Analyzing the NYC Subway Dataset

Laurent de Vito

June 12, 2015

For consistency, the improved dataset was used throughout.

1 Statistical Test

Pandas provides a simple but useful description of the NYC subway data when it is rainy and when it is not:

	no rain	rain
count	33064	9585
mean	1845	2028
std	2879	3189
min	0	0
max	32814	32289

Table 1: Summary statistics for the NYC data.

We see that on average more people ride the subway when it is rainy. However, the data is a bit more spread out away from the mean when it is rainy.

1.1

The Mann-Whitney U-statistic test was used to analyze the NYC subway data. The null hypothesis was that the underlying distributions of both groups (we just have a sample for each group) are identical. The alternative hypothesis is that the values of one group tend to be smaller or greater than the values of the other group. Notice that this is different that testing for the two populations having equal means[1]. A two-tailed p-value of .05 (or 5%) was chosen.

1.2

This test is non-parametric, meaning that we do not assume that our data is drawn from any particular probability distribution. Indeed, most tests (e.g. the Welch's t-test) are valid only if the data follows the normal distribution. As can be seen on Figure 1, our data is not normally distributed. A Shapiro-Wilk test could have been conducted to test for normality. However, the figure reveals that the underlying probability is clearly non-Gaussian.

[1] see http://discussions.udacity.com/t/welchs-t-test-vs-mann-whitney-u-test/20659

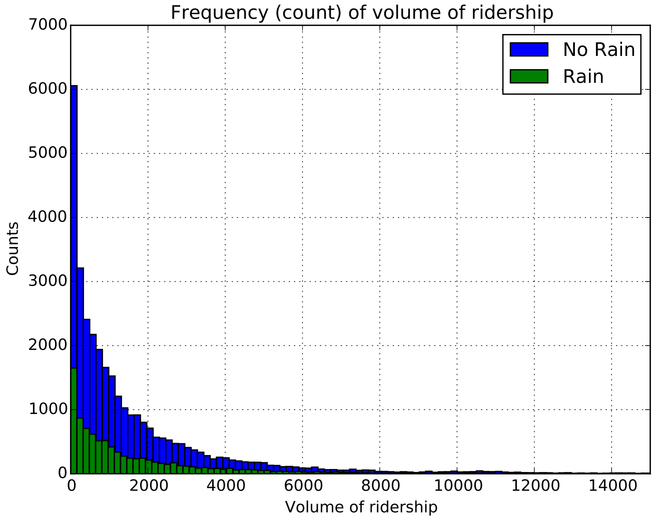

Figure 1: Histogram for entries.

1.3

The mean with rain was 2028.2 entries whereas the mean without rain was 1845.5 entries. The p-value for a one-sided hypothesis was 2 10^{-6}. Hence the two-sided p-value is 4 10^{-6}.

1.4

We reject the null-hypothesis: There is evidence at an α-level of 5% that we are dealing with two different populations. This difference in our samples cannot be ascribed to sampling error. Hence ridership increases when it rains at $p < .05$. If our alternative hypothesis was that the population of entries with rain tend to have higher values than the population of entries without rain, we would carry out a one-sided test, and at an α-level of 5%, we would reject the null-hypothesis: The effect that we perceive in our dataset, namely that ridership increases when it rains, reflect differences across the whole population.

However, we cannot argue that rain causes the ridership to increase. It could well be that traffic is highly congested when it is rainy since people drive slowly, and so more people opt for an alternative: public transports.

2 Linear Regression

2.1

To compute the coefficients of our linear model, the OLS method of Statsmodel was used. On a side note, I would have welcome some words on robust linear regression, i.e. regression that is less sensitive to outliers. I do not know whether outliers have undergone a particular treatment in the improved dataset.

2.2

The input variables were: 'rain', 'rush hour' and 'weekday'. The dummy variables were 'station': for each subway station, a new column was introduced with 1 at the rows where the data was specific to that subway station, and 0 otherwise. It acts as an indicator variable (also called a dummy variable).

2.3

The rationale behind this choice of features is as follows:

- We have previously seen that the variability in ridership is influenced by whether it's rainy. It would thus be sensible to include a variable that encode this, namely 'rain'. This is more accurately encoded through 'precipi', the precipitation in inches. However, if we retain 'precipi' and drop 'rain', the condition number is unacceptably high. So we kept 'rain' instead of 'precipi'. It is unlikely that we lose much information as 'precipi' can be seen as just a fine-tuning of 'rain'. This is supported by the fact that the R^2 statistics does not change a bit when using 'precipi' instead of 'rain'.

- It is intuitively important to know whether it is a weekday since we expect less ridership at weekends (see Figures 6 and 7). This information is encoded by 'weekday'.

- We recognize that the position (latitude and longitude) of the subway stations would be valuable for prediction, but the ridership would depend non-linearly upon those features.

- 'hour' is a bit particular, since it is evidently expected that ridership varies significantly from one day period to the other. So 'hour' must be retained, though it is dubious that the relationship be linear. By incorporating 'hour' as an indicator variable (0 or 1 for the time slots), the R^2 statistics skyrocketed but the condition number went slightly higher that the usual threshold of 30 above which the regression is said to have significant multicollinearities (see [2]). To mitigate this issue, we dropped 'hour' and introduced the binary variable 'rush hour' where rush hour is the part of the day with busy traffic occurring mostly from 6-10am and 4-8pm. Records at time 12 and 20 were considered for rush hour.

We could have increased the R^2 statistics by including the days of the week using the feature 'day_week' through dummy variables instead of 'weekday' because they provide us with a finer information. In the same vein, by including the weather conditions, represented by the feature 'conds', the R^2 statistics would increase. However, those improvements are pretty marginal.

A note on multicollinearity: Choosing the best set of features was made terribly difficult by the issue of multicollinearity that almost systematically crept in when adding new features. Why is it so ? We expect that features pertaining to nearby stations are closely related. The weather (temperature, precipitation, ...) is practically identical (or time-lagged) at two nearby stations, as is the ridership. So if we include the full set of subway stations, each time we add another feature in our model, the condition number gets worse. It would be possible to remedy this issue by selecting representative subway stations and build our model based solely on this subset of stations. This could be carried out by applying a k-means algorithm for instance and take the subways nearest to the cluster centers.

The condition number was 28.3.

2.4

We subtracted the mean of the output before training, so that there is no need for an intercept. The non-dummy features were not normalized: Since we are using a direct method to solve for the least-square estimation of the weights, it does not matter if the features are not normalized (for an iterative method though, feature scaling must be done to prevent slow convergence). However, care must be exercised when we compare one weight with another.

The weight for 'rain' was +39 and that for 'weekday' was +983.

Notice that the weight for 'rain' was not statistically significant: The 95% confidence interval includes 0. We can thus safely withdraw it from our list of features. Indeed, doing so, the R^2 statistics does not change at all (except that the condition number gets slightly better as is expected). At first sight, this is in contradiction with the conclusions we draw from our statistical analysis.

2.5

The R^2 statistics was 0.443.

2.6

First of all, before we focus excessively on the R^2 statistics to assess our model for prediction, we can check that our model residuals are well-behaved. Residuals can be thought of as elements of variation unexplained by our model. Hence they are expected to be roughly normal i.i.d with a mean of 0. Figure 2 depicts the residuals for our model. At fist glance, the underlying distribution is approximately normal with mean 0 and so we are confident that our model residuals are well-behaved. We do not see anything suspicious that would lead us to think that we miss an important structure that our model is unable to reproduce. Since the mean of the residuals is roughly 0, our model does not introduce any bias, something that we cannot assess with the R^2 statistics.

Nevertheless, a closer visual inspection would reveal that the distribution for the residuals is heavier tailed than the normal distribution. To still graphically explore this a bit further, Figure 3 depicts the probability plot associated with the residuals: Would the distribution for the residuals be normal, the samples would fall nicely on the red line (see [3]). The huge departure from the red line on the top right is linked to the high positive residuals. So, from a deeper inspection of the residuals, we conclude that our model is missing some structure, something that could be predicted, and so is inappropriate for making predictions. We now have some options for a model to explain the missing structure:

1. Stick to the linear model and try to incorporate more of the available features (after possibly some non-linear transformation).
2. Stick to the linear model but gather and incorporate new features that better explain the data, i.e. bring new information and thus share as little informa-

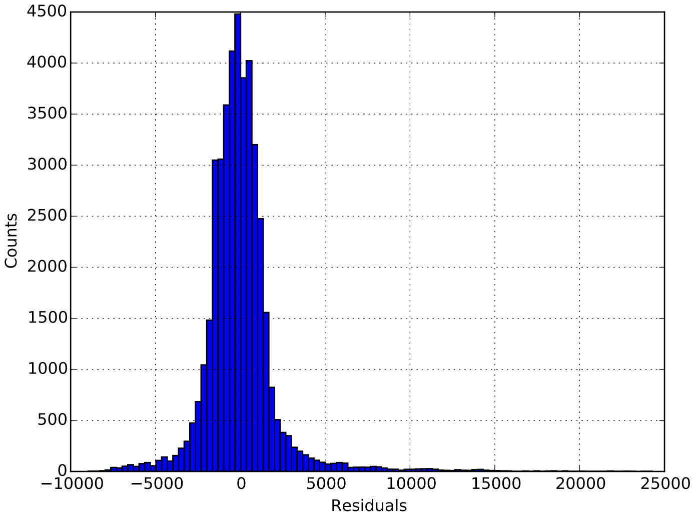

Figure 2: Histogram for residuals.

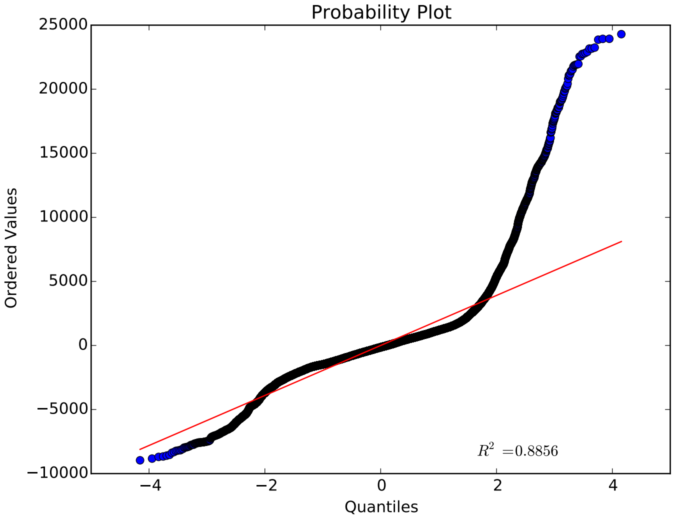

Figure 3: Measuring the agreement of the NYC data with the normal distribution.

tion as possible with the features we have already selected (multicollinearity is automatically avoided).

3. Switch to a non-linear model (gathering new features might be necessary).

Those options are left for future work.

A R^2 statistics close to 1 would mean that our model is able to capture all the variability of the response 'ENTRIESn_hourly' around its mean. The downside is that our model might over-fit if the dataset has some inherent randomness: Our model fits the noise. However, since our model is linear and the dataset is large, overfitting might not be that of a problem. A R^2 statistics of 0 would mean that our model is as trivial as the constant model that always returns the mean of the response whatever the input is.

A R^2 statistics of 0.443 means that our model captures circa half of the variability of the response around its mean. If the objective is to make predictions, we cannot judge based only on the R^2 statistics whether our model is valuable (see [1] for a full-length discussion on this point). Indeed, if the dataset has some non-negligible random noise, the model might be adequate but the R^2 statistics might be low. But the major obstacle to answer the question about the value of our model for making predictions is that we do not know the requirements for the accuracy on the predictions.

We note that the predictive power or reliability of our model for prediction is not reduced because of multicollinearity as long as the dataset is representative.

3 Visualization

3.1

See Figure 1.

3.2

We compare the ridership on non-rainy versus rainy days in Figures 4 and 5 respectively. Since there are far fewer rainy days than non-rainy days, Figure 5 looks sparser than Figure 4. Apart from this, there is no noticeable difference between the two figures. We also observe that ridership is relatively smaller on weekends (since most people commute on weekdays).

Since the points in Figures 4 and 5 are heavily cluttered around small values of ridership, we offer an alternative view, Figures 6 and 7, based on the mean volume of ridership per day[2]. It appears clearly that on average the volume is higher on rainy days.

4 Conclusion

4.1

We believe that rain does affect the ridership in NYC.

4.2

This conclusion is drawn from our statistical analysis. Our linear regression does not support this claim but the condition number is fairly high, though under the threshold of 30. It means that regression cannot give sound results about individual features like 'rain' (see [2]).

[2] Notice that the ridership at the time stamp 0 on a particular day in the dataset corresponds to the entries the eve of that day from 8 to 12pm. Hadn't we consider this particular case, we would observe a slight re-distribution of volume among the days.

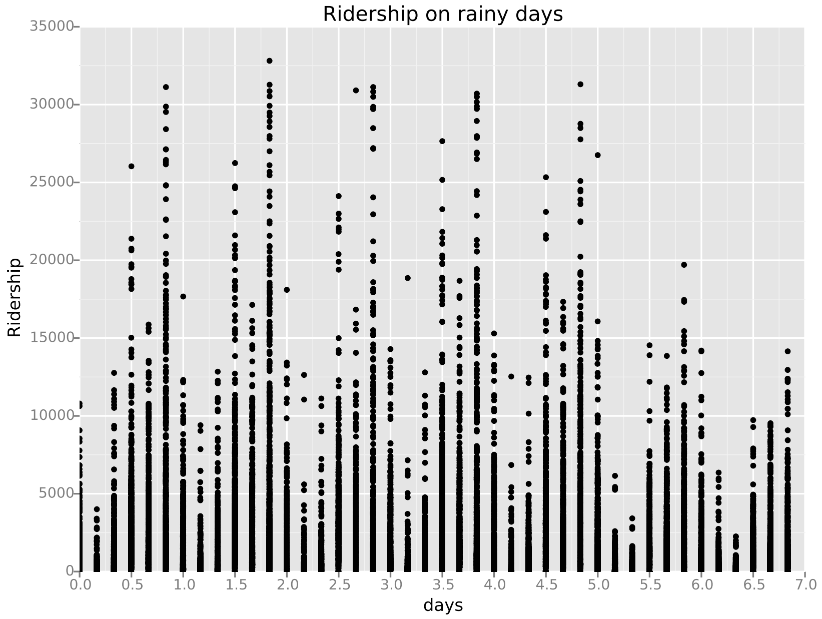

Figure 4: Ridership on non-rainy days: Monday is 0, etc.

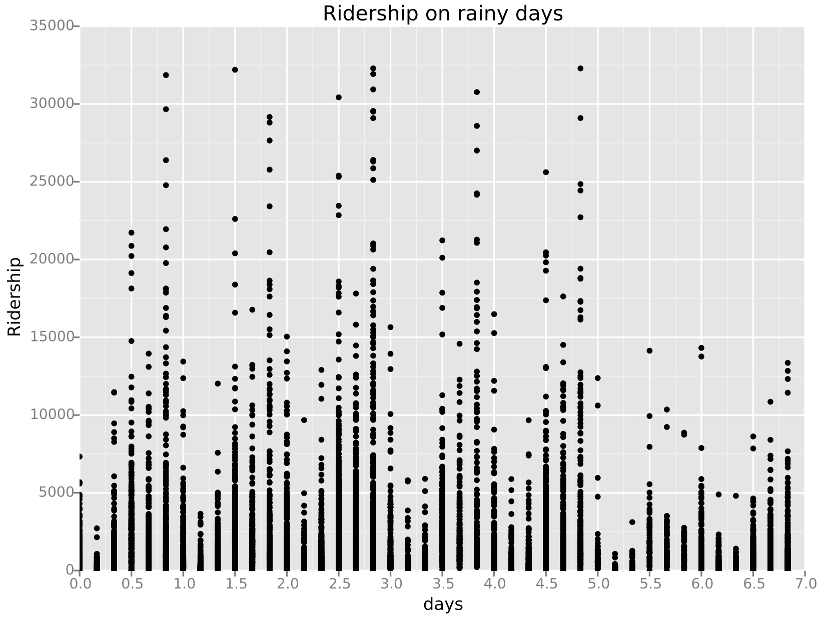

Figure 5: Ridership on rainy days: Monday is 0, etc.

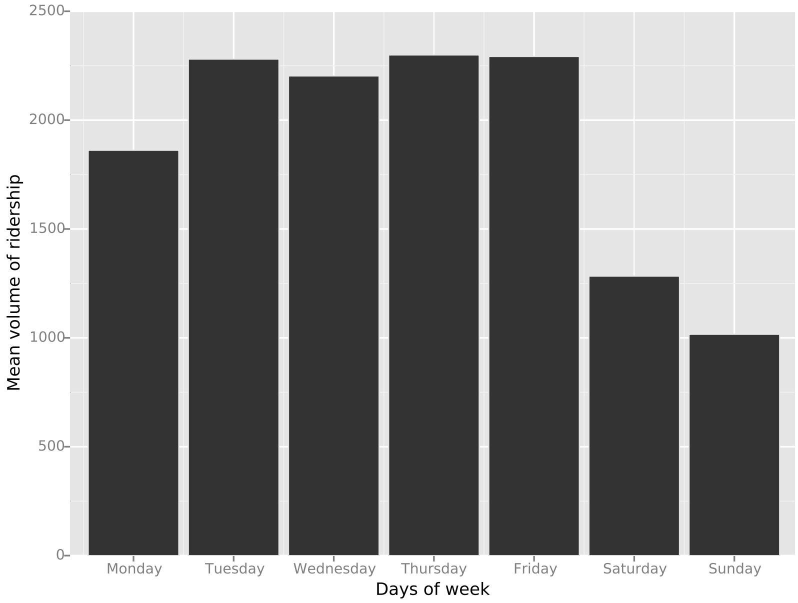

Figure 6: Mean volume of ridership on non-rainy days.

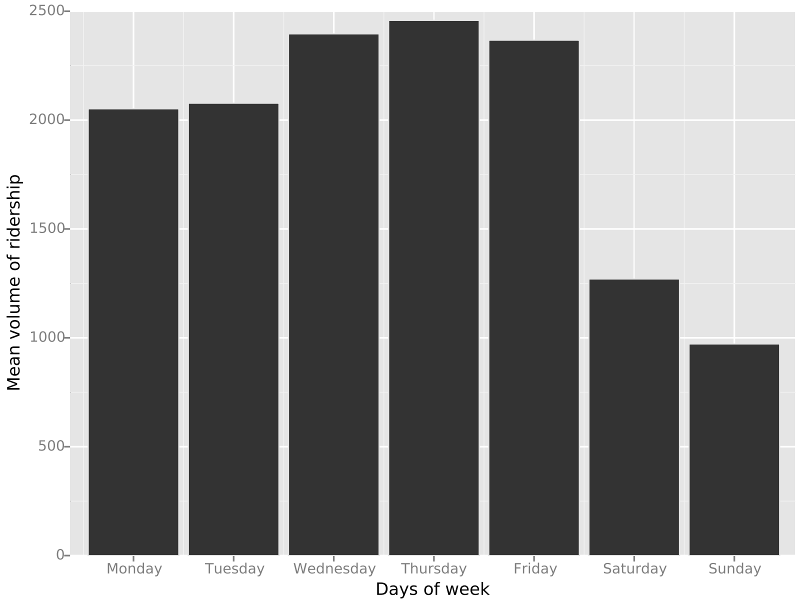

Figure 7: Mean volume of ridership on rainy days.

5 Reflection

5.1

5.1.1

The dataset does not seem to be truly random since the non-weekdays are overwhelmingly represented when it is not rainy. Consider the following table:

	no rain	rain
# of weekdays	22570	7900
# of non-weekdays	10494	1685

The problem is that we do not have the same proportion of weekdays versus non-weekdays in both columns: When it is not rainy, the data is clearly biased towards the non-weekdays and we know that the pattern during non-weekdays differs significantly from the pattern during weekdays. This might impact the validity of the statistical inference.

Do we wish more features for potentially better predictions ? We could for instance include the number of employees close to each subway station, or a ranking (e.g. from 1 to 5) for how touristic the area close to each subway station is.

5.1.2

The expressiveness of a linear model is fairly limited. We would advocate for a non-linear model:

- We expect that the ridership varies non-linearly with e.g. the position (latitude/longitude) of the subway stations and that position has a strong influence on ridership. Our too simple linear model cannot capture this kind of relationship.
- That the response varies non-linearly is illustrated in Figure 8 where we focus on a particular, randomly selected, subway station. The residuals, apart from being excessively large, exhibit a periodic structure that a linear model cannot reproduce.

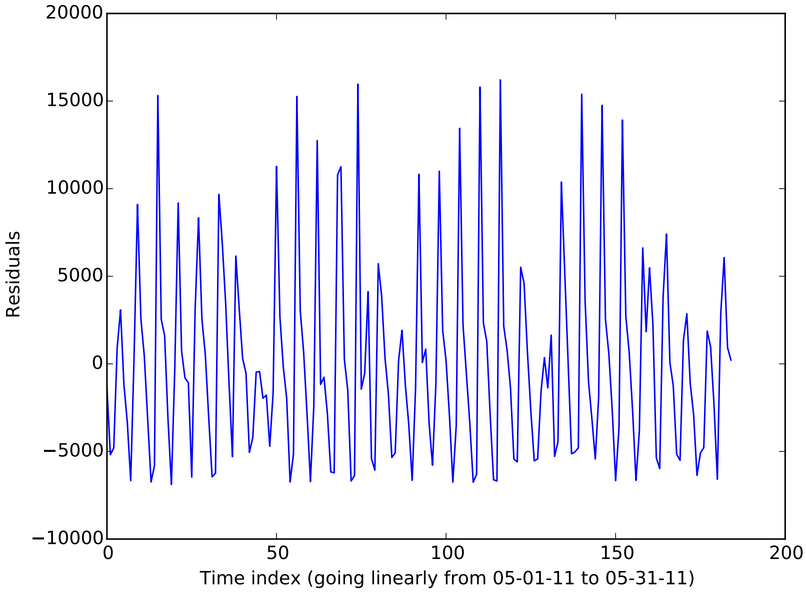

Figure 8: Residuals for the 'Roosevelt Av' subway station over time.

5.2

None.

Acknowledgments

This is the third draft of my work and I would like to thank the two anonymous reviewers for their suggestions and comments.

References

[1] J. Frost: How High Should R-squared Be in Regression Analysis ? http://blog.minitab.com/blog/adventures-in-statistics/how-high-should-r-squared-be-in-regression-analysis

[2] Wikipedia: Multicollinearity https://en.wikipedia.org/wiki/Multicollinearity

[3] Wikipedia: Q-Q plot https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot