Basic Econometrics with Stata

Carl Moody
Economics Department
College of William and Mary
2009

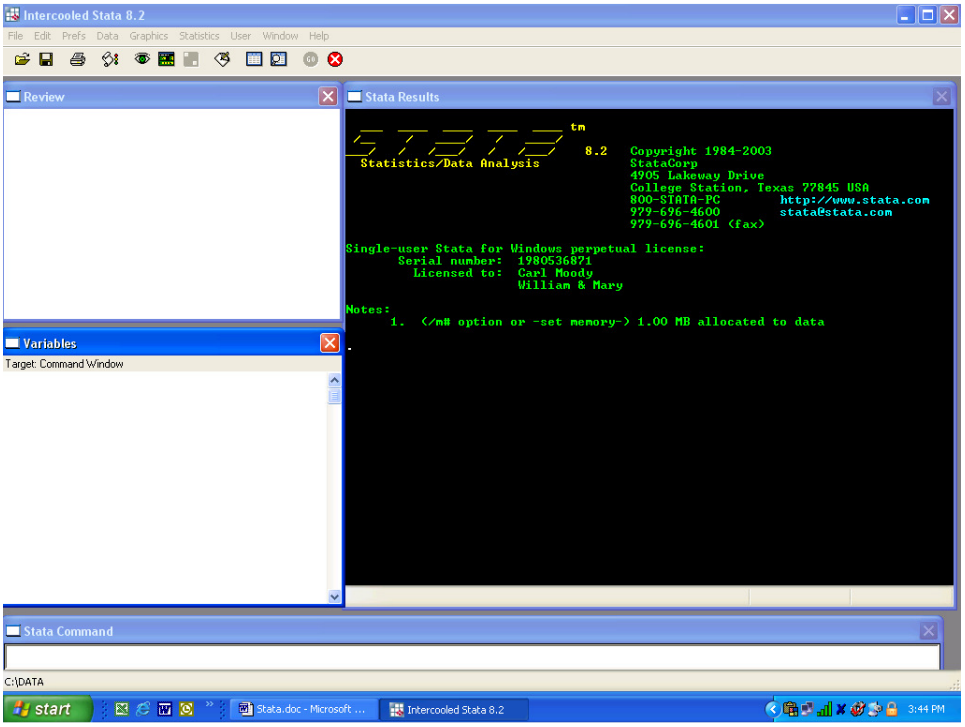
Table of Contents

1 AN OVERVIEW OF STATA

Stata is a computer program that allows the user to perform a wide variety of statistical analyses. In this chapter we will take a short tour of Stata to get an appreciation of what it can do.

Starting Stata

Stata is available on the server. When invoked, the screen should look something like this. (This is the current version on my computer. The version on the server may be different.)



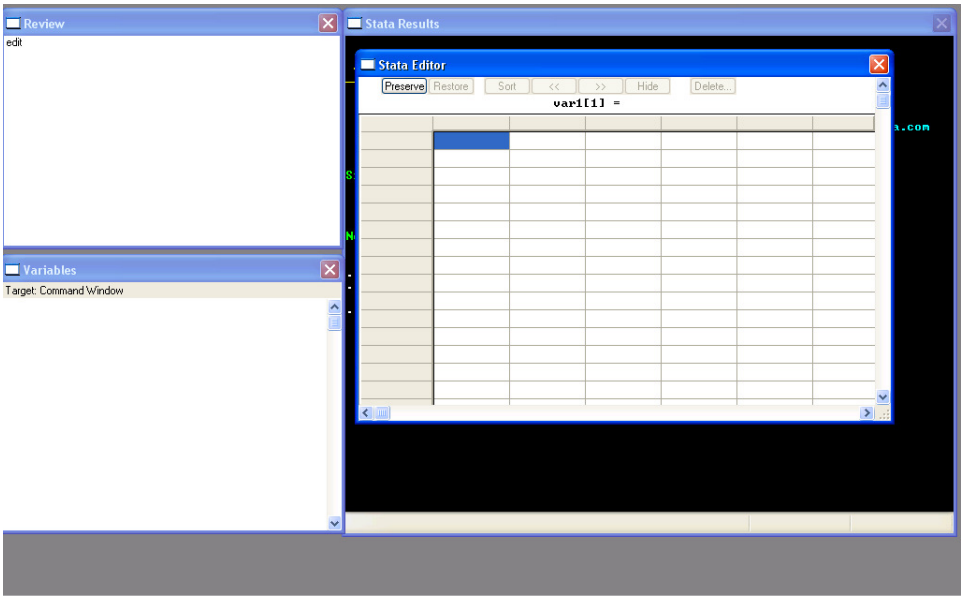
There are many ways to do things in Stata. The simplest way is to enter commands interactively, allowing Stata to execute each command immediately. The commands are entered in the "Stata Command" window (along the bottom in this view). Results are shown on the right. After the command has been entered it appears in the "Review" window in the upper left. If you want to re-enter the command you can double click on it. A single click moves it to the command line, where you can edit it before submitting. You can also use the buttons on the toolbar at the top. I recommend using "do" files consisting of a series of Stata commands for

complex jobs (see Chapter 5 below). However, the way to learn a statistical language is to do it. Different people will find different ways of doing the same task.

An econometric project consists of several steps. The first is to choose a topic. Then you do library research to find out what research has already been done in the area. Next, you collect data. Once you have the data, you are ready to do the econometric analysis. The econometric part consists of four steps, which may be repeated several times. (1) Get the data into Stata. (2) Look at the data by printing it out, graphing it, and summarizing it. (3) Transform the data. For example, you might want to divide by population to get per capita values, or divide by a price index to get real dollars, or take logs. (4) Analyze the transformed data using regression or some other procedure. The final step in the project is to write up the results. In this chapter we will do a mini-analysis illustrating the four econometric steps.

Your data set should be written as a matrix (a rectangular array) of numbers, with columns corresponding to variables and rows corresponding to observations. In the example below, we have 10 observations on 4 variables. The data matrix will therefore have four columns and 10 rows. Similarly, you could have observations on consumption, income and wealth for the United States annually for the years 1948-1998. The data matrix will consist of three columns corresponding to the variables consumption, income, and wealth, while there will be 51 rows (observations) corresponding to the years 1948 to 1998.

Consider the following data set. It corresponds to some data collected at a local recreation association and consists of the year, the number of members, the annual membership dues, and the consumer price index. The easiest way to get these data into Stata is to use Stata's Data Editor. Click on Data on the task bar then click on Data Editor in the resulting drop down menu, or click on the Data Editor button. The result should look like this.

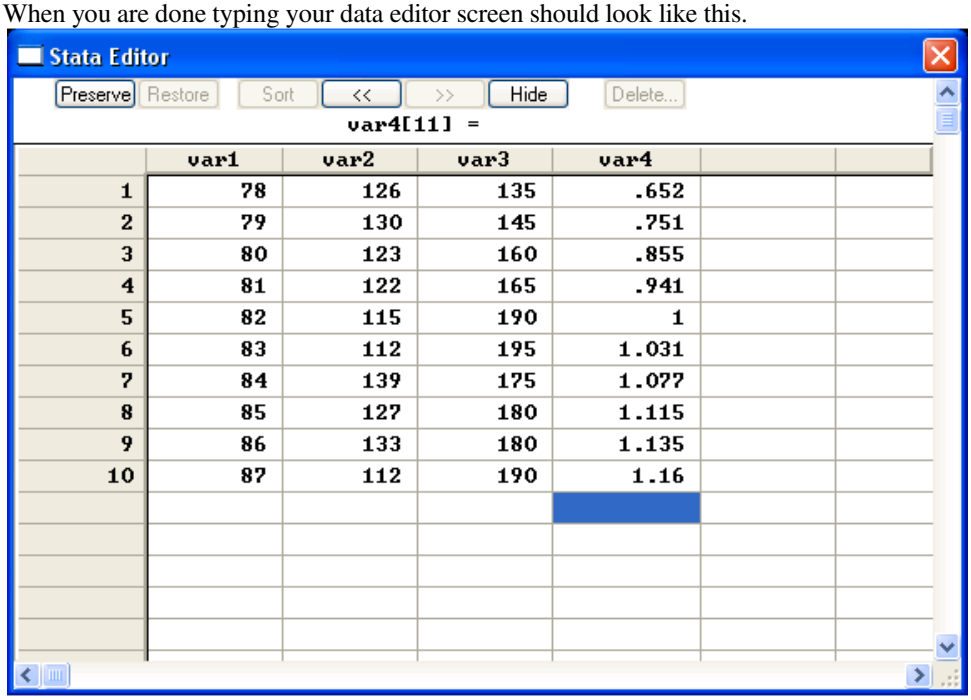


Type the following numbers into the corresponding columns of the data editor. Don't type the variable names, just the numbers. You can type across the rows or down the columns. If you type across the rows, use tab to enter the number. If you type down the columns, use the enter key after typing each value.

Year	Members	Dues	CPI
78	126	135	.652
79	130	145	.751
80	123	160	.855
81	122	165	.941
82	115	190	1.000
83	112	195	1.031
84	139	175	1.077
85	127	180	1.115

86	133	180	1.135
87	112	190	1.160

When you are done typing your data editor screen should look like this.



	var1	var2	var3	var4
1	78	126	135	.652
2	79	130	145	.751
3	80	123	160	.855
4	81	122	165	.941
5	82	115	190	1
6	83	112	195	1.031
7	84	139	175	1.077
8	85	127	180	1.115
9	86	133	180	1.135
10	87	112	190	1.16

Now rename the variables. Double click on "var1" at the top of column one and type "year" into the resulting dialog box. Repeat for the rest of the variables (members, dues, cpi). Notice that the new variable names are in the variables box in the lower left. Click on "preserve" and then the red x on the data editor header to exit the data editor.

Transforming variables

Since we have seen the data, we don't have to look at it any more. Let's move on to the transformations. Suppose we want Stata to compute the logarithms of members and dues and create a dummy variable for an advertising campaign that took place between 1984 and 1987. Type the following commands one at a time into the "Stata command" box along the bottom. Hit <enter> at the end of each line to execute the command. Gen is short for "generate" and is probably the command you will use most. The list command shows you the current data values.

```
gen logmem=log(members)
gen logdues=log(dues)
gen dum=(year>83)
list
[output suppressed]
```

We almost always want to transform the variables in our dataset before analyzing them. In the example above, we transformed the variables members and dues into logarithms. For a Phillips curve, you might want the rate of inflation as the dependent variable while using the inverse of the unemployment rate as the independent variable. All transformations like these are accomplished using the gen command. We have already seen how to tell Stata to take the logarithms of variables. The logarithm function is one of several in

Stata's function library. Some of the functions commonly used in econometric work are listed in Chapter 2, below. Along with the usual addition, subtraction, multiplication, and division, Stata has exponentiation, absolute value, square root, logarithms, lags, differences and comparisons. For example, to compute the inverse of the unemployment rate, we would write,

```
gen invunem=1/unem
```

where unem is the name of the unemployment variable.

As another example, suppose we want to take the first difference of the time series variable GNP (defined as GNP(t)-GNP(t-1)), To do this we first have to tell Stata that you have a time series data set using the tsset command.

```
tsset year
```

Now you can use the "d." operator to take first differences (Δy_t).

```
gen dgnp=d.gnp
```

I like to use capital letters for these operators to make them more visible.

gen dcpi=D.cpi

The second difference $\Delta\Delta cpi_t = \Delta^2 cpi_t = \Delta(cpi_t - cpi_{t-1}) = \Delta cpi_t - \Delta cpi_{t-1}$ be calculated as,

```
gen d2cpi=D.dcpi
```

or

```
gen d2cpi=DD.cpi
```

or

```
gen d2cpi=D2.cpi
```

Lags can also be calculated. For example, the one period lag of the CPI (cpi(t-1)) can be calculated as,

```
gen cpi_1=L.cpi
```

These operators can be combined. The lagged difference Δcpi_{t-1} can be written as

```
gen dcpi_1=LD.cpi
```

It is often necessary to make comparisons between variables. In the example above we created a dummy variable, called "DUM," by the command

```
gen dum=(year>83)
```

This works as follows: for each line of the data, Stata reads the value in the column corresponding to the variable YEAR. It then compares it to the value 83. If YEAR is greater than 83 then the variable "DUM" is given the value 1 ("true"). If YEAR is not greater than 83, "DUM" is given the value 0 ("false").

Continuing the example

Let's continue the example by doing some more transformations and then some analyses. Type the following commands (indicated by the `courier new` typeface), one at a time, ending with <enter> into the Stata command box.

```
replace cpi=cpi/1.160
```

(Whenever you redefine a variable, you must use replace instead of generate.)

```
gen rdues=dues/cpi
gen logrdues=log(rdues);
gen trend=year-78
list
summarize
```

```
Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        year |        10        82.5     3.02765         78         87
     members |        10       123.9    8.999383        112        139
       dues |        10       171.5    20.00694        135        195
         cpi |        10       .9717    .1711212       .652       1.16
      logmem |        10    4.817096    .0727516   4.718499   4.934474
-------------+--------------------------------------------------------
     logdues |        10    5.138088    .1219735   4.905275      5.273
         dum |        10          .4    .5163978          0          1
       trend |        10         5.5     3.02765          1         10
```

(Note: rdues is real (inflation-adjusted) dues, the log function produces natural logs, and the trend is simply a counter (0, 1, 2, etc. for each year, usually called "t" in theoretical discussions.) The summarize command produces means, standard deviations, etc.

We are now ready to do the analysis, a simple ordinary least squares (OLS) regression of members on real dues.

```
regress members rdues
```

This command produces the following output.

```
      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  1,     8) =    0.69
       Model |  58.1763689     1  58.1763689           Prob > F      =  0.4290
    Residual |  670.723631     8  83.8404539           R-squared     =  0.0798
-------------+------------------------------           Adj R-squared = -0.0352
       Total |       728.9     9  80.9888889           Root MSE      =  9.1564

------------------------------------------------------------------------------
     members |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       rdues |  -.1310586   .1573327    -0.83   0.429    -.4938684    .2317512
       _cons |   151.0834   32.76126     4.61   0.002     75.53583     226.631
------------------------------------------------------------------------------
```

Reading Stata Output

Most of this output is self-explanatory. The regression output immediately above is read as follows. The upper left section is called the Analysis of Variance. Under SS is the model sum of squares (58.1763689). This number is referred to in a variety of ways. Each textbook writer chooses his or her own convention. For example, the model sum of squares is also known as the explained sum of squares, ESS (or, sometimes SSE), or the regression sum of squares, RSS (or SSR). Mathematically, it is $\sum_{i=1}^{N} \hat{y}_i^2$ where the circumflex indicates that y is predicted by the regression. The fact that y is in lower case indicates that it is expressed as deviations from the sample mean The Residual SS (670.723631) is the residual sum of squares (a.k.a. RSS, SSR unexplained sum of squares, SSU, USS, and error sum of squares, ESS, SSE). I will refer to the model sum of squares as the regression sum of squares, RSS and the residual sum of squares as the error sum of squares, ESS. The sum of the RSS and the ESS is the total sum of squares, TSS. The next column contains the degrees of freedom. The model degrees of freedom is the number of independent variables, not including the constant term. The residual degrees of freedom is the number of observations, or sample size, N (=10) minus the number of parameters estimated, k (=2), so that N-k=8. The third column is the "mean square" or the sum of squares divided by the degrees of freedom. The residual mean square is also known as the "mean square error" (MSE), the residual variance, or the error variance. Mathematically it is

$$\hat{\sigma}^2 = \hat{\sigma}_u^2 = \left(\sum_{i=1}^{N} e_i^2 \Big/ (N-k) \right)$$

Where e is the residual from the regression (the difference between the actual value of the dependent variable and the predicted value from the regression:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$
$$Y_i = \hat{Y}_i + e_i$$
$$= \hat{\alpha} + \hat{\beta} X + e_i$$
$$e_i = Y_i - \hat{\alpha} - \hat{\beta} X_i$$

The panel on the top right has the number of observations, the overall F-statistic that tests the null hypothesis that the R-squared is equal to zero, the prob-value corresponding to this F-ratio, R-squared, the R-squared adjusted for the number of explanatory variables (a.k.a. R-bar squared), and the root mean square error (the square root of the residual sum of squares divided by its degrees of freedom, also known as the standard error of the estimate, SEE, or the standard error of the regression, SER). Since the MSE is the error variance, the SEE is the corresponding standard deviation (square root). The bottom table shows the variable names, the corresponding estimated coefficients, $\hat{\beta}$, the standard errors, $S_{\hat{\beta}}$ the t-ratios, the prob-values, and the confidence interval around the estimates.

The estimated coefficient for rdues is

$$\hat{\beta} = \sum xy / \sum x^2 \; (= -.1310586) .$$

{The other coefficient is the intercept or constant term, $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} \; (= 151.0834)$.}

The corresponding standard error for $\hat{\beta}$ is $S_{\hat{\beta}} = \sqrt{\dfrac{\hat{\sigma}^2}{\sum x^2}} \; (= .1573327)$.

The resulting t-ratio for the null hypothesis that β=0 is

$$T = \hat{\beta} / S_{\hat{\beta}} \; (= -0.83) .$$

The prob-value for the t-test of no significance is the probability that the test statistic T would be observed if the null hypothesis is true, using Student's t distribution, two-tailed. If this value is smaller than .05 then

we “reject the null hypothesis that β=0 at the five percent significance level.” We usually just say that β is significant.

Let’s add the time trend and the dummy for the advertising campaign and do a multiple regression, using the logarithm of members as the dependent variable and the log of real dues as the price.

```
regress logmem logrdues trend dum
```

```
      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  3,     6) =    8.34
       Model |  .038417989     3  .012805996           Prob > F      =  0.0146
    Residual |  .009217232     6  .001536205           R-squared     =  0.8065
-------------+------------------------------           Adj R-squared =  0.7098
       Total |  .047635222     9  .005292802           Root MSE      =  .03919

------------------------------------------------------------------------------
      logmem |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    logrdues |  -.6767872     .36665    -1.85   0.114    -1.573947    .2203729
       trend |  -.0432308   .0094375    -4.58   0.004    -.0663234   -.0201382
         dum |    .155846   .0608159     2.56   0.043      .007035    .3046571
       _cons |   8.557113   1.989932     4.30   0.005     3.687924     13.4263
------------------------------------------------------------------------------
```

The coefficient on logrdues is the elasticity of demand for pool membership. (Mathematically, the derivative of a log with respect to a log is an elasticity.) Is the demand curve downward sloping? Is price significant in the demand equation at the five percent level? At the ten percent level? The trend is the percent change in pool membership per year. (The derivative of a log with respect to a non-logged value is a percent change.) According to this estimate, the pool can expect to lose four percent of its members every year, everything else staying the same (*ceteris paribus*). On the other hand, continuing the advertising campaign will generate a 16 percent increase in pool memberships every year. Note that, since this is a multiple regression, the estimated coefficients, or parameters, are the change in the dependent variable for a one-unit change in the corresponding independent variable, holding everything else constant.

Saving your first Stata analysis

Let’s assume that you have created a directory called “project” on the H: drive. To save the data for further analysis, type

```
save “H:\Project\test.dta”
```

will create a Stata data set called “test.dta” in the directory H:\Project. You don’t really need the quotes unless there are spaces in the pathname. However, it is good practice. You can read the data into Stata later with the command,

```
use “H:\Project\test.dta”
```

You can also save the data by clicking on File, Save and browsing to the Project folder or by clicking on the little floppy disk icon on the button bar.

You can save your output by going to the top of the output window (use the slider on the right hand side), clicking on the first bit of output you want to save, then drag the cursor, holding the mouse key down, to the bottom of the output window. Once you have blocked the text you want to save. You can type <control>-c for copy (or click on edit, copy text), go to your favorite word processor, and type <control>-v for paste (or click on edit, paste). You could also create a log file containing your results. We discuss how to create log files below.

You can then write up your results around the Stata output.

2 STATA LANGUAGE

This section is designed to provide a more detailed introduction to the language of Stata. However, it is still only an introduction, for a full development you must refer to the relevant Stata manual, or use Stata Help (see Chapter 6).

Basic Rules of Stata

1. Rules for composing variable names in Stata.

a. Every name must begin with a letter or an underscore. However, I recommend that you not use underscore at the beginning of any variable name, since Stata frequently uses that convention to make its own internal variables. For example, _n is Stata's variable containing the observation number.

b. Subsequent characters may be letters, numbers, or underscores, e.g., var1. You cannot use special characters (&,#, etc.).

c. The maximum number of characters in a name is 32, but shorter names work best.

d. **Case matters**. **The names Var, VAR, and var are all considered to be different variables in Stata notation. I recommend that you use lowercase only for variable names.**

e. The following variable names are reserved. You may not use these names for your variables.

_all	double	long	_rc	_b	float
_n	_se	byte	if	_N	_skip
_coeff	in	_pi	using	_cons	int
_pred	with	e			

2. Command syntax.

The typical Stata command looks like this.

[**by** varlist:] command [varlist] [**=**exp] [if exp] [in range] [weight] [, options]

where things in a square bracket are optional. Actually the varlist is almost always required.

A varlist is a list of variable names separated by spaces. If no varlist is specified, then, usually, all the variables are used. For example, in the simple program above, we used the command "list" with no varlist attached. Consequently, Stata printed out all the data for all the variables currently in memory. We could have said, `list members dues rdues` to get a list of those three variables only.

The "=exp" option is used for the "generate" and "replace" commands, e.g., gen logrdues=log(dues) and replace cpi=cpi/1.160.

The "if exp" option restricts the command to a certain subset of observations. For example, we could have written

```
list year members rdues if year == 83.
```

This would have produced a printout of the data for year, members, and rdues for 1983. Note that Stata requires two equal signs to denote equality because one equal sign indicates assignment in the generate and replace commands. (Obviously, the equal sign in the expression `replace cpi=cpi/1.160` indicates that the variable cpi is to take the value of the old cpi divided by 1.160. It is not testing for the equality of the left and right hand sides of the expression.)

The "in range" option also restricts the command to a certain range. It takes the form `in number1/number2`. For example,

```
list year members rdues in 5/10
```

would list the observations from 5 to 10. We can also use the letter f (first) and l (last) in place of either number1 or number 2.

The weight option is always in square brackets and takes the form

```
[weightword=exp]
```

where the weight word is one of the following (default weight for each particular command)

> `aweight` (analytical weights: inverse of the variance of the observation. This is the most commonly used weight by economists. Each observation in our typical sample is an average across counties, states, etc. The aweight is the variance divided by the number of elements, usually population, that comprise the county, state, etc. For most Stata commands, the aweight is rescaled to sum to the number of observations in the sample.)
>
> `fweight` (frequency weights: indicates duplicate observations, an fweight of 100 indicates that there are really 100 identical observations)
>
> `pweight` (sampling weights: the inverse of the probability that this observation is included in the sample. A pweight of 100 indicates that this observation is representative of 100 subjects in the population.)
>
> `iweight` (importance weight: indicates the relative importance of each observation. The definition of iweight depends on how it is defined in each command that uses them).

Options. Each Stata command takes command-specific options. Options must be separated from the rest of the command by a comma. For example, the "reg" command that we used above has the option "robust" which produces heteroskedastic-consistent standard errors. That is,

```
reg members rdues trend, robust
```

by varlist: Almost all Stata commands can be repeated automatically for each value of the variable or variables in the by varlist. This option must precede the command and be separated from the rest of the command by a colon. The data must be sorted by the varlist. For example the commands,

```
sort dum
by dum: summarize members rdues
```

will produce means, etc. for the years without an advertising campaign (dum=0) and the years with an advertising campaign (dum=1). In this case the sort command is not necessary, since the data are already sorted by dum, but it never hurts to sort before using the by varlist option.

Arithmetic operators are:

+	add
-	subtract
*	multiply
/	divide
^	raise to a power

Comparison operators

==	equal to
~=	not equal to
>	greater than
<	less than
>=	greater than or equal to
<=	less than or equal to

Logical operators

&	and
|	or
~	not

Comparison and logical operators return a value of 1 for true and 0 for false. The order of operations is: ~, ^, -(negation), /, *, -(subtraction), +, ~=, >, <, <=, >=, ==, &, |.

Stata functions

These are only a few of the many functions in Stata, see STATA USER'S GUIDE for a comprehensive list.

Mathematical functions include:

abs(x)	absolute value
max(x)	returns the largest value
min(x)	returns the smallest value
sqrt(x)	square root
exp(x)	raises e (2.17828) to a specified power
log(x)	natural logarithm

Some useful statistical functions are the following.

Chi2(df, x) returns the cumulative value of the chi-square with df degrees of freedom for a value of x.

For example, suppose we know that x=2.05 is distributed according to chi-square with 2 degrees of freedom. What is the probability of observing a number as large as 2.05 if the true value of x is zero? We can compute the following in Stata. Because we are not generating variables, we use the scalar command.

```
. scalar prob1=chi2(2,2.05)

. scalar list prob1
     prob1 =  .64120353
```

We apparently have a pretty good chance of observing a number as large as 2.05.

To find out if x=2.05 is significantly different from zero in the chi-square with 2 degrees of freedom we can compute the following to get the prob value (the area in the right hand tail of the distribution).

```
. scalar prob=1-chi2(2,2.05)

. scalar list prob
      prob =  .35879647
```

Since this value is not smaller than .05 it is not significantly different from zero at the .05 level.

Chi2tail(df,x) returns the upper tail of the same distribution. We should get the same answer as above using this function.

```
. scalar prob2=chi2tail(2,2.05)

. scalar list prob2
     prob2 =  .35879647
```

Invchi2(df,p) returns the value of x for which the probability is p and the degrees of freedom are df.

```
. scalar x1=invchi2(2,.641)

. scalar list x1
        x1 =  2.0488658
```

Invchi2tail is the inverse function for chi2tail.

```
. scalar x2=invchi2tail(2,.359)

. scalar list x2
        x2 =  2.0488658
```

Norm(x) is the cumulative standard normal. To get the probability of observing a number as large as 1.65 if the mean of the distribution is zero (and the standard deviation is one), i.e., P(z<1.65) is

```
. scalar prob=norm(1.65)

. scalar list prob
      prob =  .9505285
```

Invnorm(prob) is the inverse normal distribution so that, if prob=norm(z) then z=invnorm(prob).

Wooldridge uses the Stata functions norm, invttail, invFtail, and invchi2tail to generate the normal, t-, F-, and chii-square tables in the back of his textbook, which means that we don't have to look in the back of textbooks for statistical distributions any more.

Uniform() returns uniformly distributed pseudo-random numbers between zero and one. Even though uniform() takes no parameters, the parentheses are part of the name of the function and must be typed. For example to generate a sample of uniformly distributed random numbers use the command

```
. gen v=uniform()

. summarize v

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           v |        51    .4819048    .2889026   .0445188   .9746088
```

To generate a sample of normally distributed random numbers with mean 2 and standard deviation 10, use the command,

```
. gen z=2+10*invnorm(uniform())

. summarize z

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           z |        51     1.39789    8.954225  -15.87513   21.83905
```

Note that both these functions return 51 observations on the pseudo-random variable. Unless you tell Stata otherwise, it will create the number of observations equal to the number of observations in the data set. I used the crime1990.dta data set which has 51 observations, one for each state and DC.

The pseudo random number generator uniform() generates the same sequence of random numbers in each session of Stata, unless you reinitialize the seed. To do that, type,

```
set seed #
```

If you set the seed to # you get a different sequence of random numbers. This makes it difficult to be sure you are doing what you think you are doing. Unless I am doing a series of Monte Carlo exercises, I leave the seed alone.

Missing values

Missing values are denoted by a single period (.) Any arithmetic operation on a missing value results in a missing value. Stata considers missing values as larger than any other number. So, sorting by a variable which has missing values will put all observations corresponding to missing at the end of the data set. When doing statistical analyses, Stata ignores (drops) observations with missing values. So, for example if you did a regression of crime on police and the unemployment rate across states, Stata will drop any observations for which there are missing values in either police or unemployment ("case wise deletion"). In other commands such as the summarize command, for example, Stata will simply ignore any missing values for each variable separately when computing the mean, standard deviations, etc.

Time series operators.

The tsset command must be used to indicate that the data are time series. For example.

```
      sort year
```

```
tsset year
```

Lags: the one-period lag of a variable such as the money supply, ms, can be written as

```
L.ms
```

So that L.ms is equal to ms(t-1). Similarly, the second period lag could be written as

```
L2.ms
```

or, equivalently, as

`LL.ms`, etc.

For example, the command

```
gen rdues_1 = L.rdues
```

will generate a new variable called rdues_1 which is the one-period lag of rdues. Note that this will produce a missing value for the first observation of rdues_1.

Leads: the one-period lead is written as

`F.ms` (=ms(t+1)).

The two-period lead is written as `F2.ms` or `FF.ms`.

Differences: the first-difference is written as
`D.ms` (=ms(t) - ms(t-1)).

The second difference (difference of difference) is

`D2.ms` or `DD.ms`. (=(ms(t)-ms(t-1)) – (ms(t-1)-ms(t-2)))

Seasonal difference: the first seasonal difference is written

`S.ms`

and is the same as

`D.ms`.

The second seasonal difference is

`S2.ms`

and is equal to ms(t)-ms(t-2).

Time series operators can be combined. For example, for monthly data, the difference of the 12-month difference would be written as `DS12.ms`. Time series operators can be written in upper or lowercase. I recommend upper case so that it is more easily readable as an operator as opposed to a variable name (which I like to keep in lowercase).

Setting panel data.

If you have a panel of 50 states (1-50) across 20 years (1977-1996), you can use time series operators, but you must first tell Stata that you have a panel (pooled time series and cross-section). The following commands sort

the data, first by state, then by year, define state and year as the cross-section and time series identifiers, and generate the lag of the variable x.

```
sort state year
tsset state year
gen x_1=L.x
```

Variable labels.

Each variable automatically has an 80 character label associated with it, initially blank. You can use the "label variable" command to define a new variable label. For example,

Label variable dues "Annual dues per family"
Label variable members "Number of members"

Stata will use variable labels, as long as there is room, in all subsequent output. You can also add variable labels by invoking the Data Editor, then double clicking on the variable name at the head of each column and typing the label into the appropriate space in the dialog box.

System variables

System variables are automatically produced by Stata. Their names begin with an underscore.

_n is the number of the current observation

You can create a counter or time trend if you have time series data by using the command

```
gen t=_n
```

_N is the total number of observations in the data set.

_pi is the value of π to machine precision.

3 DATA MANAGEMENT

Getting your data into Stata

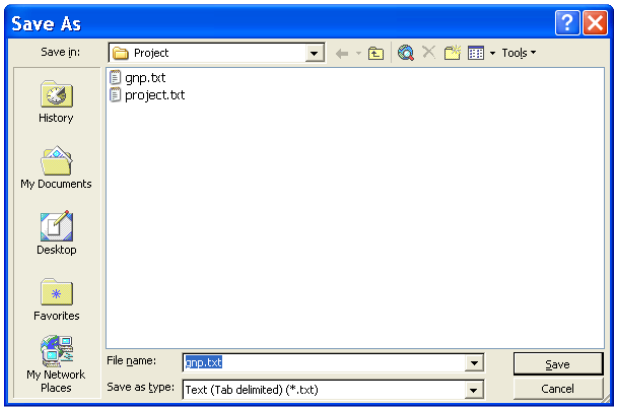
Using the data editor

We already know how to use the data editor from the first chapter.

Importing data from Excel

Generally speaking, the best way to get data into Stata is through a spreadsheet program such as Excel. For example, data presented in columns in a text file or Web page can be cut and pasted into a text editor or word processor and saved as a text file. This file is not readable as data because such files are organized as rows rather than columns. We need to separate the data into columns corresponding to variables.

Text files can be opened and read into Excel using the Text Import Wizard. Click on File, Open and browse to the location of the file. Follow the Text Import Wizard's directions. This creates separate columns of numbers, which will eventually correspond to variables in Stata. After the data have been read in to Excel, the data can be saved as a text file. Clicking on File, Save As in Excel reveals the following dialogue box.



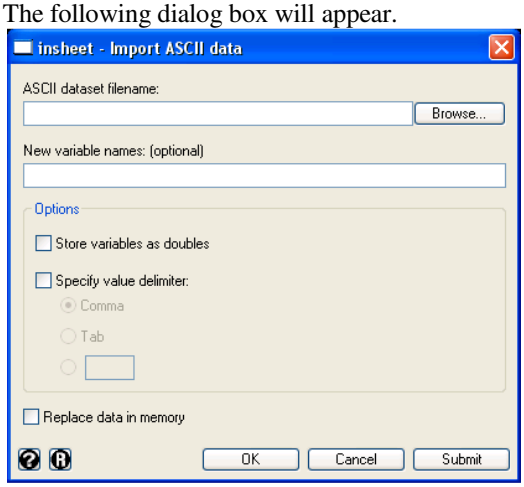
If the data have been read from a .txt file, Excel will assume you want to save it in the same file. Clicking on Save will save the data set with tabs between data values. Once the tab delimited text file has been created, it can be read into Stata using the insheet command.

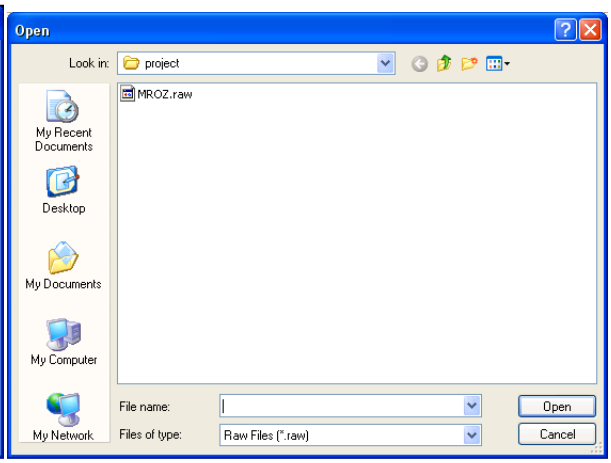
```
Insheet using filename
```

For example,

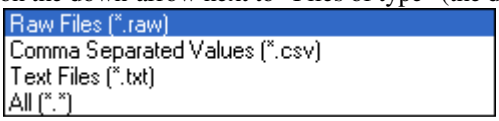
```
insheet using "H:\Project\gnp.txt"
```

The insheet command will figure out that the file is tab delimited and that the variable names are at the top. The insheet command can be invoked by clicking on File, Import, and "Ascii data created by a spreadsheet." The following dialog box will appear.





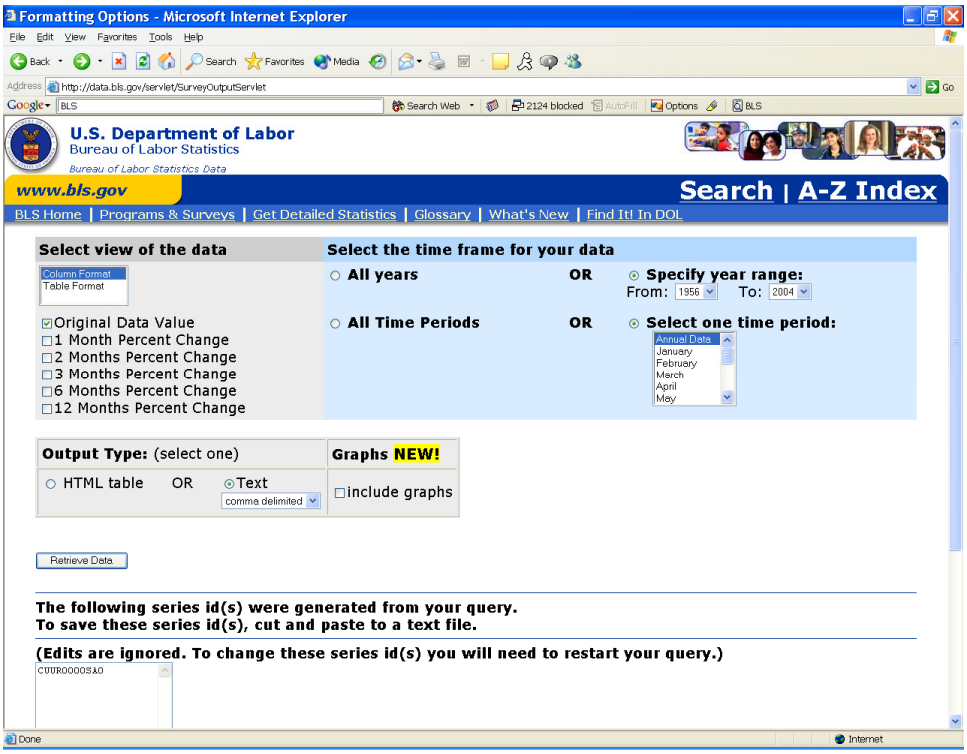
Clicking on Browse allows you to navigate to the location of the file. Once you get to the right directory, click on the down-arrow next to "Files of type" (the default file type is "Raw," whatever that is) to reveal



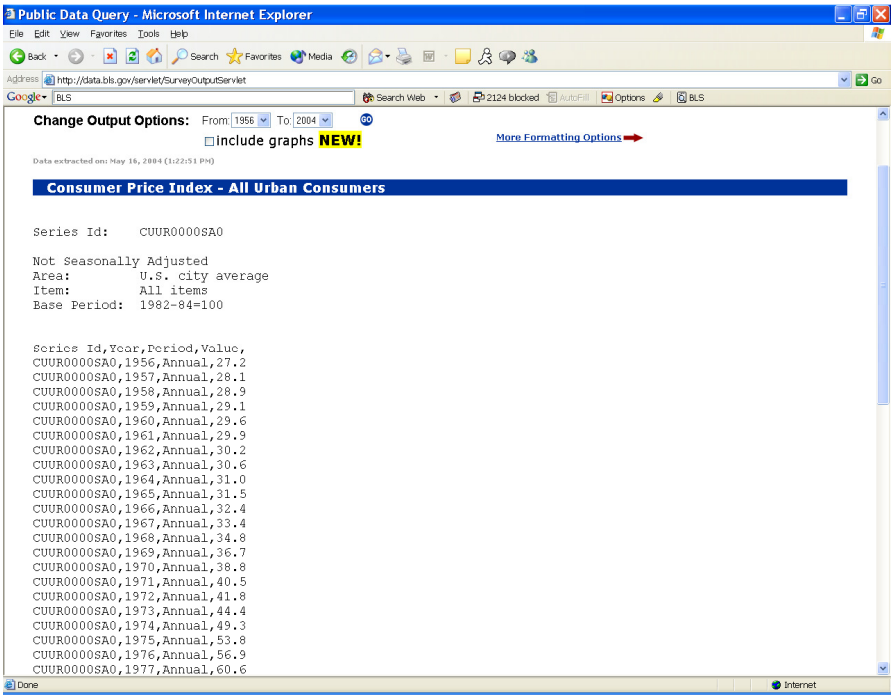
Choose .txt files if your files are tab delimited.

Importing data from comma separated values (.csv) files

The insheet command can be used to read other file formats such as csv (comma separated values) files. Csv files are frequently used to transport data. For example, the Bureau of Labor Statistics allows the public to download data from its website. The following screen was generated by navigating the BLS website http://www.bls.gov/.



Note that I have selected the years 1956-2004, annual frequency, column view, and text, comma delimited (although I could have chosen tab delimited and the insheet command would read that too). Clicking on Retrieve Data yields the following screen.



Use the mouse (or hold down the shift key and use the down arrow on the keyboard) to highlight the data starting at Series Id, copy (control-c) the selected text, invoke Notepad, WordPad, or Word, and paste (control-v) the selected text. Now Save As a csv file, e.g., Save As cpi.csv. (Before you save, you should delete the comma after Value, which indicates that there is another variable after Value. This creates a

variable, v5, with nothing but missing values in Stata. You don't really have to do this because the variable can be dropped in Stata, but it is neater if you do.) The Series Id and period can be dropped once the data are in Stata.

You can read the data file with the insheet command, e.g.,

```
insheet using "H:\Project\cpi.csv"
```

or you can click on Data, Import, and Ascii data created by a spreadsheet, as above.

Reading Stata files: the use command

If the data are already in a Stata data set (with a .dta suffix), it can be read with the "use" command. For example, suppose we have saved a Stata file called project.dta in the H:\Project directory. The command to read the data into Stata is

```
use "h:\Project\project.dta", clear
```

The "clear" option is to remove any unwanted or leftover data from previous analyses. Alternatively you can click on the "open (use)' button on the taskbar or click on File, Open.

Saving a Stata data set: the save command

Data in memory can be saved in a Stata (.dta) data set with the "save" command. For example, the cpi data can be saved by typing

```
save h:\Project\cpi.dta"
```

If you want the data set to be readable by older versions of Stata use the "saveold" command,

```
saveold "h:\Project\cpi.dta"
```

The quotes are not required. However, you might have trouble saving (or using) files without the quotes if your pathname contains spaces.

If a file by the name of cpi.dta already exists, e.g., you are updating a file, you must tell Stata that you want to overwrite the old file. This is done by adding the option, replace.

```
save "h:\Project\cpi.dta", replace.
```

Combining Stata data sets: the Append and Merge commands

Suppose we want to combine the data in two different Stata data sets. There are two ways in which data can be combined. The first is to stack one data set on top of the other, so that some or all the series have additional observations. The "append" command is used for this purpose. For example, suppose we have a

data set called “project.dta” which has data on three variables (gnp, ms, cpi) from 1977 to 1998 and another called “project2.dta.” which has data on two of the three variables (ms and cpi) for the years 1999-2002. To combine the data sets, we first read the early data into Stata with the use command

```
use “h:\Project\project.dta”
```

The early data are now residing in memory. We append the newer data to the early data as follows

```
append using “h:\Project\project2.dta”
```

At this point you probably want to list the data and perhaps sort by year or some other identifying variable. Then save the expanded data set with the save command. If you do not get the result you were expecting, simply exit Stata without saving the data.

The other way to combine data sets is to merge two sets together. This is concatenating horizontally instead of vertically. It makes the data set wider in the sense that, if one data set has variables that the other does not, the combined data set will have variables from both. (Of course, there could be some overlap in the variables, so one data set will overwrite the data in the other. This is another way of updating a data set.)

There are two kinds of merges. The first is the one-to-one merge where the first observation of the first data set is matched with the first observation of the second data set. You must be very careful that each data set has its observations sorted exactly the same way. I do not recommend one to one merges as a general principle.

The preferred method is a “match merge,” where the two datasets have one or more variables in common. These variables act as identifiers so that Stata can match the correct observations. For example, suppose you have data on the following variables: year pop crmur crrap crrob crass crbur from 1929 to 1998 in a data set called crime1.dta in the h:\Project directory. Suppose that you also have a data set called arrests.dta containing year pop armur arrap arrob arass arbur (corresponding arrest rates) for the years 1952-1998. You can combine these two data sets with the following commands

Read and sort the first data set.

```
use “H:\Project\crime1.dta”, clear
sort year
list
save “H:\Project\crime1.dta”, replace
```

Read and sort the second data set.

```
use “H:\Project\arrests.dta”, clear
sort year
list
save “H:\Project\arrests.dta”, replace
```

Read the first data set back into memory, clearing out the data from the second file. This is the “master” data set.

```
use “H:\Project\crime1.dta”, clear
```

Merge with the second data set by year. This is the “using” data set.

```
merge year using “H:\Project\arrests.dta”
```

Look at the data, just to make sure.

```
list
```

Assuming it looks ok, save the combined data into a new data set.

```
save "H:\Project\crime2.dta", replace
```

The resulting data set will contain data from 1929-1998, however the arrest variables will have missing values for 1929-1951.

The data set that is in memory (crime1) is called the "master." The data set mentioned in the merge command (arrests) is called the "using" data set. Each time Stata merges two data sets, it creates a variable called _merge. This variable takes the value 1 if the value only appeared in the master data set; 2 if the value occurred only in the using data set; and 3 if an observation from the master is joined with one from the using.

Note also that the variable pop occurs in both data sets. Stata uses the convention that, in the case of duplicate variables, the one that is in memory (the master) is retained. So, the version of pop that is in the crime1 data set is stored in the new crime2 data set.

If you want to update the values of a variable with revised values, you can use the merge command with the update replace options.

```
merge year using h:\Project\arrests.dta, update replace
```

Here, data in the using file overwrite data in the master file. In this case the pop values from the arrest data set will be retained in the combined data set. The _merge variable can take two additional values if the update option is invoked: _merge equals 4 if missing values in the master were replaced with values from the using and _merge equals 5 if some values of the master disagree with the corresponding values in the using (these correspond to the revised values). You should always list the _merge variable after a merge.

If you just want to overwrite missing values in the master data set with new observations from the using data set (so that you retain all existing data in the master), then don't use the replace option:

```
merge year using h:\Project\arrests.dta, update
list _merge
```

Looking at your data: list, describe, summarize, and tabulate commands

It is important to look at your data to make sure it is what you think it is. The simplest command is

```
list varlist
```

Where varlist is a list of variables separated by spaces. If you omit the varlist, Stata will assume you want all the variables listed. If you have too many variables to fit in a table, Stata will print out the data by observation. This is very difficult to read. I always use a varlist to make sure I get an easy to read table.

The "describe" command is best used without a varlist or options.

```
describe
```

It produces a list of variables, including any string variables. Here is the resulting output from the crime and arrest data set we created above.

```
. describe

Contains data from C:\Project\crime1.dta
  obs:            70
 vars:            13                          16 May 2004 20:12
 size:         3,430 (99.9% of memory free)
------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
year            int    %8.0g
pop             float  %9.0g                 POP
crmur           int    %8.0g                 CRMUR
crrap           long   %12.0g                CRRAP
crrob           long   %12.0g                CRROB
crass           long   %12.0g                CRASS
crbur           long   %12.0g                CRBUR
armur           float  %9.0g                 ARMUR
arrap           float  %9.0g                 ARRAP
arrob           float  %9.0g                 ARROB
arass           float  %9.0g                 ARASS
arbur           float  %9.0g                 ARBUR
_merge          byte   %8.0g
------------------------------------------------------------------------------
Sorted by:
     Note:  dataset has changed since last saved
```

We can get a list of the numerical variables and summary data for each one with the "summarize" command.

```
Summarize varlist
```

This command produces the number of observations, the mean, standard deviation, min, and max for every variable in the varlist. If you omit the varlist, Stata will generate these statistics for all the variables in the data set. For example,

```
. summarize

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        year |      70      1963.5   20.35109       1929       1998
         pop |      70      186049   52068.33     121769     270298
       crmur |      64    14231.02   6066.007       7508      24700
       crrap |      64    44291.63   36281.72       5981     109060
       crrob |      64    288390.6   218911.6      58323     687730
       crass |      64    414834.3   363602.9      62186    1135610
       crbur |      64     1736091    1214590     318225    3795200
       armur |      46    7.516304   1.884076        4.4       10.3
       arrap |      40     12.1675   3.158983          7         16
       arrob |      46    53.39348   17.57911       26.5       80.9
       arass |      46    113.1174   54.79689       49.3        223
       arbur |      46    170.5326   44.00138       97.5      254.1
      _merge |      70     2.371429   .9952267          1          5
```

If you add the detail option,

```
Summarize varlist, detail
```

You will also get the variance, skewness, kurtosis, four smallest, four largest, and a variety of percentiles, but the printout is not as neat.

Don't forget, as with any Stata command, you can use qualifiers to limit the command to a subset of the data. For example, the command

```
summarize if year >1990
```

Will yield summary statistics only for the years 1991-1998.

Finally, you can use the "tabulate" command to generate a table of frequencies. This is particularly useful for large data sets with too many observations to list comfortably.

```
tabulate varname
```

For example, suppose we think we have a panel data set on crime in nine Florida counties for the years 1978-1997. Just to make sure, we use the tabulate command:

```
. tabulate county
```

county	Freq.	Percent	Cum.
1	20	11.11	11.11
2	20	11.11	22.22
3	20	11.11	33.33
4	20	11.11	44.44
5	20	11.11	55.56
6	20	11.11	66.67
7	20	11.11	77.78
8	20	11.11	88.89
9	20	11.11	100.00
Total	180	100.00	

```
. tabulate year
```

year	Freq.	Percent	Cum.
78	9	5.00	5.00
79	9	5.00	10.00
80	9	5.00	15.00
81	9	5.00	20.00
82	9	5.00	25.00
83	9	5.00	30.00
84	9	5.00	35.00
85	9	5.00	40.00
86	9	5.00	45.00
87	9	5.00	50.00
88	9	5.00	55.00
89	9	5.00	60.00
90	9	5.00	65.00
91	9	5.00	70.00
92	9	5.00	75.00
93	9	5.00	80.00
94	9	5.00	85.00
95	9	5.00	90.00
96	9	5.00	95.00
97	9	5.00	100.00
Total	180	100.00	

Nine counties each with 20 years and 20 years each with 9 counties. Looks ok.

The tabulate command will also do another interesting trick. It can generate dummy variables corresponding to the categories in the table. So, for example, suppose we want to create dummy variables for each of the counties and each of the years. We can do it easily with the tabulate command and the generate() option.

```
Tabulate county, generate(cdum)
Tabulate year, generate(yrdum)
```

This will create nine county dummies (cdum1-cdum9) and twenty year dummies (yrdum1-yrdum20).

Culling your data: the keep and drop commands

Occasionally you will want to drop variables or observations from your data set. The command

```
Drop varlist
```

Will eliminate the variables listed in the varlist. The command

```
Drop if year <1978
```

Will drop observations corresponding to years before 1978. The command

```
Drop in 1/10
```

will drop the first ten observations.

Alternatively, you can use the keep command to achieve the same result. The command

```
Keep varlist
```

Will drop all variables not in the varlist.

```
Keep if year >1977
```

will keep all observations after 1977.

Transforming your data: the generate and replace commands

As we have seen above, we use the "generate" command to transform our raw data into the variables we need for our statistical analyses. Suppose we want to create a per capita murder variable from the raw data in the crime2 data set we created above. We would use the command,

```
gen murder=crmur/pop*1000
```

I always abbreviate generate to gen. This operation divides crmur by pop then (remember the order of operations) multiplies the resulting ratio by 1000. This produces a reasonable number for statistical operations. According to the results of the summarize command above, the maximum value for crmur is 24,700. The corresponding value for pop is 270,298. However, we know that the population is about 270 million, so our population variable is in thousands. If we divide 24,700 by 270 million, we get the probability of being murdered (.000091). This number is too small to be useful in statistical analyses. Dividing 24,700 by 270,298 yields .09138 (the number of murders per thousand population) which is still too small. Multiplying the ratio of crmur to our measure of pop by 1000 yields the number of murders per million population (91.38), which is a reasonable order of magnitude for further analysis.

If you try to generate a variable called, say, widget and a variable by that name already exists in the data set, Stata will refuse to execute the command and tell you, "widget already defined." If you want to replace widget with new values, you must use the replace command. We used the replace command in our sample program in the last chapter to re-base our cpi index.

```
replace cpi=cpi/1.160
```

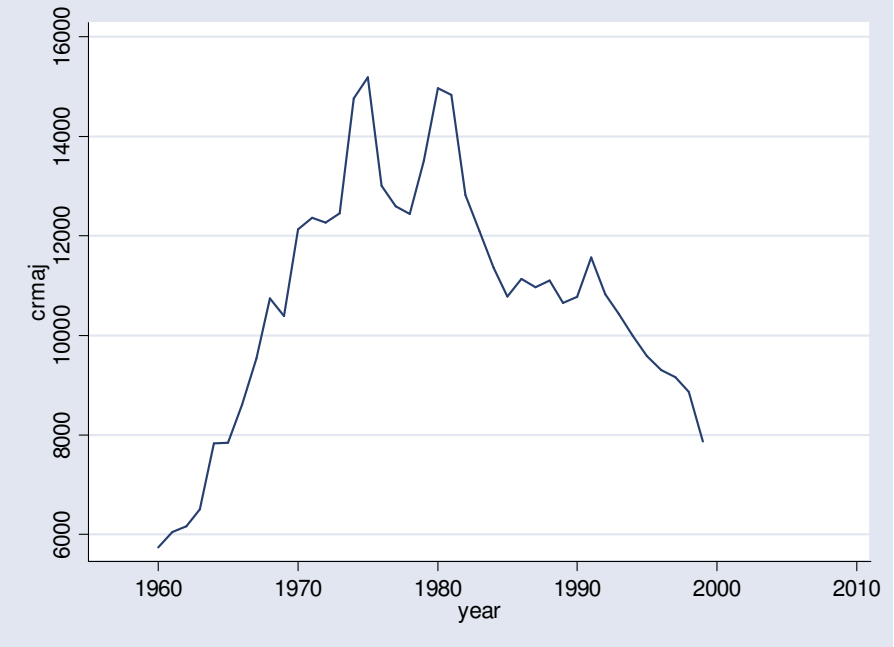
4 GRAPHS

Continuing with our theme of looking at your data before you make any crucial decisions, obviously one of the best ways of looking is to graph the data. The first kind of graph is the line graph, which is especially useful for time series.

Here is a graph of the major crime rate in Virginia from 1960 to 1998. The command used to produce the graph is

```
. graph twoway line crmaj year
```

which produces

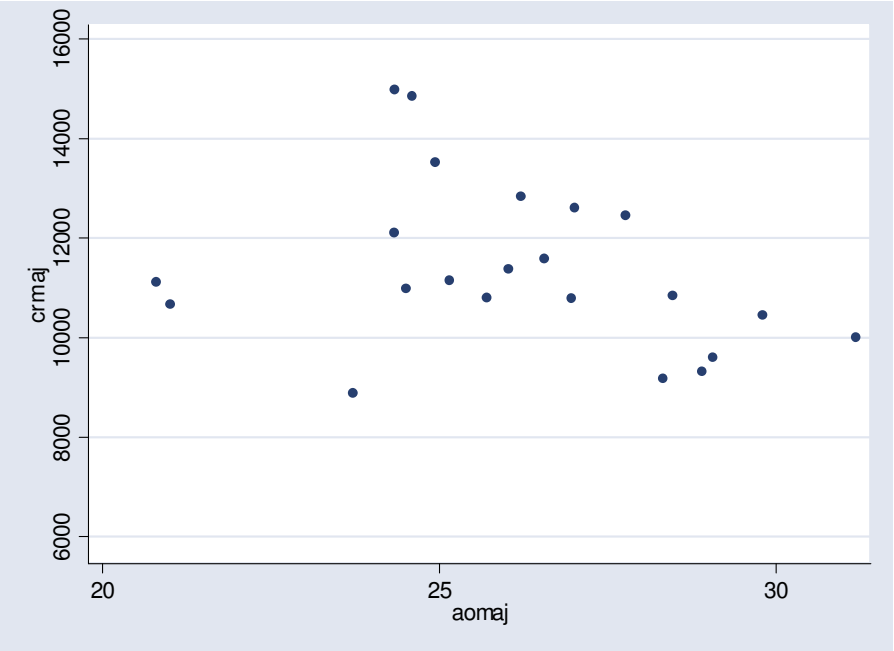


It is interesting that crime peaked in the mid-1970's and has been declining fairly steadily since the early 1980's. I wonder why. Hmmm.

Note that I am able to paste the graph into this Word document by generating the graph in State and clicking on Edit, Copy Graph. Then, in Word, click on Edit, Paste (or control-v if you prefer keyboard shortcuts).

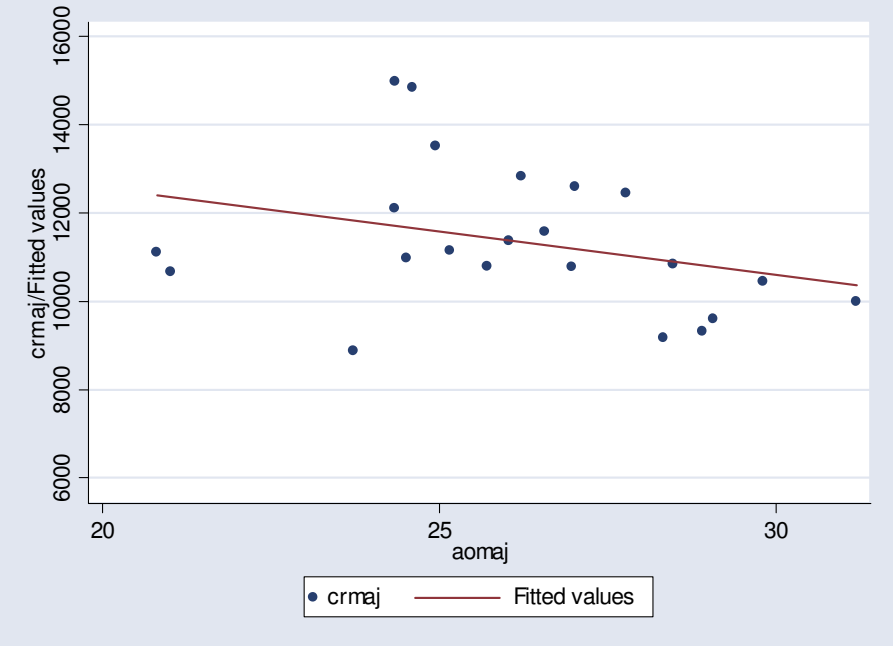
The second type of graph frequently used by economists is the scatter diagram, which relates one variable to another. Here is the scatter diagram relating the arrest rate to the crime rate.

```
. graph twoway scatter crmaj aomaj
```



We also occasionally like to see the scatter diagram with the regression line superimposed.

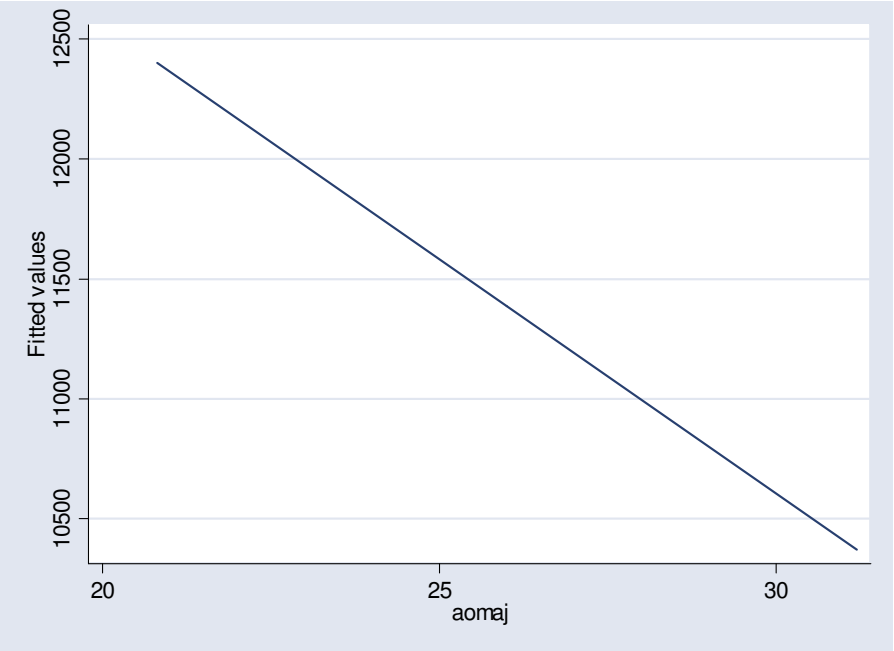
```
graph twoway (scatter crmaj aomaj) (lfit crmaj aomaj)
```



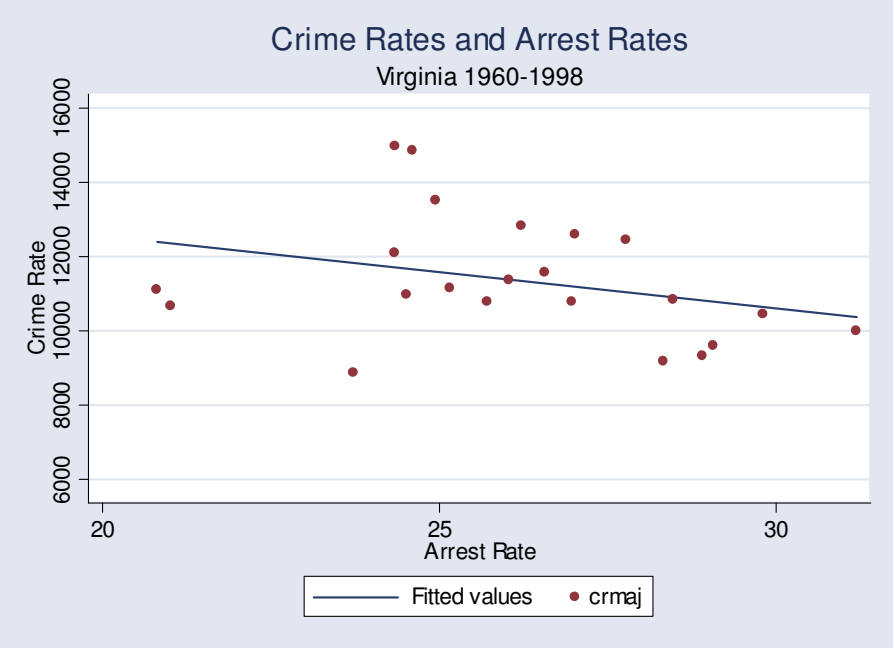
This graph seems to indicate that the higher the arrest rate, the lower the crime rate. Does deterrence work?

Note that we used so-called “parentheses binding” or “() binding” to overlay the graphs. The first parenthesis contains the commands that created the original scatter diagram, the second parenthesis contains the command that generated the fitted line (lfit is short for linear fit). Here is the lfit part by itself.

```
. graph twoway lfit crmaj aomaj
```



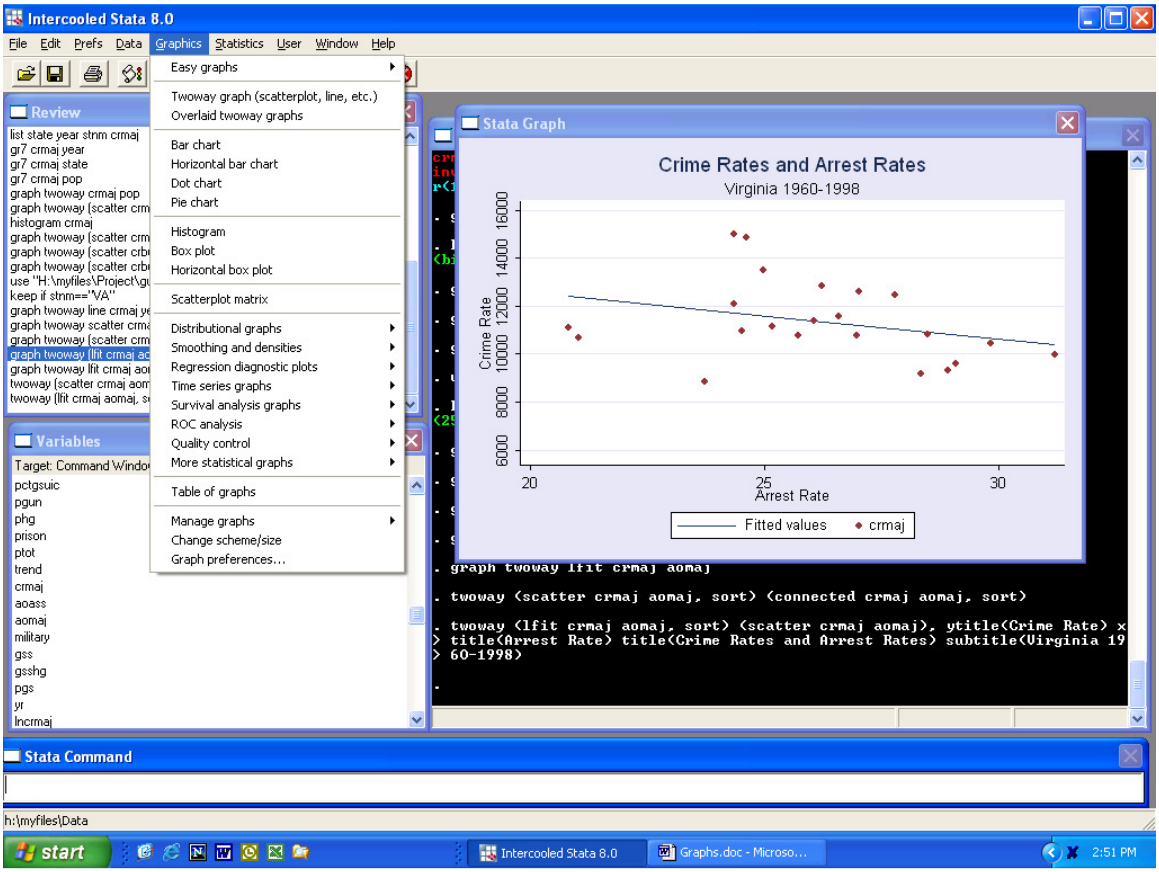
These graphs can be produced interactively. Here is another version of the above graph, with some more bells and whistles.



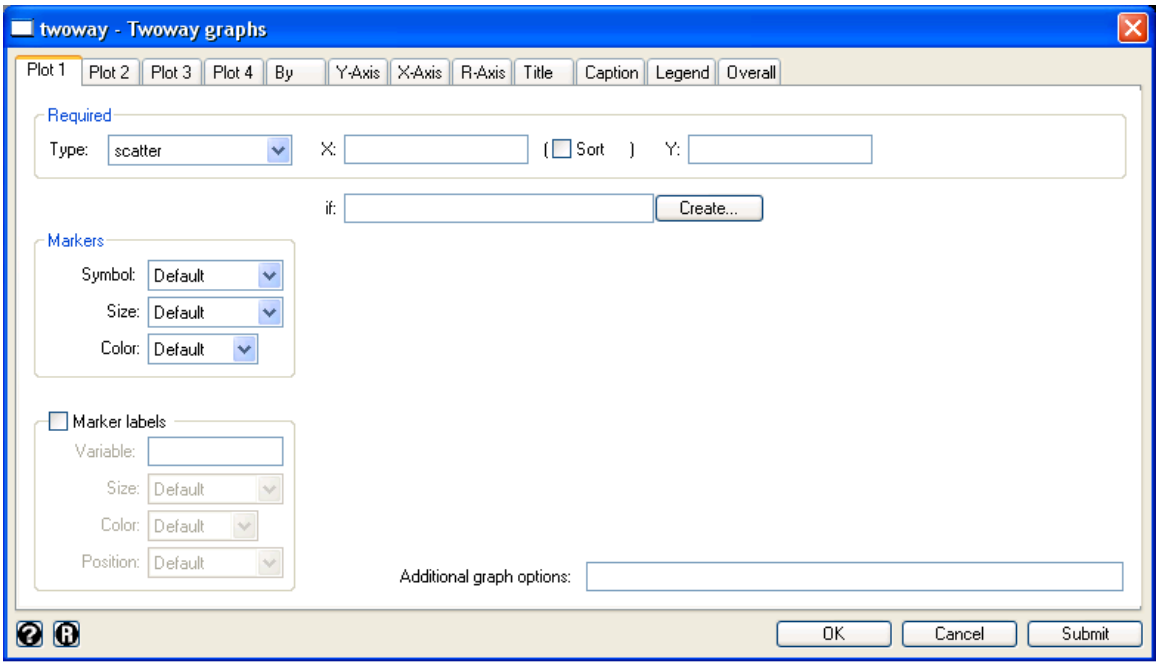
This was produced with the command,

```
. twoway (lfit crmaj aomaj, sort) (scatter crmaj aomaj), ytitle(Crime Rate) x
> title(Arrest Rate) title(Crime Rates and Arrest Rates) subtitle(Virginia 19
> 60-1998)
```

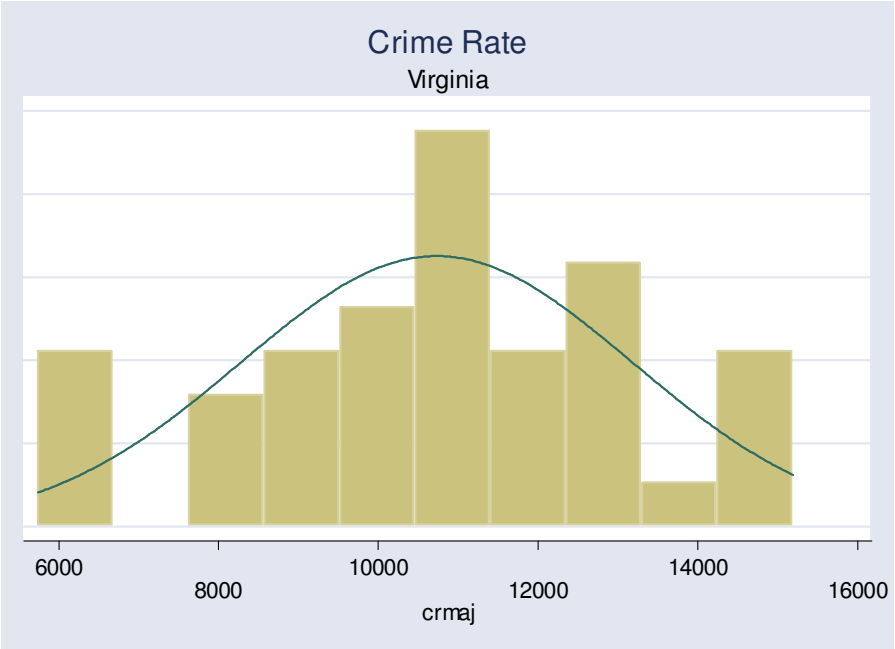
However, I did not have to enter that complicated command, and neither do you. Click on Graphics to reveal the pull-down menu.



Click on “overlaid twoway graphs” to reveal the following dialog box with lots of tabs. Clicking on the tabs reveals further dialog boxes. Note that only the first plot (plot 1) had “lfit” as an option. Also, I sorted by the independent variable (aomaj) in this graph. The final command is created automatically by filling in dialog boxes and clicking on things.



Here is another common graph in economics, the histogram.



This was produced by the following command, again using the dialog boxes.

```
. histogram crmaj, bin(10) normal yscale(off) xlabel(, alternate) title(Crime
>  Rate) subtitle(Virginia)
(bin=10, start=5740.0903, width=945.2186)
```

There are lots of other cool graphs available to you in Stata. You should take some time to experiment with some of them. The graphics dialog boxes make it easy to produce great graphs.

5 USING DO-FILES

A do-file is a list of Stata commands collected together in a text file. It is also known as a Stata program or batch file. The text file must have a ".do" filename extension. The do-file is executed either with the "do" command or by clicking on File, Do…, and browsing to the do-file. Except for very simple tasks, I recommend always using do-files. The reason is that complicated analyses can be difficult to reproduce. It is possible to do a complicated analysis interactively, executing each command in turn, before writing and executing the next command. You can wind up rummaging around in the data doing various regression models and data transformations, get a result, and then be unable to remember the sequence of events that led up to the final version of the model. Consequently, you may never be able to get the exact result again. It may also make it difficult for other people to reproduce your results. However, if you are working with a do-file, you make changes in the program, save it, execute it, and repeat. The final model can always be reproduced, as long as the data doesn't change, by simply running the final do-file again.

If you insist on running Stata interactively, you can "log" your output automatically to a "log file" by clicking on the "Begin Log" button (the one that looks like a scroll) on the button bar or by clicking on File, Log, Begin. You will be presented with a dialog box where you can enter the name of the log file and change the default directory where the file will be stored. When you write up your results you can insert parts of the log file to document your findings. If you want to replicate your results, you can edit the log file, remove the output, and turn the resulting file into a do file.

Because economics typically does not use laboratories where experimental results can be replicated by other researchers, we need to make our data and programs readily available to other economists, otherwise how can we be sure the results are legitimate? After all, we can edit the Stata output and data input and type in any values we want to. When the data and programs are readily available it is easy for other researchers to take the programs and do the analysis themselves and check the data against the original sources, to make sure we on the up and up. They can also do various tests and alternative analyses to find out if the results are "fragile," that is the results do not hold up under relatively small changes in the modeling technique or data. For these reasons, I recommend that any serious econometric analysis be done with do-files which can be made available to other researchers.

You can create a do-file with the built-in Stata do-file editor, any text editor, such as Notepad or WordPad, or with any word processing program like Word or WordPerfect. If you use a word processor, be sure to "save as" a text (.txt) file. However, the first time you save the do-file, Word, for example, will insist on adding a .txt filename extension. So, you wind up with a do-file called sample.do.txt. You have to rename the file by dropping the .txt extension. You can do it from inside Word by clicking on Open, browsing to the right directory, and then right-clicking on the do-file with the txt extension, click on Rename, and then rename the file.

It is probably easiest to use the built-in Stata do-file editor. It is a full service text editor. It has mouse control; cut, copy and paste; find and replace; and undo. To fire up the do-file editor, click on the button

that looks like an open envelope (with a pencil or something sticking out of it) on the button bar or click on Window, Do-file Editor.

Here is a simple do-file.

```
/**********************************************
*****         Sample Do-File                 *****
*****         This is a comment              *****
**********************************************/

/* some preliminaries */
#delimit;           * use semicolon to delimit commands;
set mem 10000;      * make sure you have enough memory;
set more off;       * turn off the more feature;
set matsize 150;    * make sure you have enough room for variables;

/* tell Stata where to store the output                  */
/* I usually use the do-file name, with a .log extension */
/* use a full pathname                                   */

log using "H:\Project\example.log", replace;

/* get the data */

use "H:\Project\crimeva.dta" , clear;

/************************************/
/*** summary statistics            ***/
/************************************/

summarize crbur aobur metpct ampct unrate;

regress crbur aobur metpct ampct unrate;

log close;
```

The beginning block of statements should probably be put at the start of any do-file. The #delimit; command sets the semicolon instead of the carriage return as the symbol signaling the end of the command. This allows us to write multi-line commands.

More conditions: in the usual interactive mode, when the results screen fills up, it stops scrolling and

--more—

appears at the bottom of the screen. To see the next screen, you must hit the spacebar. This can be annoying in do-files because it keeps stopping the program. Besides, you will be able to examine the output by opening the log file, so set more off.

For small projects you won't have to increase memory or matsize, but if you do run out of space, these are the commands you will need.

Don't forget to set the log file with the "log using filename, replace" command. The "replace" part is necessary if you edit the do-file and run it again. If you forget the replace option, Stata will not let you overwrite the existing log file. If you want Stata to put the log file in your "project" directory, you will need to use a complete pathname for the file, e.g., "H:\project\sample.log."

Next you have to get the data with the “use” command. Finally, you do the analysis.

Don’t forget to close the log file.

To see the log file, you can open it in the do-file editor, or any other editor or word processor. I use Word to review the log file, because I use Word to write the final paper..

The log file looks like this.

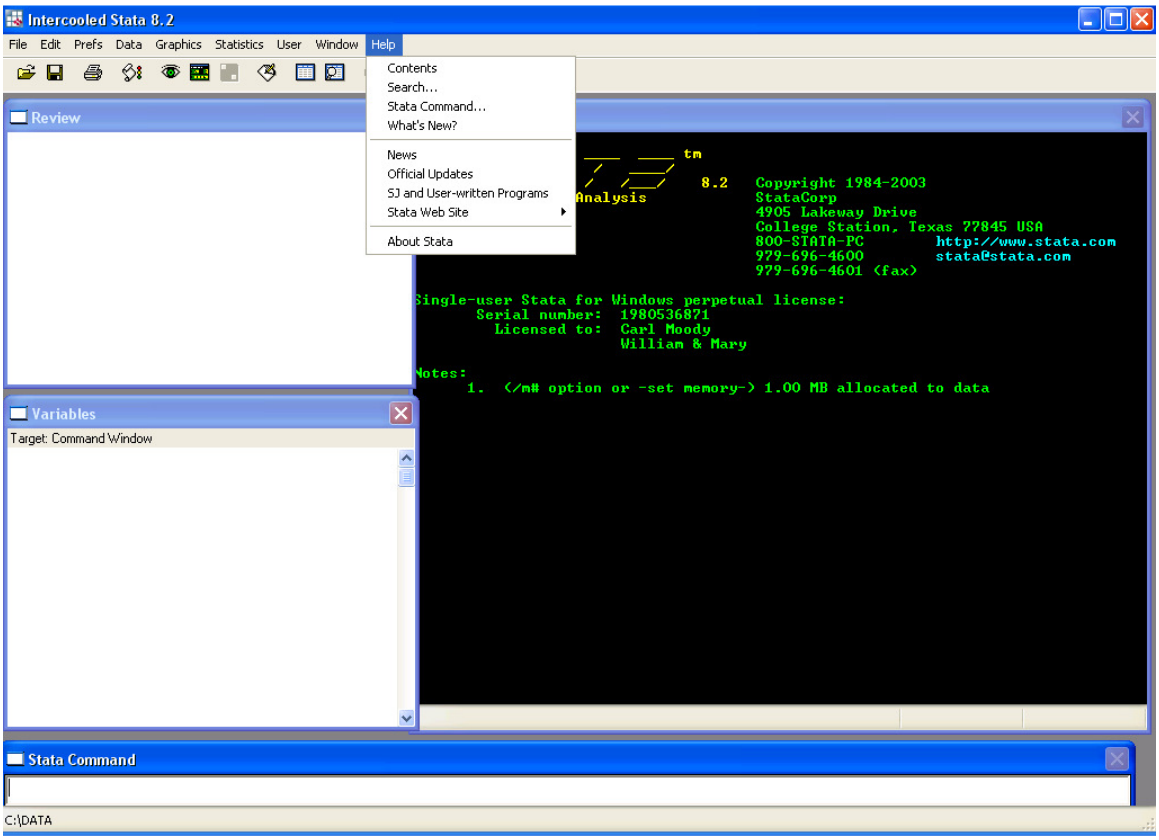
```
-------------------------------------------------------------------------------
      log:  H:\project\sample.log
  log type:  text
 opened on:   4 Jun 2004, 08:16:47

. /* get the data */
>. use "crimeva.dta" , clear;

. /************************************/
> /*** summary statistics           ***/
> /************************************/
>
> summarize crbur aobur metpct ampct unrate;

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       crbur |      40    7743.378   2166.439   3912.193   11925.49
       aobur |      22    16.99775   1.834964      12.47       19.6
      metpct |      28    71.46429   1.394291         70      74.25
       ampct |      28    18.76679   .1049885       18.5       18.9
      unrate |      30    4.786667   1.162261        2.4        7.7

. regress crbur aobur metpct ampct unrate;

      Source |       SS       df       MS              Number of obs =      21
-------------+------------------------------           F(  4,    16) =   28.74
       Model |  56035517.4     4  14008879.3           Prob > F      =  0.0000
    Residual |  7800038.17    16  487502.386           R-squared     =  0.8778
-------------+------------------------------           Adj R-squared =  0.8473
       Total |  63835555.6    20  3191777.78           Root MSE      =  698.21

------------------------------------------------------------------------------
       crbur |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       aobur |  -31.52691   132.9329    -0.24   0.816    -313.3321    250.2783
      metpct |  -986.7367   246.7202    -4.00   0.001     -1509.76   -463.7131
       ampct |   5689.995   7001.924     0.81   0.428     -9153.42    20533.41
      unrate |   71.66737   246.0449     0.29   0.775    -449.9246    593.2593
       _cons |  -27751.14   148088.1    -0.19   0.854    -341683.8    286181.6
------------------------------------------------------------------------------

. log close;
      log:  C:\Stata\sample.log
  log type:  text
 closed on:   4 Jun 2004, 08:16:48
-------------------------------------------------------------------------------
```

For complicated jobs, using lots of “banner” style comments (like the “summary statistics” comment above) makes the output more readable and reminds you, or tells someone else who might be reading the output, what you’re doing and why.

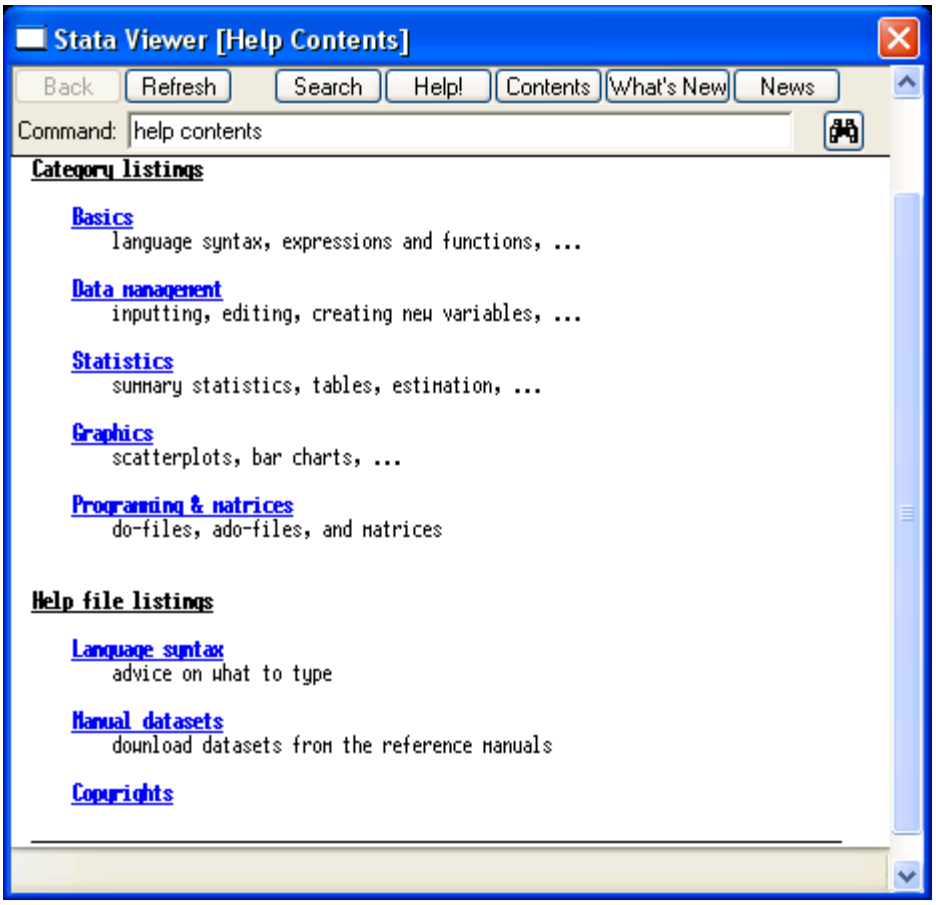
6 USING STATA HELP

This little manual can't give you all the details of each Stata command. It also can't list all the Stata commands that you might need. Finally, the official Stata manual consists of several volumes and is very expensive. So where does a student turn for help with Stata. Answer: the Stata Help menu.

Clicking on Help yields the following submenu.

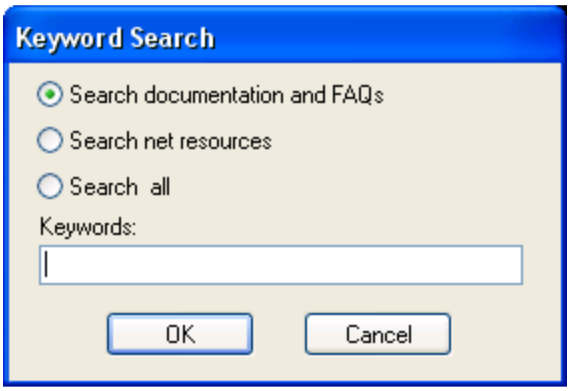


Clicking on Contents allows you to browse through the list of Stata commands, arranged according to function.

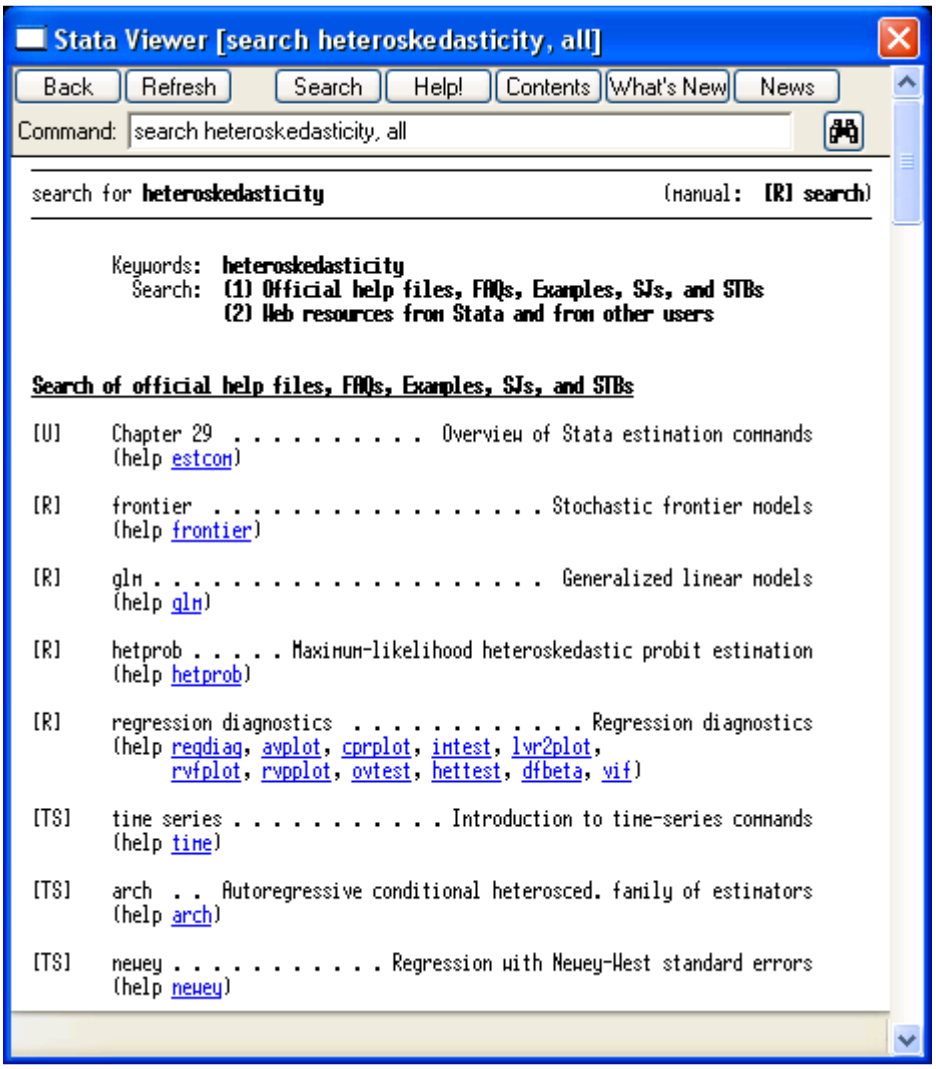


One could rummage around under various topics and find all kinds of neat commands this way. Clicking on any of the hyperlinks brings up the documentation for that command.

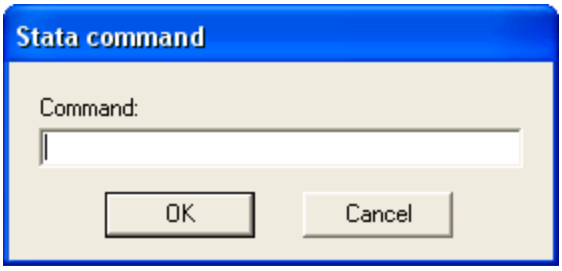
Now suppose you want information on how Stata handles a certain econometric topic. For example, suppose we want to know how Stata deals with heteroskedasticity. Clicking on Help, Search, yields the following dialog box.



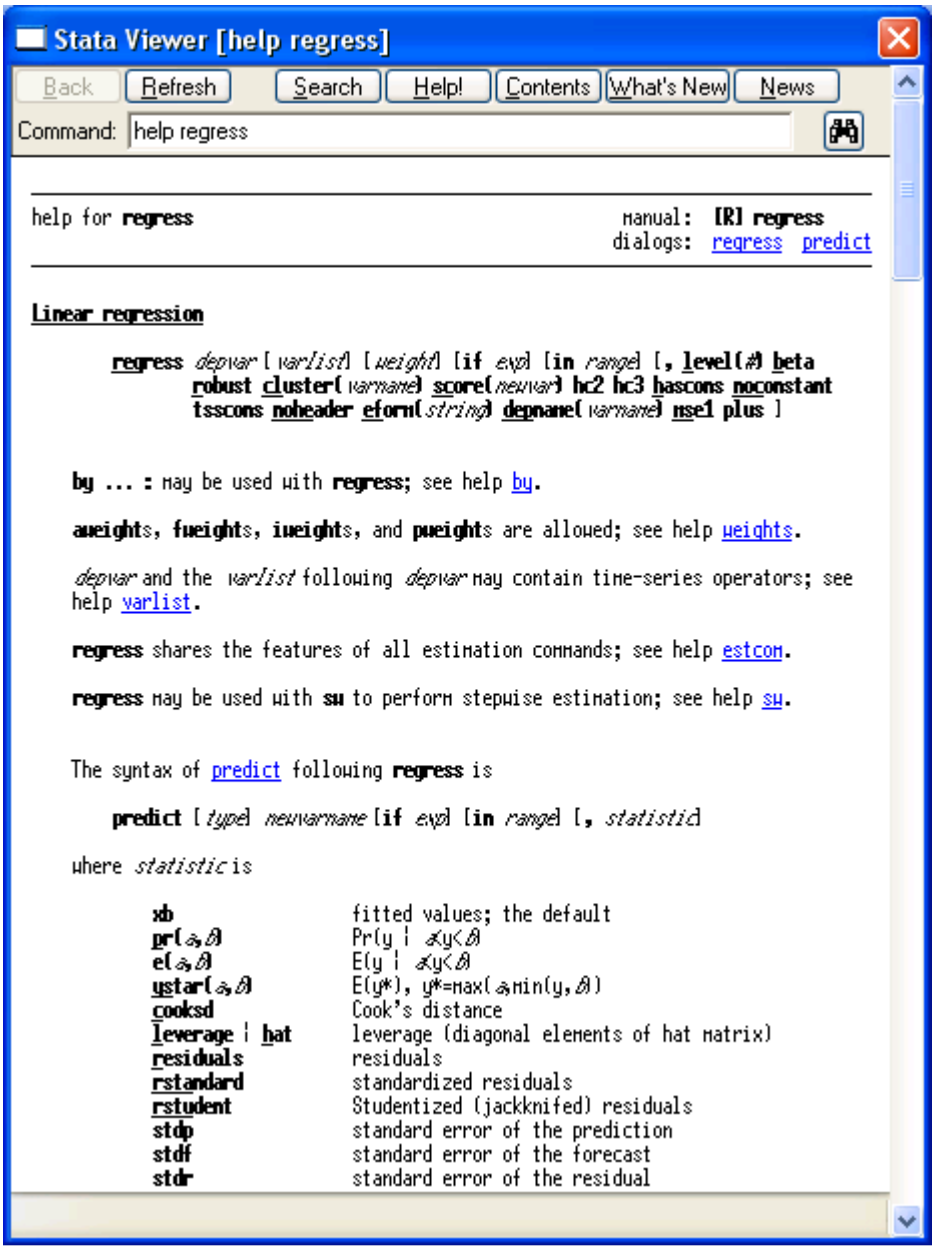
Clicking on the "Search all" radio button and entering heteroskedasticity yields



On the other hand, if you know that you need more information on a given command, you can click on Help, Stata command, yields



For example, entering “regress” to get the documentation on the regress commands yields.



This is the best way to find out more about the commands in this book. Most of the commands have options that are not documented here. There are also lots of examples and links to related commands. You should certainly spend a little time reviewing the regress help documentation because it will probably be the command you will use most over the course of the semester.

7 REGRESSION

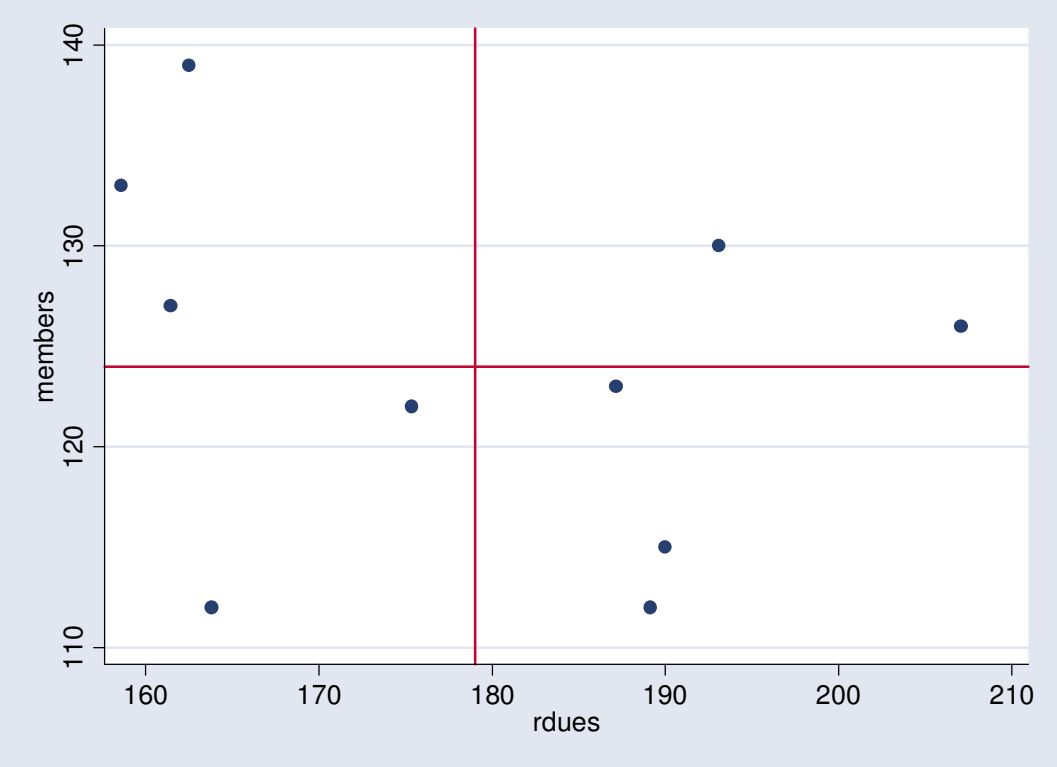
Let's start with a simple correlation. Use the data on pool memberships from the first chapter. Use summarize to compute the means of the variables.

```
. summarize

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        year |        10        82.5     3.02765         78         87
     members |        10       123.9    8.999383        112        139
        dues |        10       171.5    20.00694        135        195
         cpi |        10       .9717    .1711212       .652       1.16
       rdues |        10    178.8055    16.72355   158.5903   207.0552
-------------+--------------------------------------------------------
     logdues |        10    5.138088    .1219735   4.905275      5.273
      logmem |        10    4.817096    .0727516   4.718499   4.934474
         dum |        10          .4    .5163978          0          1
       trend |        10         4.5     3.02765          0          9
```

Plot the number of members on the real dues (rdues), sorting by rdues. This can be done by typing

```
twoway (scatter members rdues), yline(124) xline(179)
```

Or by clicking on Graphics, Two way graph, Scatter plot and filling in the blanks. Click on Y axis, Major Tick Options, and Additional Lines, to reveal the dialog box that allows you to put a line at the mean of members. Similarly, clicking on X axis, Major Tick Options, allows you to put a line at the mean of rdues. Either way, the result is a nice graph.



Note that six out of the ten dots are in the upper left and lower right quadrants. This means that high values of rdues are associated with low membership. We need a measure of this association. Define the deviation from the mean for X and Y as

$$x_i = X_i - \bar{X}$$
$$y_i = Y_i - \bar{Y}$$

Note that, in the upper right quadrant, $X_i > \bar{X}, Y_i > \bar{Y}$ so that $x_i > 0,\ y_i > 0$. Moving clockwise, in the lower right hand quadrant, $x_i > 0,\ y_i < 0$. In the lower left quadrant, $x_i < 0,\ y_i < 0$. Finally in the upper left quadrant, $x_i < 0,\ y_i > 0$.

So, for points in the lower left and upper right hand quadrants, if we multiply the deviations together and sum we get, $\sum x_i y_i > 0$. This quantity is called the sum of corrected cross-products, or the covariation. If we compute the sum of corrected cross-products for the points in the upper left and lower right quadrants we find that $\sum x_i y_i < 0$. This means that, if most of the points are in the upper left and lower right quadrants, then the variation, computed over the entire sample will be negative, indicating a negative association between X and Y. Similarly, if most of the points lie in the lower left and upper right quadrants, then the covariation will be positive, indicating a positive relationship between X and Y.

There are two problems with using covariation as our measure of association. The first is that it depends on the number of observations. The larger the number of observations, the larger the (positive or negative) covariation. We can fix this problem by dividing the covariation by the sample size, N. The resulting quantity is the average covariation, known as the covariance.

$$Cov(x,y) = \frac{\sum_{i=1}^{N} x_i y_i}{N} = \frac{\sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})}{N}$$

The second problem is that the covariance depends on the unit of measure of the two variables. You get different values of the covariance if the two variables are height in inches and weight in pounds versus height in miles and weight in milligrams. To solve this problem, we divide each observation by its standard deviation to eliminate the units of measure and then compute the covariance using the resulting standard scores. The standard score of X is $z_{X_i} = x_i / S_x$ where $S_x = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N}}$ is the standard deviation of X. The standard score of Y is $z_{Y_i} = \frac{y_i}{\sqrt{\sum_{i=1}^{N} x_i^2 \Big/ N}}$. The covariance of the standard scores is,

$$r_{XY} = \frac{\sum_{i=1}^{N} z_{X_i} z_{Y_i}}{N} = \frac{1}{N} \sum_{i=1}^{N} \frac{x_i}{S_X} \frac{y_i}{S_Y} = \frac{1}{N S_X S_Y} \sum_{i=1}^{N} x_i y_i$$

This is the familiar simple correlation coefficient.

The correlation coefficient for our pool sample can be computed in Stata by typing

```
correlate members rdues
```

```
(obs=10)

             |  members    rdues
-------------+------------------
     members |   1.0000
       rdues |  -0.2825   1.0000
```

The correlation coefficient is negative, as suspected.

With respect to regression analysis, the Stata command that corresponds to ordinary least squares (OLS) is the regress command. For example

```
Regress members rues
```

Produces a simple regression of members on rdues. That output was reported in the first chapter.

Linear regression

Sometimes we want to know more about an association between two variables than if it is positive or negative. Consider a policy analysis concerning the effect of putting more police on the street. Does it really reduce crime? If the answer is yes, the policy at least doesn't make things worse, but we also have to know by how much it can be expected to reduce crime, in order to justify the expense. If adding 100,000 police nationwide only reduces crime by a few crimes, it fails the cost-benefit test. So, we have to estimate the expected change in crime from a change in the number of police.

Suppose we hypothesize that there is a linear relationship between crime and police such that

$$Y = \alpha + \beta X$$

where Y is crime, X is police and $\beta < 0$. If this is true and we could estimate β, we could estimate the effect of adding 100,000 police as follows. Since β is the slope of the function,

$$\beta = \frac{\Delta Y}{\Delta X}$$

which means that

$$\begin{aligned} \Delta Y &= \beta \Delta X \\ &= \beta(100,000) \end{aligned}$$

This is the counterfactual (since we haven't actually added the police yet) that we need to estimate the benefit of the policy. The cost is determined elsewhere.

So, we need an estimate of the slope of a line relating Y to X, not just the sign. A line is determined by two points. We will use the slope, β, and the intercept α as the two points we need.

As statisticians, as opposed to mathematicians, we know that real world data do not line up on straight lines. We can summarize the relationship between Y and X as,

$$Y_i = \alpha + \beta X_i + U_i$$

where Ui is the error associated with observation i.

Why do we need an error term? Why don't observations line up on straight lines? There are two main reasons.

1. The model is incomplete. We have left out some of the other factors that affect crime. At this point let's assume that the factors that we left out are "irrelevant" in the sense that if we included them the basic form of the line would not change, that is, that the values of the estimated slope and intercept would not change. Then the error term is simply the sum of a whole bunch of variables that are each irrelevant, but taken together cause the error term to vary randomly.
2. The line might not be straight. That is, the line may be curved. We can get out of this either by transforming the line (taking logs, etc.), or simply assume that the line is approximately linear for our purposes.

So what are the values of α and β (the parameters) that create a "good fit," that is, which values of the parameters best describe the cloud of data in the figure above?

Let's define our estimate of α as $\hat{\alpha}$ and our estimate of β as $\hat{\beta}$. Then for any observation, our predicted value of Y is

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

If we are wrong, the error, or **residual,** is the difference between what we thought Y would be and what is actually was,

$$e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\alpha} + \hat{\beta} X_i) .$$

Clearly, we want to minimize the errors. We can't get zero errors because that would require driving the line through each point. Since they don't line up, we would get one line per point, not a summary statistic. We could minimize the sum of the errors, but that would mean that large positive errors could offset large negative errors and we really couldn't tell a good fit, with no error, from a bad fit, with huge positive and negative errors that offset each other.

To solve this last problem we could square the errors or take their absolute values. If we minimize the sum of squares of error, we are doing **least squares**. If we minimize the sum of absolute values we are doing least absolute value estimation, which is beyond our scope. The least squares procedure has been around since at least 1805 and it is the workhorse of applied statistics.

So, we want to minimize the sum of squares of error, or

$$\underset{\hat{\alpha},\hat{\beta}}{\text{Min}} \sum e_i^2 = \sum (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$$

We could take derivatives with respect to α and β, set them equal to zero and solve the resulting equations, but that would be too easy. Instead, let's do it without calculus.

Suppose we have a quadratic formula and we want to minimize it by choosing a value for b.

$$f(b) = c_2 b^2 + c_1 b + c_0$$

Recall from high school the formula for completing the square.

$$f(b) = c_2 \left(b + \frac{c_1}{2c_2} \right)^2 + \left(c_0 - \frac{c_1^2}{4c_2} \right)$$

Multiply it out if you are unconvinced. Since we are minimizing with respect to our choice of b and b only occurs in the squared term, the minimum occurs at

$$b = \frac{-c_1}{2c_2}$$

That is, the minimum is the ratio of the negative of the coefficient on the first power to two times the coefficient on the second power of the quadratic.

We want to minimize the sum of squares of error with respect to $\hat{\alpha}$ and $\hat{\beta}$, that is

$$\begin{aligned} \text{Min} \sum e^2 &= \sum (Y - \hat{\alpha} - \hat{\beta} X)^2 = \sum (Y^2 - 2\hat{\alpha} Y + \hat{\alpha}^2 + \hat{\beta}^2 X^2 - 2\hat{\alpha}\hat{\beta} X - 2\hat{\beta} XY) \\ &= \sum Y^2 - 2\hat{\alpha} \sum Y + N\hat{\alpha}^2 + \hat{\beta}^2 \sum X^2 - 2\hat{\alpha}\hat{\beta} \sum X - 2\hat{\beta} \sum XY \end{aligned}$$

Note that there are two quadratics in $\hat{\alpha}$ and $\hat{\beta}$. The one in $\hat{\alpha}$ is

$$\begin{aligned} f(\hat{\alpha}) &= N\hat{\alpha}^2 - 2\hat{\alpha} \sum Y - 2\hat{\alpha}\hat{\beta} \sum X + C \\ &= N\hat{\alpha}^2 - \hat{\alpha}(2 \sum Y - 2\hat{\beta} \sum X) \end{aligned}$$

where C is a constant consisting of terms not involving $\hat{\alpha}$. Minimizing $f(\hat{\alpha})$ requires setting $\hat{\alpha}$ equal to the negative of the coefficient of the first power divided by two times the coefficient of the second power, namely,

$$\hat{\alpha} = \frac{2\left(\sum Y - \hat{\beta} X\right)}{2N} = \frac{\sum Y}{N} - \frac{\hat{\beta} X}{N} = \bar{Y} - \hat{\beta}\bar{X}$$

The quadratic form in $\hat{\beta}$ is

$$f(\hat{\beta}) = \hat{\beta}^2 \sum X^2 - 2\hat{\beta}\sum XY + 2\hat{\alpha}\hat{\beta}\sum X = \hat{\beta}^2 \sum X^2 - \hat{\beta} 2(\sum XY + \hat{\alpha}\sum X) + C^*$$

where C* is another constant not involving $\hat{\beta}$.

Setting $\hat{\beta}$ equal to the ratio of the negative of the first power divided by two times the coefficient of the second power yields,

$$\hat{\beta} = \frac{2(\sum XY + \hat{\alpha}\sum X)}{2\sum X^2} = \frac{\sum XY + \hat{\alpha}\sum X}{\sum X^2}$$

Substituting the formula for $\hat{\alpha}$,

$$\hat{\beta} = \frac{\sum XY + (\bar{Y} - \hat{\beta}\bar{X})\sum X}{\sum X^2} = \frac{\sum XY + \bar{Y}\sum Y - \hat{\beta}\bar{X}\sum X}{\sum X^2}$$

Multiplying both sides by $\sum X^2$ yields,

$$\hat{\beta}\sum X^2 = \sum XY - \bar{Y}\sum X + \hat{\beta}\bar{X}\sum X$$
$$= \sum XY - \bar{Y}N\bar{X} + \hat{\beta}\bar{X}N\bar{X}$$

$$\hat{\beta}\sum X^2 - \hat{\beta}N\bar{X}^2 = \sum XY - N\bar{X}\bar{Y}$$
$$\hat{\beta}(\sum X^2 - N\bar{X}^2) = \sum XY - N\bar{X}\bar{Y}$$

Solving for $\hat{\beta}$

$$\hat{\beta} = \frac{\sum XY - N\bar{X}\bar{Y}}{\sum X^2 - N\bar{X}^2} = \frac{\sum xy}{\sum x^2}$$

Correlation and regression

Let's express the regression line in deviations.

$$Y = \alpha + \beta X$$

Since $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$ the regression line goes through the means:

$$\bar{Y} = \alpha + \beta\bar{X}$$

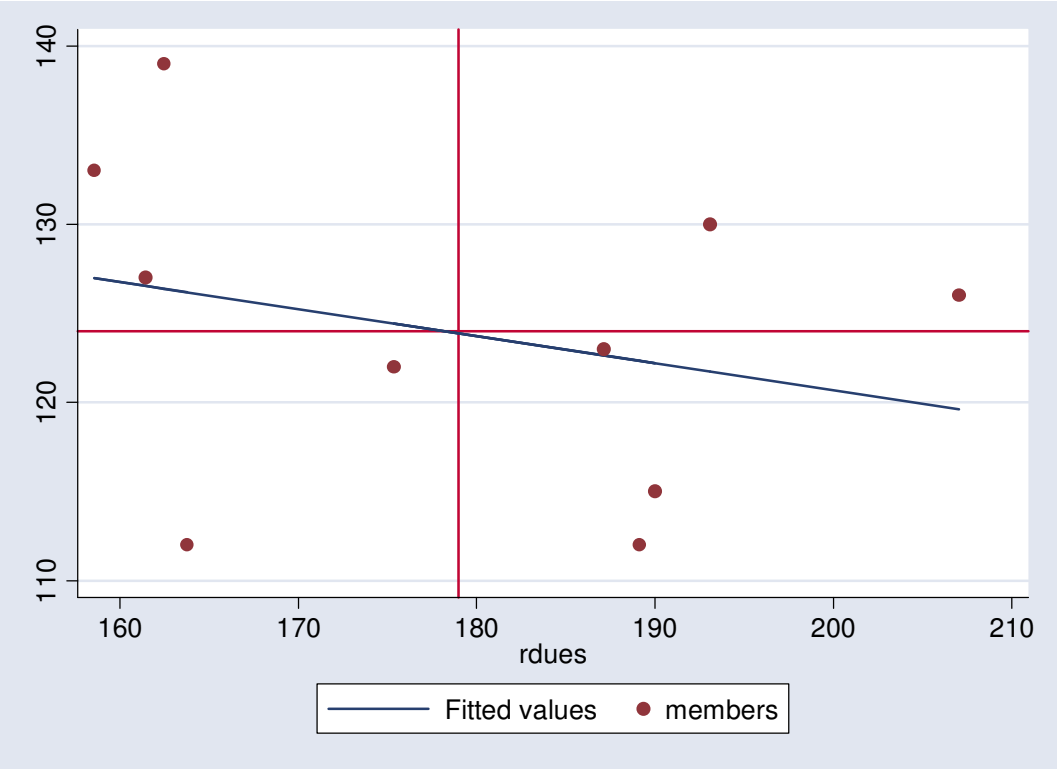
Subtract the mean of Y from both sides of the regression line,

$$Y - \bar{Y} = \alpha + \beta X - \alpha - \beta \bar{X}$$

Or,

$$y = \beta(X - \bar{X}) = \beta x$$

Note that, by subtracting off the means and expressing x and y in deviations, we have "translated the axes" so that the intercept is zero when the regression is expressed in deviations. The reason is, as we have seen above, the regression line goes through the means. Here is the graph of the pool membership data with the regression line superimposed.



Substituting $\hat{\beta}$ yields the predicted value of y ("Fitted values" in the graph):

$$\hat{y} = \hat{\beta} x$$

We know that the correlation coefficient between x and y is

$$r = \frac{\sum xy}{N S_x S_y}$$

Multiply both sides by S_y/S_x yields,

$$r \frac{S_x}{S_y} = \frac{\sum xy}{N S_x S_y} \frac{S_x}{S_y} = \frac{\sum xy}{N S_y^2}$$

But,

$$NS_y^2 = N\left(\sqrt{\frac{\sum y^2}{N}}\right)^2 = N\frac{\sum y^2}{N} = \sum y^2$$

Therefore,

$$r\frac{S_x}{S_y} = \frac{\sum xy}{\sum y^2} = \hat{\beta}$$

So, $\hat{\beta}$ is a linear transformation of the correlation coefficient; Since $S_x > 0$ and $S_y > 0$ r must have the same sign as $\hat{\beta}$. Since the regression line does everything that the correlation coefficient does, plus yield the rate of change of y to x, we will concentrate on regression from now on.

How well does the line fit the data?

We need a measure of the goodness of fit of the regression line. Start with the notion that each observation of y is equal to the predicted value plus the residual:

$$y = \hat{y} + e$$

Squaring both sides,

$$y^2 = y^2 + 2\hat{y}e + e^2$$

Summing,

$$\sum y^2 = \sum y^2 + 2\sum \hat{y}e + \sum e^2$$

The middle term on the right hand side is equal to zero.

$$\sum \hat{y}e = \sum \hat{\beta}xe = \hat{\beta}\sum xe$$

Which is zero if $\sum xe = 0$.

$$\sum xe = \sum x\left(y - \hat{\beta}x\right) = \sum xy - \hat{\beta}\sum x^2$$

$$= \sum xy - \frac{\sum xy}{\sum x^2}\sum x^2 = \sum xy - \sum xy = 0$$

This property of linear regression, namely that x is uncorrelated with the residual, turns out to be useful in several contexts. Nevertheless, the middle term is zero and thus,

$$\sum y^2 = \sum y^2 + \sum e^2$$

Or,

TSS=RSS+ESS

where TSS is the total sum of squares, $\sum y^2$, RSS is the regression (explained) sum of squares, $\sum \hat{y}^2$, and ESS is the error (unexplained, residual) sum of squares, $\sum e^2$.

Suppose we take the ratio of the explained sum of squares to the unexplained sum of squares as our measure of the goodness of fit.

$$\frac{RSS}{TSS} = \frac{\sum \hat{y}^2}{\sum y^2} = \frac{\sum \hat{\beta}^2 x^2}{\sum y^2} = \hat{\beta}^2 \frac{\sum x^2}{\sum y^2}$$

But, recall that

$$\hat{\beta} = r\frac{S_y}{S_x} = r\frac{\sqrt{\frac{\sum y^2}{N}}}{\sqrt{\frac{\sum x^2}{N}}}$$

Squaring both sides yields,

$$\hat{\beta}^2 = r^2 \frac{\frac{\sum y^2}{N}}{\frac{\sum x^2}{N}} = r^2 \frac{\sum y^2}{\sum x^2}$$

Therefore,

$$\frac{RSS}{TSS} = \hat{\beta}^2 \frac{\sum x^2}{\sum y^2} = r^2$$

The ratio RSS/TSS is the proportion of the variance of Y explained by the regression and is usually referred to as the coefficient of determination. It is usually denoted R^2. We know why now, because the coefficient of determination is, in simple regressions, equal to the correlation coefficient, squared. Unfortunately, this neat relationship doesn't hold in multiple regressions (because there are several correlation coefficients associating the dependent variable with each of the independent variables). However, the ratio of the regression sum of squares to the total sum of squares is always denoted R^2.

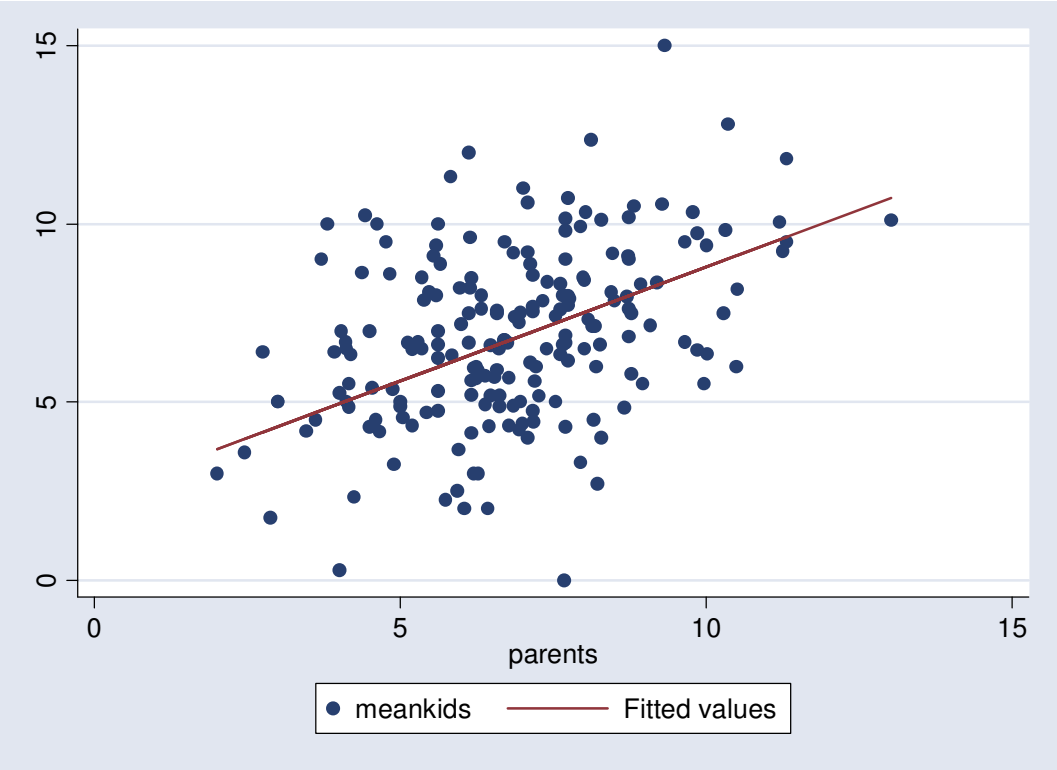
Why is it called regression?

In 1886 Sir Francis Galton, cousin of Darwin and discoverer of fingerprints among other things, published a paper entitled, "Regression toward mediocrity in hereditary stature." Galton's hypothesis was that tall parents should have tall children, but not as tall as they are, and short parents should produce short offspring, but not as short as they are. Consequently, we should eventually all be the same, mediocre, height. He knew that, to prove his thesis, he needed more than a positive association. He needed the slope of a line relating the two heights to be less than one.

He plotted the deviations from the mean heights of adult children on the deviations from the mean heights of their parents (multiplying the mother's height by 1.08). Taking the resulting scatter diagram, he eyeballed a straight line through the data and noted that the slope of the line was positive, but less than one. He claimed that, "…we can define the law of regression very briefly; It is that the height-deviate of the offspring is, on average, two-thirds of the height-deviate of its mid-parentage." (Galton, 1886,

Anthropological Miscellanea, p. 352.) Mid-parentage is the average height of the two parents, minus the overall mean height, 68.5 inches. This is the regression toward mediocrity, now commonly called regression to the mean. Galton called the line through the scatter of data a "regression line." The name has stuck.

The original data collected by Galton is available on the web. (Actually it has a few missing values and some data entered as "short" "medium," etc. have been dropped.) The scatter diagram is shown below, with the regression line (estimated by OLS, not eyeballed) superimposed.



However, another question arises, namely, when did a line estimated by least squares become a regression line? We know that least squares as a descriptive device was invented by Legendre in 1805 to describe astronomical data. Galton was apparently unaware of this technique, although it was well known among mathematicians. Galton's younger colleague Karl Pearson, later refined Galton's graphical methods into the now universally used correlation coefficient. (Frequently referred to in textbooks as the Pearsonian product-moment correlation coefficient.) However, least squares was still not being used. In 1897 George Udny Yule, another, even younger colleague of Galton's, published two papers in which he estimated what he called a "line of regression" using ordinary least squares. It has been called the "regression line" ever since. In the second paper he actually estimates a multiple regression model of poverty as a function of welfare payments including control variables. The first econometric policy analysis.

The regression fallacy

Although Galton was in his time and is even now considered a major scientific figure, he was completely wrong about the law of heredity he claimed to have discovered. According to Galton, the heights of fathers of exceptionally tall children tend to be tall, but not as tall as their offspring, implying that there is a natural

tendency to mediocrity. This theorem was supposedly proved by the fact that the regression line has a slope less than one.

Using Galton's data from the scatter diagram above (the actual data is in Galton.dta). We can see that the regression line does, in fact, have a slope less than one.

```
. regress meankids parents [fweight=number]
```
[1]

```
      Source |       SS       df       MS              Number of obs =     899
-------------+------------------------------           F(  1,   897) =  376.21
       Model |  1217.79278     1  1217.79278           Prob > F      =  0.0000
    Residual |  2903.59203   897  3.23700338           R-squared     =  0.2955
-------------+------------------------------           Adj R-squared =  0.2947
       Total |  4121.38481   898  4.58951538           Root MSE      =  1.7992

------------------------------------------------------------------------------
    meankids |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     parents |    .640958   .0330457    19.40   0.000     .5761022    .7058138
       _cons |   2.386932   .2333127    10.23   0.000      1.92903    2.844835
------------------------------------------------------------------------------
```

The slope is ..64, almost exactly that calculated by Galton, which confirms his hypothesis. Nevertheless, the fact that the law is bogus is easy to see. Mathematically, the explanation is simple. We know that

$$\hat{\beta} = r\frac{S_y}{S_x}$$

If $S_y \cong S_x$ and r > 0, then

$$0 < \hat{\beta} \cong r < 1$$

This is the regression fallacy. The regression line slope estimate is always less than one for any two variables with approximately equal variances.

Let's see if this is true for the Galton data. The means and standard deviations are produced by the summarize command. The correlation coefficient is produced by the correlate command.

```
. summarize meankids parents

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    meankids |       204    6.647265    2.678237          0         15
     parents |       197    6.826122    1.916343          2      13.03

. correlate meankids parents
(obs=197)

             | meankids  parents
-------------+------------------
    meankids |   1.0000
     parents |   0.4014   1.0000
```

So, the standard deviations are very close and the correlation coefficient is positive. We therefore expect that the slope of the regression line will be positive. Galton was simply demonstrating the mathematical theorem with these data.

[1] We weight by the number of kids because some families have lots of kids (up to 15) and therefore represent 15 observations, not one.

To take another example, if you do extremely well on the first exam in a course, you can expect to do well on the second exam, but probably not quite as well. Similarly, if you do poorly, you can expect to do better. The underlying theory is that ability is distributed randomly with a constant mean. Suppose ability is distributed uniformly with mean 50. However, on each test, you experience random error with is distributed normally with mean zero and standard deviation 10. Thus,

$$x_1 = a + e_1$$
$$x_2 = a + e_2$$
$$E(a) = 50,\ E(e_1) = 0,\ E(e_2) = 0,\ \text{cov}(a, e_1) = \text{cov}(a, e_2) = \text{cov}(e_1, e_2) = 0$$

where x_1 is the grade on the first test, x_2 is the grade on the second test, a is ability, and e_1 and e_2 are the random errors.

$$E(x_1) = E(a + e_1) = E(a) + E(e_1) = E(a) = 50$$
$$E(x_2) = E(a + e_2) = E(a) + E(e_2) = E(a) = 50$$

So, if you scored below 50 on the first exam, you should score closer to 50 on the second. If you scored higher than 50, you should regress toward mediocrity. (Don't be alarmed, these grades are not scaled, 50 is a B.)

We can use Stata to prove this result with a Monte Carlo demonstration.

```
* choose a large number of observations
. set obs 10000
obs was 0, now 10000

* create a measure of underlying ability
. gen a=100*uniform()
* create a random error term for the first test
. gen x1=a+10*invnorm(uniform())
* now a random error for the second test
. gen x2=a+10*invnorm(uniform())

. summarize

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           a |     10000    50.01914    28.88209   .0044471   99.98645
          x1 |     10000    50.00412     30.5206  -29.68982   127.2455
          x2 |     10000    49.94758    30.49167  -25.24831   131.0626

* note that the means are all 50 and that the standard deviations are about the
same

. correlate x1 x2
(obs=10000)

             |       x1       x2
-------------+------------------
          x1 |   1.0000
          x2 |   0.8912   1.0000


. regress x2 x1

      Source |       SS       df       MS              Number of obs =   10000
-------------+------------------------------           F(  1,  9998) =38583.43
```

```
       Model |  7383282.55     1  7383282.55           Prob > F      =  0.0000
    Residual |  1913206.37  9998  191.358909           R-squared     =  0.7942
-------------+------------------------------           Adj R-squared =  0.7942
       Total |  9296488.92  9999  929.741866           Root MSE      =  13.833

------------------------------------------------------------------------------
          x2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |    .890335   .0045327   196.43   0.000     .8814501    .8992199
       _cons |   5.427161   .2655312    20.44   0.000     4.906666    5.947655
------------------------------------------------------------------------------
```

```
* the regression slope is equal to the correlation coefficient, as expected
```

Does this actually work in practice? I took the grades from the first two tests in a statistics course I taught years ago and created a Stata data set called tests.dta. You can download the data set and look at the scores. Here are the results.

```
. summarize

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       test1 |        41    74.87805     17.0839         30        100
       test2 |        41    77.78049    13.54163         47        100

. correlate test2 test1
(obs=41)

             |    test2    test1
-------------+------------------
       test2 |   1.0000
       test1 |   0.3851   1.0000

. regress test2 test1

      Source |       SS       df       MS              Number of obs =      41
-------------+------------------------------           F(  1,    39) =    6.79
       Model |  1087.97122     1  1087.97122           Prob > F      =  0.0129
    Residual |  6247.05317    39  160.180851           R-squared     =  0.1483
-------------+------------------------------           Adj R-squared =  0.1265
       Total |  7335.02439    40   183.37561           Root MSE      =  12.656

------------------------------------------------------------------------------
       test2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       test1 |   .3052753   .1171354     2.61   0.013     .0683465     .542204
       _cons |   54.92207    8.99083     6.11   0.000      36.7364    73.10774
------------------------------------------------------------------------------
```

Horace Secrist

Secrist was Professor of Statistics at Northwestern University in 1933 when, after ten years of effort collecting and analyzing data on profit rates of businesses, he published his magnum opus, “The Triumph of Mediocrity in Business.” In this 486 page tome he analyzed mountains of data and showed that, in every industry, businesses with high profits tended, over time, to see their profits erode, while businesses with lower than average profits showed increases. Secrist thought he had found a natural law of competition (and maybe a fundamental cause of the depression, it was 1933 after all). The book initially received favorable reviews in places like the Royal Statistical Society, the *Journal of Political Economy*, and the *Annals of the American Academy of Political and Social Science*. Secrist was riding high.

It all came crashing down when Harold Hotelling, a well known economist and mathematical statistician, published a review in the *Journal of the American Statistical Association*. In the review Hotelling pointed out that if the data were arranged according to the values taken at the end of the period, instead of the beginning, the conclusions would be reversed. "The seeming convergence is a statistical fallacy, resulting from the method of grouping. These diagrams really prove nothing more than that the ratios in question have a tendency to wander about." (Hotelling, JASA, 1933.) According to Hotelling the real test would be a reduction in the variance of the series over time (not true of Secrist's series). Secrist replied with a letter to the editor of the JASA, but Hotelling was having none of it. In his reply to Secrist, Hotelling wrote,

> This theorem is proved by simple mathematics. It is illustrated by genetic, astronomical, physical, sociological, and other phenomena. To "prove" such a mathematical result by a costly and prolonged numerical study of many kinds of business profit and expense ratios is analogous to proving the multiplication table by arranging elephants in rows and columns, and then doing the same for numerous other kinds of animals. The performance, though perhaps entertaining, and having a certain pedagogical value, is not an important contribution to either zoology or to mathematics. (Hotelling, JASA, 1934.)

Even otherwise competent economists commit this error. See Milton Friedman, 1992, "Do old fallacies ever die?" *Journal of Economic Literature* 30:2129-2132. By the way, the "real test' of smaller variances over time might need to be taken with a grain of salt if measurement errors are getting smaller over time. Also, in the same article, Friedman, who was a student of Hotelling's when Hotelling was engaged in his exchange with Secrist, tells how he was inspired to avoid the regression fallacy in his own work. The result was the permanent income hypothesis, which can be derived from the test examples above by simply replacing test scores with consumption, ability with permanent income, and the random errors with transitory income.

Some tools of the trade

We will be using these concepts routinely throughout the semester.

Summation and deviations

Let the lowercase variable $x_i = X_i - \bar{X}$ where X is the raw data.

(1) $$\sum_{i=1}^{N} x_i = \sum (X_i - \bar{X}) = \sum X_i - N\bar{X}$$
$$= \sum X_i - N\left(\frac{\sum X_i}{N}\right) = \sum X_i - \sum X_i = 0$$

From now on, I will suppress the subscripts (which are always i=1,…, N where N is the sample size).

(2) $$\sum x^2 = \sum (X - \bar{X})^2 = \sum \left(X^2 - 2X\bar{X} + \bar{X}^2 \right)$$
$$= \sum X^2 - 2\bar{X} \sum X + N\bar{X}^2$$

Note that $\sum X = N\left(\frac{\sum X}{N}\right) = N\bar{X}$ so that

$$\sum x^2 = \sum X^2 - 2\bar{X} \sum X + N\bar{X}^2 = \sum X^2 - 2N\bar{X}^2 + N\bar{X}^2 = \sum X^2 - 2N\bar{X}$$

$$\sum xy = \sum (X - \bar{X})(Y - \bar{Y}) = \sum \left(XY - X\bar{Y} - Y\bar{X} + \bar{X}\bar{Y} \right)$$

(3)
$$= \sum XY - \bar{Y} \sum X - \bar{X} \sum Y + N\bar{X}\bar{Y}$$
$$= \sum XY - \bar{Y}N\bar{X} - \bar{X}N\bar{Y} + N\bar{X}\bar{Y} = \sum XY - N\bar{X}\bar{Y}$$

(4)
$$\sum xY = \sum (X - \bar{X})Y = \sum \left(XY - \bar{X}Y \right)$$
$$= \sum XY - \bar{X} \sum Y = \sum XY - N\bar{X}\bar{Y} = \sum xy$$

(5)
$$\sum Xy = \sum X(Y - \bar{Y}) = \sum XY - \bar{Y} \sum X$$
$$= \sum XY - N\bar{X}\bar{Y} = \sum xy = \sum xY$$

Expected value

The expected value of an event is the most probable outcome, what might be expected. It is the average result over repeated trials. It is also a weighted average of several events where the weight attached to each outcome is the probability of the event.

$$E(X) = p_1 X_1 + p_2 X_2 + ... + p_N X_N = \sum_{i=1}^{N} p_i X_i = \mu_x, \sum_{i=1}^{N} p_i = 1$$

For example, if you play roulette in Monte Carlo, the slots are numbered 0-36. However, the payoff if your number comes up is 35 to 1. So, what is the expected value of a one dollar bet? There are only two relevant outcomes, namely your number comes up and you win $35 or it doesn't and you lose your dollar.

$$E(X) = p_1 X_1 + p_2 X_2 = \left(\frac{36}{37} \right)(-1) + \left(\frac{1}{37} \right)(35)$$
$$= .973(-1) + .027(35) = -.973 + .946 = -.027$$

So, you can expect to lose about 2.7 cents every time you play. Of course, when you play you never lose 2.7 cents. You either win $35 or lose a dollar. However, over the long haul, you will eventually lose. If you play 1000 times, you will win some and lose some, but on average, you will lose $27. At the same time, the casino can expect to make $27 for every 1000 bets. This is known as the house edge and is how the casinos make a profit. If the expected value of the game is nonzero, as this one is, the game is said not to be fair. A fair game is one for which the expected value is zero. For example, we can make roulette a fair game by increasing the payoff to $36

$$E(X) = p_1 X_1 + p_2 X_2 = \left(\frac{36}{37} \right)(-1) + \left(\frac{1}{37} \right)(36) = 0$$

Let's assume that all the outcomes are equally likely, so that $p_i = 1/N$ then

$$E(X) = \frac{1}{N} X_1 + \frac{1}{N} X_2 + ... + \frac{1}{N} X_N = \frac{\sum_{i=1}^{N} X}{N} = \mu_x$$

Note that the expected value of X is the true (population) mean, μ_x not the sample mean, $\bar{X}$. The expected value of X is also called the first moment of X.

Operationally, to find an expected value, just take the mean.

Expected values, means and variance

(1) If b is a constant, then $E(b) = b$.

(2) $$E(bX) = \frac{bX_1 + bX_2 + ... + bX_N}{N} = b\sum_{i=1}^{N} X_i / N = bE(X)$$

(3) $$E(a + bX) = a + bE(X)$$

(4) $$E(X + Y) = E(X) + E(Y) = \mu_x + \mu_y$$

(5) $$E(aX + bY) = aE(X) + bE(Y) = a\mu_x + b\mu_y$$

(6) $$E\left[(bX)^2\right] = b^2 E(X)$$

(7) The variance of X is

$$Var(X) = E(X - \mu_x)^2 = \frac{\sum(X - \mu_x)^2}{N}$$

also written as $Var(X) = E(X - E(X))^2$ The variance of X is also called the second moment of X.

(8) $$Cov(X,Y) = E[(X - \mu_x)(Y - \mu_y)]^2$$

(9) $$Var(X + Y) = E[(X + Y) - (\mu_x + \mu_y)]^2 = E[(X - \mu_x) + (Y - \mu_y)]^2$$

$$= E[(X - \mu_x)^2 + (Y - \mu_y)^2 + 2(X - \mu_x)(Y - \mu_y)]$$
$$= Var(X) + Var(Y) + 2Cov(X,Y)$$

(10) If X and Y are independent, then E(XY)=E(X)E(Y).

(11) If X and Y are independent,

$$Cov(X,Y) = E(X - \mu_x)(Y - \mu_y) = E[XY - Y\mu_x - X\mu_y + \mu_x\mu_y]$$
$$= E(XY) - \mu_y\mu_x - \mu_x\mu_y + \mu_x\mu_y$$
$$= E(XY) - E(X)E(Y) = 0$$

8 THEORY OF LEAST SQUARES

Ordinary least squares is almost certainly the most widely used statistical method ever created, aside maybe from the mean. The Gauss-Markov theorem is the theoretical basis for its ubiquity.

Method of least squares

Least squares was first developed to help explain astronomical measurements. In the late 17'th Century, scientists were looking at the stars through telescopes and trying to figure out where they were. They knew that the stars didn't move, but when they tried to measure where they were, different scientists got different locations. So which observations are right? According to Adrien Marie Legendre in 1805, "Of all the principles that can be proposed for this purpose [analyzing large amounts of inexact astronomical data], I think there is none more general more exact, or easier to apply, than…making the sum of the squares of the errors a minimum. By this method a kind of equilibrium is established among the errors which, since it prevents the extremes from dominating, is appropriate for revealing the state of the system which most nearly approaches the truth."

Gauss published the method of least squares in his book on planetary orbits four years later in 1809. Gauss claimed that he had been using the technique since 1795, thereby greatly annoying Legendre. It remains unclear who thought of the method first. Certainly Legendre beat Gauss to print. Gauss did show that least squares is preferred if the errors are distributed normally ("Gaussian") using a Bayesian argument.

Gauss published another derivation in 1823, which has since become the standard. This is the "Gauss" part of the Gauss-Markov theorem. Markov published his version of the theorem in 1912, using different methods, containing nothing that Gauss hadn't done a hundred years earlier. Neyman, in the 1920's, inexplicably unaware of the 1823 Gauss paper, calls Markov's results, "Markov's Theorem." Somehow, even though he was almost 100 years late, Markov's name is still on the theorem.

We will investigate the Gauss-Markov theorem below. But first, we need to know some of the properties of estimators, so we can evaluate the least squares estimator.

Properties of estimators

An estimator is a formula for estimating the value of a parameter. For example the formula for computing the mean $\overline{X} = \Sigma X / N$ is an estimator. The resulting value is the estimate. The formula for estimating the slope of a regression line is $\hat{\beta} = \Sigma xy / \Sigma x^2$ and the resulting value is the estimate. So is $\hat{\beta}$ a good estimator?

The answer depends on the distribution of the estimator. Since we get a different value for $\hat{\beta}$ each time we take a different sample, $\hat{\beta}$ is a random variable. The distribution of $\hat{\beta}$ across different samples is called the sampling distribution of $\hat{\beta}$. The properties of the sampling distribution of $\hat{\beta}$ will determine if $\hat{\beta}$ is a good estimator, or not.

There are two types of such properties. The so called "small sample" properties are true for all sample sizes, no matter how small. The large sample properties are true only as the sample size approaches infinity.

Small sample properties

There are three important small sample properties: bias, efficiency, and mean square error.

Bias

Bias is the difference between the expected value of the estimator and the true value of the parameter.

Bias= $E\left(\hat{\beta}\right) - \beta$

Therefore, if $E\left(\hat{\beta}\right) = \beta$ the estimator is said to be unbiased. Obviously, unbiased estimators are preferred.

Efficiency

If for a given sample size, the variance of $\hat{\beta}$ is smaller than the variance of any other unbiased estimator, than $\hat{\beta}$ is said to be efficient. Efficiency is important because the more efficient the estimator, the lower its variance and the smaller the variance the more precise the estimate. Efficient estimators allow us to make more powerful statistical statements concerning the true value of the parameter being estimated. You can visualize this by imagining a distribution with zero variance. How confident can we be in our estimate of the parameter in this case?

The drawback to this definition of efficiency is that we are forced to compare unbiased estimators. It is frequently the case, as we shall see throughout the semester that we must compare estimators that may be biased, at least in small samples. For this reason, we will use the more general concept of **relative efficiency**.

The estimator $\hat{\beta}$ is efficient relative to the estimator $\tilde{\beta}$ if $Var\left(\hat{\beta}\right) < Var\left(\tilde{\beta}\right)$.

From now on, when we say an estimator is efficient, we mean efficient relative to some other estimator.

Mean square error

We will see that we often have to compare the properties of two estimators, one of which is unbiased but inefficient while the other is biased but efficient. The concept of mean square error is useful here. Mean square error is defined as the variance of the estimator, plus its squared bias, that is,

$$mse\left(\hat{\beta}\right) = Var\left(\hat{\beta}\right) + \left[Bias\left(\hat{\beta}\right)\right]^2$$

Large sample properties

Large sample properties are also known as asymptotic properties because they tell us what happens to the estimator as the sample size goes to infinity. Ideally, we want the estimator to approach the truth, with maximum precision, as $N \to \infty$.

Define the **probability limit** as the limit of a probability as the sample size goes to infinity, that is

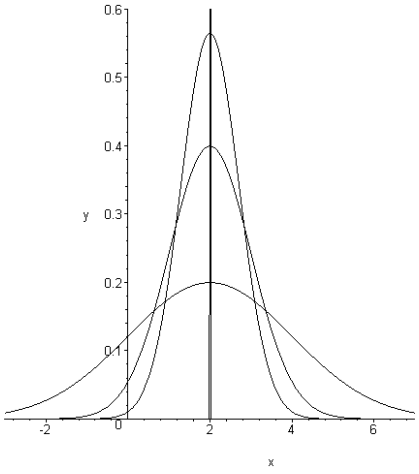
$$\text{plim} = \lim_{N\to\infty} \Pr$$

Consistency

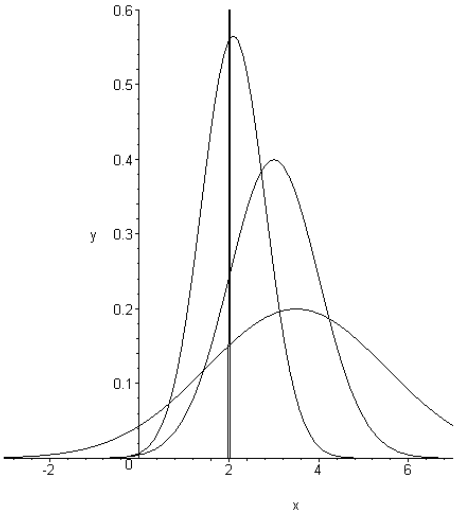
A consistent estimator is one for which

$$\text{plim}\left(\left|\hat{\beta} - \beta\right| < \varepsilon\right) = \lim_{N\to\infty} \Pr\left(\left|\hat{\beta} - \beta\right| < \varepsilon\right) = 1$$

That is, as the sample size goes to infinity, the probability that the estimator will be infinitesimally different from the truth is one, a certainty. This requires that the estimator be centered on the truth, with no variance. A distribution with no variance is said to be degenerate, because it is no longer really a distribution, it is a number.



A biased estimator can be consistent if the bias gets smaller as the sample size increases.



Since we require that both the bias and the variance go to zero for a consistent estimator, an alternative definition is mean square error consistency. That is, an estimator is consistent if its mean square error goes to zero as the sample size goes to infinity.

All of the small sample properties have large sample analogs. An estimator with bias that goes to zero as $N \to \infty$ is asymptotically unbiased. An estimator can be asymptotically efficient relative to another estimator. Finally, an estimator can be mean square error consistent.

Note that a consistent estimator is, in the limit, a number. As such it can be used in further derivations. For example, the ratio of two consistent estimators is consistent. The property of consistency is said to "carry over" to the ratio. This is not true of unbiasedness. All we know about an unbiased estimate is that its mean is equal to the truth. We do not know if our estimate is equal to the truth. If we compute the ratio of two unbiased estimates we don't know anything about the properties of the ratio.

Gauss-Markov theorem

Consider the following simple linear model.

$$Y_i = \alpha + \beta X_i + u_i$$

where α, and β, are true, constant, but unknown parameters and u_i is a random variable.
The Gauss-Markov assumptions are (1) the observations on X are a set of fixed numbers, (2) Y_t is distributed with a mean of $\alpha + \beta X_i$ for each observation i, (3) the variance of Y_i (σ^2) is constant, and (4) the observations on Y are independent from each other.

The basic idea behind these assumptions is that the model describes a laboratory experiment. The researcher sets the value of X without error and then observes the value of Y, with some error. The observations of Y come from a simple random sample, which means that the observation on Y_1 is independent of Y_2, which is independent of Y_3, etc. The mean and the variance of Y does not change between observations.

The Gauss-Markov assumptions, which are currently expressed in terms of the distribution of Y, can be rephrased in terms of the distribution of the error term, u. We can do this because, under the Gauss-Markov

assumptions, everything in the model is constant except Y and u. Therefore, the variance of u is equal to the variance of Y.

Also,
$u_i = Y_i - \alpha - \beta X_i$
so that the mean is
$E(u_i) = E(Y_i) - \alpha - \beta X_i = 0$
because $E(X_i) = X_i$ which means that $E(Y_i) = \alpha + \beta X_i$ by assumption (1).

Therefore, if the Gauss-Markov assumptions are true, the error term u is distributed with mean zero and variance σ^2 .In fact, the GM assumptions can be succinctly written as

$$Y_i = \alpha + \beta X_i + u_i, u_t \sim iid(0, \sigma^2)$$

where the tilde means "is distributed as," and iid means "independently and identically distributed." The independent assumption refers to the simple random sample while the identical assumption means that the mean and variance (and any other parameters) are constant.

Assumption (1) is very useful for proving the theorem, but it is not strictly necessary. The crucial requirement is that X must be independent of u, that is the covariance between X and u must be zero, i.e., E(X,u)=0. Assumption (2) is crucial. Without it, least squares doesn't work. Assumption (3) can be relaxed, as we shall see in Chapter 9 below. Assumption (4) can also be relaxed, as we shall see in Chapter 12 below.

OK, so why is least squares so good? Here is the Gauss-Markov theorem: if the GM assumptions hold, then the ordinary least squares estimates of the parameters are unbiased, consistent, and efficient relative to any other linear unbiased estimator. The proof of the unbiased and consistent parts are easy.

We know that the mean of u is zero. So,

$$E(Y_i) = \alpha + \beta X_i$$

The "empirical analog" of this (using the sample observations instead of the theoretical model) is $\overline{Y} = \alpha + \beta \overline{X}$ where $\overline{Y}$ is the sample mean of Y and $\overline{X}$ is the sample mean of X. Therefore, we can subtract the means to get,

$$Y_i - \overline{Y} = \alpha + \beta X_i - \alpha - \beta \overline{X}$$

Let $y_i = Y_i - \overline{Y}, x_i = X_i - \overline{X}$
where the lowercase letters indicate deviations from the mean. Then,
$y_i = \beta x_i + u_i$
(recall that the mean of u is zero so that $u - \overline{u} = u$). What we have done is translate the axes so that the regression line goes through the origin.

Define e as the error. We want to minimize the sum of squares of error.

$$e_i = y_i - \beta x_i$$
$$e_i^2 = (y_i - \beta x_i)^2$$
$$\sum_i^N e_i^2 = \sum_i^N (y_i - \beta x_i)^2$$

We want to minimize the sum of squares of error by choosing β, so take the derivative with respect to β and set it equal to zero.

$$\frac{d}{d\beta}\sum e_i^2 = \frac{d}{d\beta}\sum\left(y_i - \beta x_i\right)^2 = \frac{d}{d\beta}\sum\left(y_i^2 - 2\beta x_i y_i + \beta^2 x_i^2\right)$$

$$= \frac{d}{d\beta}\left(y_1^2 - 2\beta x_1 y_1 + \beta^2 x_1^2\right) + \frac{d}{d\beta}\left(y_2 - 2\beta x_2 y_2 + \beta^2 x_2^2\right) + \ldots + \frac{d}{d\beta}\left(y_N - 2\beta x_N y_N + \beta^2 x_N^2\right)$$

$$= -2x_1 y_1 + 2\beta x_1^2 - 2x_2 y_2 + 2\beta x_2^2 - \ldots - 2x_N y_N + 2\beta x_N^2$$

$$= -2\sum x_i\left(y_i - \beta x_i\right) = 0$$

Multiply both sides by -2, then eliminate the parentheses.

$$\sum x_i y_i - \beta\sum x_i^2 = 0$$

$$\beta\sum x_i^2 = \sum x_i y_i$$

Solve for β and let the solution value be $\hat{\beta}$:

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$$

This is the famous OLS estimator. Once we get an estimate of β, we can back out the OLS estimate for α using the means.

$$\bar{Y} = \alpha + \beta\bar{X}$$

$$\bar{Y} = \hat{\alpha} + \hat{\beta}\bar{X}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

The first property of the OLS estimate $\hat{\beta}$, is that it is a linear function of the observations on Y.

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{1}{\sum x_i^2}\sum x_i y_i = \sum w_i y_i$$

where

$$w_i = \frac{x_i}{\sum x_i^2}$$

The formula for $\hat{\beta}$ is linear because the w are constants. (Remember, by Gauss-Markov assumption (1), the x's are constant.)

$\hat{\beta}$ is an estimator. It is a formula for deriving an estimate. When we plug in the values of x and y from the sample, we can compute the estimate, which is a numerical value.

Note that so far, we have only described the data with a linear function. We have not used our estimates to make any inferences concerning the true value of α or β.

Sampling distribution of $\hat{\beta}$

Because y is a random variable and we now know that the OLS estimator $\hat{\beta}$ is a linear function of y, it must be that $\hat{\beta}$ is a random variable. If $\hat{\beta}$ is a random variable, it has a distribution with a mean and a variance. We call this distribution the sampling distribution of $\hat{\beta}$. We can use this sampling distribution to

make inferences concerning the true value of β. There is also a sampling distribution of $\hat{\alpha}$, but we seldom want to make inferences concerning the intercept, so we will ignore it here.

Mean of the sampling distribution of $\hat{\beta}$

The formula for $\hat{\beta}$ is

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$$

substitute $y_i = \beta x_i + u_i$

$$\hat{\beta} = \frac{\sum x_i (\beta x_i + u_i)}{\sum x_i^2}$$

$$= \frac{\beta \sum x_i^2 + \sum x_i u_i}{\sum x_i^2} = \frac{\beta \sum x_i^2}{\sum x_i^2} + \frac{\sum x_i u_i}{\sum x_i^2} = \beta + \frac{\sum x_i u_i}{\sum x_i^2}$$

The mean of $\hat{\beta}$ is its expected value,

$$E(\hat{\beta}) = E\left(\frac{\sum x_i y_i}{\sum x_i^2}\right) = E\left(\beta + \frac{\sum x_i u_i}{\sum x_i^2}\right)$$

$$= E(\beta) + E\left(\frac{\sum x_i u_i}{\sum x_i^2}\right) = \beta + \frac{\sum x_i E(u_i)}{\sum x_i^2} = \beta$$

because $E(u_i) = 0$ by GM assumption (1).

The fact that the mean of the sampling distribution of $\hat{\beta}$ is equal to the truth means that the ordinary least squares estimator is unbiased and thus a good basis on which to make inferences concerning the true β. Note that the unbiasedness property does not depend on any assumption concerning the form of the distribution of the error terms. We have not, for example, assumed that the errors are distributed normally. However, we will assume normally distributed errors below. It also relies only on two of the GM assumptions, that the x's are fixed and that $E(u_i) = 0$. So we really don't need the independent and identical assumptions for unbiasedness.

However, the independent and identical assumptions are useful in deriving the property that OLS is efficient. That is, if these two assumptions, along with the other, are true, then the OLS estimator is the Best Linear Unbiased Estimator (BLUE). It can be shown that the OLS estimator has the smallest variance of any linear unbiased estimator.

Variance of $\hat{\beta}$

The variance of $\hat{\beta}$ can be derived as follows.

$$Var(\hat{\beta}) = E\left(\hat{\beta} - E(\hat{\beta})\right)^2 = E\left(\hat{\beta} - \beta\right)^2$$

We know that

$$\hat{\beta} = \beta + \frac{\sum x_i u_i}{\sum x_i^2}$$

So,

$$\hat{\beta} - \beta = \frac{\sum x_i u_i}{\sum x_i^2} = \sum w_i u_i$$

and

$$E\left(\hat{\beta} - \beta\right)^2 = E\left(\frac{\sum x_i u_i}{\sum x_i^2}\right)^2 = E\left(\sum w_i u_i\right)^2$$

$$E\left(\sum w_i u_i\right)^2 = \left(w_1 u_1 + w_2 u_2 + ... + w_N u_N\right)^2$$

$$= E\left(w_1^2 u_1^2 + w_2^2 u_2^2 + ... + w_N^2 u_N^2 + w_1 w_2 u_1 u_2 + w_1 w_3 u_1 u_3 + ... + w_1 w_N u_1 u_N + ... w_{N-1} w_N u_{N-1} u_N\right)$$

$$= E\left(w_1^2 u_1^2 + w_2^2 u_2^2 + ... + w_N^2 u_N^2\right)$$

because the covariances among the errors are zero due to the independence assumption, that is,, $E(u_i u_j) = 0$ for i ≠ j . Therefore,

$$E\left(\sum w_i u_i\right)^2 = w_1^2 E(u_1^2) + w_2^2 E(u_2^2) + ... + w_N^2 E(u_N^2) = w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2 + ... + w_N^2 \sigma_N^2$$

However, $\sigma_1^2 = \sigma_2^2 = ... = \sigma_N^2 = \sigma^2$ because of the identical assumption. Thus,

$$Var\left(\hat{\beta}\right) = w_1^2 \sigma^2 + w_2^2 \sigma^2 + ... + w_N^2 \sigma^2 = \left(w_1^2 + w_2^2 + ... + w_N^2\right)\sigma^2$$

$$= \left(\frac{x_1}{\sum x_i^2}\right)^2 \sigma^2 + \left(\frac{x_2}{\sum x_i^2}\right)^2 \sigma^2 + ... + \left(\frac{x_N}{\sum x_i^2}\right)^2 \sigma^2$$

$$= \frac{1}{\left(\sum x_i^2\right)^2}\left(x_1^2 + x_2^2 + ... + x_N^2\right)\sigma^2$$

$$= \frac{\sum x_i^2}{\left(\sum x_i^2\right)^2}\sigma^2 = \frac{1}{\sum x_i^2}\sigma^2 = \frac{\sigma^2}{\sum x_i^2}$$

Consistency of OLS

To be consistent, the mean square error of $\hat{\beta}$ must go to zero as the sample size, N, goes to infinity. The mean square error of $\hat{\beta}$ is the sum of the variance plus the squared bias.

$$mse(\hat{\beta}) = Var(\hat{\beta}) + \left[E(\hat{\beta} - \beta)\right]^2$$

Obviously the bias is zero since the OLS estimate is unbiased. Therefore the consistency of the ordinary least estimator $\hat{\beta}$ depends on the variance of $\hat{\beta}$, $Var(\hat{\beta}) = \sigma^2 \Big/ \sum_{i=1}^{N} x_i^2$. As N goes to infinity, $\sum_{i=1}^{N} x_1^2$ goes to infinity because we keep adding more and more squared terms. Therefore, the variance of $\hat{\beta}$ goes to zero as N goes to infinity, the mean square error goes to zero, and the OLS estimate is consistent.

Proof of the Gauss-Markov Theorem

According to the Gauss-Markov theorem, if the G-M assumptions are true, the ordinary least squares estimator is the Best Linear Unbiased Estimator (BLUE). That is, of all linear unbiased estimators, the OLS estimator has the smallest variance.

Stated another way, any other linear unbiased estimator will have a larger variance than the OLS estimator. The Gauss-Markov assumptions are as follows.

(1) The true relationship between y and x (expressed as deviations from the mean) is

$$y_i = \beta x_i + u_i$$

(2) The residuals are independent and the variance is constant: $u_i \sim iid(o,\sigma^2)$.

(3) The x's are fixed numbers.

We know that the OLS estimator is a linear estimator

$$\hat{\beta} = \frac{\sum xy}{\sum x^2} = \sum wy$$

Where

$$w = x / \sum x^2 .$$

$\hat{\beta}$ is unbiased and has a variance equal to $\sigma^2 / \sum x^2$.

Consider any other linear unbiased estimator,

$$\tilde{\beta} = \sum cy = \sum c(\beta x + u) = \beta \sum cx + \sum cu$$

Taking expected values,

$$\begin{aligned} E(\tilde{\beta}) &= \beta \sum cx + \sum cE(u) \\ &= \beta \sum cx \end{aligned}$$

Since E(u)=0 as before. This estimator is unbiased if $\sum cx$ **=1**.

Note, since $\sum cx$ **=1**

$$\tilde{\beta} = \beta + \sum cu$$

and

$$\tilde{\beta} - \beta = \sum cu$$

Find the variance of $\tilde{\beta}$.

$$\begin{aligned} Var(\tilde{\beta}) &= E\left[\tilde{\beta} - \beta\right]^2 = E\left[\sum cu\right]^2 \\ &= E\left[c_1 u_1 + c_2 u_2 + ... + c_N u_2\right]^2 \\ &= \left[c_1^2 \sigma^2 + c_2^2 \sigma^2 + ... + c_N^2 \sigma^2\right] \\ &= \sigma^2 \sum c^2 \end{aligned}$$

According to the theorem, this variance must be greater than or equal to the variance of $\tilde{\beta}$.

Here is a useful identity.

$$c = w + (c - w)$$

$$c^2 = w^2 + (c - w)^2 + 2w(c - w)$$

$$\sum c^2 = \sum w^2 + \sum (c - w)^2 + 2\sum w(c - w)$$

The last term on the RHS is equal to zero.

$$\sum w(c - w) = \sum wc - \sum w^2$$

$$= \sum \frac{cx}{\sum x^2} - \sum \frac{x^2}{\left(\sum x^2\right)^2}$$

$$= \frac{\sum cx}{\sum x^2} - \frac{\sum x^2}{\left(\sum x^2\right)^2}$$

$$= \frac{1}{\sum x^2} - \frac{1}{\sum x^2} = 0$$

Since $\sum cx$ **=1**.

Therefore,

$$Var(\tilde{\beta}) = \sigma^2 \sum c^2$$

$$= \sigma^2 \sum (w^2 + (c - w)^2)$$

$$= \sigma^2 \sum w^2 + \sigma^2 \sum (c - w)^2$$

$$= \sigma^2 \frac{\sum x^2}{\left(\sum x^2\right)^2} + \sigma^2 \sum (c - w)^2$$

$$= \frac{\sigma^2}{\sum x^2} + \sigma^2 \sum (c - w)^2$$

$$= Var(\hat{\beta}) + \sigma^2 \sum (c - w)^2$$

So $Var(\tilde{\beta}) \geq Var(\hat{\beta})$ because $\sigma^2 \sum (c - w)^2 \geq 0$. The only way the two variances are equal is if c=w, which is the OLS estimator.

Inference and hypothesis testing

Normal, Student's t, Fisher's F, and Chi-square

Before we can make inferences or test hypotheses concerning the true value of β, we need to review some famous statistical distributions. We will be using all of the following distributions throughout the semester.

Normal distribution

The normal distribution is the familiar bell shaped curve. With mean μ and variance σ^2, it has the formula,

$$f\left(X \mid \mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(X-\mu)^2}{2\sigma^2}}$$

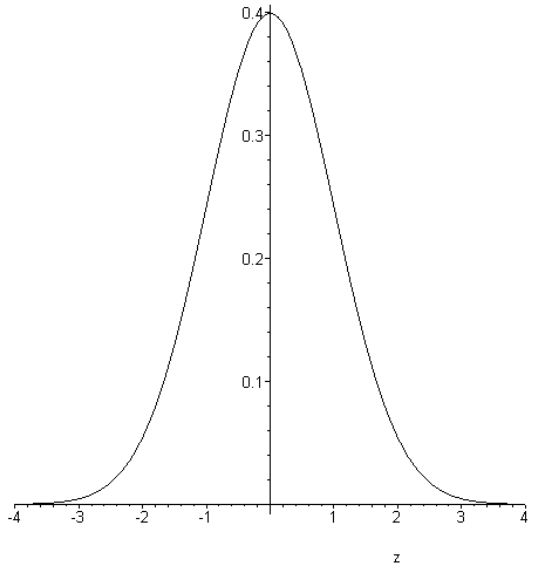
or simply $X \sim N\left(\mu,\sigma^2\right)$ where the tilde is read "is distributed as."

If X is distributed normally, then any linear transformation of X is also distributed normally. That is, if $X \sim N\left(\mu,\sigma^2\right)$ then $(a+bX) \sim N(a+b\mu,b^2\sigma^2)$ This turns out to be extremely useful, Consider the linear function of X, $z=a+bX$. Let $a=-\mu/\sigma$ and $b=1/\sigma$ then $z=a+bX=-\mu/\sigma+X/\sigma=(X-\mu)/\sigma$ which is the formula for a standard score, or z-score.

If we express X as a standard score, z, then z has a standard normal distribution because the mean of z is zero and its standard deviation is one. That is,

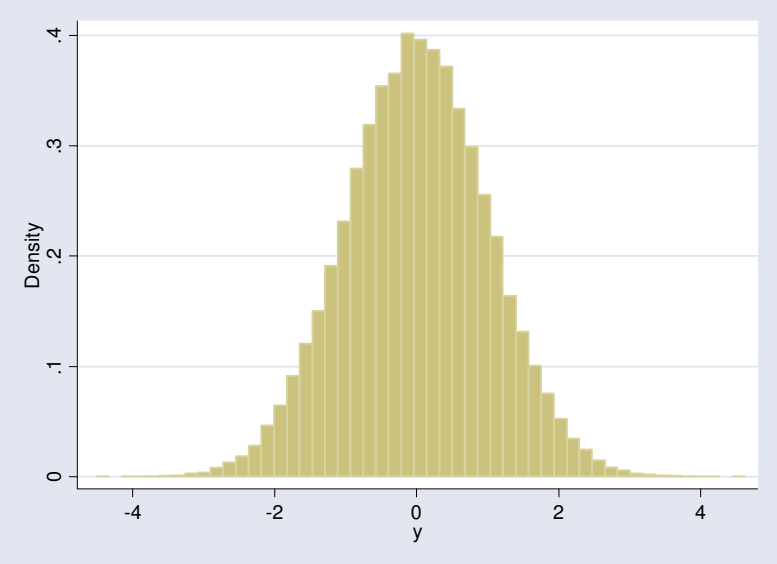
$$\varphi(z)=\frac{1}{\sqrt{2\pi}}e^{\frac{-z^2}{2}}$$

Here is the graph.



We can generate a normal variable with 100,000 observations as follows.

```
. set obs 100000
obs was 0, now 100000

. gen y=invnorm(uniform())
. summarize y

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           y |    100000    .0042645    1.002622  -4.518918   4.448323
. histogram y
(bin=50, start=-4.518918, width=.17934483)
```

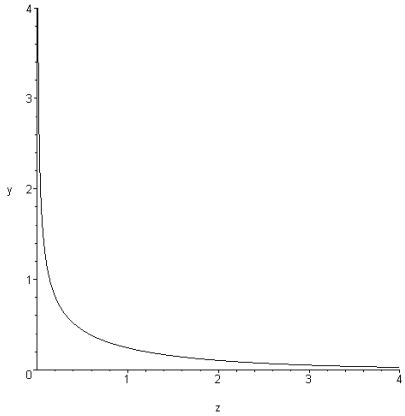


The Student's t, Fisher's F, and chi-square are derived from the normal and arise as distributions of sums of variables. They have associated with them their so-called degrees of freedom which are the number of terms in the summation.

Chi-square distribution

If z is distributed standard normal, then its square is distributed as chi-square with one degree of freedom. Since chi-square is a distribution of squared values, and, being a probability distribution, the area under the curve must sum to one, it has to be a skewed distribution of positive values, asymptotic to the horizontal axis.

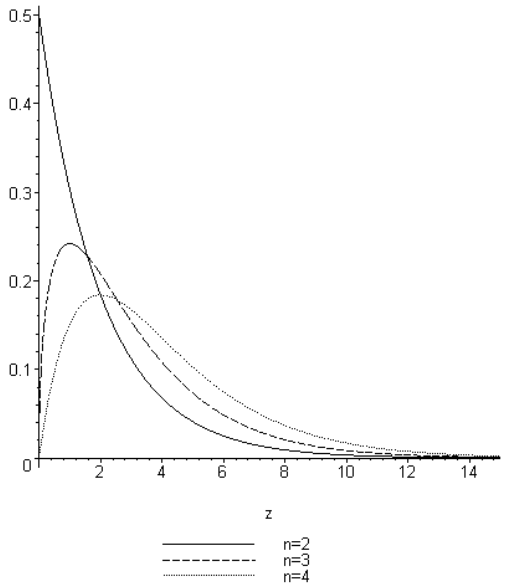
(1) If z ~ N(0,1), then $z^2 \sim \chi^2(1)$ which has a mean of 1 and a variance of 2. Here is the graph.



(2) If $X_1, X_2, ..., X_n$ are independent $\chi^2(1)$ variables, then $\sum_{i=1}^{N} X_i \sim \chi^2(n)$ with mean=n and variance = 2n

(3) Therefore, if $z_1, z_2, ..., z_n$ are independent standard normal, N(0,1), variables, then $\sum_{i=1}^{n} z_i^2 \sim \chi^2(n)$.

Here are the graphs for n=2,3,and 4.

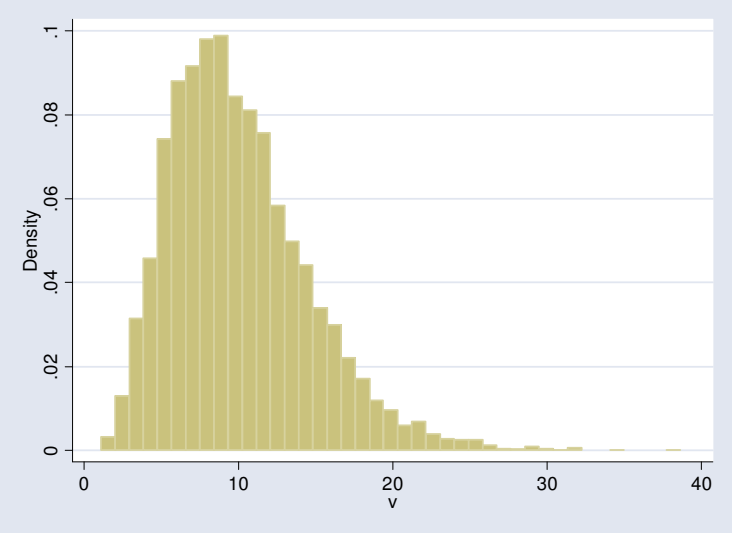


(4) If $z_1, z_2, ..., z_n$ are independent N(0,σ^2) variables then $\sum_{i=1}^{n} \left(\frac{z_i}{\sigma}\right)^2 \sim \chi^2(n)$.

(5) If $X_1 \sim \chi^2(n_1)$ and $X_2 \sim \chi^2(n_2)$ and X_1 and X_2 are independent, then $X_1 + X_2 \sim \chi^2(n_1 + n_2)$.

We can generate a chi-square variable by squaring a bunch of normal variables.

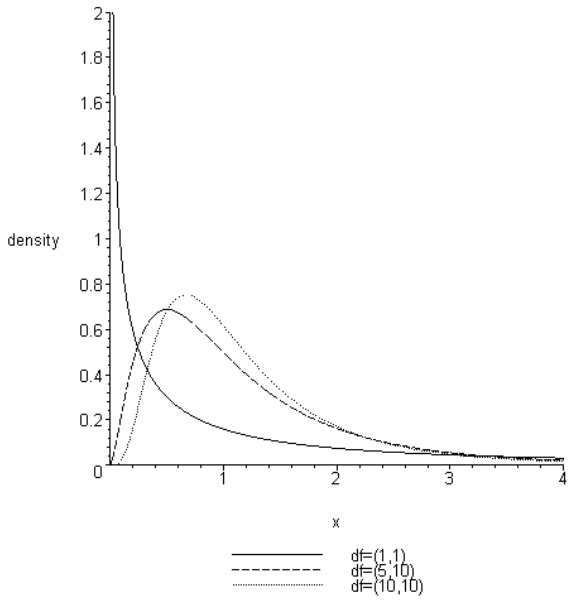
```
. gen z1=invnorm(uniform())
. set obs 10000
obs was 0, now 10000

. gen z1=invnorm(uniform())
. gen z2=invnorm(uniform())
. gen z3=invnorm(uniform())
. gen z4=invnorm(uniform())
. gen z5=invnorm(uniform())
. gen v=z1+z2+z3+z4+z5
. gen z6=invnorm(uniform())
. gen z7=invnorm(uniform())
. gen z8=invnorm(uniform())
. gen z9=invnorm(uniform())
. gen z10=invnorm(uniform())
. gen v=z1^2+z2^2+z3^2+z4^2+z5^2+z6^2+z7^2+z8^2+z9^2+z10^2
.*this creates a chi-square variable with n=4 degrees of freedom
. histogram v
(bin=40, start=1.0842115, width=.91651339)
```



F-distribution

(6) If $X_1 \sim \chi^2(n_1)$ and $X_2 \sim \chi^2(n_2)$ and X_1 and X_2 are independent, then $\dfrac{X_1/n_1}{X_2/n_2} \sim F(n_1, n_2)$

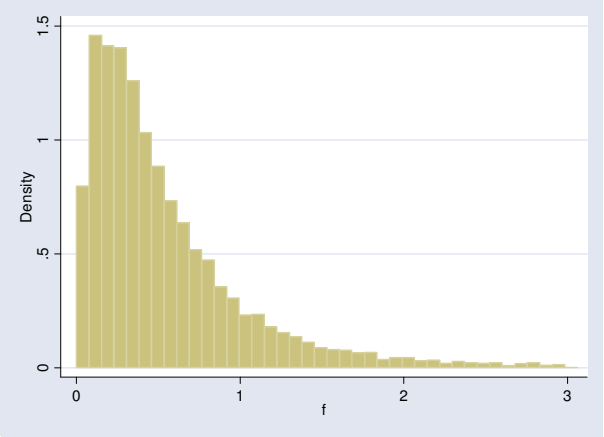
This is Fisher's Famous F-test. It is the ratio of two squared quantities, so it is positive and therefore skewed, like the chi-square. Here are graphs of some typical F distributions.



We can create a variable with an F distribution is follows: generate five standard normal variables, then sum their squares to make a chi-square variable (v) with five degrees of freedom; finally, divide v by z to make a variable distributed as F with 5 degrees of freedom in the numerator and 10 degrees of freedom in the denominator.

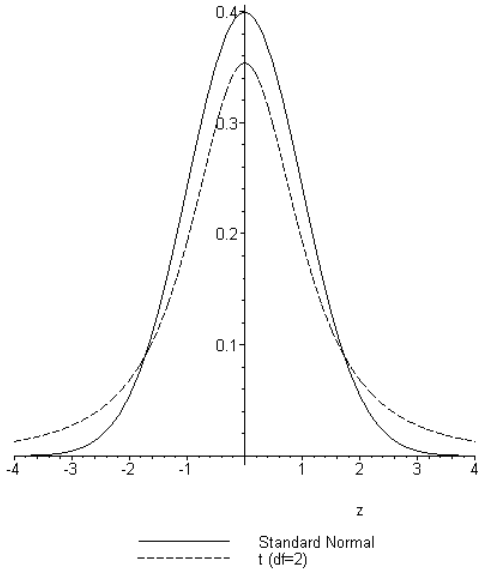
```
. gen x1=invnorm(uniform())
```

```
. gen x2=invnorm(uniform())
. gen x3=invnorm(uniform())
. gen x4=invnorm(uniform())
. gen x5=invnorm(uniform())
. gen v1=x1^2+x2^2+x3^3+x4^2+x5^2
. gen f=v1/v
. * f has 5 and 10 degrees of freedom
. summarize v1 v f

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          v1 |     10000    3.962596    4.791616  -62.87189   66.54086
           v |     10000     10.0526    4.478675   1.084211   37.74475
           f |     10000    .4877415    .7094975  -9.901778   14.72194
. histogram f, if f<3 & f>=0
```



t-distribution

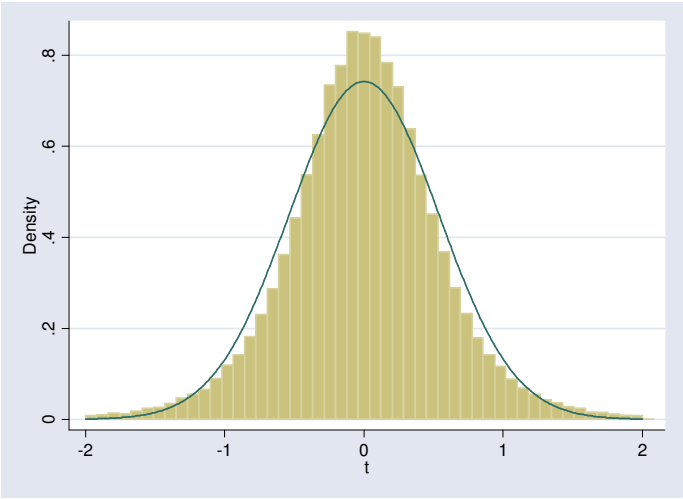
(7) If z ~ N(0,1) and $X \sim \chi^2(n)$ and z and X are independent, then the ratio $T = \frac{z}{\sqrt{X/n}} \sim t(n)$. This is Student's t distribution. It can be positive or negative and is symmetric, like the normal distribution. It looks like the normal distribution but with "fat tails," that is, more probability in the tails of the distribution, compared to the standard normal.



(8) If $T \sim t(n)$ then $T^2 = \frac{z^2}{X/n} \sim F(1,n)$ because the squared t-ratio is the ratio of two chi-squared variables.

We can create a variable with a t distribution by dividing a standard normal variable by a chi-square variable.

```
. set obs 100000
obs was 0, now 100000
. gen z=invnorm(uniform())
. gen y1=invnorm(uniform())
. gen y2=invnorm(uniform())
. gen y3=invnorm(uniform())
. gen y4=invnorm(uniform())
. gen y5=invnorm(uniform())
. gen v=y1^2+y2^2+y3^2+y4^2+y5^2
. gen t=z/sqrt(v)
. histogram t if abs(t)<2, normal
(bin=49, start=-1.9989421, width=.08160601)
```



Asymptotic properties

(9) As N goes to infinity, $t \to N(0,1)$

(10) As $n_2 \to \infty, F = \frac{X_1/n_1}{X_2/n_2} \to X_1/n_1 \sim \chi^2(n_1)$
because as $n_2 \to \infty, X_2 \to \infty$, so that the ratio goes to one.

(11) As $N \to \infty, NR^2 \to \chi^2$
where $R^2 = \frac{RSS}{TSS} = \frac{RSS/N}{TSS/N}$
and RSS is the regression (model) sum of squares.
$NR^2 = N\left[\frac{RSS/N}{TSS/N}\right]$
As $N \to \infty$, $TSS/N \to 1$ so that $NR^2 \to N\left[\frac{RSS/N}{1}\right] = RSS \sim \chi^2$

(12) As $N \to \infty, NF \sim \chi^2$. See (10) above.

Testing hypotheses concerning β

So, we can summarize the sampling distribution of $\hat{\beta}$ as $\hat{\beta} \sim iid\left(\beta, \sigma^2/\sum x_i^2\right)$. If we want to actually use this distribution to make inferences, we need to make an assumption concerning the form of the function. Let's make the usual assumption that the errors in the original regression model are distributed normally. That is, $u_i \sim iidN\left(0, \sigma^2\right)$ where the N stands for normal. This assumption can also be written as $u_i \sim IN(0, \sigma^2)$ which is translated as "u_i is distributed independent normal with mean zero and variance σ^2." (We don't need the identical assumption because the normal distribution has a constant variance.) Since $\hat{\beta}$ is a linear function of the error terms and the error terms are distributed normally, then $\hat{\beta}$ is distributed normally, that is, $\hat{\beta} \sim IN\left(\beta, \sigma^2/\sum x_i^2\right)$. We are now in position to make inferences concerning the true value of β.

To test the null hypothesis that $\beta = \beta_0$ where β_0 is some hypothesized value of β, construct the test statistic, $T = \frac{\hat{\beta} - \beta_0}{S_{\hat{\beta}}}$ where $S_{\hat{\beta}} = \sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2}}$ is the standard error of $\hat{\beta}$, the square root of its variance.

$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N-k} e_i^2}{N-k}$ where e_i^2 is the squared residual corresponding to the i'th observation, N is the sample size, and k is the number of parameters in the regression. In this simple regression, k=2 because we have two parameters, α and β. $\hat{\sigma}^2$ is the estimated error variance. It is equal to the error sum of squares divided by the degrees of freedom (N-k) and is also known as the mean square error.

Notice that the test statistic, also known as a t-ratio, is the ratio of a normally distributed random variable, $\hat{\beta}$, to a chi-square variable (the sum of a bunch of squared, normally distributed, error terms). Consequently, it is distributed as student's t with N-k degrees of freedom.

By far the most common test is the significance test, namely that X and Y are unrelated. To do this test set $\beta_0 = 0$ so that the test statistic reduces to the ratio of the estimated parameter to its standard error. $T = \frac{\hat{\beta}}{S_{\hat{\beta}}}$. This is the "t" reported in the regress output in Stata.

Degrees of Freedom

Karl Pearson, who discovered the chi-square distribution around 1900, applied it to a variety of problems but couldn't get the parameter value quite right. Pearson, and a number of other statisticians, kept requiring ad hoc adjustments to their chi-square applications. Sir Ronald Fisher solved the problem in 1922 and called the parameter **degrees of freedom**. The best explanation of the concept that I know of is due to Gerrard Dallal (http://www.tufts.edu/~gdallal/dof.htm). Suppose we have a data set with N observations. We can use the data in two ways, to estimate the parameters or to estimate the variance. If we use data points to estimate the parameters, we can't use them to estimate the variance.

Suppose we are trying to estimate the mean of a data set with X=(1,2,3). The sum is 6 and the mean is 2. Once we know this, we can find any data point knowing only the other two. That is, we used up one degree of freedom to estimate the mean and we only have two left to estimate the variance. To take another example, suppose we are estimating a simple regression model with ordinary least squares. We estimate the slope with the formula $\hat{\beta} = \sum xy / \sum x^2$ and the intercept with the formula $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$, using up two data points. This leaves the remaining N-2 data points to estimate the variance. We can also look at the regression problem this way: it takes two points to determine a straight line. If we only have two points, we can solve for the slope and intercept, but this is a mathematical problem, not a statistical one. If we have three data points, then there are many lines that could be defined as "fitting" the data. There is one degree of freedom that could be used to estimate the variance.

Our simple rule is that the number of degrees of freedom is the sample size minus the number of parameters estimated. In the example of the mean, we have one parameter to estimate, yielding N-1 data points to estimate the variance. Therefore, the formula for the sample variance of X (we have necessarily already used the data once to compute the sample mean) is

$$S_x^2 = \frac{\sum (X - \bar{X})^2}{N - 1}$$

where we divide by the number of degrees of freedom (N=1), not the sample size (N). To prove this, we have to demonstrate that this is an unbiased estimator for the variance of X (and therefore, dividing by N yields a biased estimator).

$$\begin{aligned}\sum (X - \bar{X})^2 &= \sum \left[(X - \mu) - (\bar{X} - \mu) \right]^2 \\ &= \sum \left[(X - \mu)^2 - 2(X - \mu)(\bar{X} - \mu) + (\bar{X} - \mu)^2 \right] \\ &= \sum (X - \mu)^2 - 2(\bar{X} - \mu) \sum (X - \mu) + N(\bar{X} - \mu)^2\end{aligned}$$

because $\bar{X}$ and μ are both constants. Also,

$$\sum (X - \mu) = \sum X - N\mu = N\bar{X} - N\mu = N(\bar{X} - \mu)$$

which means that the middle term in the quadratic form above is

$$-2(\bar{X}-\mu)\sum(X-\mu) = -2(\bar{X}-\mu)N(\bar{X}-\mu) = -2N(\bar{X}-\mu)^2$$

and thefore,

$$\begin{aligned}\sum(X-\bar{X})^2 &= \sum(X-\mu)^2 - 2N(\bar{X}-\mu)^2 + N(\bar{X}-\mu)^2 \\ &= \sum(X-\mu)^2 - N(\bar{X}-\mu)^2\end{aligned}$$

Now we can take the expected value of the proposed formula,

$$\begin{aligned}E\left(\frac{\sum(X-\bar{X})^2}{N-1}\right) &= E\left(\frac{1}{N-1}\sum(X-\mu)^2 - \frac{N}{N-1}(\bar{X}-\mu)^2\right) \\ &= \frac{1}{N-1}\sum E(X-\mu)^2 - \frac{N}{N-1}E(\bar{X}-\mu)^2\end{aligned}$$

Note that the second term is the variance of the mean of X:

$$E(\bar{X}-\mu)^2 = \frac{\sum(\bar{X}-\mu)^2}{N} = Var(\bar{X})$$

But,

$$\begin{aligned}Var(\bar{X}) &= Var\left(\frac{1}{N}\sum X\right) = \frac{1}{N^2}Var(X_1 + X_2 + ... + X_N) \\ &= \frac{1}{N^2}N\sigma_x^2 = \frac{\sigma_x^2}{N}\end{aligned}$$

Therefore,

$$\begin{aligned}E\left(\frac{\sum(X-\bar{X})^2}{N-1}\right) &= E\left(\frac{1}{N-1}\sum(X-\mu)^2 - \frac{N}{N-1}(\bar{X}-\mu)^2\right) \\ &= \frac{N}{N-1}\sigma_x^2 - \frac{N}{N-1}\frac{\sigma_x^2}{N} = \frac{N}{N-1}\sigma_x^2 - \frac{1}{N-1}\sigma_x^2 \\ &= \frac{N-1}{N-1}\sigma_x^2 = \sigma_x^2\end{aligned}$$

Note (for later reference),

$$E\left(\sum x^2\right) = E\left(\sum(X-\bar{X})^2\right) = (N-1)\sigma_x^2 .$$

Estimating the variance of the error term

We want to prove that the formula

$$\hat{\sigma}^2 = \frac{\sum e^2}{N-2}$$

where e is the residual from a simple regression model, yields an unbiased estimate of the true error variance.

In deviations,

$$e = y - \hat{\beta}x$$

But,

$$y = \hat{\beta}x + u$$

So,

$$e = \beta x + u - \hat{\beta}x = u - \left(\hat{\beta} - \beta\right)x$$

Squaring both sides,

$$e^2 = u^2 - 2\left(\hat{\beta} - \beta\right)xu + \left(\hat{\beta} - \beta\right)^2 x^2$$

Summing,

$$\sum e^2 = \sum u^2 - 2\sum\left(\hat{\beta} - \beta\right)xu + \left(\hat{\beta} - \beta\right)^2 \sum x^2$$

Taking expected values,

$$E\left[\sum e^2\right] = E\left[\sum u^2\right] - 2E\left[\left(\hat{\beta} - \beta\right)\sum xu\right] + E\left[\left(\hat{\beta} - \beta\right)^2 \sum x^2\right]$$

Note that the variance of $\hat{\beta}$ is

$$Var\left(\hat{\beta}\right) = E\left(\hat{\beta} - \beta\right)^2 = \frac{\sigma^2}{\sum x^2}$$

Which means that the third term of the expectation expression is

$$E\left[\sum\left(\hat{\beta} - \beta\right)^2 \sum x^2\right] = \frac{\sigma^2}{\sum x^2}\sum x^2 = \sigma^2$$

We know from the theory of least squares that

$$\hat{\beta} - \beta = \frac{\sum xu}{\sum x^2}$$

Which implies that

$$\sum xu = \left(\hat{\beta} - \beta\right)\sum x^2$$

Therefore the second term in the expectation expression is

$$-2E\left[\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)\sum x^2\right]=-2E\left[\left(\hat{\beta}-\beta\right)^2\sum x^2\right]$$
$$=-2\frac{\sigma^2}{\sum x^2}\sum x^2=-2\sigma^2$$

The first term in the expectation expression is

$$E\left[\sum u^2\right]=(N-1)\sigma^2$$

because the unbiased estimator of the variance of x is $S_x^2=\sum x^2/(N-1)$ so that the expected value of $E\left(\sum x^2\right)=(N-1)S_x^2$. In the current application x=e and $S_x^2=\sigma^2$.

Therefore, putting all three terms together, we get

$$E\left[\sum e^2\right]=(N-1)\sigma^2+\sigma^2-2\sigma^2=N\sigma^2-\sigma^2+\sigma^2-2\sigma^2=(N-2)\sigma^2$$

Which implies that the unbiased estimate of the error variance is

$$\hat{\sigma}^2=\frac{\sum e^2}{N-2}$$

$$E\left[\hat{\sigma}^2\right]=E\left[\frac{\sum e^2}{N-2}\right]=\frac{E\left[\sum e^2\right]}{N-2}=\frac{(N-2)\sigma^2}{N-2}=\sigma^2$$

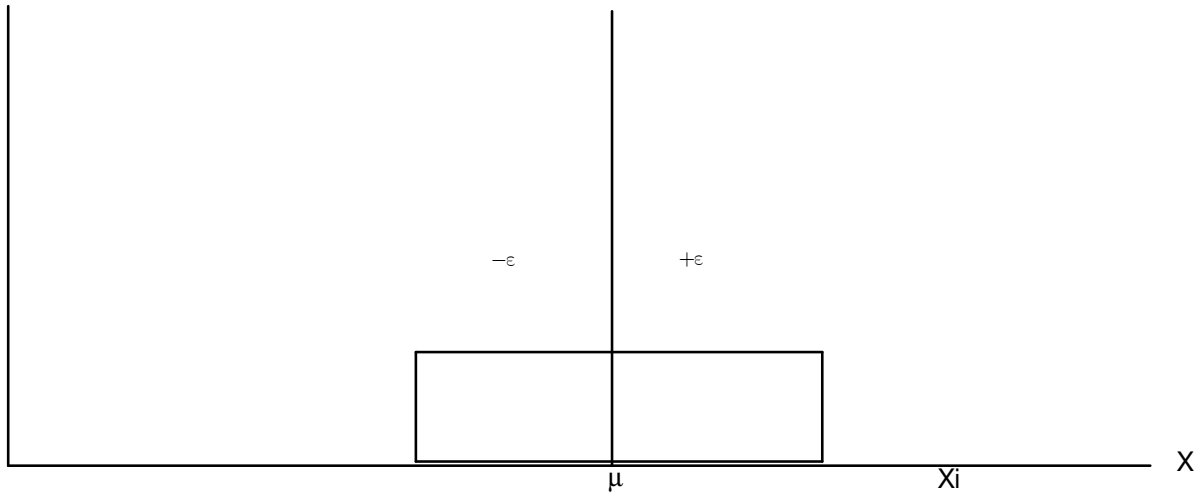
If we think of each of the squared residuals as an estimate of the variance of the error term, then the formula tells us to take the average. However, we only have N-2 independent observations on which to base our estimate of the variance, so we divide by N-2 when we take the average.

Chebyshev's Inequality

Arguably the most remarkable theorem in statistics, Chebyshev's inequality states that, for any variable x, from any arbitrary distribution, the probability that an observation lies more than some distance, ε, from the mean must be less than or equal to $\sigma_x^2 / \varepsilon^2$. That is,

$$P(|X-\mu| \geq \varepsilon) \leq \frac{\sigma_x^2}{\varepsilon^2}$$

The following diagram might be helpful.



Proof

The variance of x can be written as

$$\sigma_x^2 = \frac{\sum (X-\mu)^2}{N} = \sum (X-\mu)^2 \frac{1}{N}$$

if all observations have the same probability (1/N) of being observed.

More generally,

$$\sigma_x^2 = \sum (X-\mu)^2 P(X)$$

Suppose we limit the summation on the right hand side to those values of X that lie outside the $\pm\varepsilon$ box. Then the new summation could not be larger than the original summation, which is the variance of x, because we are omitting some (positive, squared) terms.

$$\sigma_x^2 = \sum (X-\mu)^2 P(X)$$

$$\geq \sum_{|X-\mu| \geq \varepsilon} (X-\mu)^2 P(X)$$

We also know that all of the X's in the limited summation are at least ε away from the mean. (Some of them are much further than ε away.) Therefore, replacing $(X-\mu)$ with ε,

$$\sigma_x^2 \geq \sum_{|X-\mu| \geq \varepsilon} (X-\mu)^2 P(X)$$

$$\geq \sum_{|X-\mu| \geq \varepsilon} \varepsilon^2 P(X) = \varepsilon^2 \sum_{|X-\mu| \geq \varepsilon} P(X) = \varepsilon^2 P(|X-\mu| \geq \varepsilon)$$

So,

$$\sigma_x^2 \geq \varepsilon^2 P(|X-\mu| \geq \varepsilon)$$

Dividing both sides by ε^2

$$\frac{\sigma_x^2}{\varepsilon^2} \geq P(|X-\mu| \geq \varepsilon)$$

Turning it around we get,

$$P(|X-\mu| \geq \varepsilon) \leq \frac{\sigma_x^2}{\varepsilon^2}$$

QED

The theorem is often stated as follows.

For any set of data and any constant k>1, at least $1-1/k^2$ of the observations must lie within k standard deviations of the mean.

To prove this version, let $\varepsilon = k\sigma_x$

$$P(|X-\mu| \geq k\sigma_x) \leq \frac{\sigma_x^2}{k^2\sigma_x^2} = \frac{1}{k^2}$$

Since this is the probability that the observations lie more than k standard deviations from the mean, the probability that the observations lie less than k sd from the mean is one minus this probability. That is,

$$P(|X-\mu| \leq k\sigma_x) \leq 1-\frac{1}{k^2}$$

So, if we have any data whatever, and we set k=2, we know that

$1-\frac{1}{4}=\frac{3}{4}=75\%$ of the data lie within two standard deviations of the mean.

(We can be more precise if we know the distribution. For example, 95% of normally distributed data lie within two standard deviations of the mean.)

Law of Large Numbers

The law of large numbers states that if a situation is repeated again and again, the proportion of successful outcomes will tend to approach the constant probability that any one of the outcomes will be a success. For example, suppose we have been challenged to guess how many thumbtacks will end up face down if 100 are dropped off a table. (Answer: 50.)

How would we prepare for such a challenge? Toss a thumbtack into the air and record how many times it ends up face down. Divide that into the number of trials and you have the proportion. Multiply by 100 and that is your guess.

Suppose we are presented with a large box full of black and white balls. We want to know the proportion of black balls, but we are not allowed to look inside the box. What do we do? Answer: sample with replacement. The number of black balls that we draw divided by the total number of balls drawn will give us an estimate. The more balls we draw, the better the estimate.

Mathematically, the theorem is,

Let $X_1, X_2, ..., X_N$ be an independent trials process where $E(X_i) = \mu$ and $\sigma_x^2 = Var(X_i)$.

Let $S = X_1 + X_2 + ... + X_N$. Then, for any ε > 0

$$P\left(\left|\frac{S}{N} - \mu\right| \geq \varepsilon\right) \xrightarrow[N\to\infty]{} 0$$

Equivalently,

$$P\left(\left|\frac{S}{N} - \mu\right| < \varepsilon\right) \xrightarrow[N\to\infty]{} 1$$

Proof,
We know that

$$Var(S) = Var(X_1 + X_2 + ... + X_N) = N\sigma_x^2$$

Because there is no covariance between the X's. Therefore,

$$Var\left(\frac{S}{N}\right) = Var((X_1 + X_2 + ... + X_N)/N) = \frac{N\sigma_x^2}{N^2} = \frac{\sigma_x^2}{N}$$

Also,

$$E\left(\frac{S}{N}\right) = E\left(\frac{X_1 + X_2 + ... + X_N}{N}\right) = \frac{N\mu}{N} = \mu$$

By Chebyshev's Inequality we know that, for any ε > 0,

$$P\left(\left|\frac{S}{N} - \mu\right| \geq \varepsilon\right) \leq \frac{Var\left(S/N\right)}{\varepsilon^2} = \frac{\sigma_x^2}{N\varepsilon^2} \xrightarrow[N\to\infty]{} 0$$

Equivalently, the probability that the difference between S/N and the true mean is less than ε is one minus this. That is,

$$P\left(\left|\frac{S}{N} - \mu\right| < \varepsilon\right) \xrightarrow[N\to\infty]{} 1$$

Since S/N is an average, the law of large numbers is often called the law of averages.

Central Limit Theorem

This is arguably the most amazing theorem in mathematics.

Statement:
Let x be a random variable distributed according to f(x) with mean μ and variance σ_X^2. Let $\bar{X}$ be the mean of a random sample of size N from f(x).

Let $z = \dfrac{\bar{X} - \mu}{\sqrt{\sigma_X^2/N}}$.

The density of z will approach the standard normal (mean zero and variance one) as N goes to infinity. The incredible part of this theorem is that no restriction is placed on the distribution of x. No matter how x is distributed (as long as it has a finite variance), the sample mean of a large sample will be distributed normally.

Not only that, but the sample doesn't really have to be that large. It used to be that a sample of 30 was thought to be enough. Nowadays we usually require 100 or more. Infinity comes quickly for the normal distribution.

The CLT is the reason that the normal distribution is so important.

The theorem was first proved by DeMoivre in 1733 but was promptly forgotten. It was resurrected in 1812 by LaPlace, who derived the normal approximation to the binomial distribution, but was still mostly ignored until Lyapunov in 1901 generalized it and showed how it worked mathematically. Nowadays, the central limit theorem is considered to be the unofficial sovereign of probability theory.

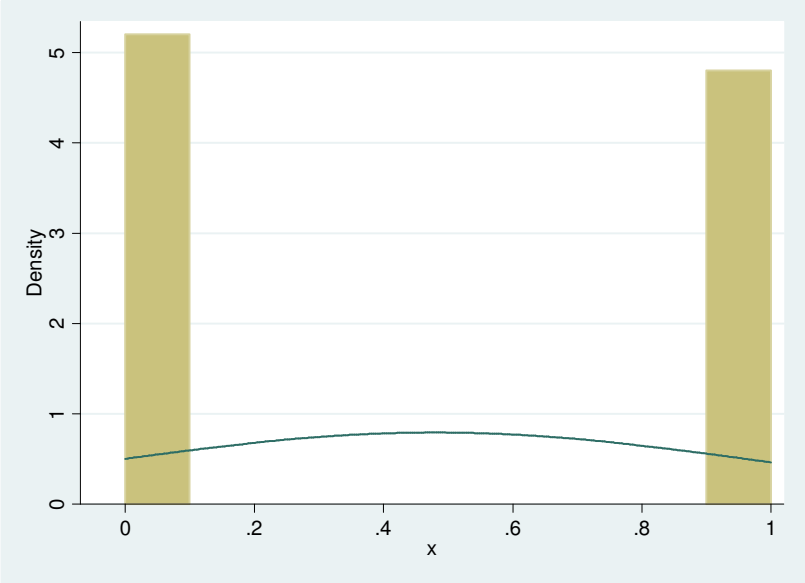
One of the few people who didn't ignore it in the 19'th Century was Sir Francis Galton, who was a big fan. He wrote in *Natural Inheritance*, 1889:

> I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error". The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement, amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshaled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along.

We can't prove this theorem here. However, we can prove it for certain cases using Monte Carlo methods.

For example, the binomial distribution is decidedly non-normal since values must be either zero or one. However, according to the theorem, the mean of a large number of sample means will be distributed normally.

Here is the histogram of the binomial distribution with an equal probability of generating a one or a zero..

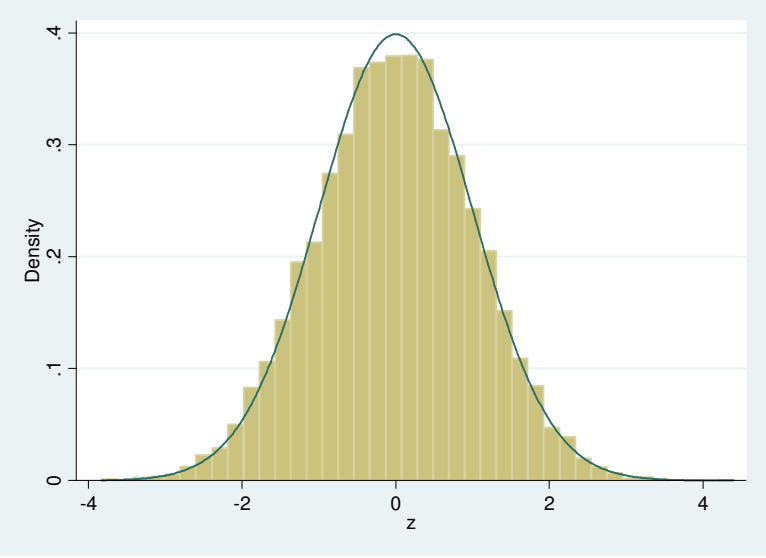


Not normal.

I asked Stata to generate 10,000 samples of size N=100 from this distribution. I then computed the mean and variance of the resulting 10,000 observations.

```
  Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         sum |     10000     50.0912    4.981402         31         72
```

I then computed the z-scores by taking each observation, subtracting the mean (50.0912) and dividing by the standard deviation. The histogram of the resulting data is:

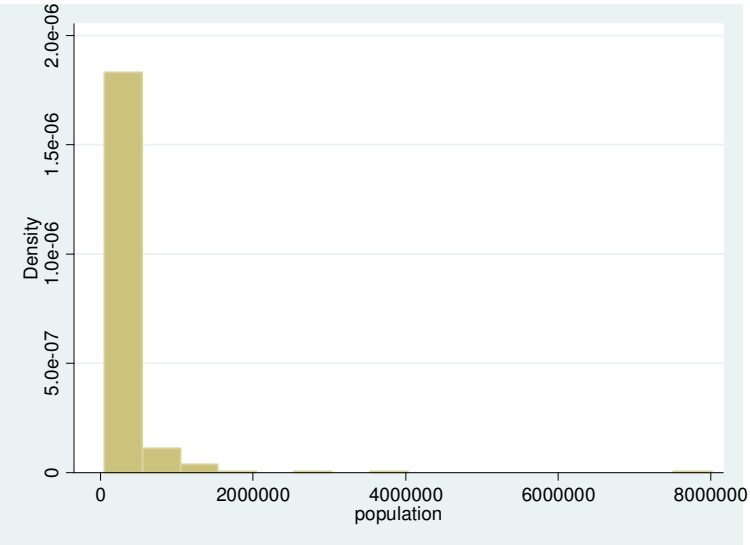


This is the standard normal distribution with mean zero and standard deviation one.

```
. summarize z
```

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           z |     10000   -3.57e-07           1  -3.832496   4.398119
```

Here is another example. This is the distribution of the population of cities in the United States with popularion of 100,000 or more.



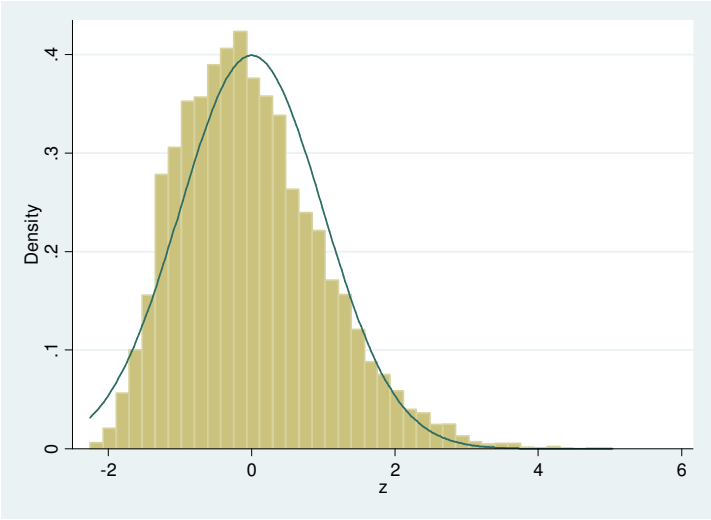
It is obviously not normal being highly skewed. There are very few very large cities and a great number of relatively small cities. Suppose we treat this distribution as the population. I told Stata to take 10,000 samples of size N=200 from this data set. I then computed the mean and standard deviation of the resulting data.

```
summarize mx

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          mx |     10000      287657    41619.64   193870.2   497314.6
```

I then computed the corresponding z-scores. According to the Central Limit Theorem, the z-scores should have a zero mean and unit variance.

```
. summarize z

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           z |     10000   -1.37e-07           1  -2.253426    5.03747
```

As expected, the mean is zero and the standard deviation is one. Here is the histogram.



It still looks a little skewed, but it is very close to the standard normal.

Method of maximum likelihood

Suppose we have the following simple regression model

$$Y_i = \alpha + \beta X_i + U_i$$
$$U_i \sim IN(0, \sigma^2)$$

So, the Y are distributed normally. The probability of observing the first Y is

$$\Pr(Y_1) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left\{-\frac{1}{2\sigma^2}U_1^2\right\}} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left\{-\frac{1}{2\sigma^2}(Y_1-\alpha-\beta X_1)^2\right\}}$$

The probability of observing the second Y is exactly the same, except for the subscript. The sample consists of N observations. The probability of observing the sample is the probability of observing Y_1 and Y_2, etc.

$$\Pr(Y_1, Y_2, ..., Y_N) = \Pr(Y_1)\Pr(Y_2)\cdots\Pr(Y_N)$$

because of the independence assumption. Therefore,

$$\Pr(Y_1, Y_2, ..., Y_N) = \prod_1^N \frac{1}{\left(2\pi\sigma^2\right)^{\frac{1}{2}}} e^{\left\{-\frac{1}{2\sigma^2}(Y_i-\alpha-\beta X_i)^2\right\}} = L(\alpha, \beta, \sigma)$$

This is the likelihood function, which gives the probability of observing the sample that we actually observed. It is a function of the unknown parameters, $\alpha, \beta,$ and σ. The symbol $\prod_1^N$ means multiply the items indexed from one to N.

The basic idea behind maximum likelihood is that we probably observed the most likely sample, not a rare or unusual sample, therefore, the best guess of the values of the parameters are those that generate the most likely sample. Mathematically, we maximize the likelihood of the sample by taking derivatives with respect to $\alpha, \beta,$ and σ, setting the derivatives equal to zero, and solving.

Let's take logs first to make the algebra easier.

$$\text{Log L} = \sum_{i=1}^{N}\left[-\tfrac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(Y_i - \alpha - \beta X_i)^2 \right]$$

This function has a lot of stuff that is not a function of the parameters.

$$\log \text{L} = C - \frac{N}{2}\log\sigma^2 - \frac{Q}{2\sigma^2}$$

where

$$C = \left(-\frac{N}{2}\right)\log(2\pi)$$

which is constant with respect to the unknown parameters, and

$$Q = \sum_{i=1}^{N}(Y_i - \alpha - \beta X_i)^2$$

where Q is the error sum of squares. Since Q is multiplied by a negative, to maximize the likelihood, we have to minimize the error sum of squares with respect to α and β. This is nothing more than ordinary least squares. Therefore, OLS is also a maximum likelihood estimator. We already know the OLS formulas, so the only thing we need now is the maximum likelihood estimator for the variance, σ^2,

Substitute the OLS values for α and β: $\hat{\alpha}$ and $\hat{\beta}$:

$$\log \text{L}(\sigma) = C - \frac{N}{2}\log(\sigma^2) - \frac{\hat{Q}}{2\sigma^2}$$

where

$$\hat{Q} = \sum_{i=1}^{N}(Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$$

is the residual sum of squares, ESS.

Now we have to differentiate with respect to σ and set the derivative equal to zero

$$\frac{d\log \text{L}(\sigma)}{d\sigma} = -\frac{N}{\sigma} + \frac{4\sigma\hat{Q}}{4\sigma^4} = 0$$

$$\frac{N}{\sigma} = \frac{\hat{Q}}{\sigma^3}$$

$$N = \frac{\hat{Q}}{\sigma^2}$$

$$\hat{\sigma}^2 = \frac{\hat{Q}}{N} = \frac{ESS}{N}$$

Now that we have estimates for $\alpha, \beta,$ and σ, we can substitute $\hat{\alpha}, \hat{\beta},$ and $\hat{\sigma}$ into the log likelihood function to get its maximum value:

$$\underset{\hat{\alpha},\hat{\beta},\hat{\sigma}}{\text{Max}} \log L = C - \frac{N}{2}\log(\sigma^2) - \frac{\hat{Q}}{2\sigma^2} = C - \frac{N}{2}\log\left(\frac{\hat{Q}}{N}\right) - \frac{N}{2}$$

Note that $\frac{\hat{Q}}{2\hat{\sigma}^2} = \frac{N\hat{\sigma}^2}{2\hat{\sigma}^2}$.

So,

$$\underset{\hat{\alpha},\hat{\beta},\hat{\sigma}}{\text{Max}} \log L = C - \frac{N}{2}\log\left(\hat{Q}\right) + \frac{N}{2}\log\left(N\right) - \frac{N}{2}$$

Since N is constant,

$$\underset{\hat{\alpha},\hat{\beta},\hat{\sigma}}{\text{Max}} \log L = const - \frac{N}{2}\log\left(\hat{Q}\right)$$

where

$$const = C + \frac{N}{2}\log\left(N\right) - \frac{N}{2}$$

Therefore,

$$MaxL = const * \hat{Q}^{-\frac{N}{2}} = const * (ESS)^{-\frac{N}{2}}$$

Likelihood ratio test

Let L be the likelihood function. We can use this function to test hypotheses such as $\beta = 0$ or $\alpha + \beta = 10$ as restrictions on the model. As in the F-test, if the restriction is false, we expect that the residual sum of squares (ESS) for the restricted model will be higher than the ESS for the unrestricted model. This means that the likelihood function will be smaller (not maximized) if the restriction is false. If the hypothesis is true, then the two values of ESS will be approximately equal and the two values of the likelihood functions will be approximately equal.

Consider the likelihood ratio,

$$\lambda = \frac{MaxL \text{ under the restriction}}{MaxL \text{ without the restriction}} \leq 1$$

The test statistic is

$$-2\log(\lambda) \sim \chi^2(p)$$

where p is the number of restrictions.

To do this test, do two regressions, one with the restriction, saving the residual sum of squares, ESS(R), and one without the restriction, saving ESS(U). So,

$$\lambda = \frac{const*(ESS(R))^{-\frac{N}{2}}}{const*(ESS(U))^{-\frac{N}{2}}} = \left(\frac{ESS(R)}{ESS(U)}\right)^{-\frac{N}{2}}$$

$$\log(\lambda) = -\frac{N}{2}\left(\log(ESS(R) - \log(ESS(U))\right)$$

The test statistic is

$$-2\log(\lambda) = N\left(\log(ESS(R) - \log(ESS(U))\right) \sim \chi^2(p)$$

This is a large sample test. Like all large sample tests, its significance is not well known in small samples. Usually we just assume that the small sample significance is about the same as it would be in a large sample.

Multiple regression and instrumental variables

Suppose we have a model with two explanatory variables. For example,

$$Y_i = \alpha + \beta X_i + \gamma Z_i + u_i, u_i \sim iid(0, \sigma^2)$$

We want to estimate the parameters α, β, and γ. We define the error as

$$e_i = Y_i - \alpha - \beta X_i - \gamma Z_i$$

At this point we could square the errors, sum them, take derivatives with respect to the parameters, set the derivatives equal to zero and solve. This would generate the least squares estimators for this multiple regression. However, there is another approach, known as instrumental variables, that is both illustrative and easier. To see how instrumental variables works, let's first derive the OLS estimators for the simple regression model,

$$Y_i = \alpha + \beta X_i + u_i$$

The crucial Gauss-Markov assumptions can be written as

(A1) $E(u_i) = 0$

(A2) $E(x_i u_i) = 0$

A1 states that the model must be correct on average, so the (true) mean of the errors is zero. A2 is the alternative to the assumption that the x's are a set of fixed numbers. It says that the x's cannot be correlated with the error term, that is, the covariance between the explanatory variable and the error term must be zero.

Since we do not observe u_i, we have to make the assumptions operational. We therefore define the following empirical analogs of A1 and A2.

(a1) $\overline{e} = \frac{\sum e_i}{N} = 0 \Rightarrow \sum e_i = 0$

(a2) $\frac{\sum \mathrm{x}_i e_i}{N} = 0 \Rightarrow \sum \mathrm{x}_i e_i = 0$

(a1) says that the mean of the residuals is zero and (a2) says that the covariance between x and the residuals is zero.

Define the residuals,
$e_i = Y_i - \alpha - \beta X_i$

Sum the residuals,

$$\sum_{i=1}^{N} e_i = \sum_{i=1}^{N} \left(Y_i - \alpha - \beta X_i\right) = \sum_{i=1}^{N} Y_i - N\alpha - \beta \sum_{i=1}^{N} X_i = 0 \text{ by (a1).}$$

Divide both sides by N.

$$\frac{\sum_{i=1}^{N} Y_i}{N} - \alpha - \beta \frac{\sum_{i=1}^{N} X_i}{N} = 0$$

Or,
$\bar{Y} - \alpha - \beta \bar{X} = 0$
$\hat{\alpha} = \bar{Y} - \beta \bar{X}$
which is the OLS estimator for α.

We can use (a2) to derive the OLS estimator for β.

$y_i = Y_i - \bar{Y} = \alpha + \beta X_i + u_i - \alpha - \beta \bar{X} = Y_i - \beta\left(X_i - \bar{X}\right) + u_i$
$y_i = \beta x_i + u_i$
The predicted value of y is
$\hat{y}_i = \hat{\beta} x_i$
$y_i = \hat{y}_i + e_i$
The observed value of y is the sum of the predicted value from the regression and the residual, the difference between the observed value and the predicted value.
$y_i = \hat{\beta} x_i + e_i$
Multiply both sides by x_i,
$x_i y_i = \hat{\beta} x^2 + x_i e_i$
Sum,
$\sum x_i y_i = \hat{\beta} \sum x^2 + \sum x_i e_i$
But
$\sum x_i e_i = 0$
by (a2). So,
$\sum x_i y_i = \hat{\beta} \sum x^2$
Solve for $\hat{\beta}$,
$\hat{\beta} = \frac{\sum x_i y_i}{\sum x^2}$
which is the OLS estimator.

Now that we know the we can derive the OLS estimators using instrumental variables, let's do it for the multiple regression above.

$Y_i = \alpha + \beta X_i + \gamma Z_i + u_i$

We have to make three assumptions (with their empirical analogs), since we have three parameters.

(A1) $E(u_i) = 0 \Rightarrow \frac{\sum e}{N} = 0$

(A2) $E(x_i u_i) = 0 \Rightarrow \frac{\sum x_i e_i}{N} = 0 \Rightarrow \sum x_i e_i = 0$

(A3) $E(z_i u_i) = 0 \Rightarrow \frac{\sum z_i e_i}{N} = 0 \Rightarrow \sum z_i e_i = 0$

From (A1) we get the estimate of the intercept.

$$\sum e_i = \sum Y_i - N\alpha - \beta \sum X_i - \gamma \sum Z_i = 0$$

$$\frac{\sum e_i}{N} = \frac{\sum Y_i}{N} - \alpha - \beta \frac{\sum X_i}{N} - \gamma \frac{\sum Z_i}{N} = 0$$

Or,

$$\bar{Y} - \alpha - \beta \bar{X} - \gamma \bar{Z} = 0$$

$$\hat{\alpha} = \bar{Y} - \beta \bar{X} - \gamma \bar{Z}$$

which is a pretty straightforward generalization of the simple regression estimator for α.

Now we can work in deviations.

$$y_i = Y_i - \bar{Y} = \alpha + \beta X_i + \gamma Z_i + u_i - \alpha - \beta \bar{X} + \gamma \bar{Z}$$

$$y_i = \beta x_i + \gamma z_i + u_i$$

The predicted values of y are

$$\hat{y}_i = \hat{\beta} x_i + \hat{\gamma} z_i$$

So that y is equal to the predicted value plus the residual.

$$y_i = \hat{\beta} x_i + \hat{\gamma} z_i + e_i$$

Multiply by x and sum to get,

$$x_i y_i = \hat{\beta} x_i^2 + \hat{\gamma} x_i z_i + x_i e_i$$

$$\sum x_i y_i = \hat{\beta} \sum x_i^2 + \hat{\gamma} \sum x_i z_i + \sum x_i e_i$$

But, $\sum x_i e_i = 0$ by (A2), so

(N1) $\sum x_i y_i = \hat{\beta} \sum x_i^2 + \hat{\gamma} \sum x_i z_i$

Now repeat the operation with z:

$$z_i y_i = \hat{\beta} z_i x_i + \hat{\gamma} z_i^2 + z_i e_i$$

$$\sum z_i y_i = \hat{\beta} \sum z x_i + \hat{\gamma} \sum z_i^2 + \sum z_i e_i$$

But, $\sum z_i e_i = 0$ by (A3), so

(N2) $\sum z_i y_i = \hat{\beta} \sum z x_i + \hat{\gamma} \sum z_i^2$

We have two linear equations (N1 and N2) in two unknowns, $\hat{\beta}$ and $\hat{\gamma}$. Even though these equations look complicated with all the summations, these summations are just numbers derived from the sample observations on x, y, and z. There are a variety of ways to solve two equations in two unknowns. Back substitution from high school algebra will work just fine. Solve for $\hat{\gamma}$ from N2.

$$\hat{\gamma} \sum z_i^2 = \sum z_i y_i - \hat{\beta} \sum z x_i$$

$$\hat{\gamma} = \frac{\sum z_i y_i}{\sum z_i^2} - \hat{\beta} \frac{\sum z x_i}{\sum z_i^2}$$

Substitute into N1

$$\sum x_i y_i = \hat{\beta} \sum x_{,}^2 + \left(\frac{\sum z_i y_i}{\sum z_i^2} - \hat{\beta} \frac{\sum z x_i}{\sum z_i^2} \right) \sum x_i z_i$$

Multiply through by $\sum z_i^2$

$$\sum x_i y_i \sum z_i^2 = \hat{\beta} \sum x_{,}^2 \sum z_i^2 + \sum z_i^2 \left(\frac{\sum z_i y_i}{\sum z_i^2} - \hat{\beta} \frac{\sum z x_i}{\sum z_i^2} \right) \sum x_i z_i$$

$$\sum x_i y_i \sum z_i^2 = \hat{\beta} \sum x_{,}^2 \sum z_i^2 + \left(\sum z_i y_i - \hat{\beta} \sum z x_i \right) \sum x_i z_i$$

$$\sum x_i y_i \sum z_i^2 = \hat{\beta} \sum x_{,}^2 \sum z_i^2 + \sum z_i y_i \sum x_i z_i - \hat{\beta} \sum z x_i \sum x_i z_i$$

Collect terms in $\hat{\beta}$ and move them to the left hand side.

$$\hat{\beta} \sum x_{,}^2 \sum z_i^2 - \hat{\beta} \sum z x_i \sum x_i z_i = \sum x_i y_i \sum z_i^2 - \sum z_i y_i \sum x_i z_i$$

$$\hat{\beta} \left(\sum x_{,}^2 \sum z_i^2 - \sum z x_i \sum x_i z_i \right) = \sum x_i y_i \sum z_i^2 - \sum z_i y_i \sum x_i z_i$$

$$\hat{\beta} = \frac{\sum x_i y_i \sum z_i^2 - \sum z_i y_i \sum x_i z_i}{\sum x_{,}^2 \sum z_i^2 - \sum z x_i \sum x_i z_i}$$

$$\hat{\beta} = \frac{\sum x_i y_i \sum z_i^2 - \sum z_i y_i \sum x_i z_i}{\sum x_{,}^2 \sum z_i^2 - \left(\sum x_i z_i \right)^2}$$

because $\sum x_i z_i = \sum z_i x_i$.

This is the OLS estimator for $\hat{\beta}$.

The OLS estimator for $\hat{\gamma}$ is derived similarly.

$$\hat{\gamma} = \frac{\sum z_i y_i \sum x_i^2 - \sum x_i y_i \sum x_i z_i}{\sum x_{,}^2 \sum z_i^2 - \left(\sum x_i z_i \right)^2}$$

Interpreting the multiple regression coefficient.

We might be able to make some sense out of these formulas if we translate them into simple regression coefficients. The formula for a simple correlation coefficient between x and y is

$$r_{xy} = \frac{\sum x_i y_i}{N S_x S_y} = \frac{\sum x_i y_i}{N \sqrt{\frac{\sum x_i^2}{N}} \sqrt{\frac{\sum y_i^2}{N}}} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2} \sqrt{\sum y_i^2}}$$

Or,

$$\sum x_i y_i = r_{xy} \sqrt{\sum x_i^2} \sqrt{\sum y_i^2}$$

And,

$$\sum x_i z_i = r_{xz} \sqrt{\sum x_i^2} \sqrt{\sum z_i^2}$$

$$\sum z_i y_i = r_{zy} \sqrt{\sum z_i^2} \sqrt{\sum y_i^2}$$

Substituting into the formula for $\hat{\beta}$

$$\hat{\beta} = \frac{r_{xy} \sqrt{\sum x_i^2} \sqrt{\sum y_i^2} \sum z_i^2 - r_{zy} \sqrt{\sum z_i^2} \sqrt{\sum y_i^2} r_{xz} \sqrt{\sum x_i^2} \sqrt{\sum z_i^2}}{\sum x_{,}^2 \sum z_i^2 - \left(r_{xz} \sqrt{\sum x_i^2} \sqrt{\sum z_i^2} \right)^2}$$

Cancel out $\sum z_i^2$ which appears in every term.

$$\hat{\beta} = \frac{r_{xy}\sqrt{\sum x_i^2}\sqrt{\sum y_i^2} - r_{zy}\sqrt{\sum y_i^2}r_{xz}\sqrt{\sum x_i^2}}{\sum x_{\cdot}^2 - \left(r_{xz}\sqrt{\sum x_i^2}\right)^2} = \frac{r_{xy}\sqrt{\sum x_i^2}\sqrt{\sum y_i^2} - r_{zy}\sqrt{\sum y_i^2}r_{xz}\sqrt{\sum x_i^2}}{\sum x_{\cdot}^2 - r_{xz}^2\sum x_i^2}$$

Factor out the common terms on the top and bottom.

$$\hat{\beta} == \frac{r_{xy} - r_{zy}r_{xz}}{1 - r_{xz}^2}\left(\frac{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}}{\sum x_{\cdot}^2}\right) = \frac{r_{xy} - r_{zy}r_{xz}}{1 - r_{xz}^2}\left(\frac{\sqrt{\sum y_i^2}}{\sqrt{\sum x_i^2}}\right)$$

$$= \frac{r_{xy} - r_{zy}r_{xz}}{1 - r_{xz}^2}\left(\frac{\sqrt{\sum y_i^2/N}}{\sqrt{\sum x_i^2/N}}\right) = \frac{r_{xy} - r_{zy}r_{xz}}{1 - r_{xz}^2}\left(\frac{S_y}{S_x}\right)$$

The corresponding expression for $\hat{\gamma}$ is

$$\hat{\gamma} = \frac{r_{zy} - r_{xy}r_{xz}}{1 - r_{xz}^2}\left(\frac{S_y}{S_z}\right)$$

The term in parentheses is simply a positive scaling factor. Let's concentrate on the ratio. The denominator is $1 - r_{xz}^2$. When x and z are perfectly correlated, $r_{xz} = 1$ and the formula does not compute. This is the case of perfect collinearity. X and Y are said to be collinear because one is an exact function of the other, with no error term. This could happen if X is total wages and Y is total income minus non-wage payments.

The multiple regression coefficient β is the effect on Y of a change in X, holding Z constant. If X and Z are perfectly collinear, then whenever X changes, Z has to change. Therefore it is impossible to hold Z constant to find the separate effect of X. Stata will simply drop Z from the regression and estimate a simple regression of Y on X.

The simple correlation coefficient r_{xy} in the numerator of the formula for $\hat{\beta}$ is the total effect of X on Y, holding nothing constant. The second term, $r_{zy}r_{xz}$ is the effect of X on Y through Z. In other words, X is correlated with Z, so when X varies, Z varies. This causes Y to change because Y is correlated with Z. Therefore, the numerator of the formula for $\hat{\beta}$ starts with the total effect on X on Y, but then subtracts off the indirect effect of X on Y through Z. It therefore measures the effect of X on Y, while statistically holding Z constant, the so-called partial effect of X on Y.

The formula for $\hat{\gamma}$ is analogous. It is the total effect of Z on Y minus the indirect effect of Z on Y through X. It partials out the effect of X on Y.

Note that, if there is no correlation between X and Z, then X and Z are said to be orthogonal and $r_{xz} = 0$. In this case, the multiple regression estimators collapse to simple regression estimators.

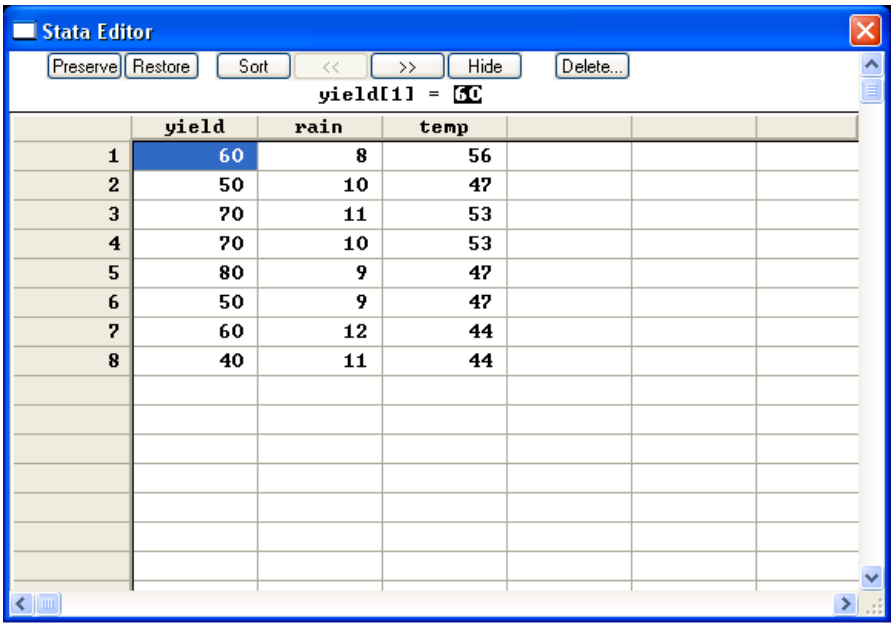
$$\hat{\beta} = \frac{r_{xy} - r_{zy}r_{xz}}{1 - r_{xz}^2}\left(\frac{S_y}{S_x}\right) = r_{xy}\left(\frac{S_y}{S_x}\right)$$

$$\hat{\gamma} = \frac{r_{zy} - r_{xy}r_{xz}}{1 - r_{xz}^2}\left(\frac{S_y}{S_z}\right) = r_{zy}\left(\frac{S_y}{S_z}\right)$$

These are the results we would get if we did two separate simple regressions, Y on X, and Y on Z.

Multiple regression and omitted variable bias

The reason we use multiple regression is that most real world relationships have more than one explanatory variable. Suppose we type the following data into the data editor in Stata.



	yield	rain	temp
1	60	8	56
2	50	10	47
3	70	11	53
4	70	10	53
5	80	9	47
6	50	9	47
7	60	12	44
8	40	11	44

Let's do a simple regression of crop yield on rainfall.

```
regress yield rain
```

```
      Source |       SS       df       MS              Number of obs =       8
-------------+------------------------------           F(  1,     6) =    0.17
       Model |  33.3333333     1  33.3333333           Prob > F      =  0.6932
    Residual |  1166.66667     6  194.444444           R-squared     =  0.0278
-------------+------------------------------           Adj R-squared = -0.1343
       Total |        1200     7  171.428571           Root MSE      =  13.944

------------------------------------------------------------------------------
       yield |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        rain |  -1.666667   4.025382    -0.41   0.693    -11.51642    8.183089
       _cons |   76.66667    40.5546     1.89   0.108    -22.56687    175.9002
------------------------------------------------------------------------------
```

According to our analysis, the coefficient on rain is -1.67 implying that rain causes crop yield to decline (!), although not significantly. The problem is that we have an omitted variable, temperature. Let's do a multiple regression with both rain and temp.

```
regress yield rain temp
```

```
      Source |       SS       df       MS              Number of obs =       8
```

```
-------------+------------------------------           F(  2,     5) =    0.60
       Model |  231.448763     2  115.724382           Prob > F      =  0.5853
    Residual |  968.551237     5  193.710247           R-squared     =  0.1929
-------------+------------------------------           Adj R-squared = -0.1300
       Total |        1200     7  171.428571           Root MSE      =  13.918

------------------------------------------------------------------------------
       yield |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        rain |   .7243816   4.661814     0.16   0.883    -11.25919    12.70796
        temp |   1.366313   1.351038     1.01   0.358    -2.106639    4.839266
       _cons |  -14.02238   98.38747    -0.14   0.892    -266.9354    238.8907
------------------------------------------------------------------------------
```

Now both rain and warmth increase crop yields. The coefficients are not significant, probably because we only have eight observations. Nevertheless, remember that omitting an important explanatory variable can bias your estimates on the included variables.

The omitted variable theorem

The reason we do multiple regressions is that most things in economics are functions of more than one variable. If we make a mistake and leave one of the important variables out, we cause the remaining coefficient to be biased (and inconsistent).

Suppose the true model is the multiple regression (in deviations),

$y_i = \beta x_i + \gamma z_i + u_i$

However, we omit z and run a simple regression of y on x. The coefficient on x will be

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum x_i \left(\beta x_i + \gamma z_i + u_i\right)}{\sum x_i^2}$$

The expected value of $\hat{\beta}$ is,

$$E\left(\hat{\beta}\right) = \frac{\beta \sum x_i^2}{\sum x_i^2} + \frac{\gamma \sum x_i z_i}{\sum x_i^2} + \frac{\sum x_i u_i}{\sum x_i^2}$$

$$= \beta + \gamma \frac{\sum x_i z_i}{\sum x_i^2}$$

Because $\sum x_i u_i = 0$ by A2. (The explanatory variables are uncorrelated with the error term.) We can write the expression for the expected value of $\hat{\beta}$ in a number of ways.

$$E\left(\hat{\beta}\right) = \beta + \gamma \frac{\sum x_i z_i}{\sum x_i^2} = \beta + \gamma \frac{r_{xz}\sqrt{\sum x_i^2}\sqrt{\sum z_i^2}}{\sum x_i^2} = \beta + \gamma r_{xz} \frac{\sqrt{\sum z_i^2 / N}}{\sqrt{\sum x_i^2 / N}}$$

$$= \beta + \gamma r_{xz} \left(\frac{S_z}{S_x}\right) = \beta + \gamma b_{xz}$$

where b_{xz} is the coefficient from the simple regression of Z on X.

Assume that Z is relevant in the sense that γ is not zero. If $\sum x_i z_i \neq 0$ then $r_{xz} \neq 0$ and $b_{xz} \neq 0$ in which case $\hat{\beta}$ is biased (and inconsistent since the bias does not go away as N goes to infinity). The direction of bias depends on the signs of r_{xz} and γ. If r_{xz} and γ. have the same sign, then $\hat{\beta}$ is biased upwards. If they have

opposite signs, then $\hat{\beta}$ is biased downward. This was the case with rainfall and temperature in our crop yield example in Chapter 7. Finally, note that if X and Z are orthogonal, then $r_{xz} = 0$ and $\hat{\beta}$ is unbiased.

The expected value of $\hat{\gamma}$ if we omit X is analogous.

$$E(\gamma) = \gamma + \beta r_{zx}\left(\frac{S_x}{S_z}\right) = \gamma + \beta b_{zx}$$

where b_{zx} is the coefficient from the simple regression of X on Z.

Now let's consider a more complicated model. Suppose the true model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$$

But we estimate

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$

We know we have omitted variable bias, but how much bias is there and in what direction?

Consider a set of auxiliary (and imaginary if we don't have data on X_3 and X_4) regressions of each of the omitted variables on all of the included variables.

$X_{3i} = a_0 + a_1 X_{1i} + a_2 X_{2i}$

$X_{4i} = b_0 + b_1 X_{1i} + b_2 X_{2i}$

Note if there are more included variables, simply add them to the right hand side of each auxiliary regression. If there are more omitted variables, add more equations.

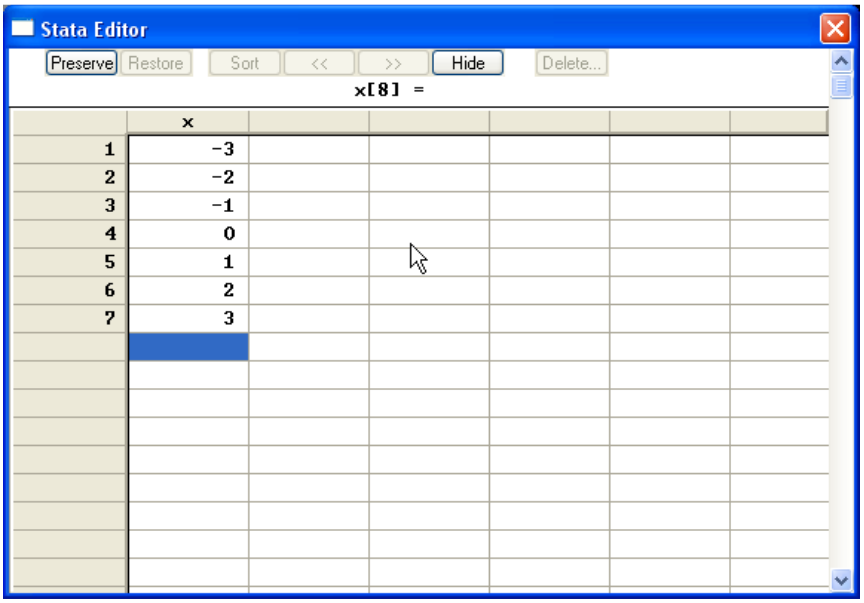
It can be shown that, as a simple extension of the omitted variable theorem above,

$$E\left(\hat{\beta}_0\right) = \beta_0 + a_0\beta_3 + b_0\beta_4$$

$$E\left(\hat{\beta}_1\right) = \beta_1 + a_1\beta_3 + b_1\beta_4$$

$$E\left(\hat{\beta}_2\right) = \beta_2 + a_2\beta_3 + b_2\beta_4$$

Here is a nice example in Stata. In the Data Editor, create the following variable.



Create some additional variables.

```
. gen x2=x^2
. gen x3=x^3
. gen x4=x^4
```

Now create y.

```
. gen y=1+x+x2+x3+x4
```

Regress y on all the x's (true model)

```
. regress y x x2 x3 x4
```

```
      Source |       SS       df       MS              Number of obs =       7
-------------+------------------------------           F(  4,     2) =       .
       Model |       11848     4        2962           Prob > F      =       .
    Residual |           0     2           0           R-squared     =  1.0000
-------------+------------------------------           Adj R-squared =  1.0000
       Total |       11848     6  1974.66667           Root MSE      =       0

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |          1          .        .       .            .           .
          x2 |          1          .        .       .            .           .
          x3 |          1          .        .       .            .           .
          x4 |          1          .        .       .            .           .
       _cons |          1          .        .       .            .           .
------------------------------------------------------------------------------
```

Now omit X3 and X4.

```
. regress y x x2
```

```
      Source |       SS       df       MS              Number of obs =       7
-------------+------------------------------           F(  2,     4) =   33.44
       Model |  11179.4286     2  5589.71429           Prob > F      =  0.0032
    Residual |  668.571429     4  167.142857           R-squared     =  0.9436
-------------+------------------------------           Adj R-squared =  0.9154
       Total |       11848     6  1974.66667           Root MSE      =  12.928

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |           8   2.443233     3.27   0.031     1.216498     14.7835
          x2 |   10.57143   1.410601     7.49   0.002     6.654972    14.48789
       _cons |  -9.285714     7.4642    -1.24   0.281    -30.00966    11.43823
------------------------------------------------------------------------------
```

Do the auxiliary regressions.

```
. regress x3 x x2
```

```
      Source |       SS       df       MS              Number of obs =       7
-------------+------------------------------           F(  2,     4) =   12.70
       Model |        1372     2         686           Prob > F      =  0.0185
    Residual |         216     4          54           R-squared     =  0.8640
-------------+------------------------------           Adj R-squared =  0.7960
       Total |        1588     6  264.666667           Root MSE      =  7.3485

------------------------------------------------------------------------------
          x3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
```

```
-------------+------------------------------------------------------------------
           x |          7    1.38873     5.04   0.007     3.144267    10.85573
          x2 |          0   .8017837     0.00   1.000    -2.226109    2.226109
       _cons |          0   4.242641     0.00   1.000    -11.77946    11.77946
------------------------------------------------------------------------------

. regress x4 x x2

      Source |       SS       df       MS              Number of obs =       7
-------------+------------------------------           F(  2,     4) =   34.01
       Model |  7695.42857     2  3847.71429           Prob > F      =  0.0031
    Residual |  452.571429     4  113.142857           R-squared     =  0.9445
-------------+------------------------------           Adj R-squared =  0.9167
       Total |        8148     6        1358           Root MSE      =  10.637

------------------------------------------------------------------------------
          x4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |           0   2.010178     0.00   1.000    -5.581149    5.581149
          x2 |   9.571429   1.160577     8.25   0.001      6.34915    12.79371
       _cons |  -10.28571   6.141196    -1.67   0.169    -27.33641    6.764979
------------------------------------------------------------------------------
```

Compute the expected values from the formulas above.

$$E\left(\hat{\beta}_0\right) = \beta_0 + a_0\beta_3 + b_0\beta_4 = 1 + 0(1) - 10.29(1) = -9.29$$

$$E\left(\hat{\beta}_1\right) = \beta_1 + a_1\beta_3 + b_1\beta_4 = 1 + 7(1) + 0(1) = 8$$

$$E\left(\hat{\beta}_2\right) = \beta_2 + a_2\beta_3 + b_2\beta_4 = 1 + 0(1) + 9.57(1) = 10.57$$

These are the values of the parameters in the regression with the omitted variables. These numbers are exact because there is no random error in the model. You can see that the theorem works and both estimates are biased upward by the omitted variables.

Target and control variables: how many regressors?

When we estimate a regression model, we usually have one or two parameters that we are primarily interested in. These variables associated with those parameters are called **target variables**. In a demand curve we are typically concerned with the coefficients on price and income. These coefficients tell us if the demand curve downward sloping and whether the good is normal or inferior. We are primarily interested in the coefficient on the interest rate in a demand for money equation. In a policy study the target variable is frequently a dummy variable (see below) that is equal to one if the policy is in effect and zero otherwise.

To get a good estimate of the coefficient on the target variable or variables, we want to avoid omitted variable bias. To that end we include a list of **control variables**. The only purpose of the control variables is to try to guarantee that we don't bias the coefficient on the target variables. For example, if we are studying the effect of three-strikes law on crime, our target variable is the three-strikes dummy. We include in the crime equation all those variables which might cause crime aside from the three-strikes law (e.g., percent urban, unemployment, percent in certain age groups, etc.)

What happens if, in our attempt to avoid omitted variable bias, we include too many control variables? Instead of omitting relevant variables, suppose we include irrelevant variables. It turns out that including

irrelevant variables will not bias the coefficients on the target variables. However, including irrelevant variables will make the estimates inefficient relative to estimates including only relevant variables. It will also increase the standard errors and underestimate the t-ratios on all the coefficients in the model, including the coefficients on the target variables. Thus, including too many control variables will tend to make the target variables appear insignificant, even when they are truly significant.

We can summarize the effect of too many or too few variables as follows.

Omitting a relevant variable biases the coefficients on all the remaining variables, but decreases the variance (increases the efficiency) of all the remaining coefficients.

Discarding a variable whose true coefficient is less than its true (theoretical) standard error decreases the mean square error (the sum of variance plus bias squared) of all the remaining coefficients. So if

$$|\tau| = \left|\frac{\beta}{S_\beta}\right| < 1$$

then we are better off, in terms of mean square error, by omitting the variable, even though its parameter is not zero. Unfortunately, we don't know any of these true values, so this is not very helpful, but it does demonstrate that we shouldn't keep insignificant control variables.

What is the best practice? I recommend the **general to specific** modeling strategy. After doing your library research and reviewing all previous studies on the issue, you will have comprised a list of all the control variables that previous researchers have used. You may also come up with some new control variables. Start with a general model, including all the potentially relevant controls, and remove the insignificant ones. Use t-tests and F-tests to justify your actions. You can proceed sequentially, dropping one or two variables at a time, if that is convenient. After you get down to a parsimonious model including only significant control variables, do one more F-test to make sure that you can go from the general model to the final model in one step. Sets of dummy variables should be treated as groups, including all or none for each group, so that you might have some insignificant controls in the final model, but not a lot. At this point you should be able to do valid hypothesis tests concerning the coefficients on the target variables.

Proxy variables

It frequently happens that researchers face a dilemma. Data on a potentially important control variable is not available. However, we may be able to obtain data on a variable that is known or suspected to be highly correlated with the unavailable variable. Such variables are known as **proxy variables**, or proxies. The dilemma is this: if we omit the proxy we get omitted variable bias, if we include the proxy we get measurement error. As we see in a later chapter, measurement error causes biased and inconsistent estimates, but so does omitting a relevant variable. What is a researcher to do?

Monte Carlo studies have shown that the bias tends to be smaller if we include a proxy than if we omit a variable entirely. However, the bias that results from including a proxy is directly related to how highly correlated the proxy is with the unavailable variable. It is better to omit a poor proxy. Unfortunately, it is impossible to know how highly correlated they are because we can't observe the unavailable variable. It might be possible, in some cases, to see how the two variables are related in other contexts, in other studies for example, but generally we just have to hope.

The bottom line is that we should include proxies for control variables, but drop them if they are not significant.

An interesting problem arises if the proxy variable is the target variable. In that case, we are stuck with measurement error. We are also faced with the fact that the coefficient estimate is a compound of the true

parameter linking the dependent variable and the unavailable variable and the parameter relating the proxy to the unavailable variable.

Suppose we know that
$Y = \alpha + \beta X$
But we can't get data on X. So we use Z, which is available and is related to X.
$X = a + bZ$
Substituting for X in the original model yields,
$Y = \alpha + \beta X = \alpha + \beta(a + bZ) = (\alpha + \beta a) + \beta bZ$
So if we estimate the model,
$Y = c + dZ$
the coefficient on Z is $d = \beta b$, the product of the true coefficient relating X to Y and the coefficient relating X to Z. Unless we know the value of b, we have no measure of the effect of X on Y.

However, we do know if the coefficient d is significant or not and we do know if the sign is as expected. That is, if we know that b is positive, then we can test the hypothesis that β is positive using the estimated coefficient, $\hat{d}$. Therefore, if the target variable is a proxy variable, the estimated coefficient can only be used to determine sign and significance.

An illustration of this problem occurs in studies of the relationship between guns and crime. There is no good measure of the number of guns, so researchers have to use proxies. As we have just seen, the coefficient on the proxy for guns cannot be used to make inferences concerning the elasticity of crime with respect to guns. Nevertheless two studies, one by Cook and Ludwig and one by Duggan, both make this mistake. See http://papers.ssrn.com/sol3/papers.cfm?abstract_id=473661 for a discussion of this issue.

Dummy variables

Dummy variables are variables that consist of one's and zeroes. They are also known as binary variables. They are extremely useful. For example, I happen to have a data set consisting of the salaries of the faculty of a certain nameless university (salaries.dta). The average salary at the time of the survey was as follows.

```
. summarize salary

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      salary |       350    72024.92    25213.02      20000     173000
```

Now suppose we create a dummy variable called female, that takes the value 1 if the faculty member is female and 0 if male. We can then find the average female salary.

```
. summarize salary if female

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      salary |       122       61050    18195.34      29700     130000
```

Similarly, we can create a dummy variable called male which is 1 if male and zero if female. The average male salary is somewhat higher.

```
. summarize salary if male

    Variable |       Obs        Mean    Std. Dev.       Min        Max
```

```
-------------+--------------------------------------------------------
      salary |       228    77897.46    26485.87      20000     173000
```

The difference between these two means is substantial. We can use the scalar command to find the difference in salary between males and females.

```
. scalar salarydif=61050-77897

. scalar list salarydif
 salarydif =     -16847
```

So, women earn almost $17,000 less than men on average. Is this difference significant given the variance in salary? Let's do a traditional test of the difference between two means by computing the variances and using the formula in most introductory statistics books. First, create two variables for men's salaries and women's salaries.

```
. gen salaryfem=salary if female==1
(228 missing values generated)

. gen salarymale=salary if female==0
(122 missing values generated)


.summarize salary*

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      salary |       350    72024.92    25213.02      20000     173000
   salaryfem |       122       61050    18195.34      29700     130000
  salarymale |       228    77897.46    26485.87      20000     173000
```

Now test for the difference between these two means using the Stata command ttest.

```
. ttest salarymale=salaryfem, unpaired

Two-sample t test with equal variances

------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
salary~e |     228    77897.46    1754.069    26485.87    74441.12     81353.8
salary~m |     122       61050    1647.328    18195.34    57788.68    64311.32
---------+--------------------------------------------------------------------
combined |     350    72024.92    1347.693    25213.02     69374.3    74675.54
---------+--------------------------------------------------------------------
    diff |            16847.46    2684.423                11567.73     22127.2
------------------------------------------------------------------------------
Degrees of freedom: 348

             Ho: mean(salarymale) - mean(salaryfem) = diff = 0

     Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
       t =   6.2760                 t =   6.2760                 t =   6.2760
   P < t =   1.0000             P > |t| =   0.0000             P > t =   0.0000
```

There appears to be a significant difference between male and female salaries.

It is somewhat more elegant to use regression analysis with dummy variables to achieve the same goal.

```
. regress salary female
```

```
      Source |       SS       df       MS              Number of obs =     350
-------------+------------------------------           F(  1,   348) =   39.39
       Model |  2.2558e+10     1  2.2558e+10           Prob > F      =  0.0000
    Residual |  1.9930e+11   348   572701922           R-squared     =  0.1017
-------------+------------------------------           Adj R-squared =  0.0991
       Total |  2.2186e+11   349   635696291           Root MSE      =   23931

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -16847.46   2684.423    -6.28   0.000     -22127.2   -11567.73
       _cons |   77897.46   1584.882    49.15   0.000     74780.31    81014.61
------------------------------------------------------------------------------
```

Note that the coefficient on female is the difference in average salary attributable to being female while the intercept is the corresponding male salary. The t-ratio on female tests the null hypothesis that this salary difference is equal to zero. This is exactly the same results we got using the standard t-test of the difference between two means. So, there appears to be significant salary discrimination against women at this university.

Since male=1-female, we can generate the bonus attributable to being male by doing the following regression.

```
. regress salary male
```

```
      Source |       SS       df       MS              Number of obs =     350
-------------+------------------------------           F(  1,   348) =   39.39
       Model |  2.2558e+10     1  2.2558e+10           Prob > F      =  0.0000
    Residual |  1.9930e+11   348   572701922           R-squared     =  0.1017
-------------+------------------------------           Adj R-squared =  0.0991
       Total |  2.2186e+11   349   635696291           Root MSE      =   23931

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   16847.46   2684.423     6.28   0.000     11567.73     22127.2
       _cons |      61050   2166.628    28.18   0.000     56788.67    65311.33
------------------------------------------------------------------------------
```

Now the female salary is captured by the intercept and the coefficient on male is the (significant) bonus that attends maleness.

Since male=1-female, female=1-male, and the intercept is simply 1, male+female=intercept, so we can't use all three terms in the same regression. This is the **dummy variable trap**.

```
regress salary male female
```

```
      Source |       SS       df       MS              Number of obs =     350
-------------+------------------------------           F(  1,   348) =   39.39
       Model |  2.2558e+10     1  2.2558e+10           Prob > F      =  0.0000
    Residual |  1.9930e+11   348   572701922           R-squared     =  0.1017
-------------+------------------------------           Adj R-squared =  0.0991
       Total |  2.2186e+11   349   635696291           Root MSE      =   23931

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   16847.46   2684.423     6.28   0.000     11567.73     22127.2
```

```
      female |  (dropped)
       _cons |       61050   2166.628    28.18   0.000     56788.67    65311.33
------------------------------------------------------------------------------
```

Stata has saved us the embarrassment of pointing out that the regression we asked it to run is impossible by simply dropping one of the variables from the regression. The result is we get the same regression we got with male as the only independent variable.

It is possible to force Stata to drop the intercept term instead of one of the other variables,

```
. regress salary male female, noconstant

      Source |       SS       df       MS              Number of obs =     350
-------------+------------------------------           F(  2,   348) = 1604.86
       Model |  1.8382e+12     2  9.1911e+11           Prob > F      =  0.0000
    Residual |  1.9930e+11   348   572701922           R-squared     =  0.9022
-------------+------------------------------           Adj R-squared =  0.9016
       Total |  2.0375e+12   350  5.8215e+09           Root MSE      =   23931

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   77897.46   1584.882    49.15   0.000     74780.31    81014.61
      female |      61050   2166.628    28.18   0.000     56788.67    65311.33
------------------------------------------------------------------------------
```

Now the coefficients are the mean salaries for the two groups. This formulation is less useful because it is more difficult to test the null hypothesis that the two salaries are different.

Perhaps this salary difference is due to a difference in the amount of experience of the two groups. Maybe the men have more experience and that is what is causing the apparent salary discrimination. If so, the previous analysis suffers from omitted variable bias.

```
. regress salary exp female

      Source |       SS       df       MS              Number of obs =     350
-------------+------------------------------           F(  2,   347) =   79.49
       Model |  6.9709e+10     2  3.4854e+10           Prob > F      =  0.0000
    Residual |  1.5215e+11   347   438471159           R-squared     =  0.3142
-------------+------------------------------           Adj R-squared =  0.3103
       Total |  2.2186e+11   349   635696291           Root MSE      =   20940

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   1069.776   103.1619    10.37   0.000     866.8753    1272.678
      female |  -10310.35   2431.983    -4.24   0.000    -15093.63   -5527.065
       _cons |   60846.72   2150.975    28.29   0.000     56616.14    65077.31
------------------------------------------------------------------------------
```

Well, the difference has been reduced to $10K, so the men are apparently somewhat older. But it is still significant and negative. Also, the average raise is a pitiful $1000 per year. It is also possible that women receive lower raises than men do. We can test this hypothesis by creating an interaction variable by multiplying experience by female and including it as an additional regressor.

```
. gen fexp=female*exp
. regress salary exp female fexp

      Source |       SS       df       MS              Number of obs =     350
-------------+------------------------------           F(  3,   346) =   53.04
       Model |  6.9888e+10     3  2.3296e+10           Prob > F      =  0.0000
    Residual |  1.5197e+11   346   439219336           R-squared     =  0.3150
```

```
-------------+------------------------------           Adj R-squared =  0.3091
       Total |  2.2186e+11   349   635696291           Root MSE      =   20958

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   1038.895   113.9854     9.11   0.000     814.7038    1263.087
      female |  -12189.86   3816.229    -3.19   0.002    -19695.79   -4683.936
        fexp |   172.0422   269.0422     0.64   0.523    -357.1217    701.2061
       _cons |   61338.93   2286.273    26.83   0.000     56842.18    65835.67
------------------------------------------------------------------------------
```

In fact, women appear to receive slightly higher, but not significantly higher, raises than men. There is a significant salary penalty associated with being female, but it is not caused by discrimination in raises.

Useful tests

F-test

We have seen many applications of the t-test. However, the F-test is also extremely useful. This test was developed by Sir Ronald Fisher (the F in F-test), a British statistical biologist in the early 1900's. Suppose we want to test the null hypothesis that the coefficients on female and fexp are jointly equal to zero. The way to test this hypothesis is to run the regression as it appears above with both female and fexp included, and note the residual sum of squares (1.5e+11). Then run the regression again without the two variables (assuming the null hypothesis is true) and seeing if the two residual sum of squares are significantly different. If they are, then the null hypothesis is false. If they are not significantly different, then the two variables do not help explain the variance of the dependent variable and the null hypothesis is true (cannot be rejected).

Stata allows us to do this test very easily with the test command. This command is used after the regress command. Refer to the coefficients by the corresponding variable name.

```
. test female fexp

 ( 1)  female = 0
 ( 2)  fexp = 0

       F(  2,   346) =    9.18
            Prob > F =    0.0001
```

This tests the null hypothesis that the coefficient on female and the coefficient on male are both equal to zero. According the F-ratio, we can firmly reject this hypothesis. The numbers in parentheses in the F ratio are the degrees of freedom of the numerator (2 because there are two restrictions, one on female and one on fexp) and the degrees of freedom of the denominator (n-k: 350-4=346).

Chow test

This test is frequently referred to as a Chow test, after Gregory Chow, a Princeton econometrician who developed a slightly different version. Let's do a slightly more complicated version of the F-test above. I also have a dummy variable called "admin" which is equal to one whenever a faculty member is also a member of the administration (assistant to the President, assistant to the assistant to the President, dean, associate dean, assistant dean, deanlet, etc., etc.). Maybe female administrators are discriminated against in

terms of salary. We want to know if the regression model explaining women's salaries is different from the regression model explaining men's salaries. This is a Chow test.

```
.gen fadmin=female*admin

. regress salary exp female fexp admin fadmin

      Source |       SS       df       MS              Number of obs =     350
-------------+------------------------------           F(  5,   344) =   35.33
       Model |  7.5271e+10     5  1.5054e+10           Prob > F      =  0.0000
    Residual |  1.4659e+11   344   426123783           R-squared     =  0.3393
-------------+------------------------------           Adj R-squared =  0.3297
       Total |  2.2186e+11   349   635696291           Root MSE      =   20643

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   1005.451    112.843     8.91   0.000     783.5015      1227.4
      female |  -11682.84    3791.92    -3.08   0.002    -19141.11   -4224.577
        fexp |   156.5447   266.3948     0.59   0.557     -367.423    680.5124
       admin |   11533.69   3905.342     2.95   0.003     3852.334    19215.04
      fadmin |   2011.575   7884.295     0.26   0.799    -13495.92    17519.07
       _cons |   60202.64   2284.564    26.35   0.000     55709.17    64696.11
------------------------------------------------------------------------------

. test female fexp fadmin

 ( 1)  female = 0
 ( 2)  fexp = 0
 ( 3)  fadmin = 0

       F(  3,   344) =    5.67
            Prob > F =    0.0008
```

We can soundly reject the null hypothesis that women have the same salary function as men. However, because the t-tests on fexp and fadmin are not significant, we know that the difference is not due to raises or differences paid to women administrators.

I also have data on a number of other variables that might explain the difference in salaries between men and women faculty: chprof (=1 if a chancellor professor), asst (=1 if an assistant prof), assoc (=1 if an associate prof), and prof (=1 if a full professor). Let's see if these variables make a difference.

```
. regress salary exp female asst assoc prof admin chprof

      Source |       SS       df       MS              Number of obs =     350
-------------+------------------------------           F(  7,   342) =   72.14
       Model |  1.3227e+11     7  1.8896e+10           Prob > F      =  0.0000
    Residual |  8.9587e+10   342   261948866           R-squared     =  0.5962
-------------+------------------------------           Adj R-squared =  0.5879
       Total |  2.2186e+11   349   635696291           Root MSE      =   16185

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   195.7694   102.5207     1.91   0.057    -5.881204      397.42
      female |  -5090.278   1952.174    -2.61   0.010    -8930.056    -1250.5
        asst |   11182.61    4175.19     2.68   0.008     2970.325    19394.89
       assoc |   22847.23   4079.932     5.60   0.000     14822.31    30872.15
        prof |   47858.19   4395.737    10.89   0.000     39212.11    56504.28
       admin |   8043.421    2719.69     2.96   0.003     2693.997    13392.85
      chprof |   14419.67   5962.966     2.42   0.016     2690.965    26148.37
       _cons |   42395.53   4042.146    10.49   0.000     34444.94    50346.13
```

Despite all these highly significant control variables, the coefficient on the target variable, female is negative and significant, indicating significant salary discrimination against women faculty. I have one more set of dummy variables to try. I have a dummy for each department. So, dept1=1 if the faculty person is in the first department, zero otherwise, dept2=1 if in the second department, etc.

```
. regress salary exp female admin chprof prof assoc asst dept2-dept30

      Source |       SS       df       MS              Number of obs =     350
-------------+------------------------------           F( 33,   316) =   20.18
       Model |  1.5047e+11    33  4.5597e+09           Prob > F      =  0.0000
    Residual |  7.1388e+10   316   225910511           R-squared     =  0.6782
-------------+------------------------------           Adj R-squared =  0.6446
       Total |  2.2186e+11   349   635696291           Root MSE      =   15030

------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   283.8026   99.07088     2.86   0.004     88.88067    478.7245
      female |  -2932.928   1915.524    -1.53   0.127     -6701.72     835.865
       admin |   6767.088   2650.416     2.55   0.011     1552.396    11981.78
      chprof |    14794.4   5622.456     2.63   0.009     3732.222    25856.58
        prof |   46121.73   4324.557    10.67   0.000     37613.17    54630.29
       assoc |    22830.1   3930.037     5.81   0.000     15097.75    30562.44
        asst |   11771.36   4036.021     2.92   0.004     3830.495    19712.23
       dept2 |   2500.519   4457.494     0.56   0.575    -6269.598    11270.63
       dept3 |   19527.87   4019.092     4.86   0.000     11620.31    27435.43
       dept4 |     1278.9   4886.804     0.26   0.794    -8335.885    10893.68
       dept5 |   9213.681   9271.124     0.99   0.321    -9027.251    27454.61
       dept6 |   -1969.66   3659.108    -0.54   0.591    -9168.953    5229.633
       dept7 |   3525.149   4881.219     0.72   0.471    -6078.646    13128.94
       dept8 |   462.3493   4317.267     0.11   0.915    -8031.871     8956.57
       dept9 |   3304.275   6448.824     0.51   0.609    -9383.782    15992.33
      dept10 |   24647.11   4618.156     5.34   0.000     15560.89    33733.33
      dept11 |   13185.66   3726.702     3.54   0.000      5853.38    20517.95
      dept12 |  -3601.157   3230.741    -1.11   0.266    -9957.638    2755.323
      dept14 |   167.8798   5985.794     0.03   0.978    -11609.17    11944.93
      dept16 |  -4854.208   6978.675    -0.70   0.487    -18584.75    8876.331
      dept17 |  -1185.402   4293.443    -0.28   0.783    -9632.748    7261.945
      dept18 |    5932.65   3473.475     1.71   0.089     -901.411    12766.71
      dept19 |   18642.71   15803.41     1.18   0.239    -12450.48     49735.9
      dept20 |   8166.382   8906.207     0.92   0.360    -9356.576    25689.34
      dept21 |  -1839.744   5208.921    -0.35   0.724    -12088.29    8408.806
      dept22 |   2940.911   11064.36     0.27   0.791    -18828.21    24710.03
      dept23 |  -1540.594    4902.64    -0.31   0.754    -11186.54    8105.348
      dept24 |  -777.8447   3916.975    -0.20   0.843    -8484.491    6928.802
      dept25 |  -14089.64   15411.33    -0.91   0.361    -44411.41    16232.14
      dept27 |  -1893.797   3397.766    -0.56   0.578      -8578.9    4791.307
      dept28 |  -4073.063   5408.214    -0.75   0.452    -14713.72    6567.596
      dept29 |     2279.7   10822.38     0.21   0.833    -19013.33    23572.73
      dept30 |  -80.38544   4879.813    -0.02   0.987    -9681.414    9520.643
       _cons |   38501.85   4173.332     9.23   0.000     30290.82    46712.88
------------------------------------------------------------------------------
```

Whoops, the coefficient on female isn't significant any more. Maybe women tend to be over-represented in departments that pay lower salaries to everyone, males and females (e.g., modern language, English, theatre, dance, classical studies, kinesiology). This conclusion rests on the department dummy variables being significant as a group. We can test this hypothesis, using an F-test, with the testparm command. We would like to list the variables as dept1-dept29, but that looks like we want to test hypotheses on the

difference between the coefficient on dept1 and the coefficient on dept29. That is where the testparm command comes in.

```
. testparm dept2-dept30

 ( 1)  dept2 = 0
 ( 2)  dept3 = 0
 ( 3)  dept4 = 0
 ( 4)  dept5 = 0
 ( 5)  dept6 = 0
 ( 6)  dept7 = 0
 ( 7)  dept8 = 0
 ( 8)  dept9 = 0
 ( 9)  dept10 = 0
 (10)  dept11 = 0
 (11)  dept12 = 0
 (12)  dept14 = 0
 (13)  dept16 = 0
 (14)  dept17 = 0
 (15)  dept18 = 0
 (16)  dept19 = 0
 (17)  dept20 = 0
 (18)  dept21 = 0
 (19)  dept22 = 0
 (20)  dept23 = 0
 (21)  dept24 = 0
 (22)  dept25 = 0
 (23)  dept27 = 0
 (24)  dept28 = 0
 (25)  dept29 = 0
 (26)  dept30 = 0

       F( 26,   316) =    3.10
            Prob > F =    0.0000
```

The department dummies are highly significant as a group. So there is no significant salary discrimination against women, once we control for field of study. Female physicists, economists, chemists, computer scientists, etc. apparently earn just as much as their male counterparts. Unfortunately, the same is true for female French teachers and phys ed instructors. By the way, the departments that are significantly overpaid are dept 3 (Chemistry), dept 10 (Computer Science), and dept 11 (Economics).

It is important to remember that, when testing groups of dummy variables for significance, that you must drop or retain all members of the group. If the group is not significant, even if one or two are significant, you should drop them all. If the group is significant, even if only a few are, then all should be retained.

Granger causality test

Another very useful test is available only for time series. In 1969 C.W.J. Granger published an article suggesting a way to test causality.[2] Suppose we are interesting in testing whether crime causes prison (that is, people in prison), prison causes (deters) crime, both, or neither. Granger suggested running a regression of prison on lagged prison and lagged crime. If lagged crime is significant, then crime causes prison. Then do a regression of crime on lagged crime and lagged prison. If lagged prison is significant, then prison

[2] Granger, C.W.J., "Investigating causal relations by econometric models and cross-spectral methods," Econometrica 36, 1969, 424-438.

causes crime (deters crime if the coefficients sum to a negative number, causes crime if the coefficients sum to a positive number).

I have data on crime (crmaj, major crime, the kind you get sent to prison for: murder, rape, robbery, assault, and burglary) and prison population per capita for Virginia from 1972 to 1999 (crimeva.dta).

```
. tsset year
```

(You have to tsset year to tell Stata that year is the time variable.) Let's use two lags.

```
. regress crmaj L.crmaj LL.crmaj L.prison LL.prison

      Source |       SS       df       MS              Number of obs =      24
-------------+------------------------------           F(  4,    19) =   31.36
       Model |  63231512.2     4  15807878.1           Prob > F      =  0.0000
    Residual |  9578520.75    19  504132.671           R-squared     =  0.8684
-------------+------------------------------           Adj R-squared =  0.8407
       Total |    72810033    23  3165653.61           Root MSE      =  710.02

------------------------------------------------------------------------------
crmaj        |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
crmaj        |
         L1 |   1.003715   .2095943     4.79   0.000     .5650287    1.442401
         L2 |  -.4399737    .206103    -2.13   0.046    -.8713522   -.0085951
prison       |
         L1 |   1087.249   1324.262     0.82   0.422    -1684.463    3858.962
         L2 |  -1772.896   1287.003    -1.38   0.184    -4466.625    920.8333
_cons        |   6309.665    2689.46     2.35   0.030     680.5613    11938.77
------------------------------------------------------------------------------

. test L.prison LL.prison

 ( 1)  L.prison = 0
 ( 2)  L2.prison = 0

       F(  2,    19) =    3.61
            Prob > F =    0.0468
```

So, lagged prison is significant in the crime equation. Does prison cause or deter crime? The sum of the coefficients is negative, so it deters crime if the sum is significantly different from zero.

```
. test L.prison+L2.prison=0

 ( 1)  L.prison + L2.prison = 0

       F(  1,    19) =    5.26
            Prob > F =    0.0333
```

So, prison deters crime. Does crime cause prison?

```
regress prison L.prison LL.prison L.crmaj LL.crmaj

      Source |       SS       df       MS              Number of obs =      24
-------------+------------------------------           F(  4,    19) =  510.43
       Model |  23.1503896     4  5.78759741           Prob > F      =  0.0000
    Residual |  .215433065    19  .011338582           R-squared     =  0.9908
-------------+------------------------------           Adj R-squared =  0.9888
       Total |  23.3658227    23  1.01590533           Root MSE      =  .10648
```

```
------------------------------------------------------------------------------
prison       |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
prison       |
          L1 |   1.545344   .1986008     7.78   0.000     1.129668     1.96102
          L2 |  -.5637167    .193013    -2.92   0.009    -.9676977   -.1597358
crmaj        |
          L1 |  -.0000191   .0000314    -0.61   0.551    -.0000849    .0000467
          L2 |   .0000159   .0000309     0.52   0.612    -.0000488    .0000806
_cons        |    .133937   .4033407     0.33   0.743    -.7102647    .9781387
------------------------------------------------------------------------------

. test L.crmaj LL.crmaj

 ( 1)  L.crmaj = 0
 ( 2)  L2.crmaj = 0

       F(  2,    19) =    0.20
            Prob > F =    0.8237

.
```

Apparently, prison deters crime, but crime does not cause prisons, at least in Virginia.

The Granger causality test is best done with control variables included in the regression as well as the lags of the target variables. However, if the control variables are unavailable or the analysis is just preliminary, then the lagged dependent variables will serve as proxies for the omitted control variables.

J-test for non-nested hypotheses

Suppose we are having a debate about crime with some sociologists. We believe that crime is deterred by punishment, mainly imprisonment, while the sociologists think that putting people in prison does no good at all. The sociologists think that crime is caused by income inequality. We don't think so. We both agree that unemployment and income have an effect on crime. How can we resolve this debate? We don't put income inequality in our regressions and they don't put prison in theirs. The key is to create a supermodel with all of the variables included. We then test all the coefficients. If prison is significant and income inequality is not, then we win. If the reverse is true then the sociologists win. If neither or both are significant, then it is a tie. You can get ties in tests of non-nested hypotheses. You don't get ties in tests of nested hypotheses. By the way, the J in J-test is for joint hypothesis.

I have time series data for the US which has data on major crime (murder, rape, robbery, assault, and burglary), real per capita income, the unemployment rate, and income inequality measured by the Gini coefficient from 1947-1998 (gini.csv). The Gini coefficient is a number between zero and one. Zero means perfect equality, everyone has the same level of income. One means perfect inequality, one person has all the money, everyone else has nothing. We create the supermodel by regressing the log of major crime per capita on all these variables.

```
. regress lcrmajpc prisonpc unrate income gini

      Source |       SS       df       MS              Number of obs =      52
-------------+------------------------------           F(  4,    47) = 3286.33
       Model |  1.55121091     4  .387802727            Prob > F      =  0.0000
    Residual |  .005546219    47  .000118005            R-squared     =  0.9964
-------------+------------------------------           Adj R-squared =  0.9961
       Total |  1.55675713    51   .03052465            Root MSE      =  .01086

------------------------------------------------------------------------------
```

```
    lcrmajpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    prisonpc |   .0186871   .0044176     4.23   0.000     .0098001    .0275741
      unrate |   .0391091   .0072343     5.41   0.000     .0245556    .0536626
      income |   .3201013   .0061851    51.75   0.000     .3076585    .3325442
        gini |   -.757686   .2507801    -3.02   0.004     -1.26219   -.2531816
       _cons |   5.200647    .175871    29.57   0.000      4.84684    5.554454
-------------r----------------------------------------------------------------
```

Well, both prison and gini are significant in the crime equation. It appears to be a draw. However, the coefficient on gini is negative, indicating that, as the gini coefficient goes up (more inequality, less equality) crime goes down. Whoops.

If there were more than one variable for each of the competing models (e.g., prison and arrests for the economists' model and giini and the divorce rate for the sociologists) then we would use F-tests rather than t-tests. We would test the joint significance of prison and arrests versus the joint significance of gini and divorce with F-tests.

LM test

Suppose we want to test the joint hypothesis that rtpi and unrate are not significant in the above regression, even if we already know that they are highly significant. We know we can test that hypothesis with an F-test. However, an alternative is to use the LM (for Lagrangian Multiplier, never mind) test. The primary advantage of the LM test is that you only need to estimate the main equation once. Under the F-test you have to estimate the model with all the variables, record the error sum of squares, then estimate the model with the variables being tested excluded, record the error sum of squares, and finally compute the F-ratio.

For example, use the crime1990.dta data set to estimate the following crime equation.

```
. regress lcrmaj prison metpct p1824 p2534

      Source |       SS       df       MS              Number of obs =      51
-------------+------------------------------           F(  4,    46) =   19.52
       Model |  5.32840267     4  1.33210067            Prob > F      =  0.0000
    Residual |   3.1393668    46  .068247104            R-squared     =  0.6293
-------------+------------------------------           Adj R-squared =  0.5970
       Total |  8.46776947    50  .169355389            Root MSE      =  .26124

------------------------------------------------------------------------------
      lcrmaj |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      prison |   .1439967   .0282658     5.09   0.000     .0871006    .2008929
      metpct |   .0086418   .0020265     4.26   0.000     .0045627    .0127208
       p1824 |  -.9140642   5.208719    -0.18   0.861    -11.39867    9.570543
       p2534 |  -1.604708   3.935938    -0.41   0.685    -9.527341    6.317924
       _cons |   9.092154    .673677    13.50   0.000     7.736113     10.4482
------------------------------------------------------------------------------
```

Suppose we want to know if the age groups are jointly significant. The F-test is easy in Stata.

```
. test p1824 p2534

 ( 1)  p1824 = 0
 ( 2)  p2534 = 0

       F(  2,    46) =    0.12
            Prob > F =    0.8834
```

To do the LM test for the same hypothesis, estimate the model without the age group variables and save the residuals using the predict command.

```
. regress lcrmaj prison metpct

      Source |       SS       df       MS              Number of obs =      51
-------------+------------------------------           F(  2,    48) =   40.39
       Model |  5.31143197     2  2.65571598           Prob > F      =  0.0000
    Residual |   3.1563375    48  .065757031           R-squared     =  0.6273
-------------+------------------------------           Adj R-squared =  0.6117
       Total |  8.46776947    50  .169355389           Root MSE      =  .25643

------------------------------------------------------------------------------
      lcrmaj |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      prison |   .1384944   .0248865     5.57   0.000     .0884566    .1885321
      metpct |   .0081552   .0017355     4.70   0.000     .0046656    .0116447
       _cons |   8.767611   .1142346    76.75   0.000     8.537927    8.997295
------------------------------------------------------------------------------
. predict e, resid
```

Now regress the residuals on all the variables including the ones we left out. If they are truly irrelevant, they will not be significant and the R-square for the regression will be approximately zero. The LM statistic is N times the R-square, which is distributed as chi-square with, in this case, two degrees of freedom because there are two restrictions (two variables left out).

```
. regress e prison metpct p1824 p2534

      Source |       SS       df       MS              Number of obs =      51
-------------+------------------------------           F(  4,    46) =    0.06
       Model |  .016970705     4  .004242676           Prob > F      =  0.9926
    Residual |  3.13936681    46  .068247105           R-squared     =  0.0054
-------------+------------------------------           Adj R-squared = -0.0811
       Total |  3.15633751    50   .06312675           Root MSE      =  .26124

------------------------------------------------------------------------------
           e |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      prison |   .0055023   .0282658     0.19   0.847    -.0513938    .0623985
      metpct |   .0004866   .0020265     0.24   0.811    -.0035924    .0045657
       p1824 |  -.9140643   5.208719    -0.18   0.861    -11.39867    9.570543
       p2534 |  -1.604708   3.935938    -0.41   0.685    -9.527341    6.317924
       _cons |   .3245431    .673677     0.48   0.632    -1.031498    1.680585
------------------------------------------------------------------------------
```

To compute the LM test statistic, we need to get the R-square and the number of observations. (Although the fact that both variables have very low t-ratios is a clue.)

Whenever we run regress Stata creates a bunch of output that is available for further analysis. This output goes under the generic heading of “e” output. To see this output, use the ereturn list command.

```
. ereturn list

scalars:
                  e(N) =  51
              e(df_m) =  4
              e(df_r) =  46
                 e(F) =  .0621663918252367
                e(r2) =  .0053767079385045
              e(rmse) =  .2612414678691742
               e(mss) =  .0169707049657037
               e(rss) =  3.139366808584275
              e(r2_a) =  -.0811122739798864
                e(ll) =  -1.276850278809347
```

```
              e(ll_0) =  -1.414326247391365

macros:
            e(depvar) : "e"
               e(cmd) : "regress"
           e(predict) : "regres_p"
             e(model) : "ols"

matrices:
                 e(b) :  1 x 5
                 e(V) :  5 x 5

functions:
            e(sample)
```

Since the numbers we want are scalars, not variables, we need the scalar command to create the test statistic.

```
. scalar n=e(N)
. scalar R2=e(r2)
. scalar LM=n*R2
```

Now that we have the test statistic, we need to see if it is significant. We can compute the prob-value using Stata's chi2 (chi-square) function.

```
. scalar prob=1-chi2(2,LM)
. scalar list
      prob =  .87187776
        LM =   .2742121
        R2 =  .00537671
         n =         51
```

The R-square is almost equal to zero, so the LM statistic is very low. The prob-value is .87. We cannot reject the null hypothesis that the age groups are not related to crime. OK, so the F-test is easier. Nevertheless, we will use the LM test later, especially for diagnostic testing of regression models.

One modification of this test that is easier to apply in Stata is the so-called F-form of the LM test which corrects for degrees of freedom and therefore has somewhat better small sample properties.[3] To do the F-form, just apply the F-test to the auxiliary regression. That is, after regressing the residuals on all the variables, including the age group variables, use Stata's test function to test the joint hypothesis that the coefficients on both variables are zero.

```
. test p1824 p2534

 ( 1)  p1824 = 0
 ( 2)  p2534 = 0

       F(  2,    46) =    0.12
            Prob > F =    0.8834
```

The results are identical to the original F-test above showing that the LM test is equivalent to the F-test. The results are also similar to the nR-square version of the test, as expected. I routinely use the F-form of the LM test because it is easy to do in Stata and corrects for degrees of freedom.

[3] See Kiviet, J.F., *Testing Linear Econometric Models*, Amsterdam: University of Amsterdam, 1987.

9 REGRESSION DIAGNOSTICS

Influential observations

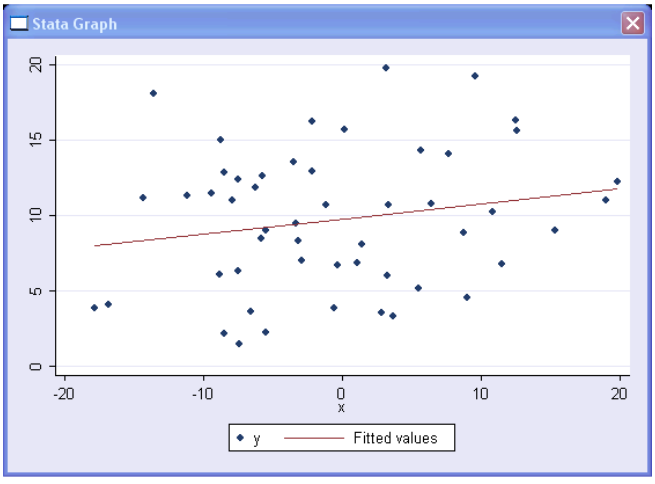
It would be unfortunate if an outlier dominated our regression in such a way that one observation determined the entire regression result.

```
. gen y=10 + 5*invnorm(uniform())
. gen x=10*invnorm(uniform())
. summarize y x
```

Variable	Obs	Mean	Std. Dev.	Min	Max
y	51	9.736584	4.667585	1.497518	19.76657
x	51	-.6021098	8.954225	-17.87513	19.83905

Let's graph these data. We know that they are completely independent, so there should not be a significant relationship between them.

```
. twoway (scatter y x) (lfit y x)
```

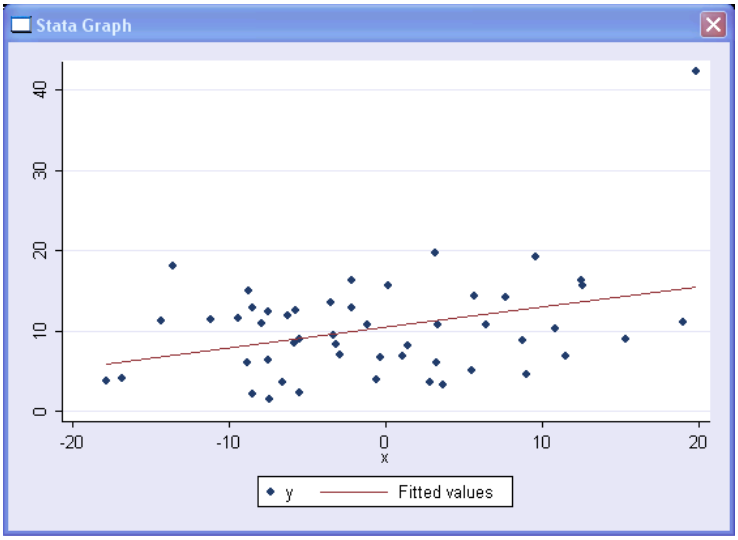


The regression confirms no significant relationship.

```
. regress y x

      Source |       SS       df       MS              Number of obs =      51
-------------+------------------------------           F(  1,    49) =    1.94
       Model |  41.4133222     1  41.4133222           Prob > F      =  0.1703
    Residual |  1047.90398    49  21.3857955           R-squared     =  0.0380
-------------+------------------------------           Adj R-squared =  0.0184
       Total |   1089.3173    50   21.786346           Root MSE      =  4.6245

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .1016382   .0730381     1.39   0.170    -.0451374    .2484138
       _cons |   9.797781    .649048    15.10   0.000      8.49347    11.10209
------------------------------------------------------------------------------
```

Now let's create an outlier.

```
. replace y=y+30 if state==30
(1 real change made)

. twoway (scatter y x) (lfit y x)
```



Let's try that regression again.

```
. regress y x

      Source |       SS       df       MS              Number of obs =      51
-------------+------------------------------           F(  1,    49) =    6.83
       Model |  259.874802     1  259.874802           Prob > F      =  0.0119
    Residual |  1864.52027    49  38.0514342           R-squared     =  0.1223
-------------+------------------------------           Adj R-squared =  0.1044
       Total |  2124.39508    50  42.4879015           Root MSE      =  6.1686

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .2546063   .0974255     2.61   0.012     .0588225    .4503901
       _cons |   10.47812   .8657642    12.10   0.000     8.738302    12.21794
------------------------------------------------------------------------------
```

So, one observation is determining whether we have a significant relationship or not. This observation is called "influential." State 30, which happens to be New Hampshire, is said to dominate the regression. It is sometimes easy to find outliers and influential observations, but not always. If you have a large data set, or lots of variables, you could easily miss an influential observation. You do not want your conclusions to be driven by a single observation or even a few observations. So how do we avoid this influential observation trap?

Stata has a function called dfbeta which helps you avoid this problem. It works like this suppose we simply ran the regression with and without New Hampshire. With NH we get a significant result and a relatively large coefficient on x. Without NH we should get a smaller and insignificant coefficient.

```
. regress y x if state ~=30

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =    1.61
       Model |  35.0543824     1  35.0543824           Prob > F      =  0.2112
    Residual |   1047.6542    48  21.8261292           R-squared     =  0.0324
-------------+------------------------------           Adj R-squared =  0.0122
       Total |  1082.70858    49  22.0960935           Root MSE      =  4.6718

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .0989157   .0780517     1.27   0.211    -.0580178    .2558493
       _cons |   9.785673   .6653936    14.71   0.000     8.447809    11.12354
------------------------------------------------------------------------------
```

By dropping NH we are back to the original result reflecting the (true) relationship for the other 50 observations. We could do this 51 times, dropping each state in turn and seeing what happens, but that would be tedious. Instead we can use matrix algebra and computer technology to achieve the same result.

DFbetas

Belsley, Kuh, and Welsch[4] suggest dfbetas (scaled dfbeta) as the best measure of influence.

$DFbetas = \frac{\hat{\beta} - \hat{\beta}_i}{S_{\hat{\beta}}}$ where $S_{\hat{\beta}}$ is the standard error of $\hat{\beta}$ and $\hat{\beta}_i$ is the estimated value of β omitting observation i.

BKW suggest, as a rule of thumb, that an observation is influential if

Dfbetas $> 2/\sqrt{N}$

where N is the sample size, although some researchers suggest that DFbetas should be greater than one (that is greater than one standard error).

After regress, invoke the dfbeta command.

```
. dfbeta
        DFx:  DFbeta(x)
```

[4] Belsley, D. A., Kuh, E. and Welsch, R. E. *Regression Diagnostics*. John Wiley and Sons, New York, NY, 1980.

This produces a new variable called DFx. Now list the observations that might be influential.

```
. list state stnm y x DFx if DFx>2/sqrt(51)

     +-----------------------------------------------+
     | state   stnm        y          x        DFx |
     |-----------------------------------------------|
 51. |    30     NH   42.282   19.83905   2.110021 |
     +-----------------------------------------------+
```

There is only one observation for which DFbetas is greater than the threshold value. Clearly NH is influential, increasing the value of the coefficient by two standard errors. At this point it would behoove us to look more closely at the NH value to make sure it is not a typo or something.

What do we do about influential observations? We hope we don't have any. If we do, we hope they don't affect the main conclusions. We might find simple coding errors. If we have some and they are not coding errors, we might take logarithms (which has the advantage of squashing variance) or weight the regression to reduce their influence. If none of this works, we have to live with the results and confess the truth to the reader.

Multicollinearity

We know that if one variable in a multiple regression is an exact linear combination of one or more variables in the regression, that the regression will fail. Stata will be forced to drop the collinear variable from the model. But what happens if a regressor is not an exact linear combination of some other variable or variables, but almost? This is called multicollinearity. It has the effect of increasing the standard errors of the OLS estimates. This means that variables that are truly significant appear to be insignificant in the collinear model Also, the OLS estimates are unstable and fragile (small changes in the data or model can make a big difference in the regression estimates). It would be nice to know if we have multicollinearity.

Variance inflation factors

Stata has a diagnostic known as the variance inflation factor , VIF, which indicates the presence of multicollearity. The rule of thumb is that multicollinearity is a problem if the VIF is over 30.

Here is an example.

```
. use http://www.ats.ucla.edu/stat/stata/modules/reg/multico, clear

. regress y x1 x2 x3 x4

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  4,    95) =   16.37
       Model |  5995.66253     4  1498.91563            Prob > F      =  0.0000
    Residual |  8699.33747    95  91.5719733            R-squared     =  0.4080
-------------+------------------------------           Adj R-squared =  0.3831
       Total |       14695    99  148.434343            Root MSE      =  9.5693

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   1.118277   1.024484     1.09   0.278    -.9155807    3.152135
```

```
          x2 |   1.286694   1.042406     1.23   0.220    -.7827429    3.356131
          x3 |   1.191635    1.05215     1.13   0.260    -.8971469    3.280417
          x4 |  -.8370988   1.038979    -0.81   0.422    -2.899733    1.225535
       _cons |   31.61912   .9709127    32.57   0.000     29.69161    33.54662
------------------------------------------------------------------------------

. vif

    Variable |       VIF       1/VIF
-------------+----------------------
          x4 |    534.97    0.001869
          x3 |    175.83    0.005687
          x1 |     91.21    0.010964
          x2 |     82.69    0.012093
-------------+----------------------
    Mean VIF |    221.17
```

Wow. All the VIF are over 30. Let's look at the correlations among the independent variables.

```
. correlate x1-x4
(obs=100)

             |       x1       x2       x3       x4
-------------+------------------------------------
          x1 |   1.0000
          x2 |   0.3553   1.0000
          x3 |   0.3136   0.2021   1.0000
          x4 |   0.7281   0.6516   0.7790   1.0000
```

x4 appears to be the major problem here. It is highly correlated with all the other variables. Perhaps we should drop it and try again.

```
. regress y x1 x2 x3

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  3,    96) =   21.69
       Model |  5936.21931     3  1978.73977           Prob > F      =  0.0000
    Residual |  8758.78069    96  91.2372989           R-squared     =  0.4040
-------------+------------------------------           Adj R-squared =  0.3853
       Total |       14695    99  148.434343           Root MSE      =  9.5518

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |    .298443   .1187692     2.51   0.014     .0626879    .5341981
          x2 |   .4527284   .1230534     3.68   0.000     .2084695    .6969874
          x3 |   .3466306   .0838481     4.13   0.000     .1801934    .5130679
       _cons |   31.50512   .9587921    32.86   0.000     29.60194    33.40831
------------------------------------------------------------------------------

. vif

    Variable |       VIF       1/VIF
-------------+----------------------
          x1 |      1.23    0.812781
          x2 |      1.16    0.864654
          x3 |      1.12    0.892234
-------------+----------------------
    Mean VIF |      1.17
```

Much better, the variables are all significant and there is no multicollinearity at all. If we had not checked for multicollinearity, we might have concluded that none of the variables were related to y. So doing a VIF check is generally a good idea.

However, dropping variables is not without its dangers. It could be that x4 is a relevant variable and the result of dropping it is that the remaining coefficients are biased and inconsistent. Be careful.

10 HETEROSKEDASTICITY

The Gauss-Markov assumptions can be summarized as

$$Y_i = \alpha + \beta X_i + u_i$$

$$u_i \sim iid(0, \sigma^2)$$

where ~iid(0,σ^2) means "is independently and identically distributed with mean zero and variance σ^2". In this chapter we learn how to deal with the possibility that the "identically" assumption is violated. Another way of looking at the G-M assumptions is to glance at the expression for the variance of the error term, σ^2 , if there is no i subscript then the variance is the same for all observations. Thus, for each observation the error term, u_i, comes from a distribution with exactly the same mean and variance, i.e., is identically distributed.

When the identically distributed assumption is violated, it means that the error variance differs across observations. This non-constant variance is called heteroskedasticity (hetero = different, skedastic, from the Greek skedastikos, scattering). Constant variance must be homoskedasticity.

If the data are heteroskedastic, ordinary least squares estimates are still unbiased, but they are not efficient. This means that we have less confidence in our estimates. However, even worse, perhaps, our standard errors and t-ratios are no longer correct. It depends on the exact kind of heteroskedasticity whether we are over- or under-estimating our t-ratios, but in either case, we will make mistakes in our hypothesis tests. So, it would be nice to know if our data suffer from this problem.

Testing for heteroskedasticity

The first thing to do is not really a test, but simply good practice: look at the evidence. The residuals from the OLS regression, e_i are estimates of the unknown error term u_i, so the first thing to to is graph the residuals and see if they appear to be getting more or less spread out as Y and X increase.

Breusch-Pagan test

There are two widely used tests for heteroskdeasticity. The first is due to Breusch and Pagan. It is an LM test. The basic idea is as follows.

Consider the following multiple regression model.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

What we want to know is if the variance of the error term is increasing with one of the variables in the model (or some other variable). So we need an estimate of the variance of the error term at each observation. We use the OLS residual, e, as our estimate of the error term, u. The usual estimator of the variance of e is $\sum_{i=1}^{n}(e_i - \bar{e})^2 \Big/ n$. However, we know that $\bar{e}$=0 in any OLS regression and if we want the variance of e for each observation, we have to set n=1. Therefore our estimate of the variance of u_i is e_i^2, the square of the residual. If there is no heteroskedasticity, the squared residuals will be unrelated to any of the independent variables. So an auxiliary regression of the form,

$$e_i^2 / \hat{\sigma}^2 = a_0 + a_1 X_{i1} + a_2 X_{i2}$$

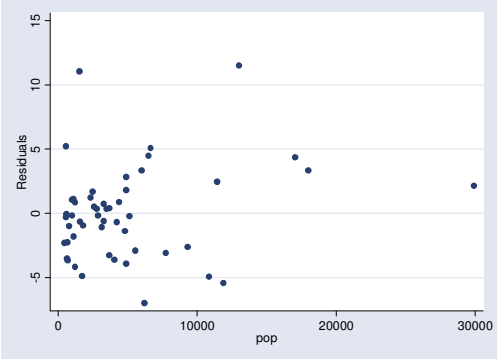
where $\hat{\sigma}^2$ is the estimated error variance from the original OLS regression, should have an R-square of zero. In fact, $nR^2 \sim \chi_2^2$ where the degrees of freedom is the number of variables on the right hand side of the auxiliary test equation. Alternatively, we could test whether the squared residual increases with the predicted value of y or some other suspected variable.

As an example, consider the following regression of major crime per capita across 51 states (including DC) in 1990 (hetero.dta) on prison population per capita, the percent of the population living in metropolitan (metpct) areas, and real per capita personal income.

```
. regress crmaj prison metpct rpcpi

      Source |       SS       df       MS              Number of obs =      51
-------------+------------------------------           F(  3,    47) =   43.18
       Model |  1805.01389     3  601.671298           Prob > F      =  0.0000
    Residual |  654.970383    47  13.9355401           R-squared     =  0.7338
-------------+------------------------------           Adj R-squared =  0.7168
       Total |  2459.98428    50  49.1996855           Root MSE      =   3.733

------------------------------------------------------------------------------
       crmaj |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      prison |   2.952673   .3576379     8.26   0.000     2.233198    3.672148
      metpct |   .1418042   .0328781     4.31   0.000     .0756619    .2079464
       rpcpi |  -.0004919   .0003134    -1.57   0.123    -.0011224    .0001386
       _cons |    6.57136   3.349276     1.96   0.056    -.1665135    13.30923
------------------------------------------------------------------------------
```

Let's plot the residuals on population (which we suspect might be a problem variable).

```
. predict e, resid

. twoway (scatter e pop)
```



Hard to say. If we ignore the outlier at 3000, it looks like the variance of the residuals is getting larger with larger population. Let's see if the Breusch-Pagan test can help us.

The Breusch-Pagan test is invoked in Stata with the hettest command after regress. The default, if you don't include a varlist is to use the predicted value of the dependent variable.

```
. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of crmaj

         chi2(1)      =     0.17
         Prob > chi2  =   0.6819
```

Nothing appears to be wrong. Let's see if there is a problem with any of the independent variables.

```
. hettest prison metpct rpcpi

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: prison metpct rpcpi

         chi2(3)      =     4.73
         Prob > chi2  =   0.1923
```

Still nothing. Let's see if there is a problem with population (since the dependent variable is measured as a per capita variable).

```
. hettest pop

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: pop

         chi2(1)      =     3.32
         Prob > chi2  =   0.0683
```

Ah ha! Population does seem to be a problem. We will consider what to do about it below.

White test

The other widely used test for autocorrelation is due to Halbert White (econometrician from UCSD). White modifies the Breusch-Pagan auxiliary regression as follows.

$$e_i^2 = a_0 + a_1 X_{i1} + a_2 X_{i2} + a_3 X_{i1}^2 + a_4 X_{i2}^2 + a_5 X_{i1} X_{i2}$$

The squared and interaction terms are used to capture any nonlinear relationships that might exist. The number of regressors in the auxiliary test equation is k(k+1)/2 where k is the number of parameters (including the intercept) in the original OLS regression. Here k=3 so, 3(4)/2 -1=6-1=5. The test statistic is

$$nR^2 \sim \chi^2_{k(k+1)/2-1}.$$

To invoke the White test use imtest, white. The im stands for information matrix.

```
. imtest, white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(9)      =      6.63
         Prob > chi2  =    0.6757

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |       6.63      9    0.6757
            Skewness |       4.18      3    0.2422
            Kurtosis |       2.28      1    0.1309
---------------------+-----------------------------
               Total |      13.09     13    0.4405
---------------------------------------------------
```

Ignore Cameron and Trivedi's decomposition. All we care about is heteroskedasticity, which doesn't appear to be a problem according to White. The White test has a slight advantage over the Bruesh-Pagan test in that it is less reliant on the assumption of normality. However, because it computes the squares and cross-products of all the variables in the regression, it sometimes runs out of space or out of degrees of freedom and will not compute. Finally, the White test doesn't warn us of the possibility of heteroskedasticity with a variable that is not in the model, like population. (Remember, there is no heteroskdeasticity associated with prison, metpct, or rpcpi.) For these reasons, I prefer the Breusch-Pagan test.

Weighted least squares

Suppose we know what is causing the heteroskedasticity. For example, suppose the model is (in deviations)

$$y_i = \beta x_i + u_i \text{ where } \text{var}(u_i) = z_i^2 \sigma^2$$

Consider the weighted regression, where we divide both sides by z_i

$$y_i / z_i = \beta(x_i / z_i) + u_i / z_i$$

What is the variance of the new error term, u_i/z_i.

$$\text{var}(u_i / z_i) = \text{var}(c_i u_i) \text{ where } c_i = 1/ z_i$$

$$\text{var}(u_i / z_i) = c_i^2 \text{ var}(u_i) = \frac{z_i^2 \sigma^2}{z_i^2} = \sigma^2$$

So, weighting by $1/z_i$ causes the variance to be constant. Applying OLS to this weighted regression will be unbiased, consistent and efficient (relative to unweighted least squares).

The problem with this approach is that you have to know the exact form of the heteroskedasticity. If you think you know you have heteroskedasticity and correct it with weighted least squares, you'd better be right. If you were wrong and there was no heteroskedasticity, you have created it. You could make things worse by using WLS inappropriately.

There is one case where weighted least squares is appropriate. Suppose that the true relationship between two variables is (in deviations) is,

$$y_i^T = \beta x_i^T + u_i^T, u_i^T \sim IN(0, \sigma^2)$$

However, suppose the data is only available at the state level, where it is aggregated across n individuals. Summing the relationship across the state's population, n, yields,

$$\sum_{i=1}^{n} y_i^T = \beta \sum_{i=1}^{n} x_i^T + \sum_{i=1}^{n} u_i$$

or, for each state,

$$y = \beta x + u.$$

We want to estimate the model as close as possible to the truth, so we divide the state totals by the state population to produce per capita values.

$$\frac{y}{n} = \beta \frac{x}{n} + \frac{u}{n}.$$

What is the variance of the error term?

$$Var\left(\frac{u}{n}\right) = Var\left(\frac{\sum_{i=1}^{n} u_i^T}{n}\right) = \frac{1}{n^2}(Var(u_i^T) + Var(u_2^T) + ... + Var(u_n^T))$$

$$= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

So, we should weight by the square root of population.

$$\sqrt{n}\frac{y}{n} = \beta\sqrt{n}\frac{x}{n} + \sqrt{n}\frac{u}{n}.$$

$$Var\left(\frac{\sqrt{n}u}{n}\right) = nVar\left(\frac{u}{n}\right) = n\frac{\sigma^2}{n} = \sigma^2$$

which is homoskedastic.

This does not mean that it is always correct to weight by the square root of population. After all, the error term might have some heteroskedasticity in addition to being an average. The most common practice in crime studies is to weight by population, not the square root, although some researchers weight by the square root of population.

We can do weighted least squares in Stata with the [weight=variable] option in the regress command. Put the [weight=variable] in square brackets after the varlist.

```
. regress crmaj prison metpct rpcpi [weight=pop]
(analytic weights assumed)
(sum of wgt is   2.4944e+05)

      Source |       SS       df       MS              Number of obs =      51
-------------+------------------------------           F(  3,    47) =   19.35
       Model |  1040.92324     3  346.974412           Prob > F      =  0.0000
    Residual |  842.733646    47  17.9305031           R-squared     =  0.5526
-------------+------------------------------           Adj R-squared =  0.5241

       Total |  1883.65688    50  37.6731377           Root MSE      =  4.2344

------------------------------------------------------------------------------
       crmaj |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      prison |   3.470689   .7111489     4.88   0.000     2.040042    4.901336
      metpct |   .1887946    .058493     3.23   0.002     .0711219    .3064672
       rpcpi |  -.0006482   .0004682    -1.38   0.173    -.0015902    .0002938
       _cons |   4.546826   4.851821     0.94   0.353    -5.213778    14.30743
------------------------------------------------------------------------------
```

Now let's see if there is any heteroskedasticity.

```
. hettest pop

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: pop

         chi2(1)      =     0.60
         Prob > chi2  =   0.4386

. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of crmaj

         chi2(1)      =     1.07
         Prob > chi2  =   0.3010
```

```
. hettest prison metpct rpcpi

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: prison metpct rpcpi

         chi2(3)      =     2.93
         Prob > chi2  =   0.4021
```

The heteroskedasticity seems to have been taken care of. If the correction had made the heteroskedasticity worse instead of better, then we would have tried weighting by 1/pop, the inverse of population, rather than by population itself.

Robust standard errors and t-ratios

Since heteroskedasticity does not bias the coefficient estimates, we could simply correct the standard errors and resulting t-ratios so our hypothesis tests are valid.

Recall how we derived the mean and variance of the sampling distribution of $\hat{\beta}$. We started with a simple regression model in deviations.

$$y_i = \beta x_i + u_i$$

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum x_i(\beta x_i + u_i)}{\sum x_i^2} = \frac{\beta \sum x_i^2}{\sum x_i^2} + \frac{\sum x_i u_i}{\sum x_i^2} = \beta + \frac{\sum x_i u_i}{\sum x_i^2}$$

The mean is

$$E(\hat{\beta}) = E\left(\beta + \frac{\sum x_i u_i}{\sum x_i^2}\right) = \beta + \frac{\sum x_i E(u_i)}{\sum x_i^2} = \beta \quad \text{because } E(u_i) = 0 \text{ for all i.}$$

The variance of $\hat{\beta}$ is the expected value of the square of the deviation of $\hat{\beta}$, $E(\hat{\beta} - \beta)^2$

$$\hat{\beta} - \beta = \frac{\sum x_i u_i}{\sum x_i^2}$$

$Var(\hat{\beta}) = E(\hat{\beta} - \beta)^2$

The x's are fixed numbers so $E(x)=x$, which means

$$Var(\hat{\beta})=E\left(\frac{\sum x_i u_i}{\sum x_i^2}\right)^2=\frac{1}{\left(\sum x_i^2\right)^2}E\left(x_1u_1+x_2u_2+\ldots+x_nu_n\right)^2$$

$$Var(\hat{\beta})=\frac{1}{\left(\sum x_i^2\right)^2}E\left(x_1^2u_1^2+x_2^2u_2^2+\ldots+x_n^2u_n^2+x_1x_2u_1u_2+x_1x_3u_1u_3+x_1x_4u_1u_4+\ldots\right)$$

We know that $E(u_i^2)=\sigma^2$ (identical) and $E(u_iu_j)=0$ for i≠j (independent), so

$$Var(\hat{\beta})=\frac{1}{\left(\sum x_i^2\right)^2}\left(x_1^2\sigma^2+x_2^2\sigma^2+\ldots+x_n^2\sigma^2\right)$$

Factor out the constant variance, σ^2, to get

$$=\frac{1}{\left(\sum x_i^2\right)^2}\left(x_1^2+x_2^2+\ldots+x_n^2\right)\sigma^2=\frac{\sigma^2\sum x_i^2}{\left(\sum x_i^2\right)^2}=\frac{\sigma^2}{\sum x_i^2}$$

In the case of heteroskedasticity we cannot factor out σ^2. So,

$$Var(\hat{\beta}) = \frac{1}{\left(\sum x_i^2\right)^2}E\left(x_1^2u_1^2+x_2^2u_2^2+\ldots+x_n^2u_n^2+x_1x_2u_1u_2+x_1x_3u_1u_3+x_1x_4u_1u_4+\ldots\right)$$

It is still true that $E(u_iu_j)=0$ for i≠j (independent), so

$$Var(\hat{\beta})=\frac{1}{\left(\sum x_i^2\right)^2}\left(x_1^2\sigma_1^2+x_2^2\sigma_2^2+\ldots+x_n^2\sigma_n^2\right)=\frac{\sum x_i^2\sigma_i^2}{\left(\sum x_i^2\right)^2}$$

To make this formula operational, we need an estimate of σ_i^2. As above, we will use the square of the residual, e_i^2.

$$\hat{\sigma}^2_{\hat{\beta}}=\frac{\sum x_i^2e_i^2}{\left(\sum x_i^2\right)^2}$$

Take the square root to produce the robust standard error of beta.

$$S_{\hat{\beta}}=\sqrt{\frac{\sum x_i^2e_i^2}{\left(\sum x_i^2\right)^2}}$$

This is also known as the heteroskedastic consistent standard error (HCSE), or White, Huber, Eicker, or sandwich standard errors. We will call them robust standard errors. Dividing $\hat{\beta}$ by the robust standard error yields the robust t-ratio.

It is easy to get robust standard errors in Stata. For example, in the crime equation above we can see the robust standard and t-ratios by adding the option robust to the regress command.

```
. regress crmaj prison metpct rpcpi, robust
```

```
Regression with robust standard errors                Number of obs =      51
                                                      F(  3,    47) =  131.72
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.7338
                                                      Root MSE      =   3.733

------------------------------------------------------------------------------
             |               Robust
       crmaj |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      prison |   2.952673   .1667412    17.71   0.000     2.617233    3.288113
      metpct |   .1418042     .03247     4.37   0.000     .0764828    .2071255
       rpcpi |  -.0004919   .0002847    -1.73   0.091    -.0010646    .0000807
       _cons |    6.57136   3.006661     2.19   0.034     .5227393    12.61998
------------------------------------------------------------------------------
```

Compare this with the original model.

```
      Source |       SS       df       MS              Number of obs =      51
-------------+------------------------------           F(  3,    47) =   43.18
       Model |  1805.01389     3  601.671298            Prob > F      =  0.0000
    Residual |  654.970383    47  13.9355401            R-squared     =  0.7338
-------------+------------------------------           Adj R-squared =  0.7168
       Total |  2459.98428    50  49.1996855            Root MSE      =   3.733

------------------------------------------------------------------------------
       crmaj |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      prison |   2.952673   .3576379     8.26   0.000     2.233198    3.672148
      metpct |   .1418042   .0328781     4.31   0.000     .0756619    .2079464
       rpcpi |  -.0004919   .0003134    -1.57   0.123    -.0011224    .0001386
       _cons |    6.57136   3.349276     1.96   0.056    -.1665135    13.30923
------------------------------------------------------------------------------
```

It is possible to weight and use robust standard errors. In fact, it is the standard methodology in crime regressions.

```
. regress crmaj prison metpct rpcpi [weight=pop], robust
(analytic weights assumed)
(sum of wgt is   2.4944e+05)

Regression with robust standard errors                Number of obs =      51
                                                      F(  3,    47) =   17.28
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.5526
                                                      Root MSE      =  4.2344

------------------------------------------------------------------------------
             |               Robust
       crmaj |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      prison |   3.470689   .6700652     5.18   0.000     2.122692    4.818686
      metpct |   .1887946   .0726166     2.60   0.012     .0427089    .3348802
       rpcpi |  -.0006482   .0005551    -1.17   0.249    -.0017649    .0004684
       _cons |   4.546826   4.683806     0.97   0.337    -4.875776    13.96943
------------------------------------------------------------------------------
```

What is the bottom line with respect to our treatment of heteroskedasticity? Robust standard errors yield consistent estimates of the true standard error no matter what the form of the heteroskedasticity. Using them allows us to avoid having to discover the particular form of the problem. Finally, if there is no heteroskedasticity, we get the usual standard errors anyway. For these reasons, I recommend always using

robust standard errors. That is, always add “, robust” to the regress command. This is becoming standard procedure among applied researchers. If you know that a variable like population is a problem, then you should weight in addition to using robust standard errors. The existence of a problem variable will become obvious in your library research when you find that most of the previous studies weight by a certain variable.

11 ERRORS IN VARIABLES

One of the Gauss-Markov assumptions is that the independent variable, x, is fixed on repeated sampling. This is a very convenient assumption for proving that the OLS parameter estimates are unbiased However, it is only necessary that x be independent of the error term, that is, that there be no covariance between the error term and the regressor or regressors, that is, $E(x_i u_i)=0$ If this assumption is violated, then ordinary least squares is biased and inconsistent.

This could happen if there are errors of measurement in the independent variable or variables. For example, suppose the true model, expressed in deviations is

$$y = \beta x^T + \varepsilon$$

where x^T is the true value of x. However, we only have the observed $x=x^T-f$ where f is random measurement error.

$$x^T = x - f$$
$$y = \beta(x - f) + \varepsilon$$
$$= \beta x + (\varepsilon - \beta f)$$

We have a problem if x is correlated with the error term (ε-βf).

$$\text{cov}(x, \varepsilon - \beta f) = E[(x^T + f)(\varepsilon - \beta f)]$$
$$= E(x^T\varepsilon + f\varepsilon - x^T\beta f - \beta f^2) = -\beta E(f^2)$$
$$= -\beta\sigma_f^2 \neq 0 \text{ if f has any variance.}$$

So, x is correlated with the error term. This means that the OLS estimate is biased.

$$\hat{\beta} = \frac{\Sigma xy}{\Sigma x^2} = \frac{\Sigma(x^T + f)y}{\Sigma(x^T + f)^2} = \frac{\Sigma(x^T + f)(\beta x^T + \varepsilon)}{\Sigma x^{T2} + 2\Sigma x^T f + \Sigma f^2}$$

$$P\lim(\hat{\beta}) = P\lim\left(\frac{\Sigma(\beta x^{T2} + x^T\varepsilon + \beta x^T f + fy)}{\Sigma x^{T2} + 2\Sigma x^T f + \Sigma f^2}\right) = P\lim\left(\frac{\beta\Sigma x^{T2}}{\Sigma x^{T2} + \Sigma f^2}\right)$$

(We are assuming that $P\lim(fy) = P\lim(x^T f) = P\lim(x^T f) = P\lim(fy) = 0)$ So,

dividing top and bottom by $\sum x^T 2$,

$$P\lim(\hat{\beta}) = P\lim\left(\frac{\beta}{1+\frac{\sum f^2}{\sum x^T 2}}\right) = \frac{\beta}{1+\frac{\text{var}(f)}{\text{var}(x^T)}}$$

Since the two variances cannot be negative, we can conclude that $p\lim\hat{\beta} \neq \beta$ if $\text{var}(f) \neq 0$ which means that the OLS estimate is biased and inconsistent if the measurement error has any variance at all. In this case, the OLS estimate is biased downward, since the true value of beta is divided by a number greater than one.

Cure for errors in variables

There are only two things we can do to fix errors in variables. The first is to use the true value of x. Since this is usually not possible, we have to fall back on statistics. Suppose we know of a variable, z, that is (1) highly correlated with x^T, but (2) uncorrelated with the error term. This variable, called an instrument or instrumental variable, allows us to derive consistent (not unbiased) estimates.

$y = \beta x + v$ where $v = (\varepsilon - \beta f)$

Multiply both sides by z.

$zy=\beta zx+zv$

The error is

$e=zy-\hat{\beta}zx$

The sum of squares of error is

$\sum e^2=\sum(zy-\hat{\beta}zx)^2$

To minimize the sum of squares of error, we take the derivative with respect to $\hat{\beta}$ and set it equal to zero.

$\frac{d\sum e^2}{d\hat{\beta}}=-2zx\sum(zy-\hat{\beta}zx)=0$ (note chain rule).

The derivative is equal to zero if and only if the expression is parentheses is zero.

$\sum zy-\hat{\beta}\sum zx=0$ which implies that

$$\hat{\beta} = \frac{\sum zy}{\sum zx}$$

This is the instrumental variable (IV) estimator. It collapses to the OLS estimator if x can be an instrument for itself, which would happen if x is not measured with error. Then x would be highly correlated with itself, and not correlated with the error term, making it the perfect instrument.

This IV estimator is consistent because z is not correlated with v.

$$p\lim\left(\hat{\beta}\right) = p\lim\left(\frac{\sum yz}{\sum xz}\right) = p\lim\left(\frac{\sum(\beta x+v)z}{\sum xz}\right)$$

$$=p\lim\left(\frac{\beta\sum xz+\sum vz)}{\sum xz}\right)=\beta+\frac{p\lim(\sum vz)}{p\lim(\sum xz)}$$

$= \beta$ since $p\lim\left(\sum vz\right) = 0$ because z is uncorrelated with the error term.

Two stage least squares

Another way to get IV estimates is called two stage least squares (TSLS or 2sls). In the first stage we regress the variable with the measurement error, x, on the instrument, z, using ordinary least squares, saving the predicted values. In the second stage, we substitute the predicted values of x for the actual x and apply ordinary least squares again. The result is the consistent IV estimate above.

First stage: regress x on z:

$$x = \gamma z + w$$

Save the predicted values.

$$\hat{x} = \hat{\gamma} z$$

Note that $\hat{x}$ is a linear combination of z. Since z is independent of the error term, so is $\hat{x}$.

Second stage: substitute $\hat{x}$ for x and apply OLS again.

$$y = \beta\hat{x} + v$$

The resulting estimate of beta is the instrumental variable estimate above.

$$\hat{\beta} = \frac{\sum \hat{x}y}{\sum \hat{x}^2} = \frac{\sum(\hat{\gamma}z)y}{\sum(\hat{\gamma}z)^2} = \frac{\hat{\gamma}\sum zy}{\hat{\gamma}^2\sum z^2} = \frac{\sum zy}{\hat{\gamma}\sum z^2}$$

$\hat{\gamma}=\frac{\sum xz}{\sum z^2}$ from the first stage regression. So,

$= \frac{\sum zy}{\frac{\sum zx}{\sum z^2}\sum z^2} = \frac{\sum zy}{\sum zx}$, the instrumental variable estimator.

Hausman-Wu test

We can test for the presence of errors in variables if we have an instrument. Assume that the relationship between the instrument z and the possibly mis-measured variable, x is linear.

$$x = \gamma z + w$$

We can estimate this relationship using ordinary least squares.

$$\hat{x} = \hat{\gamma} z$$

So,

$$x = \hat{x} + \hat{w}$$

where $\hat{w}$ is the residual from the regression of x on z.

And,

$$y = \beta x + v = \beta(\hat{x} + \hat{w}) + v$$
$$y = \beta\hat{x} + \beta\hat{w} + v$$

Since $\hat{x}$ is simple a linear combination of z, it is uncorrelated with the error term. Therefore, applying OLS to this equation will yield a consistent estimate of β as the parameter multiplying $\hat{x}$. If there is measurement error in x it is all in the residual $\hat{w}$, so the estmate of β from that parameter will be inconsistent. If these two estimates are significantly different, it must be due to measurement error. Let's call the coefficient on $\hat{w}$ δ. So we need an easy way to test whether $\beta=\delta$.

$y = \beta\hat{x} + \delta\hat{w} + v$ but $\hat{x} = x - \hat{w}$, so
$y = \beta(x - \hat{w}) + \delta\hat{w} + v$
$y = \beta x + (\delta - \beta)\hat{w} + v$

The null hypothesis that $\beta=\delta$ can be tested with a standard t-test on the coefficient multiplying $\hat{w}$. If it is significant, then the two parameter estimates are different and there is significant errors in variables bias. If it is not significant, then the two parameters are equal and there is no significant bias.

We can implement this test in Stata as follows. Wooldridge has a nice example and the data set is available on the web. The data are in the Stata data set called errorsinvars.dta. The dependent variable is the log of the wage earned by working women. The independent variable is education. The estimated coefficient is the return to education. First let's do the regression with ordinary least squares.

```
. regress lwage educ

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  1,   426) =   56.93
       Model |  26.3264237     1  26.3264237           Prob > F      =  0.0000
    Residual |  197.001028   426  .462443727           R-squared     =  0.1179
-------------+------------------------------           Adj R-squared =  0.1158
       Total |  223.327451   427  .523015108           Root MSE      =  .68003

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1086487   .0143998     7.55   0.000     .0803451    .1369523
       _cons |  -.1851969   .1852259    -1.00   0.318    -.5492674    .1788735
------------------------------------------------------------------------------
```

The return to education is a nice ten percent. However, the high wages earned by highly educated women might be due to more intelligence. Brighter women will probably do both, go to college and earn high wages. So how much of the return to education is due to the college education? Let's use father's education as an instrument. It is probably correlated with his daughter's brains and is independent of her wages. We can implement instrumental variables with ivreg. The first stage regression varlist is put in parentheses.

```
. ivreg lwage (educ=fatheduc)

Instrumental variables (2SLS) regression
```

```
      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  1,   426) =    2.84
       Model |  20.8673618     1  20.8673618           Prob > F      =  0.0929
    Residual |  202.460089   426  .475258426           R-squared     =  0.0934
-------------+------------------------------           Adj R-squared =  0.0913
       Total |  223.327451   427  .523015108           Root MSE      =  .68939

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0591735   .0351418     1.68   0.093    -.0098994    .1282463
       _cons |   .4411035   .4461018     0.99   0.323    -.4357311    1.317938
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   fatheduc
------------------------------------------------------------------------------

. regress educ fatheduc

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  1,   426) =   88.84
       Model |  384.841983     1  384.841983           Prob > F      =  0.0000
    Residual |  1845.35428   426  4.33181756           R-squared     =  0.1726
-------------+------------------------------           Adj R-squared =  0.1706
       Total |  2230.19626   427  5.22294206           Root MSE      =  2.0813

------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    fatheduc |   .2694416   .0285863     9.43   0.000     .2132538    .3256295
       _cons |   10.23705   .2759363    37.10   0.000     9.694685    10.77942
------------------------------------------------------------------------------

. predict what , resid

. regress lwage educ what

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  2,   425) =   29.80
       Model |  27.4648913     2  13.7324457           Prob > F      =  0.0000
    Residual |   195.86256   425  .460853082           R-squared     =  0.1230
-------------+------------------------------           Adj R-squared =  0.1189
       Total |  223.327451   427  .523015108           Root MSE      =  .67886

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0591735   .0346051     1.71   0.088     -.008845    .1271919
        what |   .0597931   .0380427     1.57   0.117    -.0149823    .1345684
       _cons |   .4411035    .439289     1.00   0.316    -.4223459    1.304553
------------------------------------------------------------------------------
```

The instrumental variables regression shows a lower and less significant return to education. In fact, education is only significant at the .10 level. Also, there might still be some bias because the coefficient on the residual w-hat is almost significant at the .10 level. More research is needed.

12 SIMULTANEOUS EQUATIONS

We know that whenever an explanatory variable is correlated with the error term in a regression, the resulting ordinary least squares estimates are biased and inconsistent. For example, consider the simple Keynesian system.

$$C = \alpha + \beta Y + u$$
$$Y = C + I + G + X$$

where C is consumption, Y is national income (e.g. GDP), I is investment, G is government spending, and X is net exports.

Before we go on, we need to define some terms. The above simultaneous equation system of two equations in two unknowns (C and Y) is known as the **structural model**. It is the theoretical model from the textbook or journal article, with an error term added for statistical estimation. The variables C and Y are known as the **endogenous** variables because their values are determined within the system. I, G, and X are **exogenous** variables because their values are determined outside the system. They are "taken as given" by the structural model. This structural model has two types of equations. The first equation (the familiar consumption function) is known as a **behavioral** equation. Behavioral equations have error terms, indicating some randomness. The second equation is an **identity** because it simply defines national income as the sum of all spending (endogenous plus exogenous). The terms α and β are the **parameters** of the system.

If we knew the values of the parameters (α and β) and the exogenous variables (I, G, X, and the exogenous error term u), we could solve for the corresponding values of C and Y, as follows. Substituting the second equation into the first yields the value for C.

$$C = \alpha + \beta\left(C+I+G+X\right) + u = \alpha + \beta C + \beta I + \beta G + \beta X + u$$
$$C - \beta C = \alpha + \beta I + \beta G + \beta X + u$$
$$C(1-\beta) = \alpha + \beta I + \beta G + \beta X + u$$
$$C = \frac{\alpha + \beta I + \beta G + \beta X + u}{(1-\beta)}$$
$$\text{(RF1) } C = \frac{1}{(1-\beta)}\alpha + \frac{\beta}{(1-\beta)}I + \frac{\beta}{(1-\beta)}G + \frac{\beta}{(1-\beta)}X + \frac{1}{(1-\beta)}u$$

The resulting equation is called the **reduced form** equation for C. I have labeled it RF1 for "reduced form 1." Note that it is a behavioral equation and that, in this case, it is linear. We could, if we had the data, estimate this equation with ordinary least squares.

We can derive the reduced form equation for Y by substituting the consumption function into the national income identity.

$$Y = \alpha + \beta Y + u + I + G + X$$

$$Y - \beta Y = \alpha + I + G + X + u$$

$$Y(1-\beta) = \alpha + I + G + X + u$$

$$Y = \frac{\alpha + I + G + X + u}{(1-\beta)}$$

$$\text{(RF2) } Y = \frac{1}{(1-\beta)}\alpha + \frac{1}{(1-\beta)}I + \frac{1}{(1-\beta)}G + \frac{1}{(1-\beta)}X + \frac{1}{(1-\beta)}u$$

The two reduced form equations look very similar. They are both linear, with almost the same parameters. Also, 1/(1-β) or β/(1-β). The parameters are functions of the parameters of the structural model. Both equations have the same variables on the right hand side, namely all of the exogenous variables.

Because of the fact that the reduced form equations have the same list of variables on the right hand side, all we need to write them down in **general form** is the list of endogenous variables (because we have one reduced form equation for each endogenous variable) and the list of exogenous variables. In the system above the endogenous variables are C and Y and the exogenous variables are I, G, X (and the random error term u), so the reduced form equations can be written as

$$C = \pi_{10} + \pi_{11}I + \pi_{12}G + \pi_{13}X + v_1$$

$$Y = \pi_{20} + \pi_{21}I + \pi_{22}G + \pi_{23}X + v_2$$

where we are using the "pi's" to indicate the reduced from parameters. Obviously, $\pi_{20} = 1/(1-\beta)$, etc.

We are now in position to show that the endogenous variable, Y, on the right hand side of the consumption function is correlated with the error term in the consumption equation. Look at the reduced form equation for Y (RF2) above. Note that Y is a function of u. Clearly, variation in u will cause variation in Y. Therefore, cov(Y,u) is not zero. As a result, the OLS estimate of β in the consumption function is biased and inconsistent.

On the other hand, OLS estimates of the reduced form equation are unbiased and consistent because the exogenous variables are taken as given and therefore uncorrelated with the error term, v_1 or v_2. The bad news is that the parameters from the reduced form are "scrambled" versions of the structural form parameters. In the case of the simple Keynesian model, the coefficients from the reduced form equation for Y correspond to multipliers, giving the change in Y for a one-unit change in I or G or X.

Example: supply and demand

Consider the following structural supply and demand model (in deviations).

$$\text{(S) } q = \alpha p + \varepsilon$$

$$\text{(D) } q = \beta_1 p + \beta_2 y + u$$

where q is the equilibrium quantity supplied and demanded, p is the price, and y is income. There are two endogenous variables (p and q) and one exogenous variable, y. We can derive the reduced form equations by solving by back substitution.

$$q = \alpha p + \varepsilon$$

Solve for p:

$$-\alpha p = -q + \varepsilon$$

$$p = \frac{q-\varepsilon}{\alpha}$$

Substitute into the demand equation,

$$q = \beta_1\left(\frac{q-\varepsilon}{\alpha}\right) + \beta_2 y + u$$

Multiply by α to get rid of fractions.

$$\alpha q = \beta_1(q-\varepsilon) + \alpha\beta_2 y + \alpha u = \beta_1 q - \beta_1\varepsilon + \alpha\beta_2 y + \alpha u$$

$$\alpha q - \beta_1 q = -\beta_1\varepsilon + \alpha\beta_2 y + \alpha u$$

$$(\alpha - \beta_1)q = +\alpha\beta_2 y + \alpha u - \beta_1\varepsilon$$

$$q = \frac{\alpha\beta_2 y+\alpha u-\beta_1\varepsilon}{(\alpha-\beta_1)}$$

$$q = \frac{\alpha\beta_2}{(\alpha-\beta_1)}y + \frac{\alpha u-\beta_1\varepsilon}{(\alpha-\beta_1)} \quad \text{(RF1)}$$

$$q = \pi_{11} y + v_1$$

This is RF1 in general form. Now we solve for p.

$$\alpha p + \varepsilon = \beta_1 p + \beta_2 y + u$$

$$\alpha p-\beta_1 p=\beta_2 y+u-\varepsilon$$

$$(\alpha-\beta_1)p=\beta_2 y+u-\varepsilon$$

$$p=\frac{\beta_2 y+u-\varepsilon}{(\alpha-\beta_1)}$$

$$p=\frac{\beta_2}{(\alpha-\beta_1)}y+\frac{u-\varepsilon}{(\alpha-\beta_1)} \quad \text{(RF2)}$$

$$p = \pi_{21} y + v_2$$

This is RF2 in general form. RF2 reveals that price is a function of u, so applying OLS to either the supply or demand curve will result in biased and inconsistent estimates.

However, we can again appeal to instrumental variables (2sls) to yield consistent estimates. The good news is that the structural model itself produces the instruments. Remember, instruments are required to be highly correlated with the problem variable (p), but uncorrelated with the error term. The exogenous variable(s) in the model satisfy these conditions. In the supply and demand model, income is correlated with price (see RF2), but independent of the error term because it is exogenous.

Suppose we choose the supply curve as our target equation. We can get a consistent estimate of α with the IV estimator,

$$\hat{\alpha} = \frac{\sum yq}{\sum yp}$$

This estimator can be derived as follows.

$$p = \alpha q + \varepsilon$$

Multiply both sides by y

$yp = \alpha yq + y\varepsilon$

Now sum

$\sum yp = \alpha \sum yq + \sum y\varepsilon$

But, the instrument is not correlated with the error term,

$\sum y\varepsilon = 0$

So,

$\sum yp = \alpha \sum yq$

and

$\alpha = \sum yq / \sum yp$

We can also use two stage least squares. In the first stage we regress the problem variable, p, on all the exogenous variables in the model. Here there is only one exogenous variable, y. In other words, estimate the reduced form equation, RF2. The resulting estimate of π_{21} is $\hat{\pi}_{21} = \frac{\sum py}{\sum y^2}$ and the predicted value of p is $\hat{p} = \hat{\pi}_{21} y$. In the second stage we regress q on $\hat{p}$. The resulting estimator is

$$\hat{\alpha} = \frac{\sum q\hat{p}}{\sum \hat{p}^2} = \frac{\hat{\pi}_{21} \sum qy}{\hat{\pi}_{21}^2 \sum y^2} = \frac{\sum qy}{\hat{\pi}_{21} \sum y^2} = \frac{\sum qy}{\frac{\sum py}{\sum y^2} \sum y^2} = \frac{\sum qy}{\sum py}$$

which is the same as the IV estimator above.

Indirect least squares

Yet another method for deriving a consistent estimate of α is indirect least squares. Suppose we estimate the two reduced form equations above. Then we get unbiased and consistent estimates of π_{11} and π_{21}. If we take the probability limit of the two estimators, we get the IV estimator again.

From the reduced form equations,

$$\pi_{11} = \frac{\alpha\beta_2}{\alpha - \beta_1}$$

$$\pi_{12} = \frac{\beta_2}{\alpha - \beta_1}$$

which implies that

$$\frac{\pi_{11}}{\pi_{12}} = \frac{\alpha\beta_2}{\alpha - \beta_1} \frac{\alpha - \beta_1}{\beta_2} = \alpha$$

So,

$$p\lim\left(\frac{\hat{\pi}_{11}}{\hat{\pi}_{12}}\right) = \alpha$$

Because both $\hat{\pi}_{11}$ and $\hat{\pi}_{12}$ are consistent estimates and the ratio of two consistent estimates is consistent. (Remember, consistency "carries over.")

Unfortunately, indirect least squares does not always work. In fact, we cannot get a consistent estimate of the demand equation because, no matter how we manipulate π_{11} and π_{12}, we keep getting α instead of β_1 or β_2.

Suppose we expand the model above to add another explanatory variable to the demand equation: w for weather.

$$q = \alpha p + \varepsilon$$
$$q = \beta_1 p + \beta_2 y + \beta_3 w + w$$

Solving by back substitution, as before, yields the reduced form equations.

$$q = \frac{\alpha\beta_2}{\alpha - \beta_1} y + \frac{\alpha\beta_3}{\alpha - \beta_1} w + \frac{\alpha u - \beta_1 \varepsilon}{\alpha - \beta_1}$$
$$q = \pi_{11} y + \pi_{12} w + v_1$$
$$p = \frac{\beta_2}{\alpha - \beta_1} y + \frac{\beta_3}{\alpha - \beta_1} w + \frac{u - \varepsilon}{\alpha - \beta_1}$$
$$p = \pi_{21} y + \pi_{22} w + v_2$$

We can use indirect least squares to estimate the slope of the supply curve as before.

$$\frac{\pi_{11}}{\pi_{21}} = \frac{\alpha\beta_2}{\alpha - \beta_1} \frac{\alpha - \beta_1}{\beta_2} = \alpha \text{ as before.}$$

But also,

$$\frac{\pi_{12}}{\pi_{22}} = \frac{\alpha\beta_3}{\alpha - \beta_1} \frac{\alpha - \beta_1}{\beta_3} = \alpha \text{ (again).}$$

Whoops. Now we have two estimates of α. Given normal sampling error, we won't get exactly the same estimate, so which is right? Or which is more right?

So, indirect least squares yielded a unique estimate in the case of α, no estimate at all of the parameters of the demand curve, and multiple estimates of α with a minor change in the specification of the structural model. What is going on?

The identification problem

An equation is identified if we can derive estimates of the structural parameters from the reduced form. If we cannot, it is said to be unidentified. The demand curve was unidentified in the above examples because we could not derive any estimates of β_1 or β_2 from the reduced form equations. If an equation is identified, it can be "just" identified, which means the reduced form equations yield a unique estimate of the parameters of the equation of interest, or it can be "over" identified, in which case the reduce form yields multiple estimates of the same parameters of the equation of interest. The supply curve was just identified in the first example, but over identified when we added the weather variable.

It would be nice to know if the structural equation we are trying to estimate is identified. A simple rule, is the so-called **order condition for identification**. To apply the order condition, we first have to choose the equation of interest. Let G be the number of endogenous variables in the equation of interest (including the dependent variable on the left hand side). Let K be the number of exogenous variables that are in the structural model, but **excluded** from the equation of interest. To be identified, the following relationship must hold.

$$K \geq G - 1$$

Let's apply this rule to the supply curve in the first supply and demand model. G=2 (q and p) and K=1 (y is excluded from the supply curve), so 1=2-1 and the supply equation is just identified. This is why we got a

unique estimate of α from indirect least squares. Choosing the demand curve as the equation of interest, we find that G=2 as before, but K=0 because there are no exogenous variables excluded from the demand curve. Finally, let's analyze the supply curve from the second supply and demand model. Again, G=2, but now K=2 (y and w are both excluded from the supply curve). So 2>2-1 and the supply curve is over identified. That is why we got two estimates from indirect least squares.

Two notes concerning the order condition. First, it is a necessary, but not sufficient condition for identification. That is, no equation that fails the order condition will be identified. However, there are rare cases where the order condition is satisfied but the equation is still not identified. However, this is not a major problem because, if we try to estimate an equation using two stage least squares or some other instrumental variable technique, the model will simply fail to yield an estimate. At that point we would know that the equation was unidentified. The good news is that it won't generate a wrong or biased estimate. A more complicated condition, called the rank condition, is necessary and sufficient, but the order condition is usually good enough.

Second, there is always the possibility that the model will be identified according to the order condition, but fail to yield an estimate because the exogenous variables excluded from the equation of interest are too highly correlated with each other. Say we require that two variables be excluded from the equation of interest and we have two exogenous variables in the model but not in the equation. Looks good. However, if the two variables are really so highly correlated that one is simply a linear combination of the other, then we don't really have two independent variables. This problem of "statistical identification" is extremely rare.

Illustrative example

Dennis Epple and Bennett McCallum, from Carnegie Mellon, have developed a nice example of a simultaneous equation model using real data.[5] The market they analyze is the market for broiler chickens. The demand for broiler chicken meat is assumed to be a function of the price of chicken meat, real income, and the price of beef, a close substitute. Supply is assumed to be a function of price and the price of chicken feed (corn). Epple and McCallum have collected data for the US as a whole, from 1950 to 2001. They recommend estimating the model in logarithms. The data are available in DandS.dta.

First we try ordinary least squares on the demand curve. For reasons that will be discussed in a later chapter, we estimate the chicken demand model in first differences. This is a short run demand equation relating the change in price to the change in consumption.

$$\Delta q = \beta_1 \Delta p + \beta_2 \Delta y + \beta_3 \Delta p_{beef} + u$$

Since all the variables are logged, these first differences of logs are growth rates (percent changes). To estimate this equation in Stata, we have to tell Stata that we have a time variable with the tsset year command. We can then use the D. operator to take first differences. We will use the noconstant option to avoid adding a time trend to the demand curve.

```
. tsset year
        time variable:  year, 1950 to 2001

. regress D.lq D.ly D.lpbeef D.lp , noconstant
```

[5] Epple, Dennis and Bennett T. McCallum, Simultaneous equation econometrics: the missing example, http://littlehurt.gsia.cmu.edu/gsiadoc/WP/2004-E6.pdf

```
      Source |       SS       df       MS              Number of obs =      51
-------------+------------------------------           F(  3,    48) =   19.48
       Model |  .053118107     3  .017706036           Prob > F      =  0.0000
    Residual |  .043634876    48   .00090906           R-squared     =  0.5490
-------------+------------------------------           Adj R-squared =  0.5208
       Total |  .096752983    51  .001897117           Root MSE      =  .03015

------------------------------------------------------------------------------
D.lq         |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ly           |
         D1 |   .7582548   .1677874     4.52   0.000     .4208957    1.095614
lpbeef       |
         D1 |    .305005   .0612757     4.98   0.000      .181802    .4282081
lp           |
         D1 |  -.3553965   .0641272    -5.54   0.000    -.4843329   -.2264601
------------------------------------------------------------------------------
```

Lq is the log of quantity (per capita chicken consumption), lp is the log of price, ly is the log of income, and lpbeef is the log of the price of beef. This is not a bad estimated demand function. The demand curve is downward sloping, but very inelastic, and significant at the .10 level. The elasticity of income is positive, indicating that chicken is not an inferior good, which makes sense. The coefficient on the price of beef is positive, which is consistent with the price of a substitute. All the explanatory variables are highly significant.

Now let's try the supply curve.

```
. regress D.lq D.lpfeed D.lp, noconstant

      Source |       SS       df       MS              Number of obs =      39
-------------+------------------------------           F(  2,    37) =    0.28
       Model |   .00083485     2  .000417425           Prob > F      =  0.7595
    Residual |  .055731852    37  .001506266           R-squared     =  0.0148
-------------+------------------------------           Adj R-squared = -0.0385
       Total |  .056566702    39  .001450428           Root MSE      =  .03881

------------------------------------------------------------------------------
D.lq         |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lpfeed       |
         D1 |  -.0392005    .060849    -0.64   0.523    -.1624923    .0840913
lp           |
         D1 |   .0038389   .0910452     0.04   0.967    -.1806362     .188314
------------------------------------------------------------------------------
```

We have a problem. The coefficient on price is virtually zero. This means that the supply of chicken does not respond to changes in price. We need to use instrumental variables.

Let's first try instrumental variables on the demand equation. The exogenous variables are lpfeed, ly, and lpbeef.

```
. ivreg D.lq D.ly D.lpbeef (D.lp= D.lpfeed D.ly D.lpbeef ), noconstant

Instrumental variables (2SLS) regression

      Source |       SS       df       MS              Number of obs =      39
-------------+------------------------------           F(  3,    36) =       .
       Model |  .032485142     3  .010828381           Prob > F      =       .
    Residual |  .024081561    36  .000668932           R-squared     =       .
-------------+------------------------------           Adj R-squared =       .
       Total |  .056566702    39  .001450428           Root MSE      =  .02586
```

```
------------------------------------------------------------------------------
D.lq         |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lp           |
          D1 |  -.4077264   .1410341    -2.89   0.006    -.6937568   -.1216961
ly           |
          D1 |   .9360443   .1991184     4.70   0.000     .5322135    1.339875
lpbeef       |
          D1 |   .3159692   .0977788     3.23   0.003     .1176646    .5142739
------------------------------------------------------------------------------
Instrumented:  D.lp
Instruments:   D.ly D.lpbeef D.lpfeed
------------------------------------------------------------------------------
```

Not much different, but the estimate of the income elasticity of demand is now virtually one, indicating that chicken demand will grow at about the same rate as income. Note that the demand curve is just identified (by lpfeed).

Let's see if we can get a better looking supply curve.

```
. ivreg D.lq D.lpfeed (D.lp=D.lpbeef D.ly), noconstant

Instrumental variables (2SLS) regression

      Source |       SS       df       MS              Number of obs =      39
-------------+------------------------------           F(  2,    37) =       .
       Model | -.079162381     2  -.03958119           Prob > F      =       .
    Residual |  .135729083    37  .003668354           R-squared     =       .
-------------+------------------------------           Adj R-squared =       .
       Total |  .056566702    39  .001450428           Root MSE      =  .06057

------------------------------------------------------------------------------
D.lq         |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lp           |
          D1 |   .6673431   .2617257     2.55   0.015     .1370364     1.19765
lpfeed       |
          D1 |  -.2828285   .1246235    -2.27   0.029    -.5353398   -.0303173
------------------------------------------------------------------------------
Instrumented:  D.lp
Instruments:   D.lpfeed D.lpbeef D.ly
------------------------------------------------------------------------------
```

This is much better. The coefficient on price is now positive and significant, indicating that the supply of chickens will increase in response to an increase in price. Also, increases in chickenfeed prices will reduce the supply of chickens. The supply curve is over identified (by the omission of ly and lpbeef).

We could have asked for robust standard errors, but they didn't make any difference in the levels of significance, so we used the default homoskedastic only standard errors.

Note to the reader. I have taken a few liberties with the original example developed by Professors Epple and McCallum. They actually develop a slightly more sophisticated, and more believable, model, where the coefficient on price in the supply equation starts out negative and significant in the supply equation and eventually winds up positive and significant. However, they have to complicate the model in a number of ways to make it work. This is a cleaner example, using real data, but it could suffer from omitted variable bias. On the other hand, the Epple and McCallum demand curve probably suffers from omitted variable bias too, so there. Students are encouraged to peruse the original paper, available on the web.

Diagnostic tests

There are three diagnostic tests that should be done whenever you do instrumental variables. The first is to justify the identifying restrictions.

Tests for over identifying restrictions

In the above example, we identified the supply equation by omitting income and the price of beef. This seems quite reasonable, but it is nevertheless necessary to justify omitting these variables by showing that they would not be significant if we included them in the regression. There are several versions of this test by Sargan, Basmann, and others. It is simply an LM test where we save the residuals from the instrumental variables regression and then regress them against all the instruments. If the omitted instrument does not belong in the equation, then this regression should have an R-square of zero. We test this null hypothesis with the usual $nR^2 \sim \chi^2(m-k)$ where m is the number of instruments excluded from the equation and k is the number of endogenous variables on the right hand side of the equation.

We can implement this test easily in Stata. First, let's do the ivreg again.

```
. ivreg D.lq D.lpfeed (D.lp=D.lpbeef D.ly), noconstant

Instrumental variables (2SLS) regression

      Source |       SS       df       MS              Number of obs =      39
-------------+------------------------------           F(  2,    37) =       .
       Model | -.079162381     2  -.03958119           Prob > F      =       .
    Residual |  .135729083    37  .003668354           R-squared     =       .
-------------+------------------------------           Adj R-squared =       .
       Total |  .056566702    39  .001450428           Root MSE      =  .06057

------------------------------------------------------------------------------
D.lq         |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lp           |
          D1 |   .6673431   .2617257     2.55   0.015     .1370364     1.19765
lpfeed       |
          D1 |  -.2828285   .1246235    -2.27   0.029    -.5353398   -.0303173
------------------------------------------------------------------------------
Instrumented:  D.lp
Instruments:   D.lpfeed D.lpbeef D.ly
------------------------------------------------------------------------------
```

If your version of Stata has the overid command installed, then you can run the test with a single command.

```
. overid

Tests of overidentifying restrictions:
Sargan N*R-sq test        0.154  Chi-sq(1)    P-value = 0.6948
Basmann test              0.143  Chi-sq(1)    P-value = 0.7057
```

The tests are not significant, indicating that these variables do not belong in the equation of interest. Therefore, they are properly used as identifying variables. Whew!

If there is no overid command after ivreg, then you can install it by clicking on help, search, overid, and clicking on the link that says, "click to install." It only takes a few seconds. If you are using Stata on the server, you might not be able to install new commands. In that case, you can do it yourself as follows. First, save the residuals from the ivreg.

```
. predict es, resid
(13 missing values generated)
```

Now regress the residuals on all the instruments. In our case we are doing the regression in first and the constant term was not included. So use the noconstant option.

```
. regress es D.lpfeed D.lpbeef D.ly, noconstant

      Source |       SS       df       MS              Number of obs =      39
-------------+------------------------------           F(  3,    36) =    0.05
       Model |  .000535634     3  .000178545           Prob > F      =  0.9860
    Residual |  .135193453    36  .003755374           R-squared     =  0.0039
-------------+------------------------------           Adj R-squared = -0.0791
       Total |  .135729086    39  .003480233           Root MSE      =  .06128

------------------------------------------------------------------------------
es           |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lpfeed       |
          D1 |    .005058   .0863394     0.06   0.954    -.1700464    .1801625
lpbeef       |
          D1 |  -.0536104   .1686351    -0.32   0.752    -.3956183    .2883974
ly           |
          D1 |   .1264734   .3897588     0.32   0.747    -.6639941    .9169409
```

We can use the F-form of the LM test by invoking Stata's test command on the identifying variables.

```
. test D.lpbeef D.ly

 ( 1)  D.lpbeef = 0
 ( 2)  D.ly = 0

       F(  2,    36) =    0.07
            Prob > F =    0.9313
```

The results are the same, namely, that the identifying variables are not significant and therefore can be used to identify the equation of interest. However, the prob-valuses are not identical to the large sample chi-square versions reported by the overid command.

To replicate the overid analysis, let's start by reminding ourselves what output is available for further analysis.

```
. ereturn list

scalars:
                  e(N) =  39
              e(df_m) =  3
              e(df_r) =  36
                 e(F) =  .0475437411474742
                e(r2) =  .00394634310271
              e(rmse) =  .0612811038678276
               e(mss) =  .0005356335440678
               e(rss) =  .1351934528853412
              e(r2_a) =  -.0790581283053975
                e(ll) =  55.12129587257286
macros:
            e(depvar) : "es"
               e(cmd) : "regress"
           e(predict) : "regres_p"
             e(model) : "ols"

matrices:
                  e(b) :  1 x 3
                  e(V) :  3 x 3

functions:
            e(sample)
```

```
. scalar r2=e(r2)

. scalar n=e(N)

. scalar nr2=n*r2
```

We know there is one degree of freedom for the resulting chi-square statistic because m=2 omitted instruments (ly and lpbeef) and k=1 endogenous regressor (lp).

```
. scalar prob=1-chi2(1,nr2)

. scalar list r2 n nr2 prob
        r2 =  .00394634
         n =         39
       nr2 =  .15390738
      prob =  .69482895
```

This is identical to the Sargan form reported by the overid command above. An equivalent test of this hypothesis is the J-test for overidentifying restrictions. Instead of using the nR-square statistic, the J-test uses the j=mF statistic where F is the F-test for the auxiliary regression. From the output above. J is distributed as chi-square with the same m-k degrees of freedom.

```
. scalar J=2*e(F)

. scalar probJ=1-chi2(2,J)

. scalar list J probJ
         J =  .09508748
     probJ =  .95356876
```

All of these tests agree that the identifying restrictions are valid. The fact that they are not significant means that income and beef prices do not belong in the supply of chicken equation.

We cannot do the tests for over identifying restrictions on the demand curve because it is just identified. The problem is that the residual from the ivreg regression will be an exact linear combination of the instruments. If we forget and try to do the LM test we get a regression that looks like this (ed is the residual from the ivreg estimate of the demand curve above.)

```
. regress ed D.ly D.lpbeef D.lpfeed, noconstant

      Source |       SS       df       MS              Number of obs =      39
-------------+------------------------------           F(  3,    36) =    0.00
       Model |           0     3           0           Prob > F      =  1.0000
    Residual |  .024081561    36  .000668932           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared = -0.0833
       Total |  .024081561    39  .000617476           Root MSE      =  .02586

------------------------------------------------------------------------------
          ed |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ly |
         D1. |  -5.70e-10   .1644979    -0.00   1.000    -.3336173    .3336173
      lpbeef |
         D1. |   3.94e-10   .0711725     0.00   1.000    -.1443446    .1443446
     lpfeed |
         D1. |   1.02e-11   .0364396     0.00   1.000    -.0739029    .0739029
------------------------------------------------------------------------------
```

So, if you see output like this, you know you can't do a test for over identifying restrictions, because your equation is not over identified.

Test for weak instruments

To be an instrument a variable must satisfy two conditions. It must be (1) highly correlated with the problem variable and (2) it must be uncorrelated with the error term. What happens if (1) is not true? That is, is there a problem when the instrument is only vaguely associated with the endogenous regressor? Bound, Jaeger[6], and Baker have demonstrated that instrumental variable estimates are biased by about 1/F* where F* is the F-test on the (excluded) identifying variables in the reduced form equation for the endogenous regressor. This is the percent of the OLS bias that remains after applying two stage least squares. For example, if F*=2, then 2SLS is still half as biased as OLS. In the case of our chicken supply curve, that would be a regression of lp on ly, lpbeef, and lpfeed and the F* test would be on lpbeef and ly.

```
. regress lp ly lpbeef lpfeed, noconstant

      Source |       SS       df       MS              Number of obs =      40
-------------+------------------------------           F(  3,    37) =39090.05
       Model |  801.425968     3  267.141989           Prob > F      =  0.0000
    Residual |  .252858576    37  .006834016           R-squared     =  0.9997
-------------+------------------------------           Adj R-squared =  0.9997
       Total |  801.678827    40  20.0419707           Root MSE      =  .08267

------------------------------------------------------------------------------
          lp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ly |   .1264946   .0226887     5.58   0.000     .0805228    .1724664
      lpbeef |    .628012   .0764761     8.21   0.000     .4730568    .7829672
      lpfeed |   .1196143   .1046508     1.14   0.260    -.0924285     .331657
------------------------------------------------------------------------------

. test ly lpbeef

 ( 1)  ly = 0
 ( 2)  lpbeef = 0

       F(  2,    37) =   35.75
            Prob > F =    0.0000
```

Obviously, these are not weak instruments and two stage least squares is much less biased than OLS. Actually we should have predicted this by the difference between the OLS and 2sls versions of the equations above.

Anyway, the BJB F test is very easy to do and it should be routinely reported for any simultaneous equation estimation. Note that it has the same problem as the tests for over identifying restrictions, namely, it does not work where the equation is just identified. So we cannot use it for the demand function.

Hausman-Wu test

We have already encountered the Hausman-Wu test in the chapter on errors in variables. However, it is also relevant in the case of simultaneous equations.

Consider the following equation system.

[6] Yes, that is our own Professor Jaeger.

Model A (simultaneous)

(A1) $y_1 = \beta_{12} y_2 + \beta_{13} z_1 + u_1$

(A2) $y_2 = \beta_{21} y_1 + \beta_{22} z_2 + \beta_{23} z_3 + u_2$

Model B (recursive)

(B1) $y_1 = \beta_{12} y_2 + \beta_{13} z_1 + u_1$

(B2) $y_2 = \beta_{22} z_2 + \beta_{23} z_3 + u_2$

where $\text{cov}(u_1 u_2) = 0.$

Suppose we are interested in getting a consistent estimate of β_{12}. If we apply OLS to equation (A1) we get biased and inconsistent estimates. If we apply OLS to equation (B1) we get unbiased, consistent, and efficient estimates. And it is the same equation! We need to know about the second equation, not only the equation of interest.

To apply the Hausman-Wu test to see if we need to use 2sls on the supply equation, we estimate the reduced form equation for Δlp and save the residuals.

```
. regress D.lp D.ly D.lpbeef D.lpfeed, noconstant

      Source |       SS       df       MS              Number of obs =      39
-------------+------------------------------           F(  3,    36) =   12.37
       Model |  .132113063     3  .044037688           Prob > F      =  0.0000
    Residual |  .128161345    36  .003560037           R-squared     =  0.5076
-------------+------------------------------           Adj R-squared =  0.4666
       Total |  .260274407    39  .006673703           Root MSE      =  .05967

------------------------------------------------------------------------------
D.lp         |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ly           |
          D1 |   .7530405   .3794868     1.98   0.055    -.0165944    1.522675
lpbeef       |
          D1 |   .3437728   .1641908     2.09   0.043     .0107785    .6767671
lpfeed       |
          D1 |   .2583745    .084064     3.07   0.004     .0878849    .4288641
------------------------------------------------------------------------------

. predict dlperror, resid
(13 missing values generated)
```

Now, add the error term to the original regression. If it is significant, then there is a significant difference between the OLS estimate and the 2sls estimate. In which case, OLS is biased and we need to use instrumental variables. If it is not significant, then we need to check the magnitude of the coefficient. If it is large and the standard error is even larger, then we cannot be sure that OLS is unbiased. In that event, we should do it both ways, OLS and IV, and hope we get essentially the same results.

```
. regress D.lq D.lp dlperror D.lpfeed, noconstant

      Source |       SS       df       MS              Number of obs =      39
-------------+------------------------------           F(  3,    36) =   18.43
       Model |  .034261762     3  .011420587           Prob > F      =  0.0000
    Residual |  .022304941    36  .000619582           R-squared     =  0.6057
-------------+------------------------------           Adj R-squared =  0.5728
       Total |  .056566702    39  .001450428           Root MSE      =  .02489

------------------------------------------------------------------------------
D.lq         |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lp           |
```

```
          D1 |   .6673431   .1075623     6.20   0.000     .4491967    .8854896
dlperror     |    -.94075   .1280782    -7.35   0.000    -1.200505   -.6809953
lpfeed       |
          D1 |  -.2828285    .051217    -5.52   0.000    -.3867013   -.1789557
------------------------------------------------------------------------------
```

You can also do the Hausman-Wu test using the predicted value of the problem variable instead of the residual from the reduced form equation. The following H-W test equation was performed after saving the predicted value. The test coefficient is simply the negative of the coefficient on the residuals.

```
. regress D.lq D.lp dlphat D.lpfeed, noconstant

      Source |       SS       df       MS              Number of obs =      39
-------------+------------------------------           F(  3,    36) =   18.43
       Model |  .034261763     3  .011420588           Prob > F      =  0.0000
    Residual |   .02230494    36  .000619582           R-squared     =  0.6057
-------------+------------------------------           Adj R-squared =  0.5728
       Total |  .056566702    39  .001450428           Root MSE      =  .02489

------------------------------------------------------------------------------
D.lq         |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lp           |
          D1 |  -.2734069   .0695298    -3.93   0.000    -.4144198    -.132394
dlphat       |     .94075   .1280782     7.35   0.000     .6809953    1.200505
lpfeed       |
          D1 |  -.2828285    .051217    -5.52   0.000    -.3867013   -.1789557
------------------------------------------------------------------------------
```

Applying the Hausman-Wu test to the demand equation yields the following test equation.

```
. regress D.lq D.ly D.lpbeef D.lp dlperror, noconstant

      Source |       SS       df       MS              Number of obs =      39
-------------+------------------------------           F(  4,    35) =   13.99
       Model |  .034797396     4  .008699349           Prob > F      =  0.0000
    Residual |  .021769306    35   .00062198           R-squared     =  0.6152
-------------+------------------------------           Adj R-squared =  0.5712
       Total |  .056566702    39  .001450428           Root MSE      =  .02494

------------------------------------------------------------------------------
D.lq         |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ly           |
          D1 |   .9360443   .1920032     4.88   0.000      .546257    1.325832
lpbeef       |
          D1 |   .3159692   .0942849     3.35   0.002     .1245608    .5073777
lp           |
          D1 |  -.4077264   .1359945    -3.00   0.005    -.6838099    -.131643
dlperror     |   .1343196   .1527992     0.88   0.385    -.1758793    .4445185
------------------------------------------------------------------------------
```

The test reveals that the OLS estimate is not significantly different from the instrumental variables estimate.

Seemingly unrelated regressions

Consider the following system of equations relating local government expenditures on fire and police (FP), schools (S) and other (O) to population density (D), income (Y), and the number of school age children (C). The equations relate the proportion of the total budget (T) in each of these categories to the explanatory variables, D, Y, and C.

$$FP/T = a_0 + a_1 D + a_2 Y + U_1$$
$$S/T = b_0 + b_1 Y + b_2 C + U_2$$
$$O/T = c_0 + c_1 D + c_2 Y + c_2 C + U_3$$

These equations are not simultaneous, so OLS can be expected to yield unbiased and consistent estimates. However, because the proportions have to sum to one (FP/T+S/T+O/T=1), any random increase in expenditures on fire and police must imply a random decrease in schools and/or other components. Therefore, the error terms must be correlated across equations, i.e., $E(U_i U_j) \neq 0$ for i≠j. This is information that could be used to increase the efficiency of the estimates. The procedure is to estimate the equations using OLS, save the residuals, compute the covariances between U_1, U_2, etc. and use these estimated covariances to improve the OLS estimates. This technique is known as seemingly unrelated regressions (SUR) and Zellner efficient least squares (ZELS) for Alan Zellner from the University of Chicago who developed the technique.

In Stata, we can use the sureg command to estimate the above system. Let FPT be the proportion spent on fire and police and ST be the proportion spent on schools. Because the proportion of the budget going to other things is simply one minus the sum of expenditures on fire, police, and schools, we drop the third equation. In the sureg command, we simply put the varlist for each equation in parentheses.

```
Sureg (FPT D Y) (ST Y C)
```

Note: if the equations have the same regressors, then the SUR estimates will be identical to the OLS estimates and there is no efficiency gain. Also, if there is specification bias in one equation, say omitted variables, then, because the residuals are correlated across equations, the SUR estimates of the other equation are biased. The specification error is “propagated” across equations. For this reason, it is not clear that we should use SUR for all applications for which it might be employed. That being said, we should use SUR in those cases where the dependent variables sum to a constant, such as in the above example.

Three stage least squares

Whenever we have a system of simultaneous equations, we automatically have equations whose errors could be correlated. Thus, two stage least squares estimates of a multi-equation system could potentially be made more efficient by incorporating the information contained in the covariances of the residuals. Applying SUR to two stage least squares results in three stage least squares.

Three stage least squares can be implemented in Stata with the reg3 command. Like the sureg command, the equation varlists are put in equations. In this example I use equation labels to identify the demand and supply equations.

In the first example, I replicate the ivreg above exactly by specifying each equation with the noconstant option (in the parentheses) and adding , noconstant to the model as a whole to make sure that the constant is not used as an instrument in the reduced form equations.

```
. reg3 (Demand:D.lq D.lp D.ly D.lpbeef, noconstant) (Supply: D.lq D.lp D.lpfeed
> , noconstant), noconstant

Three-stage least squares regression
------------------------------------------------------------------------------
Equation          Obs  Parms        RMSE    "R-sq"       chi2        P
------------------------------------------------------------------------------
Demand             39      3    .0265163    0.5152      36.31   0.0000
Supply             39      2    .0379724    0.0059       0.15   0.9258
------------------------------------------------------------------------------

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Demand       |
lp           |
          D1 |   -.194991   .0549483    -3.55   0.000    -.3026876   -.0872944
ly           |
          D1 |   .5204546   .1255017     4.15   0.000     .2744758    .7664334
lpbeef       |
          D1 |   .1600626   .0546049     2.93   0.003      .053039    .2670861
-------------+----------------------------------------------------------------
Supply       |
lp           |
          D1 |  -.0250497   .0835039    -0.30   0.764    -.1887144     .138615
lpfeed       |
          D1 |  -.0040577   .0468383    -0.09   0.931    -.0958591    .0877437
------------------------------------------------------------------------------
Endogenous variables:  D.lq
Exogenous variables:   D.lp D.ly D.lpbeef D.lpfeed
------------------------------------------------------------------------------
```

Yuck. The supply function is terrible. Let's try adding the intercept term..

```
. reg3 (Demand:D.lq D.lp D.ly D.lpbeef) (Supply: D.lq D.lp D.lpfeed)

Three-stage least squares regression
------------------------------------------------------------------------------
Equation          Obs  Parms        RMSE    "R-sq"       chi2        P
------------------------------------------------------------------------------
Demand             39      3    .0236988    0.2754      15.13   0.0017
Supply             39      2    .0249881    0.1944       9.56   0.0084
------------------------------------------------------------------------------

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Demand       |
lp           |
          D1 |  -.1871072   .0484276    -3.86   0.000    -.2820235   -.0921909
ly           |
          D1 |   .0948342    .109993     0.86   0.389    -.1207482    .3104166
lpbeef       |
          D1 |   .0491061   .0336213     1.46   0.144    -.0167905    .1150026
_cons        |   .0274115   .0044904     6.10   0.000     .0186104    .0362125
-------------+----------------------------------------------------------------
Supply       |
lp           |
```

```
          D1 |  -.1627361   .0547593    -2.97   0.003    -.2700624   -.0554097
lpfeed       |
          D1 |   .0018805   .0196841     0.10   0.924    -.0366997    .0404607
_cons        |   .0306582   .0042607     7.20   0.000     .0223074     .039009
------------------------------------------------------------------------------
Endogenous variables:  D.lq
Exogenous variables:   D.lp D.ly D.lpbeef D.lpfeed
------------------------------------------------------------------------------
```

Better, at least for the estimates of the price elasticities. None of the other variables are significant. Three stage least squares has apparently made things worse. This is very unusual. Normally the three stage least squares estimates are almost the same as the two stage least squares estimates.

Types of equation systems

We are already familiar with a fully simultaneous equation system. The supply and demand system

(S) $q = \alpha p + \varepsilon$

(D) $q = \beta_1 p + \beta_2 y + u$

is fully simultaneous in that quantity is a function of price and price is a function of quantity. Now consider the following general three equation model.

$$Y_1 = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + u_1$$
$$Y_2 = \beta_0 + \beta_1 Y_1 + \beta_2 X_1 + \beta_3 X_2 + u_2$$
$$Y_3 = \gamma_0 + \gamma_1 Y_1 + \gamma_2 Y_2 + \gamma_3 X_1 + \gamma_4 X_2 + u_3$$

This equation system is simultaneous, but not completely simultaneous. Note that Y_3 is a function of Y_2 and Y_1, but neither of them is a function of Y_3. Y_2 is a function of Y_1, but Y_1 is not a function of Y_2. The model is said to be **triangular** because of this structure. If we also assume that there is no covariance among the error terms, that is, $E(u_1 u_2) = E(u_1 u_3) = E(u_2 u_3) = 0$, then this is a **recursive system**. Since there is no feedback from any right hand side endogenous variable to the dependent variable in the same equation, there is no correlation between the error term and any right hand side variable in the same equation, therefore there is no simultaneous equation bias in this model. OLS applied to any of the equations will yield unbiased, consistent, and efficient estimates (assuming the equations have no other problems).

Finally, consider the following model.

$$Y_1 = \alpha_0 + \alpha_1 Y_2 + \alpha_2 X_1 + \alpha_3 X_2 + u_1$$
$$Y_2 = \beta_0 + \beta_1 Y_1 + \beta_2 X_1 + \beta_3 X_2 + u_2$$
$$Y_3 = \gamma_0 + \gamma_1 Y_1 + \gamma_2 Y_2 + \gamma_3 X_1 + \gamma_4 X_2 + u_3$$

Now Y_1 is a function of Y_2 and Y_2 is a function of Y_1. These two equations are fully simultaneous. OLS applied to either will yield biased and inconsistent estimates. This part of the model is called the **simultaneous core**. This model system is known as **block recursive** and is very common in large simultaneous macroeconomic models. If we assume that $E(u_1 u_3) = E(u_2 u_3) = 0$ then OLS applied to the third (recursive) equation yields unbiased, consistent, and efficient estimates.

Strategies for dealing with simultaneous equations

The obvious strategy when faced with a simultaneous equation system is to use two stage least squares. I don't recommend three stage least squares because of possible propagation of errors across equations. However, there is always the problem of identification and/or weak instruments. Some critics have even suggested that, since the basic model of all economics is the general equilibrium model, all variables are functions of all other variables and no equation can ever be identified. Even if we can identify the equation of interest with some reasonable exclusion assumptions and we have good instruments, the best we can hope for is a consistent, but inefficient estimate of the parameters.

For many situations, the best solution may be to estimate one or more reduced form equations, rather than trying to estimate equations in the structural model. For example, suppose we are trying to estimate the effect of a certain policy, say three-strikes laws, on crime. We may have police in the crime equation and police and crime may be simultaneously determined, but we do not have to know the coefficient on the police variable in the crime equation to determine the effect of the three-strikes law on crime. That coefficient is available from the reduced form equation for crime. OLS applied to the reduced form equations are unbiased, consistent, and efficient.
However, it may be the case that we can't retreat to the reduced form equations. In the above example, suppose we want to estimate the potential effect of a policy that puts 100,000 cops on the streets. In this case we need to know the coefficient linking the police to the crime rate, so that we can multiply it by 100K in order to estimate the effect of 100,000 more police. We have no choice but to use instrumental variables if we are to generate consistent estimates of the parameter of interest.

A strategy that is frequently adopted when large systems of simultaneous equations are being estimated is to plow ahead with OLS despite the fact that the estimates are biased and inconsistent. The estimates may be biased and inconsistent, but they are efficient and the bias might not be too bad (and 2sls is also biased in small samples). So OLS might not be too bad. Usually, researchers adopting this strategy insist that this is just the first step and the OLS estimates are "preliminary." The temptation is to never get around to the next step.

A strategy that is available to researchers that have time series data is to use lags to generate instruments or to avoid simultaneity altogether. Suppose we start with the simple demand and supply model above where we add time subscripts.

$$\text{(S) } q_t = \alpha p_t + \varepsilon_t$$

$$\text{(D) } q_t = \beta_1 p_t + \beta_2 y_t + u_t$$

Note that this model is fully simultaneous and the only way to consistently estimate the supply function is to use two stage least squares. However, suppose this is agricultural data. Then the supply is based on last year's price while demand is based on this year's price. (This is the famous cobweb model.)

$$\text{(S) } q_t = \alpha p_{t-1} + \varepsilon_t$$

$$\text{(D) } q_t = \beta_1 p_t + \beta_2 y_t + u_t$$

Presto! Assuming that $E(u_t \varepsilon_t) = 0$ and assuming we have no autocorrelation (see Chapter 14 below), we now have a recursive system. OLS applied to either equation will yield unbiased, consistent, and efficient estimates.

Even if we can't assume that current supply is based on last year's price, we can at least generate a bunch of instruments by assuming that the equations are also functions of lagged variables. For example, the supply equation could be a function of last years supply as well as last year's price. In that case, we have the following equation system.

(S) $q_t = \alpha_1 p_t + \alpha_2 p_{t-1} + \alpha_3 q_{t-1} + \varepsilon_t$

(D) $q_t = \beta_1 p_t + \beta_2 y_t + u_t$

Note that we have two new instruments (p_{t-1} and q_{t-1}) which are called **lagged endogenous variables**. Since time does not go backwards, this year's quantity and price cannot affect last year's values, these are valid instruments (again assuming no autocorrelation). Now both equations are identified and both can be consistently estimated using two stage least squares.

Time series analysts frequently estimate **vector autoregression (VAR)** models in which the dependent variable is a function of its own lags and lags of all the other variables in the simultaneous equation model. This is essentially the reduced form equation approach for time series data where the lagged endogenous variables are the exogenous variables.

Another example

Pinydyck and Rubinfeld have an example where state and local government expenditures (exp) are taken to be a function of the amount of federal government grants (aid), and, as control variables, population and income

The data are in a Stata data set called EX73 (for example 7.3 from Pindyck and Rubinfeld). The OLS regression model is

```
. regress exp aid inc pop

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  3,    46) = 2185.50
       Model |   925135805     3   308378602           Prob > F      =  0.0000
    Residual |  6490681.38    46  141101.769           R-squared     =  0.9930
-------------+------------------------------           Adj R-squared =  0.9926
       Total |   931626487    49  19012785.4           Root MSE      =  375.64

------------------------------------------------------------------------------
         exp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         aid |   3.233747    .238054    13.58   0.000     2.754569    3.712925
         inc |   .0001902   .0000235     8.10   0.000      .000143    .0002375
         pop |  -.5940753   .1043992    -5.69   0.000      -.80422   -.3839306
       _cons |  -45.69833   84.39169    -0.54   0.591      -215.57    124.1733
------------------------------------------------------------------------------
```

However, many grants are tied to the amount of money the state is spending. So the aid variable could be a function of the dependent variable, exp. Therefore aid is a potentially endogenous variable. Let's do a Hausman-Wu test to see if we need to use two stage least squares. First we need an instrument. Many open-ended grants (that are functions of the amount the state spends, like matching grants) are for education. A variable that is correlated with grants but independent of the error term is the number of school children, ps. We regress aid on ps, and the other exogenous variables, and save the residuals, $\hat{w}$.

```
. regress aid ps inc pop

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  3,    46) =  220.70
       Model |  32510247.4     3  10836749.1           Prob > F      =  0.0000
    Residual |  2258713.81    46  49102.4741           R-squared     =  0.9350
-------------+------------------------------           Adj R-squared =  0.9308
       Total |  34768961.2    49  709570.636           Root MSE      =  221.59
```

```
------------------------------------------------------------------------------
         aid |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ps |  -.8328735   .3838421    -2.17   0.035    -1.605508   -.0602394
         inc |   .0000365   .0000133     2.75   0.008     9.79e-06    .0000633
         pop |   .1738793   .1248854     1.39   0.171    -.0775019    .4252605
       _cons |   42.40924   49.68255     0.85   0.398    -57.59654     142.415
------------------------------------------------------------------------------

. predict what, resid
```

Now add the residual, what (w-hat), to the regression. If it is significant, ordinary least squares is biased.

```
. regress exp aid what inc pop

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  4,    45) = 1716.17
       Model |   925559177     4   231389794           Prob > F      =  0.0000
    Residual |  6067310.05    45  134829.112           R-squared     =  0.9935
-------------+------------------------------           Adj R-squared =  0.9929
       Total |   931626487    49  19012785.4           Root MSE      =  367.19

------------------------------------------------------------------------------
         exp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         aid |   4.522657   .7636841     5.92   0.000     2.984518    6.060796
        what |  -1.420832   .8018143    -1.77   0.083    -3.035769     .194105
         inc |   .0001275   .0000422     3.02   0.004     .0000424    .0002125
         pop |  -.5133494   .1117587    -4.59   0.000     -.738443   -.2882557
       _cons |   -90.1253   86.22021    -1.05   0.301    -263.7817    83.53111
------------------------------------------------------------------------------
```

Apparently there is significant measurement error, at the recommended 10 percent significance level for specification tests. It turns out that the coefficient on aid in the Hausman-Wu test equation is exactly the same as we would have gotten if we had used instrumental variables. To see this, let's use IV to derive consistent estimates for this model.

```
. ivreg exp (aid=ps) inc pop

Instrumental variables (2SLS) regression

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  3,    46) = 1304.09
       Model |   920999368     3   306999789           Prob > F      =  0.0000
    Residual |  10627118.5    46  231024.316           R-squared     =  0.9886
-------------+------------------------------           Adj R-squared =  0.9878
       Total |   931626487    49  19012785.4           Root MSE      =  480.65

------------------------------------------------------------------------------
         exp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         aid |   4.522657   .9996564     4.52   0.000     2.510453    6.534861
         inc |   .0001275   .0000553     2.31   0.026     .0000162    .0002387
         pop |  -.5133494   .1462913    -3.51   0.001    -.8078184   -.2188803
       _cons |   -90.1253   112.8616    -0.80   0.429    -317.3039    137.0533
------------------------------------------------------------------------------
Instrumented:  aid
Instruments:   inc pop ps
------------------------------------------------------------------------------
```

Note that ivreg uses all the variables that are not indicated as problem variables (inc and pop) as well as the designated instrument, ps, as instruments. Note that the coefficient is, in fact, the same as in the Hausman-Wu test equation.

If we want to see the first stage regression, add the option "first." We can also request robust standard errors.

```
. ivreg exp (aid=ps) inc pop, first robust

First-stage regressions
-----------------------

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  3,    46) =  220.70
       Model |  32510247.4     3  10836749.1           Prob > F      =  0.0000
    Residual |  2258713.81    46  49102.4741           R-squared     =  0.9350
-------------+------------------------------           Adj R-squared =  0.9308
       Total |  34768961.2    49  709570.636           Root MSE      =  221.59

------------------------------------------------------------------------------
         aid |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inc |   .0000365   .0000133     2.75   0.008     9.79e-06    .0000633
         pop |   .1738793   .1248854     1.39   0.171    -.0775019    .4252605
          ps |  -.8328735   .3838421    -2.17   0.035    -1.605508   -.0602394
       _cons |   42.40924   49.68255     0.85   0.398    -57.59654     142.415
------------------------------------------------------------------------------


IV (2SLS) regression with robust standard errors       Number of obs =      50
                                                       F(  3,    46) =  972.40
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.9886
                                                       Root MSE      =  480.65

------------------------------------------------------------------------------
             |               Robust
         exp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         aid |   4.522657      1.515     2.99   0.005      1.47312    7.572194
         inc |   .0001275   .0000903     1.41   0.165    -.0000544    .0003093
         pop |  -.5133494   .1853334    -2.77   0.008    -.8864062   -.1402925
       _cons |   -90.1253   86.34153    -1.04   0.302    -263.9218    83.67117
------------------------------------------------------------------------------
Instrumented:  aid
Instruments:   inc pop ps
------------------------------------------------------------------------------
```

Because the equation is just identified, we can't test the over identification restriction. However, ps is significant in the reduced form equation, which means that it is not a weak instrument, although the sign doesn't look right to me.

Summary

The usual development of a research project is that the analyst specifies an equation of interest where the dependent variable is a function of the target variable or variables and a list of control variables. Before estimating with ordinary least squares, the analyst should think about possible reverse causation with respect to each right hand side variable in turn. If there is a good argument for simultaneity, the analyst should specify another equation, or at least a list of possible instruments, for each of the potentially

endogenous variables. Sometimes this process is shortened by the review of the literature where previous analysts may have specified simultaneous equation models.

Given that there may be simultaneity, what is the best way to deal with it? We know that OLS is biased, inconsistent, but relatively efficient compared to instrumental variables (because changing the list of instruments will cause the estimates to change, creating additional variance of the estimates). On the other hand, instrumental variables is consistent, but still biased (presumably less biased than OLS), and inefficient relative to OLS. So we are swapping consistency for efficiency when we use two stage least squares. Is this a good deal for analysts? The consensus is that OLS applied to an equation that has an endogenous regressor results in substantial bias, so much so that the loss of efficiency is a small price to pay.

There are a couple of ways to avoid this tradeoff. The first is to estimate the reduced form equation only. We know that OLS applied to the reduced form equations results in estimates that are unbiased, consistent, and efficient. For many policy analyses, the target variable is an exogenous policy variable. For example, suppose we are interested in the effect of three strikes laws on crime. The crime equation will be a function of arrests, among other things, a number of control variables, and the target variable, a dummy variable that is one for states with three-strikes laws. The arrest variable is likely to be endogenous. We could specify an arrest equation, with the attendant data collection to get the additional variables, and employ two or three stage least squares, which are relatively inefficient. Alternatively, we can estimate the reduced form equation, derived by simply dropping the arrest variable. The resulting estimated coefficient on the three-strikes variable is unbiased, consistent, and efficient. Of course, if arrests are not simultaneous, these estimates are biased by omitted variables. The safest thing to do is to estimate the original crime equation in OLS and the reduced form version, dropping arrests, and see if the resulting estimated coefficient on the target variable are much different. If the results are the same in either case, then we can be confident we have a good estimate of the effect of three-strikes laws on crime.

Another way around the problem is available if the data are time series. Suppose that there is a significant lag between the time a crime is committed and the resulting prison incarceration of the guilty party. Then prison is not simultaneously determined with crime. Similarly, it usually takes some time to get a law, like the three-strikes law, passed and implemented. In that case, crime this year is a function of the law, but the law is a function of crime in previous years, hence not simultaneous. Finally, if a variable Y is known to be a function of lagged X, then, since time cannot go backwards, the two variables cannot be simultaneous. So, lagging the potentially endogenous variable allows us to sidestep the specification of a multi-equation mode. I have seen large simultaneous macroeconomic models estimated on quarterly data avoid the use of two stage least squares by simply lagging the endogenous variables, so that consumption this quarter is a function of income in the previous quarter.These arguments require that the residuals be independent, that is, not autocorrelated. We deal with autocorrelation below.

Should we use three stage least squares? Generally, I recommend against it. Usually, the other equations are only there to generate consistent estimates of a target parameter in the equation of interest. I don't want specification errors in those equations causing inconsistency in the equation of interest after I have gone to all the trouble to specify a multi-equation model.

Finally, be sure to report the results of the Hausman-Wu test, the test for over identifying restrictions, and the Bound, Jaeger, Baker F test for weak instruments whenever you estimate a simultaneous equation model, assuming your model is not just identified.

13 TIME SERIES MODELS

Time series data are different from cross section data in that it has an additional source of information. In a cross section it doesn't matter how the data are arranged, The fifty states could be listed alphabetically, or by region, or any number of ways. However, a time series must be arrayed in order, 1998 before 1999, etc. The fact that time's arrow only flies in one direction is extremely useful. It means, for example, that something that happened in 1999 cannot cause something that happened in 1998. The reverse, of course, is quite possible. This raises the possibility of using lags as a source of information, something that is not possible in cross sections.

Time series models should take dynamics into account: lags, momentum, reaction time, etc. This means that time series data have all the problems of cross section data, plus some more. Nevertheless, time series data have a unique advantage because of time's arrow. Simultaneity is a serious problem for cross section data, but lags can help a time series analysis avoid the pitfalls of simultaneous equations. This is a good reason to prefer time series data.

An interesting question arises with respect to time series data, namely, can we treat a time series, say GDP from 1950 to 2000 as a random sample? It seems a little silly to argue that GDP in 1999 was a random draw from a normal population of potential GDP's ranging from positive to negative infinity. Yet we can treat any tme series as a random sample if we condition on history. The value of GDP in 1999 was almost certainly a function of GDP in 1998, 1997, etc. Let's assume that GDP today is a function of a bunch of past values of GDP,

$$Y_t = a_0 + a_1 Y_{t-1} + \ldots + a_p Y_{t-p} + \varepsilon_t$$

so that the error term is

$$\varepsilon_t = Y_t - a_0 - a_1 Y_{t-1} - \ldots - a_p Y_{t-p}$$

So the error term is simply the difference from what the value GDP today is compared to what it should have been if the true relationship had held exactly. So a time series regression of the form

$$Y_t = a_0 + a_1 Y_{t-1} + \ldots + a_p Y_{t-p} + \varepsilon_t$$

has a truly random error term.

Linear dynamic models

ADL model

The mother of all time series models is the autoregressive-distributed lag, ADL, model. It has the form, assuming one dependent variable (Y) and one independent variable (X),

$$Y_t = a_0 + a_1 Y_{t-1} + \ldots + a_p Y_{t-p} + b_0 X_t + b_1 X_{t-1} + \ldots + b_p X_{t-p} + \varepsilon_t$$

Lag operator

Just for convenience, we will resort to a shorthand notation that is particularly useful for dynamic models. Using the lag operator, we can write the ADL model very conveniently and compactly as

$$a(L)Y_t = b(L)X_t + \varepsilon_t$$

where

$a(L)$ and $b(L)$ are "lag polynomials."

$$a(L) = a_0 + a_1 L + a_2 L^2 + \ldots + a_p L^p = \sum_{i=0}^{p} a_i L^i$$

$$b(L) = b_0 + b_1 L + b_2 L^2 + \ldots + b_k L^k = \sum_{j=0}^{k} b_j L^j$$

The "lag operator" works like this.

$$LX_t = X_{t-1}$$

$$L^2 X_t = L(LX_t) = L(X_{t-1}) = X_{t-2}$$

$$L^3 X_t = L(L^2 X_t) = L(X_{t-2}) = X_{t-3}$$

$$L^k X_t = X_{t-k}$$

A first difference can be written as

$$\Delta X_t = (1-L)X_t = X_t - X_{t-1}$$

So, using lag notation, the ADL(p,p) model

$$Y_t = a_0 + a_1 Y_{t-1} + \ldots + a_p Y_{t-p} + b_0 X_t + b_1 X_{t-1} + \ldots + b_p X_{t-p} + \varepsilon_t$$

can be written as

$$Y_t - a_0 - a_1 Y_{t-1} - \ldots - a_p Y_{t-p} = b_0 X_t + b_1 X_{t-1} + \ldots + b_p X_{t-p} + \varepsilon_t$$

$$\left(1 - a_0 - a_1 L - \ldots - a_p L^p\right) Y_t = \left(b_0 + b_1 L + \ldots + b_p L^p\right) X_t + \varepsilon_t$$

$$a(L)Y_t = b(L)X_t$$

Mathematically, the ADL model is a difference equation, the discrete analog of a differential equation. For the model to be stable (so that it does not explode to positive or negative infinity), the roots of the lag polynomial must sum to a number less than one in absolute value. Since we do not observe explosive behavior in economic systems, we will have to make sure all our dynamic models are stable. So, for a model like the following ADL(p,p) model to be stable

$$Y_t = a_0 + a_1 Y_{t-1} + \ldots + a_p Y_{t-p} + b_0 X_t + b_1 X_{t-1} + \ldots + b_p X_{t-p} + \varepsilon_t$$

we require that

$$\left|a_1 + a_2 + \ldots + a_p\right| < 1$$

The coefficients on the X's, and the values of the X's, are not constrained because the X's are the inputs to the difference equation and do not depend on their own past values.

All dynamic models can be derived from the ADL model. Let's start with the simple ADL(1,1) model.

$$Y_t = a_0 + a_1 Y_{t-1} + b_0 X_t + b_1 X_{t-1} + \varepsilon_t$$

Static model

The static model (no dynamics) can be derived from this model by setting $a_1 = b_1 = 0$

$$Y_t = a_0 + b_0 X_t + \varepsilon_t$$

AR model

The univariate autoregressive (AR) model is

$$Y_t = a_0 + a_1 Y_{t-1} + \varepsilon_t \ \ (b_0 = b_1 = 0)$$

Random walk model

Note that, if

$a_1 = 1$ and $a_0 = 0$ then Y is a "random walk"

$$Y_t = Y_{t-1} + \varepsilon_t$$

Y is whatever it was in the previous period, plus a random error term.

$$Y_t - Y_{t-1} = \Delta Y_t = \varepsilon_t$$

So, the movements of Y are completely random.

The random walk model describes the movement of a drop of water across a hot skillet. Also, because

$$\Delta Y_t = (1 - L) Y_t$$

$$(1 - L) Y_t = a(L) Y_t = \varepsilon_t$$

Y is said to have a unit root because the root of the lag polynomial is equal to one.

If $a_0 \neq 0$, then $Y_t = a_0 + Y_{t-1} + \varepsilon_t$ and $Y_t - Y_{t-1} = a_0 + \varepsilon_t$, so

$$\Delta Y_t = a_0 + \varepsilon_t$$

This is a random walk, "with drift," because if a_0 is positive, Y drifts up, if it is negative, Y drifts down.

First difference model

The first difference model can be derived by setting $a_1 = 1, b_0 = -b_1$

$$Y_t = a_0 + a_1 Y_{t-1} + b_0 X_t + b_1 X_{t-1} + \varepsilon_t$$

$$Y_t = a_0 + Y_{t-1} + b_0 X_t - b_0 X_{t-1} + \varepsilon_t$$

$$Y_t - Y_{t-1} = a_0 + b_0 (X_t - X_{t-1}) + \varepsilon_t$$

$$\Delta Y_t = a_0 + b_0 \Delta X_t + \varepsilon_t$$

Note that there is "drift" or trend if $a_0 \neq 0$.

Distributed lag model

The finite distributed lag model is $(a_1 = 0)$

$$Y_t = a_0 + b_0 X_t + b_1 X_{t-1} + \varepsilon_t$$

Partial adjustment model

The partial adjustment model is (b_1=0)

$$Y_t = a_0 + a_1 Y_{t-1} + b_0 X_t + \varepsilon_t$$

Error correction model

The error correction model is derived as follows. Start by writing the ADL(1,1) model in deviations,

(1) $y_t = a_1 y_{t-1} + b_0 x_t + b_1 x_{t-1} + \varepsilon_t$

Subtract y_{t-1} from both sides.

$$y_t - y_{t-1} = a_1 y_{t-1} + b_0 x_t + b_1 x_{t-1} + \varepsilon_t - y_{t-1}$$

Add and subtract $b_0 x_{t-1}$.

$$y_t - y_{t-1} = a_1 y_{t-1} + b_0 x_t + b_1 x_{t-1} + \varepsilon_t - y_{t-1} + b_0 x_{t-1} - b_0 x_{t-1}$$

Collect terms.

(2) $\Delta y_t = (a_1 - 1) y_{t-1} + b_0 \Delta x_t + (b_0 + b_1) x_{t-1} + \varepsilon_t$

Rewrite equation (1) as

$$y_t - a_1 y_{t-1} = b_0 x_t + b_1 x_{t-1} + \varepsilon_t$$

In long run equilibrium, $y_t = y_{t-1}$, $x_t = x_{t-1}$, and $\varepsilon_t = 0$.

$$y_t - a_1 y_t = b_0 x_t + b_1 x_t$$

$$y_t (1 - a_1) = (b_0 + b_1) x_t$$

$$y_t = \frac{(b_0 + b_1)}{(1 - a_1)} x_t = k x_t \text{ where } k = \frac{(b_0 + b_1)}{(1 - a_1)}$$

In long run equilibrium, $y_t = y_{t-1}$ so

$$y_{t-1} = \frac{(b_0 + b_1)}{(1 - a_1)} x_{t-1} = -\left[\frac{(b_0 + b_1)}{(a_1 - 1)}\right] x_{t-1} = -k x_{t-1}$$

Multiply both sides by -$(a_1 - 1)$ to get

$$-(a_1 - 1) y_{t-1} = (b_0 + b_1) x_{t-1}$$

Substituting for $(b_0 + b_1) x_{t-1}$ in (2) yields

$$\Delta y_t = (a_1 - 1) y_{t-1} - (a_1 - 1) y_{t-1} + b_0 \Delta x_t + \varepsilon_t$$

Substituting for -y_{t-1}

$$\Delta y_t = (a_1 - 1) y_{t-1} - (a_1 - 1)\left[\frac{(b_0 + b_1)}{(a_1 - 1)}\right] x_{t-1} + b_0 \Delta x_t + \varepsilon_t$$

Collect terms in $(a_1 - 1)$

$$\Delta y_t = b_0 \Delta x_t + (a_1 - 1)\left(y_{t-1} - \left[\frac{(b_0 + b_1)}{(a_1 - 1)} \right] x_{t-1} \right) + \varepsilon_t$$

Or,

$$(3)\ \Delta y_t = b_0 \Delta x_t + (a_1 - 1)\left(y_{t-1} - k x_{t-1} \right) + \varepsilon_t$$

Sometimes written as,

$$(3a)\ \Delta y_t = b_0 \Delta x_t + (a_1 - 1)\left(y - kx \right)_{t-1} + \varepsilon_t$$

The term $(y_{t-1}\text{-}kx_{t-1})$ is the difference between the equilibrium value of y (kx) and the actual value of y in the previous period. The coefficient k is the long run response of y to a change in x. The coefficient $a_1 - 1$ is the speed of adjustment of y to errors in the form of deviations from the long run equilibrium. Together, the coefficient and the error are called the error correction mechanism. The coefficient β_0 is the impact on y of a change in x.

Cochrane-Orcutt model

Another commonly used time series model is the Cochrane-Orcutt or **common factor** model. It imposes the restriction that $b_1 = -b_0 a_1$.

$$Y_t = a_0 + a_1 Y_{t-1} + b_0 X_t + b_1 X_{t-1} + \varepsilon_t$$
$$Y_t = a_0 + a_1 Y_{t-1} + b_0 X_t - b_0 a_1 X_{t-1} + \varepsilon_t$$
$$Y_t - a_1 Y_{t-1} = a_0 + b_0 (X_t - a_1 X_{t-1}) + \varepsilon_t$$
$$(1 - a_1 L) Y_t = a_0 + b_0 (1 - a_1 L) X_t + \varepsilon_t$$

Y and X are said to have common factors because both lag polynomials have the same root, a_1. The terms $Y_t - a_1 Y_{t-1}$ and $(X_t - a_1 X_{t-1})$ are called **generalized first differences** because they are differences, but the coefficient on the lag polynomial is a_1, instead of 1 which would be the coefficient for a first difference.

14 AUTOCORRELATION

We know that the Gauss-Markov assumptions require that $Y_t = \alpha + \beta X_t + u_t$ where $u_t \sim iid(0, \sigma^2)$. The first i in iid is independent. This actually refers to the sampling method and assumes a simple random sample where the probability of observing u_t is independent of observing u_{t-1}. When this assumption, and the other assumptions, are true, ordinary least squares applied to this model is unbiased, consistent, and efficient.

As we have seen above, it is frequently necessary in time series modeling to add lags. Adding a lag of X is not a problem. However adding a lagged dependent variable changes this dramatically. In the model

$$Y_t = \alpha + \beta X_t + \gamma Y_{t-1} + u_t, \;\; u_t \sim iid(0, \sigma^2)$$

the explanatory variables, which now include Y_{t-1}, cannot be assumed to be a set of constants, fixed on repeated sampling because, since Y is a random number, the lag of Y is also a random number. It turns out that the best we can get out of ordinary least squares with a lagged dependent variable is that OLS is consistent and asymptotically efficient. It is no longer unbiased and efficient.

To see this, assume a very simple AR model (no x) expressed in deviations.

$$y_t = a_1 y_{t-1} + \varepsilon_t$$

$$\hat{a}_1 = \frac{\sum y_{t-1} y_t}{\sum y_{t-1}^2} = \frac{\sum y_{t-1}\left(a_1 y_{t-1} + \varepsilon_t\right)}{\sum y_{t-1}^2} = a_1 + \frac{\sum y_{t-1}\varepsilon_t}{\sum y_{t-1}^2}$$

$$E\left(\hat{a}_1\right) = a_1 + E\left(\frac{\sum y_{t-1}\varepsilon_t}{\sum y_{t-1}^2}\right) = a_1 + \frac{Cov(y_{t-1}\varepsilon_t)}{Var(y_{t-1})}$$

Since $\varepsilon_t = y_t - a_1 y_{t-1}$ it is likely that $Cov(y_{t-1}\varepsilon_t) \neq 0$ at least in small samples, which means that $E\left(\hat{a}_1\right) \neq a_1$ and OLS is biased. However, in large samples, as the sample size goes to infinity, we expect that $Cov(y_{t-1}\varepsilon_t) = 0$ which implies that OLS is at least consistent.

It is unfortunate that all we can hope for in most dynamic models is consistency, since most time series are quite small relative to cross sections. We will just have to live with it.

If the errors are not independent, then they are said to be "autocorrelated" (i.e., correlated with themselves, lagged) or "serially correlated." The most common type of autocorrelation is so-called "first order" serial correlation,

$$u_t = \rho u_{t-1} + \varepsilon_t$$

where $-1 \le \rho \le 1$ is the autocorrelation coefficient and $\varepsilon_t \sim iid(0, \sigma^2)$

It is possible to have autocorrelation of any order, the most general form would be q'th order autocorrelation,

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \ldots + \rho_q u_{t-q} + \varepsilon_t$$

Autocorrelation is also known as moving average errors.

Autocorrelation is caused primarily by omitting variables with time components. Sometimes we simply don't have all the variables we need and we have to leave out some variables that are unobtainable. The impact of these variables goes into the error term. If these variables have time trends or cycles or other time-dependent behavior, then the error term will be autocorrelated. It turns out that autocorrelation is so prevalent that we can expect some form of autocorrelation in any time series analysis.

Effect of autocorrelation on OLS estimates

The effect that autocorrelation has on OLS estimators depends on whether the model has a lagged dependent variable or not. If there is no lagged dependent variable, then autocorrelation causes OLS to be inefficient. However, there is a "second order bias" in the sense that OLS standard errors are underestimated and the corresponding t-ratios are overestimated in the usual case of positive autocorrelation. This means that our hypothesis tests can be wildly wrong.[7]

If there is a lagged dependent variable, then OLS is biased, inconsistent, and inefficient, and the t-ratios are overestimated. There is nothing good about OLS under these conditions. Unfortunately, as we have seen above, most dynamic models use a lagged dependent variable, so we have to make sure we eliminate any autocorrelation.

We can prove the inconsistency of OLS in the presence of autocorrelation and a lagged dependent variable as follows. Recall our simple AR(1) model above.

$$y_t = a_1 y_{t-1} + u_t$$
$$u_t = \rho u_{t-1} + \varepsilon_t$$

The ordinary least squares estimator of a is,

$$\hat{a}_1 = \frac{\sum y_{t-1} y_t}{\sum y_{t-1}^2} = \frac{\sum y_{t-1}\left(a_1 y_{t-1} + u_t\right)}{\sum y_{t-1}^2} = a_1 + \frac{\sum y_{t-1} u_t}{\sum y_{t-1}^2}$$

The expected value is

$$E\left(\hat{a}_1\right) = a_1 + E\left(\frac{\sum y_{t-1} u_t}{\sum y_{t-1}^2}\right) = a_1 + \frac{Cov(y_{t-1} u_t)}{Var(y_{t-1})}$$

The covariance between the explanatory variable and the error term is

[7] In the case of negative autocorrelation, it is not clear whether the t-ratios will be overestimated or underestimated. However, the t-ratios will be "wrong."

$Cov(y_{t-1}u_t) = E(y_{t-1}u_t) = E\left[(a_1 y_{t-2} + u_{t-1})(\rho u_{t-1} + \varepsilon_t)\right]$

$= E\left[a_1\rho u_{t-1}y_{t-2} + a_1 y_{t-2}\varepsilon_t + \rho u_{t-1}^2 + u_{t-1}\varepsilon_t\right] = a_1\rho E\left[u_{t-1}y_{t-2}\right] + \rho E\left[u_{t-1}^2\right]$

Since $E(y_{t-2}\varepsilon_t) = E(u_{t-1}\varepsilon_t) = 0$ because ε_t is truly random. Now,

$E(u_t^2) = E(u_{t-1}^2) = Var(u_t)$ and $E(u_{t-1}y_{t-2}) = E(u_t y_{t-1}) = Cov(y_{t-1}u_t)$ which implies

$E\left[u_{t-1}y_{t-2}\right] = a_1\rho E\left[u_{t-1}y_{t-2}\right] + \rho E\left[u_{t-1}^2\right]$

$E\left[u_{t-1}y_{t-2}\right] - a_1\rho E\left[u_{t-1}y_{t-2}\right] = \rho E\left[u_{t-1}^2\right]$

$$E\left[u_{t-1}y_{t-2}\right] = \frac{\rho E\left[u_{t-1}^2\right]}{1-a_1\rho}$$

$$Cov(y_{t-1}u_t) = \frac{\rho Var(u_t)}{1-a_1\rho}$$

So, if $\rho \neq 0$, the explanatory variable is correlated with the error term and OLS is biased. Since ρ is constant, it doesn't go away as the sample size goes to infinity. Therefore, OLS is inconsistent.

Since OLS is inefficient and has overestimated t-ratios in the best case and is biased, inconsistent, inefficient, with overestimated t-ratios in the worse case, we should find out if we have autocorrelation or not.

Testing for autocorrelation

The Durbin Watson test

The most famous test for autocorrelation is the Durbin-Watson statistic. It has the formula

$d = \frac{\sum(e_t - e_{t-1})^2}{\sum e_t^2}$ where e_t is the residual from the OLS regression.

Under the null hypothesis of no autocorrelation, we can derive the expected value as follows.

Note that, under the null

u_t □ $IN(0,\sigma^2)$ which implies that $E(u_t) = 0$, $E(u_t^2) = \sigma^2$, $E(u_t u_{t-1}) = 0$.

The empirical analog is that the residual, e, has the following properties.

$E(e_t) = E(e_{t-1}) = E(e_t e_{t-1}) = 0$ and $E(e_t) = E(e_{t-1}^2) = \sigma^2$. Therefore,

$$E(d) = \frac{\sum E(e_t - e_{t-1})^2}{\sum E(e_t^2)} = \frac{\sum E(e_t^2 - 2e_t e_{t-1} + e_{t-1}^2)^2}{\sum E(e_t^2)} = \frac{\sum(E(e_t^2) - 2E(e_t e_{t-1}) + E(e_{t-1}^2))}{\sum E(e_t^2)}$$

$$= \frac{\sum_{t=1}^{T}(\sigma^2 + \sigma^2)}{\sum_{t=1}^{T}\sigma^2} = \frac{T\sigma^2 + T\sigma^2}{T\sigma^2} = 2$$

where T is the sample size (in a time series t=1...T).

Now assume autocorrelation so that $E(e_t e_{t-1}) \neq 0$. What happens to the expected value of the Durbin Watson statistic? As a preliminary, let's review the concept of a correlation coefficient.

$$r = \frac{1}{T}\frac{\sum xy}{S_x S_y}$$

But,

$$\text{cov}(x, y) = E(xy) = \frac{1}{T}\sum xy$$

So,

$$r = \frac{\text{cov}(x, y)}{\sqrt{\sigma_x^2}\sqrt{\sigma_y^2}}$$

In this case, $x = e_t$, $y = e_{t-1}$ and $\sigma_x^2 = \sigma_y^2 = \sigma^2$ which means that the autocorrelation coefficient is

$$\rho = \frac{\text{cov}(e_t e_{t-1})}{\sqrt{\sigma^2}\sqrt{\sigma^2}} = \frac{E(e_t e_{t-1})}{\sigma^2}$$

and

$$\rho\sigma^2 = E(e_t e_{t-1})$$

The expected value of the Durbin Watson statistic is

$$E(d) = \frac{\sum E(e_t - e_{t-1})^2}{\sum E(e_t^2)} = \frac{\sum E(e_t^2 - 2e_t e_{t-1} + e_{t-1}^2)^2}{\sum E(e_t^2)} = \frac{\sum (E(e_t^2) - 2E(e_t e_{t-1}) + E(e_{t-1}^2))}{\sum E(e_t^2)}$$

$$= \frac{\sum_{t=1}^{T}(\sigma^2 - 2\rho\sigma^2 + \sigma^2)}{\sum_{t=1}^{T}\sigma^2} = \frac{T\sigma^2 - 2T\rho\sigma^2 + T\sigma^2}{T\sigma^2} = 2 - 2\rho$$

Since ρ is a correlation coefficient it has to less than or equal to one in absolute value. The usual case in economics is positive autocorrelation, so that ρ is positive. Clearly, as ρ approaches one, the Durbin-Watson statistic approaches zero. On the other hand, for negative autocorrelation, as ρ goes to negative one, the Durbin-Watson statistic goes to four. As we already know, when ρ=0, the Durbin-Watson statistic is equal to two.

The Durbin-Watson test, despite its fame, has a number of drawbacks. The first is that the test is not exact and there is an indeterminate range in the DW tables. For example, suppose you get a DW statistic of 1.35 and you want to know if you have autocorrelation. If you look up the critical value corresponding to a sample size of 25 and one explanatory variable for the five percent significance level, you will find two values, d_l=1.29 and d_u=1.45. If your value is below 1.29 you have significant autocorrelation, according to DW. If your value is above 1.45, you cannot reject the null hypothesis of no autocorrelation. Since your value falls between these two, what do you conclude? Answer: test fails! Now what are you going to do? Actually, this is not a serious problem. Be conservative and simply ignore the lower value. That is, reject the null hypothesis of no autocorrelation whenever the DW statistic is below d_U.[8]

The second problem is fatal. Whenever the model contains a lagged dependent variable, the DW test is biased toward two. That is, when the test is most needed, it not only fails, it tells us we have no problem when we could have a very serious problem. For this reason, we will need another test for autocorrelation.

The LM test for autocorrelation

[8] If the DW statistic is negative, then the test statistic is 4-DW.

Breusch and Godfrey independently developed an LM test for autocorrelation. Like all LM tests, we run the regression and then save the residuals for further analysis. For example, consider the following model.

$Y_t = a_0 + b_0 X_t + u_t$

$u_t = \rho u_{t-1} + \varepsilon_t$

which implies that

$Y_t = a_0 + b_0 X_t + \rho u_{t-1} + \varepsilon_t$

It would be nice to simply estimate this model, get an estimate of ρ and test the null hypothesis that ρ=0 with a t-test. Unfortunately, we don't observe the error term, so we can't do it that way. However, we can estimate the error term with the residuals from the OLS regression of Y on X and lag it to get an estimate of u_{t-1} We can then run the test equation,

$Y_t = a_0 + b_0 X_t + \rho e_{t-1} + \varepsilon_t$

where e_{t-1} is the lagged residual. We can test null hypothesis with either the t-ratio on the lagged residual or, because this, like all LM tests, is a large sample test, we could use the F-ratio for the null hypothesis that the coefficient on the lagged residual is zero. The F-ratio is distributed according to chi-square with one degree of freedom. Both the Durbin-Watson and Breusch-Godfrey LM tests are automatic in Stata.

I have a data set on crime and prison population in Virginia from 1972 to 1999 (crimeva.dta). The first thing we have to do is tell Stata that this is a time series and the time variable is year.

```
. tsset year
        time variable:  year, 1956 to 2005
```

Now let's do a simple regression of the log of major crime on prison.

```
. regress lcrmaj prison

      Source |       SS       df       MS              Number of obs =      28
-------------+------------------------------           F(  1,    26) =   83.59
       Model |  .570667328     1  .570667328           Prob > F      =  0.0000
    Residual |  .177496135    26  .006826774           R-squared     =  0.7628
-------------+------------------------------           Adj R-squared =  0.7536
       Total |  .748163462    27  .027709758           Root MSE      =  .08262

------------------------------------------------------------------------------
      lcrmaj |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      prison |  -.1345066   .0147116    -9.14   0.000    -.1647467   -.1042665
       _cons |   9.663423   .0379523   254.62   0.000     9.585411    9.741435
------------------------------------------------------------------------------
```

First we want the Durbin-Watson statistic. It is ok since we don't have any lagged dependent variables.

```
. dwstat

Durbin-Watson d-statistic(  2,    28) =  .6915226
```

Now let's get the Breusch-Godfrey test statistic.

```
. bgodfrey

Breusch-Godfrey LM test for autocorrelation
------------------------------------------------------------------------------
    lags(p)  |          chi2              df                 Prob > chi2
-------------+----------------------------------------------------------------
       1     |         10.184              1                   0.0014
------------------------------------------------------------------------------
                        H0: no serial correlation
```

OK, the DW statistic is uncomfortably close to zero. (The dl is 1.32 for k=1 and T=27 at the five percent level. So we reject the null hypothesis of no autocorrelation.) And the BG test is also highly significant. So we apparently have serious autocorrelation. Just for fun, let's do the LM test ourselves. First we have to save the residuals.

```
. predict e, resid
(22 missing values generated)
```

Now we add the lagged residuals to the regression.

```
. regress lcrmaj prison L.e

      Source |       SS       df       MS              Number of obs =      27
-------------+------------------------------           F(  2,    24) =   78.55
       Model |   .64494128     2   .32247064           Prob > F      =  0.0000
    Residual |  .098530434    24  .004105435           R-squared     =  0.8675
-------------+------------------------------           Adj R-squared =  0.8564
       Total |  .743471714    26  .028595066           Root MSE      =  .06407

------------------------------------------------------------------------------
lcrmaj       |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
prison       |  -.1451148   .0118322   -12.26   0.000    -.1695353   -.1206944
e            |
          L1 |   .6516013   .1631639     3.99   0.001     .3148475    .9883552
_cons        |   9.689539   .0308821   313.76   0.000     9.625801    9.753277
------------------------------------------------------------------------------
```

The t-test on the lagged e is highly significant, as expected. What is the corresponding F-ratio?

```
. test L.e

 ( 1)  L.e = 0

       F(  1,    24) =   15.95
            Prob > F =    0.0005
```

It is certainly significant with the small sample correction. It must be incredibly significant asymptotically.

```
. scalar prob=1-chi2(1, 15.95)

. scalar list prob
      prob =  .00006504
```

Wow. Now that's significant.

We apparently have very serious autocorrelation. What should we do about it? We discuss possible cures below.

Testing for higher order autocorrelation

The Breusch-Godfrey test can be used to test for higher order autocorrelation. To test for second order serial correlation, for example, use the lags option of the bgodfrey command.

```
. bgodfrey, lags(1 2)

Breusch-Godfrey LM test for autocorrelation
------------------------------------------------------------------------------
    lags(p)  |          chi2              df                 Prob > chi2
-------------+----------------------------------------------------------------
       1     |         10.184              1                   0.0014
```

```
       2     |         14.100              2                   0.0009
-------------------------------------------------------------------------
                      H0: no serial correlation
```

Hmmm, look's bad. Let's see if there is even higher order autocorrelation.

```
. bgodfrey, lags(1 2 3 4)

Breusch-Godfrey LM test for autocorrelation
-------------------------------------------------------------------------
    lags(p)  |          chi2               df                 Prob > chi2
-------------+-----------------------------------------------------------
       1     |         10.184              1                   0.0014
       2     |         14.100              2                   0.0009
       3     |         14.412              3                   0.0024
       4     |         15.090              4                   0.0045
-------------------------------------------------------------------------
                      H0: no serial correlation
```

It gets worse and worse. These data apparently have severe high order serial correlation. We could go on adding lags to the test equation, but we'll stop here.

Cure for autocorrelation

.

There are two approaches to fixing autocorrelation. The classic textbook solution is to use Cochrane-Orcutt generalized first differences, also known as common factors.

The Cochrane-Orcutt method

Suppose we have the model above with first order autocorrelation.

$$Y_t = a_0 + b_0 X_t + u_t$$
$$u_t = \rho u_{t-1} + \varepsilon_t$$

Assume for the moment that we knew the value of the autocorrelation coefficient. Lag the first equation and multiply by ρ.

$$\rho Y_{t-1} = \rho a_0 + b_0 \rho X_{t-1} + \rho u_{t-1}$$

Now subtract this equation from the first equation.

$$Y_t - \rho Y_{t-1} = a_0 - \rho a_0 + b_0 X_t - b_0 \rho X_{t-1} + u_t - \rho u_{t-1}$$
$$Y_t - \rho Y_{t-1} = a_0(1-\rho) + b_0(X_t - \rho X_{t-1}) + \varepsilon_t$$

Since

$$u_t - \rho u_{t-1} = \varepsilon_t$$

The problem is that we don' t know the value of ρ. In fact we are going to estimate $a_0, b_0,$ and ρ jointly. Cochrane and Orcutt suggest an iterative method.

(1) Estimate the original regression $Y_t = a_0 + b_0 X_t + u_t$ with OLS and save the residuals, e_t.

(2) Estimate the auxiliary regression $e_t = \rho e_{t-1} + v_t$ with OLS, save the estimate of ρ, $\hat{\rho}$.

(3) Transform the equation using generalized first differences using the estimate of ρ.

$$Y_t - \hat{\rho} Y_{t-1} = a_0(1-\hat{\rho}) + b_0(X_t - \hat{\rho} X_{t-1}) + \varepsilon_t$$

(4) Estimate the generalized first difference regression with OLS, save the residuals.
(5) Go to (2).
Obviously, we need a stopping rule to avoid an infinite loop. The usual rule is to stop when successive estimates of ρ are within a certain amount, say, .01.

The Cochrane-Orcutt method can be implemented in Stata using the prais command The prais command uses the Prais-Winston technique, which is a modification of the Cochrane-Orcutt method that preserves the first observation so that you don't get a missing value because of lagging the residual.

```
. prais lcrmaj prison, corc

Iteration 0:  rho = 0.0000
Iteration 1:  rho = 0.6353
Iteration 2:  rho = 0.6448
Iteration 3:  rho = 0.6462
Iteration 4:  rho = 0.6465
Iteration 5:  rho = 0.6465
Iteration 6:  rho = 0.6465
Iteration 7:  rho = 0.6465

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

      Source |       SS       df       MS              Number of obs =      27
-------------+------------------------------           F(  1,    25) =   28.34
       Model |  .111808732     1  .111808732           Prob > F      =  0.0000
    Residual |  .098637415    25  .003945497           R-squared     =  0.5313
-------------+------------------------------           Adj R-squared =  0.5125
       Total |  .210446147    26  .008094083           Root MSE      =  .06281

------------------------------------------------------------------------------
      lcrmaj |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      prison |  -.1619739   .0304269    -5.32   0.000    -.2246394   -.0993085
       _cons |   9.736981   .0864092   112.68   0.000     9.559018    9.914944
-------------+----------------------------------------------------------------
         rho |   .6465167
------------------------------------------------------------------------------
Durbin-Watson statistic (original)    0.691523
Durbin-Watson statistic (transformed) 1.311269
```

The Cochrane-Orcutt approach has improved things, but the DW statistic still indicates some autocorrelation in the residuals.

Note that the Cochrane-Orcutt regression is also known as feasible generalized least squares or FGLS. We have seen above that it is a common factor model, and can be derived from a more general ADL model. It is therefore true only if the restrictions of the ADL are true. We can test these restrictions to see if the CO model is a legitimate reduction from the ADL.

Start with the CO model.

$$Y_t - \rho Y_{t-1} = a^* + b(X_t - \rho X_{t-1}) + \varepsilon_t$$

where $a^* = a_0(1-\rho)$.
Re-write the equation as

$$\text{(A) } Y_t = a^* + \rho Y_{t-1} + bX_t - \rho b X_{t-1} + \varepsilon_t$$

Now for something completely different, namely the ADL model,

(B) $Y_t = a_0 + a_1 Y_{t-1} + b_0 X_t - b_1 X_{t-1} + \varepsilon_t$

Model (A) is actually model (B) with $b_1 = -b_0 a_1$ since $a_1 = \rho$ and $b_1 = \rho b_0$. The Cochrane-Orcutt model (A) is true only if $b_1 = -b_0 a_1$ or $b_1 + b_0 a_1 = 0$ .This is a testable hypothesis.

Because the restriction multiplies two parameters, we can't use a simple F-test. Instead we use a likelihood ratio test. The likelihood ratio is a function of the residual sum of squares. So, we run the Cochrane-Orcutt model (using the ,corc option because we don't want to keep the first observation) and save the residual sum of squares, ESS(A). Then we run the ADL model and keep its residual sum of squares, ESS(B). If the Cochrane-Orcutt model is true, then ESS(A)=ESS(B). This implies that the ratio, ESS(B)/ESS(A)=1 and the log of this ratio, ln(ESS(A)/ESS(B))=0. This is the likelihood ratio. It turns out that

$$L = T \ln\left[\frac{ESS(B)}{ESS(A)}\right] \sim \chi^2_q$$

where q=K(B)-K(A), K(B) is the number of parameters in model (A) and K(B) is the number of parameters in model (B). If this statistic is not significant, then the Cochrane-Orcutt model is true. Otherwise, you should use the ADL model.

We have already estimated the Cochran-Orcutt model (model A). The residual sum of squares is .0986, but we can get Stata to remember this by using scalars.

```
. scalar ESSA=e(rss)
```

Now let's estimate the ADL model (model B).

```
. regress lcrmaj prison L.lcrmaj L.prison

      Source |       SS       df       MS              Number of obs =      27
-------------+------------------------------           F(  3,    23) =   50.18
       Model |   .64494144     3   .21498048           Prob > F      =  0.0000
    Residual |  .098530274    23  .004283925           R-squared     =  0.8675
-------------+------------------------------           Adj R-squared =  0.8502
       Total |  .743471714    26  .028595066           Root MSE      =  .06545

------------------------------------------------------------------------------
lcrmaj       |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
prison       |  -.1457706   .1082056    -1.35   0.191     -.369611    .0780698
lcrmaj       |
         L1  |   .6515369   .1670076     3.90   0.001     .3060554    .9970184
prison       |
         L1  |   .0883118   .1116564     0.79   0.437    -.1426671    .3192906
_cons        |   3.393491   1.611994     2.11   0.046     .0588285    6.728154
------------------------------------------------------------------------------
```

Now do the likelihood ratio test.

```
. scalar ESSA=e(rss)

. scalar L=27*log(ESSB/ESSA)

.. scalar list
       L = -.02934373
    ESSA =  .09863742
    ESSB =  .09853027

. scalar prob=1-chi2(2,-.029)
```

```
. scalar list prob
    prob =         1
```

Since ESSA and ESSB are virtually identical, the test statistic is not significant and Cochrane-Orcutt is justified in this case.

Curing autocorrelation with lags

Despite the fact that Cochrane-Orcutt appears to be justified in the above example, I do not recommend using it. The problem is that it is a purely mechanical solution. It has nothing to do with economics. As we noted above, the most common cause of autocorrelation is omitted time varying variables. It is much more satisfying to specify the problem allowing for lags, momentum, reaction times, etc., that are very likely to be causing dynamic behavior than using a simple statistical fix. Besides, in the above example, the Cochrane-Orcutt method did not solve the problem. The Durbin-Watson statistic still showed some autocorrelation even after FGLS. The reason, almost certainly, is that the model omits too many variables, all of which are likely to be time-varying. For these reasons, most modern time series analysts do not recommend Cochrane-Orcutt. Instead they recommend adding variables and lags until the autocorrelation is eliminated.

We have seen above that the ADL model is a generalization of the Cochrane-Orcutt model. Since using Cochrane-Orcutt involves restricting the ADL to have a common factor, a restriction that may or may not be true, why not simply avoid the problem by estimating the ADL directly? Besides, the ADL model allows us to model the dynamic behavior directly.

So, I recommend adding lags, especially lags of the dependent variable until the LM test indicates no significant autocorrelation. Use the ten percent significance level when using this, or any other, specification test. For example, adding two lags of crime to the static crime equation eliminates the autocorrelation.

```
. regress lcrmaj prison L.lcrmaj LL.lcrmaj

      Source |       SS       df       MS              Number of obs =      28
-------------+------------------------------           F(  3,    24) =   59.71
       Model |  .659768412     3  .219922804           Prob > F      =  0.0000
    Residual |   .08839505    24  .003683127           R-squared     =  0.8819
-------------+------------------------------           Adj R-squared =  0.8671
       Total |  .748163462    27  .027709758           Root MSE      =  .06069

------------------------------------------------------------------------------
lcrmaj       |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
prison       |  -.0635699   .0208728    -3.05   0.006    -.1066493   -.0204905
lcrmaj       |
          L1 |   .9444664   .2020885     4.67   0.000     .5273762    1.361557
          L2 |  -.3882994   .1902221    -2.04   0.052    -.7808986    .0042997
_cons        |   4.293413   1.520191     2.82   0.009     1.155892    7.430934
------------------------------------------------------------------------------

. bgodfrey

Breusch-Godfrey LM test for autocorrelation
---------------------------------------------------------------------------
    lags(p)  |          chi2               df                 Prob > chi2
-------------+-------------------------------------------------------------
       1     |          0.747               1                   0.3875
```

```
-------------------------------------------------------------------------------
                      H0: no serial correlation

. bgodfrey, lags (1 2 3 4)

Breusch-Godfrey LM test for autocorrelation
-------------------------------------------------------------------------------
    lags(p)  |          chi2               df                 Prob > chi2
-------------+-----------------------------------------------------------------
       1     |          0.747               1                   0.3875
       2     |          2.220               2                   0.3296
       3     |          2.284               3                   0.5156
       4     |          6.046               4                   0.1957
-------------------------------------------------------------------------------
                      H0: no serial correlation. regress lcrmaj prison
```

Heteroskedastic and autocorrelation consistent standard errors

It is possible to generalize the concept of robust standard errors to include a correction for autocorrelation.

In the chapter on heteroskedasticity we derived the formula for heteroskedastic consistent standard errors.

$$Var(\hat{\beta}) = \frac{1}{\left(\sum x_t^2\right)^2} E\left(x_1^2 u_1^2 + x_2^2 u_2^2 + \ldots + x_T^2 u_T^2 + x_1 x_2 u_1 u_2 + x_1 x_3 u_1 u_3 + x_1 x_4 u_1 u_4 + \ldots\right)$$

.

In the absence of autocorrelation, $E(u_i u_j) = 0$ for $i \neq j$ (independent). The OLS estimator ignores these terms and is unbiased. Note that the formula can be rewritten as follows.

$$Var(\hat{\beta}) = \frac{1}{\left(\sum x_t^2\right)^2} \begin{pmatrix} x_1^2 Var(u_1) + x_2^2 Var(u_2) + \ldots + x_n^2 Var(u_T) u_T^2 \\ + x_1 x_2 Cov(u_1 u_2) + x_1 x_3 Cov(u_1 u_3) + x_1 x_4 Cov(u_1 u_4) + \ldots \end{pmatrix}$$

If the covariances are positive (positive autocorrelation) and the x's are positively correlated, which is the usual case, then ignoring them will cause the standard error to be underestimated. If the covariances are negative while the x's are positively correlated, then ignoring them will cause the standard error to be overestimated. If some are positive and some negative, the direction of bias of the standard error will be unknown. Things are made more complicated if the observations on x are negatively correlated. The one thing we do know is that if we ignore the covariances when they are nonzero, then ols estimates of the standard errors in no longer unbiased. This is the famous second order bias of the standard errors under autocorrelation. Typically, everything is positively correlated, the standard errors are underestimated and the corresponding t-ratios are overestimated, making variables appear to be related when they are in fact independent.

So, if we have autocorrelation, then $E(u_i u_j) \neq 0$ and we cannot ignore these terms. The problem is, if we use all these terms, the standard errors are not consistent because, as the sample size goes to infinity, the number of terms goes to infinity. Since each tem contains error, the amount of error goes to infinity and the estimate is not consistent. We could use a small number of terms, but that doesn't do it either, because it ignores higher order autocorrelation. Newey and West suggested a formula that increased the number of terms as the sample size increased, but at a much slower rate. The result is that the estimate is consistent

because the number of terms goes to infinity, but always with fewer than T terms. The formula relies on something known as the Bartlett kernel, but that is beyond the scope of this manual.

We can make this formula operational by using the residuals from the OLS regression to estimate u. To get heteroskedasticity and autocorrelation (HAC aka Newey-West) consistent standard errors and t-ratios, use the newey command.

```
. newey lcrmaj prison, lag(2)

Regression with Newey-West standard errors          Number of obs  =        28
maximum lag: 2                                      F(  1,    26)  =     60.80
                                                    Prob > F       =    0.0000

------------------------------------------------------------------------------
             |             Newey-West
      lcrmaj |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      prison |  -.1345066   .0172496    -7.80   0.000    -.1699636   -.0990496
       _cons |   9.663423   .0498593   193.81   0.000     9.560936    9.765911
------------------------------------------------------------------------------
```

Note that you must specify the number of lags in the newey command. We can also use Newey-West HAC standard errors in our ADL equation, even after autocorrelation has presumably been eliminated.

```
. newey lcrmaj prison L.lcrmaj LL.crmaj, lag(2)

Regression with Newey-West standard errors          Number of obs  =        28
maximum lag: 2                                      F(  3,    24)  =     66.77
                                                    Prob > F       =    0.0000

------------------------------------------------------------------------------
             |             Newey-West
lcrmaj       |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
prison       |  -.0608961   .0193394    -3.15   0.004    -.1008106   -.0209816
lcrmaj       |
          L1 |   .9678325   .1822947     5.31   0.000     .5915948     1.34407
crmaj        |
          L2 |  -.0000335   .0000103    -3.25   0.003    -.0000548   -.0000123
_cons        |   .8272894    1.67426     0.49   0.626    -2.628212    4.282791
------------------------------------------------------------------------------
```

Summary

David Hendry, one of the leading time series econometricians, suggests the following general to specific methodogy. Start with a very general ADL model with lot's of lags. A good rule of thumb is at least two lags for annual data, five lags for quarterly data, and 13 lags for monthly data. Check the Breusch-Godfrey test to make sure there is no autocorrelation. If there is, add more lags. Once you have no autocorrelation, you can begin reducing the model to a parsimonious, interpretable model. Delete insignificant variables. Justify your deletions with t-tests and F-tests. Make sure that dropping variables does not cause autocorrelation. If it does, then keep the variable even if it is not significant. When you get down to a parsimonious model with mostly significant variables, make sure that all of the omitted variables are not significant with an overall F-test. Make sure that there is no autocorrelation in the final version of the model. Use Newey-West heteroskedastic and autocorrelation robust t-ratios.

15 NONSTATIONARITY, UNIT ROOTS, AND RANDOM WALKS

The Gauss-Markov assumptions require that the error term is distributed as $iid(0,\sigma^2)$. We have seen how we can avoid problems associated with non-independence (autocorrelation) and non-identical (heteroskedasticity). However, we have maintained the assumption that the mean of the distribution is constant (and equal to zero). In this section we confront the possibility that the error mean is not constant.

A time series is stationary if its probability distribution does not change over time. The probability of observing any given y is conditional on all past values, $\Pr(y_t \mid y_0, y_1, ..., y_{t-1})$. A stationary process is one where this distribution doesn't change over time:

$$\Pr(y_t, y_{t+1}, ..., y_{t+p}) = \Pr(y_{t+q}, ..., y_{t+p+q})$$

for any t, p, or q. If the distribution is constant over time, then its mean and variance are constant over time. Finally, the covariances of the series must not be dependent on time, that is

$$Cov(y_t, y_{t+p}) = Cov(y_{t+q}, y_{t+p+q}).$$

If a series is stationary, we can infer its properties from examination of its histogram, its sample mean, its sample variance, and its sample autocovariances. On the other hand, if a series is not stationary, then we can't do any of these things. If a time series is nonstationary then we cannot consider a sample of observations over time as representative of the true distribution, because the truth at time t is different from the truth at time t+p.

Consider a simple deterministic time trend. It is called deterministic because it can be forecast with complete accuracy.

$$y_t = \alpha + \delta t$$

Is y stationary? The mean of y is

$$E(y_t) = \mu_t = \alpha + \delta t$$

Clearly, the mean changes over time. Therefore, a linear time trend is not stationary.

Now consider a simple random walk.

$y_t = y_{t-1} + \varepsilon_t, \varepsilon_t \sim iid(0, \sigma^2)$

For the sake of argument, let's assume that y=0 when t=0. Therefore,

$y_1 = 0 + \varepsilon_1$

$y_2 = 0 + \varepsilon_1 + \varepsilon_2$

$y_3 = 0 + \varepsilon_1 + \varepsilon_2 + \varepsilon_3$

so that

$y_t = \varepsilon_1 + \varepsilon_2 + ... + \varepsilon_t$

Thus, the variance of y_t is

$E(y_t)^2 = E(\varepsilon_1 + \varepsilon_2 + ... + \varepsilon_t)^2 = E(\varepsilon_1)^2 + E(\varepsilon_2)^2 + ... + E(\varepsilon_t)^2$

because of the independence assumption. Also, because of the identical assumption, $\sigma_1^2 = \sigma_2^2 = ... = \sigma_t^2$ which implies that

$E(\varepsilon_t)^2 = Var(\varepsilon_t) = t\sigma^2$.

Obviously, the variance of y depends on time and y cannot be stationary. Also, note that the variance of y goes to infinity as t goes to infinity. A random walk has potentially infinite variance.

A random walk is also known as a stochastic trend. Unlike the deterministic trend above, we cannot forecast future values of y with certainty. Also unlike deterministic trends, stochastic trends can exhibit booms and busts. Just because a random walk is increasing for several time periods does not mean that it will necessarily continue to increase. Similarly, just because it is falling does not mean we can expect it to continue to fall. Many economic time series appear to behave like stochastic trends. As a result most time series analysts mean stochastic trend when they say trend.

As we saw in the section on linear dynamic models, a random walk can be written as

$y_t - y_{t-1} = \Delta y_t = (1 - L) y_t = \varepsilon_t$

The lag polynomial is (1-L), so the root of the lag polynomial is one and y is said to have a unit root. Unit root and stochastic trend are used interchangeably.

Random walks and ordinary least squares

If we estimate a regression where the independent variable is a random walk, then the usual t- and F-tests are invalid. The reason is that if the regressor is nonstationary, it has no stable distribution, so the standard errors and t-ratios are not derived from a stable distribution. There is no distribution from which the statistics can be derived. The problem is similar to autocorrelation, only worse, because the autocorrelation coefficient is equal to one, extreme autocorrelation.

Granger and Newbold published an article in 1974 where they did some Monte Carlo exercises. They generated two random walks completely independently of one another and then regressed one on the other. Even though the true relationship between the two variables was zero, Granger and Newbold got t-ratios around 4.5. They called this phenomenon "spurious regression."

We can replicate their findings in Stata. In the crimeva.dta data set we first generate a random walk called y1. Start it out at y1=0 in the first year, 1956.

```
. gen y1=0
```

This creates a series of zero from 1956 to 2005. Now create the random walk by making y1 equal to itself lagged plus a random error term. We use the replace command to replace the zeroes. We have to start in 1957 so we have a previous value (zero) to start from. The function 5*invnorm(uniform()) produces a random number from a normal distribution with a mean of zero and a standard deviation of 5.

```
. replace y1=L.y1+5*invnorm(uniform()) if year > 1956
(49 real changes made)
```

Repeat for y2.

```
. gen y2=0

. replace y2=L.y2+5*invnorm(uniform()) if year > 1956
(49 real changes made)
```

Obviously, y1 and y2 are independent of each other. So if we regressed y1 on y2 there should be no significant relationship.

```
. regress y1 y2

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =   55.23
       Model |  9328.23038     1  9328.23038           Prob > F      =  0.0000
    Residual |  8107.68518    48  168.910108           R-squared     =  0.5350
-------------+------------------------------           Adj R-squared =  0.5253
       Total |  17435.9156    49  355.835011           Root MSE      =  12.997

------------------------------------------------------------------------------
          y1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          y2 |  -.9748288   .1311767    -7.43   0.000    -1.238577   -.7110805
       _cons |  -14.82044   3.009038    -4.93   0.000    -20.87052   -8.770368
```

This is a spurious regression. It looks like y1 is negatively (and highly significantly) associated with y2 even though they are completely independent.

Therefore, to avoid spurious regressions, we need to know if our variables have unit roots.

Testing for unit roots

The most widely used test for a unit root is the Augmented Dickey-Fuller (ADF) test. The idea is very simple. Suppose we start with a simple AR(1) model.

$y_t = a_0 + a_1 y_{t-1} + \varepsilon_t$

Subtract y_{t-1} from both sides,

$y_t - y_{t-1} = a_0 + a_1 y_{t-1} - y_{t-1} + \varepsilon_t$

or

$\Delta y_t = a_0 + (a_1 - 1) y_{t-1} + \varepsilon_t$

This is the (nonaugmented) Dickey-Fuller test equation, known as a DF equation. Note that if the AR model is stationary, $|a_1| < 1$, in which case $a_1 - 1 < 0$. On the other hand, if y has a unit root then $a_1 = 1$ and $a_1 - 1 = 0$. So we can test for a unit root using a t-test on the lagged y.

Since most time series are autocorrelated, and autocorrelation creates overestimated t-ratios, we should do something to correct for autocorrelation. Consider the following AR(1) model with first order autocorrelation.

$y_t = a_1 y_{t-1} + u_t$

$u_t = \rho u_{t-1} + \varepsilon_t, \varepsilon_t \sim iid(0, \sigma^2)$

But the error term u_t is also

$$u_t = y_t - a_1 y_{t-1}$$

$$u_{t-1} = y_{t-1} - a_1 y_{t-2}$$

$$\rho u_{t-1} = \rho y_{t-1} - \rho a_1 y_{t-2}$$

Therefore,

$$y_t = a_1 y_{t-1} + u_t$$

$$y_t = a_1 y_{t-1} + \rho y_{t-1} - \rho a_1 y_{t-2} + \varepsilon_t$$

Subtract y_{t-1} from both sides to get

$$y_t - y_{t-1} = a_1 y_{t-1} - y_{t-1} + \rho y_{t-1} - \rho a_1 y_{t-2} + \varepsilon_t$$

Add and subtract $\rho a_1 y_{t-1}$

$$\Delta y_t = a_1 y_{t-1} - y_{t-1} + \rho a_1 y_{t-1} - \rho a_1 y_{t-1} + \rho y_{t-1} - \rho a_1 y_{t-2} + \varepsilon_t$$

Collect terms

$$\Delta y_t = (a_1 - 1 + \rho - \rho a_1) y_{t-1} + \rho a_1 (y_{t-1} - y_{t-2}) + \varepsilon_t$$

$$\Delta y_t = (a_1 - 1)(1 - \rho) y_{t-1} + \rho a_1 \Delta y_{t-1} + \varepsilon_t$$

$$\Delta y_t = \gamma y_{t-1} + \delta \Delta y_{t-1} + \varepsilon_t$$

If $a_1 = 1$ (unit root) then γ=0. We can test this hypothesis with a t-test. Note that the lagged difference has eliminated the autocorrelation, generating an iid error term. If there is higher order autocorrelation, simply add more lagged differences. All we are really doing is adding lags of the dependent variable to the test equation. We know from our analysis of autocorrelation that this is a cure for autocorrelation. The lagged differences are said to "augment" the Dickey-Fuller test.

The ADF in general form is

$$\Delta y_t = a_0 + (a_1 - 1) y_{t-1} + c_1 \Delta y_{t-1} + c_2 \Delta y_{t-2} + \ldots + c_p \Delta y_{t-p} + \varepsilon_t$$

I keep saying that we test the hypothesis that $(a_1 - 1) = 0$ with the usual t-test. The ADF test equation is estimated using OLS. The t-ratio is computed as usual, the estimated parameter divided by its standard error. However, while the t-ratio is standard, its distribution is anything but standard. For reasons that are explained in the last section of this chapter, the t-ratio in the Dickey-Fuller test equation has a non-standard distribution. Dickey and Fuller, and other researchers, have tabulated critical values of the DF t-ratio. These critical values are derived from Monte Carlo exercises. The DF critical values are tabulated in many econometrics books and are automatically available in Stata.

As an example, we can do our own mini Monte Carlo exercise using y_1 and y_2, the random walks we created above in the crimeva.dta data set. Dickey-Fuller unit root tests can be done with the dfuller command. Since we know there is no autocorrelation, we will not add any lags.

```
. dfuller y1, lags(0)

Dickey-Fuller test for unit root                   Number of obs   =        49

                                 ---------- Interpolated Dickey-Fuller ---------
                    Test         1% Critical       5% Critical      10% Critical
                  Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -1.761            -3.587            -2.933            -2.601
------------------------------------------------------------------------------
* MacKinnon approximate p-value for Z(t) = 0.4001
```

Note that the critical values are much higher (in absolute value) than the usual t distribution. We cannot reject the null hypothesis of a unit root, as expected. We should always have an intercept in any real world

application, but we know that we did not use an intercept when we constructed y1, so let's rerun the DF test with no constant.

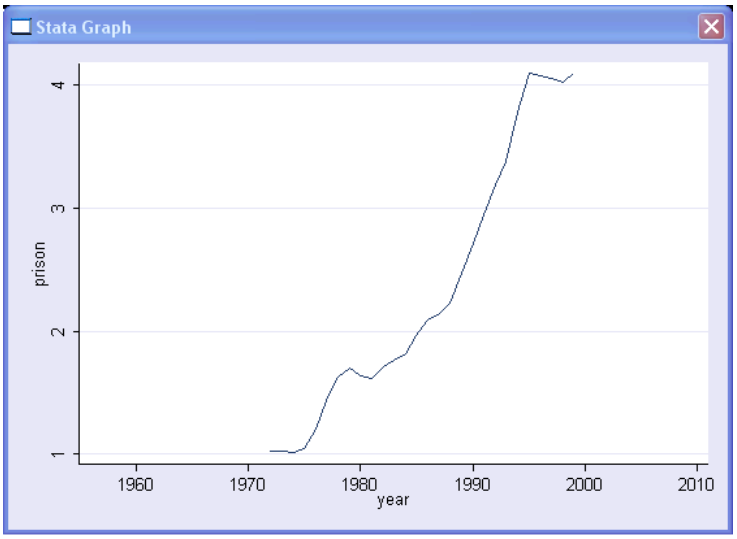
```
. dfuller y1, noconstant lags(0)

Dickey-Fuller test for unit root                    Number of obs   =        49

                                 ---------- Interpolated Dickey-Fuller ---------
                 Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -0.143            -2.622            -1.950            -1.610
```

The t-ratio is approximately zero, as expected. Note the critical values. They are much higher than the corresponding criticial values for the standard t-distribution.

So far we have tested for a unit root where the alternative hypothesis is that the variable is stationary around a constant value. If a variable has a distinct trend, and many economic variables, like GDP, have a distinct upward trend, then we need to be able to test for a unit root against the alternative that y is a stationary deterministic trend variable. To do this we simply add a deterministic trend term to the ADF test equation.

$$\Delta y_t = a_0 + (a_1 - 1) y_{t-1} + \delta t + c_1 \Delta y_{t-1} + c_2 \Delta y_{t-2} + ... + c_p \Delta y_{t-p} + \varepsilon_t$$

The t-ratio on $(a_1 - 1)$ is again the test statistic. However, the critical values are different for test equations with deterministic trends. Fortunately, Stata knows all this and adjusts accordingly. Since it makes a difference whether the variable we are testing for a unit root has a trend or not, it is a good idea to graph the variable and look at it. If it has a distinct trend, include a trend in the test equation. For example, we might want to test the prison population (in the crimeva.dta data set) for stationarity. The first thing we do is graph it.



This variable has a distinct trend. Is it a deterministic trend or a stationary trend? Let's test for a unit root using an ADF test equation with a deterministic trend. We choose two lags to start with and add the regress option to see the test equation.

```
. dfuller prison, lags(4) regress trend

Augmented Dickey-Fuller test for unit root          Number of obs   =        23
```

```
                                 ---------- Interpolated Dickey-Fuller ---------
                 Test            1% Critical       5% Critical      10% Critical
               Statistic            Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -1.328            -4.380            -3.600            -3.240
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.8809

------------------------------------------------------------------------------
D.prison     |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
prison       |
          L1 |  -.1454225   .1095418    -1.33   0.203    -.3776407    .0867956
          LD |   .9004267   .2221141     4.05   0.001      .429566    1.371288
         L2D |  -.4960933   .3426802    -1.45   0.167    -1.222543    .2303562
         L3D |   .3033914   .2941294     1.03   0.318    -.3201351    .9269179
         L4D |  -.1248609   .3584488    -0.35   0.732    -.8847383    .6350165
_trend       |   .0214251   .0127737     1.68   0.113    -.0056539     .048504
_cons        |   .0723614   .0643962     1.12   0.278    -.0641524    .2088751
------------------------------------------------------------------------------
```

How many lags should we include? The rule of thumb is at least two for annual data, five for quarterly data, etc. Since these data are annual, we start with two. It is of some importance to get the number of lags right because too few lags causes autocorrelation, which means that our OLS estimates are biased and inconsistent, with overestimated t-ratios. Too many lags means that our estimates are inefficient with underestimated t-ratios. Which is worse? Monte Carlo studies have shown that including too few lags is much worse. There are three approaches to this problem. The first is to use t-tests and F-tests on the lags to decide whether to drop them or not. The second is to use model selection criteria. The third approach is to do it all possible ways and see if we get the same answer. The problem with the third approach is that, if we get different answers, we don't know which to believe.

Choosing the number of lags with F tests

I recommend using F-tests to choose the lag length. Look at the t-ratios on the lags and drop the insignificant lags (use a .15 significance level to be sure not to drop too many lags). The t- and F-ratios on the lags are standard, so use the standard critical values provided by Stata. It appears that the last three lags are not significant. Let's test this hypothesis with an F-test.

```
. test L4D.prison L3D.prison L2D.prison

 ( 1)  L4D.prison = 0
 ( 2)  L3D.prison = 0
 ( 3)  L2D.prison = 0

       F(  3,    16) =    0.86
            Prob > F =    0.4835
```

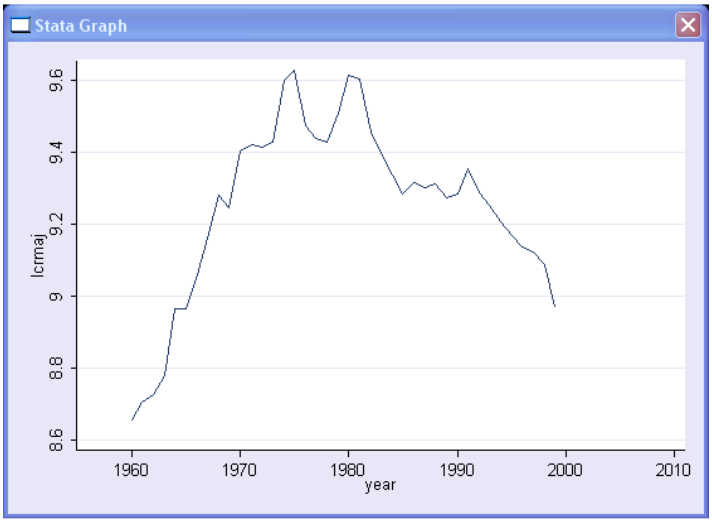
So, the last three lags are indeed not significant and we settle on one lag as the correct model.

```
. dfuller prison, lags(1) regress trend

Augmented Dickey-Fuller test for unit root         Number of obs   =        26

                                 ---------- Interpolated Dickey-Fuller ---------
                 Test            1% Critical       5% Critical      10% Critical
```

```
                 Statistic           Value              Value              Value
------------------------------------------------------------------------------
 Z(t)              -2.352            -4.371            -3.596             -3.238
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.4055

------------------------------------------------------------------------------
D.prison     |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
prison       |
          L1 |  -.1608155   .0683769    -2.35   0.028    -.3026204   -.0190106
          LD |   .6308502     .15634     4.04   0.001     .3066209    .9550795
       _trend |   .0211228   .0091946     2.30   0.031     .0020544    .0401913
       _cons |   .1145497   .0493077     2.32   0.030     .0122918    .2168075
------------------------------------------------------------------------------
```

Note that the lag length apparently doesn't matter because the DF test statistic is not significant in either model. Prison population in Virginia appears to be a random walk.

In some cases it is not clear whether to use a trend or not. Here is the graph of the log of major crime in Virginia.



Does it have a trend, two trends (one up, one down), or no trend? Let's start with a trend and four lags.

```
. dfuller lcrmaj, lags(4) regress trend

Augmented Dickey-Fuller test for unit root         Number of obs   =        35

                                 ---------- Interpolated Dickey-Fuller ---------
                    Test         1% Critical       5% Critical      10% Critical
                 Statistic           Value             Value            Value
------------------------------------------------------------------------------
 Z(t)              -1.832            -4.288            -3.560             -3.216
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.6889

------------------------------------------------------------------------------
D.lcrmaj     |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lcrmaj       |
          L1 |  -.1243068   .0678357    -1.83   0.078    -.2632618    .0146483
          LD |  -.0308934   .1746865    -0.18   0.861    -.3887225    .3269358
```

```
         L2D |  -.2231895    .171358    -1.30   0.203    -.5742004    .1278214
         L3D |  -.1747988   .1704449    -1.03   0.314    -.5239392    .1743417
         L4D |  -.0359288    .173253    -0.21   0.837    -.3908215    .3189639
_trend       |  -.0053823   .0018615    -2.89   0.007    -.0091955   -.0015691
_cons        |   1.282789    .640328     2.00   0.055    -.0288636    2.594441
------------------------------------------------------------------------------
```

The trend is significant, so we retain it. Let's test for the significance of the lags.

```
. test L4D.lcrmaj L3D.lcrmaj L2D.lcrmaj LD.lcrmaj

 ( 1)  L4D.lcrmaj = 0
 ( 2)  L3D.lcrmaj = 0
 ( 3)  L2D.lcrmaj = 0
 ( 4)  LD.lcrmaj = 0

       F(  4,    28) =    0.72
            Prob > F =    0.5879
```

None of the lags are significant, so we drop them all and run the test again.

```
. dfuller lcrmaj, lags(0) regress trend

Dickey-Fuller test for unit root                   Number of obs   =        39

                                 ---------- Interpolated Dickey-Fuller ---------
                    Test         1% Critical       5% Critical      10% Critical
                Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -1.314            -4.251            -3.544            -3.206
------------------------------------------------------------------------------
MacKinnon approximate p-value for Z(t) = 0.8845

------------------------------------------------------------------------------
D.lcrmaj     |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lcrmaj       |
          L1 |  -.0620711   .0472515    -1.31   0.197    -.1579016    .0337594
_trend       |  -.0030963   .0010268    -3.02   0.005    -.0051789   -.0010138
_cons        |   .6447613   .4310014     1.50   0.143    -.2293501    1.518873
------------------------------------------------------------------------------
```

Again, the number of lags does not make any real difference. The ADF test statistics are all insignificant, indicating that crime in Virginia appears to be a random walk.

Digression: model selection criteria

One way to think of the problem is to hypothesize that the correct model should have the highest R-square and the lowest error variance compared to the alternative models. Recall that the R-square measures the variation explained by the regression to the total variation of the dependent variable. The problem with this approach is that the R-square can be maximized by simply adding more independent variables.

One way around this problem is to correct the R-square for the number of explanatory variables used in the model. The adjusted R-square also known as the $\bar{R}^2$ (R-bar-square) is the ratio of the variance (instead of variation) explained by the regression to the total variance.

$$\bar{R}^2 = 1 - \frac{Var(e)}{Var(y)} = 1 - (1 - R^2)\frac{N-1}{T-k}$$

where T is the sample size and k is the number of independent variables. So, as we add independent variables, R^2 increases but so does N-k. Consequently, $\bar{R}^2$ can go up or down. One rule might be to choose the number of lags such that $\bar{R}^2$ is a maximum.

The Akaike information criterion (AIC) has been proposed as an improvement over the $\bar{R}^2$ on the theory that the $\bar{R}^2$ does not penalize heavily enough for adding explanatory variables. The formula is

$$AIC = \log\left(\frac{\sum e_i^2}{T}\right) + \frac{2k}{T}$$

The goal is to choose the number of explanatory variables (lags in the context of a unit root test) that minimizes the AIC.

The Bayesian Information Criterion (also known as the Schwartz Criterion),

$$BIC = \log\left(\frac{\sum e_i^2}{T}\right) + k\frac{\log(T)}{T}$$

The BIC penalizes additional explanatory variables even more heavily than the AIC. Again, the goal is to minimize the BIC.

Note that these information criteria are not formal statistical tests, however they have been shown to be useful in the determination of the lag length in unit root models. I usually use the BIC, but usually they both give the same answer.

Choosing lags using model selection criteria

Let's adopt the information criterion approach. If your version of Stata has the fitstat command, invoke fitstat after the dfuller command. If your version does not have fitstat, you can install it by doing a keyword search for fitstat and clicking on "click here to install." Fitstat generates a variety of goodness of fit tests, including the BIC. Then run the test several times with different lags and check the information criterion. The minimum (biggest negative) is the one to go with.

```
. dfuller prison, lags(2) trend

Augmented Dickey-Fuller test for unit root         Number of obs   =        25

                              ---------- Interpolated Dickey-Fuller ---------
                 Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -1.528           -4.380            -3.600            -3.240
------------------------------------------------------------------------------
* MacKinnon approximate p-value for Z(t) = 0.8182

. fitstat

Measures of Fit for regress of D.prison

Log-Lik Intercept Only:       18.085     Log-Lik Full Model:           27.328
D(20):                       -54.656     LR(4):                        18.486
                                         Prob > LR:                     0.001
R2:                            0.523     Adjusted R2:                   0.427
AIC:                          -1.786     AIC*n:                        -44.656
BIC:                        -119.034     BIC':                          -5.610
```

```
. dfuller prison, lags(1) trend

Augmented Dickey-Fuller test for unit root         Number of obs   =        26

                              ---------- Interpolated Dickey-Fuller ---------
                 Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)            -2.352           -4.371           -3.596            -3.238
------------------------------------------------------------------------------
* MacKinnon approximate p-value for Z(t) = 0.4069

. fitstat

Measures of Fit for regress of D.prison

Log-Lik Intercept Only:       18.727     Log-Lik Full Model:           27.580
D(22):                       -55.160     LR(3):                        17.706
                                         Prob > LR:                     0.001
R2:                            0.494     Adjusted R2:                   0.425
AIC:                          -1.814     AIC*n:                       -47.160
BIC:                       -126.838     BIC':                         -7.931

. dfuller prison, lags(0) trend

Dickey-Fuller test for unit root                   Number of obs   =        27

                              ---------- Interpolated Dickey-Fuller ---------
                 Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)            -1.647           -4.362           -3.592            -3.235
------------------------------------------------------------------------------
* MacKinnon approximate p-value for Z(t) = 0.7726

. fitstat

Measures of Fit for regress of D.prison

Log-Lik Intercept Only:       19.427     Log-Lik Full Model:           21.654
D(24):                       -43.308     LR(2):                         4.454
                                         Prob > LR:                     0.108
R2:                            0.152     Adjusted R2:                   0.081
AIC:                          -1.382     AIC*n:                       -37.308
BIC:                       -122.408     BIC':                          2.138
```

So we go with one lag. In this case it doesn't make any difference, so we are pretty confident that prison is a stochastic trend.

DF-GLS test

A souped-up version of the ADF which is somewhat more powerful has been developed by Elliott, Rothenberg, and Stock (ERS, 1996).[9] ERS use a two step method where the data are detrended using generalized least squares, GLS. The resulting detrended series is used in a standard ADF test.

The basic idea behind the test can be seen as follows. Suppose the null hypothesis is

H_0: $Y_t = Y_{t-1} + \delta + \varepsilon_t$ or $\Delta Y_t = \delta + \varepsilon_t$ where $\varepsilon_t \sim I(0)$

The alternative hypothesis is

H_1: $Y_t = \alpha + \beta t + \varepsilon_t$ or $Y_t - \alpha - \beta t = \varepsilon_t$ where $\varepsilon_t \sim I(0)$

If the null hypothesis is true but we assume the alternative is true and detrend,

$$Y_t - \alpha - \beta t = Y_{t-1} + (\delta - \alpha) - \beta t + \varepsilon_t$$

Y is still a random walk.

So, the idea is to detrend Y and then test the resulting series for a unit root. If we reject the null hypothesis of a unit root, then the series is stationary. If we cannot reject the null hypothesis of a unit root, then Y is a nonstationary random walk.

What about autocorrelation? ERS suggest correcting for autocorrelation using a generalized least squares procedure, similar to Cochrane-Orcutt. The procedure is done in two steps. In the first step we transform to generalized first differences.

$$Y_t^* = Y_t - \rho Y_{t-1}$$

where ρ=1-7/T if there is no obvious trend and ρ=1-13.5/T if there is a trend. The idea behind this choice of ρ is that ρ should go to one as T goes to infinity. The values 7 and 13.5 were chosen with Monte Carlo methods.

$$\begin{aligned} Y_t &= \alpha + \beta t \\ \rho Y_{t-1} &= \rho\alpha + \rho\beta(t-1) \\ Y_t - \rho Y_{t-1} &= \alpha + \beta t - \rho\alpha + \rho\beta(t-1) \\ &= \alpha(1-\rho) + \beta(t - \rho(t-1)) \end{aligned}$$

This equation is estimated with OLS.

We compute the detrended series,

[9] Elliot, G., T. Rothenberg, and J. H. Stock. 1996. Efficient tests for an autoregressive unit root. Econometrica 64: 813-836.

$Y_t^d = Y_t - \hat{\alpha} - \hat{\beta}t$

if there is a trend and

$Y_t^d = Y_t - \hat{\alpha}$
if no trend.

Finally, we use the Dickey-Fuller test (not the augmented DF), with no intercept, to test Y_t^d for a unit root.

The DF-GLS test is available in Stata with the dfgls command. Let's start with 4 lags just to make sure.

```
. dfgls lcrmaj, maxlag(4)

DF-GLS for lcrmaj            Number of obs =    35

                DF-GLS tau      1% Critical       5% Critical      10% Critical
  [lags]     Test Statistic        Value             Value             Value
------------------------------------------------------------------------------
    4            -0.752           -3.770            -3.072            -2.772
    3            -0.527           -3.770            -3.147            -2.843
    2            -0.443           -3.770            -3.215            -2.906
    1            -0.456           -3.770            -3.273            -2.960

Opt Lag (Ng-Perron seq t) = 0 [use maxlag(0)]
Min SC   = -5.028735 at lag  1 with RMSE  .0730984
Min MAIC = -5.161388 at lag  1 with RMSE  .0730984
```

What if we drop the trend?

```
. dfgls lcrmaj, maxlag(4) notrend

DF-GLS for lcrmaj            Number of obs =    35

                DF-GLS mu       1% Critical       5% Critical      10% Critical
  [lags]     Test Statistic        Value             Value             Value
------------------------------------------------------------------------------
    4            -0.965           -2.636            -2.236            -1.937
    3            -0.800           -2.636            -2.275            -1.975
    2            -0.753           -2.636            -2.312            -2.010
    1            -0.769           -2.636            -2.346            -2.041

Opt Lag (Ng-Perron seq t) = 0 [use maxlag(0)]
Min SC   = -5.056724 at lag  1 with RMSE  .0720825
Min MAIC = -5.166232 at lag  1 with RMSE  .0720825
```

Note that the SC is the Schwartz Criterion (BIC). The MAIC is the modified Akaike Information Criterion, much like the BIC. Again, all tests appear to point to a unit root for the crime variable.

I recommend the DF-GLS test for all your unit root testing needs.

Trend stationarity vs. difference stationarity

We know that a regression of one independent random walk on another is likely to result in a spurious regression where the t-ratio is highly significant according to the usual critical values. So we need to know if our variables are random walks or not. Unfortunately, the unit roots that have been developed so far are

not particularly powerful, that is, they cannot distinguish between unit roots and nearly unit roots that are nevertheless stationary. For example, it would be difficult for an ADF test equation to tell the difference between a_1=1 and a_1=.9, even though the former is a random walk and the latter is stationary. What happens if we get it wrong?

First a bit of terminology: a stationary random variable is said to be "stationary of order zero," of I(0). A random walk that can be made stationary by taking first differences is said to be "integrated of order one," or I(1). Most economic time series are either I(0) or I(1). There is the odd time series that is I(2), which means it has to be differenced twice to make it stationary. To avoid spurious regressions, we need the variables in our models to be I(0). If the variable is a random walk, we take first differences, if it is already I(0) it is said to be "stationary in levels," and we can use it as is.

Consider the following table.

Truth	Levels	First differences
I(0)	OK	Negative autocorrelation
I(1)	Spurious regression	OK

Clearly, if we have a random walk and we take first differences, or we run a regression in levels with stationary variables, we are OK. However, if we make a mistake and run a regression in levels with I(1) variables we get a spurious regression, which is really bad. On the other hand, if we mess up and take first differences of a stationary variable, it is still stationary. The error term is likely to have negative autocorrelation, but that is a smaller problem. We can simply use Newey-West HAC standard errors and t-ratios. Besides, most economic time series tend to have booms and busts and otherwise behave like stochastic trends rather than deterministic trends. For all these reasons, most time series analysts assume a variable is a stochastic trend unless convinced otherwise.

Why unit root tests have nonstandard distributions

To see this, consider the non-augmented Dickey-Fuller test equation with no intercept.

$$\Delta y_t = a_1 y_{t-1} + \varepsilon_t$$

$$\hat{a}_1 = \frac{\sum \Delta y_t y_{t-1}}{\sum y_{t-1}^2} = \frac{\frac{1}{T}\sum \Delta y_t y_{t-1}}{\frac{1}{T}\sum y_{t-1}^2}$$

Multiply both sides by $\mathrm{T}\left(=\frac{1}{\frac{1}{\mathrm{T}}}\right)$

$$T\hat{a}_1 = \frac{\frac{1}{T}\sum \Delta y_t y_{t-1}}{\frac{1}{T^2}\sum y_{t-1}^2}$$

Look at the numerator.

$$\sum y_t^2 = \sum \left(y_{t-1} + \Delta y_t \right)^2 = \sum y_{t-1}^2 + 2\sum y_{t-1}\Delta y_t + \sum (\Delta y_t)^2$$

Solve for the middle term.

$$2\sum y_{t-1}\Delta y_t = \sum y_t^2 - \sum y_{t-1}^2 - \sum(\Delta y_t)^2$$

$$\sum y_{t-1}\Delta y_t = \frac{1}{2}\left(\sum y_t^2 - \sum y_{t-1}^2 - \sum(\Delta y_t)^2\right)$$

$$\frac{1}{T}\sum y_{t-1}\Delta y_t = \frac{1}{2}\frac{1}{T}\left(\sum y_t^2 - \sum y_{t-1}^2 - \sum(\Delta y_t)^2\right)$$

The first two terms in the parentheses can be written as,

$$\sum_{t=1}^{T} y_t^2 = \left(\sum_{t=1}^{T-1} y_t^2 + y_T^2\right)$$

$$\sum_{t=1}^{T} y_{t-1}^2 = \left(y_0^2 + \sum_{t=1}^{T-1} y_t^2\right)$$

Which means that the difference between these two terms is

$$\sum_{t=1}^{T} y_t^2 - \sum_{t=1}^{T} y_{t-1}^2 = \left(\sum_{t=1}^{T-1} y_t^2 + y_T^2\right) - \left(y_0^2 + \sum_{t=1}^{T-1} y_t^2\right) = y_T^2 - y_0^2 = y_T^2$$

Because we can conviently assume that $y_0 = 0$.

Therefore,

$$\frac{1}{T}\sum y_{t-1}\Delta y_t = \frac{1}{T}\frac{1}{2}\left(y_T^2 - \sum\left(\Delta y_t^2\right)\right) = \frac{1}{2}\left(\left(\frac{y_T}{\sqrt{T}}\right)^2 - \frac{1}{T}\sum\left(\Delta y_t^2\right)\right)$$

Under the null hypothesis, $\Delta y_t = \varepsilon_t$ so that the second term in the parentheses is

$\frac{1}{T}\sum(\Delta y_t)^2 = \frac{1}{T}\sum(\varepsilon_t)^2$ has the probability limit $\frac{1}{T}\sum(\varepsilon_t)^2 \to \sigma^2$

Under the assumption that $y_0 = 0$ the first term is

$$\frac{y_T}{\sqrt{T}} = \sqrt{\frac{1}{T}\sum \Delta y_t} = \sqrt{\frac{1}{T}\sum \varepsilon_t}$$

Since the sum of T normal variables is normal,

$$\frac{y_T}{\sqrt{T}} \to N(0, \sigma^2)$$

Therefore,

$$\left(\frac{y_T}{\sqrt{T}}\right)^2 - \frac{1}{T}\sum\left(\varepsilon_t^2\right) \to \sigma^2\left(Z^2 - 1\right)$$

where Z is a standard normal variable. We know that the square of a normal variable is a chi-square variable with one degree of freedom. Therefore the numerator of the OLS estimator of the Dickey-Fuller test equation is

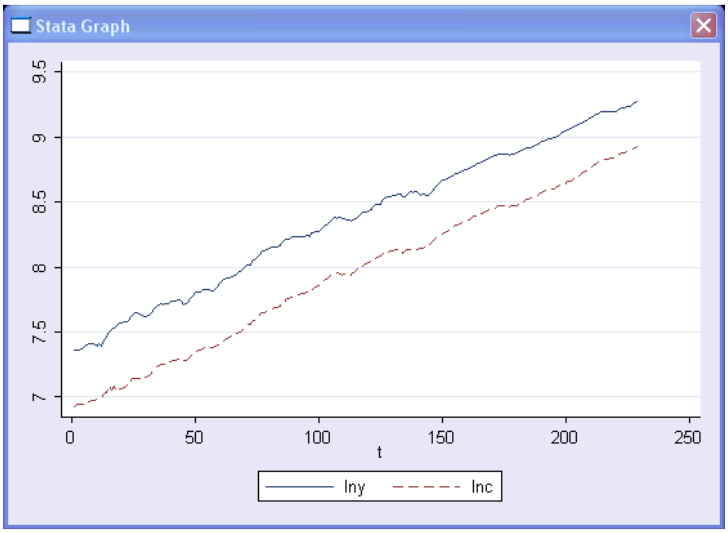
$$\frac{1}{T}\sum y_{t-1}\Delta y_t \to \frac{\sigma^2}{2}\left(\chi_1^2 - 1\right)$$

If y were stationary, the estimator would converge to a normal random variable. Because y is a random walk the numerator converges to a chi-square variable. Unfortunately, the numerator is the good news. The really bad news is the denominator. The sum of squares $\frac{1}{T}\sum y_{t-1}^2$ does not converge to a constant, but remains a random variable even as T goes to infinity. So, the OLS estimator in the Dickey-Fuller test equation converges to a ratio. The numerator is chi-square while the denominator is the distribution of a random variable. The latter distribution is different for different dependent variables.

16 ANALYSIS OF NONSTATIONARY DATA

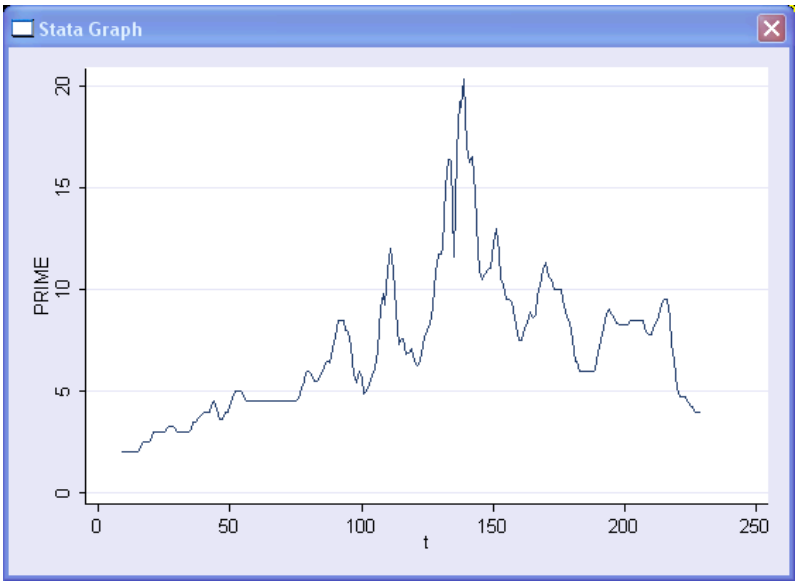
Consider the following graph of the log of GDP and the log of consumption. The data are quarterly, from 1947-1 to 2004-1. I created the time trend t from 1 to 229 corresponding to each quarter. These data are available in CYR.dta.

```
. gen t=_n

. tsset t
        time variable:  t, 1 to 229
```



Are these series random walks with drift (difference stationary stochastic trends) or are they deterministic trends?.

The graph of the prime rate of interest is quite different.



Since the interest rate has to be between zero and one, we assume it cannot have a deterministic trend. It does have the booms and busts that are characteristic of stochastic trends, so we suspect that it is a random walk.

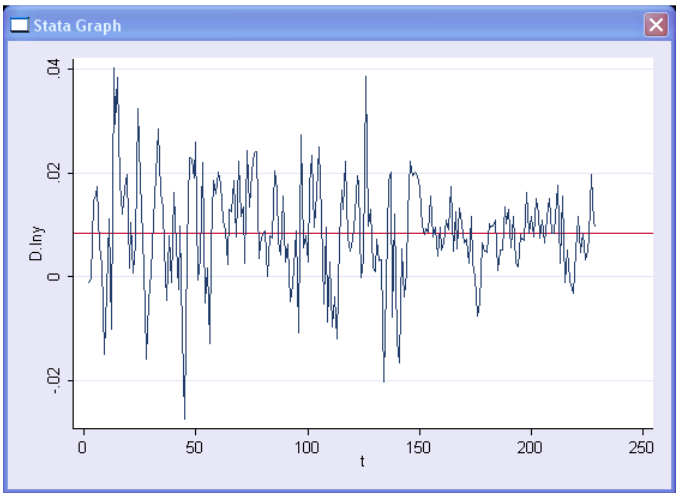
We apply unit root tests to these variables below.

```
. dfgls lny, maxlag(5)

DF-GLS for lny               Number of obs =   223

              DF-GLS tau      1% Critical       5% Critical      10% Critical
  [lags]     Test Statistic        Value             Value             Value
------------------------------------------------------------------------------
    5            -1.719           -3.480            -2.893            -2.607
    4            -1.892           -3.480            -2.901            -2.614
    3            -2.137           -3.480            -2.908            -2.620
    2            -2.408           -3.480            -2.914            -2.626
    1            -2.165           -3.480            -2.921            -2.632

Opt Lag (Ng-Perron seq t) =  1 with RMSE  .0094098
Min SC   = -9.283516 at lag  1 with RMSE  .0094098
Min MAIC = -9.289291 at lag  5 with RMSE  .0092561

. dfgls lnc, maxlag(5)

DF-GLS for lnc               Number of obs =   223

              DF-GLS tau      1% Critical       5% Critical      10% Critical
  [lags]     Test Statistic        Value             Value             Value
------------------------------------------------------------------------------
    5            -2.042           -3.480            -2.893            -2.607
    4            -2.084           -3.480            -2.901            -2.614
    3            -2.487           -3.480            -2.908            -2.620
    2            -2.482           -3.480            -2.914            -2.626
    1            -1.818           -3.480            -2.921            -2.632

Opt Lag (Ng-Perron seq t) =  4 with RMSE  .0079511
Min SC   = -9.572358 at lag  2 with RMSE  .0080462
Min MAIC = -9.589425 at lag  4 with RMSE  .0079511

. dfgls prime, maxlag(5)

DF-GLS for prime             Number of obs =   215
```

```
              DF-GLS tau      1% Critical       5% Critical      10% Critical
  [lags]     Test Statistic        Value             Value             Value
------------------------------------------------------------------------------
    5            -2.737           -3.480            -2.895            -2.609
    4            -2.179           -3.480            -2.903            -2.616
    3            -2.069           -3.480            -2.910            -2.623
    2            -1.768           -3.480            -2.917            -2.629
    1            -1.974           -3.480            -2.924            -2.635

Opt Lag (Ng-Perron seq t) =  5 with RMSE  .7771611
Min SC   =  -.390588 at lag  1 with RMSE  .8022992
Min MAIC = -.3949452 at lag  2 with RMSE  .8005883
```

All of these series are apparently random walks. We can transform them to stationary series by taking first differences. Here is the graph of the first differences of the log of income. First let's summarize the first differences of the logs of these variables.

```
. summarize D.lnc D.lny D.lnr

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
lnc          |
          D1 |       228    .008795    .0084876   -.029891     .05018
lny          |
          D1 |       228   .0084169    .0100605  -.0275517   .0402255
lnr          |
          D1 |       220   .0031507     .078973  -.3407288    .365536

. twoway (tsline D.lny), yline(.0084)
```



Note that the series is centered at .0084 (the mean of the differenced series, which is the average rate of growth, or "drift"). Note also that, if the series wanders too far from the mean, it tends to get drawn back to the center quickly. This is a typical graph for a stationary series.

Once we transform to stationary variables, we can use the usual regression analyses and significance tests.

```
regress D.lnc LD.lnc D.lny D.lnr

      Source |       SS       df       MS              Number of obs =     220
-------------+------------------------------           F(  3,   216) =   46.01
       Model |  .006260983     3  .002086994           Prob > F      =  0.0000
    Residual |  .009798288   216  .000045362           R-squared     =  0.3899
-------------+------------------------------           Adj R-squared =  0.3814
```

```
       Total |  .016059271   219   .00007333           Root MSE      =  .00674

------------------------------------------------------------------------------
D.lnc        |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lnc          |
          LD |  -.1957726   .0578761    -3.38   0.001    -.3098469   -.0816984
lny          |
          D1 |   .5741064    .049067    11.70   0.000     .4773951    .6708177
lnr          |
          D1 |  -.0141193   .0059641    -2.37   0.019    -.0258746    -.002364
_cons        |   .0057757   .0007042     8.20   0.000     .0043876    .0071637
------------------------------------------------------------------------------
```

The change in consumption is a function of lagged consumption, the change in income, and the change in interest rates. The coefficient on income is positive and the coefficient on the interest rate is negative, as expected. The negative coefficient on lagged consumption means that there is significant "reversion to the mean," that is, if consumption grows too rapidly in one quarter, it will tend to grow much less rapidly I the next quarter, and vice versa. Also, in a first difference regression, the intercept term is a linear time trend. Consumption appears to be growing exogenously at a rate of about one-half percent per quarter.

The problem with a first difference regression is that it is purely short run. Changes in lnc are functions of changes in lny. It doesn't say anything about any long run relationship between consumption and income. And economists typically don't know anything about the short run. All our hypotheses are about equilibria, which presumably occur in the long run. So, it would be very useful to be able estimate the long run relationship between consumption, income, and the interest rate.

Cointegration

Engle and Granger have developed a method for determining if there is a long run relationship among a set of nonstationary random walks. They noted that if two independent random walks were regressed on each other, the result would simply be another random walk consisting of a linear combination of two individual random walks. Imagine a couple of drunks who happen to exit Paul's Deli at the same time. They are walking randomly and we expect that they will wander off in different directions, so the distance between them (the residual) will probably get bigger and bigger. So, for example, if x(t) and z(t) are independent random walks, then a regression of y on x will yield z=a+bx+u, so that u=z-a-bx is just a linear combination of two random walks and therefore itself a random walk.

However, suppose that the two random walks are related to each other. Suppose y(t) is a random walk, but is in fact a linear function of x(t), another random walk such that $y=\alpha+\beta x+u$. The residual is $u=y-\alpha-\beta x$. Now, because x and y are related, there is an attraction between them, so that if u gets very large, there will be natural tendency for x and y to adjust so that u becomes approximately zero.

Using our example of our two drunks exiting Paul's Deli, suppose they have become best buddies after a night of drinking, so they exit the deli arm in arm. Now, even though they are individually walking randomly, they are linked in such a way that neither can wander too far from the other. If one does lurch, then the other will also lurch, the distance between them remaining approximately an arm's length.

Economists suspect that there is a consumption function such that consumption is a linear function of income, usually a constant fraction. Therefore, even though both income and consumption are individually random walks, we expect that they will not wander very far from each other. In fact the graph of the log of consumption and the log of national income above shows that they stay very close to each other. If income increases dramatically, but consumption doesn't grow at all, the residual will become a large negative number. Assuming that there really is a long run equilibrium relationship between consumption, then consumption will tend to grow to its long run level consistent with the higher level of income. This implies that the residual will tend to hover around the zero line. Large values of the residual will be drawn back to the zero line. This is the behavior we associate with a stationary variable. So, if there is a long run

equilibrium relationship between two (or more) random walks, the residual will be stationary. If there is no long run relationship between the random walks, the residual will be nonstationary.

Two random walks that are individually I(1) can nevertheless generate a stationary I(0) residual if there is a long run relationship between them. Such a pair of random walks are said to be cointegrated.

So we can test for cointegration (i.e., a long run equilibrium relationship) by regressing consumption on income and maybe a time trend, and then testing the resulting residual for a unit root. If it has a unit root, then there is no long run relationship among these series. If we reject the null hypothesis of a unit root in the residual, we accept the alternative hypothesis of cointegration.

We use the same old augmented Dickey-Fuller test to test for stationarity of the residual. This is the "Engle-Granger two step" test for cointegration. In the first step we estimate the long run equilibrium relationship. This is just a static regression (no lags) in levels (no first differences) of the log of consumption on the log of income, the log of the interest rate, and a time trend. We will drop the time trend if it is not significant. We then save the residual and run an ADF unit root test on the residual. Because the residual is centered on zero by construction and can have no time trend, we do not include an intercept or a trend term in our ADF regression.

```
. regress lnc lny lnr t

      Source |       SS       df       MS              Number of obs =     221
-------------+------------------------------           F(  3,   217) =85503.23
       Model |  69.7033367     3  23.2344456           Prob > F      =  0.0000
    Residual |  .058967068   217  .000271738           R-squared     =  0.9992
-------------+------------------------------           Adj R-squared =  0.9991
       Total |  69.7623038   220  .317101381           Root MSE      =  .01648

------------------------------------------------------------------------------
         lnc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lny |   .8053303    .032362    24.89   0.000     .7415463    .8691144
         lnr |  -.0047292    .003373    -1.40   0.162    -.0113772    .0019189
           t |   .0021389   .0002607     8.21   0.000     .0016252    .0026527
       _cons |   .9669527   .2378607     4.07   0.000     .4981397    1.435766
------------------------------------------------------------------------------

. predict error, resid
(8 missing values generated)

. dfuller error, noconstant lags(3) regress

Augmented Dickey-Fuller test for unit root         Number of obs   =       217

                             ---------- Interpolated Dickey-Fuller ---------
                 Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value            Value
------------------------------------------------------------------------------
 Z(t)             -4.841            -2.584            -1.950           -1.618

------------------------------------------------------------------------------
D.error      |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
error        |
          L1 |  -.1445648   .0298631    -4.84   0.000    -.2034297   -.0856998
          LD |  -.2037883    .064761    -3.15   0.002    -.3314428   -.0761338
         L2D |   .2077673   .0657189     3.16   0.002     .0782246      .33731
         L3D |   .2383771   .0635725     3.75   0.000     .1130652    .3636889
------------------------------------------------------------------------------
```

Unfortunately, we can't use the critical values supplied by dfuller. The reason is that the dfuller test is based on a standard random walk. We have a different kind of random walk here because it is a residual. However, Phillips and Ouliaris (Asymptotic Properties of Residual Based Tests for Cointegration, *Econometrica* 58 (1), 165-193) provide critical values for the Engle-Granger cointegration test.

Critical values for the Engle-Granger ADF cointegration statistic
Intercept and trend included

NUMBER OF EXPLANATORY VARIABLES IN STATIC REGRESSION MODEL (EXCLUDING INTERCEPT AND TREND)	10%	5%	1%
1	-3.52	-3.80	-4.36
2	-3.84	-4.16	-4.65
3	-4.20	-4.49	-5.04
4	-4.46	-4.74	-5.36
5	-4.73	-5.03	-5.58

Critical values for the Engle-Granger ADF cointegration test statistic
Intercept included, no trend

NUMBER OF EXPLANATORY VARIABLES IN STATIC REGRESSION MODEL (EXCLUDING INTERCEPT)	10%	5%	1%
1	-3.07	-3.37	-3.96
2	-3.45	-3.77	-4.31
3	-3.83	-4.11	-4.73
4	-4.16	-4.45	-5.07
5	-4.43	-4.71	-5.28

We have an intercept, a trend, and two independent variables in the long run consumption function, yielding a critical value of -4.16 at the five percent level. Our ADF test statistic is -4.84, so we reject the null hypothesis of a unit root in the residual from the consumption function and accept the alternative hypothesis that consumption, income, and interest rates are cointegrated. There is, apparently, a long run equilibrium relationship linking consumption, income, and the interest rate. Macro economists will be relieved.

What are the properties of the estimated long run regression? It turns out that the OLS estimates of the parameters of the static regression are consistent as the number of observations goes to infinity. Actually, Engle and Granger show that the estimates are “superconsistent” in the sense that they converge to the true parameter values much faster than the usual consistent estimates. An interesting question arises at this point. Why isn’t the consumption function, estimated with OLS, inconsistent because of simultaneity? We know that consumption is a function of national income, but national income is also a function of consumption (Y=C+I+G). So how can we get consistent estimates with OLS?

The reason is that the explanatory variable is a random walk. We know that the bias and inconsistency arises because of a correlation between the explanatory variables and the error term in the regression. However, because each of the explanatory variables in the regression is a random walk, they can wander off to infinity and have potentially infinite variance. The error term, on the other hand, is stationary, which means it is centered on zero with a finite variance. A variable that could go to infinity cannot be correlated with a residual that is stuck on zero. Therefore, OLS estimates of the long run relationship are consistent even in the presence of simultaneity.

Dynamic ordinary least squares

Although the estimated parameters of the Engle-Granger first stage regression are consistent, the estimates are inefficient relative to a simple extension of the regression model. Further, the standard errors and t-ratios are non-standard, so we can't use them for hypothesis testing. To fix the efficiency problem, we use dynamic ordinary least squares (DOLS). To correct the standard errors and t-ratios we use heteroskedastic and autocorrelation consistent (HAC) standard errors.(i.e., Newey-West standard errors).

DOLS is very easy, we simply add leads and lags of the differences of the integrated variables in the regression model. (If your static regression has stationary variables mixed in with the nonstationary integrated variables, just take leads and lags of the differences of the integrated variables, leaving the stationary variables alone.) It also turns out that you don't need more than one or two leads or lags to get the job done. To get the HAC standard errors and t-ratios, we use the "newey" command.

But first we have to generate the leads and lags. We will go with two leads and two lags. We can use the usual time series operators: F.x for leads (F for forward), L.x for lags, and D.x for differences.

This is the static long run model. Do not use a lagged dependent variable in the long run model.

```
. newey lnc lny lnr t FD.lny F2D.lny LD.lny L2D.lny FD.lnr F2D.lnr LD.lnr
L2D.lnr , lag(4)

Regression with Newey-West standard errors          Number of obs  =       216
maximum lag: 4                                      F( 11,   204)  =   5741.83
                                                    Prob > F       =    0.0000

------------------------------------------------------------------------------
             |             Newey-West
         lnc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lny |   .8891959   .0701308    12.68   0.000     .7509218     1.0274
         lnr |  -.0001377   .0058142    -0.02   0.981    -.0116013    .011325
           t |   .0014333   .0005705     2.51   0.013     .0003084    .002558
         lny |
         FD. |    .379992   .1678191     2.26   0.025     .0491097    .710874
        F2D. |   .1907836    .131358     1.45   0.148    -.0682098     .44977
         LD. |  -.1234908    .109279    -1.13   0.260    -.3389519    .091970
        L2D. |  -.0499793   .1226579    -0.41   0.684    -.2918191    .191860
         lnr |
         FD. |  -.0030922   .0116686    -0.27   0.791    -.0260987    .019914
        F2D. |  -.0051308   .0140005    -0.37   0.714    -.0327351    .022473
         LD. |  -.0154288   .0145519    -1.06   0.290    -.0441202    .013262
        L2D. |   -.025613   .0133788    -1.91   0.057    -.0519916    .00076
```

The lag(4) option tells newey that there is autocorrelation of up to order four. The interest rate does not appear to be significant in the long run consumption function.

Error correction model

We now have a way of estimating short run models with nonstationary time series, by taking first differences. We also have a way of estimating long run equilibrium relationships (the first step of the

Engle-Granger two-step). Is there any way to combine these techniques? It turns out that the error correction model does just this.

We derived the error correction model from the autoregressive distributed lag (ADL) model in an earlier chapter. The simple error correction model, also known as an equilibrium correction model or ECM can be written as follows.

$$\Delta y_t = b_0 \Delta x_t + (a_1 - 1)\left(y - kx\right)_{t-1} + \varepsilon_t$$

This equation can be interpreted as follows. The first difference captures the short run dynamics. The remaining part of the equation is called the error correction mechanism (ECM). The term in parentheses is the difference between the long run equilibrium value of y and the realized value of y in the previous period. It is equal to zero if there is long run equilibrium. If the system is out of equilibrium, this term will be either positive or negative. In that case, the coefficient multiplying this value will determine how much of the disequilibrium will be eliminated in each period, the "speed of adjustment" coefficient. As a result, y changes from period to period both because of short run changes in x, lagged changes in y, and adjustment to disequilibrium. Therefore, the error correction model incorporates both long run and short run information.

Note that, if x and y are both nonstationary random walks, then their first differences are stationary. If they are cointegrated, then the residuals from the static regression are also stationary. Thus, all the terms in the error correction model are stationary, so all the usual tests apply.

Also, if x and y are cointegrated then there is an error correction model explaining their behavior. If there is an error correction model relating x and y, then they are cointegrated. Further, they must be related in a Granger causality sense because each variable adjusts to lagged changes in the other through the error correction mechanism.

This model can be estimated using a two step procedure. In the first step, estimate the static long run regression (don't use DOLS). Predict the residual, which measures the disequilibrium. Lag the residual and use it as an explanatory variable in the first difference regression.

For example, consider the consumption function above. We have already estimated the long run model and predicted the residual. The only thing left to do is to estimate the EC model.

```
. regress D.lnc LD.lnc L4D.lnc D.lny D.lnr L.error

      Source |       SS       df       MS              Number of obs =     220
-------------+------------------------------           F(  5,   214) =   30.99
       Model |  .006745058     5  .001349012           Prob > F      =  0.0000
    Residual |  .009314213   214  .000043524           R-squared     =  0.4200
-------------+------------------------------           Adj R-squared =  0.4065
       Total |  .016059271   219   .00007333           Root MSE      =   .0066

------------------------------------------------------------------------------
D.lnc        |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lnc          |
          LD |  -.1823329   .0572958    -3.18   0.002    -.2952693   -.0693965
         L4D |  -.1036533   .0528633    -1.96   0.051    -.2078528    .0005462
lny          |
          D1 |   .5949995   .0485122    12.26   0.000     .4993766    .6906224
lnr          |
          D1 |  -.0131099   .0059389    -2.21   0.028    -.0248161   -.0014037
error        |
          L1 |  -.0734877   .0278959    -2.63   0.009    -.1284737   -.0185017
_cons        |   .0063855   .0008469     7.54   0.000     .0047162    .0080548
------------------------------------------------------------------------------
```

So, consumption changes because of lagged changes in consumption, changes in income, changes in the interest rate, and lagged disequilibrium. About 7 percent of the disequilibrium is eliminated each quarter.

This is an example only. There is a lot more modeling involved in estimating even a function as simple as a consumption function. If you were to do an LM test for autocorrelation on the above model, which you are welcome to do, the data are in CYR.dta, you would find that there is serious autocorrelation, indicating that the model is not properly specified. You are invited to do a better job.

17 PANEL DATA MODELS

A general model that pools time series and cross section data is

$$Y_{i,t} = \alpha_i + \beta_t + \sum_{k=1}^{K} \gamma_{i,k} X_{k,i,t} + \varepsilon_{i,t}$$

where i=1,…, N (number of cross sections, e.g., states); t=1,…, T (number of time periods, e.g., years); and K = number of explanatory variables. Note that this model gives each state its own intercept. This allows, for example, New York to have higher crime rates, holding everything else equal, than, say, North Dakota. The model also allows each year to have its own effect (β_t). The model therefore also controls for national events and trends that affect crime across all states, again holding everything else constant. Finally, the model allows each state to have its own parameter with respect to each of the explanatory variables. This allows, for example, the crime rate in New York to be affected differently by, say, an increase in the New York prison population, than the crime rate in North Dakota is affected by an increase in North Dakota's prison population.

Different models are derived by making various assumptions concerning the parameters of this model. If we assume that $\alpha_i = \alpha_2 = \ldots = \alpha_N$, $\beta_1 = \ldots = \beta_T$, and $\gamma_{1,k} = \gamma_{2,k} = \ldots = \gamma_{N,k}$ then we have the **OLS** model. If we assume that the α_i and β_t not all equal but are fixed numbers (and that the coefficients $\gamma_{i,k}$ are constant across states, i) then we have the **fixed effects (FE)** model. This model is also called the **least squares dummy variable (LSDV)** model, the **covariance** model, and the **within** estimator. If we assume that the α_i and β_t are random variables, still assuming that the γ_i are all equal, then we have the **random effects (RE)** model also known as the **variance components** model or the **error components** model. Finally, if we assume the coefficients are constant across time, but allow the $\alpha_{i,k}$ and $\gamma_{i,k}$ to vary across states and assume that $\beta_1 = \ldots = \beta_T = 0$, then we have the **random coefficients** model.

The Fixed Effects Model

Let's assume that the coefficients on the explanatory variables $x_{k,i,t}$ are constant across states and across time. The model therefore reduces to

$$Y_{i,t} = \alpha_i + \beta_t + \sum_{k=1}^{K} \gamma_k X_{k,i,t} + \varepsilon_{i,t}$$

where the error term is assumed to be independently and identically distributed random variables with $E\left[\varepsilon_{it}\right] = 0$ and $E\left[\varepsilon_{i,t}^2\right] = \sigma_\varepsilon^2$.Note that this specification assumes that each state is independent of all other states (no contemporaneous correlation across error terms), that each time period is independent of all the others (no autocorrelation), and that the variance of the error term is constant (no heteroskedasticity).

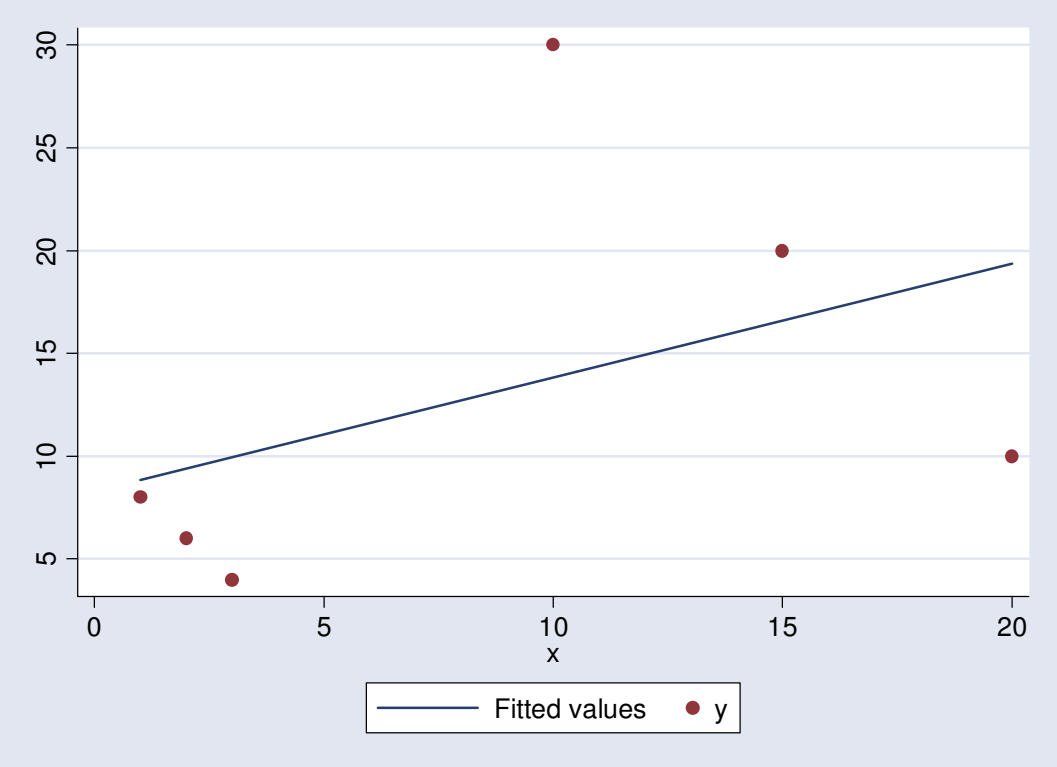
This model can be conveniently estimated using dummy variables. Suppose we create a set of dummy variables, one for each state $D_1,\ldots,D_N$ where $D_i = 1$ if the observation is from state i and $D_i = 0$ otherwise. Suppose also that we similarly create a set of year dummy variables ($YR_1,\ldots,YR_T$). Then we can estimate the model by regressing the dependent variable on all the explanatory variables X_k, on all the state dummies, and on all the year dummies using ordinary least squares. This is why the model is also called the least squares dummy variable model. Under the assumptions outlined above the estimates produced by this process are unbiased and consistent.

The fixed effects model is the most widely use panel data model. The primary reason for its wide use is that it corrects for a particular kind of omitted variable bias, known as **unobserved heterogeneity**. Suppose, for the sake of illustration, that we have data on just two states, North Dakota and New York, with three years of data on each state. Let's also assume that prison deters crime, so that the more criminals in prison, the lower the crime rate. Finally, let's assume that both the crime rate and the prison population is quite low in North Dakota and they are both quite high in New York.

These assumptions are reflected in the following example. The true relationship between crime and prison is $y_{it} = \alpha_i - 2x_{it}$ where $\alpha_1 = 10$ (ND) and $\alpha_2 = 50$ (NY) . We set x=1,2,3 for ND and x=10,15,20 for NY. The data set is panel.graph.dta.

```
. list

     +-----------------------------+
     |  x    y   state   st1   st2 |
     |-----------------------------|
  1. |  1    8       1     1     0 |
  2. |  2    6       1     1     0 |
  3. |  3    4       1     1     0 |
  4. | 10   30       2     0     1 |
  5. | 15   20       2     0     1 |
  6. | 20   10       2     0     1 |
     +-----------------------------+
```

where st1 and st2 are state dummy variables. Here is the scatter diagram with the OLS regression line superimposed.



Here is the OLS regression line.

```
. regress y x

      Source |       SS       df       MS              Number of obs =       6
-------------+------------------------------           F(  1,     4) =    0.92
       Model |  93.4893617     1  93.4893617            Prob > F      =  0.3929
    Residual |  408.510638     4   102.12766            R-squared     =  0.1862
-------------+------------------------------           Adj R-squared = -0.0172
       Total |         502     5       100.4            Root MSE      =  10.106

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .5531915    .578184     0.96   0.393    -1.052105    2.158488
       _cons |   8.297872   6.416714     1.29   0.266    -9.517782    26.11353
```

Since we don't estimate a separate intercept for each state, the fact that the observations for New York show both high crime and high prison population while North Dakota has low crime and low prison population yields a positive correlation between the individual state intercepts and the explanatory variable. The OLS model does not employ individual state dummy variables, and is therefore guilty of a kind of omitted variable bias. In this case, the bias is so strong that it results in an incorrect sign on the estimated coefficient for prison.

Here is the fixed effects model including the state dummy variables (but no intercept because of the dummy variable trap).

```
. regress y x st1 st2, noconstant
```

```
      Source |       SS       df       MS              Number of obs =       6
-------------+------------------------------           F(  3,     3) =       .
       Model |        1516     3  505.333333           Prob > F      =       .
    Residual |           0     3           0           R-squared     =  1.0000
-------------+------------------------------           Adj R-squared =  1.0000
       Total |        1516     6  252.666667           Root MSE      =       0

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |          -2          .        .       .            .           .
         st1 |          10          .        .       .            .           .
         st2 |          50          .        .       .            .           .
------------------------------------------------------------------------------
```

The fixed effects model fits the data perfectly.

If we include the intercept, we are in the dummy variable trap. Stata will drop the last dummy variable, making the overall intercept equal to the coefficient on the New York dummy. North Dakota's intercept is now measured relative to New York's. I recommend always keeping the overall interecept, and dropping one of the state dummies.

```
. regress y x st1 st2
```

```
      Source |       SS       df       MS              Number of obs =       6
-------------+------------------------------           F(  2,     3) =       .
       Model |         502     2         251           Prob > F      =       .
    Residual |           0     3           0           R-squared     =  1.0000
-------------+------------------------------           Adj R-squared =  1.0000
       Total |         502     5       100.4           Root MSE      =       0

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |          -2          .        .       .            .           .
         st1 |         -40          .        .       .            .           .
         st2 |  (dropped)
       _cons |          50          .        .       .            .           .
------------------------------------------------------------------------------
```

Since different states can be wildly different from each other for a variety of reasons, the assumption of unobserved heterogeneity across states makes perfect sense. Thus, the fixed effects model is the correct model for panels of states. The same is true for panels of countries, cities, and counties.

This procedure becomes awkward if there is a large number of cross section or time series observations. If your sample consists of the 3000 or so counties in the United States over several years, then 3000 dummy variables is a bit much. Luckily, there is an alternative approach that yields identical estimates. If we subtract the state means from each observation, we get deviations from the state means. We can then use ordinary least squares to regress the resulting "absorbed" dependent variable against the K "absorbed" explanatory variables. This is the **covariance** estimation method. The parameter estimates will be identical to the LSDV estimates.

The model is estimated with the "areg" (absorbed regression) command in Stata. It yields the same coefficients as the LSDV method.

```
. areg y x, absorb(state)
```

```
                                                Number of obs =       6
                                                F(  0,     3) =       .
```

```
                                                  Prob > F      =          .
                                                  R-squared     =     1.0000
                                                  Adj R-squared =     1.0000
                                                  Root MSE      =          0

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |         -2          .       .       .            .           .
       _cons |         30          .       .       .            .           .
-------------+----------------------------------------------------------------
       state |           F(1, 3) =          .       .           (2 categories)
```

One drawback to this approach is that we will not be able to see the separate state intercepts without some further effort[10]. Usually, we are not particularly interested in the numerical value of each of these estimated coefficients as they are very difficult to interpret. Since each state intercept is a combination of all the influences on crime that are specific to each state, constant across time and independent of the explanatory variables, it is difficult to know exactly what the coefficient is capturing. You might think of it as a measure of the "culture" of each particular state. The fact that New York's intercept is larger, and perhaps significantly larger, than North Dakota's intercept is interesting and perhaps important, but the scale is unknown. All we know is that if the state intercepts are different from each other and we have included all the relevant explanatory variables, the intercepts capture the effect of unobserved heterogeneity across states.

The way to test for the significance of the state and year effects is to conduct a test of their joint significance as a group with F-tests corresponding to each of the following hypotheses ($\alpha_i = 0$ for all i, $\beta_t = 0$ for all t). In my experience, the state dummies are always significant. The year dummies are usually significant. Do not be tempted to use the t-ratios to drop the insignificant dummies and retain the significant ones. The year and state dummies should be dropped or retained as a group. The reason is that different parameterizations of the same problem can result in different dummies being omitted. This creates an additional source of modeling error that is best avoided.

We can use the panel.dta data set to illustrate these procedures. It is a subset of the crimext.dta data set consisting of the first five states for the years 1977-1997. First let's create the state and year dummy variables for the LSDV procedure.

```
. tabulate state, gen(st)

      state |
     number |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         21       20.00       20.00
          2 |         21       20.00       40.00
          3 |         21       20.00       60.00
          4 |         21       20.00       80.00
          5 |         21       20.00      100.00
------------+-----------------------------------
      Total |        105      100.00

. tabulate year, gen(yr)

       year |      Freq.     Percent        Cum.
------------+-----------------------------------
       1977 |          5        4.76        4.76
       1978 |          5        4.76        9.52
       1979 |          5        4.76       14.29
```

[10] See Judge, et.al. (1988), pp. 470-472 for the procedure.

```
       1980 |          5        4.76       19.05
       1981 |          5        4.76       23.81
       1982 |          5        4.76       28.57
       1983 |          5        4.76       33.33
       1984 |          5        4.76       38.10
       1985 |          5        4.76       42.86
       1986 |          5        4.76       47.62
       1987 |          5        4.76       52.38
       1988 |          5        4.76       57.14
       1989 |          5        4.76       61.90
       1990 |          5        4.76       66.67
       1991 |          5        4.76       71.43
       1992 |          5        4.76       76.19
       1993 |          5        4.76       80.95
       1994 |          5        4.76       85.71
       1995 |          5        4.76       90.48
       1996 |          5        4.76       95.24
       1997 |          5        4.76      100.00
------------+-----------------------------------
      Total |        105      100.00

. regress lcrmaj lprison lmetpct lblack lrpcpi lp1824 lp2534 st2-st5 yr2-yr21

      Source |       SS       df       MS              Number of obs =     105
-------------+------------------------------           F( 30,    74) =   27.19
       Model |  5.24465133    30  .174821711           Prob > F      =  0.0000
    Residual |  .475757707    74  .006429158           R-squared     =  0.9168
-------------+------------------------------           Adj R-squared =  0.8831
       Total |  5.72040904   104  .055003933           Root MSE      =  .08018

------------------------------------------------------------------------------
      lcrmaj |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lprison |  -.4886959   .1576294    -3.10   0.003    -.8027794   -.1746125
     lmetpct |   .0850076   .8021396     0.11   0.916     -1.51329    1.683306
      lblack |  -1.006078   .4363459    -2.31   0.024    -1.875516   -.1366394
      lrpcpi |   .9622313      .3844     2.50   0.015     .1962975    1.728165
      lp1824 |  -.5760403   .4585362    -1.26   0.213    -1.489694     .337613
      lp2534 |  -1.058662   .6269312    -1.69   0.095     -2.30785    .1905254
         st2 |  -2.174764   .7955754    -2.73   0.008    -3.759982    -.589545
         st3 |  -1.984824   .9657397    -2.06   0.043    -3.909102   -.0605453
         st4 |  -.7511301    .420334    -1.79   0.078    -1.588664    .0864037
         st5 |  -1.214557   .6007863    -2.02   0.047     -2.41165   -.0174643
         yr2 |   .0368809   .0529524     0.70   0.488     -.068629    .1423909
         yr3 |   .1114482   .0549662     2.03   0.046     .0019256    .2209707
         yr4 |   .3081141   .0660811     4.66   0.000     .1764446    .4397836
         yr5 |   .3655603   .0749165     4.88   0.000     .2162859    .5148348
         yr6 |   .3656063   .0808562     4.52   0.000     .2044969    .5267157
         yr7 |   .3213356   .0905921     3.55   0.001     .1408268    .5018443
         yr8 |   .3170379   .1026446     3.09   0.003     .1125141    .5215617
         yr9 |   .3554716   .1176801     3.02   0.003     .1209889    .5899542
        yr10 |   .4264209   .1358064     3.14   0.002     .1558207    .6970211
        yr11 |   .3886578   .1501586     2.59   0.012     .0894603    .6878553
        yr12 |   .3874652   .1672654     2.32   0.023     .0541816    .7207488
        yr13 |   .4056992   .1871167     2.17   0.033      .032861    .7785375
        yr14 |   .4489243   .2073026     2.17   0.034     .0358647    .8619838
        yr15 |   .5139933   .2261039     2.27   0.026     .0634715    .9645151
        yr16 |   .4322245   .2514655     1.72   0.090    -.0688314    .9332805
        yr17 |    .387142   .2775931     1.39   0.167    -.1659743    .9402583
        yr18 |   .3239752    .304477     1.06   0.291    -.2827085    .9306589
        yr19 |   .2460049   .3344635     0.74   0.464    -.4204281    .9124379
        yr20 |   .1592112   .3602613     0.44   0.660     -.558625    .8770475
        yr21 |   .1163704   .3901125     0.30   0.766    -.6609458    .8936866
```

```
       _cons |   7.362542   4.453038     1.65   0.102     -1.51033    16.23541
------------------------------------------------------------------------------

* test for significance of the state dummies
. testparm st2-st5

 ( 1)  st2 = 0
 ( 2)  st3 = 0
 ( 3)  st4 = 0
 ( 4)  st5 = 0

       F(  4,    74) =    6.32
            Prob > F =    0.0002

* test for significance of the year dummies

. testparm yr2-yr21

 ( 1)  yr2 = 0
 ( 2)  yr3 = 0
 ( 3)  yr4 = 0
 ( 4)  yr5 = 0
 ( 5)  yr6 = 0
 ( 6)  yr7 = 0
 ( 7)  yr8 = 0
 ( 8)  yr9 = 0
 ( 9)  yr10 = 0
 (10)  yr11 = 0
 (11)  yr12 = 0
 (12)  yr13 = 0
 (13)  yr14 = 0
 (14)  yr15 = 0
 (15)  yr16 = 0
 (16)  yr17 = 0
 (17)  yr18 = 0
 (18)  yr19 = 0
 (19)  yr20 = 0
 (20)  yr21 = 0

       F( 20,    74) =    4.49
            Prob > F =    0.0000
```

As expected, both sets of dummy variables are highly significant. Let's do the same regression by absorbing the state means.

```
. areg lcrmaj lprison lmetpct lblack lrpcpi lp1824 lp2534 yr2-yr21,
absorb(state)

                                                  Number of obs =     105
                                                  F( 26,    74) =    6.12
                                                  Prob > F      =  0.0000
                                                  R-squared     =  0.9168
                                                  Adj R-squared =  0.8831
                                                  Root MSE      =  .08018

------------------------------------------------------------------------------
      lcrmaj |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lprison |  -.4886959   .1576294    -3.10   0.003    -.8027794   -.1746125
     lmetpct |   .0850076   .8021396     0.11   0.916     -1.51329    1.683306
      lblack |  -1.006078   .4363459    -2.31   0.024    -1.875516   -.1366394
      lrpcpi |   .9622313      .3844     2.50   0.015     .1962975    1.728165
      lp1824 |  -.5760403   .4585362    -1.26   0.213    -1.489694     .337613
```

```
      lp2534 |  -1.058662   .6269312    -1.69   0.095     -2.30785    .1905254
         yr2 |   .0368809   .0529524     0.70   0.488     -.068629    .1423909
         yr3 |   .1114482   .0549662     2.03   0.046     .0019256    .2209707
         yr4 |   .3081141   .0660811     4.66   0.000     .1764446    .4397836
         yr5 |   .3655603   .0749165     4.88   0.000     .2162859    .5148348
         yr6 |   .3656063   .0808562     4.52   0.000     .2044969    .5267157
         yr7 |   .3213356   .0905921     3.55   0.001     .1408268    .5018443
         yr8 |   .3170379   .1026446     3.09   0.003     .1125141    .5215617
         yr9 |   .3554716   .1176801     3.02   0.003     .1209889    .5899542
        yr10 |   .4264209   .1358064     3.14   0.002     .1558207    .6970211
        yr11 |   .3886578   .1501586     2.59   0.012     .0894603    .6878553
        yr12 |   .3874652   .1672654     2.32   0.023     .0541816    .7207488
        yr13 |   .4056992   .1871167     2.17   0.033      .032861    .7785375
        yr14 |   .4489243   .2073026     2.17   0.034     .0358647    .8619838
        yr15 |   .5139933   .2261039     2.27   0.026     .0634715    .9645151
        yr16 |   .4322245   .2514655     1.72   0.090    -.0688314    .9332805
        yr17 |    .387142   .2775931     1.39   0.167    -.1659743    .9402583
        yr18 |   .3239752    .304477     1.06   0.291    -.2827085    .9306589
        yr19 |   .2460049   .3344635     0.74   0.464    -.4204281    .9124379
        yr20 |   .1592112   .3602613     0.44   0.660     -.558625    .8770475
        yr21 |   .1163704   .3901125     0.30   0.766    -.6609458    .8936866
       _cons |   6.137487   4.477591     1.37   0.175    -2.784307    15.05928
-------------+----------------------------------------------------------------
       state |          F(4, 74) =      6.315   0.000           (5 categories)
```

Note that the coefficients are identical to the LSDV method. Also, the F-test on the state dummies is automatically produced by areg.

There is yet another fixed effects estimation command in Stata, namely xtreg. However, with xtreg, you must specify the fixed effects model with the option, fe.

```
. xtreg lcrmaj lprison lmetpct lblack lrpcpi lp1824 lp2534 yr2-yr21, fe

Fixed-effects (within) regression               Number of obs      =       105
Group variable (i): state                       Number of groups   =         5

R-sq:  within  = 0.6824                         Obs per group: min =        21
       between = 0.2352                                        avg =      21.0
       overall = 0.2127                                        max =        21

                                                F(26,74)           =      6.12
corr(u_i, Xb)  = -0.9696                        Prob > F           =    0.0000

------------------------------------------------------------------------------
      lcrmaj |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lprison |  -.4886959   .1576294    -3.10   0.003    -.8027794   -.1746125
     lmetpct |   .0850076   .8021396     0.11   0.916     -1.51329    1.683306
      lblack |  -1.006078   .4363459    -2.31   0.024    -1.875516   -.1366394
      lrpcpi |   .9622313      .3844     2.50   0.015     .1962975    1.728165
      lp1824 |  -.5760403   .4585362    -1.26   0.213    -1.489694     .337613
      lp2534 |  -1.058662   .6269312    -1.69   0.095     -2.30785    .1905254
         yr2 |   .0368809   .0529524     0.70   0.488     -.068629    .1423909
         yr3 |   .1114482   .0549662     2.03   0.046     .0019256    .2209707
         yr4 |   .3081141   .0660811     4.66   0.000     .1764446    .4397836
         yr5 |   .3655603   .0749165     4.88   0.000     .2162859    .5148348
         yr6 |   .3656063   .0808562     4.52   0.000     .2044969    .5267157
         yr7 |   .3213356   .0905921     3.55   0.001     .1408268    .5018443
         yr8 |   .3170379   .1026446     3.09   0.003     .1125141    .5215617
         yr9 |   .3554716   .1176801     3.02   0.003     .1209889    .5899542
        yr10 |   .4264209   .1358064     3.14   0.002     .1558207    .6970211
        yr11 |   .3886578   .1501586     2.59   0.012     .0894603    .6878553
```

```
        yr12 |   .3874652   .1672654     2.32   0.023     .0541816    .7207488
        yr13 |   .4056992   .1871167     2.17   0.033      .032861    .7785375
        yr14 |   .4489243   .2073026     2.17   0.034     .0358647    .8619838
        yr15 |   .5139933   .2261039     2.27   0.026     .0634715    .9645151
        yr16 |   .4322245   .2514655     1.72   0.090    -.0688314    .9332805
        yr17 |    .387142   .2775931     1.39   0.167    -.1659743    .9402583
        yr18 |   .3239752    .304477     1.06   0.291    -.2827085    .9306589
        yr19 |   .2460049   .3344635     0.74   0.464    -.4204281    .9124379
        yr20 |   .1592112   .3602613     0.44   0.660     -.558625    .8770475
        yr21 |   .1163704   .3901125     0.30   0.766    -.6609458    .8936866
       _cons |   6.137487   4.477591     1.37   0.175    -2.784307    15.05928
-------------+----------------------------------------------------------------
     sigma_u |  .89507954
     sigma_e |  .08018203
         rho |  .99203915   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(4, 74) =     6.32                 Prob > F = 0.0002
```

Again, the coefficients are identical and the F-test is automatic.

Another advantage of the fixed effects model is that, since the estimation procedure is really ordinary least squares with a bunch of dummy variables, problems of autocorrelation and heteroskedasticity, among others can be dealt with relatively simply. Heteroskedasticity can be handled with weighting the regression, usually with population, or by using heteroskedastic-consistent standard errors, or both. As we argued in the section on time series analysis, autocorrelation is best dealt with by adding lagged variables.

I recommend either the regress (LSDV) or areg (covariance) methods because they both allow robust standard errors and t-ratios, as well as a wide variety of post-estimation diagnostics. Heteroskedasticity can be addressed by using the Breusch-Pagan or White test or corrected using weighting and robust standard errors. The xtreg command does not allow robust standard errors.

Time series issues

Since we are combining cross-section and time-series data, we can expect to run into both cross-section and time-series problems. We saw above that one major cross-section problem, heteroskedasticity, is routinely handled with this model. Further, our model of choice is usually the fixed effects model, ignores all variation between states and is therefore a pure time series model, As a result, we have to address all the usual time-series issues.

The first time series topic that we need to deal with is the possibility of deterministic trends in the data.

Linear trends

One suggestion for panel data studies is to include an overall trend, as well as the year dummies. The trend captures the effect of omitted slowly changing time related variables. Since we never know if we have included all the relevant variables, I recommend always including a trend in any time series model. If it turns out to be insignificant, then it can be dropped. The year dummies do not capture the same effects as the overall trend. The year dummies capture the effects of common factors that affect the dependent variable over all the states at the same time. For example, an amazingly effective federal law could reduce crime uniformly in all the states in the years after its passage. This would be captured by year dummies, not a linear trend.

The model becomes

$$Y_{i,t} = \alpha_i + \beta_t + \delta t + \sum_{k=1}^{K} \gamma_k X_{k,i,t} + \varepsilon_{i,t}$$

A linear trend is simply the numbers 0,1,2,…,T. We can create an overall trend by simply subtracting the first year from all years. For example, if that data set starts with 1977, we can create a linear trend by subtracting 1977 from each value of year. Note that adding a linear trend will put us in the dummy variable trap because the trend is simply a linear combination of the year dummies with the coefficients equal to 0, 1, 2, etc. As a result, I recommend deleting one of the year dummies when adding a linear trend.

We can illustrate these facts with the panel.dta data set.

```
. gen t=year-1977
. list state stnm year t yr1-yr5 if state==1

     +-------------------------------------------------------------+
     | state   stnm   year    t   yr1   yr2   yr3   yr4   yr5 |
     |-------------------------------------------------------------|
  1. |     1     AL   1977    0     1     0     0     0     0 |
  2. |     1     AL   1978    1     0     1     0     0     0 |
  3. |     1     AL   1979    2     0     0     1     0     0 |
  4. |     1     AL   1980    3     0     0     0     1     0 |
  5. |     1     AL   1981    4     0     0     0     0     1 |
     |-------------------------------------------------------------|
  6. |     1     AL   1982    5     0     0     0     0     0 |
  7. |     1     AL   1983    6     0     0     0     0     0 |
  8. |     1     AL   1984    7     0     0     0     0     0 |
  9. |     1     AL   1985    8     0     0     0     0     0 |
 10. |     1     AL   1986    9     0     0     0     0     0 |
     |-------------------------------------------------------------|
 11. |     1     AL   1987   10     0     0     0     0     0 |
 12. |     1     AL   1988   11     0     0     0     0     0 |
 13. |     1     AL   1989   12     0     0     0     0     0 |
 14. |     1     AL   1990   13     0     0     0     0     0 |
 15. |     1     AL   1991   14     0     0     0     0     0 |
     |-------------------------------------------------------------|
 16. |     1     AL   1992   15     0     0     0     0     0 |
 17. |     1     AL   1993   16     0     0     0     0     0 |
 18. |     1     AL   1994   17     0     0     0     0     0 |
 19. |     1     AL   1995   18     0     0     0     0     0 |
 20. |     1     AL   1996   19     0     0     0     0     0 |
     |-------------------------------------------------------------|
 21. |     1     AL   1997   20     0     0     0     0     0 |
     +-------------------------------------------------------------+
```

Now let's estimate the corresponding fixed effects model.

```
. areg lcrmaj lprison lmetpct lblack lp1824 lp2534 t yr2-yr20 , absorb(state)

                                                  Number of obs =     105
                                                  F( 25,    75) =    5.71
                                                  Prob > F      =  0.0000
                                                  R-squared     =  0.9098
                                                  Adj R-squared =  0.8749
                                                  Root MSE      =  .08295

------------------------------------------------------------------------------
      lcrmaj |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lprison |  -.6904558   .1401393    -4.93   0.000    -.9696276   -.4112841
     lmetpct |   .2764064   .8260439     0.33   0.739    -1.369157     1.92197
```

```
      lblack |  -.9933732   .4513743    -2.20   0.031    -1.892557   -.0941895
      lp1824 |   .0528402   .3968248     0.13   0.894    -.7376754    .8433558
      lp2534 |  -.1393306   .5256308    -0.27   0.792    -1.186441    .9077796
           t |   .0431765    .012993     3.32   0.001     .0172932    .0690598
         yr2 |   .0135131   .0542071     0.25   0.804    -.0944729    .1214992
         yr3 |   .0296471   .0611756     0.48   0.629     -.092221    .1515151
         yr4 |   .1321815   .0773008     1.71   0.091    -.0218097    .2861727
         yr5 |   .1534079   .0914021     1.68   0.097    -.0286743    .3354901
         yr6 |   .1403835   .0946307     1.48   0.142    -.0481306    .3288975
         yr7 |   .1052747    .099064     1.06   0.291    -.0920709    .3026204
         yr8 |   .1169556   .0997961     1.17   0.245    -.0818484    .3157595
         yr9 |   .1689011   .0983383     1.72   0.090    -.0269988     .364801
        yr10 |   .2511994   .0968453     2.59   0.011     .0582736    .4441253
        yr11 |   .1998446   .0940428     2.13   0.037     .0125017    .3871874
        yr12 |    .195804    .093125     2.10   0.039     .0102895    .3813186
        yr13 |   .2179579   .0911012     2.39   0.019     .0364749    .3994409
        yr14 |   .2592705   .0842792     3.08   0.003     .0913777    .4271632
        yr15 |   .3172979   .0757261     4.19   0.000     .1664438     .468152
        yr16 |   .2495013    .068577     3.64   0.001     .1128888    .3861138
        yr17 |   .2169851   .0623196     3.48   0.001      .092838    .3411321
        yr18 |   .1670167   .0573777     2.91   0.005     .0527144     .281319
        yr19 |   .1087532   .0535365     2.03   0.046      .002103    .2154034
        yr20 |   .0263492   .0520785     0.51   0.614    -.0773964    .1300948
       _cons |   10.67773   4.235064     2.52   0.014     2.241046    19.11441
-------------+------------------------------------------------------------------
       state |          F(4, 75) =      4.585   0.002           (5 categories)
```

The trend is significant, and should be retained.

Another suggestion is that each state should have its own trend term as well as its own intercept. The model becomes

$$y_{i,t} = \alpha_i + \beta_t + \delta_i t + \sum_{k=1}^{K} \gamma_k x_{k,i,t} + \varepsilon_{i,t}$$

We can create these state trends by multiplying the overall trend by the state dummies.

```
. gen tr1=t*st1
. gen tr2=t*st2
. gen tr3=t*st3
. gen tr4=t*st4
. gen tr5=t*st5
```

Let's look at a couple of these series.

```
. list state stnm year t st1 tr1 st2 tr2 if state <=2

+-------------------------------------------------+
     | state   stnm   year    t   st1   tr1   st2   tr2 |
     |-------------------------------------------------|
  1. |     1     AL   1977    0     1     0     0     0 |
  2. |     1     AL   1978    1     1     1     0     0 |
  3. |     1     AL   1979    2     1     2     0     0 |
  4. |     1     AL   1980    3     1     3     0     0 |
  5. |     1     AL   1981    4     1     4     0     0 |
     |-------------------------------------------------|
  6. |     1     AL   1982    5     1     5     0     0 |
  7. |     1     AL   1983    6     1     6     0     0 |
  8. |     1     AL   1984    7     1     7     0     0 |
  9. |     1     AL   1985    8     1     8     0     0 |
```

```
 10. |      1      AL    1986     9      1      9      0      0 |
     |-----------------------------------------------------------|
 11. |      1      AL    1987    10      1     10      0      0 |
 12. |      1      AL    1988    11      1     11      0      0 |
 13. |      1      AL    1989    12      1     12      0      0 |
 14. |      1      AL    1990    13      1     13      0      0 |
 15. |      1      AL    1991    14      1     14      0      0 |
     |-----------------------------------------------------------|
 16. |      1      AL    1992    15      1     15      0      0 |
 17. |      1      AL    1993    16      1     16      0      0 |
 18. |      1      AL    1994    17      1     17      0      0 |
 19. |      1      AL    1995    18      1     18      0      0 |
 20. |      1      AL    1996    19      1     19      0      0 |
     |-----------------------------------------------------------|
 21. |      1      AL    1997    20      1     20      0      0 |
 22. |      2      AK    1977     0      0      0      1      0 |
 23. |      2      AK    1978     1      0      0      1      1 |
 24. |      2      AK    1979     2      0      0      1      2 |
 25. |      2      AK    1980     3      0      0      1      3 |
     |-----------------------------------------------------------|
 26. |      2      AK    1981     4      0      0      1      4 |
 27. |      2      AK    1982     5      0      0      1      5 |
```

Now let's estimate the implied fixed effects model.

```
. areg lcrmaj lprison lmetpct lblack lp1824 lp2534 tr1-tr5 yr2-yr20 ,
absorb(state)
```

```
                                                  Number of obs =     105
                                                  F( 29,    71) =    9.04
                                                  Prob > F      =  0.0000
                                                  R-squared     =  0.9442
                                                  Adj R-squared =  0.9183
                                                  Root MSE      =  .06706

------------------------------------------------------------------------------
      lcrmaj |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lprison |  -.4922259   .1500995    -3.28   0.002    -.7915157    -.192936
     lmetpct |   6.408965   3.265479     1.96   0.054    -.1022141    12.92014
      lblack |   .1648299   1.133181     0.15   0.885    -2.094669    2.424329
      lp1824 |  -.0461248   .3258571    -0.14   0.888    -.6958654    .6036158
      lp2534 |  -.4839899   .4826101    -1.00   0.319    -1.446287    .4783073
         tr1 |  -.0051039    .017904    -0.29   0.776    -.0408036    .0305957
         tr2 |    .033832   .0311964     1.08   0.282     -.028372    .0960359
         tr3 |  -.0140482    .016859    -0.83   0.407    -.0476641    .0195677
         tr4 |   .0214053   .0127019     1.69   0.096    -.0039216    .0467323
         tr5 |   .0197035   .0137888     1.43   0.157    -.0077906    .0471976
         yr2 |   .0292319   .0449536     0.65   0.518     -.060403    .1188667
         yr3 |   .0680641   .0533424     1.28   0.206    -.0382976    .1744258
         yr4 |   .1931018   .0702593     2.75   0.008     .0530088    .3331948
         yr5 |   .2151175   .0845614     2.54   0.013     .0465069     .383728
         yr6 |   .1876612   .0899632     2.09   0.041     .0082798    .3670427
         yr7 |   .1441216   .0959832     1.50   0.138    -.0472634    .3355067
         yr8 |    .151261    .097835     1.55   0.127    -.0438165    .3463386
         yr9 |   .1996939   .0973446     2.05   0.044     .0055943    .3937936
        yr10 |   .2782618   .0968381     2.87   0.005      .085172    .4713515
        yr11 |   .2211901   .0952123     2.32   0.023     .0313422     .411038
        yr12 |   .2153056   .0944229     2.28   0.026     .0270317    .4035796
        yr13 |   .2356666   .0920309     2.56   0.013     .0521622    .4191709
        yr14 |   .2714963   .0856649     3.17   0.002     .1006853    .4423072
        yr15 |   .3271032   .0761548     4.30   0.000     .1752548    .4789515
```

```
        yr16 |    .259055   .0667641     3.88   0.000     .1259312    .3921789
        yr17 |   .2258548   .0579972     3.89   0.000     .1102116     .341498
        yr18 |   .1751649   .0509189     3.44   0.001     .0736355    .2766942
        yr19 |   .1152277    .045183     2.55   0.013     .0251353      .20532
        yr20 |   .0302166   .0425423     0.71   0.480    -.0546102    .1150435
       _cons |  -17.61505   13.88093    -1.27   0.209    -45.29283    10.06274
-------------+------------------------------------------------------------------
       state |          F(4, 71) =      1.246   0.299           (5 categories)

. testparm tr1-tr5

 ( 1)  tr1 = 0
 ( 2)  tr2 = 0
 ( 3)  tr3 = 0
 ( 4)  tr4 = 0
 ( 5)  tr5 = 0

       F(  5,    71) =   12.13
            Prob > F =    0.0000
```

Note that we have to drop the overall trend if we include all the state trends because of collinearity. Also, the individual trends are significant as a group and should be retained. We could also test whether the state trends all have the same coefficient and could be combined into a single, overall trend.

```
. test tr1=tr2=tr3=tr4=tr5

 ( 1)  tr1 - tr2 = 0
 ( 2)  tr1 - tr3 = 0
 ( 3)  tr1 - tr4 = 0
 ( 4)  tr1 - tr5 = 0

       F(  4,    71) =   10.94
            Prob > F =    0.0000
```

Apparently the trends are significantly different across states and cannot be combined into a single overall trend.

Estimating fixed effects models with individual state trends is becoming common practice.

Unit roots and panel data

With respect to the time series issues, we need to know whether each individual time series has a unit root. For example, pooling states across years means that each state forms a time series of annual data. We need to know if the typical state series has a unit root. There are panel data unit root tests, but none are officially available in Stata. It is possible to download user written Stata modules, however, you cannot use use these modules on the server.

Even if you had easily available panel unit root tests, what would you do with the information generated by such tests? If you have unit roots, then you will want to estimate a short run model with first differences, estimate a long run equilibrium model, or estimate a error correction model. If you do not have unit roots, you probably want to estimate a model in levels. Because unit root tests are not particularly powerful, it is probably a good idea to estimate both short run (first differences) and a long run model in levels and hope that the main results are consistent across both specifications.

Pooled time series and cross section panel models are ideal for estimating the long run average relationships among nonstationary variables. The pooled panel estimator yields consistent estimates with a

normal limit distribution. These results hold in the presence of individual fixed effects. The coefficients are analogous to the population (not sample) regression coefficients in conventional regressions.[11] Thus, the static fixed effects model estimated above can be interpreted as the long run average crime equation (for the five states in the sample).

.

It turns out that another way to estimate the fixed effects model is to take first differences. The first differencing "sweeps out" the state dummies because a state dummy is constant across time, so that the value of the dummy (=1) in year 2 is the same as its value (=1) in year 1, so the difference is zero. The year dummies are also altered. The dummy for year 2 is equal to one for year 2, zero otherwise. Thus, the first difference, year2-year1=1-0=1. However the difference, year3-year2=0-1=-1. All the remaining values are zero. So the first differenced year dummy equals zero for the years before the year of the dummy, then 1 for the year in question, then negative one for the year after, then zero again.

Let's use the panel.dta data set to illustrate this proposition.

```
. tsset state year
       panel variable:  state, 1 to 5
        time variable:  year, 1977 to 1997
. gen dst1=D.st1
(5 missing values generated)

. gen dyr5=D.yr5
(5 missing values generated)

. list state year stnm st1 dst1 yr5 dyr5 if state==1

     +-----------------------------------------------+
     | state   year   stnm   st1   dst1   yr5   dyr5 |
     |-----------------------------------------------|
  1. |     1   1977     AL     1      .     0      . |
  2. |     1   1978     AL     1      0     0      0 |
  3. |     1   1979     AL     1      0     0      0 |
  4. |     1   1980     AL     1      0     0      0 |
  5. |     1   1981     AL     1      0     1      1 |
     |-----------------------------------------------|
  6. |     1   1982     AL     1      0     0     -1 |
  7. |     1   1983     AL     1      0     0      0 |
  8. |     1   1984     AL     1      0     0      0 |
  9. |     1   1985     AL     1      0     0      0 |
 10. |     1   1986     AL     1      0     0      0 |
     |-----------------------------------------------|
 11. |     1   1987     AL     1      0     0      0 |
 12. |     1   1988     AL     1      0     0      0 |
 13. |     1   1989     AL     1      0     0      0 |
 14. |     1   1990     AL     1      0     0      0 |
 15. |     1   1991     AL     1      0     0      0 |
     |-----------------------------------------------|
 16. |     1   1992     AL     1      0     0      0 |
 17. |     1   1993     AL     1      0     0      0 |
 18. |     1   1994     AL     1      0     0      0 |
 19. |     1   1995     AL     1      0     0      0 |
 20. |     1   1996     AL     1      0     0      0 |
     |-----------------------------------------------|
 21. |     1   1997     AL     1      0     0      0 |
     +-----------------------------------------------+
```

[11] Phillips, P.C.B. and H.R. Moon, "Linear Regression Limit Theory for Nonstationary Panel Data," *Econometrica*, 67, 1999, 1057-1112.

Note that if you did include the state dummies in the first difference model, it is equivalent to estimating a fixed effects in levels with individual state trends because the first difference of a trend is a series of ones.

```
. gen dt=D.t
(5 missing values generated)

. list state year stnm st1 t dt if state==1

     +-----------------------------------+
     | state   year   stnm   st1    t   dt |
     |-----------------------------------|
  1. |     1   1977     AL     1    0    . |
  2. |     1   1978     AL     1    1    1 |
  3. |     1   1979     AL     1    2    1 |
  4. |     1   1980     AL     1    3    1 |
  5. |     1   1981     AL     1    4    1 |
     |-----------------------------------|
  6. |     1   1982     AL     1    5    1 |
  7. |     1   1983     AL     1    6    1 |
  8. |     1   1984     AL     1    7    1 |
  9. |     1   1985     AL     1    8    1 |
 10. |     1   1986     AL     1    9    1 |
     |-----------------------------------|
 11. |     1   1987     AL     1   10    1 |
 12. |     1   1988     AL     1   11    1 |
 13. |     1   1989     AL     1   12    1 |
 14. |     1   1990     AL     1   13    1 |
 15. |     1   1991     AL     1   14    1 |
     |-----------------------------------|
 16. |     1   1992     AL     1   15    1 |
 17. |     1   1993     AL     1   16    1 |
 18. |     1   1994     AL     1   17    1 |
 19. |     1   1995     AL     1   18    1 |
 20. |     1   1996     AL     1   19    1 |
     |-----------------------------------|
 21. |     1   1997     AL     1   20    1 |
     +-----------------------------------+
```

Since we can estimate a fixed effects model using either levels and state dummies (absorbed or not) or using first differences with no state dummies, shouldn't we get the same answer either way? The short answer is "no." The estimates will not be identical for several reasons. If the variables are stationary, then the first difference approach, when it subtracts last year's value, also subtracts last year's measurement error, if there is any. This is likely to leave a smaller measurement error than if we subtract of the mean, which is what we do if we use state dummies or absorb the state dummies. Thus, if there is any measurement error we could get different parameter estimates by estimating in levels and then estimating in first differences.

The second reason is that, if the variables are nonstationary, and not cointegrated, then taking first differences usually creates stationary variables, which can be expected to have different parameter values. If they are nonstationary, but cointegrated, the static model in levels is the long run equilibrium model, which can be expected to be somewhat different from the short run model. Finally, we lose one year of data when we take first differences, so the sample is not the same.

Let's estimate the fixed effects model above using first differences. We use the "D." operator to generate first differenced series. Don't forget to "tsset state year" to tell Stata that it is a panel data set sorted by state and year.

```
. regress dlcrmaj dlprison dlmetpct dlblack dlrpcpi dlp1824 dlp2534 dyr2-dyr21

      Source |       SS       df       MS              Number of obs =     100
```

```
-------------+------------------------------           F( 25,    74) =    3.55
       Model |  .267228983    25  .010689159           Prob > F      =  0.0000
    Residual |   .22282077    74  .003011091           R-squared     =  0.5453
-------------+------------------------------           Adj R-squared =  0.3917
       Total |  .490049753    99  .004949998           Root MSE      =  .05487

------------------------------------------------------------------------------
     dlcrmaj |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    dlprison |  -.2317252   .1700677    -1.36   0.177    -.5705926    .1071422
    dlmetpct |   1.631629   2.166495     0.75   0.454    -2.685207    5.948465
     dlblack |  -.4023363   .8714137    -0.46   0.646    -2.138666    1.333993
     dlrpcpi |   .1796915   .3049804     0.59   0.558    -.4279951    .7873781
     dlp1824 |  -.9338755   .4775285    -1.96   0.054    -1.885372    .0176207
     dlp2534 |  -.8105992   .7116841    -1.14   0.258    -2.228661    .6074623
        dyr2 |   .0442992   .0332805     1.33   0.187    -.0220136    .1106119
        dyr3 |   .0971729   .0595598     1.63   0.107    -.0215027    .2158485
        dyr4 |   .2581997   .0972473     2.66   0.010     .0644303    .4519692
        dyr5 |   .2852936   .1208029     2.36   0.021     .0445887    .5259986
        dyr6 |   .2441781   .1275799     1.91   0.059    -.0100303    .4983865
        dyr7 |   .1863416   .1329834     1.40   0.165    -.0786336    .4513169
        dyr8 |   .1780758   .1339427     1.33   0.188     -.088811    .4449625
        dyr9 |   .2112703   .1327145     1.59   0.116    -.0531691    .4757097
       dyr10 |   .2753427   .1314994     2.09   0.040     .0133244     .537361
       dyr11 |   .2083532   .1298586     1.60   0.113    -.0503956     .467102
       dyr12 |   .1976981   .1288571     1.53   0.129    -.0590553    .4544515
       dyr13 |   .2160628   .1250786     1.73   0.088    -.0331617    .4652874
       dyr14 |   .2431188   .1160254     2.10   0.040     .0119332    .4743043
       dyr15 |   .2966003   .1038212     2.86   0.006     .0897319    .5034686
       dyr16 |   .2321504    .089506     2.59   0.011     .0538057    .4104951
       dyr17 |   .2068382   .0745739     2.77   0.007     .0582465      .35543
       dyr18 |   .1602132   .0600055     2.67   0.009     .0406496    .2797767
       dyr19 |   .1017896   .0437182     2.33   0.023     .0146792    .1888999
       dyr20 |   .0210332   .0283287     0.74   0.460     -.035413    .0774795
       dyr21 |  (dropped)
       _cons |   -.014203   .0211906    -0.67   0.505    -.0564261    .0280202
------------------------------------------------------------------------------
```

We know that the constant term captures the effect of an overall linear trend. If we want to replace the overall trend with individual state trends, we simply add the state dummies back into the model.

These results are different from the fixed effects model estimated in levels. So, I recommend that you estimate the model in levels, with no lags, and call that the long run regression. Then estimate the model in first differences, with as many lags as might be necessary to generate residuals without any serial correlation (use the LM test with saved residuals). Call that the short run dynamic model.

Let's see if we have autocorrelation in the short run model above.

```
. predict e, resid
(5 missing values generated)

. gen e_1=L.e
(10 missing values generated)

. regress e e_1 dlprison dlmetpct dlblack dlrpcpi dlp1824 dlp2534 dyr2-dyr21

      Source |       SS       df       MS              Number of obs =      95
-------------+------------------------------           F( 25,    69) =    0.22
       Model |  .015042094    25  .000601684           Prob > F      =  1.0000
    Residual |   .19261245    69  .002791485           R-squared     =  0.0724
-------------+------------------------------           Adj R-squared = -0.2636
```

```
       Total |  .207654544    94  .002209091           Root MSE      =  .05283

------------------------------------------------------------------------------
           e |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         e_1 |   .2340898   .1181197     1.98   0.051    -.0015526    .4697323
    dlprison |  -.0796463   .1829466    -0.44   0.665    -.4446148    .2853223
    dlmetpct |  -.8967098   2.120876    -0.42   0.674    -5.127742    3.334322
     dlblack |  -.6180148   .8976362    -0.69   0.493     -2.40875    1.172721
     dlrpcpi |   .2284817   .3190218     0.72   0.476    -.4079494    .8649128
     dlp1824 |  -.1962703   .4761782    -0.41   0.681     -1.14622    .7536792
     dlp2534 |  -.1558151     .72768    -0.21   0.831    -1.607497    1.295867
        dyr2 |  (dropped)
        dyr3 |   .0056397   .0377379     0.15   0.882    -.0696455    .0809248
        dyr4 |   .0226855   .0787475     0.29   0.774    -.1344115    .1797825
        dyr5 |   .0342735    .104947     0.33   0.745      -.17509    .2436371
        dyr6 |   .0459182   .1152692     0.40   0.692    -.1840377     .275874
        dyr7 |   .0486212      .1227     0.40   0.693    -.1961586     .293401
        dyr8 |   .0458445   .1247895     0.37   0.714    -.2031038    .2947929
        dyr9 |   .0422207   .1245735     0.34   0.736    -.2062966    .2907381
       dyr10 |   .0392704   .1244473     0.32   0.753    -.2089952    .2875359
       dyr11 |   .0431052   .1250605     0.34   0.731    -.2063837    .2925941
       dyr12 |   .0427233   .1252572     0.34   0.734    -.2071579    .2926046
       dyr13 |   .0404449   .1222374     0.33   0.742    -.2034122    .2843019
       dyr14 |   .0399949   .1146092     0.35   0.728    -.1886443    .2686341
       dyr15 |   .0380374   .1033664     0.37   0.714    -.1681729    .2442476
       dyr16 |    .030283   .0887014     0.34   0.734    -.1466714    .2072374
       dyr17 |   .0234023   .0735115     0.32   0.751    -.1232491    .1700537
       dyr18 |   .0163755    .058807     0.28   0.781    -.1009412    .1336922
       dyr19 |   .0080579   .0423675     0.19   0.850     -.076463    .0925788
       dyr20 |   .0045028   .0274865     0.16   0.870    -.0503314    .0593369
       dyr21 |  (dropped)
       _cons |   .0029871   .0218047     0.14   0.891    -.0405121    .0464863
------------------------------------------------------------------------------

. test e_1

 ( 1)  e_1 = 0

       F(  1,    69) =    3.93
            Prob > F =    0.0515
```

Looks like we have autocorrelation. Let's add a lagged dependent variable.

```
. regress dlcrmaj L.dlcrmaj dlprison dlmetpct dlblack dlrpcpi dlp1824 dlp2534
dyr
> 2-dyr21

      Source |       SS       df       MS              Number of obs =      95
-------------+------------------------------           F( 25,    69) =    4.12
       Model |  .285675804    25  .011427032           Prob > F      =  0.0000
    Residual |  .191239715    69   .00277159           R-squared     =  0.5990
-------------+------------------------------           Adj R-squared =  0.4537
       Total |  .476915519    94  .005073569           Root MSE      =  .05265

------------------------------------------------------------------------------
dlcrmaj      |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
dlcrmaj      |
          L1 |   .2414129   .1144276     2.11   0.039      .013136    .4696897
dlprison     |   -.275926   .1823793    -1.51   0.135    -.6397627    .0879108
dlmetpct     |   .5640736   2.123772     0.27   0.791    -3.672735    4.800882
dlblack      |  -.9904829   .8941283    -1.11   0.272     -2.77422    .7932542
```

```
dlrpcpi      |   .3920387    .318366     1.23   0.222     -.243084    1.027161
dlp1824      |   -1.14118   .4742399    -2.41   0.019    -2.087263   -.1950977
dlp2534      |  -.9256109   .7272952    -1.27   0.207    -2.376525    .5253034
dyr2         |  (dropped)
dyr3         |     .05788   .0377758     1.53   0.130    -.0174806    .1332407
dyr4         |   .2300886   .0792884     2.90   0.005     .0719126    .3882646
dyr5         |   .2429812   .1087098     2.24   0.029     .0261112    .4598513
dyr6         |   .2159285   .1187687     1.82   0.073    -.0210087    .4528656
dyr7         |   .1775044   .1243088     1.43   0.158    -.0704849    .4254938
dyr8         |   .1855627   .1247856     1.49   0.142    -.0633777    .4345031
dyr9         |   .2184576   .1244927     1.75   0.084    -.0298986    .4668138
dyr10        |    .271181   .1253369     2.16   0.034     .0211407    .5212212
dyr11        |   .1919514   .1280677     1.50   0.138    -.0635368    .4474395
dyr12        |   .1984084   .1264836     1.57   0.121    -.0539195    .4507363
dyr13        |   .2187841   .1232808     1.77   0.080    -.0271544    .4647226
dyr14        |   .2418713   .1163528     2.08   0.041     .0097539    .4739887
dyr15        |   .2866917   .1063984     2.69   0.009     .0744327    .4989507
dyr16        |   .2004976   .0946694     2.12   0.038     .0116374    .3893579
dyr17        |   .1826651    .077764     2.35   0.022     .0275302    .3378001
dyr18        |   .1345287   .0629234     2.14   0.036     .0089999    .2600575
dyr19        |    .078249    .045497     1.72   0.090     -.012515    .1690131
dyr20        |   .0059974    .029325     0.20   0.839    -.0525043    .0644992
dyr21        |  (dropped)
_cons        |  -.0137706   .0216976    -0.63   0.528    -.0570561    .0295148
------------------------------------------------------------------------------
```

The lagged dependent variable is highly significant, indicating an omitted time dependent variable. Let's re-run the LM test and see if we still have autocorrelation.

```
. drop e

. predict e, resid
(10 missing values generated)

. regress e e_1 L.dlcrmaj dlprison dlmetpct dlblack dlrpcpi dlp1824 dlp2534
dyr2-
> dyr21

      Source |       SS       df       MS              Number of obs =      95
-------------+------------------------------           F( 26,    68) =    0.00
       Model |  .000021653    26  8.3281e-07           Prob > F      =  1.0000
    Residual |  .191218062    68   .00281203           R-squared     =  0.0001
-------------+------------------------------           Adj R-squared = -0.3822
       Total |  .191239715    94  .002034465           Root MSE      =  .05303

------------------------------------------------------------------------------
           e |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         e_1 |  -.0351238   .4002693    -0.09   0.930    -.8338488    .7636012
    dlcrmaj  |
          L1 |   .0326156   .3891471     0.08   0.933    -.7439154    .8091466
    dlprison |    .005018   .1923999     0.03   0.979    -.3789099     .388946
    dlmetpct |  -.0134529   2.144696    -0.01   0.995    -4.293127    4.266221
    dlblack  |   .0047841   .9022764     0.01   0.996     -1.79568    1.805249
    dlrpcpi  |  -.0015706   .3211793    -0.00   0.996     -.642474    .6393327
    dlp1824  |  -.0021095   .4782917    -0.00   0.996    -.9565257    .9523067
    dlp2534  |   .0017255   .7328458     0.00   0.998    -1.460646    1.464097
    dyr2     |  (dropped)
    dyr3     |  -.0002407   .0381491    -0.01   0.995    -.0763661    .0758847
    dyr4     |  -.0011592     .08095    -0.01   0.989    -.1626924    .1603739
    dyr5     |  -.0048292   .1225515    -0.04   0.969    -.2493768    .2397185
    dyr6     |  -.0047852   .1314744    -0.04   0.971    -.2671382    .2575679
    dyr7     |  -.0028128   .1292503    -0.02   0.983    -.2607277    .2551021
```

```
dyr8         |   -.000533   .1258393    -0.00   0.997    -.2516413    .2505754
dyr9         |  -.0003932   .1254777    -0.00   0.998      -.25078    .2499937
dyr10        |   -.001839   .1279756    -0.01   0.989    -.2572104    .2535324
dyr11        |  -.0043207   .1380765    -0.03   0.975    -.2798481    .2712066
dyr12        |  -.0022677   .1299976    -0.02   0.986     -.261674    .2571385
dyr13        |  -.0020183   .1262891    -0.02   0.987    -.2540242    .2499876
dyr14        |  -.0028452    .121601    -0.02   0.981    -.2454962    .2398058
dyr15        |  -.0041166   .1169896    -0.04   0.972    -.2375657    .2293326
dyr16        |  -.0063943   .1200121    -0.05   0.958    -.2458748    .2330862
dyr17        |  -.0048495   .0958629    -0.05   0.960     -.196141    .1864419
dyr18        |  -.0044938   .0814841    -0.06   0.956    -.1670927    .1581052
dyr19        |  -.0034812   .0606133    -0.06   0.954    -.1244332    .1174709
dyr20        |  -.0022393   .0390345    -0.06   0.954    -.0801315    .0756529
dyr21        |  (dropped)
_cons        |  -.0001079   .0218898    -0.00   0.996    -.0437884    .0435726
------------------------------------------------------------------------------

. test e_1

 ( 1)  e_1 = 0

       F(  1,    68) =    0.01
            Prob > F =    0.9303
```

The autocorrelation is gone. This is an acceptable short run model.

The only thing you miss under this scheme is the speed of adjustment to disequilibrium. This is beyond our scope because, so far, there are no good tests for cointegration in panel data.

Estimating first difference models has several advantages. Autocorrelation is seldom a problem (unless there are mis-specified dynamics and you have to add more lags). Also, variables that are in first differences are, usually, almost orthogonal. So, if they have very low correlations, omitted variables is much less of a problem than in regressions in levels, because, as we know from the omitted variable theorem, omitting a variable that is uncorrelated with the remaining variables in a regression will not bias the coefficients on those remaining variables. For the same reason, dropping insignificant variables from a first difference regression seldom makes a noticeable difference in the estimated coefficients on the remaining variables.

Clustering

A recent investigation of the effect of so-called "shall issue" laws on crime was done by Lott and Mustard[12] Shall-issue laws are state laws that make it easier for ordinary citizens to acquire concealed weapons permits. If criminals suspect that more of their potential victims are armed, they may be dissuaded from attacking at all. In fact, Lott and Mustard found that shall-issue laws do result in lower crime rates. Lott and Mustard estimated a host of alternative models, but the model most people focus on is a fixed effects model on all 3000 counties in the United States for the years 1977 to 1993. The target variable was a dummy variable for those states in those years for which a shall-issue law was in effect.

The clustering issue arises in this case because the shall-issue law pertains to all the counties in each state. So all the counties in Virginia, which has had a shall-issue law since 1989, get a one for the shall-issue dummy. There is no within-state variation for the dummy variable across the counties in the years with the law in effect. The shall issue dummy is said to be "clustered" within the states. In such cases, OLS

[12] Lott, John and David Mustard, "Crime, Deterrence and Right-to-Carry Concealed Handguns," *Journal of Legal Studies*, 26, 1997, 1-68.

underestimates the standard errors and overestimates the t-ratios. This is a kind of second order bias similar to autocorrelation. Lott has since re-estimated his model correcting for clustering.

If the data set consists of observations on **states** and we cluster on **states**, we get standard errors and t-ratios adjusted for both heteroskedasticity and autocorrelation (because the only observations within states are across years, i.e., a time series). The standard errors are not consistent as $T \to \infty$ because as T goes to infinity, the number of covariances within each state goes to infinity and the computation cumulates all the errors. That is why Newey and West weight the long lags at zero so as to limit the number of autocorrelations that they include in their calculations. However, since T will always be finite, we do not have an infinite number of autocorrelations within each state, so we might try clustering on states to see if we can generate heteroskedasticity and autocorrelation-corrected standard errors.

```
. areg lcrmaj lprison lmetpct lblack lrpcpi lp1824 lp2534 yr2-yr21,
absorb(state)
>  cluster(state)

Regression with robust standard errors                Number of obs =     105
                                                      F(  3,    74) =    9.79
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.9168
                                                      Adj R-squared =  0.8831
                                                      Root MSE      =  .08018

                           (standard errors adjusted for clustering on state)
------------------------------------------------------------------------------
             |               Robust
      lcrmaj |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lprison |  -.4886959   .4061591    -1.20   0.233    -1.297986    .3205937
     lmetpct |   .0850076   1.391452     0.06   0.951    -2.687522    2.857537
      lblack |  -1.006078   .7206686    -1.40   0.167    -2.442041    .4298858
      lrpcpi |   .9622313    .303612     3.17   0.002     .3572712    1.567191
      lp1824 |  -.5760403   .1872276    -3.08   0.003    -.9490994   -.2029813
      lp2534 |  -1.058662   .5089163    -2.08   0.041      -2.0727   -.0446244
       _cons |   6.137487   8.352663     0.73   0.465    -10.50556    22.78053
-------------+----------------------------------------------------------------
       state |   absorbed                                     (5 categories)
```

The coefficients on the year dummies have been suppressed to save space. The standard errors are much different, and much less significant. This probably means we had serious autocorrelation in the static model.

Monte Carlo studies show that standard errors clustered by the cross-section identifier in a fixed effect panel data regression (e.g. state in a state-year panel) are approximately correct. This is true for any arbitrary pattern of heteroskedasticity. It is also true for stationary variables with any kind of autocorrelation. Finally, clustered standard errors are approximately correct for nonstationary random walks whether they are independent or cointegrated. In summary, use standard errors clustered at the cross-section level for any panel data study.

Other Panel Data Models

While the fixed effects model is by far the most widely used panel data model, there are others that have been used in the past and are occasionally used even today. But first, a digression.

Digression: the between estimator[13]

Consider another very simple model, called the "between" model. In this approach we simply compute averages of each state over time to create a single cross section of overall state means.

$$\overline{Y}_{i.} = \sum_{k=1}^{K} \gamma_k \overline{X}_{k,i.} + e_i$$

where $\overline{Y}_{i.}$ is the mean for state i. and $\overline{X}_{k,i.}$ are the corresponding state means for the independent variables. Note that the period indicates that the means have been taken over time so that the time subscript is suppressed. This model is consistent if the pooled (OLS) estimator is consistent. It has the property that it is robust to measurement error in the explanatory variables and is sometimes used to alleviate such problems.

The compression of the data into averages results in the discarding of potentially useful information. For this reason the model is not efficient. The model ignores all variation that occurs within the state (across time) and utilizes only variation across (between) states. By contrast, the fixed effects model can be estimated by subtracting off the state means. Therefore, the fixed effect model uses only the information discarded by the between estimator. It completely ignores the cross section information utilized by the between estimator. For that reason, the fixed effects model is not efficient. The fixed effects model is a pure time series model. It utilizes only the time series variation for each state. This is the reason that the FE model is also known as the "within" model.

Because the between model uses only the cross section information and the within model uses only the time series information, the two models can be combined. The OLS model is a weighted sum of the between and within estimators.

The Random Effects Model

The main competitor to the fixed effects model is the random effects model. Suppose we assume that the individual state intercepts are independent random variables with mean α and variance σ_α^2. This means that we are treating the individual state intercepts as a sample from a larger population of intercepts and we are trying to infer the true population values from the estimated coefficients. The model can be summarized as follows.

$$Y_{i,t} = \alpha + \beta_t + \sum_{k=1}^{K} \gamma_k X_{k,i,t} + \varepsilon_{i,t}$$

where $\varepsilon_{i,t} = \eta_i + \upsilon_{i,t}$. Note that the intercepts are now in the error term and are omitted from the model.

[13] See Johnston, J. and J. Dinardo, *Econometric Methods*, 4th Edition, Mc-Graw-Hill, 1997, pp. 388-411, for an excellent discussion of panel data issues.

The random effects model can be derived as a feasible generalized least squares (FGLS) estimator applied to the panel data set. While the FE model subtracts off the group means from each observation, the RE model subtracts off a fraction of the group means, the fraction depending on the variance of the state dummies, the variance of the overall error term, and the number of time periods. The random effects model is efficient relative to the fixed effect model (because the latter ignores between state variation). The random effects model has the same problem as the OLS estimator, namely, it is biased in the presence of unobserved heterogeneity because it omits the individual state intercepts. However, the random effects model is consistent as N (the number of cross sections, e.g., states) goes to infinity (T constant) .

The random effects model can also be considered a weighted average of the between and within estimators. The weight depends on the same variances. If the weight is zero then the RE model collapses to the pooled OLS model. If the weight equals one, the RE model becomes the fixed effects model. Also, as the number of time periods, T, goes to infinity, the RE model approaches the FE model asymptotically.

The random effects model cannot be used if the data are nonstationary.

Choosing between the Random Effects Model and the Fixed Effects Model.

The main consideration in choosing between the fixed effects model and the random effects model is whether you think there is unobserved heterogeneity among the cross sections. If so, choosing the random effects model omits the individual state intercepts and suffers from omitted variable bias. In this case the only choice is the fixed effects model. Since I cannot imagine a situation where I would be willing to assume away the possibility of unobserved heterogeneity, I always choose the fixed effects model. If the cross-section dummies turn out to be insignificant as a group, I might be willing to consider the random effects model.

If you are not worried about unobserved heterogeneity, the next consideration is to decide if the observations can be considered a random sample from a larger population. This would make sense, for example, if the data set consisted of observations on individual people who are resurveyed at various intervals. However, if the observations are on states, cities, counties, or countries it doesn't seem to make much sense to assume that Oklahoma, for example, has been randomly selecting from a larger population of potential but unrealized states. There are only fifty states and we have data on all fifty. Again, I lean toward the FE model.

Another way of considering this question is to ask yourself if you are comfortable with generating estimates that are conditional on the individual states. This is appropriate if we are particularly interested in the individuals that are in the sample. Clearly, if we have data on individual people, we might be uncomfortable making inferences conditional on these particular people.

On the other hand, individuals can have unique unobserved heterogeneity. Suppose we are estimating an earnings equation with wages as the dependent variable and education, among other things, as explanatory variables. Suppose there is a an unmeasured attribute called "spunk:." Spunky people are likely to get more education and earn more money than people without spunk. The random effects model will overestimate the effect of education on earnings because the omitted variable, spunk, is positively correlated with both earnings and education. If people are endowed with a given amount of spunk that doesn't change over time, then spunk is a fixed effect and the fixed effects model is appropriate.

With respect to the econometric properties of the FE and RE models, the critical assumption of the RE model is not whether the effects are fixed. Both models assume that they are in fact fixed. The crucial assumption for the random effects model is that the individual intercepts are not correlated with any of the explanatory variables, i.e., $E\left[\alpha_i x_{k,i,t}\right] = 0$ for all i,k. Clearly this is a strong assumption.

The fixed effects model is the model of choice if there is a reasonable likelihood of significant fixed effects that are correlated with the independent variables. It is difficult to think of a case where correlated fixed effects can be ruled out *a priori*. I use the fixed effects model for all panel data analyses unless there is a good reason to do otherwise. I have never found a good reason to do otherwise.

Hausman-Wu test again

The Hausman-Wu test is (also) applicable here. If there is no unobserved heterogeneity, the RE model is correct and RE is consistent and efficient. If there is significant unobserved heterogeneity, the FE model is correct, there are correlated fixed effects, and the FE model is consistent and efficient while the RE model is now inconsistent. If the FE model is correct, then the estimated coefficients from the FE model will be unbiased and consistent while the estimated coefficients from the RE model will be biased and inconsistent. Consequently, the coefficients will be different from each other, causing the test statistic to be a large positive number. On the other hand, if the RE model is correct, then both models yield unbiased and consistent estimates, so that the coefficient estimates will be similar. (The RE estimates will be more efficient.) The HW statistic should be close to zero. Thus, if the test is not significantly different from zero, the RE model is correct. If the test statistic is significant, then the FE model is correct.

This test has one serious drawback. It frequently lacks power to distinguish the two hypotheses. In other words, there are two ways this model could generate an insignificant outcome. One way is to for the estimated coefficients to be very close to each other with low variance. This is the ideal outcome because the result is a precisely measured zero. In this case one can be confident in employing the RE model. On the other hand, the coefficients can be very different from each other but the test is insignificant because of high variance associated with the estimates. In this case it is not clear that the RE model is superior. Unfortunately, this is a very common outcome of the Hausman-Wu test.

My personal opinion is that the Hausman-Wu test is unnecessary for most applications of panel data in economics. Because states, counties, cities, etc. are not random samples from a larger population, and because correlated fixed effects are so obvious and important, the fixed effects model is to be preferred on *a priori* grounds. However, if the sample is a panel of individuals which could be reasonably considered a sample from a larger population, then the Hausman-Wu test is appropriate and is available in Stata using the "hausman" command. Type "help hausman" in Stata for more information.

The Random Coefficients Model.

In this model we allow each state to have its own regression. The model was originally suggested by Hildreth and Houck and Swamy.[14] The basic idea is that the coefficients are random variables such that $\gamma_{i,k} = \gamma_k + v_{i,k}$ where the error term $v_{i,k}$ has a zero mean and constant variance. The model is estimated in two steps. In the first step we use ordinary least squares to estimate a separate equation for each state. We then compute the *mean group estimator* $\hat{\gamma}_k = \sum_{i=1}^{N} \gamma_{i,k}$. In the second step we use the residuals from the state regressions to compute the variances of the estimates. The resulting generalized least squares estimate of $\hat{\gamma}_k$ is a weighted average of the original ordinary least squares estimates, where the weights are the inverses of the estimated variances.

[14] Hildreth, C. and C. Houck, "Some Estimators for a Linear model with Random Coefficients," *Journal of the American statistical Association*, 63, 1968, 584-595. Swamy, P. "Efficient Inference in a Random Coefficient Regression Model," *Econometrica,* 38, 1970, 311-32.

As a by-product of the estimation procedure, we can test the null hypothesis that the coefficients are constant across states. The test statistic is based on the difference between the state by state OLS coefficients and the weighted average estimate.

This method is seldom employed. It tends to be very inefficient simply because we usually do not have a long enough panel to generate precise estimates of the individual state coefficients. As a result, the coefficients have high standard errors and low t-ratios. Also, even if the test indicates that the coefficients vary across states, it is not clear whether the variance is really caused by different parameters or simple random variation across parameters that are really the same. Finally, the model does not allow us to incorporate a time dependent effect, β_t, which would control for common effects across all states. These disadvantages of the random coefficients model outweigh its advantages. I prefer the fixed effects model.

The random coefficients model is available in Stata using the xtrchh procedure (cross-section time-series random coefficient Hildreth-Houck procedure). We can use this technique to estimate the same crime equation.

```
. xtrchh lcrmaj lprison lmetpct lblack lrpcpi lp1824 lp2534

Hildreth-Houck random-coefficients regression   Number of obs      =      1070
Group variable (i): state                       Number of groups   =        51

                                                Obs per group: min =        20
                                                               avg =      21.0
                                                               max =        21

                                                Wald chi2(6)       =     24.35
                                                Prob > chi2        =    0.0004

------------------------------------------------------------------------------
      lcrmaj |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lprison |  -.1703223   .1173487    -1.45   0.147    -.4003215     .059677
     lmetpct |  -52.61075   49.73406    -1.06   0.290    -150.0877    44.86622
      lblack |   .0916276   7.147835     0.01   0.990    -13.91787    14.10113
      lrpcpi |  -.4169877   .1691149    -2.47   0.014    -.7484469   -.0855285
      lp1824 |  -1.471787   .3471744    -4.24   0.000    -2.152237   -.7913379
      lp2534 |   .4542666   .3208659     1.42   0.157    -.1746191    1.083152
       _cons |   243.5776   228.6247     1.07   0.287    -204.5186    691.6737
------------------------------------------------------------------------------
Test of parameter constancy:    chi2(350) = 37470.75      Prob > chi2 = 0.0000
```

This model is similar to the others estimated above. The fact that the coefficients are significantly different across states should not be a major concern, because we already know that the state intercepts are different. For that reason we use the fixed effects model. We really don't care about the coefficients on the control variables (metpct, black, income, and the age groups) so we don't need to allow their coefficients to vary across states. If we truly believe that a target variable (prison in this case) might have different effects across states, we could multiply prison by each of the state dummies, create 51 prison-state variables and include them in the fixed effects model. This combines the best attributes of the random coefficients model and the fixed effects model.

Stick with the fixed effects model for all your panel data estimation needs.

Index