Analyzing the NYC Subway Dataset

ziheXu - January 2, 2015

1.Statistical Test

1.1 Which statistical test did you use to analyse the NYC subway data? Did you use a one-tail or a two-tail P value? What is the null hypothesis?

- We use the Mann-Whitney's U test to analyze the NYC subway data.

- I use **two tail P value**

- The null hypothesis is that the weather is not associate with ridership.

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.

- Because by observing the histogram, we find out the dataset we get doesn't fit normal distribution, and they do not have the same sample size and might not have the same variance. Therefore, we perform the Mann-Whitney's U test which can be used with such condition.

1.3 What results did you get from this statistical test? These should include the following numerical values: pvalues, as well as the means for each of the two samples under test.

- The p-values we have is: 0.024999912793489721
- the means for the number of entires with rain is: 1105.4463767458733
- the means for the number of entires without rain is: 1090.278780151855

1.4 What is the significance and interpretation of these results?

- **This p-value we found is less than 0.05, it means that the result is statistically significant and the null hypothesis would be rejected. In other word, it means the means for the entires with rain and not rain is not equal.**

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:

a. Gradient descent (as implemented in exercise 3.5)

b. OLS using Statsmodels

c. Or something different?

- I used the Gradient descent.

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

- I use rain, precipi, Hour, meantempi and the **day in a week** in my model, Yes, I did use dummy variable.

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 test these feature individually and then test several combinations, if adding any features will increase my R2 value by a large amount I will keep it. For example, by adding the meantempi features. it increased my R2 value by almost 0.1 after I including it.
- **Also from the second graph I generated about the different entries between different day in a week, we can intuitively consider adding day_in_way as a useful feature(as a result, I boost my R2 to 0.474350 (with alpha = 0.3 and iteration =75)**
- **I also test different alpha values, with same iteration = 5**
  - **if I use alpha = 0.9, I got an enormous number, the graph shows it oscillating a lot.**
  - **if I use alpha = 0.7 I got R2 = 0.466117828111**
  - **if I use alpha = 0.3 I got R2 = 0.453035414835**
  -
- **I test different alpha values, with same iteration = 75**
  - **if I use alpha = 0.7 I got R2 = 0.47435027903**
  - **if I use alpha = 0.3 I got R2 = 0.47435027903**
  -
  - **It seems like 0.7 is larger than 0.3, when number of iteration = 5, however as the number of iterations goes up, the R2 for 0.3 gets better because a small alpha means more fine adjustment, and a larger alpha means a fast adjustment. So with a small value of alpha and a large number of iteration you can get a better value. You can simply observe the graph from the each iteration to determinate how many steps you want to take.**

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

- My model’s R2 equals to **0.47435027903**

2.5 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?

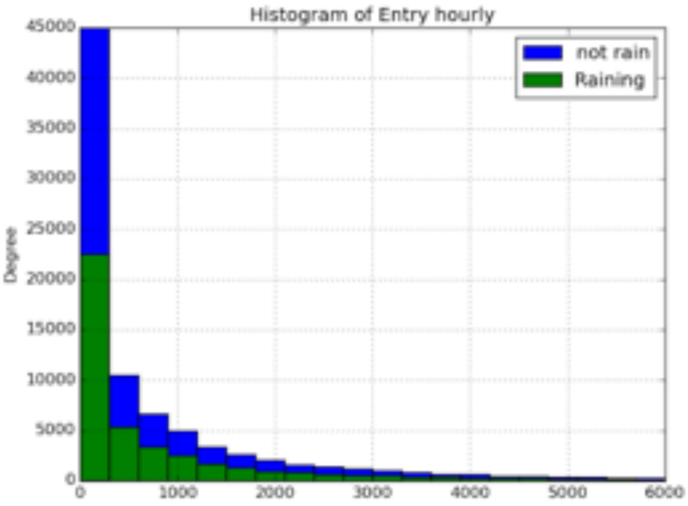
- The larger the R2 value is, the better the regression model fits actual data.
- From the R2 value I get, I think the linear model is a fair model to explain the dataset, because although we can have a decent fit for our data set, if it’s a bad fit, the R2 will close to 0. But it doesn’t get much improvement by adjusting number of iterations and other parameters. I think there might have a better model base on the results we obtained.

3.Visualization

Please include two visualizations that show the relationships between two or more variables in the NYC subway data. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots, or histograms) or attempt to implement something more advanced if you'd like. 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.

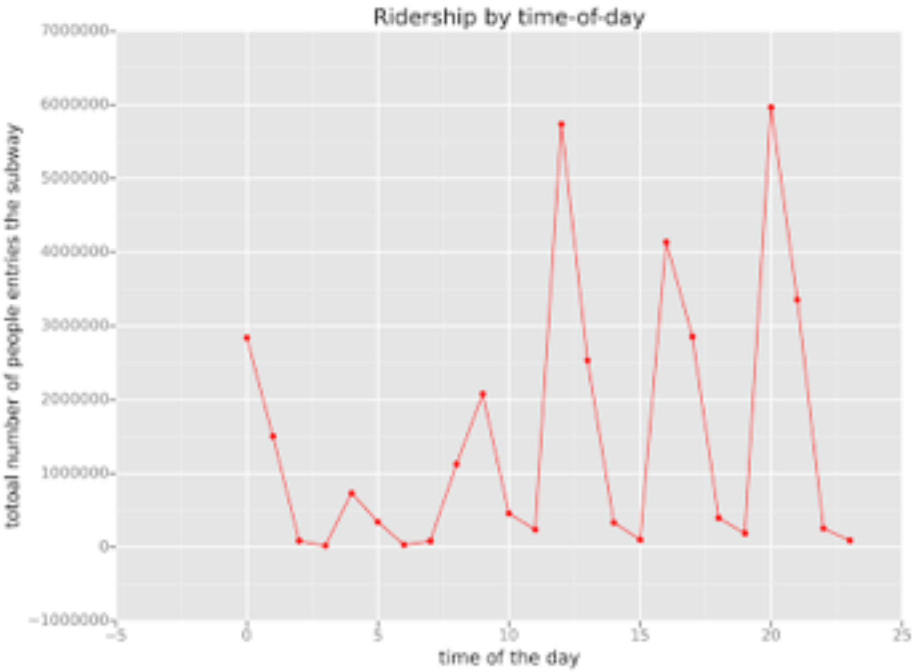
- You can combine the two histograms in a single plot or you can use two separate plots.
- If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.
- For the histograms, 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, you might have one interval (along the x-axis) with values from 0 to 1000. The height of the bar for this interval will then represent the number of records (rows in our data) that have ENTRIESn_hourly that fall into 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.



- From this graph, we can see both data sets don’t fit the normal distribution, they have the similar shape, overall, it seems like there is more people riding subway on days that not raining, however, it is because there is less raining days then the not raining days, as a result, the total number of the entry in not rain days is larger than the that of the non raining days. Therefore, we should perform a statistical test to see which mean value is statistically larger than another, or maybe they are the same.

3.2 One visualization can be more freeform. Some suggestions are:

- Ridership by time-of-day or day-of-week
- Which stations have more exits or entries at different times of day



- As the figure shows, there is more people taking the subway around 11 -12am and 7 - 8 pm. These two peaks are reasonable cause people will finish there morning around 12 and afternoon around 7. To my surprise, there is a peak at 3 - 4 pm, I am not sure why, maybe some office jobs will ended at 3.

4. Conclusion

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

- **Yes, based on the p-value, which is 5%, we reject the null hypothesis that ridership and weather are not related, as a result we find the weather and the ridership is related, from the data, we calculate the mean for rain and not rain. And the mean value shows that more people ride subway when it raining. Combine those two we can make the conclusion .**

4.2 What analyses lead you to this conclusion?

- **To make this conclusion, I used the Mann-Whitney’s U test to analyze data. We made the assumption that the ridership and the weather are not related. The p value we get is**

0.024999912793489721, which means under the assumption we make, there is a near 5% probability of obtaining a test statistic at least as extreme as ours if null hypothesis is true. As a result, we will reject the null hypothesis .

5.Reflection

5.1 Please discuss potential shortcomings of the data set and the methods of your analysis.

Possible shortcomings of the data set:

- We may not have sufficient data, we only have the date on particular month.
- We may have some false data either result of human error or malfunction of machines.
- We may have some missing data

Possible shortcomings of the analysis

- We may use a more sophisticate model than the simple linear model
- We may consider more features and do a better feature selection
- We may use a different way to compute the coefficient other than the gradient descent.

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

- I found it is interesting that there are many people use the subway around 3-4pm, I am guessing there are many jobs that will finish at 3-4.