Analyzing the NYC Subway Dataset

Questions

Section 0. References

Please include a list of references you have used for this project. Please be specific - for example, instead of including a general website such as stackoverflow.com, try to include a specific topic from Stackoverflow that you have found useful.

http://en.wikipedia.org/wiki/Student%27s_t-test
http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test
http://en.wikipedia.org/wiki/Least_squares
http://en.wikipedia.org/wiki/Coefficient_of_determination
http://en.wikipedia.org/wiki/Gradient_descent
http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html
http://math.usask.ca/~laverty/S245/Tables/wmw.pdf
http://www.statstutor.ac.uk/resources/uploaded/mannwhitney.pdf
http://bconnelly.net/2013/10/summarizing-data-in-python-with-pandas/
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?

I used Mann-Whitney U-Test.
I used a two-tailed P value because we cannot assume in advance the direction (increase or decrease) of the effect rain has on ridership.
The null hypothesis states that when drawing random values from the two populations, there is an equal chance (50%) that one distribution will generate a gigher value.
I used a p critical of 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.

This test is applicable because it does not assume the samples follow any particular statistical 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.

With rain mean 1105.4463767458733
Without rain mean 1090.278780151855
U score 1924409167.0
1 sided p value 0.024, 2 sided p value 0.049

1.4 What is the significance and interpretation of these results?

If we assume a significance level of 95% (relating to p-critical of 0.05) and given that we found a p value of 0.049 we can safely reject the null hypothesis and say that there is a statistical significant preference for riding the subway on rainy days.

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:

a. Gradient descent (as implemented in exercise 3.5)
b. OLS using Statsmodels
c. Or something different?

I used (a), Gradient descent.

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

At first I decided to use all of them and let the algorithm sort the coefficients out. But it was a poor decision, I ignored multicollinearity between the variables.
I settled on Hour, rain, mintempi, fog ,and UNIT transformed to dummy variables

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 R^2 value.”

Hour: The time of day influences all human behaviour. Subway ridership cannot be any different.
Rain: A statistical significant relation between the predicted variable and rain has already been found in the last exercise, it only makes sense to use it.
mintempi: I thought that by using minimal temperature I would have better predictions for the wee hours of the morning. The cold temperature might make a difference between riding a bike or choosing the subway.
fog: Like rain, fog might influence ridership.

I left out EXITSn_hourly. Although it improved the R^2 score it would not be possible to use it a real situation because we would not have the value before hand.

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

[468.01338394 -10.29976963 -87.36252696   67.02156388] for features [Hour, rain, mintempi, fog] respectively

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

0.464892338257

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?

This R^2 states that about 46% of the variance in the data can be explained by the predictor variables.
Altought an R^2 score of close to 50% is considered good for predicting human behaviour, the model might not be able to provide sufficiently precise results. The model might not be appropriate for predicting ridership, depending on the desired outcome.

Section 3. Visualization

Please include two visualizations that show the relationships between two or more variables in the NYC subway data.
Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

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.

The visualisation is named vis1.png.
According to the suggestions given on the last review I used a log scale for the X axis which centered the chart and allowed much better visualization.
I also managed to get a legend working this time.

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

The visualisation is named vis2.png. I chose a boxplot to display Ridership by time of day.

Section 4. Conclusion

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

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?

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.

Yes, more people ride the subway when it is raining. The first visualization made already suggested this showing higher frequencies for every bin of histogram. This was further corroborated by the statistical analysis that showed "statistical significant preference for riding the subway on rainy days".

Section 5. Reflection

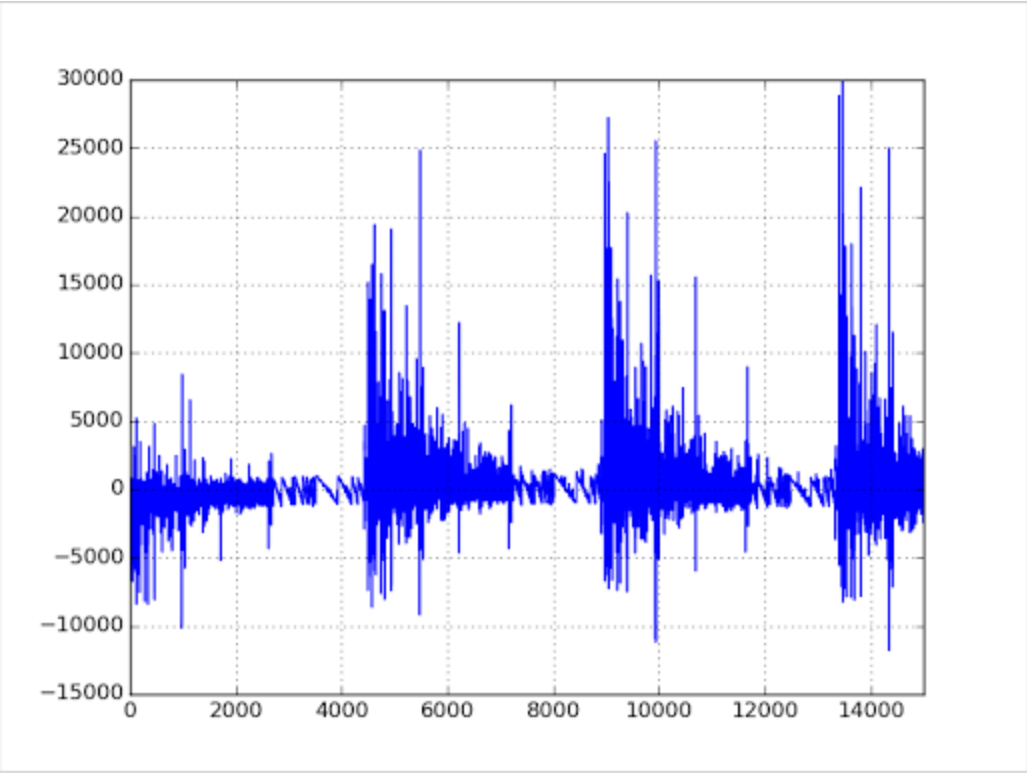
Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

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

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

The database is limited to one month of data. Having more data from other monthds might have contributed to the overall analysis. I believe this would give the model the ability to account for more seasonal events like season of the year and holidays. It would however take a lot of time (years) to generate sufficient data.

Regarding the linear regression model, I've attached below a plot showing the residuals for each data point. We can clearly see that the residuals follow a periodical pattern. At first I tried associating it with the Hour variable but it showed no correlation. This could mean that the current variables cannot (as the low R^2 value showed before) explain all the variance in the data. Or it could mean that the problem is in fact non linear, and the linear model produced is not a good fit and can only predict the values that are close the the linear region.



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