Section 0. References

(1) http://nbviewer.ipython.org/github/n-batalha/mooc-projects/blob/master/intro-data-science/subway_usage_weather-udacity_2014.ipynb
(2) http://unsupervised-learning.com/multivariate-linear-regression-gradient-descent/
(3) https://www.coursera.org/course/ml
(4) http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

Section 1. Statistical Test

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 Mann-Whiney U Test to analyze the NYC subway data.

I used a two-tailed test since we do not have any concrete proof that the ENTRIESn_hourly value during rainy days will be higher than during non-rainy days or the ENTRIESn_hourly value during non-rainy days will be higher than the rainy days.

The null hypothesis is that the observations from both groups (i.e. the ridership during rainy and non-rainy days) are statistically independent of each other. In other words, rainy or non-rainy conditions has no effect on ridership.
The Alternative hypothesis being one distribution is stochastically greater than the other.

The p-critical value is ≤ 0.05, which means that result obtained from using SciPy should be ≤ 0.025 because the result from SciPy corresponds to one-tailed test.

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.

I needed to analyze the data from two populations, the ridership during rainy and non-rainy days. These two datasets could have different sample size. Based on this, Welch's t-Test or Mann-Whitney U Test seemed ideal to check the null hypothesis, whether the ridership remains the same during rainy and non-rainy days. But to apply Welch's t-Test the sample sizes of two populations should be same. From problem set 3.1, we can very well understand that the histogram for entries during rainy and non-rainy days are not normally distributed. In this situation, Mann-Whitney U Test seemed the correct statistical test to use.

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.

The p-value = 0.0498 *(After doubling the result obtained using SciPy)*
The mean of Entries per hour during rain = 1105.46
The mean of Entries per hour during no-rain = 1090.27

1.4 What is the significance and interpretation of these results?

The difference between the mean of hourly entries during rainy and non-rainy days is 15 days. However, we cannot construe a conclusion out of this result alone. After checking the p-value (0.0498), we can say that it marginally satisfies the p-critical value of ≤ 0.05 and conclude that the null hypothesis is false and the ridership in rainy and non-rainy days are in fact different.

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:

1. Gradient descent (as implemented in exercise 3.5)
2. OLS using Statsmodels
3. Or something different?

I used gradient descent to run linear regression on the NYC subway dataset.

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

The features I used are rain, precipitation (precipi), mean wind speed (meanwindspdi), hour of the day (Hour), minimum temperature (mintempi) and dummy variable for individual station (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.

- 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 chose my features to suit the performance of the regression model. The performance of a regression model is comparatively better when the features are most relevant. In this scenario, we are observing the entries during rainy and non- rainy weather, so according to me the features which can actually affect ridership are rain, precipitation, mean wind speed, hour, and minimum temperature.

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

[15.24925667 9.11629979 54.66883165 421.36724596 -51.15719157 1100.60865797]

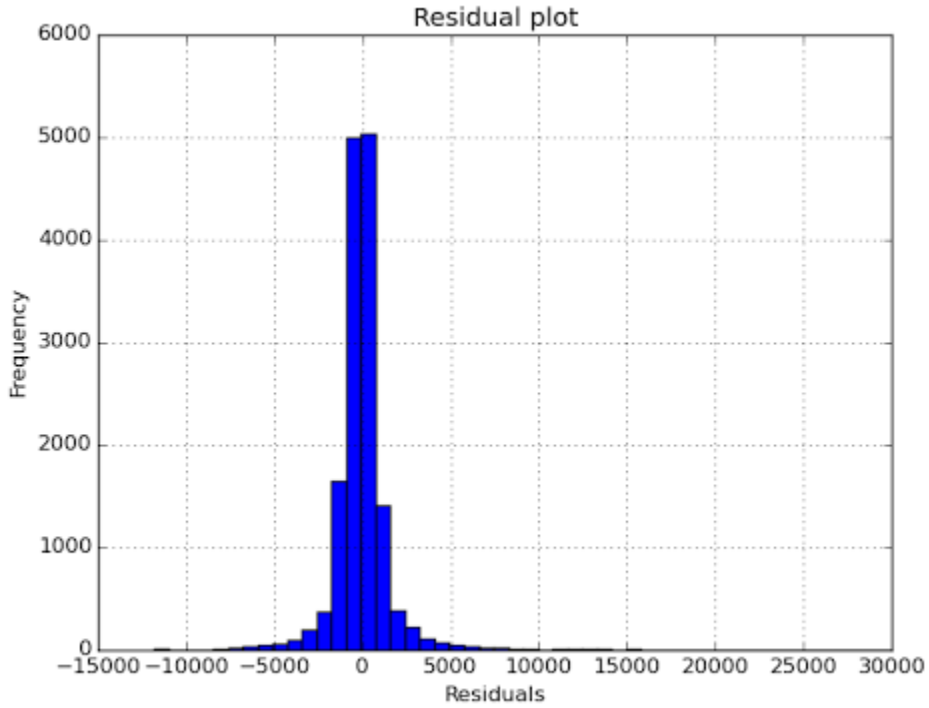
2.5 What is your model's R^2 (coefficients of determination) value?

0.4646

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?

From the R^2 value that we obtained, we can say that the model can identify 46.46% of the variation present in the data it was trained on. In other words, our models lets us predict NYC subway entries with 46.46% accuracy. But we should check residual plots also as

residual plots can reveal unwanted residual patterns that indicate biased results.

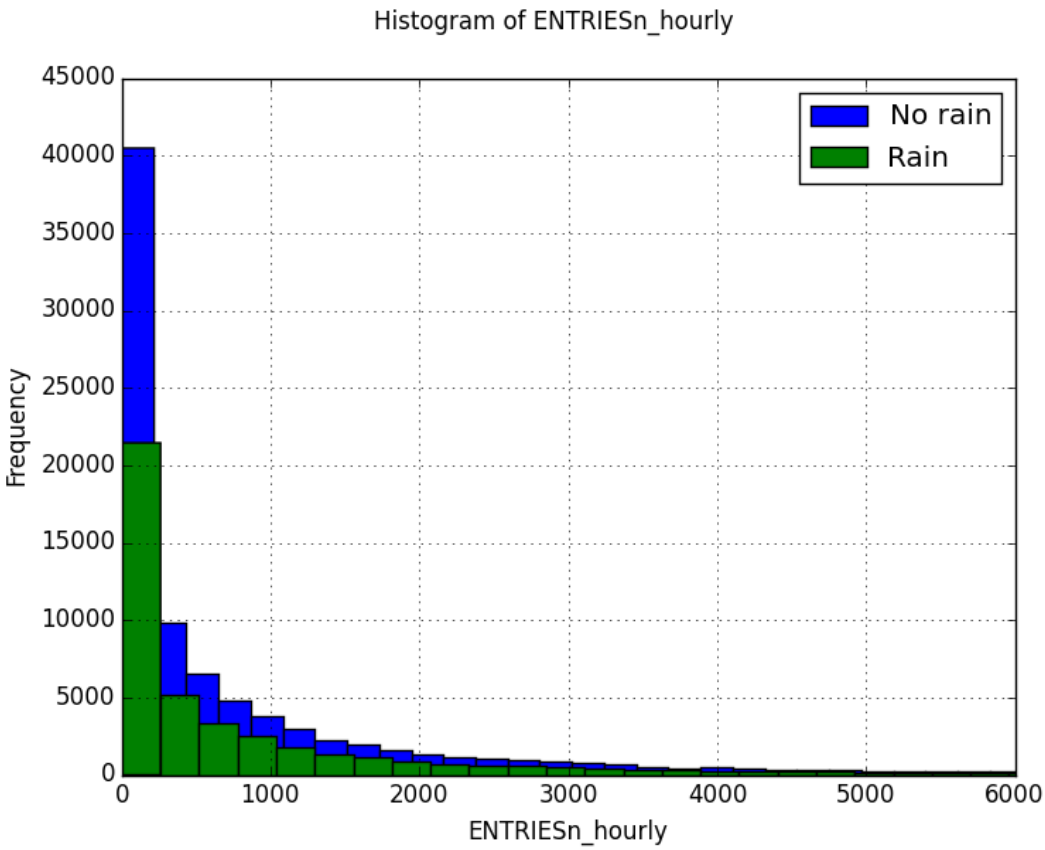


From the above histogram, we can see that residuals are quite normally distributed around zero. Qualitatively, we can say that this model is good enough to predict ridership.

Section 2. 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., scatterplots, line plots, or histograms) or attempt to implement something more advanced if you'd like.

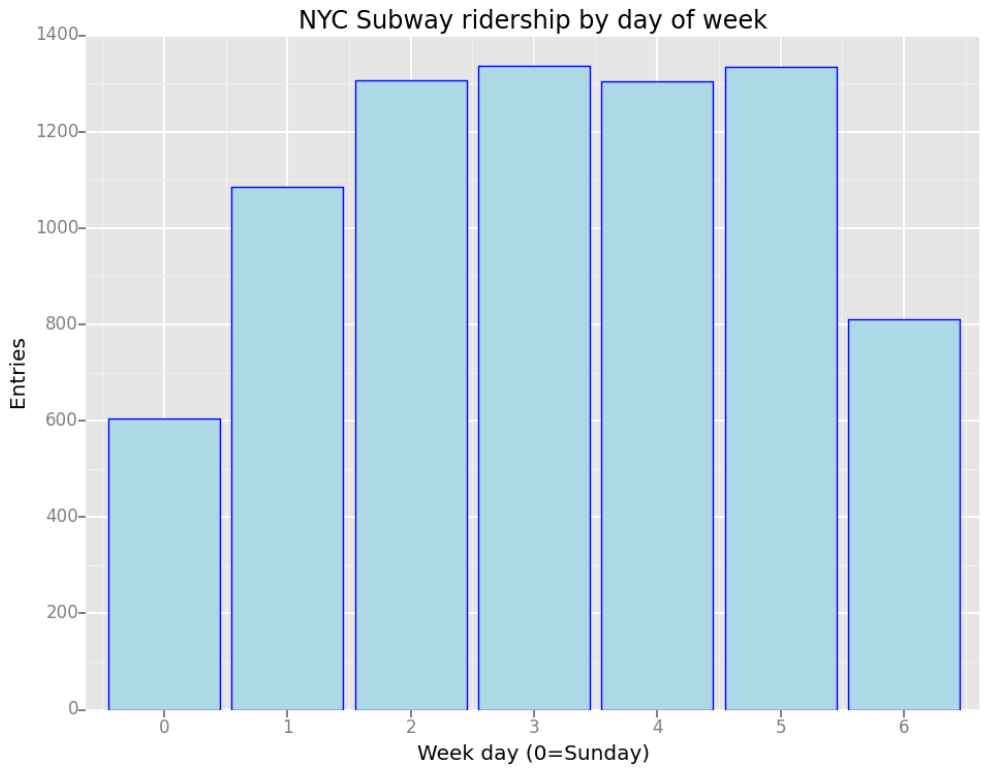
3.1 One visualization should be two histograms of ENTRIESn_hourly for rainy days and non-rainy days



In above visualization, I illustrated 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 ENTRIESn_hourly was shown in x-axis. The frequency of ENTRIESn_hourly was shown in y-axis. The figure was plotted using matplotlib.pyplot with bins = 200.

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:

- Ridership by time-of-day
- Ridership by day-of-week



In this visualization, I have illustrated the density of ridership by day of the week. It is evident that on Wednesday and Friday, the ridership is the highest. The ridership spikes to more than 83% on Mondays when compared to Sundays and remains stable at 1300-1340 levels during mid-week and then plummets to around 800 as the week comes to an end.

Section 4. Conclusion

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?

In this analysis I tested the null hypothesis is that the observations from both groups (i.e. the ridership during rainy and non-rainy days) are statistically independent of each other. I found out that the mean of ENTRIESn_hourly on rainy days (1105.46) was slightly higher than the mean of ENTRIESn_hourly on non-rainy days (1090.27). Although, coming up with a conclusion based on mean values is not enough, due to variance. Moreover, the two-tailed Mann-Whitney U Test results showed a p-value of 0.0498, which is less than 0.05. From the results obtained from Mann-Whitney U Test, we can conclude with certainty that ridership increases when it rains.

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 of rain reflects that rainy condition do increase the ridership. The mean of ENTRIESn_hourly on rainy days (1105.46) was slightly higher than the mean of ENTRIESn_hourly on non-rainy days. The difference is not much but by looking at the results from Mann-Whitney U Test we can say that there a change in ridership for rainy and non-rainy days. Hence, it would not be unwise 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.

1. ***The dataset*** only included data from May 2011, increasing sample size may change the results of analysis and conclusions. Different months may have different number of rainy days, but this dataset only included data from May. Thus, the number of rainy days or non-rainy days may be biased. To improve this, the data from the whole year should be included for analysis.

2. ***This analysis*** only considers rainy and non-rainy condition, but other weather may also impact ridership. It cannot be denied that the exclusion of “fog” or “non-fog” days biased the results of analysis, which is only based on rainy days and non-rainy days. To overcome this shortcoming, more detailed comparisons could have been used. Moreover, we did not consider if our data should use linear regression or not. The nature of some datasets might not fit very well in linear regression, sometime we need nonlinear regression to make more precise prediction