Problem Set 2: Heteroskedasticity

EC 421: Introduction to Econometrics

Due *before* midnight (11:59pm) on Saturday, 08 February 2020

DUE Your solutions to this problem set are due *before* midnight on Saturday, 08 February 2020. Your files must be uploaded to Canvas.

IMPORTANT You must submit **two files**:

1. your typed responses/answers to the question (in a Word file or something similar)
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.

1. Suppose we are interested in estimating a model on house prices

$$\text{Price}_i = \beta_0 + \beta_1 \text{Area}_i + u_i \tag{1}$$

where Area_i is a continuous variable that takes positive values.

Suppose that we know $\text{Var}(u_i|\text{Area}_i) = \sigma^2 + \delta\text{Area}$.

1a. If $\delta = 0$, do we have heteroskedasticity? Is OLS unbiased? Explain your answer.

1b. If $\delta = 1$, do we have heteroskedasticity? Is OLS still unbiased? Explain your answer.

1c. *Goldfeld-Quandt* Suppose we decide to run a Goldfeld-Quandt test to detect heteroskedasticity.

If the ratio of the two SSE's (*i.e.*, $\text{SSE}_1/\text{SSE}_2$) is exactly 1, do we have evidence of heteroskedasticity? Explain.

1d. Define the term "standard error" and explain why it is important for econometrics.

2. Now for some real data. The data in the this problem come from housing sales in Ames, Iowa. We are going to think about modeling the house price at the time of a sale.

2a. Setup: Open up Rstudio and a fresh R script. Load whichever packages you want.

Now load the data in `002-data.csv`.

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

2c. Regress the sales price (`price`) on an intercept, the area (square feet) of the house (`area`), and the quality of the house (`quality` on a 1–10 scale). 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 `quality` 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)`. To take the interaction between two variables `x1` and `x2` in `lm()`, we use the colon (`:`), *i.e.*, write `lm(y ~ x1 + x2 + x1:x2, 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(the_estimate, robust = T)` to show the heteroskedasticity-robust standard errors.

Example:

```r
# The regression
some_reg = felm(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 misspecifying our regression model. Try taking the log of price. Then plot the new residuals against area. Explain whether this change in the regression specification suggests that misspecification was creating the heteroskedasticity.

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

2j. Should we interpret the regression results in (2c)—or your preferred specification in (2i)—as *causal*? Explain your answer. If we cannot interpret the regression as causal, can we still learn something interesting here? Explain.