Training 3

In this training we will get down to the core of econometrics, estimating regressions. We will start with the basic regression model and then consider how we might add more controls and estimate different types of regressions. The purpose is not to gain mastery of econometric principles, that needs to be done on your own or through a course. If you don’t understand all the econometric principles at this point, that is okay, you will soon.

The objectives for this module are:

Objective 1: Estimate regression

Objective 2: Create Tables for regression estimates

The data we will use for this training comes from the American Community Survey and is called training\_3\_data.dta

Objective 1: Estimate a regression

**Single Variable**

Let’s first consider the single variate model

 $\left(1\right) income=β\_{0}+β\_{1}\left(age\right)+u $

The code to do this is pretty straight forward:

 reg *depvar indepvar* where y is the depvar and x is the indepvar

For this example:

reg income age

The syntax is to have your y variable first and then your x variable after that. If you do that, you get the following output:

 

Let’s start at the bottom and work our way up. The bottom panel provides the OLS estimation of the coefficients. The coefficient on age (or $\hat{β}\_{1})$ is 775.0089. Also provided is the standard error. The coefficient on age and the standard error allow you to do a t-test by hand. STATA will report results for a default t-test of the null hypothesis that the coefficient equals 0. The t-state and p-value are displayed for this t-test. The 95% confidence interval indicates that based on the data and model, we are 95% confident that the population parameter (the true $β\_{1}$ is located within that interval. The \_cons is the coefficient $β\_{0}$ often referred to as the constant.

Looking at the top right, shows the number of observations. The other commonly used statistic is the R-squared. Remember this tells you how much of the variation of y is explained by the model. In case you want to calculate r-squared by hand, the top left shows the sum of squares (SS) for the model, residual, and total.

**Multivariate Model**

Now let’s look at a model that adds additional variables.

 $\left(2\right) income=β\_{0}+β\_{1}\left(age\right)+β\_{2}\left(female\right)+β\_{3}\left(married\right)+u$

The syntax for this is the same as before, except instead of one x variable you can put in as many as you want.

 reg income age female married

The order in which you enter the variables will determine the order of the estimated coefficients, but the order of the x-variables will not change the estimated coefficients. The results should look like this:

 

Dummy Variables or Fixed Effects

You may want a model that incorporates categorical dummy variables. In this example, let’s suppose that we want to control for race. There is a variable called race. Let’s tabulate that variable to see what are the possible values the variable can be.

 

Notice that there are 9 categories. To incorporate that in the model as dummy variables, you would need to create a dummy variable for each outcome of this variable. Then in the model, you would omit one dummy variable and include the other 8. You could do this, but that seems rather tedious. There is a shortcut that you can use.

You want to estimate model (3) where race is a vector of dummy variables of the 8 categories of race

 $\left(3\right) income=β\_{0}+β\_{1}\left(age\right)+β\_{2}\left(female\right)+β\_{3}\left(married\right)+β^{'}\left(race\right)+u$

To estimate this model, you need to create those dummy variables. One quick way to do that is with the following code:

 tab race, gen(raceg)

This will tabulate race and then create a variable for each outcome. The new variable is called raceg (race group) and will take on the value of raceg1 when race==1, raceg2=1 when race==2, and so on. you can then estimate model 3 with the following code:

 reg income age female married race2-race9

The code race2-race9 is a way to include all the variables from race2 to race9 into the model. In this case, race1 is the omitted category. Because these are controls, it doesn’t really matter which one is your omitted category. This will give you the following output:



Another way to do this is with the following code:

reg income age female married i.race



It provides coefficients for the 8 categorical dummy variables without having to create those dummy variables. It automatically will use the first outcome as the omitted category. And as we would expect, the coefficients are the same.

Another way you might see a group of dummy variables included in a model is through fixed effects. Essentially, fixed effects are a group of dummy variables. The coefficients are often not important for interpretation, so we want to include them in our model but don’t need to see the coefficients in the output.

While the race variable has only 9 potential values, it is not too hard to the coefficients in the output. But what if you wanted to control for occupation, there might be thousands of potential outcomes. You don’t want to include those coefficients in the output.

One way to do a fixed effect model is with the following code:

 reg income age female married, absorb(race)

absorb says to treat the variable race as a fixed effect but absorb or don’t show the coefficients in the output. This is what the output looks like:



As we would hope, the coefficients on our variables do not change (the constant does. That’s a discussion for another day).

With absorb, you can use only a single variable as the fixed effect.

Another way to control for fixed effects is to follow this two-step process:

1. set the xt
2. use xtreg

Here is the code that we would use

 xtset race

xtreg income age female married, fe

We set the xt to race and then use xtreg. For options we need to let STATA know to use the xt as a fixed effect. This gives the following output:

 

There are many ways to incorporate categorical dummy variables or fixed effects in your model. Depending on your model, data, and your research question, one method might be better than another.

Objective 2: Create Tables

There are two common ways to create tables from regression outputs: outreg2 and estout. This training will focus on outreg2 but will give you a link where you can learn more about estout.

Within STATA are built in commands (sum, tab, reg, etc) but there are other commands that others have programmed that don’t come with STATA. These are published in the STATA journal and are available for you to install to STATA. If you are using your personal computer, once you install these commands they will remain there. But if you are using a school computer, you will have to install these each time. These files are called .ado files, outreg2 is one of these ado files.

If you know the name of the ado file you can install it in the following way. The other way to type in the command prompt “search [ado name]” and this will allow you to search for the ado file in the STATA online database.

Let’s install outreg2

 *ssc install* outreg2

Here is the full syntax to use this command. This comes from typing help outreg2 after installing the file.

outreg2 [*varlist*] [*estlist*] using *filename* [, options] [: command]

This command will allow you export your regression results into an excel file. This is super helpful. It is used immediately after estimating the regression. You also might be wondering, is this what the output folder is for? And the answer is yes. So before I begin to estimate regressions, I will change the directory to this output folder.

Here is an example of using outreg2:

 cd "E:\My Drive\Heal STATA Trainings\Training 3\output"

reg income age

outreg2 using training3.xls, se bracket

This will create and excel spreadsheet (.xls and not an .xlsx) with the main parts of the output.

The great thing with this command, is we can create a table where each column represents a different regression. In the first part of this training, we estimated 6 different regressions. We can go to our do file and use outreg2 after each regression to create a table.

When you do this, the first command will create the file. If the file already exists, you will need to use the option of replace. Then for each additional regression you want to output and include as an additional column, you will use the option append.

The code will look something like this:

 

To open you’re the excel file, there will be a link in the STATA output window. Just click on that. Or you can use file explorer to open the excel file.

Note: make sure you close the excel file before you run the code again. Outreg2 cannot export to a file that is open on your browser.

estout

The other way to create tables is to use estout. Here is a great tutorial on how to do that:

<https://stats.idre.ucla.edu/stata/faq/how-can-i-use-estout-to-make-regression-tables-that-look-like-those-in-journal-articles/>

Summary

This training showed how to estimate a regression and create tables to present the results. There are many different econometric models that you will estimate when working on research. Often times, it is your econometric model that will drive what you need out of the data. You might find that you need to clean your data in such a way to use non-linear variables (dummy and categorical dummy variables are an example), or you may need to log variables, and much, much more.

As you read research by others, pay attention to their methods section. What models do they use and how can that help you with your model? Once you understand what you model is, the next challenge is to learn how to estimate that model in STATA.

The application will give you a chance to practice estimating models and presenting results in a table.