	0.0098	0.3633	0.1097	0.0922	0.3340
214	3	3.4533	0.3560	0.1056	0.0229	0.3900	0.0483	0.0294	0.0478

215	3	4.0840	0.1587	0.0749	0.0180	0.4459	0.0887	0.0637	0.1501
225	3	3.5250	0.3296	0.1036	0.0228	0.4031	0.0528	0.0328	0.0553
226	3	3.6990	0.2693	0.0969	0.0219	0.4285	0.0641	0.0417	0.0776
453	3	3.9713	0.1875	0.0820	0.0194	0.4468	0.0819	0.0571	0.1253
454	5	4.1650	0.1399	0.0697	0.0170	0.4422	0.0932	0.0684	0.1697
495	3	4.4168	0.0913	0.0533	0.0135	0.4151	0.1043	0.0816	0.2409
496	0	2.9811	0.5410	0.1055	0.0212	0.2797	0.0236	0.0127	0.0162
526	0	2.9247	0.5633	0.1039	0.0207	0.2653	0.0214	0.0114	0.0141
527	3	3.9757	0.1863	0.0817	0.0193	0.4469	0.0822	0.0574	0.1262

Table 2.36. Predicted values of the ordered probit regression.

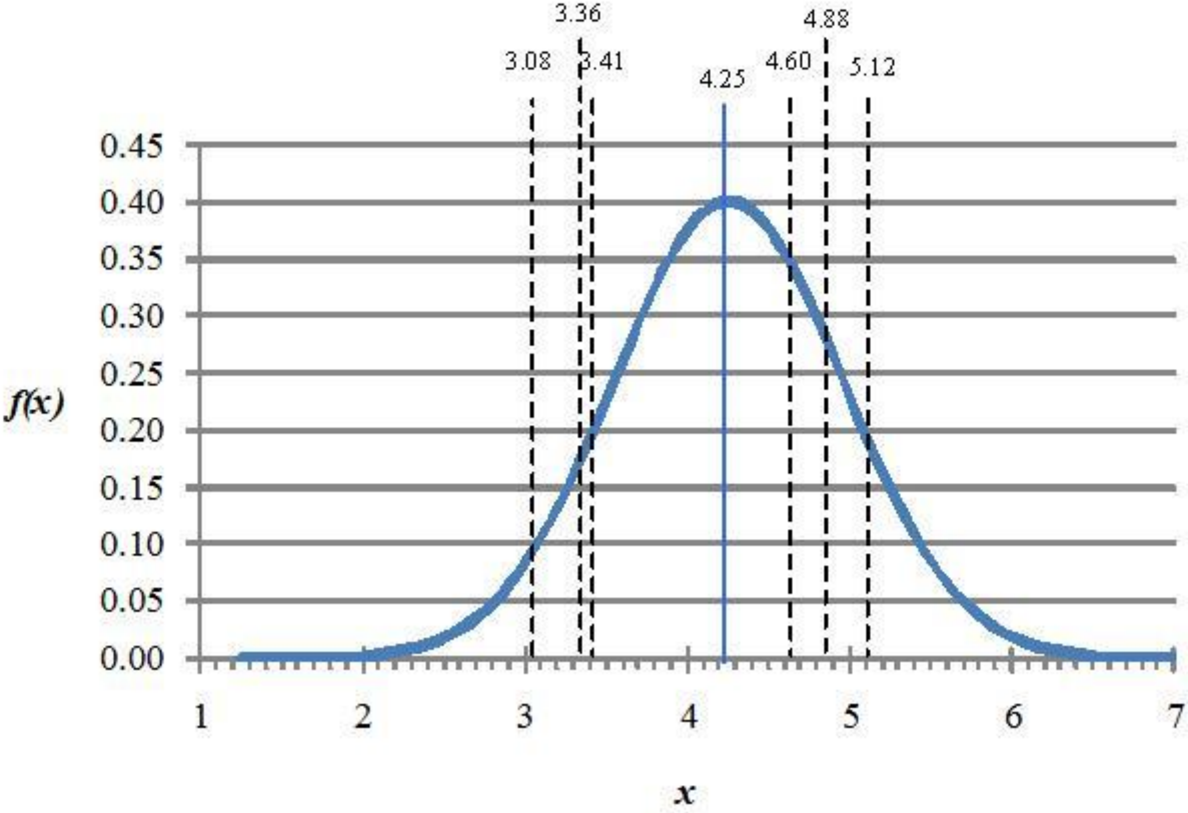
Now we are ready to calculate (6). The commands are:

```
.generate lambda0 = (-normden(cutoff1-zhat))/(norm(cutoff1-zhat)-norm(-zhat))
```

```
.generate lambda1 = (normden(cutoff1-zhat)-normden(cutoff2-zhat))/(norm(cutoff2-zhat)-norm(cutoff1-zhat))

.generate lambda2 = (normden(cutoff2-zhat)-normden(cutoff3-zhat))/(norm(cutoff3-zhat)-norm(cutoff2-zhat))

.generate lambda3 = (normden(cutoff3-zhat)-normden(cutoff4-zhat))/(norm(cutoff4-zhat)-norm(cutoff3-zhat))

.generate lambda4 = (normden(cutoff4-zhat)-normden(cutoff5-zhat))/(norm(cutoff5-zhat)-norm(cutoff4-zhat))

.generate lambda5 = (normden(cutoff5-zhat)-normden(cutoff6-zhat))/(norm(cutoff6-zhat)-norm(cutoff5-zhat))

.generate lambda6 = (normden(cutoff6-zhat))/(1-norm(cutoff6)-norm(cutoff5-zhat))

.generate lambda = m170*lambda0 + m171a*lambda1 + m172a*lambda2 + m171b*lambda3 + m172b*lambda4 + m221a*lambda5+m221b*lambda6
```

One thing to notice in these calculations is that cutoff0 is assumed to be $-\infty$ and cutoff7 is assumed to be ∞.

Figure 2.34.


3.36
4.88
3.08
3.41
4.60
5.12
4.25
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
f(x)
1
2
3
4
5
6
7
x

The probability distribution of math class category for individual 2.

10. Now we are ready to estimate our regression explaining the grade that each individual received in intermediate microeconomics. Use Table 8 to report the regression results for four specifications of the model. The first question is can the null hypothesis of sample selection bias be rejected? How does this conclusion compare with BFS's conclusions? (See Table 9.) Second, since many of the potential explanatory variables like class size and scores on the SATs do not seem to be statistically significant, it is reasonable to focus our comments on the results reported in column (4) of Table 8.

What can you conclude about the impact of calculus on how well a student will do in intermediate microeconomics? Do the final grades earned in a majority of the math classes impact the grade earned in intermediate microeconomics? Do the grades earned in any of the math classes positively and significantly affect the grade earned in intermediate microeconomics? Can you explain the impact of the freshman GPA on the grade earned in intermediate microeconomics? What, if any, is your bottom line conclusions about what matters in determining the grades earned in intermediate microeconomics?

Explanatory variables	**Model (1)**	**Model (2)**	**Model (3)**	**Model (4)**
Lambda			—	—
Sophomore		—		—
Senior		—		—
Same				
Skip		—		—
M171a				

M172a				
M171b				
M172b				
M221a				
M221b				
GE100				
GDE100				

GE101				
GDE101				—
GDE231				
Size				—
FGPA				
Female				

MSAT				**—**
VSAT				**—**
Grade in highest Math	**—**		**—**	**—**
class				
GM170		**—**		
GM171a		**—**		
GM172a		**—**		
GM171b		**—**		

GM172b		—		
GM221a		—		
GM221b		—		
Intercept				
F(28, 580)		—	—	—
Prob > F		—	—	—
F(27, 581)	—	—		—
Prob > F	—	—		—

F(20, 588)		—	—	
Prob > F		—	—	
F(19, 589)	—		—	—
Prob > F	—		—	—
R-Squared				
Root MSE				
Sample Size	609	609	609	609

Table 2.37. Determinants of Final Grade in Intermediate Microeconomics.

Robust t-ratios are in parentheses.

		MICRO-2	
Variable[a]	Expected sign	Mean (SD)	Coefficient(t-value)
Intercept	—	—	-1.64

			(3.48)
Selection bias correction	**+**	**-0.00**	**0.10**
(Predicted residual)		**(0.92)**	**(1.29)**
Level of calculus attained:			
Math 171A	**+**	**0.08**	**0.39**
		(0.27)	**(1.04)**
Math 172A	**+**	**0.02**	**-0.18**
		(0.13)	**(0.21)**
Math 171B	**+**	**0.37**	**1.02**[b]
		(0.48)	**(3.49)**
Math 172B	**+**	**0.07**	**1.52** [b]
		(0.25)	**(3.53)**
Math 221A	**+**	**0.05**	**1.33**[c]

		(0.22)	(2.27)
Math 221B or 222	+	0.14)	0.75[c]
		(0.35	(1.67)
Grade in last calculus course:			
Math 170	+	3.06	0.36[b]
		(0.70)	(4.36)
Math 171A	+	2.22	0.26[c]
		(0.86)	(2.21)
Math 172A	+	2.94	0.42
		(0.80)	(1.54)
Math 171B	+	2.62	0.10[c]
		(0.93)	(1.85)
Math 172B	+	2.63	-0.01

		(0.90)	(0.10)
Math 221A	+	3.10	-0.09
		(0.77)	(0.55)
Math 221B or 222	+	3.15	0.11
		(0.76)	(1.04)
Grade deflator of instructor in intermediate theory	+	-0.16	0.88b
course		(0.27)	(8.28)
Taken in Sophomore year	?	0.32	0.07
		(0.47)	(0.94)
Taken in Senior year	-	0.06	-0.02
		(0.24)	(0.13)
MICRO-1 and MICRO-2 in same academic year	+	0.35	0.04

		(0.48)	(0.46)
At least one semester between MICRO-1 and	-	0.27	0.13
MICRO-2		(0.44)	(1.85)
Grade in MACRO-1	+	2.73	0.20[b]
		(0.73)	(3.93)
Grade in MICRO-1	+	2.67	0.29[b]
		(0.74)	(5.93)
Instructor's grade deflator:			
MACRO-1	-	-0.32	-0.33[c]
		(0.20)	(2.20)
MICRO-1	-	-0.29	-0.11
		(0.16)	(0.53)

Class size (intermediate theory course)	**?**	**28.2**	**-0.002**
		(5.5)	**(0.45)**
Freshman Grade Point Average	**+**	**2.79**	**0.29[b]**
		(0.46)	**(3.04)**
Sex (female = 1; male = 0)	**?**	**0.39**	**0.13[c]**
		(0.49)	**(2.09)**
SAT-Math score x 10^{-2}	**+**	**6.25**	**0.12[c]**
		(0.60)	**(1.75)**
SAT-Verbal score x 10^{-2}	**+**	**5.56**	**0.04**
		(0.67)	**(0.78)**
OVERALL RESULTS			
Mean (SD) of dependent variable			

Adjusted R^2		0.44	
Number of observations		609	

Table 2.38. Results reported in BFS (p. 195).

[a] Omitted reference groups in MICRO-2 regression: attained Math 170; took MICRO-2 in Junior year; took MICRO-1 in spring, MICRO-2 next fall. [b] Significant at 0.01 level, one- or two-tailed test as appropriate. [c] Significant at 0.05 level, one- or two-tailed test as appropriate.

Exercises

Exercise 2.6.1.

Quite often health professionals request that a patient a report their perception of their health status on a scale of 0 to 10, where 0 is the lowest possible health status and 10 is the highest health status. This type of data set is best analyzed using ordered probit. In this exercise you will analyze a data set of responses to a survey made in Germany between 1984 and 1995. The question we are interested in analyzing is the respondent’s perception of their own health status.

The file Riphahn, Wambach, Million data.xls is an MS Excel file that contains 27,326 observations on 25 variables, one observation per line. The data are from Riphahn, Wambach, and Million (2003) and are also available on the web. The variables are defined in Table 10. As a first step you will need to load these data into Stata. However, due to the large sample size you will need to first expand the size of the memory that is available to Stata with the command: . set memory 1G. Here I have increased the memory to 1 gigabyte. This amount may be overkill but it seemed to be big enough on my computer to handle the data.

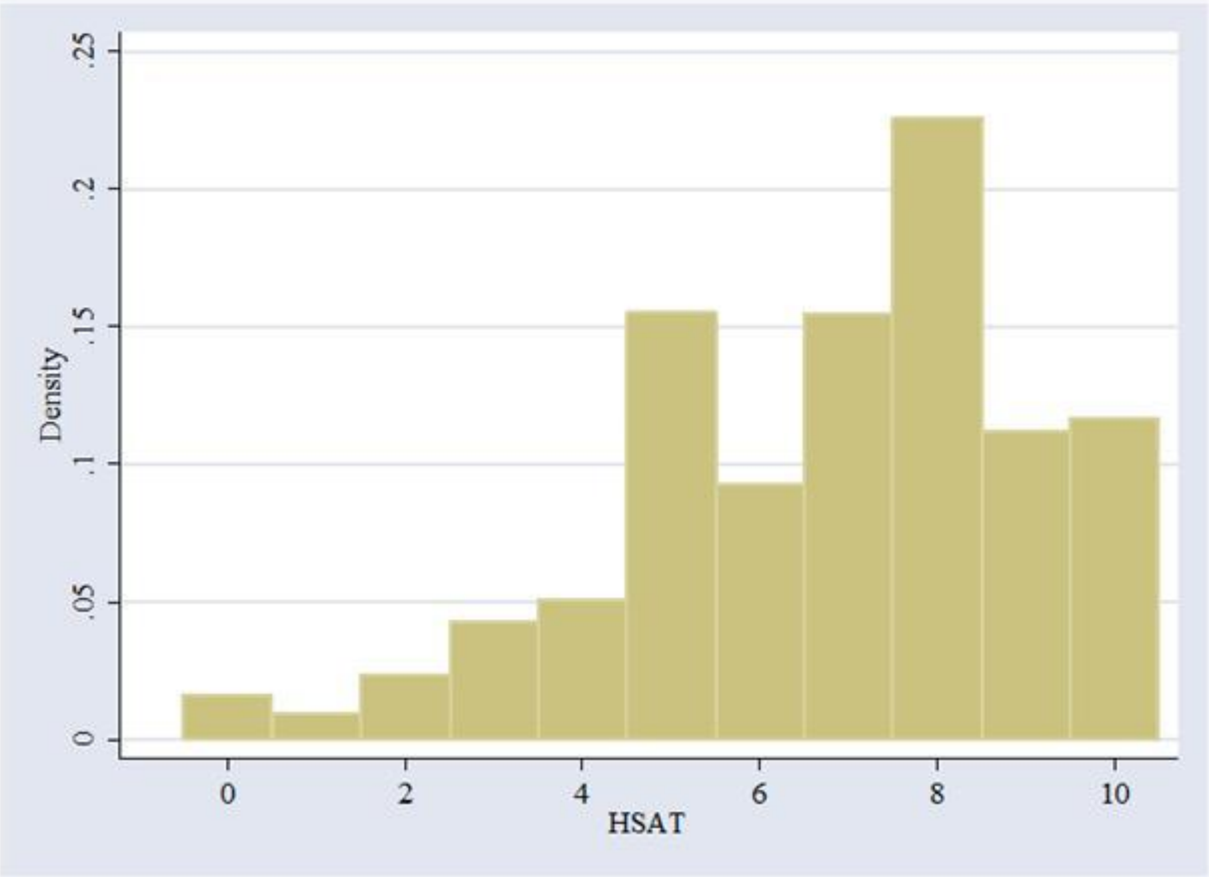
Column	Variable	Variable definition
A	ID	individual's ID number
B	Female	female = 1; male = 0
C	Year	calendar year of the observation
D	Age	age in years
E	HSAT	health satisfaction, coded 0 (low) - 10 (high)
F	Handdum	handicapped = 1; otherwise = 0

G	Handper	degree of handicap in percent (0 - 100)
H	HhnINC	household nominal monthly net income in German marks / 1000
I	HHKIDS	children under age 16 in the household = 1; otherwise = 0
J	Educ	years of schooling
K	Married	married = 1; otherwise = 0
L	Haupts	highest schooling degree is Hauptschul degree = 1; otherwise = 0
M	Reals	highest schooling degree is Realschul degree = 1; otherwise = 0
N	FachHS	highest schooling degree is Polytechnical degree = 1; otherwise = 0
O	Abitur	highest schooling degree is Abitur = 1; otherwise = 0
P	Univ	highest schooling degree is university degree = 1; otherwise = 0
Q	Working	employed = 1; otherwise = 0
R	BlueC	blue collar employee = 1; otherwise = 0
S	WhiteC	white collar employee = 1; otherwise = 0

T	Self	self employed = 1; otherwise = 0
U	Beamt	civil servant = 1; otherwise = 0
V	DocVis	number of doctor visits in last three months
W	HospVis	number of hospital visits in last calendar year
X	Public	insured in public health insurance = 1; otherwise = 0
Y	Addon	insured by add-on insurance = 1; otherwise = 0

Table 2.39. Variables in the German Socioeconomic Panel Data Set.

Figure 2.35.


Density
0
.05
.1
.15
.2
.25
0
2
4
6
8
10
HSAT

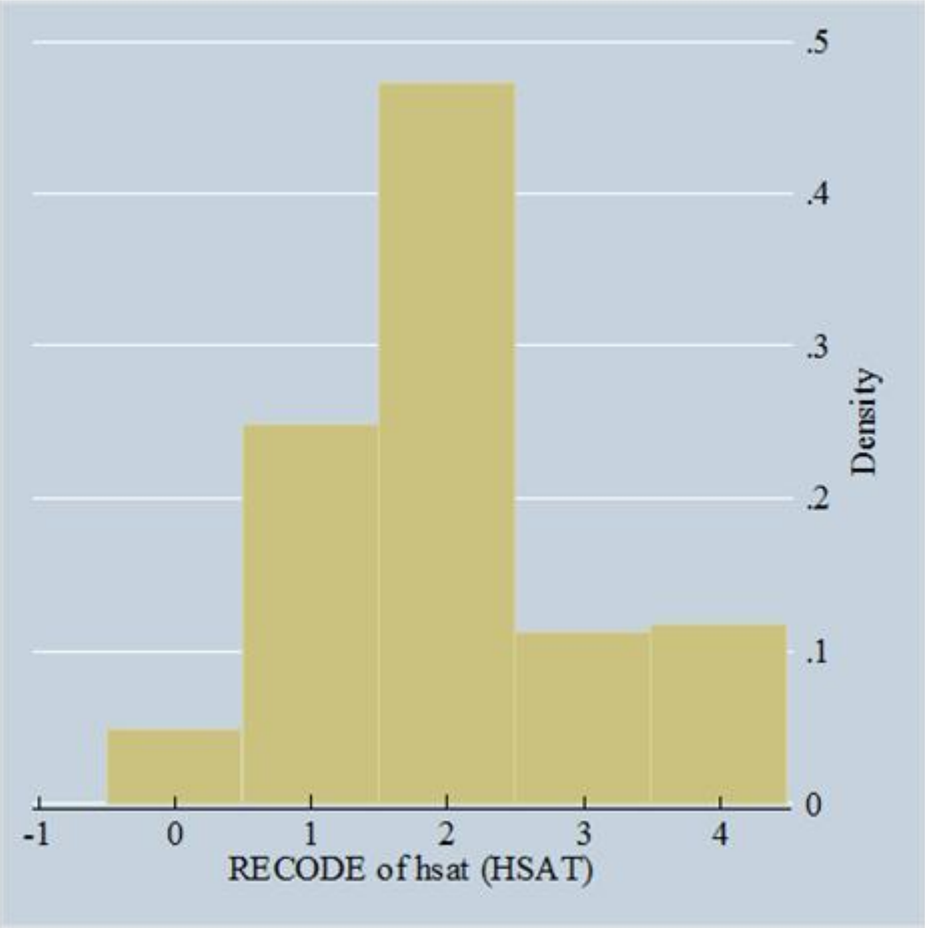
Distribution of responses on health status.

One of the major problems with survey indices is that the numbers seem to mean different things to respondents. One way to reduce this problem is to collapse the index into fewer outcomes by combining some of the responses together. However, anyway we do this is going to be ad hoc. Figure 6 shows a histogram of the responses to this question. Based on this graph, we will create 5 categories—(0) HSat = 0, 1, or 2; (1) HSat = 3, 4 or 5; (2) HSat = 6, 7, or 8; (3) HSat = 9; and (4) HSat = 10. We can create a new categorical variable called hsatnew with the command:

```
. recode hsat (0/2 = 0) (3/5 = 1) (6/8 = 2) (9 = 3) (10 = 4), generate(hsatnew)
```

Figure 7 shows the histogram of the new variable.

Figure 2.36.


.5
.4
.3
.2
.1
0
Density
-1
0
1
2
3
4
RECODE of hsat (HSAT)

The collapsed distribution of health status responses.

1. **Create a table of summary statistics for (1) health status, (2) age, (3) household income, (4) years of education, (5) marital status, and (6) number of children by year and sex. (You might want to use the command .bysort year female, list of variables).**

2. **Estimate an ordered probit regression for 1988 for health status (the new variable) using age, income, education, married, and kids as the explanatory variables. Here you might want to used the command: .oprobit hsatnew age hninc educ married hhkids if year==1988.**

3. **Use the predict newvariable, xb command to calculate the predicted mean values for each individual for the 1988 observations. Compare this histogram to one using the 1988 regression parameters to estimate xb for all years.**

4. **Estimate the ordered probit model for all of the years in the sample and put the results into a table like Table 11. (Here you might want to make use of the command: .bysort year: oprobit hsatnew varlist)**

Variable	1984	1985	1986	1987	1988	1991	1994
age							
income							
education							
married							
kids							
_cut1							
_cut2							
_cut3							
_cut4							
Observations							
LR χ^2(5)							
Prob > χ^2							

Log likelihood							
Pseudo-R^2							

Table 2.40. Sample table for part (d) of Exercise 1.

t-ratios are in parentheses.

References

Amemiya, T. (1985). *Advanced Econometrics* (Cambridge, MA: Harvard University Press).

Bourguignon, François, Martin Fournier, and Marc Gurgand (2007). Selection bias corrections based on the multinomial logit model: Monte Carlo comparisons. *Journal of Economic Surveys* 21(1): 174-205.

Butler, J. S., T. Aldrich Finegan, and John J. Siegfried (1998). Does more calculus improve student learning in intermediate micro- and macroeconomic theory?" *Journal of Applied Econometrics* 13(2): 185-202.

Chiburis, Richard and Michael Lokshin (2007). Maximum likelihood and two–step estimation of an ordered–probit selection model. *The Stata Journal* 7(2): 167-182.

Dahl, Gordon B. (2002). Mobility and the returns to education: testing a roy model with multiple markets. *Econometrica* 70(6): 2367–2420.

Dubin, Jeffrey A. and Daniel L. McFadden (1984). An econometric analysis of residential electric appliance holdings and consumption. *Econometrica* 52(2): 345–362.

Greene, William H. (1990). *Econometric Analysis* (New York: Macmillan Publishing Company).

Heckman, James J. (1979). Sample selection bias as a specification error. *Econometrica* 47(1): 153–161.

Jimenez, Emmanuel and Bernardo Kugler (1987). The earnings impact of training duration in a developing country an ordered probit selection model of Colombia's *Servicio Nacional de Aprendizaje* (SENA). *Journal of Human Resources* 22(2): 230-233.

Lee, Lung-Fei (1983). Generalized econometric models with selectivity. *Econometrica* 51(2): 507–512.

Maddala, G. S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics* (Cambridge: Cambridge University Press).

Main, B. and B. Reilly (1993). The employer size-wage gap: Evidence for Britain. *Economica* 60: 125–142.

McFadden, Daniel L. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (ed.) *Frontiers in Econometrics* (New York: Academic Press).

Newey, W. K. and Daniel L. McFadden (1994). Large sample estimation and hypothesis testing. In R. F. Engle and Daniel L. McFadden (eds.) *Handbook of Econometrics,* vol. IV (Amsterdam: North Holland).

Riphahn, Regina T., Achim Wambach, and Andreas Million (2003). Incentive effects in the demand for health care: a bivariate panel count data estimation. *Journal of Applied Econometrics* 18(4): 387-405

Schmertmann, Carl P. (1994). Selectivity bias correction methods in polychotomous sample selection models. *Journal of Econometrics* 60(1): 101–132.

Vella, Francis (1998). Estimating models with sample selection bias. *The Journal of Human Resources* 33(1): 127-169.

[7] J. S. Cramer (2003) *Logit Models from Economics and Other Fields* (Cambridge: Cambridge University Press): 10.

[8] For a full discussion of this model see Ladd, G. W. (1966) "Linear Probability Functions and Discriminant Functions," *Econometrica* 34: 873-888.

[9] The assumption that the variance is equal to 1 is due to technical considerations. See [Cramer, 22].

[10] The pdf of a logistic distribution is $f(x) = \frac{\lambda e^{-\lambda x}}{\left(1 + e^{-\lambda x}\right)^2}$, where $\lambda = \frac{\pi}{\sqrt{3}} \approx 1.814$. See Cramer, 24-26 for a fuller discussion of the logistic distribution.

[11] See Stata Library, Categorical and Count Data Analysis Utilities for useful utilities and an excellent discussion of how to interpret categorical and count regression results at http://www.ats.ucla.edu/stat/stata/library/longutil.htm/ (accessed July 19, 2009).

[12] The phrase “(Assumption: . nested in full)” tells you the name of the regression is the unrestricted model (full) and offers you a hyperlink to call this regression up to the screen.

[13] The gradient is a vector of first-derivatives. In this case it is a vector of the first-derivatives with respect to each parameter estimate $\left(\text{i.e., } \hat{\beta}_i\right)$. To obtain the ML estimate, we have to set these first-derivatives equal to zero.

[14] See StataCorp [2003:119-130] for more detail on this command.

[15] If the OLS parameter estimates are unbiased but the standard error estimates are, then applying the Cochran-Orcutt adjustment should change the estimates of the standard errors without changing the estimates of the equation parameters substantially.

[16] That is, we assume $\varepsilon_t \sim \left(0, \sigma^2\right)$, where the distribution is not specified, and $E\left(\varepsilon_i \varepsilon_j\right) = 0$ for all $i \neq j$.

[17] These methods make use of the mathematics of difference equations. See advanced texts like Enders (1995: pp. 68-77) for examples of the derivation of the conditions necessary for an ARMA(*p*, *q*) time-series to be stationary.

[18] AR(1) is the same as ARMA(1, 0)

[19] This set of graphs is from Enders (2005: p. 79).

[20] ARIMA means AutoRegressive Integrated Moving Average. See Enders (2005: 67) for a discussion of what integrated means. We can ignore it given our limited purposes.

[21] Another way to think about this point is to remember that, unlike the fixed-effects model, the random-effects does not use dummy variables to summarized the unknown characteristics; thus, there is no problem with multicollinearity.

[22] See Cameron and Trivedi (2005: 705] for a detailed discussion of the random-effects estimator.

[23] R-squared is in quotes in this line because these R-squareds do not have all the properties of OLS R-squareds.

[24] Because the mean and variance of the standard normal distribution are 0 and 1, respectively, its probability density function (pdf) is and the cumulative probability function is .

[25] A *stochastic variable* is a *random variable*—i.e., a variable whose value is determined as a result of a process involving an uncertain outcome.

[26] Greene suggested this example in 1990 when most people paid their bills with checks. Currently it would not be such a good example because of the development of electronic payment of bills.

[27] In these notes I discuss only what is known in the literature as the *order condition* for identification. The order condition is necessary for identification. Another condition—the *rank condition*—is a sufficient condition. See Greene (1990: Chapter 19, especially pp. 600-609) for a fuller discussion of simultaneous-equation models and the identification problem.

[28] Using one of the exogenous variables in an equation as an instrument will create perfect multicollinearity in the first stage regression.

[29] We exclude Equation (15) from this discussion because it is under-identified and, thus, cannot be estimated.

[30] The advantage of the ivreg command is that it allows you to estimate a single equation of a system of equations without fully specifying the equations in the rest of the model. Use the command reg3 if you want to specify the whole model or use Three-Stage Least Squares.

[31] The description of the command “ivreg depvar [varlist1] (varlist2=varlist_iv)” in the *Stata* help file is “ivreg fits a linear regression model using instrumental variables (or two-stage least squares) of depvar on varlist1 and varlist2 using varlist_iv (along with varlist1) as instruments for varlist2. In the language of two-stage least squares, varlist1 and varlist_iv are the exogenous variables and varlist2 the endogenous variables.”

[32] The model and data for this problem first appeared in Maddala, G. S. (1988) *Introductory Econometrics* (New York: Macmillan Publishing Company): 331-317.

[33] See Berndt, Ernst R. (1991) *The Practice of Econometrics* (Reading, MA: Addison-Wesley Publishing Company): 375-380.

[34] Butler, J. S., T. Aldrich Finegan, and John J. Siegfried (1998). Does more calculus improve student learning in Intermediate Micro- and Macroeconomic Theory? *Journal of Applied Econometrics* 13(2):185-202.

[35] This particular notation implies that there are *k* – 1 explanatory variables.

[36] See Greene (1990): 704.

[37] One way to make the conversion from the *Stata* output to the neater table relatively easily is to follow these steps: (1) replace each double space by a single space until there were none left; (2) replace each space with a tab (^t); (3) convert the material into a table using the "Insert/Table" command with a tab as the separator; and (4) clean up the table by moving the data into an *Excel* file, fixing the formatting, and returning the data to the *Word* file (alternatively, you can use formatting commands in *Stata* to control how the output appears).

Chapter 2. Advanced topics in econometrics

2.1. Logit and Probit Regressions[*]

Logit and Probit models

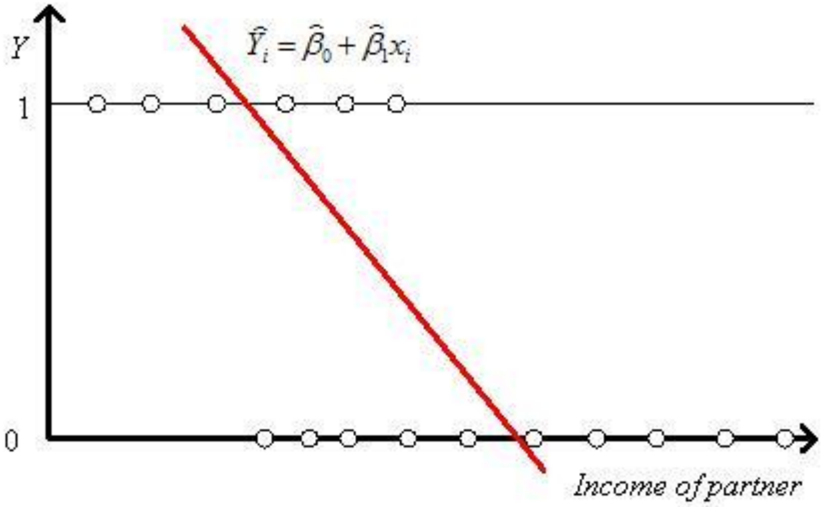
Introduction

Consider a model that “explains” whether a wife enters the work force. It is straight forward to think of potential explanatory variables—her potential wage rate, the income of her partner, the number of children under the age of 6 in the household, and the number of children in the household between the ages of 6 and 18 are candidates to be independent variables used to explain the wife’s decision to enter the labor force. The dependent variable, *Y*, however, is a dummy variable because the wife chooses either to enter the labor force (*Y* = 1) or not to enter the labor force (*Y* = 0). An OLS model of the form:

$$(2.1)\ Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

does not make sense. Figure 1 shows what the data of this model might look like when graphed against one of the explanatory variables. Figure 1 also includes the regression line that an OLS estimation of (1) will yield. It is easy to see one problem with this approach—the predicted values of *Y* that can be greater than 1 and less than 0. In addition, special properties must be attributed to the error term and it is the simple properties ascribed to the error term that make the OLS model so attractive.[7]

Figure 2.1. Linear regression line for a discrete dependent variable



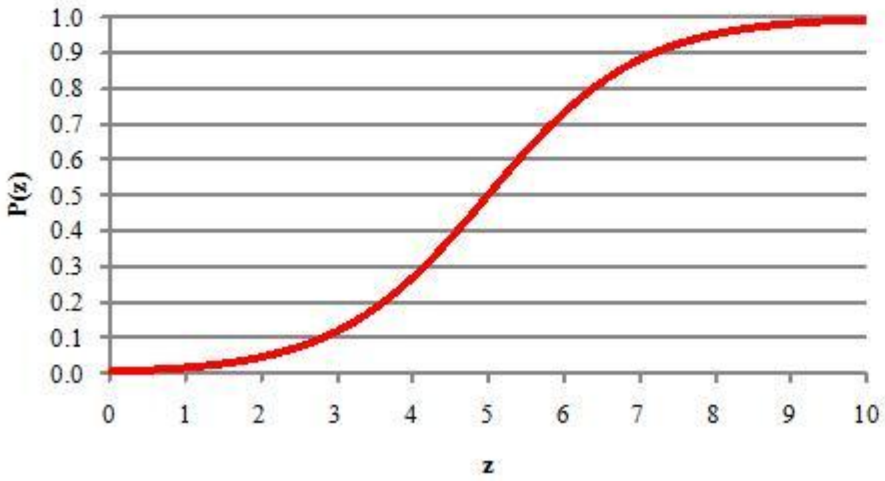
The linear regression line can be a poor representation of a discrete dependent variable.

The logit model

There does exist another approach to the modeling problem—assume that the dependent variable is *the probability that the wife is in the labor force*. For instance we might assume that we have a linear probability model of the form $\Pr(x_i) = \beta_0 + \beta_1 x_i + \varepsilon_i$. This model can be estimated reasonably successfully if the observed frequencies are well away from their bounds of 0 and 1.[8] However, is more appealing to assume that the probability varies monotonically with *x* and remains within the bounds of [0,1], as shown in Figure 2. This S-shaped curve is known as the *sigmoid curve* and can be represented algebraically for some variable *z* by:

$$\Pr(z) = \frac{e^z}{1 + e^z}.$$

Figure 2.2. The signoid function.



The signoid function forces the dependent variable to be between 0 and 1.

We can simplify our analysis by using a bit of algebra. First, the inverse probability is $1 - \Pr(z) = 1 - \frac{e^{z}}{1 + e^{z}} = \frac{1}{1 + e^{z}}.$ **Thus,**

(2.2)

$$\frac{\Pr(z)}{1 - \Pr(z)} = \frac{\frac{e^{z}}{1 + e^{z}}}{\frac{1}{1 + e^{z}}} = e^{z}.$$

Taking the natural logarithm of (2) gives $\ln\left(\frac{\Pr(z)}{1 - \Pr(z)}\right) = z.$ **Assuming that *z* is a linear function of *x* (and, more generally, of other variables) gives the *logit* model:**

(2.3)

$$\ln\left(\frac{\Pr(x_i)}{1-\Pr(x_i)}\right)=\beta_0+\beta_1 x_i+\varepsilon_i.$$

We can estimate the parameters of this model using maximum likelihood methods. In the *probit* model the error term is assumed to be normally distributed with a mean of zero and a unit variance.[9] In the logit model the error term is assumed to have a *standardized logistic distribution*. This distribution has a mean of 0 and a variance of 1 and is very similar to a normal distribution with the same mean and variance.[10] While the choice of which model to use generally is personal, it should be noted that the ratio of the parameter of a logit model to the parameter of a probit model (using the same data set) usually varies between 1.6 and 2.0. We focus on the logit model in the balance of this discussion.

Interpretation of the logit model parameters

The interpretation of the economic meaning of the parameter values in a logit model is not very obvious.[11] One simple, but not often used, interpretation comes from taking the first-derivative of (3) with respect to *x*:

(2.4)

$$\ln(\text{odds}\,Y=1)=\beta_0+\beta_1 x+\varepsilon \Rightarrow \frac{\partial \ln(\text{odds}\,Y=1)}{\partial x}=\beta_1.$$

Thus, in the labor force participation model one interpretation is that β_1 is equal to the change in the natural logarithm of the odds that the wife is in the labor force due to a one unit change in the independent variable x. This interpretation is both awkward and not really economically informative.

Stata offers two command for estimating a logit regression—logit and logistic. The logit command returns the parameter estimates as shown in (3). The logistic command returns the odds ratio rather than the parameter estimates. The odds ratio is equal to e^{β_1}. Thus, one can go from the odds ratio reported by the logistic command to the parameter estimates merely by taking the natural logarithm of the odds ratio. The interpretation of the odds ratio is straightforward. For example, assume that *y* = 1 means that the birth weight of an individual is less than 2,500 grams and *y* = 0 means that the birth weight is greater than 2,500 grams. A logit parameter estimate of -0.27 is equivalent to an odds ratio of 0.97 (i.e., $e^{-0.27} = 0.97$). An odds ratio of 0.97 means that odds of a baby being underweight are 0.97 times those of the odds of a baby being of normal weight. To see what is being said re-write (2.3) as:

$$\frac{\Pr(x)}{1-\Pr(x)} = e^{\beta_0 + \beta_1 x + \varepsilon}.$$

A one unit change in *x* implies that:

$$\frac{\Pr(x+1)}{1-\Pr(x+1)} = e^{\beta_0 + \beta_1 (x+1) + \varepsilon}$$

or

$$\frac{\Pr(x+1)}{1-\Pr(x+1)} = e^{\beta_0 + \beta_1 x + \varepsilon} e^{\beta_1}$$

or

$$\frac{\Pr(x+1)}{1-\Pr(x+1)} = e^{\beta_1} \left(\frac{\Pr(x)}{1-\Pr(x)} \right).$$

Thus, $e^{\widehat{\beta}_1}$ is equal to the percent change in the odds that *y* equals 1 (a baby is born underweight) due to a one unit change in *x*.

The logistic command reports $e^{\widehat{\beta}_1}$ while the logit command reports $\widehat{\beta}_1$. Because of the ease of interpretation of the odds ratio, *Stata* argues that the logistic command is the proper one to use.

Elasticities

Another route to follow is to try to find something that can be interpreted as an elasticity. Elasticities are important enough topic in economics for us to discuss them here in some detail. The reason they are so attractive to economists is that they have no units and, thus, can be compared across different commodities. For instance, it is quite reasonable to compare the demand elasticity for apples with the demand elasticity for pearl necklaces in spite of the fact that the units of measuring apples and necklaces are different. There are a few important ways that elasticities appear in regressions.

Linear regression elasticities

In a linear regression of the form (ignoring the subscripts and the error term)

$Y = \beta_0 + \beta_1 x$,

we would calculate the elasticity of *Y* with respect to *x* to be

$$\eta_{Yx} = \frac{x}{Y}\frac{\partial Y}{\partial x} = \beta_1 \frac{x}{Y}.$$

Clearly, researchers need to choose the levels of *Y* and *x* at which to report this elasticity; it is traditional to calculate the elasticity at the means. Thus, economists typically report

$$\eta_{Yx} = \beta_1 \frac{\bar{x}}{\bar{Y}}.$$

Constant elasticities

Consider the following demand equation:

$$q = \alpha\, p^{-\beta} e^{\varepsilon}, \tag{2.5}$$

where *q* is the quantity demanded, *p* is the price the good is sold at, *α*,*β* > 0, and *ε* is an error term. The price elasticity of demand is given by

$$\eta_{qp} = \frac{p}{q}\frac{\partial q}{\partial p} = \frac{p}{\alpha p^{-\beta} e^{\varepsilon}}\left(-\beta \alpha p^{-\beta-1} e^{\varepsilon}\right) = -\beta.$$

In other words, this demand curve has a *constant price elasticity of demand* equal to − *β*. Moreover, we can convert the estimation of this equation into a linear regression by taking the natural logarithm of both sides of (5) to get ln*q* = ln*α* − *β*ln*p* + *ε*.

The logit equation and the quasi-elasticity

It is not appropriate to use the normal formula for an elasticity with (3) because the dependent variable is itself a number without units between 0 and 1. As an alternative it makes more sense to calculate the *quasi-elasticity*, which is defined as:

(2.6)

$$\eta(x) = x\frac{\partial \Pr(x)}{\partial x}.$$

Since

$$\ln\left(\frac{\Pr(x_i)}{1-\Pr(x_i)}\right) = \beta_0 + \beta_1\, x_i + \varepsilon_i,$$

we can calculate this elasticity as follows:

$$\frac{\partial\left(\ln\left(\frac{\Pr(x_i)}{1-\Pr(x_i)}\right)\right)}{\partial x} = \beta_1.$$

Focusing on the left-hand-side, we get:

$$\frac{1-\Pr(x_i)}{\Pr(x_i)}\frac{(1-\Pr(x_i))\frac{\partial \Pr(x_i)}{\partial x} + \Pr(x_i)\frac{\partial \Pr(x_i)}{\partial x}}{(1-\Pr(x_i))^2} = \beta_1$$

or

$$\frac{1}{\Pr(x_i)(1-\Pr(x_i))}\frac{\partial \Pr(x_i)}{\partial x} = \beta_1$$

or

(2.7)

$$\frac{\partial \mathrm{Pr}(x_i)}{\partial x} = \beta_1 \mathrm{Pr}(x_i)(1 - \mathrm{Pr}(x_i)).$$

Thus, we see from (6) that the quasi-elasticity is given by:

(2.8)

$$\eta(x_i) = \beta_1 x_i \mathrm{Pr}(x_i)(1 - \mathrm{Pr}(x_i)).$$

The quasi-elasticity measures the percentage point change in the probability due to a 1 percent increase of *x*. Notice that it is dependent on what value of *x* it is evaluated at. It is usual to evaluate (8) at the mean of *x*. Thus, the quasi-elasticity at the mean of *x* is:

$$\eta(\bar{x}) = \beta_1 \bar{x} \mathrm{Pr}(\bar{x})(1 - \mathrm{Pr}(\bar{x})),$$

where

$$\mathrm{Pr}(\bar{x}) = \frac{e^{\beta_0 + \beta_1 \bar{x}}}{1 + e^{\beta_0 + \beta_1 \bar{x}}}.$$

Hypothesis testing

The researcher using the logit model (and any regression estimated by ML) has three choices when constructing tests of hypotheses about the unknown parameter estimates—(1) the Wald test statistic, (2) the likelihood ratio test, or (3) the Lagrange Multiplier test. We consider them in turn.

The Wald test

The Wald test is the most commonly used test in econometric models. Indeed, it is the one that most statistics students learn in their introductory courses. Consider the following hypothesis test:

(2.9)

$$\mathrm{H}_0 : \beta_1 = \beta$$
$$\mathrm{H}_\mathrm{A} : \beta_1 \neq \beta.$$

Quite often in these test researchers are interested in the case when *β* = 0 —i.e., in testing if the independent variable's estimated parameter is statistically different from zero. However, *β* can be any value. Moreover, this test can be used to test multiple restrictions on the slope parameters for multiple independent variables. In the case of a hypothesis test on a single parameter, the t-ratio is the appropriate test statistic. The t-statistic is given by

$$t = \frac{\widehat{\beta}_i - \beta}{\mathrm{s.e.}\left(\widehat{\beta}_i\right)} \sim t_{n-k-1},$$

where *k* is the number of parameters in the mode that are estimated. The F-statistic is the appropriate test statistic when the null hypothesis has restrictions on multiple parameters. See Cameron and Trivedi (2005: 224-231) for more detail on this test. According to Hauck and Donner (1977) the Wald test may exhibit perverse behavior when the sample size is small. For this reason this test must be used with some care.

The likelihood ratio test

The likelihood ratio test is based on a comparison of the maximum log of likelihood function for the unrestricted model with the maximum log of likelihood function for the model with the restrictions implied by the null hypothesis. Consider the null hypothesis given in (9). Let *L*(*β*) be the value of the likelihood function when *β* $_1$ be the value of the likelihood function when is restricted to being equal to *β* and $L\left(\widehat{\beta}_1\right)$ **be the value of the likelihood function when there is no restriction on the value of *β*. Then the appropriate test statistic is**

$$LR = -2\left[\ln L(\beta) - \ln L\left(\widehat{\beta}_1\right)\right].$$

The likelihood ratio statistic has the Chi-square distribution *χ* ^{2} (*r*), where *r* is the number of restrictions. Thus, using a likelihood ratio test involves two estimations—one with no restrictions on the model and one with the restrictions implied by null hypothesis. Since the likelihood ratio test does not appear to exhibit perverse behavior with small sample sizes, it is an attractive test. Thus, we

will run through an example of how to execute the test using *Stata*. The example we are using is from the *Stata* manual, volume 2, pp. 353-355.

Example 2.1. Underweight births.

In this model we estimate a model that explains the likelihood that a child will be born with a weight under 2,500 grams (low). The eight explanatory variables used in the model are listed in Table 1. The model to be estimated is:

(2.10)

$$\ln\left(\frac{\Pr(Low)}{1-\Pr(Low)}\right)=\beta_1\,Age+\beta_2\,Lwt+\beta_3\,RaceB+\beta_4\,RaceO$$

$$+\beta_5\,Smoke+\beta_6\,Ptl+\beta_7\,Ht+\beta_8\,Ui+\varepsilon.$$

Also, we want to test the null hypothesis that the coefficients on Age, Lwt, Ptl, and Ht are all zero. The first step is to estimate the unrestricted regression using the command:

```
. logistic low age lwt raceb raceo smoke ptl ht ui
```

Variable name	Definition
Age	Age of mother
Lwt	Weight at last menstrual period
RaceB	Dummy variable =1 if mother is black; 0 otherwise
RaceO	Dummy variable = 1 if mother in neither white or black; 0 otherwise
Smoke	Dummy variable = 1 if mother smoked during pregnancy; 0 otherwise
Ptl	Number of times mother had premature labor
Ht	Dummy variable = 1 if mother has a history of hypertension; 0 otherwise

Ui	Dummy variable = 1 there is presence in mother of uterine irritability; 0 otherwise
Ftv	Number of visits to physician during first trimester

Table 2.1. Definition of the explanatory variables.

The results of this estimation are shown in column 2 of Table 2. Next we save the results of this regression with the command:

```
. estimates store full
```

where "full" is the name that we will refer to when we want to recall the estimation results from this regression. Now we estimate the logistic regression with the omitting the variables whose parameters are to be restricted to being equal to zero:

```
. logistic low raceb raceo smoke ui
```

The results of this estimation are reported in column 3 of Table 2. Finally we run the likelihood ratio test with the command:

```
. lrtest full .
```

Notice that we refer to the first regression with the word "full" and to the second regression with the second period. The results of this command are as follows:

```
Likelihood-ratio test LR chi2(4) = 14.42

(Assumption: . nested in full) Prob > chi2 = 0.0061
```

The interpretation of these results is that the omitted variables are statistically significant at the 0.6 percent level.[12]

Explanatory variable	Unrestricted model	Restricted model
Age of mother	-0.9732636	—
	(-0.74)	
Weight at last menstrual period	-0.9849634	—

	(-2.19)	
Dummy variable =1 if mother is black; 0 otherwise	3.534767	3.052746
	(2.40)	(2.27)
Dummy variable = 1 if mother in neither white or black; 0 otherwise	2.368079	2.922593
	(1.96)	(2.64)
Dummy variable = 1 if mother smoked during pregnancy; 0 otherwise	2.517698	2.945742
	(2.30)	(2.89)
Number of times mother had premature labor	1.719161	—
	(1.56)	
Dummy variable = 1 if mother has a history of hypertension; 0 otherwise	6.249602	—
	(2.64)	
Dummy variable = 1 if there is presence in mother of uterine irritability; 0 otherwise	2.1351	2.419131
	(1.65)	(2.04)
Log likelihood	-100.724	-107.93404
Number of observations	189	189
pseudo-R^2	0.1416	0.0801

Table 2.2. Estimation results for (2.10).

Note: Parameter estimates are odds ratios; z statistics are shown in parentheses.

The Lagrange multiplier test

The intuition behind the Lagrange multiplier (LM) test (or score test) is that the gradient of the log of the likelihood function is equal to zero at the maximum of the likelihood function.[13] If the null hypothesis in (2.9) is correct, then maximizing the log of the likelihood function for the restricted model is equivalent to maximizing the log of the likelihood function with the constraint specified by the null hypothesis. The LM test measures how close the Lagrangian multipliers of this constrained maximization problem are to zero—the closer they are to zero, the more likely that the null hypothesis can be rejected.

Economists generally do not make use of the LM test because the test is complicated to compute and the LR test is a reasonable alternative. Thus, as a practical matter the Wald test and the LR test are reasonable alternative test statistics to use to test most linear restrictions on the parameters. Moreover, since the calculations are relatively easy, it may make sense to calculate both test statistics to be sure they produce consistent conclusions. However, when the sample size is small, the LM test probably is preferred.

Goodness-of-fit measures

The standard measure of goodness-of-fit in the linear OLS regression model is R^2 . No such measure exists for non-linear models like the logit model. Several potential alternatives have been developed in the literature and are known collectively as pseudo-R^2. Many of these measures are discussed in McFadden (1974), Amemiya (1981), and Maddala (1983). In case any reader really cares about the pseudo-- R^2 , a practical approach is to report the value that the computer program reports.

One addition measure of goodness-of-fit is a measure called percentage correctly predicted. This variable is computed in one of several ways. One way is to use the observed values of the independent variable to forecast the probability the dependent variable equal one. Then, if the predicted probability is above some critical value, you assume that the predicted value of the dependent value is one. If it is below this value, you assume the predicted value of the dependent variable is zero. Then you construct a table that compares the predicted values of the dependent variable with the actual value of the dependent as shown in Table 3.

	Predicted	
Actual	$\widehat{Y} = 0$	$\widehat{Y} = 1$

$Y = 0$	n_{00}	n_{01}
$Y = 1$	n_{10}	n_{11}

Table 2.3. Percent correctly predicted.

The percentage correctly predicted is equal to the sum of the diagonal elements, that is, $n_{00} + n_{11}$, over the sample size. The main problem with this measure is that the choice of the cutoff point is arbitrary. Traditionally, a cutoff point used has been 0.5. However, there is no reason why this cutoff is the appropriate one. Cramer (2003, 67) suggests that a more appropriate cutoff point is the sample frequency—that is, $\frac{n_{10}+n_{11}}{n_{00}+n_{01}+n_{10}+n_{11}}$. The bottom line is that the uncertainty about the proper choice of cutoff point is a major problem with using the percentage correctly predicted as a measure of goodness-of-fit.

Additional notes on binary variable models

One of the key choices in the various binary variable models involves the cumulative distribution function. The Table 4 shows the four commonly used binary outcome models along with the cumulative distribution functions:

Model	**Probability density function**	**Cumulative distribution function**	**Marginal effects,** $\frac{\partial p}{\partial x_j}$
Logit	**Logistic**	$\Lambda(\mathbf{x}'\boldsymbol{\beta}) = \frac{e^{\mathbf{x}'\boldsymbol{\beta}}}{1+e^{\mathbf{x}'\boldsymbol{\beta}}}$	$\Lambda(\mathbf{x}'\boldsymbol{\beta})[1-\Lambda(\mathbf{x}'\boldsymbol{\beta})]\beta_j$
Probit	**Normal***	$\Phi(\mathbf{x}'\boldsymbol{\beta}) = \int_{-\infty}^{\mathbf{x}'\boldsymbol{\beta}} \phi(\mathbf{x}'\boldsymbol{\beta})dx$	$\phi(\mathbf{x}'\boldsymbol{\beta})\beta_j$
Linear probability		$F(\mathbf{x}'\boldsymbol{\beta}) = \mathbf{x}'\boldsymbol{\beta}$	$\boldsymbol{\beta_j}$

Complementary log-log		$C(\mathbf{x}'\boldsymbol{\beta}) = 1 - e^{-e^{\mathbf{x}'\boldsymbol{\beta}}}$	$e^{-e^{\mathbf{x}'\boldsymbol{\beta}}} e^{\mathbf{x}'\boldsymbol{\beta}} \beta_j$

Table 2.4. Commonly used binary outcome models.

*** *φ*(·) is the probability density function (pdf) of the normal distribution.**

The logit, probit, and complementary log-log models are symmetric around zero and restrict 0 ≤ *p* ≤ 1. The linear does not impose either of these restrictions. Use of the complementary log-log regression sometimes is recommended when the sample is skewed such that there is a high proportion of ones and zeros. In general, economists use either the logit or probit models a majority of the time. Interestingly, there is no need to use robust estimation techniques for the logit and probit models if they are correctly specified. If use of the vce(robust) option produces substantially different parameter estimates than the estimates without the robust option, then it is likely that the models are misspecified. The linear model is inherently heteroskedastistic, implying that the vce(robust) option should be used.

The parameter estimates are comparable across the first three models in Table 4. In particular,

1. $\widehat{\beta}_{\text{Logit}} \approx 4\,\widehat{\beta}_{\text{Linear}}$,
2. $\widehat{\beta}_{\text{Probit}} \approx 2.5\,\widehat{\beta}_{\text{Linear}}$, and
3. $\widehat{\beta}_{\text{Logit}} \approx 1.6\,\widehat{\beta}_{\text{Logit}}$.

Example 2.2. Supplementary health insurance coverage.

These data come from wave 5 (2002) of the Health and Retirement Study (HRS), a panel survey sponsored by the National Institute of Aging. The sample is restricted to Medicare beneficiaries; there are 3,206 observations. The elderly can obtain supplementary insurance coverage either by purchasing it themselves or by joining employer-sponsored plans. The data is in the file Example.xls. The variables included are listed in Table ?.

Variable	Definition
Binary variables	
(ins	= 1 if individual has purchased supplementary insurance from any source
retire	= 1 if individual is retired
hstatusg	= 1 if individual assess his/her health status either as good, very good, or excellent
married	= 1 if married
hisp	= 1 if hispanic
female	= 1 if female
white	= 1 if white
sretire	= 1 if a retired spouse is present in household
Continuous variables	
age	Age of individual in years
hhincome	Household income
educyear	Years of education
chronic	Total number of chronic conditions
adl	Number of limitations on daily activity (up to 5)

Table 2.5. Definition of the variables used in Example 2.

Stata commands

Place the data into the editor and then create a list of the independent variables. Now create a new variable equal to the log of income:

```
.generate linc = ln(hhinc)
```

[notice that 9 observations are eliminated.]

Create list of "extra" variables in order to shorten future commands:

```
. global extralist linc female white chronic adl sretire
```

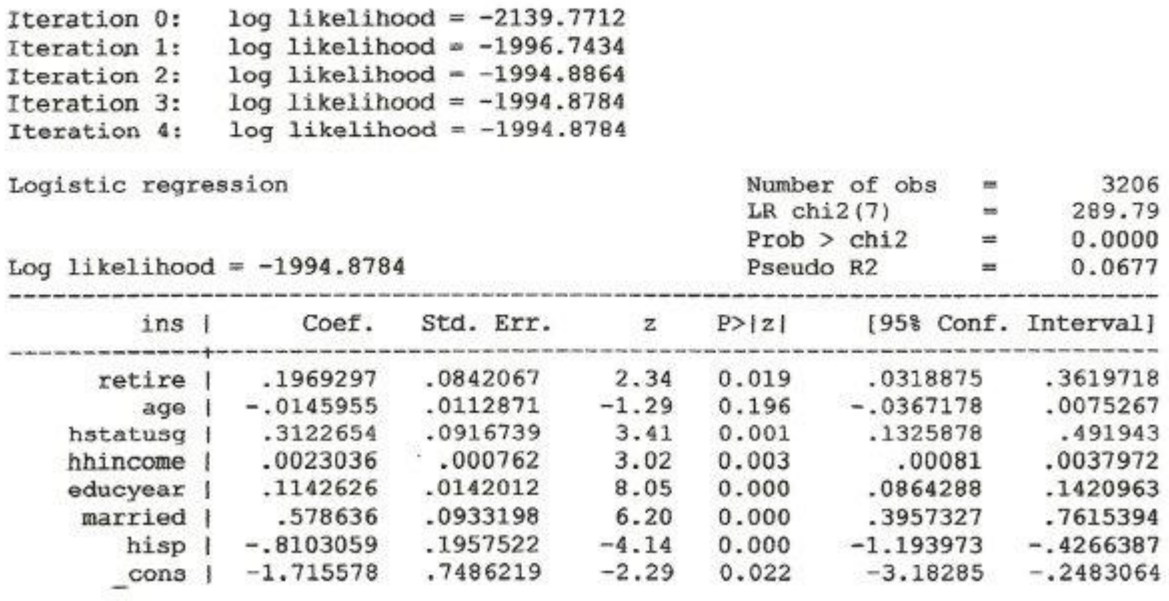
Summarize the variables in order to check for obvious typos (output is suppressed):

```
.summarize ins retire $xlist $extralist
```

Estimate logit regression (output is shown in Figure 3):

```
.logit ins retire $xlist
```

Figure 2.3. Stata regression output.



```
Iteration 0:   log likelihood = -2139.7712
Iteration 1:   log likelihood = -1996.7434
Iteration 2:   log likelihood = -1994.8864
Iteration 3:   log likelihood = -1994.8784
Iteration 4:   log likelihood = -1994.8784

Logistic regression                               Number of obs   =       3206
                                                  LR chi2(7)      =     289.79
                                                  Prob > chi2     =     0.0000
Log likelihood = -1994.8784                       Pseudo R2       =     0.0677
------------------------------------------------------------------------------
         ins |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      retire |   .1969297   .0842067     2.34   0.019     .0318875    .3619718
         age |  -.0145955   .0112871    -1.29   0.196    -.0367178    .0075267
    hstatusg |   .3122654   .0916739     3.41   0.001     .1325878     .491943
    hhincome |   .0023036    .000762     3.02   0.003       .00081    .0037972
    educyear |   .1142626   .0142012     8.05   0.000     .0864288    .1420963
     married |    .578636   .0933198     6.20   0.000     .3957327    .7615394
        hisp |  -.8103059   .1957522    -4.14   0.000    -1.193973   -.4266387
       _cons |  -1.715578   .7486219    -2.29   0.022     -3.18285   -.2483064
```

Estimate and save results from several models (the Stata command "quietly" suppresses the output from the command):

. estimates store blogit

.quietly probit ins retire $xlist

.estimates store bprobit

.quietly regress ins retire $xlist

.estimates store bols

.quietly logit ins retire $list, vce(robust)

. estimates store blogitr

.quietly probit ins retire $xlist, vce(robust)

```
.estimates store bprobitr
.quietly regress ins retire $xlist, vce(robust)
.estimates store bolsr
```

We can create table for comparing the models (output is suppressed):

```
.estimates table blogit blogitr bprobit bprobitr bols bolsr, t stats(N ll) b(%8.4f) stfmt(%8.2f)
```

We now test for the presence of interaction variables:

```
.generate age2 = age*age
.generate agefem = age*fem
.generate agewhite = age*white
.generate agechronic = age*chronic
.global intlist age2 agefem agewhite agechronic
.quietly logit ins retire $xlist $intlist
.test $intlist
( 1) [ins]age2 = 0
( 2) [ins]agefem = 0
( 3) [ins]agewhite = 0
( 4) [ins]agechronic = 0
chi2( 4) = 7.45
Prob > chi2 = 0.1141
```

Likelihood ratio test

```
.quietly logit ins retire $xlist $intlist
.estimates store B
```

```
.quietly logit ins retire $xlist
.lrtest B
Likelihood-ratio test LR chi2(4) = 7.57
(Assumption: . nested in B) Prob > chi2 = 0.1088
```

Comparison with using the logistic command:

```
. logistic ins retire $xlist
```

The marginal effects at the mean will yield more useful results when the model is non-linear:

```
.quietly logit ins retire $xlist
.mfx
```

Let's put the table comparing parameter estimates into a cleaned up table:

	Logit	Robust Logit	Probit	Robust Probit	OLS	Robust OLS
Individual retired	0.1969	0.1969	0.1184	0.1184	0.0409	0.0409
	(2.34)	(2.32)	(2.31)	(2.30)	(2.24)	(2.24)
Age of individual	-0.0146	-0.0146	-0.0089	-0.0089	-0.0029	-0.0029
	(-1.29)	(-1.29)	(-1.29)	(-1.32)	(-1.20)	(-1.25)
Health status	0.3123	0.3123	0.1977	0.1977	0.0656	0.0656
	(3.41)	(3.40)	(3.56)	(3.57)	(3.37)	(3.45)
Household income	0.0023	0.0023	0.0012	0.0012	0.0005	0.0005
	(3.02)	(2.01)	(3.19)	(2.21)	(3.58)	(2.63)
Years of education	0.1143	0.1143	0.0707	0.0707	0.0234	0.0234

	(8.05)	(7.96)	(8.34)	(8.33)	(8.15)	(8.63)
Individual married	0.5786	0.5786	0.3623	0.3623	0.1235	0.1235
	(6.20)	(6.15)	(6.47)	(6.16)	(6.38)	(6.62)
Individual is an Hispanic	-0.8103	-0.8103	-0.4731	-0.4731	-0.1210	-0.1210
	(-4.14)	(-4.18)	(-4.28)	(-4.36)	(-3.59)	(-4.49)
Intercept	-1.7156	-1.7156	-1.0693	-1.0693	0.1271	0.1271
	(-2.29)	(-2.36)	(-2.33)	(-2.40)	(0.79)	(0.83)
Sample size	3,206	3,206	3,206	3,206	3,206	3,206
Log of the likelihood function	-1994.88	-1994.88	-1993.62	-1993.62	-2104.75	-2104.75

Table 2.6. Comparison of Logit, Probit and OLS regressions with Insurance as the dependent variable. (t-ratio or z-values in parentheses.)

As a last exercise use the following commands to generate a graph of the predicted values:

```
. quietly logit ins hhincome
. predict plogit, pr
. quietly probit ins hhincome
. predict pprobit, pr
. quietly regress ins hhincome
. predict pols, xb
. summarize ins plogit pprobit pols
. sort hhincome
```

```
.twoway (scatter ins hhincome, msize(vsmall)) (line plogit hhincome, lcolor(blue) lpattern
> (solid)) (line pprobit hhincome, lcolor(red) lpattern(tight_dot)) (line pols hhincome,
> lcolor(green) lpattern(longdash_shortdash)), ytitle(Predicted Probability) xtitle(Household income)
Note: save file as a .tif file if you want to insert the graph directly into a word file.
```

Exercises

Exercise 2.1.1.

The determinants of physician advice. Physicians are expected to give lifestyle advice as a part of their normal interaction with their patients. Sometimes doctors choose not to comment on a patient’s lifestyle because they do not have time for personal comments, they feel the advice will be unwelcome, they feel that lifestyle choices are not any business of the physician, they find the discussion of lifestyle issues to be embarrassing, or they are not aware of the patient’s actual lifestyle choices. In this project we are interested in understanding when physicians choose to give advice concerning the consumption of alcohol.

The MS Excel file ktdata contains the responses to the 1990 National Health Interview Survey core questionnaire and special supplements from 2,467 males who were current drinkers in 1990. Individuals who are lifetime abstainers or who are former drinkers who have not consumed any alcohol in the past year are excluded from the sample. Table 7 contains the names and definitions of the variables collected in the survey.

Variable	Definition
Drinks	Total number of drinks taken in the past two weeks
Advice	Did your physician give you advice about alcohol consumption? Yes = 1, No = 0

Income	**Monthly income in $1,000 (there are 5 missing values denoted by a ".")**
Age30	**Dummy variable equal to 1 if 30 < Age ≤ 40and 0 otherwise**
Age40	**Dummy variable equal to 1 if 40 < Age ≤ 50 and 0 otherwise**
Age50	**Dummy variable equal to 1 if 50 < Age ≤ 60 and 0 otherwise**
Age60	**Dummy variable equal to 1 if 60 < Age ≤ 70 and 0 otherwise**
AgeGT70	**Dummy variable equal to 1 if individual's age is greater than 70 and 0 otherwise**
Educ	**Number of years of schooling (0 to 18)**
Black	**Dummy variable equal to 1 if the individual is a black and 0 otherwise**
Other	**Dummy variable equal to 1 if the individual is non-white and non-black and 0 otherwise**
Married	**Dummy variable equal to 1 if the individual is married and 0 otherwise**
Widow	**Dummy variable equal to 1 if the individual is a widow and 0 otherwise**
DivSep	**Dummy variable equal to 1 if the individual is either divorce or separated and 0 otherwise**
Employed	**Dummy variable equal to 1 if the individual is currently employed and 0 otherwise**
Unemploy	**Dummy variable equal to 1 if the individual is currently unemployed and 0 otherwise**
NE	**Dummy variable equal to 1 if the individual lives in the Northeast US and 0 otherwise**
MW	**Dummy variable equal to 1 if the individual lives in the Midwest US and 0 otherwise**
South	**Dummy variable equal to 1 if the individual lives in the South and 0 otherwise**
Medicare	**Dummy variable equal to 1 if the individual receives Medicare and 0 otherwise**
Medicaid	**Dummy variable equal to 1 if the individual receives Medicaid and 0 otherwise**
Champus	**Dummy variable equal to 1 if the individual has military insurance and 0 otherwise**

HlthIns	**Dummy variable equal to 1 if the individual has health insurance and 0 otherwise**
RegMed	**Dummy variable equal to 1 if the individual has a regular source of medical care and 0 otherwise**
DRI	**Dummy variable equal to 1 if the individual sees the same doctor and 0 otherwise**
MajorLim	**Dummy variable equal to 1 if the individual has limits on major daily activity and 0 otherwise**
SomeLim	**Dummy variable equal to 1 if the individual has limits on some daily activity and 0 otherwise**
Diabetes	**Dummy variable equal to 1 if the individual has diabetes and 0 otherwise**
Heart	**Dummy variable equal to 1 if the individual has a heart condition and 0 otherwise**
Stroke	**Dummy variable equal to 1 if the individual has had a stroke and 0 otherwise**

Table 2.7. Definition of the variables in the *Excel* worksheet ktdata.

You are to estimate a logit regression of the form: $\ln\left(\frac{p}{1-p}\right)=\beta_0+\sum_i \beta_i x_i+\varepsilon,$ **where *p* is the probability that a patient received advice about his level of consumption of alcohol and *x*$_i$ are the explanatory variables.**

Provide the following information:

1. **Make a table of the means of all of the variables.**

2. **Offer an economic justification for the inclusion of each explanatory variable you use in your regression (including a prediction of its expected sign).**

3. **Make a table reporting the results of the estimation of (1) an OLS linear estimation, (2) a probit estimation, and (3) a logit estimation. Also include a column with the ratio of each of the logit parameters to the probit parameter. Do not use the abbreviated name of the explanatory variables in the table.**

4. **Present a table of results of a logit model with all of the variables and with whatever other models you feel are suggested by your empirical results. Discuss the results of the estimation and what the estimation tells you about how physicians decide whether to give advice on alcohol consumption to their male patients.**

Exercise 2.1.2.

The Supply of Married Women in the Workforce. We are interested in understanding the decision of married women to enter the labor force. We have available two data sets, one using data from the United States and the other using data from Portugal. You are to estimate a logit regression for married women for each of the two data sets.

Variable	Definition
Working	dummy variable = 1 if a married woman works during the year
Fulltime	dummy variable = 1 if a married woman works more than 1000 hours in a year
Other	the other household income in $100 (not in $1000)
Age	age of the wife
Educ	education years of the wife
C0005	number of children for ages 0 to 5
C0613	number of children for ages 6 to 13
C1417	number of children for ages 14 to 17
NW	1 if non-white, and 0 otherwise.
HOwn	1 if the home is owned by the household, and 0 otherwise
HMort	1 if the home is on mortgage, and 0 otherwise
Prof	1 if the husband is manager or professional, and 0 otherwise

Sales	1 if the husband is sales worker or clerical or craftsman, and 0 otherwise
Farm	1 if the husband is farm-related worker
Unem	local unemployment rate in %

Table 2.8. US Data on Married Women.

Data Set 1: The data for this project are in the MS Excel file FLABOR. These data are observations on married females drawn from the 1987 wave of Michigan Panel Study of Income Dynamics (PSID). The data set has observations for 3,382 individuals.

Data Set 2: These data are from Portugal. The data set is a sample from Portuguese Employment Survey, from the interview year 1991, and has been provided by the Portuguese National Institute of Statistics (INE). The data are in the Excel file Martins. This file is organized into seven columns, corresponding to seven variables, with 2,339 observations.

Variable	Definition
Works	Dummy variable equal to 1 if the woman works, 0 otherwise
Child18	The number of children younger than 18 living in the family
Child03	The number of children younger than 3 living in the family
Age	The woman’s age
LogWomanWageRate	The log of women's hourly wage rate (measured in escudos)
Education	The women's educational level, measured in years of schooling
LogHusbandMonthlyWages	The log of the husband's monthly wage (measured in escudos)

Table 2.9. The Portuguese data set.

Answer the following questions:

1. **What factors other than wage levels determine the number of hours that a wife will spend in the work force? Remember to use economic theory in answering this question.**

2. **Clearly, one of the major factors in determining if a wife will enter the labor force is the wage level she can earn. The US data set does not include the wife's wage level. Is there any other variable in the data set that economic theory suggests will be a good proxy for wage levels?**

3. **The variable Age is a proxy for the work (or life) experience of a woman. We would expect that its effect on the probability that a woman will enter the labor force will be non-linear—that is, its marginal impact will be positive and decreasing. This reasoning suggests that you should use Age and Age^2 as explanatory variables. Can the same reasoning be used with the variable Education? What are your expectations about the signs of the parameters of these two explanatory variables? The same reasoning can be used about the number of years of education.**

4. **Estimate and report in a table the following two logit regressions: (1) US women enter the labor force at all and (2) US women enter the labor force for at least 1,000 hours if they enter the labor force,. In each of these cases, compare your results to a linear model.**

5. **The Portuguese data set has a different problem. We have reported the wage rate of women who are working, but no wage level for women who are not working. We will get around this problem by first using the data for women who actually work to estimate the relationship between wage rates and the age and education of the women. We will then use this relationship to predict the wage rate for both women who do work and women who do not work. We will then use this predicted wage rate data series as an independent variable in a logit model explaining the probability that a married woman will enter the labor force. When completing the logit regression be sure that you separate all of the children in a family into those 3 and under and those between 4 and 18. Also, include the years of education in this regression to see if a Portuguese married woman's taste for participation in the labor force increases or decreases with the level of her education.**

6. **Is it reasonable to compare your results for the two countries?**

References

Amemiya, T. (1981). Quantitative Response Models: A Survey. *Journal of Economic Literature* 19: 1483-1536.

Cramer, J. S. (2003). *Logit Models from Economics and Other Fields* (Cambridge: Cambridge University Press).

Cameron, A. Colin and Pravin K. Trivedi (2005). *Microeconometrics: Methods and Applications* (Cambridge: Cambridge University Press).

Ladd, G. W. (1966). Linear Probability Functions and Discriminant Functions. *Econometrica* 34: 873-888.

Maddala, G. S. (1983). *Limited-Dependent and Qualitative Variables in Economics* (Cambridge: Cambridge University Press).

McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior. In P. Zarembka (ed.) *Frontiers in Econometrics* (New York: Academic Press): 105-142.

Wald, A. (1943). Test of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large. *Transactions of the American Mathematical Society* 54: 426-482.

2.2. Analysis of time series[*]

Analysis of Time-Series

Introduction

This module offers a brief introduction of some of the issues that arise in the analysis of time-series. Most of the topics covered are those that we attacked first by statisticians and economists. As such they do not demand the more sophisticated tools used by the more modern approaches to time-series. In spite of these shortcomings, they should give you some understanding of the issues that arise with the use of times-series in econometric analyses. One final note of explanation is necessary. These notes are designed to give you a brief introduction to how *Stata* handles time-series data. These notes are not a substitute for reading the *Stata* manual, completing a forecasting course, or reading standard texts on the rather complicated field.

Time-series analysis in *Stata*

Throughout this module we work with US macroeconomic data included in the MS Excel file Macro data.xls. The variables are real level of investments (RINV), real gross national product (RGNP), and real interest rate (RINTRATE). The real interest rate is approximated by the difference between the nominal interest rate and the rate of change of the price index from the previous year. The data are for the years 1963 to 1982. You can replicate the analysis done here by copying this data set into a *Stata* file.

The first step after entering the data set into *Stata*, is to declare that the data set is a time-series. The command to do this is:

. tsset year

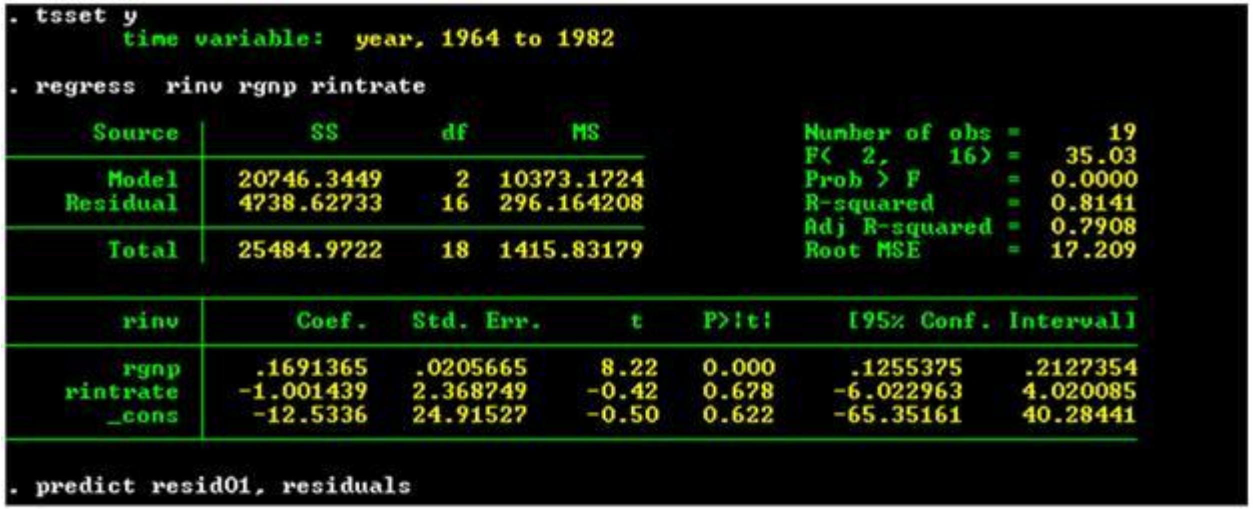
The data set can be broken into any number of time periods including daily, weekly, monthly, quarterly, halfyearly, yearly and generic.[14]

Assume that we want to estimate the following regression:

$$(2.11)\ RINV_t = \beta_0 + \beta_1 RGNP_t + \beta_2 RINTRATE_t + \varepsilon_t$$

using the data set in the appendix. Figure 1 shows this regression command and the resultant output.

Figure 2.4.

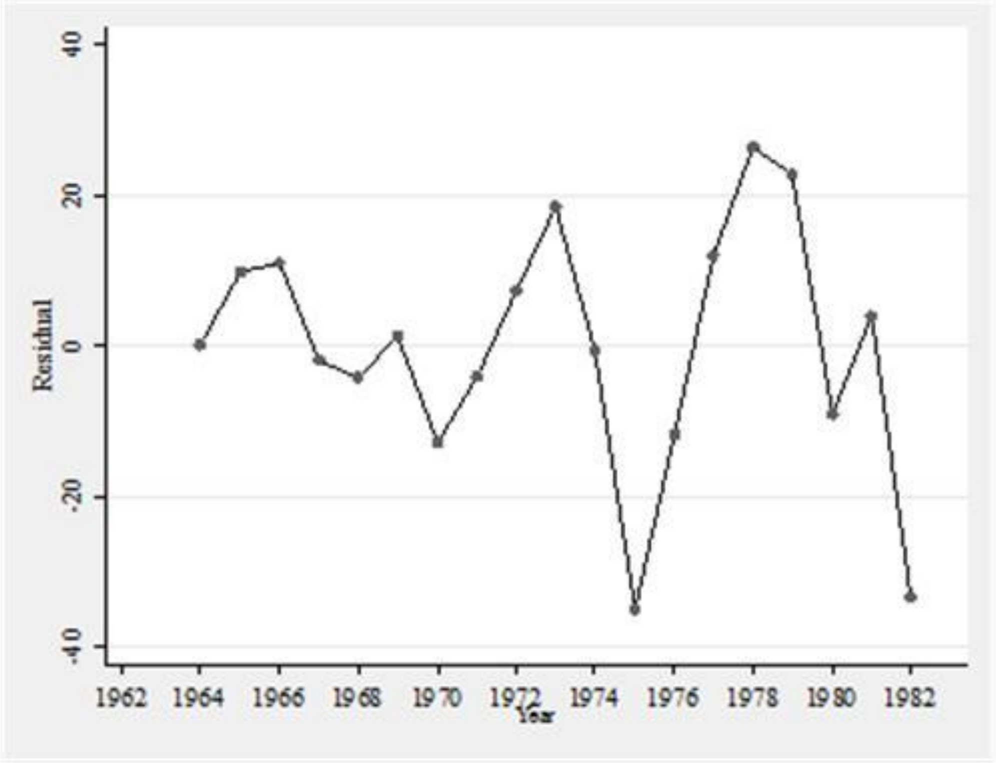


```
. tsset y
        time variable:  year, 1964 to 1982

. regress  rinv rgnp rintrate

      Source |       SS       df       MS              Number of obs =      19
-------------+------------------------------           F(  2,    16) =   35.03
       Model |  20746.3449     2  10373.1724           Prob > F      =  0.0000
    Residual |  4738.62733    16  296.164208           R-squared     =  0.8141
-------------+------------------------------           Adj R-squared =  0.7908
       Total |  25484.9722    18  1415.83179           Root MSE      =  17.209

------------------------------------------------------------------------------
        rinv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        rgnp |   .1691365   .0205665     8.22   0.000     .1255375    .2127354
    rintrate |  -1.001439   2.368749    -0.42   0.678    -6.022963    4.020085
       _cons |   -12.5336   24.91527    -0.50   0.622    -65.35161    40.28441
------------------------------------------------------------------------------

. predict resid01, residuals
```

OLS estimates for Equation (1).

On the surface the estimates seem “reasonable” because the signs on the two explanatory variables are what theory predicts they should be and the parameter for real GNP is statistically different from zero. However, an examination of the residuals shown in Figure 2 suggest that the error terms might exhibit autocorrelation.

Figure 2.5.



The residuals appear to be autocorrelated.

There are several issues that arise here. First, what sort of models can we use to account for autocorrelation? Second, what sorts of tests exist for detecting the existence of autocorrelation? We begin with the first of these questions by introducing the concept of first-order autocorrelation. Consider the following model:

$$(2.12)\ y_t = \beta_0 + \beta_1 x_t + \varepsilon_t .$$

We say that this model exhibits first-order autocorrelation if the error terms can be written as:

$$(2.13)\ \varepsilon_t = \rho\, \varepsilon_{t-1} + \mu_t ,$$

where $\mu_t \sim N\left(0, \sigma^2\right).$ Equation (3) implies that the error terms in (2) are correlated with each other. It is rather easy to show that, while the estimates of the unknown parameters are unbiased, the estimates of the standard errors are biased—downward if $1 > \rho > 0$ and upward if $-1 < \rho < 0$. This conclusion holds as long as the source of the autocorrelation is due to (3). If, on the other hand, the source of autocorrelation among the error terms in (2) is due to omitted explanatory variables (whose effects are absorbed in the error term), we have a potentially more serious problem. In particular, if the omitted explanatory variables are correlated with the included explanatory variables (as is often true in time-series), then the estimates of the unknown slope parameters are also biased.

For the moment we will assume that Equations (2) and (3) are true representations of the world. What then can we do to estimate (2)? What we need to do is find a way to transform (2) so that the error term of whatever regression we estimate does not exhibit autocorrelation. In time period $t - 1$ we have:

$$(2.14)\ y_{t-1} = \beta_0 + \beta_1 x_{t-1} + \varepsilon_{t-1} .$$

Multiply (4) by ρ to get:

$$(2.15)\ \rho\, y_{t-1} = \rho\, \beta_0 + \rho\, \beta_1 x_{t-1} + \rho\, \varepsilon_{t-1} .$$

Now subtracting (5) from (4) gives:

$$y_t - \rho y_{t-1} = \beta_0 + \beta_1 x_t + \varepsilon_t - \left(\rho\beta_0 + \rho\beta_1 x_{t-1} + \rho\varepsilon_{t-1}\right)$$

or, equivalently,

$$(y_t - \rho y_{t-1}) = \beta_0(1-\rho) + \beta_1(x_t - \rho x_{t-1}) + (\varepsilon_t - \rho\varepsilon_{t-1}).$$

Let

$y_t^* = y_{t-1} - \rho\, y_{t-1}$,

$\beta_0^* = \beta_0(1-\rho)$,

and

$x_t^* = x_{t-1} - \rho\, x_{t-1}$.

Remember that (3) implies that $\mu_t = \varepsilon_t - \rho\,\varepsilon_{t-1}$. Thus, we have:

(2.16) $y_t^* = \beta_0^* + \beta_1 x_t^* + \mu_t$,

where $\mu_t \sim N(0, \sigma^2)$. Thus, we have a regression for which the OLS estimates will be BLUE (Best Linear Unbiased Estimator) if we only knew the true value of ρ.

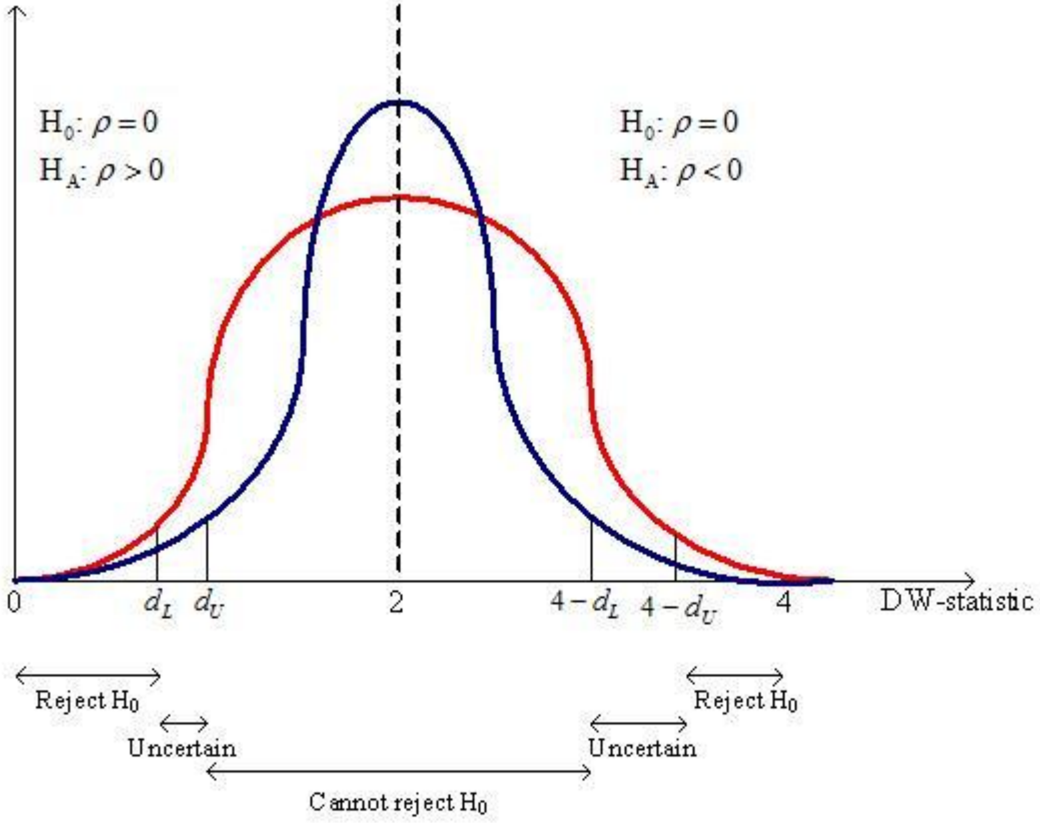
Cochran and Orcutt [1949] use this algebra to suggest one way to estimate (6). The estimation entails several steps. First, you use OLS to estimate (2). Second, you estimate (3) using the residuals from the first stage to approximate ε_t. This regression gives an estimate of ρ. In the third step, you use the estimate of ρ to construct estimates of y_t^* and x_t^*. In the fourth step, you use the estimates of y_t^* and x_t^* to estimate (6); this will yield new estimates of β_0 and β_1. You then repeat step (2) using these new estimates of β_0 and β_1 to calculate the residuals and then repeat with steps (3) and (4). You continue the process until the estimate of ρ does not change anymore (i.e., until the change in the estimate of ρ is less than some value chosen by the researcher). There are a multitude of alternative ways of estimating ρ. [See Greene (1990): Chapter 15 for a full discussion of these methods.] Once you have an estimator for ρ, there exist two major ways of completing the estimation—the Cochran-Orcutt procedure described above and the Prais-Winsten (1954) estimator. The latter estimation procedure does not involve dropping the first observation (as does the Cochran-Orcutt) estimator. In large samples these two estimation techniques are likely to be very similar. In small samples the two techniques may produce estimates that are substantially different.

We now turn to the issue of detecting the existence of autocorrelation. In what follows we focus mainly on the detection of first-order autocorrelation as shown in Equation (3). We can use the Durbin-Watson test to see if our suspicions are correct. The Durbin-Watson statistic tests the hypothesis:

(2.17)

$$\begin{aligned} &H_0 : \rho = 0 \\ &H_1 : \rho \neq 0 \end{aligned}$$

Figure 2.6.

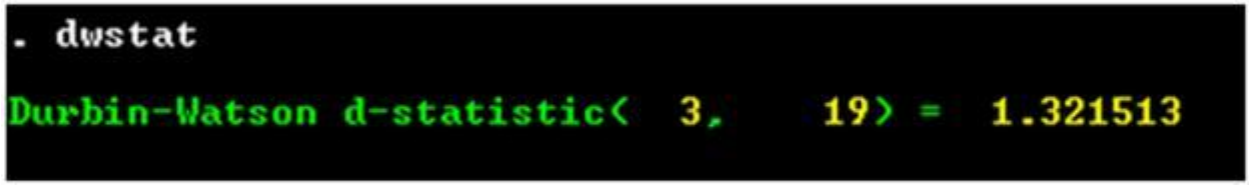


Limiting distributions for the Durbin-Watson statistic.

The details of the test statistic can be found in any econometrics textbook and need not detain us here. What you need to know about the DW-statistic are (1) it has a mean value of 2; (2) because its distribution lies between two limiting distributions, we need to look at two critical values. For this reason there are two critical values—one for each of the limiting distributions. Figure 3 illustrates the probability distribution function (pdf) for the Durbin-Watson statistic. The true pdf lies somewhere between the blue pdf and the red pdf. What is shown in the figure is the point below which, say, 5 percent of the distribution lies for each distribution. The true critical point lies somewhere between d_L and d_UThese values are relevant to testing the null hypothesis of no autocorrelation against the alternative hypothesis of positive autocorrelation (i. e., $\rho > 0$).

If $d < d_L$, we can reject the null hypothesis of no autocorrelation; if $d_U < d < 4 - d_U$, we cannot reject the null hypothesis; and if $d_L < d < d_U$, the results of the test are uncertain. Moreover, since the distributions are symmetric around 2 and between 0 and 4, the critical values for the alternative hypothesis of negative autocorrelation (i. e., $\rho > 0$) are 4 minus either the upper or lower critical values, as shown in Figure 3. Critical values for the Durbin-Watson statistic can be found in the appendices of most econometric textbooks.

Figure 2.7.



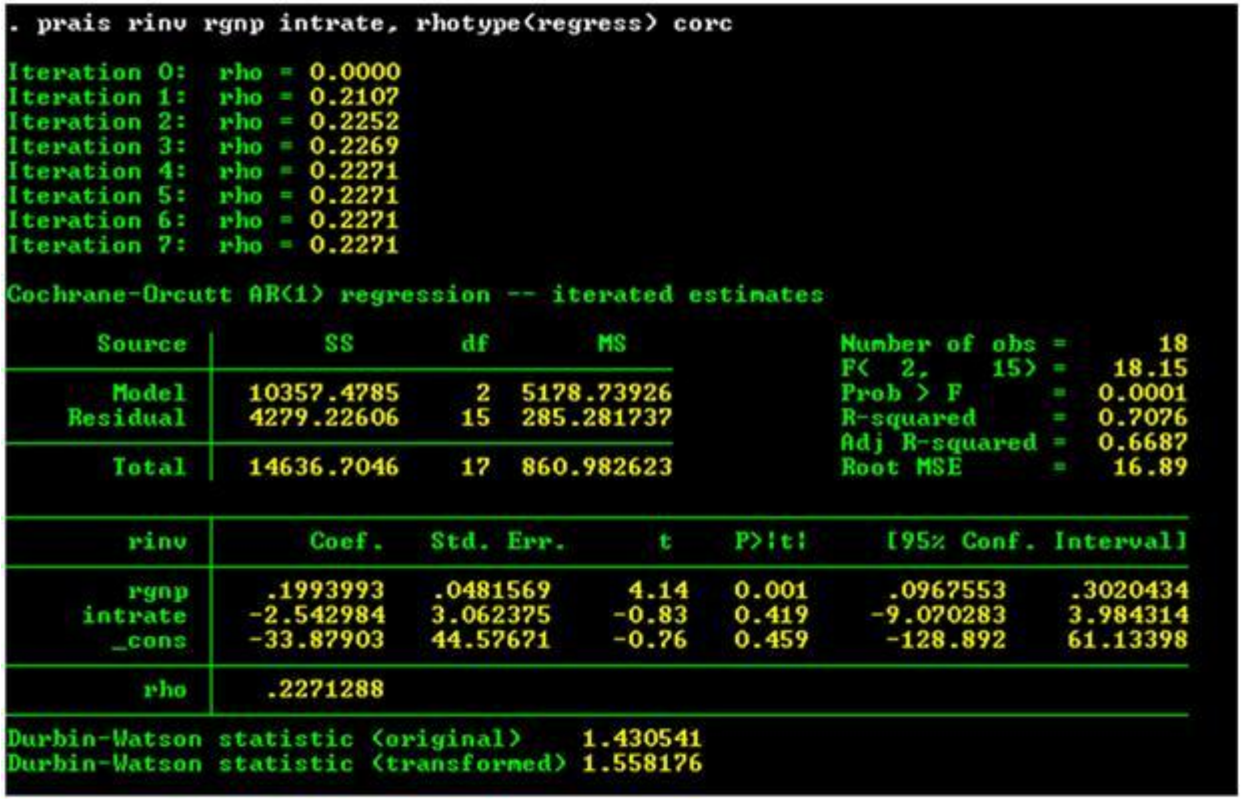
```
. dwstat

Durbin-Watson d-statistic(  3,    19) =  1.321513
```

Command for calculating the Durbin-Watson statistic in Stata.

The command for the test and the resultant DW-statistics for the estimate of Equation (2) are shown in Figure 4. The 5 percent level critical values for the Durbin-Watson statistic for a sample size of 19 with two parameters (less the intercept) estimated are 1.074 and 1.536—if the observed value of the DW-statistic is between 1.536 and 2.464, we can accept the null hypothesis that the residuals do not exhibit autocorrelation. Our value of 1.32 falls in the uncertain region where we are not sure if we can or cannot reject the null hypothesis.

At this point we can try the Cochran-Orcutt estimate. Figure 5 reports the results of using the Cochran-Orcutt estimation procedure. Notice that it took 7 iterations for the estimate of *ρ* to converge. If we use the Prais-Winsten estimation technique, we get the results shown in Figure 6. It is reassuring to see that the two estimation techniques do not yield estimates of the standard errors that are substantially different from each other.

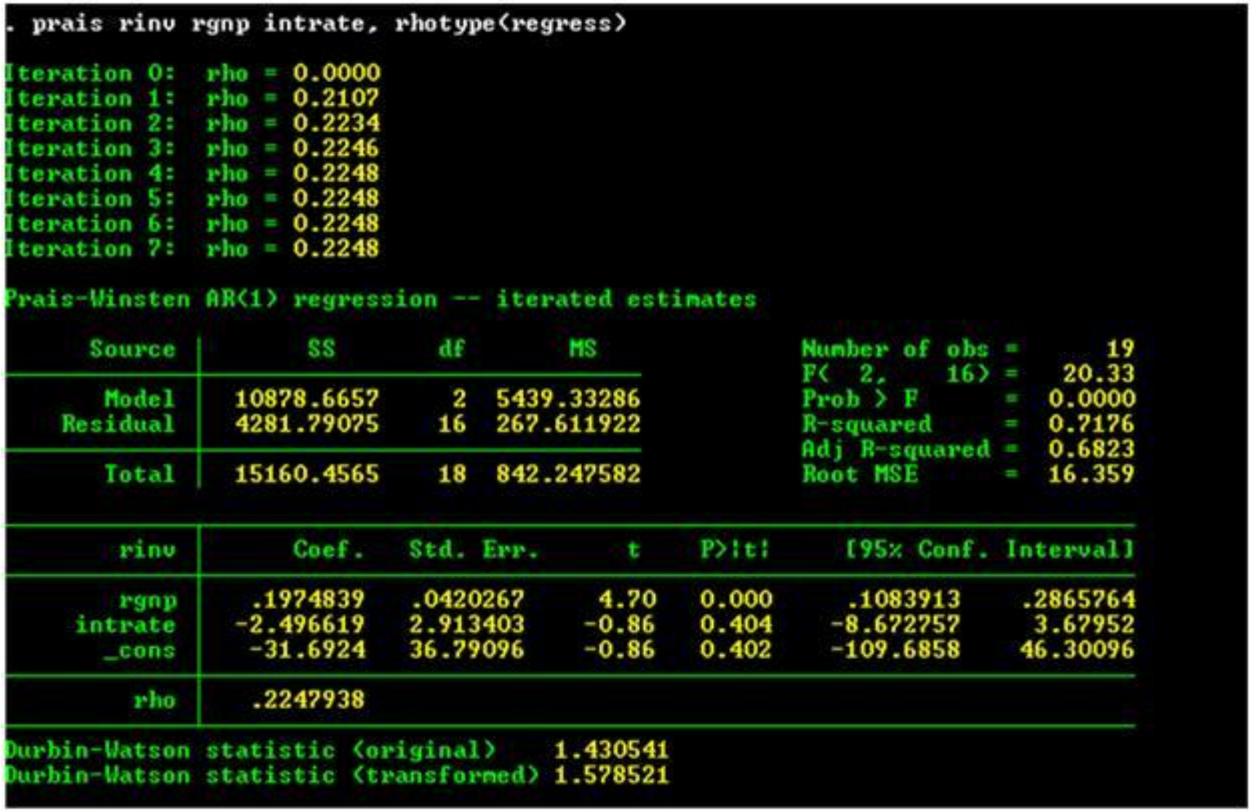
Figure 2.8.



```
. prais rinv rgnp intrate, rhotype(regress) corc

Iteration 0:  rho = 0.0000
Iteration 1:  rho = 0.2107
Iteration 2:  rho = 0.2252
Iteration 3:  rho = 0.2269
Iteration 4:  rho = 0.2271
Iteration 5:  rho = 0.2271
Iteration 6:  rho = 0.2271
Iteration 7:  rho = 0.2271

Cochrane-Orcutt AR(1) regression -- iterated estimates

      Source |       SS       df       MS              Number of obs =      18
-------------+------------------------------           F(  2,    15) =   18.15
       Model |  10357.4785     2  5178.73926           Prob > F      =  0.0001
    Residual |  4279.22606    15  285.281737           R-squared     =  0.7076
-------------+------------------------------           Adj R-squared =  0.6687
       Total |  14636.7046    17  860.982623           Root MSE      =   16.89

------------------------------------------------------------------------------
        rinv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        rgnp |   .1993993   .0481569     4.14   0.001     .0967553    .3020434
     intrate |  -2.542984   3.062375    -0.83   0.419    -9.070283    3.984314
       _cons |  -33.87903   44.57671    -0.76   0.459     -128.892    61.13398
-------------+----------------------------------------------------------------
         rho |   .2271288
------------------------------------------------------------------------------
Durbin-Watson statistic (original)    1.430541
Durbin-Watson statistic (transformed) 1.558176
```

Estimation of Equation (1) using the Cochran-Orcutt method.

Figure 2.9.



```
. prais rinv rgnp intrate, rhotype(regress)

Iteration 0:  rho = 0.0000
Iteration 1:  rho = 0.2107
Iteration 2:  rho = 0.2234
Iteration 3:  rho = 0.2246
Iteration 4:  rho = 0.2248
Iteration 5:  rho = 0.2248
Iteration 6:  rho = 0.2248
Iteration 7:  rho = 0.2248

Prais-Winsten AR(1) regression -- iterated estimates

      Source |       SS       df       MS              Number of obs =      19
-------------+------------------------------           F(  2,    16) =   20.33
       Model |  10878.6657     2  5439.33286           Prob > F      =  0.0000
    Residual |  4281.79075    16  267.611922           R-squared     =  0.7176
-------------+------------------------------           Adj R-squared =  0.6823
       Total |  15160.4565    18  842.247582           Root MSE      =  16.359

------------------------------------------------------------------------------
        rinv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        rgnp |   .1974839   .0420267     4.70   0.000     .1083913    .2865764
     intrate |  -2.496619   2.913403    -0.86   0.404    -8.672757     3.67952
       _cons |   -31.6924   36.79096    -0.86   0.402    -109.6858    46.30096
-------------+----------------------------------------------------------------
         rho |   .2247938
------------------------------------------------------------------------------
Durbin-Watson statistic (original)    1.430541
Durbin-Watson statistic (transformed) 1.578521
```

Estimation of Equation (1) using the Prais-Winsten estimator.

Using either the Cochran-Orcutt or the Prais-Winstn estimator is dependent on the assumption that the error terms exhibit first-order autocorrelation. Unfortunately, there is no particular reason (from a theoretical viewpoint) to believe in this assumption. Why, for instance, couldn't the error terms of Equation (2) exhibit second-order autocorrelation of the form:

$$\text{(2.18)}\ \varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + \mu_t ?$$

There is a more troubling possible explanation for the low Durbin-Watson statistic: the model may be misspecified. In particular, there may be important explanatory variables omitted from the regression. These omitted explanatory variables may exhibit autocorrelation and, thus, may be the source of autocorrelation in the error term. If the omitted explanatory variables are

correlated with the included explanatory variables, then the parameter estimates are biased. The large difference in the estimate of parameter for real interest rates for the OLS regression and the Cochran-Orcutt estimate is suggestive of model misspecification.[15]

More modern time-series models

ARMA models

The model we described above is assumed to have first-order autoregressive error disturbances. Such a process is referred to as AR(1). The error structure in (8) is AR(2). If we apply this concept to a data series, we would call the following an AR(*p*) process:

(2.19)

$$y_t = \alpha_0 + \sum_{i=1}^{p} \beta_i y_{t-i}.$$

Another approach available to us is to think of a data as a weighted average of some error terms that are assumed to have a mean of zero, have a fixed variance, and be uncorrelated over time[16]:

(2.20)

$$y_t = \sum_{i=0}^{q} \beta_i \varepsilon_{t-i}.$$

A data series exhibiting this pattern is called a moving average process or MA(q). The error tern is known in the literature as white noise. A data series that has both autoregressive and moving average characteristics is call an autoregressive moving average (ARMA) series; an ARMA(p, q) is:

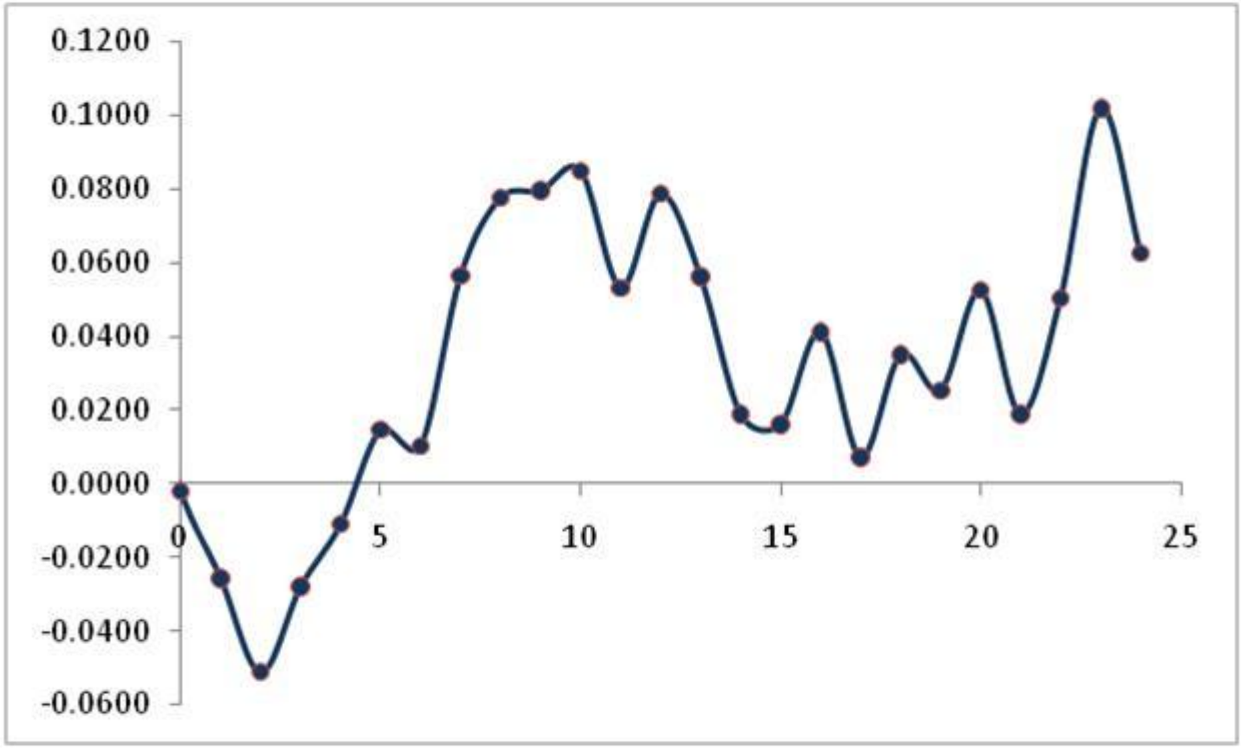
(2.21)

$$y_t = \alpha_0 + \sum_{i=1}^{p} \beta_i y_{t-i} + \sum_{i=0}^{q} \beta_i \varepsilon_{t-i}.$$

It may help to show two series constructed to have different ARMA patterns. Figure 7 shows one of the potential time series generated by the ARMA(2,1) process:

$$(2.22)\ y_t = 0.67y_{t-1} + 0.33y_{t-2} + 0.1\varepsilon_t + 0.05\varepsilon_{t-1}.$$

Figure 2.10.

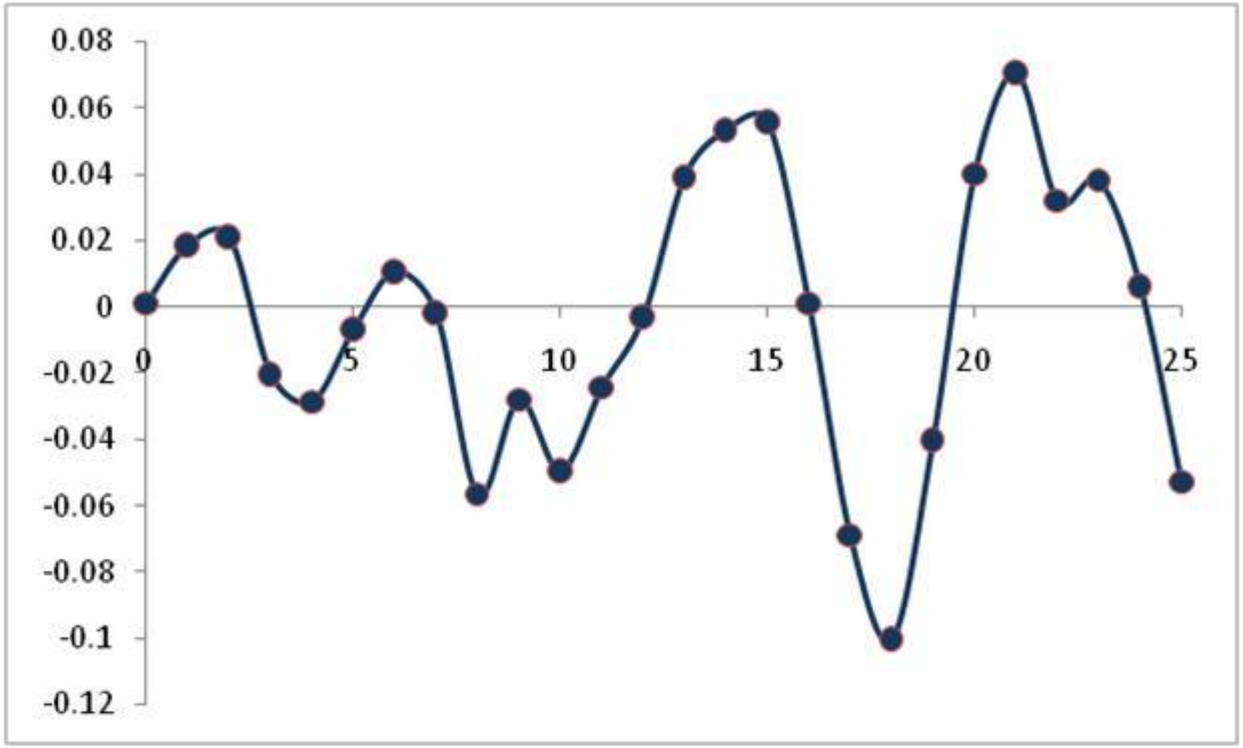


Graph of a ARMA(2,1) process.

Figure 8 shows one potential time series generated by the ARMA(1,1) process:

$$(2.23)\ y_t = 0.67y_{t-1} + 0.1\varepsilon_t + 0.05\varepsilon_{t-1}.$$

Figure 2.11.



Graph of a ARMA(1,1) process.

Stationarity

Consider the time-series y_t. We define this stochastic process as *covariance stationary* if

(2.24)

$$E(y_t) = E(y_{t-s}) = \mu,$$

(2.25)

$$E\left[(y_t-\mu)^2\right]=E\left[(y_{t-s}-\mu)^2\right]=\sigma^2,\text{ and}$$

(2.26)

$$Cov(y_t,y_{t-s})=E[(y_t-\mu)(y_{t-s}-\mu)]=E\left[\left(y_{t-j}-\mu\right)\left(y_{t-j-s}-\mu\right)\right]=\gamma_s.$$

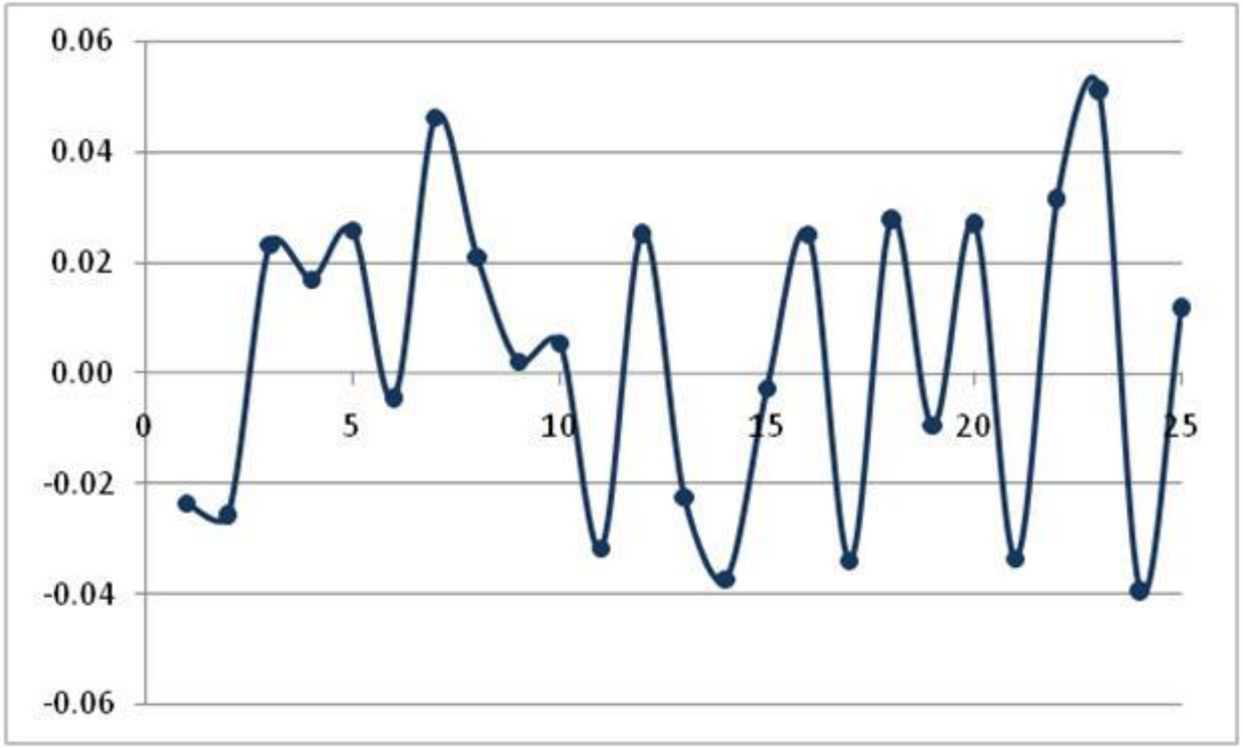
The last term, γ_s, is known as the autocovariance. A time-series is defined to be covariance stationary if its mean and all its autocovariances are unaffected by a change of time origin. We define the autocorrelation between y_t and y_{t-s} as:

(2.27)

$$\rho_s:=\frac{\gamma_s}{\gamma_0}.$$

Quite often you can create a stationary time-series from a non-stationary time-series by taking the first-differences of the non-stationary series. If the first difference does not produce a stationary series, then one continues to take first differences until you find a stationary series. For instance, the time-series shown in Figure 7 appears to be non-stationary. The first differences of this series is shown in Figure 9. Using the imperfect eye, it would appear that the first differences of (13) is stationary. However, we really cannot tell anything for sure from the graph of a data set. We need to use the restrictions of the parameters derived in advanced texts to determine if a data set is stationary.[17]

Figure 2.12.



First-differences of the time-series of the ARMA(2,1) data.

The autocorrelation function

One of the major ways to identify the structure of a time series is to look at the autocorrelation function. The autocorrelation function, ρ_s, is the correlation between y_t and y_{t-s}. Stata uses the following formula to estimate it [StataCorp: p. 60] for a time-series:

The researcher then has to compare the actual autocorrelation function with the theoretical autocorrelation for comparable data series. To see to use the autocorrelation function consider the following five time series[18]:

(2.28)

$$\widehat{\rho}_s = \frac{\sum_{i=1}^{n-s} (y_t - \bar{y})(y_{t-s} - \bar{y})}{\sum_{i=1}^{n} (y_t - \bar{y})^2}.$$

(2.29) AR(1): $y_t = 0.7y_{t-1} + \varepsilon_t$,

(2.30) AR(1): $y_t = -0.7y_{t-1} + \varepsilon_t$,

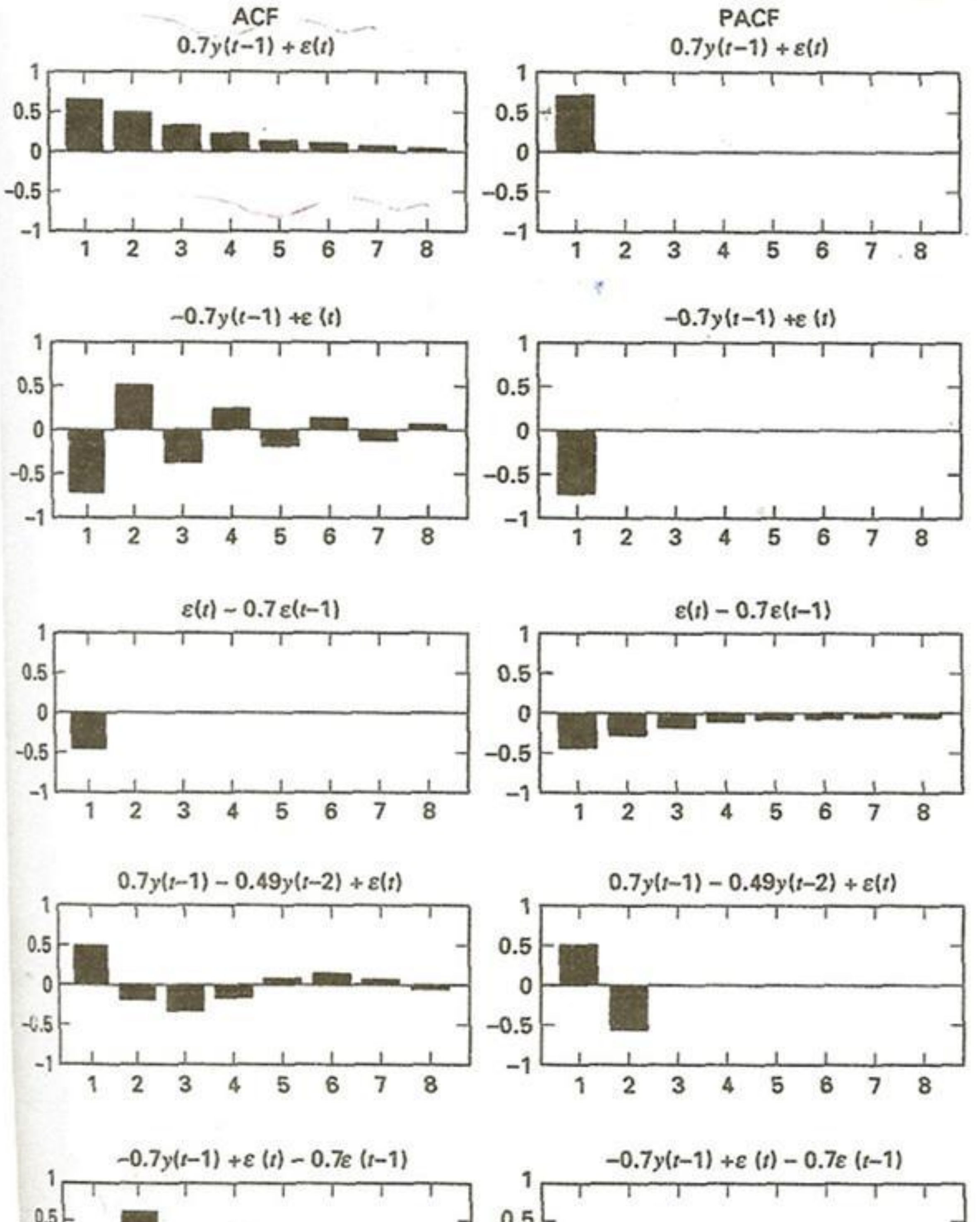
(2.31) MA(1): $y_t = \varepsilon_t - 0.7\varepsilon_{t-1}$,

(2.32) ARMA(2, 1): $y_t = 0.7y_{t-1} - 0.49y_{t-2} + \varepsilon_t$, and

(2.33) ARMA(1, 2): $y_t = -0.7y_{t-1} + \varepsilon_t - 0.7\varepsilon_{t-1}$.

Each of these functions has a theoretical autocorrelation function; graphs of these autocorrelation functions are shown in the left column of Figure 10.[19]

Figure 2.13.


ACF
PACF
$0.7y(t-1) + \varepsilon(t)$
$0.7y(t-1) + \varepsilon(t)$
$-0.7y(t-1) + \varepsilon(t)$
$-0.7y(t-1) + \varepsilon(t)$
$\varepsilon(t) - 0.7\varepsilon(t-1)$
$\varepsilon(t) - 0.7\varepsilon(t-1)$
$0.7y(t-1) - 0.49y(t-2) + \varepsilon(t)$
$0.7y(t-1) - 0.49y(t-2) + \varepsilon(t)$
$-0.7y(t-1) + \varepsilon(t) - 0.7\varepsilon(t-1)$
$-0.7y(t-1) + \varepsilon(t) - 0.7\varepsilon(t-1)$
1
0.5
0
-0.5
-1
1 2 3 4 5 6 7 8

Examples of autocorrelation and partial autocorrelation functions.

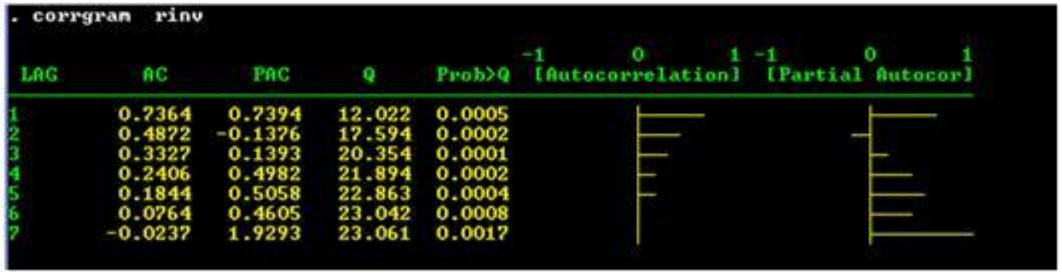
There is additional function we can use to help identify the nature of a time-series. Consider the following regressions:

$$\text{(2.34)}\quad y_t^* = \varphi_{11} y^*_{t-1} + e_t,\ y_t^* = \varphi_{21} y^*_{t-1} + \varphi_{22} y^*_{t-2} + e_t,\ \text{etc.,}$$

where $y_t^* = y_t - \bar{y}.$

Our interpretation of the φ_{ii} parameters is that they are the correlation between y_t and y_{t-i} controlling for all of the y_j where j = 2,…,(i − 1). Because these correlation coefficients control for values of y's observed between y_t and y_{t-i}, they are known as the partial autocorrelations. The theoretical partial autocorrelations are shown in the right column of Figure 10. Stata uses the command .corrgram varname to calculate the autocorrelations and partial autocorrelations for the time-series varname. Figure 11 shows the output when using this command on the real levels of investment. The autocorrelation function for this data set looks like the theoretical one for an AR(1) process. However, the partial autocorrelation function does not look like any of the partial autocorrelation functions shown in Figure 11. Thus, it would not be safe to assume that real investment follows an AR(1) process.

Figure 2.14.

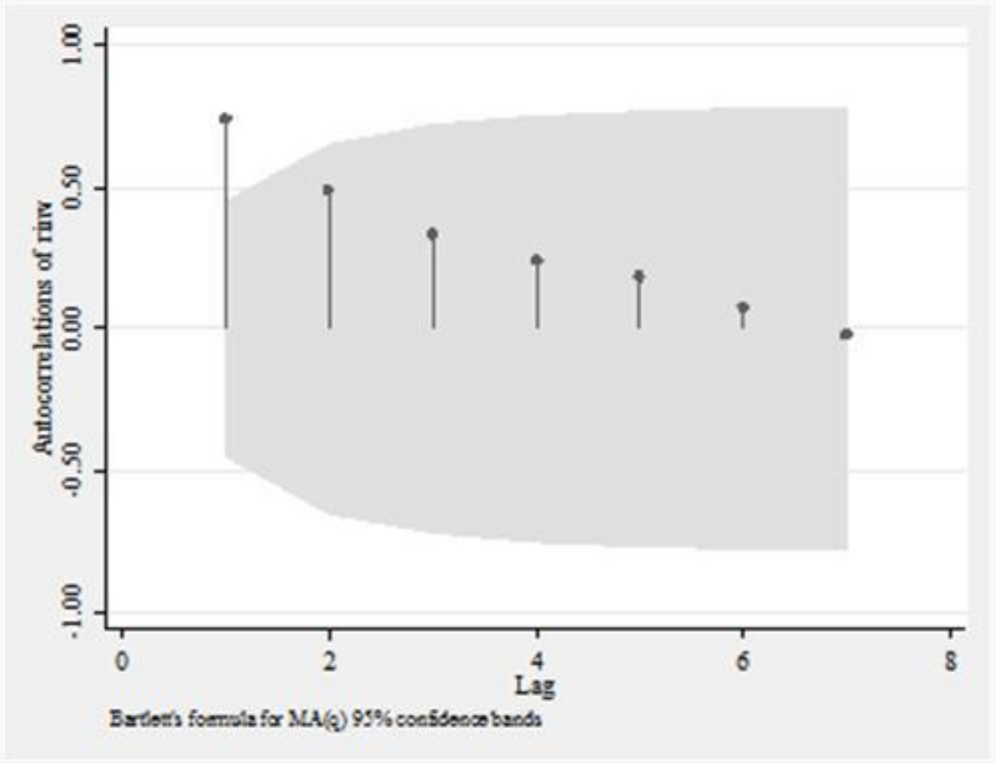


```
. corrgram  rinv

                                          -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------
1        0.7364   0.7394   12.022  0.0005          |-----             |-----
2        0.4872  -0.1376   17.594  0.0002          |---              -|
3        0.3327   0.1393   20.354  0.0001          |--                |-
4        0.2406   0.4982   21.894  0.0002          |-                 |---
5        0.1844   0.5058   22.863  0.0004          |-                 |----
6        0.0764   0.4605   23.042  0.0008          |                  |---
7       -0.0237   1.9293   23.061  0.0017          |                  |--------
```

Autocorrelation and partial autocorrelation functions for real investment.

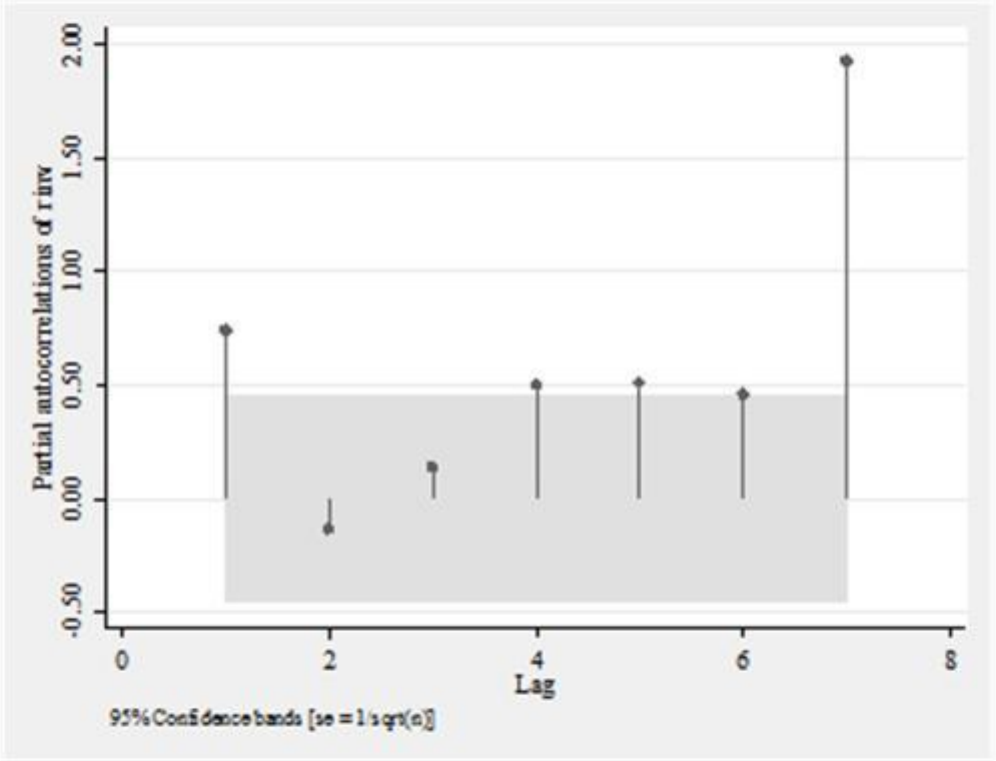
You can generate prettier graphs of the autocorrelation functions using the .ac varname command. For instance, the command .ac rinv generates the graph shown in Figure 12. The .pac varname generates a graph for the partial autocorrelations as is shown in Figure 13.

Figure 2.15.



Another graph of the autocorrelation function for real investment.

Figure 2.16.



Partial autocorrelations for real investments.

There are several generalizations one can use to help identify the process underlying a data series. Table 1 [Enders (2005): p. 85] offers a brief summary of these properties of the autocorrelation and partial autocorrelation functions.

Process	**Autocorrelation function**	**Partial autocorrelation function**
White-noise	**All $\rho_s = 0$**	**All $\varphi_{ss} = 0$**
AR(1): $\alpha_1 > 0$	**Direct exponential decay**	**$\varphi_{11} = \rho_1$; $\varphi_{ss} = 0$ for $s \geq 2$**

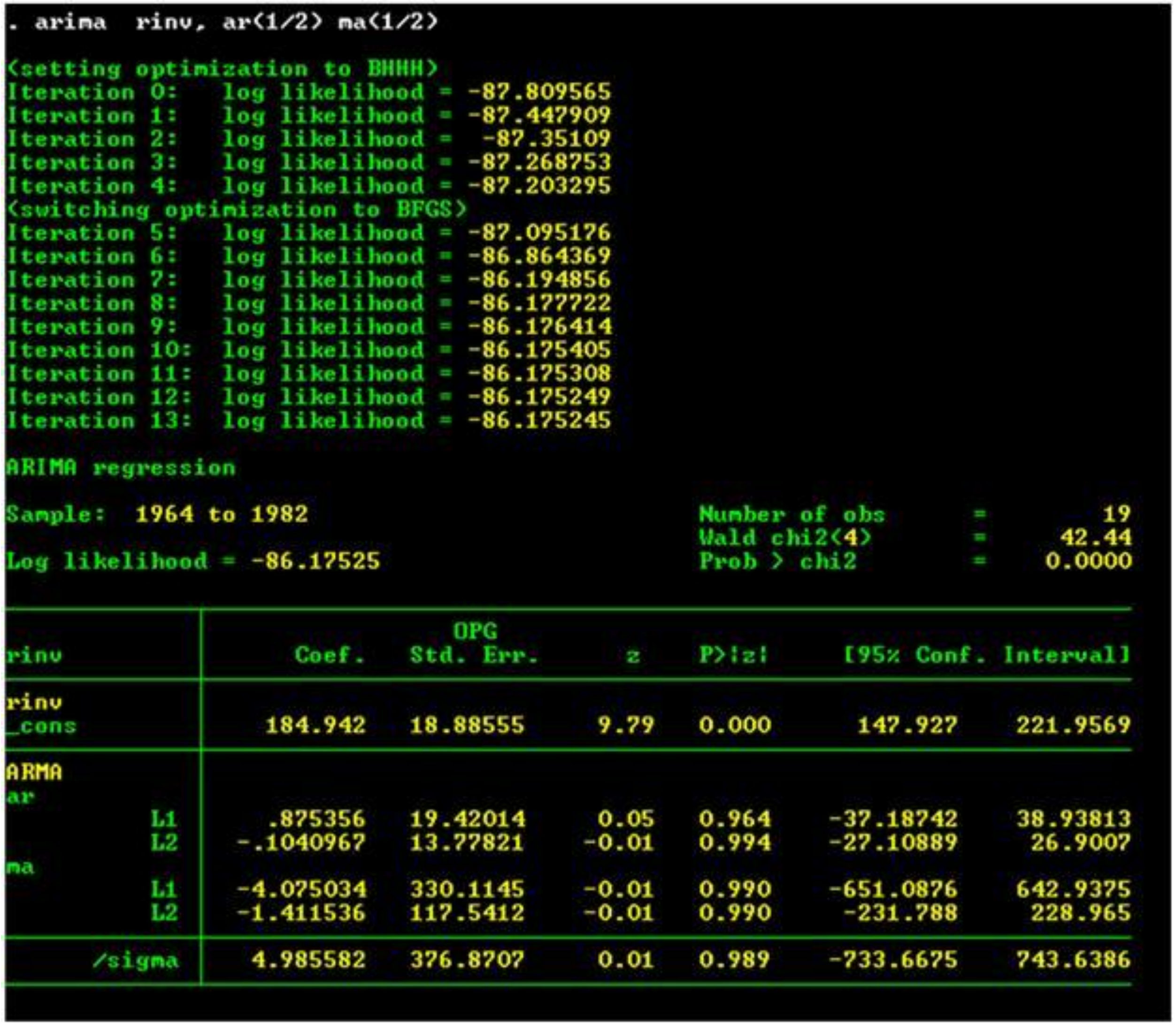
AR(1): $\alpha_1 > 0$	Decays toward zero. Coefficients may oscillate	$\varphi_{11} = \rho_1$; $\varphi_{ss} = 0$ for $s \geq 2$
AR(p)	Decays toward zero; Coefficients may oscillate	Spikes through lag p. All $\varphi_{ss} = 0$ for $s > p$
MA(1): $\beta > 0$	Negative spike at lag 1. $\rho_s = 0$ for $s \geq 2$	Oscillating decay: $\varphi_{11} < 0$
MA(1): $\beta < 0$	Positive spike at lag 1. $\rho_s = 0$ for $s \geq 2$	Decay: $\varphi_{11} > 0$
ARMA(1, 1): $\alpha_1 > 0$	Exponential decay beginning at lag 1. Sign ρ_1 = sign $(\alpha_1 + \beta)$	Oscillating decay beginning at lag 1. $\varphi_{11} = \rho_1$
ARMA(1, 1): $\alpha_1 < 0$	Oscillating decay beginning at lag 1. Sign ρ_1 = sign $(\alpha_1 + \beta)$	Exponential decay beginning at lag 1. $\varphi_{11} = \rho_1$ and sign (ϕ_{ss}) = sign (ϕ_{11})
ARMA(p, q)	Decay (either direct or oscillatory) beginning at lag q	Decay (either direct or oscillatory) beginning at lag p

Table 2.10. Properties of the autocorrelation and partial functions.

Estimation of ARMA models

The estimation of ARMA models are relatively easy in *Stata*. The basic command to estimate an ARMA model is: .arima depvar [varlist], ar(*numlist*) ma(*numlist*).[20] The first thing to notice in the command that this command can apply to either to a single variable or to an equation. If [varlist] is omitted, *Stata* will produce an estimate of the ARMA model for that variable; if the list is included, it will estimate the model with the disturbances allowed to have the ARMA structure specified in the command. Figure 14 reports the estimation of an ARMA model for real investment levels. Notice that we write AR(1/2) so that *Stata* knows to include both the first and second autoregressive term. A command of AR(2) would include only the second autoregressive term. In Figure 15 we report the ARMA (2, 1) estimation of (1).

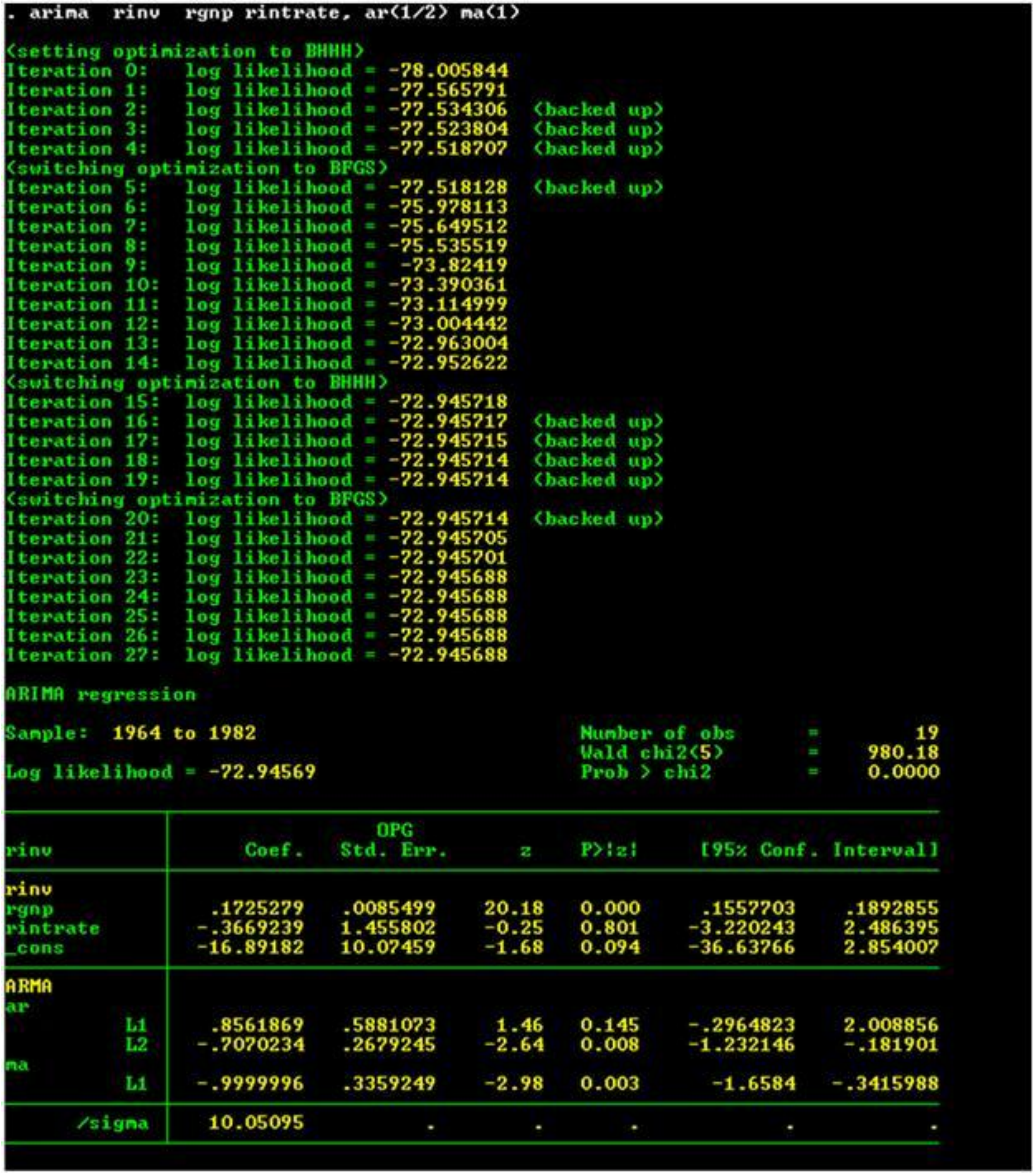
Figure 2.17.



```
. arima  rinv, ar(1/2) ma(1/2)

(setting optimization to BHHH)
Iteration 0:   log likelihood = -87.809565
Iteration 1:   log likelihood = -87.447909
Iteration 2:   log likelihood =  -87.35109
Iteration 3:   log likelihood = -87.268753
Iteration 4:   log likelihood = -87.203295
(switching optimization to BFGS)
Iteration 5:   log likelihood = -87.095176
Iteration 6:   log likelihood = -86.864369
Iteration 7:   log likelihood = -86.194856
Iteration 8:   log likelihood = -86.177722
Iteration 9:   log likelihood = -86.176414
Iteration 10:  log likelihood = -86.175405
Iteration 11:  log likelihood = -86.175308
Iteration 12:  log likelihood = -86.175249
Iteration 13:  log likelihood = -86.175245

ARIMA regression

Sample:  1964 to 1982                           Number of obs      =        19
                                                Wald chi2(4)       =     42.44
Log likelihood = -86.17525                      Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |                 OPG
rinv         |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
rinv         |
_cons        |    184.942   18.88555     9.79   0.000      147.927    221.9569
-------------+----------------------------------------------------------------
ARMA         |
ar           |
          L1 |    .875356   19.42014     0.05   0.964    -37.18742    38.93813
          L2 |  -.1040967   13.77821    -0.01   0.994    -27.10889     26.9007
ma           |
          L1 |  -4.075034   330.1145    -0.01   0.990    -651.0876    642.9375
          L2 |  -1.411536   117.5412    -0.01   0.990     -231.788     228.965
-------------+----------------------------------------------------------------
      /sigma |   4.985582   376.8707     0.01   0.989    -733.6675    743.6386
------------------------------------------------------------------------------
```

Estimation of an ARMA(2, 2) model of real investment.

Figure 2.18.



```
. arima  rinv  rgnp rintrate, ar(1/2) ma(1)

(setting optimization to BHHH)
Iteration 0:   log likelihood = -78.005844
Iteration 1:   log likelihood = -77.565791
Iteration 2:   log likelihood = -77.534306  (backed up)
Iteration 3:   log likelihood = -77.523804  (backed up)
Iteration 4:   log likelihood = -77.518707  (backed up)
(switching optimization to BFGS)
Iteration 5:   log likelihood = -77.518128  (backed up)
Iteration 6:   log likelihood = -75.978113
Iteration 7:   log likelihood = -75.649512
Iteration 8:   log likelihood = -75.535519
Iteration 9:   log likelihood =  -73.82419
Iteration 10:  log likelihood = -73.390361
Iteration 11:  log likelihood = -73.114999
Iteration 12:  log likelihood = -73.004442
Iteration 13:  log likelihood = -72.963004
Iteration 14:  log likelihood = -72.952622
(switching optimization to BHHH)
Iteration 15:  log likelihood = -72.945718
Iteration 16:  log likelihood = -72.945717  (backed up)
Iteration 17:  log likelihood = -72.945715  (backed up)
Iteration 18:  log likelihood = -72.945714  (backed up)
Iteration 19:  log likelihood = -72.945714  (backed up)
(switching optimization to BFGS)
Iteration 20:  log likelihood = -72.945714  (backed up)
Iteration 21:  log likelihood = -72.945705
Iteration 22:  log likelihood = -72.945701
Iteration 23:  log likelihood = -72.945688
Iteration 24:  log likelihood = -72.945688
Iteration 25:  log likelihood = -72.945688
Iteration 26:  log likelihood = -72.945688
Iteration 27:  log likelihood = -72.945688

ARIMA regression

Sample:  1964 to 1982                           Number of obs      =        19
                                                Wald chi2(5)       =    980.18
Log likelihood = -72.94569                      Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |                 OPG
        rinv |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
rinv         |
rgnp         |   .1725279   .0085499    20.18   0.000     .1557703    .1892855
rintrate     |  -.3669239   1.455802    -0.25   0.801    -3.220243    2.486395
_cons        |  -16.89182   10.07459    -1.68   0.094    -36.63766    2.854007
-------------+----------------------------------------------------------------
ARMA         |
ar           |
          L1 |   .8561869   .5881073     1.46   0.145    -.2964823    2.008856
          L2 |  -.7070234   .2679245    -2.64   0.008    -1.232146    -.181901
ma           |
          L1 |  -.9999996   .3359249    -2.98   0.003      -1.6584   -.3415988
-------------+----------------------------------------------------------------
      /sigma |   10.05095          .        .       .            .           .
------------------------------------------------------------------------------
```

Estimation of Equation (1) using an ARMA(2, 1) model.

	ARMA(1, 1)	ARMA(2, 1)	AR(1)	AR(2)	MA(1)
Intercept	185.307	185.6556	184.8208	185.2092	189.373
	(10.06)	(10.83)	(9.27)	(10.25)	(18.09)
AR (L1)	0.70936	1.76342	0.80307	0.95257	—
	(3.12)	(5.27)	(5.51)	(4.47)	—
AR (L2)	—	-0.81715	—	-0.18963	—
		(-3.21)		(-0.91)	
MA (L1)	0.26236	-0.99998	—	—	0.87262
	(0.90)	(-0.00)			(2.97)
Log likelihood	-86.1791	-85.8702	-86.47780	-86.21224	-88.48713
Wald χ2	26.96	422.60	30.36	31.65	8.81
Probability > χ2	0.0000	0.0000	0.0000	0.0000	0.0000
Sample size	19	19	19	19	19
(14,1)	1964-1982	1964-1982	1964-1982	1964-1982	1964-1982

Table 2.11. Estimation of various ARMA models of real investment.

The interpretation of these results is not obvious. We check the sensitivity of these results by estimation some other models. The results of these estimations are reported in Table 2 and Table 3. Based purely on ML tests, it would appear that AR(1) model in Table 2 is as good as any of the models describing the ARMA structure of real investments. On the other hand, the results reported in Table 3 suggests that the ARMA(2, 1) appears to be the best model to assume for the disturbance term in the estimates of Equation (1).

	AR(1)	**ARMA(1, 1)**	**ARMA(2, 1)**
Intercept	**-14.49489**	**-13.37455**	**-16.89182**
	(-0.26)	**(-0.23)**	**(-1.68)**
Real GNP	**0.17006**	**0.16912**	**0.17253**
	(3.96)	**(3.78)**	**(20.18)**
Real interest rate	**-0.82517**	**-0.92007**	**-0.33692**
	(-0.46)	**(-0.33)**	**(-0.25)**
AR (L1)	**0.27953**	**-0.02028**	**0.85619**
	(0.60)	**(-0.02)**	**(1.46)**
AR (L2)	**—**	**—**	**-0.70702**
			(-2.64)
MA (L1)	**—**	**0.41151**	**-1.00000**
		(0.42)	**(-2.98)**
Log likelihood	**-78.7868**	**-78.4279**	**-72.94569**
Wald χ^2	**26.30**	**31.86**	**980.18**
Probability > χ^2	**0.0000**	**0.0000**	**0.0000**

Sample size	19	19	19
Sample period	1964-1982	1964-1982	1964-1982

Table 2.12. Various ARMA estimates of Equation (1).

Other time-series concepts

There are a large number of additional time-series methods and issues that are not discussed in this module. These topics include, among others, ARCH and GARCH estimators, unit roots, the Dickey-Fuller test, and vector autoregression (VAR) models. There is no way to do justice to these topics in notes as short as these are. Moreover, it is necessary to discuss difference equations (the discrete version of differential equations) if one wants to understand many of these topic at anything more than an intuitive level. Those interested in these topics should enroll in the forecasting course (Economics 422) or, if they cannot, plan to read several textbooks on whatever econometric tool they need to understand.

Exercise

Exercise 2.2.1.

This exercise is designed to be sure you know how to use *Stata* in analyzing time-series data sets; there is no economic content in the exercise. The MS Excel file Rabun County Temperature Data reports the morning temperature (MornTemp) observed in Rabun County, Georgia for every day between March 15, 2005 to November 2, 2008. The data set includes a variable “edate” that is the daily date in *Stata* notation. The data set also includes dummy variables for the season, the month, and the year of each observation (with the Winter, the December, and the 2008 dummy variables omitted).

a. Create a graph of (a) the data set morntemp, (2) the autocorrelations of morntemp, and (3) the partial autocorrelations of morntemp (you will have to set the matrix size to some number greater than 43 using the command .set matsize #).

b. Estimate the following models:

1. ARMA(2,2) for morntemp.

2. ARMA(2,2) for morntemp as a function of the season dummy variables.

3. ARMA(2,2) for morntemp as a function of the monthly dummy variables.

4. ARMA(2,2) for morntemp as a function of the monthly dummy variables and the annual dummy variables.

5. ARMA(1,2) for morntemp as a function of the monthly dummy variables and the annual dummy variables.

6. ARMA(1,1) for morntemp as a function of the monthly dummy variables and the annual dummy variables.

c. Arrange the parameter estimates in a table and comment on them. Include the results of estimating (6) using OLS; what is the DW-statistic for this regression?

References

Cochran, D. and G. Orcutt (1949). Application of Least Squares Regression to Relationships Containing Autocorrelated Error Terms. *Journal of the American Statistical Association* 44: 32-61.

Enders, Walter (1995). *Applied Econometric Time Series* (New York: John Wiley & Sons, Inc.).

Greene, William H. (1990). *Econometric Analysis* (New York: Macmillan Publishing Company).

StataCorp (2003). *Stata Statistical Software: Release 8.0: Stata Time-Series Reference Manual* (College Station, TX: Stat Corporation).

2.3. Panel Data Models*

Equation Chapter 1 Section 1Notes on Panel Data Models

Introduction

Panel data methods are appropriate when the researcher has available observations that are both cross-sectional and time series. For example, one could form a panel data set with observations on the per capita consumption of tobacco for a set of OECD countries over the period 1960 to 2005. Usually the data is “stacked”—that is, all of the observations for country A is listed together in order of year before the data for country B, etc. It is also possible to stack the data by year—countries A to Z for 1960, countries A to Z for 1961, and so on through 2005.

Let *y* $_{\text{it}}$ be the per capita consumption of tobacco for country *i* in year *t*. We wish to model the per capita consumption of tobacco as a function of a set of observable independent variables like the price of tobacco, income, restrictions on tobacco advertising, and restrictions on tobacco consumption. Of course there are several sources of unobserved heterogeneity in that data set. In particular, we might expect that systematic differences in consumption patterns would exist due to differences in the customs and mores of the various countries in the sample. It also would be reasonable to assume that these country-level differences are be relatively stable over time. Additionally, we might expect that there would be differences the per capita consumption of tobacco over time due to changes in our understanding of the long run health effects of tobacco consumption. These changes might affect both (1) the level of consumption and (2) the responsiveness of the consumption of tobacco to changes in the explanatory variables.

In these notes we describe some of the ways of modeling panel data sets and discuss some of the issues associated with the estimation of these models. We also discuss how to use *Stata* to analyze panel data sets. We begin by considering some of the types of panel data model specifications.

Model specification

There are four general specifications of the panel data model available. The differences in these models reflect differing assumptions one might make and are listed below.

1. Slope coefficients are constant and the intercept varies over the individuals:

(2.35)

$$y_{it} = \alpha_i + \sum_{j=1}^{k} \beta_j x_{jit} + \varepsilon_{it}, i = 1, \ \ldots, N, i = 1, \ \ldots, N, \text{and} t = 1, \ \ldots, T.$$

2. Slope coefficients are constant and the intercept varies over the individuals and over time:

(2.36)

$$y_{it}=\alpha_{it}+\sum_{j=1}^{k}\beta_{j}x_{jit}+\varepsilon_{it}, i=1,\ \ldots\ ,N, \text{and} t=1,\ \ldots\ ,T.$$

3. All coefficients vary over individuals:

(2.37)

$$y_{it}=\alpha_{i}+\sum_{j=1}^{k}\beta_{ji}x_{jit}+\varepsilon_{it}, i=1,\ \ldots\ ,N, \text{and} t=1,\ \ldots\ ,T.$$

4. All coefficients vary over time and individuals:

(2.38)

$$y_{it}=\alpha_{it}+\sum_{j=1}^{k}\beta_{jit}x_{jit}+\varepsilon_{it}, i=1,\ \ldots\ ,N, \text{and} t=1,\ \ldots\ ,T.$$

These four models can be classified further, depending on whether the researcher assumes that the coefficients of the model are fixed or random. However, most research in economics is restricted to estimation of (1) and (2) because they strike a reasonable balance between being general enough without introducing unnecessary assumptions that can render estimation extremely difficult.

Estimation issues

Hsiao (2003: 27-30) discusses a convenient example of a panel data model that illustrates many of the important issues that arise with panel data. We make use of this example in what follows. Assume that we want to estimate a production function for farm production in order to determine if the farm industry exhibits increasing returns to scale. Assume the sample consists of observations for *N* farms over *T* years, giving a total sample size of *N T* . For simplicity, we assume that the Cobb-Douglas production is an adequate description of the production process. The general form of the Cobb-Douglas production function is:

(2.39) $q = \alpha_0 I_1{}^{\beta}{}_1 \cdots I_k{}^{\beta}{}_k$,

where q is output and I_j is the quantity of the j-th input (for example, land, machinery, labor, feed, and fertilizer). The parameter, β_j , is the output elasticity of the j-th input; the farms exhibit constant returns to scale if the output elasticities sum to one and either increasing or decreasing returns to scale if they sum to a value greater than or less than one, respectively. is the quantity of the *j*-th input (for example, land, machinery, labor, feed, and fertilizer). The parameter, is the output elasticity of the *j*-th input; the farms exhibit constant returns to scale if the output elasticities sum to one and either increasing or decreasing returns to scale if they sum to a value greater than or less than one, respectively.

Taking the natural logarithm of (5) gives $\ln q = \ln\alpha_0 + \beta_1 \ln I_1 + \cdots + \beta_k \ln I_k$. We can re-write this equation (adding an error term, as well as farm and year subscripts) giving:

(2.40) $y_{it} = \beta_0 + \beta_1 x_{1it} + \cdots + \beta_k x_{kit} + \varepsilon_{it}$,

where $y_{it} = \ln q_{it}$, , $\beta_0 = \ln\alpha_0$, $x_{jit} = \ln I_{jit}$, for $j = 1,\ldots,k$ and ε_{it} is an error term. One way to account for year and time effects is to assume:

(2.41) $\varepsilon_{it} = \lambda F_i + \eta P_t + \upsilon_{it}$,

where F_i is a measure of the unobserved farm specific effects on productivity and P_t is a measure of the unobserved changes in productivity that are the same for all farms but vary annually. Substitution of (7) into (6) gives:

$$y_{it} = (\beta_0 + \lambda F_i + \eta P_t) + \sum_{j=1}^{k} \beta_j x_{jit} + \upsilon_{it}$$

or

(2.42)

$$y_{it} = \alpha_{it} + \sum_{j=1}^{k} \beta_j x_{jit} + \upsilon_{it},$$

where $\alpha_{it} = \beta_0 + \lambda F_i + \eta P_t$. Thus, (8) is equivalent to (2). Moreover, if we assume that $\eta = 0$, we get

(2.43)

$$y_{it} = \alpha_i + \sum_{j=1}^{k} \beta_j x_{jit} + \upsilon_{it},$$

where $\alpha_i = \beta_0 + \lambda F_i$. Thus, (9) is equivalent to (1).

Fixed-effects models

A natural way to make (9) operational is to introduce a dummy variable, D_i, for each farm so that the intercept term becomes:

(2.44)

$$\alpha_i = \alpha_1 + \alpha_2 D_2 + \cdots + \alpha_m D_m = \alpha_1 + \sum_{j=2}^{m} \alpha_j D_j,$$

where $D_j = 1$ if $j = i$ and 0 otherwise. This substitution is equivalent to replacing the intercept term with a dummy variable for each farm and letting the farm dummy variable "sweep out" the farm-specific effects. In this specification the slope terms are the same for every farm while the intercept term is given for farm j by $\alpha_1 + \alpha_j$. Clearly, the intercept term for the first farm is equal to just α_1 . This specification is known as the *fixed effect model* and is estimated using ordinary least squared (OLS). We can extend the fixed-effects model to fit (8) by including a dummy variable for each time period except one.

In sum, *fixed-effects models* assume either (or both) that the omitted effects that are specific to cross-sectional units are constant over time or that the effects specific to time are constant over the cross-sectional units. This method is equivalent to including a dummy variable for all but one of the cross-sectional units and/or a dummy variable for all but one of the time periods.

Random-effects models

An alternative approach to treating the α_i in (1) as fixed constants over time is to treat it as a random variable. Returning to (1) where the intercepts vary due to individual level differences, we have $y_{it} = \alpha_i + \sum_{j=1}^{k} \beta_k x_{kit} + \varepsilon_{it}.$ Treating α_i as a random variable is equivalent to setting the model up as:

(2.45)

$$y_{it} = \alpha + \sum_{j=1}^{k} \beta_j x_{jit} + (\alpha_i + \lambda_t + \varepsilon_{it}).$$

For simplicity we consider only the case when $\lambda_t = 0$. Thus, the error term for (11) is $(\alpha_i + \varepsilon_{it}).$ We assume that

(2.46)

$$\begin{aligned}
&E(\alpha_i) = E(\varepsilon_{it}) = 0, \\
&E(\alpha_i \varepsilon_{it}) = 0, \\
&E(\alpha_i \alpha_j) = \begin{cases} \sigma_\alpha^2 & \text{if} i = j \\ 0 & \text{if} i \neq j \end{cases}, \quad \text{and} \\
&E(\varepsilon_{it} \varepsilon_{js}) = \begin{cases} \sigma_\varepsilon^2 & \text{if} i = j,\ t = s \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

We also assume that all of the elements of the error term are uncorrelated with the explanatory variables, x_j.

The key econometric issue is that the presence of α_i in the error term means that the correlation among the residual of the same cross-sectional unit is not zero; the error terms for one farm, for instance, are correlated with each other. Therefore, the error terms exhibit heteroskedasticity. The appropriate estimation technique is generalized-least-squares, a technique that attempts to adjust the parameter estimates (and their standard error estimates) for heteroskedasticity and autocorrelation. Alternatively one can assume that α_i and ε_{it} are normally distributed and use a ML estimator. Hsiao [2003: 35-41] and Cameron and Trivedi [2005: 699-716] offer greater detail on the estimation of the parameters of both the fixed-effects and the random-effects models. It is enough for our purposes to accept that the econometricians have found a number of ways to estimate these parameters.

Random-effects or fixed effect model?

Economists generally prefer to use fixed-effects models. The decision to use fixed-effects or random-effects does not matter when *T* is large because the two methods will yield the same estimates of the parameters. When the number of individual categories (*N*) is

large and the number of time periods (*T*) is small, the choice of which model to use becomes unclear. Hsiao summarized this somewhat arcane issue with the following observations:

> ***If the effects of omitted variables can be appropriately summarized by a random variable and the individual (or time) effects represent the ignorance of the investigator, it does not see reasonable to treat one source of ignorance () as fixed and the other source of ignorance () as random. It appears that one way to unify the fixed-effects and random-effects models is to assume from***
>
> ***the outset that the effects are random. The fixed-effects model is viewed as one in which investigators make inferences conditional on the effects that are in the sample. The random-effects model is viewed as one in which investigators make unconditional or marginal inferences with respect to the population of all effects. There is really no distinction in the "nature (of the effect)." It is up to the investigator to decide whether to make inference with respect to population characteristics or only with respect to the effects that are in the sample. Hsiao [2003: 43]***

Needless to say, Hsiao's advice may well leave many researchers without any idea of whether to use a random-effects or a fixed-effects model. *In your own research I suggest that you consult an econometrician for advice*.

There is one problem that arises when using a fixed-effects model. Assume that you have a sample of observations for a large number of individuals over a period of years. If you use a fixed-effects model, you will not be able to find parameter estimates for any variable like race or sex that do not change over the time period of the sample. The reason for this limitation is that the time-constant variables are perfectly correlated with the dummy variables used for the fixed-effects. A similar problem arises if the fixed-effects are for years (rather than individuals). You cannot include a variable is constant for all individuals in any given year. Quite often the individual-constant (or time-constant) variable is not of interest and nothing is lost by not having the parameter estimate. On the other hand, the random-effects model does not have this problem because the estimation makes use of differences amongst the individuals to estimate a parameter for the individual-constant variable.[21] We discuss in the next section an example in which this "problem" arises.

What would be nice is if there were a statistical test that allows us to decide if the random-effects model is the appropriate model? The *Hausman test* offers such a statistical test. The Hausman (specification) test exploits the fact that the parameters for the random-effects model should be not be statistically different from those found using a fixed-effects specification. If one observes a chi-squared value greater than the critical value you can conclude that the parameter estimates for the random-effects model are statistically different from the parameter estimates for a model using an assumption of fixed-effects, then you can conclude that the

random-effects model is misspecified. Unfortunately, the misspecification could be due to the fact that the fixed-effects model is appropriate or it could be due to the unobserved error terms being correlated with the included explanatory variables. If the latter is the case, then one might consider augmenting the model with an appropriate measure of the part of the unobserved effect that is correlated with the error term. What we are describing is that same thing that happens when omitted variables are correlated with the error term—the parameter estimates are biased. We include an example of how to use *Stata* to perform the Housman specification test.

Estimation of panel data models in Stata

General comments

There are three commands that matter in setting up the panel data. The first two commands precede the regression command because they establish which variable denotes the time period and which variable denotes the cross-sectional unit. These commands are:

```
.iis [variable name]
```

```
.tis [variable name]
```

The command for estimating the fixed-effects model is:

```
. xtreg depvar [varlist], fe
```

The command for estimating the random-effects model is:

```
. xtreg depvar [varlist], re
```

If the part of the command with the comma and either re or fe is omitted, *Stata* will assume that you want to estimate the random-effects model.

Understanding Stata output

To understand the *Stata* output we need to return to the algebra of the model. Assume that we are fitting a model of the following form:

(2.47)

$$y_{it} = \alpha + \sum_{j=1}^{k} \beta_j x_{jit} + \nu_i + \varepsilon_{it}, i = 1, \ \dots, N, \text{and} t = 1, \ \dots, T.$$

We can sum (13) over *t* (holding the individual unit constant) and divide by *T* to get:

(2.48)

$$\bar{y}_i = \alpha + \sum_{j=1}^{k} \beta_j \bar{x}_{ji} + \nu_i + \bar{\varepsilon}_i,$$

where $\bar{y}_i = \frac{\sum_{t=1}^{T} y_{it}}{T}$, $\bar{x}_{ji} = \frac{\sum_{t=1}^{T} x_{it}}{T}$, and $\bar{\varepsilon}_i = \frac{\sum_{t=1}^{T} \varepsilon_{it}}{T}$. Thus, (14) uses the mean values for each cross-sectional unit. We can subtract (14) from (13) to get:

(2.49)

$$(y_{it} - \bar{y}_i) = \sum_{j=1}^{k} \beta_j (x_{jit} - \bar{x}_{ji}) + (\varepsilon_{it} - \bar{\varepsilon}_i).$$

Equations (13), (14), and (15) are the basis of *Stats's* estimates of the parameters of the model. In particular, the command xtreg, fe uses OLS to estimate (15); this is known as the *fixed-effects* estimator (or the *within* estimator). The command xtreg, be uses OLS to estimate (14) and is known as the *between* estimator. The command xtreg, re—the random-effects estimator—is a weighted average of the between and within estimators, where the weight is a function of the variances of and (and respectively).[22]

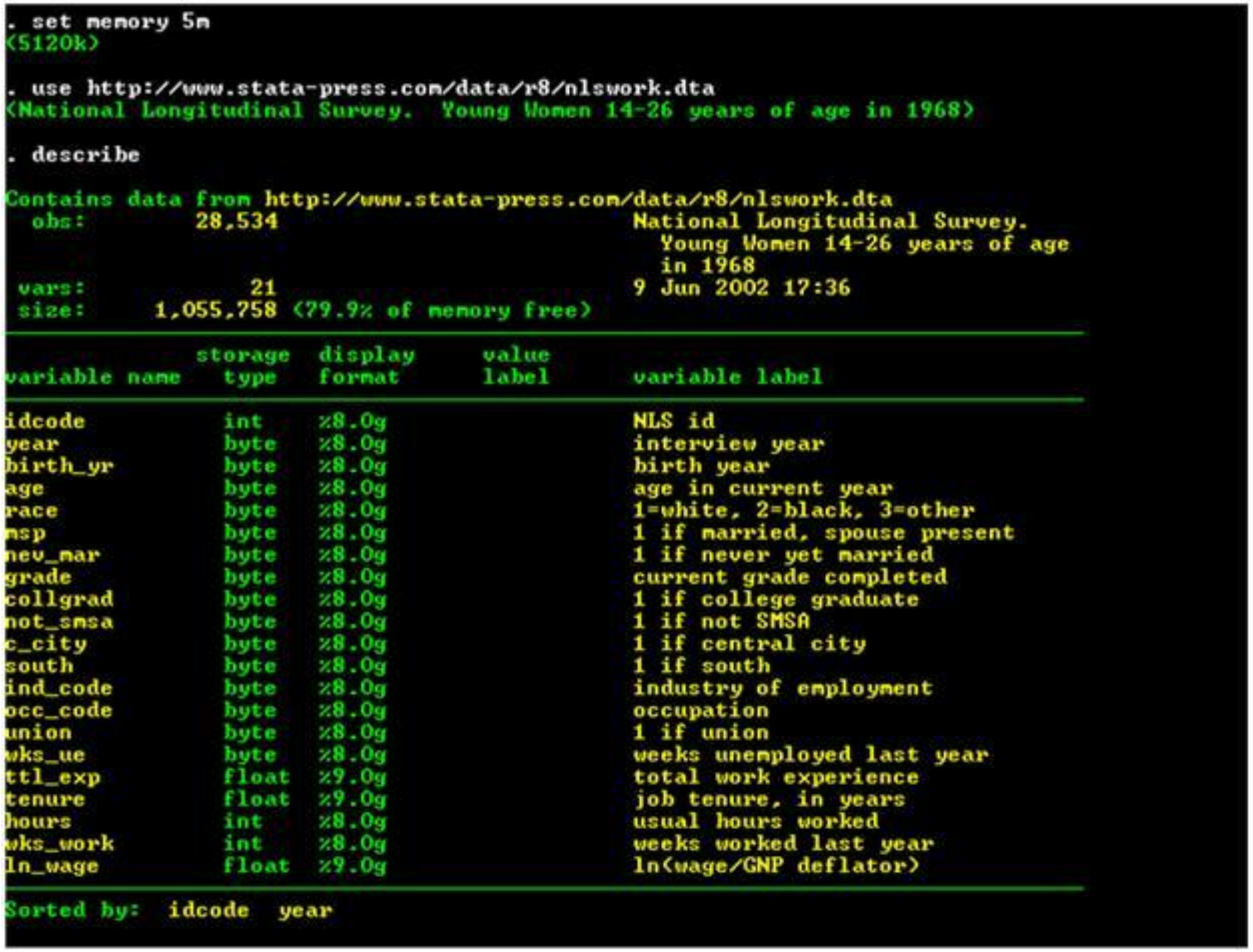
In general, you will not make use of the between estimator. However, these three equations do lie at the basis of the goodness-of-fit measures that *Stata* reports. In particular, *Stata* output reports three "R-squareds"[23]—the *overall-R²* the *between-R²* and the

***within-R*^{2} These three R-squareds are derived using one of the three equations. In particular, the overall-*R*^{2} uses (13); the between-*R*^{2} uses (14); and the within-*R*^{2} uses (15).**

Example 2.3. A panel data analysis using *Stata*

In this example we follow the example offered in the *Stata* manual and use a large data set from the National Longitudinal Survey of wage data on 28,534 women who were between 14 and 26 years of age in 1968. The women were surveyed in each of the 21 years between 1968 and 1988 except for the six years 1974, 1976, 1979, 1981, 1984, and 1986. The study is focused on the determinants of wage levels, as measured by the natural logarithm of real wages.

Figure 2.19.

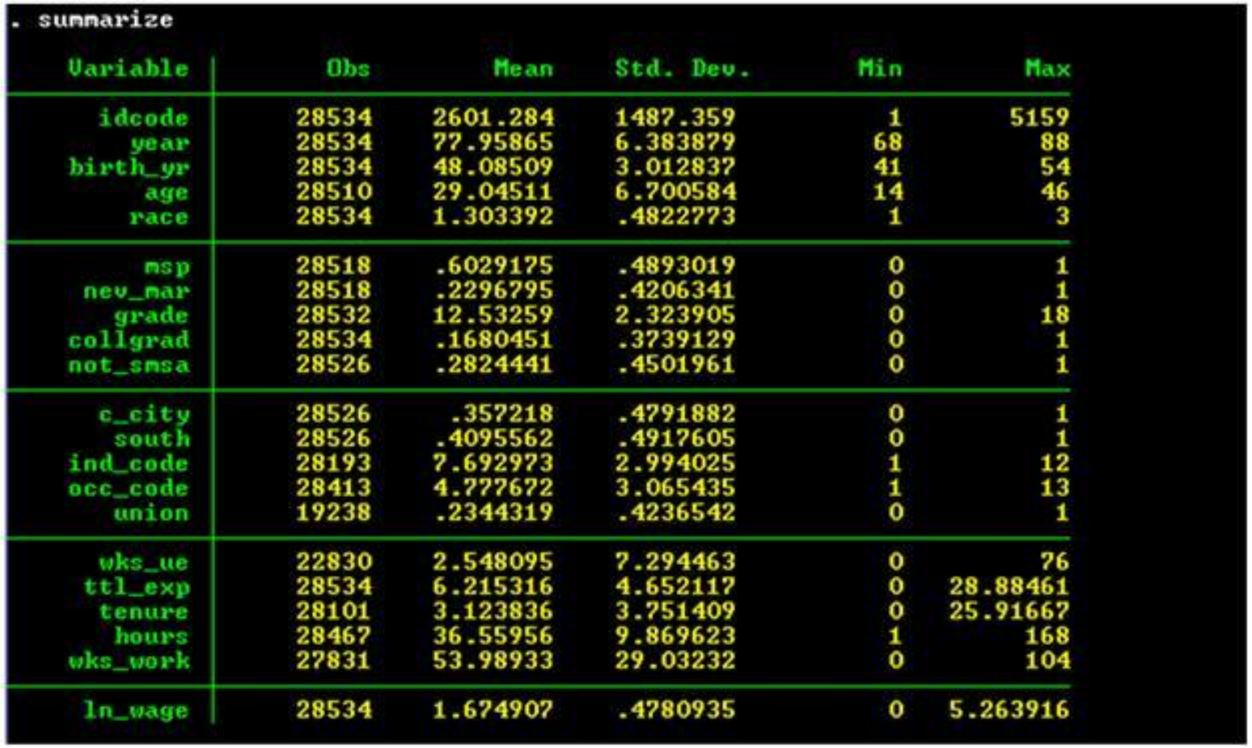


```
. set memory 5m
(5120k)

. use http://www.stata-press.com/data/r8/nlswork.dta
(National Longitudinal Survey.  Young Women 14-26 years of age in 1968)

. describe

Contains data from http://www.stata-press.com/data/r8/nlswork.dta
  obs:        28,534                           National Longitudinal Survey.
                                                 Young Women 14-26 years of age
                                                 in 1968
 vars:            21                           9 Jun 2002 17:36
 size:     1,055,758 (79.9% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
idcode          int    %8.0g                  NLS id
year            byte   %8.0g                  interview year
birth_yr        byte   %8.0g                  birth year
age             byte   %8.0g                  age in current year
race            byte   %8.0g                  1=white, 2=black, 3=other
msp             byte   %8.0g                  1 if married, spouse present
nev_mar         byte   %8.0g                  1 if never yet married
grade           byte   %8.0g                  current grade completed
collgrad        byte   %8.0g                  1 if college graduate
not_smsa        byte   %8.0g                  1 if not SMSA
c_city          byte   %8.0g                  1 if central city
south           byte   %8.0g                  1 if south
ind_code        byte   %8.0g                  industry of employment
occ_code        byte   %8.0g                  occupation
union           byte   %8.0g                  1 if union
wks_ue          byte   %8.0g                  weeks unemployed last year
ttl_exp         float  %9.0g                  total work experience
tenure          float  %9.0g                  job tenure, in years
hours           int    %8.0g                  usual hours worked
wks_work        int    %8.0g                  weeks worked last year
ln_wage         float  %9.0g                  ln(wage/GNP deflator)
-------------------------------------------------------------------------------
Sorted by:  idcode  year
```

Loading in the data set into *Stata* with a description of the data.

Figure 1 shows the commands used to put the data into *Stata*. The first command (set memory 5m) increases the size of the memory that the program uses; I did this because of the large sample size. The use command accesses that data from the *Stata* web site. The describe command calls up a description of the variables. Figure 2 presents a summary of the data using the command summerize.

Figure 2.20.



```
. summarize

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      idcode |     28534    2601.284    1487.359          1       5159
        year |     28534    77.95865    6.383879         68         88
    birth_yr |     28534    48.08509    3.012837         41         54
         age |     28510    29.04511    6.700584         14         46
        race |     28534    1.303392    .4822773          1          3
-------------+--------------------------------------------------------
         msp |     28518    .6029175    .4893019          0          1
     nev_mar |     28518    .2296795    .4206341          0          1
       grade |     28532    12.53259    2.323905          0         18
    collgrad |     28534    .1680451    .3739129          0          1
    not_smsa |     28526    .2824441    .4501961          0          1
-------------+--------------------------------------------------------
      c_city |     28526     .357218    .4791882          0          1
       south |     28526    .4095562    .4917605          0          1
    ind_code |     28193    7.692973    2.994025          1         12
    occ_code |     28413    4.777672    3.065435          1         13
       union |     19238    .2344319    .4236542          0          1
-------------+--------------------------------------------------------
      wks_ue |     22830    2.548095    7.294463          0         76
     ttl_exp |     28534    6.215316    4.652117          0   28.88461
      tenure |     28101    3.123836    3.751409          0   25.91667
       hours |     28467    36.55956    9.869623          1        168
    wks_work |     27831    53.98933    29.03232          0        104
-------------+--------------------------------------------------------
     ln_wage |     28534    1.674907    .4780935          0   5.263916
```

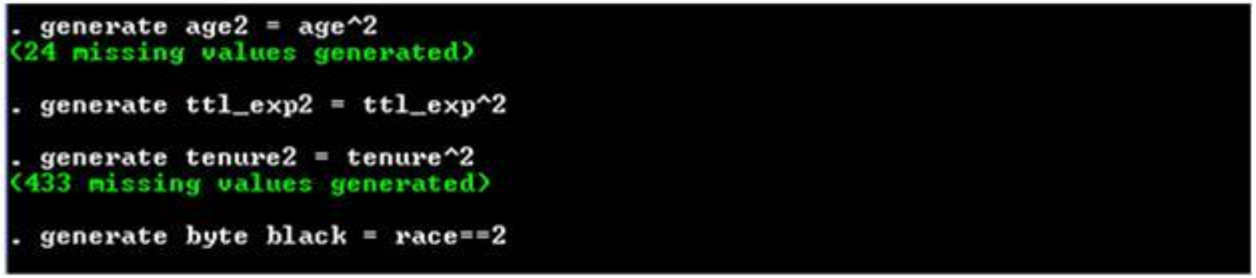
Summary of the data.

There are several transformations of the variables that we will need. In particular, we want to include the squares of several of the variables in our regression—age (age), work experience (ttl_exp), and job tenure (tenure). The reason we want to use the square of these variables is that we have reason to believe that wages have a non-linear relationship with these variables. For instance, consider the number of years a worker has been on the job, Tenure. Theory suggests that wages increase over a worker's work-life at a decreasing rate. Thus, if the equation we are estimating is $y = \ln w = \beta_0 + \beta_1 Tenure + \beta_2 Tenure^2 + \cdots$, what we expect is that: $\frac{\partial y}{\partial Tenure} = \beta_1 + 2\beta_2 Tenure > 0$ **and** $\frac{\partial^2 y}{\partial Tenure^2} = 2\beta_2 < 0.$ **The only way that this last equation can be true is if $\beta_2 < 0$.**

Moreover, if this is true, the first-derivative implies that $\beta_1 > -2\beta_2 Tenure > 0$. Also, notice that we can determine the number of years in a job when wages reach a peak; *y* reaches a maximum at the age where $\frac{\partial y}{\partial Tenure} = \beta_1 + 2\beta_2 Tenure = 0$. or when $Tenure = -\frac{\beta_1}{2\beta_2}$. The fact that $\frac{\partial^2 y}{\partial Tenure^2} = 2\beta_2 < 0$ guarantees that this point is indeed a maximum.

Additionally, because race is a categorical variable that has three potential values—1 if white, 2 if black, and 3 otherwise—we have to create a dummy variable in order to use this variable. The transformations we use are shown in Figure 3.

Figure 2.21.

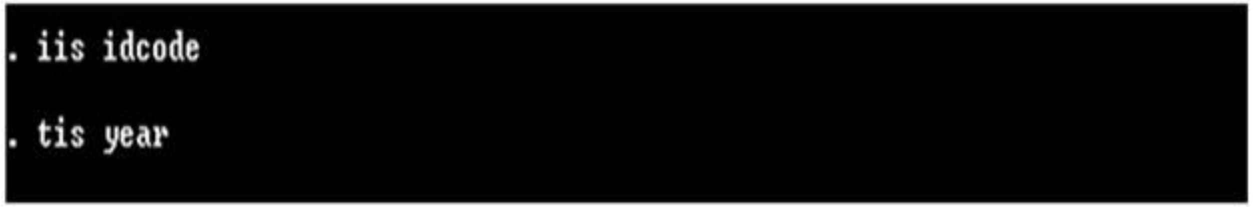


```
. generate age2 = age^2
(24 missing values generated)

. generate ttl_exp2 = ttl_exp^2

. generate tenure2 = tenure^2
(433 missing values generated)

. generate byte black = race==2
```

Transformations of the variables to create new variables.

The last step before estimating the regressions is to identify the data set as a panel data. shows the two commands that must be entered in order for *Stata* to know that idcode is the individual category and that year is the time series variable. Figure 4 shows these two commands.

Figure 2.22.

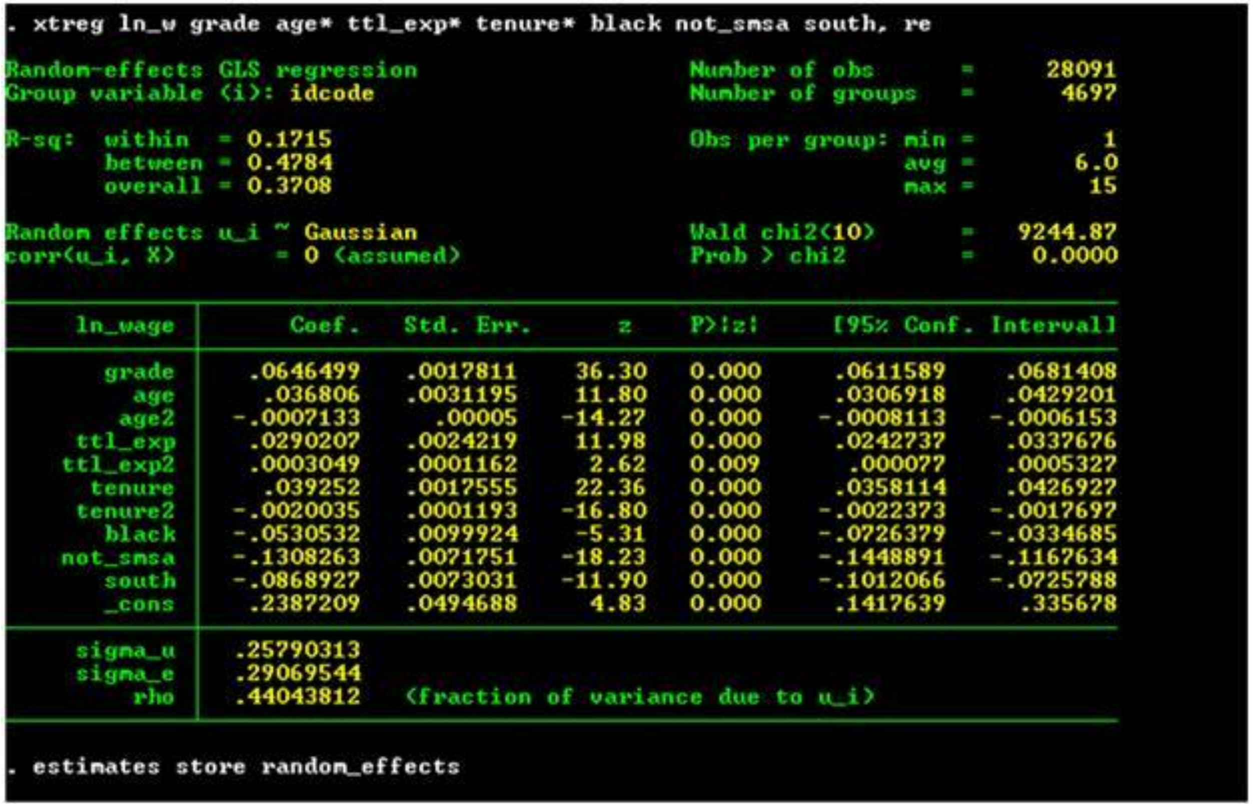


```
. iis idcode

. tis year
```

Declaring the category and time identifiers.

We are now ready to estimate the model (the natural logarithm of wages as a function of various variables). We begin with the random-effects model. Figure 5 shows the command and the results of the estimation of the random-effects model. There are several things to note here. First, in the command we are able to refer to all variables that have age in them by using age*, the * tells *Stata* to use and variable that begins with the letters age. Second, we will need to use the estimation results in the Hausman test. Thus, we have stored these results in "random_effects" using the command estimates store random_effects.

Figure 2.23.



```
. xtreg ln_w grade age* ttl_exp* tenure* black not_smsa south, re

Random-effects GLS regression                   Number of obs      =     28091
Group variable (i): idcode                      Number of groups   =      4697

R-sq:  within  = 0.1715                         Obs per group: min =         1
       between = 0.4784                                        avg =       6.0
       overall = 0.3708                                        max =        15

Random effects u_i ~ Gaussian                   Wald chi2(10)      =   9244.87
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
     ln_wage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       grade |   .0646499   .0017811    36.30   0.000     .0611589    .0681408
         age |    .036806   .0031195    11.80   0.000     .0306918    .0429201
        age2 |  -.0007133     .00005   -14.27   0.000    -.0008113   -.0006153
     ttl_exp |   .0290207   .0024219    11.98   0.000     .0242737    .0337676
    ttl_exp2 |   .0003049   .0001162     2.62   0.009      .000077    .0005327
      tenure |    .039252   .0017555    22.36   0.000     .0358114    .0426927
     tenure2 |  -.0020035   .0001193   -16.80   0.000    -.0022373   -.0017697
       black |  -.0530532   .0099924    -5.31   0.000    -.0726379   -.0334685
    not_smsa |  -.1308263   .0071751   -18.23   0.000    -.1448891   -.1167634
       south |  -.0868927   .0073031   -11.90   0.000    -.1012066   -.0725788
       _cons |   .2387209   .0494688     4.83   0.000     .1417639     .335678
-------------+----------------------------------------------------------------
     sigma_u |  .25790313
     sigma_e |  .29069544
         rho |  .44043812   (fraction of variance due to u_i)
------------------------------------------------------------------------------

. estimates store random_effects
```

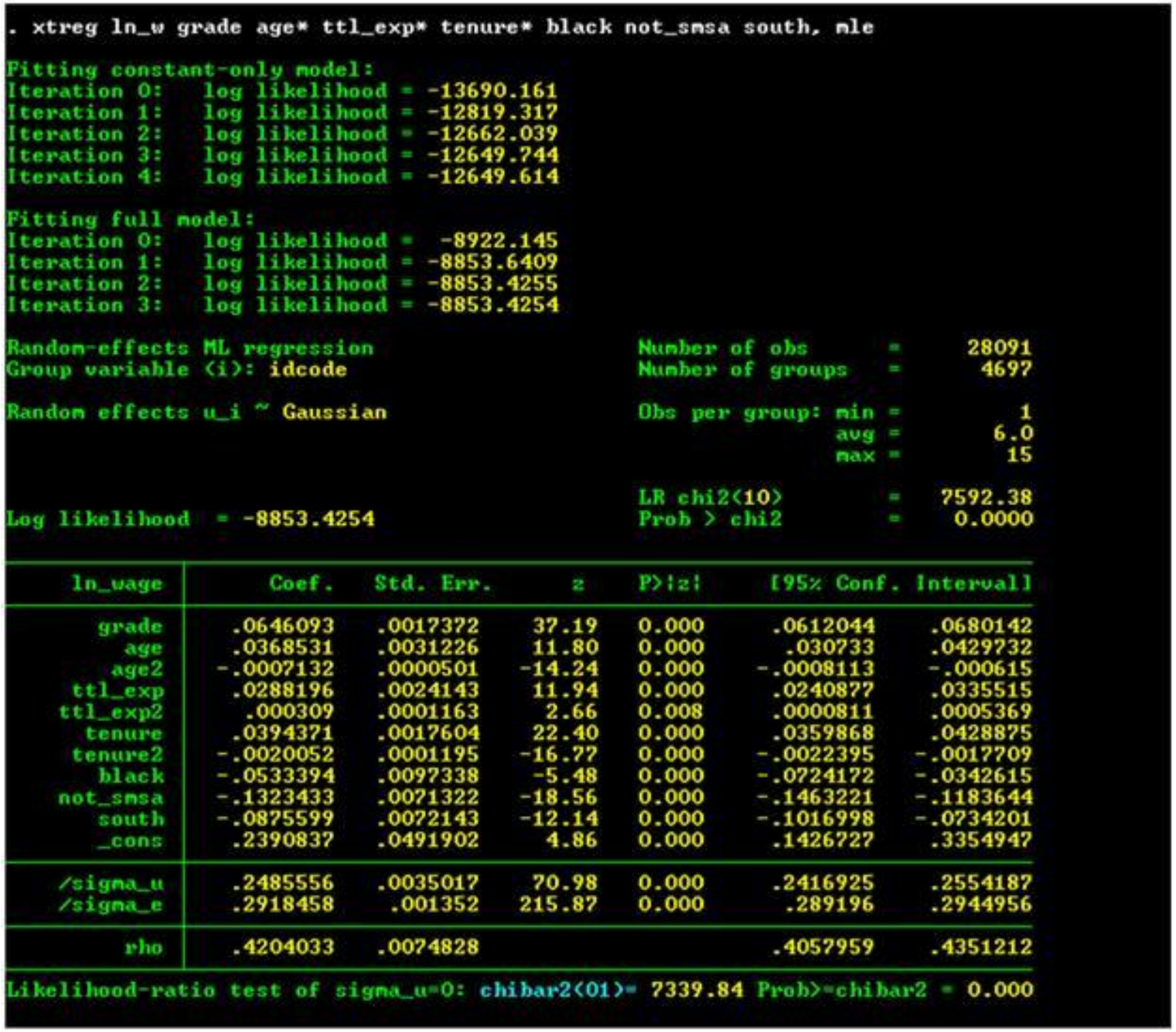
The random-effects estimation.

Notice that three R-squared values are reported in Figure 5. Also, wages reach a peak when the woman is $-\frac{0.036806}{2(-0.0007133)}=25.7998$ years old and after 9.795857 years on the job. The interpretation of the other variables demands a bit of algebra. For instance, the fact that black is a dummy variable affects our interpretation; when an individual is a black, her wage level is: $\ln w_B = \beta_0 + \beta_1 + \cdots$. When she is nonblack, her wage level is $\ln w_{NB} = \beta_0 + \cdots$. Thus, we have: $\ln w_B - \ln w_{NB} = \beta_1$ or $\frac{w_B}{w_{NB}} = e^{\beta_1} = e^{-0.0530532} = 0.94833.$ Thus, the wage level of a black is, everything else held constant, 94.8 percent of the wage level of a nonblack.

If we assume that grade is a continuous variable (it really is not), we have the following interpretation of the parameter: $\ln w = \beta_0 + \beta_1 grade + \cdots$ implies that $\frac{1}{w}\frac{\partial w}{\partial grade} = \beta_1$. Thus, in our case a increase of 1 year of schooling causes wages to increase by 6.46 percent.

We can compare the results of using the re option with using the mle option (which directs *Stata* to use maximum likelihood techniques to estimate the parameters of the system. The mle parameter estimates, shown in Figure 6, are the same as those generated using the re command. However, the estimates of the standard errors (and, thus, the z-values) are different.

Figure 2.24.



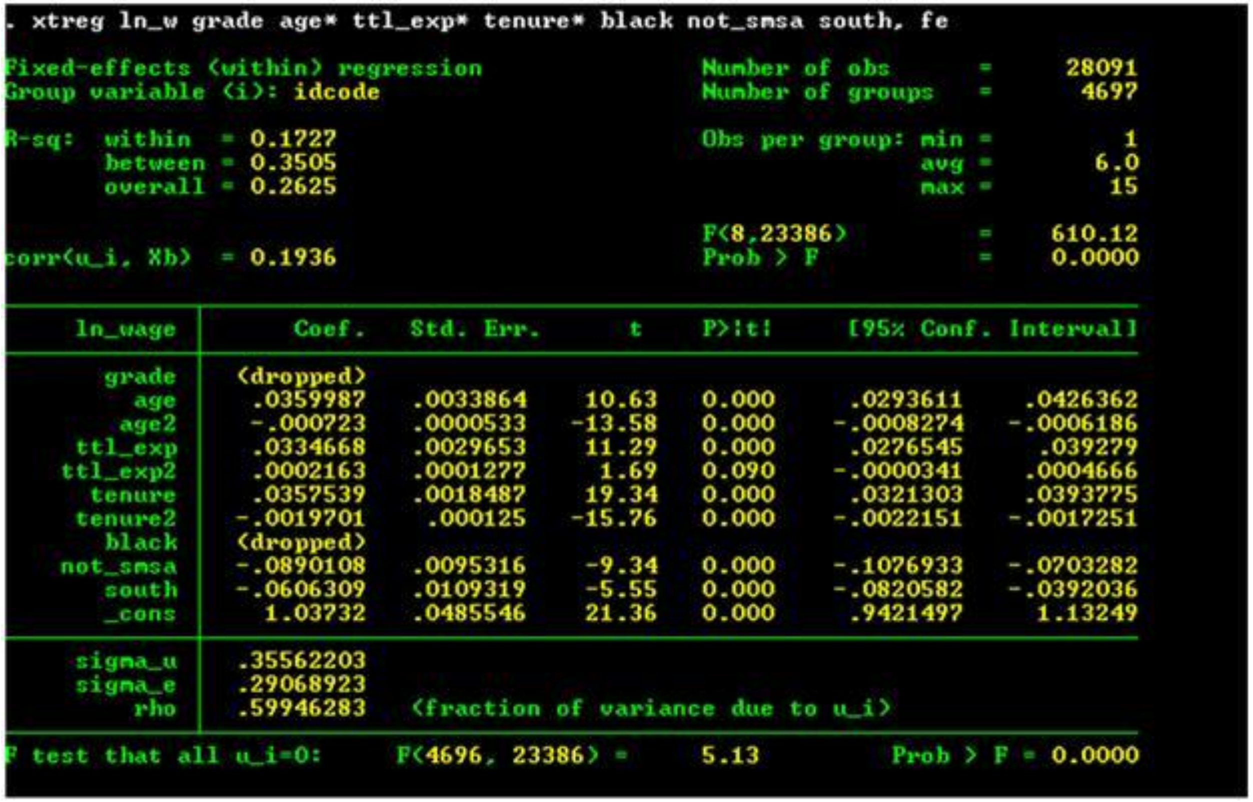
```
. xtreg ln_w grade age* ttl_exp* tenure* black not_smsa south, mle

Fitting constant-only model:
Iteration 0:   log likelihood = -13690.161
Iteration 1:   log likelihood = -12819.317
Iteration 2:   log likelihood = -12662.039
Iteration 3:   log likelihood = -12649.744
Iteration 4:   log likelihood = -12649.614

Fitting full model:
Iteration 0:   log likelihood =  -8922.145
Iteration 1:   log likelihood = -8853.6409
Iteration 2:   log likelihood = -8853.4255
Iteration 3:   log likelihood = -8853.4254

Random-effects ML regression                    Number of obs      =     28091
Group variable (i): idcode                      Number of groups   =      4697

Random effects u_i ~ Gaussian                   Obs per group: min =         1
                                                               avg =       6.0
                                                               max =        15

                                                LR chi2(10)        =   7592.38
Log likelihood  = -8853.4254                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
     ln_wage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       grade |   .0646093   .0017372    37.19   0.000     .0612044    .0680142
         age |   .0368531   .0031226    11.80   0.000      .030733    .0429732
        age2 |  -.0007132   .0000501   -14.24   0.000    -.0008113    -.000615
     ttl_exp |   .0288196   .0024143    11.94   0.000     .0240877    .0335515
    ttl_exp2 |    .000309   .0001163     2.66   0.008     .0000811    .0005369
      tenure |   .0394371   .0017604    22.40   0.000     .0359868    .0428875
     tenure2 |  -.0020052   .0001195   -16.77   0.000    -.0022395   -.0017709
       black |  -.0533394   .0097338    -5.48   0.000    -.0724172   -.0342615
    not_smsa |  -.1323433   .0071322   -18.56   0.000    -.1463221   -.1183644
       south |  -.0875599   .0072143   -12.14   0.000    -.1016998   -.0734201
       _cons |   .2390837   .0491902     4.86   0.000     .1426727    .3354947
-------------+----------------------------------------------------------------
    /sigma_u |   .2485556   .0035017    70.98   0.000     .2416925    .2554187
    /sigma_e |   .2918458    .001352   215.87   0.000      .289196    .2944956
-------------+----------------------------------------------------------------
         rho |   .4204033   .0074828                      .4057959    .4351212
------------------------------------------------------------------------------
Likelihood-ratio test of sigma_u=0: chibar2(01)= 7339.84 Prob>=chibar2 = 0.000
```

The maximum likelihood estimation.

The estimation of the fixed-effects model is straightforward and is shown in Figure 7. The command is the same as in the random-effects model but with the re replaced by fe. Notice from the results that the variables grade and black are dropped from the

estimation results. They are dropped because the amount of schooling and race of an individual is fixed over all observations. These two variables, thus, are perfectly correlated with the dummy variables that hold constant the individual level characteristics. The effects of education and race differences are absorbed into the residual.

Figure 2.25.

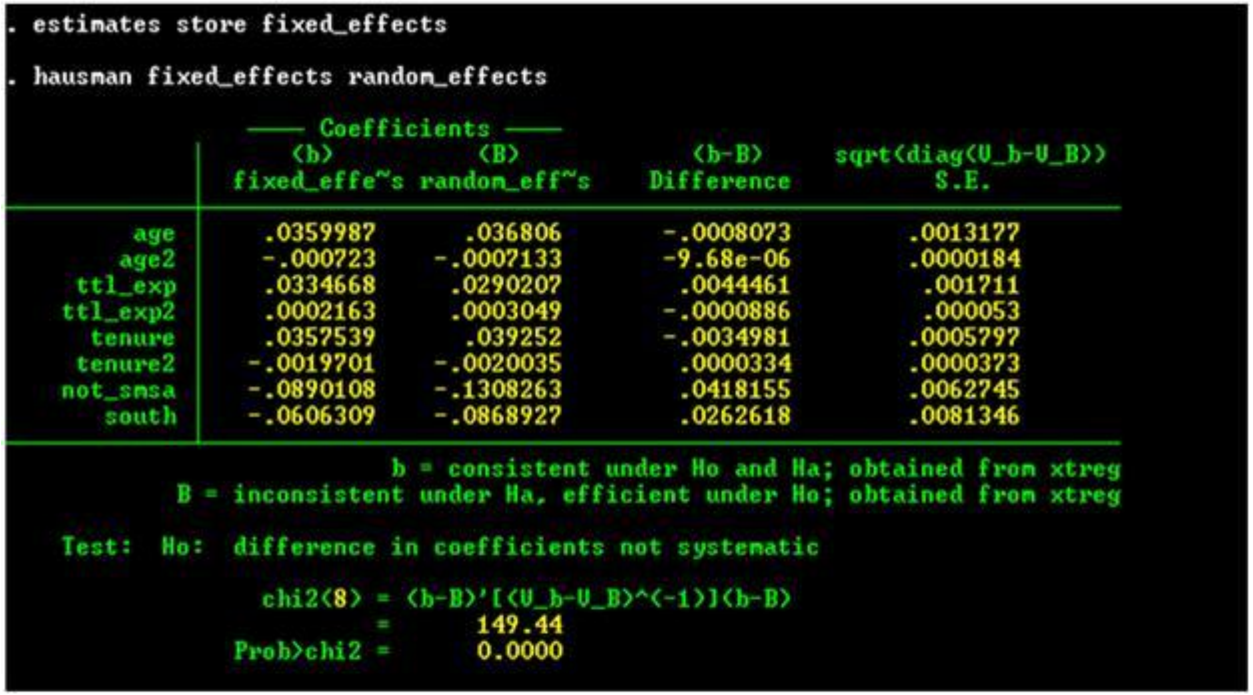


```
. xtreg ln_w grade age* ttl_exp* tenure* black not_smsa south, fe

Fixed-effects (within) regression               Number of obs      =     28091
Group variable (i): idcode                      Number of groups   =      4697

R-sq:  within  = 0.1727                         Obs per group: min =         1
       between = 0.3505                                        avg =       6.0
       overall = 0.2625                                        max =        15

                                                F(8,23386)         =    610.12
corr(u_i, Xb)  = 0.1936                         Prob > F           =    0.0000

------------------------------------------------------------------------------
     ln_wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       grade |  (dropped)
         age |   .0359987   .0033864    10.63   0.000     .0293611    .0426362
        age2 |   -.000723   .0000533   -13.58   0.000    -.0008274   -.0006186
     ttl_exp |   .0334668   .0029653    11.29   0.000     .0276545     .039279
    ttl_exp2 |   .0002163   .0001277     1.69   0.090    -.0000341    .0004666
      tenure |   .0357539   .0018487    19.34   0.000     .0321303    .0393775
     tenure2 |  -.0019701    .000125   -15.76   0.000    -.0022151   -.0017251
       black |  (dropped)
    not_smsa |  -.0890108   .0095316    -9.34   0.000    -.1076933   -.0703282
       south |  -.0606309   .0109319    -5.55   0.000    -.0820582   -.0392036
       _cons |    1.03732   .0485546    21.36   0.000     .9421497     1.13249
-------------+----------------------------------------------------------------
     sigma_u |  .35562203
     sigma_e |  .29068923
         rho |  .59946283   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(4696, 23386) =     5.13         Prob > F = 0.0000
```

The fixed-effects estimation.

The estimates of the parameter values for the fixed-effects model are very similar to those found for the random-effects model with the exception for the parameters associated with not living in an SMSA (not_smsa) and with living in the South (south). The random-effects model suggests that the wage level for someone living outside of a SMSA is 87.6 percent of the wage level of someone living in an SMSA; in the fixed-effects model, the wage level outside the SMSA is estimated to be 91.5 percent of the wage level of a woman living in a SMSA. The random-effects model estimates wages in the South are 91.6 percent the level of wages outside the South; the fixed-effects model fixes this wage premium at 91.6 percent.

Figure 2.26.



```
. estimates store fixed_effects

. hausman fixed_effects random_effects

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             | fixed_effe~s random_eff~s    Difference          S.E.
-------------+----------------------------------------------------------------
         age |    .0359987      .036806       -.0008073        .0013177
        age2 |    -.000723    -.0007133       -9.68e-06        .0000184
     ttl_exp |    .0334668     .0290207        .0044461         .001711
    ttl_exp2 |    .0002163     .0003049       -.0000886         .000053
      tenure |    .0357539      .039252       -.0034981        .0005797
     tenure2 |   -.0019701    -.0020035        .0000334        .0000373
    not_smsa |   -.0890108    -.1308263        .0418155        .0062745
       south |   -.0606309    -.0868927        .0262618        .0081346
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(8) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =      149.44
                Prob>chi2 =      0.0000
```

The Hausman test results.

The final issue we discuss in this example is the Hausman specification test. If the model is correctly specified and if v_i is uncorrelated with the explanatory variables, then the parameter estimates in the two models should not be statistically different. As shown in Figure 8, we first must same the results of the fixed-effects estimation using the command estimates store fixed_effects. The null hypothesis is that the the difference in that parameter estimates is not systematic. The appropriate test statistic is the $\chi^2(8)$, where the degrees of freedom are equal to the number of parameters in the model (8). The chi-squared statistic of 149.44 is greater than the critical value and we must reject the null hypothesis. The *Stata* offers this interpretation of this result:

> *What does this mean? We have an unpleasant choice: we can admit that our model is misspecified—that we have not parameterized it correctly—or we can hold to our specification*
>
> *being correct, in which case the observed differences must be due to the zero-correlation of and the assumption. [StataCorp: 202]*

Exercises

Exercise 2.3.1.

Estimation of a Labor Supply Function. An important issue in labor economics is the responsiveness of the number of hours worked to wages. Because labor supply curves can, in theory, be backward-bending, the sign and size of the impact of wages on the amount of labor supplied is an empirical issue. In this project you are to estimate the demand for labor curve for a cross-section of adult males.

(2.50)

The model to be estimated is:

$$y_{it} = \beta_0 + \beta_1 h_{it} + \beta_2 Age_{it} + \beta_3 Age^2_{it} + \beta_4 NC_{it} + \beta_5 HI_{it} + \varepsilon_{it}$$

where:

y_{it} = natural logarithm of individual *i*'s wage rate in year *t*,

h_{it} = natural logarithm of total number of hours worked by individual *i* in year *t*,

Age_{it} = age of individual *i* in year *t*,

NC_{it} = number of children of individual *i* in year *t*, and

HI_{it} = an dummy variable equal to 1 if individual *i* in year *t* has bad health and 0 otherwise.

The data are from Ziliak, James P. (1997) "Efficient Estimation with Panel Data When Instruments Are Predetermined: An Empirical Comparison of Moment-Condition Estimators," *Journal of Business & Economic Statistics* 15(4): 419-431. Ziliak (p. 423) describes his data as follows:

> ***The data used to estimate the life-cycle labor-supply parameters come from Waves XII-XXI (calendar years 1978-1987) of the PSID. The sample is selected on many dimensions and is similar to other research studying life-cycle models of labor supply. The sample is restricted to continuously married, continuously working, prime-age men aged 22-51 in 1978 from the Survey Research Center random subsample of the PSID. In addition the individual must either be paid an hourly wage rate or must be salaried, and he cannot be a piece-rate worker or self-employed. This selection process resulted in a balanced panel of 532 men over 10 years or 5,320 observations. The real wage rate, wit,. is the hourly wage reported by the panel participant rather than the average wage (annual earnings over annual hours) to minimize division bias (Borjas 1981).***

The data are available in the any of the three files , , and .

1. **Provide scatter plots among the dependent variable (Natural logarithm of hours) against each of the explanatory variables Natural logarithm of real wages, Age, Number of children, and Health. (Label these Figures 1 to 4.)**

2. **Present a table of the summary statistics for all of the variables in this data set (except *ID* and *Year*).**

3. **Provide a histogram of each of the following variables: Natural logarithm of hours, Natural logarithm of real wages, Age, and Number of children. (Label these Figures 5 to 8).**

4. **Estimate Equation (1) using (1) OLS (sometimes called a "pooled model"), (2) a "between" model (where the observations in the regression are the averages over the 10 years of each variable for each individual, (3) a fixed effects model, (4) a MLE random effects model and (5) a GLS random effects model. Present the results of your estimations in a single table and offer an interpretation for each parameter you estimate. Use Table 1 as shown below as a template for the table to present your results.**

	(1) Pooled	(2) Between	(3) Fixed Effects	(4) MLE Random Effects	(5) GLS Random Effects
Natural logarithm of real wages					
	( )	( )	( )	( )	( )
Age					
	( )	( )	( )	( )	( )
Age^2					
	( )	( )	( )	( )	( )
Number of children					
	( )	( )	( )	( )	( )
Health indicator					
	( )	( )	( )	( )	( )
Intercept					
	( )	( )	( )	( )	( )
R^2				—	—
σ_μ	—	—			
σ_ε	—	—			
Sample size					

Table 2.13. Hours and wages: Summary of linear panel model estimations (Dependent variable is the natural logarithm of total hours worked in a year; the observations consist of 532 adult males over the 10 year period 1978-1987).

Exercise 2.3.2.

The Effectiveness of Advertising Bans on Smoking. Anti-smoking activists often push for a total ban on cigarette advertisements. Indeed, one of the basic assumptions of the groups pushing the 1996 proposed settlement with the tobacco companies is that the amount of tobacco consumed is positively affected by the amount of tobacco advertising. There are two mechanisms that might underlie such a relationship. The first mechanism suggests that the advertising increases the amount of cigarettes smoked by *current* smokers. Many economists doubt that the tobacco advertising increases the consumption of current smokers, arguing that the total consumption of cigarettes is unresponsive to advertisement. Instead, they argue that advertising is an effort by cigarette companies to affect the brand of cigarettes that current smokers consume. The second mechanism suggests that advertising is an effort by cigarette companies to induce non-smokers (especially children) to try cigarettes. The main reason that cigarette companies want non-smokers to try smoking, so the argument goes, is that some percentage of non-smokers who try cigarettes will become addicted and will form the future demand for cigarettes.

The effect of a total ban on advertising would be completely different if cigarette companies advertise with the hope of increasing the number of people addicted to cigarettes. In particular, the ban should have a small or negligible effect on current cigarette demand. Instead, the cigarette companies would face a steadily decreasing demand for their product. Such a decrease in demand would reduce future profits for these companies. If future profits fell enough, some of the companies might be forced out of business. Clearly, it is this result that anti-smoking activists have in mind with their proposals to ban cigarette advertisements.

Finally, if advertising only induces current smokers to increase the number of cigarettes they consume, then the total ban on advertising should cause a one-time reduction in cigarette consumption that will reduce the profits of cigarette companies. However, which of these three mechanisms (if any) is correct is an empirical question.

Six European countries adopted a complete ban on cigarette advertising in the period after 1970. It this project we use annual data on smoking consumption in 22 developed countries for the 27 years between 1964 and 1990 to test the effect of a complete smoking ban on cigarette demand (giving us 594 observations). Moreover, since we have no *a priori* reason to choose one model specification over another, we check the stability of the estimated impact of an advertising ban on cigarette demand under several alternative model specifications.

We estimate three types of specifications of the model — the linear model, the log-linear model, and the log-log model. In general whether one uses a variable or the logarithm of the variable is the main difference in these three specifications. The linear model does not transform either the dependent or the independent variables. A variation on the linear models allows the use of the square and product of some of the independent variables in order to take care of any non-linearity in the data. The log-linear model takes the same form as the linear model except that the dependent variable is the logarithm of variable under study. Finally, in the log-log model both the dependent and independent variables are, if possible, in logarithm form.

For example, for this problem the dependent variable in any of these specifications is either the per capita consumption of tobacco or the logarithm of the per capita consumption of tobacco. The dependent variables might include (1) the real price of tobacco in each country for each year, (2) a measure of the per capita income level of the country for each year, (3) the unemployment rate of the country for each year, (4) a measure of the age distribution of the population to measure smoking intensity by age, (5) a trend variable to account for the rising awareness of the health costs of smoking, (6) a dummy variable equal to one for years that a country has a complete ban on cigarette advertising, and (7) a set of 21 dummy variables identifying the country. Let *T it* be the measure of per capita cigarette consumption in country *i* for year *t*; *P it*, the price of tobacco; *I it*, the measure of per capita income level; *U it*, country *i*'s unemployment rate in year *t*; *A it*, country *i*'s age distribution in year *t*; *Year*, a trend variable; *B it*, the dummy variable for the ban; and *C i*, the dummy variable for country *i*.

Examples of the three models are:

1. Linear: $T_{it} = \beta_0 + \beta_1 P_{it} + \beta_2 I_{it} + \beta U_{it} + \beta_4 A_{it} + \beta_5 Year_t + \beta_6 B_{it} + \varepsilon_{it}$

2. Log-Linear: $\ln(T_{it}) = \beta_0 + \beta_1 P_{it} + \beta_2 I_{it} + \beta U_{it} + \beta_4 A_{it} + \beta_5 Year_t + \beta_6 B_{it} + \varepsilon_{it}$

3. Log-Log: $\ln(T_{it}) = \beta_0 + \beta_1 \ln(P_{it}) + \beta_2 \ln(I_{it}) + \beta U_{it} + \beta_4 A_{it} + \beta_5 Year_t + \beta_6 B_{it} + \varepsilon_{it}$

In models (1) and (2) it is possible to include additional explanatory variables that are the square of some of the currently included explanatory variables. In all three models it is possible to include as explanatory variables the product of the ban dummy and any of the currently included explanatory variables. Finally, in equation (2) we cannot take the logarithm of the unemployment rate because the data we have report zero levels of unemployment.

The data you will use in this project are in the *MS Excel* file Smkdata.xls. The variables included in the file are as follows:

Column	Variable	Definition
A	Country	Name of country
B	Country ID	Integer from 1 to 22, each designating a country
C	Year	Year of observation (1964, …, 1990)
D	Tobacco	Total grams of tobacco sold per individual 15 years or older
E	Price	Real price of 20 grams of tobacco in 1990 US cents (= Nominal price per E 20 grams of tobacco divided by the Gross Domestic Price deflator)
F	Consump	Per capita private final consumption expenditures in 1990 US dollars
G	Unemp	Number of unemployed persons per 1000 members of the workforce
H	AgeDist	Age distribution. This variable attempts to measure the differences in intensity of smoking as a function of age. It is equal to the relative consumption rate of tobacco in the UK observed between 1966 and 1981 by age group times the percentage of the population in the country in that age group.
I	Ban	Dummy variable equal to 1 if the country has a complete ban on tobacco advertising. The six countries in the sample with a complete ban and the first year of the ban are: Iceland (1972), Norway (1976), Finland (1979), Portugal (1984), Italy (1984), and Canada (1989).
J	BanTime	The number of years since the ban was put in place (if ban went into effect in 1972, then years 1964-1972 are equal to 0, year 1973 equals 1, year 1974 equals 2, etc.)

Table 2.14. Definition of the cigarette consumption data set.

(a) How do these variables match the ones suggested in the discussion of equations (1), (2), and (3)?

(b) Estimate the fixed effects models of the following versions of equations (1), (2), and (3):

1. Equations (1), (2), and (3) as specified above.

2. Equations (1) and (2) with squared terms for the price, income, unemployment rate, and the age distribution included. This regression is designed to test for non-linearity.

3. Equations (1) and (2) with the squared terms mentioned in 2 that are statistically significant plus the following new variables: Ban*Time, Ban*Price, and Ban*Consump. (You must create these variables) This regression allows for an effect of the Ban on the slopes of the other explanatory variables.

4. Equation (3) with the following new variables: Ban*Log(Time), Ban*Log(Price), and Ban*Log(Consump).

5. Equations (1), (2), and (3) as estimated in 3 and 4 with a variable that counts the number of years that a total ban has been in effect (BanTime) and its square (BanTime^2). This regression allows for a changing impact of a ban the longer it is in effect.

Report the results of your regressions in a table that allows you to comment on the stability of your estimation results over specifications.

(c) Do these results support any of the theories suggested above?

(d) What, if any, policy conclusions would you make given your estimations?

(e) Assume for the moment that you "believe" your results you got in (5). Sketch out a strategy you would follow to forecast the impact of a ban in a country that does not currently have a ban.

Note: The data in this problem are from Stewart, Michael J. (1993) "The Effect on Tobacco Consumption of Advertising Bans in OECD Countries," *International Journal of Advertising* 12(2): 155-180. The data set can be downloaded from the author's website.

Bibliography

Cameron, A. Colin and Pravin K. Trivedi (2005). *Microeconometrics: Methods and Applications* (New York: Cambridge University Press).

Greene, W. H. (2003). *Econometric Analysis*, 5th edition (Upper Saddle River, NJ: Prentice-Hall).

Hsiao, Cheng (2003). *Analysis of Panel Data*, 2nd Edition (New York: Cambridge University Press).

StataCorp (2003). *Stata Statistical Software: Release8.0* (College Station, TX: Stata Corporation).

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data* (Cambridge, MA: MIT Press).

2.4. Sample selectivity bias[*]

Sample Selection Bias

Introduction

These notes discuss how to handle one of the more common problems that arise in economic analyses—sample selection bias. Essentially, sample selection bias can arise whenever some potential observations cannot be observed. For instance, the students enrolled in an intermediate microeconomics course are not a random sample of all undergraduates. Students self-select when they enroll in any class or choose a major. While we do not know all of the reasons for this self-selection, we suspect that students choosing to take advanced economics courses have more quantitative skills than students choosing courses in the humanities. Since we do not observe the grades that students who did not enroll in the intermediate microeconomics class would have made had they enrolled, we can never observe the grades that they would have made. Under certain circumstances the omission of potential members of a sample will cause ordinary least squares (OLS) to give biased estimates of the parameters of a model.

In the 1970s James Heckman developed techniques that will correct the bias introduced by sample selection bias. Since then, most econometric computer programs include a command that automatically used Heckman's method. However, blind use of these commands can lead to errors that would be avoided by a better understanding of his correction technique. This module is intended to provide this understanding.

In the first section I discuss the sources of sample selection bias by examining the basic economic model used to understand the problem. In the second section I present the estimation strategy first developed by Heckman. In the third section I discuss how to estimate the Heckman model in *Stata*. In the final section I examine an extended example of the technique. An exercise is included at the end of the discussion.

The model

Assume that there is an unobserved latent variable, $y_i{}^*$, and an unobserved latent index, $d_i{}^*$, such that:

(2.51)

$$y_i^* = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i \text{where} i = 1, \ldots, N;$$

(2.52)

$$d_i^* = \mathbf{z}'_i \boldsymbol{\gamma} + \nu_i \text{where} i = 1, \ldots, N;$$

(2.53)

$$d_i = \begin{cases} 1 \text{ if } d_i^* > 0 \\ 0 \text{ if } d_i^* \leq 0 \end{cases} ; \text{ and}$$

(2.54) $y_i = y_i{}^* d_i$ **.**

The matrix notation above means (1) that

1. $$\mathbf{x}'_i \boldsymbol{\beta} = \begin{bmatrix} 1 \\ x_{1i} \\ \vdots \\ x_{Ki} \end{bmatrix}' \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{bmatrix} = [1 \quad x_{1i} \quad \cdots \quad x_{Ki}] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{bmatrix} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki}$$

2. $$\mathbf{z}'_i \boldsymbol{\gamma} = [1 \quad z_{1i} \quad \cdots \quad z_{Li}] \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_L \end{bmatrix} = \gamma_0 + \sum_{j=1}^{L} \gamma_j z_{ji}.$$

Substituting (1), (2) and (3) into (4) gives:

(2.55)

$$y_i = \begin{cases} \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i & \text{if } \mathbf{z}'_i \boldsymbol{\gamma} + \nu_i > 0 \\ 0 & \text{if } \mathbf{z}'_i \boldsymbol{\gamma} + \nu_i \leq 0 \end{cases}.$$

Note that *N* is the total sample size and *n* is the number of observations for which $d_i = 1$.

Since $y_i{}^*$ is not observed for (*N* − *n*), the question becomes why are these observations missing. A concrete example of such a model is a model of female wage determination. Equation (1) would model the wage rate earned by women in the labor force and Equation (2) would model the decision by a female to enter the labor force. In this case, y_i, the wage rate woman *i* receives, is a function of the variables in $\mathbf{x}_i$; however, women not in the labor force are not included in the sample. If these missing observations are drawn randomly from the population, there is no need for concern. Selectivity bias arises if the (*N* − *n*) omitted observations have unobserved characteristics that affect the likelihood that $d_i = 1$ and are correlated with the wage the woman would receive had she entered the labor force. For instance, a mentally unstable female is likely to earn relatively low wages and might be more unlikely to enter the labor force. In this case, the error terms, ε_i and v_i would be independent and identically distributed *N*(0,∑), where

(2.56)

$$\sum = \begin{bmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon\upsilon} \\ \sigma_{\upsilon\varepsilon} & \sigma_\upsilon^2 \end{bmatrix}$$

and (ε_i, ν_i) are independent of z_i. The selectivity bias arises because $\sigma_{\varepsilon v} \neq 0$. In effect the residual ε_i includes the same unobserved characteristics as does the residual v_i causing the two error terms to be correlated. OLS estimation of equation (1) would have a missing variable—the bias created by the missing observations (due to wage data not being available for women not in the work force). As in other cases of omitted variables, the estimates of the parameters of the model, $\hat{\boldsymbol{\beta}}$, would be biased. Heckman (1979) notes in his seminal article on selectivity bias:

> *One can also show that the least squares estimator of the population variance is downward biased. Second, a symptom of selection bias is that variables that do not belong in the true structural equation (variables in not in may appear to be statistically significant determinants of when regressions are fit on selected samples. Third, the model just outlined contains a variety of previous models as special cases. ...For a more complete development of the relationship between the model*

> ***developed here and previous models for limited dependent variables, censored samples and truncated samples, see Heckman (1976). Fourth, multivariate extensions of the preceding analysis, while mathematically straightforward, are of consider-able substantive interest. One example is offered. Consider migrants choosing among K possible regions of residence. If the self selection rule is to choose to migrate to that region with the highest income, both the self selection rule and the subsample regression functions can be simply characterized by a direct extension of the previous analysis. (Notation has been altered to match the notation used in this module, see Heckman, 1979: 155)***

Estimation Strategy

Heckman (1979) suggests a two-step estimation strategy. In the first step a probit estimate of equation (2) is used to construct a variable that measures the bias. This variable is known as the “inverse Mills ratio.” Heckman and others demonstrate that

(2.57)

$$E[\varepsilon_i | \mathbf{z}_i, d_i = 1] = \frac{\sigma_{\varepsilon\nu}}{\sigma_\nu^2} \left\{ \frac{\phi(\mathbf{z}_i{}' \boldsymbol{\gamma})}{\Phi(\mathbf{z}_i{}' \boldsymbol{\gamma})} \right\},$$

where $\phi(\mathbf{z}_i{}'\boldsymbol{\gamma})$ and $\Phi(\mathbf{z}_i{}'\boldsymbol{\gamma})$ are the probability density function and the cumulative distribution functions, respectively, evaluated at $\mathbf{z}_i{}'\boldsymbol{\gamma}$. [24] The ratio in the brackets in equation (7) is known as the *inverse Mills ratio*. We will use an estimate of the inverse Mills ratio in the estimation of equation (5) to measure the sample selectivity bias.

The Heckman two-step estimator is relatively easy to implement. In the first step you use a maximum likelihood probit regression on the whole sample to calculate $\hat{\boldsymbol{\gamma}}$ from equation (2). You then use $\hat{\boldsymbol{\gamma}}$ to estimate the inverse Mills ratio:

(2.58)

$$\hat{\lambda}_i = \frac{\phi\left(\mathbf{z}_i{}' \hat{\boldsymbol{\nu}}\right)}{\Phi\left(\mathbf{z}_i{}' \hat{\boldsymbol{\nu}}\right)}.$$

In the second step, we estimate:

(2.59)

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \mu \hat{\lambda} + \eta_i$$

using OLS and where $E\left(\hat{\mu}\right) = \frac{\sigma_{\varepsilon\nu}}{\sigma_{\nu}^{2}}.$ Thus, a t-ratio test of the null hypothesis $H_0 : \mu = 0$ is equivalent to testing the null hypothesis $H_0 : \sigma_{\varepsilon\nu} = 0$ and is a test of existence of the sample selectivity bias.

An alternative approach to the sample selectivity problem is to use a maximum likelihood estimator. Heckman (1974) originally suggested estimating the parameters of the model by maximizing the average log likelihood function:

(2.60)

$$L = \frac{1}{N}\sum_{i=1}^{N} \left\{ d_i \ln\left[\int_{-(\mathbf{z}'\boldsymbol{\gamma})}^{\infty} \phi_{\varepsilon\nu}(y_i - \mathbf{x}_i{}' \boldsymbol{\beta}) d\nu \right] + (1 - d_i)\left[\ln \int_{-(\mathbf{z}'\boldsymbol{\gamma})}^{\infty} \int_{-\infty}^{\infty} \phi_{\varepsilon\nu}(\varepsilon, \nu) d\varepsilon d\nu \right] \right\},$$

where $\varphi_{\varepsilon\nu}$ is the probability density function for the bivariate normal distribution. Fortunately, Stata offers a single command for calculating either the two-step or the maximum likelihood estimators.

Estimation in *Stata*

Estimation of the two versions of the Heckman sample selectivity bias models is straightforward in *Stata*. The command is:

```
.heckman depvar [varlist], select(varlist_s) [twostep]
```

or

```
.heckman depvar [varlist], select(depvar_s = varlist_s) [twostep]
```

The syntax for maximum-likelihood estimates is:

```
.heckman depvar [varlist] [weight] [if exp] [in range], select([depvar_s =] varlist_s [, offset(varname) noconstant]) [ robust
cluster(varname) score(newvarlist|stub*) nshazard(newvarname) mills(newvarname) offset(varname) noconstant
constraints(numlist) first noskip level(#) iterate(0) nolog maximize_options ]
```

The predict command has these options, among others:

xb, the default, calculates the linear predictions from the underlying regression equation.

ycond calculates the expected value of the dependent variable conditional on the dependent variable being observed/selected; E(y | y observed).

yexpected calculates the expected value of the dependent variable (y*), where that value is taken to be 0 when it is expected to be unobserved; y* = P(y observed) * E(y | y observed). The assumption of 0 is valid for many cases where nonselection implies non-participation (e.g., unobserved wage levels, insurance claims from those who are uninsured, etc.) but may be inappropriate for some problems (e.g., unobserved disease incidence).

Examples of these two commands are:

```
. heckman wage educ age, select(married children educ age)
```

```
. predict yhat
```

These two command would use the maximum likelihood estimate of the equations (1) wage as a function of education and age using a selection equation that used marital status, number of children, education level, and age to explain which individuals are participating in the labor force. The help file in *Stata* provides additional information on the structure of the Heckman command and is well worth printing out if you are dealing with a sample selectivity bias problem.

Example 2.4. Example from *Stata*

We will illustrate various issues of selection bias using the data set available from the *Stata* site. Retrieve the data set by entering:

```
. use http://www.stata-press.com/data/imeus/womenwk, clear
```

This data set has 2,000 observations of 15 variables. We can use the describe command (.describe) to get a brief description of the data set:

obs: 2,000				
vars: 15	**9 Nov 2004 20:23**			
size: 142,000	**(86.5% of memory free)**			
Variable Name	**Storage Type**	**Display Format**	**Value Label**	**Variable Label**
c1	**double**	**%10.0g**		
c2	**double**	**%10.0g**		
u	**double**	**%10.0g**		
v	**(7,2)**	**%10.0g**		
country	**float**	**%9.0g**		
age	**int**	**%8.0g**		
education	**int**	**%8.0g**		
married	**byte**	**%8.0g**		
children	**int**	**%8.0g**		
select	**float**	**%9.0g**		
wageful	**float**	**%9.0g**		
wage	**float**	**%9.0g**		
lw	**float**	**%9.0g**		

work	float	%9.0g		
lwf	float	%9.0g		

Table 2.15. Description of variables included in the data set from http://www.stata-press.com/data/imeus/womenwk.

We are interested in only a subset of these data. Table 2 reports the definitions of variables that are relevant for our analysis. We can get further insight into the data set using the summarize command. Table 3 reports the summary statistics for the data set.

Variable name	Definition
country	County of residence (categorical variable equal to 0, 1, ..., 9)
age	Age of the woman
education	Number of years of education of the woman
married	Dummy variable equal to 1 if the woman is married and 0 otherwise
children	Number of children that the woman has in their household
wage	Hourly wage rate of the woman
lw	Natural logarithm of hourly wage rate
work	Dummy variable equal to 1 if the individual is in the workforce and 0 otherwise

Table 2.16. Definition of the relevant variables in the data set.

Variable	Obs	Mean	Std. Dev	Min	Max
Age	2000	36.208	8.28656	20	59

education	2000	13.084	3.045912	10	20
married	2000	.6705	.4701492	0	1
children	2000	1.6445	1.398963	0	5
wage	1343	23.69217	6.305374	5.88497	45.80979
lw	1343	3.126703	.2865111	1.772402	3.824498
work	2000	.6715	.4697852	0	1

Table 2.17. Summary statistics of the relevant variables in the data set (using the command: .summarize age education married children wage lw work).

We are interested in modeling two things: (1) the decision of the woman to enter the labor force and (2) determinants of the female wage rate. It might be reasonable to assume that the decision to enter the labor force by a woman is a function of age, marital status, the number of children, and her level of education. Also, the wage rate a woman earns should be a function of her age and education.

The decision to enter the labor force

We can use a probit regression to model the decision of a woman to enter the labor force. The results of this estimation are reported in Table 4. However, we can use the predict command to produce some results that we can use to be sure that we understand what the regression results mean. In particular, type in the following two commands:

.predict zbhat, xb

.predict phat, p

These two commands will predict (1) the linear prediction (zbhat) and (2) the predicted probability that the woman will be in the workforce (phat). Table 5 reports the values of these two variables for observations 1 through 10.

```
. probit work age education married children
```

Iteration 0: log likelihood = -1266.2225						
Iteration 4: log likelihood = -1027.0616						
Probit estimates Number of obs = 2000						
LR chi2(4) = 478.32						
Prob > chi2 = 0.0000						
Log likelihood = -1027.0616 Pseudo R2 = 0.1889						
work	Coef.	Std. Err.	z	P>z	[95% Conf. Interval]	
age	.0347211	.0042293	8.21	0.000	.0264318	.0430105
education	.0583645	.0109742	5.32	0.000	.0368555	.0798735
married	.4308575	.074208	5.81	0.000	.2854125	.5763025
children	.4473249	.0287417	15.56	0.000	.3909922	.5036576
_cons	-2.467365	.1925635	-12.81	0.000	-2.844782	-2.089948

Table 2.18. Probit estimation of the decision to enter the labor force.

Observation	zbhat	phat
1	-0.68900	0.24541
2	-0.20290	0.41961

3	-0.48067	0.31538
4	-0.16818	0.43322
5	0.34859	0.63630
6	0.58758	0.72159
7	0.97357	0.83486
8	0.45978	0.67716
9	0.01799	0.50718
10	0.32628	0.62790

Table 2.19. Predicted values of zbhat and phat for observations 1 through 10.

The interpretation of the numbers in Table 5 is straightforward. Consider individual 1. The z-value predicted for this individual is -0.68. Using the standard normal tables reported in Table 11 it is easy to see:

(2.61) $\Phi(z \leq -0.69)$ = Pr(Individual 1 is in the labor force)

(2.62)

$$\begin{aligned} \Phi(z\leq -0.69) &= 0.5-\Phi(0\leq z\leq 0.69) \\ &\approx 0.5-0.2549 \\ &\approx 0.2451. \end{aligned}$$

The difference between this number and the value reported for *phat* in Table 5 is due to rounding error.

A little later we will want to calculate the inverse Mills ratio. As noted in (8), the formula for the inverse Mills ratio is:

(2.63)

$$\hat{\lambda}_i = \frac{\phi\left(\mathbf{z}_i{}'\,\hat{\boldsymbol{\nu}}\right)}{\Phi\left(\mathbf{z}_i{}'\,\hat{\boldsymbol{\nu}}\right)}.$$

The variable phat is equal to $\Phi\left(\mathbf{z}_i{}'\,\hat{\boldsymbol{\nu}}\right)$. Stata offers an easy way to calculate $\phi\left(\mathbf{z}_i{}'\,\hat{\boldsymbol{\nu}}\right)$ with the function “normden(zbhat)” as follows:

```
.generate imratio = normden(zbhat)/phat
```

Table 6 repeats Table 5 with the estimate of the inverse Mills ratio for the first 10 observations.

Observation	zbhat	phat	Inverse Mills Ratio
1	-0.6889973	0.2454125	1.2821240
2	-0.2029016	0.4196060	0.9313837
3	-0.4806706	0.3153753	1.1269680
4	-0.1681804	0.4332207	0.9079438
5	0.3485867	0.6363002	0.5900134
6	0.5875849	0.7215945	0.4652062
7	0.9735670	0.8348642	0.2974918
8	0.4597758	0.6771615	0.5300468
9	0.0179909	0.5071769	0.7864666
10	0.3262833	0.6278950	0.6024283

Table 2.20. Calculation of the inverse Mills ratio for the first 10 observations.

The two Heckman estimates

One of the great advantages of using an econometrics program like *Stata* is that the authors quite often have created a command that does all of the work for the user. In our case, the commands we need to run to generate the maximum likelihood estimate of the Heckman model are:

```
. global wage_eqn wage educ age
```

```
. global seleqn married children age education
```

```
. heckman $wage_eqn, select($seleqn)
```

Notice that we have used the global command to create a shortcut for referring to each of the two equations in the estimation. The command for the Heckman two-stage estimate is:

```
.heckman $wage_eqn, select($seleqn) twostage
```

```
.predict mymills, mills
```

(1) Explanatory variable	(2) Maximum likelihood estimate	(3) Heckman two-step	(4) Probit estimate of the selection equation
Wage Equation			
Education	0.9899537	0.9825259	—
	(18.59)	(18.23)	
Age	0.2131294	0.2118695	—
	(10.34)	(9.61)	
Intercept	0.4857752	0.7340391	—

	(0.45)	**(0.59)**	
Selection equation			
Married	**0.4451721**	**0.4308575**	**0.4308575**
	(6.61)	**(5.81)**	**(5.81)**
Children	**0.4387068**	**0.4473249**	**0.4473249**
	(15.79)	**(15.56)**	**(15.56)**
Age	**0.0365098**	**0.0347211**	**0.0347211**
	(8.79)	**(8.21)**	**(8.21)**
Education	**0.0557318**	**0.0583645**	**0.0583645**
	(5.19)	**(5.32)**	**(5.32)**
Intercept	**-2.491015**	**-2.467365**	**-2.467365**
	(-13.16)	**(-12.81)**	**(-12.81)**
σ	**0.7035061**	**0.67284**	**—**
λ	**6.004797**	**5.9473529**	**—**
(Mills)λ	**4.224412**	**4.001615**	**—**
		(6.60)	
Observations	**2000**	**2000**	**2000**
Number of women not working	**657**	**657**	**657**
Number of women working	**1343**	**1343**	**1343**
Log likelihood	**-5178.304**	**—**	**-1027.0616**

Wald $\chi^2(2)$	**508.44**	**—**	**—**
Probability > χ^2	**0.0000**	**—**	**—**
Wald $\chi^2(4)$	**—**	**551.37**	**—**
Probability > χ^2	**—**	**0.0000**	**—**
***LR test of independent equations* (ρ = 0)**			
$\chi^2(1)$	**61.20**	**—**	**478.32**
Probability > χ^2	**0.0000**	**—**	**0.0000**

Table 2.21. Comparison of Heckman Maximum-Likelihood and the Heckman Two-Step Estimates with the Probit Estimates of the Selection Equation.

The second command reports the estimates of the inverse Mills ratio; we have retrieved these values in order to check our earlier calculations. Table 7 reports the results of these two estimations. Column 2 reports the maximum-likelihood estimates; Column 3 reports the Heckman two-step estimates; and Column 3 reports the probit estimate of selection equation as reported in Table 4. The estimates for the two methods are very similar. Of course, the probit estimates in Column 4 exactly match the results reported for the selection equation in Column 3. As a final check, Table 8 reports the values of the inverse Mills ratio reported in Table 6 with the values of the inverse Mills ratio calculated in the Heckman two-step method. The two estimates are identical except for some rounding errors.

Observation	**As calculated from probit estimate**	**As reported by the Heckman two-step**
1	**1.2821240**	**1.2821240**
2	**0.9313837**	**0.9313837**

3	1.1269680	1.1269680
4	0.9079438	0.9079438
5	0.5900134	0.5900134
6	0.4652062	0.4652061
7	0.2974918	0.2974918
8	0.5300468	0.5300469
9	0.7864666	0.7864666
10	0.6024283	0.6024283

Table 2.22. Inverse Mills Ratio Comparison.

Exercise

Exercise 2.4.1. The supply of married women in the workforce.

We are interested in understanding the decision of married Portugese women to enter the labor force. We have available data from Portugal. The data set is a sample from Portuguese Employment Survey, from the interview year 1991, and has been provided by the Portuguese National Institute of Statistics (INE). The data are in the Excel file Martins. This file is organized in the following way. There are seven columns, corresponding to seven variables, and 2,339 observations.

a) Estimate the following equation using OLS: $Wages = f\left(age, age^2, education\right)$ **using the observations for women actually working.**

b) What is the potential source of selection bias?

c) Estimate a wage equation for the Portuguese data three ways: (1) using OLS, (2) using the Heckman two-step method, and (3) using the ML method. Report all three estimates in a single table. For consistency, we will assume that the appropriate explanatory variables for wages are (1) age, (2) the square of age, and (3) the years of education. Further, assume that women do not enter the labor force because (1) presence of children under the age of 3, (2) presence of children between 3 and 18, (3) husband's wage level, (4) the level of education of the woman, and (5) the age of the woman.

Appendix A.

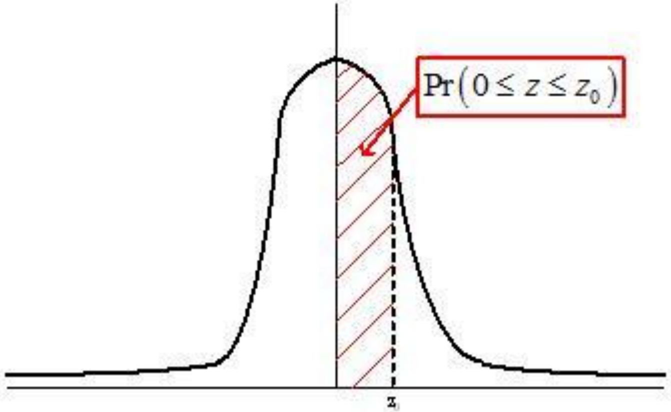
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015

1.3	**0.4032**	**0.4049**	**0.4066**	**0.4082**	**0.4099**	**0.4115**	**0.4131**	**0.4147**	**0.4162**	**0.4177**
1.4	**0.4192**	**0.4207**	**0.4222**	**0.4236**	**0.4251**	**0.4265**	**0.4279**	**0.4292**	**0.4306**	**0.4319**
1.5	**0.4332**	**0.4345**	**0.4357**	**0.4370**	**0.4382**	**0.4394**	**0.4406**	**0.4418**	**0.4429**	**0.4441**
1.6	**0.4452**	**0.4463**	**0.4474**	**0.4484**	**0.4495**	**0.4505**	**0.4515**	**0.4525**	**0.4535**	**0.4545**
1.7	**0.4554**	**0.4564**	**0.4573**	**0.4582**	**0.4591**	**0.4599**	**0.4608**	**0.4616**	**0.4625**	**0.4633**
1.8	**0.4641**	**0.4649**	**0.4656**	**0.4664**	**0.4671**	**0.4678**	**0.4686**	**0.4693**	**0.4699**	**0.4706**
1.9	**0.4713**	**0.4719**	**0.4726**	**0.4732**	**0.4738**	**0.4744**	**0.4750**	**0.4756**	**0.4761**	**0.4767**
2.0	**0.4772**	**0.4778**	**0.4783**	**0.4788**	**0.4793**	**0.4798**	**0.4803**	**0.4808**	**0.4812**	**0.4817**
2.1	**0.4821**	**0.4826**	**0.4830**	**0.4834**	**0.4838**	**0.4842**	**0.4846**	**0.4850**	**0.4854**	**0.4857**
2.2	**0.4861**	**0.4864**	**0.4868**	**0.4871**	**0.4875**	**0.4878**	**0.4881**	**0.4884**	**0.4887**	**0.4890**
2.3	**0.4893**	**0.4896**	**0.4898**	**0.4901**	**0.4904**	**0.4906**	**0.4909**	**0.4911**	**0.4913**	**0.4916**
2.4	**0.4918**	**0.4920**	**0.4922**	**0.4925**	**0.4927**	**0.4929**	**0.4931**	**0.4932**	**0.4934**	**0.4936**
2.5	**0.4938**	**0.4940**	**0.4941**	**0.4943**	**0.4945**	**0.4946**	**0.4948**	**0.4949**	**0.4951**	**0.4952**
2.6	**0.4953**	**0.4955**	**0.4956**	**0.4957**	**0.4959**	**0.4960**	**0.4961**	**0.4962**	**0.4963**	**0.4964**
2.7	**0.4965**	**0.4966**	**0.4967**	**0.4968**	**0.4969**	**0.4970**	**0.4971**	**0.4972**	**0.4973**	**0.4974**
2.8	**0.4974**	**0.4975**	**0.4976**	**0.4977**	**0.4977**	**0.4978**	**0.4979**	**0.4979**	**0.4980**	**0.4981**
2.9	**0.4981**	**0.4982**	**0.4982**	**0.4983**	**0.4984**	**0.4984**	**0.4985**	**0.4985**	**0.4986**	**0.4986**
3.0	**0.4987**	**0.4987**	**0.4987**	**0.4988**	**0.4988**	**0.4989**	**0.4989**	**0.4989**	**0.4990**	**0.4990**

Table 2.23. Normal Distribution.

z~N(0, 1). $$\Phi(z_0) = \int_{-\infty}^{z_0} \phi(z)dz = 0.5 + \Pr(0 \le z \le z_0)$$

Figure 2.27. The Normal Distribution



References

Bourguignon, François, Martin Fournier, and Marc Gurgand (2007). Selection Bias Corrections Based on the Multinomial Logit Model: Monte Carlo Comparisons. *Journal of Economic Surveys* 21(1): 174-205.

Chiburis, Richard and Michael Lokshin (2007). Maximum Likelihood and Two-Step Estimation of an Ordered-Probit Selection Model. *The Stata Journal* 7(2): 167-182.

Dahl, G. B. (2002). Mobility and the Returns to Education: Testing a Roy Model with Multiple Markets. *Econometrica* 70(6): 2367-2420.

Dubin, Jeffrey A. and Douglas Rivers (1989). Selection Bias in Linear Regression, Logit and Probit Models. *Sociological Methods and Research* 18(2 & 3): 360-390.

Heckman, James (1974). Shadow Prices, Market Wages and Labor Supply. *Econometrica* 42(4):679-694.

Heckman, James (1976) “The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models,” *The Annals of Economic and Social Measurement* 5: 475-492.

Heckman, James (1979). Sample Selection Bias as a Specification Error. *Econometrica* 47(1): 153-161.

Jimenez, Emanuel and Bernardo Kugler (1987). The Earnings Impact of Training Duration in a Developing Country: An Ordered Probit Model of Colombia's *Servicio Nacional de Aprendizaje* (SENA). *Journal of Human Resources* 22(2): 230-233.

Lee, Lung-Fei (1983). Generalized Econometric Models with Selectivity. *Econometrica* 51(2): 507-512.

McFadden, Daniel L. (1973). Conditional Logit Analysis of Qualitative Choice Behavior. In P. Zarembka *Frontiers in Econometrics* (New York: Academic Press).

Newey, W. K. and Daniel L. McFadden (1994). Large Sample Estimation and Hypothesis Testing. In R. F. Engle and D. L. McFadden (eds.) *Handbook of Econometrics* (Amsterdam: North Holland).

Schmertmann, Carl P. (1994). Selectivity Bias Correction Methods in Polychotomous Sample Selection Models. *Journal of Econometrics* 60(1): 101-132.

Vella, Francis (1998). Estimating Models with Sample Selection Bias: A Survey. *The Journal of Human Resources* 33(1):127-169.

2.5. Endogenous explanatory variables[*]

Endogenous Explanatory Variables

Introduction

One of the most common problems complicating the research of an economist is created by the inclusion of endogenous variables as an explanatory variable. The variable on the left-hand-side of a regression is an endogenous variable; its level is determined by the levels of the explanatory variables—that is, the variables on the right-hand-side of the equation. In OLS we assume that the explanatory variables are independent of the error term. However, if the level of one of these explanatory variables is determined by the levels of the other variables in the model, that explanatory variable actually is an endogenous variable. In a nutshell the problem with having endogenous explanatory variables is that these endogenous variables cause the error term in the model to be correlated with the explanatory variables thus causing the OLS estimator to be biased. This problem is also known as *simultaneous equation bias* and it is a problem that is subtly different from sample selection bias. See "What is the difference between 'endogeneity' and 'sample selection bias"'?" for an excellent discussion of the difference between these two econometric problems.

In this module we explore both the statistical and algebraic issues raised by the inclusion of endogenous explanatory variables in a model. This introduction is too sketchy to give you a thorough understanding of the many problems raised by simultaneous equation bias. Hopefully, by the time you finish the module along with the problem set, you will have an least an intuitive understanding of the problem and will be able to recognize it when you come across the problem in your own research. If you think the model you are estimating may have simultaneous equation bias, you should seek the advice of an econometrician.

The Statistical Problem

Imagine we know with certainty that the following model fully describes the true state of the supply and demand for wheat. First, the demand for wheat in any year, q_t, is a function of the price of wheat, p_t^w, the income of the average individual, I_t, and the price of corn, p_t^c. Second, in any year the price of wheat is a function of the amount of wheat brought to market, q_t, and a weather index, W_t, that is positively related to the amount of wheat that is harvested. Third, the error terms in the supply and demand functions are due purely to measurement errors—that is, there are no omitted variables in the model. Thus, we have the following two equation model:

(2.64)

Demand:

$$q_t = \alpha_0 + \alpha_1 p_t^w + \alpha_2 I_t + \alpha_3 p_t^c + \varepsilon_t$$

and

Supply:

$p_t^w = \beta_0 + \beta_1 q_t + \beta_2 W_t + \eta_t.$

We assume that the error terms each are normally distributed with a mean of zero and a constant variance. Moreover, we assume that the two error terms are independent of each other—that is, we are assuming that:

(2.65)

$$\varepsilon_t \sim N\left(0, \sigma_\varepsilon^2\right),$$
$$\eta_t \sim N\left(0, \sigma_\eta^2\right), \text{ and}$$
$$E(\varepsilon_t \eta_t) = 0.$$

Finally, we assume that income, the price of corn, and the weather index are non-stochastic variables—i.e., these variables are independent of the two error terms. Clearly, the price of wheat and the quantity of wheat are stochastic variables.[25]

What we have here is an ideal model in the sense that we know and can measure all of the variables in the model. The model as written has two *endogenous* variables—q_t and p_t^w—and three exogenous variables— I_t, p_t^c, and W_t. Equations (1) and (2) are known as *structural equations*. What makes this model useful for our purposes is that there is an endogenous explanatory variable in each of the two structural equations.

What we ultimately want to know is if we can use ordinary least squares (OLS) to obtain unbiased estimates of the parameters in Equations (1) and (2). One of the assumptions of OLS is that each of the explanatory variables are independent of the error term, ε_t; if this assumption is violated, OLS will produce biased estimates of the slope parameters. Thus, what we need to do is see if the error term in each equation is independent of the endogenous variable on the right-hand-side of that equation. That is, we want to see if $E\left(\varepsilon_t p_t^w\right) = 0$ and $E(\eta_t q_t) = 0.$

It is convenient in answering our question to use the two structural equations to find what are known as the *reduced form equations*—that is, one equation for each endogenous variable in which the endogenous variable is written as a function solely of exogenous variables and error terms. We can find the reduce form equations by solving the structural equations simultaneously for the endogenous variables. Substituting (2) into (1), we get:

$$q_t = \alpha_0 + \alpha_1 (\beta_0 + \beta_1 q_t + \beta_2 W_t + \eta_t) + \alpha_2 I_t + \alpha_3 p_t^c + \varepsilon_t$$

$q_t = \alpha_0 + \alpha_1 \beta_0 + \alpha_1 \beta_1 q_t + \alpha_1 \beta_2 W_t + \alpha_1 \eta_t + \alpha_2 I_t + \alpha_3 p_t^c + \varepsilon_t$

$$q_t - \alpha_1 \beta_1 q_t = (\alpha_0 + \alpha_1 \beta_0) + \alpha_1 \beta_2 W_t + \alpha_2 I_t + \alpha_3 p_t^c + (\varepsilon_t + \alpha_1 \eta_t)$$

or

(2.66)

$$q_t = \frac{\alpha_0 + \alpha_1 \beta_0}{1 - \alpha_1 \beta_1} + \frac{\alpha_1 \beta_2}{1 - \alpha_1 \beta_1} W_t + \frac{\alpha_2}{1 - \alpha_1 \beta_1} I_t + \frac{\alpha_3}{1 - \alpha_1 \beta_1} p_t^c + \frac{\varepsilon_t + \alpha_1 \eta_t}{1 - \alpha_1 \beta_1}.$$

Substituting (1) into (2) yields:

$$p_t^w = \beta_0 + \beta_1 (\alpha_0 + \alpha_1 p_t^w + \alpha_2 I_t + \alpha_3 p_t^c + \varepsilon_t) + \beta_2 W_t + \eta_t$$

$p_t^w = \beta_0 + \beta_1 \alpha_0 + \alpha_1 \beta_1 p_t^w + \alpha_2 \beta_1 I_t + \alpha_3 \beta_1 p_t^c + \beta_1 \varepsilon_t + \beta_2 W_t + \eta_t$

or

(2.67)

$$p_t^w = \frac{\beta_0 + \beta_1 \alpha_0}{1 - \alpha_1 \beta_1} + \frac{\alpha_2 \beta_1}{1 - \alpha_1 \beta_1} I_t + \frac{\alpha_3 \beta_1}{1 - \alpha_1 \beta_1} p_t^c + \frac{\beta_2}{1 - \alpha_1 \beta_1} W_t + \frac{\beta_1 \varepsilon_t + \eta_t}{1 - \alpha_1 \beta_1}.$$

Equations (4) and (5) are the reduced form equations for this model. We can use them to calculate $E(\varepsilon_t p_t^w) = 0$ and $E(\eta_t q_t) = 0.$ In particular,

$$E(\varepsilon_t p_t^w) = E\left[\varepsilon_t \left(\frac{\beta_0 + \beta_1 \alpha_0}{1 - \alpha_1 \beta_1} + \frac{\alpha_2 \beta_1}{1 - \alpha_1 \beta_1} I_t + \frac{\alpha_3 \beta_1}{1 - \alpha_1 \beta_1} p_t^c + \frac{\beta_2}{1 - \alpha_1 \beta_1} W_t + \frac{\beta_1 \varepsilon_t + \eta_t}{1 - \alpha_1 \beta_1}\right)\right]$$

$$E(\varepsilon_t p_t^w) = E\left[\varepsilon_t \left(\frac{\beta_0 + \beta_1 \alpha_0}{1 - \alpha_1 \beta_1} + \frac{\alpha_2 \beta_1}{1 - \alpha_1 \beta_1} I_t + \frac{\alpha_3 \beta_1}{1 - \alpha_1 \beta_1} p_t^c + \frac{\beta_2}{1 - \alpha_1 \beta_1} W_t\right) + \varepsilon_t \left(\frac{\beta_1 \varepsilon_t + \eta_t}{1 - \alpha_1 \beta_1}\right)\right]$$

or

(2.68)

$$E(\varepsilon_t p_t^w)=E\left[\varepsilon_t\left(\frac{\beta_0+\beta_1\alpha_0}{1-\alpha_1\beta_1}+\frac{\alpha_2\beta_1}{1-\alpha_1\beta_1}I_t+\frac{\alpha_3\beta_1}{1-\alpha_1\beta_1}p_t^c+\frac{\beta_2}{1-\alpha_1\beta_1}W_t\right)\right]+E\left(\frac{\beta_1\varepsilon_t^2+\eta_t\varepsilon_t}{1-\alpha_1\beta_1}\right).$$

Factoring out the non-stochastic terms from the expected value operators gives:

$$E(\varepsilon_t p_t^w)=\left(\frac{\beta_0+\beta_1\alpha_0}{1-\alpha_1\beta_1}+\frac{\alpha_2\beta_1}{1-\alpha_1\beta_1}I_t+\frac{\alpha_3\beta_1}{1-\alpha_1\beta_1}p_t^c+\frac{\beta_2}{1-\alpha_1\beta_1}W_t\right)E[\varepsilon_t]+\frac{\beta_1 E\left(\varepsilon_t^2\right)}{1-\alpha_1\beta_1}+\frac{E(\eta_t\varepsilon_t)}{1-\alpha_1\beta_1}.$$

Moreover, by assumption $E(\varepsilon_t)=0,\ E(\eta_t\varepsilon_t)=0,$ and $E\left(\varepsilon_t^2\right)=\sigma_\varepsilon^2.$ Thus, we get:

(2.69)

$$E(\varepsilon_t p_t^w)=\frac{\beta_1\sigma_\varepsilon^2}{1-\alpha_1\beta_1}\neq 0.$$

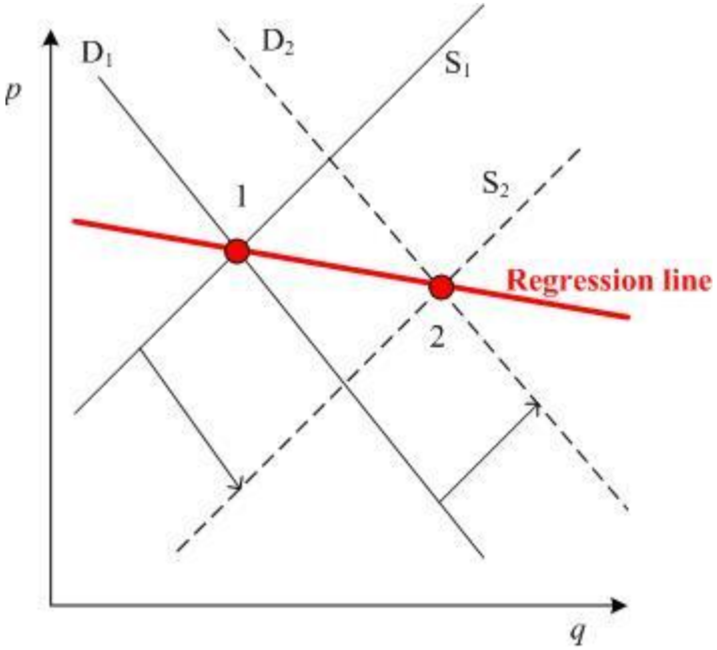
A similar analysis yields:

(2.70)

$$E(\eta_t q_t)=\frac{\alpha_1\sigma_\eta^2}{1-\alpha_1\beta_1}\neq 0.$$

Equations (6) and (7) are what create the endogeneity problem (or *simultaneous equation bias*)—using OLS to estimate the parameters of equations that have an endogenous variable as an explanatory variable yields biased estimates of the unknown parameters. Figure 1 illustrates the endogeneity problem. In this figure we have demand and supply equations that have both risen due to changes in exogenous variables. What the researcher observes are two (red) points: (1) the intersection of the old demand and supply curves and (2) the intersection of the new demand and supply curves.

Figure 2.28.



The simultaneous equation problem.

The thick red line shows the regression that would result from using OLS to estimate either of the two structural equations. As illustrated, an OLS estimate of the slope estimate will be biased. We need to use some other estimation technique than OLS.

Estimation

As noted earlier, the basic problem created by the endogeneity problem is that the endogenous explanatory variable is correlated with the error term. The most logical approach would be to replace this variable with one that is not correlated with the error term but highly correlated with the endogenous variable. Consider the value of the price predicted by the *reduced form* equation (5):

(2.71)

$$\widehat{p}_t^w = \widehat{\gamma}_0 + \widehat{\gamma}_1 I_t + \widehat{\gamma}_2 p_t^c + \widehat{\gamma}_3 W_t$$

where $\widehat{\gamma}_i$ is the OLS estimate of $\gamma_0 = \frac{\beta_0 + \beta_1 \alpha_0}{1 - \alpha_1 \beta_1}$, $\gamma_1 = \frac{\alpha_2 \beta_1}{1 - \alpha_1 \beta_1}$, $\gamma_2 = \frac{\alpha_3 \beta_1}{1 - \alpha_1 \beta_1}$, and $\gamma_3 = \frac{\beta_2}{1 - \alpha_1 \beta_1}$.

Clearly, $\widehat{p}_t^w$ is correlated with p_t^w. It also is true that the covariance between $\widehat{p}_t^w$ and ε_t goes to zero as the sample size increasing. Thus, we can use (8) to construct a variable that will produce a consistent estimator of α_1. It is this conclusion that underlies the strategy of both two-stage least squares (TSLQ) and instrumental variable (IV) estimators.

Two-stages least squares

The easiest way to understand two-stage least squares is to think of the estimation process as being in the following two steps (although the computer programs calculate the estimators in one step):

Stage 1: obtain a OLS predictions for any endogenous variable on the right-hand side of the equation to be estimated using as the explanatory variables all of the exogenous variables in the system.

Stage 2: estimate the parameters of the equation using OLS and replacing the endogenous variable on the right-hand side of the equation by the its predictions as obtained in step 1.

For obvious reasons he TSLS method works best when the full model is specified or when you know and can measure all of the exogenous variables in the system.

Instrumental variables (IV)

While the use of instrumental variable (IV) estimators is appropriate in a large number of situations, the two situations where they are most commonly used are (1) in the presence of endogenous explanatory variables and (2) in cases when errors arise in the measurement of an explanatory variable (or the *errors-in-variables* problem). Since I have already described the endogeneity problem, I now turn to a brief discussion of errors-in-variables.

Consider the following simple model:

$$(2.72)\quad y_i = \beta_1 x_i^* + \varepsilon_i \text{ and } x_i = x_i^* + \mu_i.$$

In this model the researcher observes x_i but not the desired x_i* because of some random measurement error. Using OLS to estimate (9) using the observable x_i instead of the correct x_i* is equivalent to estimating:

(2.73)

$$y_i = \beta_1 x_i + (\varepsilon_i - \beta_1 \mu_i).$$

The important thing to note in estimating (10) using OLS is that the explanatory variable, x_i, is correlated with the error term, $(\varepsilon_i - \beta_1 \mu_i).$ As was the case with the endogeneity problem, the OLS estimate of β_1 is biased. Murray (2006) summarizes the situation as follows:

> *In both examples, ordinary least squares estimation is biased because an explanatory variable in the regression is correlated with the error term in the regression. Such a correlation can result from an endogenous explanator, a mismeasured explanator, an omitted explanator, or a lagged dependent variable among the explanators. I call all such explanators "troublesome." Instrumental variable estimation can consistently estimate coefficients when ordinary least squares cannot—that is, the instrumental variable estimate of the coefficient will almost certainly be very close to the coefficient's true value if the sample is sufficiently large—despite troublesome explanators. [Murray (2006a): 112]*

Consider a regression that includes a "troublesome explanator," like x_i* in (9). Assume that there exists a variable z_i (or set of variables) that (1) is correlated with the "troublesome explanator," (2) is uncorrelated with the error term—like ε_i in (9), and (3) is not one of the explanatory variables in the equation to be estimated. Greene (1990: 300) offers the following example of such a variable. Self-reported income tends to be a very "noisy" variable because sometimes people forget to report minor sources of income and sometimes they deliberately misreport their income. If the regression you are estimating uses income as explanatory variable of consumption, OLS will yield biased estimates. On the other hand, the number of checks written in a month by the household head might serve as an instrumental variable. Clearly, the number of checks written might well be positively correlated with income and there is no reason to assume that it is correlated with the error term in the consumption equation.[26]

It is usually fairly easy to identify instances when IV estimation methods are appropriate. This is especially true when one of the explanatory variables is possibly an endogenous variable. The real problem arises in finding an instrumental variable or a set of instrumental variables. However, assuming you have one or more instrumental variables, the IV method follows the same steps as described above for TSLS. In the first stage you estimate a regression of the "troublesome variable" as a function of the instruments and the exogenous variables in the equation—i.e., you estimate the reduced form equation. In the second stage you use OLS to estimate the original equation with the value of the "troublesome variable" predicted by the first stage regression substituted for the actual values of the "troublesome variable."

In a sense TSLS is a IV estimation. The exogenous variables not in a particular regression play the role of the instruments. Thus, in the IV estimation of (1), the weather index is the instrument. In the estimation of (2) the price of corn and the income level are the IVs. Thus, in a fully specified model, the exogenous variables excluded from the regression play the role of instrumental variables. In other situations the choice of an appropriate instrument can be very difficult. The selection process demands creativity both in finding the instrument and in defending the choice.

The use either of IV or TSLS comes at a cost. First, the OLS estimators are more precise (i.e., have a smaller standard error) than the TSLS or IV estimators. Second, selecting invalid or weak instruments can create results that are not meaningful. So how does one know if they have chosen a good set of instruments? There is no easy answer to this question. Murray (2006a: 116-117) discusses some possible tests of the validity of an instrumental variable. In the end, however, the "success" of your instrument may depend more on how convincing your justifications are than any statistical test. Some economists, like Steven Levitt, make a living coming up with and justifying the use of some very creative instrumental variables. Murray (2006a) offers a detailed discussion of IV and should be read by any student planning to make use either of TSLS or IV regression estimators.

The identification problem

There is an additional issue that arises with estimating systems of equations—identification. Essentially, identification is an *algebraic* problem. Consider the reduced form equations given earlier in (4) and (5):

$$q_t = \frac{\alpha_0 + \alpha_1 \beta_0}{1 - \alpha_1 \beta_1} + \frac{\alpha_1 \beta_2}{1 - \alpha_1 \beta_1} W_t + \frac{\alpha_2}{1 - \alpha_1 \beta_1} I_t + \frac{\alpha_3}{1 - \alpha_1 \beta_1} p_t^c + \frac{\varepsilon_t + \alpha_1 \eta_t}{1 - \alpha_1 \beta_1}$$

and

$$p_t^w = \frac{\beta_0 + \beta_1 \alpha_0}{1 - \alpha_1 \beta_1} + \frac{\alpha_2 \beta_1}{1 - \alpha_1 \beta_1} I_t + \frac{\alpha_3 \beta_1}{1 - \alpha_1 \beta_1} p_t^c + \frac{\beta_2}{1 - \alpha_1 \beta_1} W_t + \frac{\beta_1 \varepsilon_t + \eta_t}{1 - \alpha_1 \beta_1}.$$

OLS estimation of both of these equations yields unbiased estimates of the parameters in the reduced form equations. Identification asks if we can retrieve the parameters of the structural equations from the reduced form equations. Say, for instance, that we re-write the reduced form equations as:

$$\text{(2.74)}\ q_t = \delta_{10} + \delta_{11} W_t + \delta_{12} I_t + \delta_{13} p_t^{c} + \gamma_1$$

and

$$\text{(2.75)}\ p_t^{w} = \delta_{20} + \delta_{21} I_t + \delta_{22} p_t^{c} + \delta_{23} W_t + \delta_2 .$$

Table 1 shows each of the parameters in (11) and (12) in terms of the parameters of the two reduced form equations. We can recover the parameters of the structural equations by algebraic manipulation of the relationships in Table 1. (This method of estimation—that is, estimating the reduced form equations of a model using OLS and then solving algebraically for the parameters of the structural equations is referred to in the literature as *indirect least squares*.) For instance,

$$\frac{\delta_{21}}{\delta_{12}} = \frac{\left(\frac{\alpha_2 \beta_1}{1-\alpha_1 \beta_1}\right)}{\left(\frac{\alpha_2}{1-\alpha_1 \beta_1}\right)} = \beta_1$$

and

$$\frac{\delta_{11}}{\delta_{23}} = \frac{\left(\frac{\alpha_1 \beta_2}{1-\alpha_1 \beta_1}\right)}{\left(\frac{\beta_2}{1-\alpha_1 \beta_1}\right)} = \alpha_1 .$$

Explanatory variable	Equation (11)	Equation (12)
Intercept	$\delta_{10} = \frac{\alpha_0 + \alpha_1 \beta_0}{1-\alpha_1 \beta_1}$	$\delta_{20} = \frac{\beta_0 + \beta_1 \alpha_0}{1-\alpha_1 \beta_1}$

I_t	$\delta_{11} = \frac{\alpha_1 \beta_2}{1 - \alpha_1 \beta_1}$	$\delta_{21} = \frac{\alpha_2 \beta_1}{1 - \alpha_1 \beta_1}$
p_t^c	$\delta_{12} = \frac{\alpha_2}{1 - \alpha_1 \beta_1}$	$\delta_{22} = \frac{\alpha_3 \beta_1}{1 - \alpha_1 \beta_1}$
W_t	$\delta_{13} = \frac{\alpha_3}{1 - \alpha_1 \beta_1}$	$\delta_{23} = \frac{\beta_2}{1 - \alpha_1 \beta_1}$
Error term	$\gamma_1 = \frac{\varepsilon_t + \alpha_1 \eta_t}{1 - \alpha_1 \beta_1}$	$\delta_2 = \frac{\beta_1 \varepsilon_t + \eta_t}{1 - \alpha_1 \beta_1}$

Table 2.24. Parameters of the structural and reduced form equations.

One can continue in a likewise manner to find formulae for other of the structural parameters. However, an interesting problem does arrive in that it is also true that $\beta_1 = \frac{\delta_{22}}{\delta_{13}}.$ Since there is no *a priori* reason to believe that $\frac{\delta_{22}}{\delta_{13}} = \frac{\delta_{21}}{\delta_{12}},$ we have two estimates of β_1 . This result illustrates the point that there are three possibilities when calculating the structural parameters from the reduced form equations—first, there may be more than one formula for a structural parameter; second, there may be only one formula for a structural parameter; or third, there may be no formula for a structural parameter. We say in the first case that the equation is over-identified; is exactly identified in the second case; and is under-identified in the third case. It turns out that in the case of an over-identified equation we can to use TSLS to estimate the structural parameters. However, in the case of an exactly identified equation, the TSLS estimators are equal to the indirect-least-squares estimators that can be calculated using estimates of the reduced form equations. Finally, an under-identified equation cannot be estimated by any technique.

Clearly, we need to know how to identify if an equation is either over-identified, exactly identified, or under-identified. A necessary rule is that the number of exogenous variables in a system of equation that are not included in a particular regression must be greater than or equal to the number of endogenous variables on the right-hand-side of the equation for the equation to be either exactly or over identified. Consider the following three-equation model, where the endogenous variables are y_1 , y_2 , and y_3 and the exogenous variables are represented by x_1 with i = 1,…,5 :

$$y_1 = \beta_{10} + \beta_{12} y_2 + \beta_{13} y_3 + \alpha_{11} x_1 + \alpha_{12} x_2 + \alpha_{15} x_5 , \tag{2.76}$$

$$y_2 = \beta_{20} + \beta_{21} y_1 + \alpha_{23} x_3 , \text{ and} \tag{2.77}$$

$$y_3 = \beta_{30} + \beta_{31} y_1 + \alpha_{31} x_1 + \alpha_{32} x_2 + \alpha_{33} x_3 + \alpha_{34} x_4 + \alpha_{35} x_5 . \tag{2.78}$$

The error terms in these three equations are omitted because they are irrelevant to determining if an equation is identified—remember, identification is an algebraic problem, not a statistical issue. There are 3 endogenous variables in the system and 3 equations in the system. Also, there are 5 exogenous variables in the system of equations. Equation (13) is exactly identified; Equation (14) is over-identified; and Equation (15) is under-identified. What this means is (1) Equation (13) can be estimated directly from the reduced form equation (using indirect-least-squares) or using TSLS; (2) Equation (14) must be estimated using TSLS; and Equation (15) cannot be estimated. Table 2 summarizes how to determine if an equation is or is not identified. Basically, if the number in column 2 equals the number in column 3, the equation is exactly identified. If the number in column 2 is less than the number in column 3, the equation is over-identified. Finally, if the number in column 2 is greater than the number in column 3, the equation is under-identified.[27]

Equation	Number of endogenous variables on right-hand-side	Number of exogenous variables excluded from the equation	Identification
$y_1 = \beta_{10} + \beta_{12} y_2 + \beta_{13} y_3 + \alpha_{11} x_1 + \alpha_{12} x_2 + \alpha_{15} x_5$	2	2	Exactly
$y_2 = \beta_{20} + \beta_{21} y_1 + \alpha_{23} x_3$	1	4	Over
$y_1 = \beta_{10} + \beta_{12} y_2 + \beta_{13} y_3 + \alpha_{12} x_2 + \alpha_{13} x_3 + \alpha_{15} x_5$	1	0	Under

Table 2.25. Identification of the equations in the example model.

One other thing to notice is the similarity of TSLS to IV estimation. The exogenous variables play the role of instruments in TSLS estimation. By implication, the instruments in an IV estimation must not include any of the exogenous variables in the equation.[28] Similarly, one of the

ways to isolate potential instruments in a regression is to think of what system of equation the equation is and then ask what exogenous variables in that system are not included in the equation. These excluded exogenous variables are potential instruments.

TSLS and IV in Stata

The command for estimating an equation in *Stata* using two-stages least squares (TSLS) is a bit tricky. Assume that you want to estimate equations (13) and (14) in the model discussed above.[29] For simplicity assume that each variable assumes the name for it in Table 2. Thus, in our *Stata* commands Y1 refers to variable Thus, in our Stata commands Y1 refers to variable y_1 and so on. The command to estimate either a TSLS or an IV regression is the same.[30] The command, ivreg, consists of three major parts—(1) the name of the dependent variable is followed by (2) a list of the names of the exogenous variables that are being used as explanatory variables and then followed in parentheses by (3) the information needed to estimate the first stage (the list of the endogenous variables that are explanatory variables along with the names of the exogenous variables in the system that are excluded from the equation or, in the case of IV, a list of the instruments).[31]

Equation to be estimated	*Stata* command
$y_1 = \beta_{10} + \beta_{12} y_2 + \beta_{13} y_3 + \alpha_{12} x_2 + \alpha_{13} x_3 + \alpha_{15} x_5$	.ivreg y1 x2 x3 x5 (y2 y3 = x1 x4)
$y_2 = \beta_{20} + \beta_{21} y_1 + \alpha_{23} x_3$	.ivreg y2 x3 (y1 = x1 x2 x4 x5)

Table 2.26. *Stata* command for estimating TSLS and IV regressions.

Example 2.5.

An example from *Stata.* The *Stata* manual offers the following example analysis. Assume that you want to use state level data from the 1980 census to estimate the following system of equations:

$$(2.79)\ hsngval = \alpha_0 + \alpha_1 fainc + \alpha_2 reg2 + \alpha_3 reg3 + \alpha_4 reg4 + \varepsilon$$

and

$$(2.80)\ rent = \beta_0 + \beta_1 hsngval + \beta_2 pcturban + v,$$

where *hsngval* is the median dollar value of owner-occupied housing; *rent* is the median monthly gross rent; *fainc* is family income; *pcturban* is the percent of the state population living in an urban area; and *reg2*, *reg3*, and *reg4* are dummy variables that designate the region of the country where the state is located. In this example we focus on estimating (17).

We begin by loading the data set and describing the data.

. use http://www.stata-press.com/data/r8/hsng2

(1980 Census housing data)

.describe

Contains data from http://www.stata-press.com/data/r8/hsng2.dta

obs: 50	1980 Census housing data
vars: 16	3 Sep 2002 12:25

size: 3,600 (99.7% of memory free)	

variable name	storage type	display format	value label	variable	label
state	str14	%	14s		State
division	int	%	8.0g	division	Census division

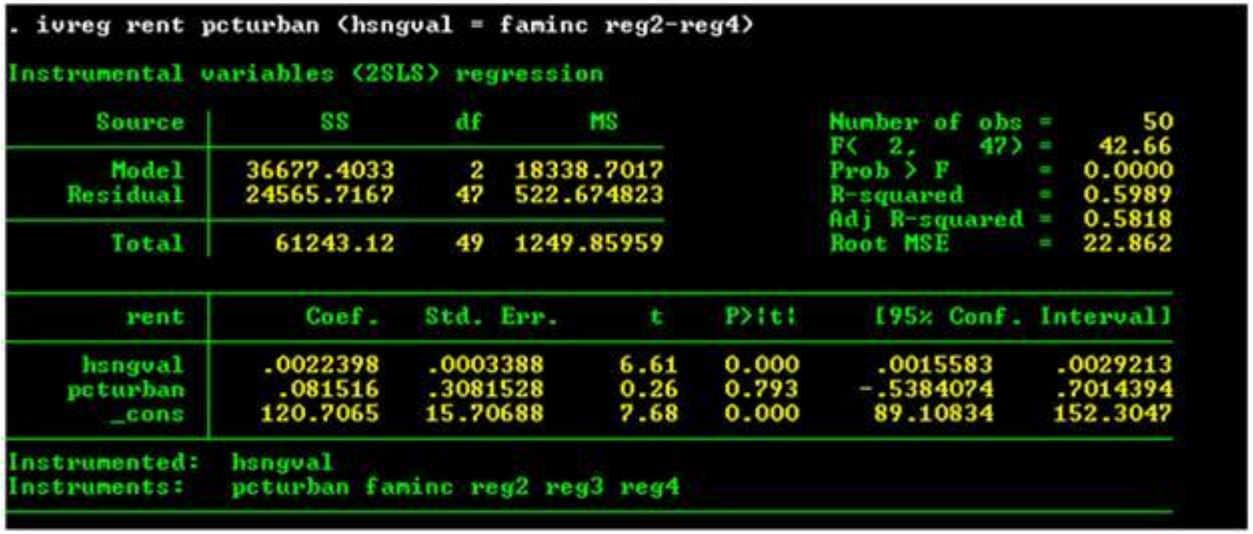
region	int	%	8.0g	region	Region
pop	long	%	10.0g		Population in 1980
popgrow	float	%	6.1f		Pop. growth 1970-80
popden	int	%	6.1f		Pop/sq. mile
pcturban	float	%	8.1f		Percent urban
faminc	long	%	8.2f		Median family inc., 1979
hsng	long	%	10.0g		Hsng units 1980
hsnggrow	float	%	8.1f		% housing growth
hsngval	long	%	9.2f		Median hsng value
rent	long	%	6.2f		Median gross rent
reg1	float	%	9.0g		
reg2	float	%	9.0g		
reg3	float	%	9.0g		
reg4	float	%	9.0g		

Sorted by: state

Table 2.27. Description of the *Stata* data set used in the example.

Now we estimate equation (17) using TSLS as shown in Figure 2.

Figure 2.29. Two-stages least square estimate of the example.



```
. ivreg rent pcturban (hsngval = faminc reg2-reg4)

Instrumental variables (2SLS) regression

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  2,    47) =   42.66
       Model |  36677.4033     2  18338.7017           Prob > F      =  0.0000
    Residual |  24565.7167    47  522.674823           R-squared     =  0.5989
-------------+------------------------------           Adj R-squared =  0.5818
       Total |    61243.12    49  1249.85959           Root MSE      =  22.862

------------------------------------------------------------------------------
        rent |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     hsngval |   .0022398   .0003388     6.61   0.000     .0015583    .0029213
    pcturban |    .081516   .3081528     0.26   0.793    -.5384074    .7014394
       _cons |   120.7065   15.70688     7.68   0.000     89.10834    152.3047
------------------------------------------------------------------------------
Instrumented:  hsngval
Instruments:   pcturban faminc reg2 reg3 reg4
------------------------------------------------------------------------------
```

The manual continues the example to include some testing of the model including the Hausman test. Students using TSLS and IV should read the discussion in the *Stata* manual thoroughly.

Exercises

Exercise 2.5.1.

Cigarette advertising and sales. A great deal of controversy exists over the issue of whether advertising expenditures affect sales. This controversy is particularly sharp when it affects policy decisions. An example of this phenomenon is the controversy over the impact of cigarette advertising on advertising sales. While many public policy experts advocate bans on cigarette advertising, a majority of economists caution against bans on cigarette advertising. The economists point out that there is little theoretical reasons to believe that cigarette advertising affects total demand for cigarettes. Instead, economists argue that cigarette advertising only affects brand choice and not the number of cigarettes that people smoke. Moreover, these economists point out that there is also little empirical evidence that supports the argument that cigarette advertising affects the demand for cigarettes. Given the

negative impact advertising bans have on freedom of speech, most economists conclude that the negative effects of cigarette advertising bans outweigh the benefits of the bans.

In this exercise we address this issue by using data used originally by Richard Schmalensee (1972) in his Ph.D. dissertation. You will use these data to estimate a simple two-equation model of the cigarette advertising industry.

We use annual data for the period 1955 to 1967 to estimate the impact of cigarette advertising on aggregate demand for cigarettes and the impact of cigarette consumption on cigarette advertising. We begin with a model of the demand for cigarettes. We assume that the demand for cigarettes is given by:

(2.81)

$$q_t = f(pc_t, y_t, A_t, D64),$$

where

q_t = cigarettes consumed per person over age 15,

pc_t = retail price of cigarettes,

y_t = real disposable personal income per capita (1958 dollars),

A_t = real advertising expenditures per individual over age 15 (1960 dollars), and

D64 = a dummy variable equal to 1 for the years 1964 through 1967 and zero otherwise.

We include the dummy variable for years after 1964 to pick up the negative impact on cigarette sales of the 1964 report of the US Surgeon General’s Advisory Committee (1964) announcing that the government believed that there was enough evidence available to conclude that cigarette smoking causes cancer. We expect the signs of the parameters with the price of cigarettes and the dummy variable to be negative. We expect that the sign of the parameters with income and advertising to be positive.

Next we turn to a model of the supply of advertising. We assume:

(2.82)

$$A_t = g(q_t, pa_t, m_t),$$

where:

pa_t = advertising price index, and

m_t = gross profits as a percentage of gross sales.

The last variable needs a bit of explaining. The amount of advertising in the industry should be a function of degree of competition in the industry. If the market were perfectly competitive, there would be no reason for any firm to advertise. If the firm were a monopoly, there also would be no reason to advertise. However, if the market is an oligopoly, then a firm would advertise in an effort to gain market share by differentiating its product from the product of its competitors.

The traditional measure of the degree of monopoly power that a firm has is the ratio of its marginal profits to its marginal cost:

(2.83)

$$m = \frac{p - mc}{mc},$$

where *p* is output price, *mc* is marginal cost, and *m* is the measure of monopoly power. Since we cannot observe the firms' marginal costs, we approximate *m* by the ratio of gross profits to gross sales. We expect the impact of the degree of monopoly to have a non-linear impact on advertising expenditures.

The data used to estimate our two equations are listed in Table 5 and are available in the MS Excel file Cigarette sales and advertising data.xls. These data are with the exception of disposable personal income from Schmalensee (1972: 273-290). The disposable personal income data are from the Department of Commerce (1975: Table F26, page 225).

Specification of the Model. Equations (18) and (19) are, as written, very general and need further specification before they can be estimated. We will assume that the two equations take a log-log form. In particular, we assume that we want to estimate:

(2.84)

$$\ln(q_t) = \alpha_0 + \alpha_1 \ln(pc_t) + \alpha_2 \ln(y_t) + \alpha_3 \ln(A_t) + \alpha_4 D64_t$$

and

(2.85)

$$\ln(A_t) = \beta_0 + \beta_1 \ln(q_t) + \beta_2 \ln(pa_t) + \beta_3 m_t + \beta_4 m_t^2.$$

Year	Cigarettes Sold per Person Over Age 15	Retail Price of Cigarettes	Real Advertising per Person Over Age 15	Advertising Price Index	Degree of Monopoly	Disposable Personal Income in 1958 dollars
1955	3163.090	93.9693	0.96100	95.4775	18.595	1659
1956	3230.517	94.7049	1.09969	94.3800	19.207	1673
1957	3313.033	94.2535	1.22180	96.2125	20.165	1683
1958	3479.063	94.7712	1.40471	97.8300	21.736	1666
1959	3584.930	98.1779	1.45816	98.2800	22.042	1735
1960	3676.912	100.0000	1.37863	100.0000	22.04	1749
1961	3743.354	99.8677	1.31871	102.0400	22.465	1756
1962	3733.504	99.6761	1.35467	102.9725	22.226	1814
1963	3775.886	101.3630	1.51345	103.9525	22.848	1867
1964	3648.211	102.3110	1.73665	103.4775	23.168	1948
1965	3710.075	105.7510	1.59761	103.7225	23.598	2047
1966	3689.386	108.0450	1.71062	104.2200	25.085	2127
1967	3652.016	109.2490	1.71444	104.6125	26.310	2164

Table 2.28. Cigarette Industry Data, 1955-1967.

Answer the following six questions:

a) Which variables in the model are exogenous and which are endogenous?

b) Check and see if equations (18) and (19) are underidentified, exactly identified, or overidentified.

c) Estimate equations (21) and (22) using ordinary least squares.

d) Estimate equations (21) and (22) using two-stage least squares. Present the results in a table that for comparison reasons includes the results from the OLS estimation. Be sure to include the R^2 and the Durbin-Watson statistic.

e) Which side of the advertising-sales controversy do your results appear to support?

f) How well-specified does your model appear to be? Why?

Exercise 2.5.2.

Exercise 2. Demand and supply of commercial loans. We are interested in estimating the demand for commercial loans by business firms and the supply of commercial loans by banks. We have available in Table 6 monthly data from the U. S. commercial loan market for the period from January, 1979 through December, 1984 and available in the MS Excel file Exercise 2.xls.[32] Define:

Q_t = total commercial loans (billions of dollars)

R_t = average prime rate charged by banks

RS_t = 3-month Treasury bill rate (represents an alternative rate of return for banks)

RD_t = Aaa corporate bond rate (represents the price of alternative financing to firms)

X_t = industrial production index (represents firms' expectation about future economic activity)

y_t = total bank deposits (billions of dollars) (represents a scale variable).

The demand and supply equations to be estimated, respectively, are as follows:

$$(2.86)\quad Q_t = \beta_0 + \beta_1 R_t + \beta_2 RD_t + \beta_3 X_t + \mu_t$$

and

$$(2.87)\quad Q_t = \alpha_0 + \alpha_1 R_t + \alpha_2 RS_t + \alpha_3 y_t + \varepsilon_t .$$

Questions

a) What are the endogenous and exogenous variables in this model?

b) Solve for the two "reduced form" equations of this model. Estimate these two equations using the data in Table 6.

c) Check the "order" condition for identification of each equation of the model.

d) Estimate equations (23) and (24) using ordinary least squares using the data in Table 6.

e) Estimate equations (23) and (24) using two-stage least squares. Report the results of the estimations for part 4 and 5 in a single table. Be sure to include the t-ratios, R^2's, and Durbin-Watson statistics for each of the equations estimated.

f) Perform the Hausman Specification Test on both equations.[33]

g) When presenting this model, Maddala notes "[T]he model postulated here is not necessarily the right model for the problem of analyzing the commercial loan market." Is there anything in the results reported above that suggests that the model may be mis-specified?

N	Date	*Q*	*R*	*RD*	*X*	*RS*	*y*
1	January-79	251.8	11.75	9.25	150.8	9.35	994.3
2	February-79	255.6	11.75	9.26	151.5	9.32	1002.5

3	March-79	259.8	11.75	9.37	152.0	9.48	994.0
4	April-79	264.7	11.75	9.38	153.0	9.46	997.4
5	May-79	268.8	11.75	9.50	150.8	9.61	1013.2
6	June-79	274.6	11.65	9.29	152.4	9.06	1015.6
7	July-79	276.9	11.54	9.20	152.6	9.24	1012.3
8	August-79	280.5	11.91	9.23	152.8	9.52	1020.9
9	September-79	288.1	12.90	9.44	151.6	10.26	1043.6
10	October-79	288.3	14.39	10.13	152.4	11.70	1062.6
11	November-79	287.9	15.55	10.76	152.4	11.79	1058.5
12	December-79	295.0	15.30	11.31	152.1	12.64	1076.3
13	January-80	295.1	15.25	11.86	152.2	13.50	1063.1
14	February-80	298.5	15.63	12.36	152.7	14.35	1070.0
15	March-80	301.7	18.31	12.96	152.6	15.20	1073.5
16	April-80	302.0	19.77	12.04	152.1	13.20	1101.1
17	May-80	298.1	16.57	10.99	148.3	8.58	1097.1
18	June-80	297.8	12.63	10.58	144.0	7.07	1088.7
19	July-80	301.2	11.48	11.07	141.5	8.06	1099.9
20	August-80	304.7	11.12	11.64	140.4	9.13	1111.1
21	September-80	308.1	12.23	12.02	141.8	10.27	1122.2
22	October-80	315.6	13.79	12.31	144.1	11.62	1161.4

23	November-80	323.1	16.06	11.94	146.9	13.73	1200.6
24	December-80	330.6	20.35	13.21	149.4	15.49	1239.9
25	January-81	330.9	20.16	12.81	151.0	15.02	1223.5
26	February-81	331.3	19.43	13.35	151.7	14.79	1207.1
27	March-81	331.6	18.04	13.33	151.5	13.36	1190.6
28	April-81	336.2	17.15	13.88	152.1	13.69	1206.0
29	May-81	340.9	19.61	14.32	151.9	16.30	1221.4
30	June-81	345.5	20.03	13.75	152.7	14.73	1236.7
31	July-81	350.3	20.39	14.38	152.9	14.95	1221.5
32	August-81	354.2	20.50	14.89	153.9	15.51	1250.3
33	September-81	366.3	20.08	15.49	153.6	14.70	1293.7
34	October-81	361.7	18.45	15.40	151.6	13.54	1224.6
35	November-81	365.5	16.84	14.22	149.1	10.86	1254.1
36	December-81	361.4	15.75	14.23	146.3	10.85	1288.7
37	January-82	359.8	15.75	15.18	143.4	12.28	1251.5
38	February-82	364.6	16.56	15.27	140.7	13.48	1258.3
39	March-82	372.4	16.50	14.58	142.7	12.68	1295.0
40	April-82	374.7	16.50	14.46	141.5	12.70	1272.1
41	May-82	379.3	16.50	14.26	140.2	12.09	1286.1
42	June-82	386.7	16.50	14.81	139.2	12.47	1325.8

43	July-82	384.4	16.26	14.61	138.7	11.35	1307.3
44	August-82	384.5	14.39	13.71	138.8	8.68	1321.7
45	September-82	395.0	13.50	12.94	138.4	7.92	1335.5
46	October-82	393.7	12.52	12.12	137.3	7.71	1345.2
47	November-82	398.9	11.85	11.68	135.7	8.07	1358.1
48	December-82	395.3	11.50	11.83	134.9	7.94	1409.7
49	January-83	392.4	11.16	11.79	135.2	7.86	1385.4
50	February-83	392.3	10.98	12.01	137.4	8.11	1412.6
51	March-83	395.9	10.50	11.73	138.1	8.35	1419.5
52	April-83	393.5	10.50	11.51	140.0	8.21	1411.0
53	May-83	391.7	10.50	11.46	142.6	8.19	1413.1
54	June-83	395.3	10.50	11.74	144.4	8.79	1443.8
55	July-83	397.7	10.50	12.15	146.4	9.08	1438.1
56	August-83	400.6	10.89	12.51	149.7	9.34	1461.4
57	September-83	402.7	11.00	12.37	151.8	9.00	1448.9
58	October-83	405.3	11.00	12.25	153.8	8.64	1459.0
59	November-83	412.0	11.00	12.41	155.0	8.76	1499.4
60	December-83	420.1	11.00	12.57	155.3	9.00	1508.9
61	January-84	424.4	11.00	12.20	156.2	8.90	1504.1
62	February-84	428.8	11.00	12.08	158.5	9.09	1499.3

63	March-84	433.1	11.21	12.57	160.0	9.52	1494.5
64	April-84	439.7	11.93	12.81	160.8	9.69	1501.5
65	May-84	447.3	12.39	13.28	162.1	9.83	1541.3
66	June-84	452.9	12.60	13.55	162.8	9.87	1532.9
67	July-84	454.4	13.00	13.44	164.4	10.12	1535.5
68	August-84	455.2	13.00	12.87	165.9	10.47	1539.0
69	September-84	459.9	12.97	12.66	166.0	10.37	1549.9
70	October-84	467.7	12.58	12.63	165.0	9.74	1578.9
71	November-84	468.7	11.77	12.29	164.4	8.61	1578.2
72	December-84	476.8	11.06	12.13	164.8	8.06	1631.2

Table 2.29. Monthly Data for the U.S. Commercial Loan Market, January 1979 to December 1984.

References

Angrist, Joshua D. and Alan B. Krueger (2001). Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic Perspectives* 15(4): 69–85.

Berndt, Ernst R. (1991). *The Practice of Econometrics* (Reading, MA: Addison-Wesley Publishing Company).

Greene, William H. (1990). *Econometric Analysis* (New York: Macmillan Publishing Company).

Murray, Michael P. (2006a). Avoiding Invalid Instruments and Coping with Weak Instruments. *Journal of Economic Perspectives* 20(4): 111-132.

Murray, Michael P. (2006b). *Econometrics: A Modern Introduction.* (Boston: Addison-Wesley): Chapter 13.

Schmalensee, Richard (1972). *The Economics of Advertising* (Amsterdam: North-Holland Publishing Company).

StataCorp (2003). *Stata Statistical Software: Release 8* (College Station, TX: Stata Corporation): Volume 2: Reference G-M, pages 186-194.

Stock, James H, and Mark W. Watson (2003). *Introduction to Econometrics* (Boston, MA: Addison-Wesley): Chapter 10.

US Department of Commerce (1975). *Historical Statistics of the United States: Colonial Times to 1970* (Washington: Government Printing Office).

US Surgeon General's Advisory Committee (1964). *Smoking and Health* (Washington: Government Printing Office).

2.6. Replication of econometric studies[*]

Replication

Introduction

One of the most important first steps in a science experiment is to replicate the results of earlier research. For a variety of reasons (most of them practical and not theoretically sound) economists generally do not undertake this step; what they tend to do is report the results of earlier papers and then compare their results with the earlier results without asking the question of whether these earlier results were reported accurately. Omitting this step in a world of honest careful researchers might seem to be a minor problem. However, there is enough casual evidence to suggest that a large portion of the econometric results reported in the journals cannot be replicated because the original researcher (1) does not have the data set used in the research because it has been lost for a variety of reasons, (2) cannot share the data set because it is proprietory, (3) is unwilling to share the data set because there are other issues they wish to investigate using the data set, or (4) just are unwilling to share the data set. For this reason much of the published econometrics research has never been replicated. In recognization of this problem several journals like the *Journal of Applied Econometrics* now require that authors submit the data set they used to the journal to be posted on the web for use by any other researcher. Whether this effort has been successful will not be clear unless someone undertakes to replicate the work in this journal to see if all of the data necessary to replicate an article have been posted and if the regressions included in the article actually can be replicated. It is very unlikely anyone would undertake such an effort given the fact that no journal will publish results that are merely a replication of previously published articles.

In this module we explore some of the difficulties that exist in replicating existing research by undertaking to replicate some of the results reported in the Butler, Finegan, and Siegfried (1998) (BFS, hereafter) article analyzing the effect of a student's calculus background on the grade he or she earns in intermediate microeconomics or in intermediate macroeconomics.[34] The goal of this module is to (1) help students to learn how to read in detail an article that appears in a typical economics trade journal, (2) introduce them to ordered probit, an advanced econometrics tool, and (3) teach them how to present and discuss the results of an estimation of a model in an economics paper. While most of the discussion in this module focuses on using *Stata* in this replication, one can use most any econometrics program they are comfortable with to replicate some of the results reported in the BFS article.

Butler, Finegan, and Siefried (1998).

The obvious first step is to find and print a copy of the article by Butler, Finegan, and Siefried. In fact, do not proceed any further in reading this module until you have read the article. We will discuss in class what the authors do in the paper and how clearly they present their conclusions. In this first pass at the article you are to pay attention to how convincing you find their arguments to be. Since everyone in the class has completed an intermediate microeconomics course, your discussion of their conclusions should reflect your own experiences. Also, you need to be able to discuss in class the estimation strategy they use in the paper. In particular, you will need to be able to identify what the source of the data is and what equations did they estimate. Also, try to determine how the estimations in the "first" stage are used in the estimations of the "second" stage. Why did the authors use a two-stage estimation strategy?

Also, what do you think the authors mean in their description of their estimation strategy by their statement about the estimation methods they use:

> *Estimation Methods and Expectations*
>
> *To cope with the selection bias problem, we use a two-stage estimation procedure. The first stage employs an ordered probit model to predict the highest level of calculus attained by each student prior to taking each intermediate economic theory course.... In the second stage, the student's grade in MICRO-2 ... (the `outcome') is regressed on the actual level of calculus attained, the grade earned in that calculus course, the predicted residual in the grade equation that we would expect on the basis of the actual level of calculus attained, and a roster of control variables reflecting ability and motivation. Individuals are*

> *the unit of observation. Ordinary least squares estimation is used because there are twelve categories of grades which are commonly interpreted as cardinal measures of performance (as is implied by the calculation of `grade point averages'). (Butler, Finegan, and Siegfried, 1998: 188)*

The ordered-probit model

In what follows you are to “replicate” the equations the authors estimate in the paper for the intermediate microeconomics course. In order to complete this assignment you will need to figure out several things including (1) what an ordered-probit model is and (2) how to use *Stata* to estimate an ordered-probit model. In this section of the module we introduce the ordered-probit model. I strongly encourage you to consult Greene (1990: 703-706) for an excellent and clear discussion of the ordered-probit model. The discussion here follows Greene closely.

It is common for surveys to have questions that require the responder to choose one of several categories that have an innate order to them. For instance, most course evaluations ask the respondent to choose an answer to a question that reflects their agreement with a statement about the course. For instance, the question might read, "The Professor was interested in the material taught in the class" where the student completing the evaluation would choose a number from 1 to 9 where a 1 indicates complete disagreement with the statement and a 9 reflects complete agreement with the statement. Thus, there is an order to the potential answers. Using a logit, probit, or multilogit model would completely ignore this order. A linear regression is inappropriate because OLS treats the difference between answers of 1 and 2 as being the same as the difference between a 7 and and 8, when in fact the numbers only provide a ranking.

Consider a latent variable, *y**, that is not observed but where $y = \boldsymbol{\beta}' \mathbf{x} + \varepsilon.$ We want to estimate the *β* $_k$' s in the vector $\boldsymbol{\beta} = (\beta_0 \;\; \beta_1 \;\; \cdots \;\; \beta_K).$ [35] We may not observe *y** but we do observe:

The μ_i's in (1) are parameters that must be estimated along with $\boldsymbol{\beta}.$ As usual, we assume that the error term *ε* is normally distributed (with a normalized mean and variance arbitrarily set to 0 and 1, respectively). It is trivial to estimate the model with the error terms having a logistic distribution, but this chance in assumptions appears to make virtually no difference in practice).[36] With the normal distribution, we have:

(2.88)

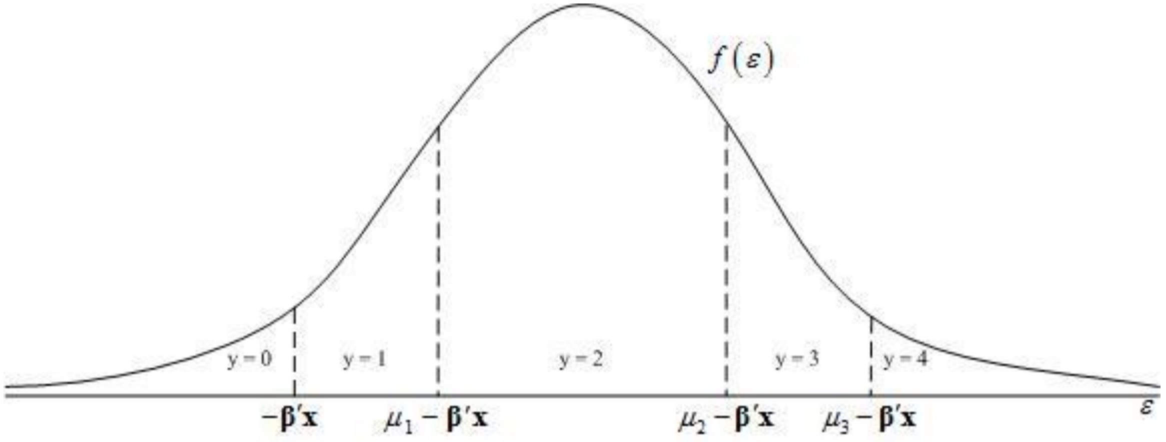
$$y=\begin{cases}0\text{if}y^{*}<0,\\ 1\text{if}0\leq y^{*}<\mu_{1},\\ 2\text{ if}\mu_{1}\leq y^{*}<\mu_{2},\\ \vdots\\ J\text{if}\mu_{J-1}\leq y^{*}.\end{cases}$$

(2.89)

$$\begin{aligned}&\Pr(y=0)=\Phi(-\boldsymbol{\beta}'\mathbf{x}),\\ &\Pr(y=1)=\Phi(\mu_{1}-\boldsymbol{\beta}'\mathbf{x})-\Phi(-\boldsymbol{\beta}'\mathbf{x}),\\ &\Pr(y=2)=\Phi(\mu_{2}-\boldsymbol{\beta}'\mathbf{x})-\Phi(\mu_{1}-\boldsymbol{\beta}'\mathbf{x}),\\ &\vdots\\ &\Pr(y=J)=1-\Phi(\mu_{J-1}-\boldsymbol{\beta}'\mathbf{x}),\end{aligned}$$

where $\boldsymbol{\Phi}(\cdot)$ is the cumulative normal function. In order for all of the probabilities to be positive, we need $\mu_1 < \mu_2 < \cdots < \mu_{J-1}$, as shown in Figure 1. One thing to note in Figure 1 is that the cutoff locations change when the values of the explanatory variables change.

Figure 2.30.


$f(\varepsilon)$
y = 0
y = 1
y = 2
y = 3
y = 4
$-\boldsymbol{\beta}'\mathbf{x}$
$\mu_1-\boldsymbol{\beta}'\mathbf{x}$
$\mu_2-\boldsymbol{\beta}'\mathbf{x}$
$\mu_3-\boldsymbol{\beta}'\mathbf{x}$
ε

Distribution of the error term in the ordered-probit model.

The estimation strategy from here follows the usual maximum likelihood method. The computer program forms the likelihood function and then chooses the values of the parameters (including the cutoffs) that maximize this likelihood function.

The estimated coefficients are not equal to the marginal effects of a change in one of the explanatory variables (as is also true with the logit and probit models). Consider the simple example Greene (1990, 704) describes. Assume that there are three categories. Then (2) becomes:

(2.90)

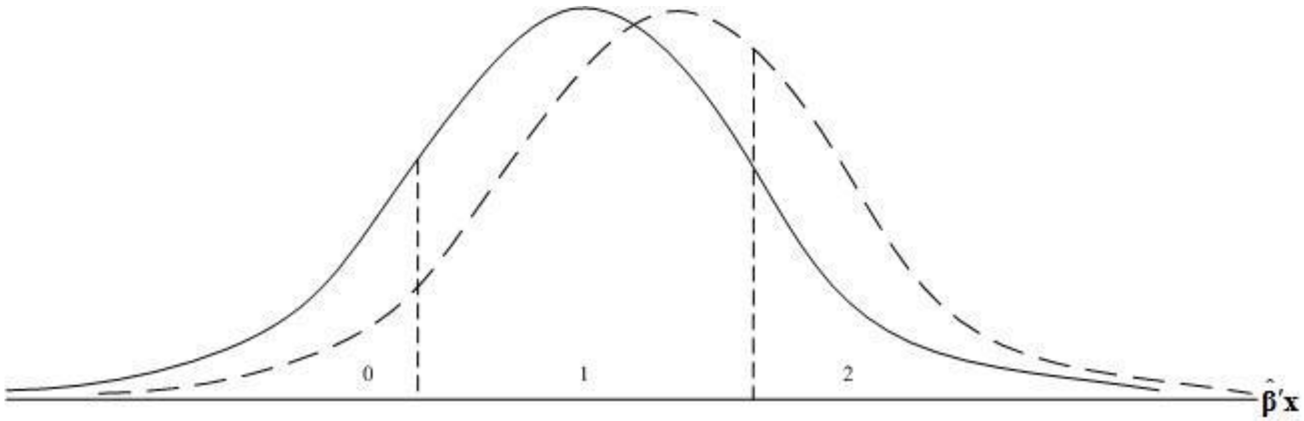
$$\begin{aligned}\Pr(y=0)&=1-\Phi(\boldsymbol{\beta}'\mathbf{x}),\\ \Pr(y=1)&=\Phi(\mu-\boldsymbol{\beta}'\mathbf{x})-\Phi(-\boldsymbol{\beta}'\mathbf{x}),\\ \Pr(y=2)&=1-\Phi(\mu-\boldsymbol{\beta}'\mathbf{x}).\end{aligned}$$

Figure 2 shows this situation. The solid curve shows the distribution of *y* and *y**. Increasing one of the x's while holding the β constant (that is, changing $\hat{\boldsymbol{\beta}}'\mathbf{x}_0$ to $\hat{\boldsymbol{\beta}}'\mathbf{x}_1)$ is the same as shifting the entire distribution of *y* and *y** to the right with $\hat{\mu}$ remaining constant. As a result the probabilities that y takes on the values of 0, 1, and 2 change. Clearly, as shown in Figure 2, Pr(*y* = 0) decreases and Pr(*y* = 2) increases. The Pr(*y* = 1), on the other hand, may increase or decrease and, thus, the effect of an increase in one of the explanatory variables is ambiguous. It is easy to show this result algebraically. The marginal effects for the 3 probabilities in (3) are, assuming $\boldsymbol{\beta}>0$:

(2.91)

$$\begin{aligned}\frac{\partial\Pr(y=0)}{\partial\mathbf{x}}&=-\phi(\boldsymbol{\beta}'\mathbf{x})\boldsymbol{\beta}<0,\\ \frac{\partial\Pr(y=1)}{\partial\mathbf{x}}&=\phi(\mu-\boldsymbol{\beta}'\mathbf{x})\boldsymbol{\beta}-\phi(\boldsymbol{\beta}'\mathbf{x})\boldsymbol{\beta},\\ \frac{\partial\Pr(y=2)}{\partial\mathbf{x}}&=\phi(\mu-\boldsymbol{\beta}'\mathbf{x})\boldsymbol{\beta}>0.\end{aligned}$$

Figure 2.31.



A rise in one of the explanatory variables whose parameter is positive will shift the probability distribution of the outcome to the right (from the solid line to the dashed line).

In general, only the sign's of the change Pr($y = 0$) and Pr($y = J$) are unambiguous. Greene (1990, 705) cautions that "“[w]e must be very careful in interpreting the coefficients in this model.... Indeed, without a fair amount o extra calculation, it is quite unclear how the coefficients in the ordered-probit model should be interpreted.”"

The BFS Dataset

The data used by BFS are available at the *Journal of Applied Econometrics* data website or in the *MS Excel* file *Vanderbilt data set.xls* . Table 1 identifies the variables in the dataset.

Column	Code	Variable definition
A	Obs	Observation number
B	SID	Student ID

C	Grade	Grade earned in Economics 231, A = 4, A- = 3.7, etc.
D	SelCorr	Variable correcting for selection bias
E	Soph	Dummy variable = 1 if student is a sophomore
F	Senior	Dummy variable = 1 if student is a senior
G	Same	Dummy variable = 1 if student took both intermediate classes the same year
H	Skip	Dummy variable = 1 if student took the intermediate classes at least one semester apart
I	HighestMath	Highest level of math attained (the dependent variable, 0-6 corresponding to Math 170, 171a, 172a, 171b, 172b, 221a, 221b)
J	M170	Dummy variable = 1 if student's highest level of math was Math 170
K	M171a	Dummy variable = 1 if student's highest level of math was Math 171A
L	M172a	Dummy variable = 1 if student's highest level of math was Math 172a
M	M171b	Dummy variable = 1 if student's highest level of math was Math 171b
N	M172b	Dummy variable = 1 if student's highest level of math was Math 172b
O	M221a	Dummy variable = 1 if student's highest level of math was Math 221a
P	M221b	Dummy variable = 1 if student's highest level of math was Math 221b
Q	GE100	Grade in Economics 100
R	GDE100	Individual instructor grade deflator in Economics 100
S	GE101	Grade in Economics 101
T	GDE101	Individual instructor grade deflator in Economics 101
U	GDE231	Individual instructor grade deflator in Economics 231
V	Size	Class size

W	**FGPA**	**Freshman GPA**
X	**Female**	**Dummy variable =1 if student is a female**
Y	**MSAT**	**Score on Math section of the SAT**
Z	**VSAT**	**Score on Verbal section of the SAT**
AA	**TE231**	**Teacher of Economics 231 (numerical code)**
AB	**SE231**	**Section of Economics 231 (numerical code)**
AC	**GM170**	**Grade in highest math class: Math 170**
AD	**GM171a**	**Grade in highest math class: Math 171a**
AE	**GM172a**	**Grade in highest math class: Math 172a**
AF	**GM171b**	**Grade in highest math class: Math 171b**
AG	**GM172b**	**Grade in highest math class: Math 172b**
AH	**GM221a**	**Grade in highest math class: Math 221a**
AI	**GM221b**	**Grade in highest math class: Math 221b**
AJ	**GHM**	**Grade in highest math class**
AK	**Foreign**	**Dummy variable = 1 if student passed foreign language proficiency test**
AL	**EMEcon**	**Dummy variable = 1 if expected major is economics**
AM	**EMOSS**	**Dummy variable = 1 if expected major is another social science**
AN	**EMNS**	**Dummy variable = 1 if expected major is a natural science**
AO	**EMH**	**Dummy variable = 1 if expected major is in the humanities**
AP	**AM1**	**Dummy variable = 1 if student completed 1 year of advanced math in high school**

AQ	AM2	Dummy variable = 1 if student completed 2 years of advanced math in high school
AR	AM3	Dummy variable = 1 if student completed 3 years of advanced math in high school
AS	Phy1	Dummy variable = 1 if student completed 1 course in physics in high school
AT	Phy2	Dummy variable = 1 if student completed 2 courses in physics in high school
AU	Chem1	Dummy variable = 1 if student completed 1 course in chemistry in high school
AV	Chem2	Dummy variable = 1 if student completed 2 courses in chemistry in high school

Table 2.30. Definition of the variables included in the Vanderbilt data set.

Replication of the Ordered Probit Regression

At this point we are ready to begin the replication. Since it is easy to get lost in the process, I have created a list of steps that include both instructions on what to do and questions you need to answer. As part of this exercise you will be asked to complete several tables of results. In order to make this effort easier, I have provided a MS Word file, Tables for ordered probit discussion.doc, with the tables to be completed in it.

1. Load the data in *Stata* from *Excel*.

2. Convert MSAT and VSAT to MSAT/100 and VSAT/100, respectively, using the commands:

```
.replace msat = msat/100

.replace vsat = vsat/100
```

3. Common sense dictates that we should calculate the means and standard deviations of the variables to be sure that there are no entry errors. We need to construct a table that compares the means and standard deviations reported in BFS with those in our dataset. Table 2, which has the means and standard deviations reported by BFS, gives a place to put the means and standard deviations for the variables in our dataset. Fill in the information missing from Table 2.

	Our data		Butler, et al.	
Variable	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**
msat			**6.25**	**0.60**
foreign			**0.11**	**0.32**
female			**0.39**	**0.49**
emecon			**0.34**	**0.48**
emoss			**0.17**	**0.38**
emns			**0.21**	**0.41**
emh			**0.07**	**0.25**
am1			**0.49**	**0.50**
am2			**0.45**	**0.50**
am3			**0.01**	**0.11**
phy1			**0.67**	**0.47**
Phy2			**0.02**	**0.14**
chem1			**0.82**	**0.39**
chem2			**0.12**	**0.32**

Table 2.31. Means and standard deviations of the data.

4. Estimate the ordered probit regression using (in *Stata*) the commands:

```
.global indvar msat foreign female emecon emoss emns emh am1 am2 am3 phy1 phy2 chem1 chem2

.oprobit highestmath $indvar
```

5. Use the result of this estimation to complete Table 3.[37]

highestmath	Coef.	Std. Err.	z	P>z	[95% Conf. Interval]	
msat1						
foreign						
female						
emecon						
emoss						
emns						
emh						
am1						
am2						
am3						
phy1						
Phy2						
chem1						
chem2						

_cut1						
_cut2						
_cut3						
_cut4						
_cut5						
_cut6						
Observations						
Log likelihood						
LR χ^2(14)						
Prob > χ^2						
Pueudo-R^2						

Table 2.32. Results of *Stata* ordered-probit regression.

6. Compare your results with the table reported in the article. The table in the article is Table II on page 193 and is reproduced in Figure 3. What we are interested in is comparing column 4 in Figure 3 with columns 2 and 4 in Table 3. Table 4 below offers a model for this comparison.

Figure 2.32.

Table II. Ordered probit estimates of level of calculus attained[a]

		Students taking MICRO-2		Students taking MACRO-2	
Variable[b]	Expected sign	Mean (SD)	Coefficient (t-value)	Mean (SD)	Coefficient (t-value)
Constant		—	−3·09 (5·48)	—	−2·62 (3·95)
SAT-math × 10^{-2}	+	6·25 (0·60)	0·50[d] (6·12)	6·25 (0·60)	0·48[d] (5·23)
Foreign lang. proficiency [1,0]	−	0·11 (0·32)	0·02 (0·14)	0·09 (0·29)	0·23 (1·22)
Sex (female = 1; male = 0)	?	0·39 (0·49)	0·25[d] (2·59)	0·36 (0·48)	0·22[e] (1·96)
Expected major:					
Economics	?	0·34 (0·48)	−0·11 (0·86)	0·36 (0·48)	−0·18 (1·31)
Other social science	?	0·17 (0·38)	−0·29[e] (1·99)	0·15 (0·36)	−0·27 (1·59)
Natural science	+	0·21 (0·41)	0·43[d] (3·10)	0·20 (0·40)	0·32[e] (2·05)
Humanities	−	0·07 (0·25)	−0·37[e] (1·78)	0·07 (0·26)	−0·39[e] (1·80)
Years of HS Advanced Math (Y_m)					
$1 \leq Y_m < 2$	+	0·49 (0·50)	0·24 (1·07)	0·49 (0·50)	−0·00 (0·02)
$Y_m = 2$	+	0·45 (0·50)	0·93[d] (4·04)	0·45 (0·50)	0·67[d] (2·83)
$Y_m > 2$	+	0·01 (0·11)	0·77[e] (1·70)	0·01 (0·11)	0·28 (0·55)
Years of HS physics (Y_p)					
$1 \leq Y_p < 2$	+	0·67 (0·47)	0·26[d] (2·71)	0·67 (0·47)	0·27[d] (2·50)
$Y_p \geq 2$	+	0·02 (0·14)	0·38 (1·07)	0·01 (0·11)	−0·11 (0·20)
Years of HS chemistry (Y_c)					
$1 \leq Y_c < 2$	+	0·82 (0·39)	−0·12 (0·69)	0·82 (0·39)	−0·18 (0·75)
$Y_c \geq 2$	+	0·12 (0·32)	0·17 (0·75)	0·13 (0·34)	0·20 (0·75)
TRUNCATION POINTS[c]					
(1)	+		0·27[d] (7·29)		0·21[d] (5·59)
(2)	+		0·33[d] (8·16)		0·27[d] (6·46)
(3)	+		1·52[d] (20·32)		1·55[d] (18·26)
(4)	+		1·79[d] (23·07)		1·88[d] (20·73)
(5)	+		2·04[d] (23·72)		2·15[d] (20·58)
OVERALL RESULTS					
Log likelihood			−886·67		−698·09
Outcomes predicted correctly			37·9%		41·2%
Number of Observations			609		490

[a] The dependent variable is the level of calculus attained, as shown by the ordered probit ranking in the lower panel of Table I.

[b] Omitted reference groups: other or unstated expected major; less than one year advanced math, physics, and chemistry in high school.

[c] In an ordered probit, an underlying, normally distributed, latent variable has a mean which is a function of observable variables. The latent variable gives rise to a set of observed dummy variables for ordered categories based on ranges

Results of ordered probit regression as reported in Butler, et al.

Table 4. Comparison of ordered probit estimations.

	Our estimates		Butler, et al. estimates	
	Estimate	z	Estimate	t-value
msat1			0.05	6.12
foreign			0.02	0.14
female			0.25	2.59
emecon			-0.11	0.86
emoss			-0.29	1.99
emns			0.43	3.10
emh			-0.37	1.78
am1			0.24	1.07
am2			0.93	4.04
am3			0.77	1.70
phy1			0.26	2.71
Phy2			0.38	1.07
chem1			-0.12	0.69
chem2			0.17	0.75

Intercept			**-3.09**	**5.48**
_cut1			**0.27**	**7.29**
_cut2			**0.33**	**8.16**
_cut3			**1.52**	**20.32**
_cut4			**1.79**	**23.07**
_cut5			**2.04**	**23.72**
_cut6				

Table 2.33. Comparison of ordered-probit estimations.

7. It is easy to see from Table 4 is that almost without exception the estimates of the parameters and their t-ratios are very similar. The exception arises with the estimates of the truncation points (_cut# in the *Stata* results). We will have to figure out what these are estimates of in order to make sense of them. Figure 1 shows the "cutoffs" that are being estimated. Footnote c in the BFS Table II on page 193 (shown in Figure 3) offers a useful observation:

> ***In an ordered probit, an underlying, normally distributed, latent variable has a mean which is a function of observable variables. The latent variable gives rise to a set of observed dummy variables for ordered categories based on ranges between unobserved but estimable truncation points which correspond to levels of effort, ability, or other factors reflected in the explanatory variables. If L categories are observed, there are L − 1 truncation points, of which the first is normalized to be zero, so that L − 2 truncation points are estimated and reported in the table. The values correspond to standard deviations of the latent normally distributed variable.***

The key idea is that the values of cutoffs are relative and can be normalized around any value. Notice that the *Stata* results do not report an intercept term but do report six cutoff values. Moreover, the difference between the estimate by *Stata* for the first cutoff (3.08402) and the estimate for the second cutoff (3.356916) is equal to 0.272896, which is itself equal to the first truncation point reported by BFS (1998: 193). Use Table 5 to report the difference between the first cutoff value and each of the cutoff points reported by *Stata*.

Cutoff	Estimate	Estimate - _cut1	BFS Truncation Points
_cut1	3.0840		
_cut2	3.3569		0.27
_cut3	3.4146		0.33
_cut4	4.6013		1.52
_cut5	4.8774		1.79
_cut6	5.1202		2.04

Table 2.34. Reconciling *Stata* estimates of cutoff points with Butler, et al.'s truncation points.

The second part of the reconciliation of the two sets of results is to compute the t-ratios. To do this we need to compute the standard deviation of the estimates of the cutoff points reported by *Stata*. To do this we need to retrieve the variance-covariance matrix from the regression. First, let's see what we are interested in computing. Let $\hat{\beta}_i$ be the estimate of the *i* th cutoff point. In column 3 of Table 5 you computed $\hat{\alpha}_i = \hat{\beta}_i - \hat{\beta}_1$ for *i* = 2,…,6 . The variance of the new variable is:

(2.92)

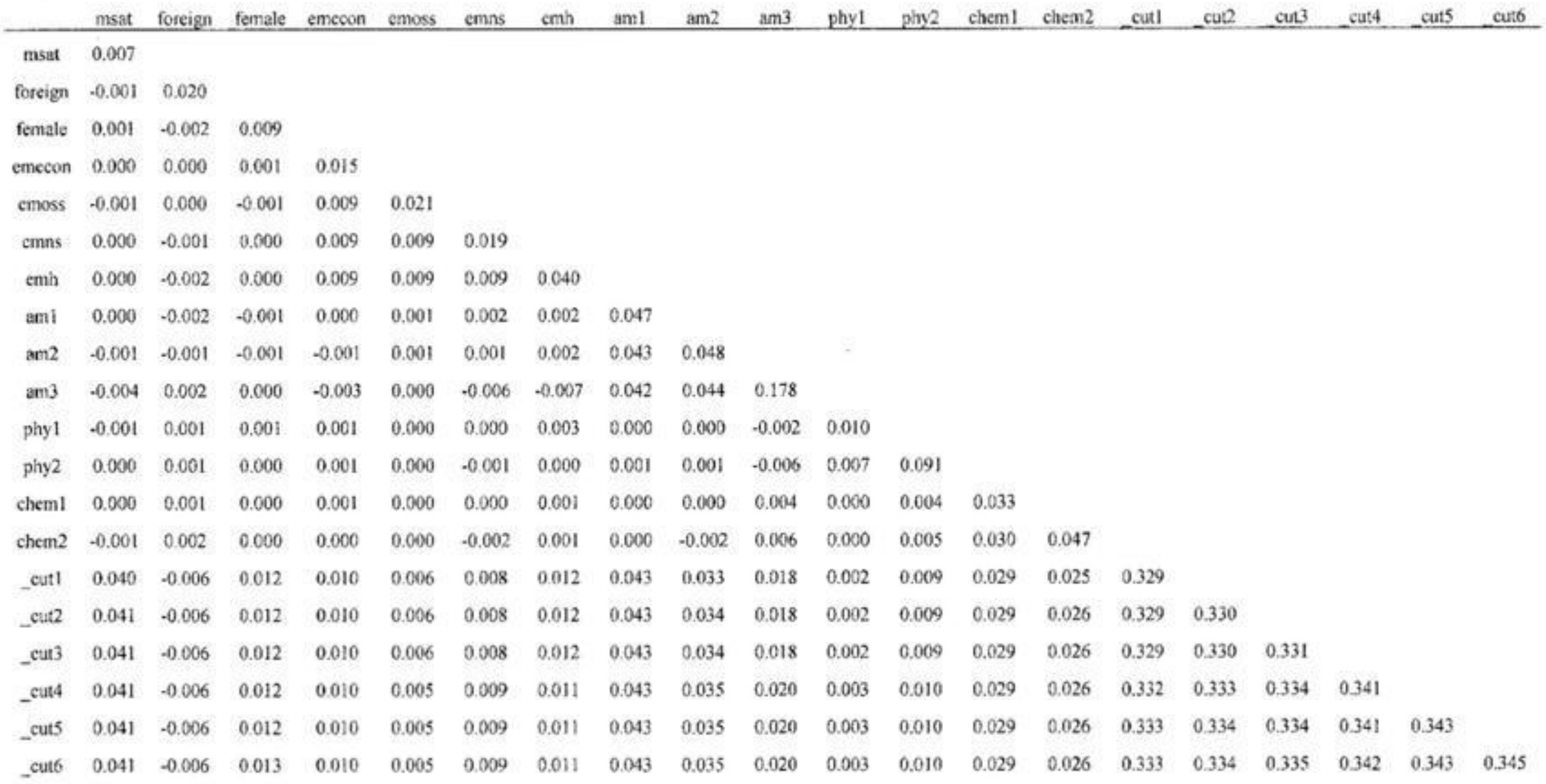
$$V\left(\hat{\alpha}_i\right) = V\left(\hat{\beta}_i\right) - 2Cov\left(\hat{\beta}_i \hat{\beta}_1\right) + V\left(\hat{\beta}_1\right) = \sigma_i^2 - 2\sigma_{i1} + \sigma_1^2$$

The variance-covariance matrix will give us estimates of these variances and covariances. When there are *j* parameters in a regression equation, this matrix is defined to be:

$$\hat{\Sigma} = \begin{bmatrix} \hat{\sigma}^2_{\beta_1} & \hat{\sigma}_{\beta_1\beta_2} & \cdots & \sigma_{\beta_1\beta_k} \\ \hat{\sigma}_{\beta_2\beta_1} & \hat{\sigma}^2_{\beta_2} & \cdots & \hat{\sigma}_{\beta_2\beta_k} \\ \vdots & \vdots & \ddots & \\ \sigma_{\beta_k\beta_1} & \sigma_{\beta_k\beta_2} & \cdots & \hat{\sigma}^2_{\beta_k} \end{bmatrix}.$$

If you type the command .vce, *Stata* will report $\hat{\Sigma}$ as shown in Figure 4. We need the section of this matrix shown in Part A of Table 6. Use equation (5) to estimate the standard errors of the estimates of the cutoff points and complete Part B of Table 6 and compares the t-ratios with the values reported by Butler, et al. (and shown in the last column 4 of Table 6). Are you satisfied that we have been able to come reasonably close to the results reported in the article?

Figure 2.33.



	msat	foreign	female	emecon	emoss	emns	emh	am1	am2	am3	phy1	phy2	chem1	chem2	_cut1	_cut2	_cut3	_cut4	_cut5	_cut6
msat	0.007																			
foreign	-0.001	0.020																		
female	0.001	-0.002	0.009																	
emecon	0.000	0.000	0.001	0.015																
emoss	-0.001	0.000	-0.001	0.009	0.021															
emns	0.000	-0.001	0.000	0.009	0.009	0.019														
emh	0.000	-0.002	0.000	0.009	0.009	0.009	0.040													
am1	0.000	-0.002	-0.001	0.000	0.001	0.002	0.002	0.047												
am2	-0.001	-0.001	-0.001	-0.001	0.001	0.001	0.002	0.043	0.048											
am3	-0.004	0.002	0.000	-0.003	0.000	-0.006	-0.007	0.042	0.044	0.178										
phy1	-0.001	0.001	0.001	0.001	0.000	0.000	0.003	0.000	0.000	-0.002	0.010									
phy2	0.000	0.001	0.000	0.001	0.000	-0.001	0.000	0.001	0.001	-0.006	0.007	0.091								
chem1	0.000	0.001	0.000	0.001	0.000	0.000	0.001	0.000	0.000	0.004	0.000	0.004	0.033							
chem2	-0.001	0.002	0.000	0.000	0.000	-0.002	0.001	0.000	-0.002	0.006	0.000	0.005	0.030	0.047						
_cut1	0.040	-0.006	0.012	0.010	0.006	0.008	0.012	0.043	0.033	0.018	0.002	0.009	0.029	0.025	0.329					
_cut2	0.041	-0.006	0.012	0.010	0.006	0.008	0.012	0.043	0.034	0.018	0.002	0.009	0.029	0.026	0.329	0.330				
_cut3	0.041	-0.006	0.012	0.010	0.006	0.008	0.012	0.043	0.034	0.018	0.002	0.009	0.029	0.026	0.329	0.330	0.331			
_cut4	0.041	-0.006	0.012	0.010	0.005	0.009	0.011	0.043	0.035	0.020	0.003	0.010	0.029	0.026	0.332	0.333	0.334	0.341		
_cut5	0.041	-0.006	0.012	0.010	0.005	0.009	0.011	0.043	0.035	0.020	0.003	0.010	0.029	0.026	0.333	0.334	0.334	0.341	0.343	
_cut6	0.041	-0.006	0.013	0.010	0.005	0.009	0.011	0.043	0.035	0.020	0.003	0.010	0.029	0.026	0.333	0.334	0.335	0.342	0.343	0.345

***Stata* estimate of the variance-covariance matrix.**

Part A. Relevant portion of the variance-covariance matrix.						
	_cut1	**_cut2**	**_cut3**	**_cut4**	**_cut5**	**_cut6**
_cut1	**0.329**					

_cut2	0.329	0.330				
_cut3	0.329	0.330	0.331			
_cut4	0.332	0.333	0.334	0.341		
_cut5	0.333	0.334	0.334	0.341	0.343	
_cut6	0.333	0.334	0.335	0.342	0.343	0.345
Part B. Calculation of the t-ratios (with comparison of values reported in BFS)						
	V($\hat{\beta}$)	St. Dev.($\hat{\beta}$	t-ratio	BFS t-ratio		
_cut2				7.29		
_cut3				8.16		
_cut4				20.32		
_cut5				23.07		
_cut6				23.72		

Table 2.35. Calculation of the t-ratios for the cutoff estimates.

8. The next step in the process is to generate the term we will use in the estimation of the grade regression to account for the potential sample selection bias. To do this we will need to find a reference in the literature that offers a clear description of what we need to do. As it turns out, a reasonable explanation of the appropriate estimation technique is available in Jimenez and Kugler (1987). Since much of what follows comes directly from this article, I highly recommend you read it yourself.

The gist of the method suggests that the potential sample bias is accounted for by an inverse Mills ratio for each of the categories. What we need to do is calculate:

(2.93)

$$\hat{\lambda}_i = \frac{\phi\left(\hat{\mu}_j - \hat{z}_i^*\right) - \phi\left(\hat{\mu}_{j+1} - \hat{z}_i^*\right)}{\Phi\left(\hat{\mu}_{j+1} - \hat{z}_i^*\right) - \Phi\left(\hat{\mu}_j - \hat{z}_i^*\right)}$$

for the category that the individual actually is in. What we will do is calculate (6) for all of the categories and then sum the product of this number and a dummy variable indicating if a course is the highest math class completed by an individual. Since the dummy variables will equal 0 for math categories an individual is not in, the resulting sum will preserve the value of (6) that is associated with the category the individual does belong to.

It is clear from (6) that we will need to retain the 6 cutoffs. We can do this with the commands:

```
. generate cutoff1 = _b[_cut1]

. generate cutoff2 = _b[_cut2]

. generate cutoff3 = _b[_cut3]

. generate cutoff4 = _b[_cut4]

. generate cutoff5 = _b[_cut5]

. generate cutoff6 = _b[_cut6]
```

Technically, this step is not necessary since the parameter estimates are preserved until the next regression is estimated; I suggest doing this purely as a precaution.

9. Preserve the predicted values of the ordered-probit using the command:

```
. predict zhat, xb

. predict phat1 phat2 phat3 phat4 phat5 phat6 phat7, p
```

These two commands will generate for each observation the predicted mean category of math classes and the probability that this individual will fall in each category. To see what is going on we will retrieve some representative values of these variables and then

graph them for one individual. Table 7 reports these values for 10 individuals in the sample. Now consider individual 2. Fitting a normal distribution with a mean of 4.25 and using the critical values from our estimation yields the probabilities that the individual is in each of the categories. For example, the probability that individual 1 will have completed no math classes is equal to 0.1223. Figure 5 illustrates the results for individual 1. The dashed vertical lines are the six cutoff values that are the same for each individual. The solid vertical line is the zhat for individual 1. The heavy blue line represents the normal probability density function for this individual. While, there is, of course, a different probability distribution for each individual, the cutoff values are the same for all members of the sample.

Observation	Highest Math Class	zhat	Pr(0)	Pr(1)	Pr(2)	Pr(3)	Pr(4)	Pr(5)	Pr(6)
1	3	3.9657	0.1890	0.0824	0.0194	0.4467	0.0816	0.0568	0.1241
2	0	4.2507	0.1217	0.0640	0.0158	0.4355	0.0975	0.0731	0.1923
165	0	3.5982	0.3036	0.1011	0.0225	0.4149	0.0575	0.0364	0.0640
166	6	4.6914	0.0540	0.0370	0.0098	0.3633	0.1097	0.0922	0.3340
214	3	3.4533	0.3560	0.1056	0.0229	0.3900	0.0483	0.0294	0.0478
215	3	4.0840	0.1587	0.0749	0.0180	0.4459	0.0887	0.0637	0.1501
225	3	3.5250	0.3296	0.1036	0.0228	0.4031	0.0528	0.0328	0.0553
226	3	3.6990	0.2693	0.0969	0.0219	0.4285	0.0641	0.0417	0.0776
453	3	3.9713	0.1875	0.0820	0.0194	0.4468	0.0819	0.0571	0.1253
454	5	4.1650	0.1399	0.0697	0.0170	0.4422	0.0932	0.0684	0.1697
495	3	4.4168	0.0913	0.0533	0.0135	0.4151	0.1043	0.0816	0.2409
496	0	2.9811	0.5410	0.1055	0.0212	0.2797	0.0236	0.0127	0.0162
526	0	2.9247	0.5633	0.1039	0.0207	0.2653	0.0214	0.0114	0.0141

527	3	3.9757	0.1863	0.0817	0.0193	0.4469	0.0822	0.0574	0.1262

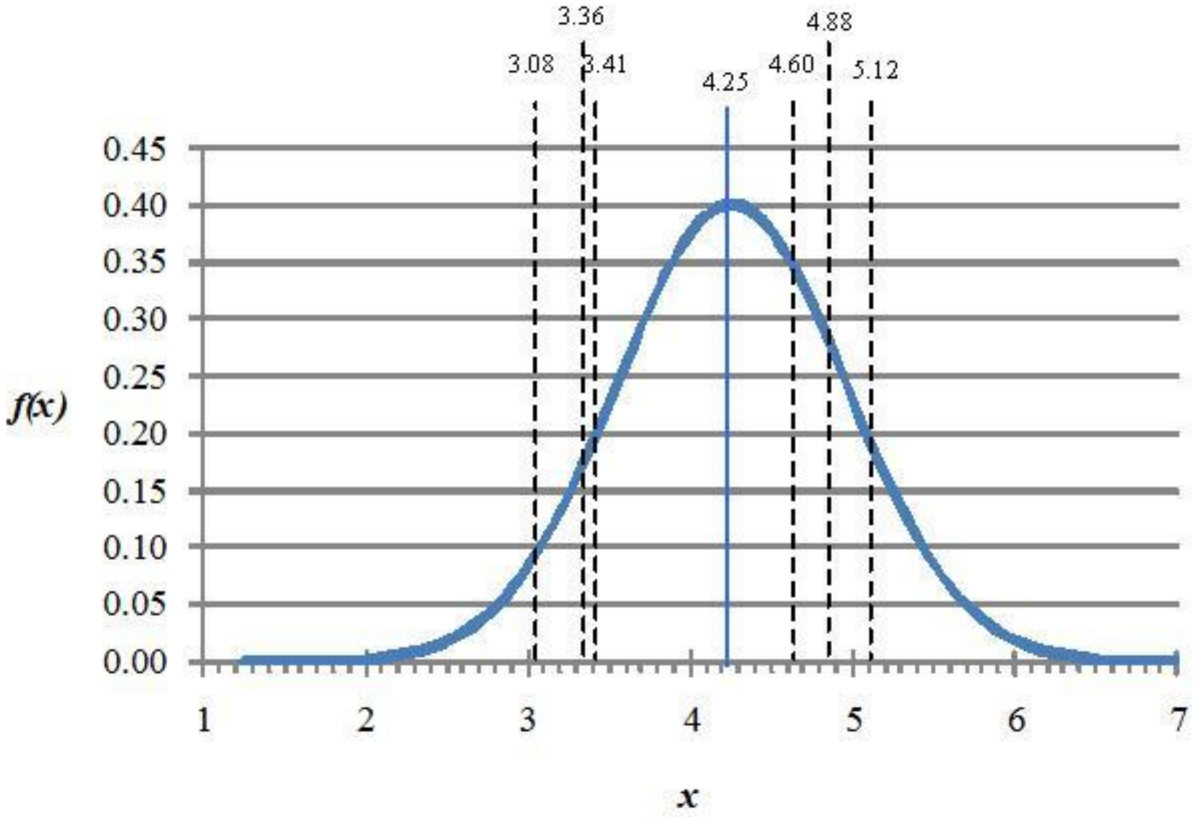
Table 2.36. Predicted values of the ordered probit regression.

Now we are ready to calculate (6). The commands are:

```
.generate lambda0 = (-normden(cutoff1-zhat))/(norm(cutoff1-zhat)-norm(-zhat))

.generate lambda1 = (normden(cutoff1-zhat)-normden(cutoff2-zhat))/(norm(cutoff2-zhat)-norm(cutoff1-zhat))

.generate lambda2 = (normden(cutoff2-zhat)-normden(cutoff3-zhat))/(norm(cutoff3-zhat)-norm(cutoff2-zhat))

.generate lambda3 = (normden(cutoff3-zhat)-normden(cutoff4-zhat))/(norm(cutoff4-zhat)-norm(cutoff3-zhat))

.generate lambda4 = (normden(cutoff4-zhat)-normden(cutoff5-zhat))/(norm(cutoff5-zhat)-norm(cutoff4-zhat))

.generate lambda5 = (normden(cutoff5-zhat)-normden(cutoff6-zhat))/(norm(cutoff6-zhat)-norm(cutoff5-zhat))

.generate lambda6 = (normden(cutoff6-zhat))/(1-norm(cutoff6)-norm(cutoff5-zhat))

.generate lambda = m170*lambda0 + m171a*lambda1 + m172a*lambda2 + m171b*lambda3 + m172b*lambda4 +
m221a*lambda5+m221b*lambda6
```

One thing to notice in these calculations is that cutoff0 is assumed to be − ∞ and cutoff7 is assumed to be ∞.

Figure 2.34.



The probability distribution of math class category for individual 2.

10. Now we are ready to estimate our regression explaining the grade that each individual received in intermediate microeconomics. Use Table 8 to report the regression results for four specifications of the model. The first question is can the null hypothesis of sample selection bias be rejected? How does this conclusion compare with BFS's conclusions? (See Table 9.) Second, since many of the potential explanatory variables like class size and scores on the SATs do not seem to be statistically significant, it is reasonable to focus our comments on the results reported in column (4) of Table 8.

What can you conclude about the impact of calculus on how well a student will do in intermediate microeconomics? Do the final grades earned in a majority of the math classes impact the grade earned in intermediate microeconomics? Do the grades earned in any of the math classes positively and significantly affect the grade earned in intermediate microeconomics? Can you explain the impact of the freshman GPA on the grade earned in intermediate microeconomics? What, if any, is your bottom line conclusions about what matters in determining the grades earned in intermediate microeconomics?

Explanatory variables	Model (1)	Model (2)	Model (3)	Model (4)
Lambda			—	—
Sophomore		—		—
Senior		—		—
Same				
Skip		—		—
M171a				
M172a				
M171b				

M172b				
M221a				
M221b				
GE100				
GDE100				
GE101				
GDE101				—
GDE231				
Size				—
FGPA				

Female				
MSAT				—
VSAT				—
Grade in highest Math	—		—	—
class				
GM170		—		
GM171a		—		
GM172a		—		
GM171b		—		
GM172b		—		
GM221a		—		

GM221b		—		
Intercept				
F(28, 580)		—	—	—
Prob > F		—	—	—
F(27, 581)	—	—		—
Prob > F	—	—		—
F(20, 588)		—	—	
Prob > F		—	—	
F(19, 589)	—		—	—
Prob > F	—		—	—
R-Squared				
Root MSE				
Sample Size	609	609	609	609

Table 2.37. Determinants of Final Grade in Intermediate Microeconomics.

Robust t-ratios are in parentheses.

		MICRO-2	
Variable[a]	Expected sign	Mean (SD)	Coefficient(t-value)

Intercept	—	—	-1.64
			(3.48)
Selection bias correction	+	-0.00	0.10
(Predicted residual)		(0.92)	(1.29)
Level of calculus attained:			
Math 171A	+	0.08	0.39
		(0.27)	(1.04)
Math 172A	+	0.02	-0.18
		(0.13)	(0.21)
Math 171B	+	0.37	1.02[b]
		(0.48)	(3.49)
Math 172B	+	0.07	1.52 [b]
		(0.25)	(3.53)
Math 221A	+	0.05	1.33[c]
		(0.22)	(2.27)
Math 221B or 222	+	0.14)	0.75[c]
		(0.35	(1.67)
Grade in last calculus course:			
Math 170	+	3.06	0.36[b]
		(0.70)	(4.36)

Math 171A	**+**	**2.22**	**0.26**[c]
		(0.86)	**(2.21)**
Math 172A	**+**	**2.94**	**0.42**
		(0.80)	**(1.54)**
Math 171B	**+**	**2.62**	**0.10**[c]
		(0.93)	**(1.85)**
Math 172B	**+**	**2.63**	**-0.01**
		(0.90)	**(0.10)**
Math 221A	**+**	**3.10**	**-0.09**
		(0.77)	**(0.55)**
Math 221B or 222	**+**	**3.15**	**0.11**
		(0.76)	**(1.04)**
Grade deflator of instructor in intermediate theory	**+**	**-0.16**	**0.88b**
course		**(0.27)**	**(8.28)**
Taken in Sophomore year	**?**	**0.32**	**0.07**
		(0.47)	**(0.94)**
Taken in Senior year	**-**	**0.06**	**-0.02**
		(0.24)	**(0.13)**
MICRO-1 and MICRO-2 in same academic year	**+**	**0.35**	**0.04**
		(0.48)	**(0.46)**

At least one semester between MICRO-1 and	-	0.27	0.13
MICRO-2		(0.44)	(1.85)
Grade in MACRO-1	+	2.73	0.20[b]
		(0.73)	(3.93)
Grade in MICRO-1	+	2.67	0.29[b]
		(0.74)	(5.93)
Instructor's grade deflator:			
MACRO-1	-	-0.32	-0.33[c]
		(0.20)	(2.20)
MICRO-1	-	-0.29	-0.11
		(0.16)	(0.53)
Class size (intermediate theory course)	?	28.2	-0.002
		(5.5)	(0.45)
Freshman Grade Point Average	+	2.79	0.29[b]
		(0.46)	(3.04)
Sex (female = 1; male = 0)	?	0.39	0.13[c]
		(0.49)	(2.09)
SAT-Math score x 10^{-2}	+	6.25	0.12[c]
		(0.60)	(1.75)

SAT-Verbal score x 10^{-2}	+	5.56	0.04
		(0.67)	(0.78)
OVERALL RESULTS			
Mean (SD) of dependent variable			
Adjusted R^2		0.44	
Number of observations		609	

Table 2.38. Results reported in BFS (p. 195).

[a] Omitted reference groups in MICRO-2 regression: attained Math 170; took MICRO-2 in Junior year; took MICRO-1 in spring, MICRO-2 next fall. [b] Significant at 0.01 level, one- or two-tailed test as appropriate. [c] Significant at 0.05 level, one- or two-tailed test as appropriate.

Exercises

Exercise 2.6.1.

Quite often health professionals request that a patient a report their perception of their health status on a scale of 0 to 10, where 0 is the lowest possible health status and 10 is the highest health status. This type of data set is best analyzed using ordered probit. In this exercise you will analyze a data set of responses to a survey made in Germany between 1984 and 1995. The question we are interested in analyzing is the respondent's perception of their own health status.

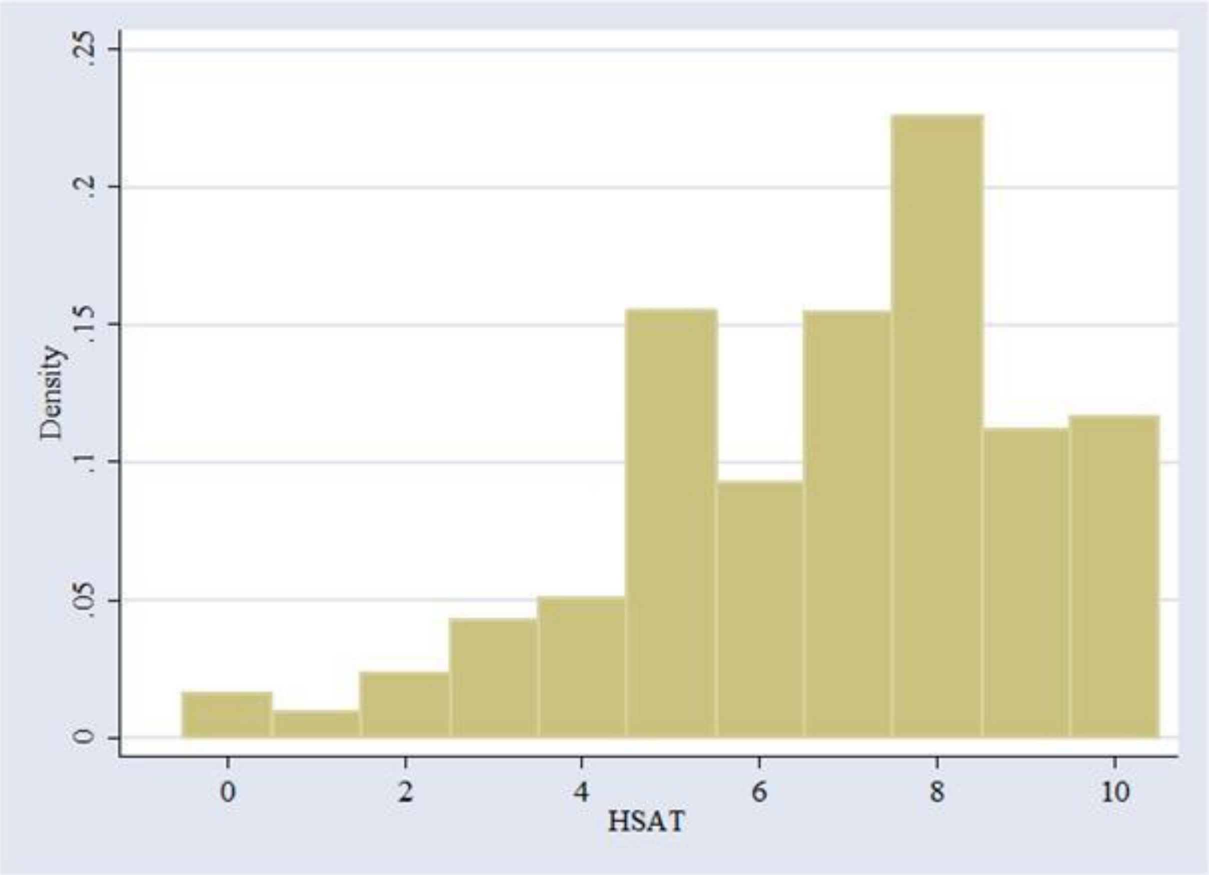
The file Riphahn, Wambach, Million data.xls is an MS Excel file that contains 27,326 observations on 25 variables, one observation per line. The data are from Riphahn, Wambach, and Million (2003) and are also available on the web. The variables are defined in Table 10. As a first step you will need to load these data into Stata. However, due to the large sample size you will need to first expand the size of the memory that is available to Stata with the command: . set memory 1G. Here I have increased the memory to 1 gigabyte. This amount may be overkill but it seemed to be big enough on my computer to handle the data.

Column	Variable	Variable definition
A	ID	individual's ID number
B	Female	female = 1; male = 0
C	Year	calendar year of the observation
D	Age	age in years
E	HSAT	health satisfaction, coded 0 (low) - 10 (high)
F	Handdum	handicapped = 1; otherwise = 0
G	Handper	degree of handicap in percent (0 - 100)
H	HhnINC	household nominal monthly net income in German marks / 1000
I	HHKIDS	children under age 16 in the household = 1; otherwise = 0
J	Educ	years of schooling
K	Married	married = 1; otherwise = 0
L	Haupts	highest schooling degree is Hauptschul degree = 1; otherwise = 0
M	Reals	highest schooling degree is Realschul degree = 1; otherwise = 0
N	FachHS	highest schooling degree is Polytechnical degree = 1; otherwise = 0
O	Abitur	highest schooling degree is Abitur = 1; otherwise = 0
P	Univ	highest schooling degree is university degree = 1; otherwise = 0
Q	Working	employed = 1; otherwise = 0
R	BlueC	blue collar employee = 1; otherwise = 0
S	WhiteC	white collar employee = 1; otherwise = 0

T	Self	self employed = 1; otherwise = 0
U	Beamt	civil servant = 1; otherwise = 0
V	DocVis	number of doctor visits in last three months
W	HospVis	number of hospital visits in last calendar year
X	Public	insured in public health insurance = 1; otherwise = 0
Y	Addon	insured by add-on insurance = 1; otherwise = 0

Table 2.39. Variables in the German Socioeconomic Panel Data Set.

Figure 2.35.



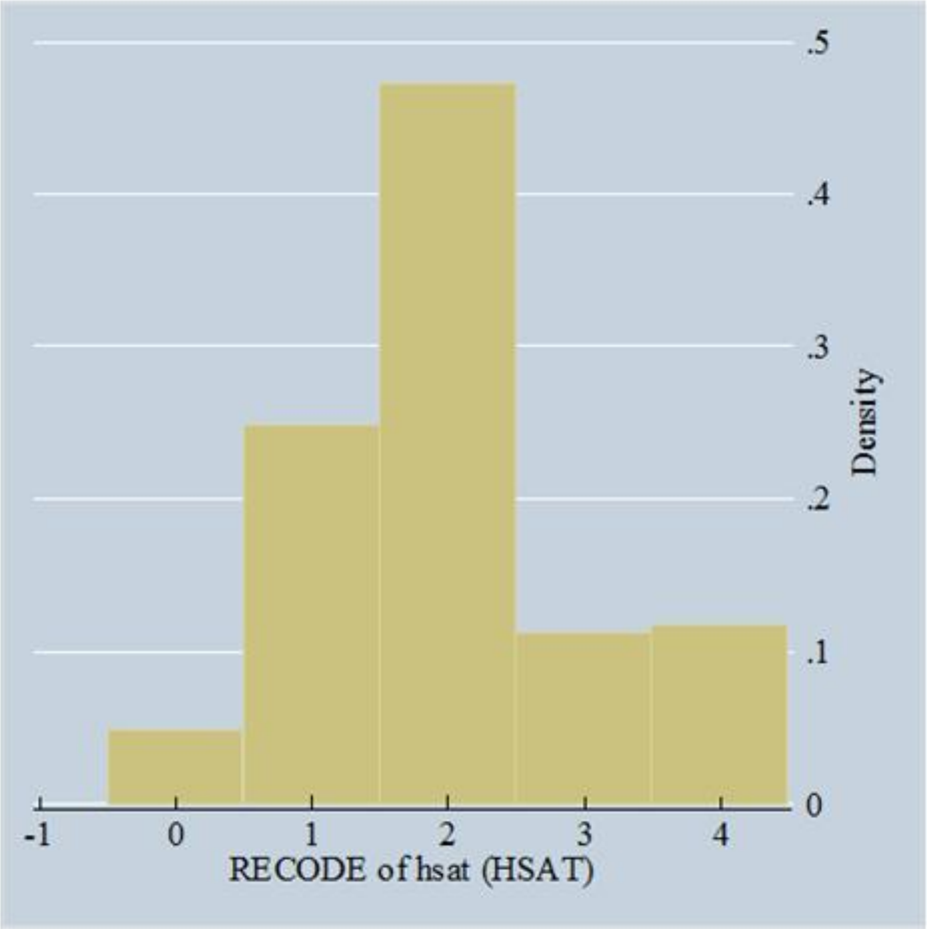
Distribution of responses on health status.

One of the major problems with survey indices is that the numbers seem to mean different things to respondents. One way to reduce this problem is to collapse the index into fewer outcomes by combining some of the responses together. However, anyway we do this is going to be ad hoc. Figure 6 shows a histogram of the responses to this question. Based on this graph, we will create 5 categories—(0) HSat = 0, 1, or 2; (1) HSat = 3, 4 or 5; (2) HSat = 6, 7, or 8; (3) HSat = 9; and (4) HSat = 10. We can create a new categorical variable called hsatnew with the command:

```
. recode hsat (0/2 = 0) (3/5 = 1) (6/8 = 2) (9 = 3) (10 = 4), generate(hsatnew)
```

Figure 7 shows the histogram of the new variable.

Figure 2.36.



The collapsed distribution of health status responses.

1. **Create a table of summary statistics for (1) health status, (2) age, (3) household income, (4) years of education, (5) marital status, and (6) number of children by year and sex. (You might want to use the command .bysort year female, list of variables).**

2. **Estimate an ordered probit regression for 1988 for health status (the new variable) using age, income, education, married, and kids as the explanatory variables. Here you might want to used the command: .oprobit hsatnew age hninc educ married hhkids if year==1988.**

3. **Use the predict newvariable, xb command to calculate the predicted mean values for each individual for the 1988 observations. Compare this histogram to one using the 1988 regression parameters to estimate xb for all years.**

4. **Estimate the ordered probit model for all of the years in the sample and put the results into a table like Table 11. (Here you might want to make use of the command: .bysort year: oprobit hsatnew varlist)**

Variable	1984	1985	1986	1987	1988	1991	1994
age							
income							
education							
married							
kids							
_cut1							
_cut2							
_cut3							
_cut4							
Observations							
LR χ^2(5)							

Prob > χ^2							
Log likelihood							
Pseudo-R^2							

Table 2.40. Sample table for part (d) of Exercise 1.

t-ratios are in parentheses.

References

Amemiya, T. (1985). *Advanced Econometrics* (Cambridge, MA: Harvard University Press).

Bourguignon, François, Martin Fournier, and Marc Gurgand (2007). Selection bias corrections based on the multinomial logit model: Monte Carlo comparisons. *Journal of Economic Surveys* 21(1): 174-205.

Butler, J. S., T. Aldrich Finegan, and John J. Siegfried (1998). Does more calculus improve student learning in intermediate micro- and macroeconomic theory?" *Journal of Applied Econometrics* 13(2): 185-202.

Chiburis, Richard and Michael Lokshin (2007). Maximum likelihood and two–step estimation of an ordered–probit selection model. *The Stata Journal* 7(2): 167-182.

Dahl, Gordon B. (2002). Mobility and the returns to education: testing a roy model with multiple markets. *Econometrica* 70(6): 2367–2420.

Dubin, Jeffrey A. and Daniel L. McFadden (1984). An econometric analysis of residential electric appliance holdings and consumption. *Econometrica* 52(2): 345–362.

Greene, William H. (1990). *Econometric Analysis* (New York: Macmillan Publishing Company).

Heckman, James J. (1979). Sample selection bias as a specification error. *Econometrica* 47(1): 153–161.

Jimenez, Emmanuel and Bernardo Kugler (1987). The earnings impact of training duration in a developing country an ordered probit selection model of Colombia's *Servicio Nacional de Aprendizaje* (SENA). *Journal of Human Resources* 22(2): 230-233.

Lee, Lung-Fei (1983). Generalized econometric models with selectivity. *Econometrica* 51(2): 507–512.

Maddala, G. S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics* (Cambridge: Cambridge University Press).

Main, B. and B. Reilly (1993). The employer size-wage gap: Evidence for Britain. *Economica* 60: 125–142.

McFadden, Daniel L. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (ed.) *Frontiers in Econometrics* (New York: Academic Press).

Newey, W. K. and Daniel L. McFadden (1994). Large sample estimation and hypothesis testing. In R. F. Engle and Daniel L. McFadden (eds.) *Handbook of Econometrics,* vol. IV (Amsterdam: North Holland).

Riphahn, Regina T., Achim Wambach, and Andreas Million (2003). Incentive effects in the demand for health care: a bivariate panel count data estimation. *Journal of Applied Econometrics* 18(4): 387-405

Schmertmann, Carl P. (1994). Selectivity bias correction methods in polychotomous sample selection models. *Journal of Econometrics* 60(1): 101–132.

Vella, Francis (1998). Estimating models with sample selection bias. *The Journal of Human Resources* 33(1): 127-169.

[7] J. S. Cramer (2003) *Logit Models from Economics and Other Fields* (Cambridge: Cambridge University Press): 10.

[8] For a full discussion of this model see Ladd, G. W. (1966) “Linear Probability Functions and Discriminant Functions,” *Econometrica* 34: 873-888.

[9] The assumption that the variance is equal to 1 is due to technical considerations. See [Cramer, 22].

[10] The pdf of a logistic distribution is $f(x) = \frac{\lambda e^{-\lambda x}}{\left(1 + e^{-\lambda x}\right)^2}$, where $\lambda = \frac{\pi}{\sqrt{3}} \approx 1.814$. See Cramer, 24-26 for a fuller discussion of the logistic distribution.

[11] See Stata Library, Categorical and Count Data Analysis Utilities for useful utilities and an excellent discussion of how to interpret categorical and count regression results at http://www.ats.ucla.edu/stat/stata/library/longutil.htm/ (accessed July 19, 2009).

[12] The phrase "(Assumption: . nested in full)" tells you the name of the regression is the unrestricted model (full) and offers you a hyperlink to call this regression up to the screen.

[13] The gradient is a vector of first-derivatives. In this case it is a vector of the first-derivatives with respect to each parameter estimate $\left(\text{i.e., } \hat{\beta}_i\right)$. To obtain the ML estimate, we have to set these first-derivatives equal to zero.

[14] See StataCorp [2003:119-130] for more detail on this command.

[15] If the OLS parameter estimates are unbiased but the standard error estimates are, then applying the Cochran-Orcutt adjustment should change the estimates of the standard errors without changing the estimates of the equation parameters substantially.

[16] That is, we assume $\varepsilon_t \sim \left(0, \sigma^2\right)$, where the distribution is not specified, and $E\left(\varepsilon_i \varepsilon_j\right) = 0$ for all $i \neq j$.

[17] These methods make use of the mathematics of difference equations. See advanced texts like Enders (1995: pp. 68-77) for examples of the derivation of the conditions necessary for an ARMA(*p*, *q*) time-series to be stationary.

[18] AR(1) is the same as ARMA(1, 0)

[19] This set of graphs is from Enders (2005: p. 79).

[20] ARIMA means AutoRegressive Integrated Moving Average. See Enders (2005: 67) for a discussion of what integrated means. We can ignore it given our limited purposes.

[21] Another way to think about this point is to remember that, unlike the fixed-effects model, the random-effects does not use dummy variables to summarized the unknown characteristics; thus, there is no problem with multicollinearity.

[22] See Cameron and Trivedi (2005: 705] for a detailed discussion of the random-effects estimator.

[23] R-squared is in quotes in this line because these R-squareds do not have all the properties of OLS R-squareds.

[24] Because the mean and variance of the standard normal distribution are 0 and 1, respectively, its probability density function (pdf) is and the cumulative probability function is .

[25] A *stochastic variable* is a *random variable*—i.e., a variable whose value is determined as a result of a process involving an uncertain outcome.

[26] Greene suggested this example in 1990 when most people paid their bills with checks. Currently it would not be such a good example because of the development of electronic payment of bills.

[27] In these notes I discuss only what is known in the literature as the *order condition* for identification. The order condition is necessary for identification. Another condition—the *rank condition*—is a sufficient condition. See Greene (1990: Chapter 19, especially pp. 600-609) for a fuller discussion of simultaneous-equation models and the identification problem.

[28] Using one of the exogenous variables in an equation as an instrument will create perfect multicollinearity in the first stage regression.

[29] We exclude Equation (15) from this discussion because it is under-identified and, thus, cannot be estimated.

[30] The advantage of the ivreg command is that it allows you to estimate a single equation of a system of equations without fully specifying the equations in the rest of the model. Use the command reg3 if you want to specify the whole model or use Three-Stage Least Squares.

[31] The description of the command "ivreg depvar [varlist1] (varlist2=varlist_iv)" in the *Stata* help file is "ivreg fits a linear regression model using instrumental variables (or two-stage least squares) of depvar on varlist1 and varlist2 using varlist_iv (along with varlist1) as instruments for varlist2. In the language of two-stage least squares, varlist1 and varlist_iv are the exogenous variables and varlist2 the endogenous variables."

[32] The model and data for this problem first appeared in Maddala, G. S. (1988) *Introductory Econometrics* (New York: Macmillan Publishing Company): 331-317.

[33] See Berndt, Ernst R. (1991) *The Practice of Econometrics* (Reading, MA: Addison-Wesley Publishing Company): 375-380.

[34] Butler, J. S., T. Aldrich Finegan, and John J. Siegfried (1998). Does more calculus improve student learning in Intermediate Micro- and Macroeconomic Theory? *Journal of Applied Econometrics* 13(2):185-202.

[35] This particular notation implies that there are $k - 1$ explanatory variables.

[36] See Greene (1990): 704.

[37] One way to make the conversion from the *Stata* output to the neater table relatively easily is to follow these steps: (1) replace each double space by a single space until there were none left; (2) replace each space with a tab (^t); (3) convert the material into a table using the "Insert/Table" command with a tab as the separator; and (4) clean up the table by moving the data into an *Excel* file, fixing the formatting, and returning the data to the *Word* file (alternatively, you can use formatting commands in *Stata* to control how the output appears).

Chapter 3. A sample Honors paper[*]

Traditionally, empirical research papers in economics journals have five or more sections. In the first section, unimaginatively known as the introduction, the researcher briefly (1) describes what question he or she is attempting to answers, (2) indicates why the reader should be interested in the answer to the questions, and (3) often summarizes what the paper's conclusions. It is traditional in the second section for authors to discuss the instidutional background to the question and provide a theoretical model to be used in the estimation process. Quite often it makes more sense to refer to the variables in conceptual terms in this section and leave the actual specification of the variables in later parts of the paper. A traditional example of this is the ubiquitous "socioeconomic variables" included in many economic models. The reason for this generality is that perfect measures of the variables conceived in most models are not available and most researchers are forced to use proxies for the variables in the model when completing their empirical work. For this reason it is traditional in the third section of the paper to discuss what variables are used as proxies for the variables mentioned in the model. For instance, many papers use this section to specify what variables will proxy the "socioeconomic variables." It is appropriate to discuss shortcoming of the data set in the third section.

Economists use the fourth section of the paper to describe the econometric model estimated along with the statistical issues created by the shortcomings of data and the model. The fourth section of the paper also usually includes a presentation of the empirical estimations and a discussion of the implications of the estimations for the central questions of the paper. The fifth section of the paper usually includes a recap of the research, a discussion of the implications of the empirical work, and suggestions for further research.

Obviously, not all economics journal articles are split into the five sections described above; every author has his or her way of organizing their arguments. Indeed, how a paper is organized will reflect the story the author is trying to tell. It is as James Joyce noted in *Protrait of an Artist as a Young Man*, in art "the whole is related to the parts and the parts are related to the whole." In a well-crafted paper the author's message dictates the organizational structure of the paper and the material in each section must relate back to this message. In what follows we will outline what might go into each of these sections, leaving it to you to fill in the missing parts.

3.1. Section 1. Introduction

In this hypothetical Honors paper we examine the impact of a law change on a desired outcome of the law. In particular, sometime during the years leading up to 2007 all of the states adopted a 0.08 *per se* rule on the blood alcohol content (BAC) of determining if a driver is drunk: after passage of the law any driver with a BAC of 0.08 or higher is presumed to be driving under the influence. Some of the states also have "zero tolerance for underaged drinking and driving" level that applies only to drivers under age 21. Defence of drivers accused of DUI is, not surprisingly, big business for lawyers. Table 1 reports the some of the current DUI laws by state as reported on the website of a law firm specializing in DUI cases.

State	***Per se* BAC Level**	**Zero Tolerance BAC Level**	**Enhanced Penalty BAC Level**	**State**	***Per se* BAC Level**	**Zero Tolerance BAC Level**	**Enhanced Penalty BAC Level**
Alabama	**0.08**	**0.02**	**N/A**	**Montana**	**0.08**	**0.02**	**0.18**
Alaska	**0.08**	**0.00**	**0.16**	**Nebraska**	**0.08**	**0.02**	**0.15**
Arizona	**0.08**	**0.00**	**0.15**	**Nevada**	**0.08**	**0.02**	**0.18**
Arkansas	**0.08**	**0.02**	**0.15**	**New Hampshire**	**0.08**	**0.02**	**0.16**
California	**0.08**	**0.01**	**0.15**	**New Jersey**	**0.08**	**0.01**	**N/A**
Colorado	**0.08**	**0.02**	**0.20**	**New Mexico**	**0.08**	**0.02**	**0.16**
Connecticut	**0.08**	**0.02**	**0.16**	**New York**	**0.08**	**0.02**	**0.18**
Delaware	**0.08**	**0.02**	**0.15**	**North Carolina**	**0.08**	**0.00**	**0.16**
DC	**0.08**	**0.00**	**0.20**	**North Dakota**	**0.08**	**0.02**	**0.18**
Florida	**0.08**	**0.02**	**0.15**	**Ohio**	**0.08**	**0.02**	**0.17**
Georgia	**0.08**	**0.02**	**0.15**	**Oklahoma**	**0.08**	**0.00**	**0.15**
Hawaii	**0.08**	**0.02**	**0.15**	**Oregon**	**0.08**	**0.00**	**N/A**

Idaho	0.08	0.02	0.20	Pennsylvania	0.08	0.02	0.16
Illinois	0.08	0.00	0.16	Rhode Island	0.08	0.02	0.15
Indiana	0.08	0.02	0.15	South Carolina	0.08	0.02	0.15
Iowa	0.08	0.02	0.15	South Dakota	0.08	0.02	0.17
Kansas	0.08	0.02	0.15	Tennessee	0.08	0.02	0.20
Kentucky	0.08	0.02	0.18	Texas	0.08	0.00	0.15
Louisiana	0.08	0.02	0.15	Utah	0.08	0.00	0.16
Maine	0.08	0.00	0.15	Vermont	0.08	0.02	N/A
Maryland	0.08	0.02	N/A	Virginia	0.08	0.02	0.15
Massachusetts	0.08	0.02	0.20	Washington	0.08	0.02	0.15
Michigan	0.08	0.02	N/A	West Virginia	0.08	0.02	N/A
Minnesota	0.08	0.00	0.20	Wisconsin	0.08	0.00	0.17
Mississippi	0.08	0.02	N/A	Wyoming	0.08	0.02	0.15
Missouri	0.08	0.02	0.15				

Table 3.1. Table 1. State drunk driving laws. (Source: http://www.totaldui.com/breathalyzers/bac/laws-by-state.aspx)

The theoretical justifications for the *per se* BAC level rule is (1) that it will provide a disincentive for individuals to drive after drinking and (2) that it will reduce the cost of prosecuting DUI drivers. In terms of economics the law aims to reduce the negative externalities created by drunk drivers. The question to be examined in this paper is whether the *per se* laws have reduce the number of automobile fatalities. Persumably, if the law is successful in reducing the number of DUI drivers, it will reduce the number of accidents they cause and, thus, reduce the number of DUI fatalities. Whether the *per se* BAC law does reduce the number of automobile fatalities—and, thus, is a useful law—is the empirical issue this paper proposes to investigate.

Exercises

1. The introduction or section 2 should include a discussion of the current state of the literature. What, if anything, is written in economics journals about the impact of DUI laws on the automobile fatality rate?

2. The introduction presented above is very "thin". How would you fill out this discussion? Is this the appropriate place to introduce a discussion of the institutional history of the adoption of the *per se* BAC law?

3. How would your introduction be affected by the results you report later in the paper?

4. *A priori*, do you think that the *per se* BAC law is an effective way of reduing drunk driving or is it just a placebo for voters upset with drunk drives (like MOM)? Does it "matter" to you as a researcher whether the *per se* BAC law is effective?

3.2. Theoretical issues

Any model of automobile fatalities is a function of the unit of observation. Since we are interested in the impact of state laws on automobile fatalities, it seems reasonable that we construct a model to explain the differences in automobile fatalities at the state level (although it is tempting to use county level data). There are interstate differences that potentially explain differences in fatalities. First, people drive more in phyically larger states and states with larger populations than they do in other states. since more driving increases the probability of an accident, we need to standardize our measure of fatalities by the vehicle miles driven in the state. It is traditional in the empirical literature to measure the number of fatalities as fatalities per 100 million vehicle miles driven rather than the number of fatalities; in the interest of simplicity we follow this tradition.

A second phyical characteristic that affects the fatality rate is the type of road used in a state. In particular, it is well-known that in the United States perhaps the safest roads are rural interestate highways. Thus, in our model we will need to hold constant the type of highway in the state. An additional variable that potentially affects the fatality rate is the mix of drivers. In particular, given the propensity of insurance companies to charge higher rates to individuals under the age of 25, it is reasonable to assume that the more young drivers in the state the higher the fatality rate. Similarly, given the tendency of the elderly to have decreased reaction rates, it is possible that the presence of more elderly drivers would drive up the automobile accident rate.

There are several behavioral variables that might affect driving habits and, thus, automobile accident rates. First, it seems reasonable to assume that the value of time and cost of death are higher for wealthier people than they are for less wealth drivers. However, the direction of the effect of income on driver behavior is unclear. A person with a higher value of time might be more willing to speed than one with a lower value of time because time spent driving is time not spent earning income or engaging in leisure. Additionally, and here the issue is very uncertain, a wealthier person may be less willing to engage in risky driving or drinking behavior because he or she has more income to lose than a poorer individual.

A second variable that affects the behavior of individuals is the cost of gasoline. Higher gas prices will cause individuals to drive less and closer to the gas efficient speed. Most often driving closer to the gas efficient speed implies a slower and safer speed. Moreover, since all drivers are driven toward the gas efficient speed, the variance in speeds on the highways should be reduced. In either case, a higher price of gasoline should cause the number of automobile fatalities to fall. Since gasoline is purchased on the world market, the major source of differences in state-level gasoline prices is diffences among the state gasoline taxes. Similarly, we would expect things like state taxes on alcohol consumption and the strictness of the of the DUI laws to reduce both the amount of alcohol comsumption and the amount of driving under the influence.

In the most general terms the model to be estimated is:

$$(3.1)\ FPVMD = f(\text{ type of roads, mix of drivers, income, cost of gasoline, state laws }),$$

where *FPVMD* is a measure of the number of automobile fatalities per vehicle mile driven annually in a state. In the next section of the paper we will make this model useable by chosing specific variables to proxy the explanatory variables

Exercises

1. The model described is incomplete (as are almost all useful models). What, if anything, would you add to the model?
2. Often the models in economics papers involve constrained optimization models that yield the predictions that are tested in the empirical part of the paper. Are there any optimization models implicit in the description above?

3.3. The data

Many of the states adopted the 0.08 BAC *per se* standard between 1994 and 2008. In fact, all states adopted this standard by 2007. Thus, a panel data set of data from all of the 50 states and the District of Columbia should offer enough variance in the this variable to enable us to evaluate the effectiveness of the law. The Department of Transportation and the Census Bureau provide enough data to enable us to construct a reasonable data set for all of the states for this period. What should ensue here is a detailed description of all of the variables in the data set along with the sources used to collect the data. However, we leave the construction of this part of the paper to you and resort to summarizing the variables included in the data set in Table 2. The data are available in the file the "Data set" sheet in Auto_fatalities_data.xls; the definition of the FIPS codes are included in a sheet named "State FIPS codes" in the same file. Table 3 defines the variables included by column in the "Data set" sheet of Auto_fatalities_data.xls.

Care needs to be taken when gathering the data because some sources list the states in alphabetical order by the full name of the state, the way that the FIPS codes orders the states. In this case Deleware preceeds the District of Columbia. In other sources the states are listed in alphabetical order of the each state's abreviated title. In these cases the District of Columbia preceeds Deleware because DC preceeds DE. This sorting of the states causes several states to appear in an order different than they appear in the FIPS codes. A similar problem occurs with working with county level data because some government sources list all county names beginning with Mc ahead of all other county names beginning with M while other sources list county names beginning with Mc after county names beginning with Ma. In both cases order all of the state or county data by their FIPS code prevents confusing the order of the observations.

Variable	Source	Period
FIPS code identifying each state	http://www.census.gov/datamap/fipslist/AllSt.txt	1994-2008
Fatalities from automobile accidents	http://www-fars.nhtsa.dot.gov/States/StatesFatalitiesFatalityRates.aspx	1994-2008
Fatalities per 100 million vehicle miles driven	www-fars.nhtsa.dot.gov	1994-2008
State gas tax rate per gallon in dollars	www.fhwa.dot.gov/policyinformation/statistics	1994-2008
Real state gas tax rate per gallon in 2009 dollars	State gas tax rate per gallon in dollars divided by the CPI with a base year of 2009	1994-2008

State cigarette tax per pack in dollars,	State Sales, Gasoline, Cigarette, and Alcohol Tax Rates by State, 2000-2010	2000-2008
State tax on spirits	State Sales, Gasoline, Cigarette, and Alcohol Tax Rates by State, 2000-2010	2000-2008
State tax wine	State Sales, Gasoline, Cigarette, and Alcohol Tax Rates by State, 2000-2010	2000-2008
State tax on beer	State Sales, Gasoline, Cigarette, and Alcohol Tax Rates by State, 2000-2010	2000-2008
Vehicle miles driven on state rural interstates	Table VM-202 for various years on: http://www.fhwa.dot.gov	1994-2008
Total vehicle miles driven on state rural roads	Table VM-202 for various years on: http://www.fhwa.dot.gov	1994-2008
Vehicle miles driven on state urban interstates	Table VM-202 for various years on: http://www.fhwa.dot.gov	1994-2008
Total vehicle miles driven on state urban roads	Table VM-202 for various years on: http://www.fhwa.dot.gov	1994-2008
Percent of the registered drivers under the age of 20	Table VM-202 for various years on: http://www.fhwa.dot.gov	1994-2008
Percent of the registered drivers under the age of 25	Table DL-22 for various years on: http://www.fhwa.dot.gov	1994-2008
Percent of the registered drivers over age 70	Table DL-22 for various years on: http://www.fhwa.dot.gov	1994-2008
Percent of the registered drivers over age 75	Table DL-22 for various years on: http://www.fhwa.dot.gov	1994-2008

Percent of the registered drivers over age 80	Table DL-22 for various years on: http://www.fhwa.dot.gov	1994-2008
Percent of the registered drivers over age 85	Table DL-22 for various years on: http://www.fhwa.dot.gov	1994-2008
State mean family income in 2009 dollars[a]	http://www.census.gov	1994-2008
Dummy variable = 1 if the state has passed the 0.08 *per se* BAC law; 0 otherwise	NHTSA, Regional Office. Updated as of December 1, 2008.	1994-2008

Table 3.2. Definitions and sources of the variables in the data set.

Column	Column title	Variable
A	FIPS	FIPS code identifying each state
B	Year	Variable denoting the year and ranges from1994 to 2008
C	Fatalities	Fatalities from automobile accidents
D	DPVM	Fatalities per 100 million vehicle miles driven
E	SGasTax	State tax on gasoline, $/gallon
F	RSGasTax	Real state tax on gasoline, 2009$/gallon
G	CigTax	State tax on cigarettes, dollars per 20-pack
H	SpTax	State tax on spirits, dollars per gallon
I	WineTax	State tax on wine, dollars per gallon
J	BeerTax	State tax on beer, dollars per gallon
K	RuralInterstateVMD	Vehicle-miles driven in a year on rural interstates, 100 million

L	RuralTotalVMD	Vehicle-miles driven in a year on all rural roadways, 100 million
M	UrbanInterstateVMD	Vehicle-miles driven in a year on urban interstates, 100 million
N	UrbanTotalVMD	Vehicle-miles driven in a year on all urban roadways, 100 million
O	PU20	Percent of licensed under the age of 20
P	PU25	Percent of licensed under the age of 25
Q	PO70	Percent of licensed over the age of 70
R	PO75	Percent of licensed over the age of 75
S	PO80	Percent of licensed over the age of 80
T	PO85	Percent of licensed over the age of 85
U	BACPS	Dummy variable equal to 1 if the state has adopted the 0.08 BAC *per se* law; 0 otherwise
V	RMFI09	Median family income in a state in 2009 dollars

Table 3.3. Data included in dataset.

Exercises

1. At this point in your thesis you would want to point out that each of the variables in the data set are proxies for the variables discussed in part 2 of your paper. As an exercise explain how each of the explanatory variables in Table 2 are proxies for the explanatory variables mentioned in the theory section.

2. It would seem that the "cleanest" variable in the whole data set is "fatalities." Lookup the official definition of how a fatality from an automobile accident is measured. Does this variable still seem to have a clear and unequivocal meaning?

3.4. Empirical estimation

Now we are almost ready to present the estimation results from the model. There are a few things we need to cover before we move to presenting the estimation results. First, what, if any, are the econometric issues raised by the model and the data set? In this case we are using a panel data set to estimate the regression:

(3.2)

$$fpvmd_{it} = \beta_0 + \sum_{j=1}^{k-1} \beta_j x_{jit} + \beta_k D_{it}^{BAC} + \varepsilon_{it},$$

where *fpvmd* $_{it}$ is the number of fatalities per 100 million vehicle miles driven in state i in year t, the *x* $_{jit}$ is the jth explanatory variable in state i in year t, and D_{it}^{BAC} is the dummy variable equal to 1 if state *i* has a 0.08 *per se* BAC law in year *t*. From a policy point of view what we are interested in is the sign of *β* $_k$ and if *β* $_k$ is statistically different from zero. At this point it would be appropriate to discuss whether you intend to use a fixed effects or a random effect model. In the interest is simplicity, we will use a fixed effects model but in your own research you would need to consider using either model.

A second issue that needs to be considered is if you plan to use a linear model as specified above or if you might use the natural logarithm of the fatality rate. Since we have no *a priori* reason to believe that the relationship between the fatality rate and the explanatory variables are linear, we will estimate both log-linear and a log-log models. In this way we can test if our policy conclusions are sensitive to the mathematical specification of our model.

Now we are ready to report the results of the estimation. The key here is to avoid writing a travelog of the estimations. Instead, report all of the regressions in one or more tables and then discuss the results presented in each table.

Exercises

1. **In our estimations we use (a) a linear model, (b) a log-linear model, and (c) a log-log linear model. What are the economic interpretation of the estimated parameters in each of the models? Be sure to discuss both dummy variables and continuous variables.**

2. **Why does it not make more sense to use an explanatory variable rather than the log of that explanatory variable when that explanatory variable is a percentage?**

3.5. Notes on the estimation of the model

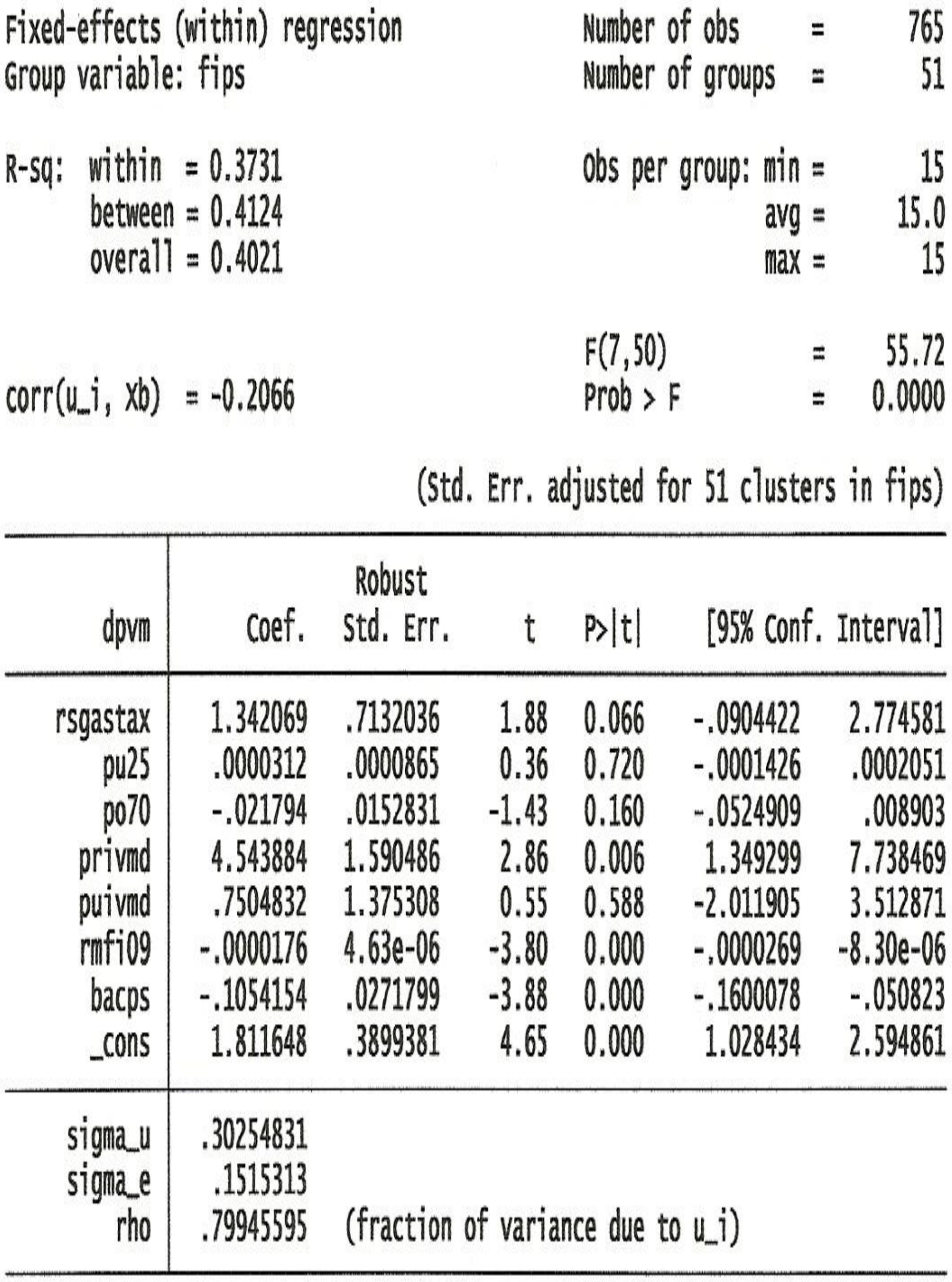
Since you will find it useful to replicate the estimation of the basic results, this section consists mainly of a set instructions in Table 4 for use with *Stata*.

	Instruction	*Stata* commands
1.	Open *Stata* and copy the data in Auto_fatalities_data.xls into the data editor. You will have 765 observations of 22 variables.	
2.	Tell *Stata* what variable denotes the state	.iis
3.	Tell *Stata* what variable denotes the year	.tis
4.	Create the new variable the percentage of the total vehicle miles driven that are on rural interstate roads	.generate privmd = ruralinterstatevmd/(ruraltotalvmd + urbantotalvmd)
5.	Create the new variable the percent of the total vehicle miles driven that are on urban interstate roads	.generate puivmd= urbaninterstatevmd/(ruraltotalvmd+urbantotalvmd)
6.	Create the logarithm transportation of all of the variables that are not percentages	.generate lz = log(z), where z = dpvmd, sgastax, rsgastax, and rmfi09
7a.	Estimate the fixed effects model for the linear model (see output in Figure 1)	.xtreg dpvm rsgastax pu25 po70 privmd puivmd rmfi09 bacps, fe vce(robust) vsquish
7b.	Estimate the fixed effects model for the log-linear model (see output in Figure 2)	.xtreg ldpvm rsgastax pu25 po70 privmd puivmd rmfi09 bacps, fe vce(robust) vsquish
7c.	Estimate the fixed effects model for the log-log model (see output in Figure 3)	.xtreg ldpvm lrsgastax pu25 po70 privmd puivmd lrmfi09 bacps, fe vce(robust) vsquish
8.	Place the results into a table making it easier to compare your results; Table 5 is one such table.	

9a.	**The results in Table 5 suggest that the *per se* 0.08 BAC is a successful way to reduce automobile deaths. However the sign on the real gasoline tax rate is the opposite of what we might reasonably expect. Let's check the sensitivity of our results by rerunning the same three regressions with the real gasoline tax replaced by the nominal gasoline tax. See Table 6 for the results of these regressions.**	**. xtreg dpvm sgastax pu25 po70 privmd puivmd rmfi09 bacps, re vce(robust) vsquish**
9b.		**.xtreg ldpvm sgastax pu25 po70 privmd puivmd rmfi09 bacps, fe vce(robust) vsquis**
9c.		**.xtreg ldpvm lsgastax pu25 po70 privmd puivmd lrmfi09 bacps, fe vce(robust) vsquish**

Table 3.4. Instructions for further investigation of the stability of the regression estimates.

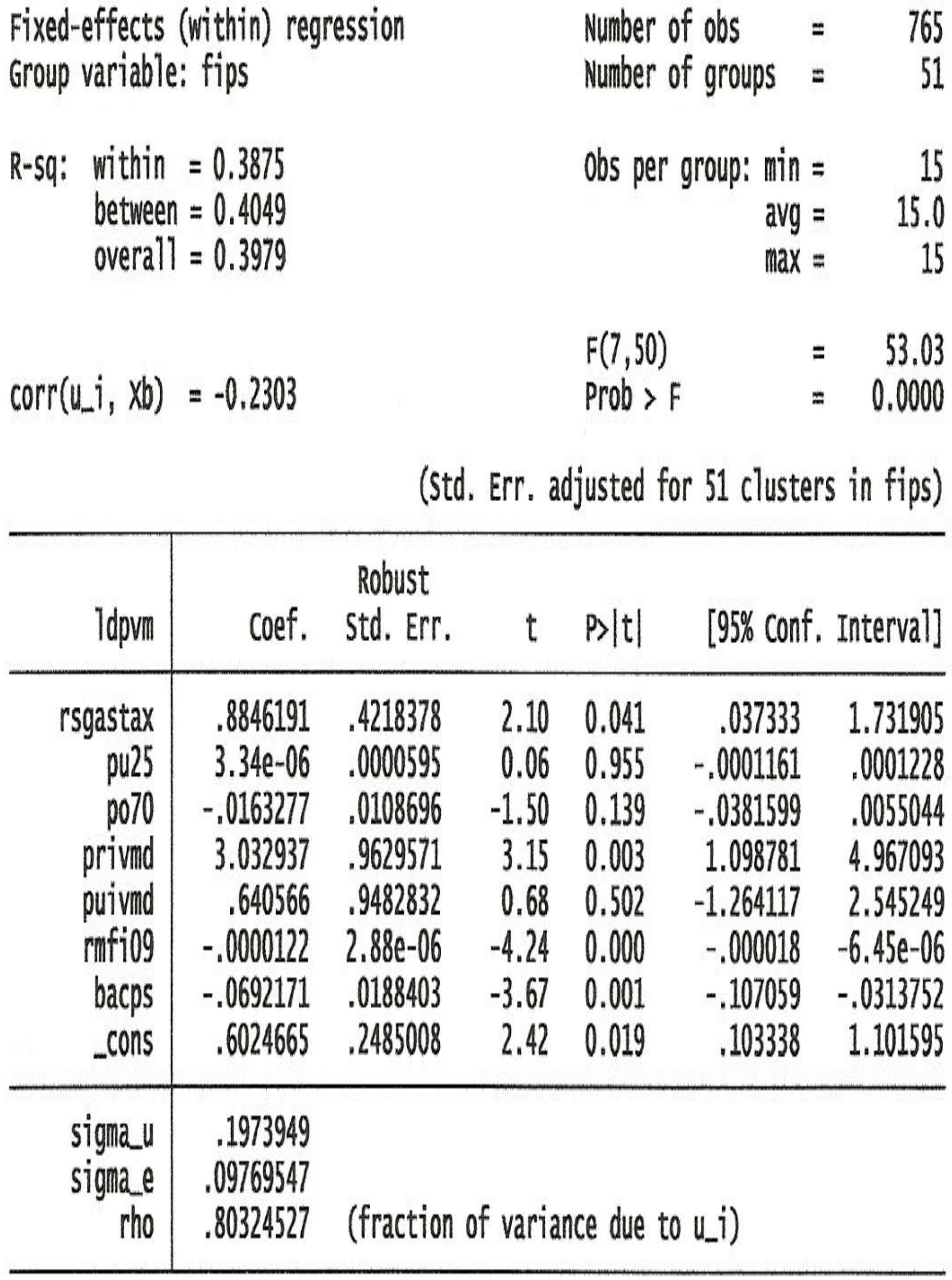
Figure 3.1.



```
Fixed-effects (within) regression               Number of obs      =       765
Group variable: fips                            Number of groups   =        51

R-sq:  within  = 0.3731                         Obs per group: min =        15
       between = 0.4124                                        avg =      15.0
       overall = 0.4021                                        max =        15

                                                F(7,50)            =     55.72
corr(u_i, Xb)  = -0.2066                        Prob > F           =    0.0000

                                  (Std. Err. adjusted for 51 clusters in fips)
------------------------------------------------------------------------------
             |               Robust
        dpvm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    rsgastax |   1.342069   .7132036     1.88   0.066    -.0904422    2.774581
        pu25 |   .0000312   .0000865     0.36   0.720    -.0001426    .0002051
        po70 |  -.021794    .0152831    -1.43   0.160    -.0524909     .008903
      privmd |   4.543884   1.590486     2.86   0.006     1.349299    7.738469
      puivmd |   .7504832   1.375308     0.55   0.588    -2.011905    3.512871
      rmfi09 |  -.0000176   4.63e-06    -3.80   0.000    -.0000269   -8.30e-06
       bacps |  -.1054154   .0271799    -3.88   0.000    -.1600078   -.050823
       _cons |   1.811648   .3899381     4.65   0.000     1.028434    2.594861
-------------+----------------------------------------------------------------
     sigma_u |  .30254831
     sigma_e |   .1515313
         rho |  .79945595   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Results of linear regression results. (t-ratios are in parentheses)

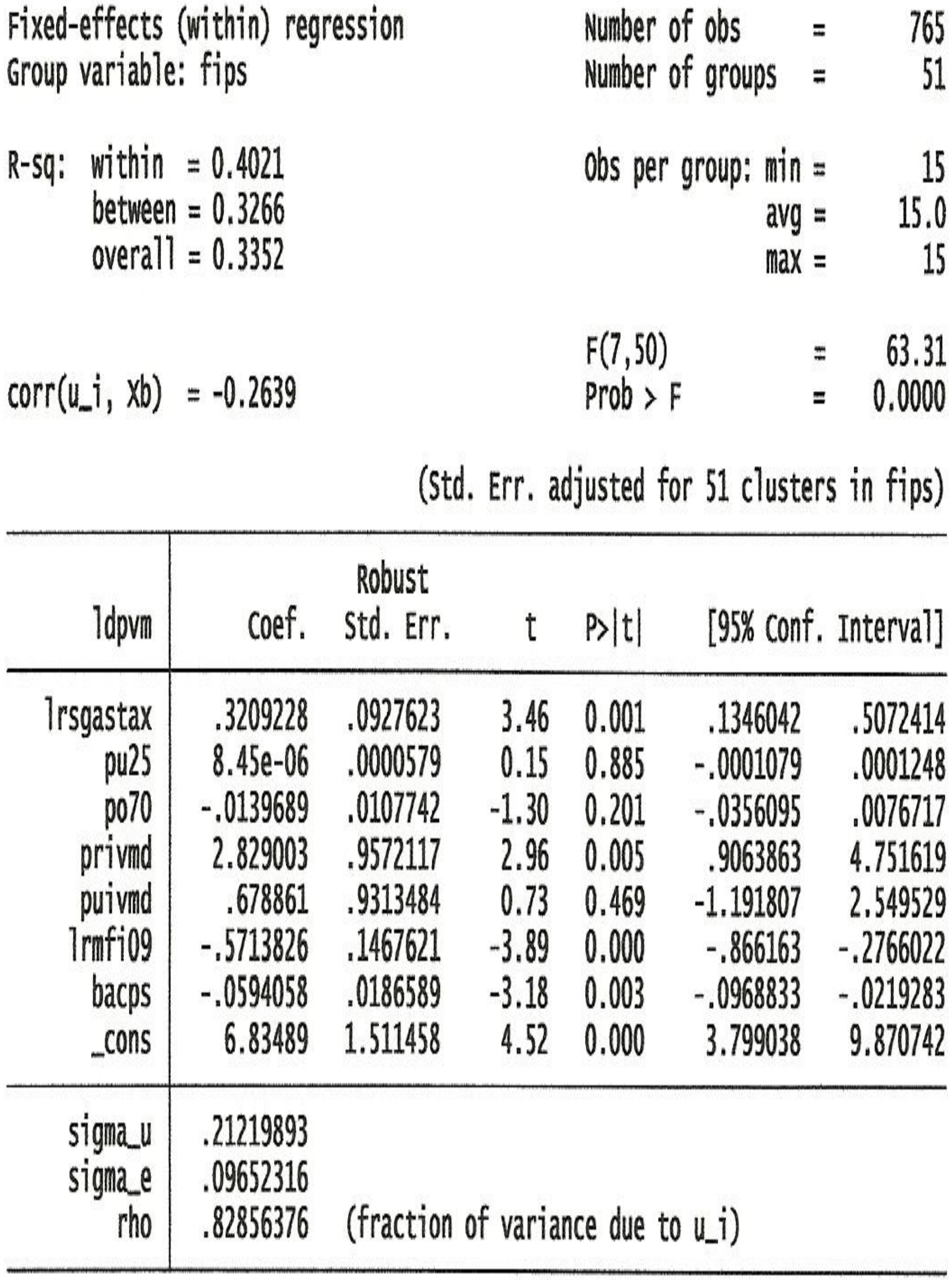
Figure 3.2.



```
Fixed-effects (within) regression               Number of obs      =       765
Group variable: fips                            Number of groups   =        51

R-sq:  within  = 0.3875                         Obs per group: min =        15
       between = 0.4049                                        avg =      15.0
       overall = 0.3979                                        max =        15

                                                F(7,50)            =     53.03
corr(u_i, Xb)  = -0.2303                        Prob > F           =    0.0000

                                   (Std. Err. adjusted for 51 clusters in fips)
------------------------------------------------------------------------------
             |               Robust
       ldpvm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    rsgastax |   .8846191   .4218378     2.10   0.041      .037333    1.731905
        pu25 |   3.34e-06   .0000595     0.06   0.955    -.0001161    .0001228
        po70 |  -.0163277   .0108696    -1.50   0.139    -.0381599    .0055044
      privmd |   3.032937   .9629571     3.15   0.003     1.098781    4.967093
      puivmd |    .640566   .9482832     0.68   0.502    -1.264117    2.545249
      rmfi09 |  -.0000122   2.88e-06    -4.24   0.000     -.000018   -6.45e-06
       bacps |  -.0692171   .0188403    -3.67   0.001     -.107059   -.0313752
       _cons |   .6024665   .2485008     2.42   0.019      .103338    1.101595
-------------+----------------------------------------------------------------
     sigma_u |   .1973949
     sigma_e |  .09769547
         rho |  .80324527   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Results of the log-linear regression. (t-ratios are in parentheses)

Figure 3.3.



```
Fixed-effects (within) regression               Number of obs      =       765
Group variable: fips                            Number of groups   =        51

R-sq:  within  = 0.4021                         Obs per group: min =        15
       between = 0.3266                                        avg =      15.0
       overall = 0.3352                                        max =        15

                                                F(7,50)            =     63.31
corr(u_i, Xb)  = -0.2639                        Prob > F           =    0.0000

                                  (Std. Err. adjusted for 51 clusters in fips)
------------------------------------------------------------------------------
             |               Robust
       ldpvm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   lrsgastax |   .3209228   .0927623     3.46   0.001     .1346042    .5072414
        pu25 |   8.45e-06   .0000579     0.15   0.885    -.0001079    .0001248
        po70 |  -.0139689   .0107742    -1.30   0.201    -.0356095    .0076717
      privmd |   2.829003   .9572117     2.96   0.005     .9063863    4.751619
      puivmd |    .678861   .9313484     0.73   0.469    -1.191807    2.549529
     lrmfi09 |  -.5713826   .1467621    -3.89   0.000     -.866163   -.2766022
       bacps |  -.0594058   .0186589    -3.18   0.003    -.0968833   -.0219283
       _cons |    6.83489   1.511458     4.52   0.000     3.799038    9.870742
-------------+----------------------------------------------------------------
     sigma_u |  .21219893
     sigma_e |  .09652316
         rho |  .82856376   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Results of the log-log regression. (t-ratios are in parentheses)

At this point is makes some sense to compare the parameter estimates for 0.08 BAC *per se* law; this comparison, shown in Table 5, suggests that the effect of the *per se* 0.08 BAC law was to reduce fatalities. Moreover, the estimates for each of the models is very stable whether one uses the real price of gasoline or the nominal price of gasoline, thus giving us some more confidence in our conclusions.

	Linear	Log-linear	Log-log
State tax of gasoline in 2009 dollars			
State has a 0.08 *per se* BAC law	-0.1054	-0.0692	-0.0594
	(-3.88)	(-3.67)	(-3.18)
State tax of gasoline in current dollars			
State has a 0.08 *per se* BAC law	-0.1191	-0.0778	-0.0762
	(-4.83)	(-4.54)	(-4.52)

Table 3.5. Comparison of the parameter estimates for each model with different measures of the cost of gasoline.

The balance of this section of the paper would be devoted to further tests of the stability of our results under varying assumptions. Among other tests one would expect to see if the choice of a fixed-effects model affects your policy conclusions.

Exercises

1. Complete the Lagrange test for random effects for each of the three models, using the nominal price of gasoline. Organize the results of this test into a table.

2. **Re-estimate the three models replacing the percent of registered drivers under the age of 25 with the percent of drivers under 20. Make the same same kind of replacement for the number of drivers over age 70 (i.e., experiment with the alternative age cutoffs—over 75, over 80, and over 85). Do any of your major conclusions change?**

3. **What, if any, explanation can you give for the differences in the parameter estimates for the price of gasoline generated when the real price of gasoline is replaced by the nominal price of gasoline?**

3.6. Conclusions and further research

This section of your paper should be devoted to a careful recapping of your results and providing suggestions for further research. Such a discussion might include some cautious guesses at why the 0.08 BAC *per se* standard appears to affect driver behavior. The discussion could also include some estimates of the number of lifes saved by the introduction of a *per se* standard.

Chapter 3. A sample Honors paper[*]

Traditionally, empirical research papers in economics journals have five or more sections. In the first section, unimaginatively known as the introduction, the researcher briefly (1) describes what question he or she is attempting to answers, (2) indicates why the reader should be interested in the answer to the questions, and (3) often summarizes what the paper's conclusions. It is traditional in the second section for authors to discuss the instidutional background to the question and provide a theoretical model to be used in the estimation process. Quite often it makes more sense to refer to the variables in conceptual terms in this section and leave the actual specification of the variables in later parts of the paper. A traditional example of this is the ubiquitous "socioeconomic variables" included in many economic models. The reason for this generality is that perfect measures of the variables conceived in most models are not available and most researchers are forced to use proxies for the variables in the model when completing their empirical work. For this reason it is traditional in the third section of the paper to discuss what variables are used as proxies for the variables mentioned in the model. For instance, many papers use this section to specify what variables will proxy the "socioeconomic variables." It is appropriate to discuss shortcoming of the data set in the third section.

Economists use the fourth section of the paper to describe the econometric model estimated along with the statistical issues created by the shortcomings of data and the model. The fourth section of the paper also usually includes a presentation of the empirical estimations and a discussion of the implications of the estimations for the central questions of the paper. The fifth section of the paper usually includes a recap of the research, a discussion of the implications of the empirical work, and suggestions for further research.

Obviously, not all economics journal articles are split into the five sections described above; every author has his or her way of organizing their arguments. Indeed, how a paper is organized will reflect the story the author is trying to tell. It is as James Joyce noted in *Protrait of an Artist as a Young Man*, in art "the whole is related to the parts and the parts are related to the whole." In a well-crafted paper the author's message dictates the organizational structure of the paper and the material in each section must relate back to this message. In what follows we will outline what might go into each of these sections, leaving it to you to fill in the missing parts.

3.1. Section 1. Introduction

In this hypothetical Honors paper we examine the impact of a law change on a desired outcome of the law. In particular, sometime during the years leading up to 2007 all of the states adopted a 0.08 *per se* rule on the blood alcohol content (BAC) of determining if a driver is drunk: after passage of the law any driver with a BAC of 0.08 or higher is presumed to be driving under the influence. Some of the states also have "zero tolerance for underaged drinking and driving" level that applies only to drivers under age 21. Defence of drivers accused of DUI is, not surprisingly, big business for lawyers. Table 1 reports the some of the current DUI laws by state as reported on the website of a law firm specializing in DUI cases.

State	***Per se* BAC Level**	**Zero Tolerance BAC Level**	**Enhanced Penalty BAC Level**	**State**	***Per se* BAC Level**	**Zero Tolerance BAC Level**	**Enhanced Penalty BAC Level**
Alabama	**0.08**	**0.02**	**N/A**	**Montana**	**0.08**	**0.02**	**0.18**
Alaska	**0.08**	**0.00**	**0.16**	**Nebraska**	**0.08**	**0.02**	**0.15**
Arizona	**0.08**	**0.00**	**0.15**	**Nevada**	**0.08**	**0.02**	**0.18**

Arkansas	0.08	0.02	0.15	New Hampshire	0.08	0.02	0.16
California	0.08	0.01	0.15	New Jersey	0.08	0.01	N/A
Colorado	0.08	0.02	0.20	New Mexico	0.08	0.02	0.16
Connecticut	0.08	0.02	0.16	New York	0.08	0.02	0.18
Delaware	0.08	0.02	0.15	North Carolina	0.08	0.00	0.16
DC	0.08	0.00	0.20	North Dakota	0.08	0.02	0.18
Florida	0.08	0.02	0.15	Ohio	0.08	0.02	0.17
Georgia	0.08	0.02	0.15	Oklahoma	0.08	0.00	0.15
Hawaii	0.08	0.02	0.15	Oregon	0.08	0.00	N/A
Idaho	0.08	0.02	0.20	Pennsylvania	0.08	0.02	0.16
Illinois	0.08	0.00	0.16	Rhode Island	0.08	0.02	0.15

Indiana	0.08	0.02	0.15	South Carolina	0.08	0.02	0.15
Iowa	0.08	0.02	0.15	South Dakota	0.08	0.02	0.17
Kansas	0.08	0.02	0.15	Tennessee	0.08	0.02	0.20
Kentucky	0.08	0.02	0.18	Texas	0.08	0.00	0.15
Louisiana	0.08	0.02	0.15	Utah	0.08	0.00	0.16
Maine	0.08	0.00	0.15	Vermont	0.08	0.02	N/A
Maryland	0.08	0.02	N/A	Virginia	0.08	0.02	0.15
Massachusetts	0.08	0.02	0.20	Washington	0.08	0.02	0.15
Michigan	0.08	0.02	N/A	West Virginia	0.08	0.02	N/A
Minnesota	0.08	0.00	0.20	Wisconsin	0.08	0.00	0.17
Mississippi	0.08	0.02	N/A	Wyoming	0.08	0.02	0.15

Missouri	0.08	0.02	0.15				

Table 3.1. Table 1. State drunk driving laws. (Source: http://www.totaldui.com/breathalyzers/bac/laws-by-state.aspx)

The theoretical justifications for the *per se* BAC level rule is (1) that it will provide a disincentive for individuals to drive after drinking and (2) that it will reduce the cost of prosecuting DUI drivers. In terms of economics the law aims to reduce the negative externalities created by drunk drivers. The question to be examined in this paper is whether the *per se* laws have reduce the number of automobile fatalities. Persumably, if the law is successful in reducing the number of DUI drivers, it will reduce the number of accidents they cause and, thus, reduce the number of DUI fatalities. Whether the *per se* BAC law does reduce the number of automobile fatalities—and, thus, is a useful law—is the empirical issue this paper proposes to investigate.

Exercises

1. **The introduction or section 2 should include a discussion of the current state of the literature. What, if anything, is written in economics journals about the impact of DUI laws on the automobile fatality rate?**

2. **The introduction presented above is very "thin". How would you fill out this discussion? Is this the appropriate place to introduce a discussion of the institutional history of the adoption of the *per se* BAC law?**

3. **How would your introduction be affected by the results you report later in the paper?**

4. ***A priori*, do you think that the *per se* BAC law is an effective way of reduing drunk driving or is it just a placebo for voters upset with drunk drives (like MOM)? Does it "matter" to you as a researcher whether the *per se* BAC law is effective?**

3.2. Theoretical issues

Any model of automobile fatalities is a function of the unit of observation. Since we are interested in the impact of state laws on automobile fatalities, it seems reasonable that we construct a model to explain the differences in automobile fatalities at the state level (although it is tempting to use county level data). There are interstate differences that potentially explain differences in fatalities. First, people drive more in phyically larger

states and states with larger populations than they do in other states. since more driving increases the probability of an accident, we need to standardize our measure of fatalities by the vehicle miles driven in the state. It is traditional in the empirical literature to measure the number of fatalities as fatalities per 100 million vehicle miles driven rather than the number of fatalities; in the interest of simplicity we follow this tradition.

A second phyical characteristic that affects the fatality rate is the type of road used in a state. In particular, it is well-known that in the United States perhaps the safest roads are rural interestate highways. Thus, in our model we will need to hold constant the type of highway in the state. An additional variable that potentially affects the fatality rate is the mix of drivers. In particular, given the propensity of insurance companies to charge higher rates to individuals under the age of 25, it is reasonable to assume that the more young drivers in the state the higher the fatality rate. Similarly, given the tendency of the elderly to have decreased reaction rates, it is possible that the presence of more elderly drivers would drive up the automobile accident rate.

There are several behavioral variables that might affect driving habits and, thus, automobile accident rates. First, it seems reasonable to assume that the value of time and cost of death are higher for wealthier people than they are for less wealth drivers. However, the direction of the effect of income on driver behavior is unclear. A person

with a higher value of time might be more willing to speed than one with a lower value of time because time spent driving is time not spent earning income or engaging in leisure. Additionally, and here the issue is very uncertain, a wealthier person may be less willing to engage in risky driving or drinking behavior because he or she has more income to lose than a poorer individual.

A second variable that affects the behavior of individuals is the cost of gasoline. Higher gas prices will cause individuals to drive less and closer to the gas efficient speed. Most often driving closer to the gas efficient speed implies a slower and safer speed. Moreover, since all drivers are driven toward the gas efficient speed, the variance in speeds on the highways should be reduced. In either case, a higher price of gasoline should cause the number of automobile fatalities to fall. Since gasoline is purchased on the world market, the major source of differences in state-level gasoline prices is diffences among the state gasoline taxes. Similarly, we would expect things like state taxes on alcohol consumption and the strictness of the of the DUI laws to reduce both the amount of alcohol comsumption and the amount of driving under the influence.

In the most general terms the model to be estimated is:

(3.1) $$FPVMD = f(\text{type of roads, mix of drivers, income, cost of gasoline, state laws}),$$

where *FPVMD* is a measure of the number of automobile fatalities per vehicle mile driven annually in a state. In the next section of the paper we will make this model useable by chosing specific variables to proxy the explanatory variables

Exercises

1. The model described is incomplete (as are almost all useful models). What, if anything, would you add to the model?

2. Often the models in economics papers involve constrained optimization models that yield the predictions that are tested in the empirical part of the paper. Are there any optimization models implicit in the description above?

3.3. The data

Many of the states adopted the 0.08 BAC *per se* standard between 1994 and 2008. In fact, all states adopted this standard by 2007. Thus, a panel data set of data from all of the 50 states and the District of Columbia should offer enough variance in the this variable to enable us to evaluate the effectiveness of the law. The Department of Transportation and the Census Bureau provide enough data to enable us to construct a reasonable data set for all of the states for this period. What should ensue here is a detailed description of all

of the variables in the data set along with the sources used to collect the data. However, we leave the construction of this part of the paper to you and resort to summarizing the variables included in the data set in Table 2. The data are available in the file the "Data set" sheet in Auto_fatalities_data.xls; the definition of the FIPS codes are included in a sheet named "State FIPS codes" in the same file. Table 3 defines the variables included by column in the "Data set" sheet of Auto_fatalities_data.xls.

Care needs to be taken when gathering the data because some sources list the states in alphabetical order by the full name of the state, the way that the FIPS codes orders the states. In this case Deleware preceeds the District of Columbia. In other sources the states are listed in alphabetical order of the each state's abreviated title. In these cases the District of Columbia preceeds Deleware because DC preceeds DE. This sorting of the states causes several states to appear in an order different than they appear in the FIPS codes. A similar problem occurs with working with county level data because some government sources list all county names beginning with Mc ahead of all other county names beginning with M while other sources list county names beginning with Mc after county names beginning with Ma. In both cases order all of the state or county data by their FIPS code prevents confusing the order of the observations.

Variable	Source	Period
FIPS code identifying each state	http://www.census.gov/datamap/fipslist/AllSt.txt	1994-2008
Fatalities from automobile accidents	http://www-fars.nhtsa.dot.gov/States/StatesFatalitiesFatalityRates.aspx	1994-2008
Fatalities per 100 million vehicle miles driven	www-fars.nhtsa.dot.gov	1994-2008
State gas tax rate per gallon in dollars	www.fhwa.dot.gov/policyinformation/statistics	1994-2008
Real state gas tax rate per gallon in 2009 dollars	State gas tax rate per gallon in dollars divided by the CPI with a base year of 2009	1994-2008

State cigarette tax per pack in dollars,	State Sales, Gasoline, Cigarette, and Alcohol Tax Rates by State, 2000-2010	2000-2008
State tax on spirits	State Sales, Gasoline, Cigarette, and Alcohol Tax Rates by State, 2000-2010	2000-2008
State tax wine	State Sales, Gasoline, Cigarette, and Alcohol Tax Rates by State, 2000-2010	2000-2008
State tax on beer	State Sales, Gasoline, Cigarette, and Alcohol Tax Rates by State, 2000-2010	2000-2008
Vehicle miles driven on state rural interstates	Table VM-202 for various years on: http://www.fhwa.dot.gov	1994-2008
Total vehicle miles driven on state rural roads	Table VM-202 for various years on: http://www.fhwa.dot.gov	1994-2008
Vehicle miles driven on state	Table VM-202 for various years on: http://www.fhwa.dot.gov	1994-2008

urban interstates		
Total vehicle miles driven on state urban roads	Table VM-202 for various years on: http://www.fhwa.dot.gov	1994-2008
Percent of the registered drivers under the age of 20	Table VM-202 for various years on: http://www.fhwa.dot.gov	1994-2008
Percent of the registered drivers under the age of 25	Table DL-22 for various years on: http://www.fhwa.dot.gov	1994-2008
Percent of the registered drivers over age 70	Table DL-22 for various years on: http://www.fhwa.dot.gov	1994-2008
Percent of the registered drivers	Table DL-22 for various years on: http://www.fhwa.dot.gov	1994-2008

over age 75		
Percent of the registered drivers over age 80	Table DL-22 for various years on: http://www.fhwa.dot.gov	1994-2008
Percent of the registered drivers over age 85	Table DL-22 for various years on: http://www.fhwa.dot.gov	1994-2008
State mean family income in 2009 dollars[a]	http://www.census.gov	1994-2008
Dummy variable = 1 if the state has passed the 0.08 *per se* BAC law; 0 otherwise	NHTSA, Regional Office. Updated as of December 1, 2008.	1994-2008

Table 3.2. Definitions and sources of the variables in the data set.

Column	Column title	Variable
A	FIPS	FIPS code identifying each state
B	Year	Variable denoting the year and ranges from1994 to 2008
C	Fatalities	Fatalities from automobile accidents
D	DPVM	Fatalities per 100 million vehicle miles driven
E	SGasTax	State tax on gasoline, $/gallon
F	RSGasTax	Real state tax on gasoline, 2009$/gallon
G	CigTax	State tax on cigarettes, dollars per 20-pack
H	SpTax	State tax on spirits, dollars per gallon
I	WineTax	State tax on wine, dollars per gallon
J	BeerTax	State tax on beer, dollars per gallon
K	RuralInterstateVMD	Vehicle-miles driven in a year on rural interstates, 100 million
L	RuralTotalVMD	Vehicle-miles driven in a year on all rural roadways, 100

		million
M	**UrbanInterstateVMD**	**Vehicle-miles driven in a year on urban interstates, 100 million**
N	**UrbanTotalVMD**	**Vehicle-miles driven in a year on all urban roadways, 100 million**
O	**PU20**	**Percent of licensed under the age of 20**
P	**PU25**	**Percent of licensed under the age of 25**
Q	**PO70**	**Percent of licensed over the age of 70**
R	**PO75**	**Percent of licensed over the age of 75**
S	**PO80**	**Percent of licensed over the age of 80**
T	**PO85**	**Percent of licensed over the age of 85**
U	**BACPS**	**Dummy variable equal to 1 if the state has adopted the 0.08 BAC *per se* law; 0 otherwise**
V	**RMFI09**	**Median family income in a state in 2009 dollars**

Table 3.3. Data included in dataset.

Exercises

1. At this point in your thesis you would want to point out that each of the variables in the data set are proxies for the variables discussed in part 2 of your paper. As an exercise explain how each of the explanatory variables in Table 2 are proxies for the explanatory variables mentioned in the theory section.

2. It would seem that the "cleanest" variable in the whole data set is "fatalities." Lookup the official definition of how a fatality from an automobile accident is measured. Does this variable still seem to have a clear and unequivocal meaning?

3.4. Empirical estimation

Now we are almost ready to present the estimation results from the model. There are a few things we need to cover before we move to presenting the estimation results. First, what, if any, are the econometric issues raised by the model and the data set? In this case we are using a panel data set to estimate the regression:

(3.2)

$$fpvmd_{it} = \beta_0 + \sum_{j=1}^{k-1} \beta_j x_{jit} + \beta_k D_{it}^{BAC} + \varepsilon_{it},$$

where *fpvmd* $_{it}$ is the number of fatalities per 100 million vehicle miles driven in state i in year t, the *x* $_{jit}$ is the jth explanatory variable in state i in year t, and D_{it}^{BAC} is the dummy variable equal to 1 if state *i* has a 0.08 *per se* BAC law in year *t*. From a policy point of view what we are interested in is the sign of *β* $_k$ and if *β* $_k$ is statistically different from zero. At this point it would be appropriate to discuss whether you intend to use a fixed effects or a random effect model. In the interest is simplicity, we will use a fixed effects model but in your own research you would need to consider using either model.

A second issue that needs to be considered is if you plan to use a linear model as specified above or if you might use the natural logarithm of the fatality rate. Since we have no *a priori* reason to believe that the relationship between the fatality rate and the explanatory variables are linear, we will estimate both log-linear and a log-log models. In this way we can test if our policy conclusions are sensitive to the mathematical specification of our model.

Now we are ready to report the results of the estimation. The key here is to avoid writing a travelog of the estimations. Instead, report all of the regressions in one or more tables and then discuss the results presented in each table.

Exercises

1. **In our estimations we use (a) a linear model, (b) a log-linear model, and (c) a log-log linear model. What are the economic interpretation of the estimated parameters in each of the models? Be sure to discuss both dummy variables and continuous variables.**

2. **Why does it not make more sense to use an explanatory variable rather than the log of that explanatory variable when that explanatory variable is a percentage?**

3.5. Notes on the estimation of the model

Since you will find it useful to replicate the estimation of the basic results, this section consists mainly of a set instructions in Table 4 for use with *Stata*.

	Instruction	***Stata* commands**

1.	**Open *Stata* and copy the data in Auto_fatalities_data.xls into the data editor. You will have 765 observations of 22 variables.**	
2.	**Tell *Stata* what variable denotes the state**	**.iis**
3.	**Tell *Stata* what variable denotes the year**	**.tis**
4.	**Create the new variable the percentage of the total vehicle miles driven that are on rural interstate roads**	**.generate privmd = ruralinterstatevmd/(ruraltotalvmd + urbantotalvmd)**
5.	**Create the new variable the percent of the total vehicle miles driven that are on urban interstate roads**	**.generate puivmd= urbaninterstatevmd/(ruraltotalvmd+urbantotalvmd)**

6.	**Create the logarithm transportation of all of the variables that are not percentages**	**.generate lz = log(z), where z = dpvmd, sgastax, rsgastax, and rmfi09**
7a.	**Estimate the fixed effects model for the linear model (see output in Figure 1)**	**.xtreg dpvm rsgastax pu25 po70 privmd puivmd rmfi09 bacps, fe vce(robust) vsquish**
7b.	**Estimate the fixed effects model for the log-linear model (see output in Figure 2)**	**.xtreg ldpvm rsgastax pu25 po70 privmd puivmd rmfi09 bacps, fe vce(robust) vsquish**
7c.	**Estimate the fixed effects model for the log-log model (see output in Figure 3)**	**.xtreg ldpvm lrsgastax pu25 po70 privmd puivmd lrmfi09 bacps, fe vce(robust) vsquish**
8.	**Place the results into a table making it easier to compare your results; Table 5 is one such table.**	

9a.	The results in Table 5 suggest that the *per se* 0.08 BAC is a successful way to reduce automobile deaths. However the sign on the real gasoline tax rate is the opposite of what we might reasonably expect. Let's check the sensitivity of our results by rerunning the same three regressions with the real gasoline tax replaced by the nominal gasoline tax. See Table 6 for the results of these regressions.	. xtreg dpvm sgastax pu25 po70 privmd puivmd rmfi09 bacps, re vce(robust) vsquish
9b.		.xtreg ldpvm sgastax pu25 po70 privmd puivmd rmfi09 bacps, fe vce(robust) vsquis
9c.		.xtreg ldpvm lsgastax pu25 po70 privmd puivmd

		lrmfi09 bacps, fe vce(robust) vsquish

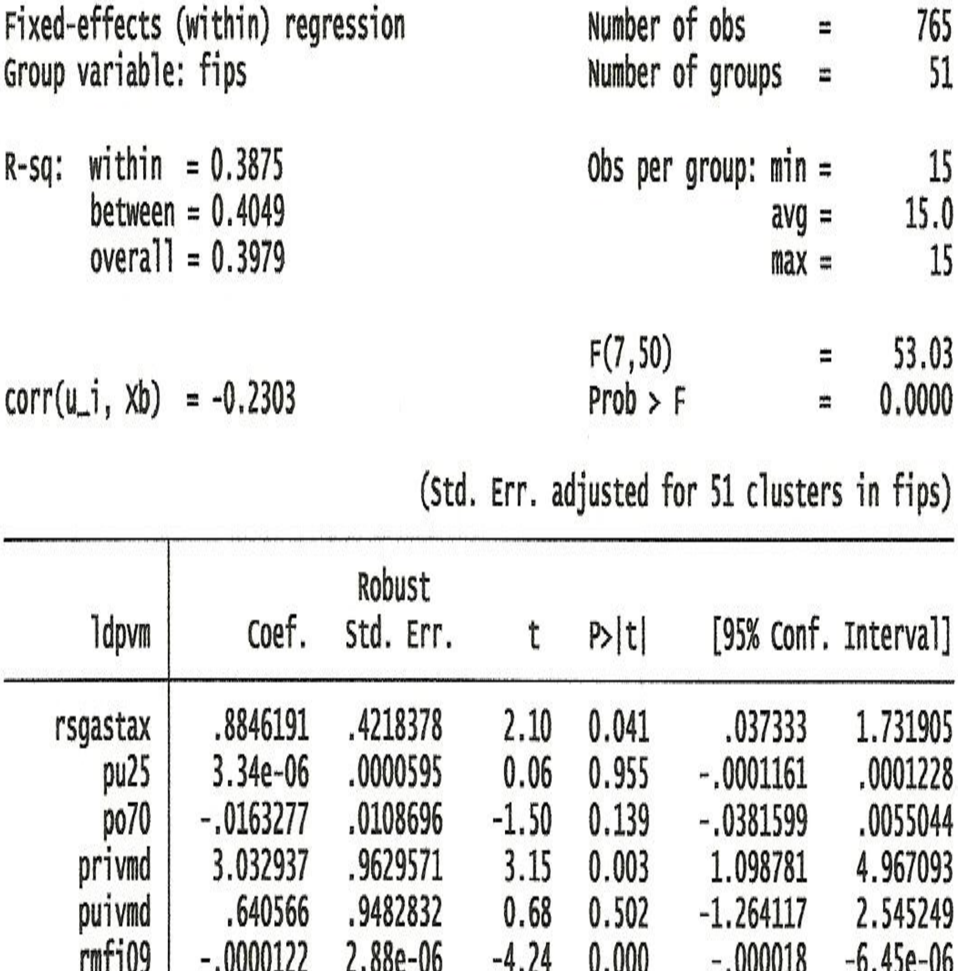
Table 3.4. Instructions for further investigation of the stability of the regression estimates.

Figure 3.1.

```
Fixed-effects (within) regression               Number of obs      =       765
Group variable: fips                            Number of groups   =        51

R-sq:  within  = 0.3731                         Obs per group: min =        15
       between = 0.4124                                        avg =      15.0
       overall = 0.4021                                        max =        15

                                                F(7,50)            =     55.72
corr(u_i, Xb)  = -0.2066                        Prob > F           =    0.0000

                                 (Std. Err. adjusted for 51 clusters in fips)
------------------------------------------------------------------------------
             |               Robust
        dpvm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    rsgastax |   1.342069   .7132036     1.88   0.066    -.0904422    2.774581
        pu25 |   .0000312   .0000865     0.36   0.720    -.0001426    .0002051
        po70 |   -.021794   .0152831    -1.43   0.160    -.0524909     .008903
      privmd |   4.543884   1.590486     2.86   0.006     1.349299    7.738469
      puivmd |   .7504832   1.375308     0.55   0.588    -2.011905    3.512871
      rmfi09 |   -.0000176   4.63e-06   -3.80   0.000    -.0000269   -8.30e-06
```

Results of linear regression results. (t-ratios are in parentheses)

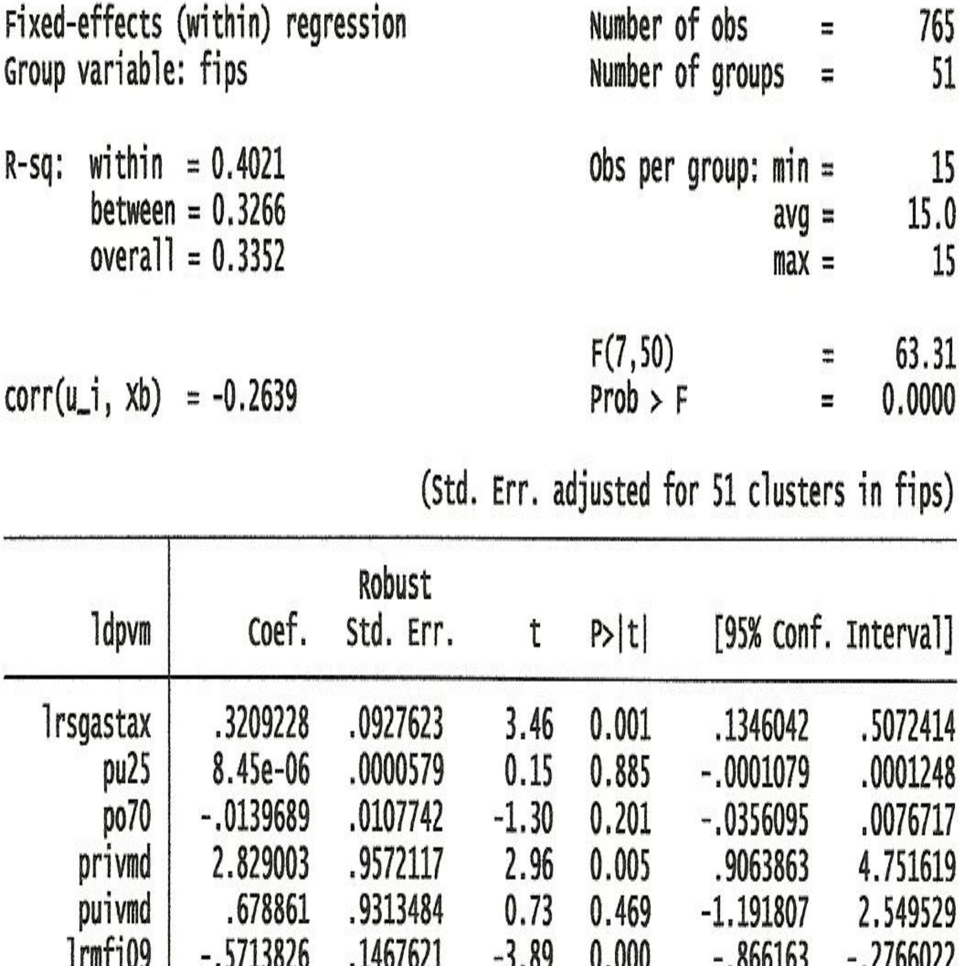
Figure 3.2.



```
Fixed-effects (within) regression               Number of obs      =       765
Group variable: fips                            Number of groups   =        51

R-sq:  within  = 0.3875                         Obs per group: min =        15
       between = 0.4049                                        avg =      15.0
       overall = 0.3979                                        max =        15

                                                F(7,50)            =     53.03
corr(u_i, Xb)  = -0.2303                        Prob > F           =    0.0000

                                  (Std. Err. adjusted for 51 clusters in fips)
------------------------------------------------------------------------------
             |               Robust
       ldpvm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    rsgastax |   .8846191   .4218378     2.10   0.041      .037333    1.731905
        pu25 |   3.34e-06   .0000595     0.06   0.955    -.0001161    .0001228
        po70 |  -.0163277   .0108696    -1.50   0.139    -.0381599    .0055044
      privmd |   3.032937   .9629571     3.15   0.003     1.098781    4.967093
      puivmd |    .640566   .9482832     0.68   0.502    -1.264117    2.545249
      rmfi09 |  -.0000122   2.88e-06    -4.24   0.000     -.000018   -6.45e-06
```

Results of the log-linear regression. (t-ratios are in parentheses)

Figure 3.3.



```
Fixed-effects (within) regression               Number of obs      =       765
Group variable: fips                            Number of groups   =        51

R-sq:  within  = 0.4021                         Obs per group: min =        15
       between = 0.3266                                        avg =      15.0
       overall = 0.3352                                        max =        15

                                                F(7,50)            =     63.31
corr(u_i, Xb)  = -0.2639                        Prob > F           =    0.0000

                                  (Std. Err. adjusted for 51 clusters in fips)
------------------------------------------------------------------------------
             |               Robust
       ldpvm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   lrsgastax |   .3209228   .0927623     3.46   0.001     .1346042    .5072414
        pu25 |   8.45e-06   .0000579     0.15   0.885    -.0001079    .0001248
        po70 |  -.0139689   .0107742    -1.30   0.201    -.0356095    .0076717
      privmd |   2.829003   .9572117     2.96   0.005     .9063863    4.751619
      puivmd |    .678861   .9313484     0.73   0.469    -1.191807    2.549529
     lrmfi09 |  -.5713826   .1467621    -3.89   0.000     -.866163   -.2766022
```

Results of the log-log regression. (t-ratios are in parentheses)

At this point is makes some sense to compare the parameter estimates for 0.08 BAC *per se* law; this comparison, shown in Table 5, suggests that the effect of the *per se* 0.08 BAC law was to reduce fatalities. Moreover, the estimates for each of the models is very stable whether one uses the real price of gasoline or the nominal price of gasoline, thus giving us some more confidence in our conclusions.

	Linear	Log-linear	Log-log
State tax of gasoline in 2009 dollars			
State has a 0.08 *per se* BAC law	-0.1054	-0.0692	-0.0594
	(-3.88)	(-3.67)	(-3.18)
State tax of gasoline in current dollars			
State has a 0.08 *per se* BAC law	-0.1191	-0.0778	-0.0762

	(-4.83)	(-4.54)	(-4.52)

Table 3.5. Comparison of the parameter estimates for each model with different measures of the cost of gasoline.

The balance of this section of the paper would be devoted to further tests of the stability of our results under varying assumptions. Among other tests one would expect to see if the choice of a fixed-effects model affects your policy conclusions.

Exercises

1. **Complete the Lagrange test for random effects for each of the three models, using the nominal price of gasoline. Organize the results of this test into a table.**

2. **Re-estimate the three models replacing the percent of registered drivers under the age of 25 with the percent of drivers under 20. Make the same same kind of replacement for the number of drivers over age 70 (i.e., experiment with the alternative age cutoffs—over 75, over 80, and over 85). Do any of your major conclusions change?**

3. **What, if any, explanation can you give for the differences in the parameter estimates for the price of gasoline generated when the real price of gasoline is replaced by the nominal price of gasoline?**

3.6. Conclusions and further research

This section of your paper should be devoted to a careful recapping of your results and providing suggestions for further research. Such a discussion might include some cautious guesses at why the 0.08 BAC *per se* standard appears to affect driver behavior. The discussion could also include some estimates of the number of lifes saved by the introduction of a *per se* standard.

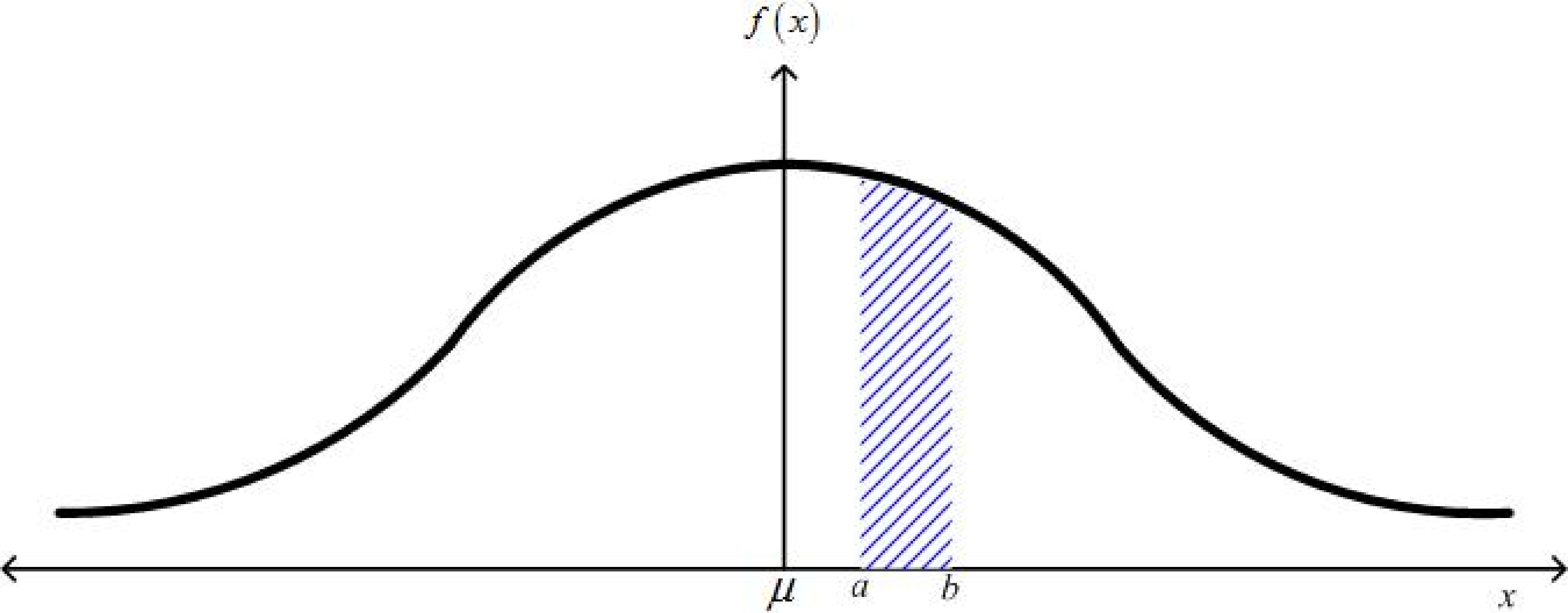
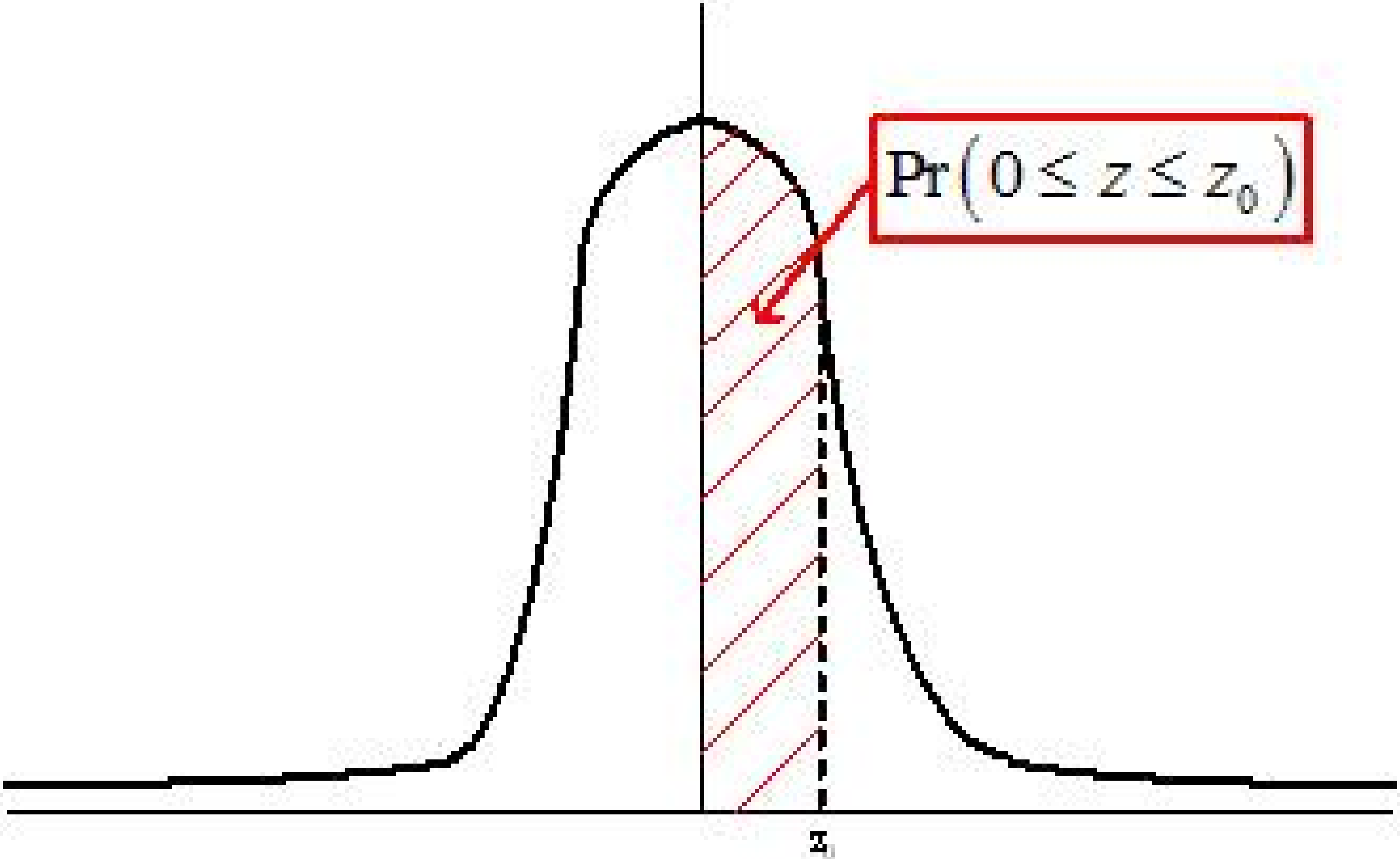
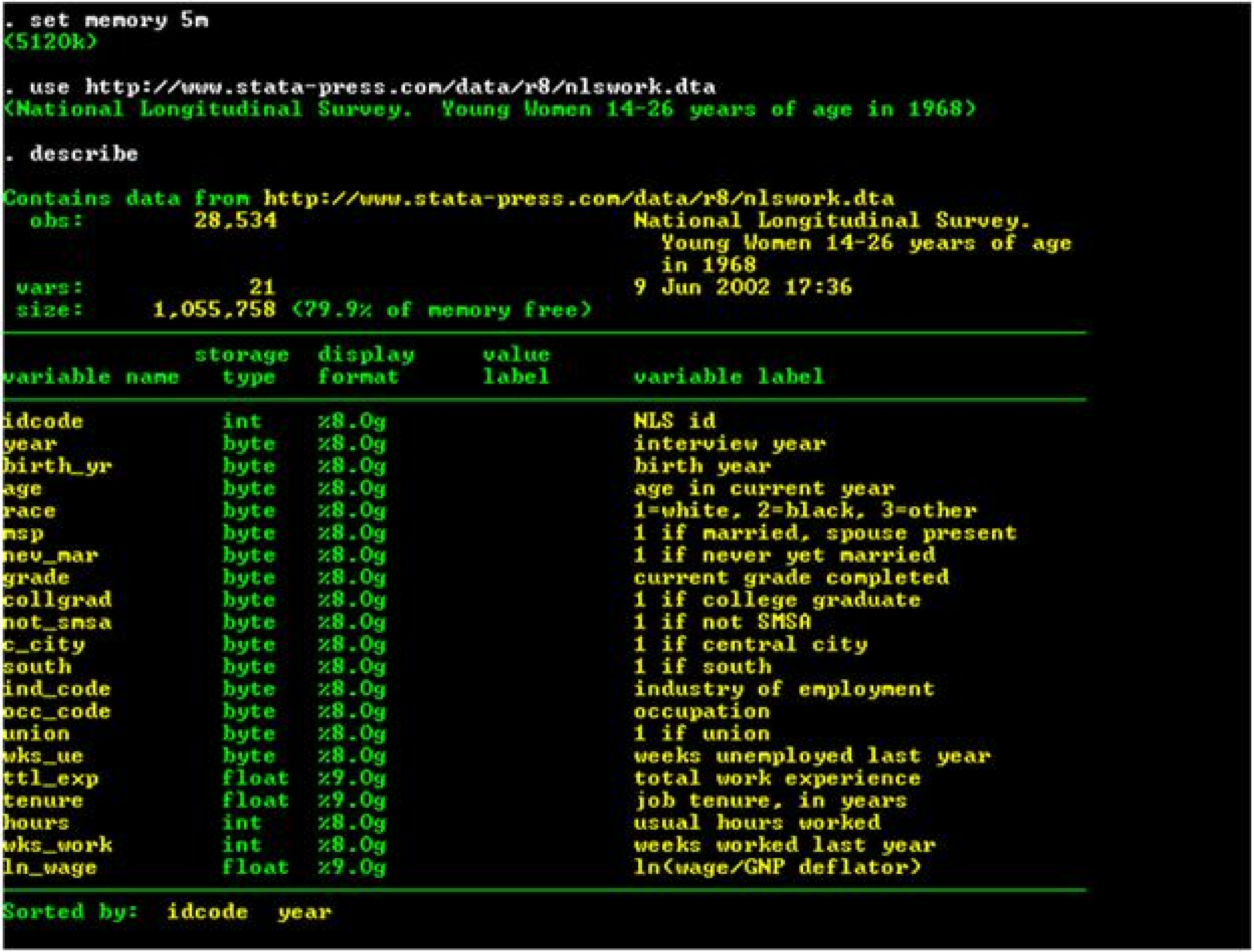

$f(x)$
μ
a
b
x

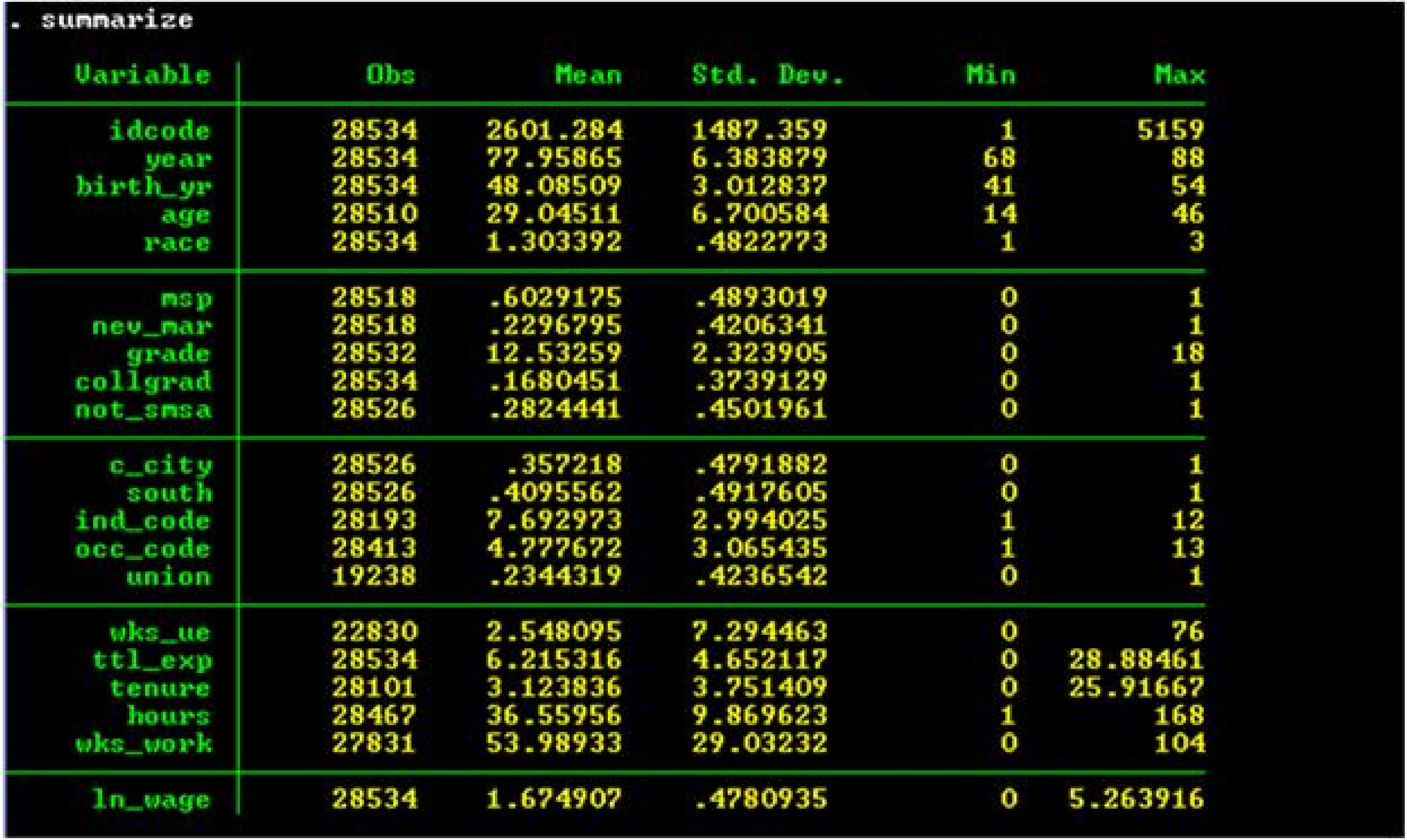
Table 2. Standard Normal Table.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

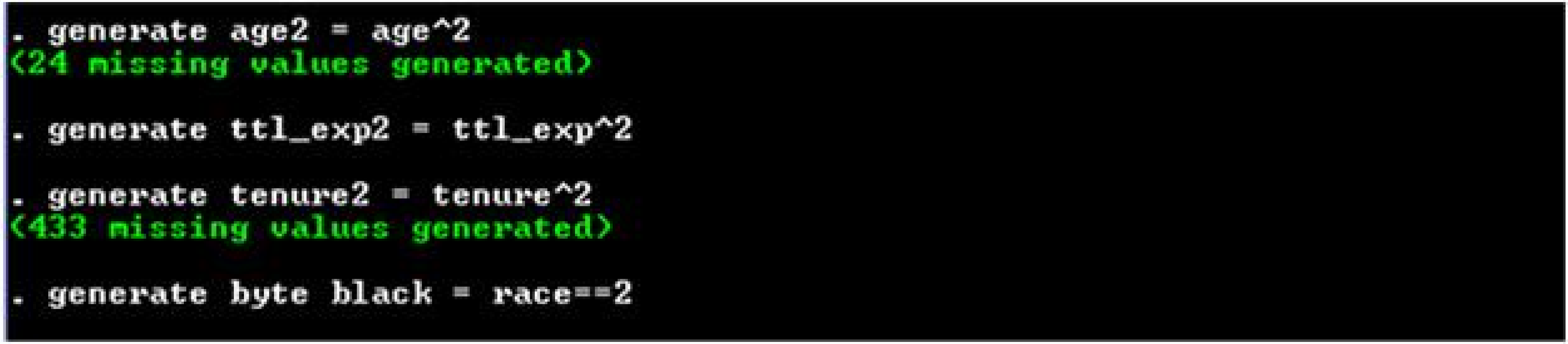

$\Pr(0 \le z \le z_0)$
z_0



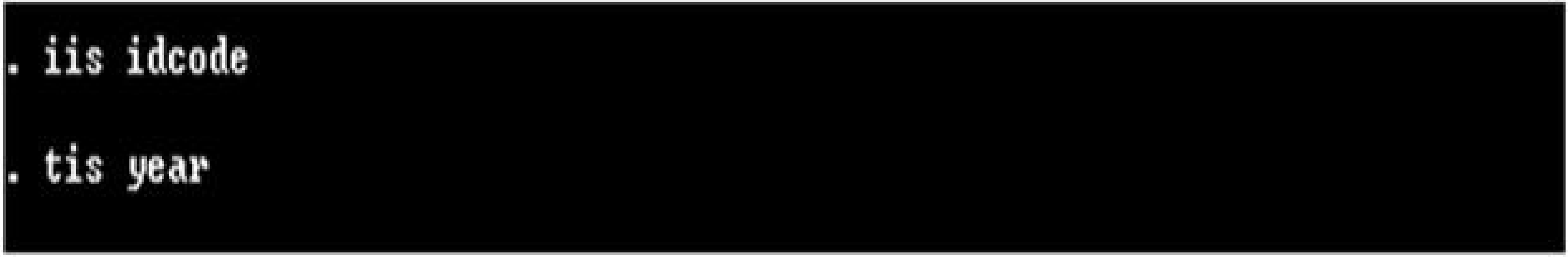
```
. set memory 5m
(5120k)

. use http://www.stata-press.com/data/r8/nlswork.dta
(National Longitudinal Survey.  Young Women 14-26 years of age in 1968)

. describe

Contains data from http://www.stata-press.com/data/r8/nlswork.dta
  obs:        28,534                          National Longitudinal Survey.
                                                Young Women 14-26 years of age
                                                in 1968
 vars:            21                          9 Jun 2002 17:36
 size:     1,055,758 (79.9% of memory free)
-------------------------------------------------------------------------------
                storage  display     value
variable name     type   format      label      variable label
-------------------------------------------------------------------------------
idcode            int    %8.0g                  NLS id
year              byte   %8.0g                  interview year
birth_yr          byte   %8.0g                  birth year
age               byte   %8.0g                  age in current year
race              byte   %8.0g                  1=white, 2=black, 3=other
msp               byte   %8.0g                  1 if married, spouse present
nev_mar           byte   %8.0g                  1 if never yet married
grade             byte   %8.0g                  current grade completed
collgrad          byte   %8.0g                  1 if college graduate
not_smsa          byte   %8.0g                  1 if not SMSA
c_city            byte   %8.0g                  1 if central city
south             byte   %8.0g                  1 if south
ind_code          byte   %8.0g                  industry of employment
occ_code          byte   %8.0g                  occupation
union             byte   %8.0g                  1 if union
wks_ue            byte   %8.0g                  weeks unemployed last year
ttl_exp           float  %9.0g                  total work experience
tenure            float  %9.0g                  job tenure, in years
hours             int    %8.0g                  usual hours worked
wks_work          int    %8.0g                  weeks worked last year
ln_wage           float  %9.0g                  ln(wage/GNP deflator)
-------------------------------------------------------------------------------
Sorted by:  idcode  year
```



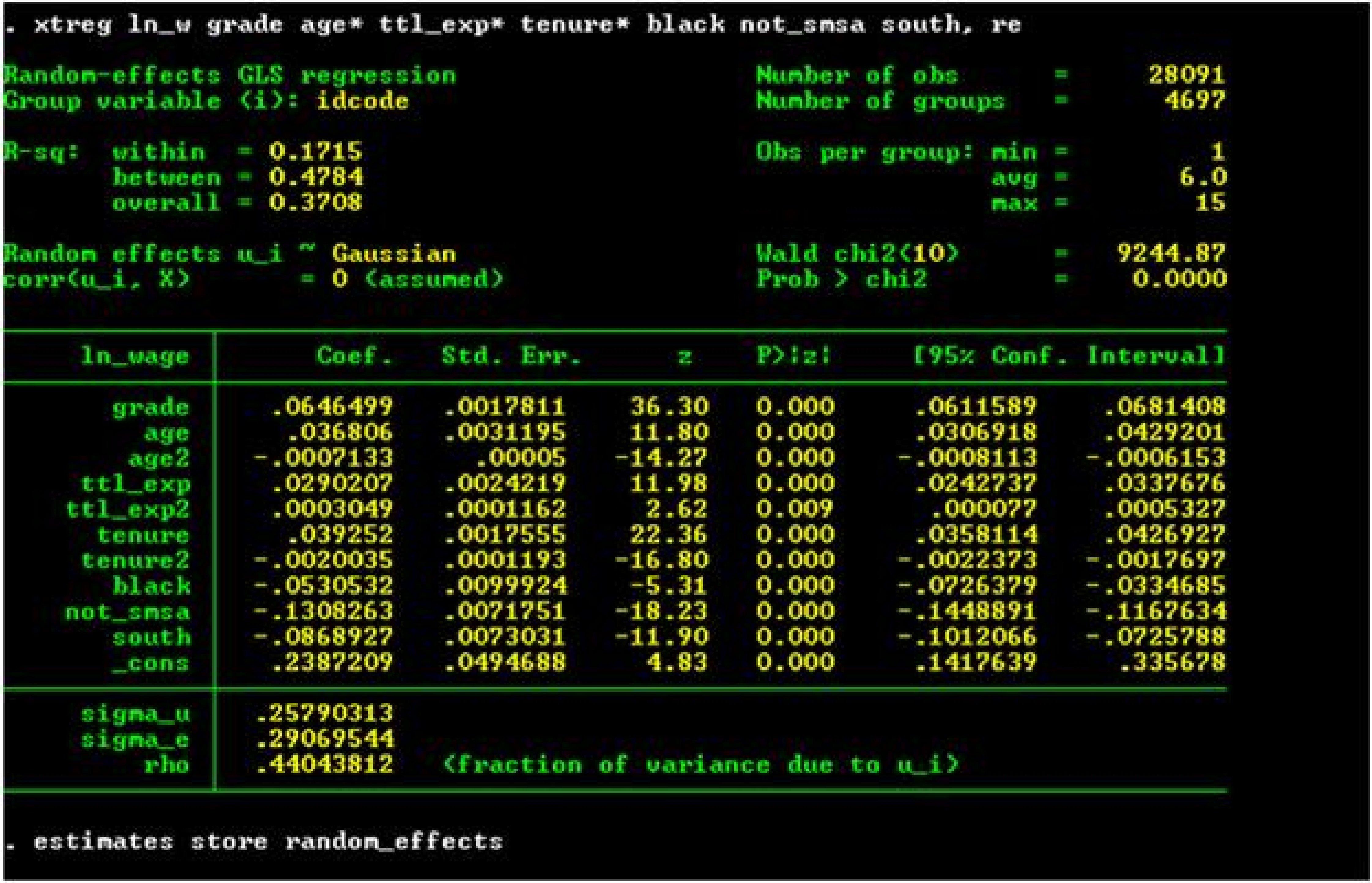
```
. summarize

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      idcode |     28534    2601.284    1487.359          1       5159
        year |     28534    77.95865    6.383879         68         88
    birth_yr |     28534    48.08509    3.012837         41         54
         age |     28510    29.04511    6.700584         14         46
        race |     28534    1.303392    .4822773          1          3
-------------+--------------------------------------------------------
         msp |     28518    .6029175    .4893019          0          1
     nev_mar |     28518    .2296795    .4206341          0          1
       grade |     28532    12.53259    2.323905          0         18
    collgrad |     28534    .1680451    .3739129          0          1
    not_smsa |     28526    .2824441    .4501961          0          1
-------------+--------------------------------------------------------
      c_city |     28526     .357218    .4791882          0          1
       south |     28526    .4095562    .4917605          0          1
    ind_code |     28193    7.692973    2.994025          1         12
    occ_code |     28413    4.777672    3.065435          1         13
       union |     19238    .2344319    .4236542          0          1
-------------+--------------------------------------------------------
      wks_ue |     22830    2.548095    7.294463          0         76
     ttl_exp |     28534    6.215316    4.652117          0   28.88461
      tenure |     28101    3.123836    3.751409          0   25.91667
       hours |     28467    36.55956    9.869623          1        168
    wks_work |     27831    53.98933    29.03232          0        104
-------------+--------------------------------------------------------
     ln_wage |     28534    1.674907    .4780935          0   5.263916
```



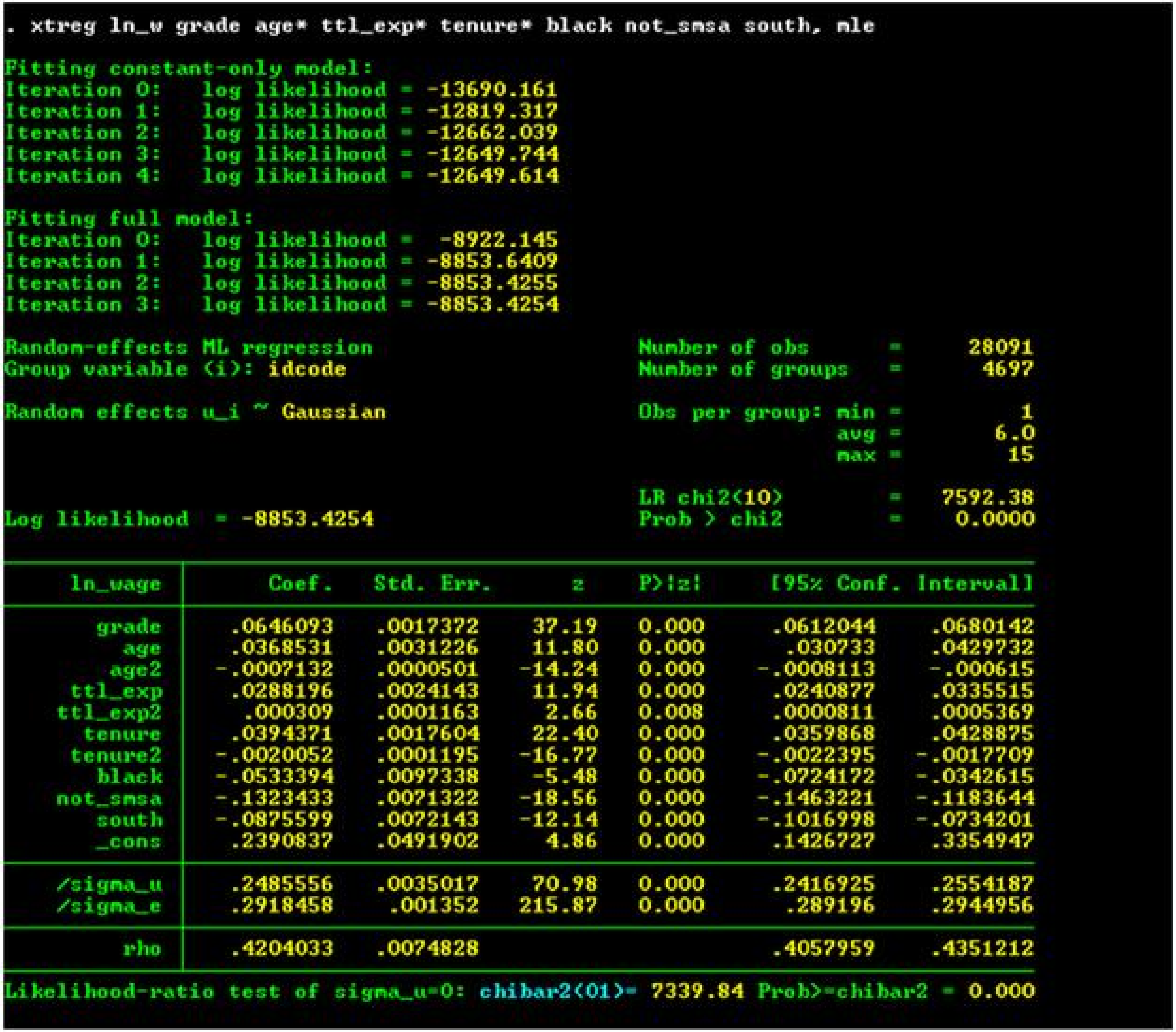
```
. generate age2 = age^2
(24 missing values generated)

. generate ttl_exp2 = ttl_exp^2

. generate tenure2 = tenure^2
(433 missing values generated)

. generate byte black = race==2
```



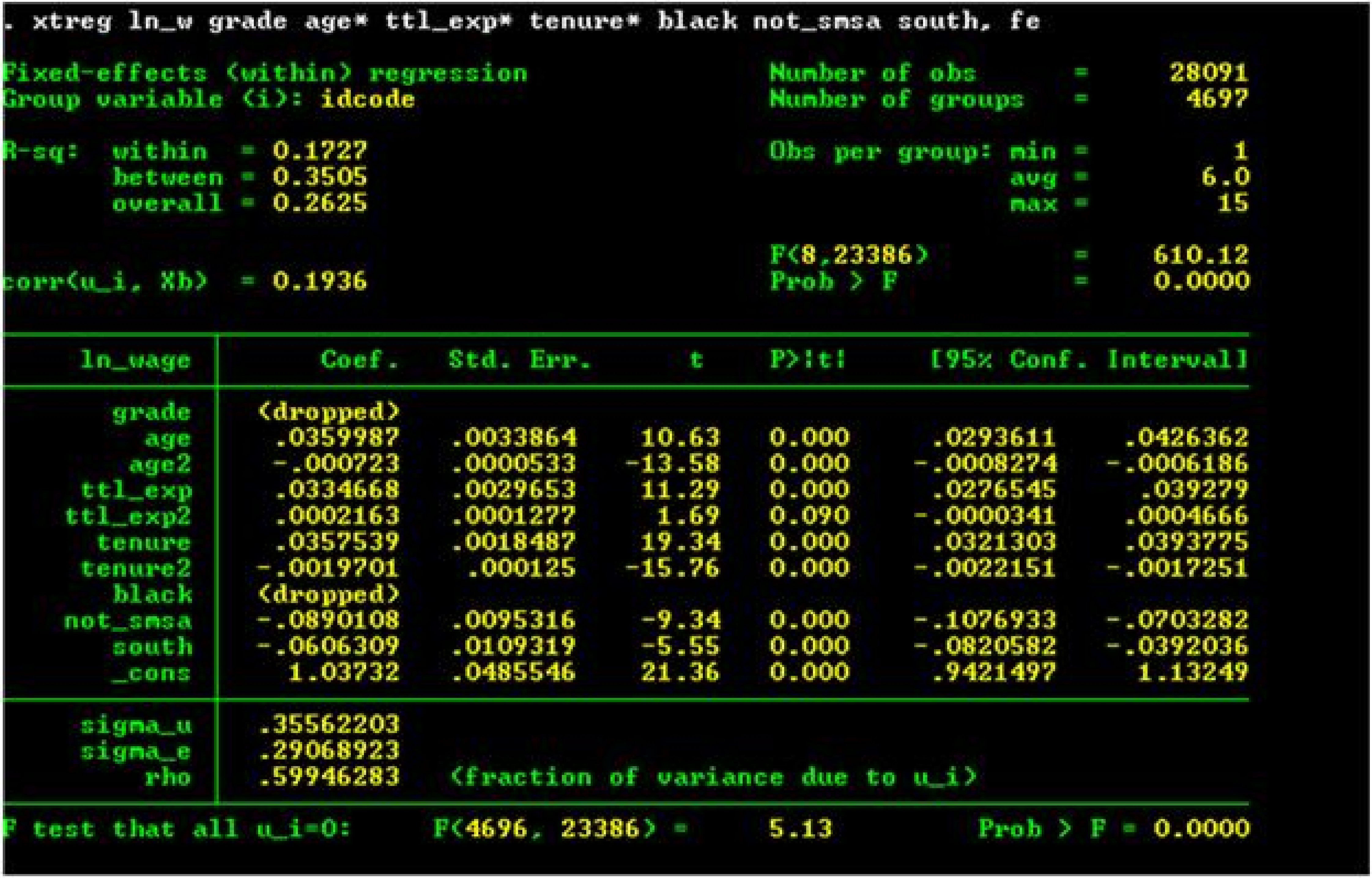
```
. iis idcode

. tis year
```



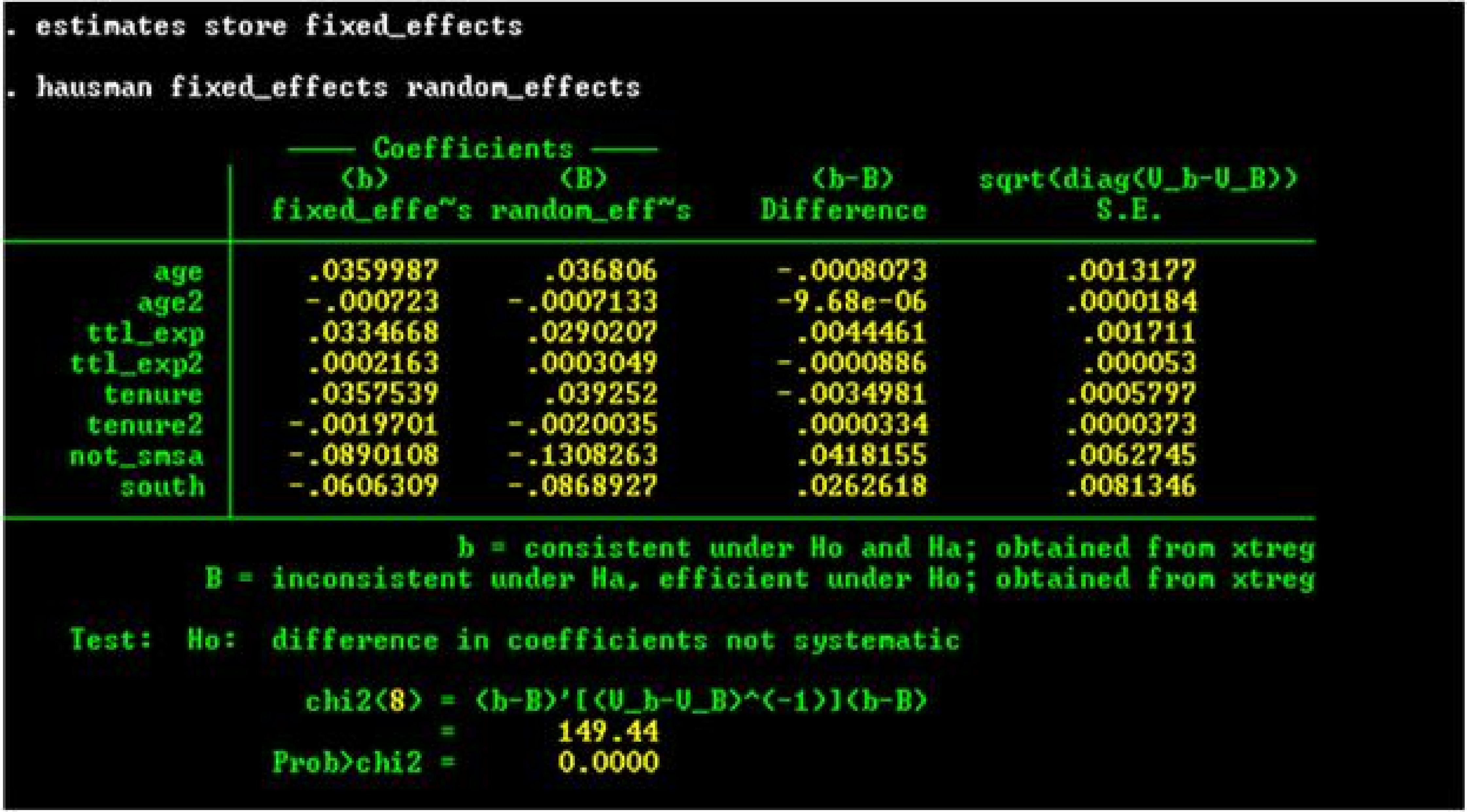
```
. xtreg ln_w grade age* ttl_exp* tenure* black not_smsa south, re

Random-effects GLS regression                   Number of obs      =     28091
Group variable (i): idcode                      Number of groups   =      4697

R-sq:  within  = 0.1715                         Obs per group: min =         1
       between = 0.4784                                        avg =       6.0
       overall = 0.3708                                        max =        15

Random effects u_i ~ Gaussian                   Wald chi2(10)      =   9244.87
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
     ln_wage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       grade |   .0646499   .0017811    36.30   0.000     .0611589    .0681408
         age |    .036806   .0031195    11.80   0.000     .0306918    .0429201
        age2 |  -.0007133     .00005   -14.27   0.000    -.0008113   -.0006153
     ttl_exp |   .0290207   .0024219    11.98   0.000     .0242737    .0337676
    ttl_exp2 |   .0003049   .0001162     2.62   0.009      .000077    .0005327
      tenure |    .039252   .0017555    22.36   0.000     .0358114    .0426927
     tenure2 |  -.0020035   .0001193   -16.80   0.000    -.0022373   -.0017697
       black |  -.0530532   .0099924    -5.31   0.000    -.0726379   -.0334685
    not_smsa |  -.1308263   .0071751   -18.23   0.000    -.1448891   -.1167634
       south |  -.0868927   .0073031   -11.90   0.000    -.1012066   -.0725788
       _cons |   .2387209   .0494688     4.83   0.000     .1417639     .335678
-------------+----------------------------------------------------------------
     sigma_u |  .25790313
     sigma_e |  .29069544
         rho |  .44043812   (fraction of variance due to u_i)
------------------------------------------------------------------------------

. estimates store random_effects
```



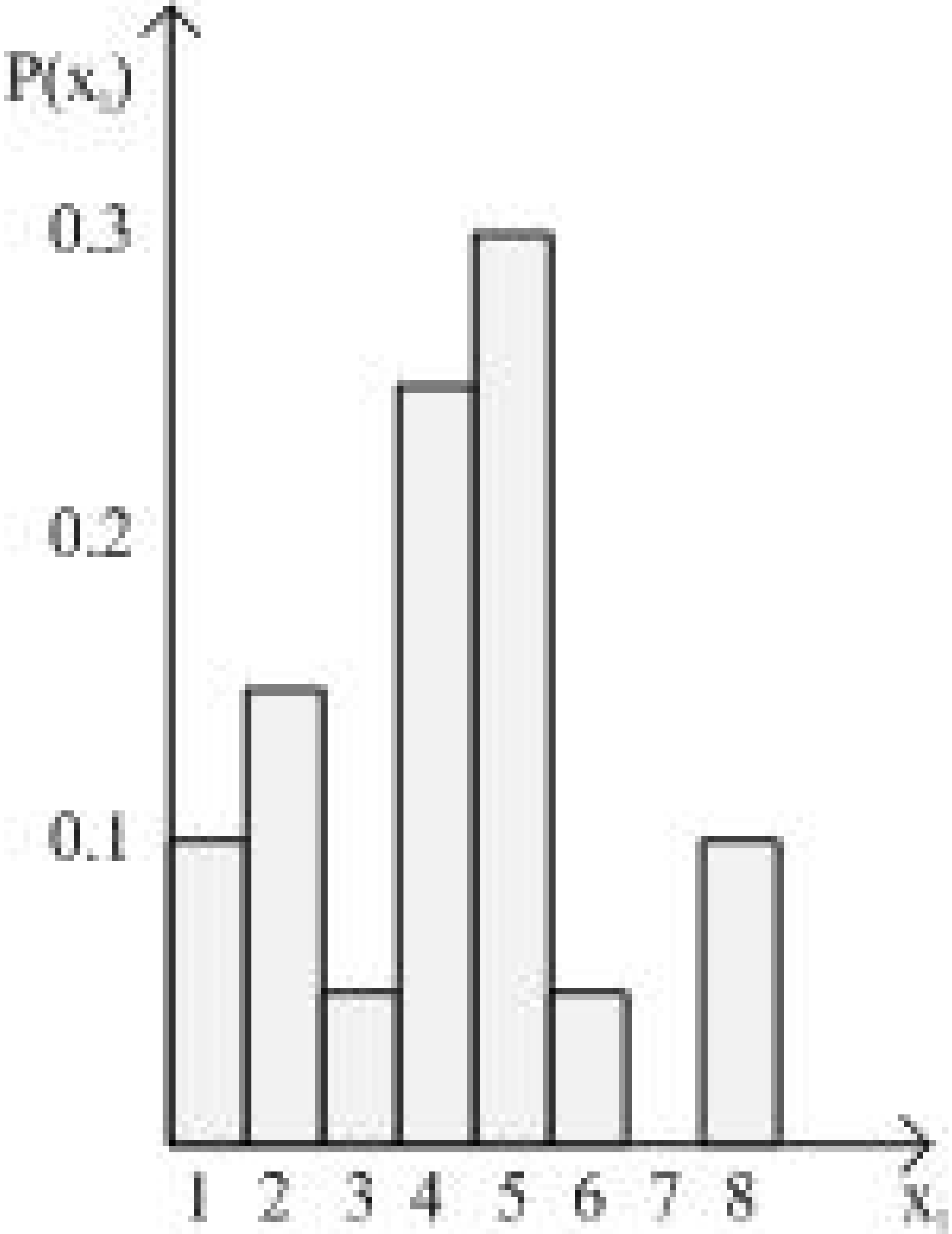
```
. xtreg ln_w grade age* ttl_exp* tenure* black not_smsa south, mle

Fitting constant-only model:
Iteration 0:   log likelihood = -13690.161
Iteration 1:   log likelihood = -12819.317
Iteration 2:   log likelihood = -12662.039
Iteration 3:   log likelihood = -12649.744
Iteration 4:   log likelihood = -12649.614

Fitting full model:
Iteration 0:   log likelihood =  -8922.145
Iteration 1:   log likelihood = -8853.6409
Iteration 2:   log likelihood = -8853.4255
Iteration 3:   log likelihood = -8853.4254

Random-effects ML regression                    Number of obs      =     28091
Group variable (i): idcode                      Number of groups   =      4697

Random effects u_i ~ Gaussian                   Obs per group: min =         1
                                                               avg =       6.0
                                                               max =        15

                                                LR chi2(10)        =   7592.38
Log likelihood  = -8853.4254                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
     ln_wage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       grade |   .0646093   .0017372    37.19   0.000     .0612044    .0680142
         age |   .0368531   .0031226    11.80   0.000      .030733    .0429732
        age2 |  -.0007132   .0000501   -14.24   0.000    -.0008113    -.000615
     ttl_exp |   .0288196   .0024143    11.94   0.000     .0240877    .0335515
    ttl_exp2 |    .000309   .0001163     2.66   0.008     .0000811    .0005369
      tenure |   .0394371   .0017604    22.40   0.000     .0359868    .0428875
     tenure2 |  -.0020052   .0001195   -16.77   0.000    -.0022395   -.0017709
       black |  -.0533394   .0097338    -5.48   0.000    -.0724172   -.0342615
    not_smsa |  -.1323433   .0071322   -18.56   0.000    -.1463221   -.1183644
       south |  -.0875599   .0072143   -12.14   0.000    -.1016998   -.0734201
       _cons |   .2390837   .0491902     4.86   0.000     .1426727    .3354947
-------------+----------------------------------------------------------------
    /sigma_u |   .2485556   .0035017    70.98   0.000     .2416925    .2554187
    /sigma_e |   .2918458    .001352   215.87   0.000      .289196    .2944956
-------------+----------------------------------------------------------------
         rho |   .4204033   .0074828                      .4057959    .4351212
------------------------------------------------------------------------------
Likelihood-ratio test of sigma_u=0: chibar2(01)= 7339.84 Prob>=chibar2 = 0.000
```

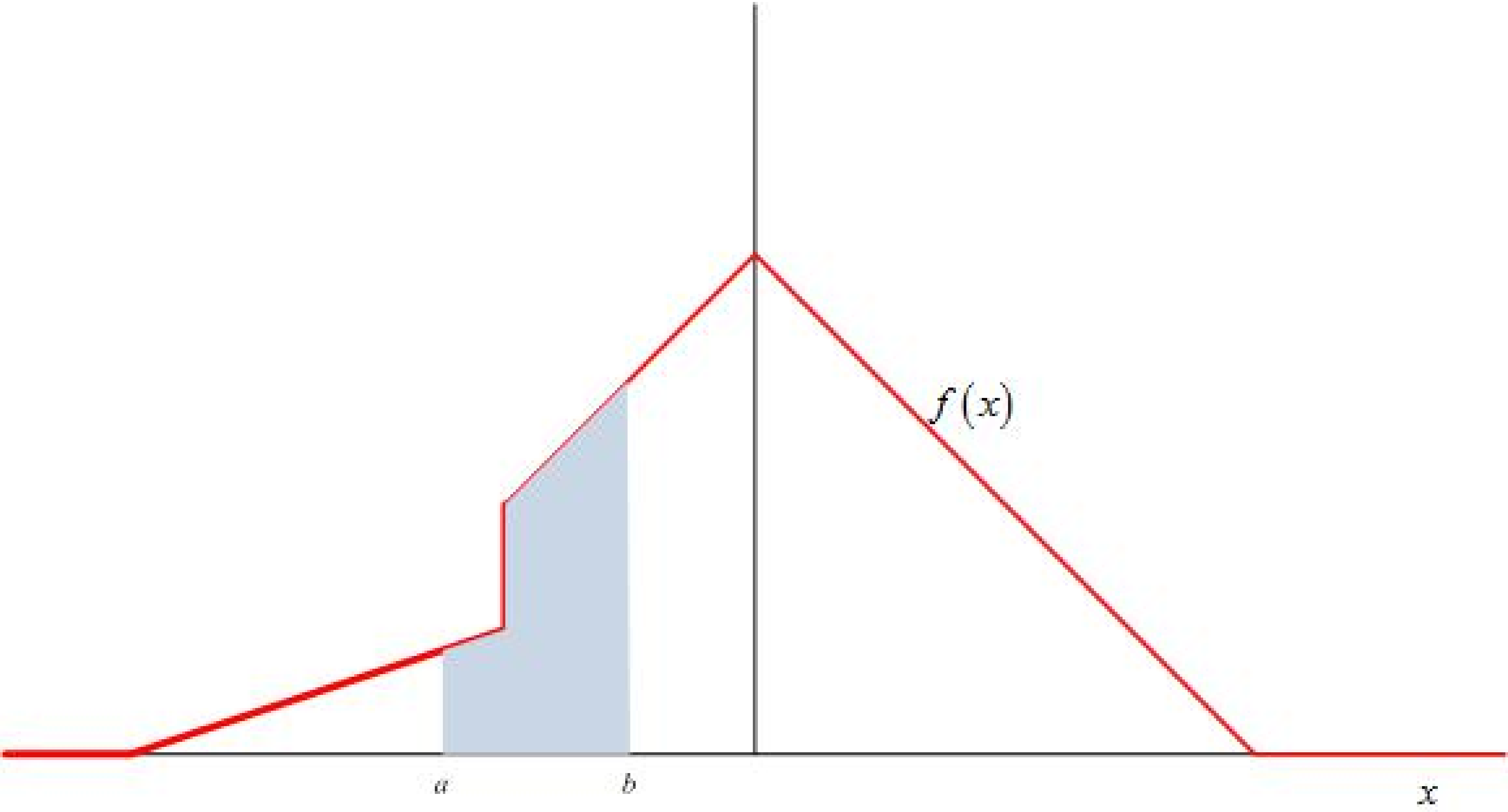


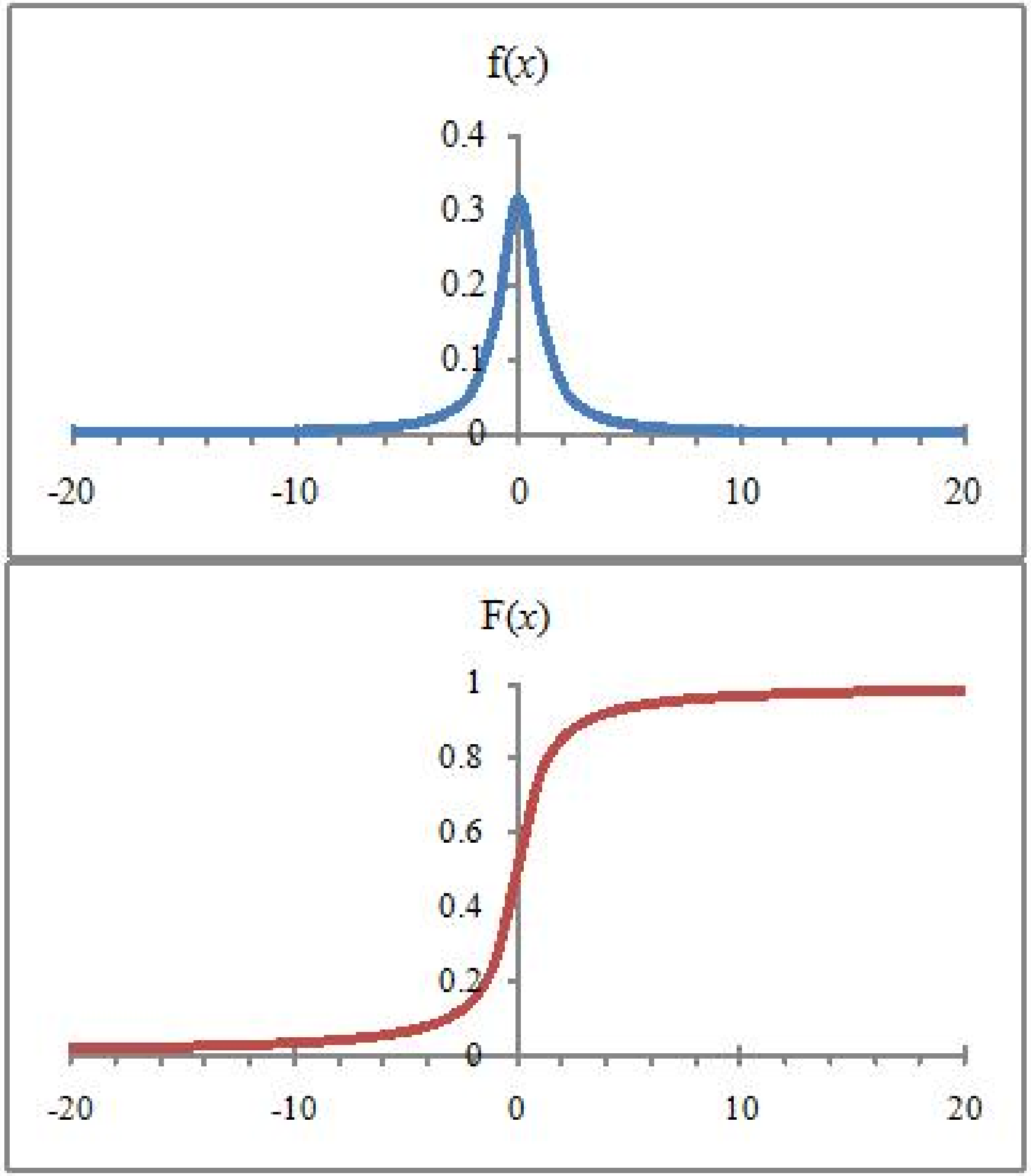
```
. xtreg ln_w grade age* ttl_exp* tenure* black not_smsa south, fe

Fixed-effects (within) regression               Number of obs      =     28091
Group variable (i): idcode                      Number of groups   =      4697

R-sq:  within  = 0.1727                         Obs per group: min =         1
       between = 0.3505                                        avg =       6.0
       overall = 0.2625                                        max =        15

                                                F(8,23386)         =    610.12
corr(u_i, Xb)  = 0.1936                         Prob > F           =    0.0000

------------------------------------------------------------------------------
     ln_wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       grade |  (dropped)
         age |   .0359987   .0033864    10.63   0.000     .0293611    .0426362
        age2 |   -.000723   .0000533   -13.58   0.000    -.0008274   -.0006186
     ttl_exp |   .0334668   .0029653    11.29   0.000     .0276545     .039279
    ttl_exp2 |   .0002163   .0001277     1.69   0.090    -.0000341    .0004666
      tenure |   .0357539   .0018487    19.34   0.000     .0321303    .0393775
     tenure2 |  -.0019701    .000125   -15.76   0.000    -.0022151   -.0017251
       black |  (dropped)
    not_smsa |  -.0890108   .0095316    -9.34   0.000    -.1076933   -.0703282
       south |  -.0606309   .0109319    -5.55   0.000    -.0820582   -.0392036
       _cons |    1.03732   .0485546    21.36   0.000     .9421497     1.13249
-------------+----------------------------------------------------------------
     sigma_u |  .35562203
     sigma_e |  .29068923
         rho |  .59946283   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(4696, 23386) =     5.13         Prob > F = 0.0000
```



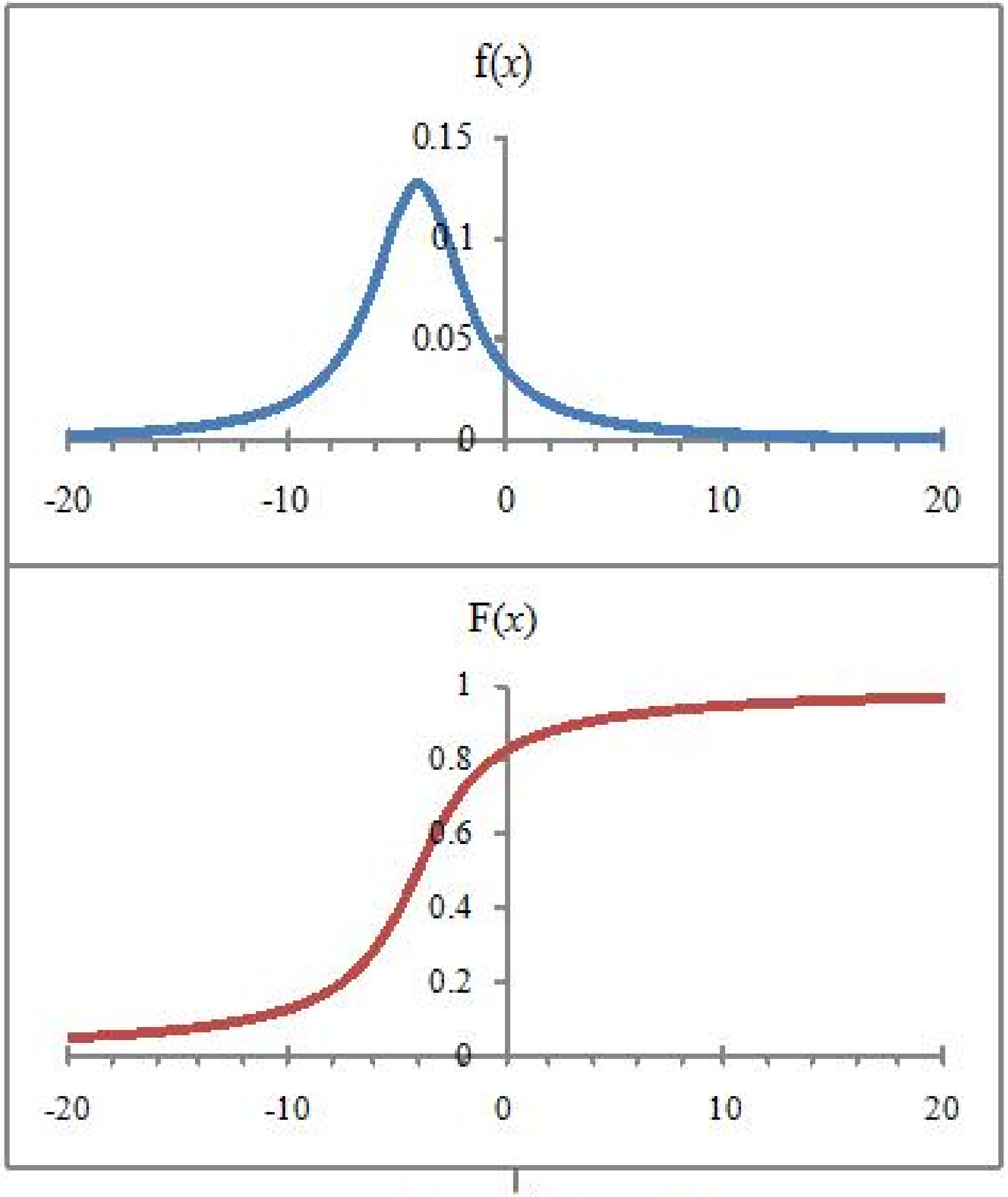
```
. estimates store fixed_effects

. hausman fixed_effects random_effects

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             | fixed_effe~s random_eff~s    Difference          S.E.
-------------+----------------------------------------------------------------
         age |    .0359987      .036806       -.0008073        .0013177
        age2 |    -.000723    -.0007133       -9.68e-06        .0000184
     ttl_exp |    .0334668     .0290207        .0044461         .001711
    ttl_exp2 |    .0002163     .0003049       -.0000886         .000053
      tenure |    .0357539      .039252       -.0034981        .0005797
     tenure2 |   -.0019701    -.0020035        .0000334        .0000373
    not_smsa |   -.0890108    -.1308263        .0418155        .0062745
       south |   -.0606309    -.0868927        .0262618        .0081346
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(8) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =      149.44
                Prob>chi2 =      0.0000
```


$P(x_i)$
0.3
0.2
0.1
1
2
3
4
5
6
7
8
x_i

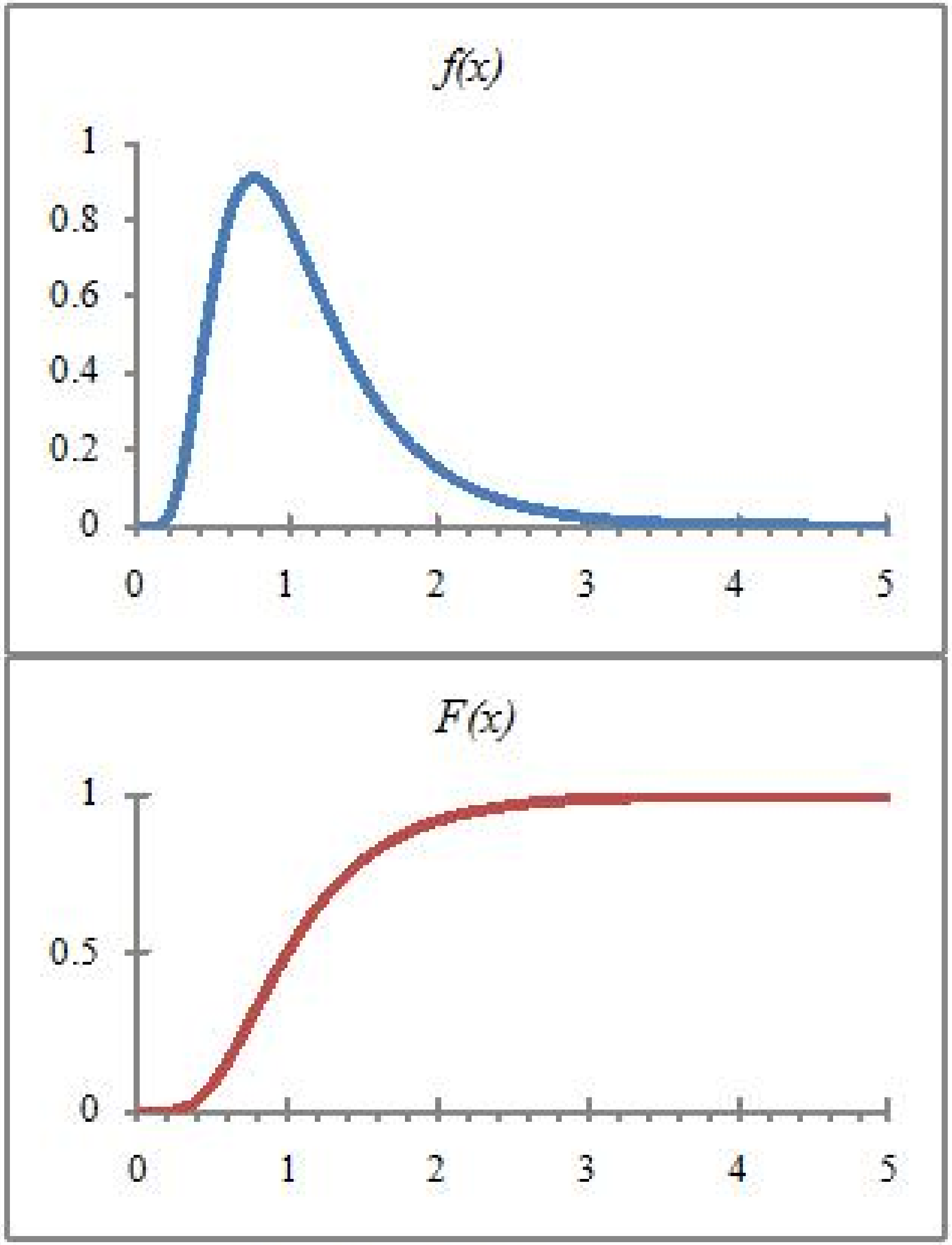

$f\left(x\right)$
a
b
x



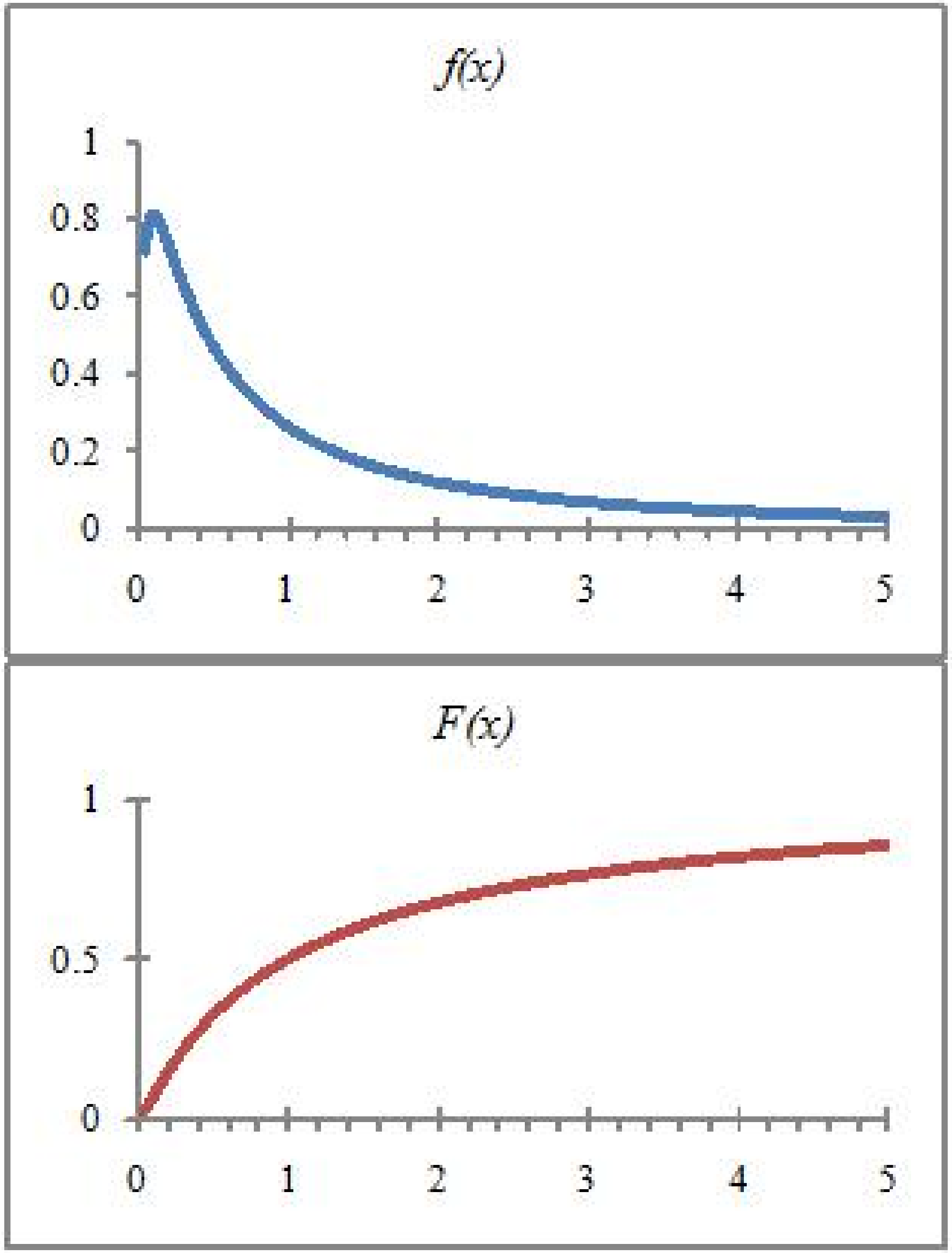
$\gamma = 1$ and $x_0 = 0$



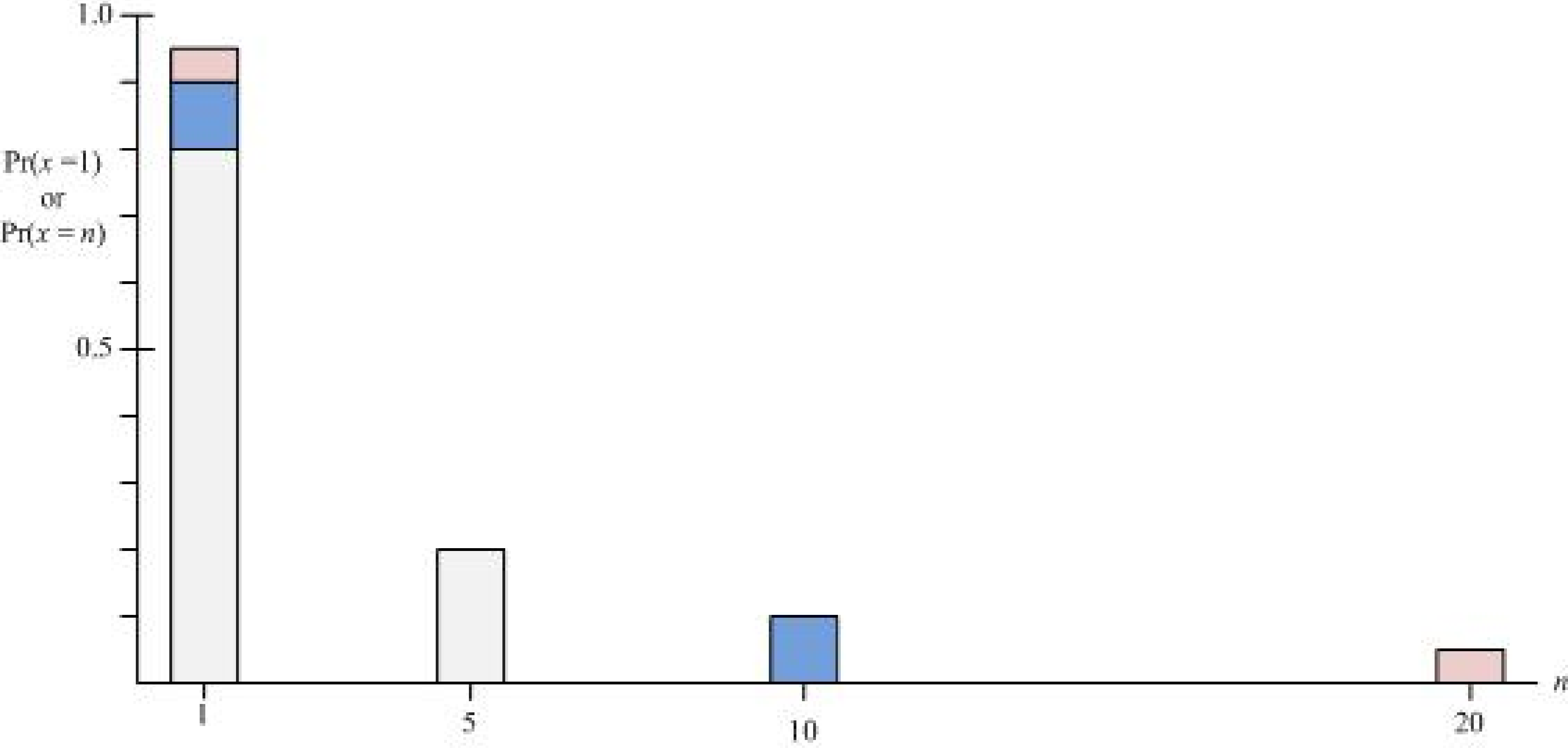
$\gamma = 2.5$ and $x_0 = -4$

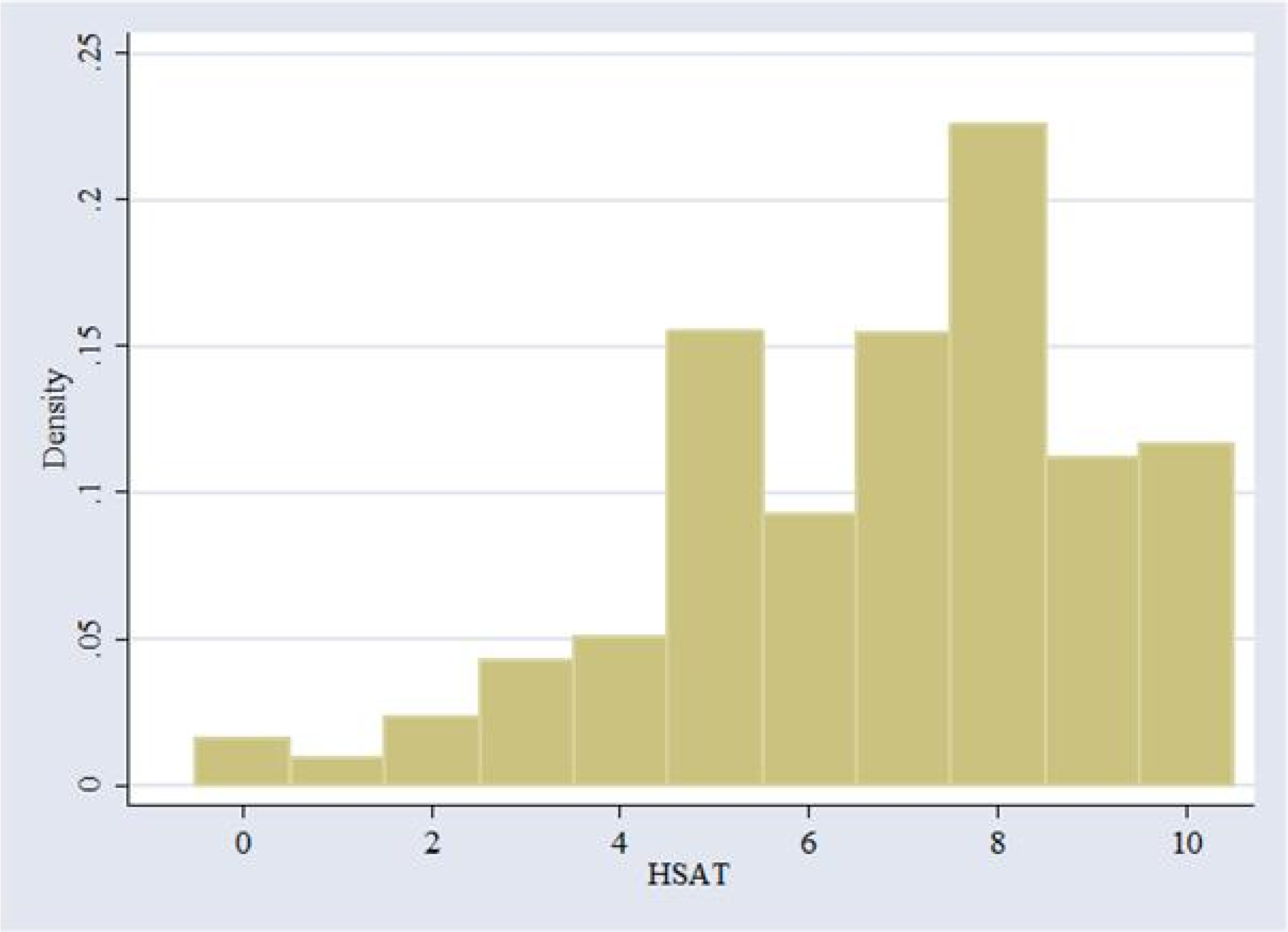


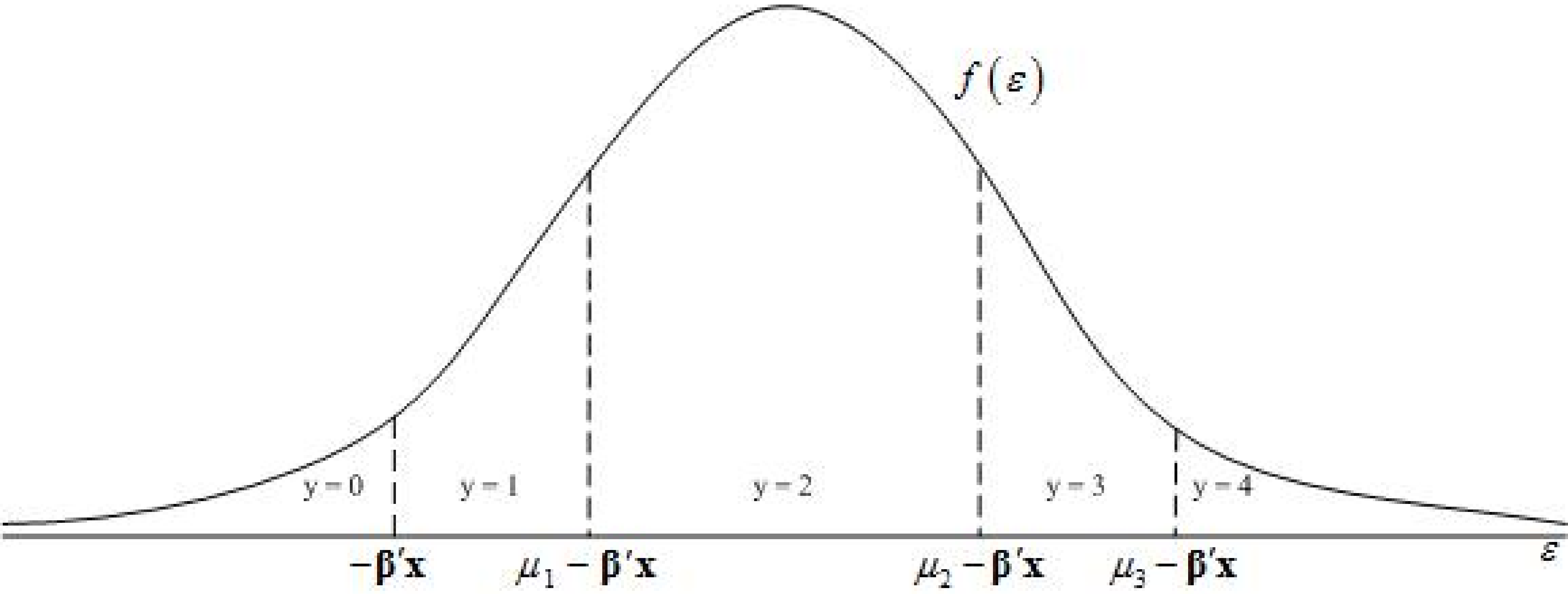
$\mu = 0$ and σ=0.5

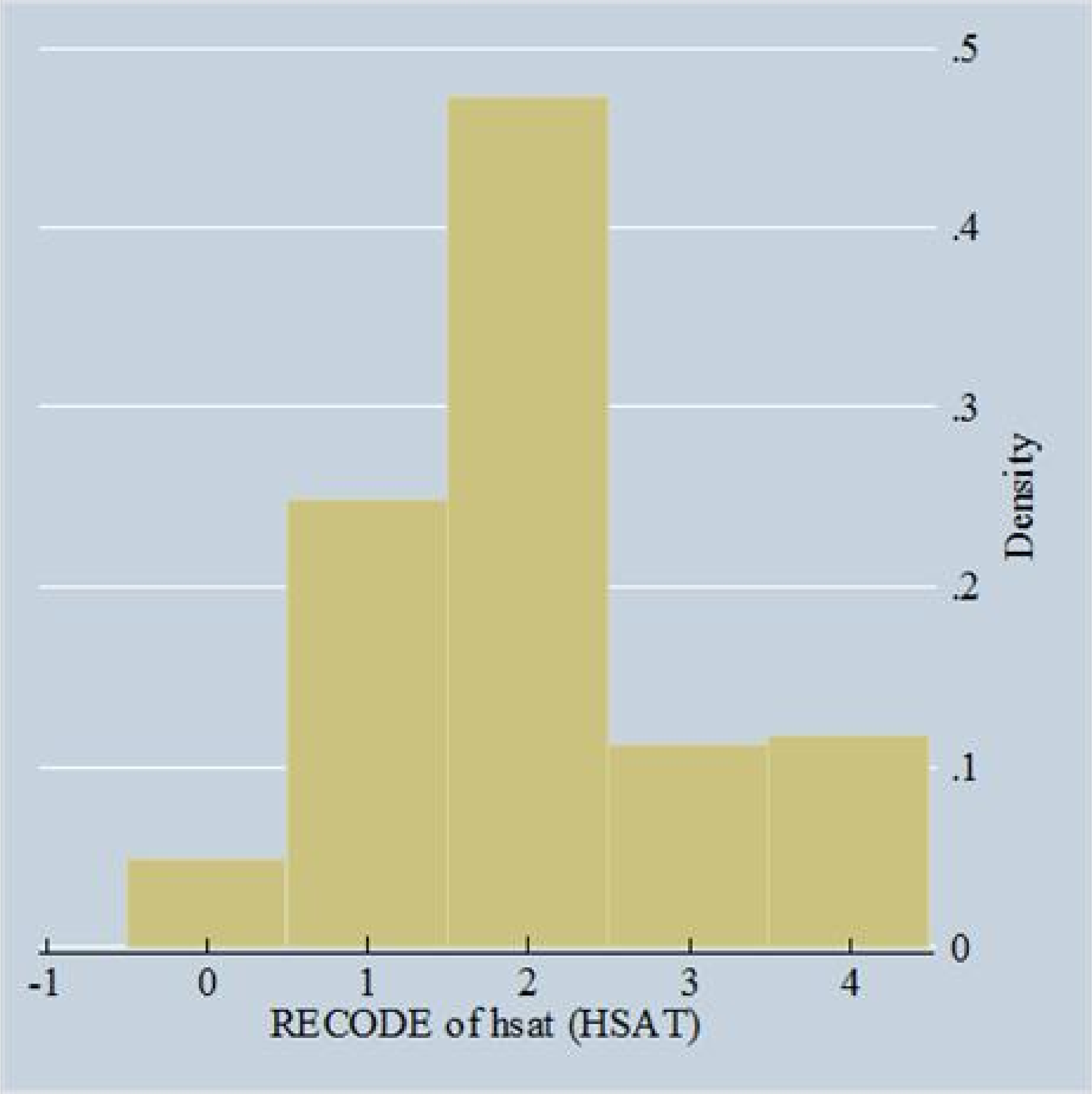


$\mu = 0$ and σ=1.5


1.0
0.5
Pr(x =1)
or
Pr(x = n)
1
5
10
20
n


Density
0
.05
.1
.15
.2
.25
0
2
4
6
8
10
HSAT


$f(\varepsilon)$
y = 0
y = 1
y = 2
y = 3
y = 4
$-\boldsymbol{\beta}'\mathbf{x}$
$\mu_1 - \boldsymbol{\beta}'\mathbf{x}$
$\mu_2 - \boldsymbol{\beta}'\mathbf{x}$
$\mu_3 - \boldsymbol{\beta}'\mathbf{x}$
ε


.5
.4
.3
.2
.1
0
Density
-1
0
1
2
3
4
RECODE of hsat (HSAT)

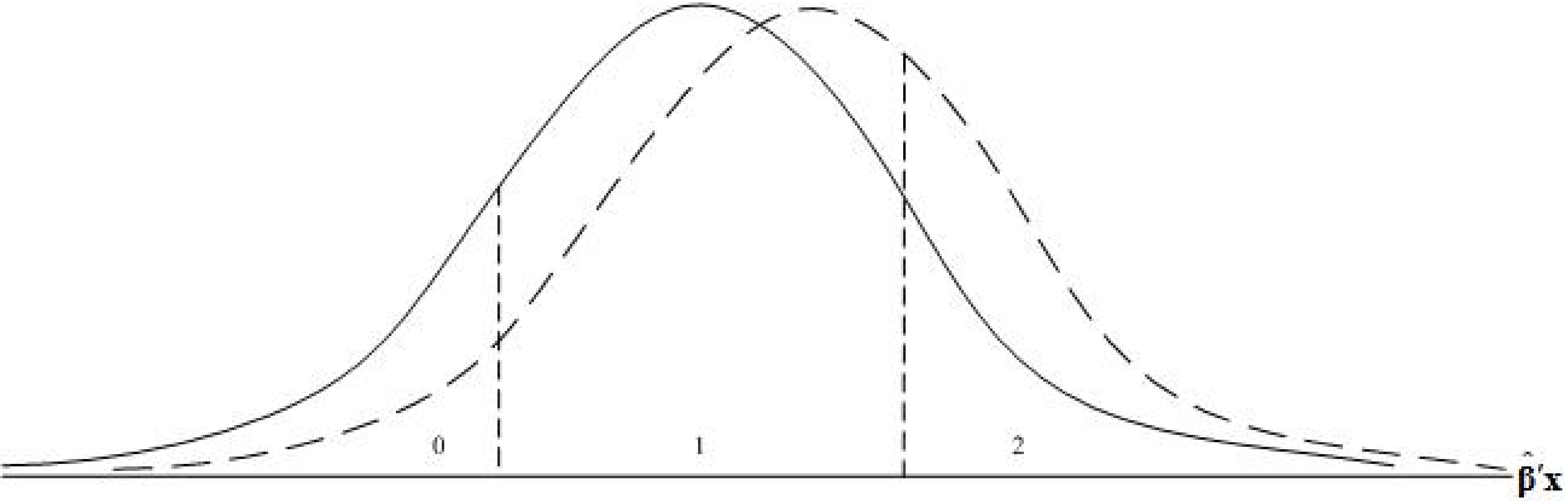

0
1
2
$\hat{\beta}'\mathbf{x}$

Table II. Ordered probit estimates of level of calculus attained[a]

		Students taking MICRO-2		Students taking MACRO-2	
Variable[b]	Expected sign	Mean (SD)	Coefficient (t-value)	Mean (SD)	Coefficient (t-value)
Constant		—	−3·09 (5·48)	—	−2·62 (3·95)
SAT-math × 10^{-2}	+	6·25 (0·60)	0·50[d] (6·12)	6·25 (0·60)	0·48[d] (5·23)
Foreign lang. proficiency [1,0]	−	0·11 (0·32)	0·02 (0·14)	0·09 (0·29)	0·23 (1·22)
Sex (female = 1; male = 0)	?	0·39 (0·49)	0·25[d] (2·59)	0·36 (0·48)	0·22[e] (1·96)
Expected major:					
Economics	?	0·34 (0·48)	−0·11 (0·86)	0·36 (0·48)	−0·18 (1·31)
Other social science	?	0·17 (0·38)	−0·29[e] (1·99)	0·15 (0·36)	−0·27 (1·59)
Natural science	+	0·21 (0·41)	0·43[d] (3·10)	0·20 (0·40)	0·32[e] (2·05)
Humanities	−	0·07 (0·25)	−0·37[e] (1·78)	0·07 (0·26)	−0·39[e] (1·80)
Years of HS Advanced Math (Y_m)					
$1 \leqslant Y_m < 2$	+	0·49 (0·50)	0·24 (1·07)	0·49 (0·50)	−0·00 (0·02)
$Y_m = 2$	+	0·45 (0·50)	0·93[d] (4·04)	0·45 (0·50)	0·67[d] (2·83)
$Y_m > 2$	+	0·01 (0·11)	0·77[e] (1·70)	0·01 (0·11)	0·28 (0·55)
Years of HS physics (Y_p)					
$1 \leqslant Y_p < 2$	+	0·67 (0·47)	0·26[d] (2·71)	0·67 (0·47)	0·27[d] (2·50)
$Y_p \geqslant 2$	+	0·02 (0·14)	0·38 (1·07)	0·01 (0·11)	−0·11 (0·20)
Years of HS chemistry (Y_c)					
$1 \leqslant Y_c < 2$	+	0·82 (0·39)	−0·12 (0·69)	0·82 (0·39)	−0·18 (0·75)
$Y_c \geqslant 2$	+	0·12 (0·32)	0·17 (0·75)	0·13 (0·34)	0·20 (0·75)
TRUNCATION POINTS[c]					
(1)	+		0·27[d] (7·29)		0·21[d] (5·59)
(2)	+		0·33[d] (8·16)		0·27[d] (6·46)
(3)	+		1·52[d] (20·32)		1·55[d] (18·26)
(4)	+		1·79[d] (23·07)		1·88[d] (20·73)
(5)	+		2·04[d] (23·72)		2·15[d] (20·58)
OVERALL RESULTS					
Log likelihood			−886·67		−698·09
Outcomes predicted correctly			37·9%		41·2%
Number of Observations			609		490

[a] The dependent variable is the level of calculus attained, as shown by the ordered probit ranking in the lower panel of Table I.

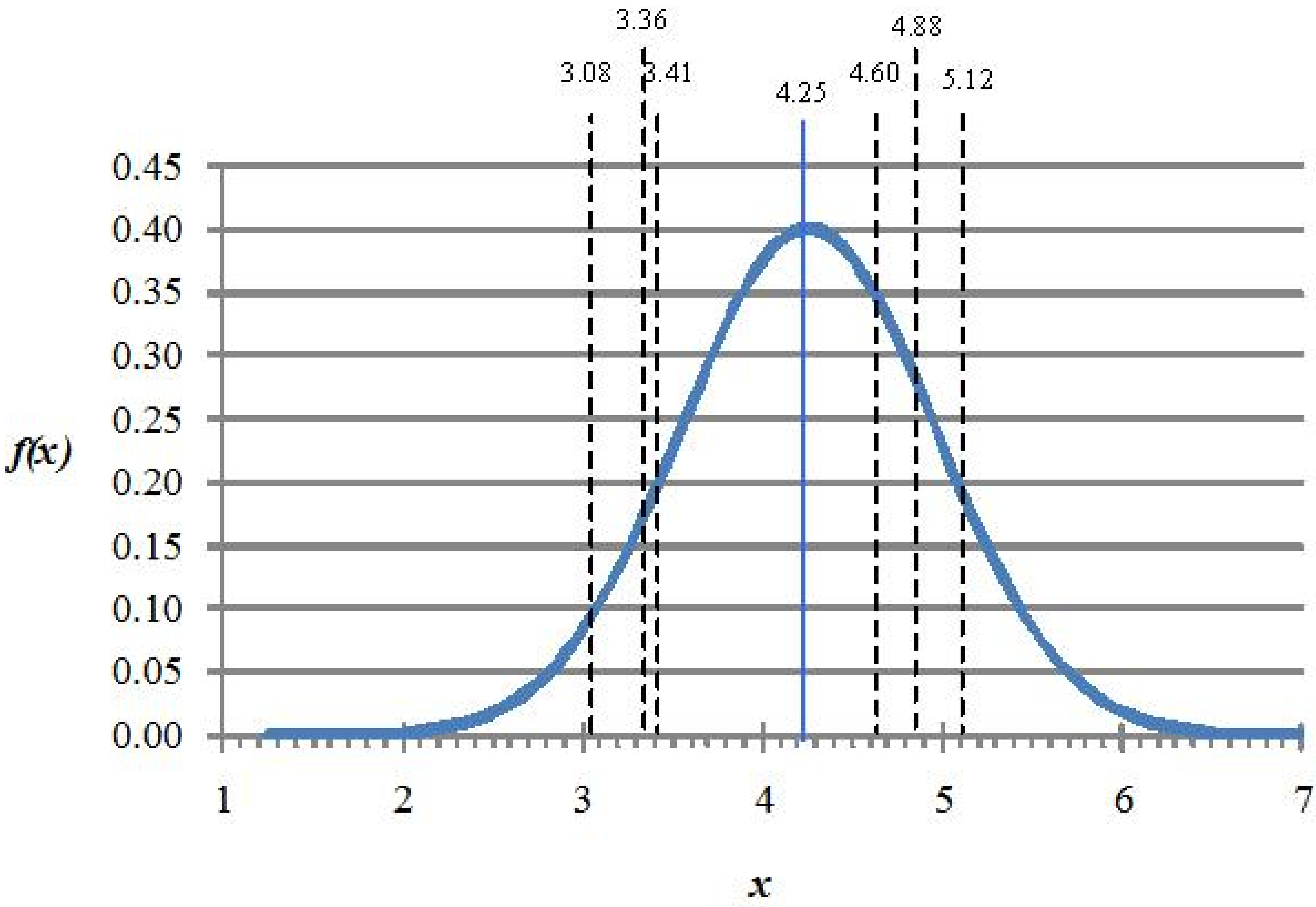
[b] Omitted reference groups: other or unstated expected major; less than one year advanced math, physics, and chemistry in high school.

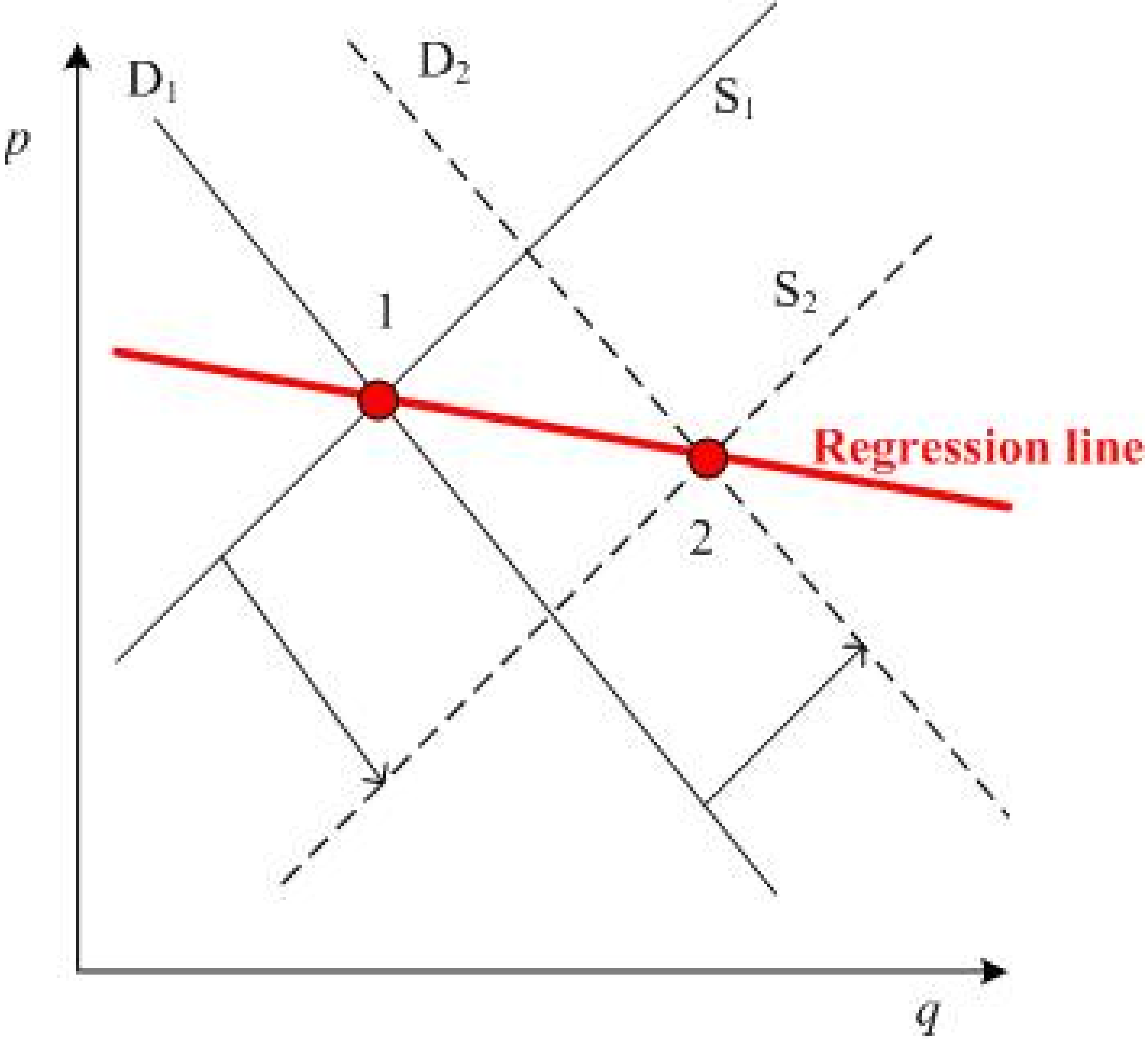
[c] In an ordered probit, an underlying, normally distributed, latent variable has a mean which is a function of observable variables. The latent variable gives rise to a set of observed dummy variables for ordered categories based on ranges between unobserved but estimable truncation points which correspond to levels of effort, ability, or other factors reflected in the explanatory variables. If L categories are observed, there are L−1 truncation points, of which the first is normalized to be zero, so that L−2 truncation points are estimated and reported in the table. The values correspond to standard deviations of the latent normally distributed variable.

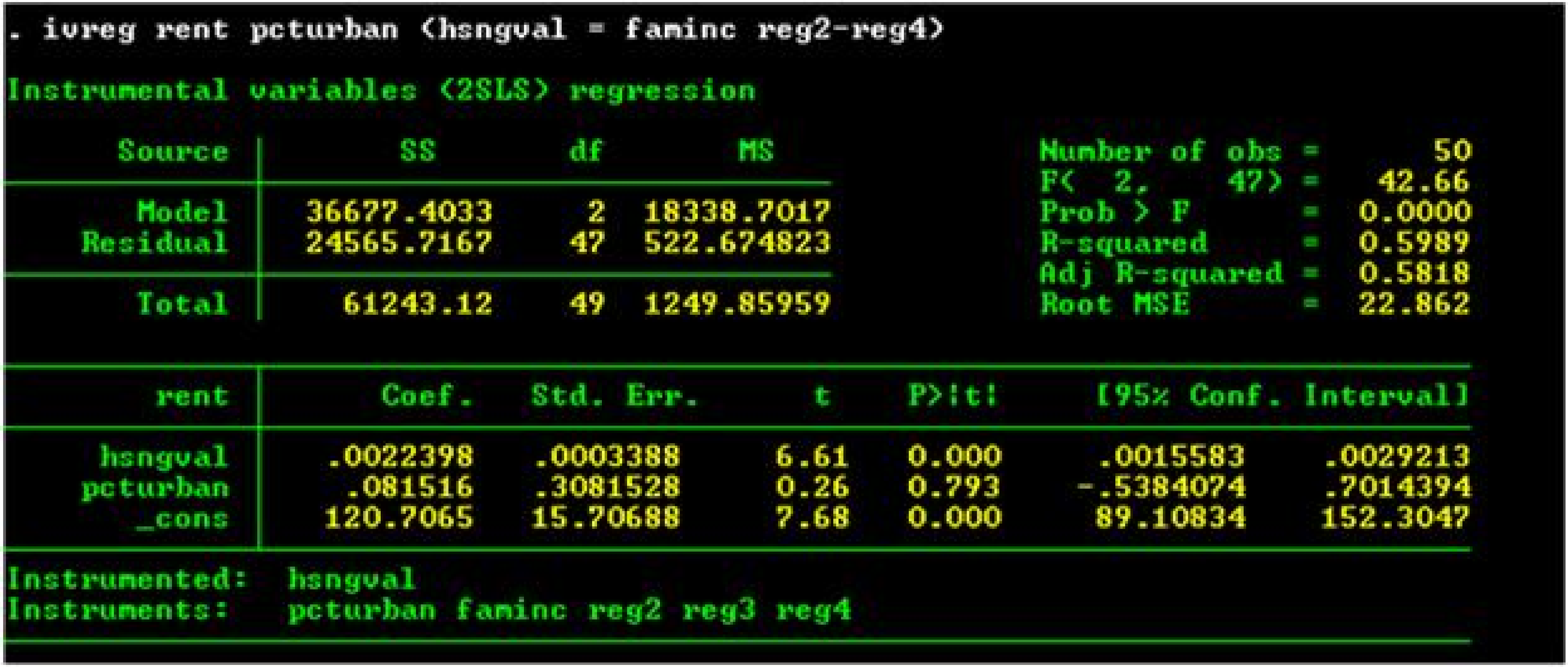
[d] Significant at 0·01 level, one- or two-tailed test, as appropriate.

[e] Significant at 0·05 level, one- or two-tailed test, as appropriate.

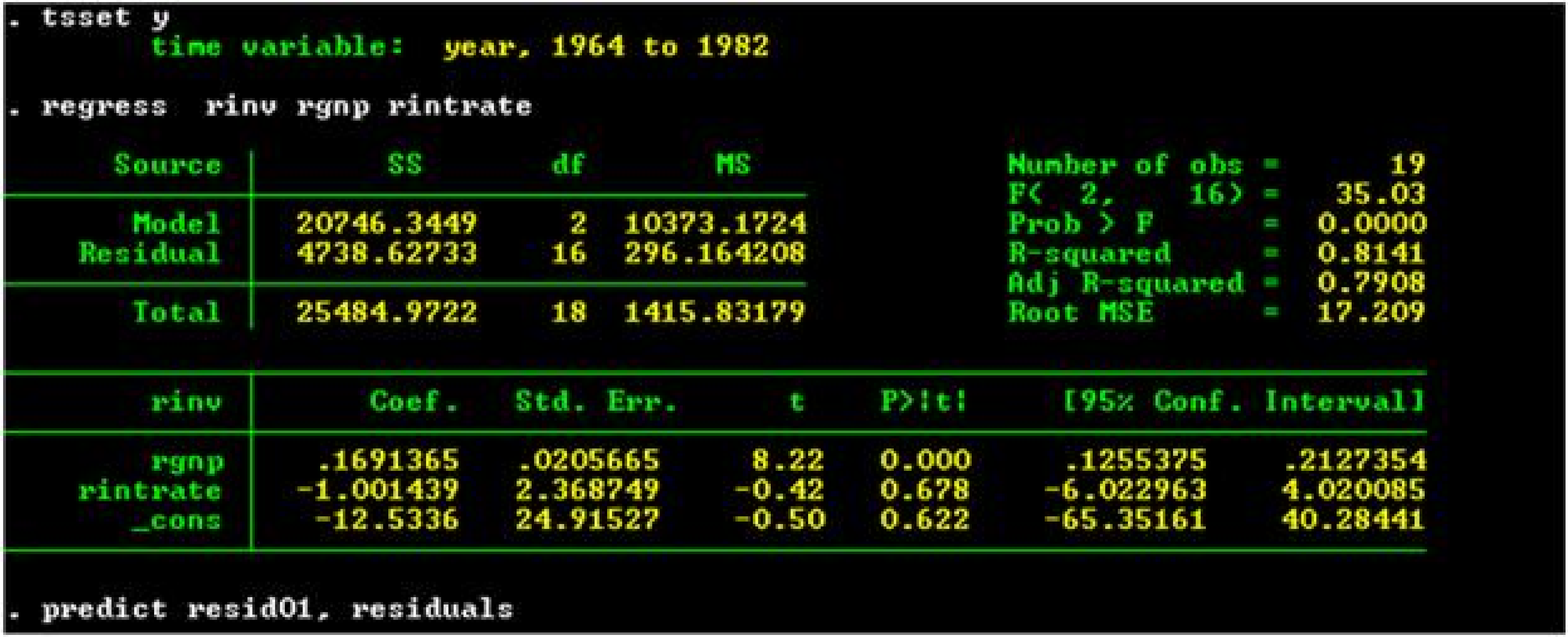
	msat	foreign	female	emecon	emoss	emns	emh	am1	am2	am3	phy1	phy2	chem1	chem2	_cut1	_cut2	_cut3	_cut4	_cut5	_cut6
msat	0.007																			
foreign	-0.001	0.020																		
female	0.001	-0.002	0.009																	
emecon	0.000	0.000	0.001	0.015																
emoss	-0.001	0.000	-0.001	0.009	0.021															
emns	0.000	-0.001	0.000	0.009	0.009	0.019														
emh	0.000	-0.002	0.000	0.009	0.009	0.009	0.040													
am1	0.000	-0.002	-0.001	0.000	0.001	0.002	0.002	0.047												
am2	-0.001	-0.001	-0.001	-0.001	0.001	0.001	0.002	0.043	0.048											
am3	-0.004	0.002	0.000	-0.003	0.000	-0.006	-0.007	0.042	0.044	0.178										
phy1	-0.001	0.001	0.001	0.001	0.000	0.000	0.003	0.000	0.000	-0.002	0.010									
phy2	0.000	0.001	0.000	0.001	0.000	-0.001	0.000	0.001	0.001	-0.006	0.007	0.091								
chem1	0.000	0.001	0.000	0.001	0.000	0.000	0.001	0.000	0.000	0.004	0.000	0.004	0.033							
chem2	-0.001	0.002	0.000	0.000	0.000	-0.002	0.001	0.000	-0.002	0.006	0.000	0.005	0.030	0.047						
_cut1	0.040	-0.006	0.012	0.010	0.006	0.008	0.012	0.043	0.033	0.018	0.002	0.009	0.029	0.025	0.329					
_cut2	0.041	-0.006	0.012	0.010	0.006	0.008	0.012	0.043	0.034	0.018	0.002	0.009	0.029	0.026	0.329	0.330				
_cut3	0.041	-0.006	0.012	0.010	0.006	0.008	0.012	0.043	0.034	0.018	0.002	0.009	0.029	0.026	0.329	0.330	0.331			
_cut4	0.041	-0.006	0.012	0.010	0.005	0.009	0.011	0.043	0.035	0.020	0.003	0.010	0.029	0.026	0.332	0.333	0.334	0.341		
_cut5	0.041	-0.006	0.012	0.010	0.005	0.009	0.011	0.043	0.035	0.020	0.003	0.010	0.029	0.026	0.333	0.334	0.334	0.341	0.343	
_cut6	0.041	-0.006	0.013	0.010	0.005	0.009	0.011	0.043	0.035	0.020	0.003	0.010	0.029	0.026	0.333	0.334	0.335	0.342	0.343	0.345


3.36
4.88
3.08
3.41
4.60
5.12
4.25
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
f(x)
1
2
3
4
5
6
7
x

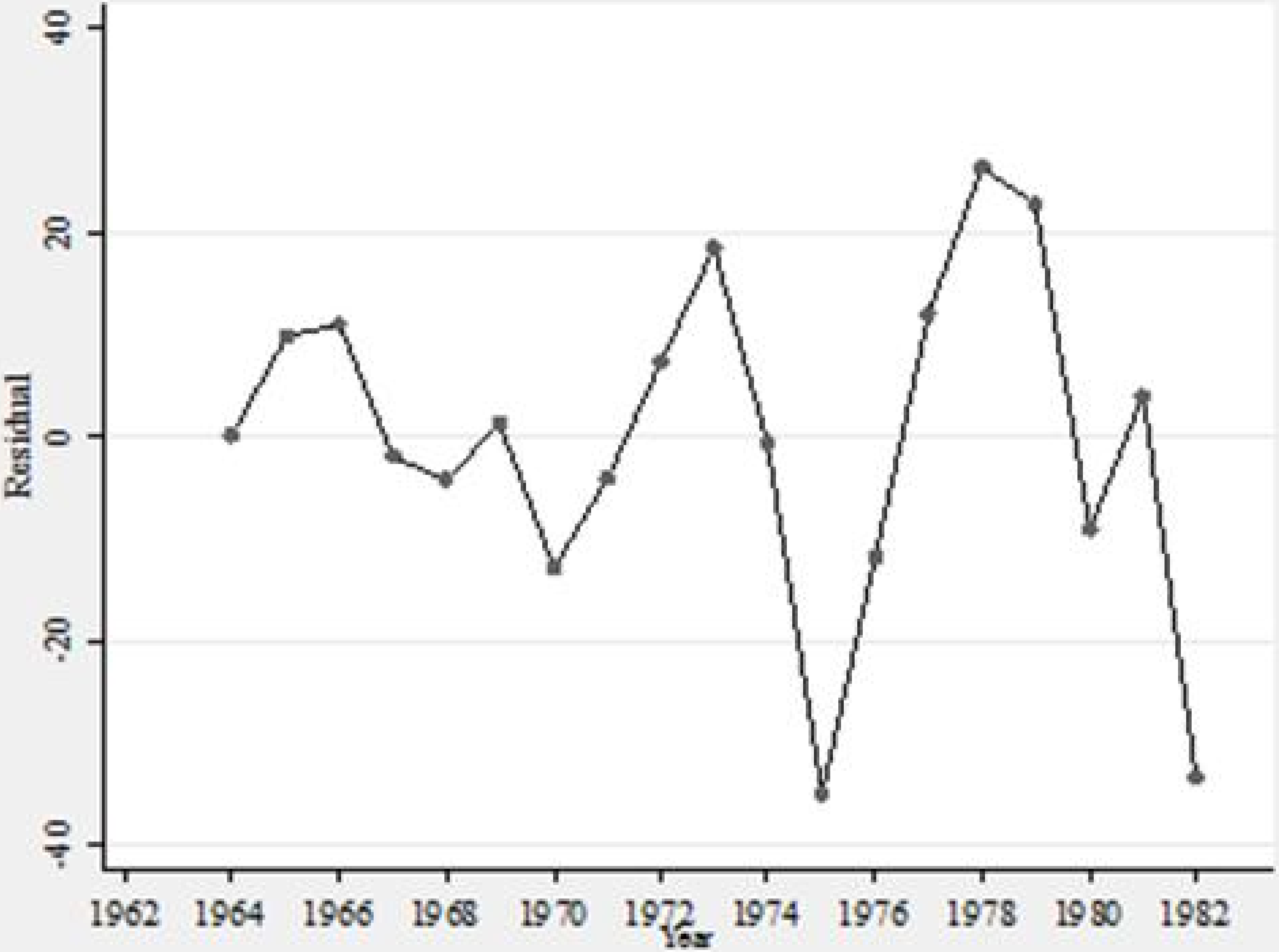

p
D$_1$
D$_2$
S$_1$
S$_2$
1
2
Regression line
q

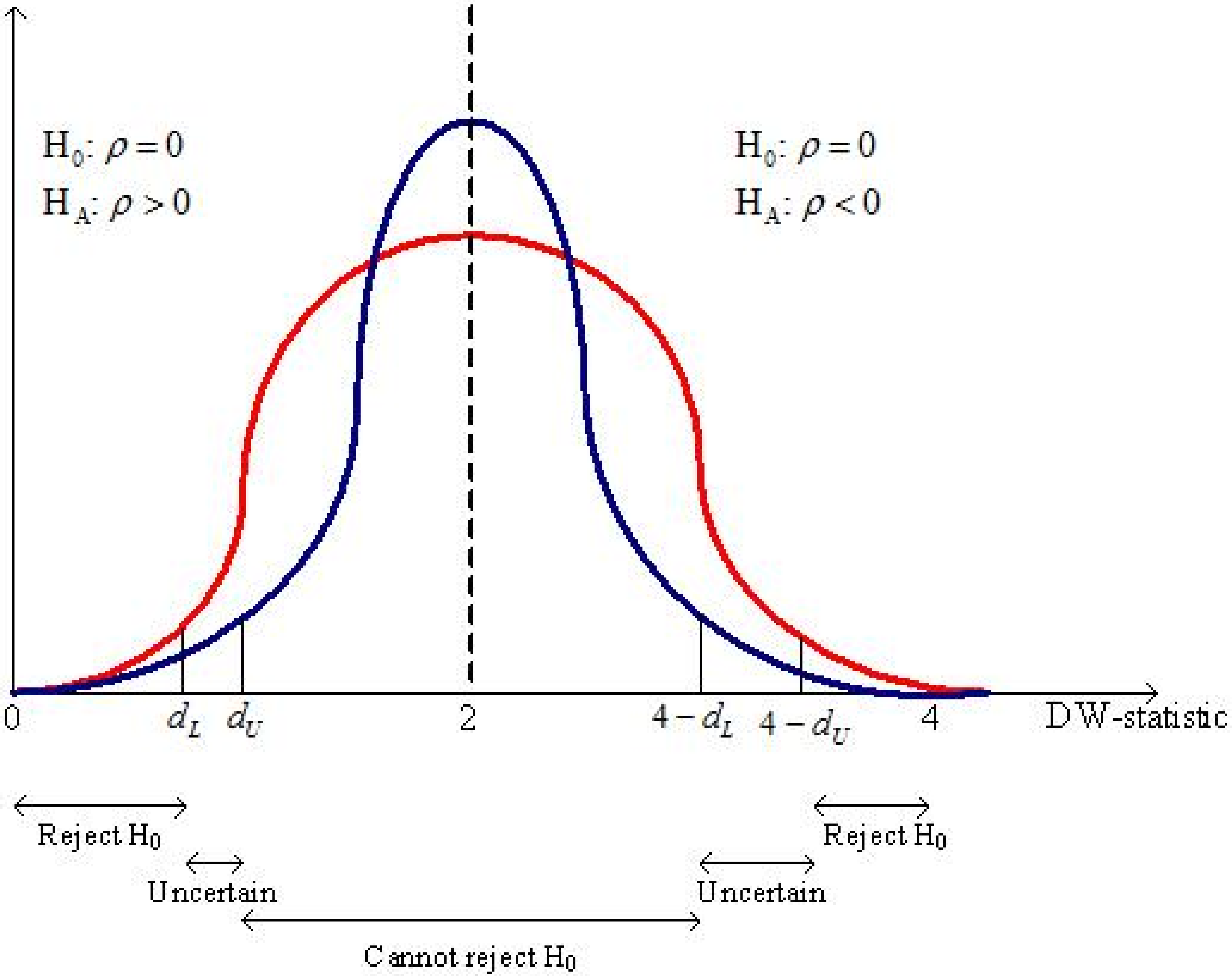


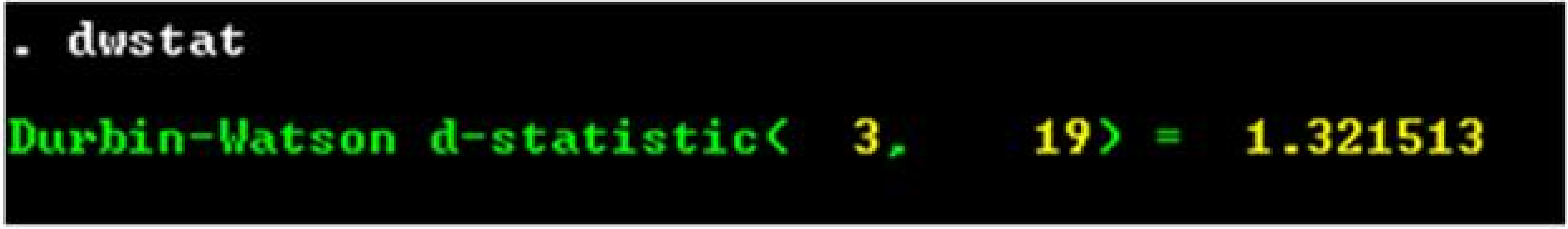
```
. ivreg rent pcturban (hsngval = faminc reg2-reg4)

Instrumental variables (2SLS) regression

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  2,    47) =   42.66
       Model |  36677.4033     2  18338.7017           Prob > F      =  0.0000
    Residual |  24565.7167    47  522.674823           R-squared     =  0.5989
-------------+------------------------------           Adj R-squared =  0.5818
       Total |    61243.12    49  1249.85959           Root MSE      =  22.862

------------------------------------------------------------------------------
        rent |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     hsngval |   .0022398   .0003388     6.61   0.000      .0015583    .0029213
    pcturban |    .081516   .3081528     0.26   0.793     -.5384074    .7014394
       _cons |   120.7065   15.70688     7.68   0.000      89.10834    152.3047
------------------------------------------------------------------------------
Instrumented:  hsngval
Instruments:   pcturban faminc reg2 reg3 reg4
------------------------------------------------------------------------------
```



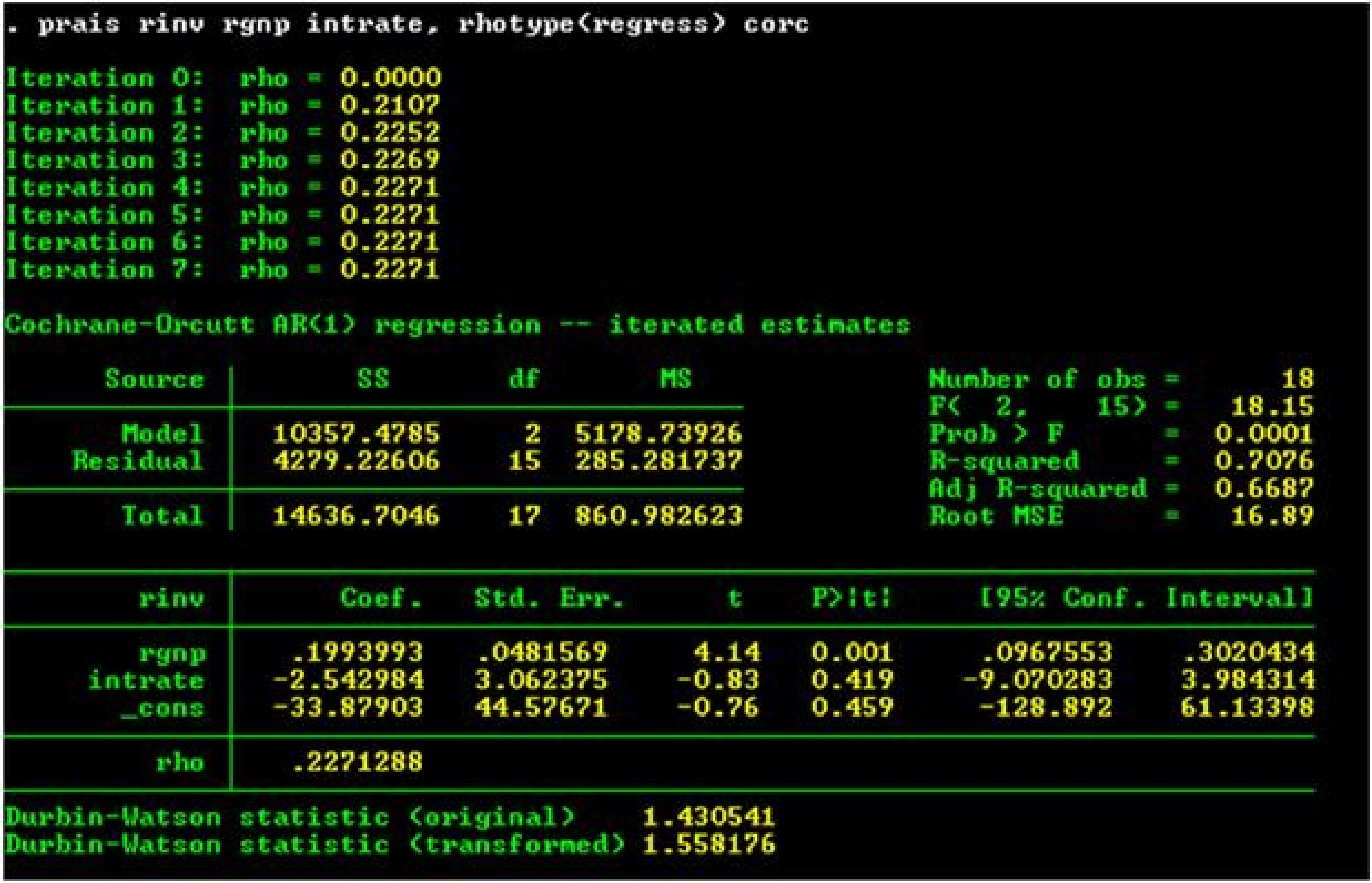
```
. tsset y
        time variable:  year, 1964 to 1982

. regress  rinv rgnp rintrate

      Source |       SS       df       MS              Number of obs =      19
-------------+------------------------------           F(  2,    16) =   35.03
       Model |  20746.3449     2  10373.1724           Prob > F      =  0.0000
    Residual |  4738.62733    16  296.164208           R-squared     =  0.8141
-------------+------------------------------           Adj R-squared =  0.7908
       Total |  25484.9722    18  1415.83179           Root MSE      =  17.209

------------------------------------------------------------------------------
        rinv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        rgnp |   .1691365   .0205665     8.22   0.000     .1255375    .2127354
    rintrate |  -1.001439   2.368749    -0.42   0.678    -6.022963    4.020085
       _cons |   -12.5336   24.91527    -0.50   0.622    -65.35161    40.28441
------------------------------------------------------------------------------

. predict resid01, residuals
```


Residual
40
20
0
-20
-40
1962
1964
1966
1968
1970
1972
1974
1976
1978
1980
1982
Year

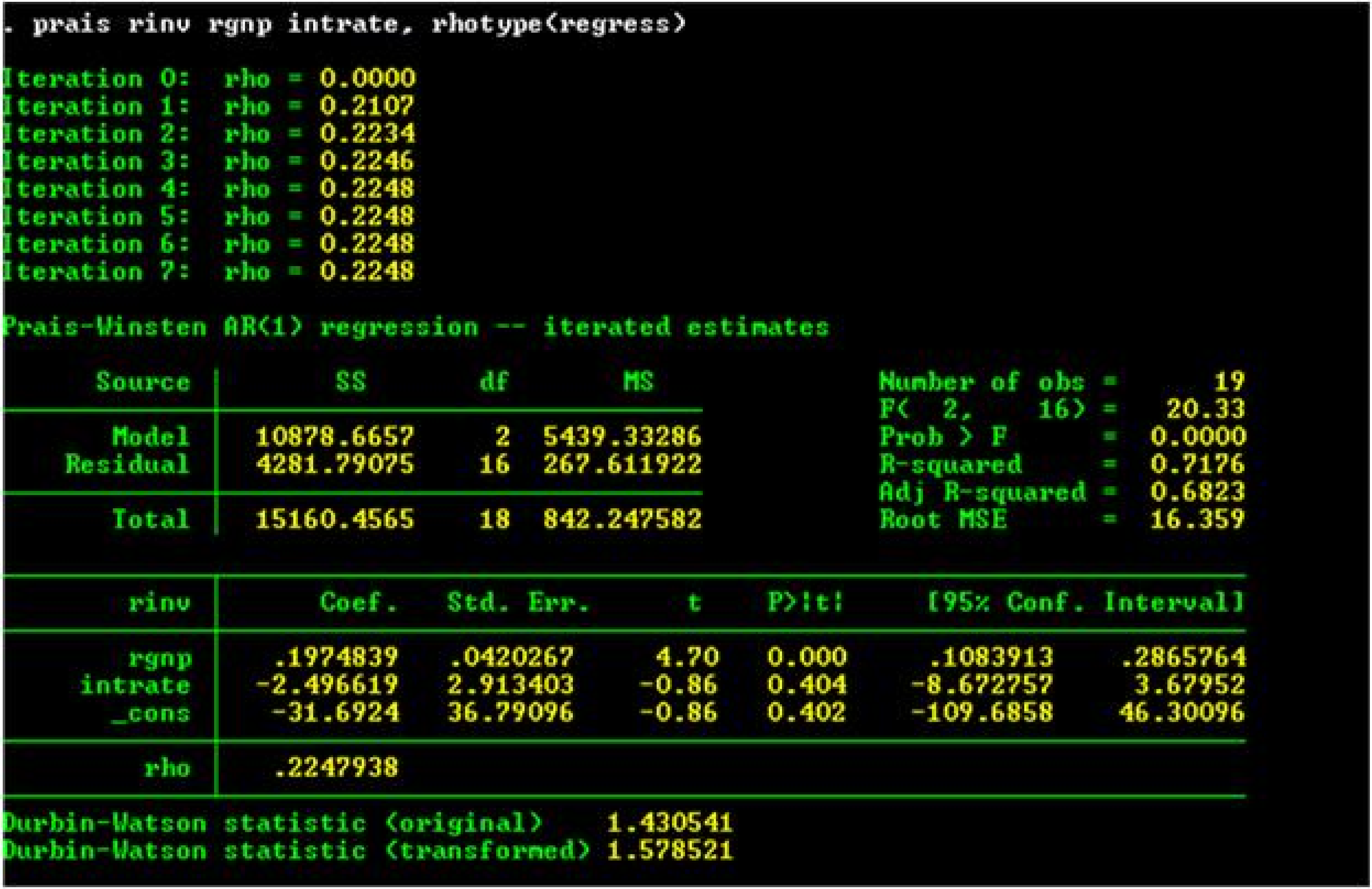

$H_0: \rho = 0$
$H_A: \rho > 0$
$H_0: \rho = 0$
$H_A: \rho < 0$
0
d_L
d_U
2
$4 - d_L$
$4 - d_U$
4
DW-statistic
Reject H_0
Uncertain
Cannot reject H_0
Uncertain
Reject H_0



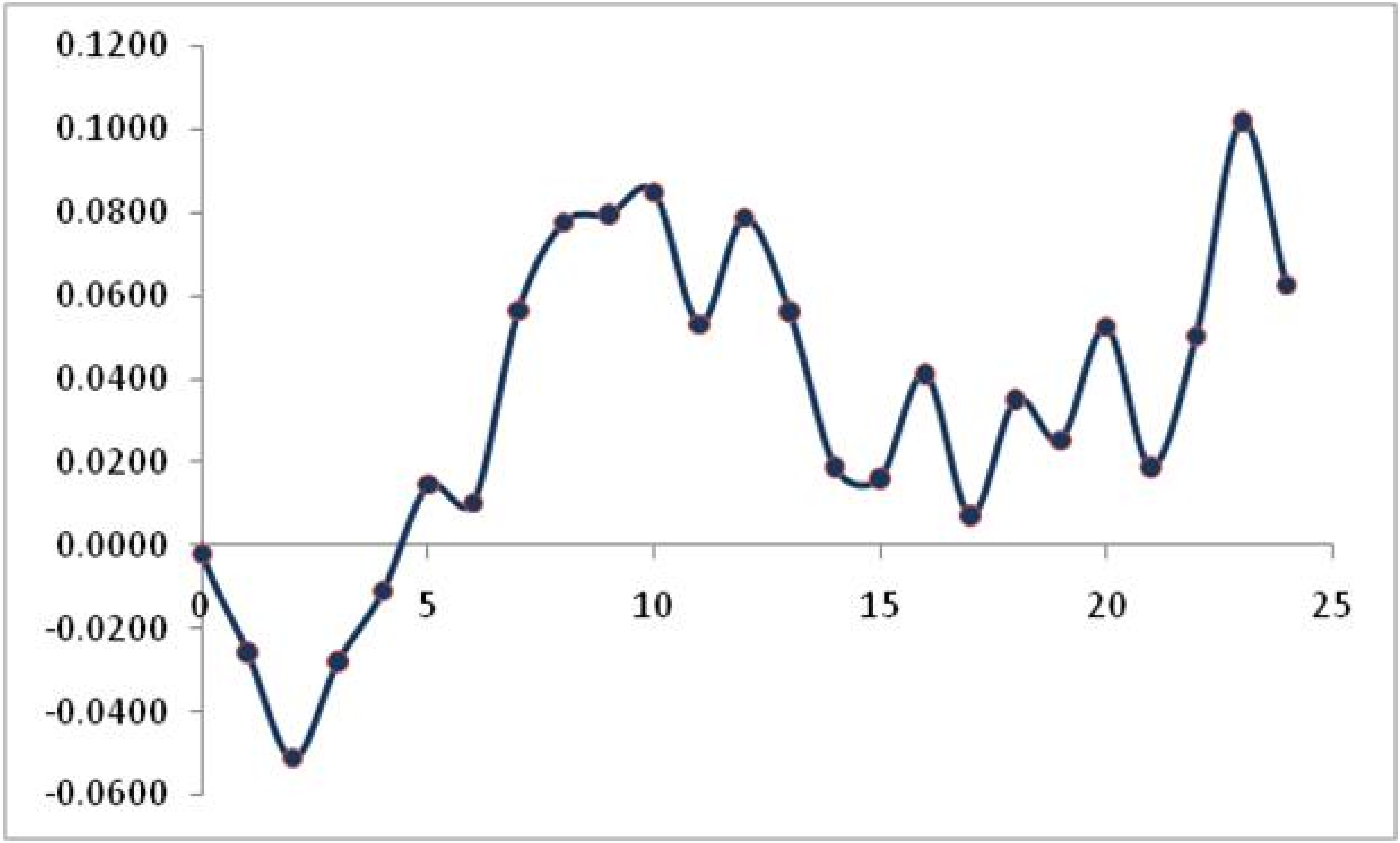
```
. dwstat

Durbin-Watson d-statistic(  3,    19) =  1.321513
```

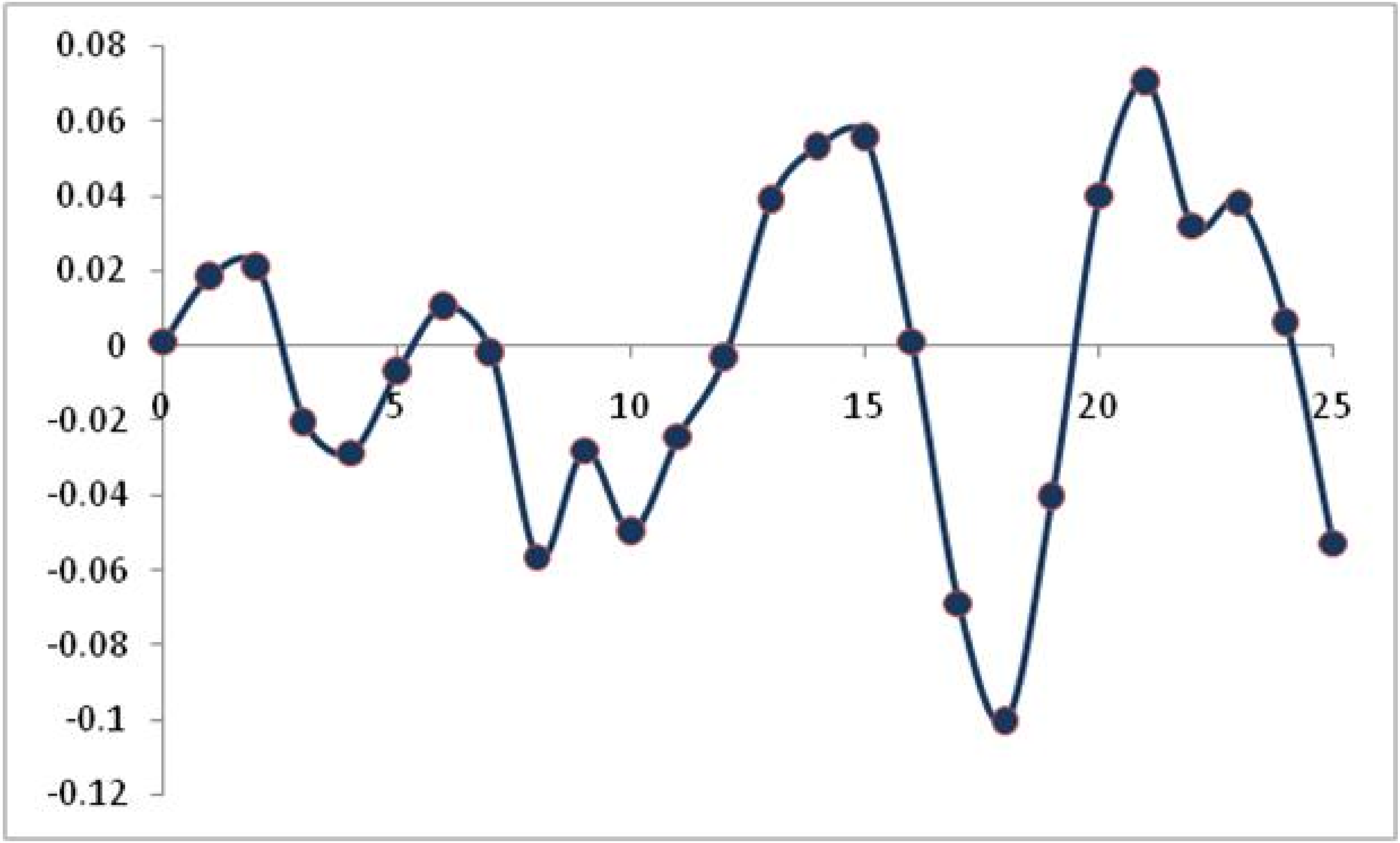


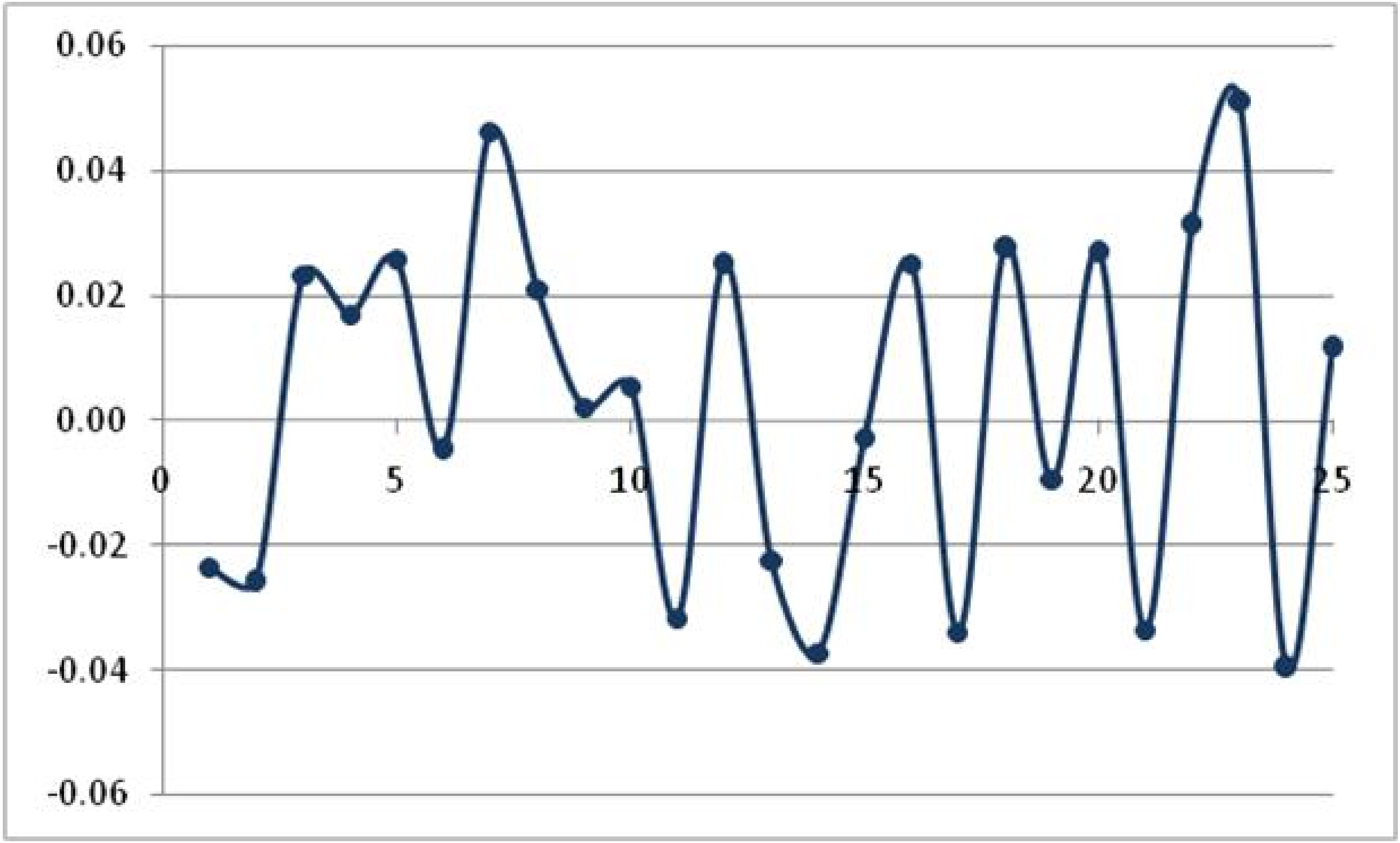
```
. prais rinv rgnp intrate, rhotype(regress) corc

Iteration 0:  rho = 0.0000
Iteration 1:  rho = 0.2107
Iteration 2:  rho = 0.2252
Iteration 3:  rho = 0.2269
Iteration 4:  rho = 0.2271
Iteration 5:  rho = 0.2271
Iteration 6:  rho = 0.2271
Iteration 7:  rho = 0.2271

Cochrane-Orcutt AR(1) regression -- iterated estimates

      Source |       SS       df       MS              Number of obs =      18
-------------+------------------------------           F(  2,    15) =   18.15
       Model |  10357.4785     2  5178.73926           Prob > F      =  0.0001
    Residual |  4279.22606    15  285.281737           R-squared     =  0.7076
-------------+------------------------------           Adj R-squared =  0.6687
       Total |  14636.7046    17  860.982623           Root MSE      =   16.89

------------------------------------------------------------------------------
        rinv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        rgnp |   .1993993   .0481569     4.14   0.001     .0967553    .3020434
     intrate |  -2.542984   3.062375    -0.83   0.419    -9.070283    3.984314
       _cons |  -33.87903   44.57671    -0.76   0.459     -128.892    61.13398
-------------+----------------------------------------------------------------
         rho |   .2271288
------------------------------------------------------------------------------
Durbin-Watson statistic (original)    1.430541
Durbin-Watson statistic (transformed) 1.558176
```

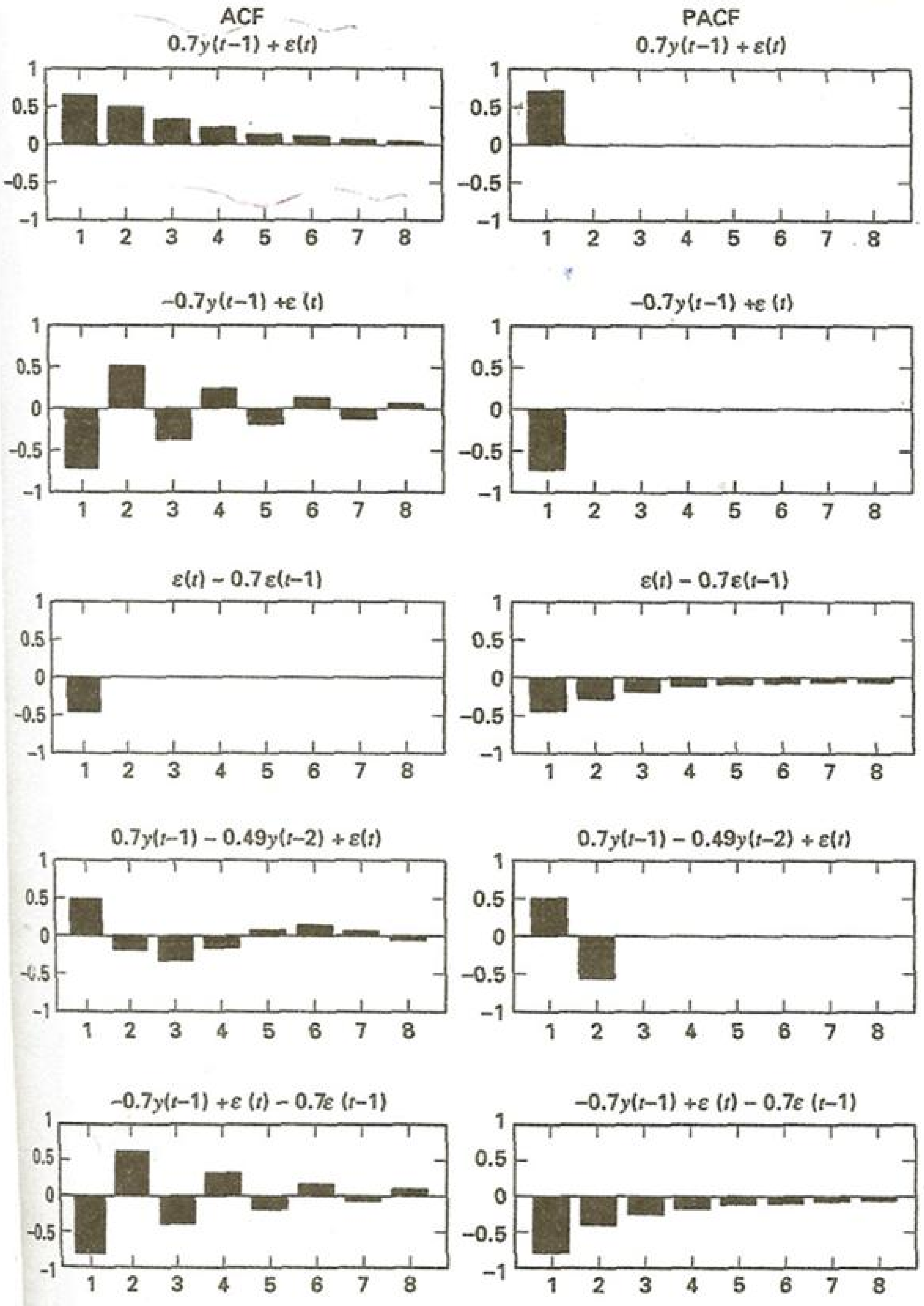


```
. prais rinv rgnp intrate, rhotype(regress)

Iteration 0:  rho = 0.0000
Iteration 1:  rho = 0.2107
Iteration 2:  rho = 0.2234
Iteration 3:  rho = 0.2246
Iteration 4:  rho = 0.2248
Iteration 5:  rho = 0.2248
Iteration 6:  rho = 0.2248
Iteration 7:  rho = 0.2248

Prais-Winsten AR(1) regression -- iterated estimates

      Source |       SS       df       MS              Number of obs =      19
-------------+------------------------------           F(  2,    16) =   20.33
       Model |  10878.6657     2  5439.33286           Prob > F      =  0.0000
    Residual |  4281.79075    16  267.611922           R-squared     =  0.7176
-------------+------------------------------           Adj R-squared =  0.6823
       Total |  15160.4565    18  842.247582           Root MSE      =  16.359

------------------------------------------------------------------------------
        rinv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        rgnp |   .1974839   .0420267     4.70   0.000     .1083913    .2865764
     intrate |  -2.496619   2.913403    -0.86   0.404    -8.672757     3.67952
       _cons |   -31.6924   36.79096    -0.86   0.402    -109.6858    46.30096
-------------+----------------------------------------------------------------
         rho |   .2247938
------------------------------------------------------------------------------
Durbin-Watson statistic (original)    1.430541
Durbin-Watson statistic (transformed) 1.578521
```


0.1200
0.1000
0.0800
0.0600
0.0400
0.0200
0.0000
-0.0200
-0.0400
-0.0600
0
5
10
15
20
25


0.08
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
-0.08
-0.1
-0.12
0
5
10
15
20
25


0.06
0.04
0.02
0.00
-0.02
-0.04
-0.06
0
5
10
15
20
25



ACF

PACF

$0.7y(t-1) + \varepsilon(t)$

$0.7y(t-1) + \varepsilon(t)$

$-0.7y(t-1) + \varepsilon(t)$

$-0.7y(t-1) + \varepsilon(t)$

$\varepsilon(t) - 0.7\varepsilon(t-1)$

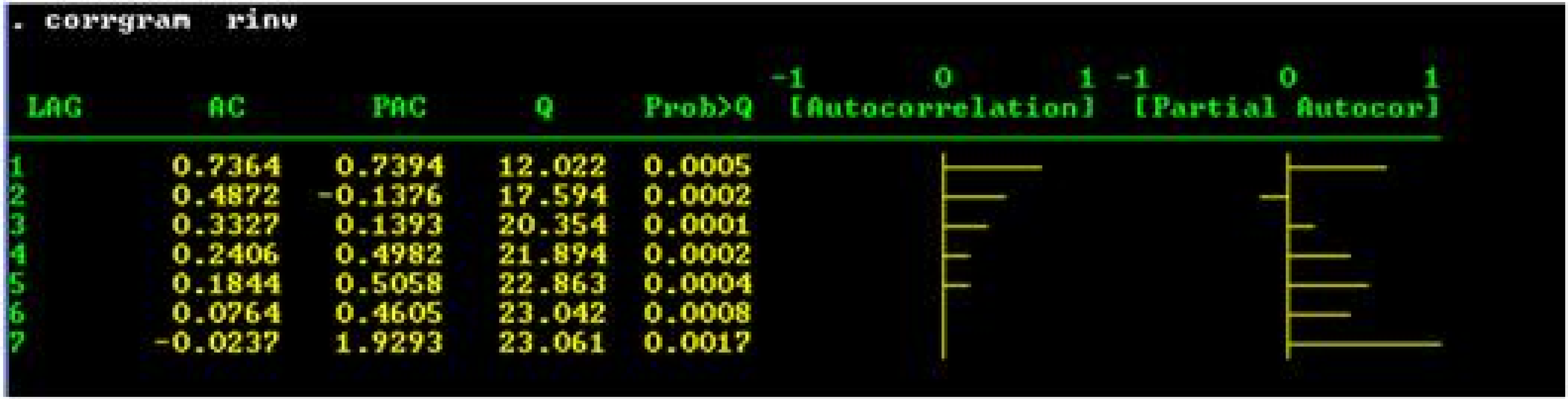
$\varepsilon(t) - 0.7\varepsilon(t-1)$

$0.7y(t-1) - 0.49y(t-2) + \varepsilon(t)$

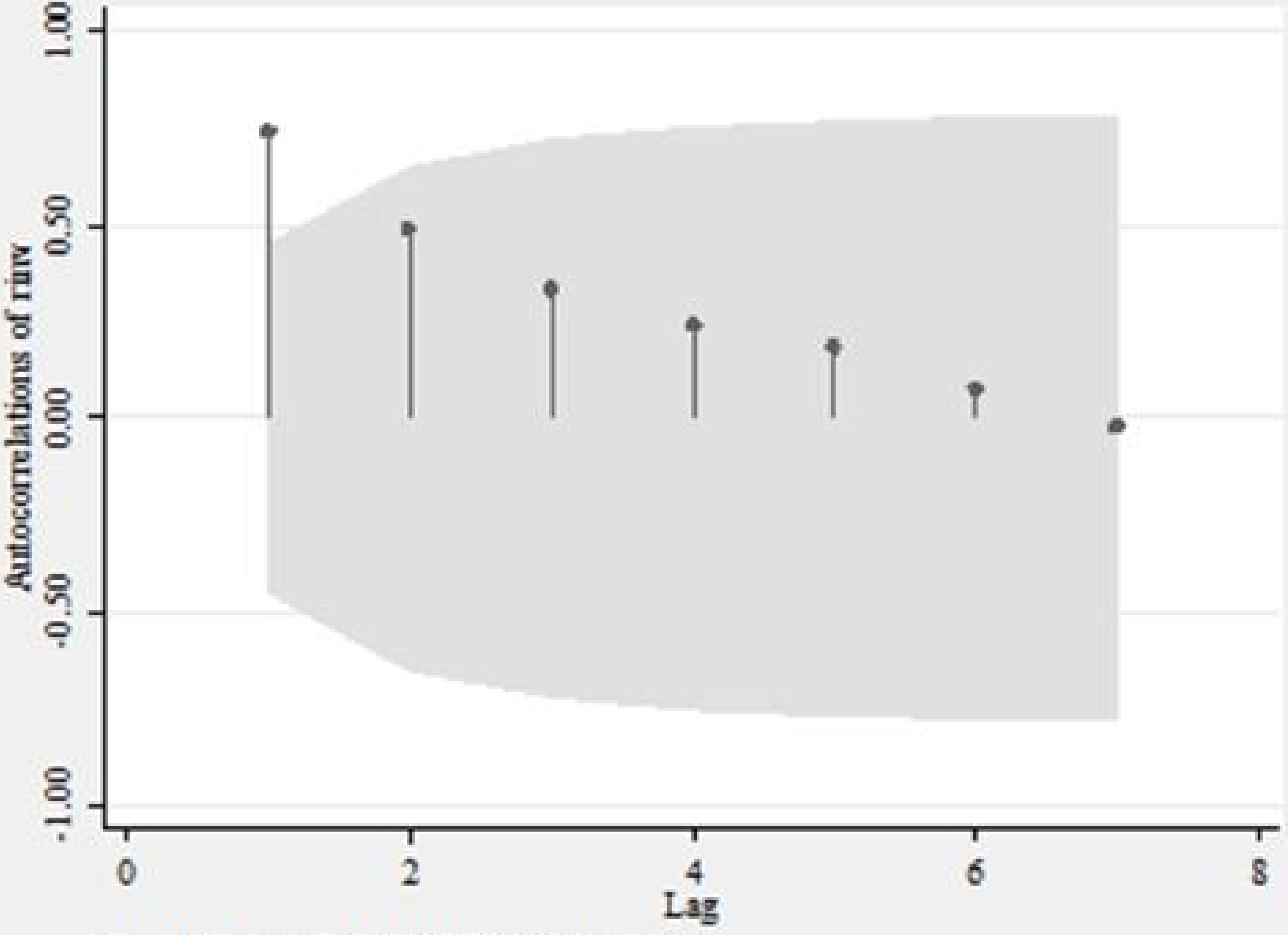
$0.7y(t-1) - 0.49y(t-2) + \varepsilon(t)$

$-0.7y(t-1) + \varepsilon(t) - 0.7\varepsilon(t-1)$

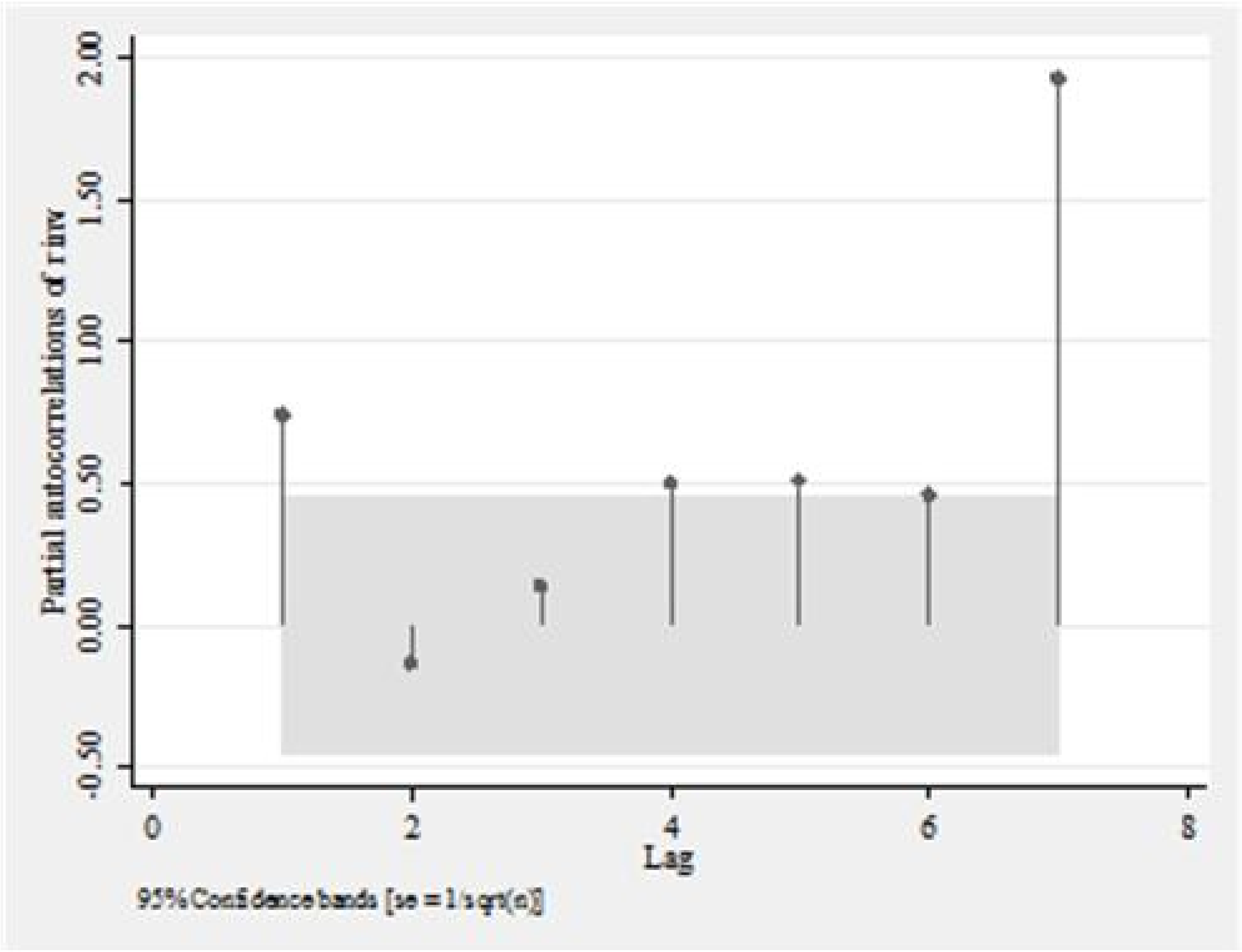
$-0.7y(t-1) + \varepsilon(t) - 0.7\varepsilon(t-1)$

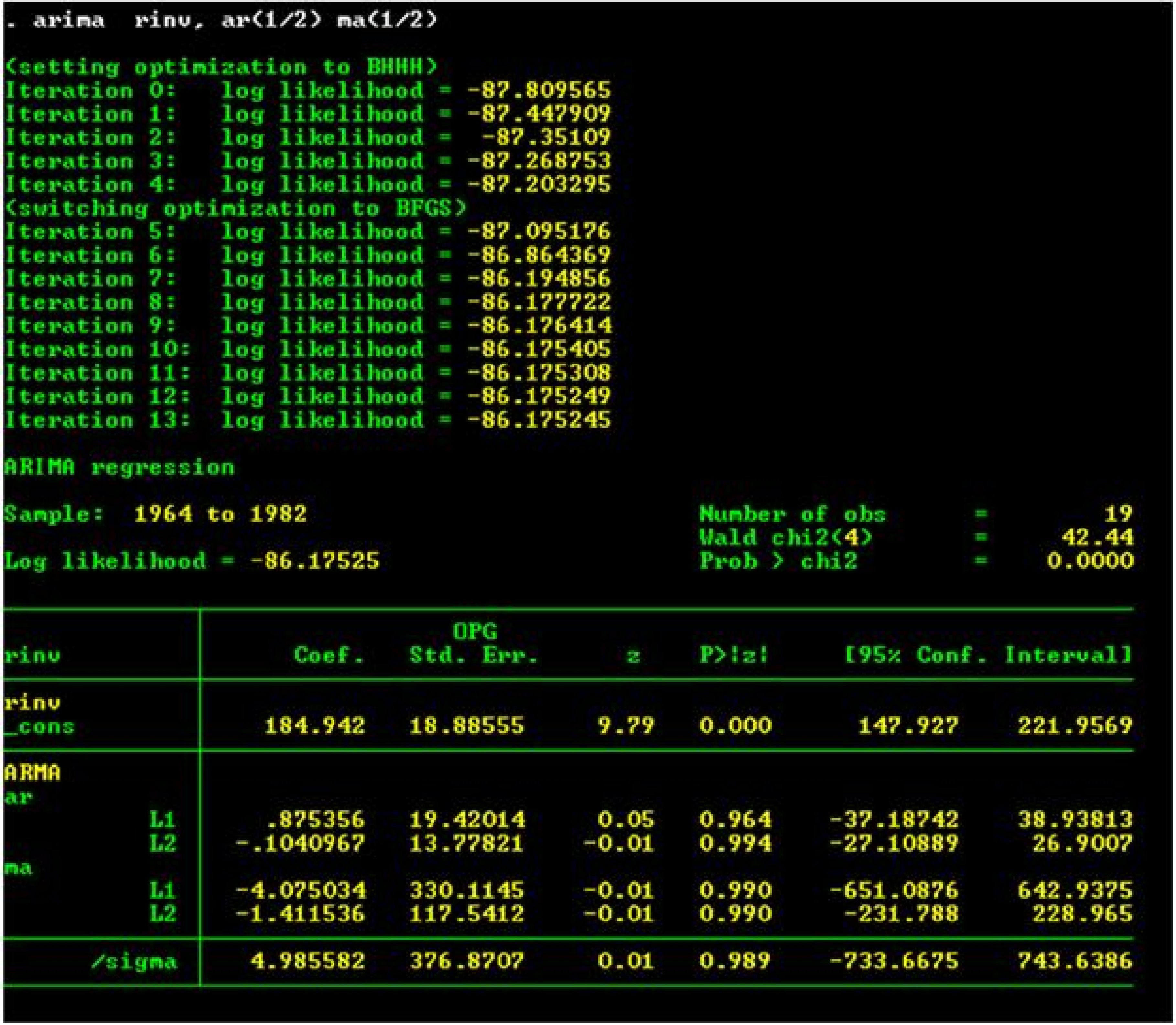


```
. corrgram  rinv

                                                    -1       0       1 -1       0       1
 LAG        AC        PAC        Q      Prob>Q  [Autocorrelation]  [Partial Autocor]
-----------------------------------------------------------------------------------------
1        0.7364    0.7394    12.022   0.0005
2        0.4872   -0.1376    17.594   0.0002
3        0.3327    0.1393    20.354   0.0001
4        0.2406    0.4982    21.894   0.0002
5        0.1844    0.5058    22.863   0.0004
6        0.0764    0.4605    23.042   0.0008
7       -0.0237    1.9293    23.061   0.0017
```

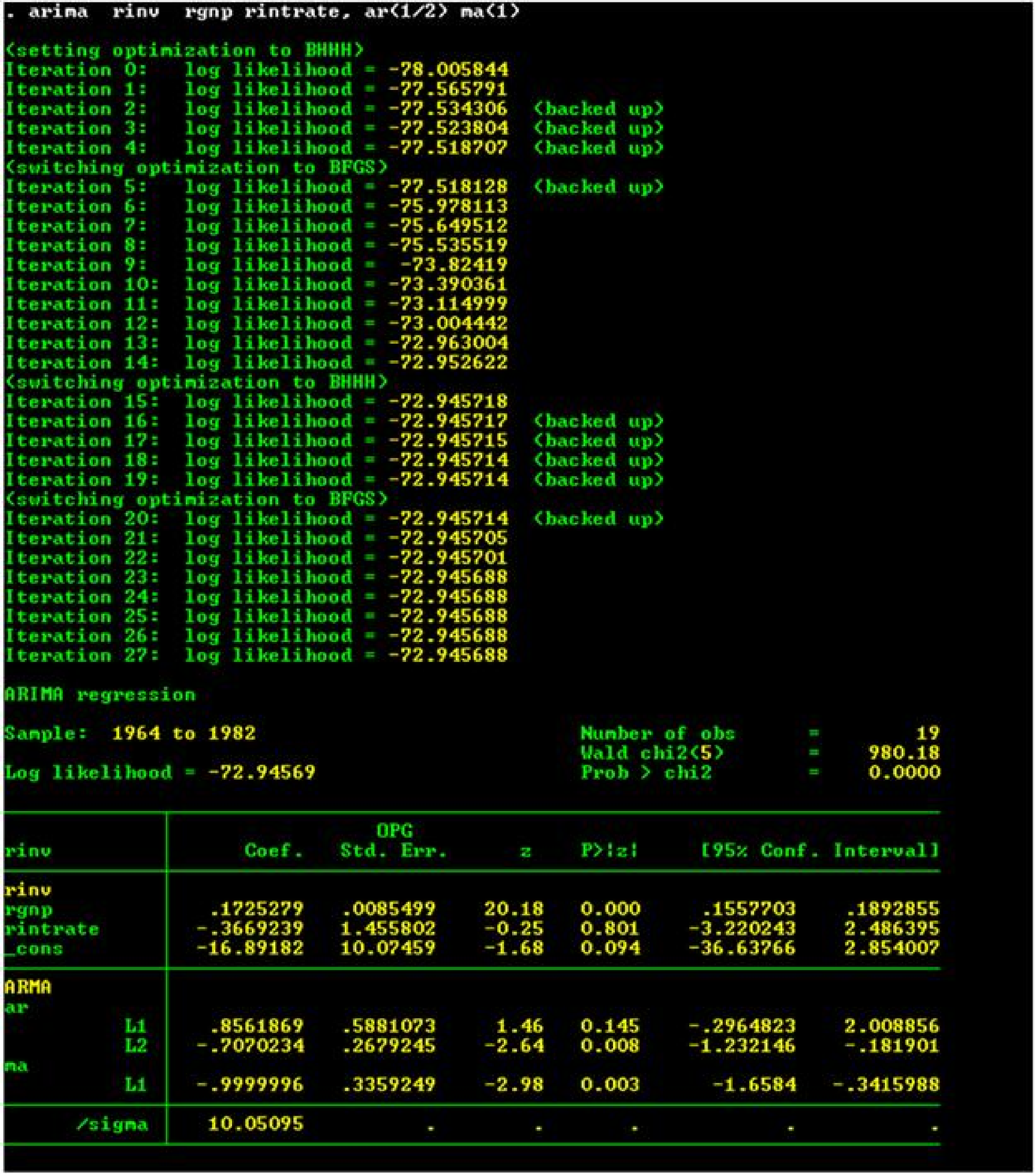


Bartlett's formula for MA(q) 95% confidence bands

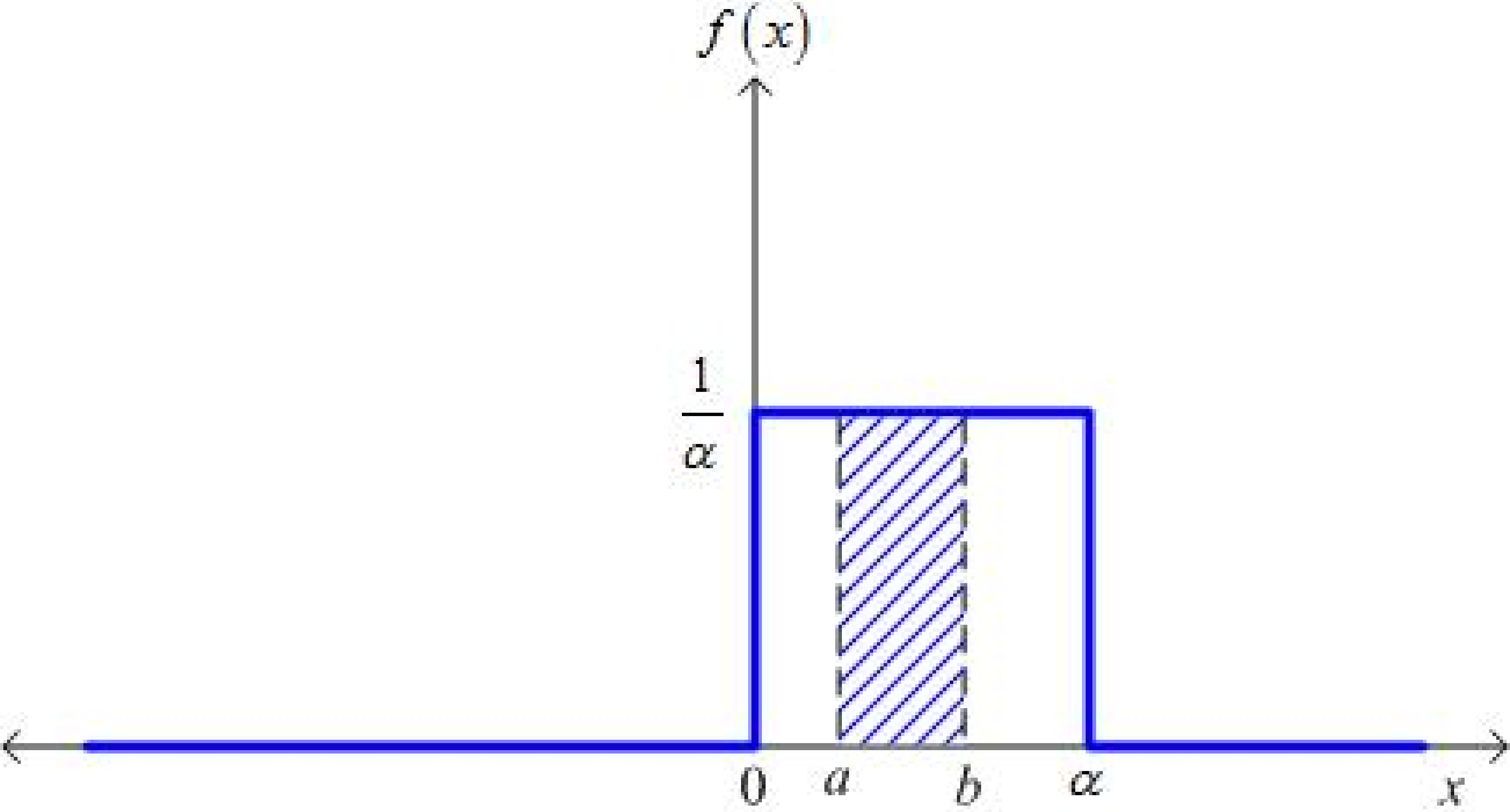

Partial autocorrelations of r inv
2.00
1.50
1.00
0.50
0.00
-0.50
0
2
4
6
8
Lag
95% Confidence bands [se = 1/sqrt(n)]



```
. arima  rinv, ar(1/2) ma(1/2)

(setting optimization to BHHH)
Iteration 0:   log likelihood = -87.809565
Iteration 1:   log likelihood = -87.447909
Iteration 2:   log likelihood =  -87.35109
Iteration 3:   log likelihood = -87.268753
Iteration 4:   log likelihood = -87.203295
(switching optimization to BFGS)
Iteration 5:   log likelihood = -87.095176
Iteration 6:   log likelihood = -86.864369
Iteration 7:   log likelihood = -86.194856
Iteration 8:   log likelihood = -86.177722
Iteration 9:   log likelihood = -86.176414
Iteration 10:  log likelihood = -86.175405
Iteration 11:  log likelihood = -86.175308
Iteration 12:  log likelihood = -86.175249
Iteration 13:  log likelihood = -86.175245

ARIMA regression

Sample:  1964 to 1982                           Number of obs      =        19
                                                Wald chi2(4)       =     42.44
Log likelihood = -86.17525                      Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |                 OPG
        rinv |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
rinv         |
       _cons |    184.942    18.88555     9.79   0.000      147.927    221.9569
-------------+----------------------------------------------------------------
ARMA         |
ar           |
         L1. |    .875356    19.42014     0.05   0.964    -37.18742    38.93813
         L2. |  -.1040967    13.77821    -0.01   0.994    -27.10889     26.9007
ma           |
         L1. |  -4.075034    330.1145    -0.01   0.990    -651.0876    642.9375
         L2. |  -1.411536    117.5412    -0.01   0.990     -231.788     228.965
-------------+----------------------------------------------------------------
      /sigma |   4.985582    376.8707     0.01   0.989    -733.6675    743.6386
------------------------------------------------------------------------------
```



```
. arima  rinv  rgnp rintrate, ar(1/2) ma(1)

(setting optimization to BHHH)
Iteration 0:   log likelihood = -78.005844
Iteration 1:   log likelihood = -77.565791
Iteration 2:   log likelihood = -77.534306  (backed up)
Iteration 3:   log likelihood = -77.523804  (backed up)
Iteration 4:   log likelihood = -77.518707  (backed up)
(switching optimization to BFGS)
Iteration 5:   log likelihood = -77.518128  (backed up)
Iteration 6:   log likelihood = -75.978113
Iteration 7:   log likelihood = -75.649512
Iteration 8:   log likelihood = -75.535519
Iteration 9:   log likelihood =  -73.82419
Iteration 10:  log likelihood = -73.390361
Iteration 11:  log likelihood = -73.114999
Iteration 12:  log likelihood = -73.004442
Iteration 13:  log likelihood = -72.963004
Iteration 14:  log likelihood = -72.952622
(switching optimization to BHHH)
Iteration 15:  log likelihood = -72.945718
Iteration 16:  log likelihood = -72.945717  (backed up)
Iteration 17:  log likelihood = -72.945715  (backed up)
Iteration 18:  log likelihood = -72.945714  (backed up)
Iteration 19:  log likelihood = -72.945714  (backed up)
(switching optimization to BFGS)
Iteration 20:  log likelihood = -72.945714  (backed up)
Iteration 21:  log likelihood = -72.945705
Iteration 22:  log likelihood = -72.945701
Iteration 23:  log likelihood = -72.945688
Iteration 24:  log likelihood = -72.945688
Iteration 25:  log likelihood = -72.945688
Iteration 26:  log likelihood = -72.945688
Iteration 27:  log likelihood = -72.945688

ARIMA regression

Sample:  1964 to 1982                           Number of obs      =        19
                                                Wald chi2(5)       =    980.18
Log likelihood = -72.94569                      Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |                 OPG
        rinv |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
rinv         |
        rgnp |   .1725279   .0085499    20.18   0.000     .1557703    .1892855
    rintrate |  -.3669239   1.455802    -0.25   0.801    -3.220243    2.486395
       _cons |  -16.89182   10.07459    -1.68   0.094    -36.63766    2.854007
-------------+----------------------------------------------------------------
ARMA         |
          ar |
         L1. |   .8561869   .5881073     1.46   0.145    -.2964823    2.008856
         L2. |  -.7070234   .2679245    -2.64   0.008    -1.232146    -.181901
          ma |
         L1. |  -.9999996   .3359249    -2.98   0.003      -1.6584   -.3415988
-------------+----------------------------------------------------------------
      /sigma |   10.05095          .        .       .            .           .
------------------------------------------------------------------------------
```


$f(x)$
$\frac{1}{\alpha}$
0
a
b
α
x