Jason Reaves
May 31, 2015

Analyzing the NYC Subway Dataset

- (0) References

I used a wide variety of web sites, both during the lessons and while writing this document.

I used The Minitab Blog for a discussion how to interpret of R2 (http://bit.ly/1lmfMOI and http://bit.ly/1FRmHwg). Wikipedia also provided useful information on this topic (http://bit.ly/1Jb3fLT).

I found an excellent discussion of the pitfalls of using a linear regression model at http://bit.ly/1JZWlKe.

- (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?

   We used the Mann–Whitney U test with a two-tailed critical P value of 0.05. The standard two-tailed hypothesis is as follows:

   H_0: $P(Entry_{Rain} > Entry_{NoRain}) = 0.5$
   H_1: $P(Entry_{Rain} > Entry_{NoRain}) != 0.5$

   Under the null hypothesis, the probability of randomly sampling a greater number of entries when it's raining is 0.5. A common assumption under the null hypothesis is that the two distributions are identical, but this is not necessarily the case when using the Mann-Whitney test.

   With the two-tailed test, we are not assuming a particular direction of the relationship between rainy weather and ridership.

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.

   We used the Mann–Whitney U test (rather than Welch's t-Test) because the distribution of ENTRIESn_hourly was non-normal., a finding that is apparent by inspection of the histogram of entries (see section 3.1 below).

   This finding of non-normality can be confirmed using the Shapiro–Wilk test:

```
W, p = scipy.stats.shapiro(with_rain)
W, p = scipy.stats.shapiro(without_rain)
```

Both of these computations give a P value of 0.0. The null hypothesis of the Shapiro-Wilk test (that the data are normally distributed) is rejected.

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.

   ```
   U, p = scipy.stats.mannwhitneyu(with_rain, without_rain)
   ```

   produced the following:

   ```
   (1105.4463767458733, 1090.278780151855, 1924409167.0,
   0.024999912793489721)
   ```

   Here, the one-sided P value is <0.025. Doubling this gives a two-sided P value of <0.05.

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

   Based on the above results ($P < 0.05$), we can reject the null hypothesis that the probability of randomly selecting a higher number of entries when it's raining is equal to the probability of randomly selecting a higher number of entries when it's not raining. However, the P value for the U statistic just barely reaches the critical value.

   Examination of the descriptive statistics supports the idea that the difference in ridership on rainy days is *very slight*:

	With rain	Without rain
Median	282	278
IQR	1062.25	1073.0

- (2) Linear Regression

1. What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model

   I used the Ordinary Least Squares regression as implemented in `statsmodels.api`.

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

The features in my model were rain, meantempi, UNIT, Hour, and DATEn.

I used dummy variables for these categorical features: rain UNIT, Hour, DATEn, and. I converted DATEn to day of the week (0-6) as follows:

```
days_of_week = dataframe[['DATEn']].applymap(lambda x:
datetime.datetime.strptime(x, '%Y-%m-%d').strftime('%w'))
```

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.

   rain - We already demonstrated in a previous lesson that ridership increases slightly when it's raining, though the effect is small.

   meantempi – More people may ride the subway when it's very cold or very hot outside. (However, this may be a non-linear feature, since people may be more likely to ride when it is either extremely hot or extremely cold.)

   Hour - Ridership is probably higher during "rush hour," as people are going to or from work.

   DATEn (after converting to day of the week) - There may be fewer riders on weekends. (On a later visualization exercise, I found this to be true.)

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

   meantempi - -10.4262

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.534
Model:                            OLS   Adj. R-squared:                  0.516
Method:                 Least Squares   F-statistic:                     29.39
Date:                Sun, 31 May 2015   Prob (F-statistic):               0.00
Time:                        01:46:58   Log-Likelihood:            -1.1600e+05
No. Observations:               13195   AIC:                         2.330e+05
Df Residuals:                   12699   BIC:                         2.367e+05
Df Model:                         495
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const       1752.8469    124.143     14.120      0.000     1509.508  1996.185
x1            -0.2358     32.624     -0.007      0.994      -64.184    63.712
x2           -10.4262      2.225     -4.687      0.000      -14.787    -6.066
```

5. What is your model's R2 (coefficients of determination) value?

   My model's R2 value was 0.534, as calculated both by the Udacity grader and by `statsmodels.api`.

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 is a statistical measure of how close the data are to the fitted regression line. Its value ranges from 0 to 1, and in general, the higher the R2, the better the model fits the data.

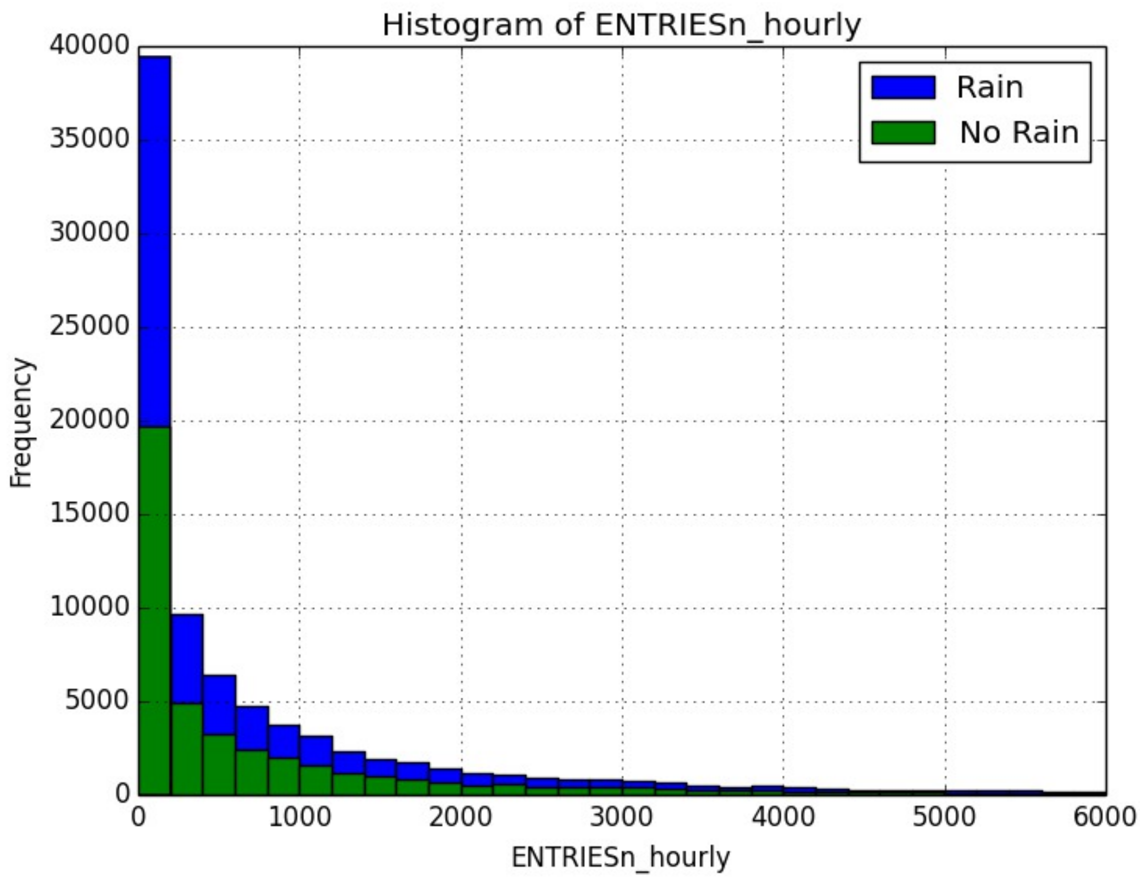
   R2 always increases as more features are added to a linear regression model. A model can have a "good" R2 just by having a lot of variables, even if those variables hold no true predictive value.

   In reality, most systems are not linear. As noted above, ridership may increase at both high and low extremes of temperature, as people choose to ride the subway rather than walk outside. A linear regression model cannot capture this non-linearity.

   A linear regression model may lead to poor prediction when the features are significantly correlated with each other (e.g., rain and temperature).

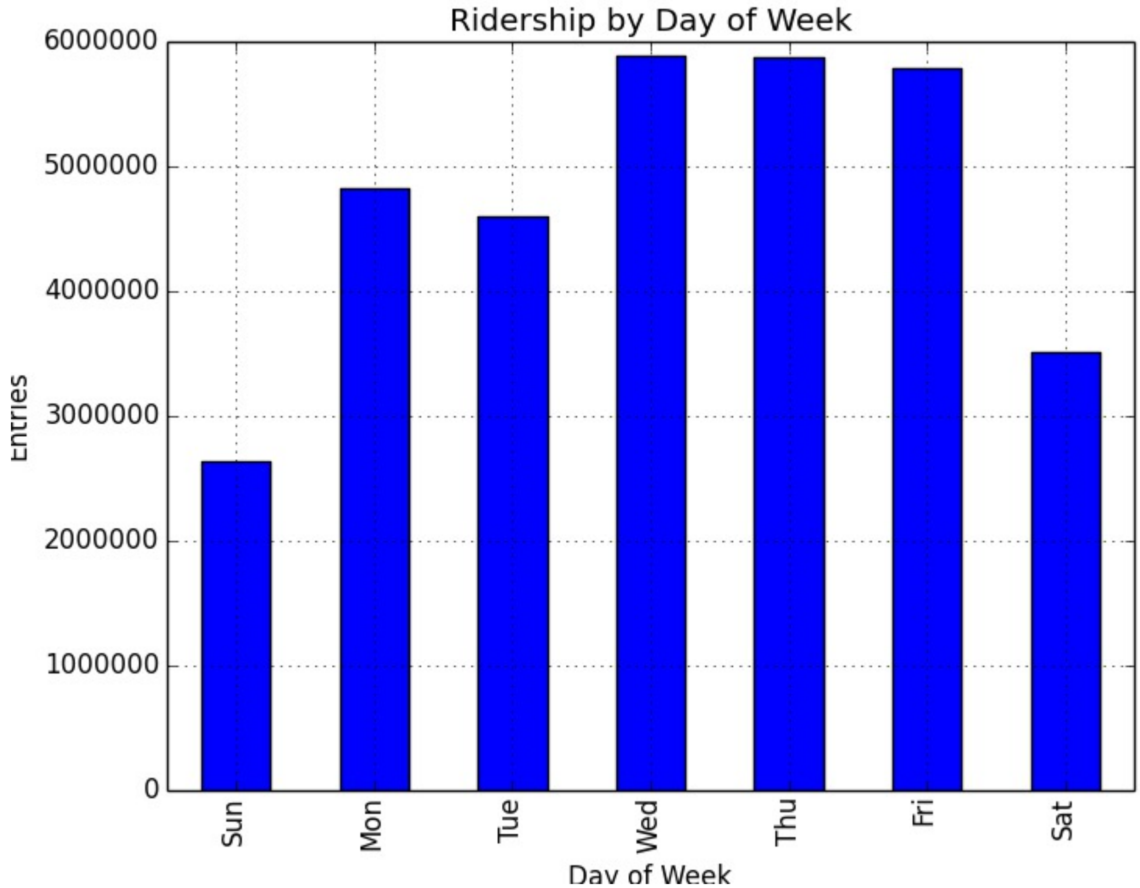
- (3) Visualization

1. Histograms of ENTRIESn_hourly for rainy and non-rainy days:



The figure above is a histogram of entries for rainy and non-rainy periods. The histograms are of the same shape and do not reflect a normal distribution.

2. Freeform visualization: Ridership by day of week



The figure above shows average ridership by day of week. The number of entries plotted on the vertical axis is the total for that weekday in the dataset, divided by the number of occurrences of that weekday in the dataset. For instance, there were 5 Mondays in the dataset, so the total Monday turnstile entries was divided by 5. As one might expect, ridership decreases on the weekend.

- (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?

   Yes, average ridership increases slightly when it's raining.

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.

   As discussed above, the results of the Mann–Whitney U test suggest a small increase in ridership when it's raining. However, in my regression model, the coefficient of the rain variable was negative, suggesting an opposite trend. In fact, the absolute value of the rain coefficient was small (0.2358), its P value was high (0.994), and it's confidence interval was wide and included 0 (-64.184 63.712). From this, one can conclude that rain did not play a significant role in the linear regression model. Other variables, particularly UNIT (turnstile), played a larger role.

- (5) Reflection

1. Please discuss potential shortcomings of the methods of your analysis.

   i. Dataset

   The dataset covered only the month of May, so the model cannot reflect variations in ridership over the course of a year. One can imagine that over the course of a year, there would be greater fluctuations in temperature and other potentially significant weather phenomena such as snow and ice. There could also be changes in ridership as the population using the subway varies (e.g., more tourists during the summer months, and the holiday season).

   ii. Analysis

   The limitations of a linear regression model, some of which were discussed above, include outliers, non-linearities, and dependence among variables.

   The statistical test showed slightly greater ridership when it's raining, but the practical significance of this finding depends on the reason the question is being asked in the first place.