Updated sections compared to submission v01: Sections 3.1, 2.6 , 5.2

Updated sections compared to submission v02: Sections 2.4, 4.2, 1.2

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.

For problem sets:

2.10

- http://pandas.pydata.org/pandas-docs/dev/generated/pandas.tseries.tools.to_datetime.html
- http://pandas.pydata.org/pandas-docs/dev/generated/pandas.DatetimeIndex.hour.html

2.11:

- https://docs.python.org/2/library/datetime.html#strftime-strptime-behavior

3.1:

- http://matplotlib.org/users/pyplot_tutorial.html

3.2:

- Source that Welch's-t-test requires a standard distribution as a requirement: http://en.wikipedia.org/wiki/Welch%27s_t_test#Assumptions

For the project:

1.2

- http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test – Requirements for Mann-Whitney-U Test

2:

- http://www.statsoft.com/Textbook/Multiple-Regression
- 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?
The Mann-Whitney U-Test is used to analyze the NYC data

Did you use a one-tail or a two-tail P value?
Two-tail P value

What is the null hypothesis?
H_0: P(x>y) ≤ 0.5, i.e. the probability that a value from the x-distribution (distribution when it is raining) is bigger than a value from the y-distribution (distribution when it is not raining) is less or equal than 50%; i.e. the probability that more people are taking the subway when it is raining not higher than 50%

What is your p-critical value?
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 Mann-Whitney U-Test is applicable, as the single values from both groups (i.e. number of riders) can be interpreted ordinal. The Mann-Whitney U-Test does not require both sets to be equally distributed. The independence of both sets is given as long as one compares values from different days.
In case of different weather conditions of rain/no rain at different stations (e.g. as given on May 19) one could argue that the values are not completely independent as the entries at one station with rain could influence the entries at another station at another point in time. This can best be described via an example: Assuming the expected value of the number of ridership doubles in case it rains and people take the subway in the morning from a station where it rains on that day. Let's say that 100 (instead of 50) are going in the morning from turnstile unit A (where it rains on that day) to turnstile unit B (where it does not rain on that whole day). These people want to go back in the evening from B to A. This means, that the expected number of people going back from B to A increases due to the fact that it rained in A. Thus the number of riders in station B in the evening depends on the number of riders in the morning in station A which depends on the weather condition of station A in the morning. Thus the number of riders of A and B are not independent, and thus the independence of both sets is not given.

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.
P-Value: 0.04999982558
Mean when it is raining: 1105.4463767458733
Mean when it is not raining: 1090.278780151855
Median when it is raining: 282

Median when it is not raining: 278

1.4 What is the significance and interpretation of these results?

With chosen p-critical value 0.05, the result is significant, meaning that with a probability of 95%, no type 1 error is made. Thus, with probability of 95%, the test result is not significant by chance, but ridership increases really significantly when it is raining.
Even though the result is barely significant with chosen p-value, the mean and median values support the result.

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?*~~

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

- *Features: Hour, rain, mintempi*
- *Dummy variables: UNIT*

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.

- *Mintempi:*
  - *I assume that less people take the subway when it is warm*
  - *Including this as a feature, the R^2 value increased; it increased for mintempi most out of all temperature values*
- *Hour:*
  - *For sure, it depends on the time how many people are taking the subway; e.g. during night, there are less people taking the subways than in the morning when everyone is going to work*
  - *The R^2 value increases, when it is included*
- *Rain:*
  - *It is included, based on intuition but also because it was tested positive in the Mann-Whitney-U-Test*
  - *It actually does not improve the R2 value by much*
- *UNIT as dummy variable:*
  - *It is relevant which station is chosen to predict the entries; there are busier stations in the city center and less busier in the sub-areas*

- *The R^2 value doubles, when including UNIT as dummy variable*

2.4 What are the coefficients (or weights) of the non-dummy features in your linear regression model?
θ_{rain} = 43.46
θ_{Hour} =65.33
$\theta_{meantempi}$ = -10.71

2.5 What is your model's R^2 (coefficients of determination) value?
0.479230519494

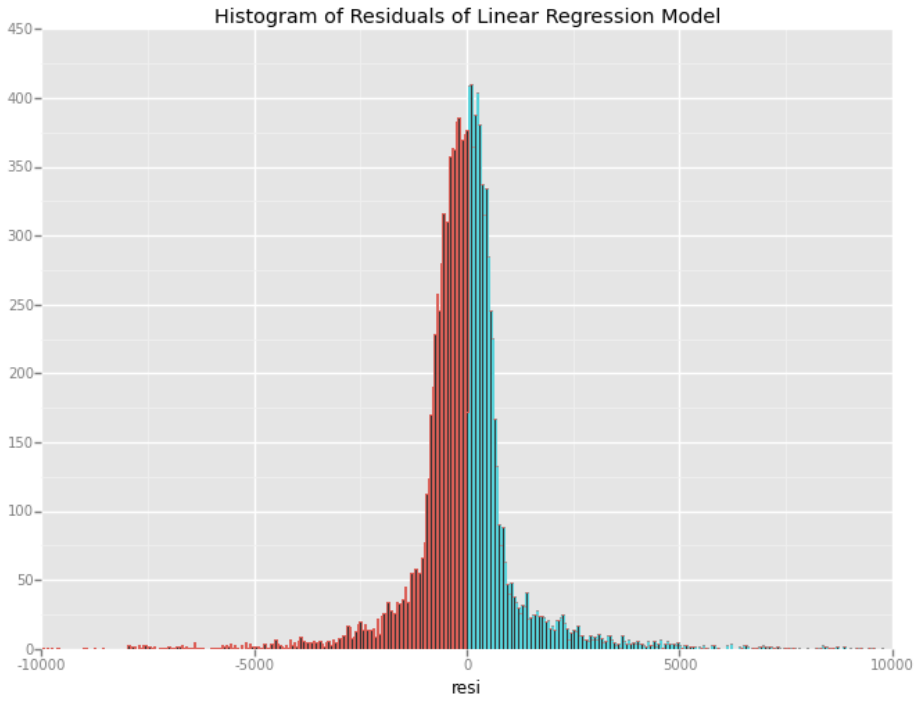
2.6 What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?
As R^2 lies between 0 (no fitting at all) and 1 (fitting to some degree), this model's R^2 indicates that the model helps to predict the ridership, but it is by far not brilliant. R^2 is 1-(ratio of residual variability) as stated at statsoft.com.[1] This means that a high R^2 value says that one has a high variability of the residuals (residual = real value – predicted value). With our resulting R^2 value, we therefore have a "medium" variability of residuals.

A very good explanation of R^2 is also given by Jim Frost in his blog at minitab.com[2]: "R-squared = Explained variation / Total variation". This actually means, that in our case, almost 50% of the variation can be explained by the model, while the other 50% cannot.

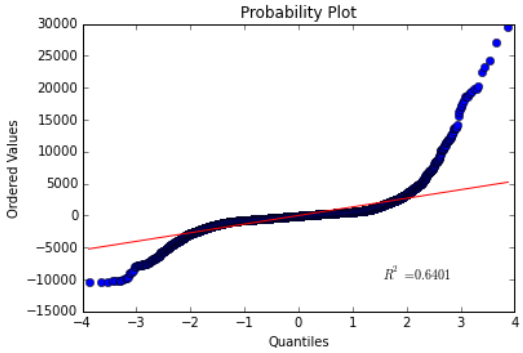
The below plot, displaying the histogram of residuals of our linear regression model, also visualizes that for quite some cases, the predicted values have been way off the real values – in both directions. The blue bars are those cases where the values have been greater than the predictions, for the red ones it is the other way around.

*From looking at the graph, one can have the impression that the model is in average underestimating the real values. The mean of the residuals seems to disprove this (i.e. mean(Residuals) = 2.88 * e^{-11}), the median (median(Residuals) = -60.38) reinforces this assumption.*

[1] http://www.statsoft.com/Textbook/Multiple-Regression as on 2015-06-14
[2] http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit as on 2015-06-14



Having a look, if the residuals are normal distributed, as it is the case for good linear model as mentioned by statsoft.com, below is a Q-Q-Plot included. While the red line displays the expected values with regards to the order of the items and the quantiles of the normal distribution, the blue dots show the real values. This plot reinforces the statement from above that there are more items with large residuals than one would expect, i.e. the tails of the distribution are longer than expected. This might prove, that there is some non-linearity in the data, which cannot be modeled by a linear regression model, but requires a non linear model.



Based on the findings as described above (long tail of the distribution, "medium" good R^2 value), the linear regression model does not to seem an ideal model and probably leaves some important points out, but it already explains some of the variability of the values.

Section 3. Visualization

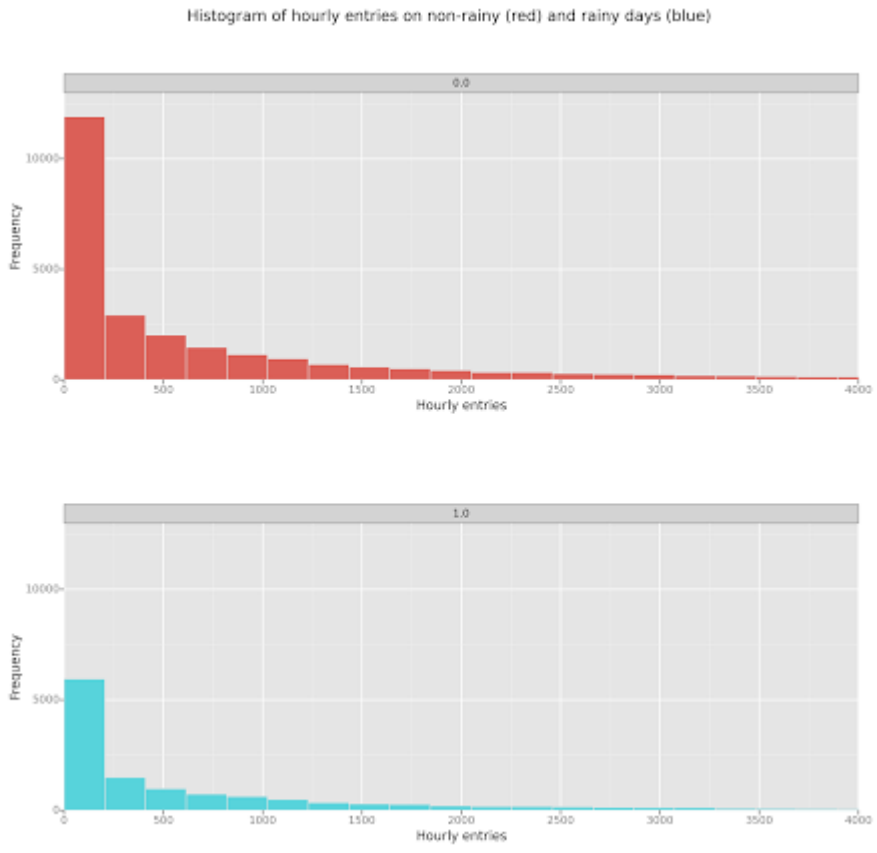
Please include two visualizations that show the relationships between two or more variables in the NYC subway data.

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.

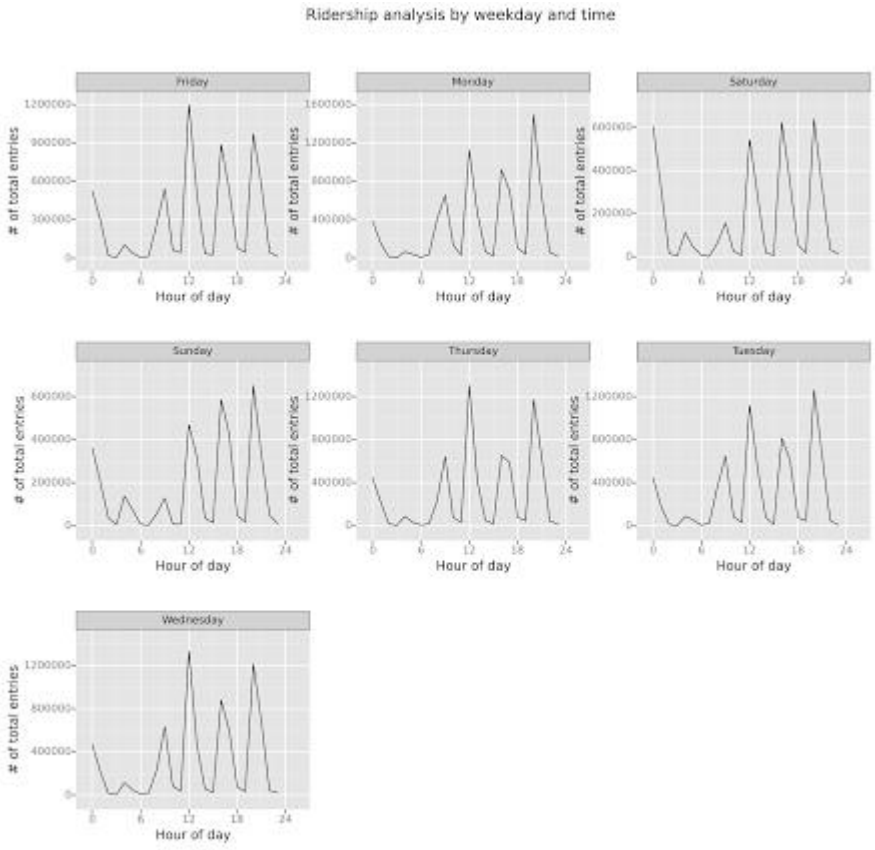


Key insights:

- *The two data sets seem to have the same distribution*

- *The most common case is that there are less than 200 entries per unit per hour which holds true when it is raining as well as when it is not raining*
- *More general speaking, the frequency of higher entries per hour is lower than the frequency of few 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. Some suggestions are:



Key insights:

- *On Saturday's and Sunday's, there is a small peak at between 0 and 4 am as some people probably come home from a party*
- *One should be careful viewing the visualization as it indicates that there are 6 peaks per days which only seems like this as the data is only provided for 6 points in time throughout the day*
- *During all weekdays the main part of riders are between 8 am and noon and 4 pm and 8 pm which exactly represent the rush hours of people going to work and coming back from work*
- *During the weekend the highest numbers are for the two time periods between 12 pm and 4 pm and 4 pm and 8 pm which represent a shift in rush hours*
- *Ridership on the weekend is also in general lower than on weekdays*

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 statistical test used in section 1, more people seem to take the subway when it is raining. On the other hand, the linear regression model as used in model 2 only improves a little bit, when including rain as a parameter. Based on these two main indicators, I tend to say that the ridership increases slightly when it is raining.

This would also follow an intuitive hypothesis that more people are taking the subway when the it is raining, as they do not want to walk. On the other hand, it could be the case that – if they have a car – people could tend to drive to omit any time outside in the rain. Or it could even be the case that people try not to go outside in case they do not have to go, because they do not want to become wet.

Combining both, the analytical part and the intuitive argumentation including its counterarguments, I feel confident to say that it seems that more people ride the NYC subway when it is raining and it is even significant but the difference is not enormously.

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 statistical test results in the assumption that there are more riders when it is raining not only as the pure significance of the result but also due to the Medians and the Means of the two data sets. Both are higher, when it is raining.

The linear regression model only improves a little bit and only from an improvement, one cannot say if more people are riding the subway on rainy than on non-rainy days. But the theta for rain is positive which indicates that on days when it is raining the model predicts more riders.

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,
   The dataset only contains information from one month and one could check if there were

other reasons for peaks or drops in volumes, like events at specific places, e.g. concerts or sport events. One could also have a look for public holidays during that time period which results in different volumes during different times. This information is not available from the dataset but it could be researched from the internet.
If there would be longer time periods available, one could compare rainy weekdays with non-rainy weekdays to have a better indicator and repeat this for the weekends.

2. Analysis, such as the linear regression model or statistical test.
   The statistical test does not take into account effects as described in 5.1.1. Thus, the significance could be by chance as there could be other circumstances which lead to slightly higher volumes when it was raining. It was also not tested if the time when it was raining coincided with specific times when the traffic was in general higher. As also pointed out in 1.2, one could argue that the two data sets are not completely independent which is a requirement for the Mann-Whitney U-Test.

   As described in 2.6, the linear regression model seems not to be the best model for the given data. Especially the long-tail of the distribution of the residuals is either an indication that a linear regression model is not the best answer to the question of predicting ridership or it could also be the case that there are peaks in reality, which are hard to predict.

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

None