Analyzing Subway Data

Arash Karami

August 17, 2015

1 Introduction

In this report, we look at the NYC Subway data and figure out if more people ride the subway when it is raining versus when it is not raining. First, we wrangle the NYC subway data and then by using appropriate statistical methods and data visualization we draw an interesting conclusion about the subway dataset.

2 References

1. MannWhitney U test from Wikipedia
2. scipy.stats.mannwhitneyu from scipy.org
3. Coefficient of determinant from Wikipedia

3 Statistical Test

1. At first we let X , Y be the populations that represent ENTRIESn-hourly for the rainy and non rainy weather respectively, and let x, y to be the random draws from those populations . Then We use the two sided Mann-Whitney U-test with the p-critical value equal of 0.05 and the null hypothesis:

   $$H_0 : p(x > y) = 0.5 \text{ where x and } y \text{ are the random draws from X, Y}$$

2. Our assumptions on data sets using Mann Whitney U-test are:

   - The samples from the populations are random
   - We don't assume any assumptions related to the distribution of our populations(We talk more about it in data visualization).
   - All the observations from X and Y are independent.
   - Dependent variable which is called response(ENTRIESn-hourly) is ordinal.

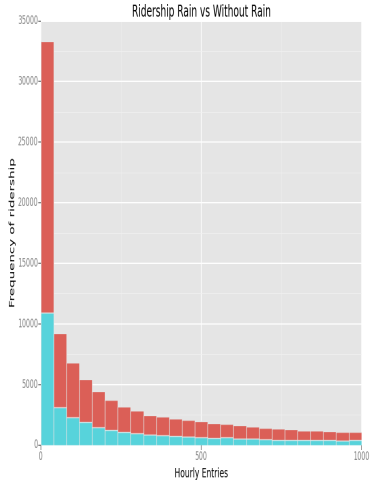
3. The results of the test are:
    - p-values= 0.049
    - U-statistics=1924409167.0
    - μ_X=Mean for X(with-rain-weather)=1105.4463767458733
    - μ_Y=Mean for Y(without-rain-weather)=1090.278780151855
4. As the two-tail p-value is about the 0.049 considering a significance level of 5% we can then conclude there is enough evidence to reject the null hypothesis. To be more precise , the mean ridership during rainy days is higher than the mean ridership without rain, which would imply more people use NYC subway during rainy days.

4 Linear regression

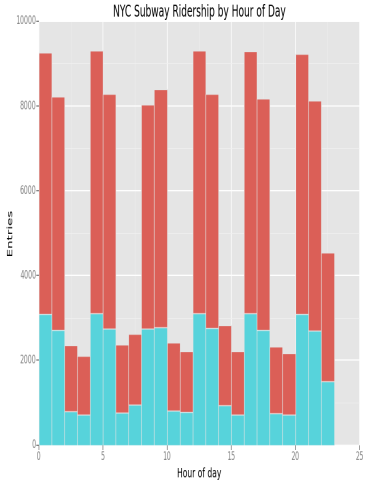
1. OLS using Statsmodels
2. Features are $\{rain, precipi, Hour, meantempi, meandewpti, meanwindspdi, UNIT, fog\}$
3. UNIT is dummy variable
4. In the last section we observed a significant increase of ridership in rainy weather. I decided to use meanwindewpti and fog because in foggy and windy weather people use subway more often. Also same reason works for the precipi. Hour is also another important factor that is so likely to influence the ridership, people using subway to go to work and school, therefore during the morning and evening we expect to have more more ridership.
5. List of coefficient:
    - rain=-1.483246e+01
    - precipi=-2.040542e+01
    - Hour=6.739965e+01
    - meantempi=-3.952262e+00
    - meandewpti =-1.870724e+00
    - meanwindspdi=2.592247e+01
    - fog=1.232434e+02
6. R^2 is equal to 0.55
7. R^2 is a statistic that will give some information about the goodness of fit of a model. In regression, the R^2 coefficient of determination is a statistical measure of how well the regression line approximates the real data points. An R^2 of 1 indicates that the regression line perfectly fits the data. In our problem we could say that our regression line predicts almost half of the real data points.

5 Visualization

First histogram shows that the distribution of ENTRIEs-hourly is not asymptotically normal so we can't use the t-test to our data set. The second histogram shows ridership for each hour of the day across all data for rainy and without rain weather. It shows there are certain pick hours of the day, morning rush hour and the rush back home hours, where the ridership increase notably



(a) Distribution of ridership in rainy weather vs without rain



(b) How ridership changes by hour in rainy and without rain weather

6 Conclusion

6.1 Statistical test

1. Based on this analysis, I see that more people ride the NYC subway when it is raining.

2. If we look at the statistical test that we implemented on the data set, we see that mean of the ENTRIESn-hourly is greater for hours with rain than without ($1,105$ vs $1,090$). Also Mann-Whitney U-test rejects our null hypothesis which it means hat the ENTRIESn-hourly sample with rain appears to be drawn from a different distribution than the ENTRIESn-hourly sample from hours without rain.

6.2 Regression analysis

1. After looking at the coefficients' p-values of the linear regression model, I realized that hour, meantempi, meandewpti, meanwindspdi, fog and most of the dummy variables are meaningful in our analysis, they have p-value less than 0.05.

2. Regression coefficients represent the mean change in our ridership for one unit of change in the predictor variable while holding rest constant. The Coefficients of the dummy variables are positive and they are bigger than 1000 but weather variables are less that 100. Also most of the dummy variables have p-values less than 0.05, they are meaningful

3. Let's use F-Test to see how well our linear model does compare to our baseline model. Our F-statistic is 349.0 and p-value for this statistic is 0.000 which means our model provides a better fit compare to the baseline model(intercept model).

7 Reflection

7.1 Shortcoming of the data set

1. The data that we are doing analysis on is a small portion of the original data set so it may not reflect all the insight from the original data.

2. We only studied the ridership during the month of May, so our study can't be generalized for ridership through whole year because it may have additional external factor for the month of May.

7.2 Shortcoming of the statistical test

1. As we already saw Hour is an important feature in the subway ridership. In our statistical test we ignored this feature. To overcome this problem I ran Mann-Whitney U test for different time of the day. To be more precise let's define X_i, Y_i to be the populations that represent ridership in rainy, without rain weather at $hour == i$. let's use Mann-Whitney U-test for each $i \in \{0, 4, 8, 12, 16, 20\}$:

   $H_{0,i} : p(x_i > y_i) = 0.5$ where x_i and y_i are the random draws from X_i, Y_i

   After computing p-value for each of this test I observed:

   $$p_8 = 0.0004,\ p_{20} = 0.024,$$

   are the only p-values less than 0.05. In other words in morning rush hour and evening rush hours we can see the significant increase of ridership in rainy weather compare to non rainy weather.

2. The p-value of our Mann-Whitney U test is 0.049 that is pretty close to our significance level and also the difference between μ_X and μ_Y is about 15 which is not significant.

7.3 Shortcoming of the regression model

1. However R^2 measures of how well this model fits a set of observations but it doesn't tell us the entire story. To have better insight we look at the distribution of residuals and we see that this distribution is asymptotically normal but our histogram has long tails with high values which indicates our model doesn't predict our data points properly.