Analyzing the NYC Subway Dataset – Short Questions to Analyzing the NYC Subway Dataset

Anna Anesiadou-Hansen

Contents

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, which is a non-parametric statistical test, given by following Python statistical function:

```
scipy.stats.mannwhitneyu(X, Y, use_continuity = True)
```

where X sample is the number of entries with rain and Y sample is the number of entries without rain.

I used a two-tail p-value. The function, scipy.stats.mannwhitneyu returns a one-tail value, which must be multiplied by 2 in order to report the proper p-value.

The null hypothesis H_0 is: the number of entries with rain and the number of entries without rain are the same or in mathematical terms, given random draws x from X and y from Y: $P(x > y) = 0.5$

My p-critical value is 0.05.

Note I used as data source `ProblemSet3/turnstile_weather_v2.csv` for answering the short questions to analyzing the NYC subway dataset.

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 a statistical test for non-normal distributions, i.e. non-parametric ones.

As a first step to verify the distribution I checked the histograms of ENTRIESn_hourly with rain and without rain. The histogram, see figure 2 on page 4, of the two samples shows that their distribution is not normal.

As a second step I used the Shapiro-Wilk test in order to check if the samples are normal distributed. For sample X the Shapiro W test statistic is 0.5938820838928223 and p is 0.0. For sample Y the Shapiro W test statistic is 0.5956180691719055 and p is also 0.0, which means that in both cases the samples are not normal distributed. Based on the results above I concluded that the samples are non-parametric. Therefore the Mann-Whitney U Test is applicable to the given dataset.

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.

My one-tailed p-value is 2.7410695712437496e-06, which multiplied by two gives the two-tailed p-value 5.482139142487499e-06. The mean value of entries with rain is 2028.1960354720918. The mean value of entries without rain is 1845.5394386644084.

1.4 What is the significance and interpretation of these results?

Based on the results of Mann-Whitney U test the null hypothesis has to be rejected. The p-value is 5.482139142487499e-06, which is less than p-critical (0.05). That means that there is a significant difference between ENTRIESn_hourly with rain and ENTRIESn_hourly without rain. Taking into account the mean values **I conclude that more people use NYC subway when it rains**, since the mean value of ENTRIESn_hourly with rain is larger than the mean value of ENTRIESn_hourly without rain.

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?

For producing predictions and computing the coefficients theta I used the gradient descent algorithm in my regression model.

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

The features that I used in my model are hour and UNIT. UNIT is a categorical variable from which many dummy variables are generated.

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.

I decided to use hour because I thought that subway ridership depends on time of day. I assumed that many people use the subway to travel to work. That should lead to ridership peaks at beginning and at end of working hours.

Additional I included UNIT in my features because I thought that ridership varies by UNIT, e.g. UNIT R179 (1,270,579) has more entries than R172 (362,755).

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

The weight of hour is 834.962837 after 75 iterations with alpha (learning rate) equal to 0.05.

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

The R2 value is 0.458442000297 after 75 iterations with alpha equal to 0.05.

Table 1: R2 values of various feature combinations

Case	Features	alpha	Iterations	R2
1	hour, UNIT	0.01	75	0.267
2	hour, UNIT	0.01	150	0.426
3	**hour, UNIT**	**0.05**	**75**	**0.458**
4	hour, UNIT	0.05	150	0.469
5	rain, UNIT	0.05	75	0.375
6	fog, UNIT	0.05	75	0.374
7	day_week, UNIT	0.05	75	0.384
8	hour, rain, UNIT	0.05	75	0.459
9	hour, fog, UNIT	0.05	75	0.458
10	hour, day_week, UNIT	0.05	75	0.468
11	hour, day_week, fog, UNIT	0.05	75	0.468
12	hour, day_week, rain, UNIT	0.05	75	0.468
13	meantempi, UNIT	0.05	75	0.375

2.6 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 approximately 0.46, which means that 46% of the variation in the dependent variable ENTRIESn_hourly can be attributed to the variation in the independent variables hour and UNIT.

The above value of R2 is the best value which I got after testing various features, see table 1, but still not very high. The best performance of R2 has been shown with alpha equal to 0.05 after 75 iterations. Weather conditions like rain, fog and mean temperature or day of the week do not have a strong impact on ridership.

The histogram of the residuals, i.e. observed values minus predicted values, illustrates an approximately normal distribution of the residuals, see figure 1. That leads to the conclusion that the linear model to predicted ridership is appropriate for this dataset.

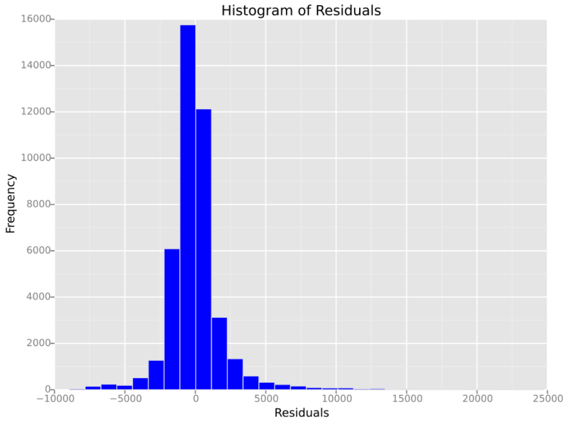

Figure 1: Histogram of Residuals

3 Visualization

3.1 Histograms of ENTRIESn_hourly for rainy days and histogram of ENTRIESn_hourly for non-rainy days.

Figure 2 on the next page contains two histograms. The histogram with blue color shows the relationship between ENTRIESn_hourly in x-axis and its frequency in y-axis, when it is not raining. In the same way, the histogram with green color shows the relationship between ENTRIESn_hourly in x-axis and its frequency in y-axis, when it is raining. The histograms illustrates that the distribution of the samples is not normal.

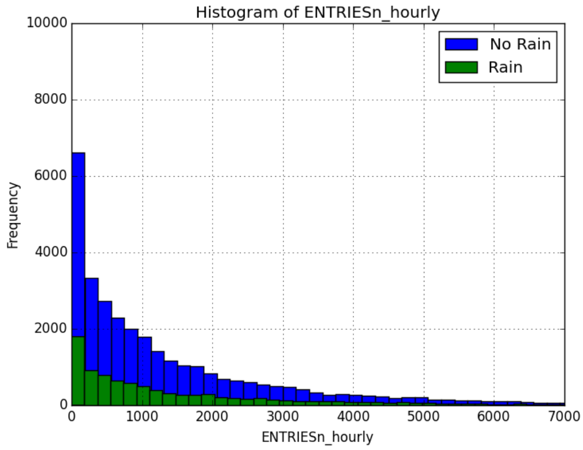

Figure 2: Histogram of ENTRIESn_hourly

3.2 Ridership by time-of-day and Ridership by day-of-week

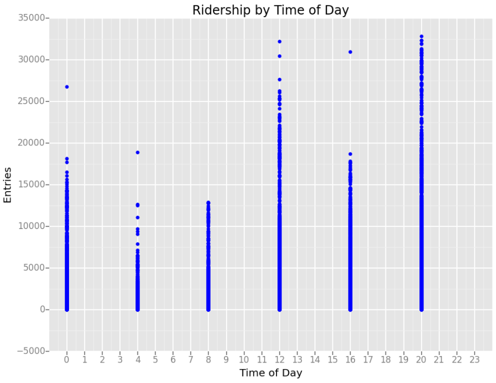

Figure 3: Ridership by Time of Day

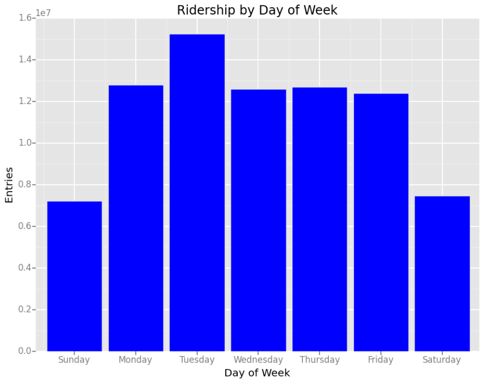

Figure 4: Ridership by Day of Week

Figure 3 visualizes ridership by time-of-day. It represents the relationship between hours and ENTRIESn_hourly. The figure below shows that at around 12 o'clock and at 20 o'clock the ridership is highest.

Figure 4 visualizes ridership by day-of-week. It represents the relationship between day of week and the number of entries. The figure shows that during the week the ridership is substantially higher than on weekends.

4 Conclusion

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?

Analyzing the results from the statistical test, the linear regression and the visualization of the dataset I conclude that more people use subway when it is 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.

Using rain as a feature in the regression model with gradient descent, the coefficient of determination R2 is relative low. It is equal to 0.375, see table 2 on the following page. The comparison of R2 values from case

3, 5 and 8, shows that the independent variable rain does not have a big impact on the dependent variable ENTRIESn_hourly — only 37.5%.

Table 2: R2 values of various feature combinations

Case	Features	alpha	Iterations	R2
3	hour, UNIT	0.05	75	0.458
5	rain, UNIT	0.05	75	0.375
8	hour, rain, UNIT	0.05	75	0.459

The Mann-Whitney U statistical test in Section 1 returns a p-value equal to 5.482139142487499e-06, which is less than the p-critical, and therefore the null hypothesis H_0 can be rejected. That means that the ridership between raining and non raining days is not the same. Additional taking in account, the means of the samples:

- mean value of entries with rain = 2028.1960354720918,
- mean value of entries without rain = 1845.5394386644084

I conclude that **more people use the NYC subway when it is raining**.

5 Reflection

The model that I used is based on data for only one month, the month May 2011. Therefore I think that my regression model is not robust enough. I have to few data to make robust predictions.

In the dataset rain is reported only on a daily basis. It might be a crucial factor for ridership to distinguish if it rains some minutes or if it rains the whole day. This is not possible to do with the given data.

People might take the decision to travel by subway, walk or drive with the car depending on the weather forecast for the traveling time and not on the actual weather condition when they commute. It might be better to base the ridership predictions on the weather forecast and not on the actual weather at travel time.

The number of entries/exits (ENTRIESn_hourly/EXITSn_hourly) are not reported per hour, but for four hours. This has a negative impact on the precision of the regression model.

The value of coefficient of determination R2 is about 0.46. It might be possible to achieve higher R2 values with a nonlinear regression model.

6 References

- Shapiro-Wilk test on Wikipedia
- statisticsviews.com: Getting to the Bottom of Regression with Gradient Descent
- Nonparametric regression on Wikipedia
- SciPy Shapiro-Wilk reference
- statisticssolutions.com: Mann-Whitney U-test
- ggplot geomline reference