Kevin Palm
5/13/14

Short Questions to Analyzing the NYC Subway Dataset

Section 0. References:

I must have utilized at least a hundred resources throughout this project. A small sampling of them from my browsing history, pretty much picked out at random:

- http://pandas.pydata.org/pandas-docs/stable/gotchas.html
- http://www.tutorialspoint.com/python/dictionary_keys.htm
- http://stackoverflow.com/questions/11033573/difference-between-numpy-dot-and-inner
- http://docs.scipy.org/doc/numpy/reference/generated/numpy.matrix.dot.html
- https://docs.python.org/2/library/json.html
- http://www.tutorialspoint.com/python/list_len.htm
- https://onlinecourses.science.psu.edu/stat414/node/274
- http://stackoverflow.com/questions/18022845/pandas-index-column-title-or-name

Section 1. Statistical Test

1.1 In analyzing whether ridership and rain were correlated, I used the Mann-Whitney U-Test. I used a one-tail P value. The null hypothesis was that rain had no effect upon ridership. My critical p-value was 0.05.
1.2 The Mann-Whitney U-Test was applicable because the data was non-normally distributed.
1.3 The mean entries recorded during rain was 1105, the mean entries recorded without rain was 1090, and the p-value returned by the U-test was 0.025.
1.4 The returned p-value was less than the critical p-value so the null hypothesis was rejected and it can be stated that there is a correlation between rain and ridership. Considering the difference in means (rain had only 1.4% higher ridership), however, the relationship is not extraordinary.

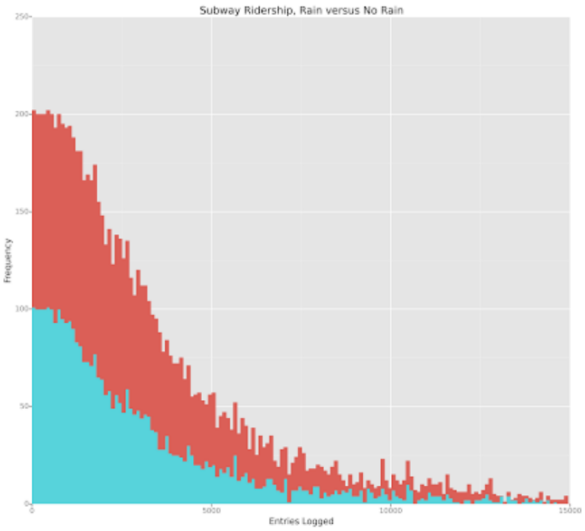
Section 2. Linear Regression

2.1 In problem set 3.8 I used scipy.stats.linregress to create a linear regression based prediction model for ridership.
2.2 In my model I used the features: 'Hour', 'EXITSn_hourly', 'maxpressurei', 'mindewpti', 'minpressurei', 'meanpressurei', 'fog', 'meanwindspdi', 'mintempi', 'meantempi', and 'maxtempi'. I did not use any dummy features in my model.
2.3 All of the features I incorporated in my model had, versus ridership, a p-value < 0.05 in a two-sided p-value hypothesis test whose null hypothesis was that the slope was zero (excluding the index and the dependant variable).
2.4 My features were weighted against their r-values (proportionate to the sum of all the model's r-values) and their individual slopes.

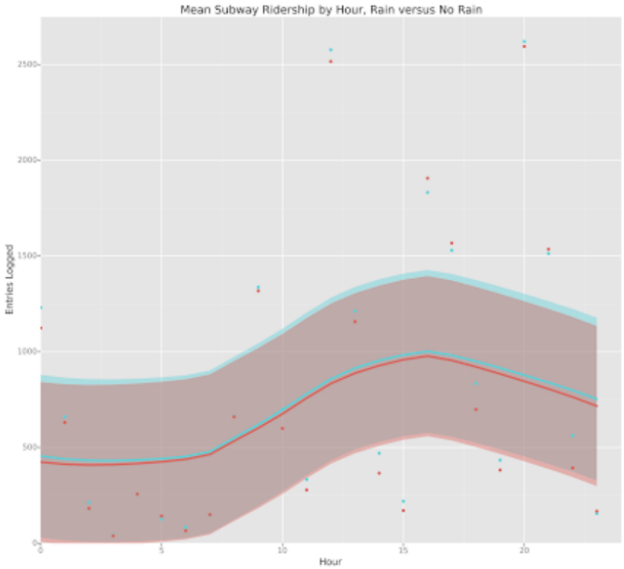
2.5 My model returned an R2 value of 0.55.
2.6 An R2 value of 0.55 means that my model explains about 55% of the observed ridership. I think a linear model is very limited in predicting ridership with the goal of incorporating weather variables. On the other hand, 55% is better than it is worse, and better than the results we achieved through gradient descent.

Section 3. Visualization

3.1



3.2



(Legends did not appear when I was using ggplot with colors, which I'm reading is a glitch.)

Section 4. Conclusion

4.1 According to this dataset, slightly more people ride the subway when it is raining.
4.2 The Mann-Whitney U-test returning a p-value less than 0.05 was the most definitive indication of a relationship between ridership and rain in this dataset. My linear regression based model did not reinforce the rain-ridership finding, as rain was not one of my input variables.

Section 5. Reflection

5.1 For a study specifically interested in rain, a variable quantifying rain as either raining or not raining is a bit crude. Theoretically, a torrential downpour and a light sprinkle should have very different effects upon a pedestrian's plans. A better variable for purpose of the analysis would have been precipitation rate.

The Mann-Whitney U-test returned a p-value within 95% confidence intervals. While this is acceptable for a social science rooted question, it's not as compelling as if in 99% intervals.

In my linear regression based model I included the variable 'EXITSn_hourly' as, versus ridership, it returned a p-value < 0.05 in a two-sided p-value hypothesis test whose null hypothesis was that the slope was zero. Expectably, 'EXITSn_hourly' was the best indicator in the dataset for entries recorded, with an R2 value of 0.74. The next highest variable was 'Hour' at R2 = 0.17. The the weather variables followed way back in the hundredths and negatives. If a researcher was really trying to model subway ridership like this, it's unlikely that he or she would have exit logs available when lacking entry logs, which makes my model very impractical. This also means that my final R2 value of 0.55 is lower than one of my 'start points' of R2 = 0.74, and that I actually hurt my model by basing my prediction off of anything except 'EXITSn_hourly'. This doesn't reflect well on the quality of my model, and I don't think I'll use this same method of selecting variables purely by p-value and weighting purely by R2 value again.