Presentations Details

Materials Due 11:59p on Monday, December 4, 2023 to the course website.
Presentations will be on Tuesday, December 5 and Thursday, December 7 in class.

Presenting Regression Results

You will have the opportunity to run a regression and present the results using data of your own choosing. Complete the following steps to prepare a 10-minute presentation of your results. To receive credit for this portion of the final, submit the dataset you used, your R code, and the deck of 5 slides to the course website by **11:59pm on Monday, December 4**. *Pro tip: don't underestimate how long data importing and cleaning can take.*

First, find data and clean it:

1. Find a dataset that contains three variables that you would like to work with. You'll use these variables to assess whether some outcome variable Y can be explained in part by independent variables X_1 and X_2; specifically, whether X_1 is related to Y controlling for X_2. The dataset should satisfy the following requirements: (a) The variable that you'll use as the dependent variable is an interval/ ratio variable that can take at least three values (is not binary). (b) The data are cross-sectional (so do not contain the same units measured at different points in time). You are welcome to make use of a subset of the dataset that satisfies these conditions, for instance, if the data are time series cross-sectional, you can take one cross section and use that.
2. Load the data into R and make a dataframe called `mydata` that contains only the three variables you plan to use.
3. Clean `mydata` so that you can run a meaningful regression. For instance, do the data add extra numbers, such as 99s instead of NAs, that will throw things off if you don't remove them? Is R correctly treating the numeric variables as numbers, or do you need to force R to do so? Are any variables stored as strings that you should convert to numerics?
4. Now, your interval/ratio variable is your dependent variable (Y), which leaves two independent variables (X_1, X_2). Pick one of the two independent variables to be a dummy variable. Recode it so that it takes the value of 1 for some value(s) of the variable and 0 otherwise. This variable will be X_2. For instance, if it measured age in years, convert it to be coded as 1 if age is above some cutoff number of years and 0 otherwise. (If the variable is already a dummy variable, then leave it and skip to the next step).

Now, analyze the data in R:

5. Regress Y on X_1 and X_2.
6. Make a well-formatted scatterplot that shows the relationship between X_1 and Y, and that uses 2 different symbols and colors to contrast points for which $X_2 = 1$ and $X_2 = 0$. Include a legend.
7. Add the two regression lines associated with each of the two values of X_2 to the X_1, Y scatterplot.
8. Note the coefficients, their levels of statistical significance, and the 95% confidence intervals from the regression– you'll use this information below.

Now, build 5 slides to accompany your presentation of the results:

9. Make a slide that will aid your narration of why we might expect a causal relationship between X_1 and Y.
10. Make a slide that will aid your narration of where the data came from and what the three variables are.
11. Make a slide that contains your scatterplot with regression lines.
12. Make a slide that contains both your regression equation as well as a table with the estimated regression coefficients, standard errors, p-values, and 95% confidence intervals for both X_1 and X_2.
13. Make a slide that will aid your narration of why, even if the results were perfectly statistically significant, we should be skeptical of a causal interpretation.

Finally, prepare a 10-minute presentation of your slides:

14. Prepare a presentation to give with your slides **that lasts no longer than 10 minutes** and covers the following:

    (a) Very briefly introduces a hypothesis for why the concept measured by X_1 could be causally related to the concept measured by Y.

    (b) Introduces the data and three variables you'll use.

    (c) Briefly introduces the scatterplot and what it suggests about relationships in the data.

    (d) Explains the regression run and introduces the regression lines in terms of sign and magnitude.

    (e) Explains what each regression line tells us and what the difference between them tells us.

    (f) Explains in what sense the regression "controls for" X_2.

    (g) Discusses the statistical significance of the results and what we should make of the hypothesis based on all of this.

    (h) Explains why regression results that perfectly supported the hypothesis would not *confirm* a causal relationship. (Think about confounders, endogeneity, etc.)

    (i) Offers one concrete way that future research could take a step toward better assessing this hypothesis. (new data? new analyses?)