STATA SESSION 2

TA : YOUNG JOON OH

STATA

Hi.

In the 2nd posting, I will cover simple, very simple regression model.

Please find **.dta** file and **do** file I attach. => click these links ; Stataforum2.do Stataforum2.dta (or find "Stataforum2.zip" and extract it)

Load dta file in your STATA, and load do file in your do-file editor in STATA.

» cd "C:\.....\ applied regression"

» use "Stataforum2.dta", clear

'cd' means changing directory. So select the folder in which you downloaded my files.

Then, just put file name for ’use’ command. Since you already set up your working directory, you don’t have to put all path for your dta file. Simply, you can just click dta file.

You can find three variables : Y, X1, X2.

To see the values for the data, click Data-Editor which is a button, having a magnifier.

» corr Y X1 X2

» corr Y X1 X2, cov

Look through correlations and covariances for these three variables. What do you find ? How about using graph.

» graph matrix Y X1 X2, half

We cannot find a significant relation between X2 and Y.

Run regression.

» regress Y X1

» reg Y X2

» reg Y X1 X2

’reg’ is the abbreviation form for regression.
Let’s compare these three models. How? you may learn this later in class.

From comparison, I decide to remove X2 in my dataset. (In this case, I just want to show how to delete a variable. But, I **DO NOT** recommend you to remove any original variable in your data set and delete any independent variable in your regression model just because the variable is insignificant.)

» drop X2

Let’s look relation between Y and X1 in graph, and run a regression.

» graph twoway scatter Y X1

» reg Y X1

Furthermore, make a graph with the regression line.

» graph twoway (scatter Y X1)(lfit Y X1)

Dots are your data, and the line is your regression line.

You can find there are differences between the dots and the line. What is it ?

They are estimated errors or ehat.

Let’s compute ehat. How can we get this ?

ehat = Y - Yhat ; remind previous lectures

» gen ehat= Y-(_b[_cons]+_b[X1]*X1)

This means I want make ehat which is Y - Yhat. To get Yhat, we need to use coefficients of the regression model. _b[_cons] is b0 and _b[X1] is b1.

Click Data-Editor to see our new variable.

Let’s add up all values of ehat.
What number you can get ?

It will be almost ZERO!!!!!.

This comes from a very important and powerful OLS assumption. The sum of error is always zero. Without this assumption, your regression result cannot be reliable. The reason is very intuitive. Think about it.

How to get MSE(Mean Square Error or conditional variance) ?

We need squared errors.

For this,
» gen ehat_sq = (ehat)^2

And sum up ehat_sq, and divide it with degree of freedom(DF). The sample size is 10, so DF is 8.

Done?

You can find the same numbers from the regression result. Look at the top-left table of the regression result for Y and X1.

You can see the numbers from 'sum-up ehat-sq' and 'sum-up ehat-sq' / DF. Furthermore you can find DF(8).

Let's compute b1 and b0.

For them, we need X deviations, Y deviations , and products of X deviation and Y deviation. ; Look at your note or textbook for the formula.

For this,

» sum X1

» gen xvar= X1-r(mean)

» gen xvar_sq = (xvar)^2

» sum Y

» gen yvar= Y-r(mean)

» gen xycorr = xvar*yvar

'xvar' is X1 - mean of X1, and its square root is 'xvar_sq'.

'yvar' is Y - mean of Y

'xycorr' is product of above both.

And compute B1 and B0 from these values. Compare your results with regression result.

For your convenience, you can use Excel.

» outsheet Y X1 ehat ehat_sq xvar xvar_sq yvar xycorr using Stataforum2.csv, comma replace

You can find 'Stataforum2.csv' in your working directory.

This is the end for 2nd posting

See you in class.