Analyzing the NYC Subway Dataset

Casimir COMPAORE

Section 0. References

List of references used for this project.

http://statsmodels.sourceforge.net/0.5.0/generated/statsmodels.regression.linear_model.OLS.html
http://scikit-learn.org/0.14/modules/generated/sklearn.linear_model.SGDRegressor.html
http://www.skymark.com/resources/tools/normal_test_plot.asp
https://en.wikipedia.org/wiki/Multicollinearity
http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.probplot.html
https://docs.google.com/document/d/1S4Gk42ZPBKAUZh7IPbqyzq_vk18oCvcniVcA4byBXCE/pub - for csv reader and writer in Problem Set 2-5
http://docs.python.org/2/library/datetime.html#datetime.datetime.strptime - for Problem Set 2-11
http://statsmodels.sourceforge.net/devel/ - to use statsmodels.formula.api in Problem Set 3-8
https://pypi.python.org/pypi/ggplot/ - for Problem Set 4-1

Section 1. Statistical Test

1.1 Statistical test used to analyze the NYC subway data?
The Mann-Whitney U statistic test is used to analyze the NYC subway data.

P value
A two-tail P value is used: the alternative hypothesis asserts that the number of entries with rain and the number of entries without rain are **different**.

Null hypothesis
The null hypothesis asserts that the population of entries per hour on rainy days and the population of entries on non-rainy days are identical.

p-critical value
The p-critical value is 0.05.

1.2 Mann-Whitney statistical test applicable to the dataset
The histograms shows that the entries data is **not** normally distributed for both samples. The Mann-Whitney U-test has greater efficiency than the t-test on non-normal distributions and all assumptions about the two samples (independance, dependent variable) are satisfied.

1.3 Results from this statistical test
p-value = 0.0249x2 = 0.0498
Mean(rainy days) = 1105
Mean(non rainy days) = 1090

Mann-Whitney U-Test = 1924409167

1.4 Significance and interpretation of these results
With p-value = 0.0498 < 0.05 that is low, then we reject the null hypothesis. Therefore the population entries per hour on rainy days and the population entries per hour on non-rainy days are significantly different.

Section 2. Linear Regression

2.1 Approach used to compute the coefficients theta and produce prediction for ENTRIESn_hourly in the regression model:
OLS using Statsmodels is used to compute the coefficients theta and produce prediction for ENTRIESn_hourly in the regression model.

2.2 Features and dummy variables used in the model
The following features are used: 'rain', 'meantempi', 'meanpressurei', 'meanwindspdi', 'fog'

Dummy variables as part of the features
Dummy data is generated from the UNIT and Hour features.

2.3 Specific reasons of selecting these features in the model
The reasons of selecting these features are based on intuition. I think that the subway ridership depends on weather means ('meantempi', 'meanpressurei', 'meanwindspdi') and foggy or not foggy days and rainy or not rainy days. I also decided to use dummy data generated from UNIT and Hour because I thought the ridership also depends on the UNIT and the Hour of the Day.

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

`rain`	`-19.532831`
`meantempi`	`-9.928900`
`meanpressurei`	`-123.315278`
`meanwindspdi`	`28.658395`
`fog`	`179.919894`

2.5 The model's R^2 (coefficients of determination) value

The model's R^2 value is: 52%

2.6 Meaning of this R^2 value for the goodness of fit for the regression model.
An R-square of 52% is meaning that the variability of the hourly entries (ENTRIESn_hourly) values around the regression line is 1-0.52=48% times the original variance; in other words we have explained 52% of the original variability, and are left with 48% residual variability.

Is this linear model appropriate to predict ridership for this dataset, given this R^2 value?

This linear model is not appropriate to predict ridership for this dataset.

The histogram of the residuals (Figure 1) has long tails. And the **probability plot of residuals (Figure 2)** shows long tails as well. It is meaning that we are seeing more variance of the residuals than we would expect in a normal distribution. The linear model might not be appropriate to predict ridership for this dataset.

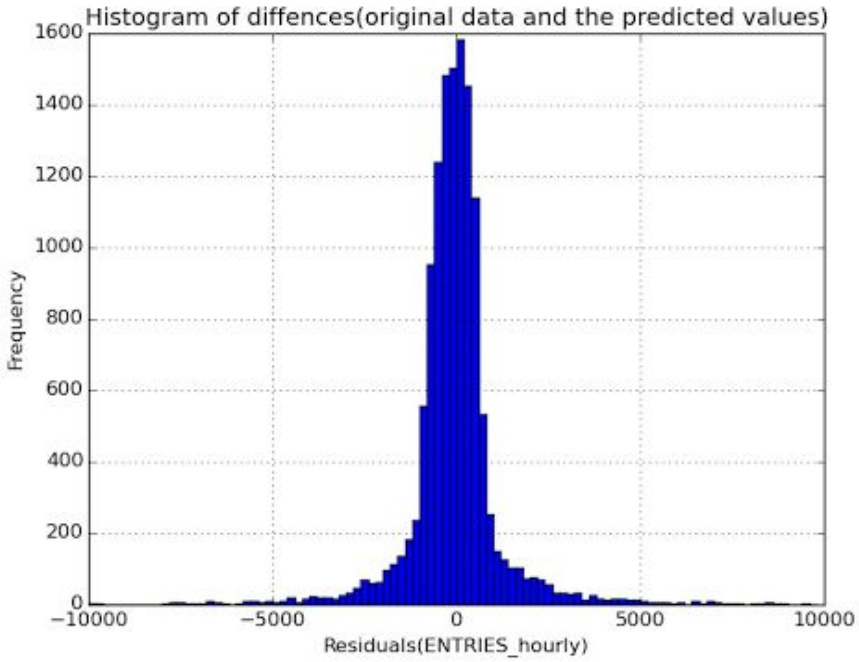

Figure 1: Histogram of the difference between the original hourly entry data and the predicted values.
Description: The histogram of the residuals (difference between the original hourly entry data and the predicted values) shows long tails.

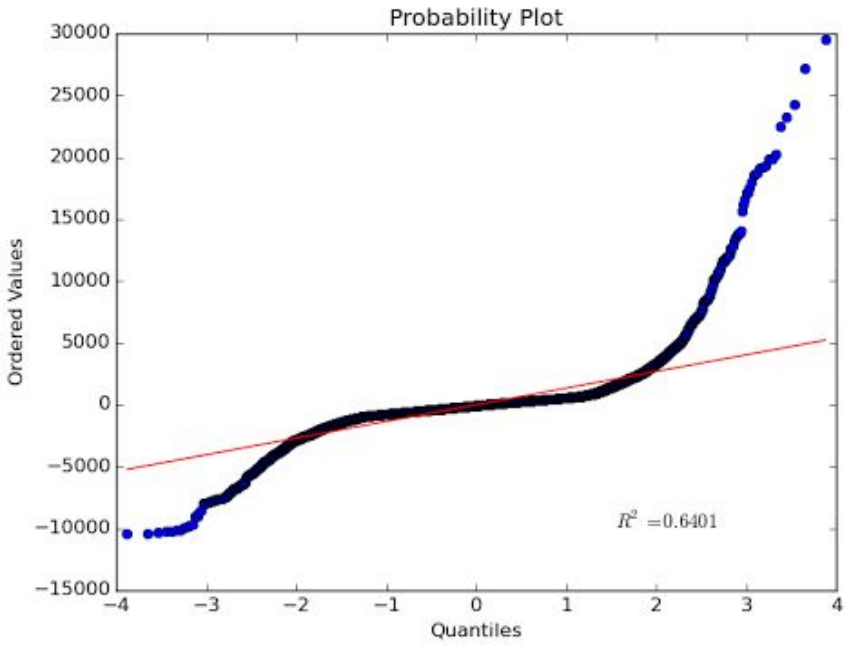

Figure 2: probability plot of residuals against the quantiles of the normal distribution.

Description: Figure 2 shows a curve which starts below the normal line, bends to follow it, and ends above it, indicating long tails.

Section 3. Visualization

Two visualizations that show the relationships between two or more variables in the NYC subway data.

3.1 Two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days. The two histograms are combined in a single plot.

For the histograms, we 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.

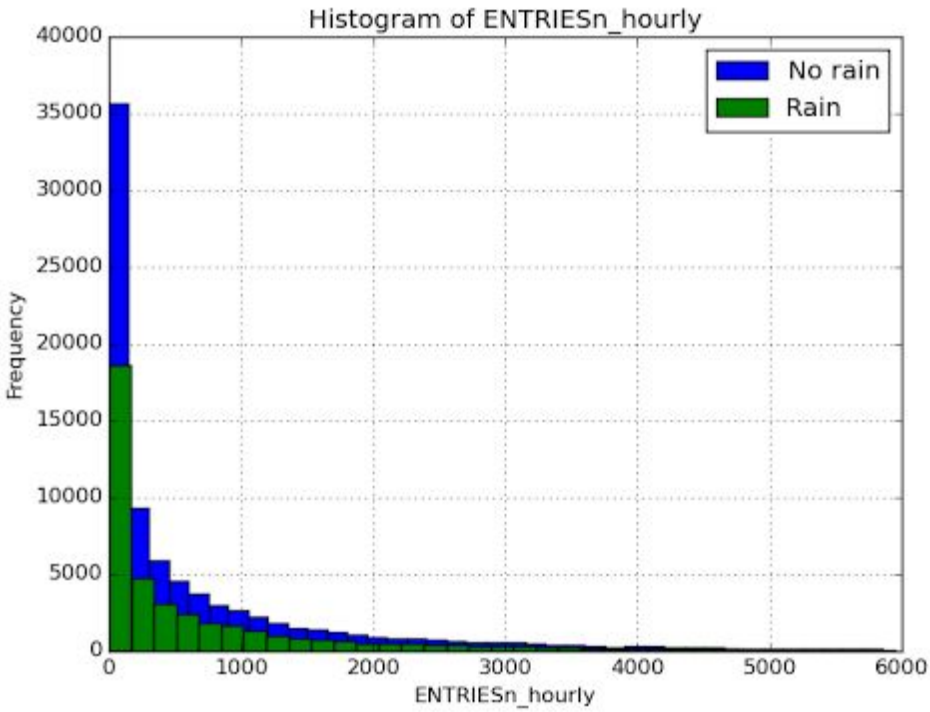

Figure 3: Histograms of ENTRIESn_hourly for rainy days (green) and non-rainy days (blue)
Description: Figure 3 plots two histograms on the same axes to show hourly entries in NYC subway when raining (green histogram) vs. when not raining (blue histogram) for the month of May 2011. The hourly entries in the NYC subway data is **not normally distributed** for both samples, rainy days and non rainy days, but have the same shape.

3.2 One visualization can be more freeform: Ridership by time-of-day

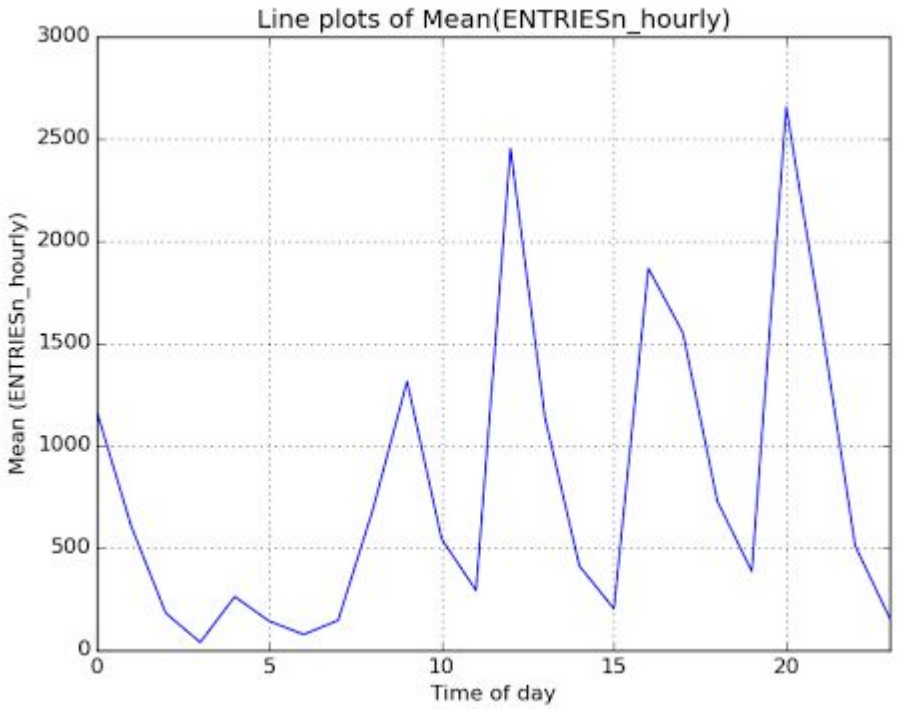

Figure 4: Line plots of ENTRIESn_hourly mean per time of day.
Description: Figure 4 plots line plots of the means of hourly entries in the NYC subway for the month of May 2011. The peak hours are 12 PM and 8 PM. And the ridership is at its minimum between 12 AM and 7 AM.

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?
We can expect more people on average riding the subway on rainy days compared to non rainy days. But it is hard to linearly predict the ridership based on the weather features.

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.
Based on the Mann Whitney U Test results, the difference of 15 in averages of ridership on rainy days and non-rainy days is statistically significant. That means we can expect higher ridership on rainy days. But with a R-square of 50% (using independent weather variables) for an OLS regression model, the ridership prediction errors is high. So it is hard to linearly predict the ridership based on the weather features.

Section 5. Reflection

5.1 Potential shortcomings of the methods of the analysis:

Time span issue: the dataset is from the month of May only. This month may have less rainy days, which is the case in the dataset. It would have been good to have data from some other months with

more rainy days. Maybe most people are having a monthly plan to ride the subway or not. And for more rainy months, they will ride more the subway. This dataset assumes that the decision to ride the subway or not is based on a daily plan. Based on experience, people know the months of year with bad weather, and are buying monthly subway tickets which may be less expensive than daily tickets.

Linear model issue: the residuals follow a cyclical pattern (Figure 5). That proves some non-linearity in the data. We should consider a nonlinear model like a polynomial model.

Multicollinearity issue: weather variables provided in the dataset to use as independent variables (min, max, mean) for linear regression might be correlated.

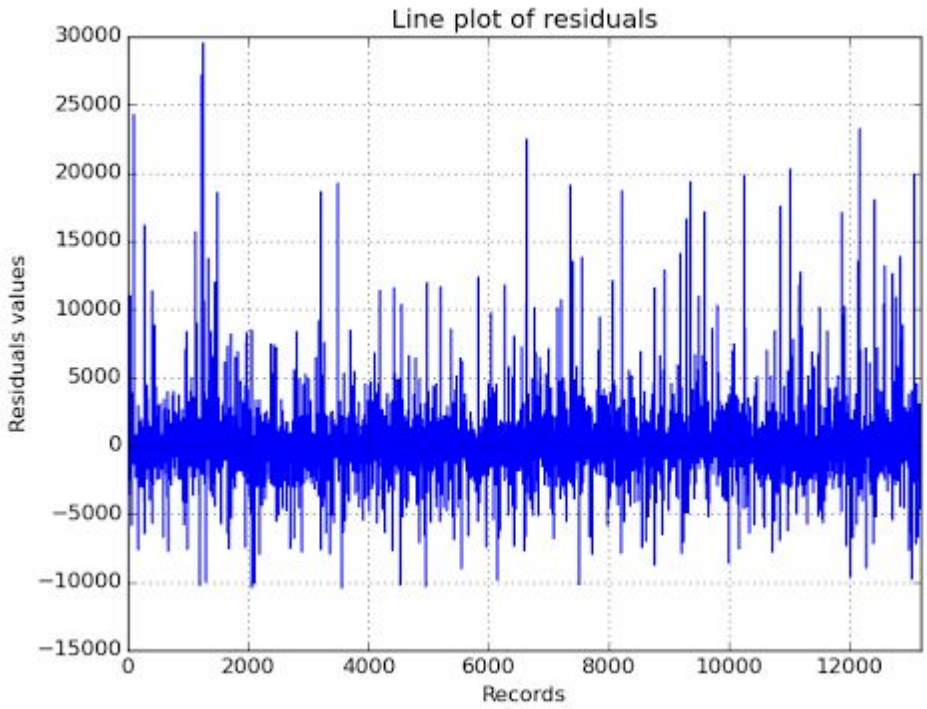

Figure 5: Line plot of residuals
Description: The residuals follow a cyclical pattern.

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