Quality of analysis

# statistical tests and linear regression

I have used the **Mann-Whitney-U test** for the analysis of the data, for finding out average (mean) entries of number of people using the subway on the rainy vs non-rainy day.

The Mann-Whitney-U test is non-parametric equivalent test to the 2 independent samples t-test.

Because it is not yet known, nor hypothesized, which data set would be higher or lower, a two-tailed test here is apt. In using the Mann-Whitney U test, the null hypothesis is that the two populations are the same, or simply put, that rain has no correlation with ridership. The p-critical value used was 0.05, or 5%.

Two tail tests are used here as they reduce the probability of type-1 error (rejecting a true null hypothesis).

Why I have considered Mann-Whitney-U test for this data set?

The data was not normally distributed as we have seen in the visualization section, neither the data for the rainy days nor the data for the non-rainy days.

Records of the data set were independent of each other. Sample is large (size is greater than 20).

As such, a non-parametric test such as Mann-Whitney U is a good fit, while a test such as Welch’s two-sample t-test is not.

**Linear Regression**

I have used the gradient descent algorithm is used. I have used the default values i.e. alpha = 0.5 and 75 iterations. Features used includes rain or no rain, precipitation, mean wind speed, hour, weekday and mean temperature. The dummy variable called UNIT is introduced.

These features are used because continuous variables in this type of data set provide more information.

Weekday influence more the ridership rather than fog or rain.

The R^2 value drawn from the regression model is essentially a quantitative measure of goodness of fit and the percentage of variance. R^2 value is found to be nearly equal to 0.56.

**Results drawn from the above test:**

Differences were observed in the ridership between the rainy days (mean entry: 1105) and non-rainy days(mean entry: 1090).

The p- value obtained is 0.025.

Hence, a conclusion can be drawn that ridership is different on rainy or non-rainy day with 95% confidence that null hypothesis is false.

More people ride the subway when the weather outside is bad (or its raining).

The positive value of the coefficient indicates that the ridership increases as the rain increases.

The effect of rain or no rain is significantly small in comparison to the weekday and the weekend days.