Intro to Data Science - Final Project

New York Subway Turnstile Data

Jamie Brand - brandjamie@hotmail.com

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?**

The first test I ran was to see if there was a statistical difference between nENTRIES_hourly for rainy and non-rainy days. For this I used the Mann-Whitney U-test. I used a two-tail P value as, while we might assume that the nENTRIES_hourly would be higher on rainy days, we would be interested to know if it is in fact lower.

The null hypothesis is that there is a 0.5 probability that a randomly selected sample from the population on a rainy day has higher ridership (nENTRIES_hourly) than a randomly selected sample from the population on a non-rainy day.

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 Mann-Whitney test is a non-parametric test and is suitable for data that does not have a normal distribution. The Shapiro Wilkes test can be used to test for normal distribution. However, it is not reliable for sample sets with over 5000 data points as was the case here. Making a histrogram to show the distribution (as seen in the visualization section) shows that the data is not normally distributed. The two distributions do appear to have the same shape which is required for the Mann-Whitney test. I chose a critical p - value of 0.05 as this is a generally accepted figure for non critical data.

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.**

with_rain_mean	2028.2
without_rain_mean	1845.54
whitman p (for two tailed test)	0.00000274106

1.4 **What is the significance and interpretation of these results?**

The p-value is much less than our critical value of 0.05. We can say there is less than 5% chance that the difference in distributions is due to random sampling. For this test we can reject the null hypothesis.

Linear Regression

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

I used the statsmodel implementation of OLS regression. .

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 used were 'meanprecipi', 'meantempi', 'fog', 'weekday' and 'meanwspdi'
I used dummy variables for the 'UNIT' feature, the 'hour' feature, and the 'conds' feature.

2.3 **Why did you select these features in your model?**

For linear regression it is best not to have redundant features (i.e. features that include the same information). For instance 'day_week' and 'weekday' contain similar information. While it may seem that the 'day_week' has more information, the linear nature of the regression means it loses the instinctive distinction between weekdays and weekend. On trying each of the features independently I got a better score using 'weekday'. The score was a tiny bit larger using both features but we want to minimise the features if we can.

I used dummy variables for the 'UNIT' feature. The 'station' feature was not used as each unit was at one station however each station could have more than one unit (presumably describing more than one exit). So, the 'UNIT' feature had all the information in the 'station' feature and more. This was shown in practice by getting a higher score using the 'UNIT' feature than with either the 'station' feature or both the 'station' and 'unit' features.

The 'precipi','rain' and 'mean_precipi' all have very similar information. Through experimentation I found the best results using 'mean_precipi'. I used mean features for other weather attributes (pressure, wind-speed and temperature). I also added dummy variables for the 'conds' feature.

It seemed that the hour feature was not very usefull, as intuitively, it is non linear. Certain times of the day are likely to be busier than others but these busy times are generally in the daytime creating a bell curve that is not well reflected in linear regression. I experimented with adding a new 'daytime' feature for hours between 7am and 7pm. This gave a slightly improved result. However, I found that using dummy variables for the hour resulted in a far greater increase in the score.

I also experimented with using dummy variables for the day of the week instead of the 'weekday' feature. While this did show a small increase in the r-squared value, the value of the constant with this model was approximately 33 million which seems like a clear indicator that the model has a problem.

2.4 **What are the parameters (also known as "coefficients" or "weights") of the non-dummy features in your linear regression model?**

The coefficients for the non-dummy variables were:

constant	2260.45
meanprecipi	2640.90
meantimpi	-21.26
fog	-622.95
meanwspidi	-16.50
weekday	993.71

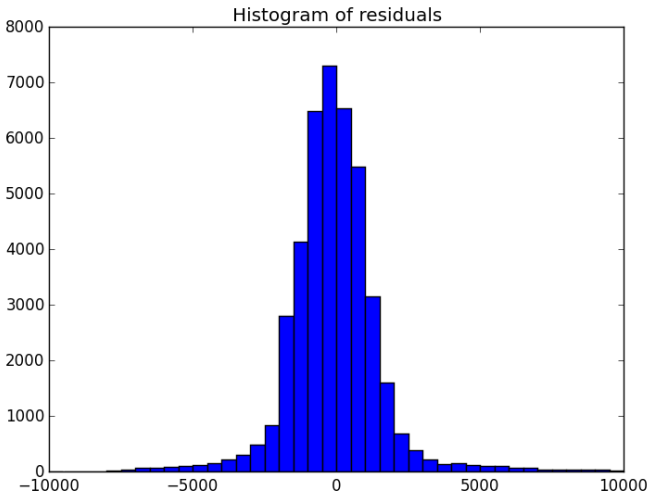
2.5 **What is your model's R2 (coefficients of determination) value?**

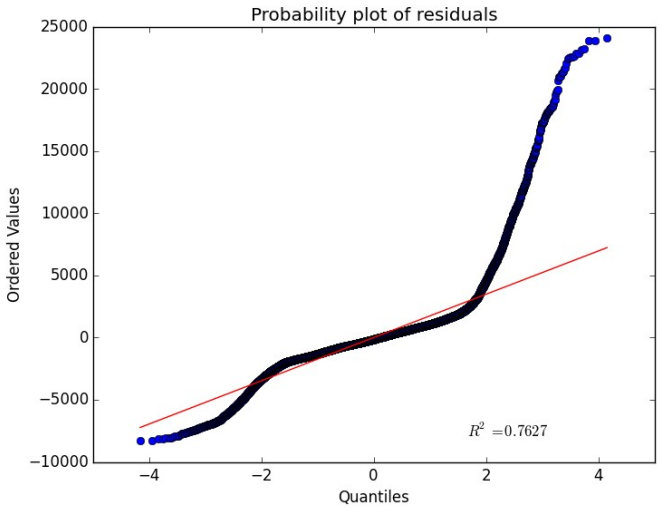
The R2 value was 0.547

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?**

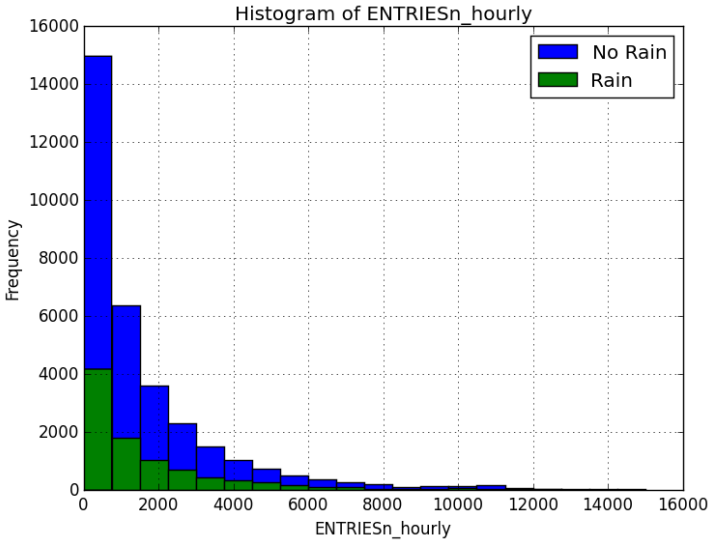
The R2 value shows that we have explained 54.5 percent of the variability of the nENTRIES_hourly feature. This shows a weak relationship between the variables and the the ridership. While this model could be used to make a very rough estimate of ridership, it could not be used to make accurate predictions as approximately 45 percent of the variability is still unexplained.

For the linear regression model to be appropriate the error terms should be normally distributed. The histogram of the residuals shows very long tails which implies there are some very large residuals. Looking at the normal probability plot of residuals confirms that the distubution is heavy tailed with extreme divergence from the normal distribution beyond the second quantile. This implies that the linear regression model is not appropriate for this dataset.

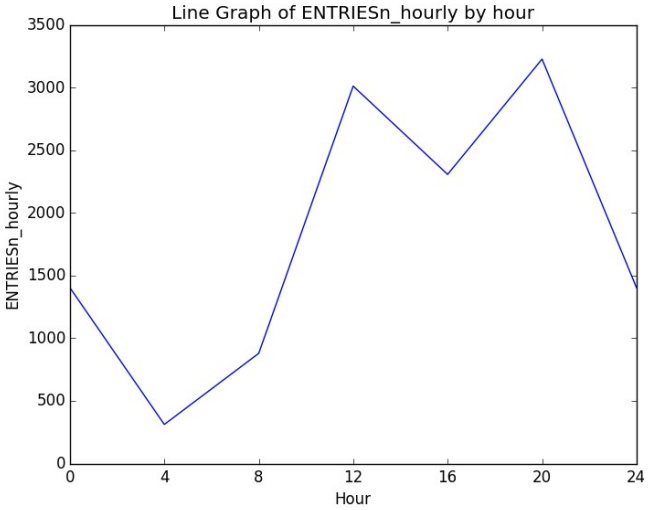




Visualization



From this visualization we can see that the distributions of nENTRIES_hourly for days that have rain and days that do not are similar. We can also see that the distributions are not normal. This was very important when considering which test to run. It can also be seen that there were many more days without rain than with rain.



This line graph shows the mean nENTRIES_hourly by hour. It shows the non-linear nature of this particular feature. It is worth noting that the linear regression would not include the last point on this plot (from 20-24) as 24:00 hours is the same as 00:00. As there are only a small number of data points, it is reasonable to create dummy variables with them. If there were more data points, another form of analysis may have been more appropriate.

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?**

The analysis shows that more people ride the NYC subway 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.**

The results of the Mann-Whitney test show that it is very unlikely the difference observed in ridership on rainy and non-rainy days occurred by chance. Looking at the means in ridership on rainy and non-rainy days, it is clear that ridership on rainy days is higher.

However, this accounts for a very small amount of the variation in ridership. The linear regression allowed us to see how different factors affect the ridership. The biggest factor was the unit (and/or Station) which is no surprise. Some stations are busier than others. The next biggest factor was the time of day with the busiest periods occurring at the times people were travelling to or from work. Of course there are more aspects to weather than just rain but all of these together did not have a big influence on the number of riders.

Reflection

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

i. Dataset

The data had a lot of information that was not independent, for instance 'rain' and 'precipi'. Thus, the number of useful features was much less that appeared at first.

After the turnstile unit feature, the most important feature was the hour. This feature was not as detailed as it could have been with the interval between data points at every four hours.

The data was all taken from a single month so the model cannot take into account any seasonal variations. The model may be less accurate for different months. This also makes the 'date' feature almost useless. For an accurate model we would really need data for at least a year.

Additional information that could have been useful would be whether the stations were in residential or business districts as they would have different busy periods.

ii. Analysis, such as the linear regression model or statistical test.

The Mann-Whitney test was very useful for answering a specific question (Whether the ridership is higher on rainy or non-rainy days). On its own it did not present a very accurate picture of the subway usage in New York City. While the linear regression did give us a better impression overall, it was not very accurate in predicting ridership. Some of the data was non-linear and while using dummy variables helped, other algorithms could have been more accurate.

References

https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test

http://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.stats.shapiro.html

http://pandas.pydata.org/pandas-docs/stable/visualization.html#histograms

http://www.graphpad.com/guides/prism/6/statistics/index.htm?stat_checklist_mannwhitney.htm

http://www.graphpad.com/guides/prism/6/statistics/index.htm?one-tail_vs__two-tail_p_values.htm

http://www.itl.nist.gov/div898/handbook/prc/section1/prc131.htm

http://statsmodels.sourceforge.net/devel/generated/statsmodels.regression.linear_model.OLS.html

http://stackoverflow.com/questions/13218461/predicting-values-using-an-ols-model-with-statsmodels

http://iquantny.tumblr.com/post/93845043909/quantifying-the-best-and-worst-times-of-day-to-hit

https://onlinecourses.science.psu.edu/stat501/node/276

https://onlinecourses.science.psu.edu/stat501/node/281