Analyzing the NYC Subway Dataset

by Eduardo Eidelwein Berlitz in fulfillment of Udacity's Data Analyst Nanodegree, Project 1

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?

The data was analyzed with the Mann-Whitney U-Test with a two-tailed p-value where the null hypothesis indicates that both samples being compared are statistically identical using a significance level of 5%. As described by the *scipy.stats.mannwhitneyu* docs the reported p-value is for a one-sided hypothesis, so to get the two-sided p-value the returned p-value must be multiplied by 2.

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 assumption that the test is making about the distribution of ridership in the two samples is whether subway ridership varies with the weather. So visualizing the histogram of the data in rainy and non-rainy days we can see that data is not normally distributed. The Mann-Whitney U-Test is a non-parametric test which does not assume any particular distribution, as opposed to Welch's t-test, making this test appropriate.

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.

Mean entries in groups of rainy days and non-rainy days were 1105.45 and 1090.28; the distributions in the two groups differed significantly (Mann–Whitney U = 1924409167.0, n1 = 44104,n2 = 87847, P < 0.05 two-tailed). Where two-tailed p-value equals to 0.04999982558.

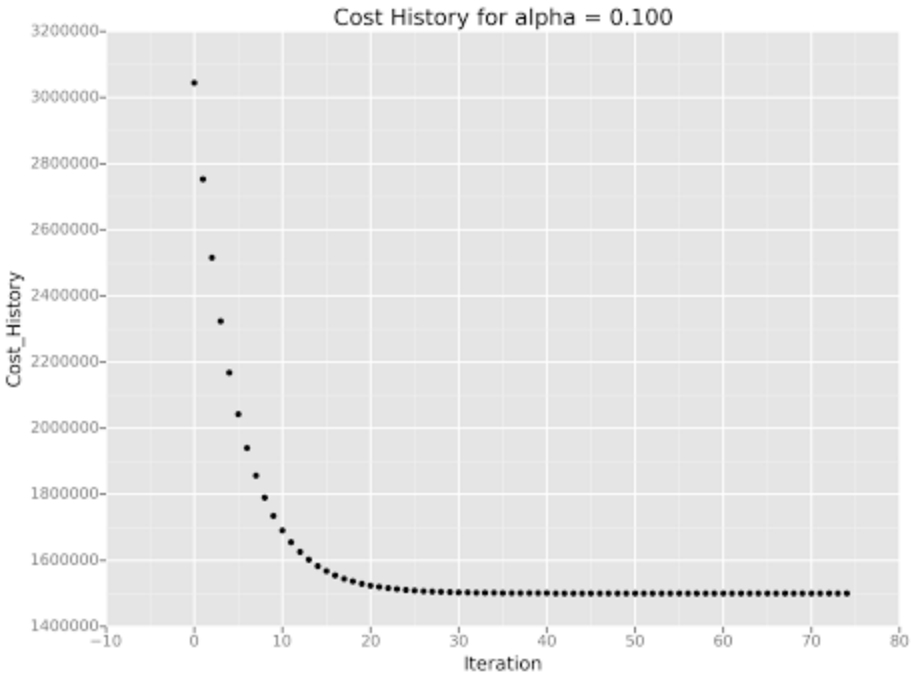
1.4 What is the significance and interpretation of these results?

As the two-tailed p-value is smaller than the significance level of 5% (0.05) we can reject the null hypothesis that the difference is due to random sampling, and conclude instead that the populations are distinct. So we can conclude that the distribution of entries is statistically different between rainy and non-rainy days.

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 Gradient descent to compute the coefficients theta and prediction of the ENTRIESn_hourly. The default values for alpha and the number of iterations were used and as seen by plotting the cost history by the number of iterations the regression model converge on a local minimum.



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 for Gradient Descent were: rain, precipitation (precipi), hour, mean temperature (meantempi) and dummy variables for individual stations (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.

I decided to use features that relate in some way whether, hour of day and ridership. Based on exploration and experimentation of the features the most suitable were those described before because they improved my R^2 value. First I included *Hour* because the ridership changes through day, and it can give us a more detailed insight. Then I added *meantempi* as a weather feature in my model and it improved my R^2 value. I tried to change this feature by *meandewpti***,** but it did not affect the R^2 value as *meantempi* did. So I decided to use *meantempi* as temperature is a component of the weather that affects people's decision. I also used the *rain* and *precipi* feature because I thought that when it is raining outside people might decide to use the subway more often. Also, as *Unit* is a non-numeric feature it was added more columns with dummy data to get it in a format we can work with.

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

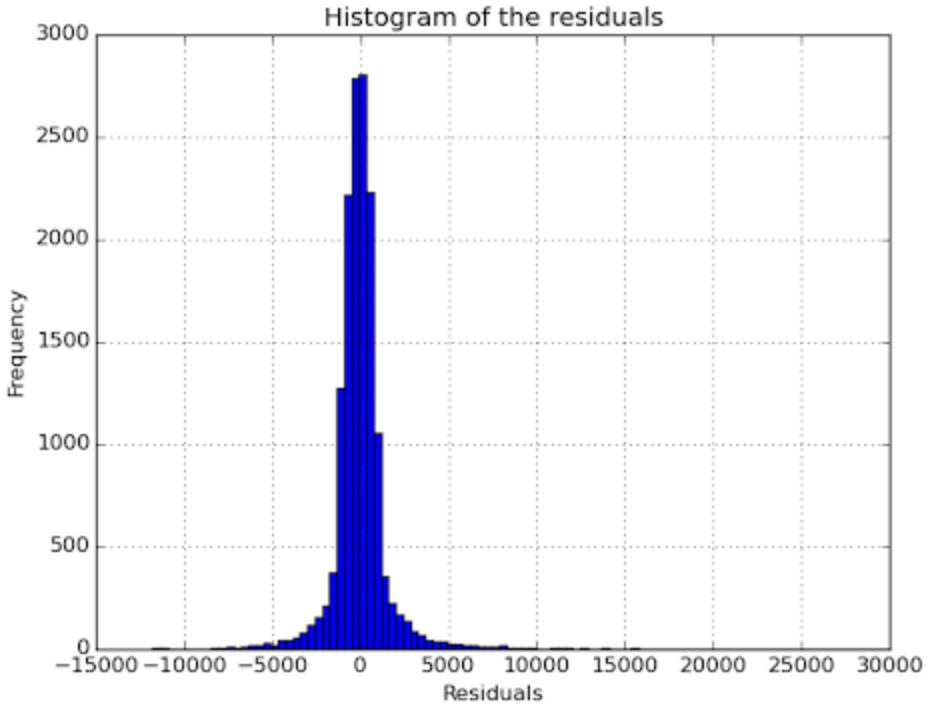
- rain: 2.92398062e+00
- precipi: 1.46526720e+01
- Hour: 4.67708502e+02
- meantempi: -6.22179395e+01

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

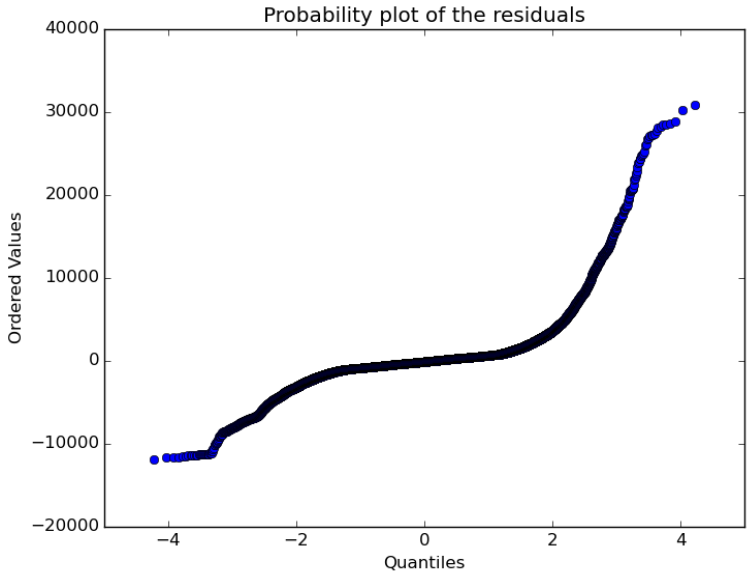
0.463968815042

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?

In general, the goodness of fit for a linear regression model can be determinate if the differences between the observed values and the model's predicted values are small and unbiased. The regression model accounts for 46.4% of the variance based on the R-squared which is a statistical measure of how close the data are to the fitted regression line. But it cannot determine whether the coefficient estimates and predictions are biased, which is why the residual plots should be assessed.



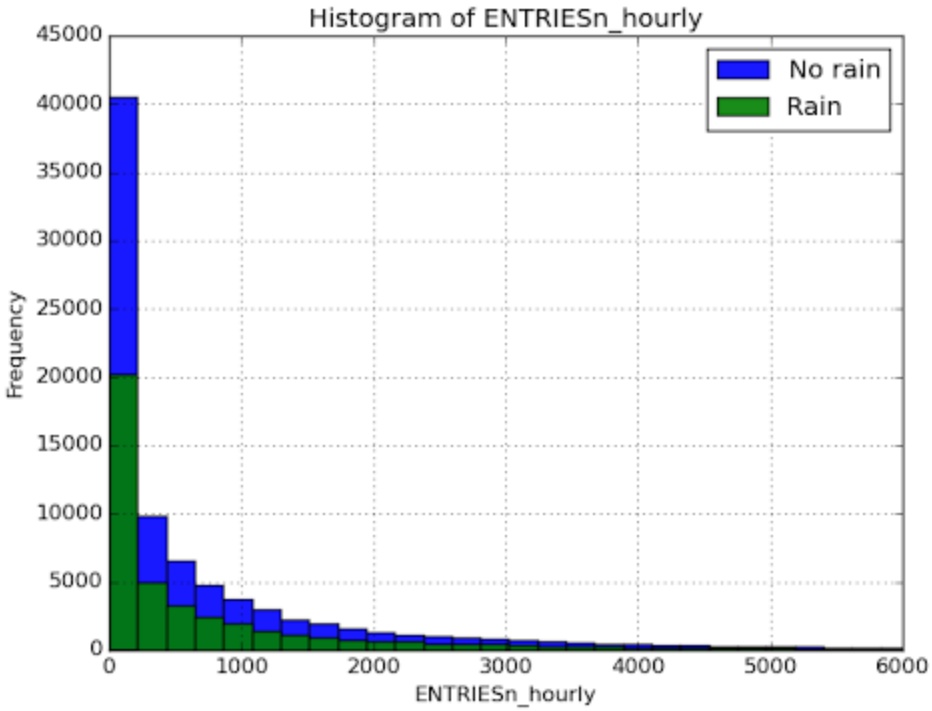
Note that the histogram of the residuals has long tails, which suggests that there are some very large residuals. We can use the normal plot of residuals to verify the assumption that the residuals are normally distributed. The normal probability plot of the residuals should approximately follow a straight line.



The probability plot of residuals does not approximately follow a straight line, which means that the residuals does not follow a normal distribution. Thus, prediction intervals may be inaccurate. So based on this and the R-squared value, that is relatively low, I conclude that this linear regression model is probably inappropriate to predict ridership for this dataset. Further and advanced study could check for other problems with the model, such as missing terms or a time order effect, which could lead to include more features or the use of polynomial regression to improve the goodness of fit for this model.

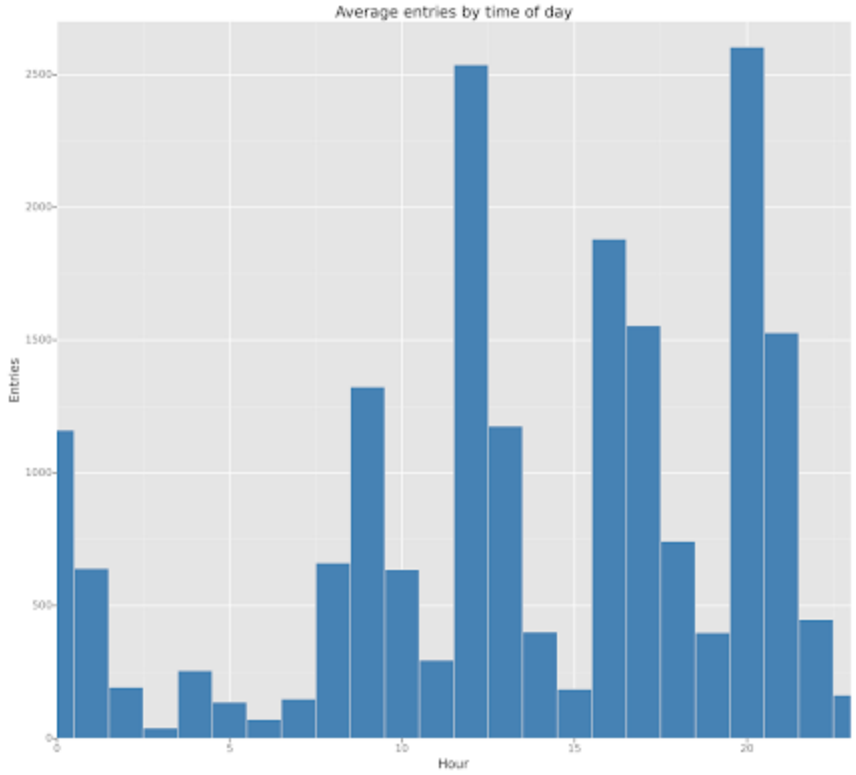
Section 3. Visualization

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



Plotting this two histograms of subway entries for rainy and non-rainy days we can determine that both distributions are not normally-distributed. Also, as seen by these plots, there are many more samples for non-rainy days than for 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.



The above chart shows the average hourly ridership by time of day. The sum of ENTRIESn_hourly by Hour was divided by the count of rows for a given time of day. The chart shows that the average ridership is higher at noon and at 20pm.

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 this analysis, more people ride the NYC subway when it is raining.

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.

Despite the fact that the means of both populations (rainy and non-rainy days) are almost the same, based on the results from the Mann-Whitney U test (where the two-tailed p-value is smaller than the significance level of 5%) we can conclude that the

distribution of entries is statistically different between rainy and non-rainy days, in other words, rain affects the ridership. Also the positive coefficient (section 2.4) for the rain parameter in the regression model indicates that the presence of rain contributes to increased ridership. So we can conclude that more people ride the NYC subway when it is raining.

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.

One of the possible shortcomings of this analysis may be the fact that the NYC subway data were joined with the weather data on a daily basis and not on an hourly basis. Also the dataset only included data from one month (May 2011), increasing the dataset could change this results and conclusions. Also special days like holidays or sport games could change the way of how we predict values. This prevent a more detailed analysis of how weather can affect ridership within a day.
The linear regression model could certainly be improved, although it was suitable for the purpose of study. As noted in Section 2.6, the inclusion of more features or polynomial combinations could have increased the model accuracy. However, this can lead to a significant over-fitting, and the model may fail to new data sets. In this case, regularization would be a good method to attenuate any over-fitting.

Section 6. References

- GraphPad - How the Mann-Whitney test works
- Probability Values
- The Minitab Blog - Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit?
- The Minitab Blog - Why You Need to Check Your Residual Plots for Regression Analysis: Or, To Err is Human, To Err Randomly is Statistically Divine
- ggplot from ŷhat docs
- scipy.stats.mannwhitneyu docs
- The Minitab Blog - Interpret all statistics and graphs