Gov 2018 Lab 2: Prediction with regression

Adeline Lo

February 1, 2022

Please upload your lab *with comments*, in .rmd *and* knitted forms on Gradescope so we can verify it runs without problem. For any problems that require calculations, please show your work. Finally, when doing simulations or draws from a random distribution in `R`, please set your seed immediately using `set.seed(12020)`.

This assignment will analyze vote returns for California House elections and vote choice in a presidential election.

2006 California Congressional Election Results

Our goal in this exercise is to calculate the probability that a Democratic candidate wins a House seat in a "swing district": one where the support for Democratic and Republican candidates is about equal and the incumbent is a Democrat.

Question 1

Load the data set `ca2006.csv` from your computer into R. `ca2006` a slightly modified version of the 2006 House election return data from the PSCL library.

Create a plot of the proportion of votes for the Democratic candidate, against the proportion of the two-party vote for the Democratic presidential candidate in 2004 (John Kerry) in the district. Be sure to clearly label the axes and provide an informative title for the plot

The data set contains the following variables:

- `district`: California Congressional district
- `prop_d`: proportion of votes for the Democratic candidate
- `dem_pres_2004`: proportion of two-party presidential vote for Democratic candidate in 2004 in Congressional district
- `dem_pres_2000`: proportion of two-party presidential vote for Democratic candidate in 2000 in Congressional district
- `dem_inc`: An indicator equal to 1 if the Democrat is the incumbent
- `contested`: An indicator equal to 1 if the election is contested

Question 2

Regress the proportion of votes for the Democratic candidate, against the proportion of the two-party vote for the Democratic presidential candidate in 2004 in the district. Print the results and add the bivariate regression line to the plot.

Question 3

Using the bivariate regression and a function you have written yourself (not `predict`!), report the expected vote share for the Democratic candidate if `dem_pres_2004` = 0.5.

Question 4

Now, regress `prop_d` on `dem_pres_2004`, `dem_pres_2000` and `dem_inc`.

Using the multivariate regression and a function you have written yourself, report the expected vote share for the Democratic candidate if:

- `dem_pres_2004` = 0.5
- `dem_pres_2000` = 0.5
- `dem_inc` = 1

Question 5

We are often interested in characterizing the uncertainty in our estimates. One nonparametric approach to estimating uncertainty is called the bootstrap. Here, we will walk through the steps to produce a function that will produce a bootstrap distribution, from which you can find the bootstrap estimates (and 95% confidence bands).

Do the following 1000 times (in a for loop; we'll learn how to vectorize/parallelize later):

a) Randomly sample 53 rows – the number of districts in California in 2006 – with replacement (set the seed). This is your bootstrap sample.
b) Using the bootstrap sample from a), fit the bivariate and multivariate regressions.
c) Using the regressions from b), predict the expected vote share for the Democratic candidate from both regressions, using the values and functions from 3) and 4)
d) Store the predictions from both regressions.
e) Repeat a-d 1000 times.

Question 6

Create histograms for both predictions adding lines for the prediction (each from Question 3 and 4) and 95% confidence bands.

We will say the model predicts that the incumbent wins if the predicted vote share is greater than 50%. What proportion of time does each model above predict the incumbent will win?