Gov 2002: Problem Set 10

Spring 2021

Problem Set Instructions

This problem set is due on April 29, 11:59 pm Eastern time. Please upload a PDF of your solutions to gradescope. We will accept hand-written solutions for problem 1 but we strongly advise you to typeset your answers in Rmarkdown. Problems 2-4 should be typeset. Please list the names of other students you worked with on this problem set.

The two datasets you'll need to do this problem set can be found at:

- http://gov2002.mattblackwell.org/data/prezElectionSubRepublicLevel.csv
- http://gov2002.mattblackwell.org/data/prezElectionRepublicLevel.csv

Question 1: OLS Confidence Intervals

The standard output from OLS will give the standard errors for the estimated coefficients, but often we want to obtain measures of uncertainty for the predicted value of Y_i given some value of X_i (that is, the conditional expectation function). Using the example from lecture, we might be interested in the average wait times to vote for individuals making \$25,000, \$50,000, or \$100,000 in annual income, along with measures of uncertainty around those estimates. In this problem we will look at how to calculate interval estimates for these predicted values. Assume the following *true* population model for $Y_i|X_i$:

$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

where the X_i are random variables and u_i are i.i.d. random variables with $E[u_i \mid X_i] = 0$ and $Var(u_i \mid X_i) = \sigma^2$. Suppose we observe a random sample of n paired observations $\{Y_i, X_i\}$. Assume the Gauss-Markov assumptions hold and that we have a large sample. Our goal is to estimate the predicted value at some value $X_i = x$:

$$\mu(x) = E[Y_i \mid X_i = x] = \beta_0 + \beta_1 x.$$

(a)

Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be OLS estimators of the regression of Y on X. Use what you know about the unbiasedness of OLS estimates to show that $\widehat{\mu}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ is an unbiased estimator of the population quantity $\mu(x) = E[Y_i \mid X_i = x]$.

(b)

Find the conditional variance of $\hat{\beta}_0$, $Var(\hat{\beta}_0 \mid X_1, \ldots, X_n)$, using these two facts:

$$Cov(\overline{Y}, \hat{\beta}_1 \mid X_1, \ldots, X_n) = 0 \quad \text{and} \quad Var(\hat{\beta}_1 \mid X_1, \ldots, X_n) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

You answer should be in terms of σ^2 and functions of X_i.

(c)

Find the covariance of the OLS estimates given our X values, $Cov(\hat{\beta}_0, \hat{\beta}_1 | X_1, \ldots, X_n)$, again in terms of σ^2 and functions of the X_i. (Hint: It's not zero!)

(d)

Using what you found in (b) and (c), find the standard error of $\widehat{\mu}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$.

(e)

Assume that we don't know σ^2 and instead construct our estimate of the standard error by plugging in for σ^2 our unbiased estimate s^2 using the residuals.

Give the formula for a large-sample 95% confidence interval estimator for $\mu(x) = E[Y \mid X = x]$ using what you found above and substituting s^2 for σ^2. How do we interpret this confidence interval?

Question 2: Your Own `lm()`

Last problem set, you wrote a custom function to perform OLS on `subprime` data. In this problem, you will code the rest of `lm()`'s core functionality.

(a) Write a generic function that takes two arguments: a `formula` (such as `income ~ loan.amount`), and `data` (a data frame). You may copy the function from last problem set.

Your function should return a `list()` object, with the following elements:

- `coefficients`: the coefficient estimates (you did this already)
- `R.squared`: the R^2 of your model (you did this already)
- `standard.errors`: standard errors of your coefficients (assuming homoskedasticity)
- `t.stats`: t-statistics for your coefficients
- `p.vals`: p-values for your coefficients

Do not use `lm()` in your function. We want you to code the estimates using `R`'s matrix operations, such as `t()` (transpose), `%*%` (multiply matrices), `solve()` (invert), `diag()` (extract diagonal).

(b) Now, test your function again on the subprime data. Run the following regression model `income ~ loan.amount + black + woman` using your function from part (a) to `lm()` and compare the outputs.

(c) Finally, run your regression model with an interaction term, `income ~ loan.amount + black + woman + black:woman`. Compare the outputs to `lm()`. Does the substantive interpretation of your results change from part (b)?

Question 3

For this problem, we are going to look at data on the 2012 Russian Presidential Election. This election was held a year after the 2011 Duma (Parliamentary) Election that, according to OCSE observers, was considered "slanted in favor of the ruling party" - Vladimir Putin's United Russia.[1] Observers noted a number of irregularities in the voting process, including evidence of ballot-box stuffing at some polling stations. The elections were followed by a series of protests in major cities against the government denouncing the election fraud. As a consequence, the 2012 elections were subject to greater domestic observation efforts, but were still highly skewed in favor of the ultimate winner - Vladimir Putin. As OCSE observers noted, while voting procedures were relatively well followed, the vote count showed many procedural irregularities at around 1/3 of polling stations.[2]

(a) We will first look at the election returns at the sub-regional level (roughly equivalent to county-level). Load the dataset `prezElectionSubRepublicLevel.csv` into R. Create a new variable `putinvote` which is equal to the number of votes for Putin (`Putin`) divided by the number of valid ballots cast (`Number.of.Valid.Ballots`) multiplied by 100 (to scale to a percentage from 0 to 100). Further, create another variable `turnout` which is equal to the number of valid ballots cast (`Number.of.Valid.Ballots`) divided by the number of registered voters (`Number.of.Registered.Voters`).

(b) Make a scatterplot of Putin's share of the vote on the percentage turnout in the sub-region. Run a linear regression of Putin's vote share on percent turnout and overlay the regression line on top of the points. Make sure the regression line is clearly visible.

[1] http://www.osce.org/odihr/elections/86981
[2] http://www.osce.org/odihr/elections/88661

Report your estimated regression coefficients and standard errors in a neatly formatted table and interpret the coefficient on % turnout.

Hint: Since there are a lot of datapoints, in order to reduce clutter, you may want to change how the points look (e.g. set `pch=18` if you're using base `R` or `shape` in ggplot) and their size (e.g. set `cex = .5` in base or `size = 0.5` in ggplot). You might also want to change the color of the points (e.g. `col = "darkgrey"` in base or `color = "darkgrey"` in ggplot).

(c) Take a close look at your scatterplot. Just eyeballing it, do you detect any evidence of nonlinearity here?

(d) Now, test your intuition by making a plot of the residuals from your regression in B) against the fitted values from that regression. Add a smooth loess regression curve to the plot (using `geom_smooth` in ggplot, or the `scatter.smooth` function in base R). Does there appear to be a pattern in the residuals? What does this suggest?

(e) How might you specify a model for Putin's vote share that accounts for the pattern in (c-d)? Run a regression using that model specification and report your results in a nicely formatted table below.

(f) Make a plot of the residuals from your regression in (e) against the fitted values from the same regression. Add a smooth loess regression curve to the plot as in (d). Does the pattern from (c-d) remain?

Question 4

Now, we are going to see if the inverse relationship generally holds for the opposition candidates. In particular, we will look at the vote share for Mikhail Prokhorov, former CEO of Norilsk Nickel and current owner of the Brooklyn Nets, who ran against Putin as an independent candidate on (roughly) an economically liberal/pro-business/pro-Western platform.

To get a better sense of which regions comprise the data points, we are going to look at the election data aggregated at the regional level (roughly 84 "federal subjects" plus votes from Russians living abroad). Load in the `prezElectionRepublicLevel.csv` file. You should have 85 observations in this dataset.

(a) As before, make a variable for the % share of votes received by Prokhorov (the number of raw votes for Prokhorov is found in the variable `Prokhorov`) and for turnout.

(b) Estimate a regression of % votes for Prokhorov on % turnout, report your coefficient and standard error estimates in a neatly formatted table and substantively interpret the coefficient on % turnout. Add the regression line to your scatterplot.

(c) Eyeballing the scatterplot suggests a number of major outliers. Compute the leave-one-out prediction errors for your regression in (a) using the `residuals()` and `hatvalues()` functions. Which three regions have the highest leave-one-out prediction errors? Hint: To get region names, look at the `Region.Name.English` variable.

(d) Use the Cook's Distance for each of the three outliers you identified in (c) to identify which one is the most influential data point.

(e) Remove the most influential outlier point from your data and estimate a regression of % votes for Prokhorov on % turnout. Report the estimated coefficients and standard errors and interpret your coefficient on % turnout. Add the regression line from this regression to your scatterplot from (a) (you should have two lines on the plot - differentiate them by color). Compare your estimated coefficient on % turnout with the estimate you obtained in (a) and explain why you see any discrepancies.

(f) When you removed the outlier point in (e), how did you change the population of interest? Hint: What observations remained in the sample and what observations did not? Look at the name of the region you removed and consider that a sizeable amount of Prokhorov's support came from Russians living abroad.