Problem Set 2: Heteroskedasticity

EC 421: Introduction to Econometrics

Due *before* midnight on Saturday, 06 February 2020

DUE Upload your answer on Canvas *before* midnight on Saturday, 06 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).

Answer:

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

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`.

Answer:

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

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.

Answer:

```r
# Summary of 'ps_df' variables
summary(ps_df)
# Skim the dataset
# skim(ps_df)
```

continued on next page...

```
#>     state              i_urban           age          i_asian           i_black      
#>  Length:5000        Min.   :0.000   Min.   :16.0   Min.   :0.0000   Min.   :0.0000  
#>  Class :character   1st Qu.:0.000   1st Qu.:31.0   1st Qu.:0.0000   1st Qu.:0.0000  
#>  Mode  :character   Median :1.000   Median :43.0   Median :0.0000   Median :0.0000  
#>                     Mean   :0.614   Mean   :43.2   Mean   :0.0566   Mean   :0.0826  
#>                     3rd Qu.:1.000   3rd Qu.:55.0   3rd Qu.:0.0000   3rd Qu.:0.0000  
#>                     Max.   :1.000   Max.   :94.0   Max.   :1.0000   Max.   :1.0000  
#>    i_hispanic     i_indigenous       i_white         i_female         i_male     
#>  Min.   :0.000   Min.   :0.0000   Min.   :0.000   Min.   :0.000   Min.   :0.000  
#>  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:0.000   1st Qu.:0.000  
#>  Median :0.000   Median :0.0000   Median :1.000   Median :0.000   Median :1.000  
#>  Mean   :0.148   Mean   :0.0084   Mean   :0.785   Mean   :0.486   Mean   :0.514  
#>  3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:1.000  
#>  Max.   :1.000   Max.   :1.0000   Max.   :1.000   Max.   :1.000   Max.   :1.000  
#>    education    i_grad_college    i_married     personal_income   i_foodstamps   
#>  Min.   : 7.0   Min.   :0.000   Min.   :0.000   Min.   :  0.00   Min.   :0.0000  
#>  1st Qu.:12.0   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:  2.40   1st Qu.:0.0000  
#>  Median :13.0   Median :0.000   Median :1.000   Median :  4.20   Median :0.0000  
#>  Mean   :13.8   Mean   :0.367   Mean   :0.544   Mean   :  6.02   Mean   :0.0718  
#>  3rd Qu.:16.0   3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:  7.00   3rd Qu.:0.0000  
#>  Max.   :17.0   Max.   :1.000   Max.   :1.000   Max.   :135.34   Max.   :1.0000  
#>  i_health_insurance   i_internet     time_depart    time_arrive   time_commuting 
#>  Min.   :0.000      Min.   :0.000   Min.   :  15   Min.   :  39   Min.   :  1.0  
#>  1st Qu.:1.000      1st Qu.:1.000   1st Qu.: 392   1st Qu.: 419   1st Qu.: 15.0  
#>  Median :1.000      Median :1.000   Median : 452   Median : 474   Median : 20.0  
#>  Mean   :0.911      Mean   :0.949   Mean   : 495   Mean   : 524   Mean   : 27.2  
#>  3rd Qu.:1.000      3rd Qu.:1.000   3rd Qu.: 512   3rd Qu.: 544   3rd Qu.: 35.0  
#>  Max.   :1.000      Max.   :1.000   Max.   :1425   Max.   :1434   Max.   :188.0  
```

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?

Answer: The mean and median of commute time are 27.244 and 20, respectively. Because the mean is quite a bit larger than the median it tells us that the right tail of the distribution of household size is skewed---meaning there are a small number of individuals with very long commutes.

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?

Answer: The minimum, maximum, and mean of `i_health_insurance` are 0, 1, and 0.911, respectively.

The mean of a binary indicator variable tells us the share of individuals whose value equals one. Here: We learn that in the sample, approximately 91% of individuals had some type of health insurance.

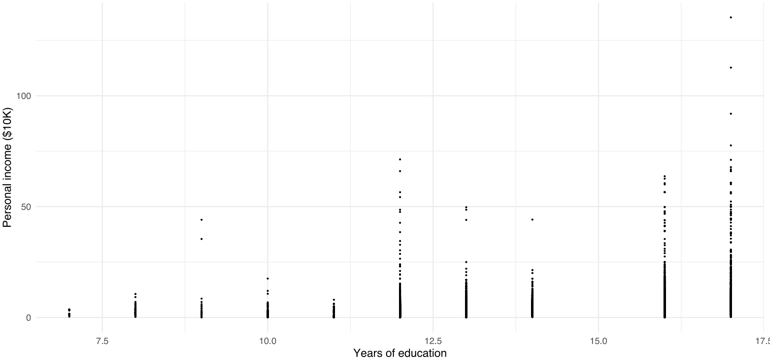
What's the value of an education?

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.

Answer:

```
ggplot(data = ps_df, aes(x = education, y = personal_income)) +
geom_point(size = 0.25) +
scale_y_continuous("Personal income ($10K)") +
scale_x_continuous("Years of education") +
theme_minimal()
```



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.

Answer: We may very well have heteroskedastic disturbances in the described 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. There are also certainly levels of education with more variance than others (*e.g.*, 12 years and 16 years).

Q08. What issues can heteroskedasticity cause? (*Hint:* There are at least two main issues.) Does it bias OLS when estimating coefficients?

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.

On the other hand, heteroskedasticity **does not** bias OLS when estimating linear regression 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).

Answer:

```r
# Regression
est09 = lm(personal_income ~ education + i_female, 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)    -8.39    0.642      -13.1 2.04e- 38
#> 2 education       1.14    0.0460      24.8 3.64e-128
#> 3 i_female       -2.76    0.201      -13.7 3.53e- 42
```

We find statistically significant relationships between individuals' incomes and each of our explanatory variables---both education and our indicator for "female."

- The intercept tells us the expected income (-8.3918) for **a man** with **zero education** (which we do not observe in the actual data).
- The coefficient on `education` tells us that a each additional year of education is significantly associated with approximately $1,140 additional dollars of income (holding all else constant).
- The coefficient on `i_female` tells us that women in the sample, on average, make $2,763 less than the men in the sample (holding education constant).

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:

```r
# Regression for BP test
est10 = lm(residuals(est09)^2 ~ education + i_female, 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)   -181.      33.2      -5.46 5.01e- 8
#> 2 education       18.6      2.38      7.82 6.47e-15
#> 3 i_female       -52.3     10.4      -5.03 5.12e- 7
```

continued on next page...

```r
# BP test statistic
lm10 = summary(est10)$r.squared * nrow(ps_df)
# Test against Chi-squared 2
pchisq(lm10, df = 2, lower.tail = F) %>% round(5)
```

```
#> [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`.
- 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).

Answer:

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

```
#> # A tibble: 5 x 5
#>   term               estimate std.error statistic     p.value
#>   <chr>                 <dbl>     <dbl>     <dbl>       <dbl>
#> 1 (Intercept)          553.     180.         3.07 0.00214
#> 2 education           -103.      26.4       -3.89 0.000100
#> 3 i_female             285.      66.7        4.27 0.0000198
#> 4 I(education^2)         4.84     0.961      5.03 0.000000508
#> 5 education:i_female   -24.3      4.76      -5.11 0.000000340
```

```r
# BP test statistic
lm11 = summary(est11)$r.squared * nrow(ps_df)
# Test against Chi-squared 4
pchisq(lm11, df = 4, 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 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.

Answer:

```
# Arrange the dataset by commute time
ps_df = ps_df %>% arrange(education)
# Create the two subsets (first and last 8,000 observations)
g1 = head(ps_df, 1600)
g2 = tail(ps_df, 1600)
# Run the two regressions
est12_1 = lm(personal_income ~ education + i_female, data = g1)
est12_2 = lm(personal_income ~ education + i_female, data = g2)
# Find the SSE from each regression
sse1 = sum(residuals(est12_1)^2)
sse2 = sum(residuals(est12_2)^2)
# GQ test statistic
gq = sse2 / sse1
# p-value
pf(gq, df1 = 1600, df2 = 1600, lower.tail = F)
```

```
#> [1] 4.523e-151
```

Using the Goldfeld-Quandt test for heteroskedasticity, we again reject the null hypothesis of *homoskedasticity* with a *p*-value of approximately 0.

When we looked at the figure at the beginning of the problem set, it definitely seemed like there was possibly a funnel-like heteroskedasticity. This is the type of heteroskedasticity that the Goldfeld-Quandt test is capable of picking up, so it makes sense that we were able to detect it.

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
est13 = lm_robust(
  personal_income ~ education + i_female,
  data = ps_df,
  se_type = "HC2"
)
# Print results
est13 %>% summary()
```

continued on next page...

```
#>
#> Call:
#> lm_robust(formula = personal_income ~ education + i_female, data = ps_df,
#>     se_type = "HC2")
#>
#> Standard error type:  HC2
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper   DF
#> (Intercept)    -8.39     0.6995   -12.0 1.04e-32    -9.76    -7.02 4997
#> education       1.14     0.0579    19.7 3.77e-83     1.03     1.25 4997
#> i_female       -2.76     0.2077   -13.3 1.01e-39    -3.17    -2.36 4997
#>
#> Multiple R-squared:  0.131 ,    Adjusted R-squared:  0.13
#> F-statistic:  203 on 2 and 4997 DF,  p-value: <2e-16
```

The heteroskedasticity-robust standard errors are slightly slightly larger than the OLS standard errors. The increase is especially "large" for education---increasing by approximately 26%. That said, the statistical significance of the term has not changed meaningfully.

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. 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.

continued on next page...

Answer:

```
# The new model
est16 = lm_robust(
  personal_income ~ education + i_female + education:i_female,
  data = ps_df,
  se_type = "HC2"
)
# The results
summary(est16)
```

```
#>
#> Call:
#> lm_robust(formula = personal_income ~ education + i_female +
#>     education:i_female, data = ps_df, se_type = "HC2")
#>
#> Standard error type:  HC2
#>
#> Coefficients:
#>                    Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper   DF
#> (Intercept)         -12.131     1.2075  -10.05 1.58e-23  -14.498   -9.764 4996
#> education             1.415     0.0965   14.66 1.14e-47    1.226    1.604 4996
#> i_female              5.236     1.4301    3.66 2.54e-04    2.432    8.039 4996
#> education:i_female   -0.578     0.1126   -5.14 2.93e-07   -0.799   -0.358 4996
#>
#> Multiple R-squared:  0.137 ,    Adjusted R-squared:  0.137
#> F-statistic:  170 on 3 and 4996 DF,  p-value: <2e-16
```

In this new model, the interaction between female and education is statistically significant at the 5-percent level with a coefficient of approximately -0.58. This coefficient tests whether the relationship between education and earnings appears to differ for females and non-females (in this sample: non-female means male). In more "economics" terms: We are testing whether the returns to education are different for women (relative to rest of the sample—men). The coefficient tells us that the returns to education for females in the sample make is approximately $5,784.09 **less** than males in the sample (for each additional year of education).

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?

Answer: The expected income for women with 16 years of education is approximately $64,861. The expected income for men with 16 years of education is approximately $105,049.

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.

Answer:

```r
# Get residuals from the model in 16
resid16 = ps_df$personal_income - est16$fitted.values
# Regression for BP test
est18 = lm(
  resid16^2 ~
  education + i_female +
  education:i_female +
  I(education^2) + I(education^2):i_female,
  data = ps_df
)
# Results
est18 %>% tidy()
# BP test statistic
lm18 = summary(est18)$r.squared * nrow(ps_df)
# Test against Chi-squared 5
pchisq(lm18, df = 5, lower.tail = F) %>% round(3)
```

```
#> # A tibble: 6 x 5
#>   term                    estimate std.error statistic     p.value
#>   <chr>                      <dbl>     <dbl>     <dbl>       <dbl>
#> 1 (Intercept)               866.      230.        3.76 0.000171
#> 2 education                -149.       34.1      -4.37 0.0000127
#> 3 i_female                 -475.      362.       -1.31 0.189
#> 4 I(education^2)              6.50      1.25      5.22 0.000000186
#> 5 education:i_female         87.4      53.1       1.65 0.0999
#> 6 i_female:I(education^2)    -4.01      1.92     -2.09 0.0368
```

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

Even with this new interaction (our new specification to try to address misspecification), we still have very strong evidence of heteroskedasticity (*i.e.*, highly statistically significant). Thus, it does not seem like the interaction "helped" resolve the heteroskedasticity---though it does seem like an important part of the model (given its statistical significance and economic meaning).

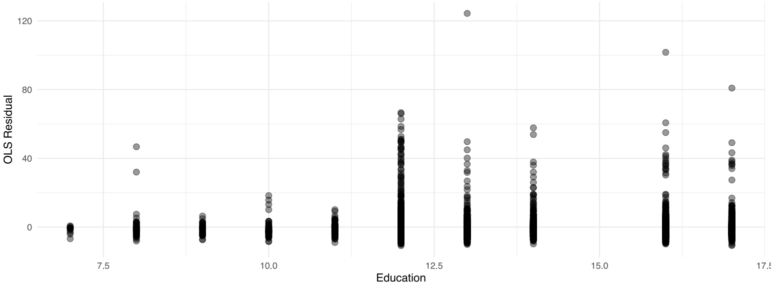
Q19. Based upon your findings from the preceding questions: Do you think heteroskedasticity is present? If so: Does heteroskedasticity appear to matter in this setting?

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

Answer:

```r
# Plotting the residuals from our OLS regression against education
ggplot(
  data = data.frame(
    education = ps_df$education,
    residual = est09$residuals
  ),
  aes(x = education, y = residual)
) +
geom_point(size = 2.5, alpha = 0.4) +
xlab("Education") +
ylab("OLS Residual") +
theme_minimal()
```

continued on next page...



Heteroskedasticity does appear to be present---it appeared likely in our original plot, it was highly significant in our tests, and the figure above seems to suggest that variance (in the residuals) changes with values of education.

This heteroskedasticity appears to be causing us to over-estimate our precision---especially for the relationship between education and personal income. For example, our t statistic drops from 24.8018 to 19.6856 when we use heteroskedasticity-robust standard errors. However, the t statistic of 19.6856 is still highly significant, so adjusting for heteroskedasticity doesn't really change our results/understanding much in this setting.

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.*

Answer: It is very likely that there is omitted variable bias here---there are many variables that affect personal income and that interact with education, sex, or their interaction.

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`).