ANALYZING THE NYC SUBWAY DATASET

Done by,
Sheethal Mohan

SHORT ANSWERS

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?**

We have used Mann-Whitney U test to analyze the NYC subway data. MWU test was used to measure the difference between the entries of rainy and non-rainy days. We have chosen MWU test since the distribution of entries is right skewed.

We have used two-tail p value.

Null hypothesis is a statement that we are trying to disprove by running our test. Here, Null hypothesis is that there is no difference in ridership on rainy days vs. non-rainy days.

P-critical value = 0.05

P-value from MWU test = 0.024999912793489721 = 0.0249(approx.)
To get two tailed p value,
P-value = 2*0.0249 = 0.0498

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**

Mann Whitney U-test can be used for normal and non-normal distribution. MWU test is a non-parametric test. It does not assume any particular distribution. We have to analyze two populations, i.e., the ridership on rainy days and the ridership of non-rainy days. These two populations have different sample sizes. We know that when we perform Welch's t-test, we have to assume that the data is drawn from a normal population and that the two samples must be independent.

Here, we have used MWU test to determine if there was a statistically significant difference between the number of entries on rainy and non-rainy days. Both rainy and non-rainy data sets are non-normally distributed (from the histogram we can see that data is non-normal-right skewed). On using the Shapiro-Wilks test to check if the population came from normally distributed population normality, it was against the null hypothesis that the populations were normally distributed. Since the

assumptions of Welch's t-test have failed, I have used MWU test as it has an advantage which allow us to use it for both normal and non-normal distributions.

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.**

Null hypothesis:
Ridership on rainy days and non-rainy days are same.

Alternate hypothesis:
Ridership on rainy days and non-rainy days are not same

The results are:

P-value = 0.0498
U-value = 1924409167.0

The mean on rainy days = 1105.4463767458733
The mean on non-rainy days = 1090.278780151855

Both the datasets have different mean. So, the NULL hypothesis is rejected because our p value is less than the p-critical value.

1.4 **What is the significance and interpretation of these results?**

From the mean values of both rainy and non-rainy days we understand that more people use the NYC Subway on rainy days as compared to non-rainy days.

We understood that the probability of obtaining a test statistic at least as extreme as ours if NULL hypothesis was true is around 0.025.

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:**

a. **Gradient descent (as implemented in exercise 3.5)**
b. **OLS using Statsmodels**
c. **Or something different?**

I have used gradient descent in exercise 3.5 and OLS in exercise 3.8, to compute the coefficients theta.

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

Yes, I have used dummy variables in gradient descent.

Gradient descent:

Features: ‘rain’, ‘precipi’, ‘Hour’
and ‘meantempi’
Dummy variable: ‘UNIT’

OLS Model:

Features: ‘EXITSn_hourly’, ‘Hour’, ‘maxpressurei’, ‘maxdewpti’, ‘mindewpti’, ‘minpressurei’, ‘meandewpti’, ‘meanpressurei’, ‘fog’, ‘rain’, ‘meanwindspdi’, ‘mintempi’, ‘maxtempi’, ‘precipi’, ‘thunder’

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 have selected these features based on my intuition that these features have more chances in effecting the Subway ridership. These features can be classified into three categories:

1. Time per day (‘Hour’)
2. Transit ridership per hour in a day (‘EXITSn_hourly’)
3. Weather conditions (‘percipi’, ‘meantempi’, ‘meanwindspdi’, ‘rain’, ‘fog’, ‘meanpressurei’). These features can take up numeric value.

Through experimentation, we saw that by using only few features (like ‘rain’, ‘precipi’, ‘Hour’ and ‘meantempi’) in gradient descent method, we got an R^2 value of 0.46(approx.).
But when we used OLS method, we saw that the R^2 value increased to 0.55(approx.). Thus we can say that using all these features in fact helps us increasing the predictive power of the model.

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

For gradient descent the features were ‘rain’, ‘percipi’, ‘Hour’, ‘meantempi’ and their weights are 5.346, 21.656, 420.881 and -52.427 respectively.

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

Gradient descent:

$R^2 = 0.463968815042$

OLS:

$R^2 = 0.554803842649$

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?**

R^2 = .46

R^2 = 1-residual variability

= 1-.46

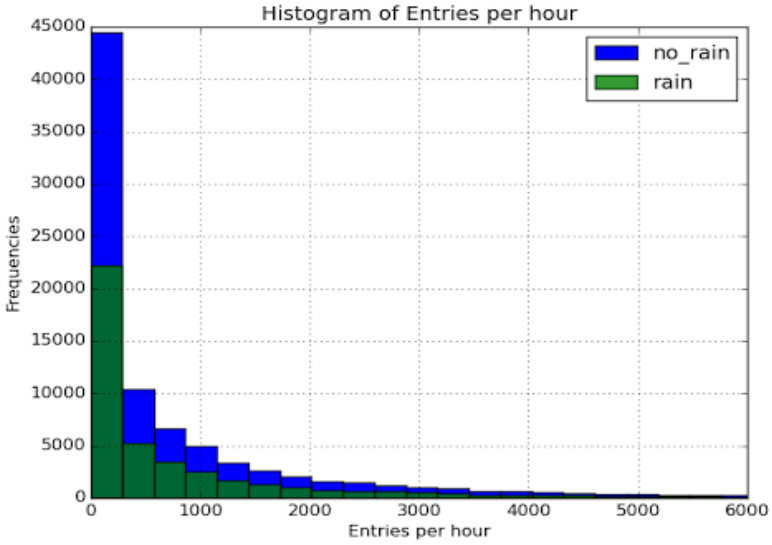
= .54

Since there is more residual variability (46%) than original variability (54%), predictions from linear model is not good.

Section-3: Visualisation

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

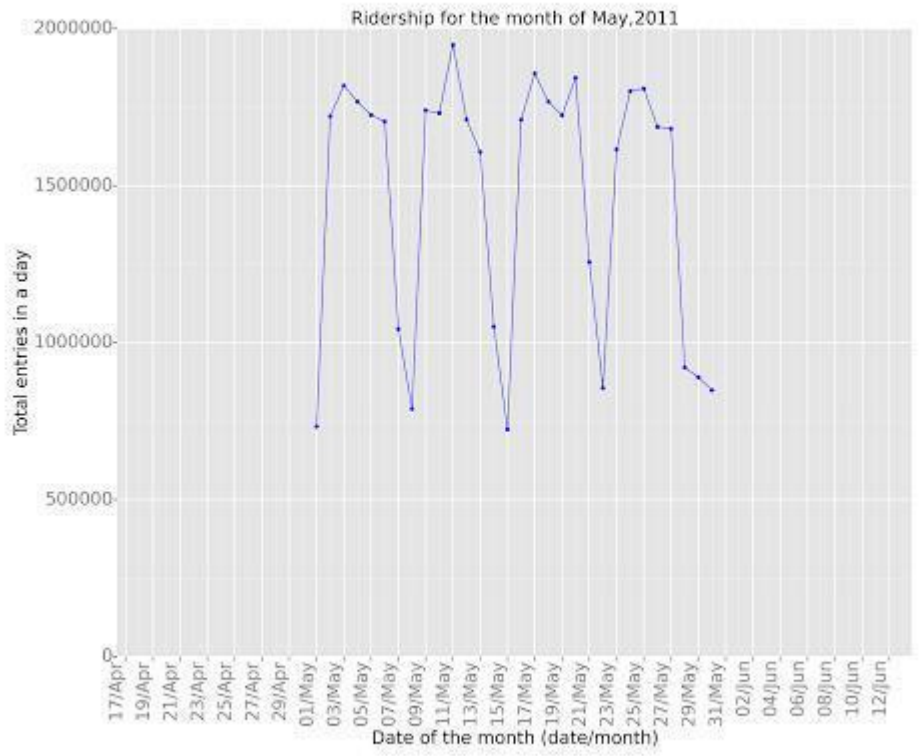
- **You can combine the two histograms in a single plot or you can use two separate plots.**
- **If you decide to use two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.**
- **For the histograms, 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**



It is quite evident from the above histogram that the shapes of the two samples are similar and there are fewer rainy days than non-rainy days. Another point to be noted is that this histogram is not for the entire axis. In fact, the actual tail goes beyond 40k+ on the x-axis.

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:

- **Ridership by time-of-day**
- **Ridership by day-of-week**



On joining the lines in this scatter plot, we get a better understanding of our dataset. We observe that the peak entries were observed during May 4th – May 11th, 2011 and then we observed a rapid fall in entries until around May 20th, after which there was an increase in the level of entries until the month end.

Section 4: Conclusion

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?

Null hypothesis: Ridership on rainy days and non-rainy days are same.

Alternate hypothesis: Ridership on rainy days and non-rainy days are not the same.

The mean of ENTRIESn_hourly on rainy days = 1105.4463767458733. The mean of ENTRIESn_hourly on non-rainy days = 1090.278780151855

That is, mean of ENTRIESn_hourly on rainy days is slightly greater than that of non-rainy days From MWU test, since P <P critical, we can reject the null hypothesis and conclude that ridership on rainy days and non-rainy days are significantly different. From the means of both rainy and non-rainy days, it is clear that more people prefer to ride the NYC Subway on a rainy day in May 2011.

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.

Null hypothesis: Ridership on rainy days and non-rainy days are same.
Alternate hypothesis: Ridership on rainy days and non-rainy days are not the same.

The mean of ENTRIESn_hourly on rainy days = 1105.4463767458733

The mean of ENTRIESn_hourly on non-rainy days = 1090.278780151855

That is, mean of ENTRIESn_hourly on rainy days is slightly greater than that of non-rainy days

Since p<p critical, we reject the null hypothesis and conclude that ridership on rainy days and non-rainy days are significantly different.

When we performed linear regression through gradient descent, we obtained a R^2 value of 0.461. In gradient descent method we had only used few features like 'rain', 'precipi', 'meantempi' and 'Hour'. But when we used all the features in the OLS method, our R^2 value increased to 0.54 i.e., we achieved an original variability of 54%. Thus we can come to a conclusion that when we use all the features, it helps to improve the predictive power of the model.

Section 5: Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis, including:
1. Dataset,
2. Analysis, such as the linear regression model or statistical test.

There are some shortcomings in the dataset. The dataset is limited only to a few months in

2011. We could have come to a better conclusion if the sample size was increased.

Another shortcoming is that we discussed only about rainy and non-rainy days, and how it had an impact on the ridership. For example, we could have also involved “fog” and “non-fog” days along with rainy and non-rainy days.

We are all aware that the weather conditions vary over a period of year. So according to me, I don’t feel it’s good to come to a conclusion of the ridership rates for rainy and non-rainy days by just taking into account of some of the months in a year.

References

- http://pandas.pydata.org/pandas-docs/stable/
- https://bitbucket.org/hrojas/learn-pandas
- https://pypi.python.org/pypi/ggplot/
  http://docs.scipy.org/doc/scipy/reference/stats.html
- https://docs.python.org/2/tutorial/controlflow.html#lambda-expressions
- http://www.sqlite.org/lang.html
- https://pypi.python.org/pypi/pandasql
- http://people.duke.edu/~rnau/testing.htm
- http://www.statsoft.com/Textbook/Multiple-Regression#cresidual
- http://blog.minitab.com/blog/adventures-in-statistics/how-high-should-r-squared-be-i n-regression-analysis