Homework #3: OLS Regression

POS 3930 (Spring 2021)

Dr. Ahler

In this short homework assignment, you will use R or Stata to investigate an important research question: "What affects the descriptive representation of women at the state level?" (That is, what causes a state to have higher or lower percentages of its legislature composed by women?) You'll use the `states` dataset, which I have attached.

Problem 1: Load the dataset into R or Stata.

- It's in .dta (Stata) format, so you should be able to read it into Stata with the command `use` and into R with the command `read.dta13`. If you go with R, make sure you have installed the package `readstata13`.
- In either case, don't forget to set a working directory (where the data is stored) using `cd` in Stata or `setwd` in R.
- Don't forget quotation marks where necessary.

Congratulations! If you've made it this far, you already have 6/10 on the assignment. Finding and loading data is sometimes half the trouble.

Problem 2: For this exercise, Y_i (women's descriptive representation) is operationalized as a variable called `womleg_2011`. Make a density plot showing the distribution of this variable. In Stata, use the command `kdensity`. In R, use the command `plot(density(`Y_i`))`.

Got the right density plot? You now have 7/10 on the assignment.

Problem 3: Predict Y_i with one other variable, the percentage of the state's workforce belonging to a labor union—that is, run a bivariate regression of `womleg_2011` on `union10`.

- Interpret the output of the model, filling in the blanks and finishing this statement: "Using bivariate OLS regression, we expect the percentage of women in the legislature would change _% when..."
- Interpret the p-value related to $\hat{\beta}_1$ by finishing this statement: "If the null hypothesis were true, we would expect to see beta-hat this big (or larger in absolute value) with a probability of..."

Nailed it? You're at 8/10.

Problem 4: Assess whether the apparent relationship you found in Problem 3 is actually causal or, alternatively, spurious. One obvious confounding variable is the state's political climate—states that are more liberal/Democratic will tend to have more women in the

legislature (because women are more likely to serve as Democrats) and also to be more supportive of unions. The Cook Political Index is a variable that measures how "blue" or "red" a state is. More specifically, it estimates the advantage the Democratic Party should expect to have in a typical election in the state. For example, a state with a CPI of -2.5 would be expected to elect a Republican governor or senator by 2.5 points over the Democratic candidate in a typical election; a state with a CPI of 13 would be expected to elect a Democratic candidate by 13 points.

Include `cook_index` as a control variable in the regression you just ran (the bivariate regression of `womleg_2011` on `union10`).

- Interpret the coefficients (the $\hat{\beta}$s, or "beta-hats"). What do they tell you? For each of the two X_i variables, finish the statement: "Holding all else constant with multiple OLS regression, we would expect the percentage of women in the legislature to change..."
- Interpret the p-values as you did in #3. Based on this analysis, what can we infer about the relationship between these three variables?

Getting the hang of it? If you got that one, you're at 8.5/10.

Problem 5: Finally, include one more variable as a control—the percentage of the state with a college degree (`BA_or_higher`). You should have a regression model with one Y_i and 3 X_i variables.

- Interpret the coefficients (the $\hat{\beta}$s, or "beta-hats"). What do they tell you? For each of the three X_i variables, finish the statement: "Holding all else constant with multiple OLS regression, we would expect the percentage of women in the legislature to change..."
- Interpret the p-values as you did in #3 and #4. Based on this analysis, what can we infer about the relationship between these three variables?

Did you master multivariate OLS? If so, you're at 9/10.

Problem 6: For the final point, calculate the expected value of Y_i (or the "fitted value") for each of the following hypothetical states, based on the analysis in #5...

A. A state with a Cook Index of -5, 9% of its workforce unionized, and 25% of its adult population holding a BA or higher

B. A state with a Cook Index of 2, 15% of its workforce unionized, and 17% of its adult population holding a BA or higher

C. A state with an even Cook Index (0), 0% of its workforce unionized, and 0% of its adult population holding a BA or higher