Brian Chase
11/16/2015

Analyzing the NYC Subway Dataset

Section 0. References

http://matplotlib.org/api/pyplot_api.html
http://www.bertplot.com/visualization/?p=229
https://pypi.python.org/pypi/ggplot/
http://ggplot.yhathq.com/docs/geom_histogram.html
http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm
https://storage.googleapis.com/supplemental_media/udacityu/4332539257/MannWhitneyUTest.pdf
http://www.theanalysisfactor.com/interpreting-regression-coefficients/
https://en.wikipedia.org/wiki/Multicollinearity

Section 1. Statistical Test

1.1 Which statistical test did you use to analyze the NYC subway data?
Mann-Whitney U-test

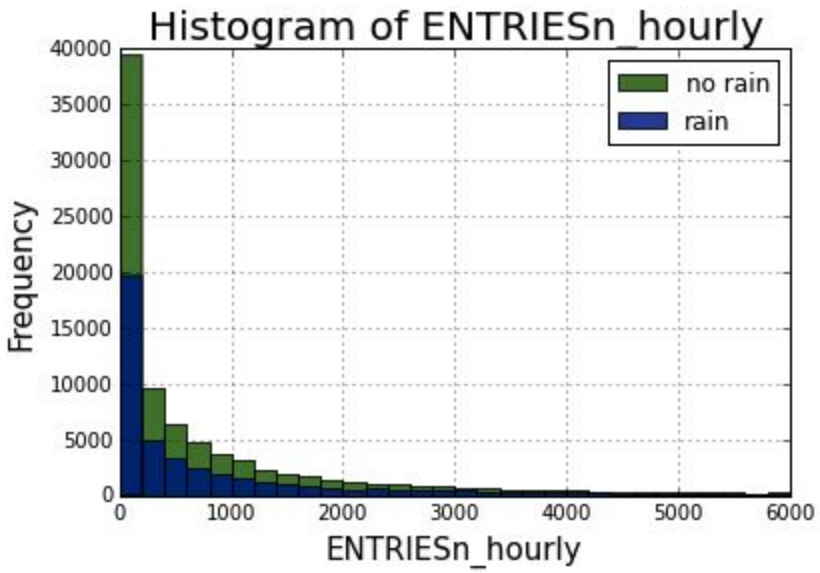
Did you use a one-tail or a two-tail P value?
I used a two-tail P value to test for statistical significance in both directions.

What is the null hypothesis?
The null hypothesis H_0 is the probability of a random draw from population 'rain' exceeding a random draw from the second population 'no rain' equals the probability of a random draw from population 'no rain' exceeding a random draw from population 'rain.'

$H_0 : P(x > y) = 0.5$

What is your p-critical value?
p-value = 0.05

1.2 Why is this statistical test applicable to the dataset?
The Mann-Whitney U-test is used because the dataset does not follow a normal distribution (see below) and the samples come from the same population (NYC subway riders).



1.3 What results did you get from this statistical test?
p-value = 0.049999825586979442
Mean (with rain) = `1105.4463767458733`
Mean (without rain) = `1090.278780151855`

1.4 What is the significance and interpretation of these results?
We consider the distribution of the number of entries per hour on to NYC subway stations of rain days versus non-rain days statistically significant and different since the p-value (0.049999825586979442) < p-critical (0.05).

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?
OLS using Statsmodels

2.2 What features (input variables) did you use in your model?
I used rain (whether it was raining or not), precipi (how much precipitation), Hour (hour of the day), meantempi (mean temperature) and Unit (proxy for trainstation used as a dummy).

Did you use any dummy variables as part of your features?
Yes. I used the UNIT column as a dummy which turned out to be critical for getting a good (greater than 0.4) R^2 value. This allowed for the model to differentiate between the different subway station locations.

2.3 Why did you select these features in your model?
For ‘rain’ and precipi’, I figured that there would be a change in the people going to ride the subway based on days that it rained and that the magnitude of the rain mattered.

For ‘meantempi’, my guess was that when it was extremely hot or cold there would be less riders as they would try to stay home.

For ‘hour’, I thought it that ridership would be different during different times of the day based on events like rush hour for work.

For ‘UNIT’ used as a dummy, the different subway stations are physically located in different places with different people so this was important to be made as a feature of the model.

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

rain -20.195400
precipi 43.697290
Hour 67.396606
meantempi -7.271767

Here we see that the parameter for rain is negative---meaning that the model subtracts 20.1954 from the prediction ENTRIESn_hour when rain = 1. It is interesting to see that the output of the Mann-Whitney U-test function for mean also estimated a lower mean for rainy days ~15 entries per hour.

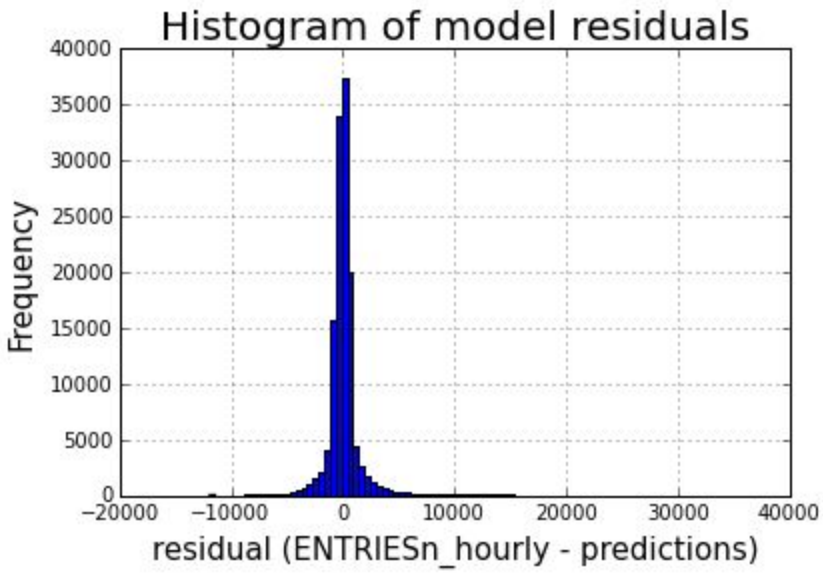
Also important to note here is that the coefficient for rain will changed based on the variables included in our model. Predictor variables are almost always associated and can explain the same variation. To see the effect of just the rain variable on our model, we would only include rain as a variable.

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

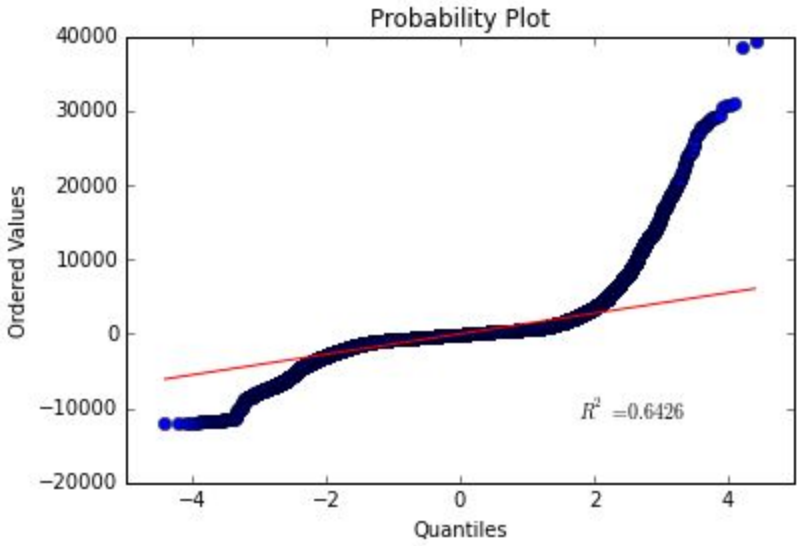
R^2 = `0.47924770782`

2.6 What does this R^2 value mean for the goodness of fit for your regression model?

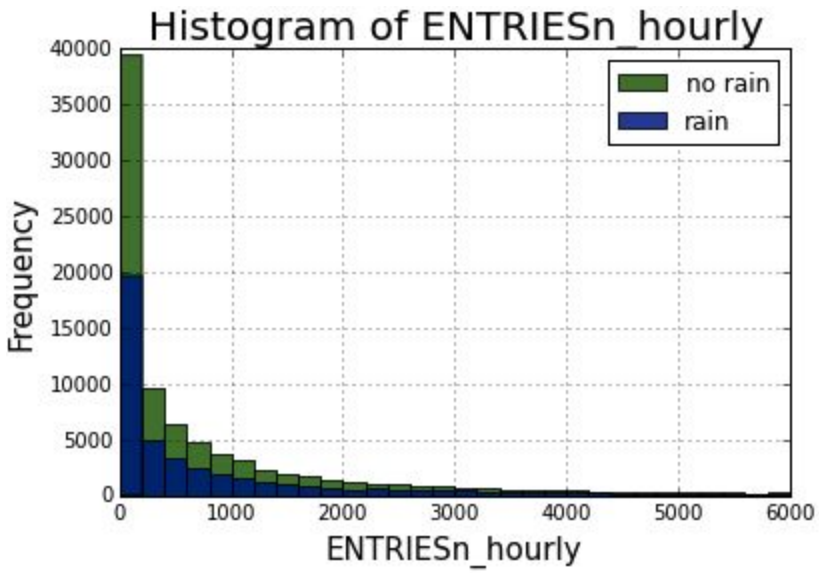
This R^2 value means that we have explained 48% of the original variability with this linear model. To assess the performance of this R^2 value, we plot a histogram of the model residuals and see there are some very long tail values.



We have reason to question our linear regression model based on the very large residual values. Using a probability plot below, we confirm that the residuals are not of normal distribution and that a non-linear model would be more appropriate for this dataset.



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



Code:

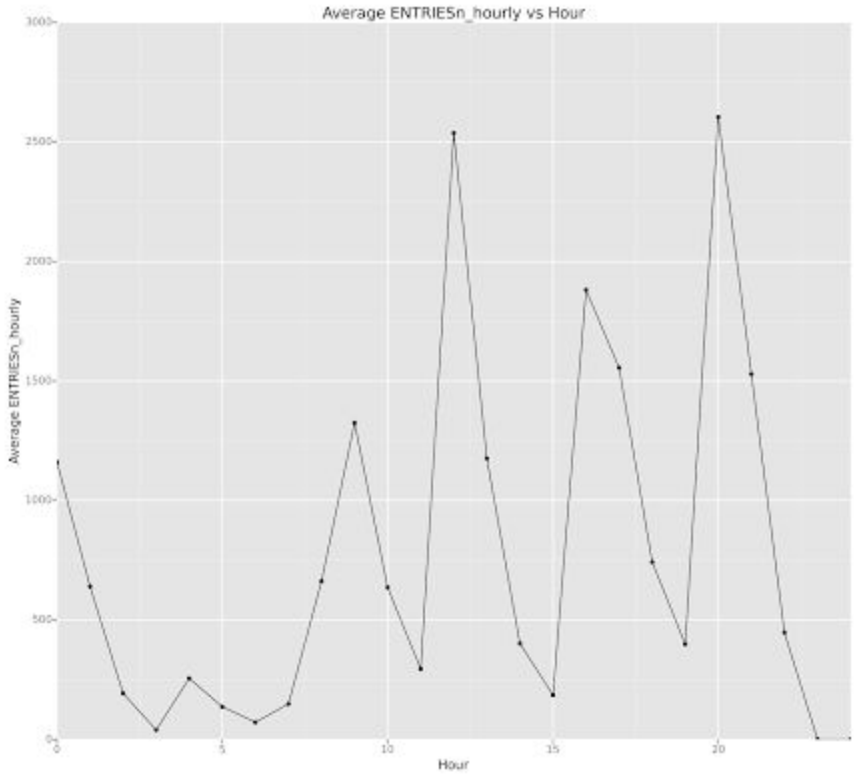
```
plt.figure()
turnstile_weather['ENTRIESn_hourly'][turnstile_weather['rain'] == 0].hist(bins=30, color = 'DarkGreen', alpha = 0.8, label = 'no rain', range=(0,6000)) # your code here to plot a historgram for hourly entries when it is not raining

turnstile_weather['ENTRIESn_hourly'][turnstile_weather['rain'] == 1].hist(bins=30, color = 'DarkBlue', alpha = 0.8, label = 'rain', range=(0,6000)) # your code here to plot a historgram for hourly entries when it is raining

plt.title('Histogram of ENTRIESn_hourly', fontsize=20)
plt.ylabel('Frequency',fontsize=15)
plt.xlabel('ENTRIESn_hourly',fontsize=15)
plt.legend()
```

Key insight:
The distribution for ENTRIESn_hourly is not a normal distribution. And the frequency of the lower number of ENTRIESn_hourly bins are larger in magnitude with the ‘no rain’ case having a higher frequency than the ‘rain’ case.

3.2 One visualization can be more freeform.



Key insights:
The peak entries occur at 9am, 12pm, 3pm, and 8pm---while the morning times seem reasonable, the evening hours does not correspond with my idea of rush hour subway traffic (5pm). The chart also shows us that while there is a positive correlation in ENTRIESn_hourly later in the day---a linear model for this data would not do a great job predicting solely on Hour.

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?

The statistical test performed confirms that more people ride the subway when it rains. The two-tailed P value is less than the chosen P critical value and reject the null hypothesis.

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.
Firstly, the Mann-Whitney U-test rejected the null hypothesis that there is no difference between days that it rains and days that it does not---admittedly, it barely passes the criteria for the p-critical value. This also deems the difference between the means statistically significant.

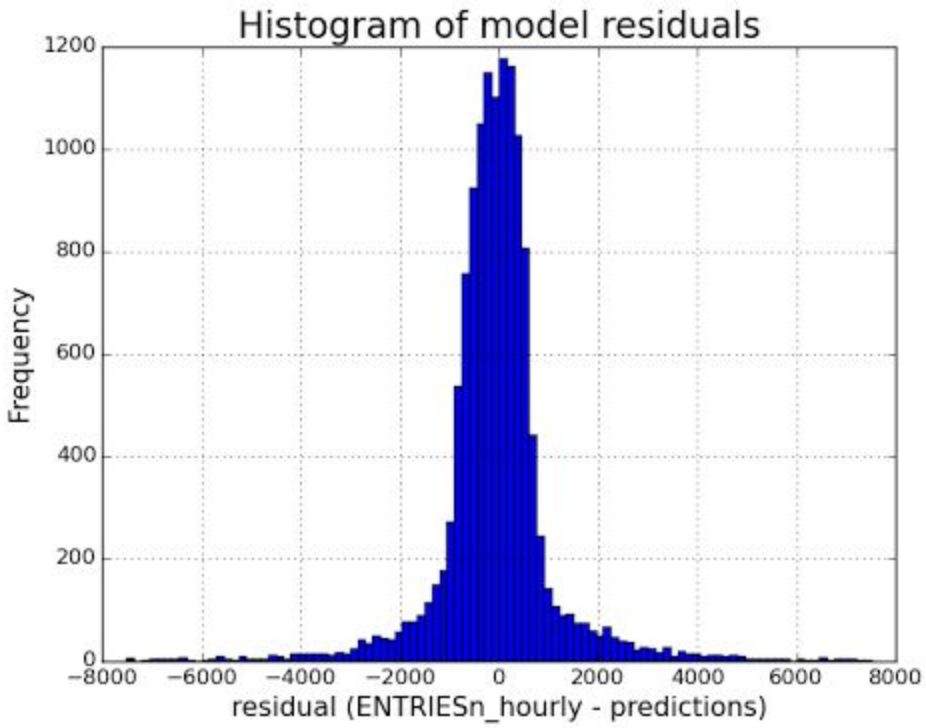
Both the difference in means for ENTRIESn_hour for days with rain and without rain (-15.1676) ---and the parameter of rain in our linear regression model (-20.1954) reflect the same negative magnitude in relation to rain days.

Also looking into the residuals for the linear regression model predictions vs actual data gives extra confidence that the model (and the negative parameter for rain) is reasonable, since it displays a normal distribution (histogram shown below).

Code for histogram:

```
plt.figure()
(turnstile_weather['''ENTRIESn_hourly'''] - predictions).hist(bins = 100, range = (-7500, 7500))

plt.title('Histogram of model residuals', fontsize=20)
plt.ylabel('Frequency',fontsize=15)
plt.xlabel('residual (ENTRIESn_hourly - predictions)',fontsize=15)
```



Section 5. Reflections

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.

The dataset does not come with any supplemental document that describes how the data was taken and the accuracy of the measurements. For example, are the temperatures recorded accurately between the different locations using the same test sample methods? Another thing I noticed when trying to determine

whether to use entries or exits hourly to estimate ridership the sums do not equal each other (sum of ENTRIESn_hourly minus sum of EXITSn_hourly = 27506194).

The dataset timespan is also rather short at only 15 days of data (from 2011-05-01 to 2011-05-15). A higher quality dataset and analysis would use a significantly longer timespan.

Using a linear model here is not appropriate as we reported from the findings of very large residuals shown in the probability plot and histogram (in section 2.6).

There are many variables in the dataset that are most likely highly linearly correlated, for example---mintempi, meantempi, and maxtempi---that could lead to multicollinearity. In that case, it may not be valid to use the individual predictors by themselves to glean information on how it affects the output variable.