STATA SESSION 3

TA : YOUNG JOON OH

STATA

Hi..

I will cover creating dummy variable and Endogeneity problem(advanced level).

Please find the **do-file** and **.dta file**(dta file is only for endogeneity) Stataforum3.do Stataforum3.dta.

I will use the same data set as I used in the last posting.

```
» sysuse auto.dta, clear
```

Look at the variables.

I want to know whether or how much country of car manufacturers affects the price.

I can create a dummy for USA manufacturers by using another dummy variable ’foreign’

» gen usa = .

» replace usa = 1 if (foreign == 0)

» replace usa = 0 if (foreign == 1)

The first line means creating usa and allocating ’.’ which means missing value. If you look at the data, there is no missing value, but most of data set have missing values.

When you make a new variable, you had better allocate all values are missing values to avoid error. If your data set is perfect(no missing value, like this sample case), you do not have to do this.

The second line means if foreign has 0, which means domestic car, allocate usa = 1. The third line is for the opposite case.

Check your data with Data Editor in STATA.

However, there is a problem for foreign car, because ’make’ is not a numeric variable.

If you see the previous posting, you may know how to transform to a numeric variable. But I will use another way.

The easiest way to make dummy variables is using ’tab’. You learned it in the last week.

```
» tab make if foreign == 1, gen (model_dummy)
```

This means that if foreign = 1, create dummy variables, naming model_dummy with "make".

Check you variables window. You can find many dummy variables by every foreign make and model.

I want to categorize it into three cases (Japan, Germany and other(Italy and France))

```
» gen japan = 0
```

```
» replace japan = 1 if (model_dummy4==1 | model_dummy5==1 |model_dummy6==1
|model_dummy7==1 | /// - model_dummy9==1 |model_dummy10==1
|model_dummy11==1 | model_dummy14==1 | model_dummy15==1 | -
/// model_dummy16==1 | model_dummy17==1 )
```

In the first line, I create japan = 0, because there is no missing value in this data set.

If the condition is correct, japan can have value 1. For example, model_dummy4 is Datsun 200(I think this is Japanese car). If model_dummy4 is 1, which means Datsun200, variable japan can be 1.

′ | ′ means ′or′. If you want to use ′and′ condition, use ′&′.

′ /// ′ means if the command is too long, we can use this for multiple lines of one command.

```
» gen germany = 0
```

```
» replace germany = 1 if (model_dummy1==1 | model_dummy2==1
|model_dummy3==1 | /// - model_dummy18==1 | model_dummy19==1
| model_dummy20==1 |model_dummy21==1 )
```

```
» gen other = 0
```

```
» replace other = 1 if ( usa == 0 & japan == 0 & germany == 0 )
```

With the same way, I create germany and other.

. Furthermore, I decide mpg and gear_ratio as my additional independent variables.

Thus, my regression model is :

```
» reg price mpg gear_ratio usa japan germany other
```

Look at your result.

You will find one of your dummy variable is omitted by STATA !!! Why?

You will learn the reason and interpretation in the class.

Interpretation for Dummy is different.. Well... you may think it is same as other variables. But it is different.

If you want to know more about it mathematically. Use ’expected value’ Look at the text book page 215.

Next thing is about Endogeneity.

Dr. Barnes briefly explained it two weeks ago. You may remember it.

This is advanced level, so if you do not want to know it. Just skip it.

Load dta file I attached.

```
» reg y y1 x1 x2
```

This is my regression model. The result seems good., But I made a terrible mistake.

I omitted x3 and, instead, I put y1. Furthermore, x3 is an independent variable of y1. y1 is called endogenous variable.

Mathematically, y = b0 + b1 y1 + b2 x1 + b3 x2 + e1 ; e1 is error term.

y1 = z0 + z1 x3 + e2 ; e2 is error term.

So, we can rewrite this
y = b0 + b1 (z0 + z1 x3 + e2) + b2 x1 + b3 x2 + e1.

Why is it problem ? What makes it problem ?

I will show you the problem.
y = b0 + b1 (z0 + z1 x3) + b2 x1 + b3 x2 + e1+b1 e2.

Do you find it ? Look at the error term for my regression model.

The error term is e1+b1 e2 !!!!

b1 e2 is related with (z0 + z1 x3). This is a violation of OLS assumptions.

Error term should be random, but, in this case, it is not random.

To solve this problem, you need to use instrumental variable, like x3.

In practice, x3 is unobserved variable or omitted variable, so I used y1.

So we try to find x3. In reality, there may be many different types of x3... For example, if y1 is Party ID, x3 may be race or income or ideology, etc.

How to find endogeneity.

It is difficult, but there is a simple way in STATA. Use Hausman-Wu test

```
» ivreg y x1 x2 (y1 = x3 )
```

```
» ivendog
```

(For ivendog, in some versions of Stata, it may not be found. So, type, findit ivendog, and install "ivendog")

In the first line, I suppose y1 has a endogenous problem and to solve this, I use x3 for instrumental variable for y1.

This is call 2SLS(Two-Stage Least Squares) regression. In the second line is Hausman-Wu test.

The result is rejecting the null which is exogenous. Therefore, the regression model has endogeneity, so the OLS estimates are inconsistent. We need to use 2SLS. Sound like difficult.... I know...

Just try to understand what is endogeneity, how to find it, what is instrumental variable.

I think once you understand them, when you face this problem, you can solve it with papers or advanced books.

```
» reg y1 x1 x2 x3

» predict y1_res, res

» reg y y1 x1 x2 y1_res

» test y1_res
```

Compare the result with the result of "ivendog"
What do you find ? Consider them, you may find better understanding about endogeneity and the test.

This is the end of this posting.