Timothee ROBIN
timothee.robin@axa-lifeinvest.com

Introduction to Data Science

Analyzing the NYC Subway Dataset – Questions

Section 1. Statistical tests

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?*

Our first test was to determine if subway ridership was the same in NYC on rainy days or on sunny (eg non-rainy days in the following) days, eg. if the number of entries was the same on rainy days or on sunny days. Our own experience and intuition is that subway ridership is greater on rainy days. Our "null" hypothesis is that both datasets come from the same sample, thereby having the same average. We will try to dismiss this hypothesis with the tests below.

When plotting the observed distribution of ridership (on rainy or on sunny days), we found that the distribution cannot be approximated by a Gaussian distribution. So, a Welch's t-test would make no sense in this instance.

As a consequence, we implemented a non-parametric test, which does not assume any specific distribution for the observations. We implemented a Mann-Whitney U test (see below).
The test itself returns a 1-sided p-value. However, we actually need to know if the 2 sets of observations (ridership on rainy or on sunny days) have the same average (our null hypothesis), and not if one is greater than the other. So, we would actually need a 2-sided test. We could multiply the p-value by 2 for a 2-sided test.

The p-critical value we choose is 5%.

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 all the observations from both groups are independent of each other, and because it is known to provide with greater efficiency than other tests (t-test for example).

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.*

The results I got for the Mann-Whitney U test are the following:

- Value of the U statistics equal to: 1924409167
- A one-sided p-value of: 0.0249999, and a 2-sided p-value of: 0.499998
- An average number of entries for the sample of rainy days equal to: 1105

- An average number of entries for the sample of sunny days equal to: 1090

4) *What is the significance and interpretation of these results?*

The way I interpret the results is the following. Given the 2 samples, we have a 2-sided p-value which is slightly under 5%. If we assume that the 2 samples are similar (this is our "null hypothesis"), this means that the likelihood to observe values for U which are as extreme as the value we got is under the 5% p-critical threshold. This suggests that we should reject the null hypothesis, and consider that the difference of average ridership on the 2 samples (rainy and sunny days) is actually statistically significant.

Section 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 computed the explanatory variables "theta" using the gradient descent algorithm as described in the class.

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

I mainly used an intercept, hour, rain, meantempi, fog and the dummy variables

3) *Why did you select these features in your model?*

I tried several combinations of explanatory variables, mainly a combination of:
- time of the day 'Hour': I observed that there was a peak due to usual rush hours in the morning, lunchtime and evenings. So the relationship between ridership and "Hour" might not really be linear in reality, and maybe this variable on its own is not sufficient;
- rain, measured by "precipi": we shall expect more entries on rainy days;
- fog, for the same reason;
- outdoor temperature: with 'meantempi'. The assumption here is that people would be more likely to use the tube when it is cold outside. I also tested squares of these variables to test whether there was a particular threshold below which people would tend to use the tube more often.

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

The regression coefficients that I got for my variables (rain, precipi, Hour, meantempi) are the following:

- Rain: 2.92
- Precipi: 14.65
- Hour: 467.71
- Meantempi: -62.21
- Intercept: 1100.61

5) *What is your R^2?*

My R2 is 0.464.

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?*

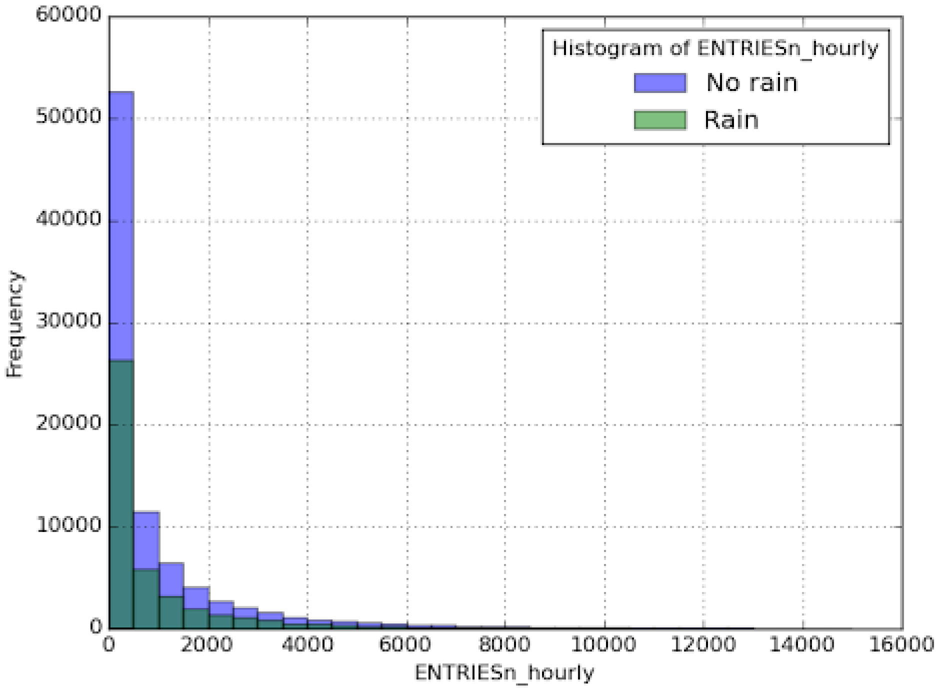
The R2 is usually a measure of the quality of regression: the higher (and closer to 1), the better.

R2 is actually equal to 1 – sum_{square of residuals}/variance_of_data. So ideally, if the regression is good, the residuals are equal to 0, and R2 is 1.

In my case, the quality of the regression is not too bad with this set of variables. When removing the “dummy variables”, I noted that the quality of the regression becomes very poor.

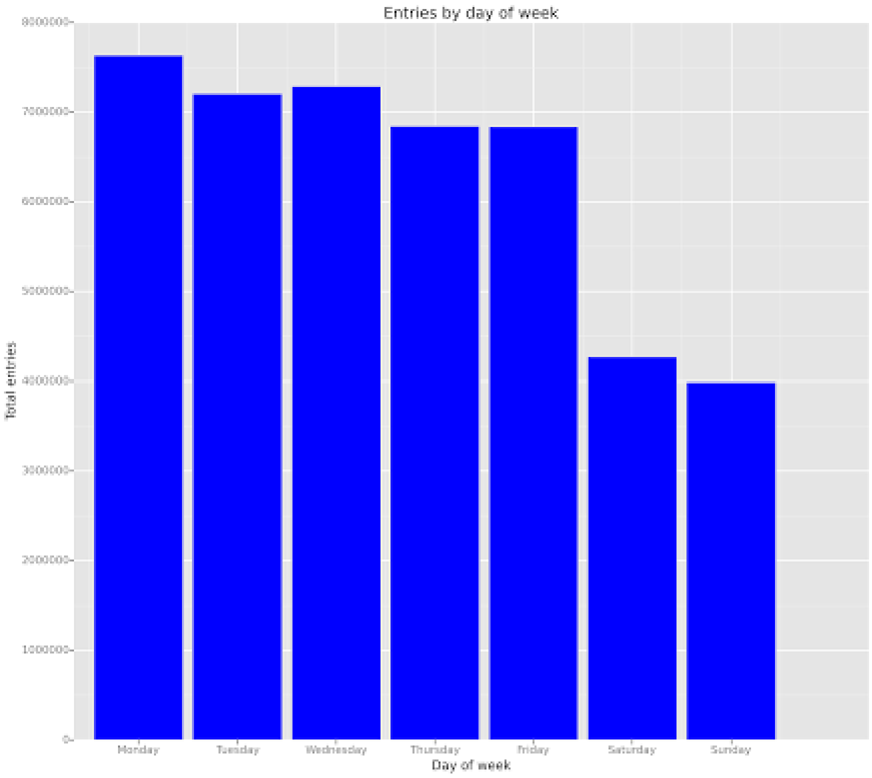
Section 3. Vizualization

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



The histogram shows the distribution of entries in the NYC subway, on rainy days and on sunny days. We can see that the distribution really does not look “Gaussian”, but more like a “power” law. This is why we dismissed the Welch’s test as it assumes that the distributions have normal distributions.

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 graph depicts the total of entries in the NYC subway in the dataset, split by day of the week. Assuming that there are equal numbers of weeks/weekends in the dataset (May 2011), the graph tends to show that there are more entries in the week days than on weekend days. As a matter of fact, the tube looks more crowded with workers that commute to their workplace.

Section 4. Conclusion

I would conclude that more people ride the NYC subway on rainy days than on sunny days.

First we can definitely see this in the data itself and on the plot (even if it comes from a reduced sample of data points) in Section 3.1. The graphs suggests that people use the tube less often on sunny days.

Then, the statistical tests that we ran (the Mann-Whitney U-test) tends to show, as discussed, that we should reject the "null hypothesis" (hypothesis the 2 data samples, rainy days and sunny days, yield similar patterns) with a level of confidence of 5% (p-critical = 5%).

Moreover, the linear regression shows that the coefficient for variable ‘rain’ is positive. This means that the number of entries increases when the variable ‘rain’ increases. So the result confirms that the intuition that there are more entries when it is raining.

Section 5. Reflection

Going back to the dataset itself, we noticed that the “Unit” variable added in the regression analysis has a significant impact on the output results of the regression. Each unit records the number of entries on a particular time of the day. But then, I would imagine that the entries near Times Square or near Wall Street station in the business district have a very large number of records. So the dataset would actually be dominated by these units on a certain period of the day. Even if it causes no problem to build a regression analysis to describe the overall patterns in the data, we have to be careful not to draw conclusions about entries on a particular turnstile or entries on a particular time of the day.

Moreover, for our reduced dataset for testing purposes, we only have data for May 2011. I would expect to results to be even more convincing for winter days: people would rather use the subway than bike or walk…

For the linear regression, I think the coefficients are strongly dependent on the dataset itself. As discussed above, I would expect the coefficient for ‘rain’ or for ‘meantempi’ to increase significantly on winter days for example. I think one other test that needs to be discussed is whether the coefficients (and their signs) are significant. For example: is there a test to prove that the coefficient for ‘rain’ is significantly positive? I think a t-test, where we would test if the null hypothesis (coefficient = 0 or not for example) can be used, or can help derive confidence intervals. So that we can positively conclude that the number of entries does increase when it is raining.