Analyzing Subway Data

Arash Karami

August 5, 2015

1 References

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

2 Statistical Test

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

   $$H_0 : \mu_X = \mu_Y \text{ where } \mu_X, \mu_Y \text{ are the means of X, Y respectively.}$$

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

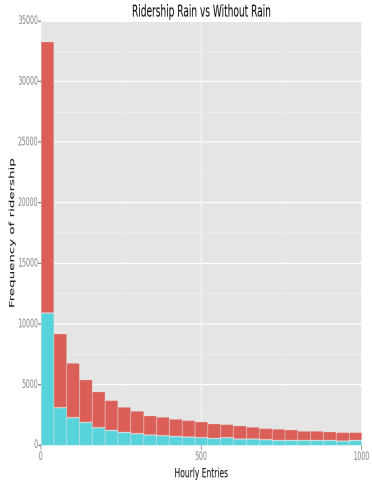
4. As the two-tail p-value is about the 0.049 considering a significance level of 5% we can then conclude there is enough evidence to reject the null hypothesis. To be more precise , the mean ridership during rainy days is higher than the mean ridership without rain, which would imply more people use NYC subway during rainy days.

3 Linear Regression

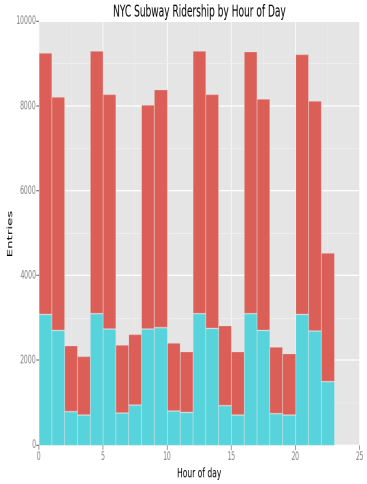
1. OLS using Statsmodels
2. Features are $\{rain, precipi, Hour, meantempi, meandewpti, meanwindspdi, UNIT, fog\}$
3. UNIT is dummy variable
4. In the last section we observed a significant increase of ridership in rainy weather. I decided to use meanwindewpti and fog because in foggy and windy weather 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 0.480386191073
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.

4 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

5 Conclusion

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 $(1,105 \text{vs} 1,090)$. 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 Reflection

- 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. Also

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.

- In my linear regression model I haven't used the feature DATEn. The below visualization shows that ridership on Monday and Sunday is higher than other days of week. I guess I would obtain a better model by adding this feature to my linear regression model.