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).

Answer:

```
# Load packages
library(pacman)
p_load(tidyverse, broom, here)
```

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`.

Answer:

```
# Load dataset
ps_df = here("002-data.csv") %>% read_csv()
```

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

Answer:

```
# Summary of 'ps_df' variables
summary(ps_df)
```

```
#>      fips              hh_size         hh_income       cost_housing     n_vehicles
#>  Length:25000       Min.   : 1.00   Min.   :  0.0   Min.   :   4   Min.   :0.00
#>  Class :character   1st Qu.: 2.00   1st Qu.:  4.6   1st Qu.: 700   1st Qu.:1.00
#>  Mode  :character   Median : 2.00   Median :  8.0   Median :1100   Median :2.00
#>                     Mean   : 2.83   Mean   : 10.6   Mean   :1278   Mean   :2.04
#>                     3rd Qu.: 4.00   3rd Qu.: 13.0   3rd Qu.:1600   3rd Qu.:3.00
#>                     Max.   :17.00   Max.   :143.6   Max.   :7400   Max.   :6.00
#>  hh_share_nonwhite    i_renter         i_moved        i_foodstamp       i_smartphone
#>  Min.   :0.000     Min.   :0.000   Min.   :0.000   Min.   :0.0000   Min.   :0.000
#>  1st Qu.:0.000     1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:1.000
#>  Median :0.000     Median :0.000   Median :0.000   Median :0.0000   Median :1.000
#>  Mean   :0.233     Mean   :0.376   Mean   :0.189   Mean   :0.0844   Mean   :0.936
#>  3rd Qu.:0.400     3rd Qu.:1.000   3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:1.000
#>  Max.   :1.000     Max.   :1.000   Max.   :1.000   Max.   :1.0000   Max.   :1.000
#>    i_internet     time_commuting
#>  Min.   :0.000   Min.   :  0.2
#>  1st Qu.:1.000   1st Qu.: 15.0
#>  Median :1.000   Median : 30.0
#>  Mean   :0.948   Mean   : 36.7
#>  3rd Qu.:1.000   3rd Qu.: 47.5
#>  Max.   :1.000   Max.   :376.0
```

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?

Answer: The mean and median of household size are 2.834 and 2, respectively. Because the median is relatively larger than the mean it tells us that the right tail of the distribution of household size is skewed—meaning there are a small number of very large households.

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?

Answer: The minimum, maximum, and mean of `i_moved` are 1, 17, and 2.834, respectively.

The mean of a binary indicator variable tells us the share of individuals whose value equals one (here: the share of households that moved in the last year).

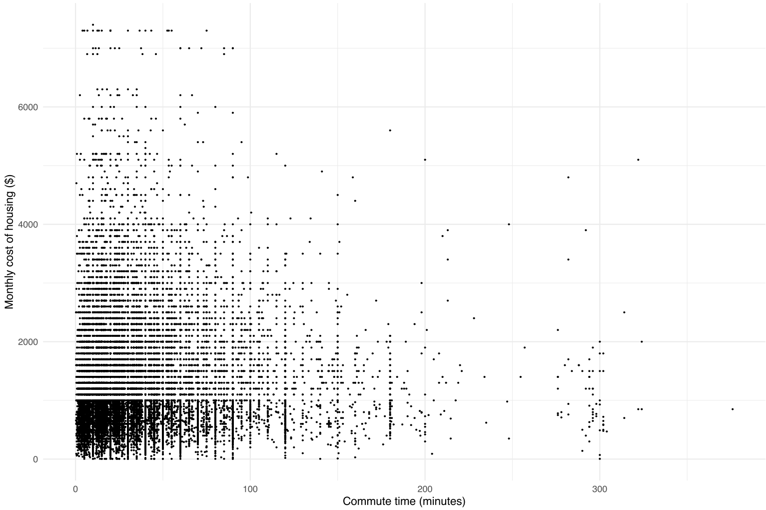
Time and money

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.

Answer:

```r
ggplot(data = ps_df, aes(x = time_commuting, y = cost_housing)) +
geom_point(size = 0.25) +
labs(x = "Commute time (minutes)", y = "Monthly cost of housing ($)") +
theme_minimal()
```



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.

Answer: We may very well have heteroskedastic disturbances in the given regression: it appears as though the variance of our outcome variable (which depends upon the variance of the disturbance) grows as our explanatory variable grows.

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

Answer: Heteroskedasticity causes our standard errors to be biased (which affects inference—*e.g.*, hypothesis tests, confidence intervals). Heteroskedasticity also makes OLS regression less efficient for estimating coefficients.

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).

Answer:

```
# Regression
est09 = lm(cost_housing ~ time_commuting + hh_income, data = ps_df)
# Results
est09 %>% tidy()
```

```
#> # A tibble: 3 x 5
#>   term           estimate std.error statistic  p.value
#>   <chr>             <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)     801.        8.27      96.8  0
#> 2 time_commuting    0.473     0.140      3.37 0.000756
#> 3 hh_income        43.4       0.451     96.0  0
```

We find statistically significant relationships between the cost of housing and each of our explanatory variables—commute time and household income.

- The intercept tells us the expected cost of housing (800.8162) for someone with zero commute time and zero income.
- The coefficient on `time_commuting` tells us an additional minute of commuting is significantly associated with a $0.473 increase in the cost of housing.
- The coefficient on `time_commuting` tells us an additional $10K of household income (1 unit of `hh_income`) is significantly associated with a $43.352 increase in the cost of housing.

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`.

Answer:

```
# Regression for BP test
est10 = lm(residuals(est09)^2 ~ time_commuting + hh_income, data = ps_df)
# Results
est10 %>% tidy()
```

```
#> # A tibble: 3 x 5
#>   term           estimate std.error statistic p.value
#>   <chr>             <dbl>     <dbl>     <dbl>   <dbl>
#> 1 (Intercept)      37529.    16924.      2.22  0.0266
#> 2 time_commuting    -684.      287.     -2.38  0.0172
#> 3 hh_income        49624.      923.     53.7   0
```

```
# BP test statistic
lm10 = summary(est10)$r.squared * nrow(ps_df)
# Test against Chi-squared 2
pchisq(lm10, df = 2, lower.tail = F) %>% round(3)
```

```
#> [1] 0
```

The *p*-value is extremely small—nearly zero, so we reject the null hypothesis and conclude that there is statistically significant evidence of heteroskedasticity.

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`.

Answer:

```
# Regression for BP test
est11 = lm(
    residuals(est09)^2 ~
    time_commuting + hh_income +
    I(time_commuting^2) + I(hh_income^2) +
    time_commuting:hh_income,
    data = ps_df
)
# Results
est11 %>% tidy()
```

```
#> # A tibble: 6 x 5
#>   term                     estimate std.error statistic  p.value
#>   <chr>                       <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)              175747.    24663.       7.13 1.06e-12
#> 2 time_commuting            -1518.      662.      -2.29 2.19e- 2
#> 3 hh_income                 31974.     2156.      14.8  1.47e-49
#> 4 I(time_commuting^2)           8.22      3.12     2.63 8.51e- 3
#> 5 I(hh_income^2)              330.       29.6     11.1  1.01e-28
#> 6 time_commuting:hh_income    -30.4      27.0     -1.13 2.59e- 1
```

```
# BP test statistic
lm11 = summary(est10)$r.squared * nrow(ps_df)
# Test against Chi-squared 5
pchisq(lm11, df = 5, lower.tail = F) %>% round(3)
```

```
#> [1] 0
```

The *p*-value is still extremely small—nearly zero, so we reject the null hypothesis and conclude that there is statistically significant evidence of heteroskedasticity. The result did not change because we already found strong evidence of heteroskedasticity, and the White test is just a more flexible test for heteroskedasticity.

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.

Answer:

```
# Arrange the dataset by commute time
ps_df = ps_df %>% arrange(time_commuting)
# Create the two subsets (first and last 8,000 observations)
g1 = head(ps_df, 8000)
g2 = tail(ps_df, 8000)
# Run the two regressions
est12_1 = lm(cost_housing ~ time_commuting + hh_income, data = g1)
est12_2 = lm(cost_housing ~ time_commuting + hh_income, data = g2)
# Find the SSE from each regression
sse1 = sum(residuals(est12_1)^2)
sse2 = sum(residuals(est12_2)^2)
# GQ test statistic
gq = sse1 / sse2
# p-value
pf(gq, df1 = 8000, df2 = 8000, lower.tail = F)
```

```
#> [1] 0.3621
```

Using the Goldfeld-Quandt test for heteroskedasticity, we fail to reject the null hypothesis of *homoskedasticity* with a *p*-value of approximately 0.362.

It makes since that we are finding as different result as the Goldfeldt-Quandt test for heteroskedasticity can be very sensitive to the type of heteroskedasticity or to the variable that we choose to consider. In this case, we are considering **only** commute time, when the previous tests also included income.

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?

Answer:

```
# Load estimatr package
p_load(estimatr)
# Estimate het-robust standard errors
lm_robust(
    cost_housing ~ time_commuting + hh_income,
    data = ps_df,
    se_type = "HC2"
) %>% summary()
```

```
#>
#> Call:
#> lm_robust(formula = cost_housing ~ time_commuting + hh_income,
#>     data = ps_df, se_type = "HC2")
#>
#> Standard error type:  HC2
#>
#> Coefficients:
#>                Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper    DF
#> (Intercept)     800.816     10.445   76.67  0.00000  780.342  821.290 24997
#> time_commuting    0.473      0.146    3.24  0.00118    0.187    0.759 24997
#> hh_income        43.352      1.003   43.23  0.00000   41.386   45.317 24997
#>
#> Multiple R-squared:  0.271 ,    Adjusted R-squared:  0.271
#> F-statistic:  973 on 2 and 24997 DF,  p-value: <2e-16
```

The heteroskedasticity-robust standard errors are larger than the OLS standard errors—especially the standard error for household income. The standard error for household income more than doubles.

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?

Answer: Our coefficients have not changed because we are still using OLS to estimate the coefficients. The thing that has changed is how we calculate the *standard errors* (not the coefficients).

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.

Answer: With WLS, we would expect our coefficients and standard error to differ from OLS. We expect this because WLS is a different estimator than OLS, which produces different estimates, different residuals, and different standard errors.

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

Answer: Heteroskedasticity does appear to be present. It is causing us to over-estimate our precision—especially for the relationship between commute time and income. For example, our t statistic drops from 96 to 43. However, the t statistic of 43 is still highly significant, so adjusting for heterskedasticity doesn't really change our results/understanding much in this setting.

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.