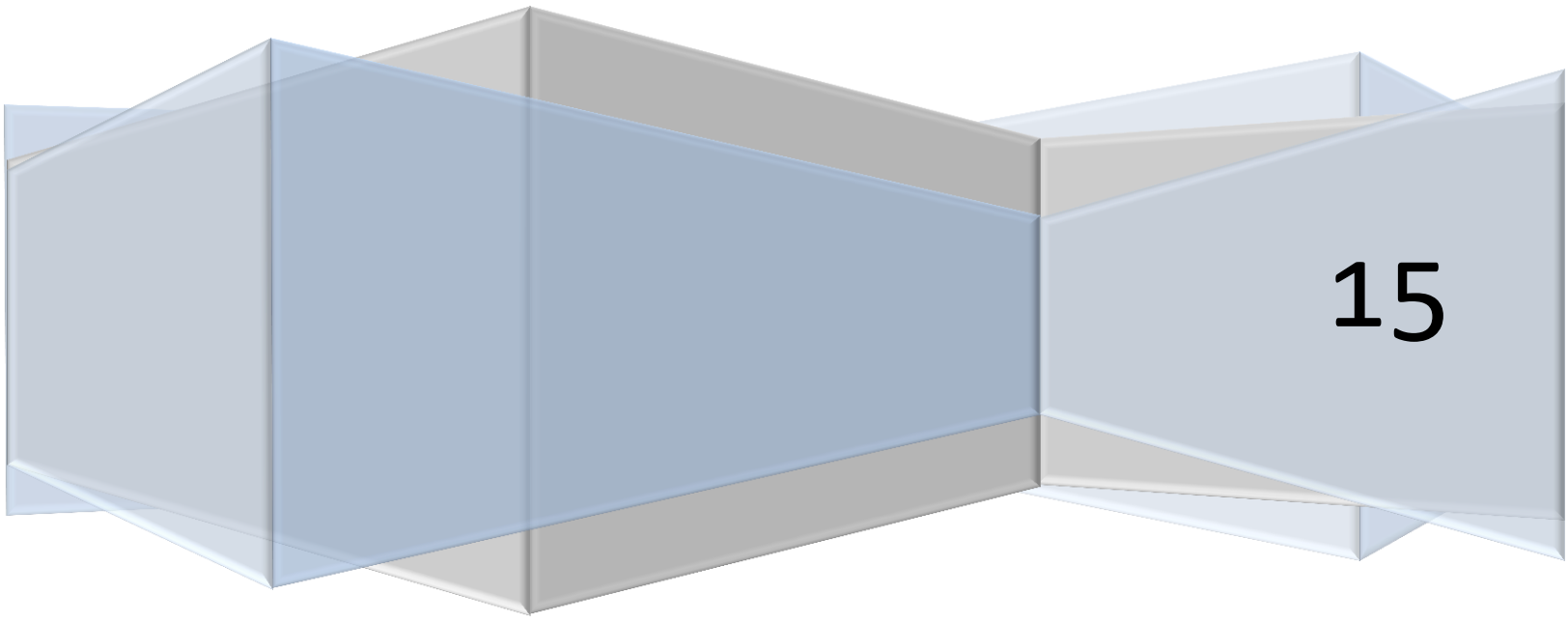
Analyzing the NYC Subway Dataset

Data Analyst Nanodegree – Intro to Data Science

Alex Schaal



Section 0. References

- "Interpreting results: Mann-Whitney test"
  - http://www.graphpad.com/guides/prism/6/statistics/index.htm?how_the_mann-whitney_test_works.htm
- "Coefficient of Determination"
  - http://stattrek.com/statistics/dictionary.aspx?definition=coefficient_of_determination
- "Points, as for a scatterplot"
  - http://docs.ggplot2.org/0.9.3/geom_point.html
- "Working with DataFrames"
  - http://www.gregreda.com/2013/10/26/working-with-pandas-dataframes/
- Residual Plot
  - http://stattrek.com/statistics/dictionary.aspx?definition=residual+plot

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?

I used the Mann-Whitney U test to analyze the NYC subway data. Since it is not known which data set has a higher / lower mean, I used a two-tail P value. The null hypothesis we are testing is that the two populations (in this case, rain & non rain populations) are the same so that an observation from a value randomly selected from one population exceeds the observation randomly selected from the other population. The p-critical value is 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.

I performed a histogram plot of the hourly ridership data with rain vs. without rain. Through this graph I determined the data was non-normal. A Shapiro-Wilk test could also be run to verify the non-normality of data; however with this particular data set (> 5,000 records) this test would not be accurate.

The Mann-Whitney U test is a non-parametric test and would be appropriate when analyzing data that is not normally distributed.

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.

Mean with rain: 1105.45
Mean without rain: 1090.28
U-value: 1924409167
One tail P-value: 0.024999912793489721
Two tail P-value: 0.049999825586979442

1.4 What is the significance and interpretation of these results?

The generated P value above is smaller than our p-critical (0.05) value. This allows us to reject the null hypothesis and conclude that hourly subway ridership is different when raining vs. when not raining.

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 descent

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

I used rain, precipitation, mean temperature, and hour as input variables. I used UNIT as a dummy variable

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 used the above input variables based on my own experience riding the NYC subway. I suspected that poor weather conditions would have a linear effect on ridership.

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

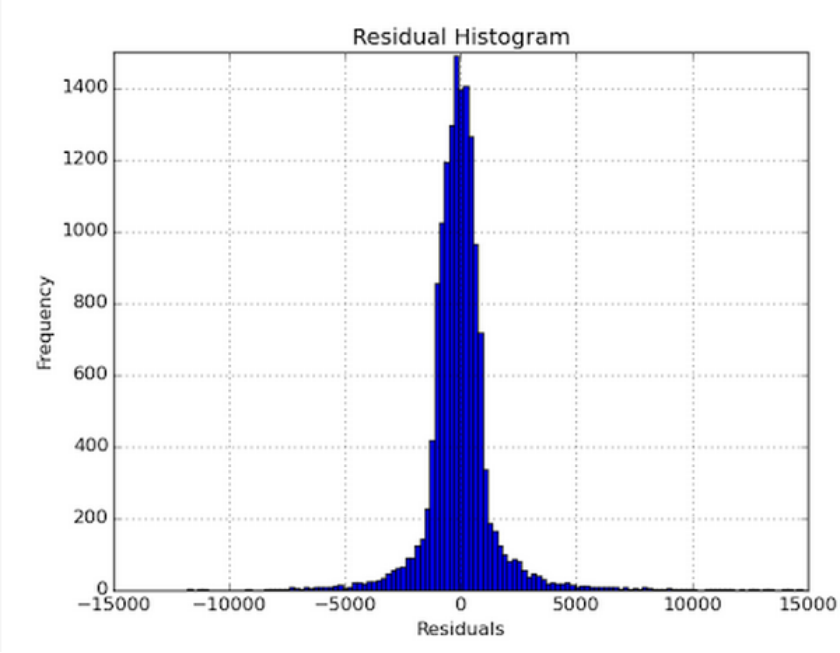
Rain: 8.35e-01
Hour: 4.64e+02
Meantempi: -4.96e+01

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

0.463968815042

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?

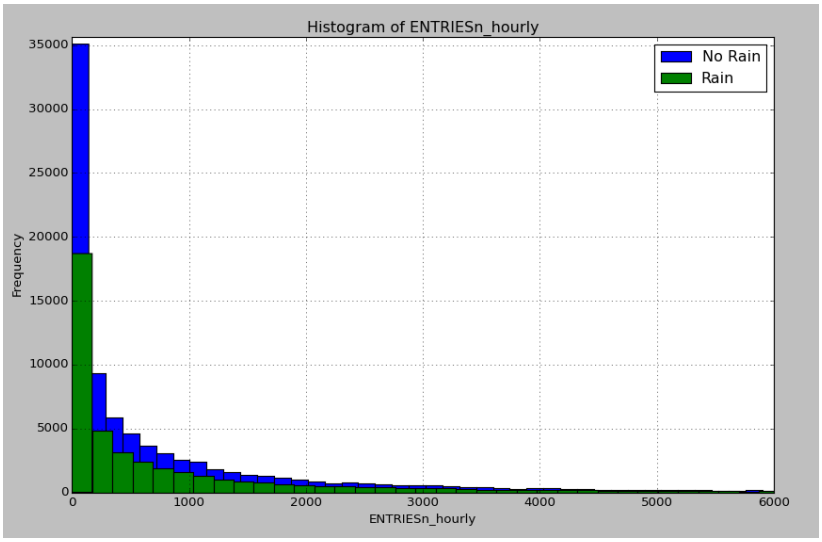
Given that the R2 value indicates how well data fit a statistical model (i.e. the closer R2 is to 1, the better the fit). I do not believe my model is appropriate as it only explains 46.3% of the variation. To further analyze this data we can plot the residuals.



Based on the residual plot above we can see the data appears non-random in a cyclical pattern (inverse U shaped). In this type of residual plot a non-linear model would be a more appropriate fit.

Section 3. Visualization

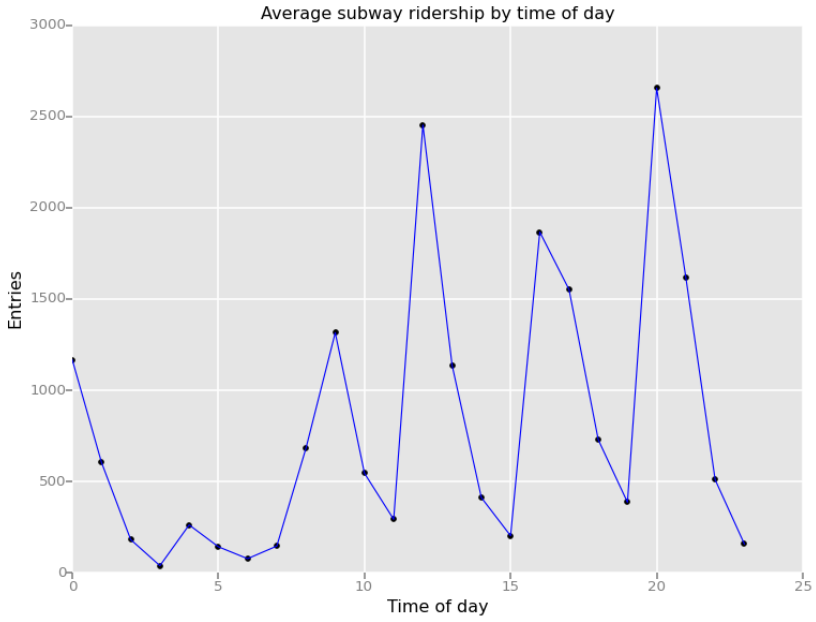
1.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.



The above image represents a histogram plot of hourly subway entries with & without rain. From the plot we can determine the following:

- Both data sets are not normally distributed
- The number of data points for days without rain is greater than data points with rain

1.2 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.



The above image represents the average subway entries per time of day. We can see from the above image that ridership increases in the afternoon and evening. It's unsurprising that ridership is low between 2–7am as most people would be asleep. However, I expected ridership peaks at 7-8am and 5-6pm which indicates the typical start and end of the work day. Surprisingly we see peaks at 1pm and 4pm. To further analyze this data for trends we should increase our input variables to include weather, days of the week, holiday schedule, and demographics.

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? What analyses lead you to this conclusion? You should use results from both your statistical tests and your linear regression to support your analysis.

From my analysis I can say that ridership of the NYC subway increases when raining vs. when not raining. I have come to this conclusion based on tests performed in section 1 (Mann-Whitney U) & section 2 (Linear Regression).

The Mann-Whitney U test allowed us to reject our null hypothesis based on our P value being lower than the P-critical value. This tells us that there is an impact on subway ridership when raining vs. when not raining. Further analysis using linear regression testing with Gradient Descent shows that the rain coefficient has a positive impact on subway ridership (i.e. increases ridership).

Section 5. Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis, including:

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

One area of concern when analyzing the turnstile data was the number of entries vs. exits. From the dataset we can see ~19% less subway exists than entries. This leads me to question the accuracy of the dataset and the mechanism used to capture the data.

The use of the Mann-Whitney U test to disprove our null hypothesis was the best approach given that the data was non-normal. However, this test doesn't take into account some units being more active despite weather conditions (i.e. special events, tourism, holidays, etc…).

I also question the use of a linear regression model to analyze this particular data set. In section 2.6, I explained some of the reasons why this particular data set would not be appropriate to analyze with a linear regression model. Based on the residual plot, a non-linear model would be a better fit for testing this data set.