Analyzing the NYC Subway Dataset

Short Questions

Blane Grann

February 18th, 2015

**Section 1. Statistical Test**

1. Which statistical test did you use to analyse the NYC subway data?

In Section 1 exercises, the Mann-Whitney U-test was used to evaluate differences in subway ridership on rainy vs. non-rainy days.

1. Why is this statistical test appropriate or applicable to the dataset?

The Mann-Whitney U-test is a non-parametric statistical test which is useful for this data because the Mann-Whitney U-test does not depend on normally distributed data as, for example, the student t-test does. It also does not require equal sample sizes which is the case of rainy vs non-rainy days.

1. What results did you get from this statistical test?

The Mann-Whitney U test was used to test the null hypothesis that there was no difference between NY subway ridership on rainy vs non-rainy days. The mean number of trips on the NYC subway when it is raining is 1105 trips per hour, while the mean number of trips on non-rainy days is 1090 per hour. Mann-Whitney U = 1924409167.0; rain sample size = 44104; no-rain sample size = 87847; two-tailed p = 0.050.

1. What is the significance of these results?

These results suggest a significant difference between ridership on rainy vs non-rainy days with a p-value of 0.050. Importantly, however, the above statistical test does not consider other important explanatory factors such as day of the week.

**Section 2. Linear Regression**

1. What approach did you use to compute the coefficients theta and produce prediction in your regression model:
2. Gradient descent (as implemented in exercise 3.5)
3. OLS using Statsmodels
4. Or something different?

I used gradient descent in exercise 3.5, and then I used OLS statsmodels.forumlas.api in exercise 3.8

1. What features did you use in your model?

I found the term “features” to be confusing when referring to independent variables. The independent variables I included were: Hour (as a categorical variable), and UNIT (as a categorical variable).

1. Why are these features appropriate?

Intuitively these features seem appropriate because they are all, either directly or indirectly, likely to lead to increased transit ridership. Obviously during rush Hour transit is busier. Specific stations, and therefore UNIT’s, are also obviously busier than others.

1. What is your model’s R2 (coefficients of determination) value?

These variables provided an R^2 of 0.524 using statsmodel.formula.api on Udacity vs 0.502 on my own machine which likely indicates that slightly different data sets were used.

1. What does this R2 value mean for the goodness of fit for your regression model?

R^2 is a statistical measure describing how close the data is to the model. The R^2 value of close to 0.5 is a fairly good fit for modelling transit ridership since the behaviour of people is rather difficult to predict.

**Section 3. Visualization**

Please include two visualizations that show the relationships between two or more variables in the NYC subway data. You should feel free to implement something that we discussed in class (e.g., scatterplots, line plots, or histograms) or attempt to implement something more advanced if you'd like.

1. One visualization should be two histograms of ENTRIESn\_hourly for rainy days and non-rainy days

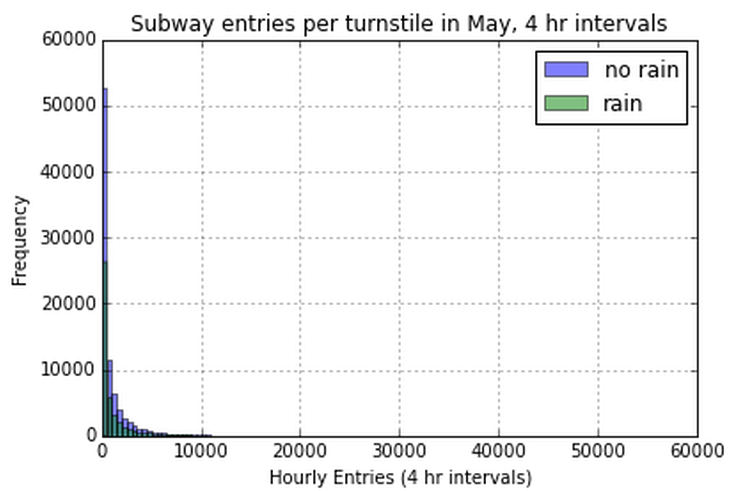


Figure 1

For the month of May, 2011, Figure 1 shows the frequency of 4 hr turnstile intervals (y axis) where the number of hourly entries falls within the specified bins size on the horizontal axis. For example, looking across all the turnstiles on on non-rainy days, more than 50,000 4 hr turnstile periods had 0-500 entries. The data has a non-normal distribution because the majority of the 4hr turnstile intervals have less than 1000 entries.

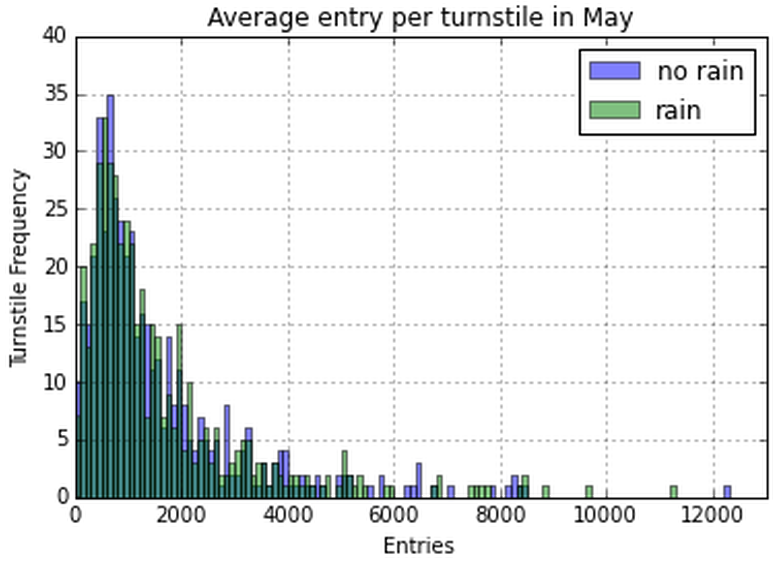


Figure 2

Figure 2 shows the turnstile distrubtion according to the average number of entries during the month of May (rather than 4hr intervals per turnstiles). It is difficult to derive anything conclusive from this plot since the frequency of turnstiles in each ridership “entry” bin varies throughout the spectrum.

1. One visualization can be more freeform, some suggestions are:
2. Ridership by time-of-day or day-of-week
3. How ridership varies by subway station
4. Which stations have more exits or entries at different times of day

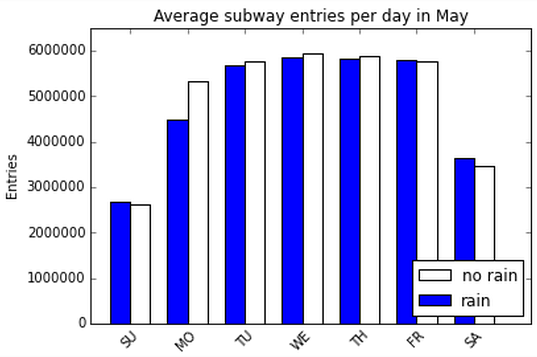


Figure 3

Figure 3 contains a box plot of average, total subway entries by day of week for rainy and non-rainy days. According to the figure, it looks like the average daily entries is lower on rainy days for tuesdays, wednesdays and thursdays, whereas on Friday, Saturday and Sunday there are more subway riders on rainy days. On Mondays there seems to be a lot fewer subway riders on rainy days. However, the underlying data only contains 9 rainy days during the month, and one monday had a particularly low ridership level suggesting that this was probably a holiday. Since the day of the week is an important indicator for ridership levels, looking at average subway entries over the month is probably not going to give us very much insight in to how ridership is affected by rain since a higher frequency of rainy days falling on the weekend would mask increased ridership on week days.

**Section 4. Conclusion**

1. From your analysis and interpretation of the data, do more people ride  
   the NYC subway when it is raining versus when it is not raining?

The Mann-Whitney U test was used to test the null hypothesis that there was no difference between NY subway ridership on rainy vs non-rainy days. The mean number of trips on the NYC subway when it is raining is 1105 trips per hour, while the mean number of trips on non-rainy days is 1090. Mann-Whitney U = 1924409167.0; rain sample size = 44104; no-rain sample size = 87847; two-tailed p = 0.050.

The Mann-Whitney U-test was used to test the null hypothesis that there was no difference in NY subway ridership on rainy vs non-rainy days. With a rainy day mean of 1105 trips per hour and a non-rainy day mean of 1090 trips per hour, results from this test show that subway ridership was statistically higher on rainy compared to non-rainy days (Mann-Whitney U = 1924409167.0; rain sample size = 44104; no-rain sample size = 87847; two-tailed p = 0.050).

1. What analyses lead you to this conclusion?

In this case, the Mann-Whitney U-test is used to test the null hypothesis that the distribution of ridership is the same on rainy and non-rainy days. A two-tailed p-value of 0.50 provides a high level of confidence for rejecting the null hypothesis, because ridership on rainy days is statistically higher.

**Section 5. Reflection**

1. Please discuss potential shortcomings of the dataset and the methods   
   of your analysis.

As mentioned above, the main limitation of the regression analysis is due to the limited number of days for which we have data for. Grouping the data by weekdays, Saturdays, and Sundays/holidays would likely improve the model fit which would in turn help in clarifying the influence of rain in terms of R2.

Another limitation is that the analysis relies on a linear model. There are likely to be other models that would provide a better fit to the data which would help to determine the influence rain has on transit ridership.