Andrew Burruss
Udacity Data Science
Project 2: New York City Subway Data
30 November, 2015

Section 0: References

(1) *Elementary Statistics*, Mario F. Triola, 12th ed. section 8.2: *Basics of Hypothesis Testing*
(2) matplotlib documentation:
    hist() method
        matplotlib.org/api/pyplot_api.html?highlight=hist#matplotlib.pyplot.hist
    set_alpha
        matplotlib.org/api/artist_api.html?highlight=alpha#matplotlib.artist.Artist.set_alpha
    Legend Guide
        matplotlib.org/users/legend_guide.html
(3) nyc_subway_weather_descriptions.pdf. Udacity: Introduction to Data Science: Lesson 3: Data Analysis: Downloadables
(4) Pandas 0.17.0 documentation. API reference. Pandas.get_dummies
    pandas.pydata.org/pandas-docs/version/0.17.0/generated/pandas.get_dummies.html
(5) SciPy Documentation: Mann-Whitney U test
    docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html
(6) StatsModels Documentation: Ordinary Least Squares
    statsmodels.sourceforge.net/devel/examples/notebooks/generated/ols.html
(7) *Understanding the Mann-Whitney U Test*, Udacity: Introduction to Data Science: Lesson 3: Data Analysis: Downloadables. Udacity 2014
(8) Wikipedia: Coefficient of Determination. en.wikipedia.org/wiki/Coefficient_of_determination
(9) Wikipedia: Dummy Variable (statistics).
    en.wikipedia.org/wiki/Dummy_variable_(statistics)#Linear_probability_model
(10) Wikipedia: Multicollinearity. en.wikipedia.org/wiki/Multicollinearity
(11) Engineering Statistics Handbook. 5.2.4. *Are the model residuals well-behaved?*
    www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm
(12) Stack Overflow:
    Adding a legend to PyPlot in Matplotlib
        stackoverflow.com/questions/19125722/adding-a-legend-to-pyplot-in-matplotlib-in-the-most-simple-manner-possible
    How to truncate axes of a subplot in Matplotlib
        stackoverflow.com/questions/28452989/how-to-easily-truncate-axes-of-a-subplot-in-matplotlib
    Python pyplot histogram: Adjusting bin width, Not number of bins
        stackoverflow.com/questions/28101623/python-pyplot-histogram-adjusting-bin-width-not-number-of-bins

Section 1: Statistical tests

The New York City Subway system data (turnstile_data_master_with_weather.csv) was analyzed with a comparison of subway ridership on rainy days versus non-rainy days. In particular, we tested if the incidence of rain had any influence (increase or decrease) on the number of subway riders. The data values of 'ENTRIESn_hourly' were sampled. The given documentation (ref. 3) defined 'ENTRIESn_hourly' to be the difference between regular readings (usually 4 hour intervals) of cumulative turnstile entries at each NYC subway unit.

A preliminary histogram of the frequency of the set of 'ENTRIESn_hourly' data values, considered separately for days with rain and days without rain, indicated that the data set was not normally distributed. (Please see section 3: Visualizations.) Since the data did not appear to come from a normal probability distribution, we ruled out the use of Welch's T-test. The Mann-Whitney U test (ranked-sum test) was applied because it does not depend a normal distribution of the data sets, nor does it assume that the either of the data sets are sampled from the same population. Moreover, we note that there were more days without rain than with rain. The Mann-Whitney U test can also be used in cases where one data sample pool is larger than the other. A two-tailed Mann-Whitney U test at the 95% significance level was conducted using the SciPy.Stats library with the following hypotheses.

Let *x* be a randomly drawn value from the set of 'ENTRIESn_hourly' on a day when rain occurred, and let *y* be a randomly drawn 'ENTRIESn_hourly' on a day when rain did not occur.

$$H_0 : P(x > y) = 0.5$$

$$H_1 : P(x > y) \neq 0.5$$

In summary, the null hypothesis (H_0) states that for two randomly chosen 'ENTRIESn_hourly', one chosen from a day with rain (*x*) and one chosen from a day without rain (*y*), there will be an equal probability of *x* being greater than *y*, compared to the probability of *y* being greater than *x*. The alternate hypothesis (H_1) states that there is some significant difference between the two data sets, and that on average, we can expect either the case where *x* is greater than *y*, or the case where *y* is greater than *x*. The statistical analysis using SciPi.Stats Mann-Whitney U test yielded the following results.

U test statistic	1924409167.0
p-value (one-sided)	0.02499991
p-value (two-sided)	0.04999982

Since the SciPi.Stats Mann-Whitney U test generates a one-sided *p*-value (reference 5), then to test the significance of the test statistic, we doubled the reported *p*-value to yield a two-sided *p*-value. The computed two-sided p-value is almost identical to 0.05 (less than 0.000001 different from 0.05), the critical rejection level. We therefore failed to reject the null hypotheses. That is, the Mann-Whitney U test statistic did not indicate that subway ridership was significantly more likely to be greater on a rainy day than a non-rainy day.

Mean value of 'ENTRIESn_hourly'

Days with rain	1105.45
Days without rain	1090.28

For further analysis, we evaluated separately the means of 'ENTRIESn_hourly' (listed above) on days with rain and days without rain, and found there was a 1.38% difference between the two means. The small difference between these two means appears to agree with the hypothesis test conducted with the Mann-Whitney U test. That is, we found no statistically significant evidence to indicate that a randomly chosen 'ENTRIESn_hourly' from a rainy day would be likely to have a value greater or less than a randomly chosen 'ENTRIESn_hourly' from a day without rain.

Therefore we concluded from both of these statistical analyses that the data set indicated that there was no significant increase or decrease in NYC subway ridership on rainy days versus non-rainy days.

Section 2: Linear Regression

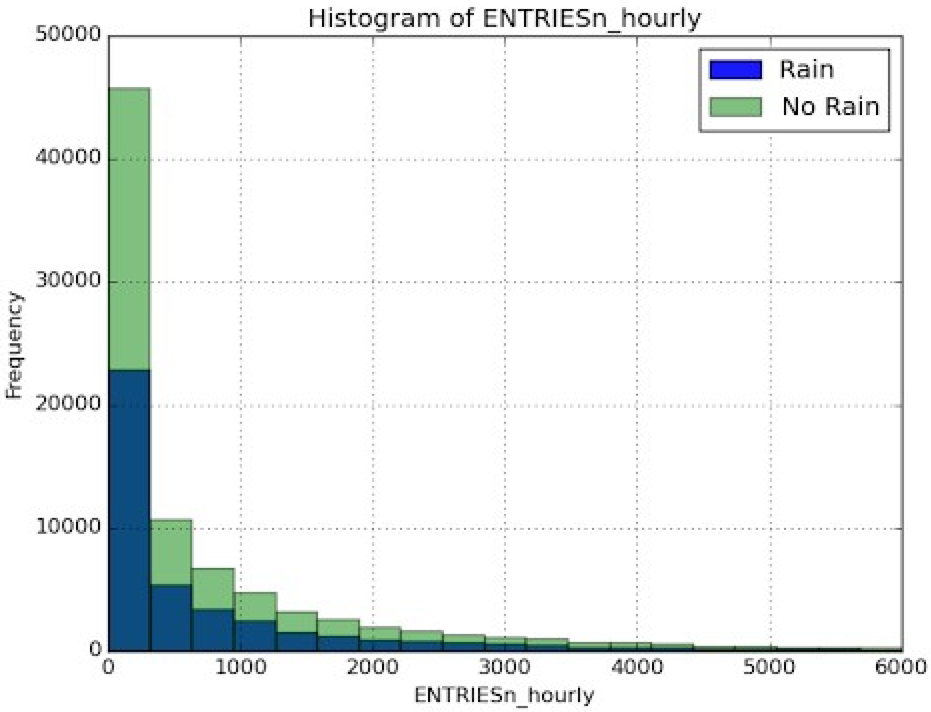
A predictive linear regression model was built with the data using the OLS (Ordinary Least Squares) method implemented with Statsmodels.api in Python. The model attempted to predict the 'ENTRIESn_hourly' in the given data set, turnstile_data_master_with_weather.csv. Several different OLS models were built using different combinations of input variables. The relative effectiveness of different versions of the model were evaluated by comparing the r^2 values. The r^2 value is interpreted as being a coefficient which measures what proportion of the variance of the output variable, 'ENTRIESn_hourly', can be explained by the input variables (features) of the model. The most effective model included the 'Hour' data value as an input variable, and used 'UNIT' as a dummy (indicator) variable. The combined model yielded an r^2 value of 0.478.

The use of a dummy variable in a linear regression model effectively splits the data into subgroups for evaluating the coefficients (weights) associated with the features of a given subgroup (ref. 4, 9). We note that in this data set, 'UNIT' took 552 unique values. Since the NYC Subway data was collected from these remote units' turnstiles, there was no overlap in the 'UNIT' variables. That is, each data entry in the set was recorded from one and only one 'UNIT' entry. This was a computationally effective method of partitioning the data set in the OLS model because of the absence of overlap in the 'UNIT' variable.

Weather variables, such as 'rain' did not provide a substantial increase to the r^2 value of the model. This agreed with the conclusion from the statistical test that the incidence of rain did not increase subway ridership. By itself as an input variable, 'rain' a yielded a negligible r^2 value of 0.0002729. Both 'meantempi' and 'fog', each by themselves, yielded r^2 values of 0.001 which made no significant contribution to the predictive effectiveness of the model. While there were a number of weather variables available for use in the model, in order to reduce redundancy in the model input variables, only these three were considered (ref. 10).

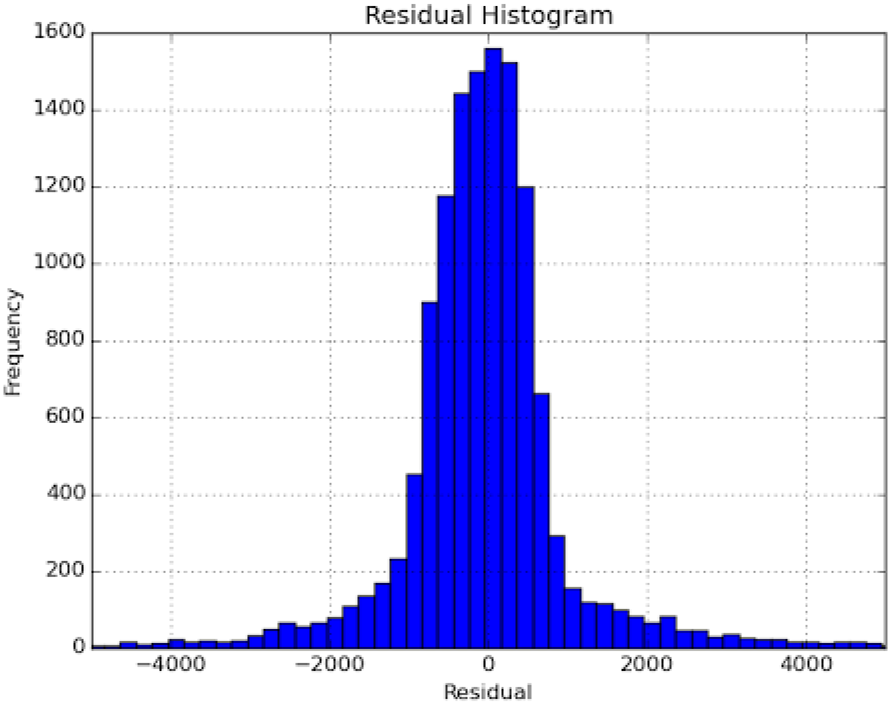
Section 3: Visualizations

This histogram plots the frequency of rider entries (at four-hour intervals) to the NYC Subway systerm on rainy days (blue) compared with the frequency of rider entries on non-rainy days (green). We note that there were many more data records for days without rain. The curve of the histogram indicated that this data set was not normally distributed, and therefore a non-parametric test, the Mann-Whitney U Test, was used to analyze the data. Note that the horizontal axis, 'ENTRIESn_hourly' was truncated at 6000 to make the distribution curve appear more clearly. The outliers in 'ENTRIESn_hourly' extended beyond 50000.



Section 3: Visualizations

The following histogram plots the distribution of the residuals of the OLS regression model. The *residuals* are defined to be the differences between the observed and predicted values of the model (ref. 11). The histogram indicates that the residuals appear to be normally distributed with a mean of zero. The distribution of the residuals was concentrated around zero which indicates that the majority of predictions made by the model were close to the observed values of 'ENTRIESn_hourly'. Note that the horizontal axis of residuals was truncated between (-5000,5000) to make the distribution curve appear more clearly. Outliers of the residuals occurred between -15000 and 30000.



Section 4: Conclusion

Statistical analysis conducted with the Mann-Whitney U-test indicated that there was not sufficient statistical evidence at the 95% confidence level to reject the null hypothesis that passengers are no more likely to ride the New York City subway on rainy days than non-rainy days. Comparison of the two different mean number of riders on rainy and non-rainy days also indicated that there was no significant statistical influence of rain on New York City subway ridership.

Predictive modeling with OLS (Ordinary Least Squares) linear regression indicated that rain and other weather variables were almost negligible predictors of the hourly number of riders entering the subway. The number of riders exiting subway stations and the hour of the day contributed the most useful predictors of the number of riders entering subway stations.

Section 5.1: Reflection

Time was incremented in four-hour intervals in this data set. It could be more valuable for predictive modeling to have more time data at finer (possibly hourly) time intervals. Inclusion of the 'Hour' input variable in the linear regression model improved the r^2 value by 0.004, but it was limited to the four-hour increments of all time values reported in the data set. It may be possible to improve the r^2 value of the linear regression model with more time values recorded at finer intervals.

The analysis would benefit from more geographical relational data. Station identifiers ('UNIT') were given, but there was no accompanying mapping data which related these identifiers to their geographical position in the city. It may be possible to generate more accurate predictive models using more detailed geographical data. For example, to relate rush hour ridership to downtown locations, or to identify stations which were outliers to more dominant trends in the subway system.

The data set is also limited by the fact that the primary outcome were measured by turnstile data, namely hourly entries and exits. It would be valuable to have more detailed data which plotted the journeys taken by individual riders, such as stations of entry and exit to the subway system, and the times at which these journeys occurred. With this type of data it may be possible to predict trends of riders' journeys, such as daily commuters versus occasional riders, or determine what proportion of subway riders did not change their travels due to weather variables.

It would also be useful to have a larger data set which sampled from dates throughout the year in New York City. All of the data values appear from the month of May, 2011. It would be interesting to compare the influence of weather variables in cold months versus warm months. We may see the same type of trend where precipitation encourages more subway riders, but the relative influence of cold weather and snow may be different than rain in May.