

P1: Analyzing the NYC Subway Dataset

Section 0. References

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?

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.

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.

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 R2 (coefficients of determination) value?

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?

Section 3. Visualization

3.1 Histograms of ridership for rainy and non-rainy days (limited at 2000)

3.2 Average amount of riders per hour

Section 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?

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.

Section 5. Reflection

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?

Section 6. Supporting Graphs

[6.1 R2 obtained for choices of polynomial degree in the interval \[20, 30\) for the hour input feature](#)
[6.2 Execution time obtained for choices of polynomial degree in the interval \[20, 30\) for the hour input feature](#)
[6.3 Residuals plot obtained after applying the model to predict the value for the ENTRIESn_hourly column](#)

Section 0. References

- Dataset used for local testing
 - https://www.dropbox.com/s/meyki2wl9xfa7yk/turnstile_data_master_with_weather.csv
- API references
 - <http://docs.scipy.org/doc/scipy/reference/stats.html>
 - <http://docs.scipy.org/doc/numpy/reference/>
 - <http://ggplot.yhathq.com/docs/index.html>
- Statistical Tests
 - <http://www.itl.nist.gov/div898/handbook/prc/section1/prc131.htm>
 - https://en.wikipedia.org/wiki/Null_hypothesis
 - https://en.wikipedia.org/wiki/Mann%E2%80%93U_test
 - http://www.graphpad.com/guides/prism/6/statistics/index.htm?how_the_mann-whitney_test_works.htm
- Linear Regression
 - http://en.wikipedia.org/wiki/Ordinary_least_squares
 - [http://en.wikipedia.org/w/index.php?title=Linear_least_squares_\(mathematics\)](http://en.wikipedia.org/w/index.php?title=Linear_least_squares_(mathematics))
 - http://en.wikipedia.org/wiki/Polynomial_regression
 - https://en.wikipedia.org/wiki/Coefficient_of_determination
 - <http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm>
 - <http://blog.minitab.com/blog/adventures-in-statistics/why-you-need-to-check-your-residual-plots-for-regression-analysis>
 - <http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit>

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?

- **Statistical Test:** Mann Whitney U-test.
- **Null hypothesis:** Is the hourly entries of rainy and not rainy days sampled from populations with identical distributions? This is a two tailed hypothesis.
- **p-critical value:** I used a significance level of 0.05 (5%)

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.

After plotting the histograms of both samples, I decided to use the Mann Whitney U-test as the samples didn't seem normally distributed but shared a similar shape.

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.

- **Mean hourly entries for rainy days:** 1105.4463767458733
- **Mean hourly entries for non-rainy days:** 1090.278780151855
- **Mann-Whitney statistics value:** 1924409167.0
- **p-value:** 0.024999912793489721

1.4 What is the significance and interpretation of these results?

Applied the Mann–Whitney U test over the two independent samples and obtained $U = 1924409167.0$ and $p = 0.024999912793489721$. For a two tailed test we have $2 * p\text{-value} = 0.049999826 < 0.05$, so we can say samples are statistically different.

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.

I tried both Gradient Descent and Ordinary Least Squares (OLS) and they produced similar results when using the same input features. I then proceeded to improve the OLS model to generate the final predictions for `ENTRIESn_hourly`.

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

- Columns from the dataset used: *fog*, *rain*, *meantempi*.
- 25th degree polynomial from the input variable *hour*.
- Dummy feature for each type of UNIT.

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 started by including the following columns: *fog*, *rain*, *hour*, *meantempi*. I also included a dummy feature for each type of unit. This selection was heavily based on intuition and some minor experimentation. For all these features, it was fair for me to assume they would impact the use of subway, e.g. people would travel more during certain hours of the day or some units would get more traffic than others given the location. After trying them out I could see that the coefficients obtained reflected this, e.g. the predicted coefficients for some dummy features were larger than others and the coefficient for the hours had the largest magnitude of the non-dummy coefficients.
- Thanks to a hint in the assignment, I decided to explore the use of polynomial terms for the column *hour* as it was the non-boolean feature with the largest coefficient value and my intuition suggested the model would benefit from this. I ran local tests where only the degree of the polynomial was changed and recorded the R^2 values and execution time obtained, reflected by the supporting graphs [6.1](#) and [6.2](#) respectively. From this I concluded that using a 25th degree polynomial was the simplest model within the explored interval of [20, 30).

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

| Feature | Coefficient |
|-----------|-------------|
| fog | 4.11E+01 |
| rain | -1.82E+01 |
| meantempi | -6.03E+01 |
| Hour1 | -8.65E+06 |
| Hour2 | 4.94E+08 |
| Hour3 | -1.03E+10 |
| Hour4 | 1.12E+11 |
| Hour5 | -7.15E+11 |
| Hour6 | 2.89E+12 |

| | |
|--------|-----------|
| Hour7 | -7.76E+12 |
| Hour8 | 1.42E+13 |
| Hour9 | -1.80E+13 |
| Hour10 | 1.54E+13 |
| Hour11 | -8.57E+12 |
| Hour12 | 2.80E+12 |
| Hour13 | -4.07E+11 |
| Hour14 | -1.46E+05 |
| Hour15 | 2.21E+03 |
| Hour16 | 1.79E+03 |
| Hour17 | -3.77E+03 |
| Hour18 | 1.07E+04 |
| Hour19 | -1.12E+05 |
| Hour20 | -3.32E+05 |
| Hour21 | -2.21E+09 |
| Hour22 | 1.69E+09 |
| Hour23 | 7.06E+08 |
| Hour24 | -1.56E+08 |
| ones | 1.10E+03 |

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

$R^2 = 0.526331531875$

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?

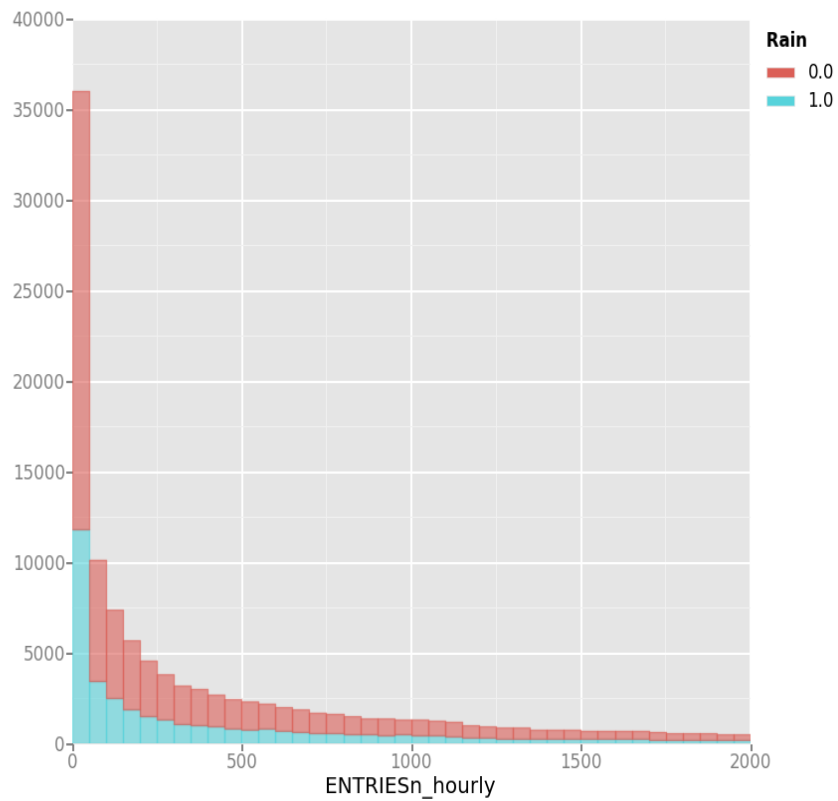
A higher R^2 value means our model fits our sample data better, in this case it means we can better predict the value for the ENTRIESn_hourly column on our sample with the hope of generalizing to unknown samples.

Is hard to say if the obtained value of $R^2 = 0.526331531875$ is appropriate for this dataset without following a more formal benchmark process. In the light of this limitation, I decided to also examine the residuals plot obtained after applying my model to the sample data provided in the class resources dashboard, referenced by the supporting graph [6.3](#).

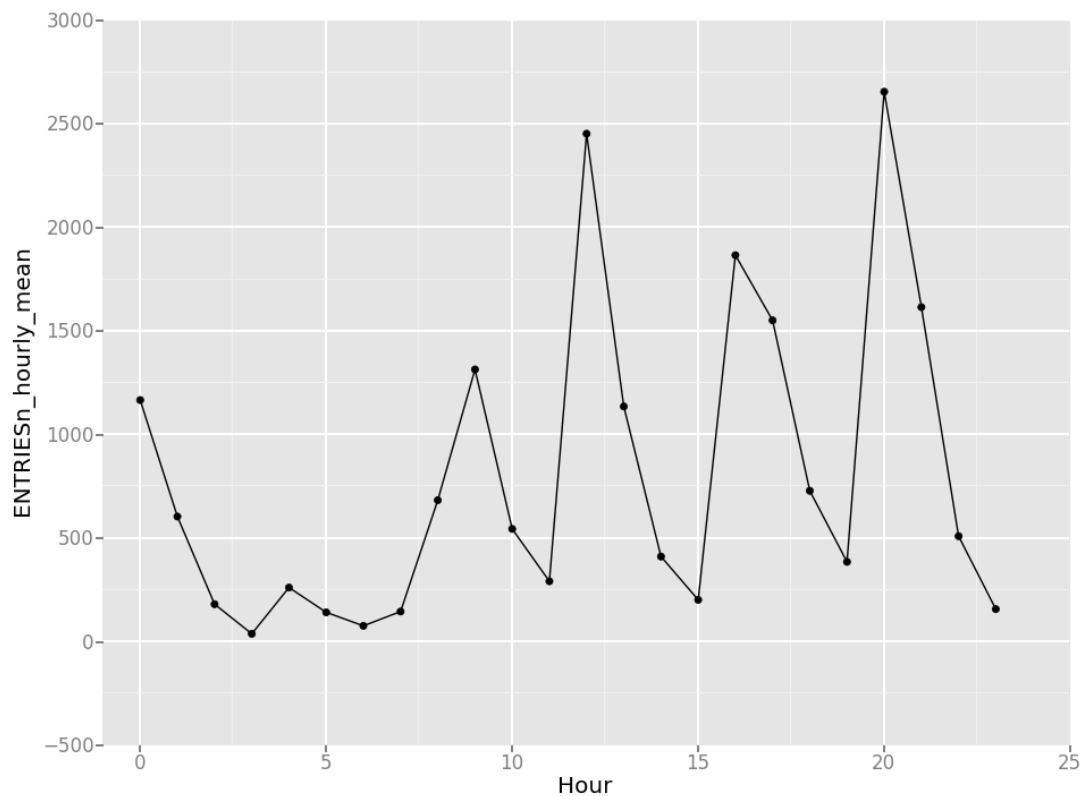
Given that the residuals distribution seems normally distributed and we obtained a value for R^2 greater than 0.5 in the context of what looks like a highly noisy real world problem, then I would conclude that this is a satisfactory result for a linear model and one that can have a strong predictive power in the real world.

Section 3. Visualization

3.1 Histograms of ridership for rainy and non-rainy days (limited at 2000)



3.2 Average amount of riders per hour



Section 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?

Based on my analysis, I can say with a good amount of confidence that more people ride the NYC subway 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.

1. The results obtained from running the [Mann Whitney U-test](#) led me to conclude with a 95% of confidence that the hourly entries of rainy and not rainy days are sampled from populations with non-identical distributions.

2. After observing the [histogram](#) of ridership on rainy and non-rainy days, is easy to observe the higher frequency of ridership when is not raining.
3. The coefficient obtained for the boolean input feature *rain* was $-1.82E+01$, which is a negative value that only applies if *rain*=1.

Section 5. Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis, including: Dataset, Analysis, such as the linear regression model or statistical test.

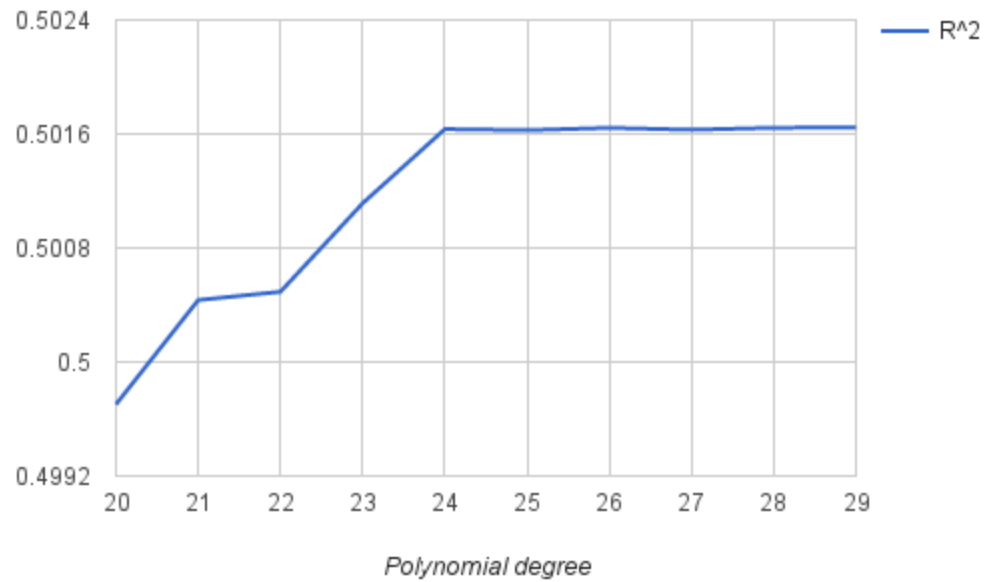
- A linear regression model might be too simple for this dataset, but is unclear unless we compare the performance of this method against others with a benchmark dataset.
- The values reported for the column *ENTRIESn_hourly* could be highly noisy, so it could be useful to do our analysis after removing outliers.
- The sample size in the real world context will be much more higher, so the method will also need to take into account performance and take advantage of a distributed environment.
- We only worked with one null hypothesis, but we could work with many others that I can imagine have a great potential for real world impact. For example: Does a particular unit gets more passengers than others? Does a particular date had more riders than usual?

5.2 (Optional) Do you have any other insight about the dataset that you would like to share with us?

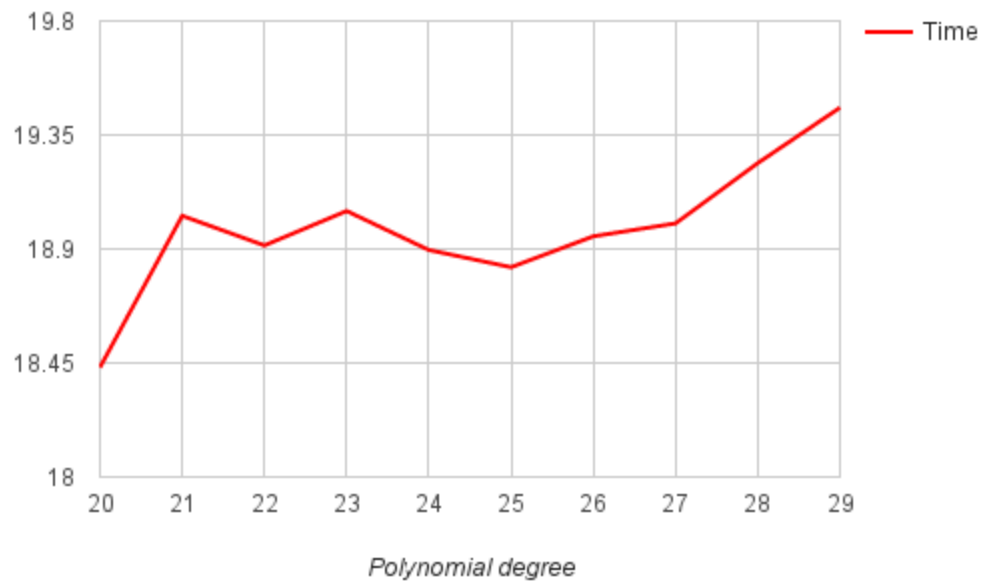
- According to the [graph](#) for the average amount of riders per hour, more riders ride the subway during 8pm, followed closely by 12pm. More analysis is required, but just by looking at the graph is easy to see that there's no clear linear relationship between the ridership and time of day, instead reflecting the real life behavior where we come to expect more riders during particular peak hours.

Section 6. Supporting Graphs

6.1 R^2 obtained for choices of polynomial degree in the interval $[20, 30)$ for the *hour* input feature



6.2 Execution time obtained for choices of polynomial degree in the interval [20, 30) for the *hour* input feature



6.3 Residuals plot obtained after applying the model to predict the value for the ENTRIESn_hourly column

