Questions

The Canada fuel consumption data set is an open public data. Various variables are listed in the CSV file. For the sake of global warming policies, CO2 emissions and sources that cause emission of CO2 need to be monitored in very nation.

(1) Investigate a **multiple linear regression** model fit in which the CO2 emission is the outcome variable. The last 20 observations of the data set are kept for the sake of testing the prediction accuracy.

1.1. Plot the CO2 emission

1.2. Construct the best model with all relevant factors that influence the increase of CO2 in Canada.

1.3. Conduct the regression diagnosis and discuss about the goodness of fit

1.4. Provide the corresponding Anova Table

1.5. Write the prediction model

1.6. How is your prediction model with respect to the testing data?

Variable Description

There are totally 1085 observations and 8 variables in the data. One of them is categorical, and the rest is continuous. The type and the levels/range of each variable is summarized as follows:

Variable type	Variable name	Range/Levels
Categorical	Fuel type	D ` Z
Continuous	Cylinders	2 ~ 16
Continuous	Fuel consumption city (l/100km)	4.7 ~ 30.3
Continuous	Fuel consumption hwy (l/100km)	4.9 ~ 20.8
Continuous	Fuel consumption comb (l/100km)	4.8 ~ 25.8
Continuous	Fuel consumption comb (mpg)	11 ~ 59
Continuous	CO2 emissions (g/km))	110 ~ 547

The Data set is divided into two sub-sets: Data_Train, and Data_Test. The Data_Train includes the first 1065 observations, and Data_Test includes the last 20 observations.
Next, we remove missing data and replace outliers in the train data sub-set by value at a quantile of 95%. (Figure 1)

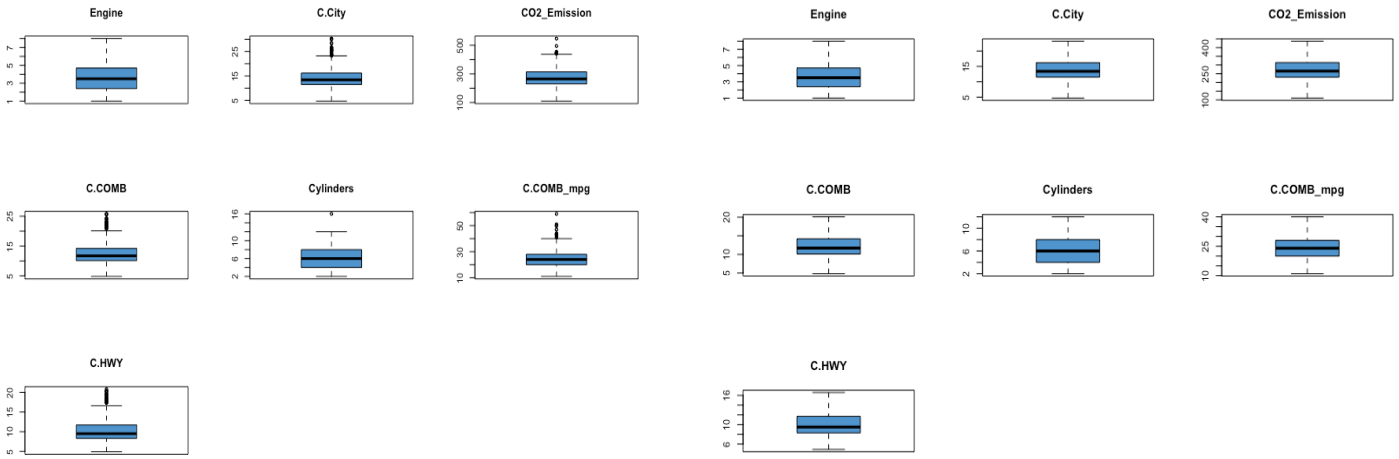


Figure 1. Data with outliers vs. Data without outliers

Descriptive Statistic

Variable	Mean	SD	Engine	Cylinders	City	HWY	COMB	COMB_mpg	CO2_E
Engine	3.588	1.357	1						
Cylinders	6.047	1.861	0.908	1					
City	14.17	3.805	0.803	0.763	1				
HWY	10.110	2.470	0.759	0.696	0.965	1			
COMB	12.350	3.185	0.794	0.745	0.995	0.985	1		
COMB_mpg	24.180	6.084	-0.800	-0.743	-0.934	-0.897	-0.928	1	
CO2_E	273.800	61.153	0.850	0.825	0.879	0.843	0.873	-0.890	1

Table 1. Summary statistics and correlation matrix

We have scatter plots of CO2 Emission versus other variables is as follows:

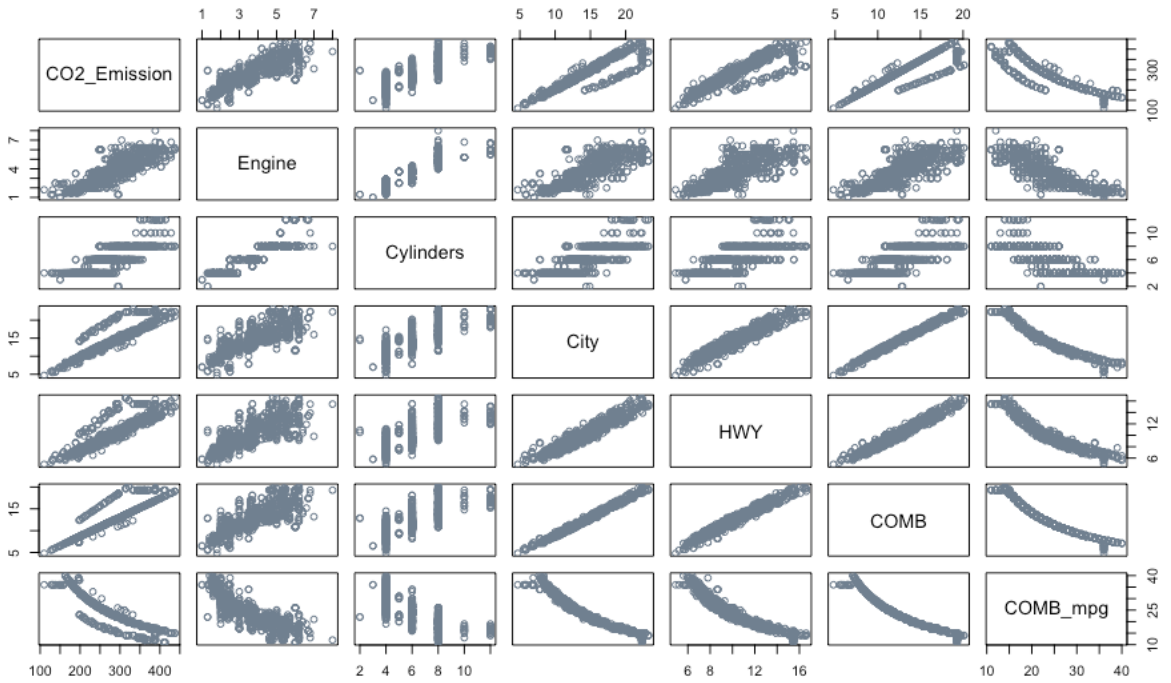


Figure 2: A pairs plot for the numeric variables these dataset

Graphically, there are linear relationships between: Engine and CO2_Emission,
Cylinders and CO2_Emission,
City and CO2_Emission,
HWY and CO2_Emission,
COMB and CO2_Emission.

1. CO2 Emission Plot

We have the scatter plot of CO2 Emission is as follows:

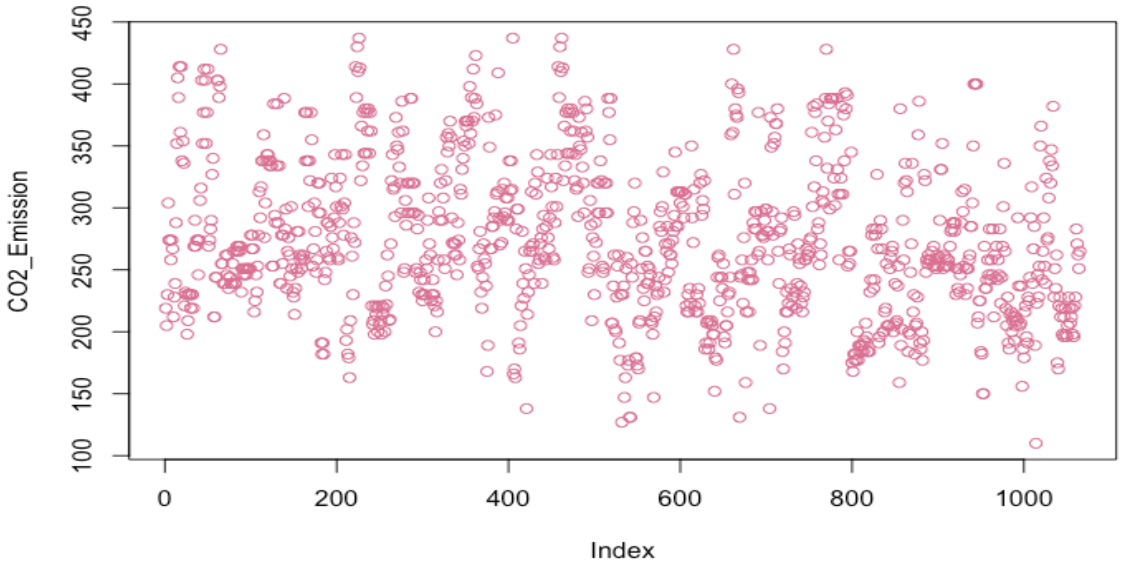


Figure 3: CO2 Emission Scatter Plot

It seems that CO2 Emission is not clustered and distributes spread

2. Proposal Multiple Linear Regression Model

Dependent Variable: CO2_Emission
Independent Variables: Engine, Cylinders, Fuel Type, and COMB

Motivation:

+ Engine, Cylinders, Fuel Type, and COMB graphically has a linear relationship with CO_Emission

+ City, HWY, and COMB has a strong correlation with each other; simultaneously, the information in COMB overlaps City and HWY since COMB combines the fuel consumption of City and HWY. If we give all variables into the model, multicollinearity could be a big challenge.

Proposal Model:

$$CO2\_Emission = \beta_0 + \beta_1 Engine + \beta_2 Cylinders + \beta_3 Fuel\ Type + \beta_4 Comb + u$$

We have estimation results from the regression models as follows:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   52.9510     2.8449  18.613  < 2e-16 ***
Engine         0.5973     0.6044   0.988    0.323
Cylinders      1.8487     0.4073   4.539 6.29e-06 ***
COMB          19.9006     0.1974 100.800  < 2e-16 ***
FuelE       -141.8118     3.0323 -46.766  < 2e-16 ***
FuelX        -30.3748     2.6164 -11.609  < 2e-16 ***
FuelZ        -29.9924     2.6265 -11.419  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 4: Estimation results from regression model

$R^2 = 0.9756$, adj-$R^2 = 0.9755$
As the result above, except for Engine, all variables are statistically significant. Insignificance of Engine means we have no prove to show there is relationship between Engine and CO2 Emission at the population level.

At the categorical variable – Fuel Type, it can be seen that:

- Being from type E is significantly associated with an average decrease of 141.8118 in CO2 Emission compared to Type D
- Being from type X is significantly associated with an average decrease of -30.3748 in CO2 Emission compared to Type D
- Being from type E is significantly associated with an average decrease of 29.9924 in CO2 Emission compared to Type D

It seems that Type E has least CO2 emission relatively.

Model Adequacy Test:

$$H0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$
$$H1: \beta_1^2 + \beta_2^2 + \beta_3^2 + \beta_4^2 \neq 0$$

We have P_value = 2.2e-16 < 0.05 → At the level of 5, we reject H0. There is at least one variable is statistical significance and the model has an appropriate structure.

3. Regression diagnosis and the goodness of fit

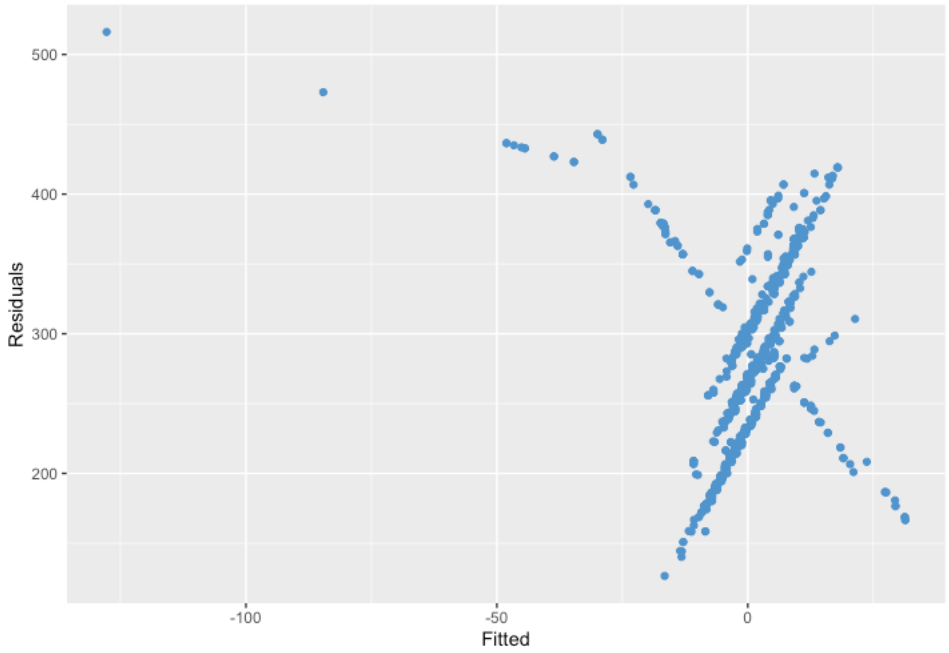


Figure 4: Residuals plot

The residuals vs. fitted values plots suggest that the mean of residuals possibly is not zero and there are several alternative patterns in data that indicate heteroscedastic. The model will become more challenging if assumption 1 (Mean of the error term is expected to be zero) and assumption 2 (variance of the error term is constant) of General Linear Model assumptions are violated.

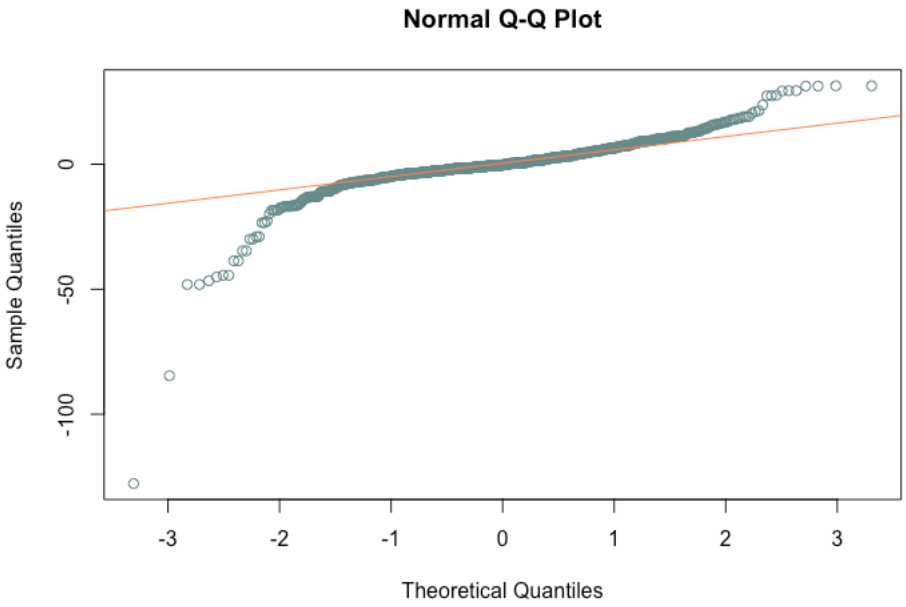


Figure 5: QQ Plot

The Q-Q plot shows that the data possibly is normally distributed as almost data is not far from the line. We can use this model to predict CO2 Emission.

The total variance in data is derived from 2 sources: systematic variation (~predicted value) and unexplained variation (residuals).

Source of variation	Sum of Squares (SS)
Model	SSM = 145241.4
Residual	SSE = 3882367
Total	SST = 4027608

Table 2: Analysis of Variance Table

4. ANOVA Table

```
Analysis of Variance Table

Response: CO2_Emission
            Df  Sum Sq Mean Sq  F value    Pr(>F)
Engine       1 2889100 2889100 31493.81 < 2.2e-16 ***
Cylinders    1   54005   54005   588.71 < 2.2e-16 ***
COMB         1  387884  387884  4228.29 < 2.2e-16 ***
Fuel         3  550958  183653  2001.98 < 2.2e-16 ***
Residuals 1058   97056      92
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 3: ANOVA Table

As results above, all of the variables are statistically significant, which indicates that all variables have the power to explain the dependent variable. In other words, all variables can be used to predict the dependent variable.

5. Write Prediction Model

Based on the estimation result, we have the prediction model is as follows:

$$\widehat{CO2\ Emission} = 52.9510 + 0.9573\ Engine + 1.8487\ Cylinders + 19.9006\ COMB - 141.8118\ Fuel\ Type\ E - 30.3748\ Fuel\ Type\ X - 29.9924\ Fuel\ Type\ Z$$

=

$$\begin{cases} 52.9510 + 0.9573\ Engine + 1.8487\ Cylinders + 19.9006\ COMB \\ \quad \text{if Fuel E = Fuel X = Fuel Z = 0, Fuel Type = D} \\ 52.9510 + 0.9573\ Engine + 1.8487\ Cylinders + 19.9006\ COMB - 141.8118 \\ \quad \text{if Fuel Type = E} \\ 52.9510 + 0.9573\ Engine + 1.8487\ Cylinders + 19.9006\ COMB - 30.3748 \\ \quad \text{if Fule Type = X} \\ 52.9510 + 0.9573\ Engine + 1.8487\ Cylinders + 19.9006\ COMB - 29.9924 \\ \quad \text{if Fuel Type = Z} \end{cases}$$

6. Model Accuracy

We conduct out-of-sample prediction for 20 observations and here is the result.

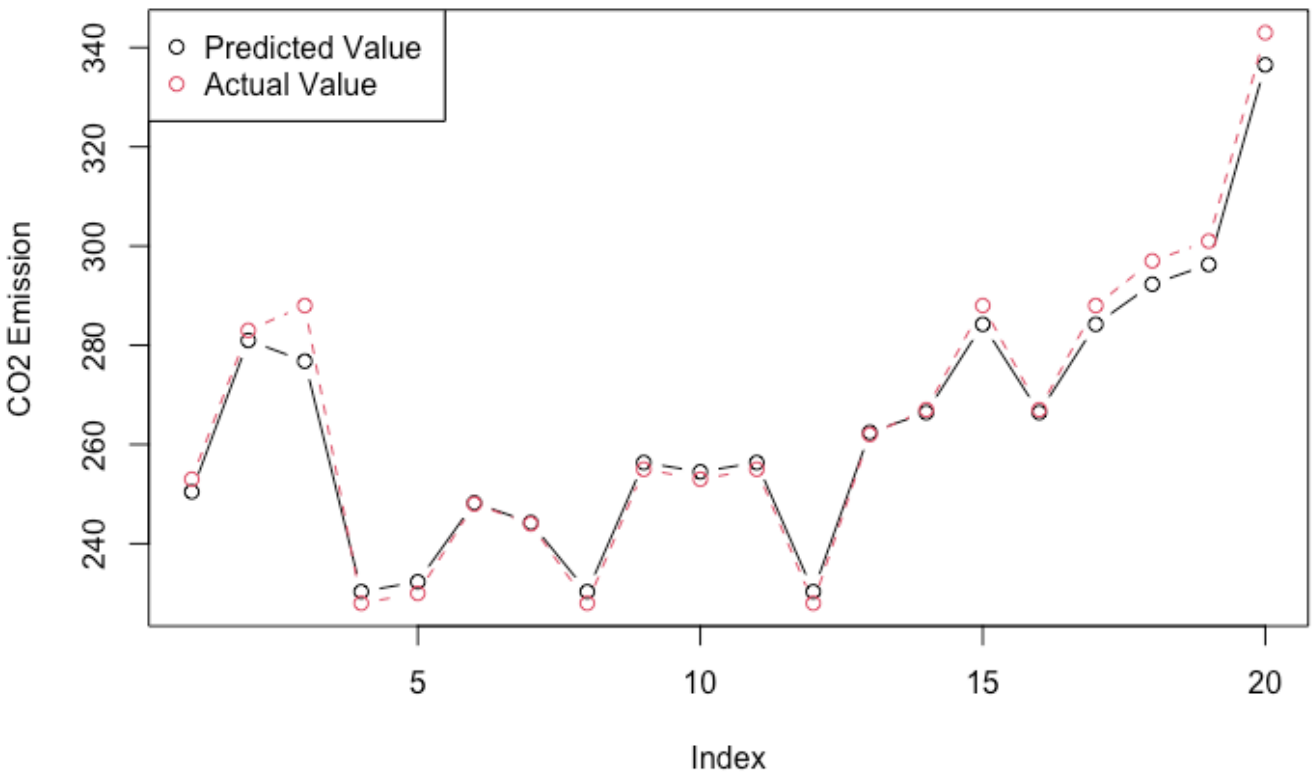


Figure 6: Predicted Value vs. Actual Value

We can see that, all predicted values are very close to actual value, which indicates that our model is good.

(2) Investigate a **polynomial linear regression** model fit in which the CO2 emission is the outcome variable. We seek to examine the relationship between the CO2 emission and Engine size.

2.1. Plot CO2 Emission vs Engine size. Do you think a polynomial regression can be suggested? If yes, fit the data set to an appropriate polynomial regression model by using orthogonal polynomial

2.2. Discuss about the goodness of fit and the highest degree for the postulated model.

2.3. Write the prediction model

2.4. Compare actual vs fitted values by using graphical

2.1 Model Fitting

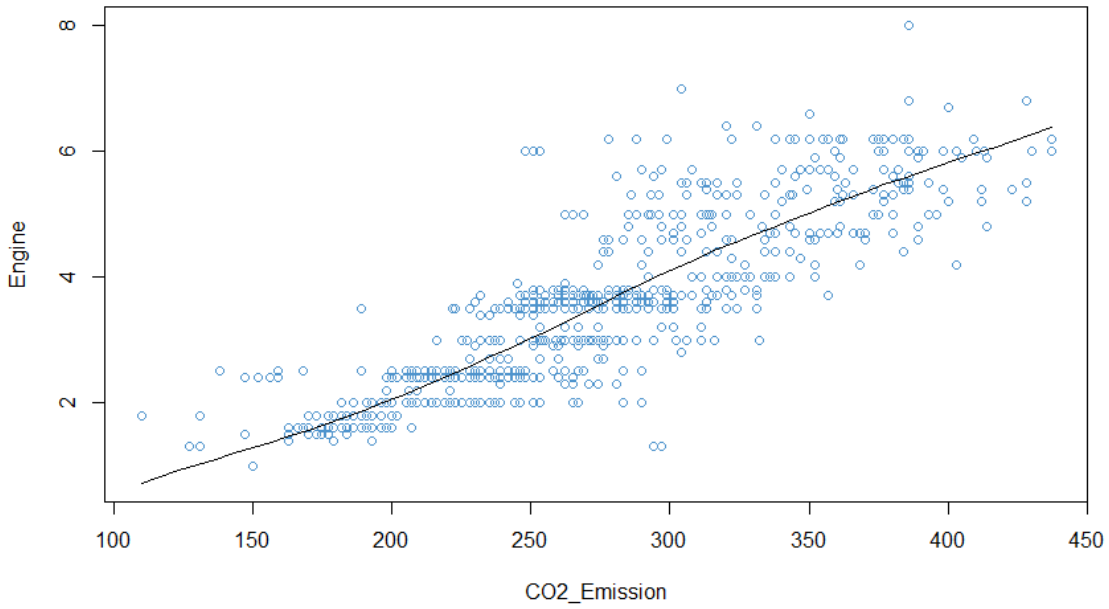


Figure 1: Engine vs. CO2 Emission

The scatter plot of Engine against CO2 Emission shows different patterns, curvilinear effects appear in data rather than linear relationship.

- The polynomial linear regression is suggested.

A general polynomial regression can be expressed as:

$$Y_{i1} = \beta_0 + \sum_{j=1}^{k} \beta_j X_i^j + \varepsilon_i \quad i = 1,2,\ldots,n$$

	(Intercept)	Engine	I(Engine^2)	I(Engine^3)	I(Engine^4)
(Intercept)	1.0000000	-0.9893244	0.9635786	-0.9277934	0.8837649
Engine	-0.9893244	1.0000000	-0.9916235	0.9693631	-0.9355955
I(Engine^2)	0.9635786	-0.9916235	1.0000000	-0.9926038	0.9715223
I(Engine^3)	-0.9277934	0.9693631	-0.9926038	1.0000000	-0.9928577
I(Engine^4)	0.8837649	-0.9355955	0.9715223	-0.9928577	1.0000000

Figure 2: Correlation between variables

A big challenge of Polynomial Linear Regression is multicollinearity. Figure 2 shows a high correlation between Engine and a higher degree of Engine. Otherwise, we don't know which degree of a polynomial is appropriate at the beginning, so we need to fit the model by trying to add new terms, and for each tentative, modeling will start from scratch. To handle this situation, we apply the orthogonal polynomial model to the data.

Orthogonal Polynomial model:

$$Y_{i1} = \alpha_0 \emptyset 1(\mathrm{Xi}) + \alpha_1 \emptyset 1(\mathrm{Xi}) + \alpha_2 \emptyset_2(\mathrm{X}_2) + \ldots + \alpha_k \emptyset_k(\mathrm{X}_i) + \varepsilon_i \quad \text{with } i = 1,2,\ldots,n$$

Proposal Model

- Model 1: $Y_i = \alpha_0 + \alpha_1 \emptyset_1(X_i) + \varepsilon_I$ #Linear
- Model 2: $Y_i = \alpha_0 + \alpha_1 \emptyset_1(X_i) + \alpha_2 \emptyset_2(X_i) + \varepsilon_I$ #Quadratic
- Model 3: $Y_i = \alpha_0 + \alpha_1 \emptyset_1(X_i) + \alpha_2 \emptyset_2(X_i) + \alpha_3 \emptyset_3(X_i) + \varepsilon_I$ #Cubic
- Model 4: $Y_i = \alpha_0 + \alpha_1 \emptyset_1(X_i) + \alpha_2 \emptyset_2(X_i) + \alpha_3 \emptyset_3(X_i) + \alpha_4 \emptyset_4(X_i) + \varepsilon_i$ #Quartic

We have the estimation result is as follows:

Variable	**Model 1**	**Model 2**	**Model 3**	**Model 4**
Intercept	273.80075 (0.98119) (2.2e-16)***	273.80075 (0.96326) (2.2e-16)***	273.80075 (0.96326) (2.2e-16)***	273.80075 (0.96326) (2.2e-16)***
$\emptyset_1$**(Engine)**	1699.73519 (32.02044) (2.2e-16)***	1699.73519 (31.43541) (2.2e-16)***	1699.73519 (31.43541) (2.2e-16)***	1699.73519 (31.43541) (2.2e-16)***
$\emptyset_2$**(Engine)**		-201.12386 (31.43541) (2.356e-10)***	-201.12386 (31.43541) (2.356e-10)***	-201.12386 (31.43541) (2.356e-10)***
$\emptyset_3$**(Engine)**			-54.90466 (31.40502) (0.08071)+	-54.90466 (31.40502) (0.08071)+
$\emptyset_4$**(Engine)**				6.71819 (31.41915) (0.83072)

Table 1: Estimation Result

Significance code: '***': 0.001
'**': 0.01
'*': 0.05
'+': 0.1

According to result above, $\emptyset_3$Engine and $\emptyset_4$Engine are insignificant at the level of 5, so we choose Model 2 is as the best model.

2.2 The goodness of fit

Source of variation	Sum of Squares (SS)
Model	SSM/SSR = 2930231
Residual	SSE = 1161962
Total	SST = 4092192

We construct a Baseline model and obtain SSE of 4244452 which is far greater than SSE of model 2. Otherwise, with model 2 we have $R^2 = 73.63\%$. and Adj-$R^2 = 73.58\%$

F-Test

H0: $\alpha_1 = \alpha_2 = 0$

H1: $\alpha_1 + \alpha_2 \neq 0$

According to estimation result, P_Value = 2.2e-16 < 0.05 → at the level of 5, we reject H0 → Model 2 has appropriate structure.

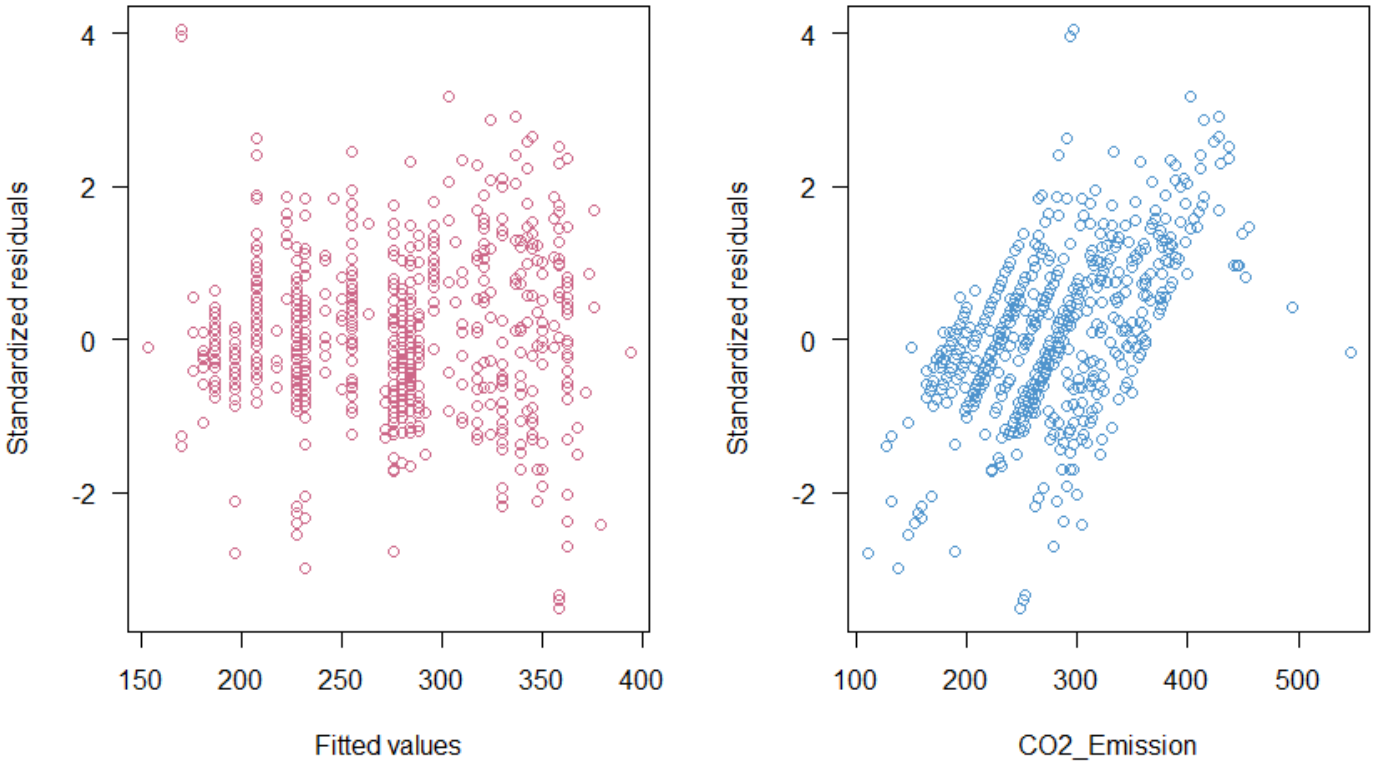


Figure 3: Standard Residuals

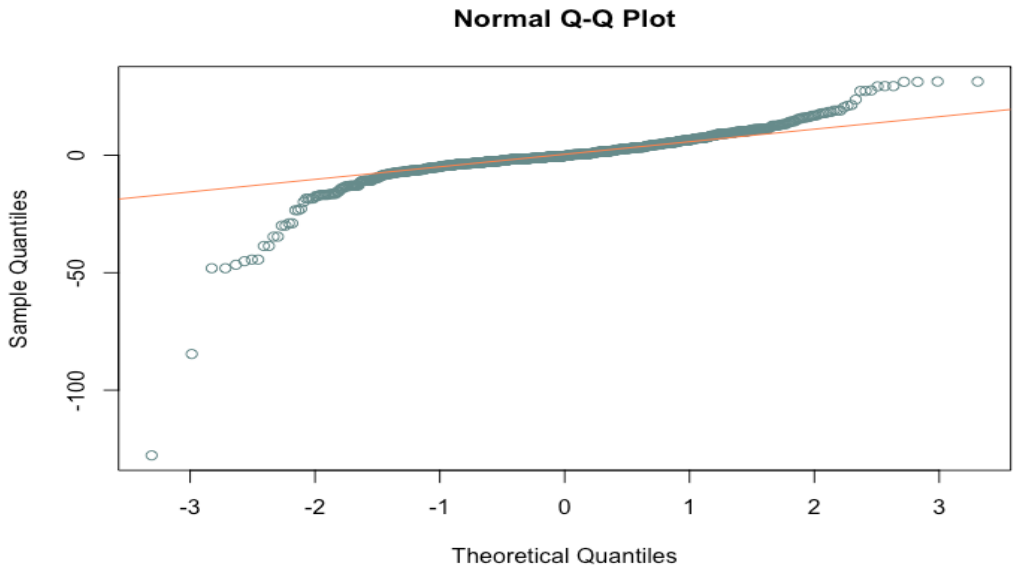


Figure 4: QQ Plot

We can see that, almost data lies on the line

2.3 Prediction Model

According to result above, we have prediction model is as follows:

$$\widehat{Y\imath} = \alpha 0 + 1699.73519 \emptyset 1(Xi) - 201.12386\ \emptyset 2(Xi) + \varepsilon i$$

2.4 Fitted Value and Actual Value

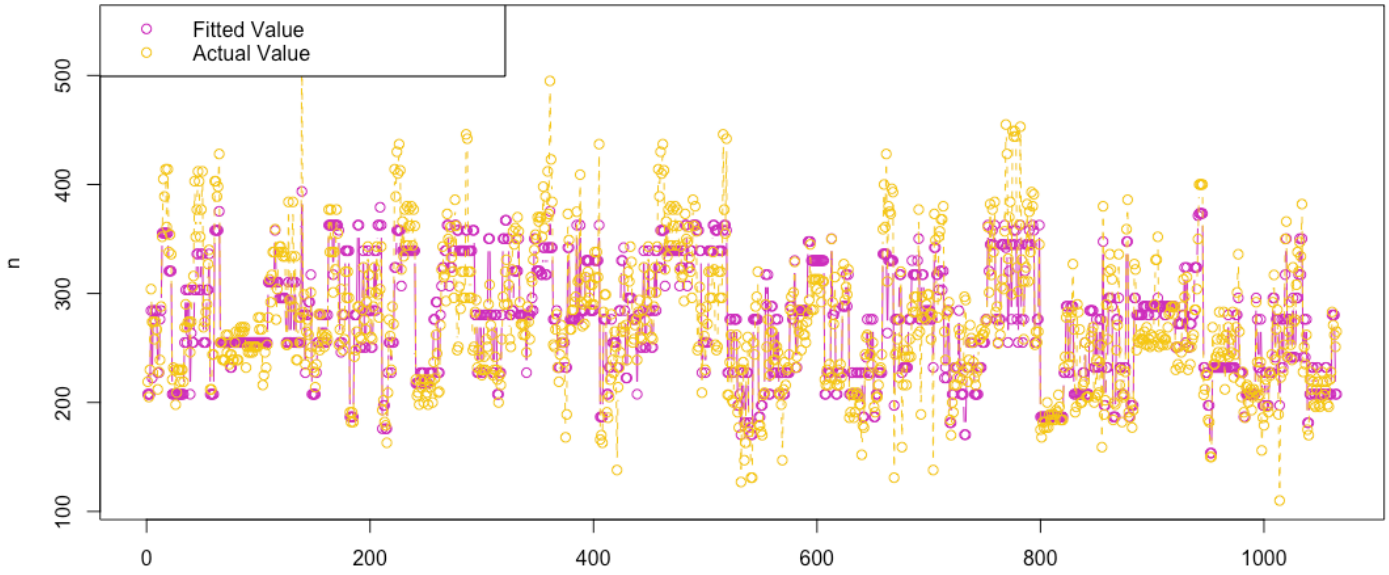


Figure 5: Fitted Value vs. Actual Value Graph

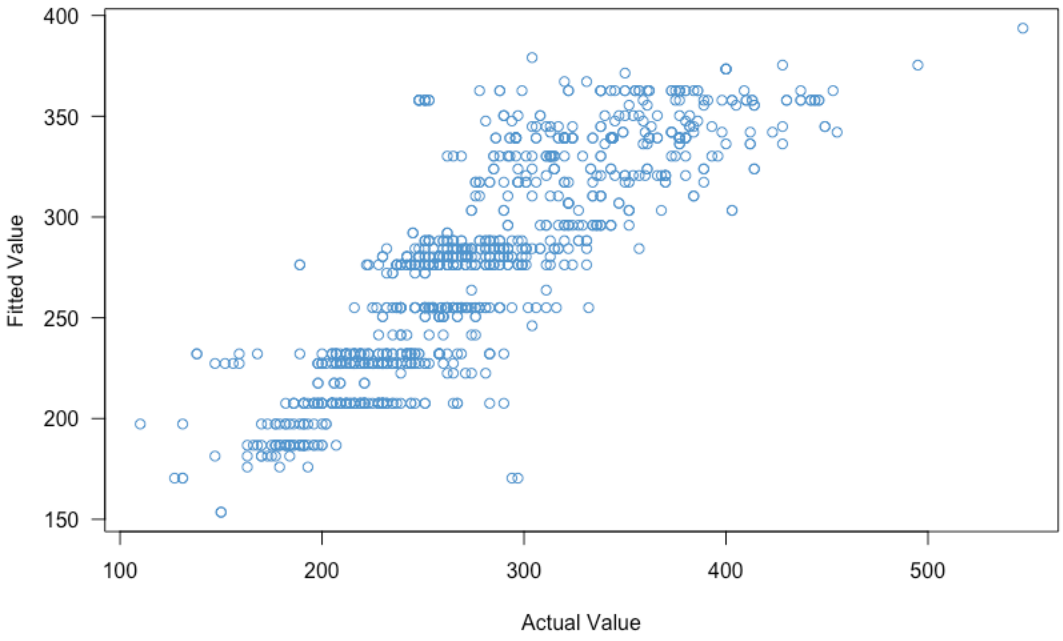
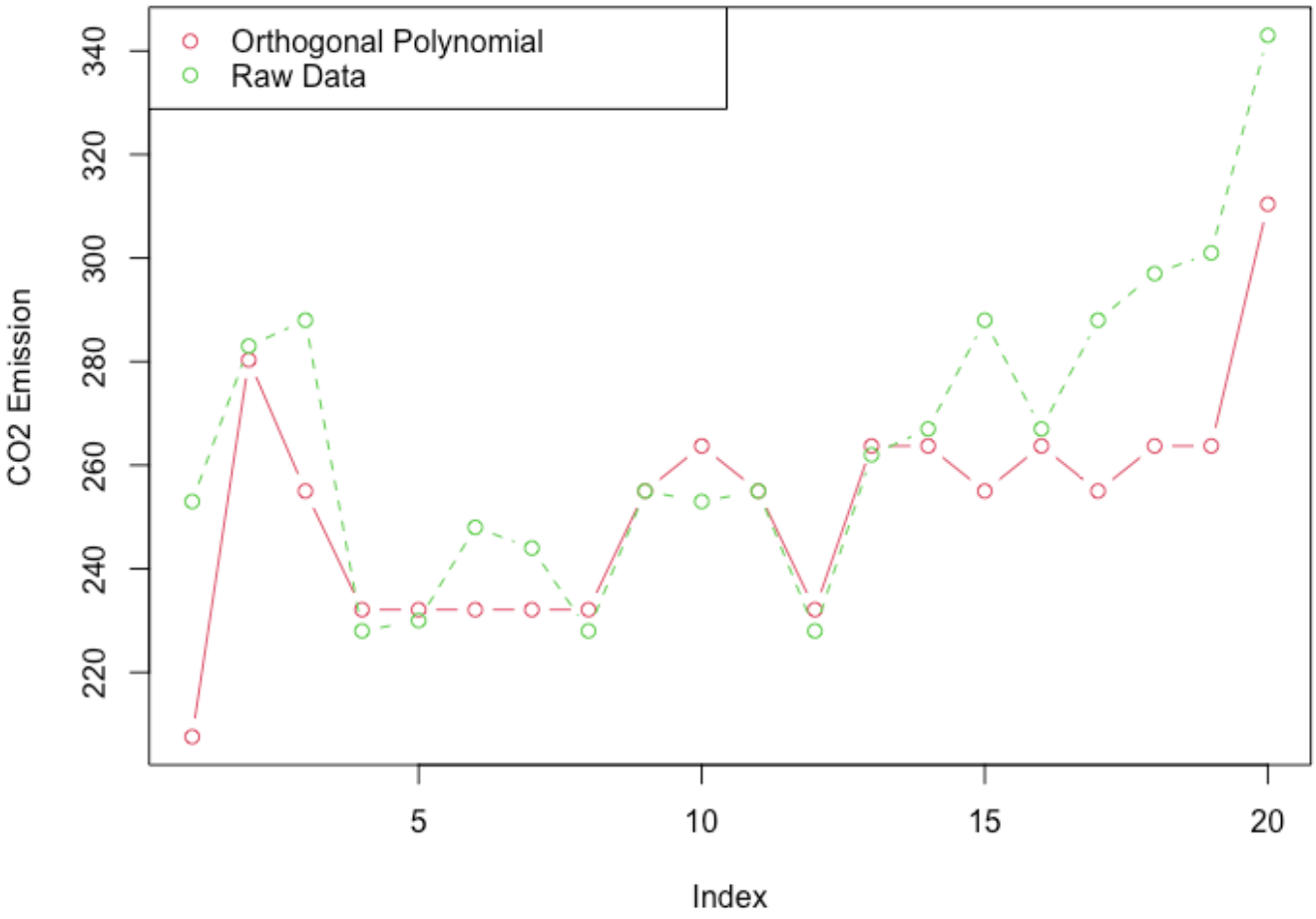


Figure 6: Scatter plot of Fitted Value and Actual Value

We can see from figure 5 that, the fitted value is very near to actual value

We predict out-of-sample for the last 20 observation in the whole dataset. The result is as follows:



(3) Investigate the possibility of building a response surface model model in which the CO2 emission is the outcome variable Engine size and Cylinder are regressors

3.1. Write down the response surface model

3.2. Does the response surface make sense? why?

3.3. Fit the model to the data and discuss about the goodness of fit

3.1 The response surface model

The response surface model fits a polynomial regression model with cross-product terms of variables that may be raised up to the third power.

The first order model:

$$Yj = \beta_0 + \beta_1 X1j + \beta_2 X2j + \beta_3 X1j X2j + \varepsilon_j$$

The second order model:

$$Yj = \beta_0 + \beta_1 X1j + \beta_2 X2j + \beta_3 X1j X2j + \beta_4 X^2 1j + \beta_5 X^2 2j \quad \varepsilon_j$$

3.2 Does response surface model make sense

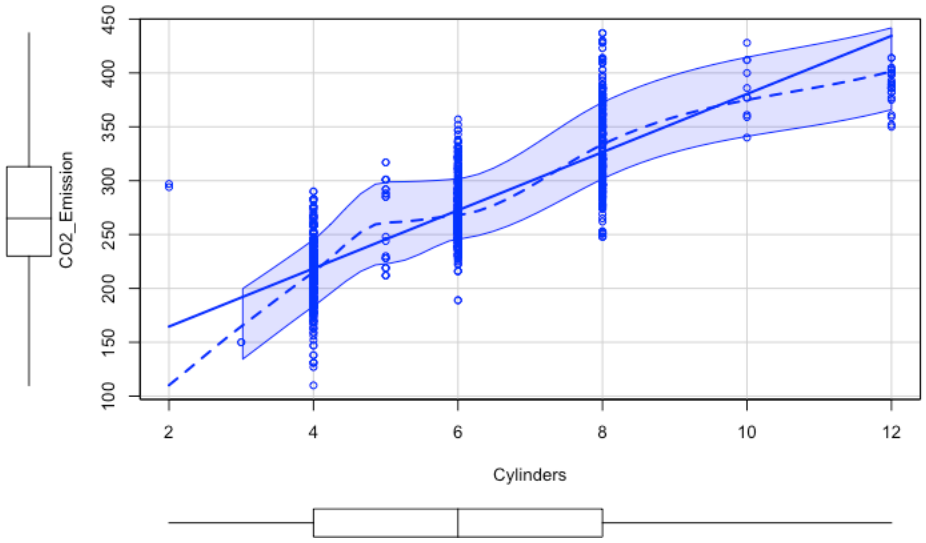


Figure 1: CO2 Emission vs Cylinders

From figure 1, we can see that the relationship between CO2 Emission and Cylinder is nonlinear. Otherwise, based on question 2, we find a second-order polynomial regression to explain the relationship between CO2 Emission and Engine. Hence, the response surface model makes sense because the main goal of response surface analysis is to find a polynomial approximation of the true nonlinear model.

There are three proposal models with the estimation results below:

Variable	Model 1	Model 2	Model 3
Intercept	92.1369 (9.6788) (2e-16) ***	103.6070 (10.1869) (2.2e-16)***	100.8780 (9.5653) (2e-16) ***
Engine	33.9473 (2.7394) (2e-16) ***	76.7432 (7.4125) (2.2e-16) ***	78.5400 (7.0441) (2e-16) ***

Cylinders	15.0194 (2.1369) (3.69e-12) ***	-11.8298 (4.4567) (0.008063) **	-11.9057 (4.4549) (0.00764) **
Engine:Cylinders	-1.2848 (0.3928) (0.00111) **	-5.1172 (2.8697) (0.074838)	-7.2291 (0.9496) (5.85e-14) ***
Engine^2		-1.4955 (1.9176) (0.435633)	3.4151 (0.4988) (1.27e-11) ***
Cylinders^2		2.7808 (0.9541) (0.003635) **	

3.3 Goodness of Fit

Model 1 report R^2 = 75.2%, Adj-R^2 = 75.08%
Model 2 report R^2 = 75.32%, Adj-R^2 = 75.2%
Model 3 R^2 = 75.18%, Adj-R^2 = 75.09%

All model report P_Value = 2.2e-16, which means all above models have appropriate structure. However, at model 2, Engine^2 and interaction Engine with Cylinders report insignificances.

Here is variance analysis of model 2:

Source of variation	Sum of Squares (SS)
Model	SSM/SSR = 3001594
Residual	SSE = 1093605
Total	SST = 4095199