Bill Su
Udacity Project 2

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?

Because distributions of customer entries with rain and without rain are not normal and heavily skewed to the right, we have decided to use Mann-Whitney U test instead of a T-test. Since we are unsure about the direction of the differences of means, we have decided to use a two-tail P value (p result of the test *2). The null hypothesis is that there are no differences between two datasets, inferring that rain does not have effect on subway entrance number, the alternative hypothesis is that there is a significant difference between rainy and non-rainy days relative to amount of people entering the subway. For this test we have set the p-critical value at .05, a standard alpha value.

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.

As explained above, in order to conduct a t-test, we must make the assumption that both rainy and non-rainy dataset are distributed normally. However, after plotting, we have realized that this assumption cannot hold true. Ergo, we have decided to use Mann Whitney U Test, which does not assume any specific types of 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.

The mean for the rainy dataset was 1105.45, the mean for the non-rainy dataset was 1090.28, the U test returned a U result of u = 1924409167, p =.05. Therefore we just have enough evidences to reject the null hypothesis in alpha = .05.

1.4 What is the significance and interpretation of these results?

Since we have rejected the null per result of the previous question, we now may conclude that our data suggest there is a significant difference between means of entry numbers for rainy days and non-rainy days.

2.1 What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model:

1. **OLS using Statsmodels or Scikit Learn**
2. **Gradient descent using Scikit Learn**
3. **Or something different?**

I have used Gradient Descent via Scikit Learn to compute the model.

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

I have select rain, hours of day, subway units, and mean temperature as my features. In which Unit is a dummy variable.

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.

Because I have assumed that when it is raining people will usually enter the subway to avoid rain or take a subway instead of walking, thereby increasing entrance number. Same rationale will apply for the assumption that higher mean temperature will increase subway ridership since subways are air-conditioned. Different subway stations (Wall Street or otherwise) will also see different ridership just based on their locations. Finally, I know for sure that there are certain hours of the day (rush hours) when more people will ride the subway to get to work or get home.

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

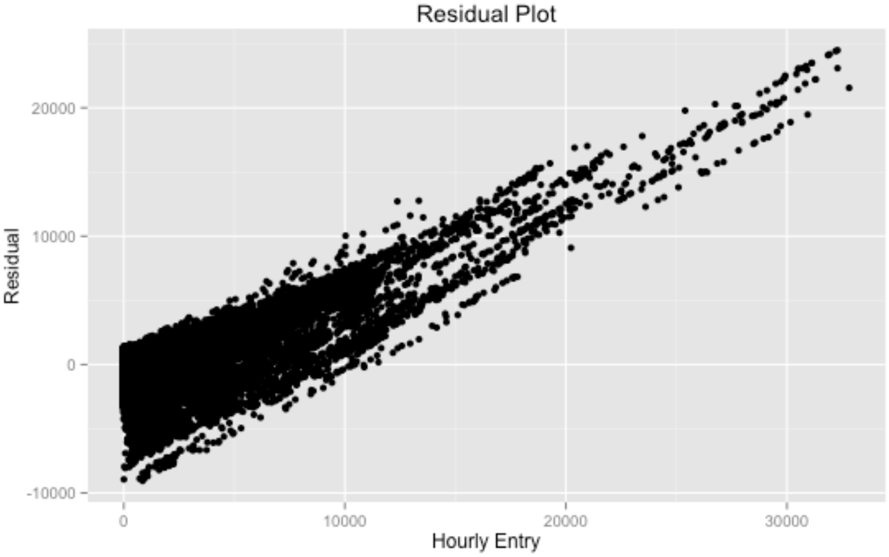
The coef for rain is -27.00, for hour is 395.56, for mean temp is -60.99.

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

The r^2 returns a value of .413

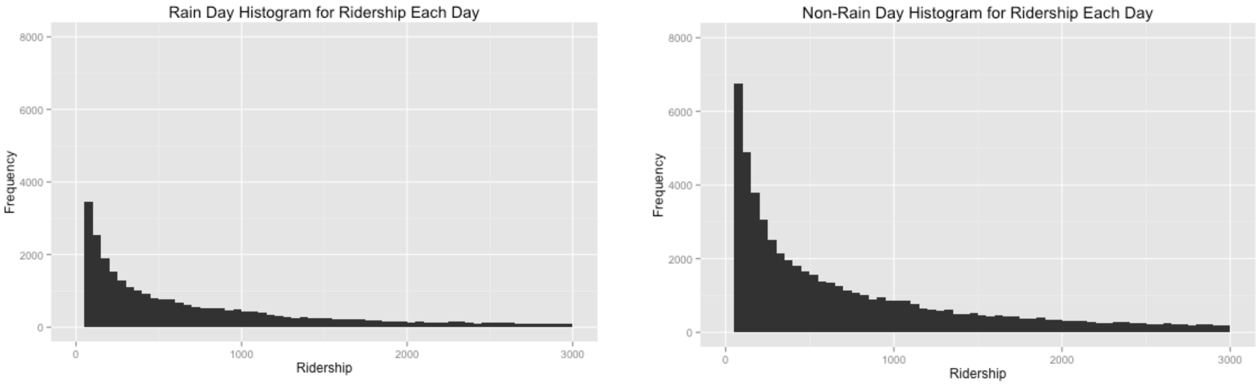
2.6 What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?

The r^2 means that 41.3% of the variance in the data can be explained through my prediction model. This model will be a good predictor of ridership in NYC subway because the R^2 value is in a moderate range. However, as illustrated in the residual graph displayed below. There is a clear linear pattern in the residual plot, meaning that linear model alone is not the best model to fit through this dataset. The residual plot is displayed in the next page.



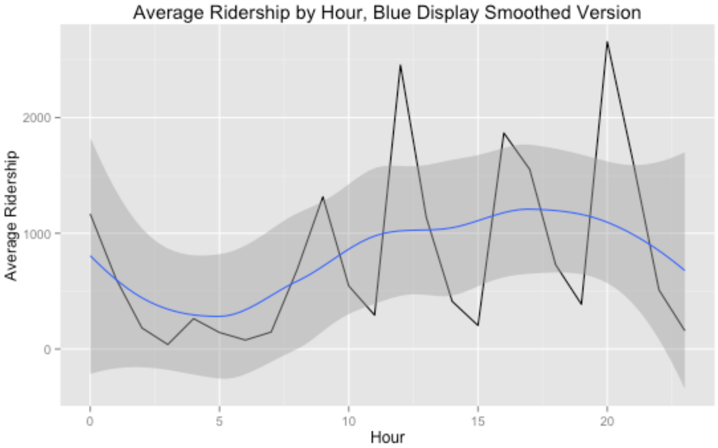
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

Figure 1: The graph of dataset from rainy and nonrainy days presents two similar distributions



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.

Figure 2: The smoothed line graph of means shows that ridership is reaches its peak between 15:00 and 20:00 while reaches its low point at around 5:00



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?

More people ride subway when it is not raining, but this conclusion need more solid support.

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.

From frequency distribution of both graphs, it is not entirely obvious which mean is greater. However, after the U test, we have concluded that there is a difference between means of ridership in rainy and nonrainy days. Meanwhile, we also found that the mean of rainy days is greater than the mean of nonrainy days. Overall, after all observations, I have concluded that rain must increase ridership. A.k.a, more people ride the NYC subway when it is raining.

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.**

If we were to look at the second visualization example, we will realize that subway ridership by hour is for sure not a continuous variable as we have imagined in the linear regression analysis. For later analysis, the variable hour should be used as a categorical variable instead of a continuous one. Furthermore, we only recorded whether during it day it rained, I would imagine that if the rain occurred at night it would not effect ridership at all since no one is riding the subway anyway. Therefore I would suggest record rain by hour instead by day.

At the same time, I also think that we should exclude last night hours where subway ridership is close to 0 and instead only focus on 5am – 10pm. We might get a different distribution and the filtered data will enable us to better understand subway ridership pattern.