Analyzing the NYC Subway Dataset

Mark Rehbein

Section 0. References

The following list of references was used. Where possible, usage of the reference is noted in the text by including the reference number in braces, for example, (1).

1. http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test
2. https://storage.googleapis.com/supplemental_media/udacityu/649959144/MannWhitneyUTest.pdf
3. http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html
4. http://en.wikipedia.org/wiki/Null_hypothesis
5. http://matplotlib.org/1.2.1/examples/pylab_examples/histogram_demo.html
6. http://statsmodels.sourceforge.net/devel/generated/statsmodels.regression.linear_model.OLS.html
7. http://nbviewer.ipython.org/urls/s3.amazonaws.com/datarobotblog/notebooks/ordinary_least_squares_in_python.ipynb
8. http://nbviewer.ipython.org/urls/s3.amazonaws.com/datarobotblog/notebooks/multiple_regression_in_python.ipynb
9. https://www.moresteam.com/whitepapers/download/dummy-variables.pdf
10. http://blog.yhathq.com/posts/logistic-regression-and-python.html
11. http://en.wikipedia.org/wiki/List_of_New_York_City_Subway_services
12. https://surfstat.anu.edu.au/surfstat-home/4-1-6.html
13. http://en.wikipedia.org/wiki/Memorial_Day
14. http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

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?

The Mann-Whitney U test was used to analyse the NYC subway data. This statistical test is a non-parametric test of the null hypothesis that two samples come from the same population (1). Rejecting or disproving the null hypothesis concludes that there are grounds for believing that there is a relationship between two phenomena (4).

The Mann-Whitney U test is usually two-tailed (2). However, the scipy.stats.mannwhitneyu implementation used in the analysis reports a one-tailed p-value. To get the two-sided value, the one-sided value is multiplied by 2 (3).

The p-critical value of 0.05 is used which is a 5% significance level and by convention, moderate evidence against the null hypothesis. (12)

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.

Plotting a histogram of the data (refer section 3.1) shows that the both samples are a non-normal distribution. The Mann-Whitney U test is a non-parametric test that can be used on sample populations with unknown distributions and thus does not assume a normal distribution. (2)

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 Mann-Whitney U test results with the means of the two samples are:

With rain mean:	`1105.4463767458733`
Without rain mean:	`1090.278780151855`
Mann-Whitney Test Statistic (U):	`1924409167`
One-sided p-value:	`0.024999912793489721`

```
(1105.4463767458733, 1090.278780151855, 1924409167.0, 0.024999912793489721)
```

1.4 What is the significance and interpretation of these results?

Because the two-sided p-value is below p-critical we can reject the null hypothesis, therefore:

Mean ridership per hour, per unit on the New York subway for rainy and non-rainy days during May 2011 were 1105 and 1090; the distributions in the two groups differed significantly (Mann-Whitney U = 1924409167, P < 0.05 two-tailed).

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:

1. Gradient descent (as implemented in exercise 3.5)

The Gradient descent algorithm was implemented in exercise 3.5 using the following features:

- 'meanpressurei', 'precipi', 'Hour', 'mintempi' with ‘UNIT’ as a dummy variable.

This model reported an r-squared value of:

```
Your R^2 value is 0.46474289883
```

2. OLS using Statsmodels

Additionally, an ordinary least squares model implementation (6) was used with the following features:

- 'meanpressurei', 'precipi', 'Hour', 'mintempi' with UNIT as a dummy variable

This model reported an r-squared value of:

```
Your R^2 value is: 0.48407218775
```

Adding and additional dummy variable for Day of the Week improves the model, reporting r-squared value of:

```
Your R^2 value is: 0.494790311722
```

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

In the improved OLS model, the following features were used:

- 'meanpressurei'
- 'precipi'
- 'Hour'
- 'mintempi'

Additional dummy variables used were:

- UNIT
- Day of Week

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.

The rationale for including the features is:

1. Mean pressure because this is an indication of weather quality – stormy weather has low pressure, good weather has high pressure. There may be more riders in threatening weather
2. Precipitation quantity (rather than RAIN boolean) because more rain could lead to more riders
3. Minimum temperature because if it's cold, there may be more riders
4. Hour of day and Day of the Week as the number of riders is influenced by the Hour and Day of the Week because of work and lifestyle patterns.
5. UNIT because the model is predicting entries at each unit.

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

The coefficients of the non-dummy features are highlighted in the table below.

```
OLS Regression Results
==============================================================================
Dep. Variable:        ENTRIESn_hourly   R-squared:                       0.495
Model:                            OLS   Adj. R-squared:                  0.470
Method:                 Least Squares   F-statistic:                     19.68
Date:                Tue, 31 Mar 2015   Prob (F-statistic):               0.00
Time:                        23:08:09   Log-Likelihood:                -88049.
No. Observations:               10000   AIC:                         1.770e+05
Df Residuals:                    9525   BIC:                         1.805e+05
Df Model:                         474
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const          9226.7977   3586.651      2.573      0.010      2196.198  1.63e+04
meanpressurei  -302.9864    136.344     -2.222      0.026      -570.250   -35.723
precipi           6.5510     46.952      0.140      0.889       -85.484    98.586
Hour             62.2722      2.462     25.295      0.000        57.447    67.098
mintempi        -10.2918      2.739     -3.758      0.000       -15.660    -4.923
```

2.5 What is your model's R^2 (coefficients of determination) value?

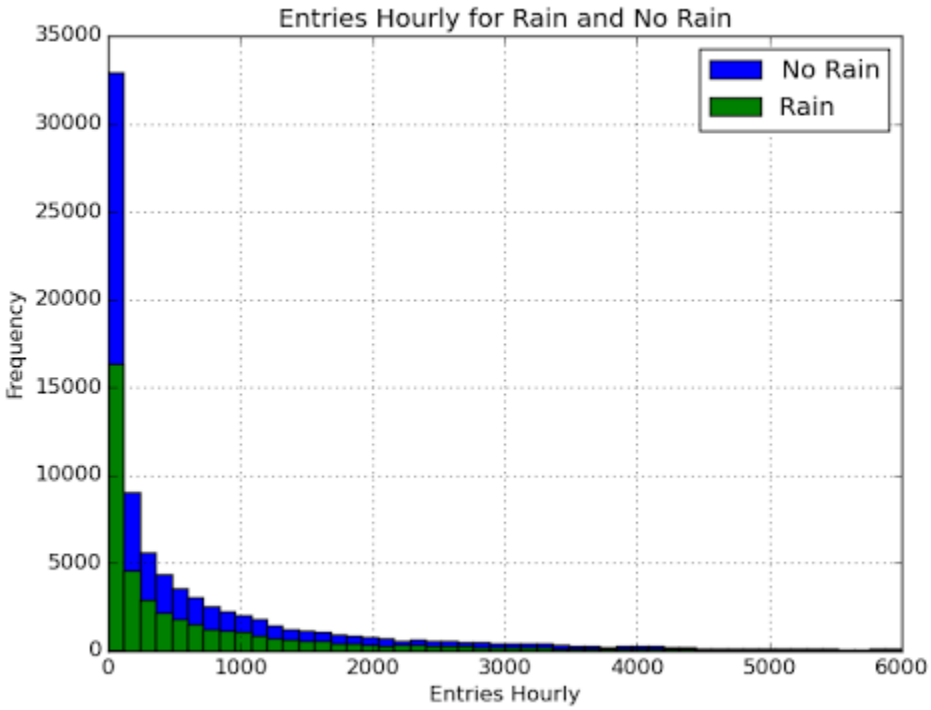
```
R-squared:                       0.495
```

2.6 What does this R^2 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^2 value?

In general, the higher the R-squared, the better the model fits your data (14). The low R-squared value would indicate that this model isn't able to precisely predict the ridership.

Section 3. Visualization

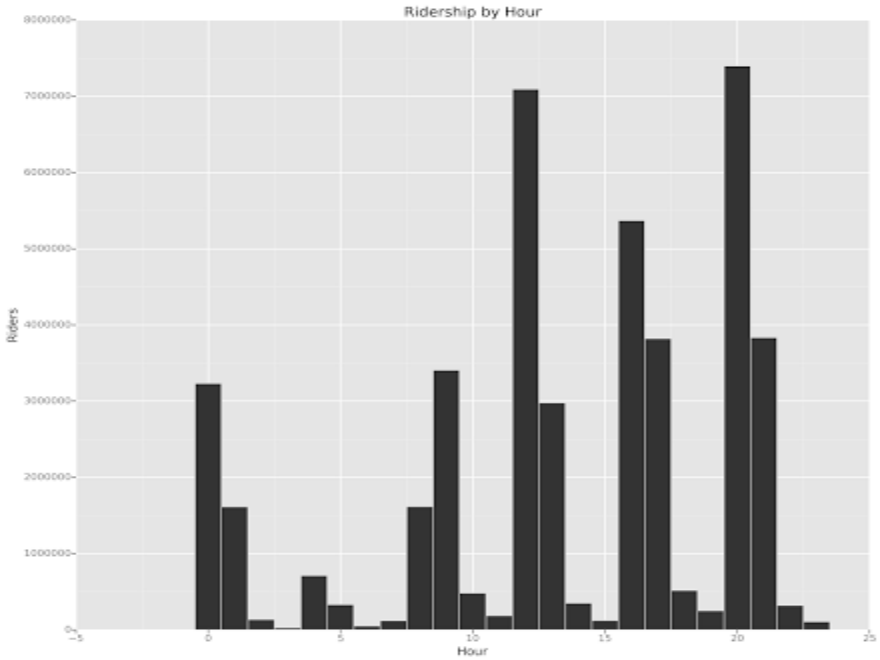
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.



The histograms of ridership for both rainy and non-rainy days appear to show a non-normal distribution. Overall, there are less data points on rainy days then non-rainy days, because there are only 8 rainy days in the data set, compared to 22 non rainy days.

3.2 One visualization can be more freeform.

Ridership by time-of-day

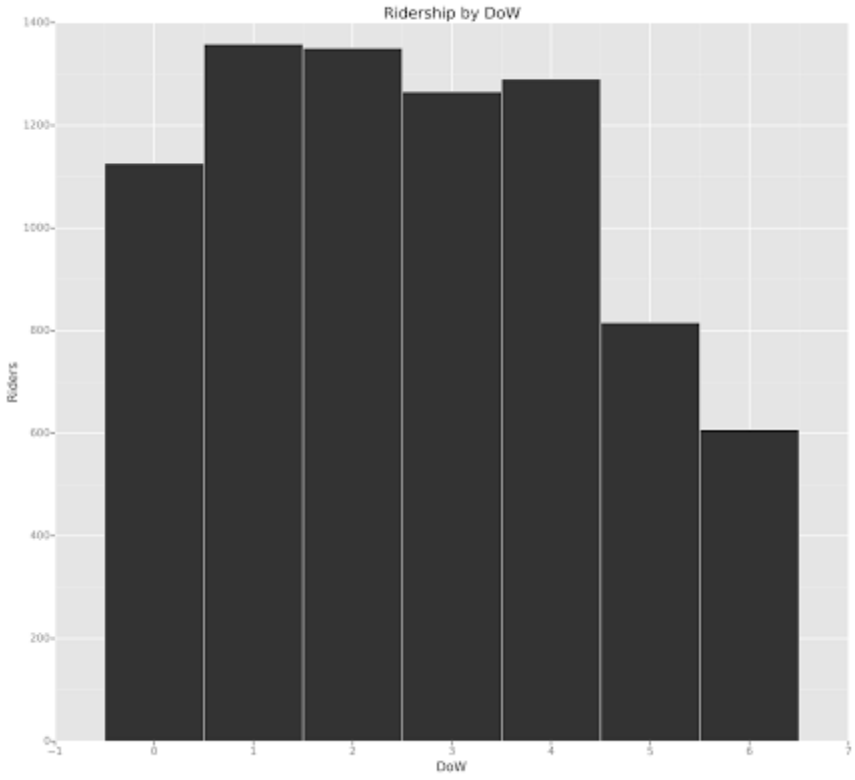


Peak times for travel are clearly shown by this visualisation. Showing:

- rush hours (9am, 4pm)
- middays (noon)
- evenings (8pm)
- late night (midnight)

This corresponds with known ridership patterns for the New York subway (11)

Average Ridership by day-of-week



This figure shows average hourly ridership for each day of the week (0 is Monday, 6 is Sunday). Ridership is much higher during weekdays than weekends. Average ridership for Monday is less than the other weekdays because of a Monday public holiday in May.

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?

A valid statistical comparison supported by the Mann-Whitney U test indicates a small increase (1.3%) in mean ridership on the NYC subway when it is a rainy day.

Mean ridership per hour, per unit on the New York subway for rainy and non-rainy days during May 2011 were 1105 and 1090; the distributions in the two groups differed significantly (Mann-Whitney U = 1924409167, P < 0.05 two-tailed).

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 U test showed a small but statistically significant rise in ridership on days that had rain. However, linear regressions using both Gradient Descent and Ordinary Least Squares approaches yield an R-squared value of ~0.5 suggesting that weather is a poor predictor of ridership.

Section 5. Reflection

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

1. Dataset,

In this dataset, daily weather data are used against hourly subway data. A more thorough analysis would be possible if hourly weather data could be matched to the subway data. For example, with the current dataset, it is possible that it rained only in the early morning or for a short period which is then recorded as rain for the whole day.

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

The linear regression models yielded an R-squared value of ~0.5. It would be interesting to see if a non-linear regression model improves the prediction. Additionally, increasing the size of the dataset to include multiple years and thus rainy days to include in the statistical analysis and regression models may provide additional insight. Filtering a larger dataset into work day vs non-work day may uncover a stronger relationship.

5.2 (Optional) Do you have any other insight about the dataset that you would like to share with us?

The dataset is only for the single month of May 2011 with only 8 days rain. The Monday public holiday was 30 May 2011 (13) which was a rainy day. This probably would have reduced the average ridership for rainy days as the need to travel due to work commitments is less. Increasing the analysis to include multiple years of data would reduce the effects of these events. Alternatively, a re-analysis of the data excluding the public holiday data or treating it as a weekend and performing a comparison analysis of the effect of rain on weekday's vs weekend's may be interesting.