Prediction of car price using regression technique & Principle component

ANURAG YADAV

[FEB 2022 – APR 2022]

PROBLEM STATEMENT

A Chinese **automobile company** Geely Auto aspires to enter the US market by setting up their manufacturing unit there and producing cars locally to give competition to their US and European counterparts.

They have contracted an automobile consulting company to understand the factors on which the pricing of cars depends. Specifically, they want to understand the **factors affecting the pricing of cars in the American market**, since those may be very different from the Chinese market.

The company wants to know:

- Which variables are significant in predicting the price of a car.
- How well those variables describe the price of a car.

Based on various market surveys, the consulting firm has gathered a large dataset of different types of cars across the American market.

GIVEN DATA

The dataset contains 14 columns with the "Price" column, taken as response variable and 13 predictor variables for n = 205 cars. The data dictionary for each of the variables can be found here:

Data Dictionary - carprices.xlsx

DEFINING VARIABLES

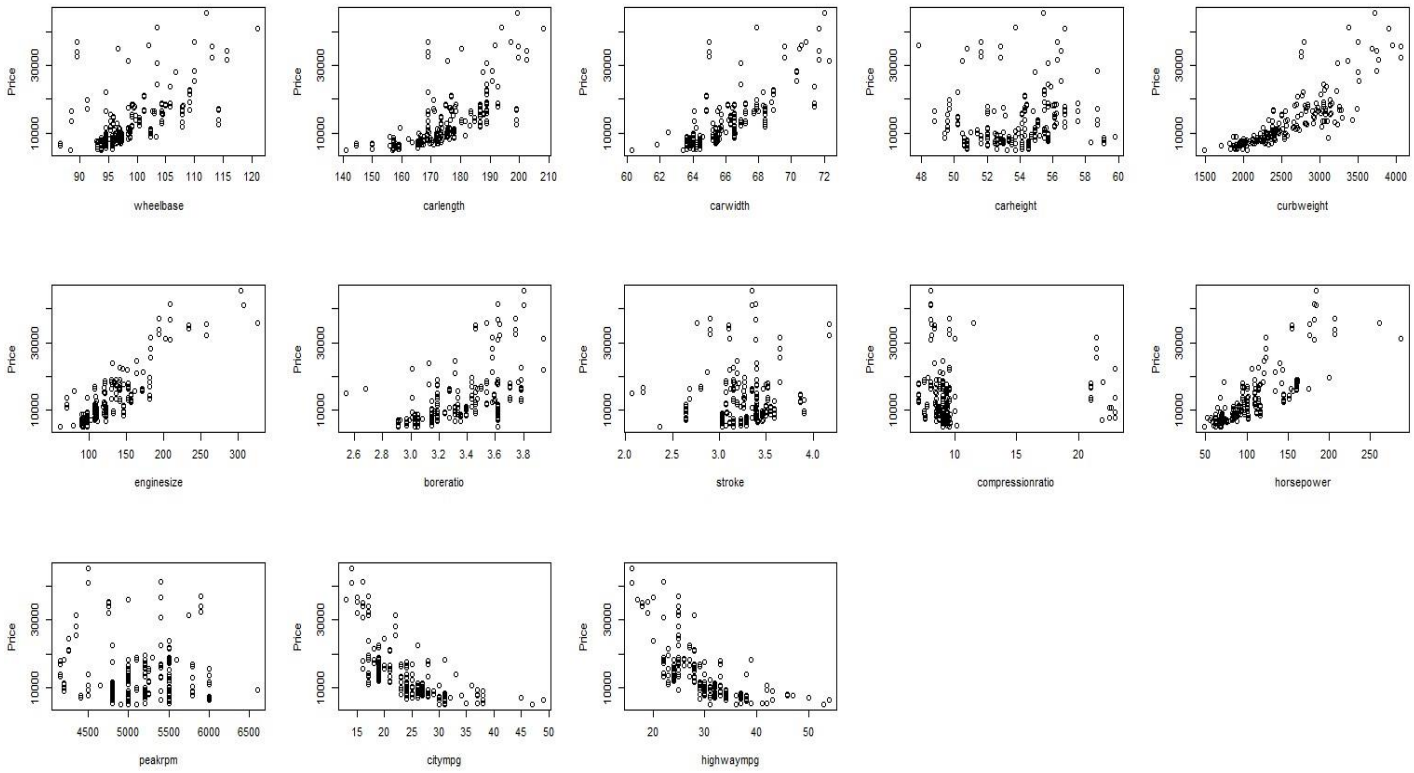
Let the Price of the car be the response variable Y and predictor variables are as mentioned in the data dictionary. These predictors (as mentioned in the data dictionary) are respectively denoted by X_1, X_2, …, X_{13}. We observe data {(y_i, x_{i1}, x_{i2}, …, x_{i13}): 1≤ i ≤ 205}.

NOTATION

Denote the observed response vector **y** = (y_1, y_2,.., y_{205})', the observed vector X_j= (x_{1j}, x_{2j}, . . . , x_{205j})' of the j-th predictor, 1_n is the vector of length n with all entries 1, and the design matrix **X** = [1_n X_1, X_2, . . , X_{13}]. **β**= (β_0, β_1, …. β_{13})' is the unknown parameter vector of the model. **E=** (ε_1, ε_2, …. ε_{205})' is the vector corresponding to the random error. Also let 0_n be the null vector of length n and I_n be the identity matrix of size n.

SCATTER PLOTS

From the data, the pairwise scatter plots of the response and predictors are found out and visualised:



Comment: From the scatter plots, we can assume a linear relationship visible between price and some predictors like wheelbase, carlength, carwidth, curbweight, enginesize, boreratio, horsepower, citympg, highwaympg. For rest of the predictor variables, judging by scatter plots, there seems no/little linearity with the response variable.

MULTIPLE LINEAR REGRESSION MODEL

Multiple linear regression model is given by

$$\mathbf{Y} = \mathbf{X}\,\beta + \mathbf{E}$$

where $\mathbf{E} \sim \mathbf{N}_n(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$ and the predictor variables $x_1, x_2, \ldots, x_p$ are assumed to be nonstochastic.

Fitting: Here we use least squares estimator of β which is given by $\beta = (X'X)^{-1}\mathbf{X'Y}$. For the given data, it turns out to be as shown in the table below.

Coefficients:

	Estimate
(Intercept)	6.024e+00
wheelbase	3.222e-03
carlength	6.875e-04
carwidth	2.036e-02
carheight	4.066e-03
curbweight	2.708e-04
enginesize	2.679e-03
boreratio	-1.734e-02
stroke	-1.155e-01
compressionratio	2.262e-02
horsepower	3.050e-03
peakrpm	9.387e-05
citympg	-3.170e-02
highwaympg	1.645e-02

And the fitted multiple linear regression model is:

$$\hat{y} = X\beta$$

R^2 and Adjusted R^2: For the given data and model, the value of R^2 and adjusted R^2 turn out to be 0.8508 and 0.8406 respectively. Therefore, 85.08% of the total variation of the response variable is explained by the above least-squares fitted multiple linear regression model.

TRANSFORMATION ON THE RESPONSE AND PREDICTOR VARIABLES FOR CORRECT SPECIFICATION OF THE MODEL

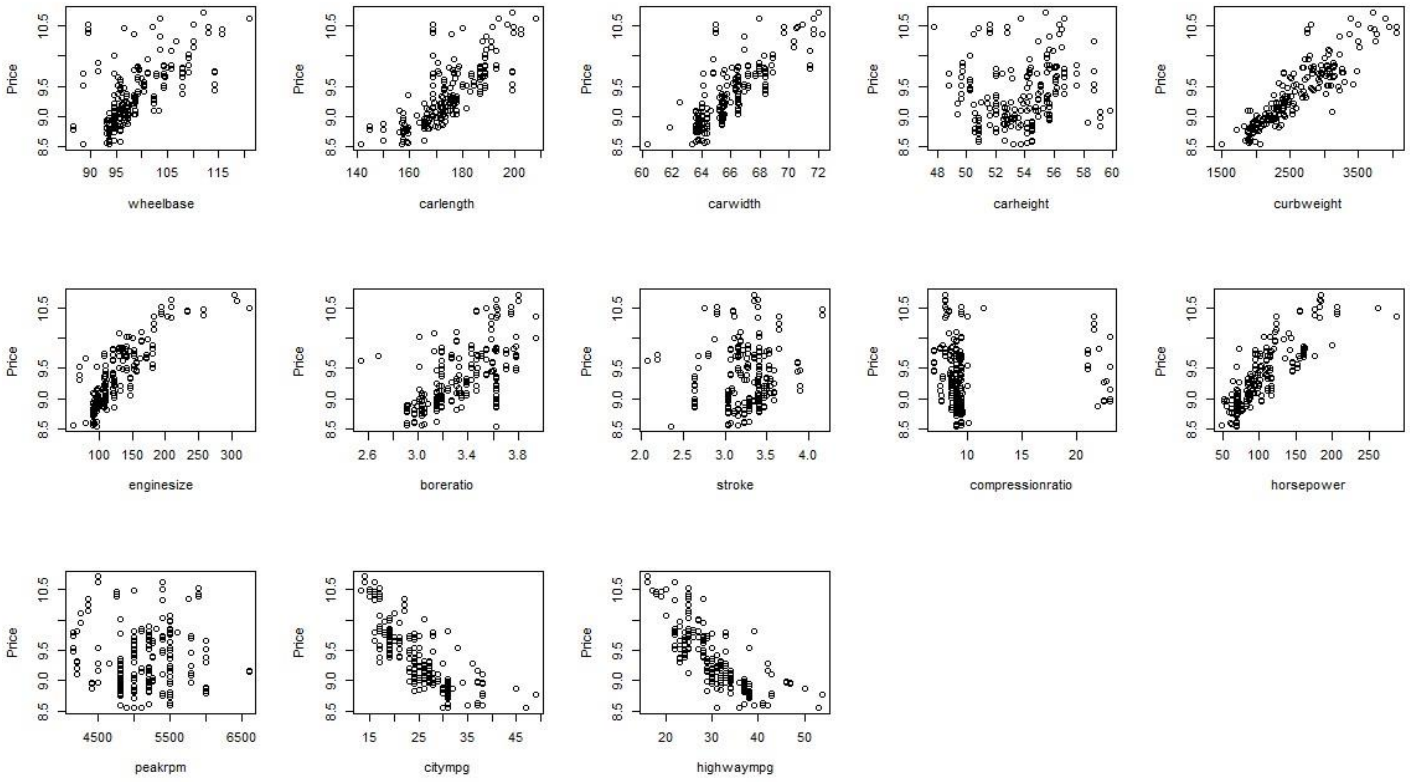
Transformation on y: Using the idea of Box-Cox method, we take the power transformation y^λ as the response variable. A suitable value of λ is chosen such that the residual sum of squares from the fitted model under this assumed model, $SS_{Res}(\lambda)$ is minimum. By theory, if λ ≠ 0, we take y^λ as the response variable, else for λ = 0, we take log(y) as the response variable.

For our data, it was observed that as λ→ 0 , the value of $SS_{Res}(\lambda)$ decreases and Adj. R^2 increases. And hence λ = 0 was assumed and log(y) was taken to be the suitable response variable.

Transformation on x: On taking log(y) as the response variable, the pairwise scatterplot of response and predictors was observed (Scatterplot diagram is mentioned below). There seems almost linear relationship between log(y) and the predictor variables. Also, for few variables, a higher degree of the predictor variables was took into account and fit into the MLRM but there was no significant decrease in SSE on the wage that a complicated model was chosen. And hence, the predictor variables were kept as it was.

SCATTER PLOTS

From the data, the pairwise scatter plots of the new response and predictors are found out and visualised:



Comment: From the scatter plots, we observe that there are few predictor variables which assume a more linear relationship with the response in the new model than it did it earlier. Overall, there are predictor variables like wheelbase, carlength, carwidth, curbweight, etc. which shows a more or less linear relationship than the other variables like carheight, stroke, peakrpm, etc. which doesn't show any relationship with the response variable. Also, it is seen that compression ratio has several outliers which may lead to failure of validity of assumptions.

MULTIPLE LINEAR REGRESSION MODEL

Multiple linear regression model is given by

$$\log \mathbf{Y} = \mathbf{X}\,\beta + \mathbf{E}$$

which can be remodelled as : $\mathbf{Y^*} = \mathbf{X}\,\beta + \mathbf{E}$ where $Y^* = \log(Y)$

where $\mathbf{E} \sim \mathbf{N}_n(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$ and the predictor variables $x_1, x_2, \ldots, x_p$ are assumed to be nonstochastic.

Fitting: Here we use least squares estimator of β which is given by $\hat{\beta} = (X'X)^{-1}\mathbf{X'Y^*}$. For the given data, it turns out to be as shown in the table below.

```
Coefficients:
                  Estimate
(Intercept)      6.024e+00
wheelbase        3.222e-03
carlength        6.875e-04
carwidth         2.036e-02
carheight        4.066e-03
curbweight       2.708e-04
enginesize       2.679e-03
boreratio       -1.734e-02
stroke          -1.155e-01
compressionratio 2.262e-02
horsepower       3.050e-03
peakrpm          9.387e-05
citympg         -3.170e-02
highwaympg       1.645e-02
```

And the fitted multiple linear regression model is:

$$\hat{y}_* = X\beta$$

R^2 and Adjusted R^2: For the given data and model, the value of R^2 and adjusted R^2 turn out to be 0.8792 and 0.871 respectively. Therefore, 87.92% of the total variation of the response variable is explained by the above least-squares fitted multiple linear regression model.

Note: For sake of ease in computation and notation, we write Y* as Y and scrap the original Y (response variable), unless explicitly mentioned

Testing of significance of regression : Here we wish to test **H_0:** $\beta_i = 0$ for all i = 0(1)13 against **H_1:** not **H_0**.

Test statistic: F_1 = MSR/MSE where MSR and MSE are respectively the mean of square and mean residual squares due to the least-squares fitted linear regression model.

Test rule: We reject the null hypothesis at 5% level of significance if $P(F_{p,n-p-1} > (F_1)_{observed}) < \alpha=0.05$

Since $P(F_{p,n-p-1} > (F_1)_{observed}) = 2.2e\text{-}16 < 0.05$, we reject the null hypothesis at 5% level of significance, i.e., that is the fitted multiple linear regression model is significant.

We now move to check the significance of **individual regression coefficients**.

H_0: $\beta_j = 0$ against **H_1:** $\beta_j \neq 0$; for all j = 0(1)13

Test statistic: $T_j = \frac{\beta_j - 0}{\sqrt{MSE * C_{JJ}}}$

where $C = ((C_{ij})) = (X'X)^{-1}$ and MSE is as defined earlier.

Test rule: We reject the null hypothesis at 5% level of significance if

$P(|t_{n-p-1}| > |(T_j)_{observed}|) < \alpha=0.05$

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.024e+00  8.663e-01   6.953 5.51e-11 ***
wheelbase         3.222e-03  5.700e-03   0.565 0.572591
carlength         6.875e-04  3.152e-03   0.218 0.827575
carwidth          2.036e-02  1.396e-02   1.458 0.146384
carheight         4.066e-03  7.700e-03   0.528 0.598093
curbweight        2.708e-04  9.857e-05   2.747 0.006585 **
enginesize        2.679e-03  7.851e-04   3.413 0.000785 ***
boreratio        -1.734e-02  6.785e-02  -0.256 0.798530
stroke           -1.155e-01  4.418e-02  -2.615 0.009648 **
compressionratio  2.262e-02  4.704e-03   4.808 3.08e-06 ***
horsepower        3.050e-03  9.201e-04   3.315 0.001096 **
peakrpm           9.387e-05  3.806e-05   2.466 0.014541 *
citympg          -3.170e-02  1.009e-02  -3.143 0.001938 **
highwaympg        1.645e-02  9.064e-03   1.815 0.071063 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can observe from the table given above that the p-value for the intercept and predictor variables like 'curbweight', 'enginesize', 'stroke', 'compressionratio', 'horsepower', 'peakrpm' and 'citympg' are significant at 5% level of significance.

Since all predictor variables in the model are not significant, we need to run variable selection algorithm to get a reasonable set of significant predictors.

VARIABLE SELECTION ALGORITHM

Here, we don't search among all possible subset linear regression model since then we need to search through 2^{13}-1 models which is a naïve method. So we use the variable selection algorithms such as:

1. Forward selection method
2. Backward elimination method
3. Step-wise selection method

After applying all above mentioned algorithms and fitting the models, we observe that both forward selection method and backward elimination method give the maximum value of R^2 and adjusted R^2 which are 0.8789 and 0.8727 respectively, whereas the stepwise selection method had R^2 and adjusted R^2 equal to 0.8781 and 0.8725 respectively. Both the models (obtained from forward and backward selection) retain the same set of predictors at their completion as can be seen from the output.

This means that either of these two can be used to get a reasonable set of significant predictors. Also, note the value of g, for the forward and backward selection, are close to 1 where g = adjusted R^2(M)/ adjusted R^2(whole model) where M is the model under consideration.

We drop the variables 'boreratio', 'carheight' and 'carlength' from our data after applying variable selection algorithms and update the matrices X and X_{CS}. The matrix X_{CS} is needed for checking whether multicollinearity is present among the predictor variables.

Check the presence of MULTICOLLINEARITY in the updated model

First, we apply centring and scaling on our updated data set. Now we again fit the MLRM to the centred and scaled data set. As the R^2 and adjusted R^2 doesn't change for the centred and scaled dataset this proves that standardization has no impact on the MLRM.

Determinant of X_{CS} and correlation matrix of predictor variables: Now we compute determinant of $X_{cs}'X_{cs}$ and observe that it is equal to 1.35e-05 ~ 0. So, we suspect presence of multicollinearity. Also, correlation matrix gives high pairwise correlation among various predictor variables.

Variance Inflation Factor: Now we compute **Variance Inflation Factor** for our scaled fit model and use 10 as our cut off.

$$VIF_j = (1 - {R_x}^2{}_j)^{-1}$$

where ${R_x}^2{}_j$ is the coefficient of determination obtained when X_j is linearly regressed on the remaining (p-1) predictor variables.

VIFs are as follows:

wheelbase	carwidth	curbweight	enginesize
4.67	5.21	15.01	6.51
stroke	compressionratio	horsepower	peakrpm
1.12	2.17	7.79	1.90
citympg	highwaympg		
24.30	23.19		

From the results we obtain that the regression coefficients corresponding to predictor variables 'curbweight', 'citympg' and 'highwaympg' are poorly estimated because of multi-collinearity.

Condition Indices: We now compute **Condition Indices** with cut off as 25.

$$C_j = \frac{\lambda_1(R)}{\lambda_j(R)}, j = 1, 2, \dots, p$$

The condition Indices are (1, 2.67, 5.56, 9.37, 10.11, 17.49, 35.95, 62.96, 101.24, 243.46).

We observe that last four condition Indices exceed our preset cut off of 25. Therefore, we suspect that last four principal components of the predictor variables may be responsible for multicollinearity.

Measures Based on Variance Decomposition:

$$p_{jk} = \frac{v_{jk,R}^2 / \lambda_k}{\sum_{k=1}^{p} v_{jk,R}^2 / \lambda_k(R)}$$

p_{jk} indicates the proportion of the contribution of k^{th} principal component on the variance of $\beta_{j,cs}$

If p_{jk} is large (say more than 0.5) and C_k lies in the danger region (say more than 25), then $\beta_{j\,CS}$ is suspected to be affected by the k-th principal component.

On computing these measures, we make the following observations:

Regression coefficient corresponding to

- 'carwidth' is suffering because of the 7^{th} principal component. The measure is 0.79
- 'curbweight' is suffering because of the 9^{th} principal component. The measure is 0.97
- 'stroke' is suffering because of the 3^{rd} principal component. The measure is 0.81
- 'horsepower' is suffering because of the 8^{th} principal component. The measure is 0.73
- 'citympg' is suffering because of the 10^{th} principal component. The measure is 0.93
- 'highwaympg' is suffering because of the 10^{th} principal component. The measure is 0.92

PRINCIPAL COMPONENT ANALYSIS

We fit the PCA model to the given data to work around the multi-collinearity detected above.

Let $\mathbf{V_R} = [V_1(\mathbf{R}), V_2(\mathbf{R}), \ldots, V_p(\mathbf{R})]$ where $V_j(\mathbf{R})$ is the eigen vector corresponding to the ith highest eigen value of **R**. Obtain the data matrix $\mathbf{X_{PCA}} = \mathbf{X_{CS}V_R}$ corresponding to the

principal components of $\mathbf{X_{CS}}$. Let $\mathbf{X_{j,PCA}}$ be the j-th column of $\mathbf{X_{PCA}}$. The i-th component of $\mathbf{X_{j,PCA}}$ is the i-th observation of the j-th principal component of Xcs. In principal component regression, we use the following algorithm.

1. Ycs is linearly regressed on all those principal components in $\mathbf{X_{1,PCA}}$, $\mathbf{X_{2,PCA}}$, ..., $\mathbf{X_{p,PCA}}$ whose sample variance is positive.
2. Test for significance of each principal component separately.
3. Obtain those principal components which are significant. Let they be the l_1, l_2, …, l_m th principal components.
4. Finally, Ycs is linearly regressed on $\mathbf{X_{l1,PCA}}$, $\mathbf{X_{l2,PCA}}$, ..., $\mathbf{X_{lm,PCA}}$.

Implementing the above algorithm on the data, we get the 1st, 4th, 5th and the 10th principal components of the predictor variables to be significant. The fitted values of principal component regression are:

The linear combination of each principal component is as shown:

	PC1	PC4	PC5	PC10
wheelbase	0.32	-0.27	-0.53	-0.03
carwidth	0.38	-0.26	-0.09	0.03
curbweight	0.42	-0.05	0.05	-0.05
enginesize	0.38	0.28	0.40	-0.08
stroke	0.08	0.18	-0.19	0.02
compressionratio	-0.01	-0.48	0.44	-0.02
horsepower	0.36	0.06	0.50	0.14
peakrpm	-0.07	-0.71	0.08	-0.02
citympg	-0.37	-0.03	0.14	0.71
highwaympg	-0.39	-0.01	0.19	-0.69

Hence, by PCA and by performing the fitting with the selected Principal Components only, it is observed that we achieve a value of adjusted R^2 = 0.8744. In this case, multiple R^2 = 0.8769, i.e., 87.69% of the total variation of the response variable is explained by the PCA fitted multiple linear regression model.

Note: Ridge regression was not carried because the tuning parameter under the model came out to be zero.

VERIFICATION OF ASSUMPTIONS:

Normality assumption:

We perform normality tests such as Shapiro-Wilk test, Anderson Darling test, etc. to test the null hypothesis

H_0: Error terms are normally distributed, against, **H_1**: Error terms are not normally distributed

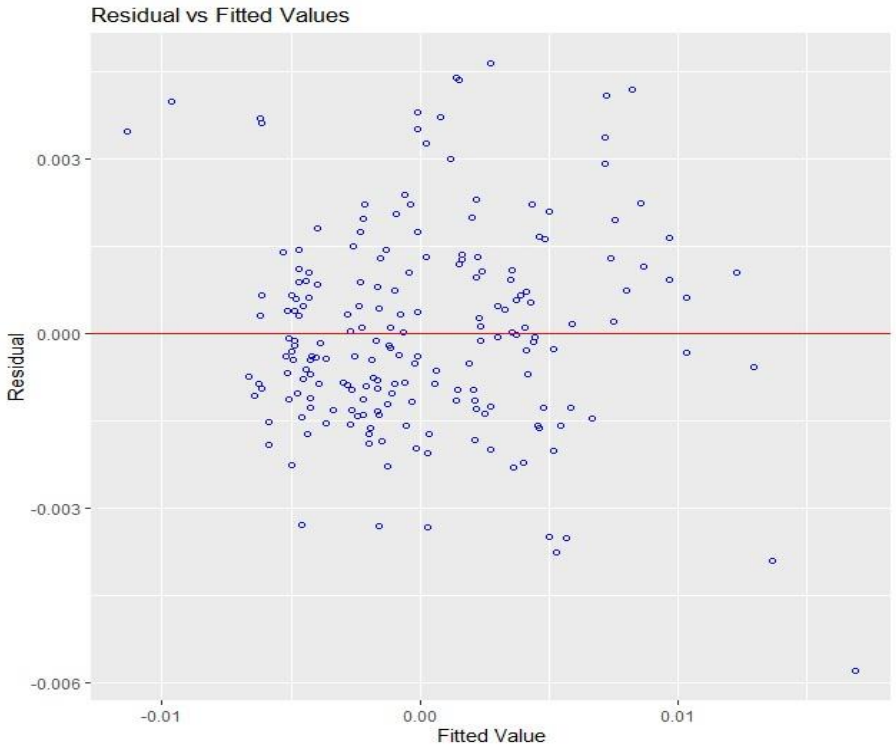
With the help of R and judging by the p-values of each tests, since p<0.05 for all tests, we conclude on the basis of the data that we reject **H_0** , that is to say, Error terms are not normally distributed.

The table of the test statistic values along with p-values is provided in the table below:

Test	Statistic	pvalue
Shapiro-Wilk	0.9746	9e-04
Kolmogorov-Smirnov	0.0601	0.4487
Cramer-von Mises	68.0947	0.0000
Anderson-Darling	1.6648	3e-04

Homoscedasticity assumption:

The best way to detect homoscedasticity is to construct residual vs fitted value plot and look for patterns. As clustering is clearly observed in the residual plot, we can conclude that **homoscedasticity assumption doesn't hold.** Also, tests like Bartlett's test or Breusch test require the error distribution to be normal which by above, clearly didn't and hence these tests can't be computed.



Assumption of Random errors being uncorrelated:

Since the sample size is sufficiently large, we draw the correlogram plot, which plots the autocorrelation of the sample with respect to order.

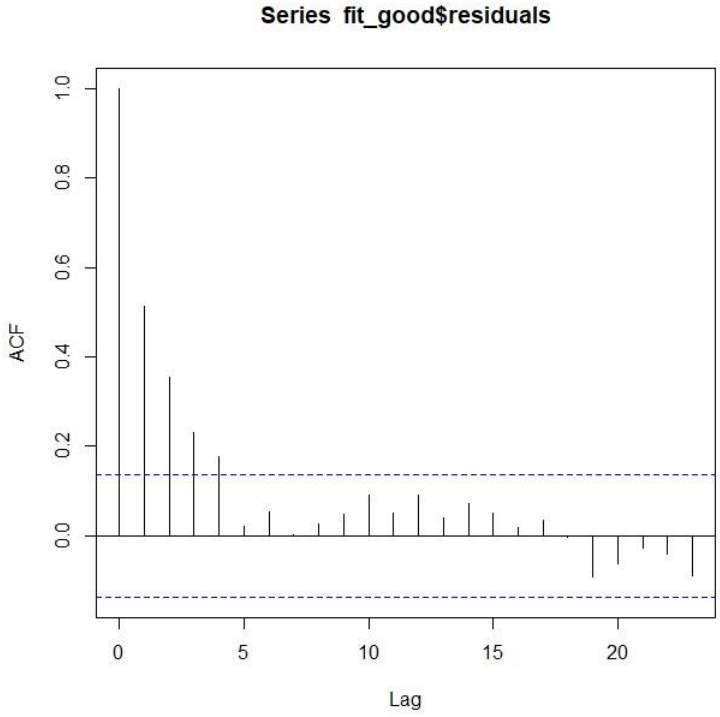
The sequence of autocorrelation are as follows:

$$\hat{p}_u = \sum_{i=1}^{n-u} e_i e_{i+u} \Big/ \sum_{i=1}^{n} e_i^2 \quad \forall\, u = 0,1,2,\ldots,n-1$$

where $\hat{p}_u$ is the autocorrelation of order u and $e_i's$ are the i'th residuals

The plot $\{(u,\hat{p}_u): u = 0,1,2\ldots,n-1\}$ is known as Correlogram.

If the assumption of the random errors being uncorrelated holds then, all the plotted autocorrelations of order more than 0 should be close to zero.



Clearly, from the diagram, it is evident that the plotted autocorrelations of order uptil 4 are significantly more than 0, and hence autocorrelation is present.

DETECTION OF OUTLIERS/ LEVERAGE/ INFLUENTIAL POINTS

After our final PC regression on the data, we found out that assumptions for the error distribution to be normal, failed, along with homoscedasticity and uncorrelated error terms. Part of this failure in assumptions may be attributed to presence of outliers in the model. And hence we check the presence of Outliers/Leverage/Influential points in the present model (model got after PC regression) and hence we have p=4, corresponding to 4 principal components of the predictors, and n=205

Notation:

Let $\hat{y}_{j,(i)}$ be the predicted value of y at $x_1 = x_{j1}, x_2 = x_{j2}, \ldots, x_p = x_{jp}$ when fitting of multiple linear regression model is done after deleting the i-th observation.

Let $Y_{(i)} = (\hat{y}_{1,(i)}, \hat{y}_{2,(i)}, \ldots, \hat{y}_{n,(i)})'$.

Let H be hat matrix, with (i, j)-th element denoted by h_{ij}.

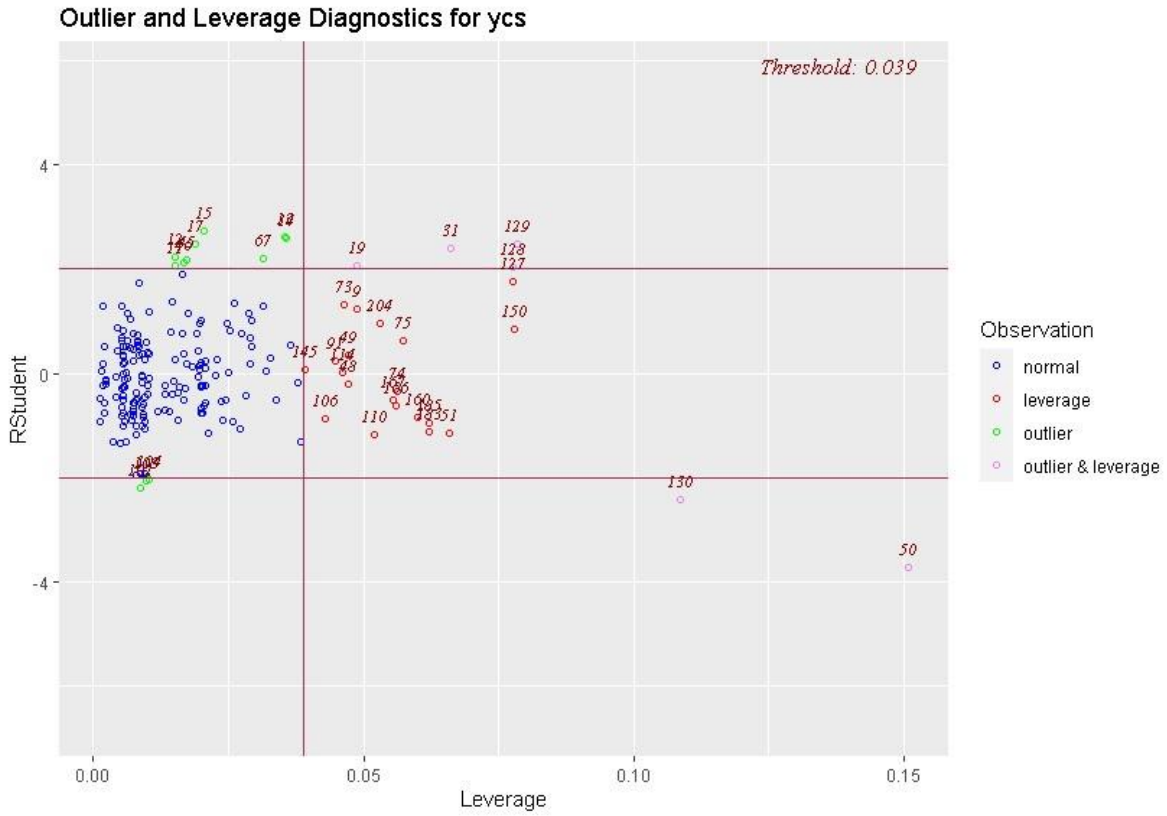
Let $\beta_{j,(i)}$ be the least square estimator of β_j when fitting of the multiple linear regression model is done after deleting i-th observation. Therefore, the corresponding least squares estimates of β is

$$\beta_{(i)} = (\beta_{0,(i)}, \beta_{1,(i)}, \ldots, \beta_{p,(i)})$$

SSE after deleting the i- th observation turn out to be

$$S^2_{(i)} = \sum_{j \neq i} \left(y_j - \hat{\beta}_{0,(i)} - \hat{\beta}_{1,(i)} x_{j1} - \hat{\beta}_{2,(i)} x_{j2} - \cdots - \hat{\beta}_{p,(i)} x_{jp}\right)^2$$

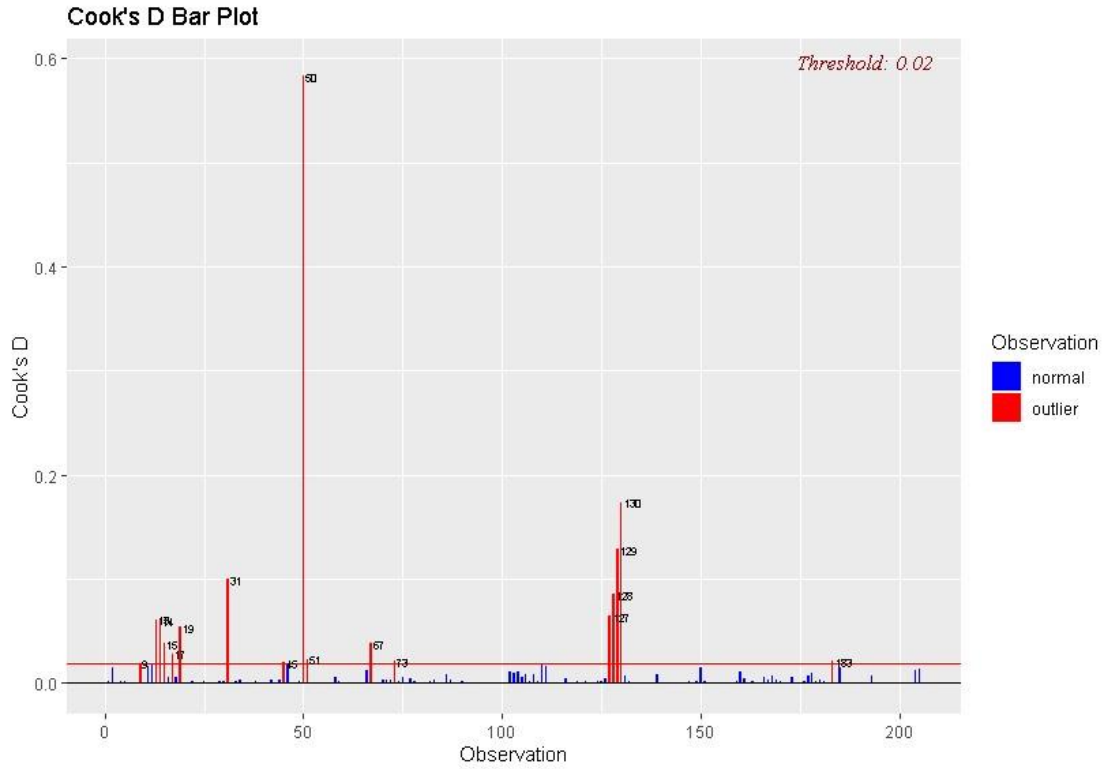
The large value of h_{ii} indicates that the i-th observation is a leverage point. The plot of $\{h_{ii} : 1 \leq i \leq n\}$ is given in Figure below. It shows that 9th, 48th, 49th, 51th, 73th, 74th, 75th, 91st, 106th, 110th, 114th, 127th, 145th, 150th, 160th, 166th, 167th, 183rd,185th and 204th observations as leverage points and 11th, 12th, 13th, 14th, 15th, 17th, 45th, 46th, 67th, 102nd, 103rd and 104th observations as outliers.



The Cook's distance statistics, also known as Cook's D-statistic, is defined by

$$D_i = \frac{\left(\hat{\beta} - \beta_{(i)}\right)(X'X)\left(\hat{\beta} - \hat{\beta}_{(i)}\right)}{(p+1)MSE} = \frac{\left(\hat{Y} - \hat{Y}_{(i)}\right)\left(\hat{Y} - \hat{Y}_{(i)}\right)}{(p+1)MSE} \quad , \quad , \quad \forall\, i = 1,2, \dots, n.$$

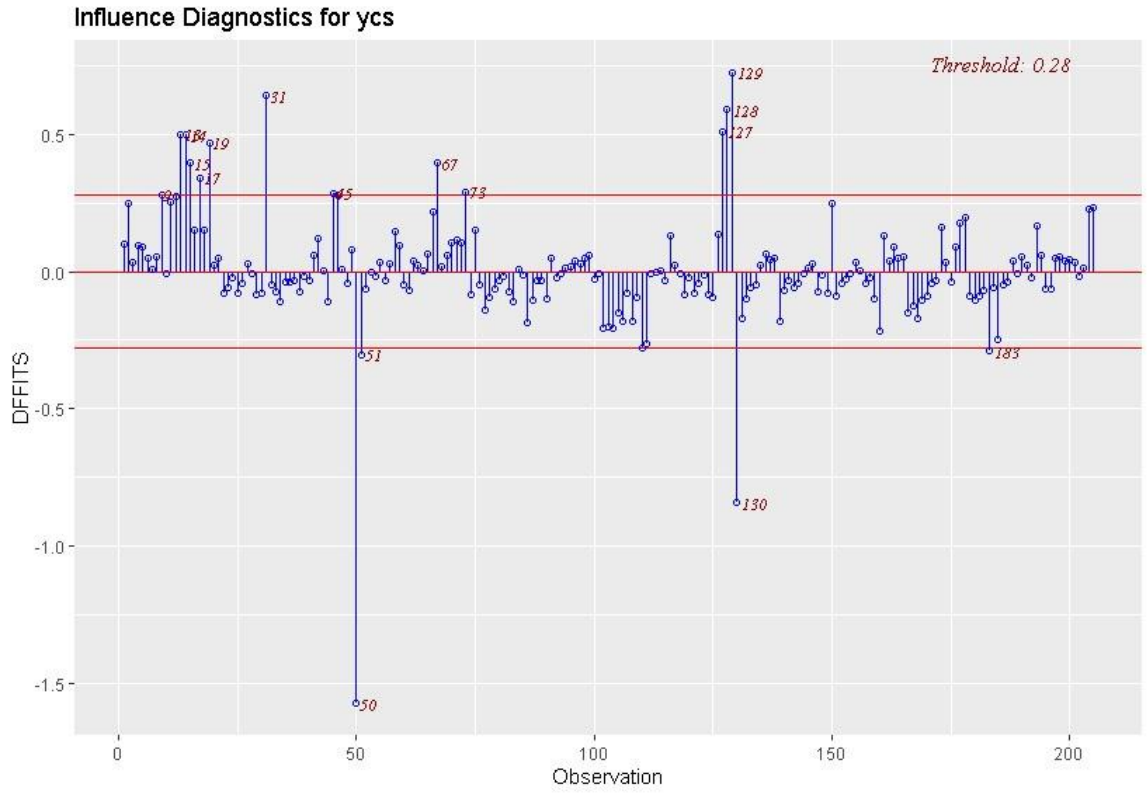
It is measure of distance between the least-squares estimate based on all n observations in β and the estimate $\beta_{(i)}$ obtained by deleting the i-th point. We consider i-th observation to be influential point if D_i is large. The plot of {D_i : 1 ≤ i ≤ n} are given in the figure below. By Cook's distance statistic we obtain 9th, 13th, 14th, 15th, 17th, 19th, 31st, 45th, 50th, 51st, 67th, 73rd, 127th, 128th, 129th, 130th and 183rd observations as influential points.



DFFITS statistic corresponding to the i-th observation is given by

$$DFFITS_i = \frac{\hat{y} - \hat{y}_{i,(i)}}{\sqrt{S^2_{(i)} h_{ii}}} \quad \forall\, i = 1,2, \dots, n.$$

We usually consider the i-th observation to be an influential point if the corresponding DFFITS is large. The plot of $\{DFFITS_i : 1 \le i \le n\}$ are given in the figure below. By DFFITS we obtain 9th, 13th, 14th, 15th, 17th, 19th, 31st, 45th, 50th, 51st, 67th, 73rd, 127th, 128th, 129th, 130th and 183rd were found out to be influential points.

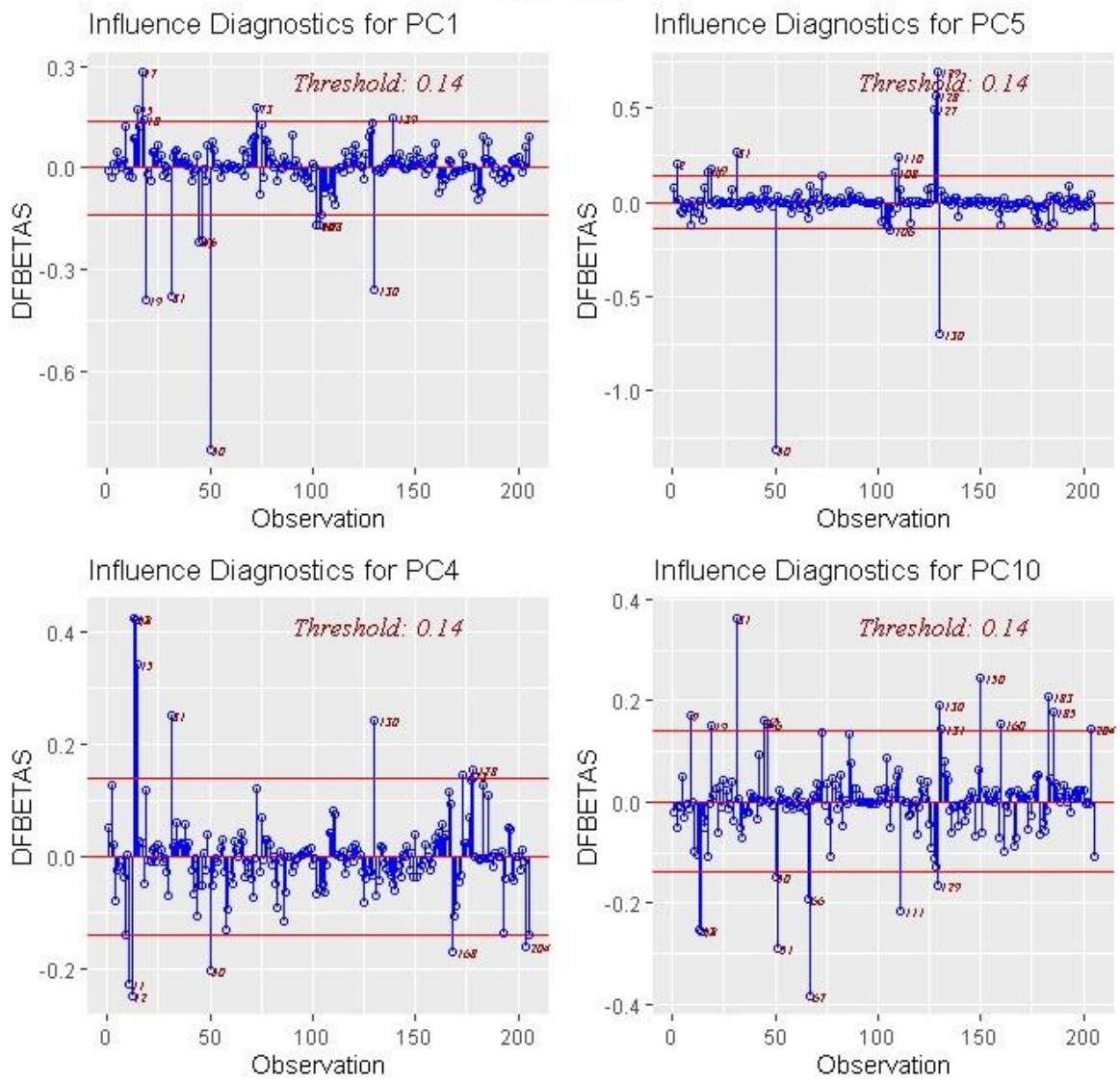


This statistic is

$$DFBETAS_{j,i} = \frac{\beta_j - \hat{\beta}_{j,(i)}}{\sqrt{S^2_{(i)}}}$$

Where $(C_{ij}) = ((X'X)^{-1})_{ij}$

Large magnitude of $DFBETAS_{ji}$ indicates that i-th observation has considerable influence on the j-th regression coefficient. Plots of DFBETAS statistics are given in Figures below. It can be noted that there are several data points which fall outside the threshold lines and hence those observations have considerable influence on the principal regression coefficients.



CONCLUSION

We had the dataset on cars, with price as response variables and 13 quantitative predictor variables. We fit the MLRM on the data, improved on the form of specification of Model, did feature selection to get rid of insignificant predictors. Then, we moved on to detect multicollinearity and taking suitable measures against it by performing the PC regression. In the final model, the assumptions of the error distribution and such, were checked which lead to failure, hence though the adjusted R^2 ~ 87% but the model is not suitable for this data. This failure can be attributed to the presence of outliers, and since the number of outliers is considerably high, one may think of the error distribution to follow a heavy tailed distribution. Finally, for a further scope of the project, it can be said that for our data, regression procedures like robust regression is more suitable.