Huy Do

Analyzing the NYC Subway Dataset

Section 0: References

http://www.statisticssolutions.com/mann-whitney-u-test/

http://www.ftpress.com/articles/article.aspx?p=2248639&seqNum=6

http://www.ehow.com/info_8562780_disadvantages-linear-regression.html

http://www.statsoft.com/Textbook/Multiple-Regression#cresidual

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?

To analyze the NYC subway data, I used the two-tail Mann Whitney U test on 2 independent variables (ENTRIESn_hourly when it rains and ENTRIESn_hourly when it doesn't rain).

Null hypothesis: the distributions of the 2 groups (the number of subway entries on rainy days and the number of subway entries on non-rainy days) are equal

P-critical value: 0.05

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.

The Mann Whitney U test is applicable to this dataset since the data is not normally distributed and thus, is not appropriate for the t-test. The Mann Whitney U test is used to compare 2 population means that come from the same population. In this case, the 2 population means are the average number of hourly entries on rainy days and average number of hourly entries on non-rainy days.

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.

P value: 0.05

Mean of entries on rainy days: 1105.45

Mean of entries on non-rainy days: 1090.28

U = 1924409167

1.4 What is the significance and interpretation of these results?

Since p value <= p critical, the result is statistically significant. The distribution of the number of subway entries is statistically different between rainy days and non-rainy days. The mean of entries on rainy days (1105) is higher than the mean of entries on non-rainy days (1090).

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?

I used Gradient descent in Problem Set 3.5 to compute the coefficients theta and produce predictions for ENTRIESn_hourly.

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

The input variables used were: rain, Hour, maxtempi, and mintempi. The dummy variables are generated for UNIT.

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.

I chose these features based on both intuition and experimentation. I assumed that people would rather take the train than walk when weather conditions are unfavorable. After that, I play with several features to find the best R value.

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

rain	2.396
hour	4.679
maxtempi	2.619
mintempi	-9.284

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

R2 = 0.464

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?

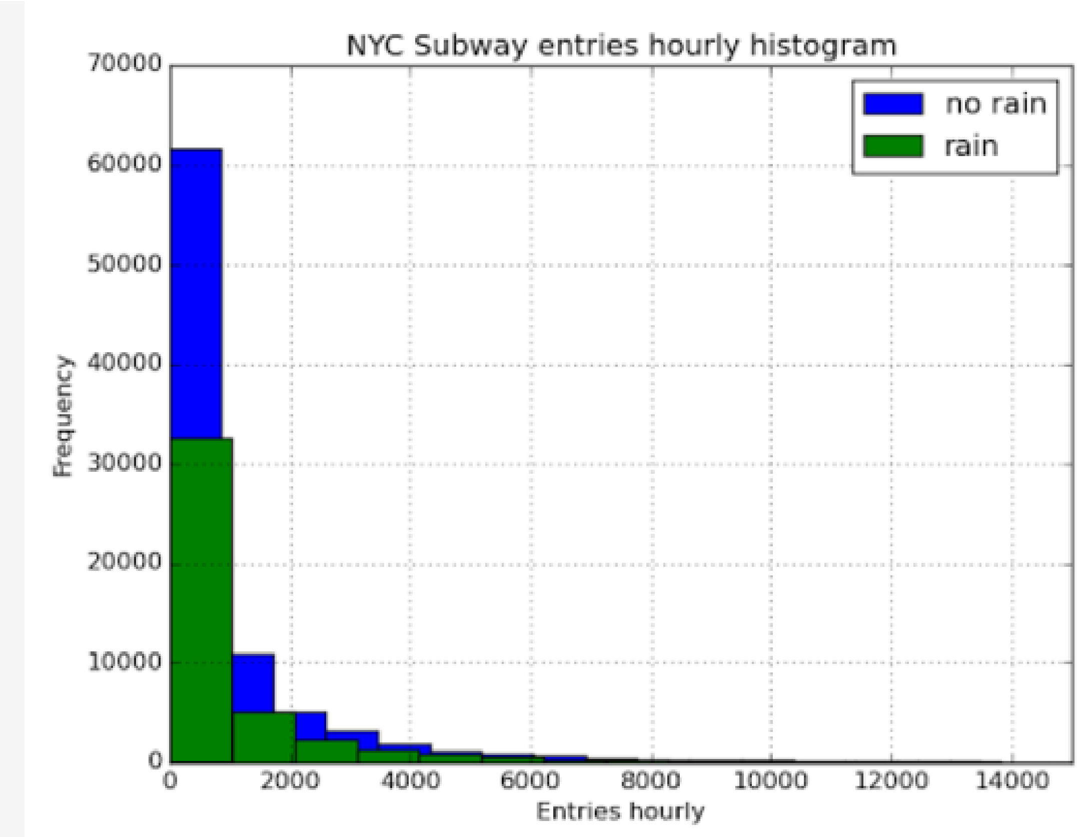
The model is considered better if R squared is closer to 1. As R squared is currently at 0.46, the model is not a good predictor of subway ridership based on weather data. The value of R squared shows that the variability of the Y values (number of entries hourly) around the regression line is 0.54 times the original variance. The regression model has explained 46% of the original variability, and still has 54% residual variability.

Adding more data that accounts for factors besides weather conditions (e.g. location of subway station, holiday, major events in NY, working schedule of different groups of NY residents, etc…) may improve the accuracy of the model to predict subway ridership.

Section 3: Visualization

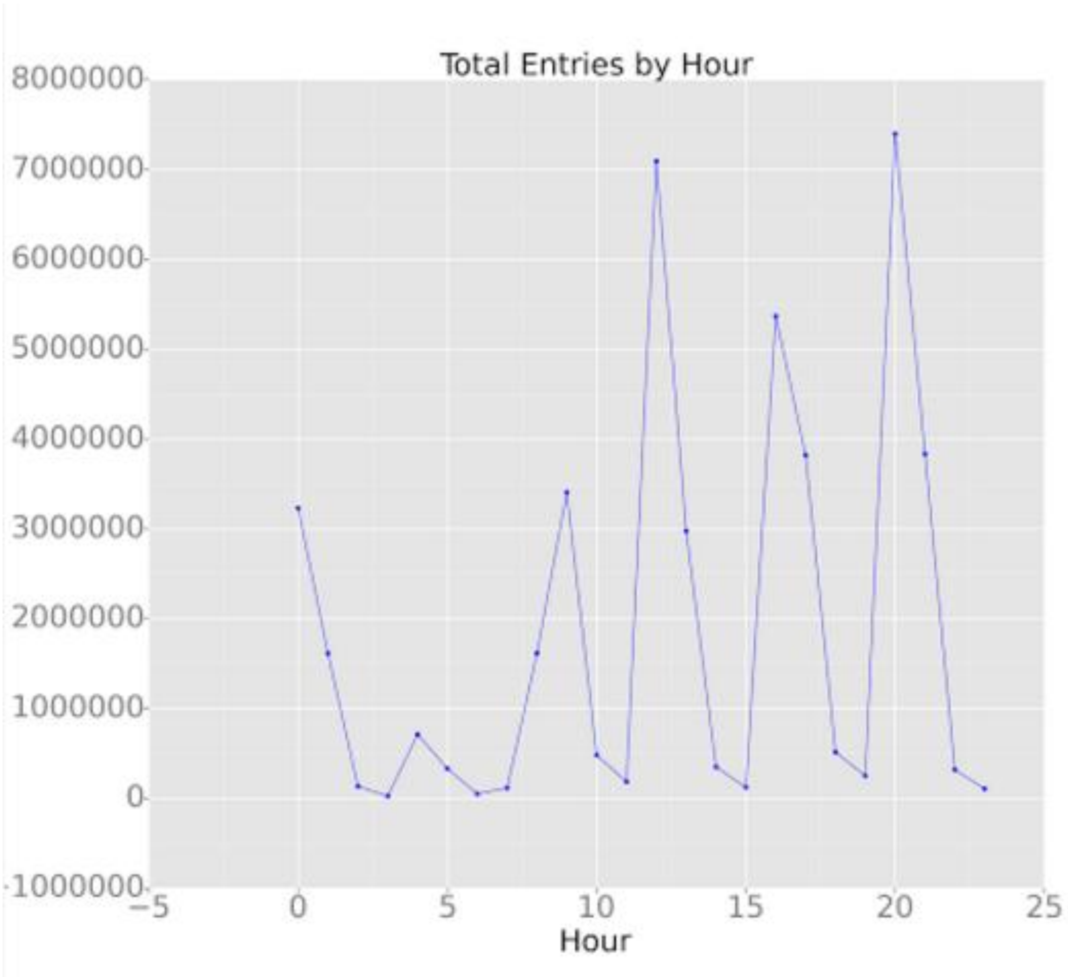
Please include two visualizations that show the relationships between two or more variables in the NYC subway data.

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 histogram shows that the data for subway entries on rainy days and non-rainy days are not normally distributed. Overall, the total entries for non-rainy days seem to be higher than the total entries for rainy days.

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. Some suggestions are:



The line graph plots total subways entries by hour. It shows that subway entries are significantly higher around certain times of the day.

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?

The results from statistically analysis and regression analysis show that people tend to ride the NYC subway more 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.

The Mann Whitney U test shows that the number of entries is statistically different between rainy days and non-rainy days. In addition, the mean of entries is slightly higher on rainy days (1105) than the mean of entries on non-rainy days (1090). The p value (0.05) is roughly equal to the p critical value (0.05)

The linear regression model has a positive coefficient for rain (2.396). This means that the number of subway entries is likely to increase when it's raining.

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.

Potential shortcomings of the methods of my analysis:

- Dataset: the dataset is small. It only has the month of May in 2011, when it rained 10 out of 30 days. Because of the limit of the server for the grader, I only test a portion of the dataset for my model. A better dataset to test the effect of weather on subway entries may include a larger time period and limit to a single subway station (to control the effect of subway location on subway entries)
- Mann Whitney U test: while the test result is statistically significant, the p value is barely smaller than the target p critical value. This is not a strong result.
- The use of dummy data: dummy data is generated for UNIT so that it can be used in the linear regression model. While it does allow the model to apply a weight for each station, it is not ideal. Perhaps a separate model can be used with a larger dataset to measure the amount of traffic at each subway stations. Then model can compute a proper weight for each station to be used to predict subway entries.
- Linear regression model: linear regression assumes that there is only a linear relationship between dependent and independent variables. With weather data, this may not be appropriate. For example, a small increase/decrease in temperature may have no effect on subway usage. I would assume that subway entries significantly increase at the extremes (i.e. when it's too hot or too cold to walk).