Kevin Palm
5/17/14

Short Questions to Analyzing the NYC Subway Dataset

Section 0. References:

I've probably utilized at least a hundred resources throughout this project while I brushed up on my statistics and Python. A small sampling of them from my browsing history:

- 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 two-tail P value, so as to account for the possibility that ridership during rain is probable to be higher *or* lower than without rain. The null hypothesis for this test states that the probability of two values, one from 'rain' and one from 'no rain', each being randomly selected from the dataset are equally likely to be greater than or less than each other. My critical p-value was 0.05.
1.2 The Mann-Whitney U-Test was applicable because the data was non-normally distributed - each data set appeared non-normally distributed when plotted as a histogram, and testing each set of data using scipy.stats.mstats.normaltest and scipy.stats.shapiro each returned clear indications that it was non-normal. Welch's T-Test is not appropriate to non-normally distributed data set and the two data sets appeared roughly the same shape when plotted as a histogram, making the Mann-Whitney U-Test an appropriate choice of test.
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 0.0499999.
1.4 The returned p-value was less than the critical p-value so the null hypothesis was rejected and it can be stated there is a statistically greater chance during rain that ridership will be higher. 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: 'tempi', 'hour', 'weekday', 'fog', 'precipi', 'wspdi', 'pressurei'; I also used ‘UNIT’ encoded as a dummy variable.
2.3 All of the features which I included affected my R2 value with at least a marginal increase. I excluded like variables as much as possible to reduce concerns of multicollinearity, with a priority for variables that were specific to the time and location (as opposed to pertaining to the entire day).
2.4 The coefficients for each of the non-dummy features included in my model are: 'tempi': 1.1199 (normalized: 8.8615), 'hour': 119.5155 (normalized: 829.0548), 'weekday': 981.0338 (normalized: 443.1419), 'fog': -355.8550 (normalized: -34.8362), 'precipi': -3089.1030 (normalized: -80.0269), 'wspdi': 12.3553 (normalized: 56.1020), 'pressurei': -521.7616 (normalized: -71.9621).
2.5 My model returned an R2 value of 0.483. Correcting negative predictions to zero and recalculating yielded an R2 value of 0.496.
2.6 R2 values are equal to 1 minus the ratio of residual variability, which is created by comparing the variability of the original data set and the predictions. An R2 value of 0.496 means that my model explains 49.6% of the observed variability, which leaves a residual variability of 50.4%. Plotting my model’s residuals as a histogram shows the values to be normally distributed and with relatively short tails. This good form appears to be primarily due to the dummy-variables ‘UNIT’, which seem to dominate the model in contributed R2 values as well. Removing ‘UNIT’ from the model decreases the R2 value to 0.108 and causes the residuals histogram to become non-normally distributed, suggesting that for the majority of the variables a linear model is not appropriate.

Section 3. Visualization

3.1 (Legends did not appear when I was using ggplot with colors, which I’m reading is a glitch. Color codes are listed in the figure description.)

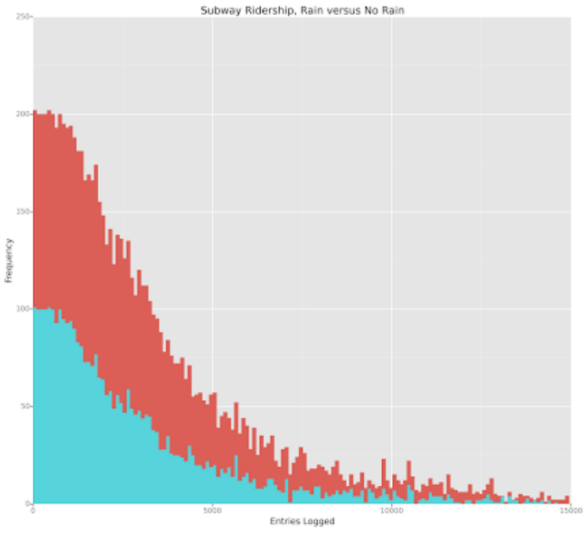

Figure 1. Subway Ridership, Rain versus No Rain: Red = Rain and Blue = No Rain - This graph portrays how that each count of rider entries was often logged more frequently during rain. This graph portrays a pronounced relationship between frequency of counts and rain. It's important to note that the counts of entries extend much farther out along the x axis - where that the 'entries logged' are higher values and more impactful on mean ridership. The end result is actually a tiny *decrease* in mean ridership during rain.

3.2

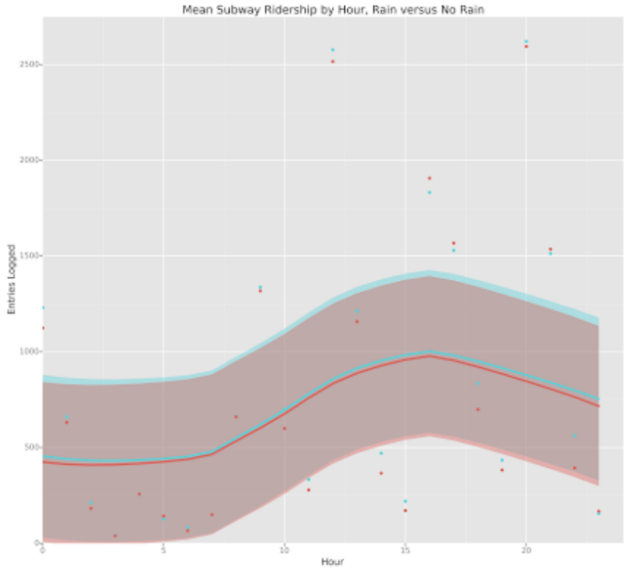

Figure 1. Mean Subway Ridership by Hour, Rain versus No Rain: Red = Rain and Blue = No Rain - This graph offers two representations of ridership throughout the average day, one for rain and one for no rain. It illustrates the small but consistent overall decreased ridership during rain, despite the more frequently logged values in rain illustrated in the previous graph.

Section 4. Conclusion

4.1 According to this dataset, slightly less people ride the subway on days 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. Under the presupposition that 'rain' and 'precipi' are generalizable to one another, my linear regression based model reinforces the U-test conclusion. According to my linear regression model, about 3000 less people ride the subway for each inch of rainfall.

Section 5. Reflection

5.1 The improved dataset and its additional variables accounted for most of my concerns about the dataset. The only potential concern I'm still noticing is in the variable

‘ENTRIESn_Hourly’, as in its description it’s stated that it ‘occasionally resets to 0.’ This is very vague, and as an analyst I’m not certain of how often exit logs are recounted across data points.

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

A linear regression based model seemed particularly not optimal to the weather related variables versus ridership, and the linear model seemed only just serviceable to ‘hour’, ‘weekday’ and ‘UNIT’ versus ridership. All except one of the regression models based off individual weathers in my model returned non-normal residuals (‘tempi’ was the only weather variable which appeared to have normally distributed residuals), most of them had a different histogram shape in exploratory comparisons against ‘ENTRIESn_Hourly’, and all of the weather variables in their own regression model returned R2 values less than 0.1. These things seem to hint that a different type of model would have been more successful.

Finally, adding the ‘UNIT’ dummy variables to the explanatory variables causes statsmodels.OLS.summar() to raise the following warning: “The smallest eigenvalue is 6.83e-27. This might indicate that there are strong multicollinearity problems or that the design matrix is singular.” This seems to be a reminder that all data points are perfectly accounted for between the dummy-variables. Given the comparatively huge contribution of the ‘UNIT’ dummy variables to the final R2 value, this could potentially cast the entire model into question. I’m still working to understand if my final R2 value is being inflated because of multicollinearity, but my current inkling is that there are so many dummy variables being created from ‘UNIT’ that multicollinearity shouldn’t be a problem.