Revised 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. The null hypothesis in this instance is that there is no difference in the distribution of the number of riders when it is raining (X) compared to the distribution of riders when it is not raining (Y). This means that the probability of a randomly chosen value from X that is larger than a randomly chosen value from Y is 0.5, that is, the same probability of getting heads on a coin toss. 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.

> The Mann-Whitney U test is applicable because the distributions of the two samples (rain vs no-rain) are not normally distributed, and therefore, tests like the t-test will not work. To test whether or not we can reject the null hypothesis as stated in question 1.1, the Mann-Whitney U test makes no assumptions about the distribution of ridership in the two samples. Some assumptions that this test does make are that the observations are independent of each other and are ordinal (meaning the observations in the samples can be put into rank order). The Mann-Whitney U test depends explicitly on the observations being ordinal as the U statistic is calculated based on sums of the ranks of the values in each sample.

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.

Using Scipy to carry out the Mann-Whitney U test in python returned ‘nan’ for the p-value. As a consequence, I used R’s equivalent wilcox.test(x,y) to carry out this test. The code for this test can be found in the file ‘mannwhitneytest.R’. The test results and associated means are:

mean (rain) = 2028
mean (no rain) = 1845
U = 153635121
P (two-tailed) = 5.5x10^(-6)

1.4 What is the significance and interpretation of these results?

This result means that we can reject the null hypothesis above the 95% confidence level. That is, the difference in the distributions is statistically significant (the means come from two different distributions).

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 modules and with statsmodels OLS method to carry out the linear regression. In the end, I chose to use Scikit Learn’s LinearRegression method.

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

My final model included the features 'hour', ‘meantempi’, and ‘tempi’. The dummy variables that I included in this model were derived from the 'UNIT', ‘day_week’, and ‘conds’ variables. This resulted in a total of 260 features.

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 the file ‘feature_selection.py’ and ‘final_model.py’.

Step 1
I first ran through all combinations of the dummy variable units, day_week, station, and conds, noting the value of R^2 and checking to see if the model coefficients were very

large. I arbitrarily chose ‘units’ as the baseline against which to compare each of the dummy variable combinations. I then selected the feature that gave the highest percentage increase in R^2 over the baseline. This is noted in bold in the table below.

To avoid collinearity in the dummy variables, however, I excluded the *n*-th variable in each instance. So, for ‘units’ I removed the feature 'unit_R464', for ‘day_week’ I removed ‘day_6’, and for ‘conds’ I removed ‘conds_Scattered Clouds’. The reason for this is that for any n dummy variables, at least one will be linearly dependent on the others. This is easy to see if we consider a dummy variable based on gender. In such a scenario we do not need to include both 'male' and 'female' as dummy variables as a '0' in the female column is identical to a '1' in the male column.

Step 1(dummy variables)	**R^2**	**Large Coefficients**	**Percent increase over baseline**
units (baseline)	0.375	No	--
day_week	0.025	No	-93.3
station	0.324	No	-13.6
conds	0.010	No	-73.3
units + day_week	0.401	No	6.9
units + station	0.375	**Yes**	0
units + conds	0.381	No	1.6
day_week + station	0.350	No	-6.7
day_week + conds	0.032	No	-91.5
station + conds	0.330	No	-12.0
units + day_week + station	0.401	**Yes**	6.9
units + day_week + conds	***0.405***	***No***	***8.0***
units +station + conds	0.381	No	1.6
day_week + station + conds	0.354	No	-5.6
units + day_week + station + conds	0.405	**Yes**	8.0

Step 2.
I then started adding a feature at a time from the remaining features, choosing the feature combination that gave the best increase over the baseline of including the single feature, ‘hour’. In this step I also excluded the features ‘latitude’ and ‘longitude’ as they consistently introduced very large coefficients into the model. For each set of features I have noted if the model coefficients became very large or not. As there was little to no gain in R^2 after adding the first 3 features, I decided that that would be the best choice for the final feature set. The table below shows the progression of feature selection:

Step 2 (all include units, day_week, and conds)	**R^2**	**Large Coefficients**	**Percent increase over baseline**
hour (baseline)	0.488	No	--
hour + meantempi	0.490	No	0.4
hour + meantempi + tempi	***0.494***	***No***	***1.2***
hour + meantempi + tempi + pressurei	0494	No	1.2
hour + meantempi + tempi + pressurei + meanprecipi	0.494	No	1.2
hour + meantempi + tempi + pressurei + meanprecipi + precipi	0.494	No	1.2
hour + meantempi + tempi + pressurei + meanprecipi + precipi +rain	0.495	No	1.4
hour + meantempi + tempi + pressurei + meanprecipi + precipi +rain + wspdi	0.495	No	1.4
hour + meantempi + tempi + pressurei + meanprecipi + precipi +rain + wspdi + meanwspdi	0.495	No	1.4
hour + meantempi + tempi + pressurei + meanprecipi + precipi +rain + wspdi + meanwspdi + fog	0.495	No	1.4
hour + meantempi + tempi + pressurei + meanprecipi + precipi +rain + wspdi + meanwspdi + fog + meanpressurei	0.495	No	1.4

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
hour	107.45
tempi	48.19
meantempi	-66.79

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

The model's R^2 value is 0.494.

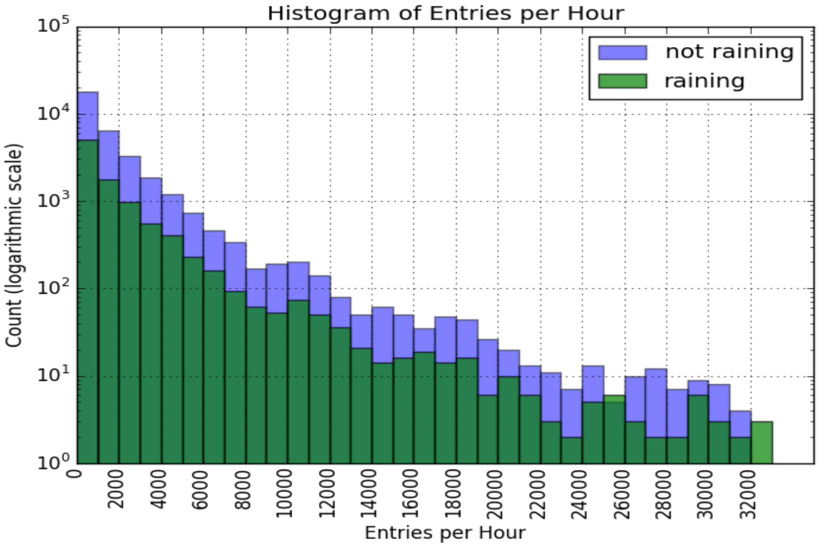
2.6 What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?

An R^2 value of 0.494 says that the model is able to explain just over 49% of the variance of the value we are trying to predict, which, in this case is 'ENTRIESn_hourly'. The model seems to provide an ok fit for the dataset. The main reasons for not thinking that this is a good fit is that the model produces many predictions below zero, which is unreasonable, and that some of the predictors seem to have a nonlinear relationship with the target values of 'ENTRIESn_hourly'.

Section 3. Visualization

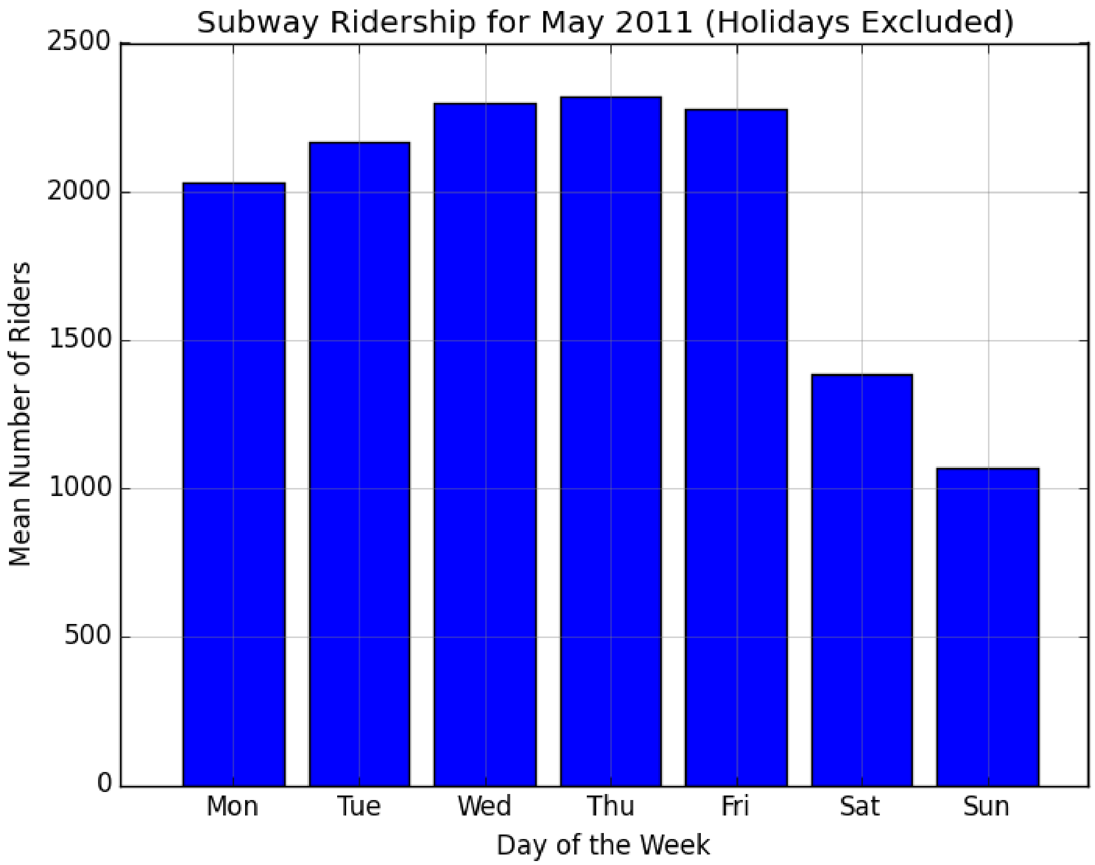
Please include two visualizations that show the relationships between two or more variables in the NYC subway data. Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

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 ‘entries_histograms.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 average number of subway rider entries based on the day of the week. As one would expect, the entries 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. There is a slight increase from Monday through Thursday with an expected large drop going into the weekend. The slightly lower value on Monday could be due to people taking time off work, as May could be considered the start of the summer season. This would have to be borne out by looking at similar data throughout the year. The code for this plot can be found in file ‘entries_day_week_plot.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 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 are a couple potential shortcomings with the methods of this analysis. The first is that the model produces a significant number of negative predictions. Such results are nonsensical give the predicted value is meant to represent subway ridership, which has a natural lower bound of zero. The second is that the current model does not account for the fact that at least a few of the features seem to be correlated with ridership in a nonlinear fashion.

As far as the dataset goes, the issues I see are: (1) only one month of data is covered, and (2) many of the features have little variability across the entire dataset. 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).

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
http://stats.stackexchange.com/questions/65405/what-to-do-when-a-mann-whitney-u-assumption-is-violated?rq=1
http://discussions.udacity.com/t/mann-whitney-test-what-does-nan-mean/16373/5
http://www.statmethods.net/stats/nonparametric.html
http://www.algosome.com/articles/dummy-variable-trap-regression.html
http://en.wikipedia.org/wiki/Coefficient_of_determination#Inflation_of_R2
http://en.wikipedia.org/wiki/Multicollinearity