PS405: Linear Models

Problem Set 1

Professor: Nicole Wilson TA: Artur Baranov

Due: Thursday, January 23, 2025

Please submit your write-up, including *any* code that you ran, via Canvas. We suggest using Quarto or `RMarkdown`, but any software that renders R code into a PDF document will work. I have provided an `RMarkdown` template if you prefer to use it (note that the process/format is very similar to Quarto, if you are familiar with that). This should be completed *before* class begins at 9:30 AM. Please save your write-up with easy-to-recognize file names (e.g., `ps1_wilson.pdf` or `ps1_baranov.pdf`). Late submissions will not be accepted without special permission from the course instructors. Where instructed to follow a specific coding procedure in `R`, such as the creation of a new function or a data generating process, please include the code in a code chunk in your write up. For example, you can use `echo = TRUE` in the chunk header in Quarto or `RMarkdown` to ensure your code is visible in the output. You do not need to submit the raw file you use to produce your PDF, but the course instructors may request it on a case-by-case basis.

Please post all questions to the Canvas anonymous discussion forum.

Problem 1

Imagine we're analyzing data obtained through random sampling from a population for which the true data generating process is $y_i = \beta_0 + \beta_1 x_i + u_i$. We're interested in estimating the relationship between x and y, and so we run a simple ordinary least squares (OLS) regression of y_i on x_i, giving us: $y_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i + \widehat{u}_i$. Define each of the following quantities both in words and with the notation used in class. Also specify each time whether we can *calculate* the quantity, and if so, how you would calculate it.

(a) Regression residuals

(b) The population error

(c) The population regression slope coefficient

(d) The estimator of the regression slope coefficient

Problem 2

In this problem, we are going to explore some of the assumptions that are necessary to make OLS an unbiased and consistent estimator for β_0 and β_1. We'll look especially at the zero conditional mean assumption. You can assume throughout this problem that OLS assumptions I, II, and III as defined in lecture (variation in X, random sampling, linearity in parameters) all hold.

(a) The zero conditional mean assumption can be expressed as $\mathbb{E}[u|X] = 0 \longrightarrow Cov(X, u) = 0$. Translate this into plain English, making sure to define both X and u. If the zero conditional mean assumption is *not* satisfied, how does this impact our estimate of β_1 (i.e. the coefficient on X)?

(b) Now, load the example data, `example.csv`. For now, we are going to deal with the first three variables in `example.csv`: a predictor variable `x`, a response variable `y`, and an error term, `u`. Note the true data generating process that was used to create the values of Y in the dataset: $Y = 1 + 0.5X + u$.

> **NOTE:** This setup is unrealistic in two ways. First, in the real social science world, we never know the true data-generating process. Second, remember that u is unobserved by definition in real life; no dataset you ever work with in the real world will include the true error term.

i. Plot the outcome Y against the regressor X. Add two lines to your scatterplot. First, add a line with the *known* regression coefficients $\beta_0 = 1$ and $\beta_1 = 0.5$. Second, regress Y on the *observed* regressor X using the command lm(). Do **not** include u because it is unobserved. Then, plot a line with the regression coefficients you extract from the object. Add a legend to your plot identifying which line is which. You are welcome do this in `ggplot` but it might be easier using the `plot()` command in base `R`. You can also make use the `abline()` and `legend()` commands.

ii. Compare the line defined by the *true* coefficients and the line defined by the `lm()` fit. How similar are they? If they are different, which one fits the scatterplot of the data better? If the line defined by the *true* coefficients represents our estimand, and the `lm()` line represents an estimate generated by the estimator $Y = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{u}$, would you say the estimator is biased or unbiased for the parameters β_0 and β_1? Why do you think this might be the case?

(c) Let's take advantage of the one-time opportunity this problem affords us to observe u directly.

i. Plot the error term u against x, making sure to title and label the plot appropriately. Does the zero conditional mean assumption seem to hold for our data? While we're looking at the plot, let's also evaluate whether the homoskedasticity assumption holds. Explain what you're looking for in the plot to evaluate zero conditional mean and homoskedasticity.

ii. Now use `lm()` to regress u on X and report the intercept, the coefficient on X, and the corresponding standard errors. Compare the regression coefficient on X to the regression coefficient you found using the `lm()` command in part b), and to the *true* coefficient on X given in the text of part b). What do you notice about the relationship between these three coefficients? What, if anything, does this relationship tell you about how the relationship between X and u relates to the bias of the OLS estimator for β_1?

iii. Consider what you found in (c)i and (c)ii along with your plot and explanation from part (b). Would failing to satisfy the zero conditional mean assumption explain your finding in part (b)? What about failing to satisfy the homoskedasticity assumption? Re-interpret, in a sentence or two, your finding from part (b) in light of your new knowledge about the relationship between u and X.

(d) As noted in the beginning of Problem 2, we never actually observe u when we are running a regression in the real world. If we never have the data necessary to make plots like the ones in (c)i, how are we supposed to assess whether the zero conditional mean assumption holds in the real world? A friend hears you wonder out loud about how to evaluate zero conditional mean, and jumps in to help. "Easy!" she says. "When you run a regression, you get a vector of residuals $\hat{u}$ that quantifies the difference between the actual value of Y and the predicted value $\hat{Y}$ for each point. If you regress those residuals on X, and see that they're un-correlated, then you know that zero conditional mean holds!" Your friend is wrong, but you decide to humor her.

i. Extract the residuals from the regression you ran in part (b) and plot them against X. Calculate the covariance between the $\hat{u}$ and X. Describe what you see. Compare this plot to your plot from (c)i, which showed the relationship between X and the *true* errors. Would you draw the same conclusion from the residual plot as from the *true* error plot if you were trying to evaluate the zero conditional mean assumption?

ii. Explain to your friend why the residuals from a regression will *always* be un-correlated with the predictor even if the zero conditional mean assumption is violated.

Problem 3

On November 5, 2024, analysts on "decision desks" were wrangling through data trying their best to project who won the U.S. presidential election. They looked at a variety of data sources, including exit polls, incoming returns, and historical data. The 2020 election cycle saw the biggest and most dramatic shift in how Americans vote in American history. Because of permanent and temporary changes made to state election laws amidst a global pandemic, the percentage of voters casting ballots by mail in 2020 doubled compared to 2016. The 2022 election saw some backing off this surge, but overall, voting by mail in 2022 was well ahead of the trend that had been established over the previous two decades. Moreover, Democratic voters, as in 2020, continued to be more likely to vote by mail than Republican voters in 2022.

Key to projecting a winner in 2024 was be understanding what percentage of all votes cast will be cast on Election Day, during early voting, or by absentee/mail ballot. In this problem, we will pretend we are analysts in November and implement a series of simple OLS models to make predictions about the usage of different voting modalities in the important state of North Carolina. To run these models, you will use the `nc_precincts.csv` file available on the assignment page on Canvas. This file provides registration and turnout data at the precinct-level (the smallest

geographic unit established by governments for conducting election) from the 2022 election to make predictions about 2024. The file contains the following columns:

- `precinct_id`: unique identifier for each precinct
- `total_reg_2022`: total number of registered voters in the precinct as of November 2022
- `total_reg_2024`: total number of registered voters in the precinct as of November 2024
- `pct_GOP_2022`: percentage of registered voters in the precinct who were registered as Republicans in November 2022
- `pct_GOP_2024`: percentage of registered voters in the precinct who were registered as Republicans in November 2024
- `pct_turnout_2022`: percentage of registered voters in the precinct who voted in the November 2022 election
- `pct_absentee_vbm_2022`: percentage of registered voters in the precinct who voted absentee or by mail in the November 2022 election
- `pct_early_2022`: percentage of registered voters in the precinct who voted early in the November 2022 election
- `pct_election_day_2022`: percentage of registered voters in the precinct who voted on Election Day in the November 2022 election

(a) Load `nc_precincts.csv`. Several analysts point to the relationship between the percentage of registered voters who are Republican in a precinct and both the turnout rate and relative usage of each voting modality (absentee/mail, early, and Election Day) within each precinct. We will first run a series of simple OLS models using data from the 2022 election to examine the linear relationship between these variables.

i. Run four separate OLS regressions, one for each of the following dependent variables: `pct_turnout_2022`, `pct_absentee_vbm_2022`, `pct_early_2022`, and `pct_election_day_2022`. In each regression, use `pct_GOP_2022` as the independent variable.

ii. Report the estimated coefficients for `pct_GOP_2022` as well as key diagnostic statistics from each regression in a *single* table. Your table should be well formatted with readable lables for relevant variables and for each corresponding model. *Hint: you may wish to use the* `modelsummary` *package to create a table of regression results, but you can use any package you prefer as long as your table is legible, nicely formatted, and well-labelled.*

iii. Interpret the coefficient for `pct_GOP_2022` in each regression. What does the sign and magnitude of the coefficient tell you about the relationship between the percentage of registered voters who are Republican in a precinct and the dependent variable in each regression? (Note: for now, set aside whether or not they are "statistically significant'' since we haven't really talked about that yet)

iv. Interpret the R^2 statistic for each regression. What does the R^2 statistic tell you about the fit of each model?