Computing Assignment 5

Predicting Presidential Elections

Due November 6 before class.

In this assignment, you'll be using a data set on the 12 U.S. presidential elections from 1952 to 1996 to "predict" the five presidential elections from 2000 to 2016. Of course, you already know the outcomes, so this isn't really a prediction at all, just practice. There are 16 different variables in this dataset. The variables included are the following:

1. `Year` - election year
2. `inc1` - Two-party vote share received in the presidential election by the party currently holding office.
3. `q2gdp` - GDP growth in the second quarter of the election year (Abramowitz, 2012)
4. `cumpt8pclei2016` - Measure of Leading Economic Indicators, see Erikson and Wlezien (2012)
5. `LogTWH` - logged time in the White House for a party (Lockerbie, 2012)
6. `G` - growth rate of real GDP per capita in the first three quarters of the election year (Fair, 2012)
7. `P` - absolute value of the GDP deflator in the first 15 quarters of the administration (Fair, 2012)
8. `Z` - number of quarters in the first 15 quarters of the administration in which the growth rate of real per capita GDP is greater than 3.2% (Fair, 2012)
9. `ECONHALVED80` - Qtr. 2 real GDP growth (Campbell, 2012)
10. `gnpchan` - Gross National Product, as percentage change non-annualized in GNP constant dollars from the fourth quarter of the year prior to the election to the second quarter of the election year, data from the Survey of Current Business (Lewis-Beck and Tien, 2012)
11. `Unemployment` - average unemployment rate in the months from January to August in the election year (Bureau of Labor Statistics, 2016)
12. `Inflation` - average inflation rate in the months from January to August in the election year (US Inflation Calculator, 2016)
13. `ViolentCrimeRate` - violent crime rate per 100,000 people in the year prior to the election (Federal Bureau of Investigation, 2016)
14. `MurderRate` - Murder and non-negligent manslaughter rate per 100,000 people in the year prior to the election (Federal Bureau of Investigation, 2016)
15. `AssaultRate` - Aggravated assault rate per 100,000 people in the year prior to the election (Federal Bureau of Investigation, 2016)
16. `OwnershipShare` - Rate of homeownership in April of the election year (Federal Reserve Bank of St. Louis, 2016)

In class we mentioned that the most predictive model isn't too simple, but also isn't too complicated. It should use a mixture of measures-of-fit (e.g., r.m.s. error, out-of-sample r.m.s. error, BIC) and theoretical intution.

This assignment asks you to build three regression models:

1. A **too-simple model** that works well, but could be improved by adding additional explanatory variables. Use about one explanatory variable.
2. A **just-right model** that includes the same varibles as the too-simple model, but include additional explanatory variables that *improve the predictive power of the model*. Use about three explanatory variables, but use BIC and your intuition to figure out the right combination.
3. A **too-complicated model** that includes the same varibles as the just-right model, but include additional explanatory variables that *worsen the predictive power of the model*. Use about eight explanatory variables. Note that where least squares allows *at most* $n-1$ explanatory variables, n is the number of observations in the data set used to fit the model.

In order to complete the assignment, do the following:

1. Download the code for the assignment from here. Copy-and-paste it into RStudio and make the following changes.
   - Make sure you have the tidyverse package installed.
   - On line 3, hange the working directory to from my `pols-209` folder to your `pols-209` folder.
   - On line 13, change `my_name <- "Josh A."` to your first name and last initial. If you share a name with a classmate, then use your full last name, like `"Will Barton"`.
2. Run the code to make sure it works.
3. On lines 24-26, change the formulas for the three regression models on lines to achieve the too-simple, just-right, and too complicated goals discussed above. Note that each time you run the *entire* script, RStudio spits out the measures of fit for each model. Feel free to experiment a bit, using the measures of fit and your theoretical intuition to find the right combination of variables. Notice that you can include multiple explanatory variables on the right-hand side of the formula by separating them with a `+` (e.g., `y ~ x1 + x2`).
4. Once you are satisfied that you have the right models, run the entire script one last time. This creates a file in your `data` subfolder called `election-predictions-josh-a.csv`, where my first name and last initial are replaced with yours. You need to send this file to me via Dropbox by clicking **here**, then clicking *Choose files*, navigating to and selecting this file in the `data` subfolder of `pols-209`.
5. Last, click the white notebook to compile a report and submit that via eCampus as usual.

Last, there is a friendly competition with the predictions. The author of the model with the lowest out-of-sample root mean squared error, will earn an extra 2% on the final grade. Feel free to collaborate to troubleshoot problems, but work separately on specifying the models! If lots of people make the same predictions, I may not apply the bonus.