Analyzing the NYC Subway Dataset

Short Questions

Section 1. Statistical Test

1.1 Which statistical test did you use to analyse 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?

The question at issue is whether the mean of the probability distribution for ridership on rainy days is higher than mean of the probability distribution for ridership on non-rainy days. The null hypothesis is that the sample comes from distributions for which the rainy day mean is less than or equal to the non-rainy day mean. The alternative hypothesis is that the samples come from distributions for which the rainy day mean is greater than the non-rainy day mean. The Mann-Whitney U-test was used to analyze the data.

A one-tail P value was used because we are interested in the probability that the rainy day mean is larger than the non-rainy day mean.

The one-sided p-critical value corresponding to a 95% significance level is 0.025.

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 non-parametric test. It does not assume that the samples have any particular probability distribution. A non-parametric test does not assume that the probability distributions of either sample are normal.

Histograms of the ridership samples from the rainy and non-rainy days looked far from Gaussian.

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.

The observed U-statistic was approximately 1.924 billion. The (one-sided) p-value was 0.025. The one-sided p-value is equal to the critical p-value. The means for the rainy day and non-rainy day samples were 1105 and 1090.

1.4 What is the significance and interpretation of these results?

U is positive and the (one-sided) p-value is equal to the critical value of 0.025. Assuming that the rainy day mean is equal to the non-rainy day mean, the probability of getting a sample with a U-statistic that is equal to or greater than the one observed is 0.025. We can reject this assumption with a confidence level of 95%.

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.

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

The input variables used in my model were rain, Hour, meanwindspdi, mintempi, maxtempi, and minpressurei. UNIT was used as a dummy variable. This introduces hundreds of pseudo-features, one per station in the input.

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.

An initial set of features was selected using the intuition that cold, wet, and windy days would increase ridership. Hour was added to account for rush hours. I tried a number of different combinations looking for one with a high R-squared. Eventually, minpressurei was added as a proxy for stormy weather and precipi was dropped because it had no influence on the answer when rain was also included.

Each row of the table that follows contains the results of an experiment that was performed. The R-squared value for each experiment is at the far right.

rain	precipi	Hour	meantempi	meanwindspi	mintempi	maxtempi	minpressurei	R-squared
7.235	4.819	458.7	-42.87	54.83				0.46431
6.999	5.783		-43.94	50.85				0.42614
10.809	7.705	458.8		68.69				0.46403
11.910	5.780	458.8		50.61	-57.80			0.46457
13.968	7.477	458.8		53.07	-66.02	19.35		0.46466
1.866	7.624	459.1		45.5	-80.14	24.23	-66.86	0.46544
5.911		459.1		46.78	-79.74	23.50	-66.66	0.46544

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

Feature	Coefficient
rain	5.91
Hour	459.10
meanwindspi	46.78
mintempi	-79.74
maxtempi	23.50
minpressurei	-66.66

Some of these are reasonable. Ridership increases when it is rainy and windy. The negative coefficient for mintempi means that ridership is high in very cold weather. Ridership also increased when maxtempi is high. My guess is that this means that there were some very hot days that pedestrians tended to avoid. The negative coefficient for min pressure says that ridership increases when air pressure is low; that is, in stormy weather.

The high and positive weight for Hour must have something to do with rush hour traffic, but I would not expect the traffic to be a linear function of the hour of the day.

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

0.46544.

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?

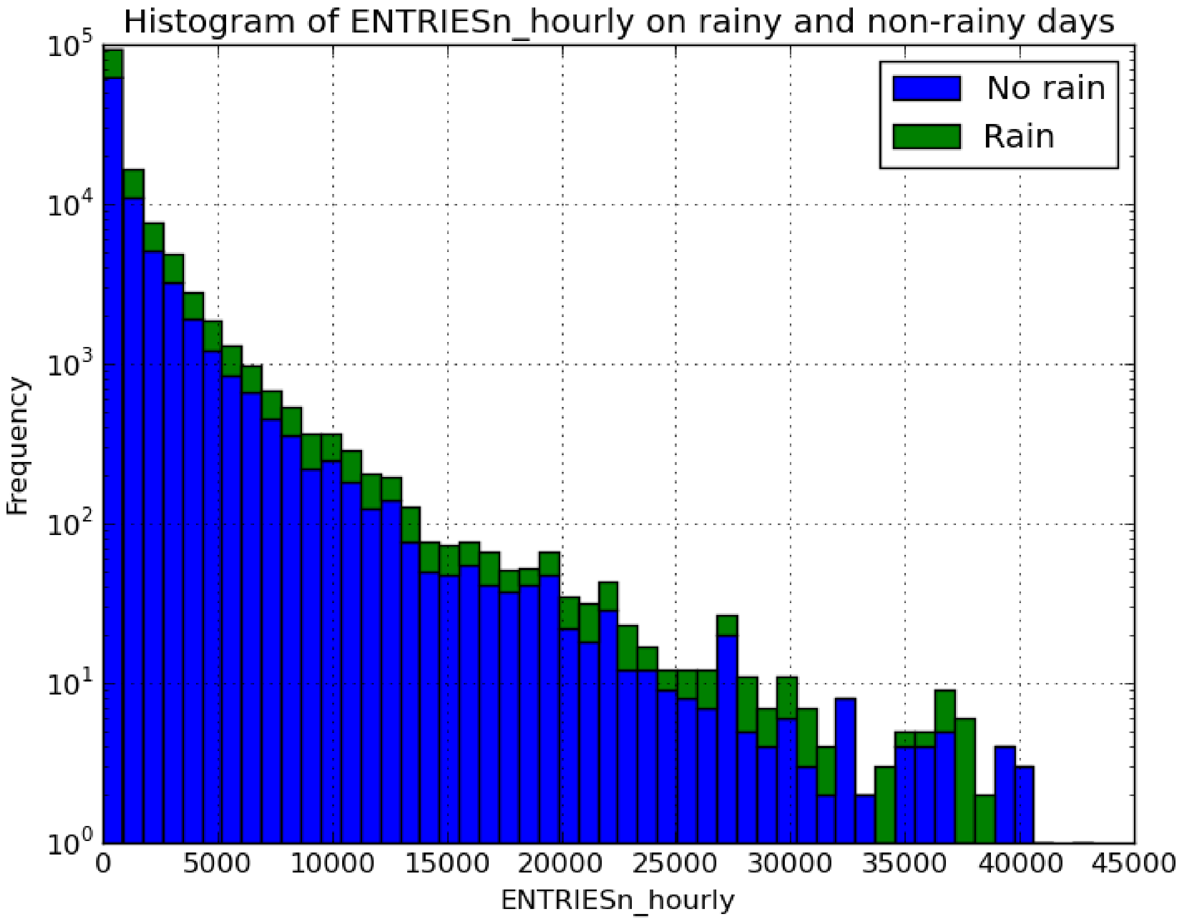
A value of 0.4 for R2 means that the fit is neither great nor horrible.

I do not think a linear model that includes Hour is all that appropriate. However, leaving Hour out of the linear fit worsened the prediction; simply ignoring the time of day does not make its importance disappear.

Since R2 is a good measure for goodness of fit for linear models, an R2 value of 0.46 means something, but not as much as it could mean.

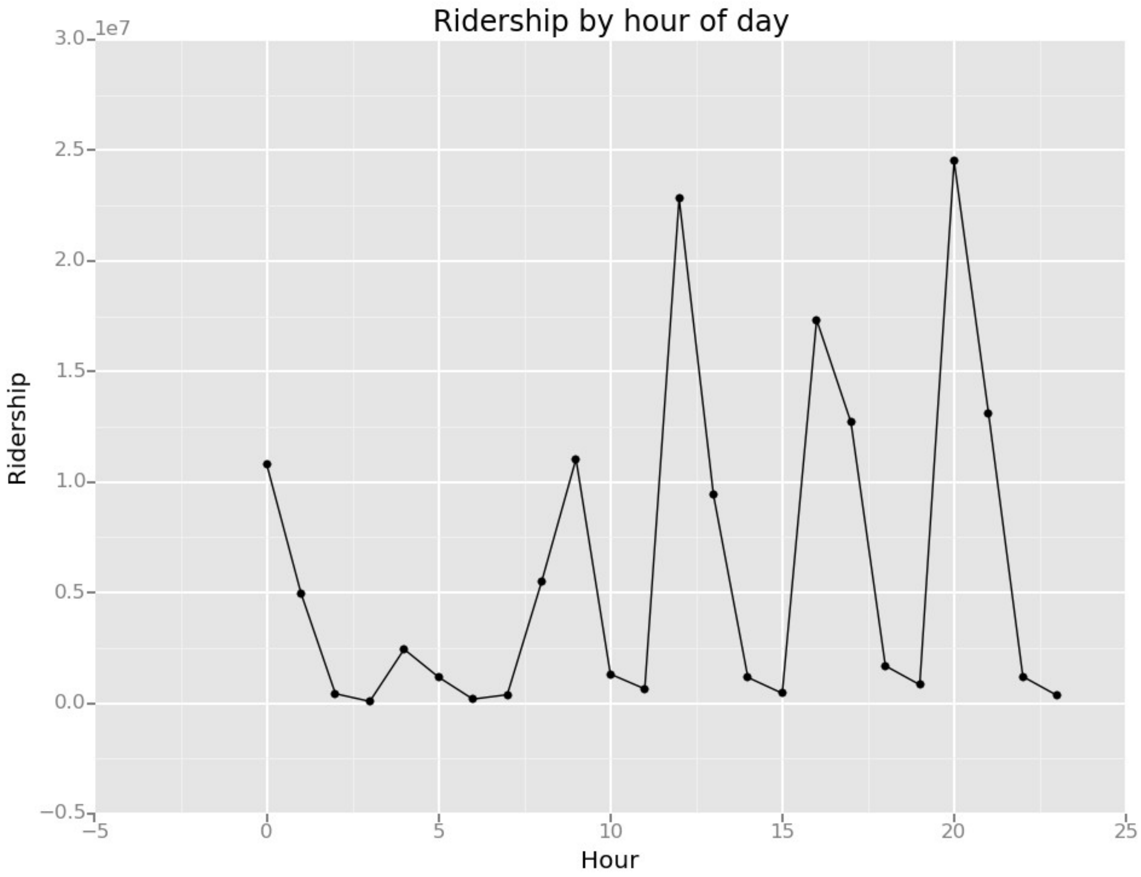
Section 3. Visualization

3.1



Ridership histograms are plotted for rainy and non-rainy days from May 2011. The ridership is on the horizontal axis and the histogram counts are on the vertical logarithmic axis. The values for rainy days (in green) are generally above those for non-rainy days (in blue).

3.2



The average ridership as a function of a time of day was constructed by averaging over all days in May 2011 and all NYC subway stations. The importance of this plot is that it shows that ridership is a non-linear function of the time of day. Ridership is higher later in the day than earlier; explaining the large and positive coefficient observed in linear regression modelling.

Section 4. Conclusion

4.1 and 4.2 The single-tail Mann-Whitney U-test showed that the hypothesis that ridership is not higher on rainy days than it is on non-rainy days can be rejected with a confidence level of 95%. This is consistent with our conclusion that more people ride the subway when it is raining than when it is not raining.

One of the features used in the linear regression model for ridership is rain. The coefficient of the rain variable is positive (5.91) which means that, all things being equal, ridership increases when it is raining (rain=1) than when it is not (rain=0).

Section 5. Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis, including:

1. Dataset,
2. Linear regression model,
3. Statistical test.

The dataset is limited to a single month. Rainy day behavior is likely to be very different in March when it is cold than it is in August when it is humid and hot.

The dataset is limited to a single month of a single year. If the ridership is not a stationary process then the conclusions we reach may be wrong in the future. Ridership will change with people's ability to afford the subway, for instance.

The precision with which rain was recorded in the dataset is limiting. Rain was represented as a categorical variable, either 0 or 1. Precipitation was measured with only one digit of accuracy.

The linear regression model was dominated by time of day. Visual inspection of average ridership versus time of day shows that the relationship is not linear.

The use of a one-sided statistical test becomes problematic when the means of the alternative distributions get too close to one another.