Analyzing the NYC Subway Dataset

Elena Korotkina
elena@elkore.com

Overview

This project consists of two parts. In Part 1 of the project, you should have completed the questions in Problem Sets 2, 3, 4, and 5 in the Introduction to Data Science course.

This document addresses part 2 of the project. Please use this document as a template and answer the following questions to explain your reasoning and conclusion behind your work in the problem sets. You will attach a document with your answers to these questions as part of your final project submission.

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://www.timothylim.info/blog/2015/4/3/sqlite-naming-conventions - very important! otherwise using SQL queries on the dataset produces the SQLLike naming convention error (since the dataset columns are not lower case)
http://matplotlib.org/users/legend_guide.html
http://stackoverflow.com/questions/6352740/matplotlib-label-each-bin - labeling
http://stackoverflow.com/questions/12125880/changing-default-x-range-in-histogram-matplotlib - setting range

http://en.wikipedia.org/w/index.php?title=Linear_least_squares_(mathematics) - the most simple explanation of the sum of squares
http://statsmodels.sourceforge.net/devel/example_formulas.html#ols-regression-using-formulas
http://www.datarobot.com/blog/ordinary-least-squares-in-python/
http://statsmodels.sourceforge.net/0.6.0/examples/notebooks/generated/formulas.html
http://docs.ggplot2.org/current/scale_continuous.html

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?

Mann-Whitney U test was used to analyze the NYC subway data. Two-tail P value was used, since we needed to know if the two samples are different. The null hypothesis was that there is no statistical difference in NYC subway ridership on rainy days versus non-rainy days. The P value for one-tail test was 0.0249, multiplied by 2 the P value for the two-tailed test is 0.0499 (just a bit less than the alpha value 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.

The datasets representing the subway ridership on rainy and non-rainy days are not distributed normally. Therefore we cannot use the Welch's t test to analyse this data.

Mann-Whitney does not assume the normal distribution of the data and is used to test if the two samples that are being compared come from the same population.

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 results of the Mann-Whitney tests are as follows:

with_rain_mean: 1105.4463767458733
without_rain_mean: 1090.278780151855
U: 1924409167.0
p-value: 0.024999912793489721

1.4 What is the significance and interpretation of these results?

The P value for one-tail test was 0.0249, multiplied by 2 the P value for the two-tailed test is 0.0499 (just a bit less than the alpha value of 0.05). This means that we can reject the null hypothesis and assume that the distribution of the NYC subway ridership is statistically different for rainy and non-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:

1. Gradient descent (as implemented in exercise 3.5)
2. OLS using Statsmodels
3. Or something different?

I used both the gradient descent and OLS (using Stats models).

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

For the gradient descent method, I used Rain column as my features. Dummy variables UNIT and Hour were used.

For the OLS (Ordinary Least Squares) method, I used Hour, UNIT and Rain as features.

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 felt that rain would influence people’s decision to ride a subway (versus walk).
The UNIT seemed like an obvious influential factor (some subway stations are busier than others due to their location), and a quick glance at the data confirms that. “Hour” is also important, since Subway ridership heavily depends on the time of the day and rush hours. Since the sample contains data for the month of May only, I did not feel like the temperature would be much of a factor (the data in the temperature columns contained the temperature of at least 50F, which is not cold to walk).

Also, removing or adding different features changed the R2 value. For instance, with OLS method the following formular produces the R2 value of 0.52: "ENTRIESn_hourly ~ C(Hour) + UNIT + rain", and removing the UNIT from the list of features reduces the R2 value to only 0.117.

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

```
[  2.76230988e+01  -3.44287255e+01 ]
```

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

For the gradient descent method the R2 value was 0.504790944024
The R2 value returned with the OLS method was 0.524418465187

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 means the percentage of the of the Subway ridership that can be explained by the variation of the parameters in the model. The R2 values of my models are not enough to make these models a good fit.

A residual is a difference between a predicted value and the real value. When the variability of the residuals is relatively small, the model is effective at predicting the future values. When the variability of the residuals is large, it may mean that there is no correlation between the x and y, or the selected features have no or little impact to the y values.
In the R2 equation the dividend is the sum of the differences between the real and expected values, squared. The smaller the difference is (as compared to the varience), the closer to zero the equation is, which makes the R2 value closer to 1. The closer the R2 value is to 1, the better is the model at predicting values.
The R2 value of 0.52 is right in the middle between 0 and 1, meaning that the model is possibly weak.

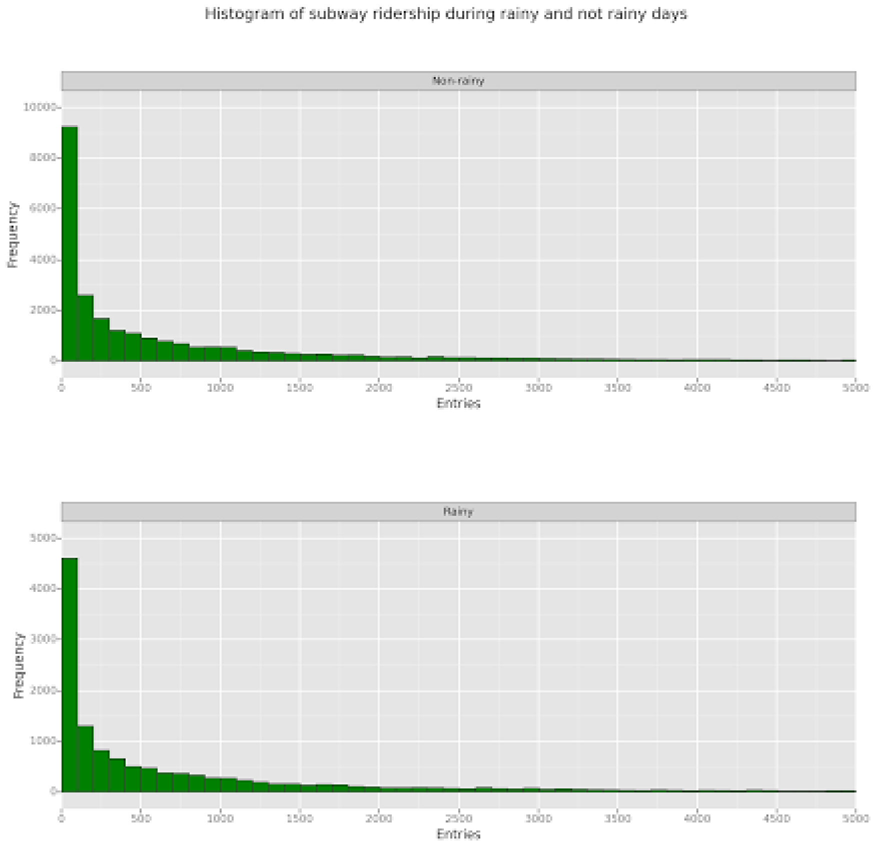
Plotting the residuals (see the plot image at the bottom of this document) shows that the residuals are fairly normally distributed around 0, but also contain a few large values, particularly as the values get larger. We also see that the residuals follow a cyclical pattern (the errors are decreasing and increasing at certain intervals), which is a sign that the model may not be a good fit for this relationship.

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.


Histogram of subway ridership during rainy and not rainy days

Non-rainy

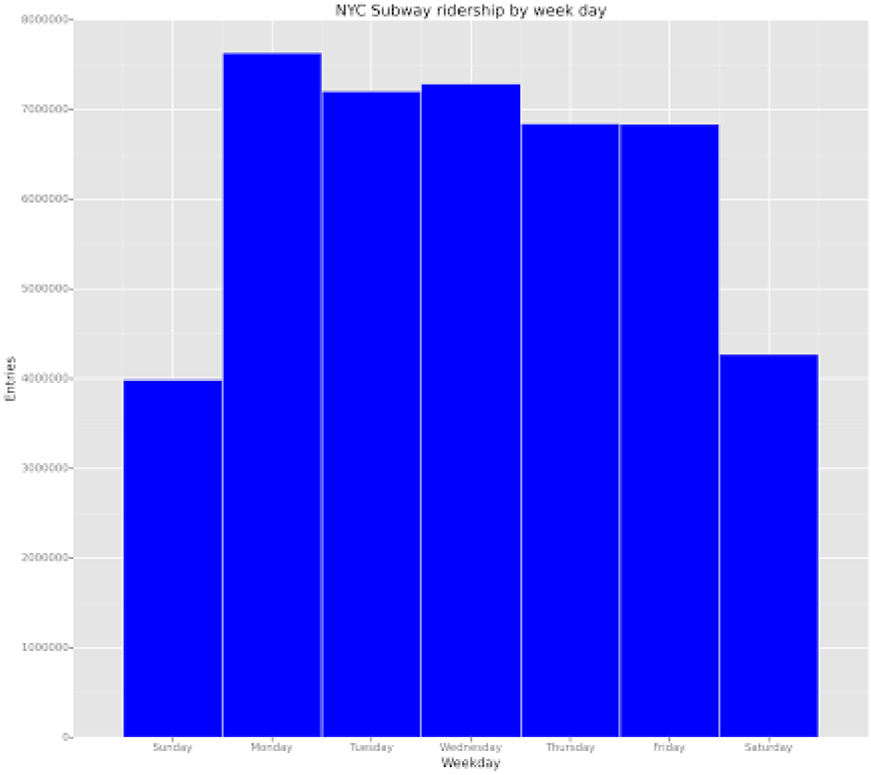
Rainy

This histogram visualizes Subway ridership hourly counts during non-rainy and rainy days. The shapes of both distributions are skewed to the left (fewer hoursly entrances are more frequent). I limited the x-axis to the maximum value of 5000 to get a better view on the data. The value of the Y-axis for the non-rainy day are higher than the values of the Y-axis for the rainy days (which can be explained by the fact that there are more non-rainy days in the dataset than rainy days).

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 visualization represents ridership by day of the week:



This histogram shows the total number of entries by the day of the week. Based on this visualization we can see that Monday in May had the most number of Subway entries, and Sunday was the slowest 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?

Based on the Mann-Whitney test, we can conclude that more people ride the NYC subway when it is raining than what it is not raining however that link is rather weak.

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 P-value of the Mann-Whitney test was 0.0499, which barely allows us to reject the null hypothesis (that Subway ridership is the same in rainy and non-rainy days) based on the alpha of 0.05. This allows me to conclude that there is a correlation between rain and heavier Sunway ridership, but the correlation is weak.

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:

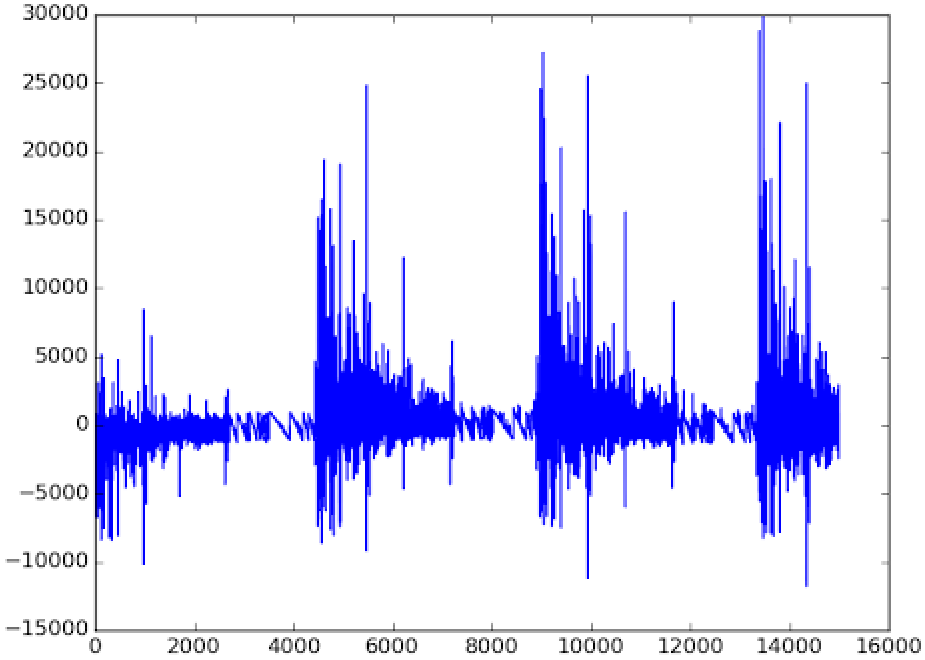
1. Dataset,
2. Analysis, such as the linear regression model or statistical test.

The main shortcoming in the given dataset I see is that the “Rain” column provides potentially erroneous data, considering that Subway ridership is analyzed by the hour, and the rain data is provided by day. It may rain in the morning but not for the rest of the day, and vice versa. When analyzing the influence of rainy weather on Subway ridership the availability of the rain data by the hour is critical in my opinion. Otherwise we are basing our analysis on inaccurate data.

As far as the shortcomings of the analysis, based on what I’ve read about the Mann-Whitney test, the disadvantage of this test is that it is less accurate than parametric tests (if there is a difference between two groups, Mann-Whitney test is less likely to find it)

The analysis of the data suffered from the above mentioned dataset limitation (that the Rain column included the data by the day and not by the hour).

Looking at the plot of residuals shows that there are a few rather large residuals, particularly as teh values of the entries get larger:



This means that the model's fit is less accurate for larger values.
Also, residuals seem to follow a cyclical pattern, with the predicion errors decreasing and increasing at certain intervals. This may mean that the relationship may not be exactly linear (e.g., a line with curves may be a better fit to represent this relationship).

The downside of the linear regression model is the expectation that the relationship is linear, while in reality many relationships are not (for example, the relationship between income and age).

Also, linear regression only looks at the mean and variance of the dependent variable. In reality, we should consider the mean of the variance and the extreams of the dependent variable as well.

And linear regression may show the correlation between variables, but it does not show the cause and effect.

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

Interestingly, there are almost 30 million more Subway entries than exits in total, and 30 million is almost 20% of the total number of entries.
I cannot think of any other reasons for this difference other than errors in the data. (And 20% of erroneous data may be significant enough to invalidate the results of the above tests).