Analyzing the NYC Subway Dataset

Overview

In this analysis, I am taking data from May 2011 NYC MTA Subway as provided in the Udacity 'Introduction to Data Science' class and testing the hypothesis that whether rainy days impacts ridership based on problem sets 2 to 5.

Section 1 - Statistical Test

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?

I used the Mann-Whitney U test to analyze the NYC subway data. Because it is not yet known, nor hypothesized, which data set would be higher or lower, a two-tailed test here is apt. In using the Mann-Whitney U test, the null hypothesis is that the two populations (rainy and non-rainy) are the same, or simply put, that rain has no correlation with ridership. The p-critical value used was 0.05 or 5%.

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.

As shown in Section 3.1, neither the rain or no-rain histograms are normally-distributed. As such, a non-parametric test such as Mann-Whitney U is a good fit, while a test such as Welch's two-sample t-test is not. To quantitatively capture and confirm that neither data sets are normally-distributed, a Shapiro-Wilk test could have been conducted.

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.

Mean entries with rain	Mean entries without rain	U-statistics	p-value
1105.45	1090.28	1924409167.0	0.025

1.4 What is the significance and interpretation of these results?

Comparing the means yields 1.4% more subway entries when it rains. This statistic alone is insufficient in drawing conclusion or correlation. The U-statistic has a high value, very close to the maximum value of 1937202044.0, or half the product of the number of values in each data set. A U-statistic of half the maximum would indicate that the null hypothesis is true. Of note, the p-value 0.025 satisfies the p-critical value, and the conclusion can be drawn with 95% confidence that the null hypothesis is false and that ridership is different with vs. without rain.

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:

a. Gradient descent (as implemented in exercise 3.5)
b. OLS using Statsmodels
c. Or something different?

A machine learning algorithm, batch Gradient Descent, was used to train the liner regression coefficients. I used the default values of learning rate (alpha) 0.1 and 75 iterations, and also kept the mean normalization feature scaling. The given values were sufficient in converging on a local minimum, as confirmed by plotting the cost history vs. number of iterations.

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

Features used included rain (0 or 1), precipitation, mean wind speed, hour, and mean temperature. Per the default configuration, dummy variables were introduced for features 'UNIT' (the turnstile location/identification number), which were categorical in nature. They were initialized with Boolean (0 or 1) features with prefix 'unit', and each data point would have a '1' in the feature that it "belonged" to. It did not make sense to apply linear regression to the raw 'UNIT' parameters quantitatively; however, it was important to keep track of it as there was a wide variation between different subway stops and account for it first. If this was not done, the differences between different turnstiles would mask the markedly smaller changes due to rain, precipitation, hour or temperature.

2.3 Why did you select these features in your model? We are looking for specific reasons that lead you to believe that the selected features will contribute to the predictive power of your model.

- **Your reasons might be based on intuition. For example, response for fog might be: “I decided to use fog because I thought that when it is very foggy outside people might decide to use the subway more often.”**
- **Your reasons might also be based on data exploration and experimentation, for example: “I used feature X because as soon as I included it in my model, it drastically improved my R^2 value.”**

I maintained rain, precipitation, hour and mean temperature out of experimentation; I was unable to find R^2 values that were better. Broadening the hypothesis that "people use the subway more often when it's raining" to "people use the subway more often when there bad weather outside". I also included wind speed. I saw a slight increase in my R^2 values.

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

[-4.24736210e+00, 8.60518981e+00, 4.64832083e+01, 4.64163466e+02, -3.19114921e+01, 1.08898857e+02]

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

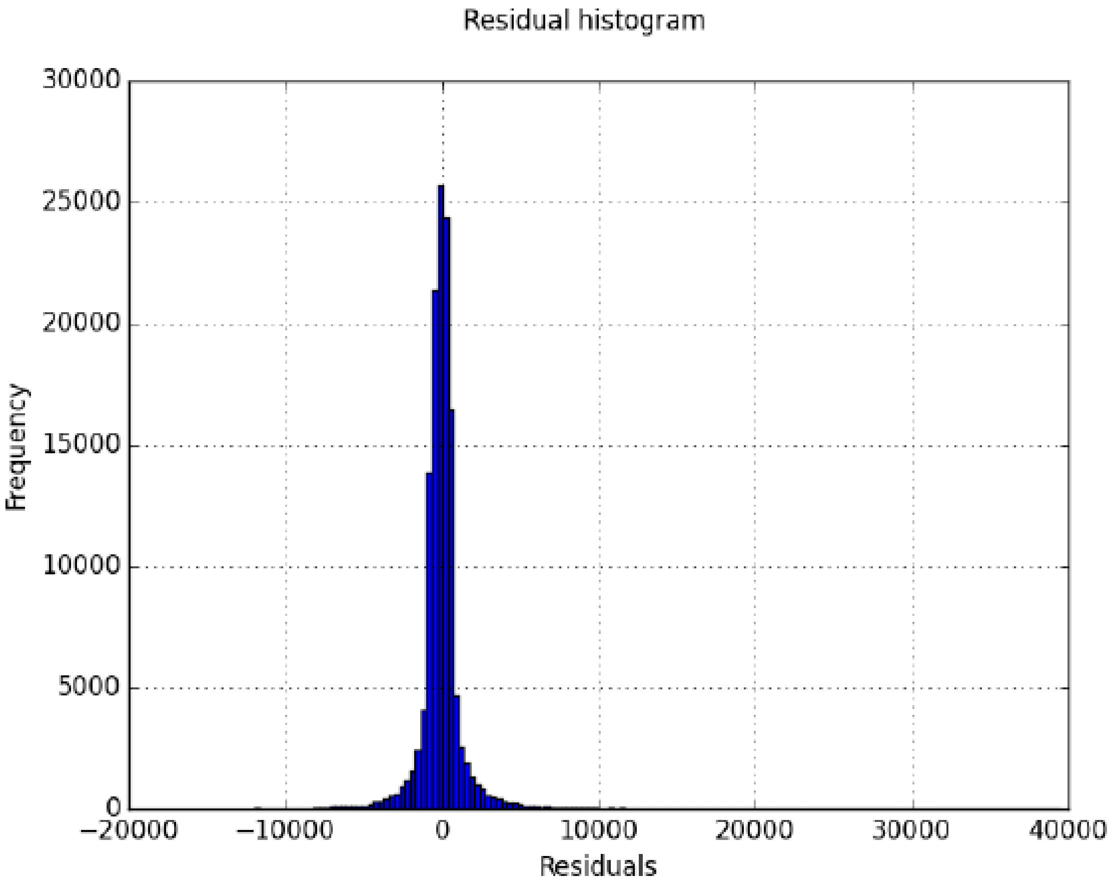
0.458375428174

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?

R^2 is essentially the percentage of variance that is explained, and is a quantitative measure of the "goodness of fit". While it only explains 45.8% of variation, I think a better metric is to plot the residuals.

To decisively conclude whether or not this model was a good fit certainly depends on the context and use case for the prediction data. If this were a use case that had safety and security concerns, it would certainly be insufficient! The residual plot shows that most of the residuals were close to 0 +/-5,000. Qualitatively, and for the objective of being able to "ballpark" ridership, the liner model is sufficient.

Further and advance study could include more features or utilize polynomial regressions. However, this might lead to significant over-fitting, and the model may fail on new data sets. In that case, regularization would be a good method to attenuate any over-fitting.



Section-3: Visualization

Please include two visualizations that show the relationships between two or more variables in the NYC subway data. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots, or histograms) or attempt to implement something more advanced if you'd like.

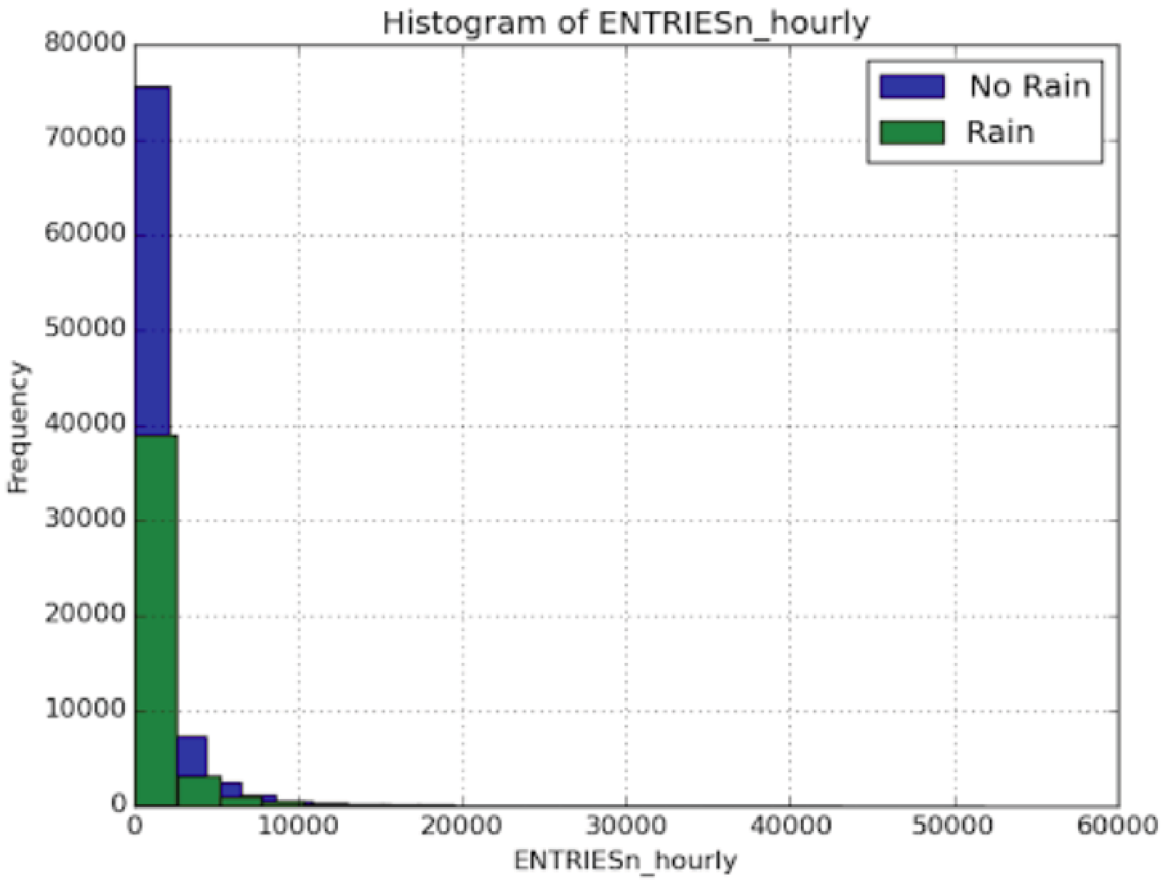
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

In this visualization below, the ENTRIESn_hourly for rainy days and non-rainy days .

Overall, the frequency of ENTRIESn_hourly for non-rainy days (blue bars) was higher than rainy days in each range.

The frequency of ENTRIESn_hourly for non-rainy days (blue bar) and rainy days (green bar) on May 2011 (NYC subway). The ENTRIESn_hourly was shown in x-axis. The frequency of

ENTRIESn_hourly was shown in y-axis. The figure was plotted using ggplot and python with bins = 20.



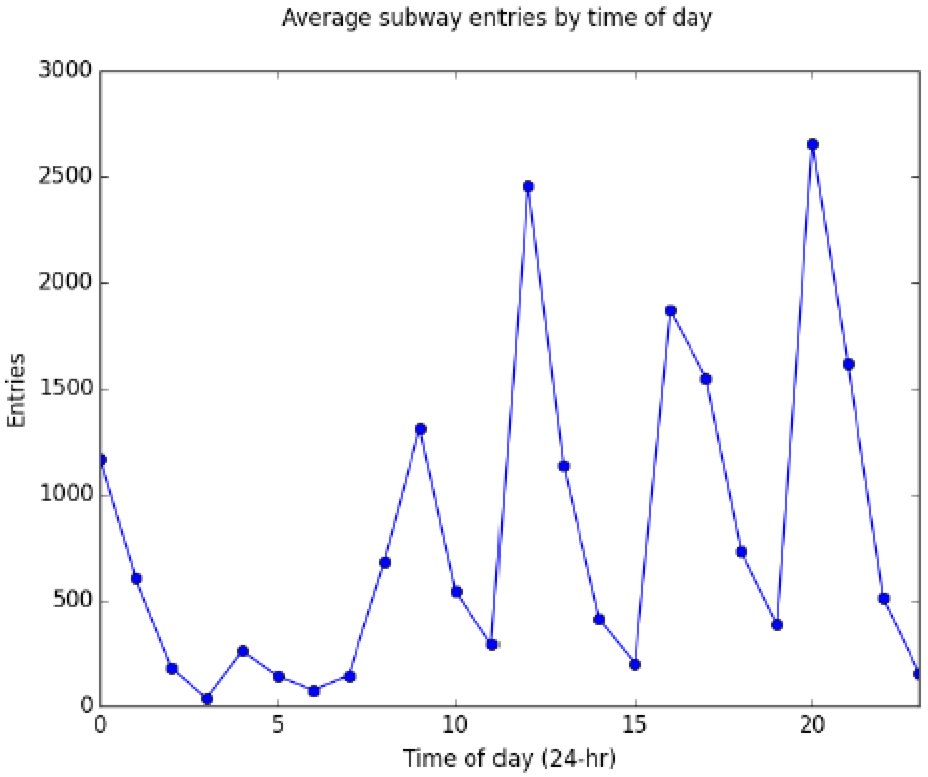
3.2 One visualization can be more freeform. Some suggestions are:

Ridership by time-of-day

By plotting the average number of subway entries at each hour, it's clear that there are several peaks throughout the day, with the most prominent ones being at noon and 8pm. Interestingly, these peaks are large than those during rush hours (8-9am and 5-6pm). It raises some interesting questions about the demographics and characteristics of NYC subway riders:

- Assuming a 9am - 5pm workday, why are more subway entries occurring at 5pm vs. 9am?
- Are more people going out for lunch (12pm) and dinner (8pm), or is that there work schedule?

Without any demographic data, it would be impossible to determine these questions from the current data set.



Section-4: Conclusion

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

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?

Particularly given the results from the Mann-Whitney U test (p-value: 0.025), we can say with a high level of certainty that more people ride the NYC subway when it is raining. It is important to note that simply looking at the means of both data sets is insufficient, due to variance. The Mann-Whitney U test is needed to quantitatively confirm that the two sets are statistically different.

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.

The positive coefficient for the rain (0 or 1) parameter indicates that the presence of rain contributes to increased ridership. This may have not been the case for all data points, with the R^2 being approximately 46%; however, the small residuals show relatively high accuracy, given our objectives. Although the means of both data sets are not that different from each other, the Mann-Whitney U test did indicate that there was a statistically significant change in ridership for rain vs. no rain day. It is conscientious to claim that rain increases subway ridership.

Section-5: Reflection

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

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.**

One immediate red flag that was presented while exploring the data was that there were more 'entries' than there were 'exits'. The only logical explanations could be that there were miscounts, or some turnstiles/stations were not included in the data set. Presumably, this would have had an equivalent effect on both rain and no-rain data sets, so for the purposes of this study, it likely had little to no effect.

A combination of increased sample size (larger data set) and normalization by location/turnstile ID could have potentially increased the confidence of both the Mann-Whitney U test and the linear regression model. As we saw from examining the 'UNIT' column, ridership varied greatly. Simply put, some stations and turnstiles were naturally more active than others. The Mann-Whitney U test did not take this into account, and only looked at the subway entry distributions for rain and no-rain. Examining how the same stations at the same day and time varied by rain could have increased the fidelity of the test.

The linear regression model was adequate for the purpose of the study, but could certainly have been improved. It's possible that the region of study had a linear relationship, but it is still an assumption and simplification. Considering the extreme, subway ridership certainly has an asymptotic limit; only so many riders can get on the subways! As mentioned in Section 2.6, the inclusion of more features or polynomial combinations could have increased the accuracy of the model. Given more data, it would have also been appropriate to split the data into a training data set (~60%), as cross-validation data set (~20%), and a testing set (~20%). This could have illuminated any errors with high variance, high bias, and any over/under-fitting.

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

I think that an interesting investigation would be to use gradient descent with logistics regression to see if one might be able to predict if it rained or not given various parameters, to include turnstile location/ID, time of day and subway entries. Intuitively, this might produce false positives or negatives on special days (e.g. holidays, city events, marathon runs, etc).

REFERENCES

- Welch's t-test
  http://en.wikipedia.org/wiki/Welch%27s_t_test

- Mann-Whitney test
  http://www.real-statistics.com/non-parametric-tests/mann-whitney-test/
  http://ocw.umb.edu/psychology/psych-270/other-materials/RelativeResourceManager.pdf
  http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html

- R2 - linear regression
  http://www.graphpad.com/guides/prism/6/curve-fitting/index.htm?r2_ameasureofgoodness_of_fitoflinearregression.htm