Week 2 Live Session

Paul Laskowski

September 6, 2016

1. **Slack Channel**
   If you are not on the course-wide slack channel yet, please

   1. make sure you have joined the ucbischool.slack.com slack team
   2. enter your slack ID into the short answer poll

2. **Fertility and Education EDA:** A selection of your critiques

   (a) Michael Amodeo (Section 1) - Given changes in demographics and medicine over the 25 years since Sander's study, it would be interesting to see a re-evaluation of the selected age range. It is more common to have children later (especially with women who have achieved post-graduate education), so the selected age range might not be meeting the desired criteria of finding a sample of women who have completed both education and childbearing.

   Jonah Smith and David Harding make similar observations.

   (Follow-up question: If it's true that many women in the study have not finished having children or have not finished their education, how might this be reflected in the observed relationship between fertility and education?)

   (b) Sam Goodgame (Section 2) - Removing the 'NA' values makes an assumption which may not be valid: that the pool of respondents who listed 'NA' are representative of the sample as a whole. There is probably a relationship between the respondents who listed 'NA' as survey responses and what the true data for those fields actually is. In other words, if a respondent didn't understand a question about years of education–and as a result listed 'NA'–then it is plausible that that respondent's level of education is different from the mean...

   (Follow-up question: describe a hypothetical worst-case scenario, in which the missing values would dramatically alter the conclusion of your analysis)

   (Your instructor will introduce the related assumption of selection-on-observables)

   (c) Jessica Economou (Section 4) - I thought the educational binning (breaks of 0-11 yrs, 12-15 yrs, and 15+ yrs) could have been improved. The high school graduate bin was somewhat misleading, as it included those with 14 years of education, meaning they completed two years of college/technical school. I would also want to include a bin for those who went further than getting the standard 4-year college degree.

   (Follow-up question: How do you decide when to bin levels of a variable together?)

   (d) Cynthia Hu (Section 5) - I think we had better split the first group 0-11 into 0-5 and 6-11 as data points with education years less than 6 show a different pattern.

   (Comment: This is a nice example of finding features in the data that we can use in modeling. When we learn about linear model specification, we will discuss ways to model this type of discontinuity with indicator variables.)

   (Follow-up question: Can you trust the data points you see with 0-5 years of education?)

3. **Crime and House Prices:** An in-class EDA exercise

The file Boston_w203.csv contains data on neighborhoods in the Boston area. You are given the following codebook.

crim - per capita crime rate by town

zn - proportion of residential land zoned for lots over 25,000 sq.ft.

indus - proportion of non-retail business acres per town

chas - Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)

nox - nitrogen oxides concentration (parts per 10 million)

rm - average number of rooms per dwelling

age - proportion of owner-occupied units built prior to 1940

dis - weighted mean of distances to five Boston employment centres

rad - index of accessibility to radial highways

tax - full-value property-tax rate per $10,000

ptratio - pupil-teacher ratio by town

black - $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town

lstat - lower status of the population (percent)

medv - median value of owner-occupied homes in $1000

You are to begin an exploratory analysis with the objective of understanding how the price of a home relates to neighborhood characteristics, with an emphasis on crime.

```
Boston = read.csv("Boston_w203.csv")
```

1. Generate a scatterplot matrix for all metric variables. Take a few minutes to draw as many insights as you can about the relationships in the data.
2. Examine the main output variable, medv. Comment on any unusual values you find, and any features that might be important for statistical modeling.
3. Examine the main independent variable of interest, crim. What transformation could you apply to this variable to aid in visualizing it? Comment on any unusual features you find.
4. Examine the bivariate relationship between medv and crime. What type of relationship do these variables have?
5. (As time permits) Continue your exploratory data analysis. Be prepared to share interesting findings with the class.