Analyzing the NYC Subway Dataset

Questions

Overview

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

Statistics for Management and Economics, 10e, Keller: Chapter 17-Multiple Regression, Chapter 19 – Non parametric tests

http://www.algosome.com/articles/dummy-variable-trap-regression.html

http://blog.minitab.com/blog/adventures-in-statistics/how-high-should-r-squared-be-in-regression-analysis

http://blog.minitab.com/blog/adventures-in-statistics/how-to-interpret-regression-analysis-results-p-values-and-coefficients

http://www.originlab.com/doc/Origin-Help/Residual-Plot-Analysis

http://blog.minitab.com/blog/adventures-in-statistics/why-you-need-to-check-your-residual-plots-for-regression-analysis

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 (Wilcoxon Rank Sum Test) for distributions of 2 populations, data are interval and 2 data sets are independent.
- p-critical value is 0.05 (5%)
- Two–tailed P value was used to test the difference of ridership on rainy days and on non-rainy days.
    - H_o= P(rainy>non-rainy)=0.5: the two population locations are the same. The distribution of entries per hour on rainy day is at the same location of the distribution of entries per hour on non-rainy day
    - H_1= P(rainy>non-rainy)≠0.5: For the population. The location of distribution of entries per hour on rainy day is different from that of entries per hour on non-rainy day.

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.

- There are two populations of interest: (1) ridership (entries per hour) on rainy day; (2) ridership (entries per hour) on non-rainy day. The two samples are independent (not influenced by each other)
- The histogram of entries per hour of these two data sets show positively skewed distribution (not normally distribution) and entries per hour are interval so a non-parametric test (Mann-Whitney) should be used instead of a t-test for independent samples.

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.

- Means of entries per hour (sample) with rain: 1105.446
- Means of entries per hour (sample) without rain: 1090.279
- U statistic = 1924409167.0
- p(one-tailed) = 0.02499
- p(two-tailed)= 0.049999

1.4 What is the significance and interpretation of these results?

p(two-tailed) < p-critical value so we can reject H_o.

The data provide sufficient evidence to infer that there is difference between ridership on rainy day and on non-rainy day. Although the test doesn't check the direction of difference but as mean of ridership on rainy days is larger than that on non-rainy day. We assume that the number of entries on rainy days is higher.

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

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

Input variable: fog, meantempi, meanwindspdi

Dummy variable: Unit, hour

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.

- Intuitively, I used fog because when it's foggy, people tend to use subway more. Maybe it's more dangerous to drive a car
- Based on my intuition, I think when temperature increases, the subway ridership tends to decrease. And when I tested it in the model, it did increase the R^2 value
- I used mean wind speed as it increases my r square. It's possible that when it's windy outside, it might be more comfortable for people to use subway underground.
- After testing several models, I didn't include rain (and precipitation) in my model as it has probable correlation with other variables such as wind speed, temperature, which makes the model more prone to multicollinearity. On the other hand, p value of rain in the model is too high (larger than 0.05) to be statistically significant. (The similar patterns happen to precipitation as rain and precipitation refer to the same phenomenon)

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

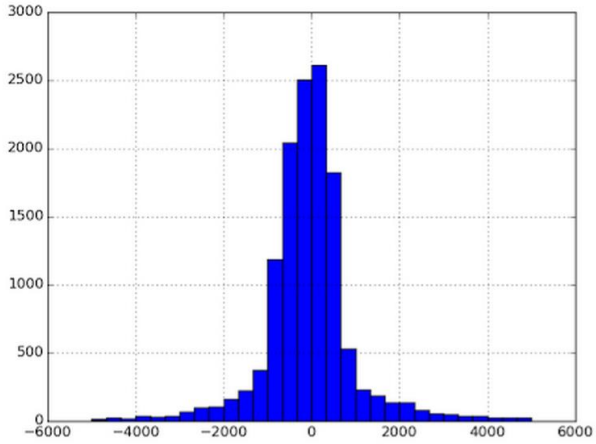
	coefficients	Standard error	t	P value
Constant	1786.2029			
Fog	181.9476	39.373	4.621	0.000
Mean temperature	-9.5090	2.361	-4.027	0.000
Mean wind speed	29.8171	7.920	3.765	0.000

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

R^2 value is 0.522671797413

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?

- The R^2 value of 0.5227 means 52.3% of the ridership (entries per hour) is explained by the set of features (fog, mean temperature, mean wind speed and precipitation). Given the R^2 value of more than 50%, the model is acceptable to predict ridership.
- Histogram bellows show the distribution of residuals. This proves that the required conditions for error variable (difference between predicted and observed values) are satisfied: normality, mean = 0, stand deviation is constant and errors are independent. Hence the multiple linear regression model is valid.



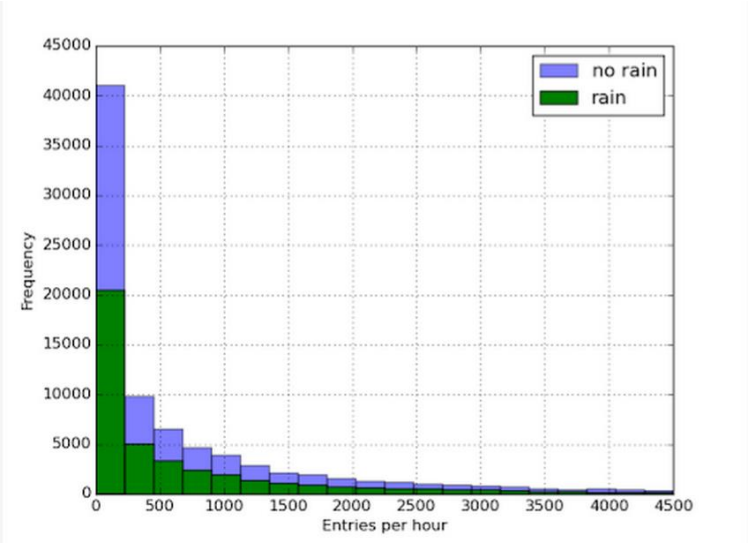
Section 3. Visualization

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.

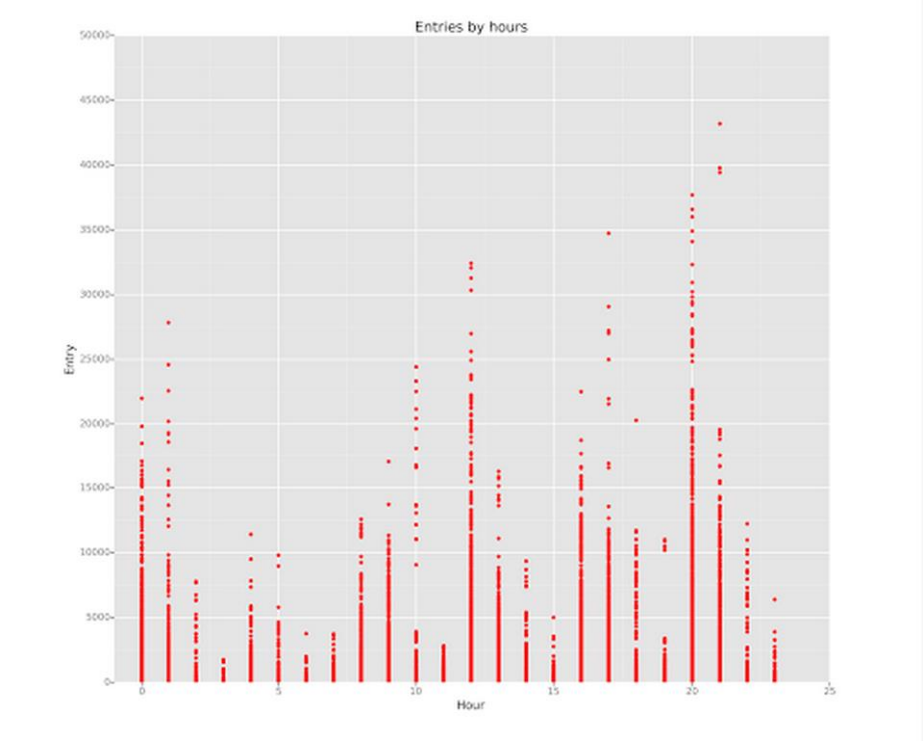
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

- Histogram of Entries hourly on rainy days and Entries hourly on non-rainy days are positively skewed. For both rainy and non-rainy days, there are mostly 0-500 entries per hour.



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.

The scatter plots below shows the number of entries by hours. Most ridership occur at 12pm, 4pm-6pm and 8pm-9pm.



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?

It seems more people ride the NYC subway when it is raining based on the Mann-Whitney U test and linear regression of rain on ridership, but we cannot conclude if the rain truly correlates with this variability in ridership.

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 Mann-Whitney U test, there is difference between ridership when it's rainy and when it's not rainy with p value smaller than 0.05. However, from linear regression, the result of p value and coefficient of rain fluctuates when we put different features into the models.

For example, if the model simply includes rain as feature, the coefficient is 78.46 and p value is 0.011. In this case, rain is positively correlated with ridership with p value significant

	coefficients	Standard error	t	P value	R square
Rain	78.4592	30.924	2.537	0.011	0.520

With rain and fog in the model, rain is still positively correlated with ridership but p value is no longer significant (increases to 0.368)

	coefficients	Standard error	t	P value	R square
Rain	30.9088	34.367	0.899	0.368	0.521
fog	138.1864	43.636	3.167	0.002	

When other features are included in the model, rain is negatively correlated with ridership (coefficient is negative) and p value increases. The change in sign of rain's coefficient may be caused by multicollinearity (rain is correlated with other features such as mean temperature or mean wind speed)

	coefficients	Standard error	t	P value	R square
Rain	-14.4979	35.041	-0.414	0.679	0.523
fog	190.3090	44.257	4.300	0.000	
meantempi	-9.6792	2.397	-4.038	0.000	
meanwindspdi	29.9611	7.928	3.779	0.000	

As the results from the test and linear models are not consistent and p value of rain is not significant, we don't have enough evidence to conclude whether rain is linear correlated with ridership.

Section 5. Reflection

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

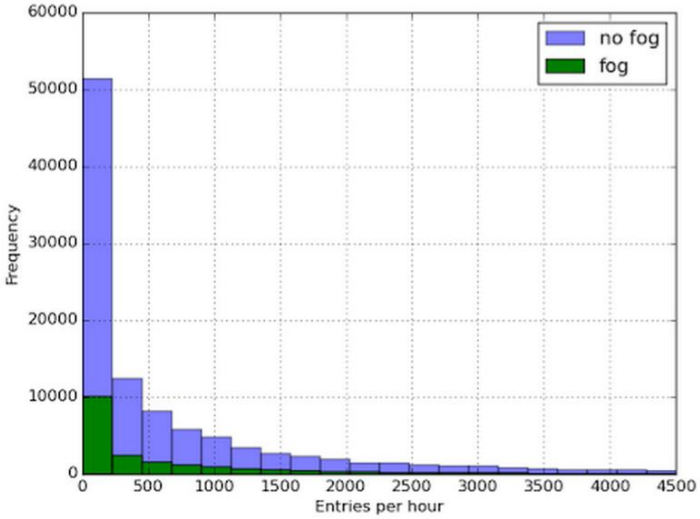
5.1 Please discuss potential shortcomings of the methods of your analysis, including:

Statistic test: The non-parametric test such as Mann-Whitney only tests if the two population are different from one another, whether their locations are different. It doesn't mean the difference can be

explained by or correlated with a specific feature in the populations. I conclude that the statistic tests are used to make comparison of two populations without considering all other factors.

Linear regression checks the correlation between an independent variable and one or several features in the population. Multiple linear regression can be prone to errors when multicollinearity exists. Different models can yield different results; and p-value of a feature can be significant in one model and not significant in another. Besides, if one or more of the required conditions such as normality of residuals distribution are violated, the results may be invalid.

5.2 (Optional) Do you have any other insight about the dataset that you would like to share with us?
While rain data shows weak correlation with ridership. Fog seems to have strong and consistent correlation with numbers of entries in NYC subway.



Histogram of entries per hour on foggy days and non-foggy days shows non-normal distribution.

Mann-Whitney U test for fog:

- H_o= P(foggy>non-foggy)≠0.5: the two population locations are the same. The distribution of entries per hour on foggy day is at the same location of the distribution of entries per hour on non-foggy day
- H_1= P(foggy>non-foggy)≠0.5: For the population. The location of distribution of entries per hour on foggy day is different from that of entries per hour on non-foggy day.

Mean of sample with fog: 1154.66
Mean of sample without fog: 1083.45
U statistic = 1189034717.5
p-value (one-tail)= 6.1×10^{-6}
p-value (two-tail) = 1.22×10^{-5}
p-value is statistically significant, which concludes ridership in the population on foggy and non-foggy days is different.

Linear Regression model: p-value of fog remains lower than significant level, showing the strong correlation between this feature and the ridership.