Analyzing the NYC Subway Dataset

Questions

Overview

This project consists of two parts. In Part 1 of the project, you should have completed the questions in Problem Sets 2, 3, and 4 in the Introduction to Data Science course.

This document addresses part 2 of the project. Please use this document as a template and answer the following questions to explain your reasoning and conclusion behind your work in the problem sets. You will attach a document with your answers to these questions as part of your final project submission.

# **Section 0. References**

<https://en.wikipedia.org/wiki/Mann–Whitney_U_test>

<https://www.coursera.org/course/ml>

[http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html#scipy.stats.mannwhitneyu](http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html" \l "scipy.stats.mannwhitneyu)

<http://blog.minitab.com/blog/adventures-in-statistics/how-to-interpret-regression-analysis-results-p-values-and-coefficients>

<http://blog.minitab.com/blog/adventures-in-statistics/why-you-need-to-check-your-residual-plots-for-regression-analysis>

# **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?

The statistical test used to analyze the NYC Subway data:  **Mann-whitney U test**

One-tail or a two-tail P value : Because it is not yet known, nor hypothesized, which data set would be higher or lower, a two-tailed test here is used.

The p-value returned by  Mann-whitney U test is one-tailed as noted here:

<http://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.mannwhitneyu.html>

In order to use a two-tailed test, the one-tailed p-value returned by Mann-whitney U must be multiplied by 2

Null hypothesis: That the two populations are the same(Rain has no correlation with ridership)

p-critical value: 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.

**Mann-Whitney U test** was used because the data is not normally distributed. When we have such kind of data, a non-parametric test(**Mann-Whitney U test**) is best suited to give us the best answer. Another benefit of **Mann-Whitney U** test is that it does not have assumptions about the distribution of the selected samples. For comparison, **Welch’s** two-sample t-test assumes the the data is normally distributed. In order to make sure that neither data set is normally distributed a **Shapiro-Wilk test** could have been conducted as well.

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.

“Good job! Your calculations are correct.  
Here's your output:  
(1105.4463767458733, 1090.278780151855, 1924409167.0, 0.024999912793489721)“

Mean entries with rain: 1105.446  
Mean entries without rain: 1090.279  
U-statistic: 1924409167.0  
p-value: 0.025

1.4 What is the significance and interpretation of these results?

The p-value is too small. Therefore, we reject the Null Hypothesis of the **Mann-whitney U test.**

**Test:**alpha = 0.05 # as set in 1.1  
# two-tailed test  
if (p \* 2) < alpha:  
   print 'Reject the Null hypothesis'  
else:  
   print 'Fail to reject null hypothesis'

**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?

**Gradient Descen**t **algorithm** was used to train the **regression coefficients**. I kept the mean normalization feature scaling , and used the default values of learning rate ( **Alpha** = 0.5, **N** = 75 iterations). The given values reached the local minimum, which was confirmed by the plot of **cost history** VS **N 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?

Rain,precipitation,mean wind speed,hour,mean temperature. Dummy variable used was UNIT( turnstile location/identification number- categorical). UNIT variable is boolean(0,1) features with prefix ‘unit’. Each data point have a 1  in the feature that it belonged to.

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.

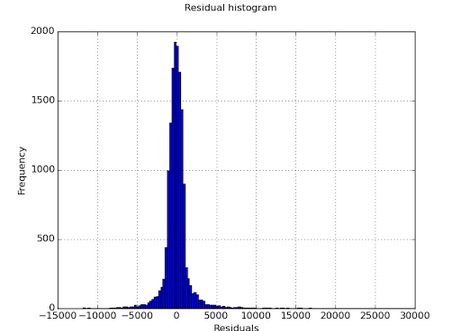
There were not better R^2 values than the one used(rain, precipitation, hour, and mean temperature). Those values were were chosen through experimentation and eventually stopped once I saw that the R^2 value did not get better. Noticed increase of r^2 value after including wind speed. I chose the wind since I tried to change the reasoning that “people use the subway more often when there is bad weather outside” than just when it is raining. However, the result was not that big.

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

Coefficients[ rain -4.24736210e+00, no rain 8.60518981e+00, precipitation 4.64832083e+01,mean wind speed 4.64163466e+02, Hour -3.19114921e+01, Temperature 1.08898857e+02]

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

Your r^2 value is 0.479247705439

2.6 What does this R2 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 R2  value?

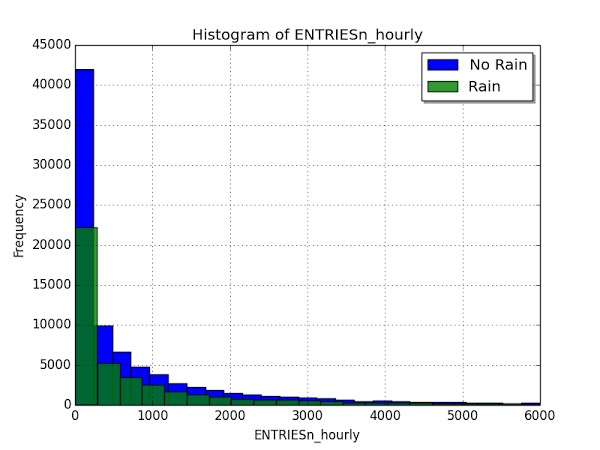
**R^2** is the the percentage of **variance** that is explained , and is measure of goodness of fit. My model explains only **~48%** of the variation.A better metric is to plot the residuals, that can be seen on the figure on the left. The long tails on the histogram suggest that there are large values that might be a reason to question the model.If this were a use case that had safety and security concerns, it would certainly be insufficient.

# **Section 3. Visualization**

Please include two visualizations that show the relationships between two or more variables in the NYC subway data.

Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

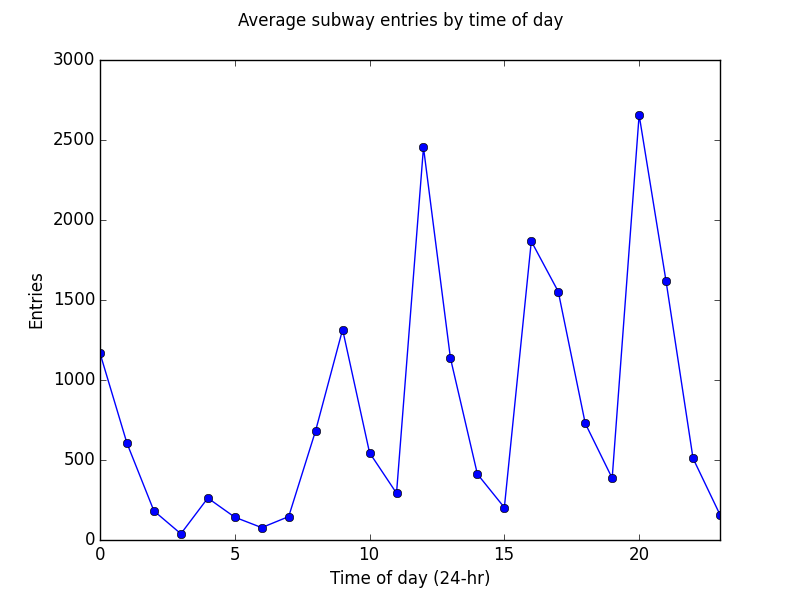
3.1 One visualization should contain two histograms: one of  ENTRIESn\_hourly for rainy days and one of ENTRIESn\_hourly for non-rainy days.



The above overlapping histogram shows that the entries over rainy and not rainy hours are not normally distributed. We cannot draw conclusion about ridership from this histogram

3.2 One visualization can be more freeform. You should feel free to implement something that we

discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like.

* 

By plotting the average entries per hour, we can see that there are times of the day when the entries are significantly more than others. The variations mostly occur from noon to 8 pm. Some peaks are larger than during peak hours(8-9 am and 5-6 pm). That might be strange but it is hard to explain without having demographic data of the area.

# **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?

By simply looking at the means of both data sets, it is is insufficient to make conclusions due to the variance.

However, Mann-Whitney U test  does confirm that the two sets are statistically different. With the results given from the **Mann-Whitney U test**, we can conclude that more people ride the NYC subway when it is raining.

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.

Since positive coefficient of the rain parameter is observed, it can be concluded that the presence of rain to some extend drives the increase of ridership. The ~48% **R^2** leads toward the conclusion that the above statement might not have been the case of all data points.However, the small residual is has relatively high accuracy, having in mind the objectives. Overall, even though the means of the two datasets are not significantly different, the **Mann-Whitney U test** does show statistically significant change in number of people riding the NYC subway when it is raining vs when it is not. And I can conclude that the the rain drives the ridership of the subway.

# **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:

* Dataset,
* Analysis, such as the linear regression model or statistical test.

While looking at the data, I noticed that there were markedly more entries than there were exits. The reason for that could be that there were miscounts, or some /stations were out of the data set. This would have effect on both rain and no-rain data sets,so it likely had little to no effect on this study.

A combination of bigger sample size and normalization by location ID could have potentially increased the confidence of both the **Mann-Whitney U** test and the **linear regression model**.It could be seen when we examined the ‘UNIT’ column, ridership varied greatly. It could be seen that some stations and turnstiles were naturally more active than others. The **Mann-Whitney U test** only looked at the subway entry distributions for rain and no-rain and did not take take in consideration the activeness. 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 could have been improved, however,was adequate for the purpose of the study.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%), a 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.

When plotted the residuals, it could be seen that there were values that were extreme and would not be appropriate for linear model.

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

Not at this moment