Analyzing the NYC Subway Dataset

Name: Tianzhixi Yin
Deadline: JUL 06 '15

Section 0. References

The Mann–Whitney U test in scipy return a "nan" value for the p-value, therefore I calculate the p-value by myself with the help of the webpage: http://stats.stackexchange.com/questions/116315/problem-with-mann-whitney-u-test-in-scipy

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 the NYC subway data. I used a two-tail P value. The null hypothesis is that if we draw hourly entries randomly from a raining day (x) and a day without rain (y), the probability of x being higher than y is 0.5. My p-critical value is 0.05.

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 is a nonparametric test, it makes no assumption about the distribution of the population, therefore this statistical test is applicable to the dataset. Whereas the two-sample t test assumes normality and by looking at the histograms I know this assumption is invalid. That's why I did not choose the t test.

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.

   There is evidence to reject the null hypothesis. My p-value is $5.48269387142e-06 < 0.0001$. The mean of ridership on raining days is 2028.20 and the mean of ridership on days without rain is 1845.54.

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

   The test indicates that there is a significant difference between the two distributions. The means show that it is the ridership on raining days higher than the ridership on days without rain.

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 OLS using Statsmodels or Scikit Learn

   b Gradient descent using Scikit Learn

   c Or something different?

   I used OLS in Statsmodels.

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

   I finally chose "UNIT", "hour", "weekday", 'fog", "rain", "tempi", and "wspdi" as my features. Among them, "UNIT" is treated as a dummy variable.

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.

   At first, there are several features I thought might be important to the ridership:

- “UNIT”: different remote units should have different ridership
- “hour”: different time during a day should have different ridership
- “weekday”: I think the ridership on weekday and weekend should be different
- “station”: different station should have different ridership
- “conds”: the weather condition is important
- “fog”: whether there is fog or not should affect the ridership
- “precipi”: how much precipitation should affect the ridership
- “rain”: whether there is rain or not should affect ridership
- “tempi”: the temperature should affect the ridership
- “wspdi”: how strong the wind is should affect the ridership

I don’t like the “mean” features because weather can be quite different at different time during a day. Obviously “UNIT” and “station” are highly correlated, and by switching between the two I found that “UNIT” provided higher R^2 value, therefore I chose “UNIT”. Also, “rain” and “fog” are highly correlated with “conds”, since I care specifically about “rain”, I kept “rain” and “fog” and removed “conds”. Also, since “precipi” is highly correlated with “rain”, I did not include it.

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

   - “rain”: 67.1850
   - “hour”: 121.4417
   - “weekday”: 962.7689
   - “fog”: -580.0869
   - “tempi”: 3.5344
   - “wspdi”: 6.5104

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

   0.481937666554

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 value is around 0.48, which means our model explains around 48% of the total variation. This R^2 value is neither very good nor very bad. However, when examining the distribution of the residuals in the QQ plot, I observed large deviation from the theoretical line for residuals greater than 5000. And there are long tails in the histogram of the residuals. These large residuals should be a reason to question the linear regression model. Also, by plotting the residuals per data point I noticed that there was a cyclic nonlinear effect that the linear model failed to describe. Therefore I would say this linear model is not appropriate to predict ridership for this dataset. Figure 1 is the QQ plot, Figure 2 is the histogram of the residuals, and Figure 3 is the plot of residuals per data point.

Section 3. Visualization

1. Histograms

   Figure 4 is the histogram for rainy days and Figure 5 is the histogram for non-rainy days.

   Obviously the distribution of ridership does not follow a normal distribution. It is more like a Poisson distribution.

   For the rainy days, most hourly entries are less than 2000, but still the entries that are greater than 1000 appear quite frequently and small entries are not the dominating force. For non-rainy days, small entries happen much more often than the medium or great entries. I think the frequent low entries lower the mean and median entries on non-rainy days, while on rainy days, medium entries is a big part thus its mean and median are both higher.

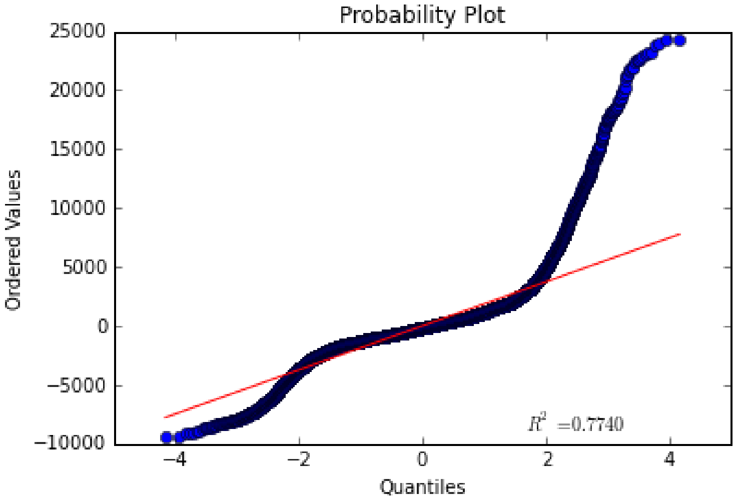

Figure 1: QQ plot of the residuals

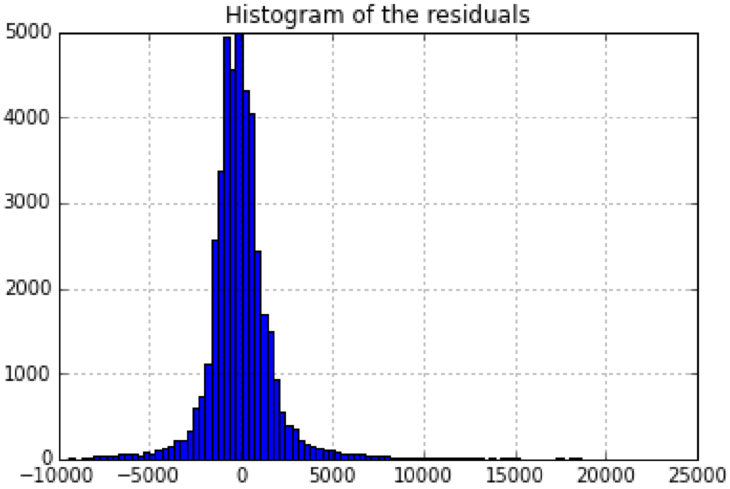

Figure 2: Histogram of the residuals

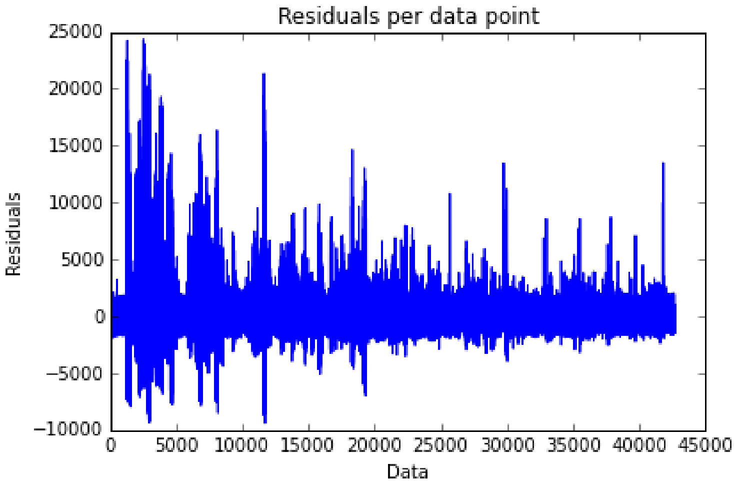

Figure 3: Residuals per data point

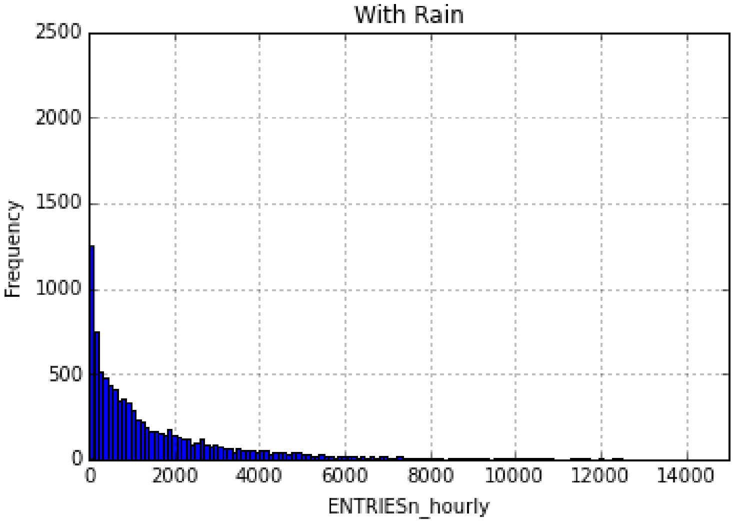

Figure 4: Histogram for rainy days

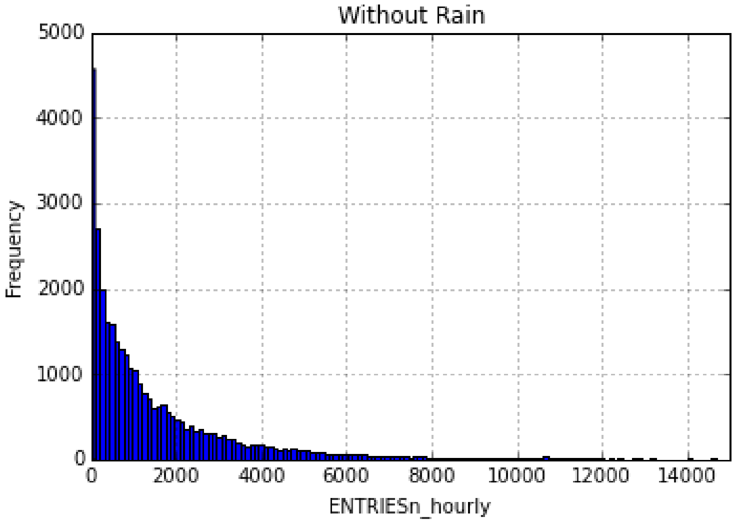

Figure 5: Histogram for non-rainy days.

2. Ridership by time-of-day

   I plotted the ridership by time-of-day for the sum of all the units in Figure 6. In the plot we can see ridership reaches its daily peak around evening (8pm), the second highest peak is at noon (12pm). This makes sense since at noon people are going out for lunch and in the evening people are going home or going out for fun.

Section 4. Conclusion

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?

   I would say more people ride the NYC subway when it is raining. If you look at the total sum, then of course non-rainy days have more total entries, that's because we have much more non-rainy days than rainy days. But if you only look at the rainy days, medium entries happen quite a lot, which means on rainy days many stations are used moderately by people. While on non-rainy days, there are a lot of low entries, which means many stations are barely used by people. That's why we have higher mean and median ridership for rainy days than non-rainy days. My interpretation would be, if there is no rain, many people would choose to walk for a short distance, but if there is rain, many people would use the subway even when the distance is short to avoid getting wet.

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.

   By looking at the Mann–Whitney U test and the means and quartiles of ridership for both rainy and non-rainy days, I concluded that there is a greater ridership on rainy days. I was able to reject the null hypothesis of the two-tailed Mann–Whitney U test, which means one population is significantly greater than the other, and since the means and all the quartiles of rainy days are greater than the ones of non-rainy days, evidently rainy days have a greater ridership.

   From my linear regression analysis, since the p-value of the coefficient for "rain" is $0.011 < 0.05$, rain does play a significant role in determining the ridership. Moreover, the coefficient estimation is 67.1850, a positive number, which means that when there is rain (rain=1), the ridership is expected to increase 67.1850 comparing to no rain (rain=0). This estimation further confirmed my conclusion that more people ride the NYC subway when it is raining.

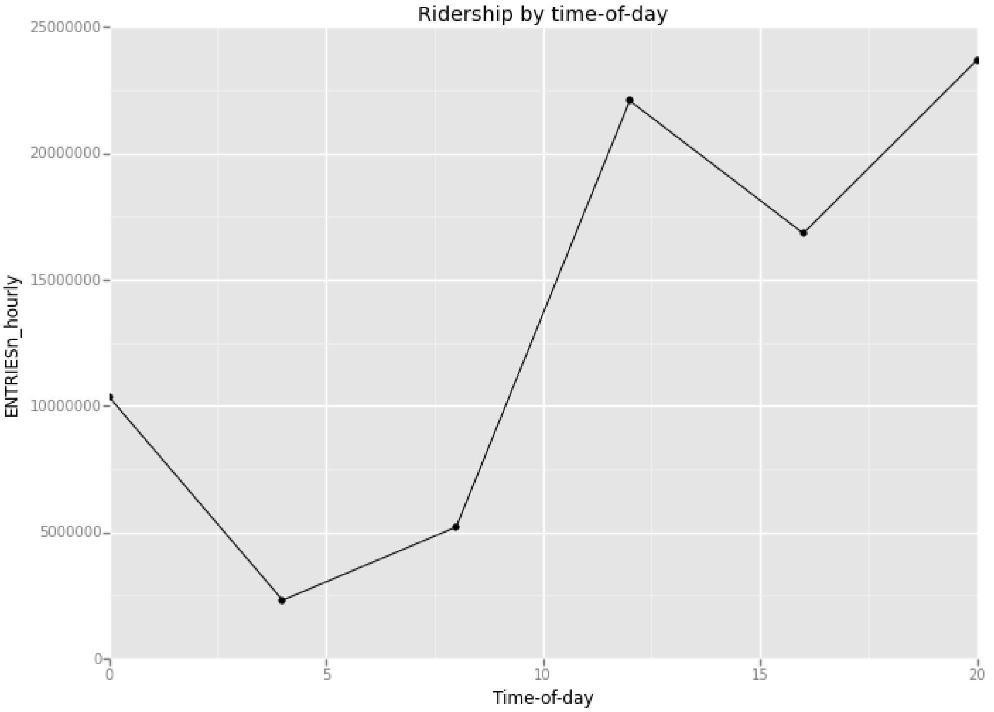

Figure 6: Ridership by time-of-day

Section 5. Reflection

1. Please discuss potential shortcomings of the methods of your analysis, including:

1 Dataset,

I think in terms of sample size, we have a decent sample size, more than 400 units at different locations and more than 40000 records totally. However, are 6 samples enough to describe the ridership of a day? Are these 6 samples the best representations of ridership of a day? These 6 samples are taken every four hours, wouldn't it be better if we take more samples within the time period when most traffic happens to describe the ridership of a day?

Also, there is heavy rain and there is drizzle, and drizzle might not affect people's travel a lot, so if we have more specific categories for rain, maybe we can gain more insight into what kind of rain will most significantly increase the ridership.

2 Analysis, such as the linear regression model or statistical test.

My linear regression model obtains an R^2 around 0.48, this is not a great result. Also, by checking the residuals I would say the ability to predict ridership using this model is not very plausible. There are a lot of large residuals and there is a cyclic nonlinear effect within the data not represented in the linear model.

I might need to do more in terms of variable selection and model structure to obtain a satisfactory model. The temporal correlation between observations and the spatial correlation between units that are close to each other apparently exist, therefore only a linear regression model might not be adequate to explain all the variation within the ridership.

Although I tried to removed variables that are intuitively correlated with "rain", I did not check the multicollinearity among the remaining variables, I just hoped that it would not be a big problem. To really make sure about it, maybe I can check the VIFs of these variables.