Intro to Data Science
Udacity Nanodegree

Daniel Nitiutomo

Section 1: Statistical Test

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?***

I used the Mann-Whitney U test to analyze whether features of the NYC subway data set were significant in predicting the number of entries. This test was the most appropriate as the data set was not normally distributed and Mann-Whitney U does not assume a normal distribution unlike t-tests and z-tests. I tested whether subway entries changed significantly when it rained. I used a two-tailed p-value because I was testing whether or not the 'rain' and 'no rain' distributions would yield different results when selected from randomly. In other words, I was testing for any difference in the two data sets, and did not care which one was greater than or less than the other. My null hypothesis was:

$$H_0: P(x > y) = 0.5\text{, where:}$$

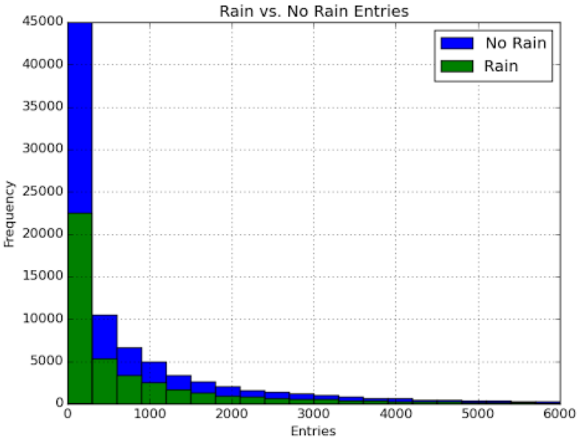
x is a random sample from the distribution of subway entries when it rains AND

y is a random sample from the distribution of subway entries when it does not rain

The p-value I calculated from the Mann-Whitney U test was 0.0249999 (one-tailed) or approximately 0.0499 for the proper two-tailed test.

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 Mann-Whitney U test was applicable to this dataset because it does not make any assumptions about the underlying distribution; unlike the other statistical tests such as t-tests and z-tests which assume the underlying distributions are normal. This test also does not assume that the variances of the two populations are equal. Two assumptions that the test does make are (1) that the two samples are independent and (2) the observations are ordinal or able to be ranked. After graphing the distribution of results in a histogram, it was clear that the data from the rain and no rain sample populations were not normal, as shown below. Therefore, the Mann-Whitney U test was the most appropriate test to analyze the data set.



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.***

P-Value = 0.0249999, P-value (two-tailed) = 0.0499, Rain Mean = 1105.4464, No Rain Mean = 1090.2788

4. ***What is the significance and interpretation of these results?***

Using a 5 percent significance, the null hypothesis was rejected as the p-value for the two tailed test was approximately 0.0499. A rejection of the null can be interpreted as the two distributions (rain and no rain) will produce meaningfully different values if each was drawn from randomly. In other words, the differences in the samples are statistically significant. Therefore, rain could be interpreted as a meaningful feature to include in prediction models such as linear regression. The p-value, however, would not reject the null under a more strict significance value such as 2.5 percent.

Section 2: Linear Regression

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 both Gradient Descent and Ordinary Least Squares in statsmodel to conduct linear regression and calculate coefficient thetas.

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

I tested many features when conducting Linear Regression. Features that I used included ‘Hour’, ‘Day’, ‘Mean/Max Temperature’, ‘Rain’, ‘Precipitation’, and ‘UNIT.’ The final features which I chose to include were ‘Hour’, ’Day’, and ‘UNIT.’

All three of the inputs in my final predictions were Dummy Variables:

- Hour_[X]: 1 = it was this hour, 0 = it was not this hour
- UNIT_[X]: 1 = it was this Unit, 0 = it was not this UNIT
- Day_[X]: 1 = it was this Day, 0 = it was not this Day

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.***
   a. ***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.”***
   b. ***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.”***

Initially, I used intuition to determine which features to include in the model. I started with a few features (rain, mean temperature, hour) to establish a base case R^2 value:

- I chose Rain and Temperature because I thought if there was extreme weather people would be less willing to walk to the subway and would prefer to either not travel or take a cab.
- I chose hour because I know that during certain times of the day people are much more likely to be riding the subway such as rush hour or the morning/afternoon commute.

To determine which were significant I added and removed features, ran the regression again, and compared the initial R^2 value to the one generated after each run. If the R^2 value improved by more than 0.01, I chose to include it within my model.

When I included the UNIT feature, the regression's R^2 value improved significantly. This confirmed the reasoning that stations located in a more populated or commercial area will receive higher traffic as a result of more people living and working there. I kept the 'HOUR' and 'DAY' variables as they increased the R^2 value by a significant amount as well.

I chose to remove all of the weather variables because none of their correlations with subway entries were above 0.03 (using absolute value) and they had a negligible effect on R^2 when included in the regression. This led me to believe they do not have predictive power.

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

Gradient Descent:

- Constant = 112.2378

Ordinary Least Squares:

- Constant = 1.419e+14

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

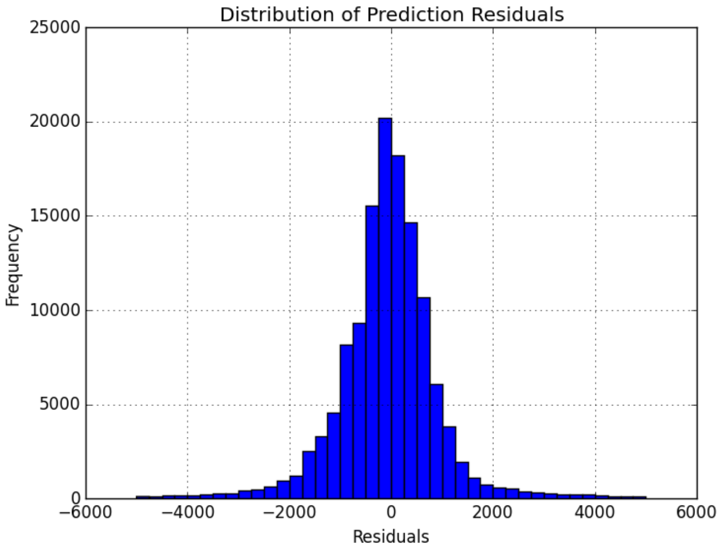
Gradient Descent R^2: 0.51382

Ordinary Least Squares R^2: 0.51378

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 is a measure of how much the variance within the data set is described by the model. The higher the R^2 value the more variance is explained by the model. While a higher R^2 value is an approximation for how well the model fits the data, one must be careful of using it as the only method of judging how successful the model is in prediction. There will always be noise that is associated with any data set. If the R^2 value is too high then this could be an indication of over-fitting where the noise in the data will also be included within the future predictions; ultimately, hurting the effectiveness of the model. Also, adding variables to the linear regression will increase the R^2 value even if that variable is only loosely associated with the prediction variable because the model will have more data to fit to. Therefore, if a variable only increases the R^2 slightly it indicates that the variable can be left out of the regression test.

The R^2 values indicated that these models only explained approximately half of the variance seen in the models. I would consider this an approximate fit or an estimate and not one which will have a low prediction error. Therefore, when predicting subway entries this model can be used for estimation purposes and cannot be relied on for low error predictions.



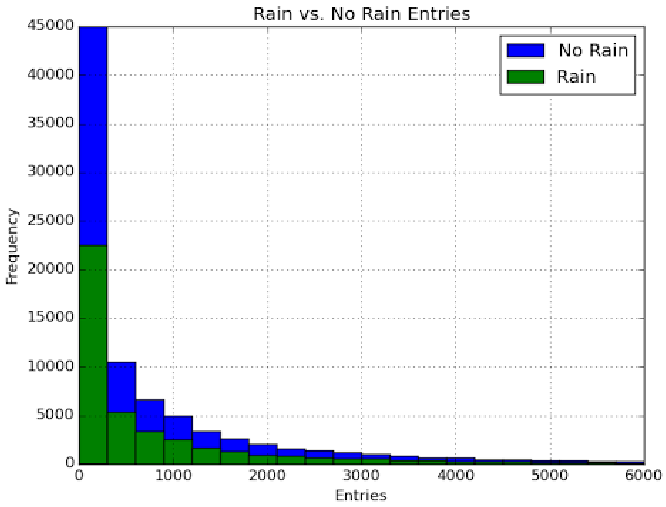
Looking at a distribution plot of the residual errors for gradient descent, however, the residual errors are distributed somewhat evenly around 0 indicating that linear regression could be a suitable model for this analysis.

Section 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.

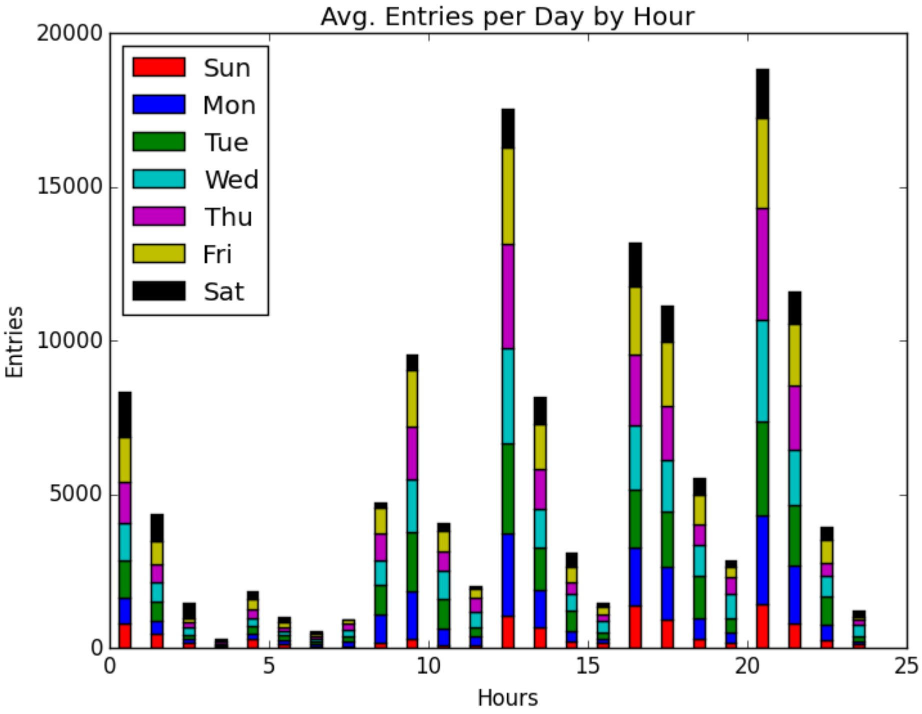
Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

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



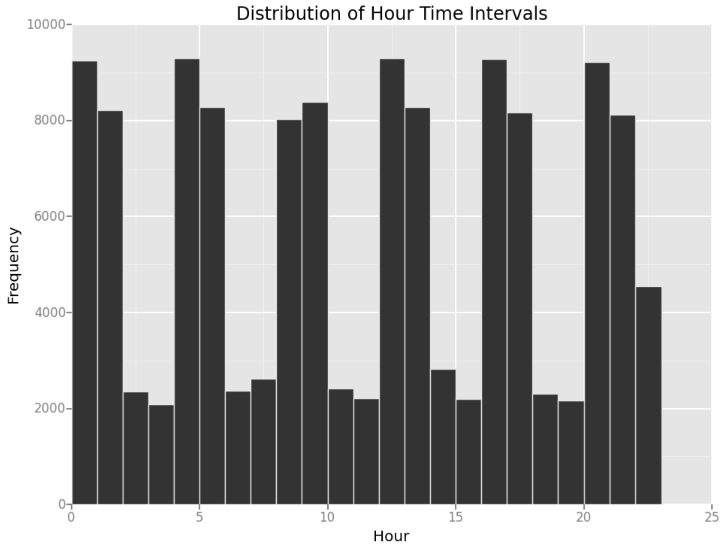
The graph above indicates that the distributions for subway entries when it is raining and not raining are almost identical and non-normal. This tells me that any test which assumes normality to identify the significance of the "rain" feature should not be used. Instead a hypothesis test that does not assume normality should be used such as the Mann-Whitney U test. This graph also leads me to believe that the rain feature may not be significant since the data distribution does not change when it rains vs. when it is not raining.

2. ***One visualization can be more freeform.***



The chart above shows the average number of subway entries per day, stacked by hour. This stacked bar chart indicates that there is significant variation in subway entries depending on the time of day. It also shows a pattern of sudden rises in ridership for certain hours and gradual declines from those sudden rises. For example, the sum of the average daily entries for 11 AM is roughly 2,000; while the sum just an hour later jumps to about 17,500. In the subsequent hours (13, 14, 15) there is a gradual decline in ridership and then a sudden increase again at 16. This pattern continues throughout the day and appears to be consistent over days. A visible and consistent pattern indicates there is predictive power within the data and that the hour variable should be included within any regression.

I made the choice to average the ridership over day instead of summing them because of the inconsistent time intervals the entries used. Since the 4 hour intervals were measured at slightly different times, one time interval could be represented far more than others causing a skewed bias to the results. I confirmed this by looking at the distribution of Hour within the data set.



It is clear that there is an uneven distribution of observations across time intervals, making it difficult to compare totals.

Section 4: Conclusion

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

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?***

No, I do not believe that rain is significant when predicting ridership of the NYC subway.

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***

In order to come to the conclusion that rain is not an important variable in predicting NYC subway ridership, I used the results from the linear regression models and compared the distributions of rain vs. no-rain subway entries.

I used the Mann Whitney U test to assess whether the number of subway entries changed based on rain. The hypothesis that the rain vs. no-rain data sets produced approximately the same values was rejected at the 5% significance level. The two tailed p-value was 0.04999 or approximately 5. This initially led me to believe that the rain variable was significant.

Whenever 'rain' was added as a feature for both Gradient Descent and OLS regressions, however, the increase in the R^2 value was negligible or less than 0.01. This indicated that it could be left out of the regression as it did not add much as a predicting variable.

Looking at the graph in 3.1, it is obvious that the distributions of the rain and no rain data sets are highly similar if not identical. The means and variances for hourly entries on the rain and no rain data sets are also similar. The likeness of the two data sets indicates that the 'Rain' dummy feature did not add much predictive power to the regression.

- Means: rain = 1105.44637675, no rain = 1090.27878015
- Variances: rain = 5619274.04184, no rain = 5382361.64079

While I chose to initially include the 'rain' feature within linear regression, further analysis indicated that the feature is not significant when it comes to predicting NYC subway ridership.

Section 5: Reflection

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

1. ***Please discuss potential shortcomings of the methods of your analysis, including:***
    a. ***Dataset***
    b. ***Linear Regression Model***
    c. ***Statistical Test***

Dataset

- The data is gathered in inconsistent time intervals between Units. For example, the entries for some units were measured at hours 0:00, 4:00, 8:00, 12:00, 16:00, and 20:00; while the entries for other units were measured at 3:00, 7:00, 11:00, 15:00, 19:00, and 23:00. Since entries were not measured consistently across units, statistics derived must be seen as approximations when compared.
- The weather variables are all measured over a full day instead of being measured in consistent time intervals as the entries data. This inconsistency made it difficult to analyze the changes in subway ridership behavior based on weather. The analysis of weather variables effect on subway entries would have been much more substantial if it had been measured at the same times as the entries were.

Linear Regression Model

- Linear regression assumes that there is a linear relationship between subway entries and the prediction features. Testing multiple different sets of features, however, I was not able to run a regression that returned an R^2 value of greater than 0.52 using both gradient descent and OLS. This R^2 value implies that there is still significant amount of variance within the data that is not explained by the model. This could be because a linear model may not be the best to use to predict subway entries given the data.

- Linear Regression also does not give insight into why a variable may be statistically significant making it difficult to separate noise from meaningful relationships. The process of rationalizing a statistically significant result after the fact can be misleading.

Statistical Test

- Initially I used the Mann Whitney U statistical test to aid me in drawing a conclusion of whether the rain variable was significant. I pre-determined that it was if the null hypothesis was rejected at a 5% significance level. The null was barely rejected at a p-value of 0.049999 and I chose to include it in my tests. After further analysis, I removed the feature as it did not prove to add predictive significance to the regression models. While statistical tests are important in allowing scientists to draw conclusions based on probability, they cannot be used as a definitive measure of significance.

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

From my analysis, UNIT is by far the most significant feature to use in prediction of NYC subway ridership. This makes sense logically and is proven by the data. The R squared value of both the regressions improved significantly when I included a proxy UNIT feature into the models, indicating that it is highly significant.