

OLS Regression Using Stata – A Very Basic Introduction

The Stata dataset MRW.DTA contains data on real GDP per capita and related variables from the Penn World Table for a sample of 121 countries. This dataset was used to estimate growth equations by N.G.Mankiw, D.Romer and D.N.Weil, “A contribution to the empirics of economic growth”, *Quarterly Journal of Economics*, 1992.

The dataset can be downloaded from:

<http://www.nuffield.ox.ac.uk/teaching/economics/bond/>

The dataset contains 11 variables:

number	a country identifier between 1 and 121
country	country name (a string variable)
n	a dummy variable equal to one if the country is included in the non-oil sample
i	a dummy variable equal to one if the country is included in the intermediate sample
o	a dummy variable equal to one if the country is included in the oecd sample
rgdpw60	real GDP per working age population in 1960
rgdpw85	real GDP per working age population in 1985
gdpgrowth	average annual growth rate of real GDP per working age population between 1960 and 1985
popgrowth	average annual growth rate of the working age population between 1960 and 1985
i_y	real investment as a share of real GDP, averaged over the period 1960-85
school	% of working age population in secondary school

Opening the dataset

The dataset is already in Stata format. It can be opened in various ways, for example by opening Stata and using File Open to locate and open the dataset; or by double clicking the file in My Computer. If the dataset is opened correctly, a list of the variable names should appear in the variables window.

Some useful commands

<u>d</u> escribe [varlist]	- lists variable names, types and labels
<u>s</u> ummarize [varlist]	- basic descriptive statistics
<u>s</u> ummarize [varlist], <u>d</u> etail	- more descriptive statistics
<u>l</u> ist [varlist]	- displays values
<u>l</u> ist [varlist] if [condition]	- displays values if the specified condition is true

Examples

describe

summarize

su rgdpw85 if n==1

list if country=="Algeria"

Try typing these commands into the command window (one at a time), then hitting return. Hit the spacebar if the results window pauses, and prompts for -more-

Stata's help facility gives much more information on all these commands. e.g.

click Help Stata Command... and then type the name of the command

click Help Search... and then the topic

type help followed by the command name in the command window

Another useful command is `generate`. This allows new variables to be created by transforming the existing ones. e.g.

`generate lrgdpw85=ln(rgdpw85)`

`gen li_y=ln(i_y/100)`

`g lpop=ln(0.05+popgrowth/100)`

OLS regression

`regress` is Stata's basic command to compute OLS estimates.

Having generated the variables `lrgdpw85`, `li_y` and `lpop` as above, we can use the command

`regress lrgdpw85 li_y lpop if n==1`

Most estimation commands in Stata have a similar format. The first variable in the variable list is treated as the dependent variable. The remaining variables are the explanatory variables. The default is to include an intercept. The default standard errors reported are not robust to conditional heteroskedasticity.

This command should reproduce the results for the unrestricted specification in the first column of Table 1 in Mankiw, Romer and Weil (1992).

The standard output includes R^2 and \bar{R}^2 . Root MSE (mean squared error) is the estimate of $\hat{\sigma}$, denoted s.e.e. (standard error of the equation) in Table 1.

To reproduce the results for the restricted specification, we generate the variable

```
gen li_lpop=li_y-lpop
```

We can then use the command

```
reg lrgdpw85 li_lpop if n==1
```

Heteroskedasticity-robust standard errors

These can be obtained using the , vce(robust) option.

```
regress lrgdpw85 li_y lpop if n==1, vce(r)
```

In this example the robust standard errors are similar to the default standard errors, suggesting that conditional heteroskedasticity may not be important in this model.

Hypothesis tests

95% confidence intervals are reported for each estimated coefficient, as are t-statistics and p-values for tests of the null hypothesis that the corresponding coefficient is zero. The F-statistic tests the null hypothesis that all the slope coefficients are zero.

Other joint tests can be obtained using the test command.

e.g. to test the restriction that the coefficient on lpop is -1 times the coefficient on li_y, we could type

```
test lpop = -li_y
```

in the command window, after estimating the model of interest. Note that Stata reports an F(p,df) version of the test statistic, where p is the number of restrictions being tested, and df is the degrees of freedom N-K.

The test command allows other kinds of linear restrictions to be tested.

The testnl command allows non-linear restrictions to be tested.

Predicted values and residuals

Predicted values ($\hat{y} = X\hat{\beta}$) can be obtained after estimation, using

```
predict yhat if n==1
```

Residuals ($\hat{u} = y - X\hat{\beta}$) can be obtained after estimation, using

predict uhat if n==1, _residuals

Using

su lrgdpw85 yhat uhat if n==1

should then confirm that the mean of the predicted values equals the mean of the actual values, and the mean of the residuals is (essentially) zero.

Storing results

To store all results printed to the results window in a plain text file, first open a log file by typing

log using filename.log

in the command window. To stop subsequent results being saved in the log file, type

file log close

in the command window.

Alternatively to save results in a Stata output file, either click File Log Begin, or click the Begin Log button on the toolbar. To stop subsequent results being saved in this way, either click File Log Close, or click the Close Log button on the toolbar. The resulting files can be viewed using Stata's viewer.

Storing commands

For more serious use, commands can be stored in a simple program called a filename.do file, and executed from within the Do File Editor.

A simple example of a .do file could contain the commands

```
set more off
use c:\temp\mrw, clear
log using c:\temp\mrwexample, replace
gen lrgdpw85=ln(rgdpw85)
gen li_y=ln(i_y/100)
gen lpop=ln(0.05+popgrowth/100)
reg lrgdpw85 li_y lpop if n==1
test lpop = -li_y
log close
exit
```