Analyzing the NYC Subway Dataset

BY: SUDEEP BHATTARAI

Statistical Test

Which statistical test did you use to analyze the NYC subway data? Did you use a one-tail or a two-tail P value? What is the null hypothesis? What is your p-critical value?

I used Mann-Whitney U test to analyze the NYC subway data, using two tailed test. The null hypothesis is - there is no effect in ridership due to rain. The p-critical value used was 0.05.

Why is this statistical test applicable to the dataset? In particular, consider the assumptions that the test is making about the distribution of ridership in the two samples.

Mann-Whitney test is used to compare the significance of two different samples having not normal distribution, unequal variations and sample sizes. Population of Ridership on rainy and not rainy days are not normally distributed. These sample also have different sample sizes and variances. In this case Mann-Whitney test is most applicable.

What results did you get from this statistical test? These should include the following numerical values: p-values, as well as the means for each of the two samples under test.

```
Mean with rain: 1105.446,

Mean without rain: 1090.278,

U-stat: 1924409167.0,

P-value: 0.02499 (0.4998: two tailed)
```

What is the significance and interpretation of these results?

The null hypothesis is false for 95% confidence interval. With p-value of 0.4998, which is less than 0.05, Mann-Whitney statistical test signifies that there is difference between ridership on rainy and not rainy days.

Linear Regression

What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model

I used gradient descent algorithm to compute the coefficients theta. For learning rate of 0.1, I choose total number of iterations of 50, causing cost history converge to minimum.

What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?

Features I used in my model are rain, precipitation, hour, mean temperature, mean wind speed and fog. I also used dummy variables for UNIT to represent different subway units. Each unit is prefixed as unit to represent dummy variables. If the ridership entries belongs to the unit, it has the value of 1 and all other have the value of 0.

Why did you select these features in your model? We are looking for specific reasons that lead you to believe that the selected features will contribute to the predictive power of your model.

I selected are rain, precipitation, hour, mean temperature, mean wind speed, fog and dummy variables for UNIT as features in my model. I added and removed features and test on R square value. I believed that, fog and wind temperature may also effect on ridership in spite of rain, so I added those features which resulted in slight increase of R square value.

Unit was used as dummy variables because units with greater traffic has more ridership.

Hour was chosen because there is high positive correlation between hour and ridership (np.corrcoef(Hour, ridership): 0.17543045) which drastically improve the value of R square.

What are the coefficients (or weights) of the non-dummy features in your linear regression model?

Coefficients of the non-dummy features are:

```
-9.84753832e+00, -2.04970559e+01, 4.64048265e+02, -5.22847669e+01, 6.25669951e+01,
6.87402445e+01
```

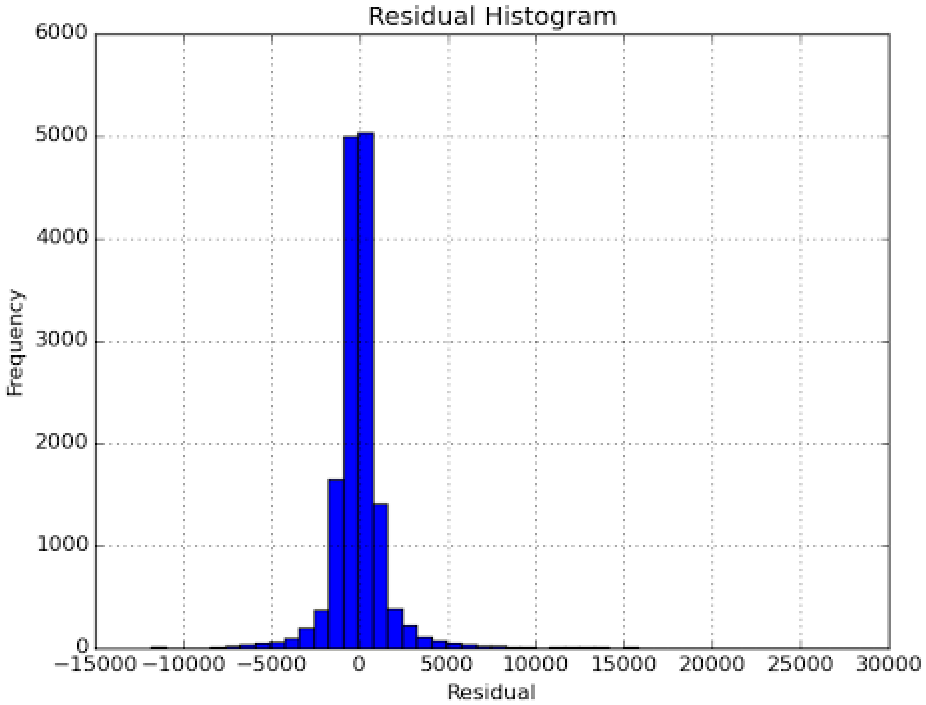
What is your model's R2 (coefficients of determination) value?

```
R square value is 0.465065352458
```

What does this R square value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R Square value?

R square = 0.465065352458, which means that46.5% of the total variation can be explained by the linear relationship between the features and Ridership value.

For checking if the model is appropriate, let's plot the residuals:

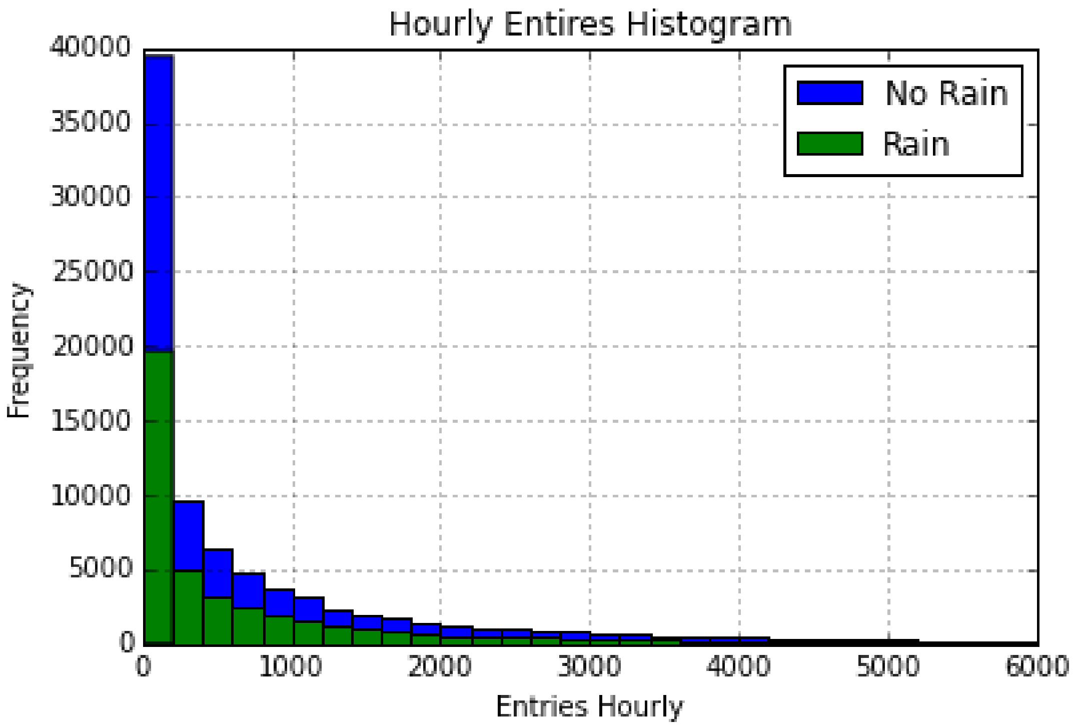


The residuals should not contain any predictive information. The residuals should be centered on zero throughout the range of fitted values.

From the figure above the residuals are normally distributed, and centered to zero and most of them are closed to +/- 5000. So, I think this linear model is appropriate for this dataset.

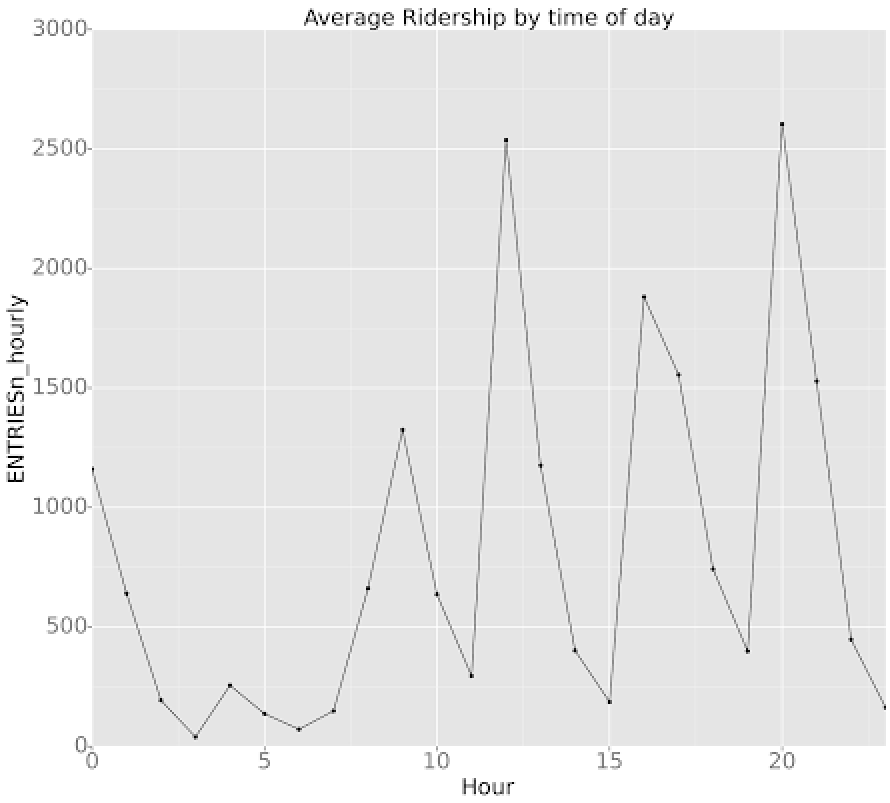
Visualization

One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.



This histogram shows Hourly Entries for Rainy and Non Rainy days. Both of the distributions are not normal. From the histogram we can see that that the number of subway entries in rainy days is less than the number of subway entries in not rainy days. But, however we cannot say that the Ridership is more in non-rainy days because number of rainy days are less than the number of not rainy days.

One visualization can be more freeform. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like.



Graph shows average subway entries in different times of day. We can see that most subways entries are occurring at 12 noon and 8 PM and least subway entries are from 3 AM to 7 AM. Peaks are also observed in office hours (9 AM and 4-5 PM).But it is interesting to know that ridership is more in noon and 8 PM than in office hours.

Conclusion

From your analysis and interpretation of the data, do more people ride the NYC subway when it is raining or when it is not raining?

After performing Mann-Whitney U test, with 95% confidence, we can say that ridership is different in rainy and non-rainy days. I calculated the average ridership on rainy and non-rainy days and found that, average ridership for rainy day is more than the average ridership in non-rainy days. So, from both analysis, it can be concluded that, more people ride NYC subway when it is raining than when it is not raining.

What analyses lead you to this conclusion? You should use results from both your statistical tests and your linear regression to support your analysis

While preforming linear regression, R square value was around 0.465, which shows that 46.5% of the total variation can be explained by the linear relationship between the features and Ridership value. This R square value is not relevant for this analysis because this R square value depend on all features that were used. But however, we can see a small positive correlation between rain and ridership (np.corrcoef(rain, ridership): 0.00306159), so we can say, rain does have some effect on ridership.

On the other hand, as population distribution for ridership in rainy and non-rainy days are not normal, Mann-Whitney U test was performed for the null hypothesis that there is no difference in ridership in rainy and non-rainy days, which was ultimately rejected with 95% confidence interval. This showed that, there is difference in ridership in rainy and non-rainy days. From the mean calculation, I got ridership in rainy days is more than ridership in non-rainy days. Using statistical test, it can be conclude that, ridership in rainy days is more than ridership in non-rainy days.

Reflection

Please discuss potential shortcomings of the methods of your analysis, including:

Dataset,

Analysis, such as the linear regression model or statistical test

The data size was small, there would be better analysis if the sample size was large. It would have been better to take random sample for whole year to find out the better analysis of number of ridership in rainy days and non-rainy days.

Unit was the main factor for variation of number of ridership. In some Subway units there are more ridership and in some there are not at all. But, Mann-Whitney U test didn't consider it and statistical test was performed based on Rain vs No Rain only. As the sample size was small, there could be high difference on ridership in rainy and non-rainy days on only one or two Units having large number of Entries just by chance.

The increase of features, or application of some other polynomial regressions may have been increased the accuracy of linear regression model. Also, there may have been relationship between different features, where the comparison could be made. Both features linearly dependent to each other may have been included. Dependent features can be excluded while performing analysis so that same accuracy can be achieved with less number of features.

Histogram only may not be the best choice for judging the distribution of residual. More sensitive graph like normal probability plot can be used to study residual behavior.

References

http://mathbits.com/MathBits/TISection/Statistics2/correlation.htm

https://bespokeblog.wordpress.com/2011/07/11/basic-data-plotting-with-matplotlib-part-3-histograms/

http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test