Analyzing the NYC Subway Dataset

Section 0. References

GraphPad Software Inc. Interpreting results: Mann-Whitney test. Available from: < http://www.graphpad.com/guides/prism/6/statistics/index.htm?how_the_mann-whitney_test_works.htm >. [3 July 2015].

Wikipedia. Coefficient of Determination. Available from: < http://en.wikipedia.org/wiki/Coefficient_of_determination>. [3 July 2015]

Section 1. Statistical Test

1.1 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?

In the analysis of the NYC subway data, the following was analyzed: Test whether the following samples of the dataset were similar: subway ridership when it was raining and ridership when it was not raining. The null hypothesis was to test whether the two samples came from the same population, meaning if the distribution of the samples were statistically similar.

In the exploratory analysis of the data it was discovered that the ridership-dataset was non-normally distributed. Therefore the null hypothesis was tested using the Mann-Whitney U-test which is a good fit for testing non-normally distributed samples.

Because the equality of distributions were tested, a two-tailed p-value was used. The p-value of the test approximates to the following:

$$p \approx 0.05$$

The confidence level of the test was decided as the following:

$$a = 0.95$$

Therefore:

$$p \leq (1 - a)$$

This means that the p-value of the test fell within the critical region, therefore the null hypothesis has to be rejected.

1.2 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.

Because the exploratory analysis of the ridership-dataset concluded that the samples were distributed non-normally, a statistical test that performed well with non-normal samples was needed.

To this end, the use of Mann-Whitney U-test was a natural choice as it is generally concluded in the scientific community that a Mann-Whitney U-test has a significantly better performance on non-normal data than, for example, the t-test.

1.3 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.

$$\mu_{rain} = 1105.45$$
$$\mu_{norain} = 1090.28$$
$$U = 1924409167.0$$
$$p = 0.02499 * 2 \approx 0.05$$
$$a = 0.95$$

Therefore since **p ≤ (1 - a)**, we have to reject the null hypothesis of ridership during rain and not during rain being identically distributed. Therefore it must be concluded that the means of the two samples have significant differences in their distribution and therefore can not be regarded as part of a same population.

1.4 What is the significance and interpretation of these results?

The results of the Mann-Whitney test indicate that the NYC subway ridership distributions will be different when it is raining and when it is not raining.

Section 2. Linear Regression

2.1 What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model:
Gradient descent (as implemented in exercise 3.5)
OLS using Statsmodels
Or something different?

For linear regression, the use of OLS/Normal equation offers a simpler solution to reach a global optimum for the model parameters, without the need for the learning rate **α**.

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

The features I selected were: 'precipi', 'meantempi', 'fog', ‘meanwindspdi’, ‘unit’, ‘Hour’, ‘weekday’. Of these, ‘unit’, ‘weekday’ and ‘Hour’ were implemented as dummy features as these features can be defined as categorical. Hour could have also been represented as an ordinal variable but it didn’t seem to follow a linear function so it was necessary to use it as a dummy variable (the fact that linear regression gave significantly better results with Hour as a dummy variable seemed to support this decision).

2.3 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.
Your reasons might be based on intuition. For example, response for fog might be: “I decided to use fog because I thought that when it is very foggy outside people might decide to use the subway more often.”
Your reasons might also be based on data exploration and experimentation, for example: “I used feature X because as soon as I included it in my model, it drastically improved my R2 value.”

The features above were initially selected by intuition. The intuition was that subway riders may choose to use the subway based on whether it's foggy, hot/cold, windy or significant precipitation (raining, snowing etc.). That included all the weather-related features in order to determine how weather affects the subway ridership. Additionally, the hour of the day and day of the week was assumed to have an impact on subway ridership, as was the subway station itself.

After some experimentation, it was deemed that the features 'rain', 'precipi', 'meanpressurei', 'meantempi', 'fog', 'meanwindspdi' had very little effect on the performance of the model. When using only features 'Hour', 'weekday' and 'unit', the R^2-value didn't drop at all.

In fact, only using the 'unit' feature leads to a R^2-value of 0.418 which is enough to pass the requirements of the exercises.

To conclude, it seemed that the weather-related features did not really affect the R^2-value of the model at all.

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

Table 1 - Results summary of non-dummy features

feature	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
const	1,328E+13	1,72E+13	0,774	0,439	-2,03E+13	4,69E+13
rain	-72,6115	13,478	-5,387	0,000	-99,028	-46,195
precipi	-12,6805	15,648	-0,810	0,418	-43,351	17,990
meantempi	-8,7754	1,047	-8,384	0,000	-10,827	-6,724
meanpressurei	-273,6103	37,756	-7,247	0,000	-347,612	-199,609
meanwindspdi	-7,8268	3,231	-2,422	0,015	-14,160	-1,493

Table 1 displays the results summary of the weather-related non-dummy features. From the table it seems that the weights (or coefficients) of the weather-related features are all negative. Therefore in the regression model, the dependant value y (ridership) would decrease as the weather-values increase. But when comparing to the coefficients of the dummy features, the effect of these features to the model are completely negligible. The P>|t|-value of the table also indicates that the rain-feature has no effect on the dependant variable.

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

R^2 = 0.514

2.6 What does this R2 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 R2 value?

R^2 or coefficient of determination determines how well the model in question represents the data. In other words the R^2-coefficient represents the percentage of the variation in the dependant variable y that can be explained by its correlation with the features (or independent variables) in the proposed model. The R^2-value of 1 would mean the model would explain the variation of y

perfectly. Even though the value 1 is practically impossible for real-world data, this model falls quite far from this. Therefore it would seem that this model is a poor fit for the data.

In Figure 1 the residuals of the regression model are plotted. In the figure it is apparent that there are quite large residuals in the model. The residual plot also seems to follow a cyclic pattern where positive residuals are followed by negative residuals. Data that follows a pattern like this is not possible to reliably represent with a simple linear model. All in all, the residual plot supports the fact that the linear regression model is not a good fit for the data.

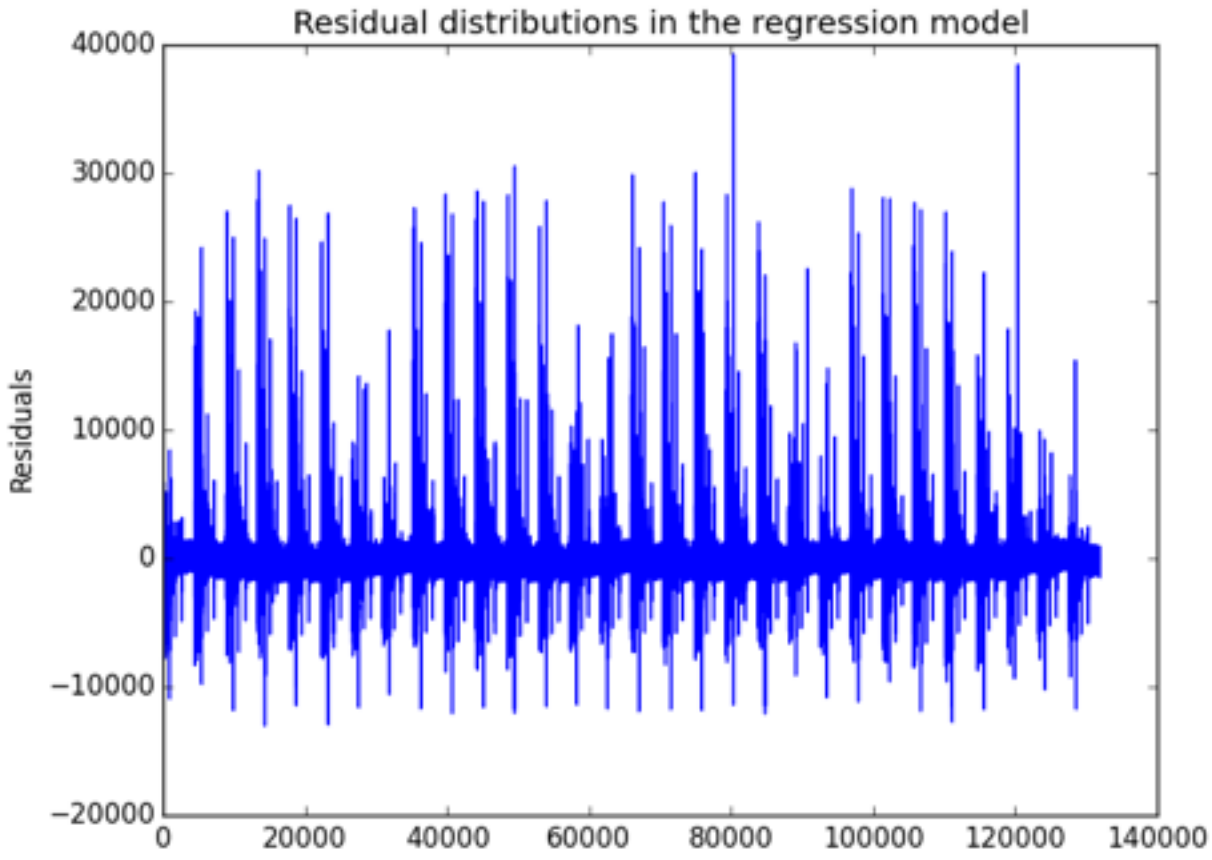

Figure 1 - Residual plot of the model

Section 3. Visualization

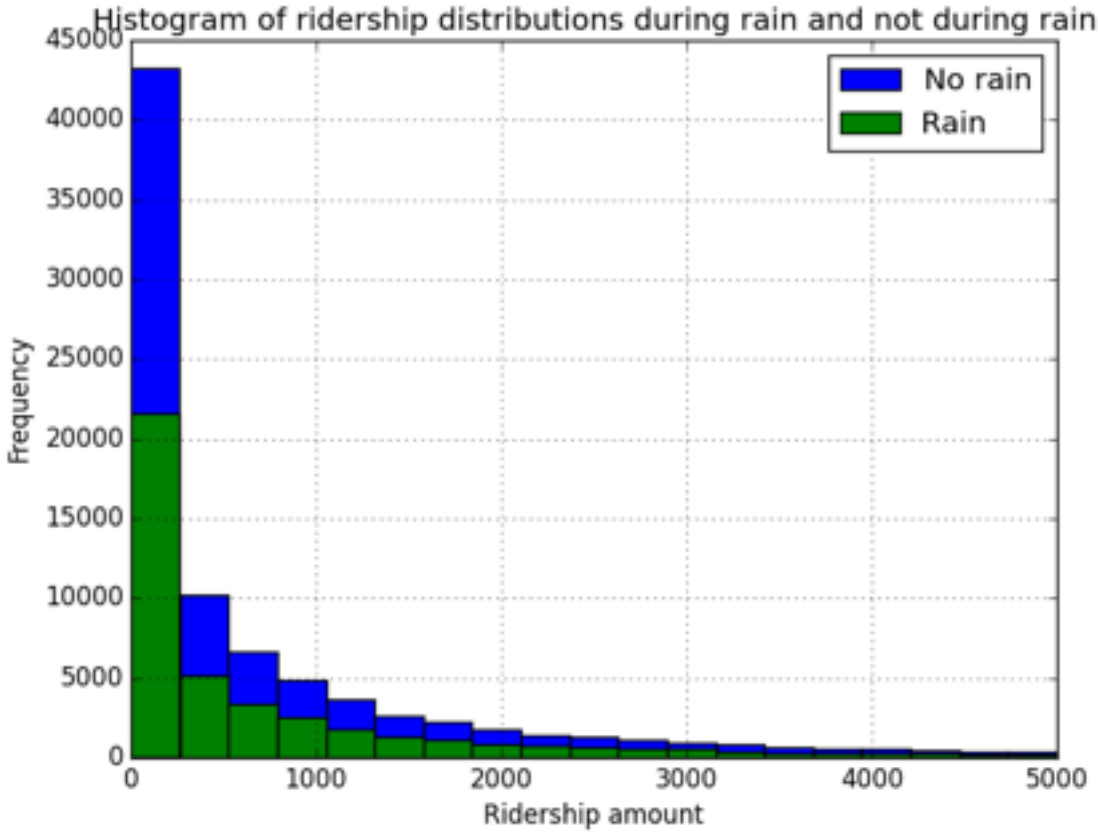

Figure 2 - The distribution of ridership in rain

Figure 2 represents the distributions of ridership amount frequencies when raining and when not raining. The visualisation illustrates quite well that the distributions of the two samples are very similar.

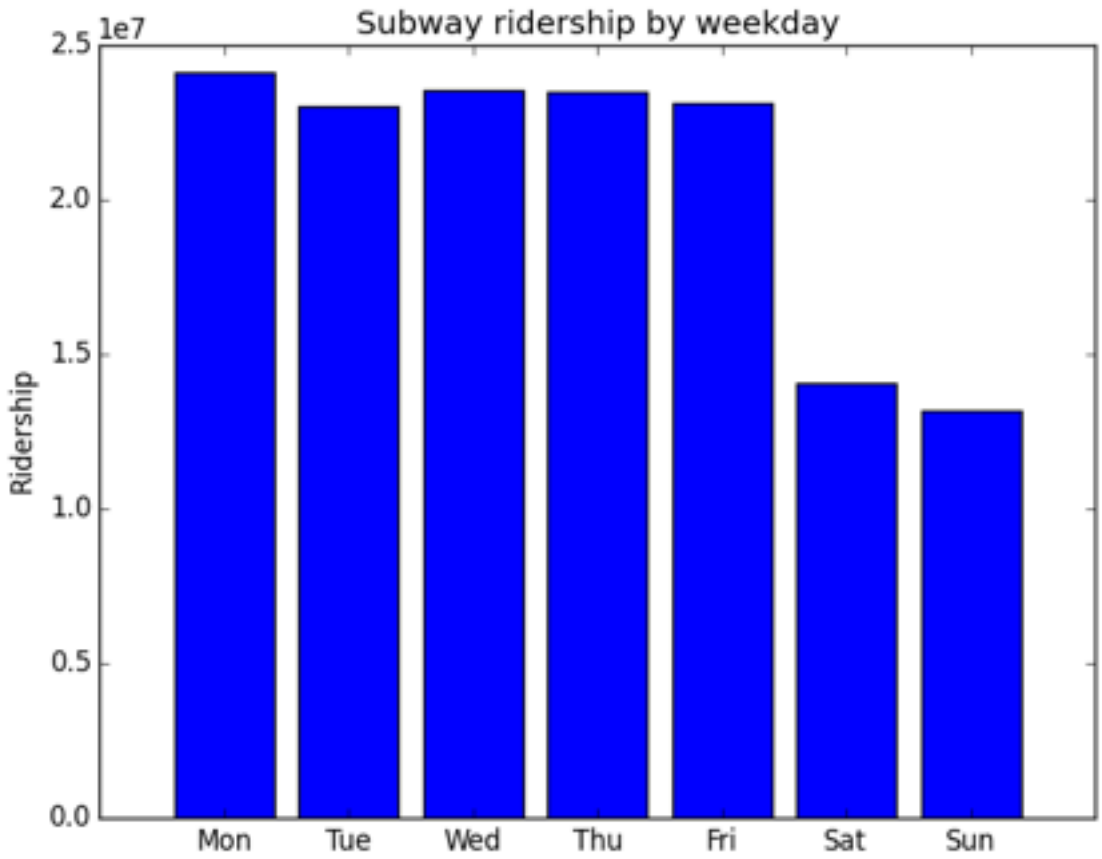

Figure 3 - Subway ridership by weekday

Figure 3 displays the subway ridership by weekday. It seems that the ridership has differences between weekdays and weekend days. According to this visualisation and the linear regression results earlier, it seems that the weekdays does have an effect on subway ridership. But the effect is only significant between weekdays and weekend days, not individual days.

Section 4. Conclusion

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

In the analysis, the visualizations and the regression models seemed to indicate that rain has little to no effect on subway ridership. In Figure 2 the distributions of ridership during and not during rain seemed very similar. Additionally, the features ‘rain’, ‘precipi’, ‘meantempi', ‘meanpressurei’, ‘meanwindspdi’ had very little effect to the R^2-value in linear regression. It should, however, be noted that the linear regression only gave an R^2-value of 0.514, which indicates that the model is not very accurate.

Even though the effect of weather to regression model was very close to none, looking at the weather coefficients in Table 1 it would seem that the

The Mann-Whitney test, however, did not support this claim. It seemed that according to test, with a 95% confidence level the null hypothesis had to be rejected as can be seen in the results of the test below.

$$\mu_{rain} = 1105.45$$
$$\mu_{norain} = 1090.28$$
$$U = 1924409167.0$$
$$p = 0.02499 * 2 \approx 0.05$$
$$a = 0.95$$

Therefore, according to the Mann-Whitney test, the distributions of ridership did differ depending on whether it rained or not. (GraphPad software inc) defines the null hypothesis of Mann-Whitney test in their documentation quite handily, which is as follows:

If the groups are sampled from populations with identical distributions, what is the chance that random sampling would result in the mean ranks being as far apart (or more so) as observed in this experiment?

The identical distribution is hence defined so that there is a 50% probability that an observation from a value randomly selected from one population exceeds an observation randomly selected from the other population. Therefore the result of the test indicates that the distributions of ridership will be different depending on whether it's raining or not raining.

Since the Mann-Whitney test determined that the sample distributions are different, the question whether ridership differs depending on rain is proven as true. And since the Mann-Whitney test proved that the distributions are different and because μ_{rain} is higher than μ_{norain}, I have to conclude that people ride the subway slightly more during rain.

Section 5. Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis, including: Dataset, Analysis, such as the linear regression model or statistical test.

When using linear regression, the ridership-values in the dataset had heavy correlation with the subway station, but had little correlation with anything other feature in the dataset except time of day and whether it was weekend or not.

Due to this fact the regression model was a very poor fit to determine the variation of ridership in the data, as determined by the low R^2 value and the large residuals in Figure 1.