Name: Long Nguyen

Course: Udacity Data Analyst Nanodegree

Project: P2: Analyzing the NYC Subway Dataset

1. The statistical test that I used to analyze the NYC Subway data is the Mann-Whitney-U Test. This test is chosen after taking a look at the histogram of the number of entries in NYC subway in raining and not raining conditions. The histograms shows that the distributions of the number of entries does not follow a normal distribution. The Mann-Whitney-U test is sufficient when we want to see if the difference between too samples is statistically significant without assuming the normality of population distribution or taking large sample size to approximate normality.

   I used the one-tail P values because we are guessing that the number of entries will increase when it rains. The null hypothesis is that when it rains, the number of people riding the subway will not be significantly different than when it doesn't rain. That means if we were to count every person entering the subway when it rains and when it doesn't rain, we will always find approximately the same number. Our alternate hypothesis is that there will be a difference. That means that when it rains the total amount of people riding subway will be significantly higher than when it doesn't rain.

   In mathematical term:

   Null hypothesis: $\mu_R - \mu_{NR} = 0$
   Alternate hypothesis: $\mu_R - \mu_{NR} > 0$

   We will perform this test using significance level of $\alpha = 0.05$. This makes our p-critical value smaller than 0.05, which means that if our p value is smaller than 0.05, we can reject the null hypothesis.

   The result I got from Mann-Whitney-U test:

   With rain mean = 1105.446
   Without rain mean = 1090.279
   U value: 1924409167.0
   P value: 0.0249

   These results mean that assuming the total number of people riding subway when it rains is not different from the total number of people riding subway when it doesn't rain, the chance of selecting a sample difference of 15.167 (1105.446-1090.279) from the sampling distribution of sample means is only 0.0249 (2%), which is less than our threshold of 5%. Because of this, we can conclude that the difference in ridership does not happen by chance, but rather by an actual difference in the total number of people riding subway when it rains versus when it doesn't rain. We also know that the total number of people riding subway when it rains will be higher because we get a higher mean in our sample.

2. I used OLS using Statsmodels to compute the coefficients theta and produce prediction for ENTRIESn_hourly in my regression model.

   The features that I used in my model are: rain, precipi, Hour, meantempi, fog. I used UNIT as a dummy variable.
   I chose rain, precipitation, and fog because I believe intuitively that when the outside condition is not ideal for driving or walking, then subway will be a good idea, which explains why the ridership in subway rises under these conditions.

   Hour is chosen because it intuitively makes sense to think that people uses subway more at certain hour in the day (such as when they get to and out of work)

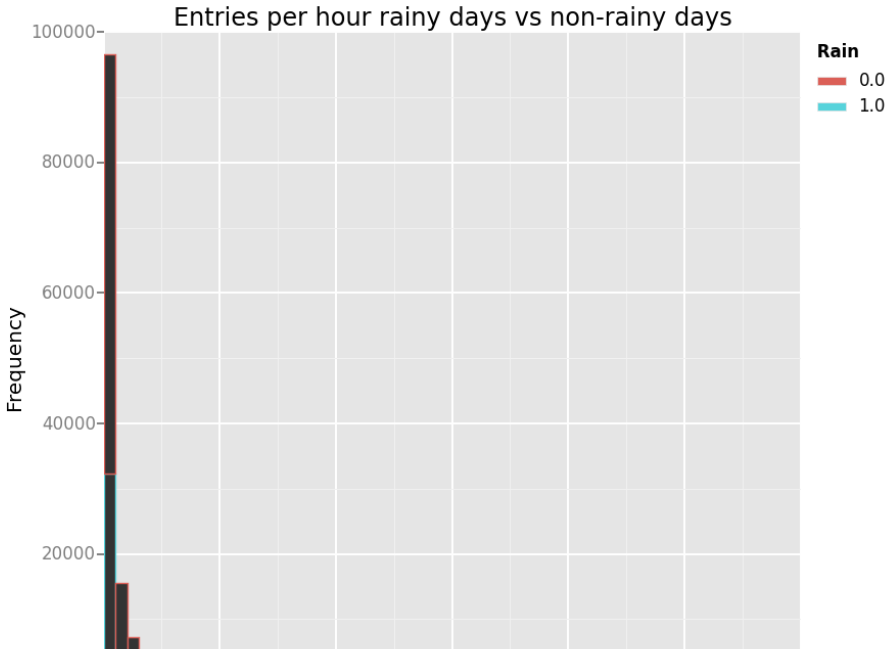
   UNIT is chosen as a dummy variable because its value can't be interpreted quantitatively, and also it's helpful to include UNIT as a variable to see which unit adds the most weight to the number of subway entries.

   The coefficients for non-dummy variables are as follow:

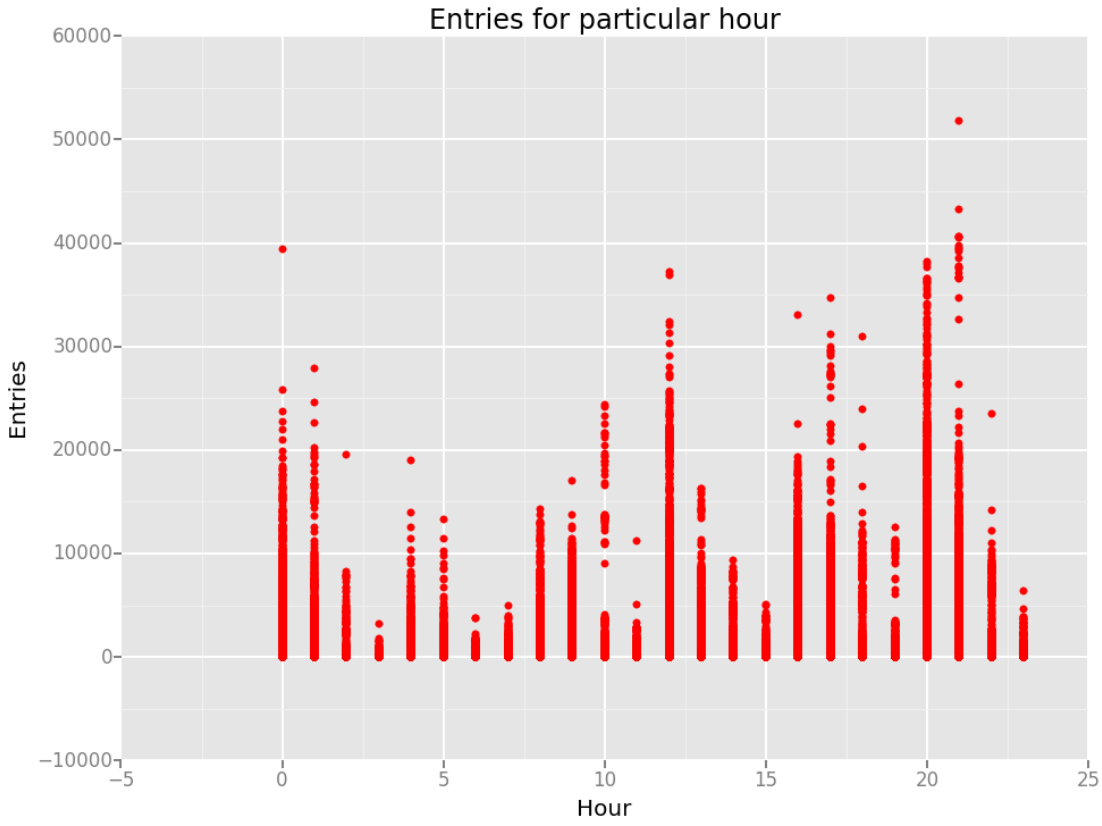
   Rain: -8.4885
   Precipi: -32.6507
   Hour: 65.3139
   Meantempi: -12.4992
   Fog: 188.0755

   The model's R-squared is 0.480. This value means that our model is able to explain 48% the variability of our data. To think if this model is appropriate to predict ridership for this dataset, we have to look at other information such as standard error and adjusted R-square to fully interpret the significance of our R-squared.

3.



Ridership by time of day:



4. Conclusion

From my analysis, it is shown that more people ride the NYC subway when it is raining then when it is not raining. When we perform the Mann-Whitney-U test to see if this difference is significant, we found out that the difference is statistically significant with $p < 0.05$. This has led me to conclude that the ridership when it's raining is definitely higher than when it's not raining.

However, in our regression, why are we seeing the coefficient for rain is -8? While it may seem to be against our statistical test, it is not. Let's look at the first few rows in our regression result:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        ENTRIESn_hourly   R-squared:                       0.479
Model:                            OLS   Adj. R-squared:                  0.460
Method:                 Least Squares   F-statistic:                     25.03
Date:                Sun, 04 Oct 2015   Prob (F-statistic):               0.00
```

Time:	06:36:03	Log-Likelihood:	-1.1673e+05
No. Observations:	13195	AIC:	2.344e+05
Df Residuals:	12726	BIC:	2.379e+05
Df Model:	468		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
const	1539.1268	154.481	9.963	0.000	1236.320	1841.933
rain	29.4645	39.018	0.755	0.450	-47.016	105.945
precipi	28.7264	44.323	0.648	0.517	-58.154	115.607
Hour	65.3346	2.204	29.645	0.000	61.015	69.655
meantempi	-10.5318	2.331	-4.519	0.000	-15.100	-5.963

As we can see, the t-statistic for rain is quite small: only 0.755. Also, the p value is 0.45 > 0.05. This means that we cannot reject the null hypothesis that the coefficient of rain is 0. We have reason to believe that if we were to calculate the rate of change in rain variable per unit of change in entries of the entire population, and take average of them, then the mean change will not be significantly different from 0. Therefore, our data shows that the number of entries when it's raining is significantly higher than when it's not raining, yet raining itself is a poor indicator to predict the number of entries.

5. Some shortcomings from the data are that the data doesn't include enough features for an effective prediction. In reality, people rely on a lot of factors more than weather to decide whether or not to use the subway. Some variables that can be added to improve our analysis include the geographical area that each subway routes cover, a dummy variable which indicates if there are any special events going on near one of the stations, or the amount of cars purchased last year to predict the potential traffic. All of these variables would add much more insight and predictive capability in our model.

   One of the shortcomings of our analysis is that we treat hour variable as a quantitative variable while it makes more sense to treat it as a dummy variable. Since people often join traffic in certain hour in the day (rush hours), rate of change in hour doesn't always translate to rate of change in ridership. For example, our coefficient indicates that a unit of increase in hour will increase the entries by around 65. This is not appropriate in the sense that at later time, there may be less people using subway. A better approach is to treat hour as a dummy variable with 1 and 0 for rush and non-rush hour or 1 and 0 for certain particular hour.

References

N/A