Analyzing the NYC Subway Dataset

by Ian Edington

Section 0. References

A list of references used for this project:

A. https://en.wikipedia.org/wiki/Mann–Whitney_U_test

B. https://statistics.laerd.com/spss-tutorials/mann-whitney-u-test-using-spss-statistics.php

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

D. http://docs.scipy.org/doc/numpy/reference/generated/numpy.mean.html

E. http://docs.scipy.org/doc/numpy/reference/generated/numpy.sum.html

F. http://pandas.pydata.org/pandas-docs/stable/visualization.html#histograms

G. http://pandas.pydata.org/pandas-docs/stable/groupby.html

H. http://pypi.python.org/pypi/ggplot/

I. http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

J. http://stackoverflow.com/questions/12190874/pandas-sampling-a-dataframe

K. http://stackoverflow.com/questions/19711943/pandas-dataframe-to-dictionary-value

L. http://stackoverflow.com/questions/7001606/json-serialize-a-dictionary-with-tuples-as-key

M. http://statsmodels.sourceforge.net/0.5.0/generated/statsmodels.regression.linear_model.OLS.html

N. http://statsmodels.sourceforge.net/0.5.0/generated/statsmodels.regression.linear_model.OLS.fit.html

O. http://statsmodels.sourceforge.net/0.5.0/generated/statsmodels.regression.linear_model.RegressionResults.html

P. http://wiki.scipy.org/Cookbook/Matplotlib/BarCharts

Q. http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm

Section 1. Statistical Test

1.1 Which statistical test did you use to analyze the NYC subway data?

The Mann-Whitney U test was used to determine if ridership on days where there was rainfall was significantly different than ridership on days where there was no rainfall. This same test was used to determine if ridership on days where there was fog was significantly different than ridership on days where there was no fog.

Did you use a one-tail or a two-tail P value?

I used two-tailed P values in order to determine directionality.

What is the null hypothesis?

Ridership does not change based on if it's raining or not that day.

$$H_0: \mu_{rain} = \mu_{no\ rain} \quad \& \quad H_0: \mu_{fog} = \mu_{no\ fog}$$

What is your p-critical value?

0.025

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.

This test is applicable because the assumptions made by the Mann-Whitney U test are true about this data and because the question we were asking can be answered by this test.

Based on Wikipedia's article[A] these are the assumptions made by the Mann-Whitney U test and evidence that this data set conforms to these assumptions.

1. All the observations from both groups are independent of each other:

   We assume that the ridership of one hour is not based on the ridership of the previous hour or the previous day. This is a reasonable assumption since ***

2. The responses are at least ordinal:

   The dependent variable ENTRIESn_hourly is a continuous range of positive whole numbers.

3. The null hypothesis H_0 is "The distributions of both populations are equal"
4. The alternative hypothesis H_1 is "the probability of an observation from the population X exceeding an observation from the second population Y is different from the probability of an observation from Y exceeding an observation from X : P(X>Y)≠P(Y>X)."

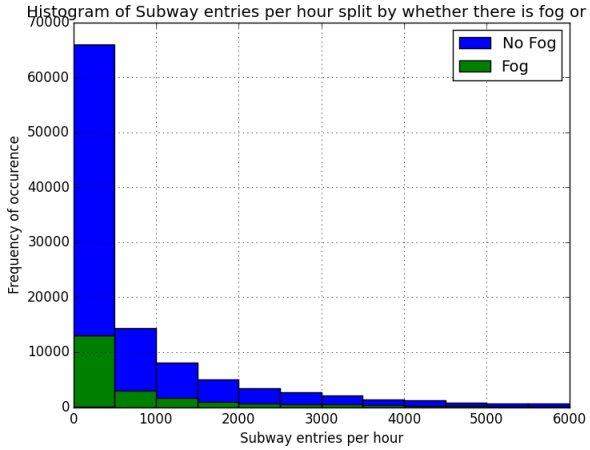
From Laerd Statistics[B] we have two additional assumptions:

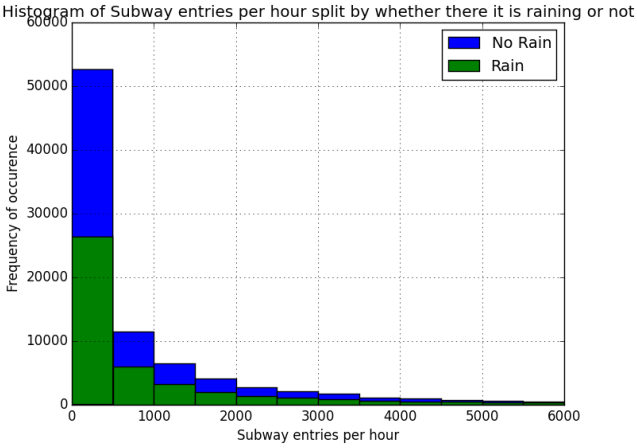
5. You have one independent variable that consists of two categorical, independent groups:

   This is true for both rain and fog variables: 1 and 0

6. You must determine whether the distribution of scores for both groups of your independent variable have the same shape or a different shape.

   We can see based on the histograms for both rain and fog that the distributions have the same shape.





1.3 What results did you get from this statistical test? P-values and the means for each of the two samples under test.

	P-Value	Mean of good weather days	Mean of rainy of fogy days
Rain	0.0249999	1090	1105
Fog	0.0000061	1083	1155

1.4 What is the significance and interpretation of these results?

For the case of both Rain and Fog we reject the null hypothesis. This means that when it is raining the NYC subway is likely to have a hirer ridership than when it is not raining. Likewise for fog.

Section 2. Linear Regression

2.1 What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model:

OLS using Statsmodels

2.2 What features (input variables) did you use in your model?

I used UNIT, Hour, day_of_week as the input variables of my model.

Did you use any dummy variables as part of your features?

For UNIT, Hour, day_of_week I used dummy variables based on the mean of the Entries.

2.3 Why did you select these features in your model?

I used these features because they had the greatest +ve impact on the r squared value. To calculate the r squared value I used parameters generated by the OLS model to make predictions of a set of test data, then compared the predictions to the recorded data.

Using only UNIT, Hour and day_of_week I got an r-squared value of 0.49370326835188771 compared to 0.49469115125765351 for 8 features (UNIT_means, Hour_means, day_of_week_means, maxtempi, precipi, fog, rain and maxdewpti).

I felt that the added features didn't add enough benefit to justify including them.

2.4 What are the parameters (also known as "coefficients" or "weights") of the non-dummy features in your linear regression model?

I made a transformation to all my features before implementing OLS by calculating the mean entries for each of these features.

Intercept: -1817.4548075398768
day_of_week_means: 0.9687637785441505

UNIT_means: 0.9427523353608755
Hour_means: 0.7492225446047587

2.5 What is your model's R2 (coefficients of determination) value?

R2 value: 0.49370326835188771

2.6 What does this R2 value mean for the goodness of fit for your regression model?

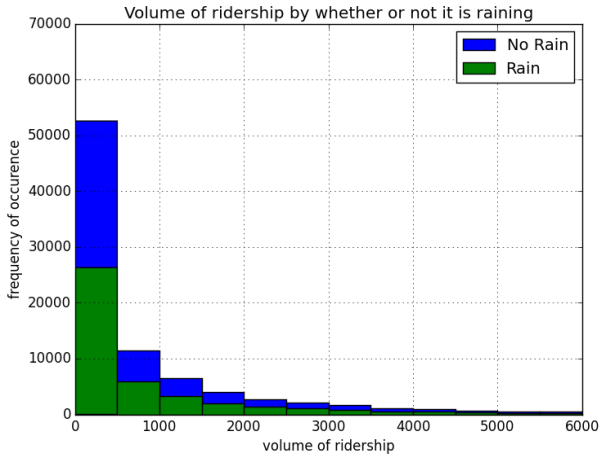
R2 on its own isn't a good indicator of a models fit. More graphical methods need to be used to determine if it is a good fit or not. Specifically, looking at the residuals.

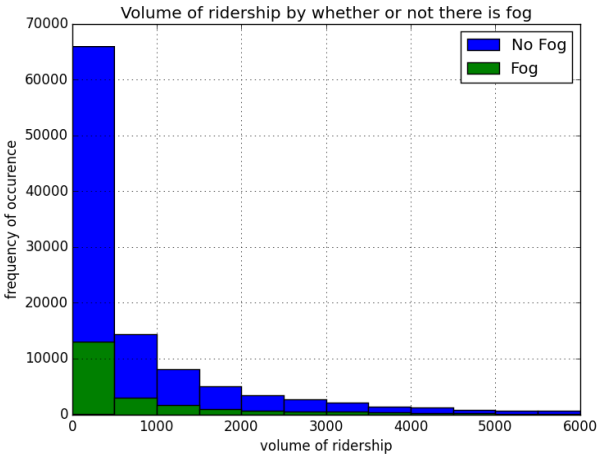
Do you think this linear model to predict ridership is appropriate for this dataset, given this R2 value?

R2 doesn't tell us if we have the right model, only how much our data varies from the model. If a large portion of the data can't be explained by a model the R2 will always be low wether or not we have a good fit. However, based on graphs of the data I don't think a linear model is the best way to predict ridership for most features in this dataset because most of the data appears to be non-linear.

Section 3. Visualization

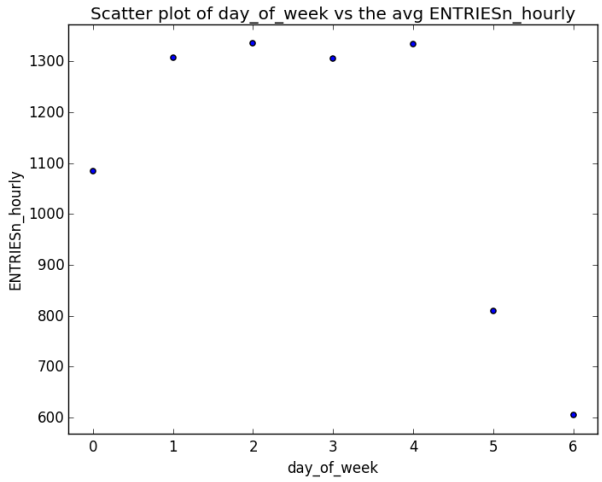
3.1 histogram of ridership for rainy/foggy days and non rainy/foggy days

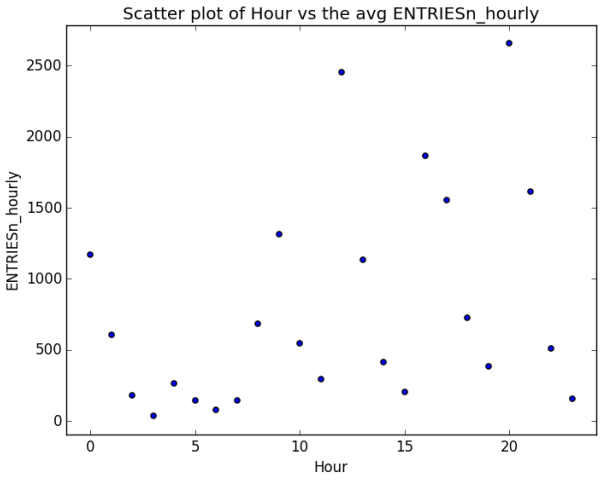




In these histograms we can see that the datas subset of Rain==0 (no rain) and Rain==1 (rain) have a similar distribution. This is important because one of the Mann-Whitney U-test assumptions is that the two sets it compares have similar distributions. This is true for Fog an No Fog as well.

3.2 scatter plot of Ridership by day-of-week and Ridership by hour of day





From these scatter plots we can see that there are peak times during the day and peak days during the week. These large correlations where why it was possible to more easily predict ridership based on them.

Section 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?

> We are 95% confident that more people ride the NYC subway when it is raining then when it is not raining. This is also true for when there is fog.

4.2 What analyses lead you to this conclusion? You should use results from both your statistical tests and your linear regression to support your analysis.

> From the Mann-Whitney U-test we know that more people ride the subway when it is raining. However, when we started the regression model rain fall was a very poor indicator of ridership. My understanding is that even though we are 95% confident that rain does have an effect on ridership the effect is so small that it isn't useful for predicting ridership.

Section 5. Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis, including: 1. Dataset, 2. Analysis, such as the linear regression model or statistical test.

Some potential shortcomings of this analysis is that the data is strongly skewed to the right. Meant that the means of the data were not necessarily good indicators of what the data looked like. It is difficult for me to know exactly how this affected the analysis.

Another shortcoming due to analysis was feature selection. In the analysis I used the resulting R2 values in order to select which features to use, however, this is not necessarily the best way to select features. For example one features may have a higher R2 but we are not able to know them a head of time. Unfortunately I don't know enough yet about feature selection to say if I did it well.

5.2 (Optional) Do you have any other insight about the dataset that you would like to share with us?

Weather is a very poor indicator of ridership. People's habits in terms of subway use seem to be more linked to their plans for the day than by if it will be raining. As much as rain affects their plans for the day it might affect ridership, however, whether or not it rains I still need to get to work :D