INTRO TO DATA SCIENCE

Project-1: Analysing the NYC Subway Dataset

By,
Vaishnav Mohanan Mavilakandy

Short Questions

Section-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? What is your p-critical value?**

Answer:

The statistical test that we use to analyse the NYC subway data is the Mann-Whitney U-test or the Mann-Whitney Wilcoxon Test. We are using a two-tailed P-value. We know the equation,

U, p=scipy.stats.mannwhitneyu(x, y)

'U' here returns the Mann-Whitney test statistic. This test returns the NULL hypothesis that the two populations are the same. NULL hypothesis is generally a statement that we are trying to disprove by running our test.

$P_{critical} = 0.05$

P-value from Mann-Whitney U test= 0.0249. This is a one-tail p-value.
Therefore, our two-tail p-value= 0.0249 * 2 =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.**

Answer:

This is because our rainy and non-rainy datasets are not normally distributed. We know that the Mann-Whitney U-test does not assume that our data is drawn from any particular underlying probability distribution. So, the best way to analyse the NYC subway dataset is using Mann-Whitney U-test.

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.**

Answer:

The p-value obtained from Mann-Whitney U-test is: 0.02499

Both the datasets don’t have the same mean. So, the NULL hypothesis is rejected. The Mann-Whitney U-test is a test of the NULL hypothesis that the two populations are the same.
The mean on rainy days is: 1105.446
The mean on non-rainy days is: 1090.278

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

Answer:

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:

1. **Gradient descent (as implemented in exercise 3.5)**
2. **OLS using Statsmodels**
3. **Or something different?**

Answer:

We have used Gradient Descent in exercise 3.5 to find out the coefficient theta and in exercise 3.8 we are supposed to use 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?

Answer:

In Gradient Descent method, the features used were ‘rain’, ‘precipi’, ‘Hour’, ‘meantempi’. Yes, we did use a dummy variable and the variable is ‘UNIT’.

In OLS Model the features used were: ‘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 R^2 value.”**

Answer:

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. But when we used OLS method, we saw that the R^2 value increased to 0.55. 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?

Answer:

The coefficient of 'rain' is 5.346
The coefficient of 'precipi' is 21.656
The coefficient of 'Hour' is 420.811
The coefficient of 'meantempi' is -52.427

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

Answer:

The R^2 value for the gradient descent method is 0.461 and for OLS model the R^2 value is 0.55.

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?

Answer:

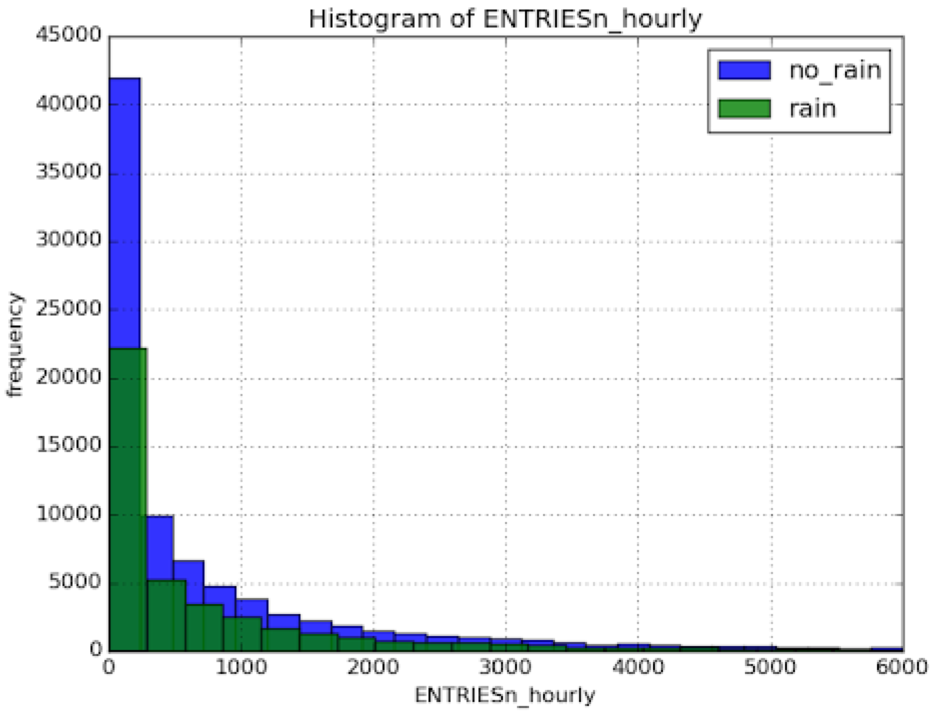
The R^2 value for the gradient descent method is only 0.461. We know that, R^2 = 1- ratio of residual variability. That is with an original variability of 46.1%, we are left with 53.1% residual variability. Since there is more of residual variability than original variability, the predictions from the regression 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.**

Answer:

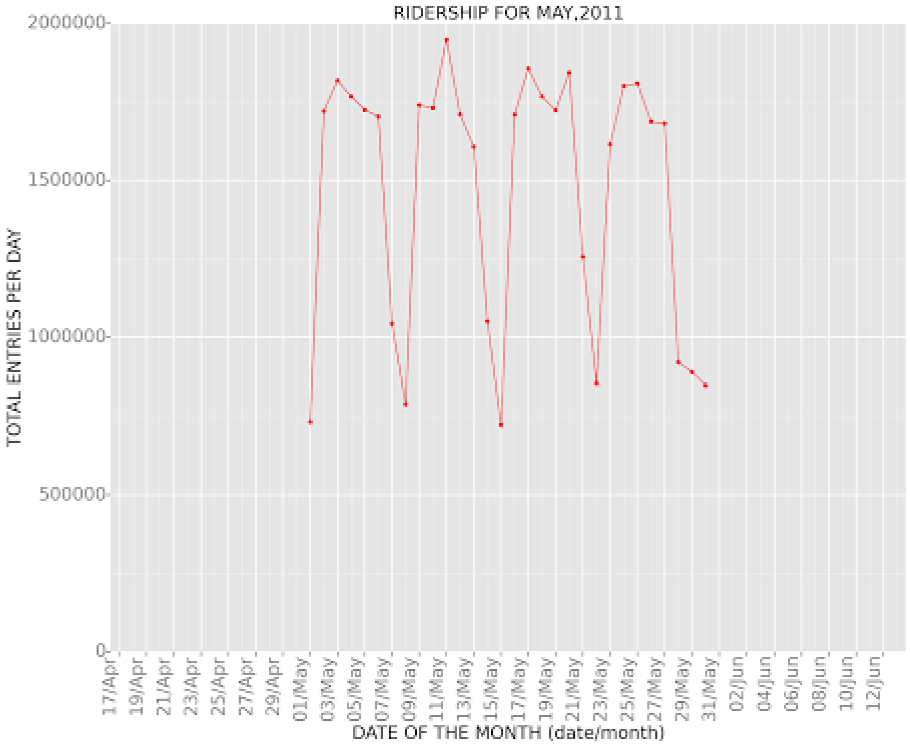


We can see from the above histogram that the shapes of the two samples are quite similar and there is certainly a difference in the sample size of both the samples (rainy and non-rainy days).

3.2) One visualisation can be more freeform. You should feel free to implement something that we discussed in class (eg., 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**

<u>Answer</u>:



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?**

Answer:

Our NULL hypothesis is that the ridership on rainy days and non-rainy days is the same.
This implies that our alternate hypothesis is that the ridership on rainy and non-rainy days isn't the same.
We found that the mean of ENTRIESn_hourly on rainy days is 1105.446 and that of non-rainy days is 1090.279. That is the mean of ENTRIESn_hourly on a rainy day is more than that of a non-rainy day.
From MWU test, we obtained the p-value as 0.0249.
$P_{critical} = 0.05$
Since $P < P_{critical}$, the NULL hypothesis can be rejected and we can conclude that the ridership on rainy and non-rainy days are significantly different and from the mean values it is clear that more people prefer to travel in 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.**

Answer:

The mean of ENTRIESn_hourly on rainy days = 1105.446
The mean of ENTRIESn_hourly on non-rainy days = 1090.279
This tells us that more people prefer to ride the NYC Subway on a rainy day.
From MWU test, p-value = 0.0249
$P_{critical} = 0.05$

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

So, to conclude, from the Mann-Whitney U-test we get to understand that there is a significant statistical difference between the two samples (rainy and non-rainy days) of the dataset. We also spotted the differences in the means calculated for both rainy and non-rainy days. All these results led us to this conclusion.

Section-5: Reflection

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

- **Dataset,**
- **Analysis, such as the linear regression model or statistical test.**

Answer:

One of the shortcomings according to me is that the sample size of the dataset is not large. The dataset is limited to only a few months in 2011. If we had a dataset with larger sample size, then we could have come to a better conclusion.

Secondly, we discussed only about rainy and non-rainy days, and how it had an impact on the ridership. This is a shortcoming. We should have also considered 'fog' and non-fog' days along with rainy and non-rainy days. We all know 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

- https://pypi.python.org/pypi/pandasql
- http://www.sqlite.org/lang.html
- http://pandas.pydata.org/pandas-docs/stable/
- https://bitbucket.org/hrojas/learn-pandas
- https://docs.python.org/2/tutorial/controlflow.html#lambda-expressions
- http://docs.scipy.org/doc/scipy/reference/stats.html
- http://www.statsoft.com/Textbook/Multiple-Regression#cresidual
- http://blog.minitab.com/blog/adventures-in-statistics/how-high-should-r-squared-be-in-regression-analysis
- http://www.bzst.com/2009/03/what-r-squared-is-and-is-not.html
- http://www.ats.ucla.edu/stat/mult_pkg/faq/general/tail_tests.htm