Analyzing the NYC Subway Dataset

Section 0. References

[1] http://web.mta.info/developers/resources/nyct/turnstile/ts_Field_Description_pre-10-18-2014.txt (http://web.mta.info/developers/resources/nyct/turnstile/ts_Field_Description_pre-10-18-2014.txt)

[2] https://s3.amazonaws.com/uploads.hipchat.com/23756/665149/05bgLZqSsMycnkg/turnstile-weather-variables.pdf (https://s3.amazonaws.com/uploads.hipchat.com/23756/665149/05bgLZqSsMycnkg/turnstile-weather-variables.pdf)

[3] https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test (https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test)

[4] http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html (http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html)

[5] http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDRegressor.html (http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDRegressor.html)

[6] https://en.wikipedia.org/wiki/Stochastic_gradient_descent (https://en.wikipedia.org/wiki/Stochastic_gradient_descent)

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?

I used Mann-Whitney U-statistic[3][4] to compare the number of entries with rain and the number of entries without rain. I used a two-tail P value. The null hypothesis is there are no difference of number of entries of subway in days with rain and without rain. This null hypothesis is for general population which is all the subways in New York. My p-critical value is 0.05.

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.

Because Mann-Whitney U-statistic belongs to tests which is applicable to no-parametric distribution. I have plot the distribution of numbers of entries. For my best knowledge, this shape cannot generated from a statistical model with parameters. The Mann-Whitney U-statistic is suitable for testing if two samples are generated from the same no-parametric distribution.

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 average number of entries in rainy day is 1105.45 and day without rain is 1090.278780151855. P-value for Mann-Whitney U-statistic is 0.019.

1.4 What is the significance and interpretation of these results?

P-value for Mann-Whitney U-statistic is 0.019 which is less than my p-critcal value. This result reject the null-hypothesis test. The entries in rainy days and days without rain are statistically different with significance.

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: OLS using Statsmodels or Scikit Learn, Gradient descent using, Scikit Learn or something different?

I use stochastic gradient descent implemented in Scikit Learn[5][6].

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

This is my features: 'rain'(rain or not), 'maxtempi'(maxim temperture),'meantempi'(average temperture), 'Hour'(time),'meanwindspdi'(average wind speed),'precipi'(precipition). I also use dummy variables created from 'UNIT'(the ID of turnstile).

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. Your reasons might be based on intuition. For example, response for fog might be: “I decided to use fog because I thought that when it is very foggy outside people might decide to use the subway more often.” Your reasons might also be based on data exploration and experimentation, for example: “I used feature X because as soon as I included it in my model, it drastically improved my R2 value.”

I chose these features by two means. First, I select a lot number of features(more than 80% of all features) based on my intuition. I just exclude the most irrelevant features based on my intuition. For example, I exclude 'maxpressurei'(maxim pressure), because I think most people are not sensitive to small changes of pressures from air. Then, I use a method that continually shrink the model. I will exclude a feature if adding this feature into the model increase the smallest amount of R2 value. In the last, only 1/3 features will be kept in the model.

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

'rain': 21.02098317
'maxtempi': 22.57601299

'meantempi': -42.4147152

'Hour': 34.00429581

'meanwindspdi': 38.18270378

'precipi': 30.77365465

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

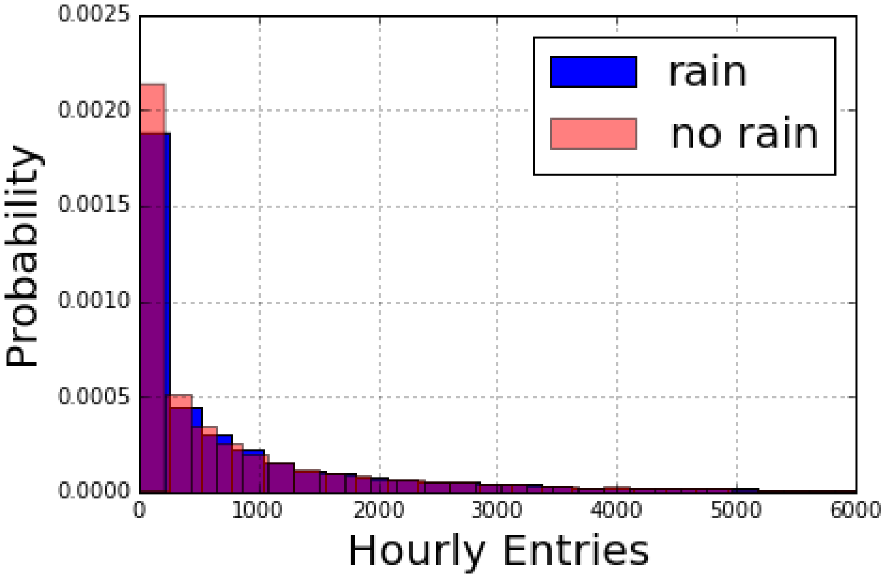
My model's R2 value is 0.397 calculated based on all the data.

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?

R2 value that equals 0.397 means that in average 39.7% in each value(number of entries) can be explained by the regression model. The remaining 60.3% of the values may be due to other factors which are not included in the regression model. I don't think the linear model is appropriate for the dataset given this R2 value.

Section 3. Visualization

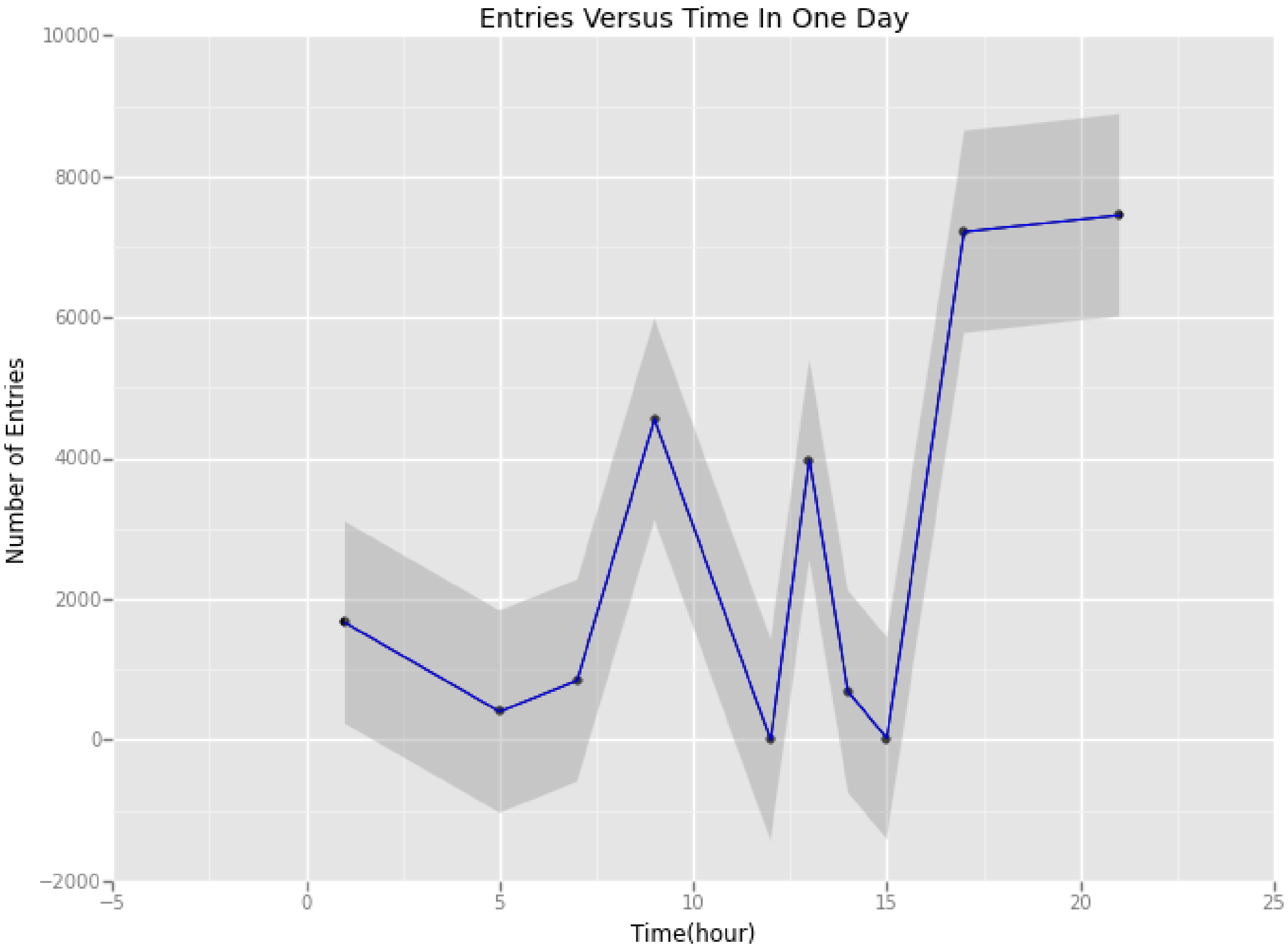
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days. You can combine the two histograms in a single plot or you can use two separate plots. If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case. For the histograms, you should have intervals representing the volume of ridership (value of ENTRIESn_hourly) on the x-axis and the frequency of occurrence on the y-axis. For example, each interval (along the x-axis), the height of the bar for this interval will represent the number of records (rows in our data) that have ENTRIESn_hourly that falls in this interval. Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.



Because number of days without rain is more than number of days with rain. I used the probabilities instead of absolute count of number of entries. There are many 0s in both distributions. However, it is obvious that the percentage of small numbers(less than 200) of rainy days is less than that of days without rain. I think this is the main reason why the average number of entries in rainy day is bigger than that in days without rain.

3.2 One visualization can be more freeform. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like.

Some suggestions are: Ridership by time-of-day, Ridership by day-of-week



This is the ridership by day plot of a particular turnstile. From the plot we can see, more people enter this turnstile in the afternoon than morning. Also, we can find the sign of rush hour. There are peaks at about 9:00 and 17:00 in the plot.

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?

In general, 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.

From the statistical tests, the null hypothesis that assume same number of people ride the NYC subway is reject by a small p-value. The number of people is different. The average number of people who ride subway in raining days is higher than that in days without rain. The conclusion of statistical test support that more people ride subway when it is raining.
The histograms of number of hourly entries of days with and without rain may reveal the difference of ridership of subway in days with and without rain more intuitively. In rain day, the small amount of hourly entries(most of them is 0s) decrease. This histogram means that number of tiny hourly entries in the rainy day reduced and the average number of hourly entries in rainy day increases.

*Also, the conclusion that more people ride the NYC subway can be made based on the coefficients of my linear model. In my model, 'rain' is a feature in which 1 stands for rainy days and 0 days without rain. The coefficient of the 'rain' feature is 21.02 which is poistive. Every time it is raining, the feature will be 1 and the predicted hourly entries would gain a number which is 1*21.02 and when it is not raining, the feature will be 0, the predicted hourly entries would gain nothing. So, the poistive coefficient of 'rain' feature reflects that the hourly entries of subway would be bigger in rainy days than days without rain.*

Section 5. Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis, including: Dataset, Analysis, such as the linear regression model or statistical test.

Drawbacks of the dataset: 1, Dates of the dataset are not representative enough. It is only the data from May. Within one month, the temperatures or other weather conditions very too little to draw a conclusion from them.
2, The data was systematic diversely distributed. The data was collected every 4 hours. Many details according to time may be missed.There are only 6 point in one day. And it is fairly difficult to predict on the time when the data was not produced using such diverse data point.

3, The calculation of hourly entries can be improved. The original hourly entries were calculated using differece of readings[1][2]. However, the intervals of times between adjacent records are not always 4 hours. The calculated hourly entries were based on different intervals of time. This irregulate values of hourly entries made training and test more inaccurate.

Drawbacks of the simple linear model:

1, Linear model alway assume the features are independent. However, this may not hold true for our data. I have calculated the difference of ridership between days with and without rain in several different times(1:00, 9:00 and 21:00). In some times, number of entries in rainy day are higher and other times not. This means that the ridership of rainy days depend on time. This is the evidence of dependent.

2, Prediction of linear model is linear on features. But I think that the ridership of subway may not linearly depend on some features. For example, the number of people who take subway may increase when there is small or not so large amount of precipitation. But is rain or snow was to heavy for people going out. the number of people who take subway may decrease also.

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

Please see Answer To 5.2 (https://rawgit.com/adenguo/work2/master/Answer%20To%205.2.html)