Problem Set 2: Heteroskedasticity

EC 421: Introduction to Econometrics

Due *before* midnight on **Sunday, 06 February 2022**

Instructions

DUE Upload your answer on Canvas *before* midnight on Sunday, 06 February 2022.

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 a single file, but it must be a `html` or `pdf` file with **both** your R code **and** your answers.

If we ask you to create a figure or run a regression, then the figure or the regression results should be in the document that you submit (not just the code—we want the actual figure or regression output with coefficients, standard errors, *etc.*).

INTEGRITY If you are suspected of cheating, then you will receive a zero—for the assignment and possibly for the course. We may report you to the dean. **Cheating includes copying from your classmates, from the internet, and from previous assignments.**

README! We are using an extended version of the dataset as the last problem set. The data in this problem set come from three sources—all of which I downloaded from NHGIS.

1. the 1860 US Census
2. the 2010 US Census
3. the American Community Survey (ACS), 2006-2010

The table no the last page of the problem set describes each variable in the data.

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

KEY TOPICS In this problem set, you will continue to investigate the link between historical repressive/discriminatory institutions and current outcomes—with a focus on inference and the assumptions behind our estimates and inferences. Specifically, we will look at the modern effects of slavery in the United States—focusing on states that previously allowed slavery. Most of these states unsuccessfully attempted to secede, while a small set joined the 'Union'. In this problem set, you will investigate how the intensity of enslavement in 1860 correlates with (or affects) economic outcomes for Black and White individuals 150 years later (in 2010).

Part 1: Setup

Q1.1 Load your R packages. You're probably going to need/want `tidyverse` and `here` (among others). Also: `pacman` and it's `p_load()` function make package management easier. The `fixest` package will also be helpful.

Q1.2 Now load the data (stored in `002-data.csv`).

The updated dataset is saved as a CSV, so you'll want to use a function that can read CSV files—for example, `read_csv()` in the `readr` package, which is part of the `tidyverse`.

Q1.3 How many states are in the dataset this time?

Part 2: Conceptual questions

2.1 What is the difference between *homoskedasticity* and *heteroskedasticity*?

2.2 How does homoskedasticity differ from exogeneity?

2.3 Which of OLS's assumptions/requirements does heteroskedasticity violate?

2.4 What problems does heteroskedasticity cause for OLS?

2.5 What are our options for "dealing with" heteroskedasticity?

Part 3: Testing for heteroskedasticity

3.1 Regress counties' median Black household income in 2010 (`income_black_2010`) on two explanatory variables:

- the percent of county's 1860 population who were enslaved (`pct_pop_enslaved_1860`)
- the county's 2010 population (`pop_total_2010`)

Store this regression's output. **Interpret** the coefficient on `pct_pop_enslaved_1860`.

3.2 Using the residuals from the previous problem (**3.1**), create two scatter plots:

- Plot the residuals from **3.1** on the y-axis and `pct_pop_enslaved_1860` on the x-axis
- Plot the residuals from **3.1** on the y-axis and `pop_total_2010` on the x-axis

Hint: You can grab the residuals from a regression object using the `residuals()` function. For example, if you named the output of `lm() my_reg`, then you can access its residuals via `residuals(my_reg)`. (You might even try adding them to a dataset with `df$resid = residuals(my_reg)`).

3.3 Do the scatter plots in **3.2** suggest that our disturbances may be heteroskedastic? Explain.

3.4 Conduct a Goldfeld-Quandt test where your data are arranged by `pct_pop_enslaved_1860` and you have 316 observations in each of the two groups you are using for the test (meaning you omit the 'middle' 316). Report your test stastic, *p-value*, and final conclusion of the test.

Hint 1: You can arrange your data using the `arrange()` function (in the `tidyverse`).
Hint 2: Put the larger of the two SSEs in the numerator.
Hint 3: The notes in lecture `04` walk you through this test.

3.5 If you had conducted the Goldfeld-Quandt test using `pop_total_2010`, then you would have found a *p-value* of approximately 0.4097. Does this *p-value* suggest that there is heteroskedasticity? How does this answer, when compared to your answer in **3.4**, illustrate a weakness of the Goldfeld-Quandt test?

3.6 Conduct a White test for heteroskedasticity. Report your test statistic, *p-value*, and final conclusion of the test.

Part 4: Correcting for heteroskedasticity

4.1 Estimate the regression in **3.1** using weighted least squares (WLS) rather than OLS, where you weight by the county's Black population in 2010 (`pop_black_2010`).biggest

Report the results of the regression and discuss any changes in the coefficients and standard errors.

Hint: You can estimate WLS by providing `lm()` with a `weights` argument (specifically: the variable you want to use for weight). For example, to weight by the `w` variable in the `fake_df` dataset: `lm(y ~ x, data = fake_df, weights = w)`. We did something very similar in the lecture notes.

4.2 Why do the estimated coefficients change when we move from OLS (**3.1**) to WLS (**4.1**)?

4.3 Now estimate the coefficients with OLS but use **heteroskedasticity-robust standard errors**. Are these heteroskedasticity-robust standard errors very different from the standard errors that assume homoskedasticity?

Hint: The lecture and lab notes teach you how to do this with `fixest` and `feols()`.

Part 5: Correcting for correlated disturbances

5.1 Using `feols()` (from the `fixest` package): Re-estimate the regression from **3.1** but with **cluster-robust** standard errors (i.e., standard errors that are robust to correlated disturbances).

Allow the errors to cluster (correlate) at the state level (use the `state` variable).

How do your standard errors change?

Hint 1: You can allow for correlated disturbances using `feols` in two different ways:

- `feols(y ~ x, data = fake_df, cluster = 'state')`
- `feols(y ~ x, data = fake_df) %>% summary(cluster = 'state')`

In both cases, we're allowing correlation at the state level (using a variable named `state`).

Hint 2: The lecture slides also show you how to do this *clustering*.

Part 6: Interpreting indicators and interactions

Important: You should use heteroskedasticity-robust standard errors for the rest of the assignment.

6.1 Regress the median Black household's income in 2010 (`income_black_2010`) on two explanatory variables:

1. the county's percent of the population that was enslaved in 1860 (i.e., `pct_pop_enslaved_1860`)
2. the indicator for whether the county's state joined the Confederacy (`was_confederate`)

Report your regression results (with het.-robust standard errors).

6.2 Interpret the coefficient on `was_confederate` *and* comment on whether it is statistically significant.

6.3 Regress the median Black household's income in 2010 (`income_black_2010`) on three explanatory variables:

1. the county's percent of the population that was enslaved in 1860 (i.e., `pct_pop_enslaved_1860`)
2. the indicator for whether the county's state joined the Confederacy (`was_confederate`)
3. the **interaction** between the two previous variables.

Report your regression results (just the regression table—no need to interpret anything yet).

Hint: You can take the interaction between two variables in `lm` (and other regression functions) using `:`, for example `y ~ x1 + x2 + x1:x2`.

6.4 Interpret the coefficient on the **interaction** between the `pct_pop_enslaved_1860` and `was_confederate`.

6.5 In Confederate states: Is a larger 1860 population share of enslaved people (larger values of `pct_pop_enslaved_1860`) positively or negatively associated with 2010 Black income? Explain your answer.

Description of variables and names

Variable	Description
state	State name
county	County name
gisjoin	Code for GIS joining
pop_enslaved_1860	County population of enslaved persons in 1860 (Census)
pop_total_1860	County population in 1860 (Census)
pop_total_2010	County population in 2010 (Census)
pop_black_2010	County black population in 2010 (Census)
pop_white_2010	County white population in 2010 (Census)
income_black_2010	County median income for Black households 2006-2010 (ACS)
income_white_2010	County median income for White households 2006-2010 (ACS)
pct_pop_enslaved_1860	Percent of county population enslaved in 1860 (Census)
had_rosenwald_school	Indicator variable: Did the county have a Rosenwald school?
was_confederate	Indicator variable: Did the state join the Confederacy?