# Problem Set 2: Heteroskedasticity

## EC 421: Introduction to Econometrics

Due *before* midnight on Sunday, 07 February 2020

**DUE** Upload your answer on Canvas *before* midnight on Sunday, 07 February 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 (it's the same dataset as the first problem set with one new variable: education). 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 `r ncol(ps_df)`` variables. You might also want to check out the `skim()` function from the `skimr` package---it's a really useful function.

**Q04.** Based upon your answer to **Q03**: What are the mean and median of commute time (`time_commuting`)? 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 the individual has health insurance (`i_health_insurance`)? What does the mean of of this binary indicator variable (`i_health_insurance`) tell us?

# What's the value of an education?

**Q06.** Suppose we are interested in the "classic" labor regression: the relationship between an individual's education and her income. Plot a scatter plot with income on the y axis and approximate years of education on the x axis.

For the scatterplot, you might try `geom_point()` from ggplot2. Make sure you label your axes.

**Q07.** Based your plot in **Q06.**, if we regress personal income on education, 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.) Does it bias OLS when estimating coefficients?

**Q09.** Time for a regression.

Regress *personal income* (`personal_income`) on *education* (`education`) and our indicator for *female* (`i_female`). Report your results---interpreting the intercept and the coefficients and commenting on the coefficients' statistical significance.

*Reminder:* The personal-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`.
- The square of a binary variable is the same binary variable (and you don't want to include the same variable in a regression twice).

**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 personal income on education and the indicator for female).
- Sort the dataset on **education**. The `arrange()` should be helpful for this task.
- Create you two groups for the Goldfeld-Quandt test by using the first **1,600** and last **1,600** 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.** As we discussed in class, a misspecified model can cause heteroskedasticity. Let's see if that's the issue here.

Update your original model by adding an interaction between education and the indicator for female. In other words: In this new econometric model, you will regression personal income on an intercept, education, the indicator for female, and the interaction between education and female. Use heteroskedasticity-robust standard errors.

Interpret the coefficient on the interaction between `education` and `i_female` and comment on its statistical significance.

**Q17.** Based upon the model you estimated in **Q16.**, what is the expected personal income for women with 16 years of education? What about a man with 16 years of education?

**Q18.** Back to heteroskedasticity! Use the residuals from **Q16.** (where we attempted to deal with misspecification) to conduct a White test. Did changing our model specification "help"? Explain your answer.

**Q19.** Based upon your findings from the preceding questions: Do you think heteroskedasticity is present? If so: Does heteroskedasticity appear to matter in this setting?

Explain your answer/reasoning. **Include a plot of the residuals in your answer.**

**Q20.** In this assignment, we've largely focused on heteroskedasticity. But let's think a bit about the regressions you actually ran. Do you think the regression that we ran could suffer from omitted-variable bias? If you think there is omitted-variable bias, explain why and provide an example of "valid" omitted variable that would cause bias. If you do not think there is omitted-variable bias, justify your answer *using all of the requirements for an omitted variable.*

# Description of variables and names

| Variable | Description |
| --- | --- |
| state | State abbreviation |
| i_urban | Binary indicator for whether home county is 'urban' |
| age | The individual's age (in years) |
| i_asian | Binary indicator for whether the individual identified as Asian |
| i_black | Binary indicator for whether the individual identified as Black |
| i_hispanic | Binary indicator for whether the individual identified as Hispanic |
| i_indigenous | Binary indicator for whether the individual identified with a group indigenous to North Am. |
| i_white | Binary indicator for whether the individual identified as White |
| i_female | Binary indicator for whether the individual identified as Female |
| i_male | Binary indicator for whether the individual identified as Male |
| education | (Approximate) years of education |
| i_grad_college | Binary indicator for whether the individual graduated college (estimated) |
| i_married | Binary indicator for whether the individual was married at the time of the sample |
| personal_income | Total (annual) personal income (tens of thousands of dollars) |
| i_foodstamps | Binary indicator for whether the individual uses 'foodstamps' (SNAP) |
| i_health_insurance | Binary indicator for whether the individual has health insurance |
| i_internet | Binary indicator for whether the individual has access to the internet |
| time_depart | The time that the individual typically leaves for work (in minutes since midnight) |
| time_arrive | The time that the individual typically arrives at work (in minutes since midnight) |
| time_commuting | The length of time that the individual typically travels to work (in minutes) |

Variables that begin with i_ denote binary/indicator variables (taking on the value of 0 or 1).