Lab 3: Linear Regression

Materials

- `graduation.dta`
- Do-file template `labtemplate_f21.do`

Objectives

By the end of this tutorial you should be able to complete the following tasks in Stata:

- Estimate and interpret a simple (two-variable) linear regression in levels, using continuous and binary variables, and use heteroskedasticity-robust standard errors.
- Identify $\hat{\beta}_0$, $\hat{\beta}_1$, standard errors, SST, SSE, SSR, and R^2 in Stata output and interpret them
- Calculate predicted values and residuals
- Create scatter plots
- Estimate a multivariate linear regression

Key commands

command	description
Estimation commands	
`regress var1 var2`	Estimate a regression, with `var1` as the dependent variable and `var2` as the independent variable(s)
`regress var1 var2, robust`	Estimate a regression with heteroskedasticity-robust standard errors
`correlate var1 var2 ... varn`	Calculate correlation coefficients of all listed variables, from `var1` to `varn`.
`graph twoway scatter var1 var2`	make a scatter plot with `var1` on the y-axis and `var2` on the x-axis.
Post-estimation commands[1]	
`predict newvar, xb`	Use estimated regression coefficients to predict $\widehat{y}$. It will generate `newvar`[2]

[1] Post-estimation commands must be run *immedately* after a regression, while the regression results are still held in your local variables

[2] Here, `newvar` equals $\widehat{newvar}_i = \widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$

command	description
`predict newvar, residuals`	Use estimated regression coefficients to predict residuals, generating `newvar`[3]
Working with data, missing values	
`count if var1 == 1`	count observations if the expression `var1 == 1` is true
`count if !missing(var1)`	count observations if `var1` is not missing
`drop if missing(var1)`	drop all observations where `var1` is missing
`tab var1, missing`	Include missing values in tabulation

Lab 3 Exercise

What do I submit?

- Your written up answers to exercise questions (1) - (13). This can be typed or written out then scanned (or photographed), in any reasonable format.
- The do-file you've created that runs this analysis
- A log file that contains the results from this exercise.

Questions

1. Download the do-file template and data files. Personalize the file paths so that you can run it and open your `graduation.dta` file. You can also work with a blank data file if you're more comfortable - just make sure you remember to include commands to start and close your log file.

2. Take a look at `graduation.dta`. How many observations are there? What is the distribution of treatment arms?[4]

3. There are six *continuous* food security variables[5]. You can look for them with `lookfor fs`. Pick one variable and write out a population model to determine the relationship assignment to graduation and food security. For the rest of this lab, I refer to the variable you chose as `foodsecurity`. If that's going to irritate you, you can rename your variable like this: `rename fsec5 foodsecurity`, using the variable name that you've chosen in place of `fsec5`.

4. Tabulate your food security value and check for missing observations. Drop any observations for which you have missing values of `foodsecurity` (see above for how to do this). How many observations are remaining?

5. Make a scatter plot of the relationship between your chosen food security variable and graduation (Include this in your submitted problem set). Is this easy to interpret? Calculate and report the associated correlation coefficient.

[3] Here, `newvar` equals $\widehat{newvar}_i = u_i = y_i - \widehat{\beta}_0 + \widehat{\beta}_1 x_i$

[4] There are a few variables here, including `treatment_arm`

[5] Not `fsec7`, which is categorical, or `fsec` which is always equal to 1

6. Conduct a t-test of whether the mean of `foodsecurity` is different between those who did and did not receive the graduation program[6]

7. Estimate the relationship between your chosen food security variable, `foodsecurity` and assignment to graduation, `graduation` using simple linear regression, with standard (homoskedasticity-assumed) standard errors. How do your t-statistics compare to what you found in the previous t-test? What was the impact of assignment to the graduation program on food security, based on your regression?

8. Re-estimate your regression, and this time adjust your standard errors to be heteroskedasticity-robust. Fill in the chart below with your estimates.

Variable	Estimate	Variable	Estimate
$\hat{\beta}_0$		$\hat{\beta}_1$	
R^2		TSS	
ESS		SSR	
d.f.		SER	

9. After that regression estimate, generate a new variable, `predict_fs` equal to the predicted value of your food security variable. Generate a second variable, `resid_fs` equal to the residual.

10. What is the mean of each variable? How does the mean of `predict_fs` compare to mean of `foodsecurity` in your sample?[7]

11. Examine the predicted value of your food security variable, `predict_fs`, for the *youngest* person in your sample.[8] What is its residual?

12. When we estimate a linear regression with no coefficients, sometimes we'll say we are "regressing on a constant." Regress `foodsecurity` *only* on a constant. What is $\hat{\beta}_0$, and how does it compare to overall mean?

13. For this final step, I'd like you to play around with the data. Pick **one** continuous dependent variable and **one** continuous *or* binary independent variable.[9] You can look at the correlation between two variables, or you can look at the impact of one of the program dimensions (group coaching, group livelihood, etc) on an *continuous* outcome of interest.

    a. Write a population model you want to estimate.
    b. Estimate it using OLS, adjusting your standard errors to be heteroskedasticity robust. Write an equation that reflects your estimated model in the form $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, replacing y and x with your chosen varables and replacing $\hat{\beta}_0$ and $\hat{\beta}_1$ with your estimates.
    c. In 1-2 sentences, , what do your results tell you, collectively?

[6] Hint; `ttest var1,by(var2)` will run a t-test of the mean of `var1` are equal for two groups determined by `var2`.

[7] If they differ, you should make sure you have dropped all missing values of `foodsecurity`! Try `sum predict_fs foodsecurity` to see if the sample sizes are the same

[8] Now is a good time to try out `lookfor age`

[9] Categorical variables that take on a just few observations, like the identity of your head of household, won't work here. You'll need to tabulate the variables to see what you're working with