Problem Set 2: Heteroskedasticity

EC 421: Introduction to Econometrics

Due *before* midnight on Friday, 01 May 2020

DUE Upload your answer on Canvas *before* midnight on Friday, 01 May 2020.

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.

If you are using RMarkdown, you can turn in one file, but it must be an HTML or PDF that includes your responses and R code.

README! As with the first problem set, the data in this problem set come from the 2018 American Community Survey (ACS), which I downloaded from IPUMS. The last page has a table that describes each variable in the dataset(s).

OBJECTIVE This problem set has three purposes: (1) reinforce the topics of heteroskedasticity and statistical inference; (2) build your R toolset; (3) start building your intuition about causality within econometrics/regression.

INTEGRITY If you are suspected of cheating, then you will receive a zero. We may report you to the dean.

Setup

Q01. Load your packages. You'll probably going to need/want `tidyverse` and `here` (among others).

Q02. Now load the data. This time, I saved the same dataset as a single format: a `.csv` file. Use a function that reads `.csv` files—for example, `read.csv()` or `read_csv()` (from the `readr` package in the `tidyverse`.

Q03. Check your dataset. Apply the function `summary()` to your dataset. You should have 12 variables.

Q04. Based upon your answer to **Q03**: What are the mean and median of household size (`hh_size`). What does this tell you about the distribution of the variable?

Q05. Based upon your answer to **Q03** What are the minimum, maximum, and mean of the indicator for whether a household moved in the last year (`i_moved`)? What does the mean of a binary indicator variable (such as `i_moved`) tell us?

Time and money

Q06. Suppose we are interested in the relationship between a household's housing costs and its time spent commuting. Plot a scatter plot/) (*e.g.*, using `geom_point()` from `ggplot2`) with housing cost (`cost_housing`) on the y axis and commute time (`time_commuting`) on the x axis.

Make sure you label your axes.

Q07. Based your plot in **Q06.**, if we regress housing costs on commute time, do you think we could have an issue with heteroskedasticity? Explain/justify your answer.

Q08. What issues can heteroskedasticity cause? (*Hint:* There are at least two main issues.)

Q09. Time for a regression.

Regress *housing cost* (`cost_housing`) on *commute time* (`time_commuting`) and *household income* (`hh_income`). Report your results—interpreting the intercept and coefficients and commenting on their statistical significance.

Reminder: The household income variable is measured in tens of thousands (meaning that a value of `3` tells us the household's income is $30,000).

Q10. Use the residuals from your regression in **Q09.** to conduct a Breusch-Pagan test for heteroskedasticity. Do you find significant evidence of heteroskedasticity? Justify your answer.

Hints

1. You can get the residuals from an `lm` object using the `residuals()` function, *e.g.*, `residuals(my_reg)`.
2. You can get the R-squared from an estimated regression (*e.g.*, a regression called `my_reg`) using `summary(my_reg)$r.squared`.

Q11. Now use your residuals from **Q09** to conduct a White test for heteroskedasticity. Does your conclusion about heteroskedasticity change at all? Explain why you think this is.

Hints: Recall that in R

- `lm(y ~ I(x^2))` will regress `y` on `x` squared.
- `lm(y ~ x1:x2` will regress `y` on the interaction between `x1` and `x2`.

Q12. Now conduct a Goldfeld-Quandt test for heteroskedasticity. Do you find significant evidence of heteroskedasticity? Explain why this result makes sense.

Specifics:

- We are still interested in the same regression (regressing the cost of housing on commute time and household income).
- Sort the dataset on **commute time**. The `arrange()` should be helpful for this task.
- Create you two groups for the Goldfeld-Quandt test by using the first **8,000** and last **8,000** observations (after sorting on commute time). The `head()` and `tail()` functions can help here.
- When you create the Goldfeld-Quandt test statistic, put the larger SSE value in the numerator.

Q13. Using the `lm_robust()` function from the `estimatr` package, calculate heteroskedasticity-robust standard errors. How do these heteroskedasticity-robust standard errors compare to the plain OLS standard errors you previously found?

Hint: `lm_robust(y ~ x, data = some_df, se_type = "HC2")` will calculate heteroskedasticity-robust standard errors.

Q14. Why did your coefficients remain the same in **Q13.**—even though your standard errors changed?

Q15. *If* you run weighted least squares (WLS), which the following four possibilities would you expect? Explain your answer.

1. The same coefficients as OLS but different standard errors.
2. Different coefficients from OLS but the same standard errors.
3. The same coefficients as OLS *and* the same standard errors.
4. Different coefficients from OLS *and* different standard errors.

Note: You do not need to run WLS.

Q16. Does heteroskedasticity appear to matter in this setting? Explain your answer/reasoning.

Description of variables and names

Variable	Description
`fips`	County FIPS code
`hh_size`	Household size (number of people)
`hh_income`	Household total income in $10,000
`cost_housing`	Household's reported monthly cost of housing (dollars)
`n_vehicles`	Household's number of vehicles
`hh_share_nonwhite`	Share of household members identifying as non-white ethnicities
`i_renter`	Binary indicator for whether any household members are renters
`i_moved`	Binary indicator for whether a household member moved in prior 1 year
`i_foodstamp`	Binary indicator for whether any household member participates in foodstamps
`i_smartphone`	Binary indicator for whether a household member owns a smartphone
`i_internet`	Binary indicator for whether the household has access to the internet
`time_commuting`	Average time spent commuting per day by each household member (minutes)

In general, I've tried to stick with a naming convention. Variables that begin with `i_` denote binary indicatory variables (taking on the value of `0` or `1`). Variables that begin with `n_` are numeric variables.