Project P1

Short Questions

answered by Mark Cassar

Overview

This project consists of two parts. In Part 1 of the project, you should have completed the questions in Problem Sets 2, 3, 4, and 5 in the Introduction to Data Science course.
This document addresses part 2 of the project. Please use this document as a template and answer the following questions to explain your reasoning and conclusion behind your work in the problem sets. You will attach a document with your answers to these questions as part of your final project submission.

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 U Test to analyze the NYC subway data. I used a two-tail P value (doubling the value return by the scipy.stat.mannwhitheyu() function). The null hypothesis in this instance is that there is no difference in the means for the number of hourly subway entries when it is raining compared to when it is not raining. I used a P-critical value of 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.

> This statistical test is applicable because the distributions of the two samples (rain vs no-rain) are not normally distributed, are independent of each other, and are ordinal (meaning they can be put into rank order).

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.

> mean (rain) = 1105.45
> mean (no rain) = 1090.23
> U = 1924409167.0
> P (one-sided) = 0.025 (was actually just under: 0.0249999)

1.4 What is the significance and interpretation of these results?

This result means that we can reject the null hypothesis at the 95% confidence level. That is, the difference in the means is statistically significant (the means come from two different distributions) and is not due to random variations.

Section 2. Linear Regression

2.1 What approach did you use to compute the coefficients theta and produce predictions for ENTRIESn_hourly in your regression model:

I experimented with Scikit Learn’s LinearRegression, Support Vector Machine, and Random Forest, but I ended up using statsmodels OLS method to compute the coefficients.

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

The final features for my model were: fog, pressurei, tempi, wspdi, precipi, rain, meantempi, menwspdi, meanprcipi, and hour.

The dummy variables I included were: units, day_week, and conds.

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.

In order to choose the features I used the following method. The code for this can be found in file ‘turnstile_weather_predict_svm.py’.

Step 1
Starting with the dummy variable units, day_week, station, and conds, I ran through my full feature list (hour, tempi, meantempi, pressurei, meanpressurei, precipi, meanprecipi, rain, wspdi, meanwspdi, fog, latitude and longitude) adding one at a time and then calculating R^2. I then selected the feature that gave the highest R^2 value. Using this single feature, I then ran through all combinations of the dummy variables and chose the combination giving the highest R^2 value.

Step 2.
I then started adding a feature at a time from the remaining features, choosing each time the next feature that gave the highest R^2, and continued adding features in this fashion until R^2 decreased, at which point I stopped. The table below shows the progression of feature selection:

Step 1 (all include ‘hour’ feature)	**R^2**
baseline (dummy variable units, station, day_week, conds)	0.45868
units	0.45883
day_week	0.10754
station	0.40773
conds	0.09137
units + day_week	0.48487
units + station	0.45835
units + conds	0.46389
day_week + station	0.43366
day_week + conds	0.11387
station + conds	0.41283
units + day_week + station	0.48420
units + day_week + conds	***0.48765***
units +station + conds	0.46336
day_week + station + conds	0.43644

Step 2 (all include units, day_week, and conds)	**R^2**
baseline (hour)	0.48765
hour + meantempi	0.48926
hour + meantempi + tempi	0.493831
hour + meantempi + tempi + pressurei	0.494093
hour + meantempi + tempi + pressurei + meanprecipi	0.494268
hour + meantempi + tempi + pressurei + meanprecipi + precipi	0.494368
hour + meantempi + tempi + pressurei + meanprecipi + precipi +rain	0.494487
hour + meantempi + tempi + pressurei + meanprecipi + precipi +rain +	0.494547

wspdi	
hour + meantempi + tempi + pressurei + meanprecipi + precipi +rain + wspdi + meanwspdi	0.494664
hour + meantempi + tempi + pressurei + meanprecipi + precipi +rain + wspdi + meanwspdi + fog	***0.494683***
hour + meantempi + tempi + pressurei + meanprecipi + precipi +rain + wspdi + meanwspdi + fog + latitude	0.494678

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

The coefficients for the non-dummy features are:

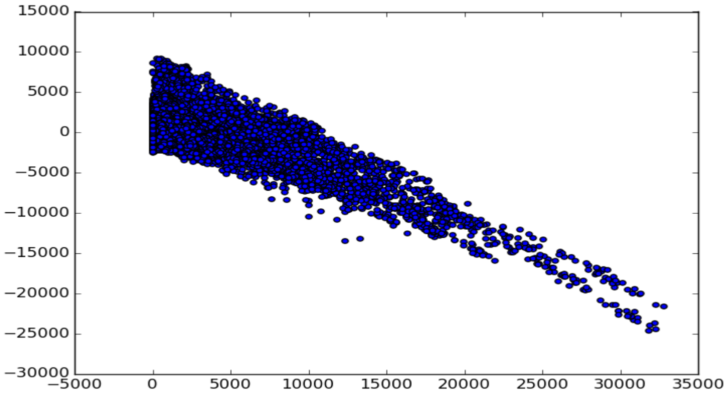
Feature	Coefficient
fog	-20.5212
pressurei	-59.7924
tempi	387.7800
wspdi	55.9091
precipi	-87.9870
rain	-52.5932
meantempi	-452.4107
meanwspdi	-65.0782
meanprecipi	107.0991
hour	728.2143

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

The model's R^2 value is 0.495.

2.6 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?

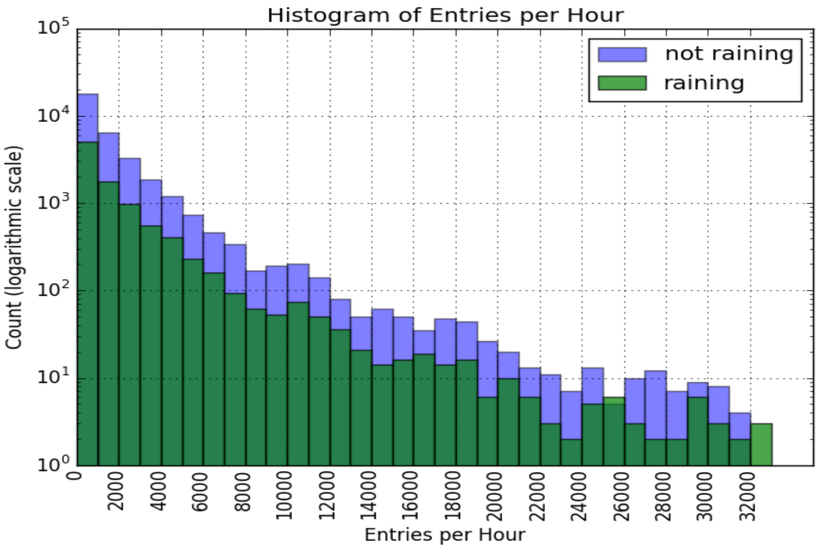
The model fit does not seem to be very good as the plot of the residual against fitted value (see below) does not show even scatter about 0. The model seems to overestimate for smaller values of hourly entries and underestimate for larger ones. The error in the estimates also follow a decidedly decreasing linear trend.



Section 3. Visualization

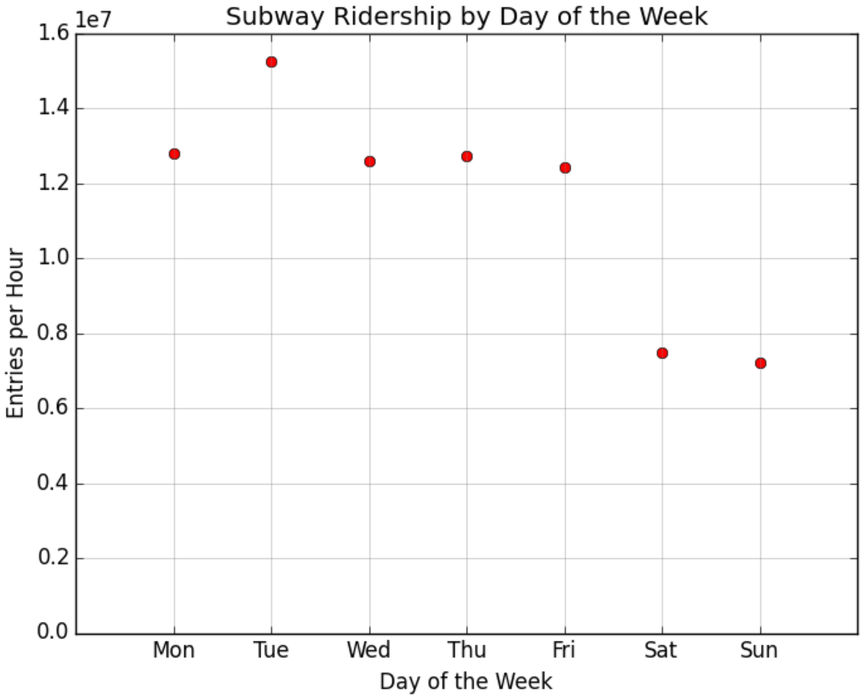
Please include two visualizations that show the relationships between two or more variables in the NYC subway data. Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

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



The above plot shows how many times the specified entries per hour occurred in the given dataset. First note that I have used a logarithmic scale along the y-axis. This allows us to see more easily the entire distribution, particularly for higher entries per hour. The drawback to this is that the differences between the counts for days with rain compared to those without are visually much closer than they actually are. The two main things I would point out are: the pattern of counts is very similar between the two distributions and, as one would expect, the counts for days without rain are much higher than those with rain. The code for this plot can be found in file ‘PS_3_1_project_1.py’.

3.2 One visualization can be more freeform. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like.



The above plot shows the total entries per hour based on the day of the week. As one would expect, the entries per hour are higher for the weekdays, which could be explained by the fact that the typical workweek is Monday through Friday and that the subway is used as a mode of transportation to and from work. The other interesting thing to note is the increase in subway use on Tuesdays. I would have expected Monday through Friday to show no meaningful differences. It would be interesting to explore this a little more. The code for this plot can be found in file ‘PS_4_1_project_1.py’.

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?
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.

Based upon my analysis, I would conclude that the mean number of people riding the subway per hour is higher on any given day when it is raining than when it is not. However, in terms of absolute numbers of riders, more people ride the subway when it is not raining. This is a result of the means being close but but the ratio of days without rain to those with rain is about 2 to 1. These conclusions are only valid for the time period covered by the dataset, which is May 2011.

I am basing these conclusions on the applicability of the Mann-Whitney U test (valid assumptions noted above in Section 1) in this scenario, the histogram noted above (and its non-logarithmic counterpart), and basic calculations done on the data (sum of ENTRIESn_hourly for days with rain and without, etc.).

In its current guise, I do not believe that the regression analysis can be trusted. To point to one reason for being skeptical, the coefficients for the dummy variables, in particular, are large with even larger standard deviations. The results seem to point to linear dependence in the feature matrix.

Section 5. Reflection

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

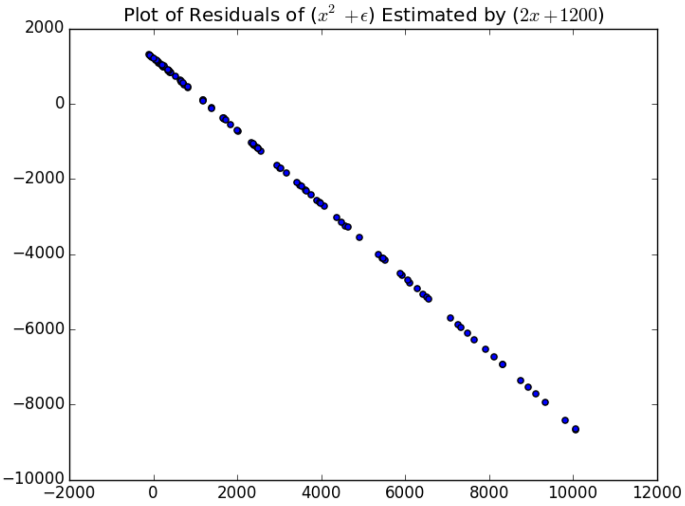
5.1 Please discuss potential shortcomings of the methods of your analysis, including:

Dataset,

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

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

There seems to be two problems with the linear regression model: linear dependence of the feature matrix and possible non-linear relationship between ENTRIESn_hourly and the features. However, while adding non-linear features (e.g., hour*hour, hour*pressure) or doing the regression with ENTRIESn_hourly replaced with log(ENTRIESn_hourly) seems to increase the value of R**2, it does not get rid of the large coefficients with larger standard deviations. The existence of nonlinearity stems from the fact that I can recreate a similar residuals plot by estimating a quadratic function with a linear one, as seen in the plot below (the code for this plot can be found in file ‘over_under.py’):



By restricting the feature set so that it does not include any of the dummy variables, I can get reasonable coefficients for the model (i.e., coefficients between about -10 and 10 with standard deviations at the 0.04 level). In doing so, however, R**2 hovers at about 0.2. But these does not correct the lack of randomness in the residual plot, which seems to be a more serious issue than a low R**2 value.

As far as the dataset goes, the issues I see are: (1) only one month of data is covered, (2) many of the features have little variability across the entire dataset, and (3) some of the features are definitely correlated. Issue (1) could be a problem because rain has seasonal variations and these are not accounted for in the dataset. It would be better to have data that covered March through October. Issues (2) could be a problem because we are trying to use these features to predict an outcome but they contain little information and that information may be uncorrelated with the outcome, thereby increasing the level of noise the model is trying to sort through (e.g., fog is a feature that has very little variation). Issue (3) is a problem, and is most likely the explanation for the linear dependence mentioned above, as multiple features are essentially encoding the same information (e.g., unit and station and [latitude, longitude], etc.).

https://statistics.laerd.com/premium-sample/mwut/mann-whitney-test-in-spss-2.php
http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test
http://stackoverflow.com/
http://www-psychology.concordia.ca/fac/kline/601/osborne.pdf