

GV903: Advanced Research Methods

Lab 6: Inference and Prediction

Download the dataset `jeremymiles_grades` from moodle. This is data on 40 Politics students, containing their grades on a module, the hours they attended the class for the module, and a dummy whether they have studied 2 or more books for the course (against 0 or 1 book).

1. Use `des`, `list`, `sum`, `tab` and `kdensity` to **explore the dataset**. Before we continue with regressions you should have a feeling about how the data looks like, the coding and some descriptives.

2. **Make a scatterplot and run a regression of *grade* on *hours*. Fit the line in the plot.** Instead of using `lfit` use `function` instead. Here you will need to plug the numbers you got from the regression table. So, the second part of the `twoway` should look like

```
(function y = a + b*x , range(hours))
```

where `a` is the intercept and `b` the slope.

- **Test the $H_0 : \beta_h = 0$ using a *t*-test.**

3. **Run a regression of *grade* on *hours* and *books*.**

- **Make a scatterplot, but now add two lines.** Use the previous command but twice (that is adding another set of brackets), one plugging the numbers when `books` is 0 and one when `books` is 1. For some of the following tests you will need the command `test` after you run the regression.

- **Test the $H_0 : \beta_h = 0$ using a *t*-test.**
- **Test the $H_0 : \beta_b = 0$ using a *t*-test.**
- **Test the $H_0 : \beta_h = 0$ and $\beta_b = 0$ using a *F*-test.**
- **Test the $H_0 : \beta_h = \beta_b$ using a *F*-test.**

4. **Redo the scatterplot using different colors** for when `books` is 0 and 1. You need a `twoway` and the `if` condition for each case inside the brackets.

5. **Run the same regression but adding an interaction term** of `hours` and `books`. You can do that by

```
xi: reg grade i.books*hours
```

See that Stata automatically creates some variables for this. Could we have done that on our own? How? Redo the previous plot and add the fitted lines. Commands are exactly the same, but plugging the numbers needs more attention.

What about testing? Are t -tests helpful here? Think the multiple restrictions!

6. Here we introduce some regression diagnostics. Type `help reg post` to see the help page on that.

- Use the `predict` command to **get the fitted values and the residuals** (there are a lot more you can get though, but not now).

- **Test the normality of the residuals.** Use the `kdensity`, `pnorm`, `qnorm` and `swilk` commands for that. Ask Stata `help` if you need help on running them.

- Plot the residuals on the fitted values and add a horizontal line at 0.

- Use `estat hetttest` to perform the Breusch-Pagan **test for heteroskedasticity**.

- Use the Ramsey Regression Equation Specification Error Test (RESET) to **test that the model is correctly specified**.

To do so, get the fitted values (\hat{y}) from the model and generate \hat{y}^2 . You could also use higher-degree powers in addition to \hat{y}^2 (e.g. \hat{y}^3 and \hat{y}^4 too).

Then fit the model by adding the powers of the \hat{y} in the right part of the equation:

$$y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi} + \delta_1 \hat{y}_i^2 + \delta_2 \hat{y}_i^3 + \delta_3 \hat{y}_i^4 + u_i$$

Deciding on whether to put only \hat{y}_i^2 , \hat{y}_i^2 and \hat{y}_i^3 , or \hat{y}_i^2 , \hat{y}_i^3 and \hat{y}_i^4 depends on the sample size and the degree of mis-specification. In linear regression the power of 2 should be OK. Try all cases and perform the appropriate F -tests.

The RESET test is not robust to the presence of autocorrelation and heteroskedasticity, so it is generally advised to use robust errors when implementing the test.

Remember the first assumption you made in the lectures when you talked about OLS. There is a reason this was the first one! Having the correct specification is fundamental for the model, so never forget to test for mis-specification in the model you used with RESET.

7. Discuss and write a small paragraph on your conclusions from the analysis we did today. What would you advise a student who will undertake the same course based on your analysis on this 40 Politics students class?