Project 1

June 2, 2015

Section 0 - References

- Scikit-Learn documentation: http://scikit-learn.org/0.14/modules/generated/sklearn.linear_model.LinearRegression.ht
- http://en.wikipedia.org/wiki/Goodness_of_fit
- http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test

Section 1 - Statistical Test

1.1

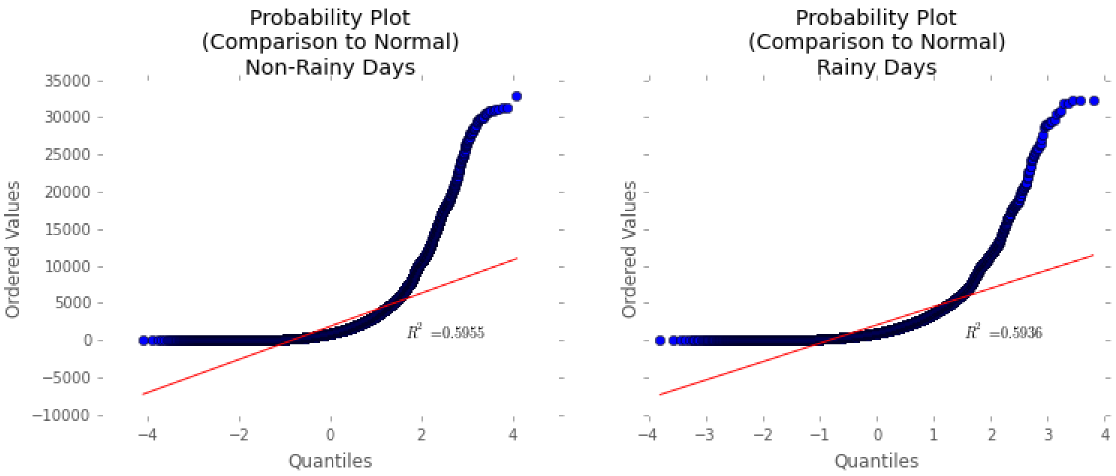
A Mann-Whitney U test was used to test the null hypothesis that the two samples (with and without rain) are from the same population. A two-tailed test was used to estimate the probability of the null hypothesis relative to the alternative hypothesis that the number of hourly entries is more extreme (higher or lower) when it is raining. The p-critical value is 0.05, indicating a maximum of 5% chance that the two samples are from the same population.

1.2

The number of hourly entries is likely to be bounded on both sides- they must be above zero because negative entries are nonsensical, and there must be some upper limit to the number of entries per hour in each turnstile because riders can only move so quickly (thus limiting the number of entries monitored by a single unit).

The hourly entries were split into Non-Rainy and Rainy subsets ('rain' = 0 or 1 respectively). Probability plots of these subsets are shown below. Neither is normal largely because of the lower bound on the probability distribution, but also because there is a long upper tail of measurements with high values. Therefore statistical tests that assume normality (like Welch's t-test) cannot not be used. The Mann-Whitney U test was used because it doesn't assume normality and its other assumptions are met- the samples are unpaired with similar distributions and the data is ordinal.

```
Out[3]: <matplotlib.text.Text at 0x7f5552082b00>
```



1.3

The calculated test statistic has a two-tailed p-value of 0.00000548, far below the critical p-value. The occurence of rain increased the average number of subway entries per hour by about 9.9%- from 1845.54 per hour to 2028.20 per hour.

```
U = 153635120.5
p = 0.00000548
Mean (with Rain) = 2028.20
Mean (without Rain) = 1845.54
```

1.4

There appears to be a statistically significant difference in entries to the subway when it is raining. Specifically, there is about a 10% increase on average.

Section 2- Linear Regression

2.1

A - OLS Linear Regression with Scikit-learn:

```
Coefficient of Determination (R^2) = 0.5457
13269.6411333 [-2017.29385914  -349.17224788   -15.53203689]
```

2.2

Features:

- precipi - precipitation in inches at the time and location
- pressurei - barometric pressure in inches HG at the time and location
- tempi - temperature in degrees F at the time and location

Dummy Variables:

- UNIT - represents several turnstiles, possibly from the same station
- day_week - day of the week
- hour - hour of the day

2.3

All three features used were aspects of the weather. In particular, they were specific measurements rather than relative measurements. For example, the precipitation in inches was a bit more informative (resulted in a slightly better coefficient of determination) than the 'rain' variable. This isn't surprising since the rain variable doesn't distinguish between a very light rain at a different time in the same day and a downpour at the time of measurement.

The dummy variables related to typical use of the turnstiles. It is reasonable to expect that a given group of turnstiles (within a unit) will have similar traffic at a given time and day each week. The 'UNIT' feature in particular was very effective at increasing the Rˆ2 value.

2.4

- precipi = -2017.294
- pressurei = -349.172
- tempi = -15.532

2.5

Rˆ2 = .5457

2.6

An Rˆ2 value of .5457 means that the fitted model explains about 55% of the variance in the outcome variable (number of entries per hour). This could be useful for applications that don't need to be too specific, but shouldn't be used to attempt highly accurate predictions of ridership numbers. Examining the residuals (done below in section 5) would be more useful to judge the prediction accuracy.

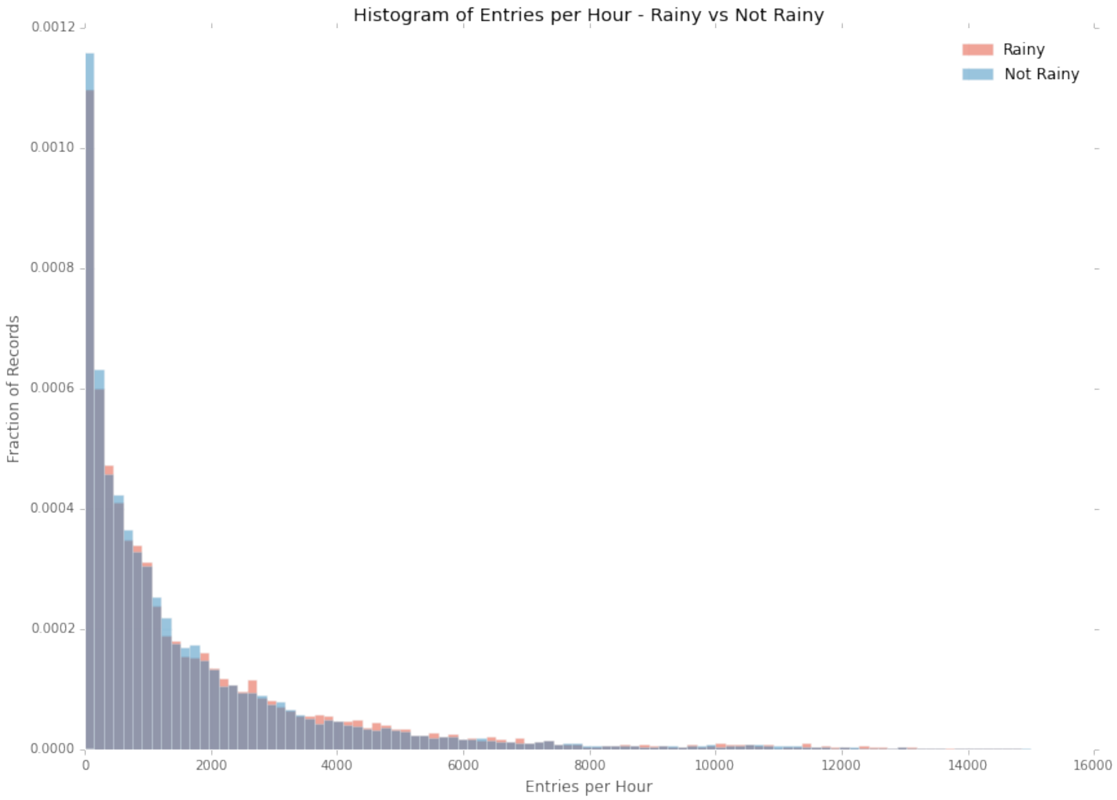
An important consideration is that a large portion of the Rˆ2 value was contributed by looking at which unit was collecting data. This means:

- A much simpler model (looking at average entries for a given UNIT) may be appropriate to get rough estimates of ridership that are almost as good.
- This model is tightly fit to the given dataset. Data collected from other units would require a newly fit model.

Section 3 - Visualization

Figure 1

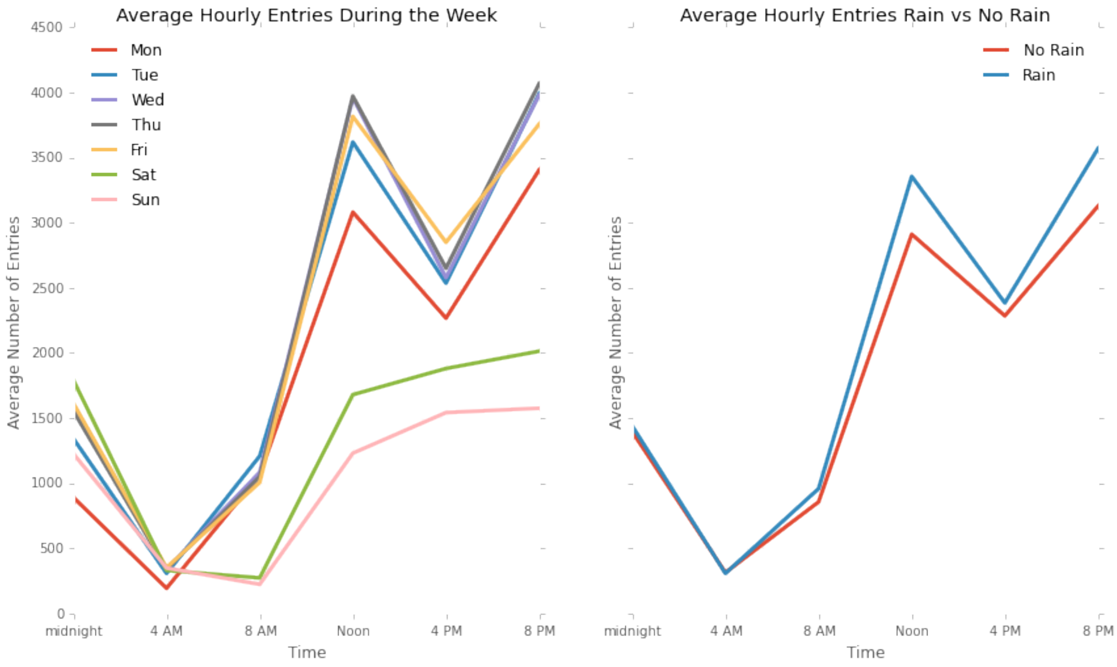
```
Out[6]: <matplotlib.text.Text at 0x7f554e2494a8>
```



This histogram comparison shows the distribution of records with different values for the number of entries per hour. The area is normalized to be equal for both records from days with rain and days without rain. The key insight is that rainy days tend to have a lower proportion of records with a small number of entries (for example, 0 to 300). Other than this the difference is subtle, but it does seem to provide some basic evidence that rain increases the number of entries.

Figure 2

```
Out[7]: <matplotlib.text.Text at 0x7f5551f82fd0>
```



The left portion of this plot shows how the average number of entries per hour changes over time- during the week and throughout the day. As might be expected, ridership is much higher during the week. During the week ridership is highest at Noon and 8pm, while weekend ridership follows a different trend- increasing from 8 AM before leveling off between 8 PM and midnight.

The right portion of the plot shows a similar comparison but uses the rain variable to distinguish the trends instead of the day of the week. Interestingly, the effect of rain appears to small in the early morning hours and larger during the day. However, the extent of the increase appears to be similar to or even less than the difference in ridership from day to day during the week.

Section 4 - Conclusion

4.1

Based on the available data there appears to be a small (roughly 10% on average) but statistically significant increase in ridership when it rains. However, this conclusion applies only to ridership throughout May 2011 at the stations that were measured. It is possible that other variables influence ridership (the linear regression only captured about 55% of the variance).

4.2

The statistical test that was used (Mann-Whitney U) showed that the null hypothesis is unlikely compared to the alternative hypothesis (that the distribution of ridership values is different when it rains). The linear regression model actually has a large negative coefficient for the precipitation feature. However, when fewer features are used (specifically day of the week and time of day) the coefficient is positive. In either case when the rain variable (or precipi, which provides similar but more detailed information) is used the linear regression is more predictive (higher coefficient of determination). The second visualiation was especially clear, showing that average ridership during rainy days is greater than or equal to average ridership on dry days no matter the time of day.

Section 5 - Reflection

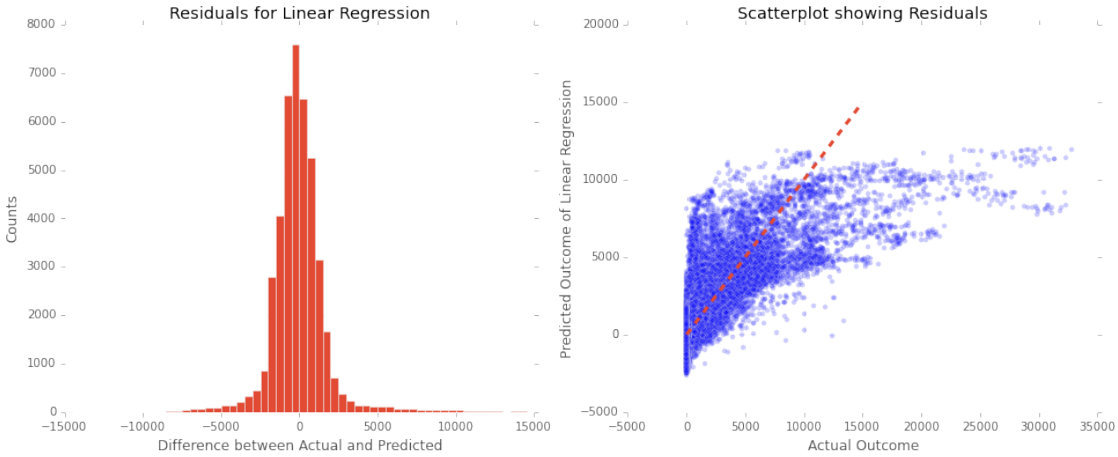
5.1

The dataset is limited in that it only contains data over the course of a single month (May 2011). It's possible that temporary effects (holidays, sporting events, etc) could have affected ridership numbers, or that this period of time doesn't provide a representative sampling of possible weather conditions.

It is reasonable to expect that the time of measurement (date and time of day) has a large effect on the ridership numbers. However, measuring over the course of a month is limiting for looking at the day of the week since each day will only occur 4 to 5 times in the sample. When combined with the weather conditions this is especially limiting. If rainy days happened to occur during the week and not on the weekend this would give the appearance of increased ridership on rainy days simply because weekend ridership is lower. Similarly, the effect of rain may be understated if rain tended to occur on the weekends.

The linear regression model also has some shortcomings. The residuals are not normally distributed- a long tail on the upper end indicates that many measurements were much higher than expected. Looking at the results in greater detail using a scatterplot of actual vs predicted values reveals two more significant issues. The model sometimes predicts negative values (which is nonsensical for hourly entries) and the highest prediction is 12,045 hourly entries, which is roughly a third of the highest actual value (32,814). This could indicate that one or more features do not have a linear relationship with the number of hourly entries.

```
Out[8]: <matplotlib.text.Text at 0x7f554c5be160>
```


Residuals for Linear Regression
Counts
Difference between Actual and Predicted
Scatterplot showing Residuals
Predicted Outcome of Linear Regression
Actual Outcome