

Analyzing the NYC subway dataset

Section 0:

http://www.graphpad.com/guides/prism/6/statistics/index.htm?stat_checklist_mannwhitney.htm

<http://www.statisticssolutions.com/mann-whitney-u-test-2/>

http://users.phhp.ufl.edu/akirpich/chapter08_mine.pdf

Section 1: Statistical Test

1.1 I used the Mann Whitney U Test for analyzing the NYC subway dataset.

I used the one tailed P value.

Null hypothesis: There is no significant difference in the ridership whether it is rainy day or non-rainy day.

p-critical : 0.05

1.2 As this data is positively skewed, it is decided to go for the non-parametric test such as Mann-Whitney U Test(Since Welch's T-test is good for the normally distributed data)

This test is assuming that the population from which the data of the 2 samples came from are distributed in the same way (i.e) they are following the same distribution.

1.3 I get the following results from the test

Mean of first sample(with rain) = 1105.45

Mean of the second sample(without rain) = 1090.28

U value : 1924409167.0

P value : 0.0249

1.4 As the P value(0.0249) is less than the p critical value of 0.05 we reject the null hypothesis that there is no change in ridership due to rain(i.e) There is a significant difference in the number of riders in subway during rainy and non-rainy days.

Section 2: Linear Regression

2.1 a. Gradient descent

2.2 The input variables that I used in my model are

Hour

Rain

Fog

The dummy variable used – UNIT

2.3

1. fog – I used this variable because I thought that if the fog is there people would prefer to ride in the subway than to drive in poor visibility.

2. Hour – I used this variable because during office hours most of the people will ride subway to avoid traffic and to get to office on time.

3. rain – I used this variable because if it is not raining heavily ,people would prefer to go in the subway.

2.4 $\Theta^{(\theta)}$ is the weight of the non-dummy variables. The theta values of input variables are

Fog 4.92006858e+01

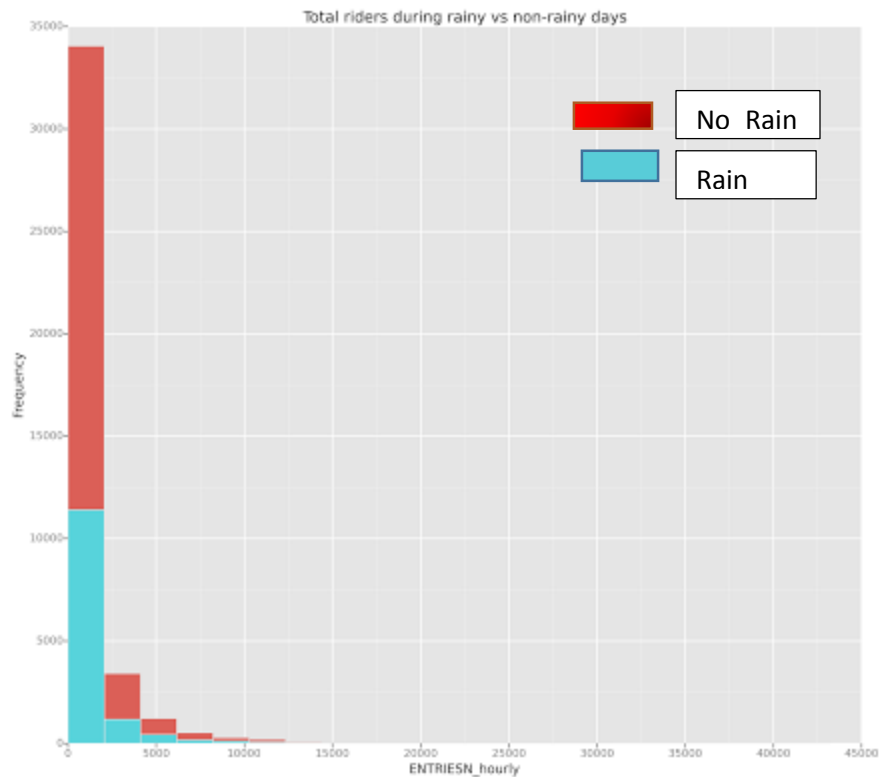
Hour 4.67811555e+02

Rain -1.16751682e+00

2.5 R^2 value = 0.46359

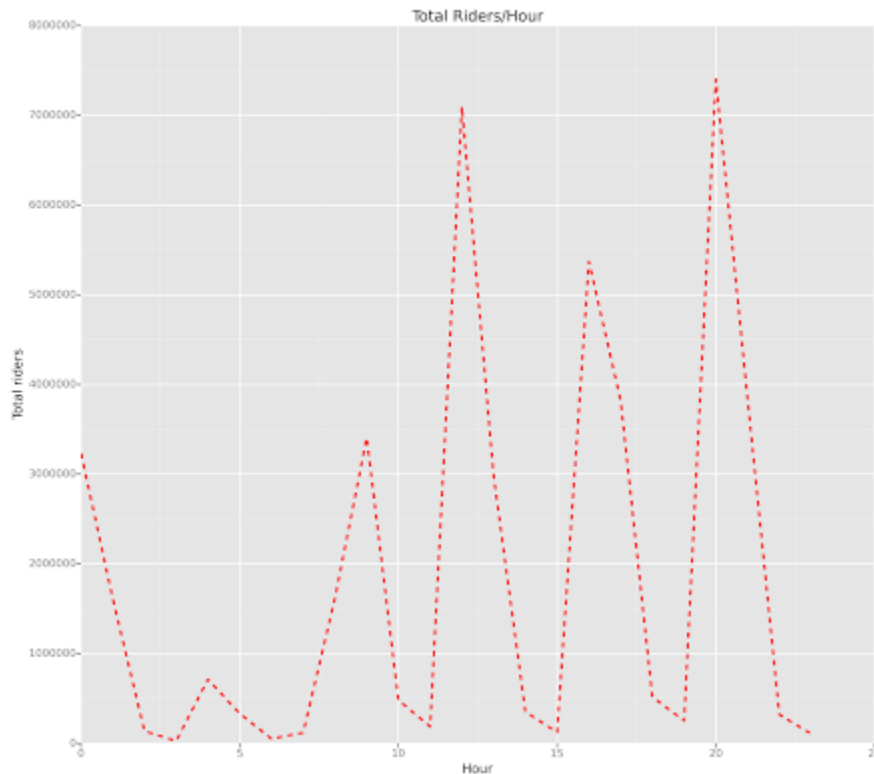
2.6 If R^2 value is closer to 1 then we can consider it as a good model. As the value is 0.46359 we can consider this model as appropriate for this dataset. And also the residuals (difference between actual and predicted value) value when plotted, follows the normal distribution. This also proves that this model is good for this data set.

3.1



The above diagram shows the histogram of the number of people riding in the subway during the rainy and non-rainy days. This clearly shows that there are more people riding in the subway in the non-rainy days than the rainy days. And also the distribution is positively skewed.

3.2



In this line graph I plotted hour Vs the number of riders. This shows that the maximum number of people ride the subway at 12 noon and at 8 P.M .Very less people are using the subway during early morning hours from 2 – 7 A.M.

4.1

More people ride the subway when it is not raining.

4.2

As the p value from Mann- whitney U test is 0.02, which is less than the p critical value of 0.05 it proves that the number of riders in the subway differs during rainy and non-rainy

days. In linear regression with gradient descent, If R^2 value is closer to 1 then we can consider it as a good model. As the value is 0.46359 we can consider this model as appropriate for this dataset. Also the residuals (difference between actual and predicted value) when plotted, follows the normal distribution. This also proves that this model is good for this data set. In gradient descent, we are using the input variables fog, rain, hour to predict the output variable 'ENTRIESn_hourly' and this also proves that change in the input variable 'rain' will have an effect on the output variable 'ENTRIESn_hourly' (i.e) if there is change in weather condition(RAIN) then there will be change in ridership (ENTRIESn_hourly) . Finally by plotting the data as histogram, it is clear that more people ride in the subway on non-rainy days than rainy days.

5.1

1. The potential shortcoming in this dataset is that it is not happening in any controlled environment. There is a great chance that there will be influence of the lurking variable. Though we say that there will be more riders in the subway in non-rainy days than rainy days, some other factor may be influencing this result.
2. A nonparametric hypothesis test(in this case, Mann-Whitney U-test) will make slightly more type II errors than the corresponding traditional method when the assumptions for that traditional method(mostly normality of data)are met.

Complicated Confidence Intervals : Nonparametric hypothesis tests are usually easy to construct and understand, but nonparametric procedures for confidence intervals can be more complicated, both computationally and conceptually

The principal disadvantage of linear regression is that many real-world phenomena simply do not correspond to the assumptions of a linear model(In this case we are considering the inputs fog,rain and hours alone has impact on the number of riders and we are assuming

there is no impact by the other factors). In these cases, it is difficult or impossible to produce useful results with linear regression.

Another disadvantage is that if there are some outliers, it will have a bad impact on the model. Then we will not get the line of best fit. We would be predicting entirely wrong values.