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 two 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(As it is a two-tailed P value, 0.025 on both the sides)

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 or equal to the p critical value of 0.025 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

The dummy variable used – UNIT

2.3

1. 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.

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

2.4 Theta(θ) is the weight of the non-dummy variables. The theta values of input variables are

```
Hour  4.67703941e+02
```

Rain 2.03470329e+01

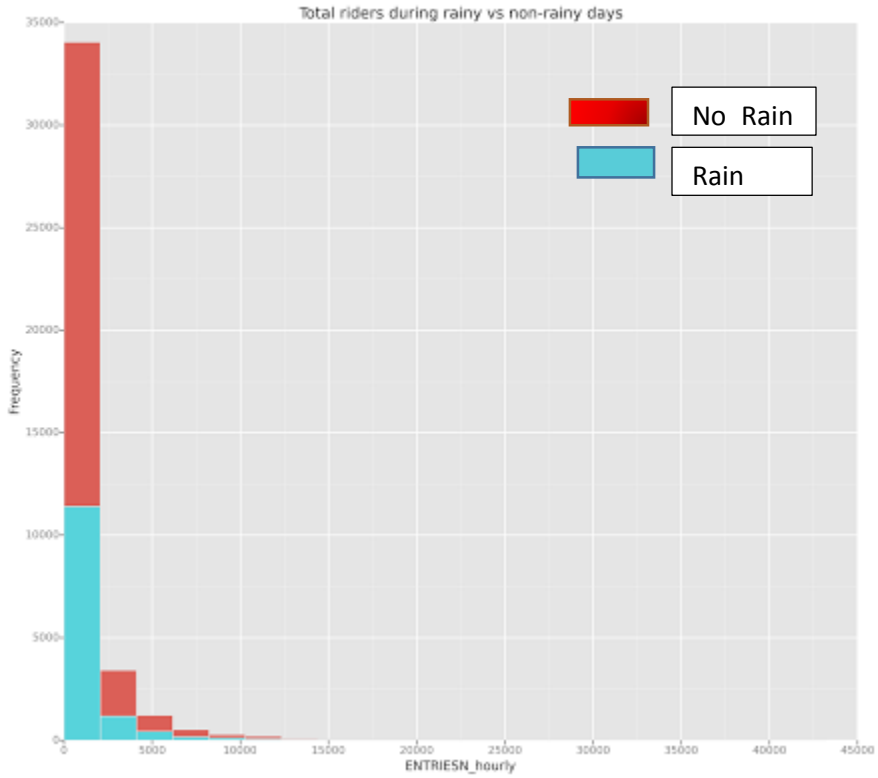
2.5 R^2 value = 0.463247669262

2.6 *R-Square*, also known as the *Coefficient of determination* is a commonly used statistic to evaluate model fit. The *R-square* value is an indicator of how well the model fits the data (e.g., an *R-square* close to 1.0 indicates that we have accounted for almost all of the variability with the variables specified in the model). *R-square* is 1 minus the *ratio of residual variability*. When the variability of the residual values around the regression line relative to the overall variability is small, the predictions from the regression equation are good. If we have an *R-square* of 0.46(as in this case) then we know that the variability of the Y values around the regression line is 1-0.46 times the original variance; in other words we have explained 46% of the original variability, and are left with 54% residual variability (i.e) we have accounted for nearly half of all of the variability with the variables specified in the model.

The degree to which two or more predictors or *X* variables (in this case: rain,hour) are related to the dependent or Y variable (in this case: ENTRIESn_hourly)is expressed in the correlation coefficient *R*, which is the square root of *R-square*. In multiple regression, *R* can assume values between 0 and 1. To interpret the direction of the relationship between variables, look at the signs (plus or minus) of the regression or *B* coefficients or Theta(θ). If a *B* coefficient is positive, then the relationship of this variable with the dependent variable is positive ; if the *B* coefficient is negative then the relationship is negative . Of course, if the *B* coefficient is equal to 0 then there is no relationship between the variables. Here in this model we are having a positive theta value for ‘rain’ and ‘Hour’. This confirms that the input variables chosen in this model is having relationship with the output

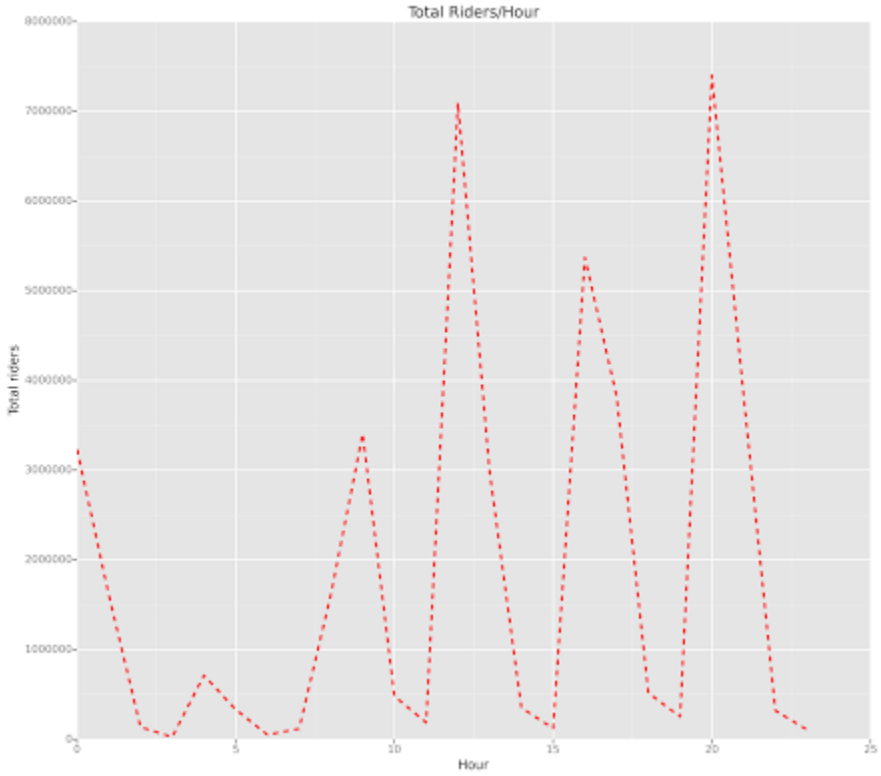
variable(ENTRIESn_hourly) and the R^2 value of 0.46 confirms that we have explained nearly half of the variability with the variables specified in the model. Thus we can say that this model fits for this data set.

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


Total Riders/Hour
Total riders
Hour
8000000
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
0
5
10
15
20
25

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 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 0.46324 we can consider this model as appropriate for this dataset. 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. In gradient descent, we are using the input variables rain, hour to predict the output variable 'ENTRIESn_hourly' and theta value for rain is positive(2.03470329e+01). If theta value is positive then it proves that there is a positive relationship between the variable rain and ENTRIESn_hourly (i.e) if there is rain then there will be more number of riders and if there is no rain then there will be less number of riders. This proves that more number of riders will be riding the subway during rainy days.

To confirm this, the average riders during rainy days(1105.45) is greater than the average riders during dry days(1090.28).

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.