Problem Set 2

Unbiasedness, Consistency, and Heteroskedasticity

EC 421: Introduction to Econometrics

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

DUE Your solutions to this problem set are due *before* 11:59pm on Friday, 03 May 2019 on Canvas. **Your answers must include two files** (**1**) your responses/answers to the question (*e.g.*, a Word document) and (**2**) the R script you used to generate any answers in R. Each student must turn in her/his own answers.

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

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**?

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

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

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

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

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

Estimator **A** is unbiased and inconsistent.

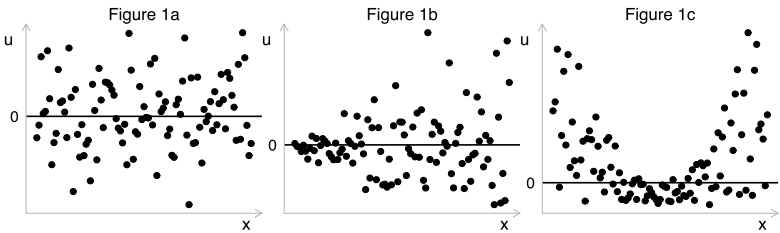
Estimator **B** is biased and consistent.

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.)

Figure 1



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

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

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

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

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

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.

Problem 3: Data and heteroskedasticity

We're now going use an actual dataset to think examine the tests and 'solutions' for heteroskedasticity.

The data come from a very influential paper "Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States" by Chetty, Hendren, Kline, and Saez—published in *The Quarterly Journal of Economics* (QJE) in 2014. Our outcome variable will be *the probability that an individual born to parents in the bottom fifth of the income distribution makes it into the top fifth of the income distribution*. This measure differs from the main outcome in the paper, but it is also very interesting—and it helps simplify our problem set. An individual observation in this dataset represents a commuting zone (sort of like cities) in the United States.

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

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?

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.

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

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)`.

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.

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)`.

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)
```

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 middleclass (`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.

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.

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.