Analyzing the NYC Subway Dataset

Mohamed Khodeir

August 4, 2015

1 References

1. Udacity Intro to Data Science Downloadable - Understanding the Mann-Whitney U-test
2. http://people.duke.edu/ rnau/rsquared.htm
3. http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.probplot.html
4. http://en.wikipedia.org/wiki/Q%E2%80%93Q_plot

2 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?

I used the one-tailed Mann-Whitney U-test to compare the distributions of hourly entries when raining against hourly entries when not raining.
The null hypothesis states that hourly entries from either sample are equally likely to be greater than each other. That is: $P(x > y) = P(y > x) = 0.5$ where $y \in Y$ and $x \in X$. (i.e. Whether or not it is raining does not affect hourly entries positively or negatively.)
At an alpha level of 0.05 and with sample sizes of $n_x = 44104$ (with rain) $n_y = 87847$ (without), we need to use the normal approximation to the distribution of the U statistic to calculate the p-value. The critical U-value (with p = 0.025 for a one-tailed test) can be obtained by standardizing the approximating normal distribution using mean $\mu = n_x n_y/2$ and standard deviation $\sigma = \sqrt{(n_x n_y (n_x + n_y + 1)/12)}$, and using the z-table to perform the calculation: $U_{critical} = z_{critical} * \sigma + \mu$

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.

The two samples do not look normally distributed when shown on a histogram, so the non-parametric Mann-Whitney test is more appropriate than the t-test, which assumes normally distributed populations.

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.

The results I obtained were in favor of rejecting the null hypothesis at an alpha level of 0.05.

$$\mu_{with\_rain} = 1105.45$$

$$\mu_{without\_rain} = 1090.28$$

$$U_{without\_rain} = 1924409167$$

$$p = 0.025 <= 0.025$$

1.4 What is the significance and interpretation of these results?

The conclusion then, is that we are 95% certain that hourly entries are higher in the rain than otherwise.

3 Linear Regression

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

In order to compute the regression coefficients, I used gradient descent on the 'sum of squared errors' cost function. I later checked my results using statsmodel's OLS implementation.
After the specified number of iterations, a simple scalar product of each data case's features with the coefficients produces the prediction for that entry.

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

These are the features I used:
Dummy Variables:
'clear', 'weekday', 'unit'
Numerical Variables:
'hour', 'wspdi', 'meanwspdi', 'tempi', 'meantempi', 'precipi', 'meanprecipi'

2.3 Why did you select these features in your model?

Firstly, I tried features that seemed appropriate/meaningful. This set included ('rain', 'weekday', 'overcast', 'fog', 'day_week', 'hour', 'meantempi').
However, I got better results (in terms of R^2) by including all possible features. I then tried to reduce the set of features by greedily removing those ones that made less than a threshold (0.001) of difference to my r-squared score.
Interestingly, the p-values (obtained from statsmodel's OLS) for each of these features is near 0.

2.4 What are the parameters of the non-dummy features in your regression model?

'hour' : 724.58 'wspdi' : 78.09 'meanwspdi' : -103.53 'tempi' : 116.47 'meantempi' : -324.25 'precipi' : -139.76 'meanprecipi' : 116.47

2.5 What is your model's R^2 value?

$$R^2 = 0.487$$

2.6 What does this R^2 value mean for the goodness of fit for your regression model? Do you think using this linear model to predict reidership is appropriate for this dataset, given this R^2 value?

This R^2 means that roughly half of the variance in the dependent variable can be explained by variance in the independent variables. I think predictions made by this model could be useful to a degree.
In terms of absolute residuals, the model seems to be less reliable when the turnstile (UNIT) in question has higher than average ridership, as shown in figure 3.
Looking at the histogram in figure 1, it seems the 95% confidence interval (assuming normality) for the prediction should be around 5000 wide, so for a given turnstile, our prediction is likely to be within 2500 of the true value.
However this is not necessarily the case. A comparison of the distribution of the residuals to the normal distribution (figure 2), shows that both the largest and smallest values in the residuals distribution are more extreme than would be expected under normality. This effect is much more pronounced on the positive tail of the distribution, so that largest residuals are much larger than would be expected under a normal distribution.
This means that the residuals distribution is long-tailed, and we cannot easily/robustly predict a confidence interval.

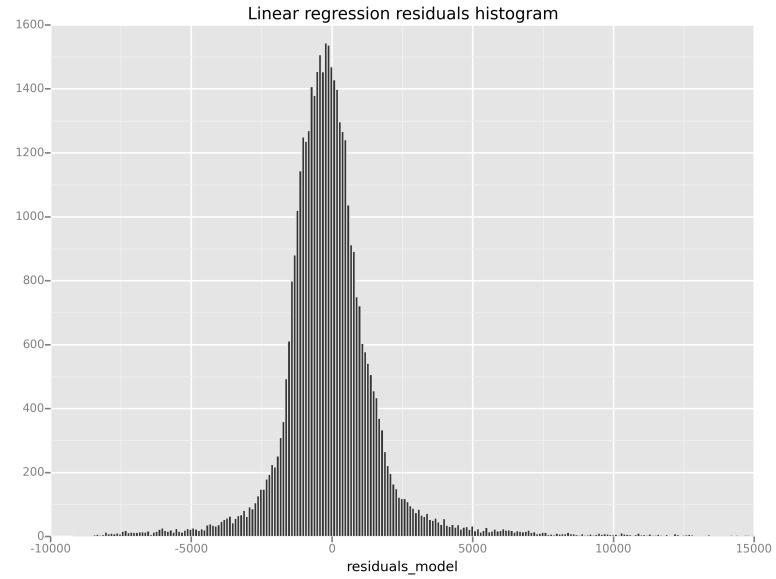


Figure 1: Histogram of residuals from linear regression model.

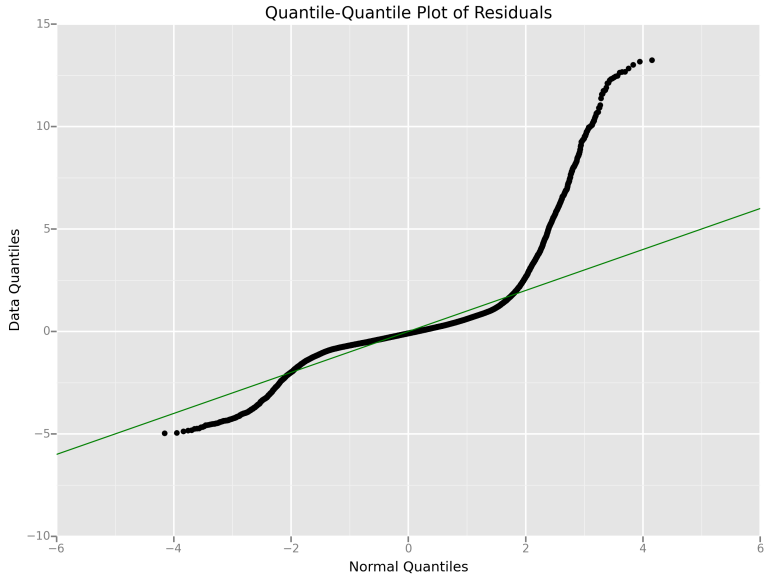


Figure 2: Quantile-quantile plot of the residuals as compared to the respective normal distribution.

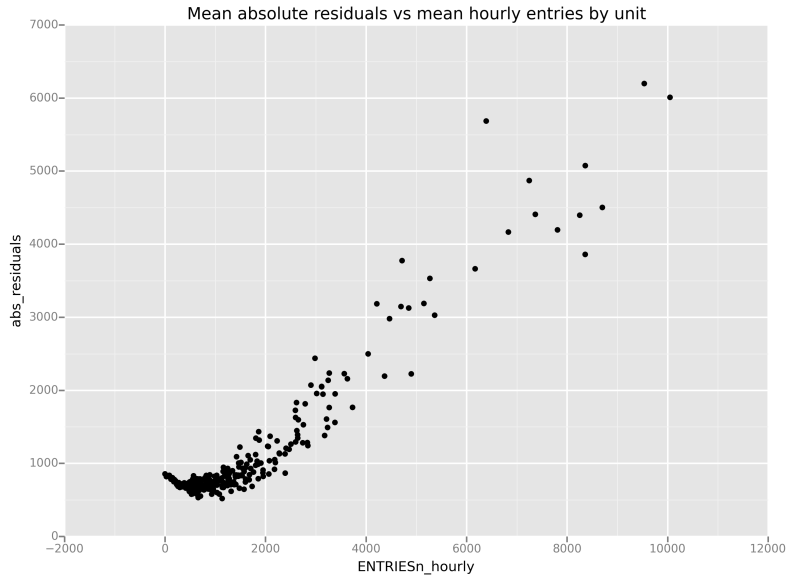
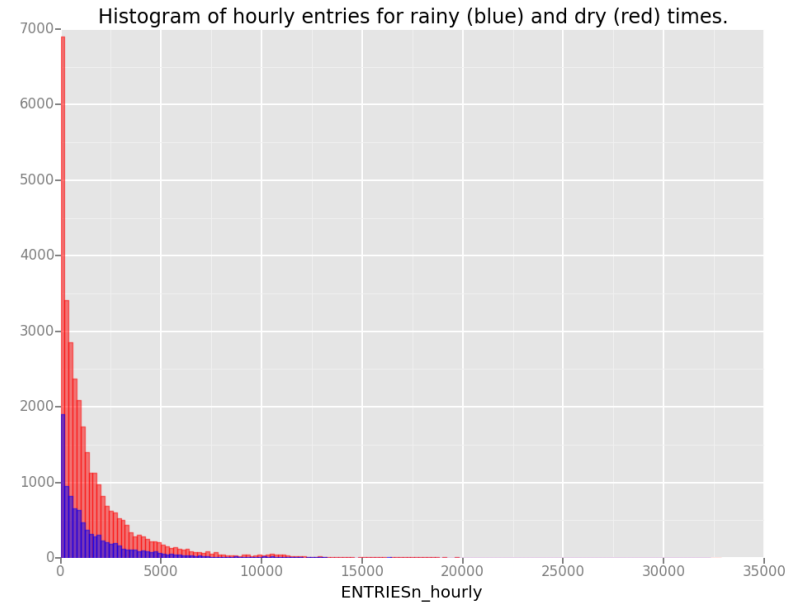


Figure 3: Mean absolute residuals vs mean hourly entries for each turnstile unit.

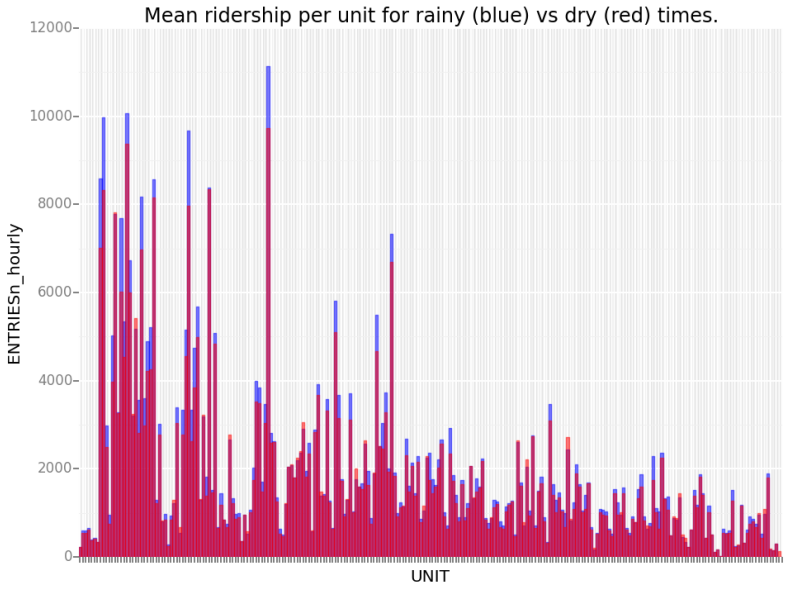
4 Visualizations

Figure 4: Hourly entries histogram for rainy vs non-rainy days.



The histogram of hourly entries in figure 4 demonstrates the non-normality of both ridership distributions. The height of the bars shows the frequency of each bin. The blue bars represent the frequency within rainy days, and the red within non-rainy days. All the red bars are higher, because there are about 3.5 times more non-rainy days in the dataset.

Figure 5: Bar chart showing mean ridership for rainy vs non-rainy days grouped by turnstile unit.



This bar chart compares the mean hourly entries of each of the 240 turnstiles in the dataset on rainy vs non-rainy days. A large proportion of the turnstiles show a higher blue bar (represnting the mean during the rain), showing that on a turnstile-basis, more people tend to use the subway during the rain than otherwise.

5 Conclusions

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?

I concluded that people ride the NYC subway more when it is raining.

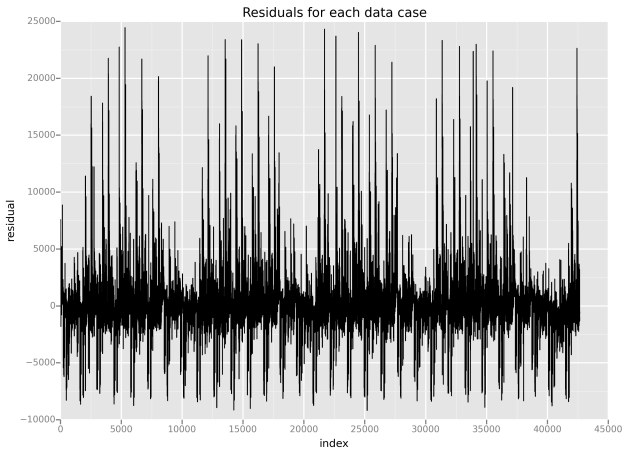
4.2 What analyses lead you to this conclusion? You should use results from both your statistical tests and your linear regression to support your analysis.

From the positive result of the Mann-Whitney U-Test in Section 1, we can conclude that ridership is higher during the rain. This conclusion is also supported by figure 5.
A contrary result is the coefficient of 'precipi' from the linear regression. Furthermore, the dummy variable 'rain' does not have an especially large coefficient when included in the regresssion, nor is its t-statistic especially significant. However, these statistics are not obtained from direct hypothesis tests of the variable in question, and so are not sufficient to contradict the positive result from section 1. Having said this, they do motivate further investigation.

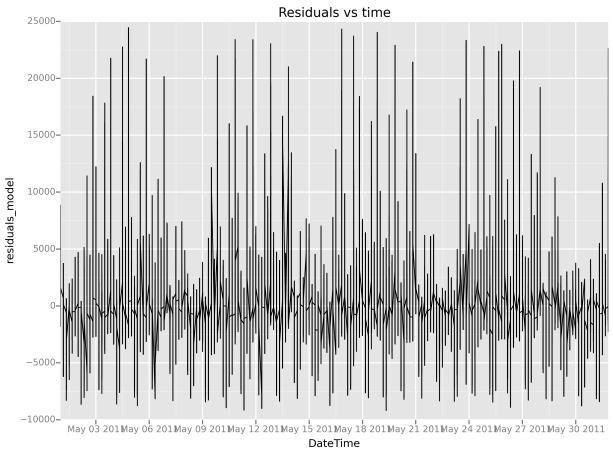
6 Reflection

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 carrying out the linear regression, my feature selection process was not very rigorous. Since there is some relationship between the different features (such as precipi, rain, meanprecipi), it is not optimal to greedily remove or add features in any order.
Linear regression was also not an especially good fit for the problem at hand, as it frequently predicted negative values for ridership. Furthermore, looking at the residuals from the predictions made by my linear model, as in figure 6a, we can see that there is some cyclic trend in the residuals. This trend is also evident in a graph of residuals with respect to time in figure 6b, which means that the model has failed to capture some of the nonlinear regularity in the data. This result motivates the use of nonlinear models.



(a) Residuals for each data case in order.



(b) Residuals with respect to time for each data case.

Figure 6: Cyclic trends in residuals.

A possible shortcoming in performing the statistical analysis of rain's effect on ridership is the fact that there are about 3.5 times less instances of rain in the dataset.