Project 2: Analyzing the NYC Subway Dataset

Christine Stoller

September 30, 2015

0 References

NIST/SEMATECH e-Handbook of Statistical Methods: Are the model residuals well-behaved?
http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm

pandas 0.16.2 documentation: pandas.get_dummies.
http://pandas.pydata.org/pandas-docs/stable/generated/pandas.get_dummies.html

SciPy: scipy.stats.mannwhitneyu.
http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html

The Minitab Blog: How to Interpret Regression Analysis Results: P-values and Coefficients. http://blog.minitab.com/blog/adventures-in-statistics/how-to-interpret-regression-analysis-results-p-values-and-coefficients

Udacity Discussion Forums: Lesson4Problem1
https://discussions.udacity.com/t/lesson4problem1/15071

Udacity Discussion Forums: Mann-Whitney U Test on improved dataset yields p=NaN? https://discussions.udacity.com/t/mann-whitney-u-test-on-improved-dataset-yields-p-nan/4470

Wikipedia: Dummy variable (statistics).
https://en.wikipedia.org/wiki/Dummy_variable_(statistics)

Wikipedia: Mann-Whitney U-test.
https://en.wikipedia.org/wiki/Ordinary_least_squares

Yhat: ggplot for python. http://blog.yhathq.com/posts/ggplot-for-python.html

1 Statistical Test

For this project, we investigate ridership on the New York City subway, including the effects of weather patterns on ridership. We will examine subway data from the New York City Metropolitan Transit Authority as well as weather data from Weather Underground. The data under investigation is dated from May 1-30, 2011. We will use hourly entries as a measure of ridership.

1.1 Which statistical test did you use to analyze the NYC subway data? Did you use a one-tail or a two-tail p value? What is the null hypothesis? What is your p-critical value?

We apply a two-tailed Mann-Whitney U-test with a p-critical value of 0.05 to analyze the New York City subway and weather data. Our null hypothesis is that if we draw randomly from the set of entry numbers on rainy days and then from the set of entry numbers on non-rainy days, it is equally likely that higher values will come from the first set as from the second set. The alternative hypothesis for this test is that the likelihoods in each case are different; that is, that rain impacts ridership on the New York City subway.

1.2 Why is this statistical test applicable to the dataset? In particular, consider the assumptions that the test is making about the distribution of ridership in the two samples.

We note that the number of hourly entries to the subway is not normally distributed. For this reason, we use the Mann-Whitney U-test because it is a non-parametric test, meaning it does not require the assumption that the dataset has any particular probability distribution.

1.3 What results did you get from this statistical test? These should include the following numerical values: p-values, as well as the means for each of the two samples under test.

The mean number of entries on days with rain is approximately 1105.45, and the mean number of entries on days with no rain is approximately 1090.28. The U-statistic for this test is $U = 1924409167.0$ and the corresponding two-sided p-value is 0.05.

1.4 What is the significance and interpretation of these results?

Based on the results of our statistical test, we may reject the null hypothesis with a confidence level of 0.05. We conclude that one of these sets has a higher probability of yielding higher values than the other; that is, that there is a statistically significant difference in distributions of the number of entries between rainy and non-rainy days.

2 Linear Regression

We construct a model that will predict ridership on the New York City subway by way of predicting the number of hourly entries, ENTRIESn_hourly.

2.1 What approach did you use to compute the coefficients θ and produce predictions for ENTRIESn_hourly in your regression model?

To do this, we use the ordinary least squares method of linear regression, employing the OLS StatsModels library.

2.2 What features did you use in your model? Did you use any dummy variables as part of your features?

In the model, we use Hour and UNIT as dummy variables, and as non-dummy variables, we use the following:

- rain – 0 if it did not rain, 1 if it rained that day.
- meantempi – Daily average temperature in Fahrenheit.
- meanwindspdi – Daily average windspeed in mph.
- WEEKDAY – 0 for Saturday/Sunday, 1 for Monday-Friday.

2.3 Why did you select these features in your model?

We select features by using reasoning about human behavior and by comparing the effectiveness of various models by computing the coefficient of determination for each model.

The number of entries is likely to depend on the location of the station, and so we use UNIT (a proxy for subway station) as a feature in our model. The variable Hour would be helpful in predicting subway ridership because subway traffic varies greatly depending on time of day. Both UNIT and Hour were used as dummy variables, as we do not want any numerical value from these variables to be interpreted mathematically in our model.

Furthermore, entry counts are likely to depend on whether the day is a weekday or a weekend day. Many New Yorkers (and visitors) use the subway differently on the weekends than they do during the work week.

Finally, while many commuters in New York City likely have a typical routine when it comes to riding the subway, it seems reasonable that averse weather conditions - namely rain, extreme temperatures, and high windspeed - would increase the number of subway riders. Because we determined that rain impacted the number of

entries, we include rain as a variable. We also reason that mean temperature and mean windspeed would affect the number of entries. When all three of these features are included in the model, the coefficient of determination increases slightly compared with using only UNIT, Hour, and WEEKDAY.

2.4 What are the parameters of the non-dummy features in your linear regression model?

The parameters in the resulting model are as follows:

Feature	θ **value**
constant	1429.27
rain	-76.46
meantempi	-6.56
meanwindspdi	-7.83
WEEKDAY	566.01

2.5 What is your model's R^2 (coefficient of determination) value?

For this model, the value of R^2 is approximately 0.513.

2.6 What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?

The value of R^2 tells us how much of the variability in the data is explained by the features included in the model. In this case, only about 51.3% of the variability in hourly ridership of the subway is explained by our linear regression model.

Let's examine another characteristic of our model to further investigate goodness of fit. When we calculate the residuals by subtracting the actual number of hourly entries from the corresponding predicted values, we see that the range of the residuals is approximately $[-39254.49, 12997.56]$. This means that some of our predicted values were off by a very large amount.

On the following page, Figure 1 shows that the majority of the residuals are relatively close to 0, but the long tails in the distribution is cause for great concern. While it appears that not many of the residuals reach the extremes of the range, we can see that the visible portion of the tails in the plot extends has a range close to $[-10000, 10000]$, which is still too wide, particularly considering that the mean number of hourly entries is about 1095.

Given the dataset and task at hand, this result indicates that our model has room for improvement. However, we cannot expect weather, location, and timing to explain all the variability in subway ridership; there are many more factors to consider for a better overall model.

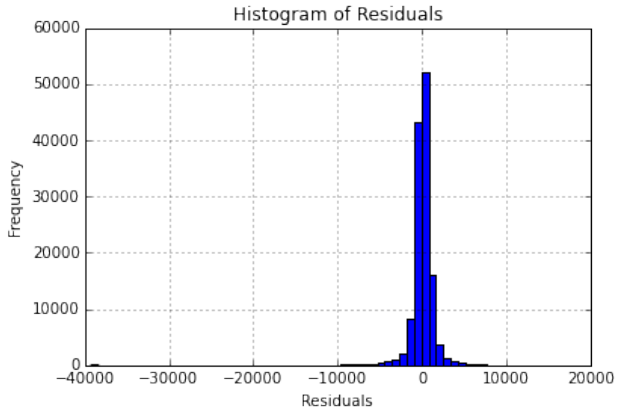

Figure 1: Distribution of residuals for linear regression model

3 Visualization

We now examine some visualizations of the relationships among variables in the dataset.

3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

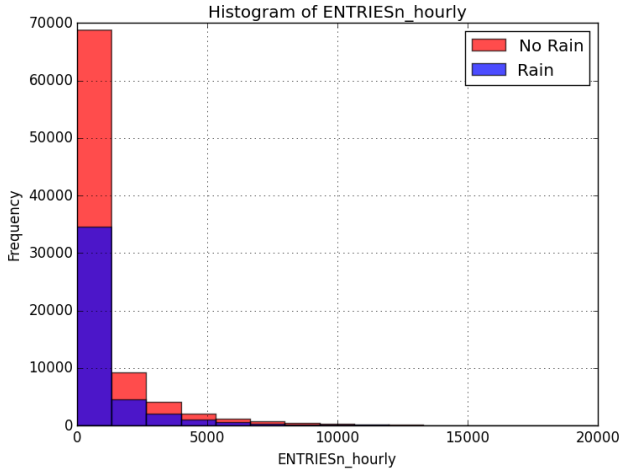

Figure 2: Distribution of hourly entries, Rain vs. No Rain

In Figure 2 above, we can see that the distribution of hourly entry counts is not normal for either sample. We observe that the data is positively skewed in both samples, and that there appear to be many more data points in the No Rain sample than in the Rain sample.

3.2 One visualization can be more freeform.

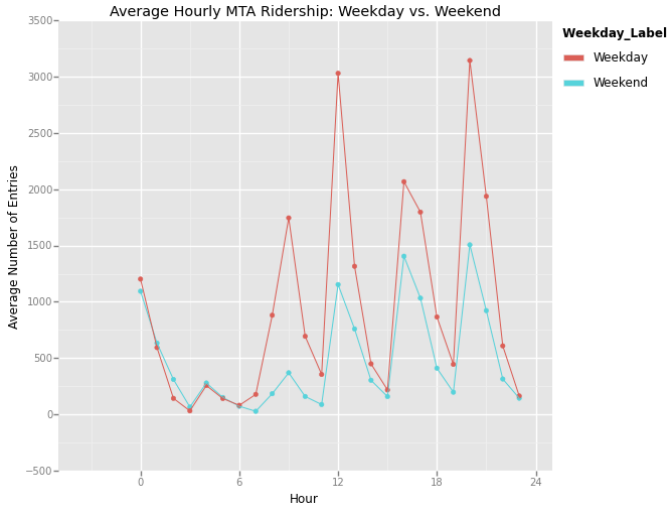

Figure 3: Average number of hourly entries on weekdays vs. weekends

Figure 3 depicts the average number of entries per hour on weekdays versus on weekends. We observe that the overall pattern of ridership is similar throughout the week, but that the average number of riders is much greater during peak times during the work week than it is on the weekends.

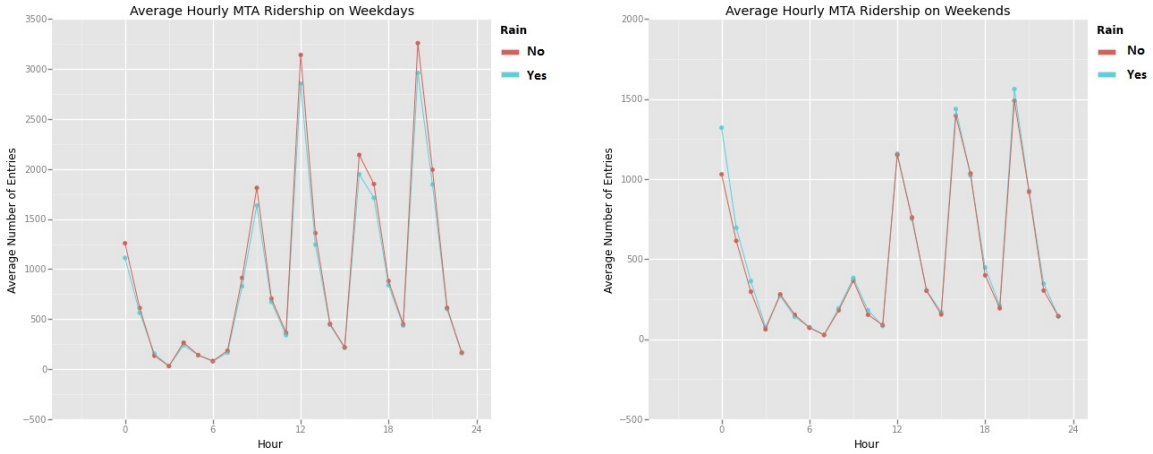

Figure 4: Average number of hourly entries with rain data (note the differing scales)

Figure 4 breaks down the data depicted in Figure 3 by splitting it into two plots, one for weekdays and one for weekends, and comparing ridership when it rained to when it did not rain. In the plot on the right, we can see that on the weekends ridership tends to be about the same or a bit higher when it is raining. However, the plot on the left shows that during the week, the average number of subway riders actually tends to be slightly *lower* when it is raining.

4 Conclusion

4.1 From your analysis and interpretation of the data, do more people ride the NYC subway when it is raining or when it is not raining?

Based on the above investigation, it is unclear whether more people ride the subway when it rains in New York City.

4.2 What analyses lead you to this conclusion?

Let's first consider the statistical test we implemented in Section 1. We found that, at a confidence level of 0.05, there is a statistically significant difference in ridership between days when it rains and days when it does not rain. Our test did not indicate under which weather condition ridership is actually higher, and so we examine the mean number of hourly entries in each case. As stated earlier, the mean number of entries on days with rain is roughly 1,105, and on days with no rain, it is roughly 1,090. The difference of 15 riders is rather small considering the overall number of riders. Furthermore, the p-value obtained from the test just barely passes below the p-critical value, which means that our results are on the border between retaining or rejecting the null hypothesis that there is no difference. This behavior in the data invites further investigation.

Next, we consider the linear regression model constructed in Section 2. While we noted that including rain as a feature in our model improved the overall effectiveness, the gain is very small. To compare, let's examine the R^2 values rounded to 4 decimal places. In our original model, R^2 is about 0.5129. Removing rain from the features to obtain a new model, we see that R^2 decreases to about 0.5127. This represents in a loss in effectiveness of only about 0.02%, suggesting that the impact of rain on subway ridership may not be significant after all.

Finally, we take into account the information from the line plots in Section 3. As seen in Figure 4, ridership appears to be somewhat lower when it is raining during the week. From Figure 3, we know that average ridership is much higher during the week. Together, this visual information strengthens the argument that there is more to the relationship between rain and ridership than can be explained by what appeared at first to be a fairly straightforward statistical test.

5 Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis, including the dataset and the analysis, such as the linear regression model or statistical test.

A shortcoming of the dataset is the limited time period of thirty consecutive days in the spring. The weather in New York City varies with the season, and much of this variation is not included in the dataset. Another is that the weather data in the dataset is not specific to each subway station or to time of day. In other words, our data does not tell us whether it rained near a particular stop, only that it rained somewhere in the city sometime during the day. This issue in particular highlights a point of weakness in the statistical test we performed, which is that the Mann-Whitney U-test only accounts for two variables (rain and hourly entries), ignoring the variation that is explained by other factors.

Our linear regression model does account for more factors, but it still has its drawbacks. One of these drawbacks is that the pool of features in the dataset is limited to turnstile and weather data. Even after comparing many combinations of features to obtain the reported model, almost half of the variability in ridership is left unexplained. This is not surprising, given that the scope of this project excludes social and economic factors like holidays, events, the price of fuel, the number of tourists in the city, and so on.