Questions from the NYC Subway project (Resubmission, most important new sections are underlined)

Section 0. References

Please include a list of references you have used for this project. Please be specific - for example, instead of including a general website such as stackoverflow.com, try to include a specific topic from Stackoverflow that you have found useful.

-Course notes
- my programs from the "Introduction to Computer Science" course from Udacity,
- stack Overflow for various google searches that I didn't document since I didn't know I was keeping track of references. I'll keep a better record for the next course
- Udacity downloadable pdf for data analysis for information on the Mann Whitney test and function call.
- http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm for residual analysis.
- https://pypi.python.org/pypi/ggplot/ for ggplot info
- http://www.itl.nist.gov/div898/handbook/prc/section1/prc131.htm for statistical test details

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 used the Mann Whitney test to determine if the number of subway entries was likely to be larger when it was raining versus when it wasn't raining. I used a 1 tail P value as I assumed people would want to stay out of the rain and ride the subway versus waiting for a cab or walking a few block when it was raining versus when it wasn't. Thus my alternative hypothesis was not that the entries would be statistically different (which would have used a two tailed test), but was that rainy day hour entries would be greater than non-rainy days

H0: Prob(subway entries when raining > subway entries when not raining) = .5
(in words the probability that entries when raining are more than when not raining is ½ , meaning there is also a ½ chance the entries when raining is not greater than when not raining)

H1: Prob(subway entries when raining > subway entries when not raining) > .5

The P-critical value or alpha used for this test was .05. This means that we will choose to reject the null hypothesis even with a 5% chance we are making a false decision

The test produced a p-value = 0.025 meaning that we can reject the null hypothesis for a p-critical value of .05. This means that with a random draw of these populations of entries with rain and entries without rain there is only a 2.49% chance that the null hypothesis was actually true and I obtained this data by just an unlucky random draw.

The mean value of entries during the rain is 1105.4 and the mean values without rain is 1090.3 and thus follows the conclusion that the subway entries during during rain are higher.

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 test is a non-parametric test that makes no assumptions about the underlying population distributions of entries when raining or when not raining. However, the P value was estimated assuming a normally distributed U statistic. This assumption is easily met with greater than 20 values from the raining set and 20 values from the non-raining set which is a condition of the assumption

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.

p-value = 0.025 and our p-critical value was .05 which means that at this 95% confidence level we can reject the null hypothesis with the given subway data. The Null hypothesis restated is that the probability that subway hourly entries when raining are more than when not raining is just 50-50; meaning it is just as likely that the number of entries when not raining will be higher than raining. Our data has shown we can reject this hypothesis and that there is an increase in hourly ridership during the rain (in May 2011 see analysis below)
The mean value of subway entries during the rain is 1105.4 and the mean values without rain is 1090.3

1.4 What is the significance and interpretation of these results?

At a 95% significance we can reject the null hypothesis. We can say with significance from our data that for a random choosing of subway entries when raining versus when not raining, the probability that entries are larger when raining is not = .5 but larger and thus it is more likely that hourly entries will increase on the NYC subway when it is raining in May, 2011.
See my discussion of the problem with this dataset. Basically we only have weather data for 30 days in May 2011. We are not random sampling data from a variety of seasons and years but have all the data from 1 month sampled once per day. We shouldn't extrapolate our results to other months, seasons or other years. This dataset is very limited

1.5 other interesting statistical tests I experimented with.

Mann whitney test on fog with a p-critical value of .01 rejected the null hypothesis also. and the means were separated even more from the rain condition
1154.65934963 1083.44928209 (means with fog and without
(1189034717.5, 1.9570617095483498e-06) (U, 1 sided p-value)

Mann whitney test on windspeed > 4 gave an even stronger result.
1170.78426415 1087.06689768 (means with windspeed >=4 and with windspeed < 4)

(752894357.0, 1.8287181009848463e-09)) (U, 1 sided p-value)
(Not sure how the means of both groups could be higher than the test above??)

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:

1. Gradient descent (as implemented in exercise 3.5)
2. OLS using Statsmodels
3. Or something different?

I used the gradient descent method as implemented in 3.5

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

I used the variables ‘Hour', 'meanwindspdi', 'meanpressurei', and 'meantempi'’ as inputs to my model for predicting the number of subway entries per hour. I used all the UNITs as dummy variables. Using the dummy variables was very important to the predictive power of the model as hourly subway ridership is heavily dependent on the subway location in the city; i.e midtown stations always have more entries than the last station on the line. Thus using the UNIT as a dummy variable allowed the linear regression model to predict the change from the mean value at a particular subway station and the UNIT variable could just be assigned the mean value by the regression model. If you pull the dummy variable out of the model it fails miserably in trying to model these different mean rider levels between subway stations based on weather variables. I pulled the dummy variables out of the model originally and got a R2 value of about .03.

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.

My hypothesis was that bad weather would cause riders to stay below ground more and increase ridership. I ran a few tests with the variables individually and found that the “hour” variable has the most predictive power by far providing an R2 = .4631 itself.
Here’s some of my R2 findings:
Dummy variables alone: .4249
Dummy + windspeed: .4258

Dummy + hour : .4631

Dummy + precipi: .4251 (noticed precipitation or rain gives the lowest gain over just the dummy variables)

Thus I surmised the hour variable was the most important and that seems reasonable as going to work 8-9 AM, lunchtime 11-1 and going home from work 4-6 would seem to drive the most subway ridership. Also the weather data is only sampled once per data and not hourly. So I started with the hour variable and added to it one variable at a time and looked at the power of the prediction through R2. R2 measures the percentage of variance of the signal (entries per hour) that can be modeled by the linear predictive model.

I found 'meanwindspdi', 'meanpressurei', and 'meantempi'' to be the next largest contributors in that order.

This fits with my hypothesis that bad weather conditions or heavier winds or colder temperatures would encourage more people to take the subway. Interestingly, “precipi” or the measured rain in inches didn’t add any predictive power as measured by R2 so I dropped it from the model.

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

The coefficients for my non-dummy variables for ‘Hour', 'meanwindspdi', 'meanpressurei', and 'meantempi' are:

Hour: 458.785
Meantempi: -48.280
Meanpressurei: -43.025
Meanwindspdi: 50.037

Notice that the coefficient for the hour variable is an order of magnitude larger than the others and contributes most to the model. However, the linear relationship between hour and ridership isn’t logical as the hourly entries doesn’t grow linearly with the hour of the day (see reflection section and the visualization sections below).

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

R2 = .4646

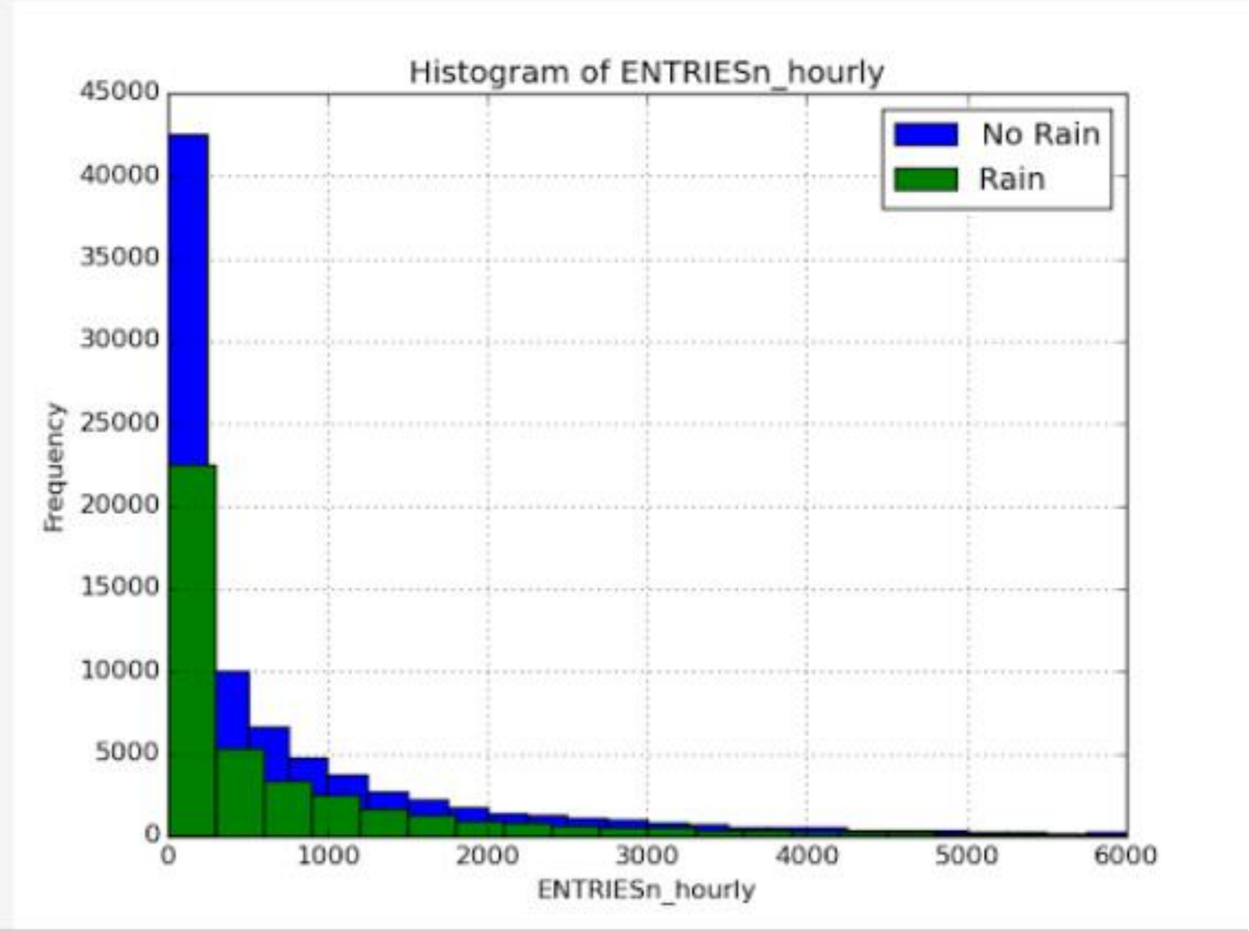
2.6 What does this R2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R2 value?

R2 measures the percentage of variance of the signal (entries per hour) that can be modeled by the linear predictive model. Thus my model above accounts for 46% of the variance of ridership from the mean value.

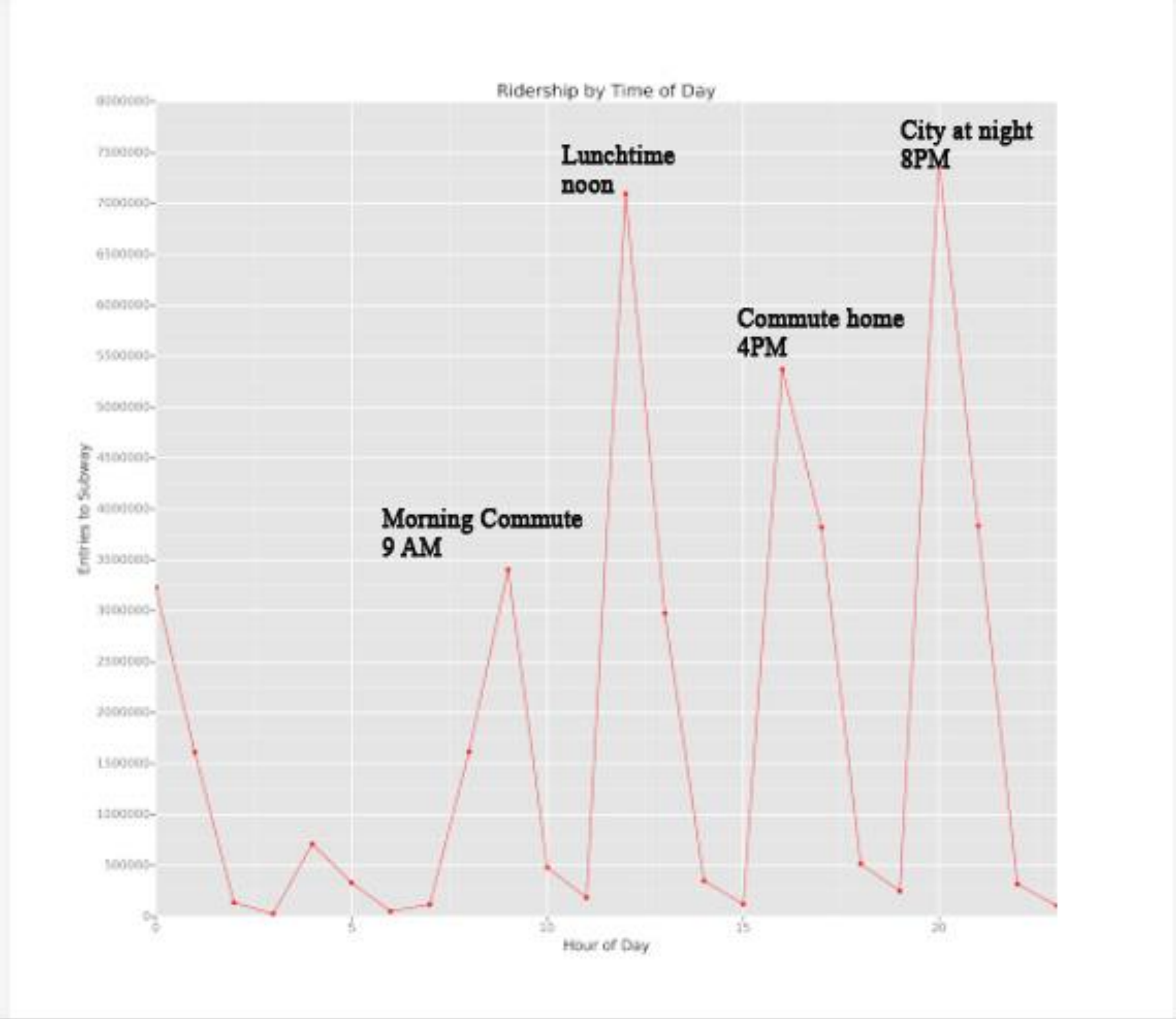
The R2 value isn't very good so the predictive power is not great. However, we plotted the residual in the next assignment and certainly the error signal from the residual was pretty close to a zero mean normally distributed distribution, so I would say that the linear model is reasonably appropriate for the dataset just and that there are other variables besides weather that contribute to the variance of ridership levels that we don't have access to. Also we are trying to predict "hourly" entries to the subway while the only variable that changes hourly is "hours". The weather data in the dataset I downloaded only changes once per day. Thus from that perspective the model is pretty weak.

Section 3 Visualizations: (added plots)

1.)



While the histograms have generally the same shape, the histogram of entries during rainy days is shifted slightly to the right on the curve indicating a slightly higher mean. There also seems to be over twice as many hourly entries for rainless days than for rainy days



Ridership by the hour. Note that no linear relationship exists between these two variables as might be suggested by the regression model. See reflections below:

Section 4. Conclusion

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.
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?

I can't say whether more people ride the NYC subway when it is raining from the data given. A statistical test requires a random sampling of the population we are trying to make inferences on. If we are trying to make such inferences we would need more than 30 datapoints of weather versus ridership for one month of one year (May 2011). Given our limited data I would be able to make statements about ridership levels in May 2011 only. Given the data and the results from the Mann Whitney test described above, with a p-critical value of .05, I would say that more people ride the subway when it is raining then when it is not rainy in May 2011.
As a side and from 1.5 above I can make the same statement about fog and windspeed >=4 as I can make about rain, only with greater confidence as I can reject the null at a p-critical value of .01 for these tests

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.

The Mann Whitney test assumptions about the underlying population distribution and sample numbers were all satisfied. The one sided P-value was .025; meaning that with a p-critical value of .05, I can conclude that there is likely to be more riders on rainy days than non-rainy days in May of 2011.
My linear regression model did not show a strong relationship between rain and hourly ridership. I ran the linear model with 1 feature: hour and got an R2 of .4630. When I added rain as a features the R2 rose ever so slightly to R2=.4631. When I constructed my model in section 3, I didn't even use rain as a feature as the other weather parameters: windspeed, meanpressure (indicating rain was coming), temperature provided a higher R2 of .4646

Section 5. Reflection

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.
5.1 Please discuss potential shortcomings of the methods of your analysis, including:

1. Dataset,

I'm glad I had the opportunity to review all the programs I wrote in this unit and play with the results some more. As mentioned above, I noticed in scrolling through the turnstile data that the only parameters that changed by the hour were "hour" and exists and entries. Thus we are predicting hourly entries with data that doesn't change hourly. Looking closer I realized we only had 30 days of weather data and all the data was from the same month, woefully inadequate to extrapolate and make predictions about ridership levels based on weather for all seasons and months. I would think that if we wanted to isolate the variability of ridership based on weather and we could only get weather readings once a day, then we should only predict the ridership levels on a day basis and not on an hourly basis. We could see that in the prediction of hourly riders was largely based on hour with a tiny amount of prediction from the weather parameters.

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

In continuing with the theme above, in order to see how meaningful the relationship is between weather and ridership levels, the entries should be averaged for an entire day if weather readings are only acquired once per day. If weather could be acquired 4 times a day then the ridership levels should be found on that same interval. Also, weather varies tremendously over the course of a year in New York City, thus weather data would be needed for at least the 4 different seasons to see a meaningful relationship in the linear prediction model between ridership and weather parameters

The linear model used for predicting hourly entries is also questionable. For instance, the major feature producing the R2 of .4630 out of a total of R2 =.4646 was the hour of the day feature. From the graph above we can see that hourly entries is not linear with respect to hour. The graph doesn’t show a steady increase or decrease as hour increases during the day. Instead the graph is basically multi-modal showing 5 peaks during the day, morning commute, lunchtime, evening commute, and nightlife trips after dark in the city. Thus there is not a linear relationship between hour and hourly entries and this leads to the weakness in using this linear regression model to predict hourly entries

5.2 (Optional) Do you have any other insight about the dataset that you would like to share with us?