Udacity Data Analyst Nanodegree Project 1 Short Answers

Brian Hurn
April 1, 2015

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.

Stackoverflow.com: Python basics (loops, conditionals, etc.), Pandas dataframe basics, numpy array handling, ggplot options, error message lookup
docs.python.org: Python basics ((loops, conditionals, etc.), datetime library functions
Graphpad.com: *Interpreting Results: Mann-Whitney Test*
Scipy.org: *scipy.stats.mannwhitneyu* reference
StatsModels.sourceforge.net: *statsmodels.regression.linear_model.OLS* reference
Wikipedia: Ordinary Least Squares, Mann-Whitney U Test
The Minitab Blog: *Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit?*
Bespoke Blog: *Plotting with Matplotlib*
matplotlib.org: *Pyplot Tutorial*

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

To determine if there was a significant difference between the distributions of ridership (entries) in hours with rain and without, I used the Mann-Whitney U test. This test is applicable because the two distributions are not normal. We can use this test to help determine if two distributions are identical or not. In this case, I used a two-tailed test given that the null hypothesis was that the distributions are the same. One could make an argument for using a one-tailed test because the distributions appear to have similar shapes and have a lower bound of zero on one side, however.

My p-critical value is 0.05. This value corresponds to a significance level of 95%, which is typically used in research and across a range of industries. Because the scipy function returns a one-sided value, I doubled the returned value of 0.0249999 to obtain a p-value of 0.049998. Because p is less than p-critical, we can reject the null hypothesis and conclude that there is a statistically significant difference in the with-rain and without-rain distributions.

The means were 1105.45 (with rain, n = 44,104) and 1090.28 (without rain, n = 87,847), so it appears that the positive effect of rain on ridership is rather small (just over 1.3%) on average in this particular data set.

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:

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

2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your 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. Your reasons might be based on intuition. For example, response for fog might be: “I decided to use fog because I thought that when it is very foggy outside people might decide to use the subway more often.” Your reasons might also be based on data exploration and experimentation, for example: “I used feature X because as soon as I included it in my model, it drastically improved my R^2 value.”

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

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

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?

For the linear regression problem, I found that an Ordinary Least Squares (OLS using Statsmodels) approach yielded slightly better results than gradient descent. I used the following input variables: a constant, 'rain', 'meanwindspdi', 'Hour', 'Day' (the day of the week calculated from the date), the square of “meantempi’, and a dummy variable for the ‘UNIT’.

I selected these features for the following reasons:

constant: The coefficient for the constant term essentially acts as a starting point for the model.

rain: This variable had slight positive impact on ridership (as determined in the statistical test exercise).

meanwindspeedi: I hypothesized that this variable served as a measure of weather severity or unpleasantness, and it delivered higher R^2 values.

Hour: It is a natural assumption that ridership varies over the course of the day.

Day: I created this variable that contained an integer (0-6) representing the day of the week to account for variations in traffic between the days of the week (weekends vs weekdays and variations therein)

meantempi: I used the square of the mean temperature to accentuate the differences in temperature. Squaring this value only improved R^2 slightly, however.

UNIT: By accounting for differences in entries across units, the model became significantly more accurate. With a larger dataset (across additional time periods), one could develop a model per unit. Each of these models would be more accurate because the other variables in the model may not behave uniformly across units. For example, riders in more affluent sections of the city may take a cab on rainy days or pursue other options not available to those of lesser means.

The coefficients of my model are shown in the second column of the table below.:

```
                   coef    std err          t      P>|t|      [95.0% Conf. Int.]
--------------------------------------------------------------------------------
const         1698.7705    140.919     12.055      0.000      1422.538  1975.003
rain           -60.1270     37.800     -1.591      0.112      -134.224    13.970
meanwindspdi   -18.4111     10.326     -1.783      0.075       -38.653     1.831
Hour            62.4949      2.474     25.256      0.000        57.644    67.345
Day            -92.0134      9.396     -9.793      0.000      -110.431   -73.596

meantempi^2     -0.0988      0.021     -4.656      0.000        -0.140    -0.057
```

There is one surprising finding here. The coefficient for the rain variable is a negative number (-60.127). This indicates that the model calculates a lower predicted value for hours with rain (‘rain’ = 1) than it does for hours without rain (‘rain’ = 0) if all other factors are equal. In other words, in a model with this particular set of input variables, the model’s accuracy improves with a negative coefficient for ‘rain,’ but it may not be a true reflection of the contribution that it makes to the actual results. As we saw in the previous section, the average for hours with rain is actually higher than that for hours without.

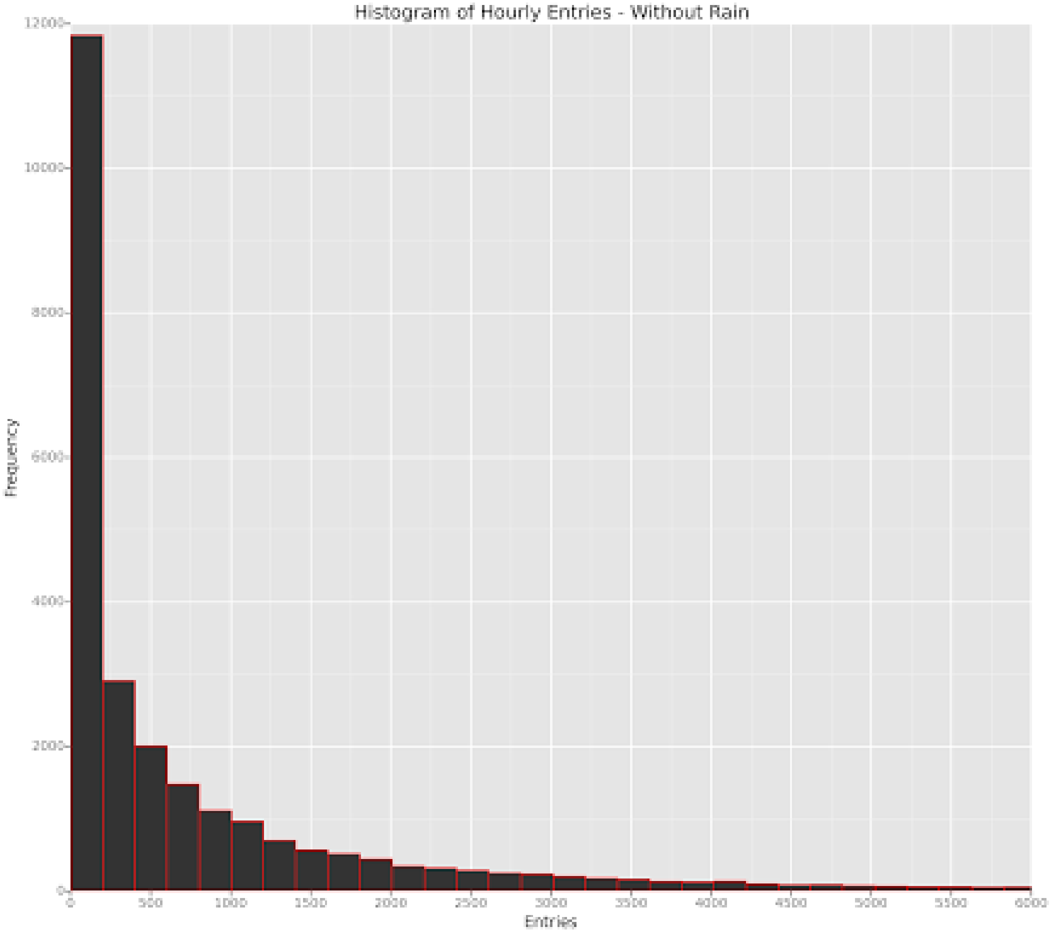
In an alternative model that I created that uses only a constant, ‘rain,’ and ‘UNIT’, the coefficient for ‘rain’ is positive (58.487), indicating that rain, when considered alone, indeed has a positive effect on ridership. The divergent results from these two models casts doubt on the ability of predictive modeling alone to generate specific conclusions regarding the contributions of individual variables, especially when one considers models with multiple inputs and relatively low coefficients of determination (R^2).

The R^2 value of my model is 0.489. This represents a decent result considering that the model is attempting to predict human behavior. The model would have limited value, however, in a scenario where precise revenue projections or resource requirements (security or janitorial personnel scheduling, for example) were required. To further refine the model, I would try to find additional relevant input variables and seek data from larger time periods to build unit-specific and seasonal models. The alternative model mentioned above that only uses 'rain' and 'UNIT' as input variables had a lower R^2 value of 0.448.

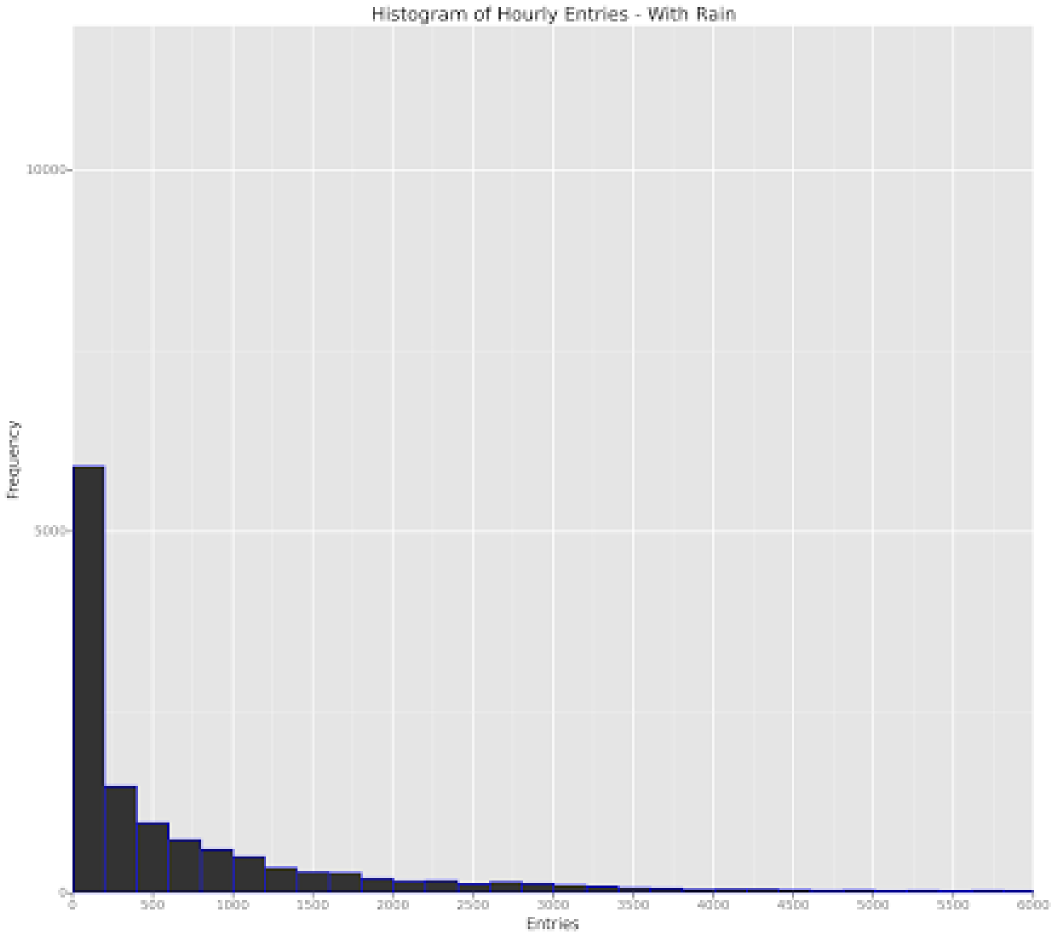
Section 3. Visualization

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.
You can combine the two histograms in a single plot or you can use two separate plots.
If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.
For the histograms, you should have intervals representing the volume of ridership (value of ENTRIESn_hourly) on the x-axis and the frequency of occurrence on the y-axis. For example, each interval (along the x-axis), the height of the bar for this interval will represent the number of records (rows in our data) that have ENTRIESn_hourly that falls in this interval.
Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.
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. Some suggestions are:

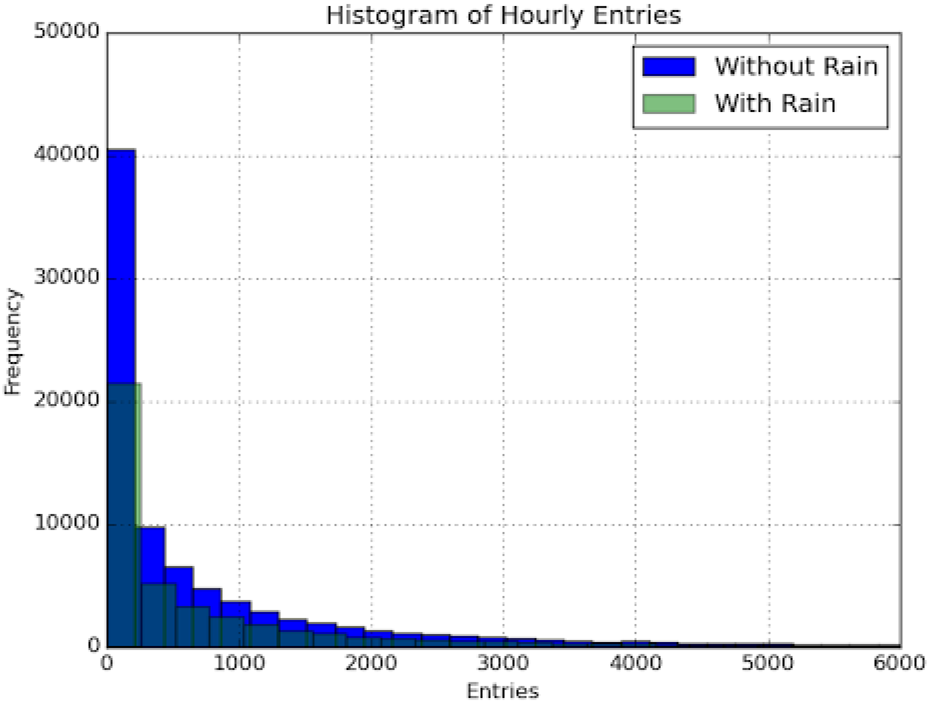
Ridership by time-of-day
Ridership by day-of-week



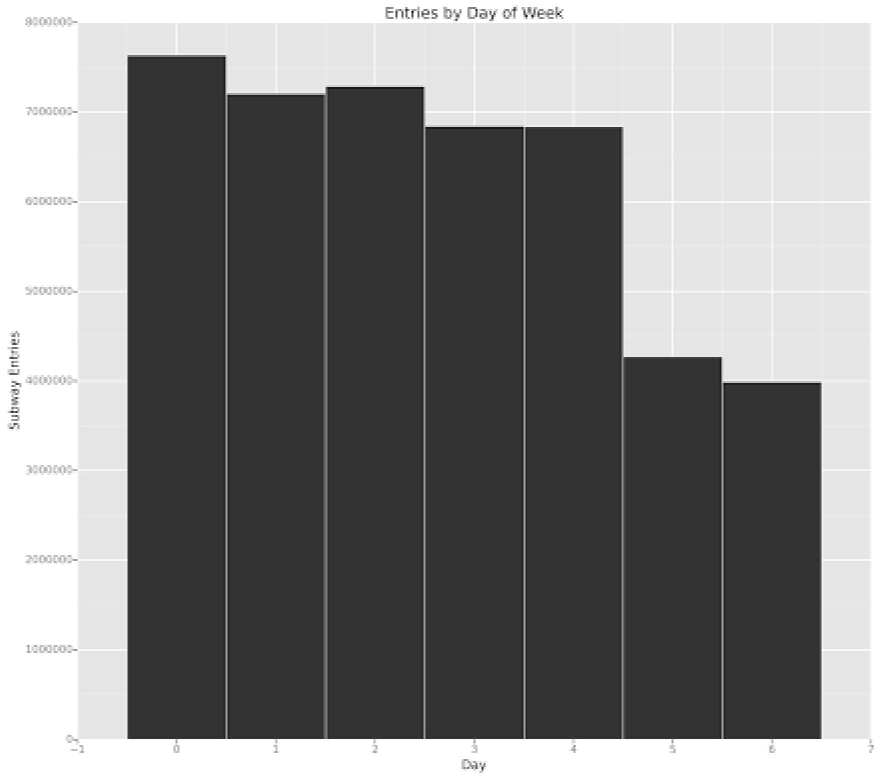
The chart above shows the ggplot histogram of ENTRIESn_hourly (from Problem Set 4) for hours without rain (‘rain’ = 0). Note that the x-axis was limited to values less than 6000 and bin width = 200. Limiting the x-axis improves readability and scaling while still showing the majority of the distribution; the downside is that larger values and right-tail outliers are not shown.



The chart above shows the ggplot histogram of ENTRIESn_hourly (from Problem Set 4) for hours with rain (‘rain’ = 1). Note again that the x-axis was limited to values less than 6000 and bin width = 200. From these two ggplot plots, it appears that the shapes of the distributions of ENTRIESn_hourly are very similar when one groups by the two values of the rain variable.



This third chart shows the matplotlib (Problem Set 3) version of the two histograms, one overlaid on the other. This chart again shows that the two distributions are similar in shape, but the *With Rain* (‘rain’ = 1) distribution clearly has fewer samples. Note that the *With Rain* plot uses an alpha channel (transparency) value of 0.5 to improve visibility of the detail in the underlying *Without Rain* distribution.



The chart above shows total ridership by the day of the week. Note that the days of the week are represented by integers; Monday is day 0 and Sunday is day 6. In this dataset, on average the ridership peaks on Mondays and is at its lowest levels on the weekend.

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.

My analysis indicates that more people ride the subway when it is raining, but the effect is quite small (on the order of 1.3%). The Mann-Whitney U test results indicated that there is a statistically significant difference in the with-rain and without-rain distributions. The means were 1105.45 (with rain, n = 44,104) and 1090.28 (without rain, n = 87,847).

My best-fit multi-variable OLS regression model had a negative coefficient for 'rain,' leading one at first glance to reach a contradictory conclusion that rain has a negative effect on hourly ridership. To determine the impact of rain in isolation from the other input variables, I created an alternative model that uses only a constant, 'rain,' and 'UNIT'. In this model, the coefficient for 'rain' is positive, indicating that rain, when considered alone, indeed has a positive effect on ridership.

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?

The dataset for this project had several potential shortcomings. The first is its size. With only a single month's data, a number of confounding factors could easily arise, including exceptional entertainment, sports, security, or political events; odd or severe weather patterns; or other unique seasonal factors that can impact the representativeness of the sample. The second is the content and range of input variables. Without further knowledge of the sources, accuracy, precision and resolution of each of the included factors, it is difficult to develop actionable insights with confidence. One would normally want to spend a significant amount of time researching each source and scrubbing each input to eliminate invalid values, unlikely patterns, and outliers.

In this case I used the the Mann-Whitney U test to compare the two distributions, which is well-suited to non-normal sample distributions (as we have here) and has a higher statistical efficiency than the t-test. Because this test makes no assumptions about the two distributions, it is less likely to find a difference than parametric tests. In fact, in this case we see that the p-value is just slightly less than p-critical, indicating that the null hypothesis (no difference) was nearly confirmed by the data.

For my regression model, I used Ordinary Least Squares. OLS can perform poorly with outliers, and this dataset certainly has some large values (namely maximums of 51,839 with rain and 43,199 without, which fall well outside the range of the bulk of the values show in the histograms above) as well as a good amount of small or zero values. In this case, the model delivered a correlation of determination of 0.489, meaning that the model predicts just under half of the variation in the hourly entries.