Problem Set 2: Heteroskedasticity

EC 421: Introduction to Econometrics

Solutions

DUE Upload your answer on Canvas *before* midnight on Sunday, 09 May 2021.

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, skimr, here)
```

Q02. Now load the data (it's the same dataset as the first problem set with one new variable:education). I saved the same dataset again as a two different formats: a `.csv` file or an `.rds` file. Use a function that reads `.csv` files or `rds` files---for example, `read.csv()` / `read.rds()` or `read_csv()` / `read_rds` (from the `readr` package in the `tidyverse`).

Answer:

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

#### ALTERNATIVELY
ps_df = here("ps-002-data.rds") %>% read_rds()
```

Q03. Check your dataset. Apply the function `summary()` to your dataset. You should have `25` variables. You might also want to check out the `skim()`function from the`skimr` package---it's a really useful function.

Answer:

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

continued on next page...

```
#>     state               marrno           age          i_urban        i_citizen    
#>  Length:8000        Min.   :0.000   Min.   :24.0   Min.   :0.000   Min.   :0.000  
#>  Class :character   1st Qu.:1.000   1st Qu.:34.0   1st Qu.:0.000   1st Qu.:1.000  
#>  Mode  :character   Median :1.000   Median :46.0   Median :1.000   Median :1.000  
#>                     Mean   :0.985   Mean   :45.8   Mean   :0.618   Mean   :0.931  
#>                     3rd Qu.:1.000   3rd Qu.:56.0   3rd Qu.:1.000   3rd Qu.:1.000  
#>                     Max.   :3.000   Max.   :92.0   Max.   :1.000   Max.   :1.000  
#>                                                                                   
#>   i_noenglish     i_only_english  i_drive_to_work    i_asian          i_black      
#>  Min.   :0.0000   Min.   :0.000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000  
#>  1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:0.0000  
#>  Median :0.0000   Median :1.000   Median :1.000   Median :0.0000   Median :0.0000  
#>  Mean   :0.0345   Mean   :0.801   Mean   :0.919   Mean   :0.0621   Mean   :0.0865  
#>  3rd Qu.:0.0000   3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:0.0000   3rd Qu.:0.0000  
#>  Max.   :1.0000   Max.   :1.000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000  
#>                                                                                    
#>   i_indigenous       i_white         i_female         i_male      i_grad_college 
#>  Min.   :0.0000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000  
#>  1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000  
#>  Median :0.0000   Median :1.000   Median :0.000   Median :1.000   Median :0.000  
#>  Mean   :0.0076   Mean   :0.783   Mean   :0.473   Mean   :0.527   Mean   :0.386  
#>  3rd Qu.:0.0000   3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:1.000  
#>  Max.   :1.0000   Max.   :1.000   Max.   :1.000   Max.   :1.000   Max.   :1.000  
#>                                                                                  
#>  i_grad_highschool   i_married     i_married_mult personal_income i_health_insurance
#>  Min.   :0.000     Min.   :0.000   Min.   :0.00   Min.   : 0.0    Min.   :0.000     
#>  1st Qu.:1.000     1st Qu.:0.000   1st Qu.:0.00   1st Qu.: 2.6    1st Qu.:1.000     
#>  Median :1.000     Median :1.000   Median :0.00   Median : 4.5    Median :1.000     
#>  Mean   :0.924     Mean   :0.609   Mean   :0.18   Mean   : 6.3    Mean   :0.919     
#>  3rd Qu.:1.000     3rd Qu.:1.000   3rd Qu.:0.00   3rd Qu.: 7.3    3rd Qu.:1.000     
#>  Max.   :1.000     Max.   :1.000   Max.   :1.00   Max.   :85.7    Max.   :1.000     
#>                                                                                     
#>    i_internet     time_depart    time_arrive   time_commuting    education   
#>  Min.   :0.000   Min.   :  15   Min.   :  19   Min.   :  2.0   Min.   : 0.0  
#>  1st Qu.:1.000   1st Qu.: 392   1st Qu.: 414   1st Qu.: 14.0   1st Qu.:12.0  
#>  Median :1.000   Median : 452   Median : 474   Median : 22.0   Median :13.0  
#>  Mean   :0.945   Mean   : 488   Mean   : 517   Mean   : 29.1   Mean   :13.8  
#>  3rd Qu.:1.000   3rd Qu.: 512   3rd Qu.: 539   3rd Qu.: 37.0   3rd Qu.:16.0  
#>  Max.   :1.000   Max.   :1425   Max.   :1439   Max.   :202.0   Max.   :17.0  
#>  NA's   :30
```

Table: Data summary

Name	ps_df
Number of rows	8000

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 29.145 and 22, respectively. Because the mean is quite a bit larger than the median it tells us that the right tail of the distribution of commute time 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 0.919, 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 92% of individuals had some type of health insurance.

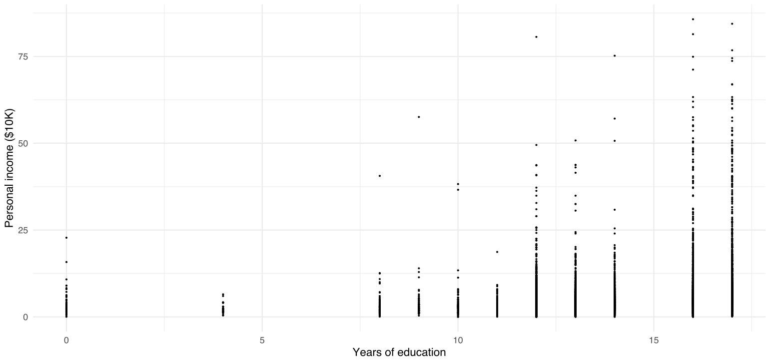
What's the value of an education?

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`) our indicator for *citizenship status* (`i_citizen`) 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:

```
# Regression
est09 = lm(personal_income ~education+ i_female + i_citizen, data = ps_df)
# Results
est09 %>% tidy()
```

```
#> # A tibble: 4 x 5
#>   term        estimate std.error statistic   p.value
#>   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
#> 1 (Intercept)   -3.42     0.451      -7.60 3.42e- 14
#> 2 education      0.784    0.0289     27.1  7.84e-155
#> 3 i_female      -2.89     0.154     -18.8  1.48e- 77
#> 4 i_citizen      0.323    0.305       1.06 2.91e-  1
```

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

- The intercept tells us the expected income (-3.4225) for **an immigrant male** 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 $784 additional dollars of income (holding all else constant).
- The coefficient on `i_female` tells us that women in the sample, on average, make $2,894 less than the men in the sample (holding education and citizen status constant).
- The coefficient on `i_citizen` tells us that citizens in the sample, on average, make $323 more than the non-citizens in the sample (holding education and gender 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:

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

```
#> # A tibble: 4 x 5
#>   term        estimate std.error statistic  p.value
#>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)   -64.9      17.4     -3.72  1.97e- 4
#> 2 education      10.0       1.12     8.94  4.97e-19
#> 3 i_female      -41.2       5.94    -6.94  4.29e-12
#> 4 i_citizen      -7.20     11.8     -0.609 5.42e- 1
```

continued on next page…

```
# 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---so small that the computer reports 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:

```
#slightly modified regression for BP test
mod_est09 = lm(personal_income ~education+ i_female +i_citizen, data = ps_df)
# Regression for BP test
est11 = lm(
  residuals(mod_est09)^2 ~
education+ i_female + i_citizen+
  I(education^2) +
 education:i_female + i_citizen:education + i_citizen:i_female + i_citizen:i_female:education,
  data = ps_df
)
# Results
est11 %>% tidy()
```

```
#> # A tibble: 9 x 5
#>   term                         estimate std.error statistic  p.value
#>   <chr>                           <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)                     93.2      41.5      2.25  2.48e- 2
#> 2 education                      -25.2       4.66    -5.39  7.06e- 8
#> 3 i_female                        38.9      68.1      0.571 5.68e- 1
#> 4 i_citizen                      -60.5      47.7     -1.27  2.05e- 1
#> 5 I(education^2)                   1.57      0.182    8.62  8.33e-18
#> 6 education:i_female              -4.91      5.30    -0.927 3.54e- 1
#> 7 education:i_citizen              5.32      3.75     1.42  1.56e- 1
#> 8 i_female:i_citizen             140.       76.7      1.82  6.80e- 2
#> 9 education:i_female:i_citizen   -11.1       5.85    -1.89  5.88e- 2
```

```
# 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 (reported as 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, so this result is expected.

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 and citizenship status).
- 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 **1,100** and last **1,100** observations (after sorting on education). 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, 1100)
g2 = tail(ps_df, 1100)
# Run the two regressions
est12_1 = lm(personal_income ~education+ i_female + i_citizen, data = g1)
est12_2 = lm(personal_income ~education+ i_female + i_citizen, 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 = 1100, df2 = 1100, lower.tail = F)
```

```
#> [1] 2.595e-146
```

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 + i_citizen,
  data = ps_df,
  se_type = "HC2"
)
# Print results
est13 %>% summary()
```

continued on next page...

```
#>
#> Call:
#> lm_robust(formula = personal_income ~ education + i_female +
#>     i_citizen, data = ps_df, se_type = "HC2")
#>
#> Standard error type:  HC2
#>
#> Coefficients:
#>             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper   DF
#> (Intercept)   -3.423     0.4519   -7.57  4.04e-14   -4.308   -2.537 7996
#> education      0.784     0.0352   22.32 4.48e-107    0.716    0.853 7996
#> i_female      -2.894     0.1563  -18.51  5.74e-75   -3.201   -2.588 7996
#> i_citizen      0.323     0.2749    1.17  2.41e-01   -0.216    0.861 7996
#>
#> Multiple R-squared:  0.114 ,    Adjusted R-squared:  0.114
#> F-statistic:  210 on 3 and 7996 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 21%. 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:

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

```
#>
#> Call:
#> lm_robust(formula = personal_income ~ education + i_female +
#>     education:i_female + i_citizen, data = ps_df, se_type = "HC2")
#>
#> Standard error type:  HC2
#>
#> Coefficients:
#>                    Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper   DF
#> (Intercept)          -4.908     0.6634   -7.40 1.53e-13   -6.208   -3.607 7995
#> education             0.896     0.0522   17.18 5.53e-65    0.794    0.999 7995
#> i_female              0.869     0.8496    1.02 3.06e-01   -0.796    2.534 7995
#> i_citizen             0.296     0.2756    1.08 2.82e-01   -0.244    0.837 7995
#> education:i_female   -0.272     0.0660   -4.13 3.70e-05   -0.402   -0.143 7995
#>
#> Multiple R-squared:  0.117 ,    Adjusted R-squared:  0.116
#> F-statistic:  185 on 4 and 7995 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 -0.27. 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 $2,723.04 **less** than males in the sample (for each additional year ofeducation).

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 $147,731. The expected income for men with 16 years of education is approximately $91,608.

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 + i_citizen +
 education:i_female +i_citizen:education +i_citizen:i_female +
  I(education^2) + I(education^2):i_female +
   i_citizen:i_female:education,
  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 = 9, lower.tail = F) %>% round(3)
```

```
#> # A tibble: 10 x 5
#>    term                        estimate std.error statistic  p.value
#>    <chr>                          <dbl>     <dbl>     <dbl>    <dbl>
#>  1 (Intercept)                   132.      42.0       3.14  1.71e- 3
#>  2 education                     -33.8      5.24     -6.45  1.21e-10
#>  3 i_female                      -73.6     72.8      -1.01  3.12e- 1
#>  4 i_citizen                     -37.5     48.3      -0.775 4.38e- 1
#>  5 I(education^2)                  1.98     0.225     8.80  1.60e-18
#>  6 education:i_female             19.7      9.06      2.18  2.94e- 2
#>  7 education:i_citizen             3.62     3.79      0.955 3.40e- 1
#>  8 i_female:i_citizen             66.1     81.5       0.810 4.18e- 1
#>  9 i_female:I(education^2)        -1.15     0.379    -3.03  2.44e- 3
#> 10 education:i_female:i_citizen   -5.79     6.17     -0.938 3.49e- 1
```

```
#> [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).

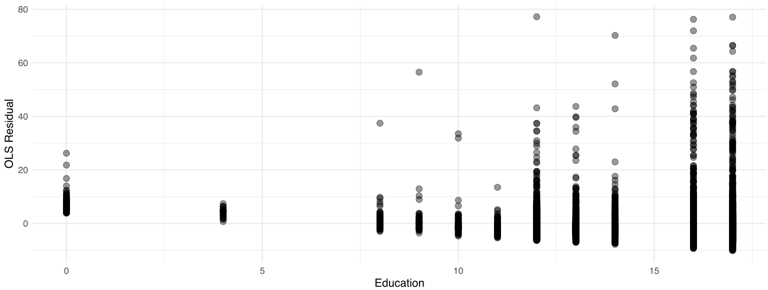
Q19. Based upon your findings from the preceding questions: Do you think heteroskedasticity is present? If so: Does heteroskedasticity appear to matter in this setting?

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

Answer:

```
# 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 27.102 to 22.3169 when we use heteroskedasticity-robust standard errors. However, the t statistic of 22.3169 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.

[Continues...]

Estimate WLS

Q21. Often, we as researchers have no idea the form of heteroskedasticity, but we'd really like to run WLS - our answer then is a procedure known as *feasible generalized least squares* or **FGLS**. Let's walk through how to do this.

Our first step is to set up an estimating equation. Let's regress `personal_income` (y) on `i_citizen`, `education`, `marrno` *(number of marriages)*, `i_female`, and `i_female` interacted with `education`. What is the significance of number of marriages here? How should we explain the causal effect of this variable (ie, does having more marriages cause an individual to get more money)? *Hint: Think about what variables are NOT in the model*

Answer:

```
#>
#> Call:
#> lm(formula = personal_income ~ education + i_citizen + marrno +
#>     i_female + i_female:education, data = ps_df)
#>
#> Residuals:
#>    Min     1Q Median     3Q    Max
#> -10.58  -3.14  -1.30   1.05  77.44
#>
#> Coefficients:
#>                    Estimate Std. Error t value Pr(>|t|)
#> (Intercept)         -5.9023     0.5543  -10.65  < 2e-16 ***
#> education            0.9052     0.0372   24.35  < 2e-16 ***
#> i_citizen            0.0965     0.3037    0.32     0.75
#> marrno               1.0834     0.1048   10.33  < 2e-16 ***
#> i_female             0.5185     0.8088    0.64     0.52
#> education:i_female  -0.2486     0.0575   -4.33  1.5e-05 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 6.77 on 7994 degrees of freedom
#> Multiple R-squared:  0.128,    Adjusted R-squared:  0.128
#> F-statistic:  236 on 5 and 7994 DF,  p-value: <2e-16
```

`Marrno` is statistically significant, with a p-value equal to 7.0371\times 10^{-25}. `Marrno` makes more sense here likely as a proxy for both age and also jobs with long working hours rather than as a direct causal factor of income growth. A person with more marriages (and therefore more divorces) is likely to be either older (so has had more opportunities for promotion), have less time at home, or both.

Q22. Our next step is to estimate $h(x_i)$. In our version of FGLS - we assume that $\sigma_i^2 = \sigma^2 h(x_i)$, but unlike our general approach to WLS, we will ALSO assume $h(x)$ is equal to $e^{\delta_0+\delta_1 x_1 \dots \delta_k x_k}$ - we can find the values for δ_0 and δ_1 by regressing our variables x on the log of the squared residuals †. In our case, we are assuming $h(x_i) = e^{\delta_0+\delta_1 education_i+\dots+\delta_5 education_i * female_i}$. We need to transform our equation to find something OLS can estimate - a sensible option is to use the log-linear specification - $log(h(x_i)) = log(e^{\delta_0+\delta_1 education_i+\dots+\delta_5 education_i * female_i}) = \delta_0 + \dots$

We can estimate the coefficients δ_0 through δ_5 by running a regression of the independent variables, $X_1 \dots X_5$ from the regression in **Q21** on the *logged and squared* `residuals` of our estimates from **Q21** (ie - `log(residuals^2)` in R.)

Then we need to create a weights variable equal to $h(x)$ for our data by extracting the fitted values of the regression above and **exponentiating them**, ie, $e^{fitted.values}$. You can do this in R once you have produced your fitted values by running the command - `weight = exp(my_fitted_values)`.

[*Hint:* You can access the fitted values of an `lm` object by using `not_a_real_lm$fitted.values`]

Q23. Now, all that is left is to estimate *WLS*. We can do this by taking the weights we calculated in **Q22** and include them in a new regression like so - `lm( ... ,weights = 1/weight)`. Use the same regression parameters you used for **Q21**. Have R report the results for you, and include your findings. Comment on the significance of `i_female` and `marrno`.

Answer

```
#> # A tibble: 6 x 5
#>   term                estimate std.error statistic  p.value
#>   <chr>                  <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)          -0.149     0.402     -0.371 7.11e- 1
#> 2 education             0.453     0.0282    16.1   3.48e-57
#> 3 i_citizen             0.242     0.253      0.956 3.39e- 1
#> 4 marrno                0.849     0.0903     9.40  6.96e-21
#> 5 i_female             -1.17      0.463     -2.52  1.18e- 2
#> 6 education:i_female   -0.0803    0.0353    -2.27  2.31e- 2
```

```
#> # A tibble: 6 x 5
#>   term              estimate std.error statistic  p.value
#>   <chr>                <dbl>     <dbl>     <dbl>    <dbl>
#> 1 intercept_wls      -0.149     0.402     -0.371 7.11e- 1
#> 2 education_wls       0.453     0.0282    16.1   3.48e-57
#> 3 i_citizen_wls       0.242     0.253      0.956 3.39e- 1
#> 4 marrno_wls          0.849     0.0903     9.40  6.96e-21
#> 5 i_female_wls       -1.17      0.463     -2.52  1.18e- 2
#> 6 i_female_educ_wls  -0.0803    0.0353    -2.27  2.31e- 2
```

It appears as if `i_female` is now significant at the 5% level, with a p-value equal to 0.0118. This means we must adjust our conclusion a bit - there are likely decreased returns to education for women (at least in our sample) but low-education women appear to out-earn their male counterparts.

Even after adjusting the weights of our observations, `marrno`, ie, the number of marriages a person has still seems significant with a p-value equal to 6.9581\times 10^{-21}. It is likely we can include age in our regression and reduce the significance of this effect, but heteroskedasticity appears to have no impact on the significance of the effect of an additional marriage on personal income.

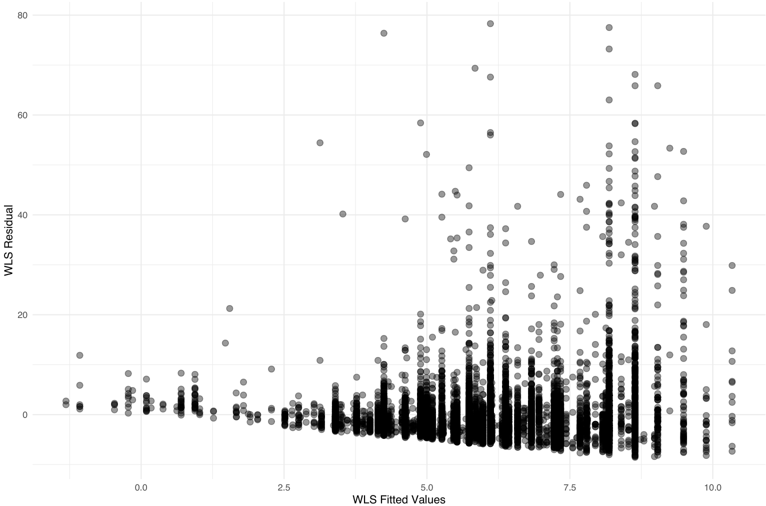
† There are a LOT of ways to estimate heteroskedasticity using fgls - this is only one of them

Q24. Explain why a critical econometrician might not trust these results. Plot your new fitted values against your new residuals. Do you think they're correct to not trust these results?

Lastly - FGLS as an estimator is not *unbiased*, but it is a *consistent* estimator and *asymptotically* more efficient than OLS for β. Explain what these three statements mean using your own words.

(*Hint: what do we need for WLS to work properly?*)

```
ggplot(
  data = data.frame(
  fitted = wls_est$fitted.values,
    residual = wls_est$residuals
  ),
  aes(x =fitted, y = residual)
) +
geom_point(size = 2.5, alpha = 0.4) +
xlab("WLS Fitted Values") +
ylab("WLS Residual") +
theme_minimal()
```



A critical econometrician might say that - if the functional form of $\sigma^2 h(x_i)$ is NOT well-approximated by $e^{\delta_0+\delta_1 \hat{y}_{fgls}}$ then we have misspecified our functional form for WLS, and we still will suffer from heteroskedasticity. When we do this, we can induce bias and inefficiency in our estimates. A cursory examination of our WLS residuals plotted against the fitted values appears to imply we did NOT fix our heteroskedasticity.

Not required for full credit: FGLS is biased for our standard errors because we estimated σ_i^2, but FGLS is efficient *asymptotically*- meaning as our sample size goes to infinity, it will perform "better" than OLS. Specifically - given an infinite sample, we expect coefficients from *fGLS* to have lower standard errors.

FGLS being **biased** means that $E(\hat{\beta}_{fgls}) \neq \beta$. FGLS being **consistent** means that $\lim_{n\to\infty} P(|\hat{\beta}_{fgls} - \beta| > \varepsilon) = 0$ for any $\varepsilon > 0$. IE - as our sample size approaches infinity, the probability our FLGS estimator differs from the true population value by more than some small number ε is zero.

Description of variables and names

Variable	Description
state	State abbreviation
marrno	number of marriages individual has had
age	The individual's age (in years)
i_urban	Binary indicator for whether home county is 'urban'
i_citizen	Binary indicator for whether the individual is a citizen (naturalized or born.)
i_noenglish	Binary indicator for whether the individual speaks English
i_only_english	Binary indicator for whether the individual speaks ONLY English
i_drive_to_work	Binary indicator for whether the individual drives to work or takes a personal car
i_asian	Binary indicator for whether the individual identified as Asian
i_black	Binary indicator for whether the individual identified as Black
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
i_grad_college	Binary indicator for whether the individual graduated college
i_grad_highschool	Binary indicator for whether the individual graduated high school
i_married	Binary indicator for whether the individual was married at the time of the sample
i_married_mult	Binary indicator for whether the individual has been married multiple times at the time of the sample
personal_income	Total (annual) personal income (tens of thousands of dollars)
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)
education	Number of years in education