**Violent Crime Project Notes**

**10/4/17**

Here are some ideas for projects to work on:

* Kaggle house pricing prediction
  + <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>
  + This has the advantage of allowing you to compare the performance of your model with the performance of others’ models
  + However, the subject itself isn’t hugely interesting for me
* Violent crime rates in the 1990’s in the US
  + <https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized>
  + This could be more interesting, although there are a huge number of variables that need to be sorted through and everything is aggregated—it is by locality instead of individual
* Race times…haven’t yet located an ideal and reliable data source

If I am going to jump into the nitty-gritty and wrangle with the data before I analyze, I would probably want it to be a data set I acquired myself or at least something that is extremely interesting, like apostasy prediction

* None of these quite hits home in that regard
* Part of the fun is in making discoveries/conclusions that others might not have before, either by using a unique dataset that no one else has analyzed and/or by analyzing the dataset in a unique way

Right now, we need something that is a continuous measure we are interested in predicting

* Run race time could be it—it is a continuous measure of athletic performance
  + However, for some reason, it doesn’t get me super-excited

A morbid part of me wants to look at predicting either murder rate or murder success rate

* Or, you could look at total murders/attempted murders
* How many data sets have statistics for attempted murders, though?

The problem with the Kaggle data set is that it is not unique—you could find in-depth workthroughs, so it wouldn’t necessarily be my work

* Nevertheless, it could be valuable, since it will allow me to see how my model compares with the performance of other people’s models
* Yes, this is the clincher in my mind
* However, this wouldn’t necessarily be the data set I would publish in my blog, it would just be the one I would use to prove I know what I am doing
  + It would allow me to figure out a template to use, which I will then present in my personal projects

Use R markdown for the regression problem

I could also comb the GSS for interesting continuous variables

* Of course, with the GSS, we have time series, which can always be interesting

Subject matter expertise always helps

* What areas do I have expertise?
  + Physics (a little, but actually not much)
  + Politics (a little, but it might be good to steer clear of controversial topics)
  + Martial arts (again, a little)
  + Running
  + Weight training (a little)
  + Catholicism (some)
    - Being covered in the apostasy prediction project
  + Philosophy (a little)

So you have things like

* Race time
  + Limited to sex and age unless you dig
* Striking power?
* Success rate in tournaments
* The Portuguese academic performance data set looked kind of interesting
* HKs in a battle ground in WoW
* Survival time…in what?

**10/11/17**

So, for regression, I think I will settle on predicting crime rates. Now, all I have to do is decide whether to use the 1990s data or to try to piece together a more recent data set (using county-level data).

* If I decide to piece together my own data, this could get rather expensive in terms of time.
* Start by looking through the 1990s data for interesting variables
* This will be an excellent data set for model selection due to the large number of ready-to-use variables
  + Can experiment with different model selection techniques and determine which results in the optimal model
  + Most of the variables in this dataset refer to the following:
    - Population
    - Household size
    - Race
    - Age
    - Urban
    - Income
    - Education
    - Employment
    - Family structure
    - Immigration
    - Language (% speaking English)
    - Living quarters (renting/owning/%occupation)
    - Living quarters costs
    - % New residents
    - Number of police/per capita
    - Racial match between police and community as a whole
    - Number of police assigned to drug units
    - Population density
    - Police budget
    - Police cars
    - Police on patrol at a given time
  + Many of the non-race variables above also have racial breakdown

**10/13/17**

**Project Plan**

Goals:

* Determine which demographic and police force variables are **associated** with violent crime rate in municipalities across the US, while controlling for all of the other variables (examine coefficients in regression model)
* **Predict** violent crime based on demographic and police variables

Data: <https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized>

* Source: UCI Machine Learning Repository

This project will be completed in R using R-Markdown

**Steps**:

* **Dataset check**
  + Dimensions
  + Valid values
* **Descriptive statistics**
  + Select variables for use in the analysis
  + Eliminate rows with NA values
  + Visualize geographic locations of communities, with size of circle based on population?
    - This would be cool, but I think it is going to be too time-consuming to do right now
  + Distributions/quartiles
  + Scatter plots
  + Correlation matrix
  + Correlations with the DV
  + Create additional variables based on subject matter knowledge / inference and summarize as above
  + Justify the modeling approach you will use (OLS, etc.: continuous response, n > p, etc.)
* **Inferential Modeling**
  + Select relevant variables by hand based on intuition/theory (if you have time to read)
  + Fit regular OLS model using these variables
  + Check model assumptions (use methods discussed in Spring course)
    - Linear relationship (residuals, possibly including standardized residuals, Ramsey RESET test)
    - Outliers/high leverage points (leverage plots, Cook’s D, etc.)
    - Multicollinearity (correlation matrix, VIFs)
    - Heteroskedasticity (residuals, Breusch-Pagan / Cook-Weisberg test)
    - Non-normal residuals (histogram, Shapiro-Wilk, Shapiro-Francia, Skewness/Kurtosis)
  + Make attempts to address and violations of model assumptions using variable transformations, removing outliers, etc.
  + Once assumptions of model are met, fit the model using standardized and unstandardized coefficients, and comment on what we can infer from this model and from the descriptive statistics—which variables are significantly associated with violent crime rate (and what is their magnitude and direction) and why
* **Predictive Modeling**
  + Use cross validation to compare different model selection approaches and to select the optimal parameters for each approach (look at cost function, bias, variance)
    - Variable transformations
    - Regular unmodified linear regression
    - Best subset (or explain why I can’t use)
    - Forward stepwise
    - Backward stepwise
    - Ridge regression
    - Lasso regression
    - Principal Components Regression (PCR)
    - Partial Least Squares Regression (PLS)
    - K-Nearest Neighbors Regression?
    - Look at each of these techniques by themselves, then, also try combining them
  + Also, possibly look at statistics like Cp, AIC, BIC, adjusted R^2, etc.
  + Compare the performance of each of these techniques with the inferential model chosen in the inferential modeling part of the project
  + Select the approach (or more likely, combination of approaches) that has optimal performance using cross validation
  + Estimate the performance of the optimal model on new data
* Summarize conclusions

When I work on this project, I should have one workbook for work, where I will do literally EVERYTHING I need to do for the analysis. Then, when I publish the results, I will take excerpts from the workbook, as well as excerpts of text from the these notes, and combine them into a user-friendly summary. I will also publish the full workbook (in case a viewer desires greater detail), which will be linked to the summary workbook.

Now, for the rest of tonight, review the course material so far, and summarize anything of import in the spring notebook. Look both at the material for the current course, as well as the material for the course back in the spring.

**10/14/17**

Plan: keep track of project using version control

Initialized project in the following repository: **Violent-Crime-Regression-Project**

Problem: There are many police-related variables that have "?" for the majority of communities--the LEMAS survey must have been conducted on a much smaller number of communities than the communities for which we have demographic data (~300 vs ~1900)

* Maybe you should just go ahead with selecting the communities

With such a large number of missing observations, should I switch over to a different data set?

* Nah, let’s just keep working with the dataset

Another thought just came to me—if I want to generate visually-pleasing visualizations in R, I should probably learn how to use ggplot2—this will take some time—at least a day

* On the other hand, ggplot2 isn’t necessarily essential for this project, either
* Start with standard visualizations, then add more advanced ones in the final published version

I could potentially create 49 dummy variables, one for each state (minus the control state)

* I’m not sure if this is worth it, though
* Might be better to just create a dummy variable for the four overarching census regions: Northeast, Midwest, South, and West
  + Let’s just do this to keep things simple
* Or, I could use more refined geographic regions

Since we have so many predictor variables, it may make more sense to use model selection techniques to select the best model.

Yes, let’s do this for the predictive modeling part. However, for the inferential modeling part, the variables should be selected by hand based on theoretical knowledge of how they affect the response.

Since I will be working on the test tomorrow, let’s start with the predictive modeling part.

Now, we get into the nitty gritty.

* So, what is the cost function we will be seeking to minimize?

Thought: do I want to set up an automated process for deciding whether to transform the variables?

* Problem is, I need to fit a model first!

Another question: when I run k-fold cross validation to estimate the errors, do I just run it once, or should I repeat the k-fold cross validation a number of times?

* Stack Exchange threads seem to indicate you should, although the number of times appears to be conditional on what exactly you’re trying to do
* I could select an arbitrary number of repeats
* OR, I could just divide the data into training and test from the beginning and repeat the whole process multiple times in order to get an estimate of the error on new data

However, when I’m doing the descriptive statistics, etc., I wouldn’t be looking at the full data—the summaries I see could be subject to random fluctuation

The Elements of Statistical Learning book makes the good point that if I select all my parameters using the full data set, then use cross validation to estimate the error, the error estimate may be excessively low, because the process used to generate the parameters has already seen the test data in each of the cross validation steps, so it will be more accurate.

The best approach would be to use the PROCESS I used to generate the model in each of the cross validation steps, which would then give an accurate estimate, since each model would not have “seen” the test data in advance.

But, how do I decide what to do?

* You’ll have to make the model selection process you used fully automated and feed it into the cross validation system to get the expected error
* This involves selecting the best model, selecting the best variables, selecting the best parameters (like regularization, number of principal components, etc.)

To get to that point, I will start by looking at the whole data set and use 10-fold cross validation repeated 10 times to pick the best model selection methods, models, parameters, etc.

Then, once the best model is selected, feed your model selection process into cross validation to get an estimate of the error on new data.