Statistical Methods for Discrete Response, Time Series, and Panel Data (W271): Lab 4

W271 Instructional Team

December 9, 2017

Instructions:

- **Due Date: 12/19/2017 (11:59 p.m. PST)**
- Submission:
  - Submit your own assignment via ISVC
  - Submit 2 files:
    1. A pdf file including the summary, the details of your analysis, and all the R codes used to produce the analysis. Please do not suppress the codes in your pdf file.
    2. R markdown file used to produce the pdf file
  - Each group only needs to submit one set of files, but please make sure the names of all members on the team are listed on the first page of your report.
  - Use the following file naming convensation
    - FirstNameLastNameFirstInitial_Lab5.fileExtension
    - For example, if you are in Section 1 and have two students named John Smith and Mary Doe, you should name your file the following
      - JohnS_MaryD_Lab5.Rmd
      - JohnS_MaryD_Lab5.pdf
  - Although it sounds obvious, please write the name of each members of your group on page 1 of your pdf report.
  - This lab can be completed in a group of up to 3 people. Each group only needs to make one submission. Although you can work by yourself, we encourage you to work in a group.
- Other general guidelines:
  - Please read the instructions carefully.
  - Please read the questions carefully.
  - Use only techniques and R libraries that are covered in this course.
  - If you use R libraries and/or functions to conduct hypothesis tests not covered in this course, you will have to explain why the function you use is appropriate for the hypothesis you are asked to test
  - Thoroughly analyze the given dataset. Detect any anomalies, including missing values, potential of top and/or bottom code, etc, in each of the variables.
  - Your report needs to include a comprehensive Exploratory Data Analysis (EDA) analysis, which includes both graphical and tabular analysis, as taught in this course.
  - Your analysis needs to be accompanied by detailed narrative. Remember, make sure your that when your audience (in this case, the professors and your classmates) can easily understand your your main conclusion and follow your the logic of your analysis. Note that just printing a bunch of graphs and model results, which we call "output dump", will likely receive a very low score.

  - Your rationale of any decisions made in your modeling needs to be explained and supported with empirical evidence. Remember to use the insights generated from your EDA step to guide your modeling step, as we discussed in live sessions.
  - All the steps to arrive at your final model need to be shown and explained very clearly.
- Students are expected to act with regards to UC Berkeley Academic Integrity.

Description of the Lab

In this lab, you are asked to answer the question **"Do changes in traffic laws affect traffic fatalities?"** To do so, you will conduct the tasks specified below using the data set *driving.Rdata*, which includes 25 years of data that cover changes in various state drunk driving, seat belt, and speed limit laws.

Specifically, this data set contains data for the 48 continental U.S. states from 1980 through 2004. Various driving laws are indicated in the data set, such as the alcohol level at which drivers are considered legally intoxicated. There are also indicators for "per se" laws—where licenses can be revoked without a trial—and seat belt laws. A few economics and demographic variables are also included. The description of the each of the variables in the dataset is come with the dataste.

Exercises:

1. Load the data. Provide a description of the basic structure of the dataset, as we have done in throughout the semester. Conduct a very thorough EDA, which should include both graphical and tabular techniques, on the dataset, including both the dependent variable *totfatrte* and the potential explanatory variables. Remember, graphs must be well-labeled. You need to write a detailed narrative of your observations of your EDA. *Reminder: giving an "output dump" (i.e. providing a bunch of graphs and tables without description and hoping your audience will interpret them) will receive zero point in this exercise.*
2. How is the our dependent variable of interest *totfatrte* defined? What is the average of this variable in each of the years in the time period covered in this dataset? Estimate a very simple regression model of totfatrte on dummy variables for the years 1981 through 2004. What does this model explain? Describe what you find in this model. Did driving become safer over this period? Please provide a detailed explanation.
3. Expand your model in *Exercise 2* by adding variables *bac08, bac10, perse, sbprim, sbsecon, sl70plus, gdl, perc14_24, unem, vehicmilespc*, and perhaps *transformations of some or all of these variables.* Please explain carefully your rationale, which should be based on your EDA, behind any transformation you made. If no transformation is made, explain why transformation is not needed. How are the variables *bac8* and *bac10* defined? Interpret the coefficients on *bac8* and *bac10.* Do *per se laws* have a negative effect on the fatality rate? What about having a primary seat belt law? (Note that if a law was enacted sometime within a year the fraction of the year is recorded in place of the zero-one indicator.)
4. Reestimate the model from *Exercise 3* using a fixed effects (at the state level) model. How do the coefficients on *bac08, bac10, perse, and sbprim* compare with the pooled OLS estimates? Which set of estimates do you think is more reliable? What assumptions are needed in each of these models? Are these assumptions reasonable in the current context?
5. Would you perfer to use a random effects model instead of the fixed effects model you build in *Exercise 4*? Why? Why not?
6. Suppose that *vehicmilespc*, the number of miles driven per capita, increases by 1,000. Using the FE estimates, what is the estimated effect on totfatrte? Be sure to interpret the estimate as if explaining to a layperson.

7. If there is serial correlation or heteroskedasticity in the idiosyncratic errors of the model, what would be the consequences on the coefficient estimates and their standard errors?