Analyzing the NYC Subway Dataset

Olivier Bézie

2015-05-04

Contents

Overview

This project consists of two parts. In Part 1 of the project, I 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. I have used this document as a template and answered the following questions to explain my reasoning and conclusion behind my work in the problem sets. I have attached this document with my answers to these questions as part of my final project submission.

Section 0. References

This is a list of references I have used for this project, including a specific topic from Stackoverflow that you have found useful.

- http://pandas.pydata.org/pandas-docs/stable/visualization.html#scales hist() is equivalent to plot(kind='hist')
- http://matplotlib.org/1.4.3/api/pyplot_api.html#matplotlib.pyplot.axis limit x coordinate to 6000, label of the axis, title of the histogram, customize legend
- https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test#Assumptions_and_formal_statement_of_hypotheses helps me define the null and alternate hypothesis
- http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html to check that this function returns one-sided p-value
- http://stackoverflow.com/questions/24109779/running-get-dummies-on-several-dataframe-columns has been useful when adding several dummy variables for the linear model
- http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit to interpret the R-squared value

Section 1. Statistical Test

1.1 Which was statistical test use to analyze the NYC subway data? Was a one-tail or a two-tail P value used? What is the null hypothesis? What is the p-critical value?

A histogram of the data (see Section 3.) shows that the distribution is not normal and one-tailed. Welsh's ample T-test cannot be used. Mann-Whitney U-Test has been used instead.

Null Hypothesis: there is no significant difference in the number of entries between rainy and non-rainy days. The alternate hypothesis is that one of the populations (ridership during rainy or non-rainy day) is greater than the other. The p critical value is 0.05.

1.2 Why is this statistical test applicable to the dataset?

The general assumptions used to formulate the sample U-Test applies to the distribution of ridership during days with rain or days without rain:

- All the observations from both groups are independent of each other: The ridership during rainy or non-rainy days does not influence each other's
- The responses are ordinal: the turnstile data counts the number of entries or exists
- The distributions of both groups are equal under the null hypothesis: The histogram shows that the 2 distribution of number of entries per hours follow the same asymptotic trend.
- Under the alternative hypothesis, the probability of an observation from the non-rainy days exceeding an observation from the rainy days (after exclusion of ties) is not equal to 0.5.

1.3 What were the results from this statistical test?

$\mu_{wt} = 1105$ Mean of the hourly ridership with_rain_mean
$\mu_{wo} = 1090$ Mean of the hourly ridership without_rain_mean
p = 0.019309634413792565 Sample U-Test p-value

1.4 What is the significance and interpretation of these results?

$p < 0.05$, the null hypothesis is rejected. There is a significant difference in the number of entries between the rainy days and the non-rainy days.

$\mu_{wo} < \mu_{wt}$, the ridership during rainy days is statistically greater than the ridership during non-rainy days.

Section 2. Linear Regression

2.1 What was the approach used to compute the coefficients theta and produce prediction for ENTRIESn_hourly in the regression model:

I have tried both Gradient descent (as implemented in exercise 3.5) and OLS using Statsmodels. Both of them gave me about the same R2 value with the same features and dummy variables.

2.2 What were the features (input variables) used in the model? Were any dummy variables as part of your features?

The features used were: the stations (UNIT), the day of the week (weekday) calculated from the date (DATEn), the hour (Hour), the rain event (rain), the fog event (fog) and the average temperature (meantempi). The features UNIT, weekday and hour were used as dummy variables.

2.3 Why were these features selected in the model?

I initially selected the rain, the hour, the temperature and the fog features by intuition, reflecting my own experience of using the public transportation. Rain as the subway provides a good shelter☺; the hour as during pick hours there are more people in the subway; temperature as rain generally brings a lower temperature and the subway could represent a warmer place; the lack of visibility in the fog can make people uncomfortable and the subway look safer.

My initial selection was refined by trying different combination in the regression model:

UNIT	weekday	fog	rain	precipi	Hour	meantempi	mintempi	OLS		Gradient descent	
								R^2	Delta	R^2	Delta
			x					0.000		0.000	
x			x					0.418	0.418	0.418	0.418
x				x				0.418	0.000	0.418	0.000
x		x		x				0.418	0.000	0.418	0.000
x		x		x	x			0.501	0.083	0.501	0.083
x		x		x	x	x		0.502	0.001	0.502	0.001
x		x		x	x		x	0.502	0.000	0.502	0.000
x	x	x		x	x		x	0.514	0.012	0.514	0.012
x	x		x		x			0.514	0.000	0.514	0.000
x	x		x		x	meandewpti		0.514	0.000	0.514	0.000
x	x	x			x			0.514	0.000	0.514	0.000

The above results show that in fact the regression model is mostly influenced by 3 features: the station, the hour of the day and the day of the week. The weather features have very little effect in the prediction of ridership.

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

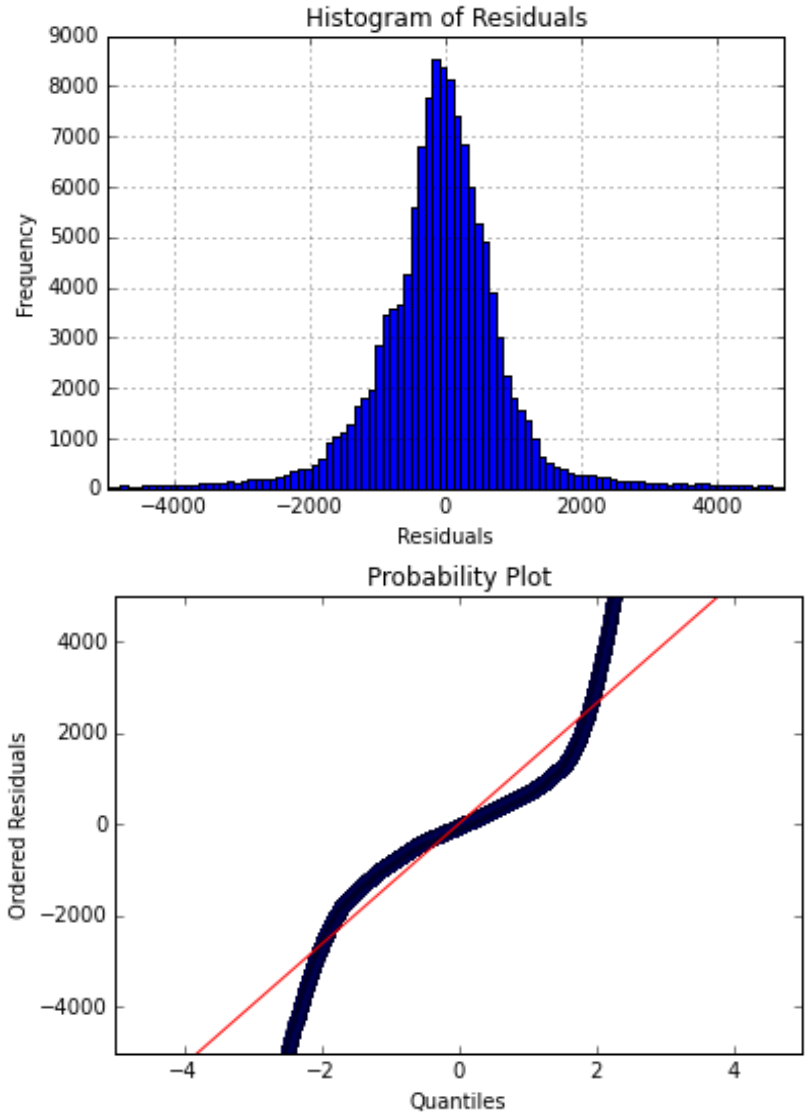
	rain	fog	meantempi
Gradient descent coef.	-4.17E+01	3.73E+01	-5.25E+01
OLS coef.	1.25E+13	-97.498	107.944

2.5 What is the model's R2 (coefficients of determination) value?

R2 = 0.514

2.6 What does this R2 value mean for the goodness of fit for the regression model? Is this linear model to predict ridership appropriate for this dataset, given this R2 value?

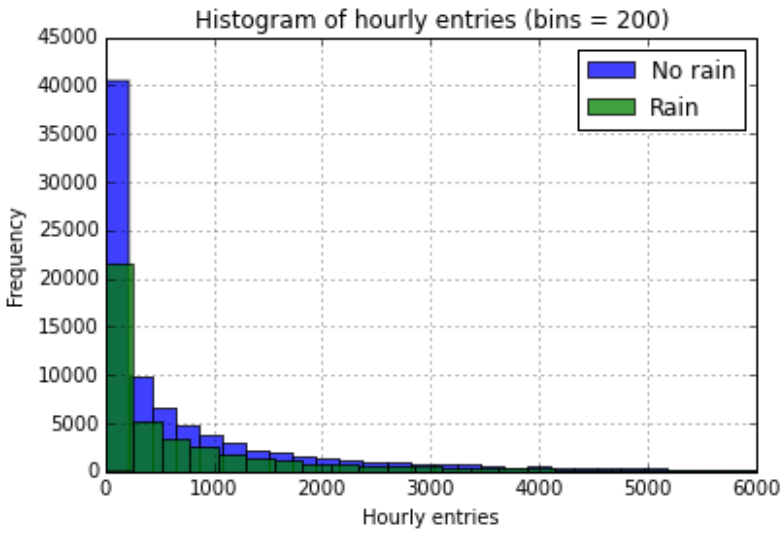
The goodness of fit of the regression model is not good for the weather features. Only about 51% of the total variation of the number of entries in the subway can be explained by the linear relationship between the station, the hour of the day and the day of the week.



The histogram of residuals or probability plot gives a visual confirmation of the non-linearity of the model: The distribution of the residuals is not normal. There are uncontrolled or unknown features which influences the ridership. In other words the linear model is not appropriate for this dataset.

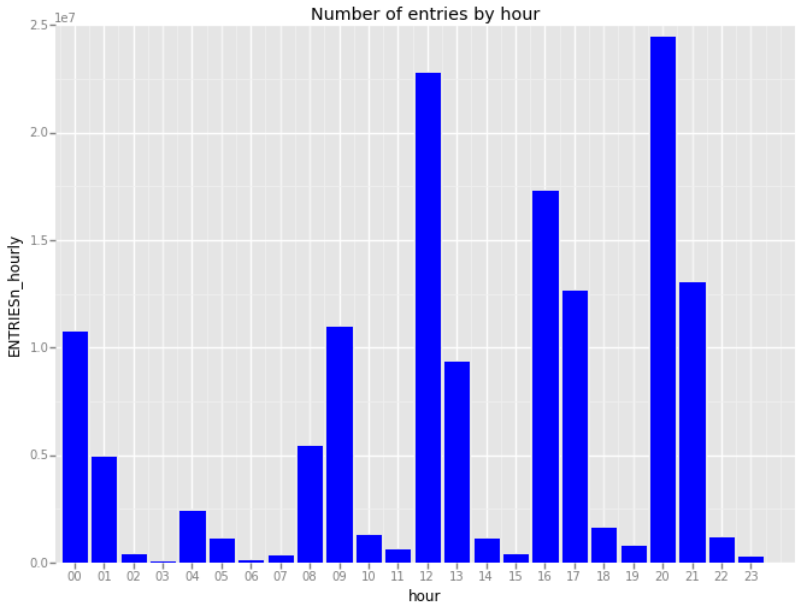
Section 3. Visualization

4.1 Histograms of ENTRIESn_hourly for rainy days and for non-rainy days

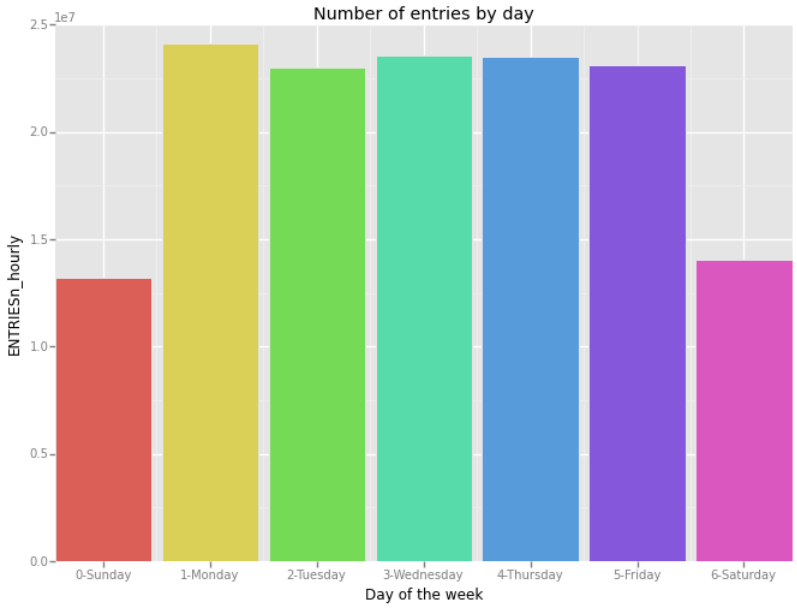


The rainy days and the non-rainy days distributions are very similar (asymptotic shape).

4.2 Ridership by time-of-day and day-of-week



The above diagram shows the ridership by time-of-day. The number of entries per hour is very different from one hour to the other. From 8am to 12am picks of entries can be observed every 4 hours.



The above diagram shows the ridership by day-of-week. The number of entries per day is homogenous during the working days and shows a significant decrease (about 40%) during the week-ends.

Section 4. Conclusion

4.1 From the above analysis and interpretation of the data, do more people ride the NYC subway when it is raining or when it is not raining?

Statistically speaking people tends to ride more the NYC subway when it is raining. However no linear relationship between rainy/non-rainy days and the number of entries could be found to establish a prediction model. Inputs to the linear prediction model like the subway stations, the day of the week and the hour of the day have much more influence on the ridership than any of the weather features especially the rain one.

Overall the influence of the rain on ridership in the NYC subway is not significant despite it exists. On one hand the statistical test shows an increase though rather small of the ridership when it rains, on the other hand rainy days have an insignificant influence on the prediction of ridership. This could be explained by the fact (pure speculation/intuition here, This could be the purpose of another analysis) that there is only a minority of people separated with a walkable distance from their usual destinations that would take the subway during rainy days, the others having to take the subway anyway, rain or not.

4.2 What analyses lead to this conclusion?

The main driver of the "people tends to ride more the NYC subway when it is raining" is the result of the Mann-Whitney U-Test with a p value of $0.019 < 0.05$. The null hypothesis (no significant difference in the number of entries between rainy and non-rainy days) is rejected. With a rainy average (1105) greater than the non-rainy average (1090) the alternate hypothesis leads to the conclusion that there are in average more entries during rainy days. The difference between the 2 average is small (15 entries more during rainy days) but it is higher during the week-ends (+41 entries).

Two linear regression methods have been used, the Gradient descent and OLS, both of them giving identical R-squared value (0.51). The step by step approach to identify the features with the biggest influence on the ridership clearly dismissed the weather features (rain, fog or temperature) from the equation: They add no more than $1^{o}/_{oo}$ to the R-squared value.

Section 5. Reflection

5.1 Potential shortcomings of the methods of the analysis

I think that it would have been better to work with hourly weather data (if they exists) and map them with the hourly entries. Mapping hourly entries with daily weather data – if it rains once it rains for the whole day – is not representative of the reality.

It does not rain the same way everywhere and the location/distance of the weather station to the station could bring more weight to the rain feature.

The U-test rejects the null hypothesis but the increase in the average ridership does not look that significant. Other method of analysis to minimize the variability between stations, days of the week and hours would be welcome though at this point I don't have too much ideas of what they could be.

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