```
Analyzing the NYC Subway Dataset
```

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?

- I use the Mann Whitney U-test on the ENTRIESn_hourly column in the turnstile_weather dataframe.
- I use a two-tail P value.
- The null hypothesis is that: The distributions of the number of entries in the population are not statistically different between rainy & non rainy days
- My p-critical value is 0.05

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 Mann Whitney U-test assumes [Ref1] that

- All the observations from both groups are independent of each other
- The responses are ordinal
- The distributions of both groups are equal under the null hypothesis

In our case, it satisfies these assumptions. In particular, Mann Whitney U-test doesn't require samples are normally distributed.

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 P value is 0.024999908316389704 and my critical value is 0.05. Since I use two-tailed test, and 0.024999908316389704 is for one-tailed test, double 0.0249999 and get 0.0499, which is smaller than 0.05. I reject the null hypothesis. So the distributions of the number of entries in the population are statistically different between rainy & non rainy days
- The mean for rainy days is 1105.4463767458733
- The mean for non rainy days is 1090.278780151855

1.4 What is the significance and interpretation of these results?

I used the critical value 0.05, and I get the P value 0.024999908316389704. 2 * 0.0249999 = 0.0499, which is smaller than 0.05. That means it falls into the critical region. So I can reject my null hypothesis. So the two distributions for rainy days and non rainy days are statistically different.

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 use a. OLS using Statsmodels

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

- I use 'rain', 'precipi', 'Hour', 'meantempi' as input features.
- Yes. I use 'UNIT' as 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

- I use feature 'rain' because I think on rainy days, people would ride the subway more often
- I use 'precipi', 'Hour' and 'meantempi' because by my tests, I find this combination has a higher r^2 value than others(like 'rain','fog','precipi' and 'meantempi').
- I use dummy variable "UNIT" because this UNIT variable will have a large influence on the Entries_hourly. Also, the values for "UNIT" variable are categories so I use dummy variable.

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

- The weight for 'rain' is 29.46452873
- The weight for 'precipi' is 65.33456526
- The weight for 'Hour' is 28.72638025
- The weight for 'meantempi' is -10.53182494

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

My R2 is 0.47924770782

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?

We know

$$R^2 = 1 - \frac{\sum(y_i - y_i')^2}{\sum(y_i - y_m)}$$

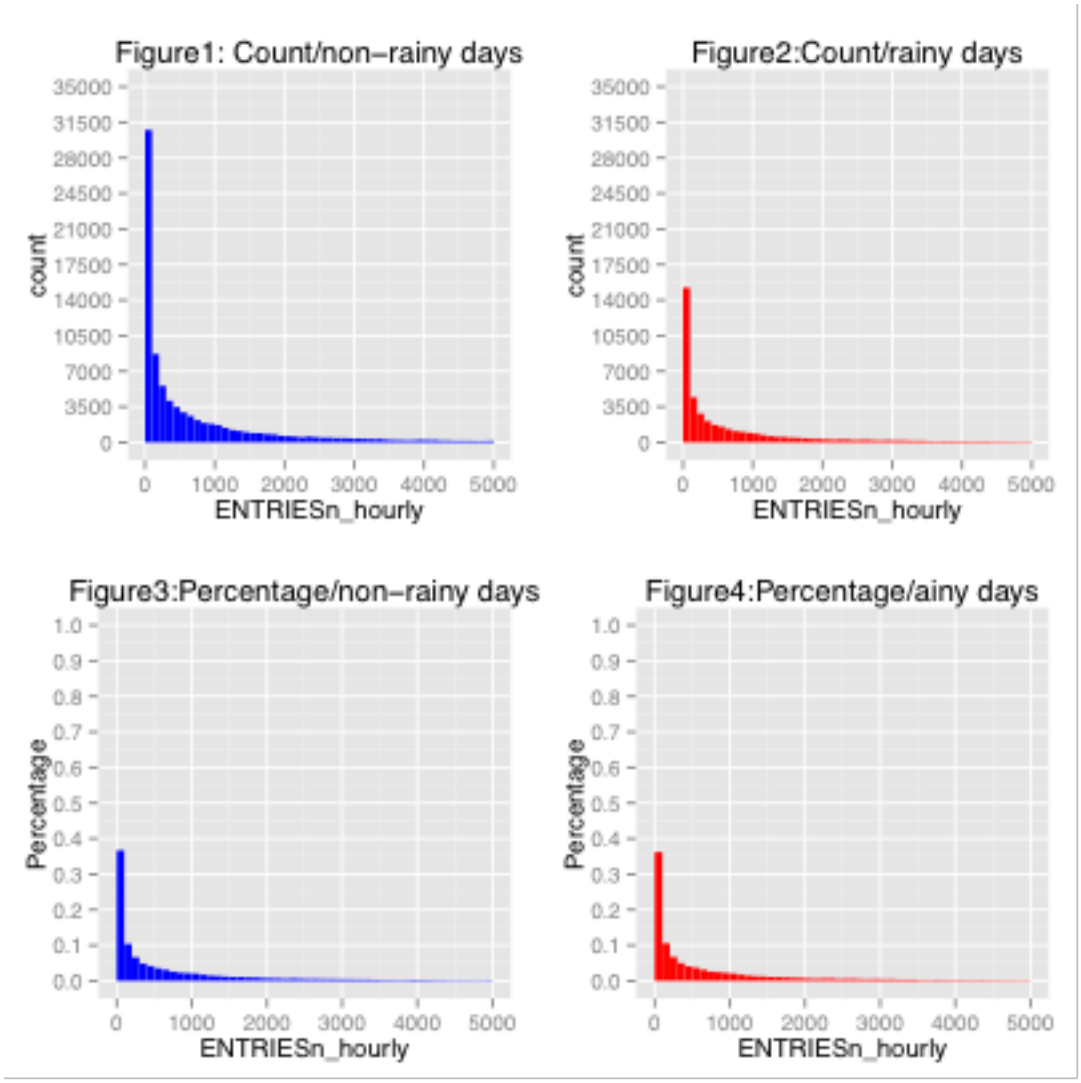
My R2 is 0.479 means that we know that the variability of the ENTRIESn_hourly around my regression line is (1 - 0.479) times the origianl variance. i.e.

$$\frac{\sum(y_i - y_i')^2}{\sum(y_i - y_m)} = 0.521$$

So 47.9% of the total variance in ENTRIESn_hourly is explained by my linear model. The variance of the predicted ENTRIESn_hourly by my linear model equals about 52% of the total variance of the originalw ENTRIESn_hourly.

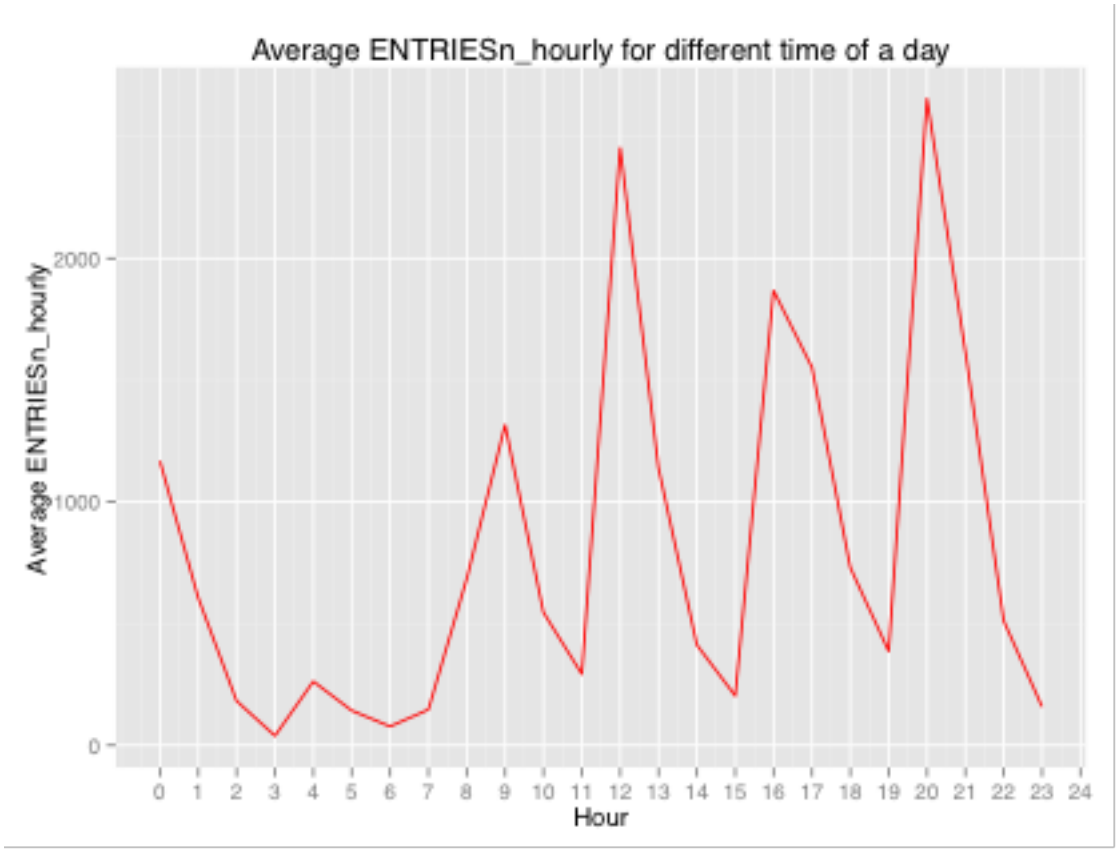
By examining the distribution of the residuals (Question 6 in Problem Set 3: Analyzing Subway Data), I find the histogram of the residuals has long tails. That means there are some very large residuals. This may say that my linear model is not very appropriate for this dataset.

Section 3. Visualization



Fiture "histograms of ENTRIESn_hourly for rainy days and non-rainy days" shows the count and the percentage of ridership for non-rainy days (0.0) and rainy days(1.0). For all four histograms, x-axis is the values of ENTRIESn_hourly and y-axis is the values of frequency or percentage. Binwidth is 100 and range for x-axis is 0 to 5000, range for count is 0 to 35000, range for percentage is 0 to 1.

From the count histogram for rainy days and non-rainy days, it seems that more people ride the subway on non-rainy days. But notice that there are 87847 observations for non-rainy days and just 44104 observations for rainy days. Figure 3 and Figure 4 show the percentage for each bin.



Fiture

"Average ENTRIESn_hourly for different time of a day" shows the relationship between ENTRIESn_hourly and Hour. x-axis is Hour (different time of the day) and y-axis is the average ENTRIESn_hourly for that Hour. We can see that more people ride subways on some particular time of the day (such as 9am, 12 p.m. 4 p.m. and 8 p.m.)

Section 4.

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 and visualization, 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.

From the results of the linear regression, we know the weight for feature 'rain' in the linear regression model is positive. That means when it's raining (1.0), the value for ENTRIESn_hourly is bigger than not raining (0.0). So more people would ride the subways on rainy days.

From the results of the statistical tests, I reject the null hypothesis and choose the alternate hypothesis that the distributions of the number of entries in the population are statistically different between rainy & non rainy days

So based on results from statistical tests and linear regression, I think more people ride the NYC subways on rainy days.

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.

The shortcomings for Dataset: * I'm not sure about the source of the dataset I used. So the dataset may be not reliable. * Besides, the size of the dataset may be not big enough. If the dataset doesn't contain many data, the analysis may lead to some wrong results. * The data are mainly for 2011/05, we need to collect more data from different months. Because in May, maybe some people are on vacation. * There are some extreme values in the dataset. We should handle these values before we use this dataset.

The shortcomings for analysis: * I just used linear regression model. One question is that whether non-linear model can predict better? * Also, I don't know how many features can predict better. I think I should use some data to build different models and then use other data to test these different models, then I can see which model performs better.

References

1. topic "Mann–Whitney U test" in Wikipedia https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test
2. topic "r2, a measure of goodness-of-fit of linear regression" http://www.graphpad.com/guides/prism/6/curve-fitting/index.htm?

r2_ameasureofgoodness_of_fitoflinearregression.htm