Problem Set 2

Heteroskedasticity, Consistency, and Time Series

EC 421: Introduction to Econometrics

Due *before* 11:59pm on Friday, 08 February 2019

DUE Your solutions to this problem set are due *before* 11:59pm on Friday, 08 February 2019. Your files must be uploaded to Canvas—including (1) your responses/answers to the question and (2) the R script you used to generate your answers. Each student must turn in her/his own answers.

OBJECTIVE This problem set has three purposes: (1) reinforce the econometrics topics we reviewed in class; (2) build your R toolset; (3) start building your intuition about causality and time series within econometrics.

Heteroskedasticity

1. We are interested in using OLS to estimate the model

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{1}$$

where x_i is a categorical variable that takes the values 1, 2, or 3.

Suppose that we know $\text{Var}(u_i|x_i = 1) = 3$ and $\text{Var}(u_i|x_i = 2) = 3$. We do not know $\text{Var}(u_i|x_i = 3)$, *i.e.*, $\text{Var}(u_i|x_i = 3) = \sigma_3^2$ for some unknown parameter σ_3^2.

1a. What value must σ_3^2 take for our model to be homoskedastic?

1b. If $\sigma_3^2 \neq 3$, is OLS still unbiased? Is it still the most efficient linear estimator? Explain your answer.

1c. *Goldfeld-Quandt* In order to test whether the data we will use to estimate (1) are homoskedastic/heteroskedastic, we will run a Goldfeld-Quandt test.

We estimate (1) for the upper one third of the dataset (sorted on x) and find SSE$_3$=1,000. We estimate (1) on the middle third and find SSE$_2$=800. Finally, we estimate (1) on the lower third and find SSE$_1$=600. Each of these three groups has 100 observations.

Conduct a Goldfeld-Quandt test. State your hypotheses, calculate the G-Q test statistic, determine the *p*-value, state your conclusion.

Hint: You can use the function `pf(q, df1, df2, lower.tail = F)` to calculate the probability of observing a value of `q` or greater in an F distribution with `df1` numerator degrees of freedom and `df2` denominator degrees of freedom.

2. The dataset in this questions comes from "Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States" by Chetty, Hendren, Kline, and Saez—published in *The Quarterly Journal of Economics* (QJE) in 2014. Our outcome variable will be *the probability that an individual born to parents in the bottom fifth of the income distribution makes it into the top fifth of the income distribution*. This measure differs from the main outcome in the paper, but it is also very interesting—and it helps simplify our problem set. In individual observation in this dataset represents a commuting zone in the United States.

2a. Open up Rstudio, an R script, load whichever packages you want, and load the dataset contained in dataPS02.csv.

2b Describe the distribution of our main variable of interest (`prob_q5_q1`). You can provide statistical or graphical descriptions of this variable—try `summary(dataset$variable)` and `hist(dataset$variable)`, among others.

2c. Regress the probability an individual moves from the bottom fifth of income to the top fifth of income (`prob_q5_q1`) on an intercept and the share of the commuting zone that is *middle class* (`share_middleclass`). Report your findings—the coefficients, brief interpretations of the coefficients, and whether the coefficients are statistically significant.

2d. Does it make sense to interpret the intercept in this case? Explain.

2e. Plot the residuals from your regression in (2c) on the y axis and `share_middleclass` on the x axis. Do you see evidence of heteroskedasticity? Explain.

Hint$_1$: You can grab the residuals from a saved `lm` object by (1) using the `residuals()` function or (2) adding the suffix `$residuals` to the end of the `lm` object, *e.g.*, `my_reg$residuals` grabs the residuals from the `lm` object `my_reg`.

Hint$_2$: `plot(x = dataset$variable1, y = dataset$variable2)` makes quick and simple plots. You can also try `qplot()` from the package `ggplot2`, *i.e.*, `qplot(x = variable1, y = variable2, data = dataset)`.

2f. Conduct a Breusch-Pagan test for heteroskedasticity in the regression model in (2c). Describe your hypotheses, the test statistic, the *p*-value, and your conclusion.

2g. Conduct a White test for heteroskedasticity in the regression model in (2c). Describe your hypotheses, the test statistic, the *p*-value, and your conclusion.

Hint: To square the variable `x` in `lm()`, we write `lm(y ~ x + I(x^2), data = dataset)`.

2h. Let's imagine that we think heteroskedasticity is present. Estimate heteroskedasticity-robust standard errors. Do your standard errors change? What about the coefficients? Why is this the case?

Hint: To do this, use the `felm()` function in the `lfe` package. `felm()` takes a regression formula just like `lm()`. Then use `summary(., robust = T)` to show the heteroskedasticity-robust standard errors.

Example:

```
# The regression
some_reg ← lm(y ~ x, data = fake_data)
# Print the coefficients w/ het-robust standard errors
summary(some_reg, robust = T)
```

2i. As we discussed in class, we can introduce heteroskedasticity by mis-specifying our regression model. Try adding the additional variables from this dataset into the regression (possibly also adding interactions or squared explanatory variables). Then plot the new residuals against share middleclass (`share_middleclass`). *Briefly* describe which regressions you ran and whether it affected the appearance of heteroskedasticity.

Note: You do not need to formally test for heteroskedasticity.

2j. Should we take the regression in (2c) be *causal*? Explain your answer. If we cannot interpret the regression as causal, can we still learn something interesting here? Explain.

Consistency

3. Compare/contrast the concepts **expected value** and **probability limit**.

4. What does it mean if the estimator $\hat{\theta}$ is consistent for θ?

5. What is required for an omitted variable to make the OLS estimator $\hat{\beta}_j$ inconsistent for β_j?

6. Imagine that we are interested in the following model

$$\text{Health}_i = \beta_0 + \beta_1 \text{Money}_i + \beta_2 \text{Happiness}_i + u_i$$

but we are unable to measure an individual's *happiness*.

6a. If we simply omit *happiness* and estimate the equation

$$\text{Health}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{Money}_i + e_i$$

in which direction should we expect our estimate for β_1 to be biased? Explain your answer.

6b. Instead of omitting happiness, we decide to use a proxy for happiness—an individual's self-reported feeling of happiness (on a scale 1–10).

$$\text{Health}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{Money}_i + \hat{\beta}_2 (\text{Survey Happiness})_i + e_i$$

Should we expect our estimate $\hat{\beta}_2$ to over- or under-estimate the true value of β_2. Explain.

Time Series

7. Give an example in which a **static time-series model** might be appropriate. Briefly explain why it is appropriate.

8. Give an example in which a **static time-series model** would not be appropriate. Write down a **dynamic model** that would deal with the shortcomings of the failed static model.

9. Why are dynamic models with lagged dependent variables biased with OLS? Which of our assumptions do they violate? To answer this question, write out an ADL(1,0) model for time t and time $t + 1$.