Analyzing the NYC Subway Dataset

Questions

Section 0. References

1. Allen B. Downey.Think Stats: Probability and Statistics for Programmers.Needham, Massachusetts: Green Tea Press, 2011.
2. Diez, David M, Christopher D Barr and Mine Cetinkaya-Rundel. OpenIntro Statistics. 2nd ed. openintro.org, 2014. Web.
3. Explorable.com (Apr 27, 2009). Mann-Whitney U-Test. Retrieved Jul 06, 2015 from Explorable.com: https://explorable.com/mann-whitney-u-test
4.

Section 1. Statistical Test

1.1 I used two-tail Mann-Whitney U-test with null hypothesis set as 'probability that randomly drawn value 'ENTRIESn_hourly' on rainy day is bigger than randomly drawn value on non-rainy day equals 0.5' and p-critical value = 0.05

1.2 This test is applicable because it does not assume that the difference between the samples is normally distributed, or that the variances of the two populations are equal. Assumptions for this test are: two samples and observations in them should be independent, observations should be ordinal.

1.3 I received following results: first sample mean = 1105.45, second sample mean = 1090.28, U-statistic = 1924409167.0 and one-tail p-value =0.02499991.

1.4 Test shows significant difference in samples because two-tail p-value (0.02499991*2 = 0.499982) is lower than p-critical.

Section 2. Linear Regression

2.1 I implemented OSL method using Statsmodels library.

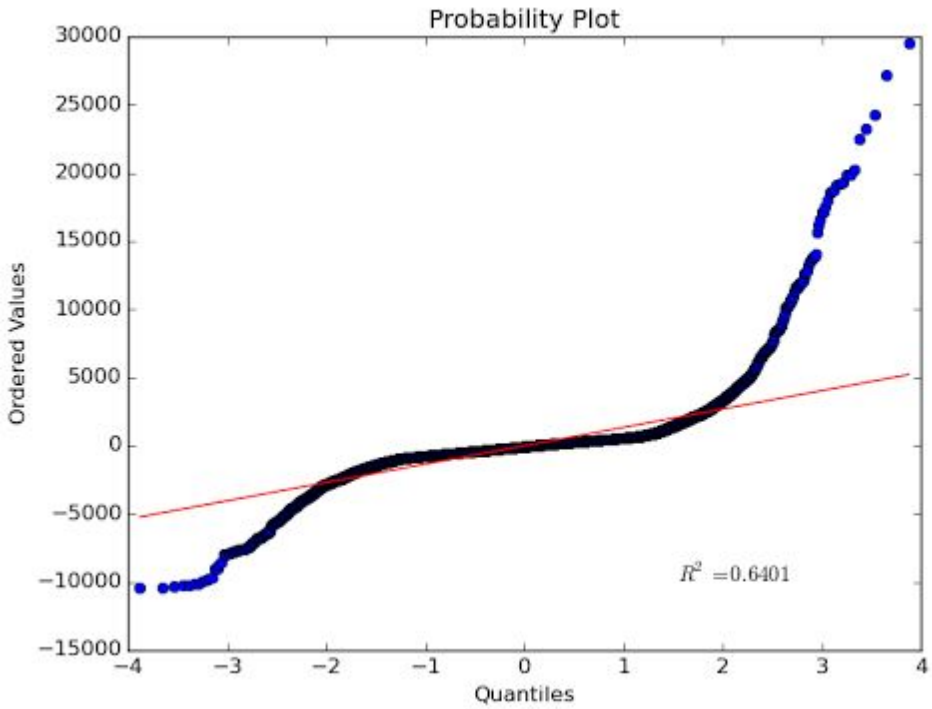
2.2 I used 'rain','precipi', 'Hour', 'mintempi','fog','meanwindspdi' as input variables and 'UNIT' as dummy variables

2.3 I included 'rain' variable because of statistical test results, then I selected 'UNIT' , 'precipi' and 'Hour' because it greatly improved my R^2 value. I used 'fog', 'mintempi' and 'meanwindspdi' because I thought that people use subway more often in foggy, cold or windy conditions.

2.4 I received following coefficients for linear regression model: (1.32412528e+01; -7.65650490e+01; 6.53714097e+01; -1.26108775e+01; 2.28599477e+02; 3.04473345e+01)

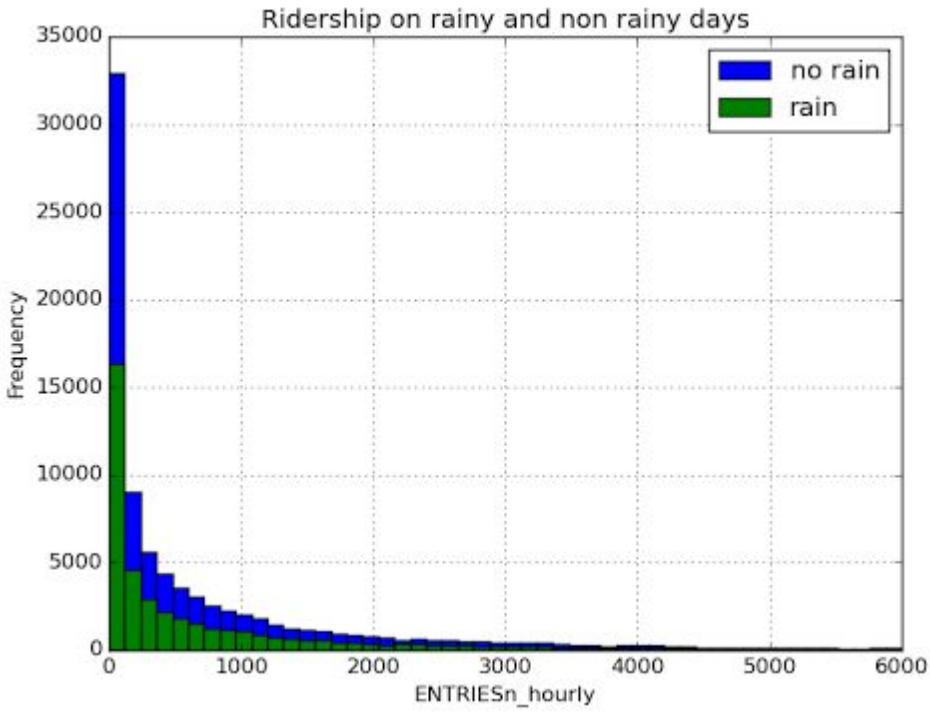
2.5 I received R^2 = 0.480858900965

2.6 This mean that 48% of data variation are explained by the model, other 52% should be explained by other variables, not included in our dataset. Probability plot shows that distribution of prediction residuals has long tails so linear model do not explain some outlying data points.



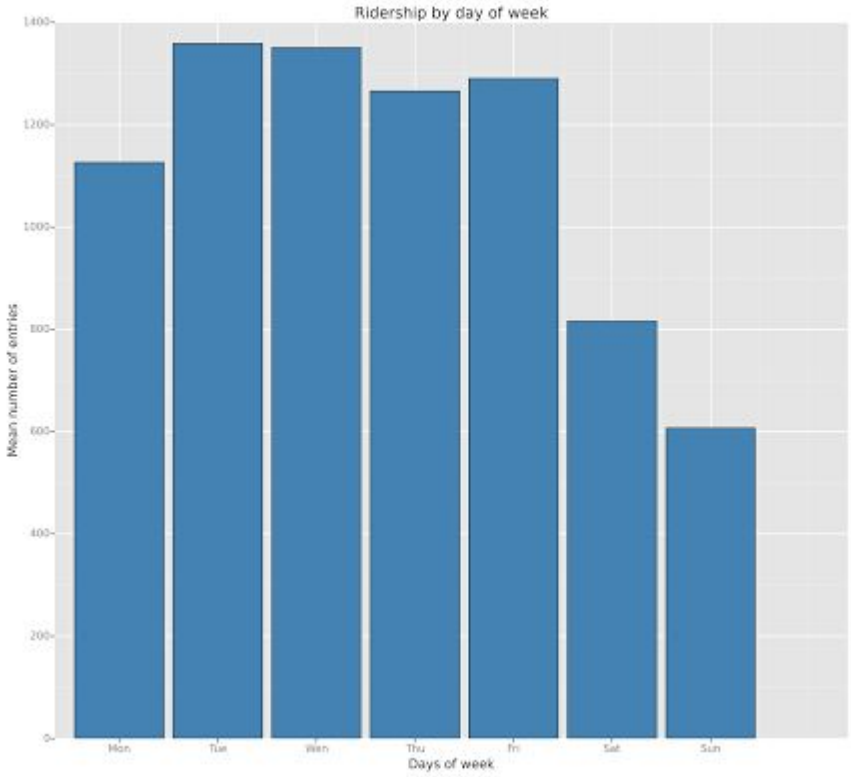
Section 3. Visualization

3.1



This graphics shows that distribution of ENTRIESn_hourly can't be described as normal for both non-rainy and rainy days, it looks more like exponential one.

3.2



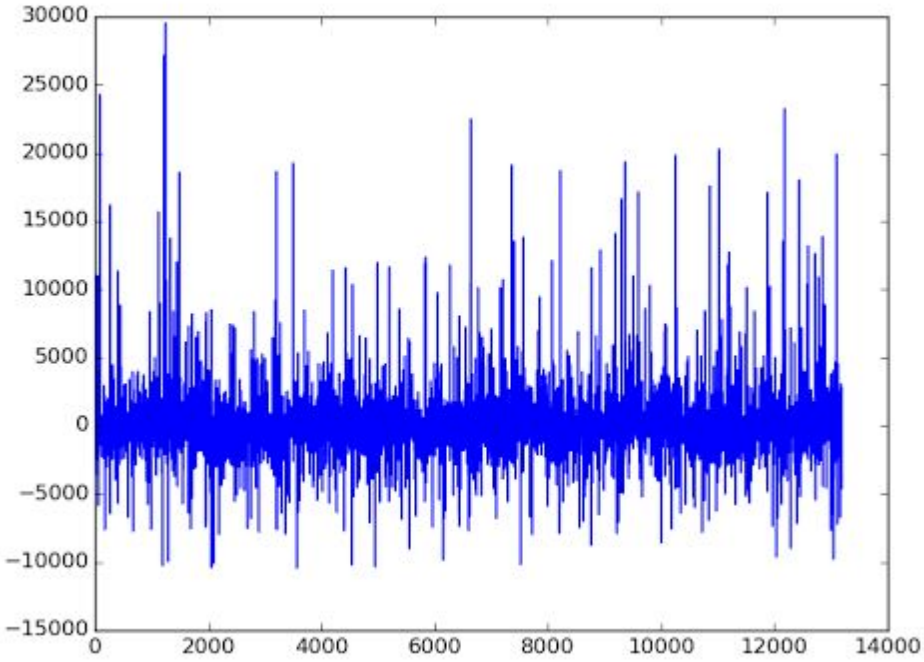
This visualization shows average number of entries on different days of week. It can be seen that more people use metro on working days, especially Tuesday - Friday.

Section 4. Conclusion

The analysis of NY Subway data shows that more people use subway then it is raining. The Mann-Whitney U test shows that there is sufficient evidence against null-hypothesis that non-rainy and rainy days samples have the same mean number of passengers, so it is possible to conclude that people use subway more on rainy days. Results of the linear regression also support this conclusion because ‘rain’ variable has positive coefficient in received linear model.

Section 5. Reflection

Linear model is not very good at explaining data variance as shown on residuals probability plot. Also residuals plot shows some cyclical patterns that suggest non-linear model may fit better for this data.



The dataset covers only one month of observations so it may not include some larger scale changes in data. Some of input variables are closely interconnected (for example ‘meantempi’ and ‘maxtempi’, ‘ENTRIESn_hourly’ and ‘EXITSn_hourly’) that can cause problems with computing accurate coefficients because of multicollinearity.
The shortcoming of chosen OSL method is that it can be very slow in case of many input variables.