Homework 2

STUDENT NAME

Due 7/5 11:59pm

NOTE: If you would like to answer the written portions using RMarkdown, simply type your answers below the questions. If you would like to print out the pdf to do the written portions, you should add `\vspace{50mm}` to add space in the knitted pdf for written work (adjust the 50mm as necessary). Alternatively, you can do the written portions on a separate paper.

Part 1: Theory

1

(Wooldridge 3.7) Which of the following can cause OLS estimators to be biased? Explain why or why not.

a) Heteroskedasticity
b) Omitting a relevant variable
c) Two of the independent variables included in the model are highly correlated, e.g. $\text{Corr}(x_1, x_2) = 0.96$

2

Compute the variance of $\hat{\beta_0}$ in a simple linear regression as a function of the population parameter σ^2. Recall that, in doing this, you can treat the x_i terms as fixed.

Properties that may be useful:

1. For random variables A and B, $\text{Var}(A - B) = \text{Var}(A) + \text{Var}(B) - \text{Cov}(A, B))$
2. For constants c and d, $\text{Cov}(cA, dB) = cd\text{Cov}(A, B)$
3. For constants a, b, c, and d and random variables W, X, Y, and Z, $\text{Cov}(aW + bX, cY + dZ) = ac\text{Cov}(W, Y) + ad\text{Cov}(W, Z) + bc\text{Cov}(X, Y) + bd\text{Cov}(X, Z)$

Part 2: Application

This exercise is inspired by and uses data from *Does Hazardous Waste Matter? Evidence from the Housing Market and the Superfund Program* by Michael Greenstone and Justin Gallagher (2008, *Quarterly Journal of Economics*). The paper implements the hedonic model, which is a model that links the prices of a good (e.g. a house) to consumer's willingness-to-pay for individual attributes (e.g. number of rooms, but also the neighborhood environmental quality). The econometric goal of the paper is to measure the willingness to pay for environmental cleanup of hazardous waste sites (Superfund sites). They do this by comparing changes in the house prices of nearby houses before and after a cleanup to control neighborhoods with no hazardous waste cleanup. Some hazardous waste sites were placed on a "National Priorities List," which meant they were legally obligated to undergo remediation.

The data you will be using is a sample of 450 census tracts located within 2 miles of a Superfund site from the year 2000. The variables are:

- fips: Federal Information Processing Standards (FIPS), a census tract ID
- npl: Indicator for whether this tract was on the National Priorities List
- lnmdvalhs: log(median house value) for the census tract
- owner_occupied: percentage of housing that is owner occupied (i.e. not rented)
- pop_den: population density
- ba_or_better: percentage of the population with at least a bachelor's degree
- unemprt: unemployment rate
- povrat: poverty rate
- bedrms1-bedrms5: percentage of housing with the stated number of bedrooms
- bedrms_3orless: percentage of housing with 3 or less bedrooms
- blt0_10yrs: percentage of housing less than 10 years old

Preliminaries

Load the necessary packages and data. Note that because the data file is a `.dta` file, you will need to use the package `haven` to read it.

1

In the analysis, we want to compare census tracts that were on the National Priorities List to those that were not. (Every census tract in this sample has a hazardous waste site, but only those on the National Priorities List were required to clean up.)

1.1

Make a histogram that shows the distribution of `lnmdvalhs` separately for each group. It does not need to be a work of art, but make it presentable (give it a title, make sure the axis labels are meaningful, etc).

1.2

Compute means for the control variables (i.e. all variables in the dataset except `fips`, `npl`, and `lnmdvalhs`) within both groups and report them. Are they notably different? (You do not need to formally check the statistical significance of the differences.)

For the most part they seem quite similar, though tracts on the NPL list seem to have lower population density.

2

2.1

Regress log median home value on whether or not the tract was on the NPL, population density, the poverty rate, and number of bedrooms variables 1-4. (Technically we *could* include `bedrms5` as well and not have perfect multicollinearity because an extremely small number of houses have > 5 bedrooms, but we will pretend these don't exist and use > 4 bedroom houses as the reference group.) Report your regression results.

2.2

Interpret (including sign, significance, and size) the parameter estimate on `npl`.

2.3

Do the signs of the parameter estimates on the bedrooms variables make sense? Why or why not?

2.4

Construct a 95% confidence interval for the parameter estimate on `npl`.

2.5

Suppose a politician that supports the Superfund program claims that cleanup can increase surrounding home prices by 10%. Take this claim as a null hypothesis. Given your parameter estimate and standard error, how do you evaluate this null hypothesis?

2.6

If the coefficients on the `bedrms1`, `2`, and `3` variables are the same, we could replace them with the `bedrms_3orless` variable. Write the null and alternative hypothesis you would use to test for this, then conduct the test.

2.7

In this question, you will compute two variations of the regression in 2.1. In the first variation, replace `pop_den` with its log. In the second variation, keep `pop_den` as is and add `pop_den^2` as an additional regressor. How would you determine which of these regressions provides a “better” fit while penalizing more complicated models? Make that determination.