Section 0: References

Whitley E, Ball J. Statistics review 6: Nonparametric methods. *Critical Care*. 2002;6(6):509-513. Accessed by: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC153434/

Bewick V, Cheek L, Ball J. Statistics review 10: Further nonparametric methods. *Critical Care*. 2004;8(3):196-199. Accessed by: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC468904/

"Hypothesis testing and p-values" Khan Academy. https://www.khanacademy.org/math/probability/statistics-inferential/hypothesis-testing/v/hypothesis-testing-and-p-values

"FAQ: What are the differences between one-tailed and two-tailed tests" http://www.ats.ucla.edu/stat/mult_pkg/faq/general/tail_tests.htm

"Introduction to linear regression analysis" http://people.duke.edu/~rnau/regintro.htm

"Statistical Models" http://work.thaslwanter.at/Stats/html/statsModels.html

"Interpreting regression coefficients" http://www.theanalysisfactor.com/interpreting-regression-coefficients/

"Multicollinearity in regression models" http://sites.stat.psu.edu/~ajw13/SpecialTopics/multicollinearity.pdf

"Role of categorical variables in multicollinearity in the linear regression model" http://epub.ub.uni-muenchen.de/2081/1/report008_statistics.pdf

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

"GGPlot" https://pypi.python.org/pypi/ggplot/

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

"Non-parametric tests comparing two dependent samples" http://stats.stackexchange.com/questions/65775/non-parametric-tests-comparing-dependent-samples

"Wilcoxon signed-rank test" https://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test

"Wilcoxon signed-rank test" http://vassarstats.net/textbook/ch12a.html

"6 ways to address collinearity in regression models" http://learnitdaily.com/six-ways-to-address-collinearity-in-regression-models/

Section 1: Statistical Test

1.1 I used the Mann Whitney test to analyze the difference in ridership levels on the NYC subway system on rainy and non-rainy days. I used the two-tailed, and set p-critical to 0.05. The null

hypothesis is that the ridership on the subway on rainy and non-rainy days is not statistically dissimilar, and that any variation we observe between the two population means is due to chance.

1.2 This statistical test is applicable to this dataset because it does not assume that the data has any specific (ie: normal) distribution. The Mann-Whitney test can be used to compare two populations regardless of the distribution of data in those populations, assuming that the populations are independent.

1.3 The mean ridership on non-rainy days is 1090.28 riders per hour. The mean ridership on rainy days is 1105.45 riders per hour. The p-value for the two-tailed test is 0.0499999.

1.4 Using a two-tailed test, the p-value of the Mann-Whitney test is less than the p-critical value of 0.05. Thus I can reject the null hypothesis that ridership on rainy and non-rainy days is statistically identical. One reasonable alternative hypothesis that I can accept is that ridership increases on rainy days. This interpretation of the data is supported by the sample means as well. There is a small but statistically significant increase in the mean number of riders on rainy days as compared to non-rainy days.

Section 2: Linear Regression

2.1 I used OLS implemented in statsmodel to produce the prediction for ENTRIESn_hourly in my regression model.

2.2 The features used in my model are: rain, Hour, mintempi, meanwindspdi. Since rain is a binary variables and meanwindspdi and mintempi can be two orders of magnitude higher than the values for rain, I normalized the features prior to performing the linear regression. I included dummy variables for UNIT in the model, but did not include a constant in my model in order to guard against multicollinearity.

2.3 The reasons I used these features in the model were in part intuition and in part empirical analysis. Because of rush hour, I thought the hour of day may play a large role in subway ridership. More people are likely to be riding the subway between 8am and 10pm than between 10pm and 8am, for example. (See the visualization below.) Likewise, I imagine that when it is rainy, cold, or windy outside those people who would usually walk to work might be more inclined to take the subway. I also looked at the correlation between features and did not include features in my model that were highly correlated. Finally, I observed that when dummy variables were included the model my R-squared statistic dramatically increased. Hence the inclusion of UNIT in the model.

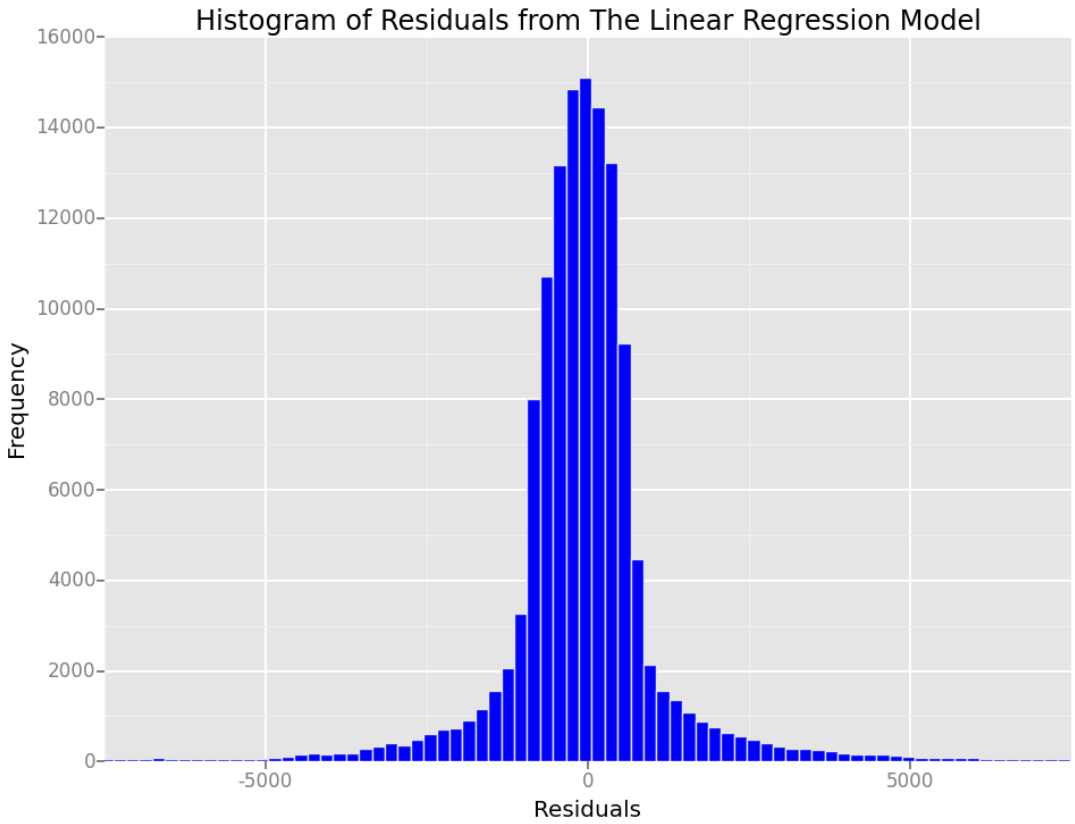
2.4 The coefficients of the non-dummy variables are as follows:

Rain: 4.6047
Hour: 464.5506
Mintempi: -49.7291
Meanwindspdi: 42.5131

2.5 My model's R-squared value is 0.556, and the adjusted R-squared is 0.555.

2.6 This R-squared value indicates that the model is a predictor of subway ridership. It means that 55% of the variance in ENTRIESn-hourly is described by the model. An R-squared of 0.556 adequately describes subway ridership given the number of degrees of freedom in the model (469) and in the residuals (13482). A plot of the residuals from the model (below) can also offer insights into the goodness-of-fit of the model to the data. The residuals of the model are normally distributed with a

mean of 1.023 x 10^-12, and a standard deviation of 1719.564. While the majority of the residuals are centered around 0 (indicated by the magnitude of the mean), the larger than expected standard deviation, as well as a number of residuals greater than 5000 in magnitude indicate that this model does not completely account for all the data in the dataset. This is consistent with an R-squared of 0.556, which indicates that the model accounts for 55% of the variance in the data. In summary, this model is a good predictive model, but there are some outliers in the data that are not well fit for in the model.



Section 3: Visualization

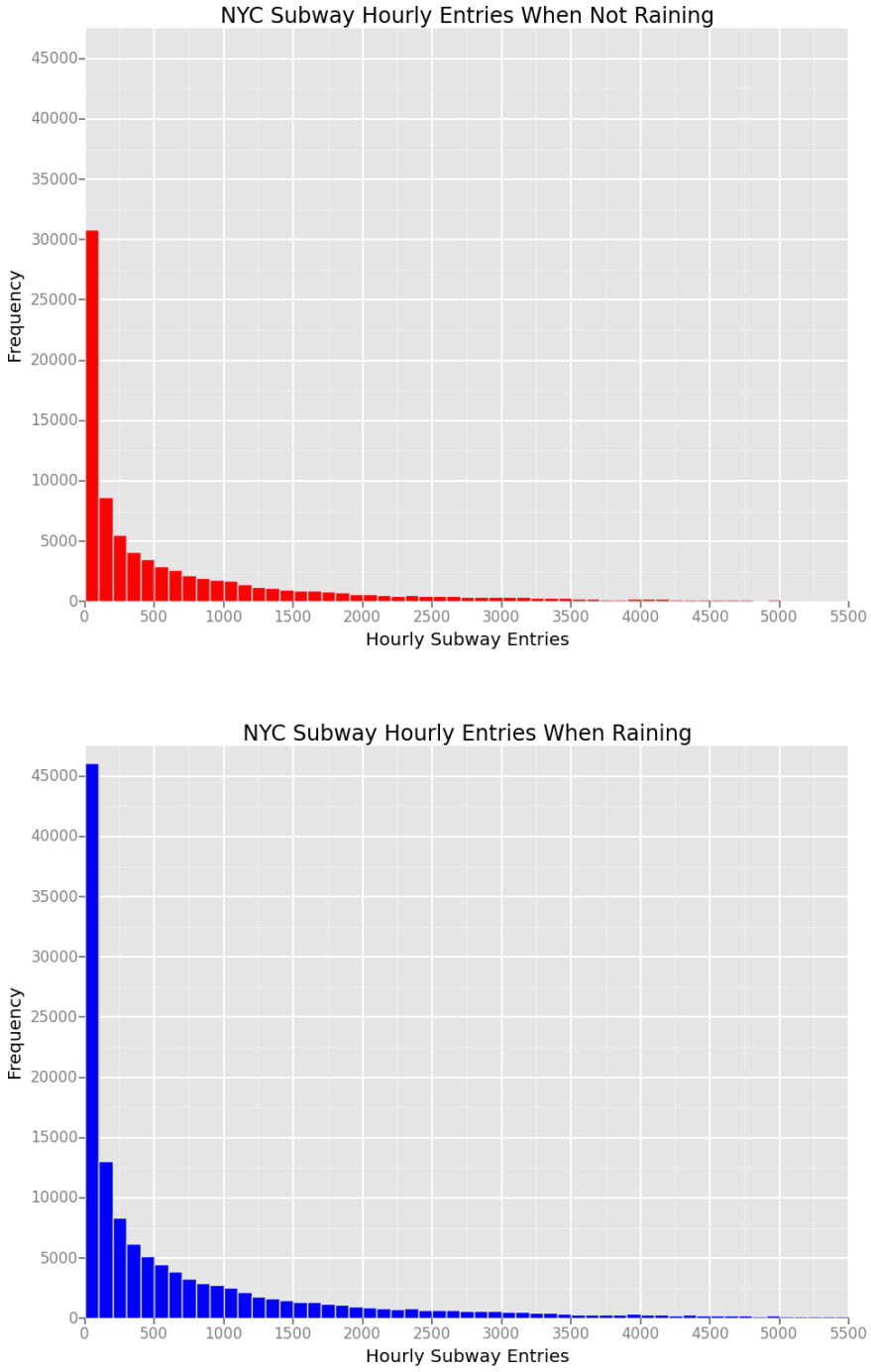


Figure: (Top) Hourly entries on the New York City subway system when it is not raining. (Bottom) Hourly entries on the New York City subway system when it is raining. There are more entries into the subway system when it is raining than when it is not raining.

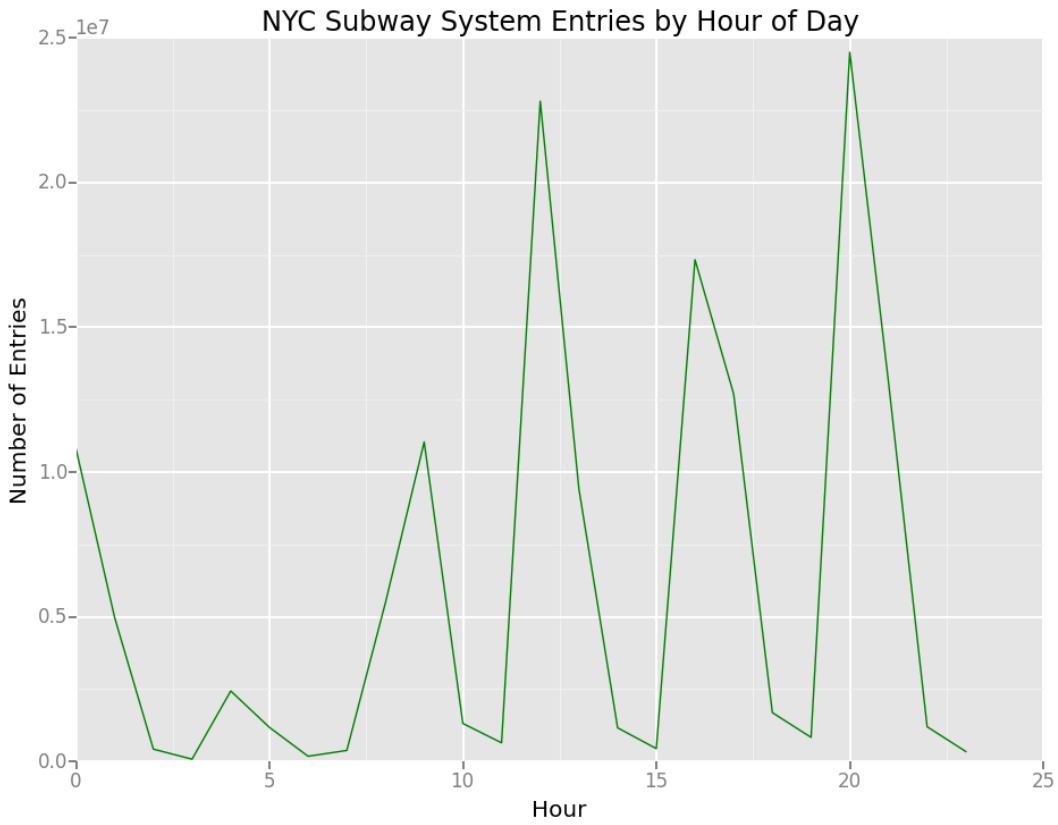


Figure: New York City subway system total entries by hour of the day. Notice that entries into the subway system peak at 09:00, 12:00, 16:00, and 20:00.

Section 4: Conclusion

We can infer that more people ride the subway when it is raining than when it is not raining. The mean subway ridership per hour is 15 riders higher on rainy days than on non-rainy days. This difference was determined to be statistically significant by the Mann Whitney test. We were able to reject the null hypothesis that there is no difference in ridership between the rainy and non-rainy days since the p-value of the Mann Whitney test fell below the p-critical value of 0.05. Furthermore, although regression analysis is used to predict, not infer, information about our data, by looking at the coefficients of the model we can see that rain and wind (which often accompany severe rain storms) positively affect the slope of the linear model, suggesting a positive correlation between rain, wind, and subway ridership. Since the data was normalized prior to the linear regression, we can also consider the magnitude of the coefficients when interpreting the linear regression model. It is interesting to note that, when compared to Hour, rain is only a weak predictor of subway ridership. According to my model, the time of day is a much greater predictor of ridership than rain. Regardless, rain is a predictor in the linear regression model, and the regression model, combined with the results of the Mann Whitney test indicate that ridership is increased on the NYC subway on rainy days compared to non-rainy days.

Section 5: Reflection

There are some shortcomings in both the dataset and the analysis used in this project. First, it is assumed that there is no relationship between riders on rainy days and non-rainy days. The decision to ride the subway on rainy days is considered to be independent of the decision to ride on non-rainy days. In other words, the probability of taking the subway on a non-rainy day does not depend on the probability of taking the subway on a rainy day. However, I would argue that these two datasets are not independent. Although this may be an independent decision for some riders (those who live close to their destination, for example), other riders, particularly those that live a long distance from their intended destination, will ride the subway regardless of the weather conditions. Thus, for these riders, and hence a certain subset of our data, ridership may not be dependent on rain. Hence, independence between the two datasets is lost. It would be interesting to re-do this analysis if we could somehow exclude those riders who are "regular" subway riders (for instance, exclude riders who ride the subway 4 or more days a week). Alternatively, instead of altering the dataset to remove possible dependence between ridership on rainy and non-rainy days, we may choose instead to use a different statistical test to test the populations. The Wilcoxon signed-rank test may be appropriate for this data. Unlike the Mann Whitney test, which assumes that the two populations are independent, the Wilcoxon signed-rank test is a nonparametric test that is used to test for correlated samples. Since at least some portion of our data may be correlated, this test may be more appropriate.

There are also some shortcomings in the linear regression model used to interpret the data. Before performing the linear regression I looked at the correlation coefficients between variables, and excluded features in the model that were highly correlated with one another from the analysis. This is one method of avoiding multi-collinearity before performing the regression analysis. I also excluded a constant from the linear regression model to avoid multi-collinearity, since dummy variables were included in the model. However, the condition number of the model is 46.4, which indicates that there is still some multicollinearity in the model. (Texts and scholarly research articles differ in what is considered to be an appropriate condition number "cut-off" and this cut-off appears to be data-dependent). The higher than optimal condition number may affect both the magnitude and direction of the coefficients. So, although the model may be a good predictive model (as evidenced by an R-squared of 0.556), the inference that hour is a great predictor of subway ridership than rain may not necessarily be true. The use of further statistical analysis, such as PCA, to select uncorrelated features to be included in the linear regression model may help avoid multicollinearity, but I believe that is outside the scope of this course.