Homework 2:

Advanced Data Analysis in Python

Due before class Monday, March 9, 2020

The purpose of this homework is to familiarize you with gathering data, working with NumPy and Pandas, performing linear regression "by hand," and writing reports. There are obviously many modules that perform linear regression available, but to develop your understanding of linear algebra and the simple linear model, I want you to only use NumPy and Pandas for matrix manipulation and calculations. You will also need a module to calculate the t-statistic.

More specifically, you are to write a function that takes as an input an outcome variable (target) and covariate(s). These can either be one data set input with a clearly marked outcome variable, or two separate arguments. The function should return in any convenient format the regression estimates ($\hat{\boldsymbol{\beta}}$), their standard errors, and 0.95 credible intervals. It should also perform list-wise deletion for handling NaN values. Save this function in its own .py file, named for the function within it. A test file for this function should be included that deals with bad input handling. As a reminder, the following equations should be used:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}, \\
\mathrm{Var}(\hat{\boldsymbol{\beta}}) &= \sigma^2(\boldsymbol{X'X})^{-1}, \\
\sigma^2 &= \frac{\boldsymbol{e'e}}{n-k-1},
\end{aligned}$$

where

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}},$$
$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}.$$

In a separate file, using either sraping or an API, retrieve an outcome variable of interest, an explanatory variable, and at least one causally prior control variable. These variables should make sense, but do not need to be strongly theoretically motivated or exhaustive by any means. If you are feeling uncreative, just use some economic variables from the World Bank API we went over in class. Write the resulting data set to a csv, and include the csv in the submission.

Next, import the function you wrote for linear regression, and run a model on your data. Write up a hypothesis (again, it does not have to be strongly theoretically motivated), a very brief justification for the inclusion of the control(s), and the results (null results are fine). The results should include both a brief discussion, and either a regression table or a plot of regression coefficients with credible intervals. The regression table should contain the estimates, the standard errors, and the 0.95 credible intervals. There is no need for using stars when including this information. Not counting the regression table or plot, this report should not exceed one page. Be concise. Submit this report as a pdf. Even if you use Word, convert it to a pdf.

To summarize, the submission should include the following:

(1) A .py file containing the linear regression function

(2) A .py test file for the linear regression function that tests for bad input

(3) A .py file (or ipython session) containing the code used to gather the data, write the csv, and run the regression

(4) The csv file of the data

(5) A pdf report of the hypothesis and findings

As always, if any of this in unclear or you need assistance, come to my office hours. Feel free to collaborate, even use the same data, but submit your own files.