Analyzing the NYC Subway Dataset

Name: Milan Patel
Email: patelmm79@yahoo.com

Section 0. References

http://en.wikipedia.org/wiki/Exponential_distribution

http://en.wikipedia.org/wiki/Multicollinearity

https://review.udacity.com/#!/reviews/9472

http://blog.minitab.com/blog/adventures-in-statistics/how-high-should-r-squared-be-in-regression-analysis

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

Section 1. Statistical Test

1.1 I used Mann-Whitney's test to analyze the NYC subway data, using a two-tailed P value. The null hypothesis is that hourly entries when it is raining and it is not raining come from the same population—that is, the presence of rain does not affect hourly entries into the subway. P critical value = 0.05.

1.2 Based on the histogram visualizations below, the data are not in a normal distribution. Since the Mann-Whitney test does not make an assumption of distribution, it is appropriate to use for this data.

1.3
p-value is 0.000005482
mean for hourly entries when raining: 2028
mean for hourly entries when not raining: 1846

1.4 The significance is that the probability that the null hypothesis is correct is 0.0005482%. So it is extremely likely that hourly entries when raining and not raining come from different data sets. Based on the means, it seems that more people use the subway when raining. Based on this information, we reject the null hypothesis.

Section 2. Linear Regression

2.1 I used the Gradient descent model, as used in exercise 3.5

2.2 & 2.3: What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?
I used the following features:

- Hour: I used this variable because usage of subway clearly changes depending on the time of day; more people use the subway during daylight hours, and especially during rush hour.
- Wspdi: I used windspeed because I thought that subway usage would increase in windier conditions.
- Tempi: I used temperature because I thought that subway usage would increase in colder conditions.
- I created new variable "difftemp", which is difference between mean temperature and actual temperature. I did so based on the idea that when weather conditions changed radically within the same day, people are more likely to use the subway.

I also used the following dummy variables:

- Conds: I used this set of dummy variables, because the column was an excellent representation of weather conditions during the day. This variable seemed more useful than comparable variables such as "rain", because the "conds" variable indicated severity of weather, such as "Light rain", "Light drizzle", "Heavy rain", etc.
- Units: This is the variable that represents the subway station used; clearly some stations are used more intensely than others; Grand Central Station, as a transit hub in the middle of Manhattan, is more active than Far Rockaway, at the end of the line in Queens.
- Day_week: Subway usage is likely to be more intense during working days during the week. However, I felt it was not sufficient to use "weekday" variable because usage of the subway would be more extensive on Friday nights, and would likely exhibit different patterns on Saturday and Sunday.

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

Variable	Coefficient
'hour',	738
'wspdi',	36.7
'tempi',	-151
'difftemp',	-315

2.5 R-squared value is 0.493954838641

2.6

The R^2 value sits almost exactly between the two possible extreme values—0, meaning that the model explains none of the variability around the mean, and 1, which indicates that the model explains the variability around the mean perfectly. So, the model that I've created explains about 49% of the variability of entries into the subway system—meaning that 51% of the variability of entries remains unexplained. Or in other words, the model produces accurate predictions about half the time. Given this R-squared value, I believe that this model is appropriate to this dataset, since the model is able to explain much of the variability of subway usage. "Appropriateness" for this data set would ultimately depend on how such a model would be used; it is highly doubtful that any model would be able to explain 100% of variability, but could explain general patterns of subway usage.

Section 3. Visualization

3.1

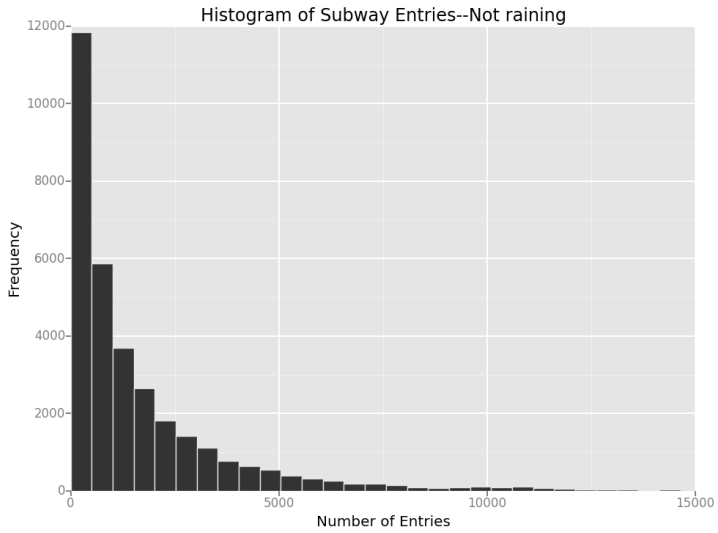

Figure 3.1.1 : Histogram of Hourly Subway Entries, when not raining.

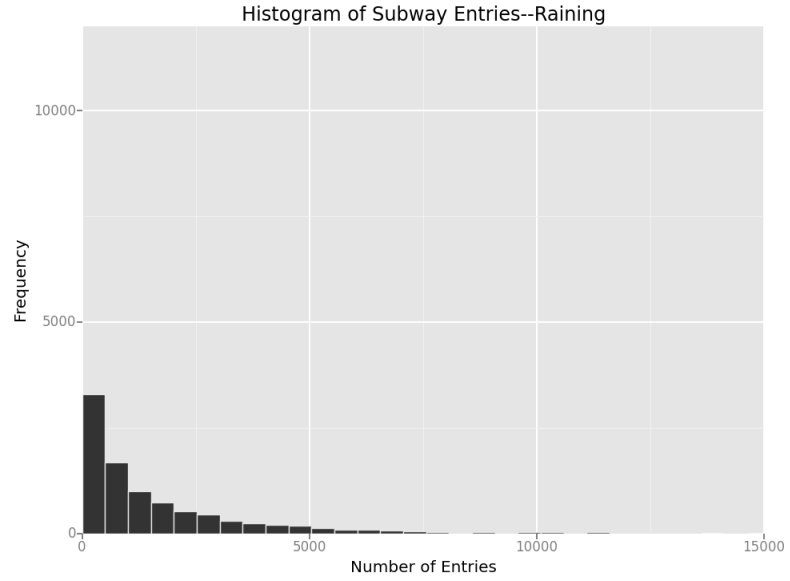

Figure 3.1.2: Histogram of Hourly Subway Entries, when not raining.

Above figures display distribution of hourly subway entries for two populations—one when raining, one when not raining. Figures clearly indicate a non-normal distribution, whose pattern is similar; in fact, the patterns for both most closely resemble an exponential distribution.[1] Note that the relatively decreased frequency in the "rain" population indicates that there were fewer observations of rain vs. no rain.

[1] http://en.wikipedia.org/wiki/Exponential_distribution

3.2

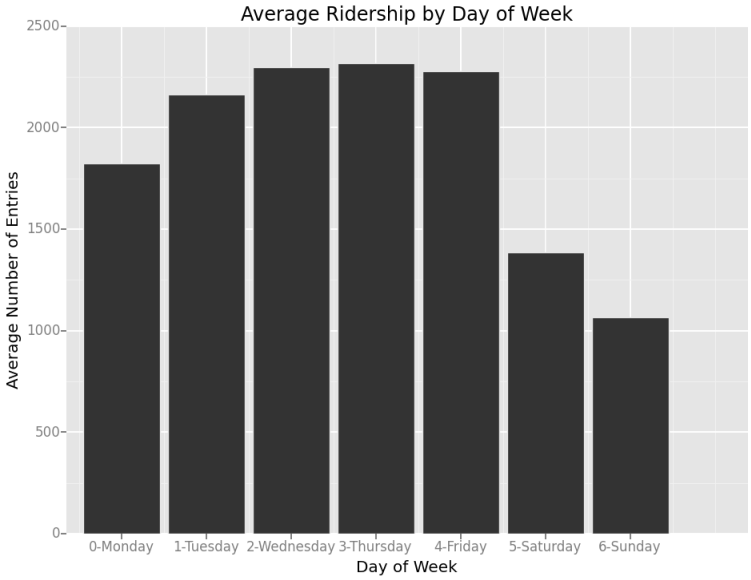

Figure 3.1: Average Ridership (Number of entries per hour) by day of week. One can see significantly decreased usage of subway on weekend days (Saturday and Sunday). Also, Monday exhibits markedly decreased activity compared to the other weekdays. This phenomenon may be because a number of public holidays often occur on Mondays, so a number of Mondays form part of a long weekend.

Section 4. Conclusion

Based on the statistical tests, we determined that subway usage shifts significantly when it is raining. Because the mean of subway entries when raining is greater the mean when not raining, we can see that subway entries will tend to increase when it is raining.

However, the regression gives a more nuanced picture of what happens with subway ridership when it rains, especially when considering other variables. The most salient conclusion from the regression analysis is that ridership may depend on the severity of the rain. Using coefficients from main regression described above, it is apparent that ridership tends to increase when weather conditions are "Light Rain" (coefficient 76.5); ridership tends also to be higher under "Clear" (55) or "Scattered Clouds" (47.9) conditions. However, ridership tends to decrease when weather conditions are "Rain" (-102) or "Heavy Rain" (-48.7).

Using a regression with fewer variables (1 dummy variable (hour)), and non-dummy values including Conds, Units, and day_week), in order to account for possible multicollinearity[2], the effect of ridership based on severity of the weather condition seems to stay roughly the same; ridership still increases with "Light Rain" (coefficient 84.3854068), and decreases with "Rain" (-81.6665947) or "Heavy Rain" (-49.9044673). [3]

[2] http://en.wikipedia.org/wiki/Multicollinearity
[3] Additional analysis conducted based on comments from previous submission

Section 5. Reflection

The dataset is primarily focused on variables related to weather; while weather conditions may certainly drive usage of the subway to a certain extent, a more complete model would consider a more diverse array of variables. Further, much of the explanation of usage is based on which stations were used; however, we have little insight as to why those particular stations were used. Additional variables may consider population who reside or work close to the station, or the subway lines serviced by the stations.

Additionally, subway ridership may depend on the degree of certain weather events, for which binary variables such as "rain" or "fog" are not designed to address. Our model used the "conds" dummy variable to reflect the intensity of rain; another useful variable may be rain gauge measurements.

In terms of analysis, the key limitation of using a linear regression is that it does not accurately reflect non-linear relationships that may exist in the data. Additionally, the contrast between the results of the statistical test (indicating that ridership increases when it is raining) and the regression (which suggests that ridership may very depending on severity of rain, and may actually decrease in certain rainy conditions) suggests the difficulty of any one analytical test to fully establish a completely clear pattern within data. It does seem however that the regression analysis may better suited to finding the nuances in data relationships. Ultimately, the quality of the analytical process does hinge on the quality of the data being analyzed.

5.2

The processes used to determine the conclusions of this analysis have been useful to determine the "what" of data patterns—that is, what is happening? The "why" requires further exploration.

One interesting question to answer would be, why does subway ridership increase in periods of light rain, but decrease in heavier rain? One explanation might be that people would tend to rush to the subway at the initiation of rain, anticipating a heavier downpour. And during heavy rain, people may tend to postpone their journey, remaining inside until such rain passes. Or they may choose alternative methods of transport, such as taxis. As anyone who has lived in New York would tell you, it is very hard to get taxis during heavy rain, due to higher demand.