MSE 491 Lab 1 – Regression

Due Date: Feb 21

Written by

Cheng Jie “Michael” Huang

For

Amir Kabir

Yasaman Vaghei

Mohammad Nirimani

Table of Content

Table of Content 2
Table of Figures 2
1. Getting Started 3
1.3 3
1.4 3
A 3
B 4
C 4
2. Simple Linear Regression 4
2.1 4
2.2 4
2.3 5
3. Multiple Linear Regression 5
3.1 5
3.2 6
4. Feature Selection 6
5. Polynomial Regression 7
A 7
B 8

Table of Figures

Figure 1: NOx histogram with 100 bins 3
Figure 2: TIT vs TAT Scatter Plot 3
Figure 3: TIT vs TEY Scatter Plot 4
Figure 4: Effect of Training/Testing Split on RMSE for TIT vs CO 5
Figure 5: All Features Multiple Linear Regression RMSE Bar Chart for TEY and CO 6
Figure 6: Top Four Correlated Features Multiple Linear Regression RMSE Bar Chart for TEY and CO 7
Figure 7: RMSE and R2 Bar Charts for Different Targets at Different Degrees 8

1. Getting Started

1.3

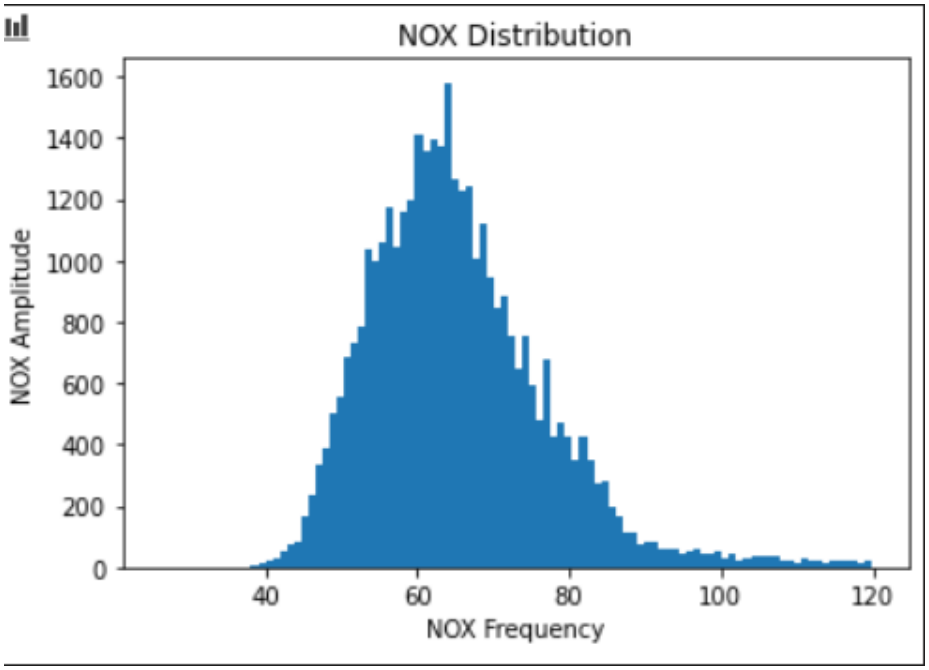

Figure 1: NOx histogram with 100 bins

Yes, the generated histogram of 100 bins resembles that of a normal distribution.

1.4

A

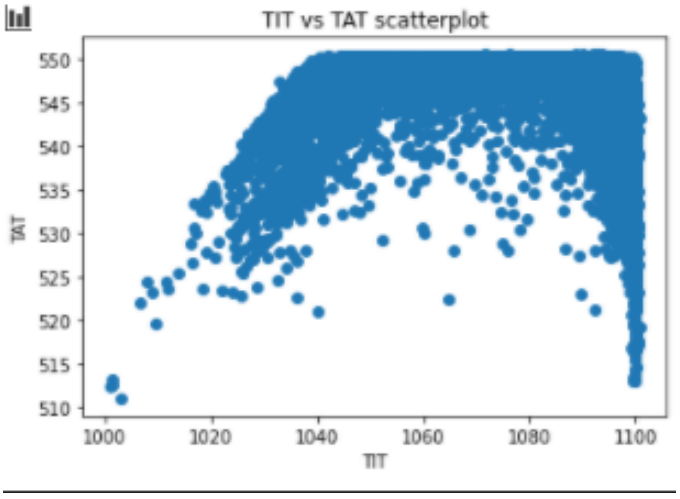

Figure 2: TIT vs TAT Scatter Plot

We can visually confirm that the association between TIT and TAT is not linear in nature by looking at the graph.

B

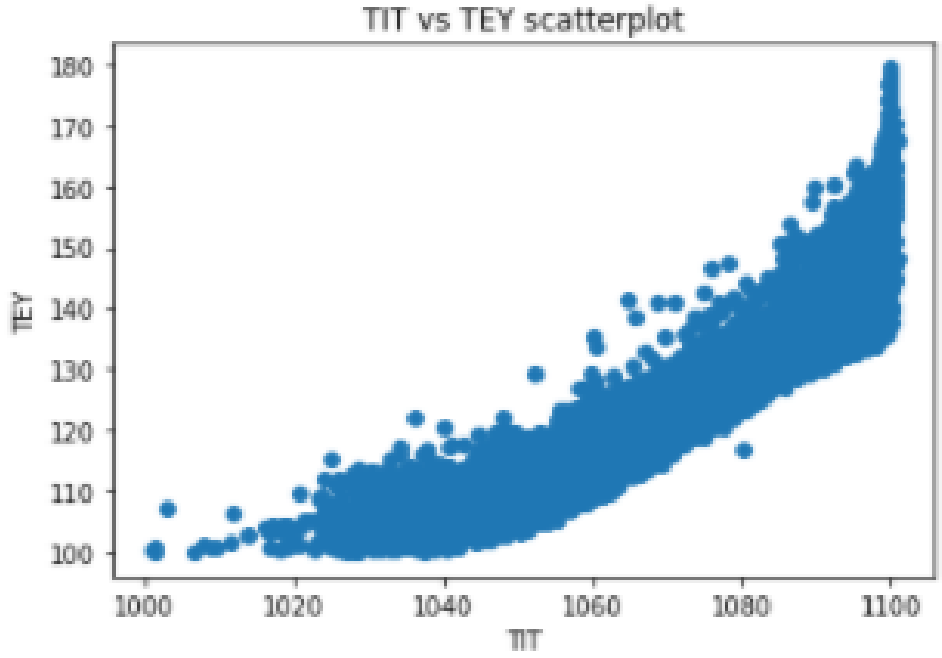

Figure 3: TIT vs TEY Scatter Plot

The scatterplot visually shows a linear association between TIT and TEY.

C

Pearson's correlation coefficient for A: -0.380862393620118

Pearson's correlation coefficient for B: 0.9102972462822895

Pearson correlation measures the linear association between datasets, thus a high magnitude closer to one would indicate a more linear association.

2. Simple Linear Regression

2.1

[CPD:TEY], [TIT:CO], and [AT:NOX] have the highest correlation.

2.2

The following are the RMSE & R2 value for the training and test pair of datasets with the highest correlation.

CDP , TEY rmse & r2 Test: (2.3811635588161972, 0.9772555954210641)

CDP , TEY rmse & r2 Train: (2.4095982438984658, 0.9760653675501568)

TIT , CO rmse & r2 Test: (1.7179512152301406, 0.42128932248941375)

TIT , CO rmse & r2 Train: (1.7263011690979155, 0.41845147092702306)

AT , NOX rmse & r2 Test: (9.665488809679491, 0.3134402878228443)

AT , NOX rmse & r2 Train: (9.71457485105685, 0.30839583892249256)

2.3

Below is the generated RMSE values for training & test datasets for 50% training, 70% training, 90% training:

Table 1: TIT vs CO Simple Linear Regression Results with Various Splits

%Test	50	30	10
Train RMSE	1.6955864241466287	1.7304605100382842	1.7285307758748658
Test RMSE	1.7514710055225535	1.7126026617535277	1.6891694062325164

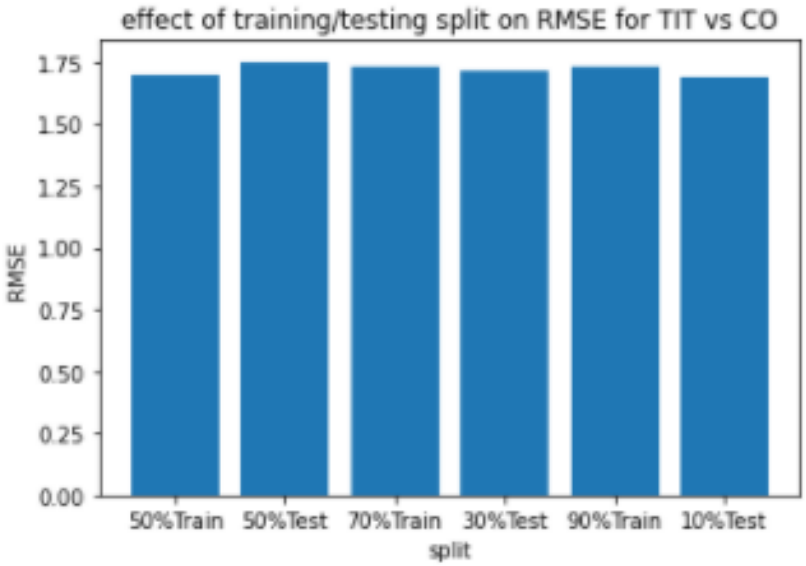

Figure 4: Effect of Training/Testing Split on RMSE for TIT vs CO

The difference in RMSE values presented by different splits in training and testing proportion is minimum, where the training RMSE results see a slight increase at 30% test data. While Test RMSE results decrease in magnitude with decrease in % test data.

3. Multiple Linear Regression

3.1

Table 2: All Features Multiple Linear Regression RMSE and R2 value for targets TEY and CO

	TEY	CO
RMSE	1.1404564262675947	1.5984369893226642
R2	0.9947826074895793	0.49900789339183194

Below is the RMSE value bar graph using all features:

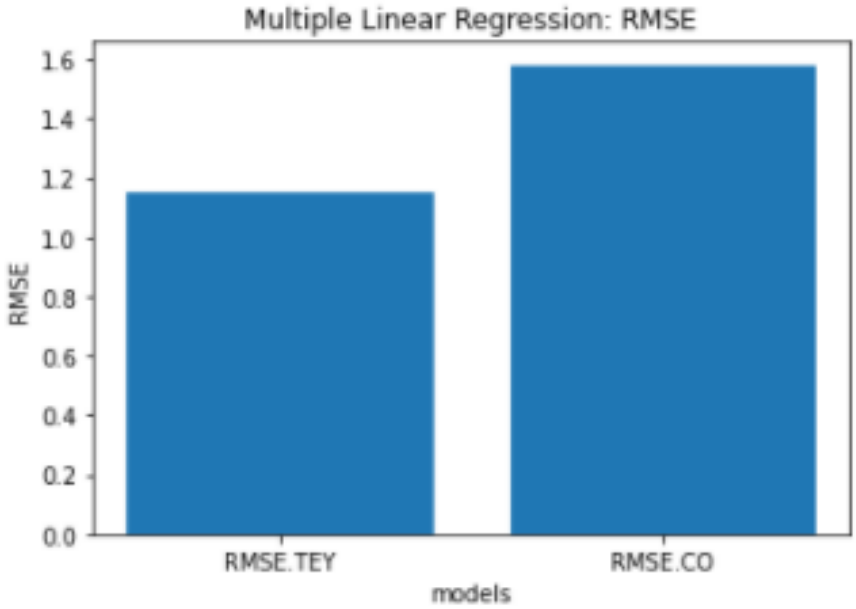

Figure 5: All Features Multiple Linear Regression RMSE Bar Chart for TEY and CO

3.2

Comparing the results of 2.2 and 3.1, we can see that using all features lowered the RMSE value and is therefore more accurate than a single feature prediction in this case. However, it should be noted that increasing the number of features used does not always improve the performance of the model.

4. Feature Selection

Table 3: Top Four Correlated Features Multiple Linear Regression RMSE and R2 value for targets TEY and CO

	TEY4 [CDP, GTEP, TIT, TAT]	CO4 [TIT, CDP, GTEP, AFDP]
RMSE	2.333135432701119	1.5984369893226642
R2	0.9781638526901284	0.49900789339183194

Comparing the RMSE and R2 evaluation metrics against the results discovered in 3.1 shows that using the top four features actually decreased the accuracy of prediction due to the higher RMSE and lowered R2 results.

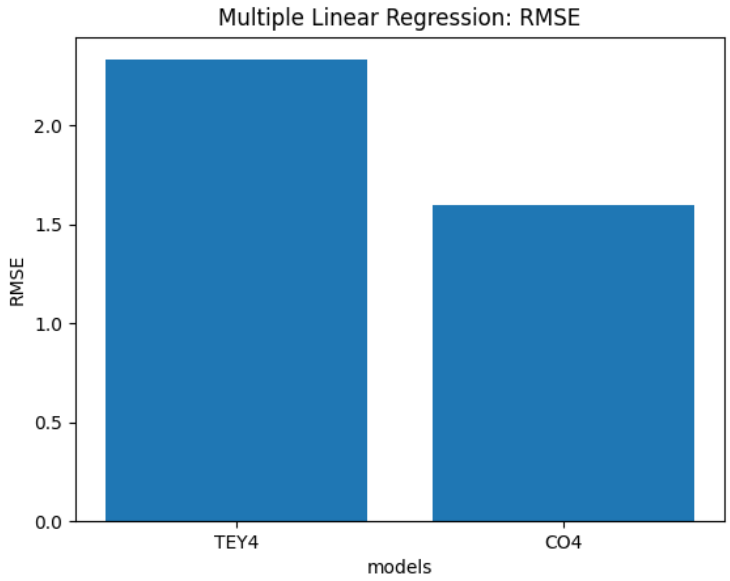

Figure 6: Top Four Correlated Features Multiple Linear Regression RMSE Bar Chart for TEY and CO

The graphics here indicates that using the highest 4 Pearson's coefficient results in higher RMSE than using all features combined, which seems counter intuitive as the four features selected should be the primary features more impactful on the dynamic of the predictive system.

5. Polynomial Regression

A

Due to the memory restraint placed upon both my personal PC as well as the school's lab PC on using all eight features with a 11-degree polynomial, below are the results of the polynomial regression in the format of [degree, target, RMSE, R2] using the 4 labels ['AT','AP','AH','AFDP'].

```
[2, 'TEY', 10.097880135559533, 0.5909687029084739]
[2, 'CO', 1.938386546014797, 0.26324920187797474]
[2, 'NOX', 8.770350366775538, 0.434718780382632]
[5, 'TEY', 9.676598578163519, 0.6243861760665472]
[5, 'CO', 1.8841930755224032, 0.30386952233866804]
[5, 'NOX', 8.314374669482765, 0.49196942225120266]
[11, 'TEY', 20.623572357078103, -0.706176305647344]
[11, 'CO', 4.44363979686184, -2.8718411987052392]
[11, 'NOX', 11.632259823248528, 0.005605167347127016]
```

Table 4: Ordered List of RMSE and R2 value for Targets based on Degrees

Degree	2	5	11
TEY RMSE	10.097880135559533	9.676598578163519	20.623572357078103
CO RMSE	1.938386546014797	1.8841930755224032	4.44363979686184
NOX RMSE	8.770350366775538	8.314374669482765	11.632259823248528
TEY R2	0.5909687029084739	0.6243861760665472	-0.706176305647344
CO R2	0.26324920187797474	0.30386952233866804	-2.8718411987052392
NOX R2	0.434718780382632	0.49196942225120266	0.005605167347127016

B

Below is the graphical comparison of the various RMSE and r2 values obtained for degrees 2,5, and 11 used in the polynomial regression case. Based on the fluctuations in value at the degree of 11 with large RMSE and r2 values, we can see that the data is overfitted. Ways to avoid overfitting includes regularization, feature number reduction, training data increase, as well as K-fold cross-validation. In this case the large degree of 11 used for regression is unnecessary, as lower degree regressions performed well in comparison. It is also interesting to consider that the R2 value provided by the sklearn.metrics kit provided a r2 value of -2.5 which should not be possible given the nature of the r2 function.

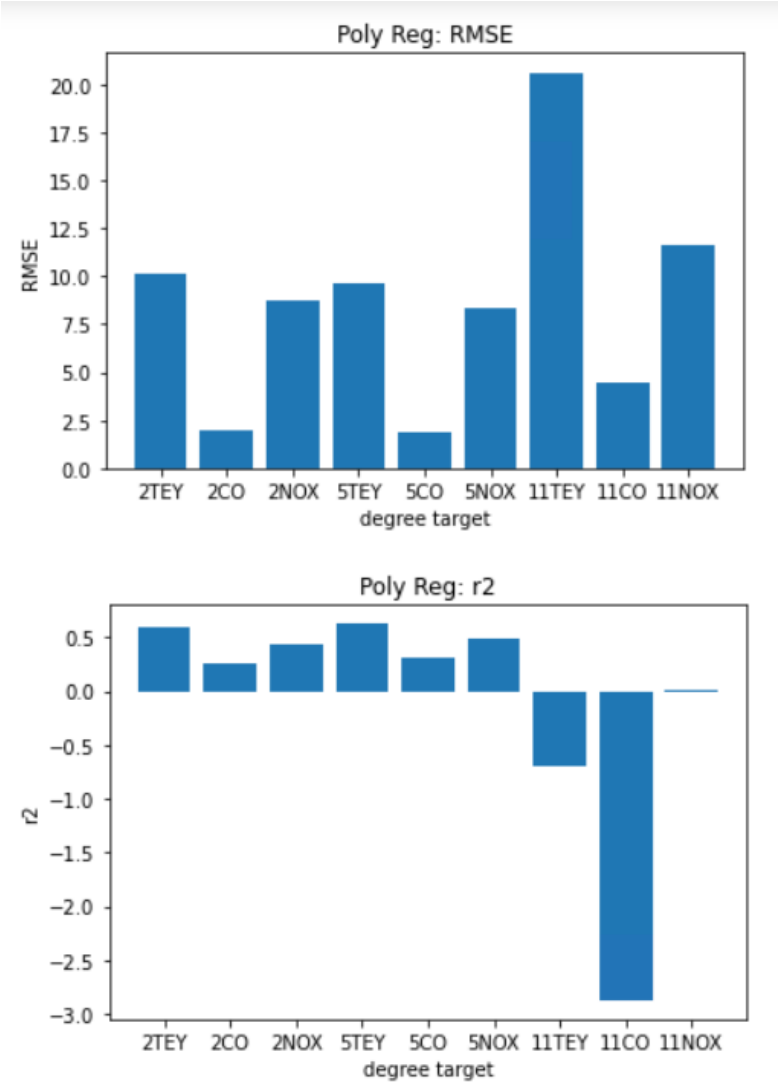

Figure 7: RMSE and R2 Bar Charts for Different Targets at Different Degrees