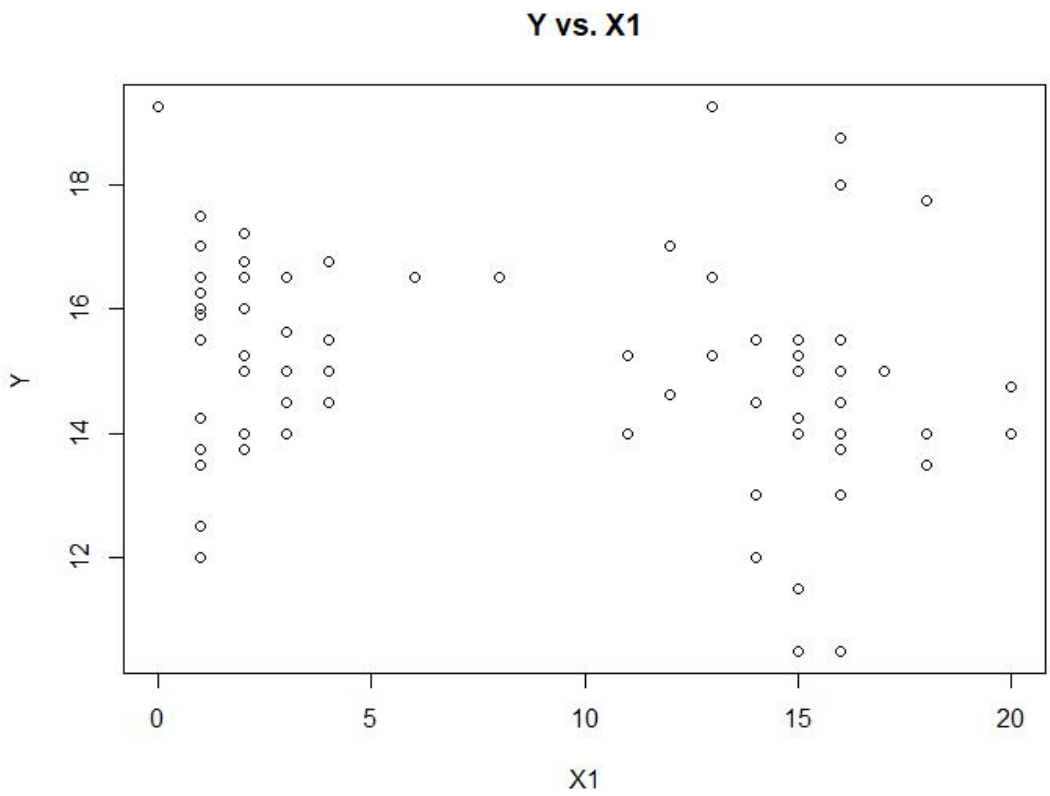
2、

(1)

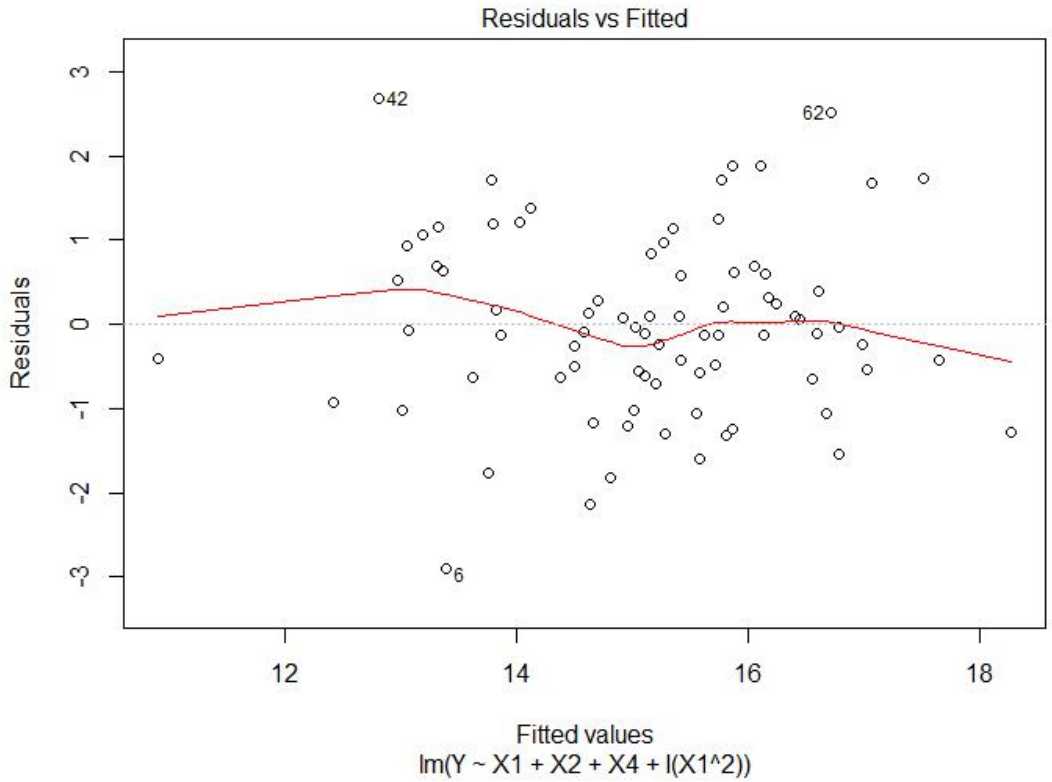


There is no obvious relationship between X1 and Y

(2)

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \beta_{11} X_1^2 + \xi_i$$

Y = 10.19 - 0.1818*X1 + 0.314*X2 + 8.046*e^-6*X4 + 0.01415*X1^2



There is no non-linearity and the variance is the same. So, it is a good fit.

(3)

In model 2 , the R^2 = 0.583, Ra^2 = 0.5667

In this model , the R^2 = 0.6131, Ra^2 = 0.5927

This model fits better than the model 2.

(4)

$$H0: \ \beta_{11} = 0 \qquad H1: \ \beta_{11} \neq 0$$

$$T^* = \frac{\hat{\beta}_{11}}{s\{\hat{\beta}_{11}\}}$$

The null distribution is T distribution.

Reject condition : pValue < 0.05

pValue = 0.0174, under the significant of 0.05, we should reject the null hypothesis.

(5)

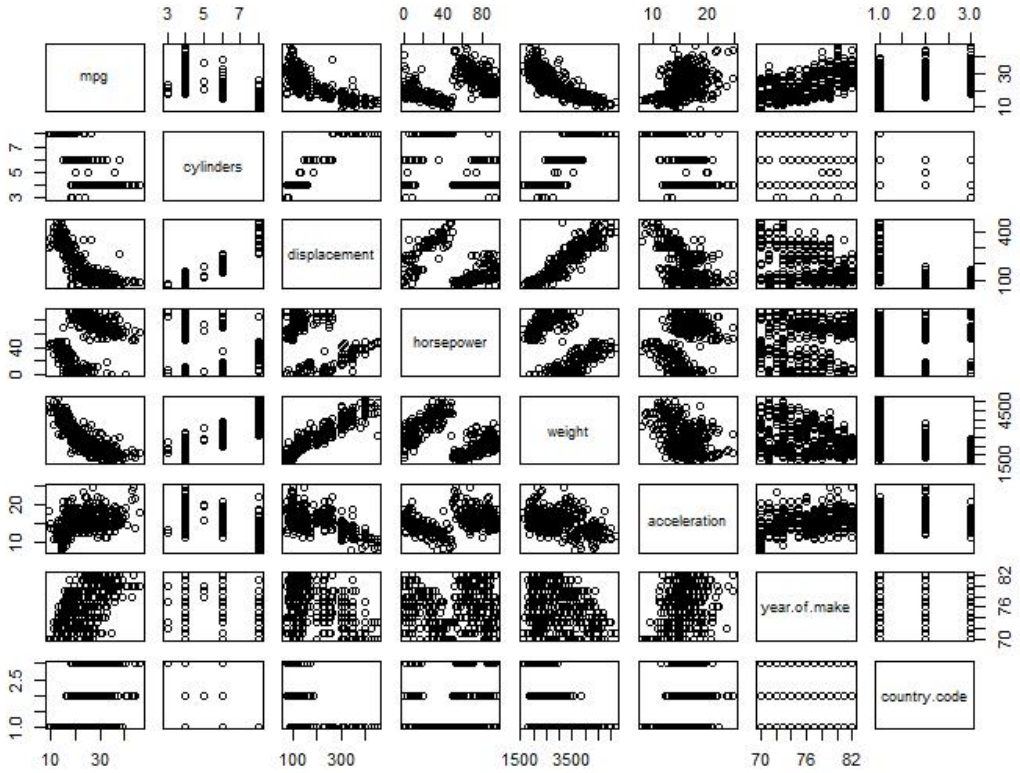
The prediction value is 14.88699

```
> predict(fit5,newdata = newData,interval = "prediction")
       fit      lwr      upr
1 15.11985 12.83659 17.40311
> predict(fit,newdata = newData1,interval = "prediction")
       fit      lwr      upr
1 14.88699 12.66453 17.10946
```

The prediction interval is smaller than the model 2.

3、

(1)



The cylinders variable, year of make and country.code variable are qualitative variables.

(2)

The cylinders variable, year of make and country.code variable are qualitative variables.

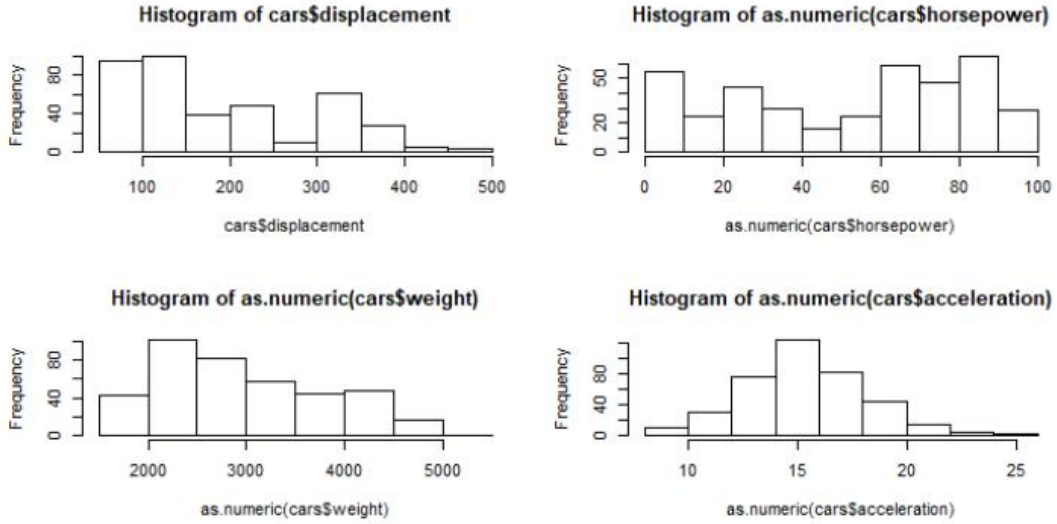
The others are quantitative variables.

(3)

remove the Na value.

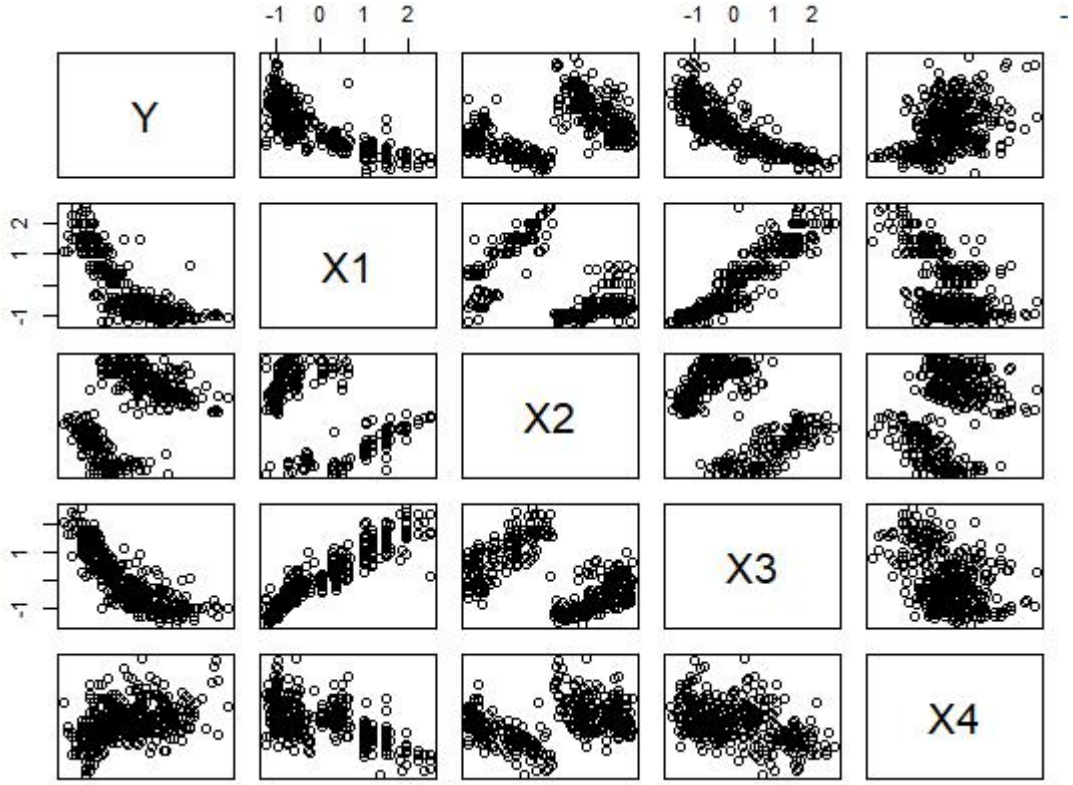
We should also transform those qualitative variables to indicator variables.

(4)



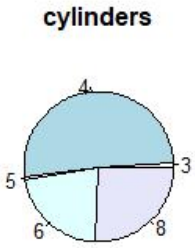
I think we need Z-Score transformation. Because we need to exclude the influence of units.

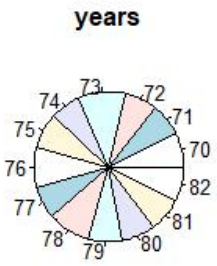
(5)

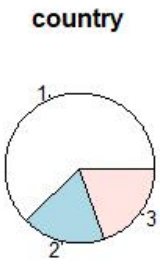


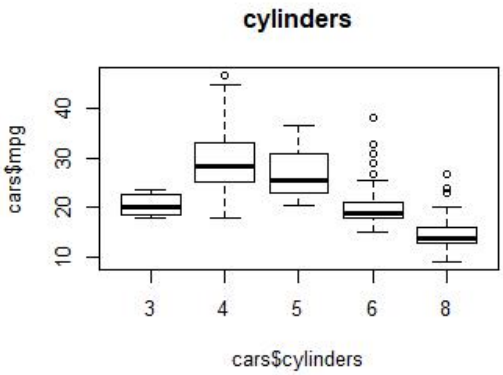
The displacement, horsepower and the weight are nonlinear relationship with response variable. We can transform response variable .

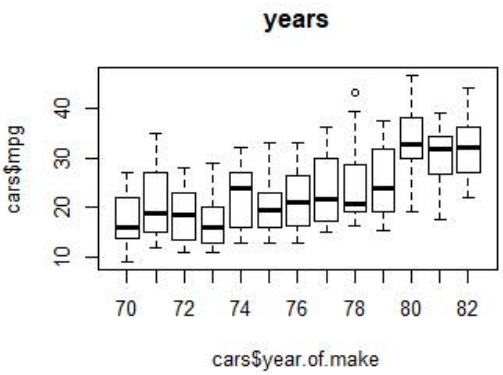
(6)

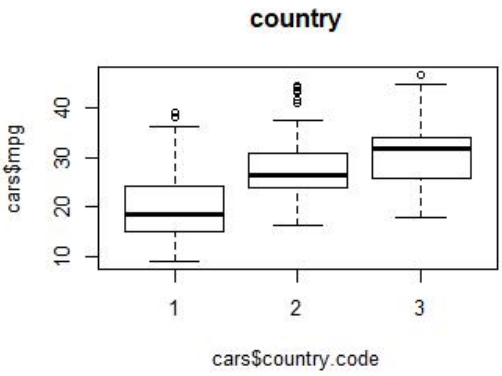












The mpg is smaller if cylinders are larger. The mpg is larger if years are larger. The mpg is larger if the country code is larger.

(7)

4、

(1)

$$Y_i = \beta_0 + \beta_1 \widetilde{X}_i + \beta_2 \widetilde{X}_i^{\,2} + \beta_3 \widetilde{X}_i^{\,3} + \xi_i$$
$$where: \widetilde{X}_i = X_i - \overline{X}$$

(2)

$$Y_i = \beta_0 + \sum_{k=1}^{K} \beta_k \widetilde{X}_{ik} + \sum_{k=1}^{K} \beta_k \widetilde{X}_{ik}^{\,2} + \sum_{1 \le k < k' \le K} \beta_{kk'} \widetilde{X}_{ik} \widetilde{X}_{ik'} + \xi_i$$

5、

(1)

There are no correct model among the models being considered. Because the key X variables are sin(x) and sin(2x). But the polynomial model dose not contain any key X variables.

(2)

Under the error variance of 0.5:
The in-sample variance of 1 model is : 0.5147214
The in-sample variance of 2 model is : 0.7689784
The in-sample variance of 3 model is : 1.001592
The in-sample variance of 5 model is : 1.472747
The in-sample variance of 7 model is : 1.987825
The in-sample variance of 9 model is : 2.51138
The variance will be changed if the error variance change.

(3)

The model with high polynomial will have the lower bias.
The bias will not change if the error variance change.

(4)

The model variance is the dominant component in the (in-sample) mean-squared-estimation-error if the polynomial is high.
The model bias is the dominant component in the (in-sample) mean-squared-estimation-error if the polynomial is low.
Because there are many nuisance X variables in the model with high polynomial, the variance will be larger and the bias will be smaller.
There is no nuisance X variables in the model with low polynomial, the variance will be smaller and the bias will be larger.

(5)

Model under error variance is 5:

```
> apply(bias^2,2,sum)
[1] 27.9942927 28.0956279 12.4328067  1.4196269  0.3010812  0.3082494
> apply(variance,2,sum)
[1]  51.47214  76.89784 100.15920 147.27472 198.78252 251.13805
> apply(err2.mean,2,sum)
[1]  79.41496 104.91657 112.49185 148.54708 198.88482 251.19516
```

Model under error variance is 0.5:

```
> apply(bias^2,2,sum)
[1] 27.983279928 27.984293280 12.321180293  1.189743521  0.031818203  0.003319873
> apply(variance,2,sum)
[1] 0.5147214 0.7689784 1.0015920 1.4727472 1.9878252 2.5113805
> apply(err2.mean,2,sum)
[1] 28.497487 28.752503 13.321771  2.661018  2.017656  2.512189
```

Model under error variance is 2:

```
> apply(bias^2,2,sum)
[1] 27.98494853 28.00116216 12.33809339  1.22457434  0.07261563  0.04952131
> apply(variance,2,sum)
[1]  8.235543 12.303655 16.025472 23.563956 31.805204 40.182088
> apply(err2.mean,2,sum)
[1] 36.21226 40.29251 28.34754 24.76497 31.84601 40.19143
```

It depends on the value of error variance.

If the error variance is large, we should choose the lower polynomial model .

If the error variance is small, we should choose the higher polynomial model.

If the error variance is in the medium value, we should choose the medium polynomial model.

Because the mean-squared-estimation-error is equal with variance plus squared bias. The bias of each model are not change but variance will change hugely.

(6)

Model under error variance is 0.5:

```
> SSE.mean
[1] 34.876531 34.621515 18.712425  7.229235  5.559057  5.013139
```

Model under error variance is 2:

```
> SSE.mean
[1] 139.67949 135.59924 116.16362  97.96621  88.58094  80.20314
```

Model under error variance is 5:

```
> SSE.mean
[1] 727.8384 702.3368 663.2994 606.1921 553.5687 501.2640
```

E(SSE) will be smaller the high polynomial model if the error variance is same.

E(SSE) will be larger if the error variance is larger and the model is the same. Because mean-squared-estimation-error is equal with variance plus squared bias. the variance of this model will larger but the bias will not change significantly.So, the E(SSE) will be larger.