Udacity data Analyst Nanodegree
Project 1 – Review 1
Diego Menin
11/06/2015

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?

To analyse the data, I used a two-tailed Mann Whitney U-test. A two tailed test was chosen because I don't' know if advance which data set would be higher or lower. "Picking a one-tailed test means that we assume in advance (before we collect the data) that rain will not be associated with lower ridership, which is a very strong assumption."

The null hypothesis is that there is no difference on Subway ridership between rainy and not rainy days (or that rain doesn't affect the Subway ridership). If we are talking on a more "statistics" language, we can say that, after analysing two populations, the null hypothesis is that the two populations are the same.

The p-critical value used was 0.05, or 5%.

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 data on this dataset in not normally distributed. When the data is not normal, we have to use what's called "nonparametric tests" to compare two samples, for example, the Mann-Whitney U test used.

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.

The results were:

- *Mean for "without rain" dataset: 1105.446*
- *Mean for "rain" dataset: 1090.278*
- *U statistic: 1924409167.0*
- *p-value:* 0.0499998

1.4 What is the significance and interpretation of these results?

Since the p-value is smaller than the p-critical, we can conclude that the distribution of the number of entries is statistically different between rainy and non-rainy days. Therefore we can reject the null hypothesis.

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?

I used something different. I calculated the predictions using Linear Regression with Ordinary Least Squares. The function used was "smf.ols" from the "statsmodels.formula.api" library.

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

The features used were:

UNIT
Fog
Rain
Meanwindspdi
Maxtempi
Precipi

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.

I tried a couple of different combinations. Since the goal was to produce an r squared bugger that 0.4, I stopped when I got 0.44 but what mainly drove the feature selection was:

- *Intuition: clearly the weather affect whether a person will or will not ride the subway so "fog", "rain", and "precipi" were selected.*
- *"UNIT" was added because no matter what combinations of features I selected, I couldn't produce a R^2 bigger than 0.1 without it. I guess it does make sense because "UNIT" tells where a particular station is located and it is intuitive that, when trying to predict ridership of the subway, the biggest driver is the subway's location. A subway located in the hearth of the city will clearly have more visitors than one far away on the suburbs, regardless of weather conditions*

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

```
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
fog            199.2309     50.696      3.930      0.000        99.858   298.603
rain             1.2000     41.977      0.029      0.977       -81.081    83.481
meanwindspdi    35.7216      8.667      4.121      0.000        18.732    52.711
maxtempi        -4.9874      2.230     -2.237      0.025        -9.358    -0.616
precipi        -74.6009     49.953     -1.493      0.135      -172.517    23.315
```

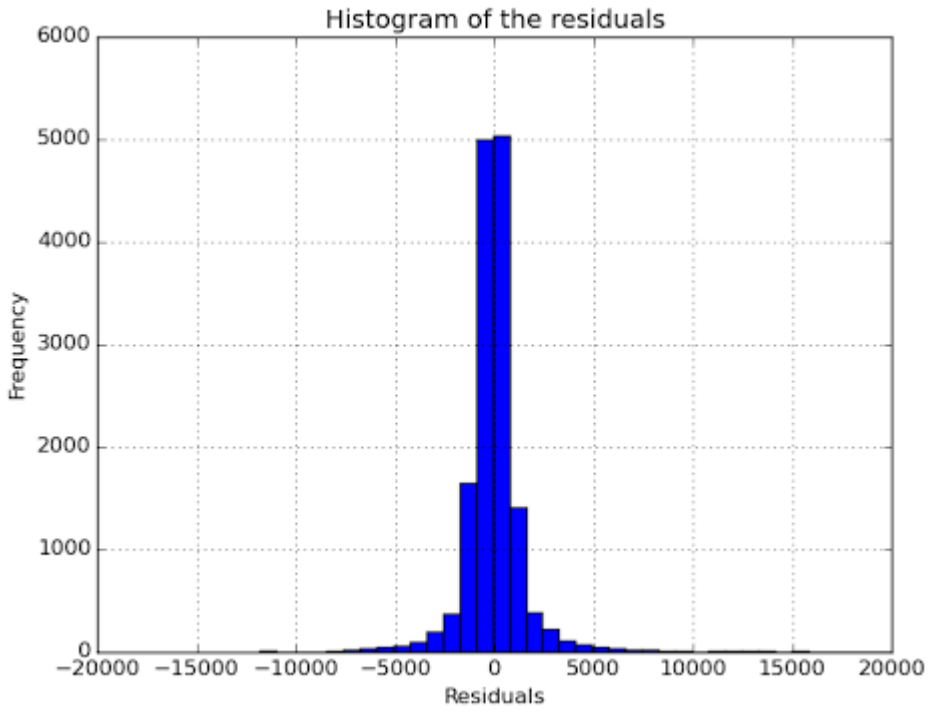
2.5 What is your model's R2 (coefficients of determination) value?

The R2 value is 0.444086721292

2.6 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 R2 is a way to evaluate the effectiveness of our model. The closer R2 is to one, the better our model. And the closer it is to zero, the poorer our model. This model, where R2 is 0.44, is a fairly poor model.

One important question to be asked is if a linear model is an appropriate model for this dataset. To assess that we can take a look at the residuals of the prediction (the difference between the predicted and the actual values)

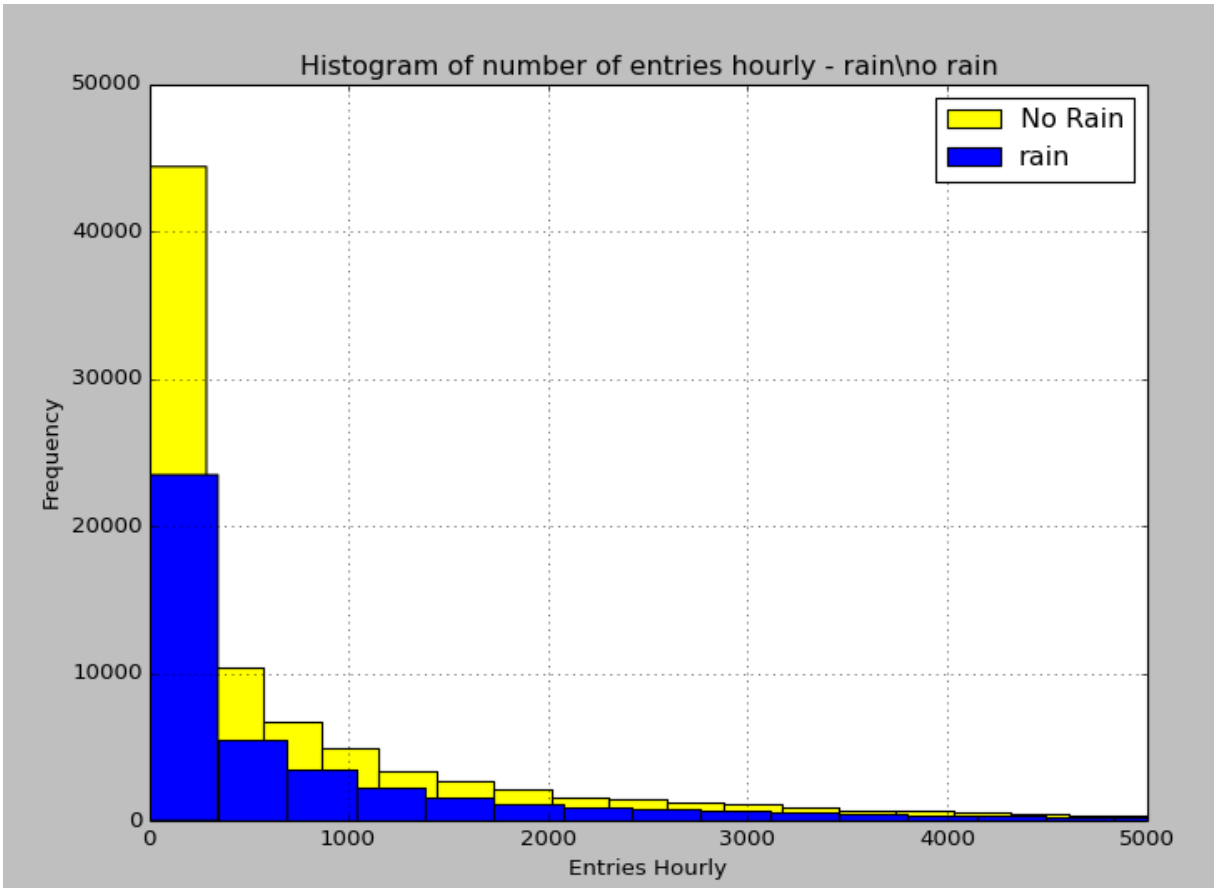


We can see that the histogram of the residuals has long tails, which suggests that there are some very large residuals, which is an indication that a linear regression model may not be the best option for this dataset.

Section 3. Visualization

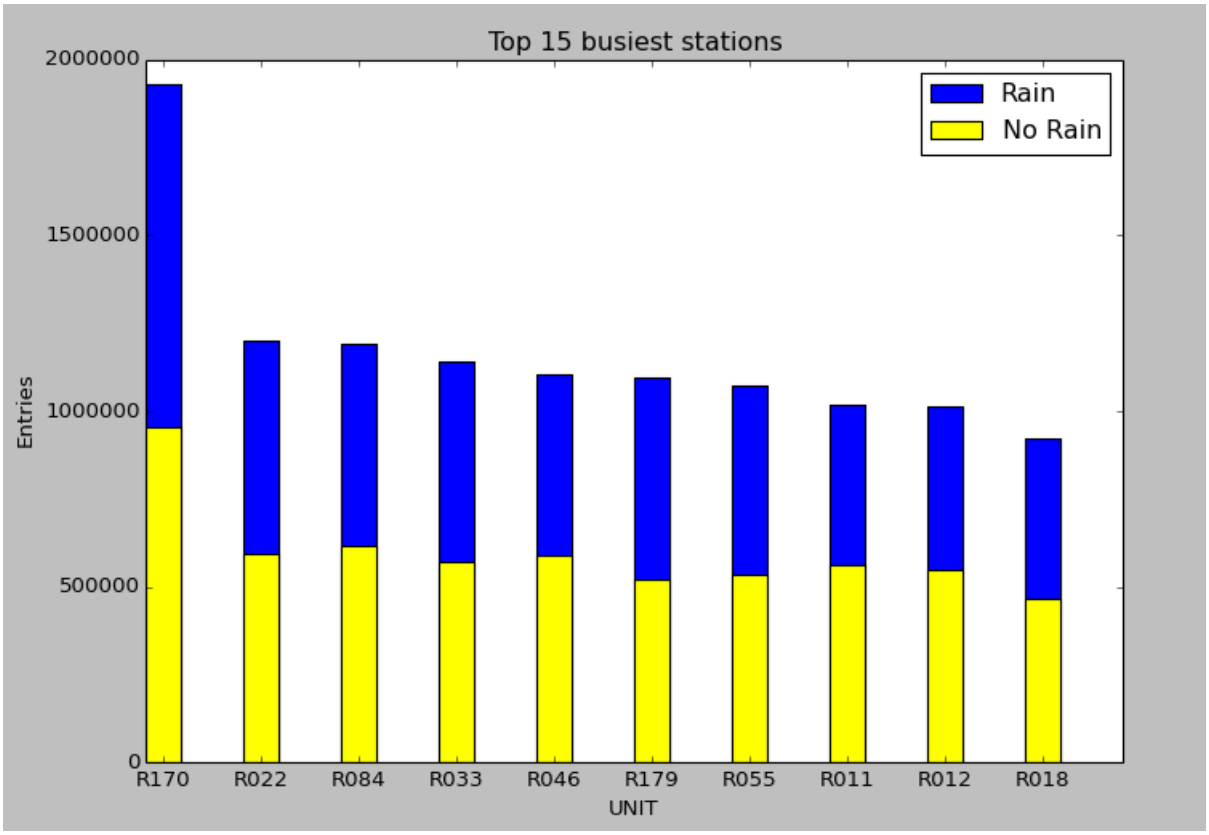
Please include two visualizations that show the relationships between two or more variables in the NYC subway data. Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

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



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.

I chose to plot the top 15 busiest stations on the month, separate by whether it was raining or not. I chose that because I could see by the regression parameters that the UNIT is the main driver on the prediction model, clearly, central units will always have more riders than others, and I was curious to see what that difference was.



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?

Based on my analysis on the data provided, I can conclude that more people ride the NYC subway when it is raining. I was able to get to this conclusion by examining the hourly entries in several UNITS of the NYC subway. I separated the data into two dataset (one with entries on rainy days and the seconds with entries on non-raining days) and then performed a Mann-Whitney U-Test with a p-critical value of 0.05.

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 result of the test was a p-value of 0.024999 which is smaller than the critical value of 0.05, suggesting strong evidence to reject the hypothesis that rain doesn't affect the Subway ridership.

On another subject, a linear regression model was fit to the data and the positive coefficient for the rain parameter indicates that the presence of rain contributes to increased ridership.

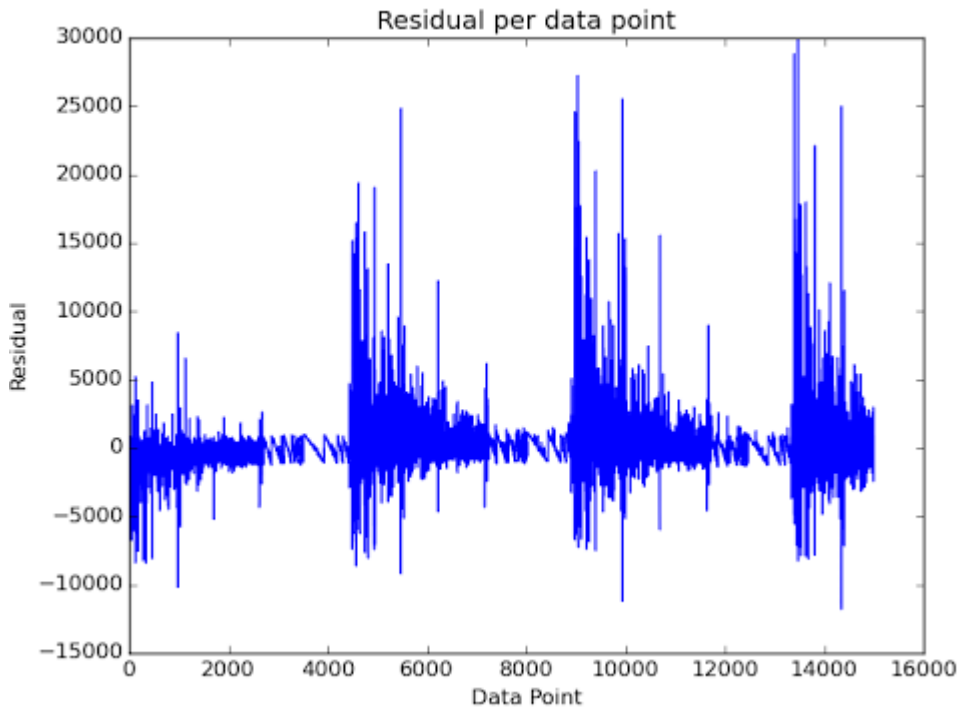
Section 5. Reflection

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

I believe that there are a few variables that could have been added to the dataset to help analysis, like holidays for example. A sunny holidays would probably have more riders than a regular rainy weekday and we can't tell the difference, which can lead to bias.

The dataset itself also doesn't have a good time span of data to be analysed. It contains only the month of May, which will probably have a very different ridership pattern than august or December where the weather is significantly different.

Regarding the model used for regression, as we discussed on section 2.6, linear regression is not the best fit for this dataset due to the very large residuals produced (please refer to the histogram on section 2.6 for details). For further details on this, we can plot the residuals per data point where we can see that, even though there are a few good predictions (where the residual is very small), there are several predictions where the residual is quite large.



Section 0. References

Udacity's U-test explanation

Mann–Whitney U test Wiki page