Problem Set 2 *Solutions*

Unbiasedness, Consistency, and Heteroskedasticity

EC 421: Introduction to Econometrics

Due *before* midnight (11:59pm) on **Friday, 03 May 2019**

Problem 1: Unbiasedness and consistency

Throughout this course, we will use the OLS estimator $\hat{\beta}$ to estimate β. We will continue to discuss situations in which the estimator (or other estimators) are (1) unbiased or (2) consistent.

1a. What is the formal (mathematical) definition of **bias**?

Answer Formally, $\text{Bias}_\beta\left(\hat{\beta}\right) = E\left[\hat{\beta}\right] - \beta$

1b. Give a more intuitive definition of **bias** (no expected values).

Answer Bias tells us whether, on average, our estimator gets the answer right (whether it hits its mark, on average).

1c. Why do we care if the OLS estimator (or any estimator) is **biased**?

Answer If our estimator is biased, then it will regularly estimate the wrong number, which can make it harder for us to learn about the unknown parameter that we are trying to estimate.

1d. What does it mean for an estimator to be **consistent?**

Multiple potential answers

Answer$_1$ (more formal) If our estimator is consistent, (1) the estimator has a probability limit and (2) the probability limit is the parameter that the estimator is trying to estimate.

Answer$_2$ (more intuitive) If our estimator is consistent, then as the sample size approaches infinity ($n \to \infty$) , the estimator's distribution collapses to a point located at the parameter the estimator is trying to estimate.

1e. True/False Unbiasedness is a property for finite-sized samples, while consistency refers to an esimator as sample sizes approach infinity.

Answer True.

1f. Which of the following two estimators would you choose? Explain your reasoning.

Estimator **A** is unbiased and inconsistent.
Estimator **B** is biased and consistent.

Answer There are many possible answers here.

All else equal, we likely prefer unbiased estimators to consistent estimators if we have a fairly small sample (since consistency refers to *large* samples). However, if the bias is fairly small and/or our sample size is very large, we might opt of the biased and consistent estimator.

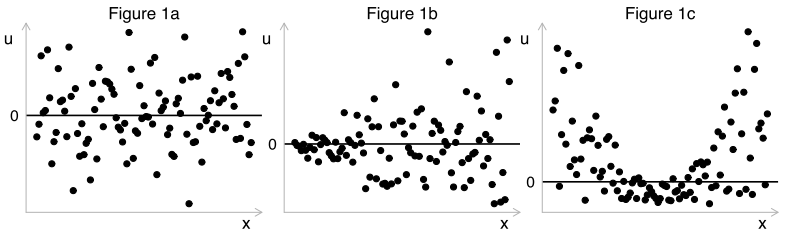
Problem 2: Heteroskedasticity

Now we turn to Heteroskedasticity.

2a. In which of the subfigures in **Figure 1** (below) is u_i likely heteroskedastic? Briefly explain your answer. (***Hint*** There may be more than one.)

Answer u_i is likely heteroskedastic in subfigures **1b** and **1c**. We can see clear trends (relationships) between the variance of u_i (its dispersion) and x_i.

Figure 1



2b. In the presence of heteroskedasticity, is OLS still unbiased?

Answer Yes.

2c. What issues does heteroskedasticity cause for our standard OLS setting?

Answer Heteroskedasticity makes (1) OLS inefficient and (2) biases our estimated standard errors.

2d. Which ways can we "fix" (or "live with") heteroskedasticity?

Answer We discussed three strategies for living with heteroskedasticity:

1. Check that misspecification has not caused the heteroskedasticity.
2. Use the WLS (weighted least squares) estimator.
3. Use heteroskedasticity-robust standard errors.

2e. Imagine that we want to use OLS to estimate the model

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{1}$$

where x_i is a categorical variable that takes the values 1, 2, or 3.

Suppose that we know $\text{Var}(u_i|x_i = 1) = 15$ and $\text{Var}(u_i|x_i = 2) = 15$. We do not know $\text{Var}(u_i|x_i = 3)$, *i.e.*, $\text{Var}(u_i|x_i = 3) = \sigma_3^2$ for some unknown parameter σ_3^2.

What value must σ_3^2 take for our model to be homoskedastic?

Answer For the model to be homoskedastic, it must be the case that $\sigma_3^2 = 15$.

2f. *Goldfeld-Quandt* In order to test whether the data we will use to estimate equation (1) are homoskedastic/heteroskedastic, we will run a Goldfeld-Quandt test.

We estimate (1) for the upper one third of the dataset (sorted on x) and find SSE_3=100. We estimate (1) on the middle third and find SSE_2=80. Finally, we estimate (1) on the lower third and find SSE_1=70. Each of these three groups has 100 observations.

Conduct a Goldfeld-Quandt test. State your hypotheses, calculate the G-Q test statistic, determine the *p*-value, state your conclusion.

Hint: The function `pf(q, df1, df2, lower.tail = F)` calculates the probability of observing a value of `q` or greater in an F distribution with `df1, df2` numerator and denominator degrees of freedom.

Answer The hypotheses for our test are

H_0: $\sigma_1^2 = \sigma_3^2$ (homoskedasticity) *vs.* H_a: $\sigma_1^2 \neq \sigma_3^2$ (heteroskedasticity)

For the Goldfeld-Quandt test, we test this null hypthesis using the test statistic

$$F = \frac{SSE_3}{SSE_1} = \frac{100}{70} \approx 1.4286$$

Under the null hypothesis, this test statistic has an F distribution with 98 (=100-2) degrees of freedom in the numerator and denominator. Using R we can calculate the *p*-value:

```
# p-value
pf(100/70, df1 = 100-2, df2 = 100-2, lower.tail = F)
```

```
#> [1] 0.03951597
```

This *p*-value is less than 0.05, so we reject the null hypothesis and conclude that that there is statistically significant evidence of heteroskedasticity (at the 5-percent level).

Problem 3: Data and heteroskedasticity

3a. Open up Rstudio, an `R` script, load whichever packages you want, and load the dataset contained in `ps02_data.csv`.

Answer

```
# Load 'pacman'
library(pacman)
# Load additional packages
p_load(tidyverse, broom, magrittr, ggplot2, ggthemes)
# Load the data
ps2_df ← read_csv("ps02_data.csv")
# Check data
ps2_df %>% head()
```

```
#> # A tibble: 6 x 6
#>   prob_q5_q1 i_urban share_black share_middlecla… share_divorced
#>        <dbl>   <int>       <dbl>            <dbl>          <dbl>
#> 1     0.0621       1      0.0208            0.548          0.110
#> 2     0.0537       1      0.0198            0.538          0.116
#> 3     0.0731       0      0.0146            0.467          0.113
#> 4     0.0563       1      0.0564            0.504          0.114
#> 5     0.0446       1      0.174             0.500          0.0924
#> 6     0.0519       0      0.224             0.538          0.0956
#> # … with 1 more variable: share_married <dbl>
```

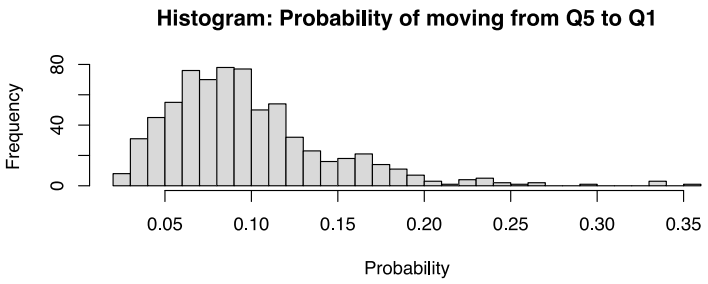
3b. Describe the distribution of our main variable of interest (`prob_q5_q1`). You can provide statistical or graphical descriptions of this variable—try `summary(dataset$variable)` and `hist(dataset$variable)`, among others. What do you see?

Answer The probability that an individual moves from the bottom 20% to the top 20% is fairly low, on average, but there is a decent amount of variation (ranging from almost 0% to 35%).

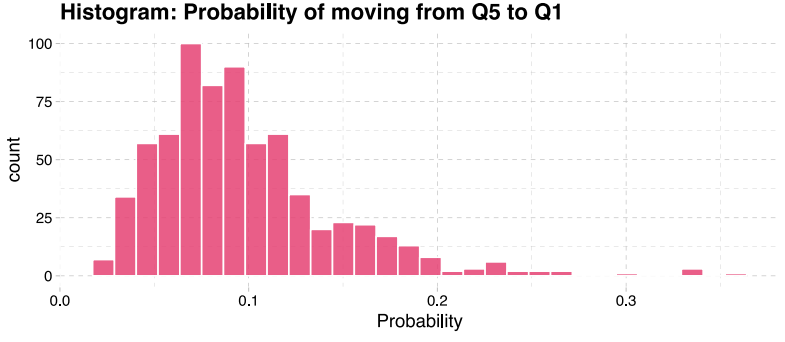
```
# Summarize variable
summary(ps2_df$prob_q5_q1)
```

```
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#> 0.02210 0.06588 0.08889 0.09761 0.11715 0.35714
```

```r
# A histogram using 'hist'
hist(
  ps2_df$prob_q5_q1,
  breaks = 25,
  col = "grey85",
  xlab = "Probability",
  main = "Histogram: Probability of moving from Q5 to Q1"
)
```




```r
# A histogram using 'ggplot'
ggplot(data = ps2_df, aes(x = prob_q5_q1)) +
  geom_histogram(fill = red_pink, color = "white", alpha = 0.85) +
  xlab("Probability") +
  ggtitle("Histogram: Probability of moving from Q5 to Q1") +
  theme_pander()
```



3c. Regress the probability an individual moves from the bottom fifth of income to the top fifth of income (`prob_q5_q1`) on an intercept and the share of the commuting zone that is **married** (`share_married`). Report your findings—the coefficients, brief interpretations of the coefficients, and whether the coefficients are statistically significant.

Answer

```
# Estimate the model
reg_3c ← lm(prob_q5_q1 ~ share_married, data = ps2_df)
# Report the results
reg_3c %>% tidy()
```

```
#> # A tibble: 2 x 5
#>   term          estimate std.error statistic  p.value
#>   <chr>            <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)     -0.198    0.0196     -10.1 1.34e-22
#> 2 share_married    0.517    0.0341      15.2 3.88e-45
```

We estimate that the coefficient on share married is approximately 0.517. This coefficient says that if the share married in a commuting zone increased by 1 percentage point (*e.g.*, from 23% to 24%), then we would expect the probability of moving from the bottom fifth to the top fifth of income to increase by 0.52%. Our estimate is statistically significant (different from zero) at the 5% level.

3d. Does it make sense to interpret the intercept in this case? Explain.

Answer It does not make sense to interpert the intercept in this setting. The interpretation would be "the average mobility probability for a commuting zone with zero marriage." In our data, the share married population ranges from 37% to 69%—zero percent is not reasonable (also evidenced by the fact that the intercept would suggest a negative probability).

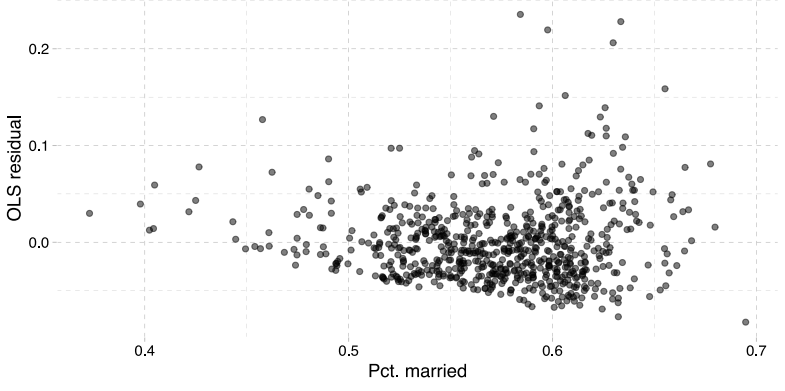
3e. Plot the residuals from your regression in (3c) on the y axis and `share_married` on the x axis. Do you see evidence of heteroskedasticity? Explain.

Hint$_1$: You can grab the residuals from a saved `lm` object by (1) using the `residuals()` function or (2) adding the suffix `$residuals` to the end of the `lm` object, *e.g.*, `my_reg$residuals` grabs the residuals from the `lm` object `my_reg`.

Hint$_2$: `plot(x = dataset$variable1, y = dataset$variable2)` makes quick and simple plots. You can also try `qplot()` from the package `ggplot2`, *i.e.*, `qplot(x = variable1, y = variable2, data = dataset)`.

Answer Based upon the funnel-like figure below, heteroskedasticity seems likely.

```
# Add residuals to the dataset
ps2_df %<>% mutate(e_3c = residuals(reg_3c))
# Plot with ggplot
ggplot(data = ps2_df, aes(x = share_married, y = e_3c)) +
  geom_point(alpha = 0.5) +
  labs(
    x = "Pct. married", y = "OLS residual",
    main = "Visual inspection for heteroskedasticity in 2c."
  ) +
  theme_pander()
```



3f. Conduct a Breusch-Pagan test for heteroskedasticity in the regression model in (2c). Report your hypotheses, the test statistic, the *p*-value, and your conclusion.

Answer

```
# B-P regression
reg_3f ← lm(e_3c^2 ~ share_married, data = ps2_df)
# B-P test statistic
lm_3f ← summary(reg_3f)$r.squared * 709
# B-P p-value
pchisq(q = lm_3f, df = 1, lower.tail = F)
```

```
#> [1] 7.149603e-05
```

Hypotheses Our Breusch-Pagan test here tests the hypotheses H_0 $\alpha_1 = 0$ *vs.* H_a $\alpha_1 \neq 0$ for $e_i^2 = \alpha_0 + \alpha_1 x_i + v_i$ (where we are using e_i^2 to estimate u_i^2, which gives us an estimate for σ_i^2.) If we reject H_0, then we have evidence of heteroskedasticity.

Test statistic We calculate a B-P test statistic of approximately 15.77.

***p*-value** Under the distribution of a χ_1^2, the implied *p*-value for our LM statistic (the probability of seeing this test statistic or greater) is approximately 0.000071.

Conclusions Because our *p*-value is less than our standard significance of 0.05, we reject the null hypothesis ($\alpha_1 = 0$)—there is statistically significant evidence at the 5% level that $\alpha_1 \neq 0$, meaning there is statistically significant evidence of a relationship between e_i^2 and `share_married` (the commuting zone's share of married residents). Therefore, we have statistically significant evidence of heteroskedasticity.

3g. Conduct a White test for heteroskedasticity in the regression model in (2c). Report your hypotheses, the test statistic, the *p*-value, and your conclusion.

Hint: To square the variable `x` in `lm()`, we write `lm(y ~ x + I(x^2), data = dataset)`.

Answer

```
# White regression
reg_3g ← lm(e_3c^2 ~ share_married + I(share_married^2), data = ps2_df)
# White test statistic
lm_3g ← summary(reg_3g)$r.squared * 709
# White p-value
pchisq(q = lm_3g, df = 2, lower.tail = F)
```

```
#> [1] 5.535562e-06
```

Hypotheses Our White test in this question tests the hypotheses H_0 $\alpha_1 = \alpha_2 = 0$ *vs.* H_a $\alpha_1 \neq 0$ *or* $\alpha_2 \neq 0$, where $e_i^2 = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + v_i$ (where, again, we are using e_i^2 to estimate u_i^2, which gives us an estimate for σ_i^2.) If we reject H_0, then we have evidence of heteroskedasticity.

Test statistic We calculate a White test statistic of approximately 24.21.

***p*-value** Under the distribution of a χ_2^2, the implied *p*-value for our LM statistic (the probability of seeing this test statistic or greater) is approximately 0.0000055.

Conclusions Because our *p*-value is less than our standard significance of 0.05, we reject the null hypothesis $(\alpha_j = 0)$—there is statistically significant evidence at the 5% level that either $\alpha_1 \neq 0$ or $\alpha_2 \neq 0$. Therefore we find statistically significant evidence of a relationship between e_i^2 with `share_married` and `share_married`^{2} (the commuting zone's share of married residents). We have statistically significant evidence of heteroskedasticity.

3h. Let's imagine that we think heteroskedasticity is present. Estimate heteroskedasticity-robust standard errors. Do your standard errors change? What about the coefficients? Why is this the case?

Hint: To do this, use the `felm()` function in the `lfe` package. `felm()` takes a regression formula just like `lm()`. Then use `summary(., robust = T)` to show the heteroskedasticity-robust standard errors.

Example:

```
# The regression
some_reg ← felm(y ~ x, data = fake_data)
# Print the coefficients w/ het-robust standard errors
summary(some_reg, robust = T)
```

Answer

```
# Load the 'lfe' package
p_load(lfe)
# Same regression as in (3c)—but with 'felm'
reg_3h ← felm(prob_q5_q1 ~ share_married, data = ps2_df)
# Print the coefficients w/ and w/out het-robust standard errors
reg_3h %>% summary(robust = T)
reg_3h %>% summary(robust = F)
```

```
#> Coefficients:
#>                Estimate Robust s.e t value Pr(>|t|)
#>  (Intercept)   -0.19845    0.02070  -9.586   <2e-16 ***
#>  share_married  0.51708    0.03703  13.963   <2e-16 ***

#> Coefficients:
#>                Estimate Std. Error t value Pr(>|t|)
#>  (Intercept)   -0.19845    0.01960  -10.13   <2e-16 ***
#>  share_married  0.51708    0.03412   15.16   <2e-16 ***
```

The estimated coefficients are the same across the two sets of estimates (with and without heteroskedasticity-robust standard errors), because they both use OLS to estimate the coefficients.

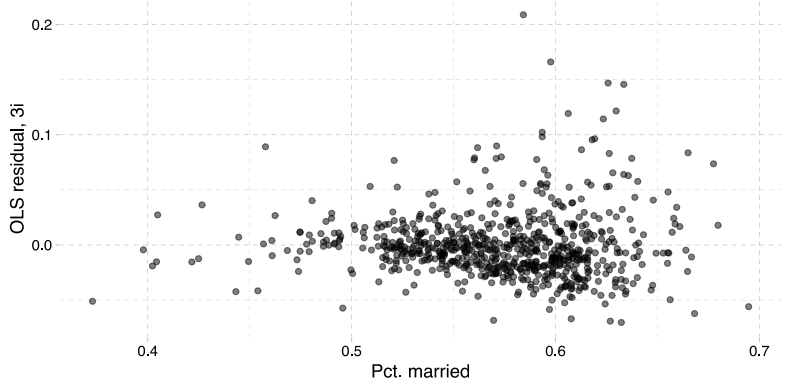
The standard errors change because they use different estimators for the standard errors—a heteroskedasticity-robust estimator and an estimator that assumes homoskedasticity. The heteroskedasticity-robust standard errors are slightly larger.

3i. As we discussed in class, we can introduce heteroskedasticity by mis-specifying our regression model. Try adding the additional variables from this dataset into the regression (possibly also adding interactions, squared explanatory variables, or transformed variables). Then plot the new residuals against share married (`share_married`). *Briefly* describe which regressions you ran and whether it affected the appearance of heteroskedasticity. Which of your specifications appears to do the best?

Note: You do not need to formally test for heteroskedasticity.

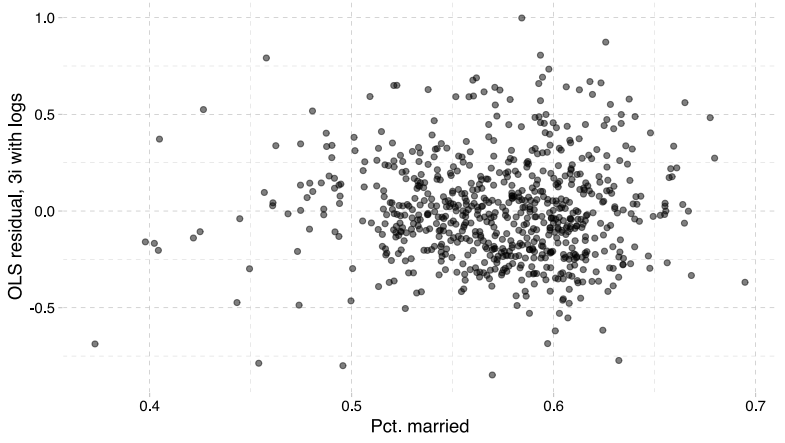
Answer If we stick with the outcome variable as a level (not logged), then heteroskedasticity appears likely, even if we include all of the variables in the dataset, their squares, and the two-way interactions.

```
# Regression with all variables, quadratics, and interactions
reg_3i ← lm(
  prob_q5_q1 ~
  i_urban +
  share_black + I(share_black^2) +
  share_middleclass + I(share_middleclass^2) +
  share_divorced + I(share_divorced^2) +
  share_married + I(share_married^2) +
  share_black:share_middleclass + share_black:share_divorced + share_black:share_married +
  share_middleclass:share_divorced + share_middleclass:share_married +
  share_divorced:share_married,
  data = ps2_df
)
# Add residuals to dataset
ps2_df$e_3i ← residuals(reg_3i)
# Plot residuals against share_married
# Plot with ggplot
ggplot(data = ps2_df, aes(x = share_married, y = e_3i)) +
  geom_point(alpha = 0.5) +
  labs(
    x = "Pct. married", y = "OLS residual, 3i",
    main = "Visual inspection for heteroskedasticity in 3i."
  ) +
  theme_pander()
```



However, if we take the log of our previous outcome variable, things start to look much more homoskedastic.

```
# Regression with all variables, quadratics, and interactions
reg_3i_log ← lm(
  log(prob_q5_q1) ~
  i_urban +
  share_black + I(share_black^2) +
  share_middleclass + I(share_middleclass^2) +
  share_divorced + I(share_divorced^2) +
  share_married + I(share_married^2) +
  share_black:share_middleclass + share_black:share_divorced + share_black:share_married +
  share_middleclass:share_divorced + share_middleclass:share_married +
  share_divorced:share_married,
  data = ps2_df
)
# Add residuals to dataset
ps2_df$e_3i_log ← residuals(reg_3i_log)
# Plot residuals against share_married
# Plot with ggplot
ggplot(data = ps2_df, aes(x = share_married, y = e_3i_log)) +
  geom_point(alpha = 0.5) +
  labs(
    x = "Pct. married", y = "OLS residual, 3i with logs",
    main = "Visual inspection for heteroskedasticity in 3i."
  ) +
  theme_pander()
```



3j. Should we interpret the regression results in (3c)—or your preferred specification in (3i)—as *causal*? Explain your answer. If we cannot interpret the regression as causal, can we still learn something interesting here? Explain.

Answer We probably should not apply a causal interpretation to our estimated coefficients in (3c). There are likely many omitted variables that are (1) correlated with *share married* and (2) affect the probability an individual moves from the first fifth to the upper fifth of the income distribution. One example may be school quality within the commuting zone.

Another potential example is the share of the commuting zone that is 'middle class'. For example, the correlation between share married and share middleclass is 0.53. If share middle class also affects our outcome variable (the probability an individual growing up in the lowest fifth of the income distribution moves into the top fifth), then our estimate on share married will suffer from omitted-variable bias. Specifically, if we think share middleclass positively affects our outcome variable, then our coefficient should be an overestimate of the true effect of *share married*. Let's try including *share middleclass*.

```
# The results with only share_middle
reg_3c %>% tidy()
```

```
#> # A tibble: 2 x 5
#>   term          estimate std.error statistic  p.value
#>   <chr>            <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)     -0.198    0.0196     -10.1 1.34e-22
#> 2 share_married    0.517    0.0341      15.2 3.88e-45
```

```
# The results from adding in share_middleclass
lm(prob_q5_q1 ~ share_married + share_middleclass, data = ps2_df) %>% tidy()
```

```
#> # A tibble: 3 x 5
#>   term              estimate std.error statistic  p.value
#>   <chr>                <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)         -0.208    0.0177    -11.7  3.93e-29
#> 2 share_married        0.274    0.0362      7.56 1.27e-13
#> 3 share_middleclass    0.270    0.0212     12.7  1.45e-33
```

Just as we predicted: Including *share middleclass* **decreases** the estimated 'effect' of *share married*.

We might guess that *share black* would also (1) correlate with *share married* and (2) affect our outcome variable. Because the correlation between *share married* and *share black* is negative (correlation of -0.5), and because *share black* may have a downward effect on the probability an individual moves from the lowest to the highest fifth of the income distribution, we would again expect the estimated effect of *share middleclass* to overstate the actual effect due to omitted variable bias. Let's see.

```
#> # A tibble: 4 x 5
#>   term              estimate std.error statistic  p.value
#>   <chr>                <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)        -0.137     0.0219     -6.25 6.95e-10
#> 2 share_married       0.225     0.0368      6.11 1.61e- 9
#> 3 share_middleclass   0.204     0.0242      8.45 1.72e-16
#> 4 share_black        -0.0796    0.0151     -5.28 1.74e- 7
```

Again, we see that the estimated coefficient on share middleclass drops.

Data description

Each row in the dataset gives records statistics for one of 709 commuting zones.

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 share_ give the share.

Variable	Description
prob_q5_q1	The probability someone born in the lowest 20% of income moves to the highest 20% of income.
i_urban	Binary variable (0,1) for whether the commuting zone is considered urban.
share_black	The share of the zone's population who identify as black.
share_middleclass	The share of the zone's population who are middleclass.
share_divorced	The share of the zone's population who are divorced.
share_married	The share of the zone's population who are married.