Multiple Linear Regression Model for Used Cars

Harper Guan (h32guan@uwaterloo.ca)
Qicheng Zhao (q83zhao@uwaterloo.ca)
Colina Dai (a8dai@uwaterloo.ca)

December.6.2021

1. Abstract:
In this report, we are going to construct a best MLRM (multiple linear regression model) through performing the relevant diagnostic tests for the candidate models generated based on the dataset of Used Cars obtained on Kaggle (see Reference 1.1). The detailed source of the dataset is not specified on Kaggle. The variables include price, engine size and curb weight, etc.

Our interest is to explore variables which have significant influence on the price of used cars. For model comparison, we used the AIC and adjusted R^2. We also used automatic model selection procedures such as forward selection, backward elimination and stepwise selection for our model selection

Overall, we found that there are 8 variables (wheel base, length, engine size, highway.L.100km, diesel, rear wheel, low price, and median price) which have significant influence on the prices of used cars.

2. Introduction:
Nowadays, besides brand-new vehicles, used cars are gaining popularity in North America, including Canada. This situation might be due to the reasons that purchasing used cars could save money and is environmentally-friendly. Thus, our interest is to explore the variables which have significant influence on the price of used cars.

The Used Cars dataset contains 201 observations and 33 variables (1 response variable y + 32 explanatory variables x's). The observations are not measured over time. Generally, we consider the price as our response variable y. The dataset was obtained on Kaggle, a professional website (see Reference 1.1). The detailed source of the dataset is not specified on Kaggle. The theme is closely related to real-world car markets. The following is an summary to the y-variable and x-variables and related operations (see Appendix 2.1):

- y: price
- <u>We removed those x's due to a small percentage of missing values or totally repeated values or low relation to y:</u> normalized losses, symboling, number of doors, stroke, bore, horsepower, peak.rpm, make, body style, engine location, number of cylinders, fuel system, engine type (removed a total of 13 x-variables).
- <u>We removed those x's due to duplication with length, width and height:</u> normalized length, normalized width, normalized height (removed a total of 3 x-variables).
- <u>We created indicator variables for those categorical variables: drive wheels, price binned, aspiration:</u>
    1) drive wheels => fwd (standing for forward wheels), rwd (standing for rear wheels) and 4wd (4 wheels)
    2) price binned => low (standing for low price), median(standing for median price) and high (standing for high price)
    3) aspiration => turbo (standing for turbo aspiration), std (standing for std aspiration) (added a total of 3+3+2=8 x-variables)
- <u>We removed those variables because they are changed into indicator variables:</u> drive wheels, price binned, aspiration (removed a total of 3 x-variables).

- <u>Those variables are treated as baselines of the corresponding indicator variables in R Studio:</u> fwd (standing for forward wheel), high (standing for high price), std (standing for std aspiration)

So far in our full-model, we have 1 response variable (price) and a total of 32-13-3+8-3= 21 explanatory variables.

<u>3. Analysis:</u>

3.1 Multicollinearity

By inspection, we could infer the following pairs of x-variables might have relatively strong linear relationships: (city.L.100km and highway.L.100km), (length and width), (length and height), (width and height), (highway.mpg and city.mpg), (curb weight and length).

We used corr in R Studio to compute their correlation coefficient. The following table summarizes the output (see Appendix 3.1):

Pair	Correlation coefficient ρ	>0.8 ?	If Yes, remove one of each two x-variables
city.L.100km, highway.L.100km	0.9583056	Yes	Remove city.L.100km
length, width	0.8571703	Yes	Remove width
length, height	0.4920625	No	/
width, height	0.3060022	No	/
highway.mpg, city.mpg	0.9720437	Yes	Remove city.mpg
curb weight, length	0.8806648	Yes	Remove curb weight

We choose ρ=0.8 as our criterion. For those pairs which have ρ>0.8 (close to 1), we think that the two x's in that pair are linearly correlated and multicollinearity occurs. To address the problem, we need to remove one of the two x's in that pair (removed a total of 4 x-variables).

So far in our full-model, we have 1 response variable (price) and a total of 21- 4= 17 explanatory variables. Thus, our full-model includes 1 response variable and 17 explanatory variables. That is yi= β0 +β1*xi1 +β2*xi2 +β3*xi3 +β4*xi4 +β5*xi5 +β6*xi6 +β7*xi7 +β8*xi8 +β9*xi9 +β10*xi10 +β11*xi11 +β12*xi12 +β13*xi13 +β14*xi14 +β15*xi15 +β16*xi16 +β17*xi17 +**ε**i, where **ε**i~i.i.d G (0, σ), for i=1, …, n

(**yi**= price, xi1=wheel base, xi2 = length, xi3 = height, xi4 = engine size, xi5 = compression ratio, xi6 = highway mpg, xi7 = highway.L.100km, xi8 = diesel, xi9 = gas, xi10 = front wheel, xi11 =rear wheel, xi12 =four wheel, xi13 =low price, xi14 =median price, xi15 = high price, xi16 =turbo aspiration, xi17 = std aspiration)

3.2 Interaction

When the effect of one variable depends on the other variables, an interaction effect occurs. We check the interaction between engine size and wheel base, length and height, engine size and compression ratio, wheelbase and compression ratio, length and compression ratio, because we detected that variables in these groups are highly likely to depend on the other.
The outcome shows these groups all without interaction effect, because p-value for each group are larger than the benchmark 0.05, engine size:wheel base (0.5192), length:height (0.873), engine size:compression ratio (0.3597), wheel base:compression ratio (0.900382), length:compression ratio (0.862122), (see Appendix 3.2.1～3.2.5). Thus, we ignore the interaction effect in this project.

3.3 Full Model

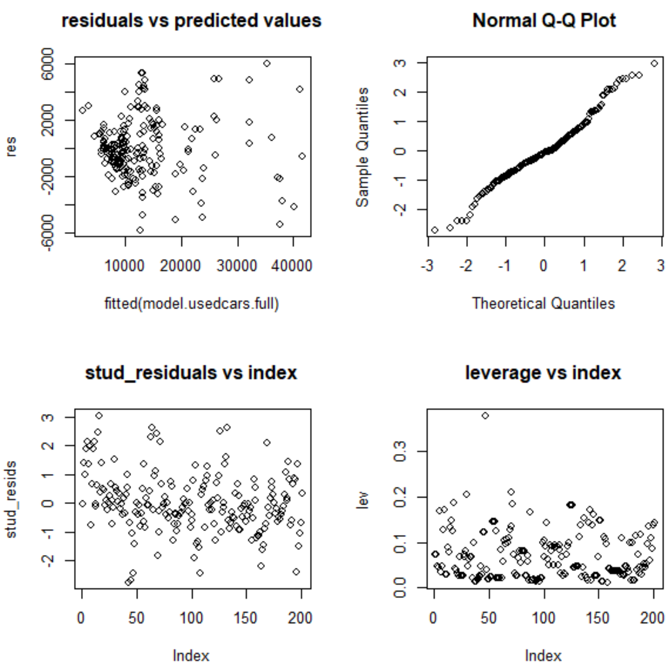
3.3.1 Summary

We take $\alpha=0.05$ as the significance level. For the summary of the full model, we could observe that the following x-variables have p-values that are less than 0.05: length, engine size, highway.L.100km, rear wheel, low price, median price (see Appendix 3.3.1). Since their p-values are less than $\alpha=0.05$, then there is evidence against their corresponding null hypothesis of B=0 (insignificant). Thus, those x-variables are significant. We will use the y and those 6 x's to form a new model. That is $y_i = \beta_0 + \beta_2 * x_{i2} + \beta_4 * x_{i4} + \beta_7 * x_{i7} + \beta_{11} * x_{i1} + \beta_{13} * x_{i13} + \beta_{14} * x_{i14} + \boldsymbol{\varepsilon}_i$, where $\boldsymbol{\varepsilon}_i \sim$ i.i.d $G(0, \sigma)$, for i=1, …, n (**yi** is price, xi2 is length, xi4 is engine size, xi7 is highway.L.100km, xi11 is rear wheel, xi13 is low price, xi14 is median price)

3.3.2 Diagnostic Tests

Next, we perform diagnostic tests for the full-model (see Appendix 3.3.2). The output looks like the following:

Diagnostic tests for the full-model



In plot 1, residuals vs. predicted values, we could observe that the data points are distributed along the zero horizontal line. Thus, the assumption of mean zero is satisfied. We could also observe that as the fitted values increase, the range of residuals also slightly increases. Thus, it violates the assumption of constant variance.

In plot 2, normal QQ plot, we could observe that almost all the points are distributed along the straight line. Thus, the standardized residuals follow a normal distribution; agree with the assumption of normality.

In plot 3, studentized residuals vs index, we could observe that the absolute value of Studentized residuals is close to 3. Thus, there are no unusual studentized residuals i.e. no outliers. Thus, the assumption of studentized residuals is satisfied.

In plot 4, leverage vs. index, we could observe that there is one point with leverage over 0.3, which is greater than others. Thus, it violates the assumption of leverage.

We also plot the residuals vs. each of the 13 x-variables (the other 4 x's are used as baselines in R Studio). The plots are shown in Appendix (see Appendix 3.3.3 ~ 3.3.6). We only see weak linear relationships between y and the following x's respectively: length, engine size, highway.L.100km, wheel base, height, highway mpg. The assumption of linearity is violated.

3.4 New Model (candidate Model 1): Exclude insignificant variables from Full Model

For our new model, it includes 1 response variable (price) and 6 explanatory variables (length, engine size, highway.L.100km, rear wheel, low price, median price). By summary in R, we could observe that their p-values are all less than $\alpha=0.05$, which means there is evidence against their corresponding null hypothesis of B=0 (insignificant). Thus it is proved that those 6 x-variables are all significant (see Appendix 3.4). Thus, the final version of the new model, model1:

yi $=\beta0 + \beta2 *xi2 + \beta4 *xi4 + \beta7 *xi7 + \beta11 *xi11 + \beta13 *xi13 + \beta14 *xi14 + \boldsymbol{\varepsilon}i$, where $\boldsymbol{\varepsilon}i$~i.i.d G $(0, \sigma)$, for i=1, …, n

(**yi**= price, xi2 = length, xi4 = engine size, xi7 = highway.L.100km, xi11 =rear wheel, xi13 =low price, xi14 =median price)

3.5 Forward Selection Model (candidate Model 2):

The forward selection model we constructed is a type of automatic model selection. The model starts with the null model, then adds 7 variables (engine size, low price, highway.L.100km, median price, compression ratio, rear wheel, length)
to reduce AIC the most. At the end, the forward selection method generates a model: price ~ engine.size + low_price + highway.L.100km + median_price + compression.ratio + rear_wheel + length (see Appendix 3.5.1). Checking the VIF of the model, we found that the VIF of all explanatory variables in this model are less than 10 (see Appendix 3.5.2), which means variables are not multicollinearity in this model. Thus, the final model for forward selection model is **yi** $= \beta0 + \beta4 * xi4 + \beta13 *xi13 + \beta7 * xi7 + \beta14 * xi14 + \beta5 *xi5 + \beta11 * xi11 + \beta2 * xi2 + \boldsymbol{\varepsilon}i$, where $\boldsymbol{\varepsilon}i$~i.i.d G $(0, \sigma)$, for i=1, …, n

(**yi**= price, xi2 = length, xi4 = engine size, xi5 = compression ratio, xi7 = highway.L.100km, xi13 =low price, xi14 =median price,)

3.6 Backward Selection Model (candidate Model 3):

The backward selection model we constructed is a type of automatic model selection. The model start with the full p-variable model (New Model), then removes 9 variables (std_aspiration, high_price, four_wheel, gas, turbo_aspiration, highway.mpg, front_wheel, compression.ratio, height) to reduce AIC the most. Finally, the backward selection method generates a model: price ~ wheel.base + length

+ engine.size + highway.L.100km + diesel + rear_wheel + low_price + median_price (see Appendix 3.6.1). Checking the VIF of the model, we found that the VIF of all 8 explanatory variables in this model are less than 10 (see Appendix 3.6.2), which means variables are not multicollinearity in this model. Thus, the final model for backward selection model is model3:

yi = $\beta 0$ + $\beta 1$ * xi1 + $\beta 2$ *xi2 + $\beta 4$ *xi4 + $\beta 7$ *xi7 + $\beta 8$ *xi8 + $\beta 11$ *xi11 + $\beta 13$ *xi13 + $\beta 14$ *xi14 +**εi**, where **εi**~i.i.d G (0, σ), for i=1, …, n

(**yi**= price, xi1=wheel base, xi2 = length, xi4 = engine size, xi7 = highway.L.100km, xi8 = diesel, xi11 =rear wheel, xi13 =low price, xi14 =median price)

3.7 Stepwise Selection Model (candidate Model 4):

The Stepwise selection model we constructed is a type of automatic model selection. The model starts with the null model, then adding and eliminating variables together, which leaves with 7 variables (engine.size, low_price, highway.L.100km, median_price, compression.ratio, rear_wheel, length) at the end (see Appendix 3.7.1). Checking the VIF of these explanatory variables, we found that the VIF of all 7 explanatory variables in this model are less than 10 (see Appendix 3.7.2), which means these 7 variables are not multicollinearity in this model. Thus, the final model for stepwise selection model is model 4:

yi = $\beta 0$ + $\beta 4$ *xi4 +$\beta 13$ * xi13 +$\beta 7$ * xi7 + $\beta 14$ *xi14 +$\beta 5$ *xi5 + $\beta 11$ *xi11 +$\beta 2$ *xi2+**εi**, where **εi**~i.i.d G (0, σ), for i=1, …, n

(yi= price, xi2 = length, xi4 = engine size, xi5 = compression ratio, xi7 = highway.L.100km, xi11 =rear wheel, xi13 =low price, xi14 =median price,)

4. Model Comparison

4.1 Comparing models with AIC:

We compute AIC for candidate models (see Appendix 4.1), and summary into the following table:

	Model 1 (New Model)	Model 2 (Forward)	Model 3 (Backward)	Model 4 (Stepwise)
AIC	3681.938	3670.744	3669.536	3670.744

We prefer to choose a model with the smallest AIC value, and the model 3 (backward selection model) with the smallest AIC (3669.536) among the 4 candidate models. According to AIC criteria, we choose model 3 as the best model.

4.2 Comparing models with adjusted R^2

We compute adjusted R^2 for candidate models (see Appendix 4.2.1~4.2.4), and summary into the following table:

	Model 1 (New Model)	Model 2 (Forward)	Model 3 (Backward)	Model 4 (Stepwise)
Adjusted R^2	0.9199	0.9246	0.9255	0.9246

We prefer to choose a model with the largest adjusted R^2, and the model 3 (backward selection model) with the the largest adjusted $R^2 (0.9255)$ among the 4 candidate models. According to adjusted R^2 criteria, we choose model 3 as the best model.

5. Preferred Model: Model 3 (Backward Selection Model)

Regarding to both AIC and adjusted R^2 criterias in (part 4.1 and 4.2), the best model best model is the model 3 (backward selection model):

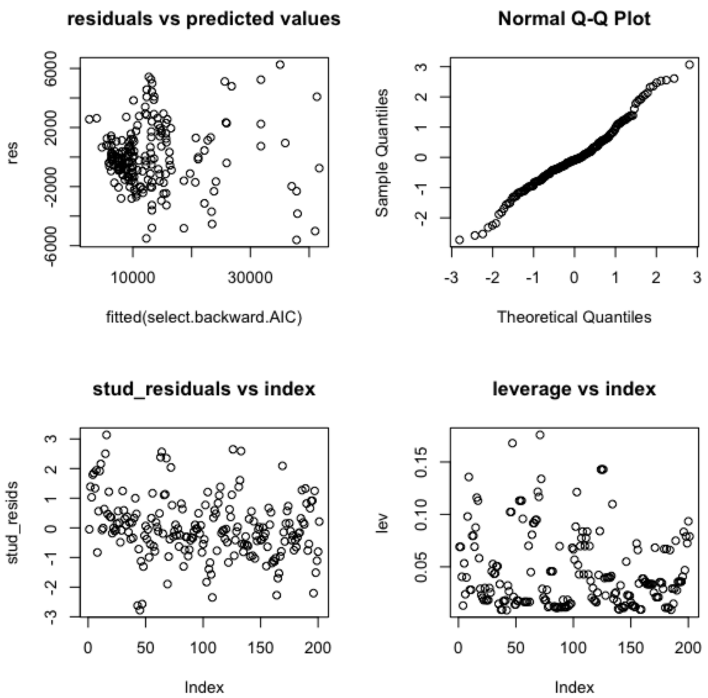
yi $=\beta 0+ \beta 1* xi1 + \beta 2*xi2 + \beta 4*xi4 + \beta 7*xi7 + \beta 8*xi8 + \beta 11*xi11 + \beta 13*xi13 + \beta 14*xi14+$**εi**,
where **ε**i~i.i.d G (0, σ), for i=1, …, n

(**yi**= price, xi1=wheel base, xi2 = length, xi4 = engine size, xi7 = highway.L.100km, xi8 = diesel, xi11 =rear wheel, xi13 =low price, xi14 =median price)

5.1 Diagnostic Tests

Next, we perform diagnostic tests for the model 3 (see Appendix 5.1.1). The output looks like the following:

Diagnostic tests for the full-model



In plot 1, residuals vs. predicted values, we could observe that the data points are mostly concentrated along the zero horizontal line. Thus, the assumption of mean zero is satisfied. We could also observe that as the fitted values increase, the range of residuals keeps almost constant. Thus, the assumption of constant variance is also satisfied.

In plot 2, normal QQ plot, we could observe that almost all the points are distributed along the straight line. Thus, the standardized residuals follow a normal distribution; agreed with the assumption of normality.

In plot 3, studentized residuals vs index, we could observe that the absolute value of Studentized residuals is close to 3. Thus, there are no unusual studentized residuals i.e. no outliers. Thus, the assumption of studentized residuals is satisfied.

In plot 4, leverage vs. index, we could observe that there are no outstanding points with very large leverage. Thus, the assumption of leverage is satisfied.

We also plot the residuals vs. each of the 7 x-variables. The plots are shown in Appendix (see Appendix 5.1.2 & 5.1.3). We notice that, for all x-variables, the range of residual almost stays the same within the certain range for each x-variables, and there is no pattern. Thus, the assumption of linearity is satisfied.

6. Result and Summary

The final model we choose is the model 3 (backward selection model), because it is the best model within the 4 candidate models we constructed in part 3, based on AIC and adjusted R^2 as criterias. Afterwards, we check the assumptions of this model, in part 5.1, and notice that assumptions of mean zero, constant variance, normality, studentized residuals, leverage and linearity are all satisfied. Thus, the final model is valid.

Based on the summary of model 3 (see Appendix 4.2.3), the final linear regression model is:

yi =3996.64 -76.62* xi1 + 109.43*xi2 + 26.27*xi4 + 954.40*xi7 + 2509.17*xi8 + 2099.13*xi11 − 15904.66*xi13 -8890.62*xi14 + **ε**i, where **ε**i~i.i.d G (0, σ), for i=1, …, n

(**yi**= price, xi1=wheel base, xi2 = length, xi4 = engine size, xi7 = highway.L.100km, xi8 = diesel, xi11 =rear wheel, xi13 =low price, xi14 =median price)

Now, we will explain the implication of these variables to the final model.

- xi1, each additional unit increase of wheel base (xi1) will decrease price (yi) by 76.62 units on average.
- xi2, each additional unit increase of length (xi2) will increase price (yi) by 109.43 units on average.
- xi4, each additional unit increase of engine size (xi4) will increase price (yi) by 26.27units on average.
- xi7, each additional unit increase of highway.L.100km (xi7) will increase price (yi) by 954.40 units on average.
- xi8, each additional unit increase of diesel (xi8) will increase price (yi) by 2509.17 units on average.
- xi11, each additional unit increase of rear wheel (xi1) will increase price (yi) by 2099.13 units on average.
- xi13, each additional unit increase of low price (xi13) will decrease price (yi) by 15904.66 units on average.
- xi14, each additional unit increase of median price (xi14) will decrease price (yi) by 8890.62 units on average.

7. Conclusion

The final model we selected:

yi =3996.64 -76.62* xi1 + 109.43*xi2 + 26.27*xi4 + 954.40*xi7 + 2509.17*xi8 + 2099.13*xi11 − 15904.66*xi13 -8890.62*xi14 + **ε**i, where **ε**i~i.i.d G (0, σ), for i=1, …, n

(**yi**= price, xi1=wheel base, xi2 = length, xi4 = engine size, xi7 = highway.L.100km, xi8 = diesel, xi11 =rear wheel, xi13 =low price, xi14 =median price)

The final model is the model 3 selected from 4 candidate models, based on AIC (smallest) and adjusted R^2 (largest).

Our main question is “ Which variables have a significant influence on the price of used cars?” The answer is: wheel base, length, engine size, highway.L.100km, diesel, rear wheel, low price, and median price are significant variables to the price of used cars.

References:

1.1 Used Cars dataset

Ammaraahmad. (2021, October 01). Used cars dataset. Retrieved from https://www.kaggle.com/ammaraahmad/used-cars-dataset

Appendix:

(all codes used in R Studio)

2.1

```
usedcars_dataset$normalized.losses<-NULL
usedcars_dataset$symboling<-NULL
usedcars_dataset$num.of.doors<-NULL
usedcars_dataset$stroke<-NULL
usedcars_dataset$bore<-NULL
usedcars_dataset$horsepower<-NULL
usedcars_dataset$peak.rpm<-NULL
usedcars_dataset$make<-NULL
usedcars_dataset$body.style<-NULL
usedcars_dataset$engine.location<-NULL
usedcars_dataset$num.of.cylinders<-NULL
usedcars_dataset$fuel.system<-NULL
usedcars_dataset$engine.type<-NULL


usedcars_dataset$normalized_length<-NULL
usedcars_dataset$normalized_width<-NULL
usedcars_dataset$normalized_height<-NULL



#create indicator variable for drive-wheels
usedcars_dataset$front_wheel<-ifelse(usedcars_dataset$drive.wheels=="fwd",1,0)
usedcars_dataset$rear_wheel<-ifelse(usedcars_dataset$drive.wheels=="rwd",1,0)
usedcars_dataset$four_wheel<-ifelse(usedcars_dataset$drive.wheels=="4wd",1,0)

#create indicator variable for price_binned
usedcars_dataset$low_price<-ifelse(usedcars_dataset$price_binned=='Low',1,0)
usedcars_dataset$median_price<-ifelse(usedcars_dataset$price_binned=='Median',1,0)
usedcars_dataset$high_price<-ifelse(usedcars_dataset$price_binned=='High',1,0)

#create indicator variable for aspiration
usedcars_dataset$turbo_aspiration<-ifelse(usedcars_dataset$aspiration=='turbo',1,0)
usedcars_dataset$std_aspiration<-ifelse(usedcars_dataset$aspiration=='std',1,0)


usedcars_dataset$aspiration<-NULL ##indicator
usedcars_dataset$drive.wheels<-NULL ##indicator
usedcars_dataset$price_binned<-NULL ###indicator
```

3.1

```
> cor(usedcars_dataset$city.L.100km,usedcars_dataset$highway.L.100km)#>0.8
[1] 0.9583056
> cor(usedcars_dataset$length,usedcars_dataset$width)#>0.8
[1] 0.8571703
> cor(usedcars_dataset$length,usedcars_dataset$height)
[1] 0.4920625
> cor(usedcars_dataset$width,usedcars_dataset$height)
[1] 0.3060022
> cor(usedcars_dataset$highway.mpg,usedcars_dataset$city.mpg)#>0.8
[1] 0.9720437
> cor(usedcars_dataset$curb.weight,usedcars_dataset$length)#>0.8
[1] 0.8806648
```

```
> usedcars_dataset$width<-NULL
> usedcars_dataset$city.L.100km<-NULL
> usedcars_dataset$city.mpg<-NULL
> usedcars_dataset$curb.weight<-NULL
```

3.2.1

```
Call:
lm(formula = price ~ engine.size * wheel.base, data = usedcars_dataset)

Residuals:
    Min      1Q  Median      3Q     Max 
-8387.0 -2273.8  -352.4  1525.2 14712.0 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)  
(Intercept)           -3.005e+04  1.243e+04  -2.419   0.0165 *
engine.size            2.035e+02  7.875e+01   2.584   0.0105 *
wheel.base             2.401e+02  1.259e+02   1.908   0.0579 .
engine.size:wheel.base -4.954e-01  7.672e-01  -0.646   0.5192  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3820 on 197 degrees of freedom
Multiple R-squared:  0.7724,	Adjusted R-squared:  0.7689 
F-statistic: 222.8 on 3 and 197 DF,  p-value: < 2.2e-16
```

3.2.2

```
Call:
lm(formula = price ~ length * height, data = usedcars_dataset)

Residuals:
   Min     1Q Median     3Q    Max 
 -9769  -3164  -1151   1776  24739 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -11465.385 130451.171  -0.088    0.930
length           411.960    747.180   0.551    0.582
height         -1258.934   2411.873  -0.522    0.602
length:height      2.196     13.768   0.159    0.873

Residual standard error: 5477 on 197 degrees of freedom
Multiple R-squared:  0.5321,	Adjusted R-squared:  0.525 
F-statistic: 74.68 on 3 and 197 DF,  p-value: < 2.2e-16
```

3.2.3

```
Call:
lm(formula = price ~ engine.size * compression.ratio, data =
usedcars_dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-11155.1  -2177.0   -490.1   1367.6  14870.9

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -6407.684   2890.967  -2.216   0.0278 *
engine.size                     148.209     21.105   7.022 3.45e-11 ***
compression.ratio              -161.577    283.754  -0.569   0.5697
engine.size:compression.ratio     1.895      2.064   0.918   0.3597
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3889 on 197 degrees of freedom
Multiple R-squared:  0.7641,    Adjusted R-squared:  0.7605
F-statistic: 212.7 on 3 and 197 DF,  p-value: < 2.2e-16
```

3.2.4

```
Call:
lm(formula = price ~ wheel.base * compression.ratio, data =
usedcars_dataset)

Residuals:
   Min     1Q Median     3Q    Max
-13180  -3268  -1775   1174  31065

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -61065.083  20393.580  -2.994 0.003103 **
wheel.base                     768.980    201.046   3.825 0.000176 ***
compression.ratio             -385.106   1805.622  -0.213 0.831328
wheel.base:compression.ratio     2.191     17.481   0.125 0.900382
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6466 on 197 degrees of freedom
Multiple R-squared:  0.3479,    Adjusted R-squared:  0.338
F-statistic: 35.03 on 3 and 197 DF,  p-value: < 2.2e-16
```

3.2.5

```
Call:
lm(formula = price ~ length * compression.ratio, data = usedcars_dataset)

Residuals:
   Min     1Q Median     3Q    Max 
-11950  -3470  -1203   1835  26152 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)              -61389.876  17682.296  -3.472 0.000635 ***
length                      433.398     98.822   4.386 1.88e-05 ***
compression.ratio          -369.224   1667.291  -0.221 0.824971    
length:compression.ratio      1.603      9.220   0.174 0.862122    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5782 on 197 degrees of freedom
Multiple R-squared:  0.4786,	Adjusted R-squared:  0.4707 
F-statistic: 60.28 on 3 and 197 DF,  p-value: < 2.2e-16
```

3.3.1

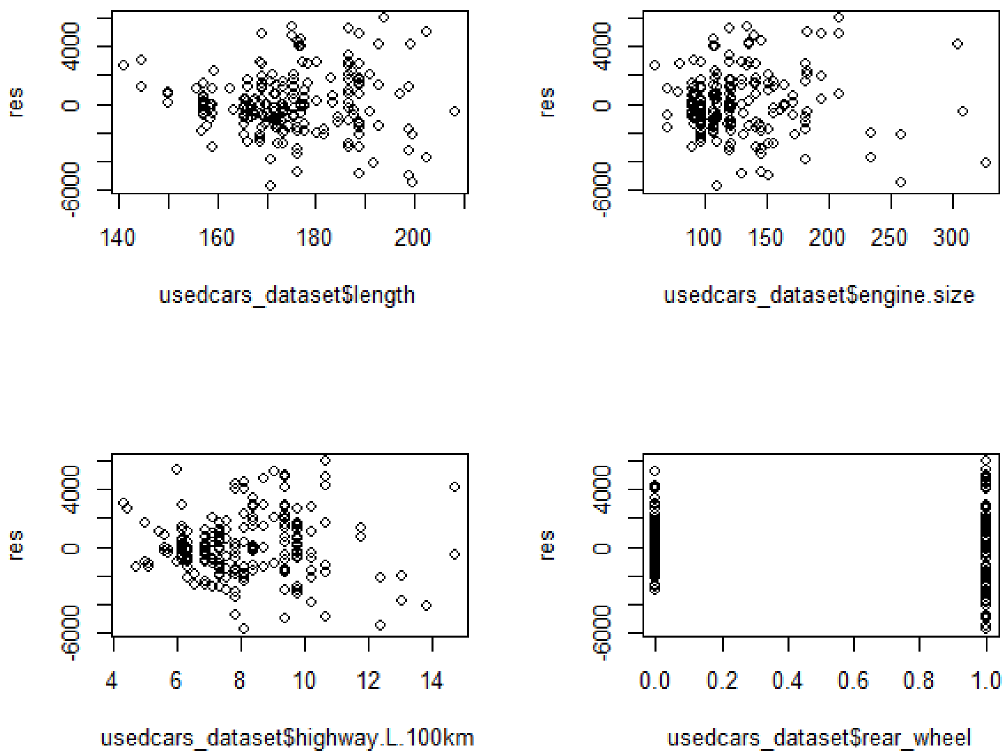
```
##fULL MODEL

model.usedcars.full<-lm(price~.,data=usedcars_dataset)
model.usedcars.full
summary(model.usedcars.full)
```

```
Call:
lm(formula = price ~ ., data = usedcars_dataset)

Residuals:
   Min     1Q Median     3Q    Max 
 -5768  -1235   -129   1175   6030 

Coefficients: (4 not defined because of singularities)
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)         6559.945   8185.041   0.801  0.42389    
wheel.base          -100.140     63.716  -1.572  0.11772    
length               100.775     34.560   2.916  0.00398 ** 
height                87.298     97.881   0.892  0.37361    
engine.size           27.849      8.682   3.208  0.00157 ** 
compression.ratio   -166.669    316.867  -0.526  0.59952    
highway.mpg          -41.052     93.292  -0.440  0.66042    
highway.L.100km      809.637    383.008   2.114  0.03585 *  
diesel              4714.932   4375.677   1.078  0.28263    
gas                       NA         NA      NA       NA    
front_wheel          462.638    881.832   0.525  0.60046    
rear_wheel          2638.792    941.406   2.803  0.00560 ** 
four_wheel                NA         NA      NA       NA    
low_price         -16120.644   1137.223 -14.175  < 2e-16 ***
median_price       -9156.028   1128.553  -8.113 6.39e-14 ***
high_price                NA         NA      NA       NA    
turbo_aspiration      44.678    622.006   0.072  0.94281    
std_aspiration            NA         NA      NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2189 on 187 degrees of freedom
Multiple R-squared:  0.9291,	Adjusted R-squared:  0.9241 
F-statistic: 188.4 on 13 and 187 DF,  p-value: < 2.2e-16
```

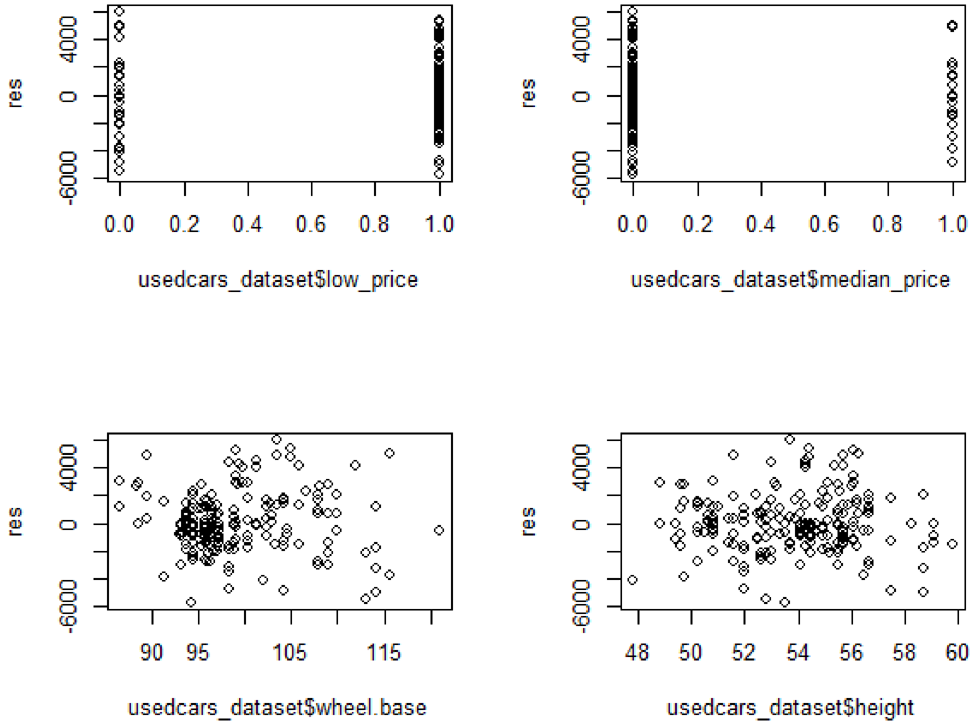
3.3.2

```
##diagnostics for full model
attach(mtcars)
par(mfrow=c(2,2))
res <- resid(model.usedcars.full)
plot(fitted(model.usedcars.full), res,main="residuals vs predicted values")
standard_res <- rstandard(model.usedcars.full)
qqnorm(standard_res)
stud_resids <- studres(model.usedcars.full)
plot(stud_resids,main="stud_residuals vs index")
lev<-hatvalues(model.usedcars.full)
plot(lev,main="leverage vs index")

## x-varaible vs res
plot(usedcars_dataset$length,res)
plot(usedcars_dataset$engine.size,res)
plot(usedcars_dataset$highway.L.100km,res)
plot(usedcars_dataset$rear_wheel,res)
plot(usedcars_dataset$low_price,res)
plot(usedcars_dataset$median_price,res)
plot(usedcars_dataset$wheel.base,res)
plot(usedcars_dataset$height,res)
plot(usedcars_dataset$compression.ratio,res)
plot(usedcars_dataset$highway.mpg,res)
plot(usedcars_dataset$turbo_aspiration,res)
plot(usedcars_dataset$front_wheel,res)
plot(usedcars_dataset$diesel,res)
```

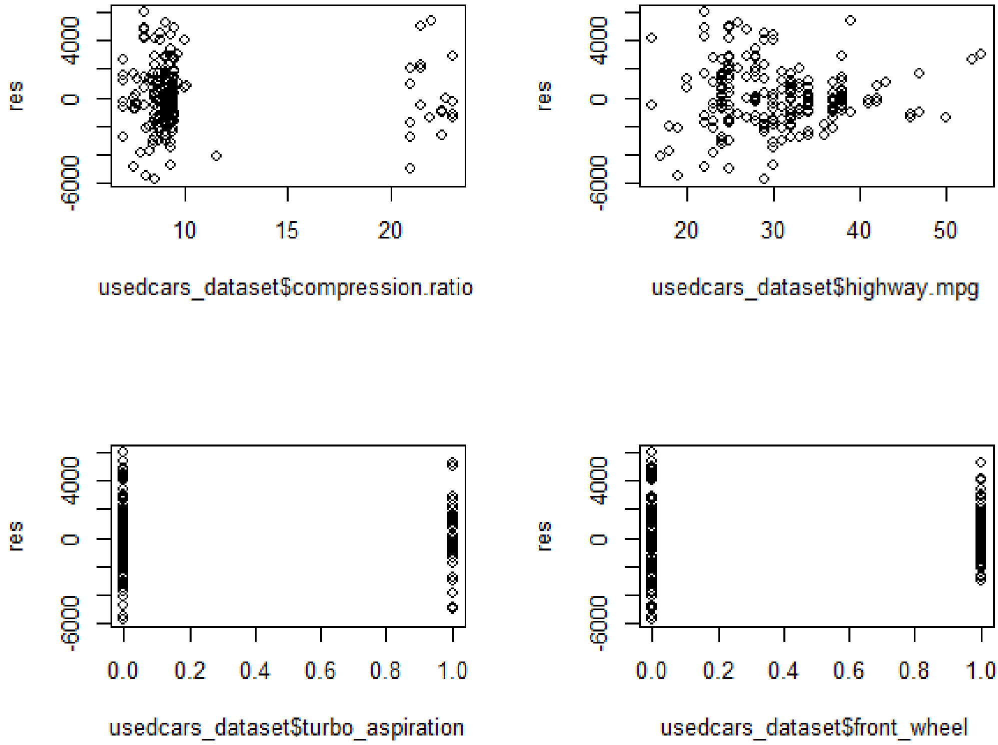
3.3.3



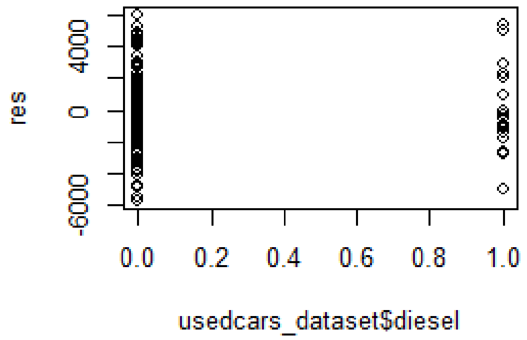
3.3.4


res
-6000
0
4000
0.0
0.2
0.4
0.6
0.8
1.0
usedcars_dataset$low_price
usedcars_dataset$median_price
90
95
105
115
usedcars_dataset$wheel.base
48
50
52
54
56
58
60
usedcars_dataset$height

3.3.5


res
usedcars_dataset$compression.ratio
res
usedcars_dataset$highway.mpg
res
usedcars_dataset$turbo_aspiration
res
usedcars_dataset$front_wheel

3.3.6



3.4

```
>
> partial_data<-usedcars_dataset[,c('price',c('length',"engine.size",
+                                            "highway.L.100km",'rear_wheel',
+                                            'low_price','median_price'))]
> model.new<-lm(price~.,data=partial_data)
> summary(model.new)

Call:
lm(formula = price ~ ., data = partial_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-6196.1 -1151.7  -214.7   965.4  6275.6 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -861.294   2988.668  -0.288 0.773512    
length             104.922     20.550   5.106 7.84e-07 ***
engine.size         32.007      8.369   3.825 0.000177 ***
highway.L.100km    633.464    154.744   4.094 6.23e-05 ***
rear_wheel        2272.002    422.694   5.375 2.18e-07 ***
low_price       -15809.782   1074.895 -14.708  < 2e-16 ***
median_price     -8531.075   1046.864  -8.149 4.43e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2249 on 194 degrees of freedom
Multiple R-squared:  0.9223,	Adjusted R-squared:  0.9199 
F-statistic: 384.1 on 6 and 194 DF,  p-value: < 2.2e-16
```

3.5.1

```
Initial Model:
price ~ 1

Final Model:
price ~ engine.size + low_price + highway.L.100km + median_price + 
    compression.ratio + rear_wheel + length


                Step Df   Deviance Resid. Df  Resid. Dev      AIC
1                                        200 12631172689 3611.182
2      + engine.size  1 9611926358       199  3019246331 3325.519
3        + low_price  1 1214029662       198  1805216669 3224.139
4  + highway.L.100km  1  322145429       197  1483071240 3186.629
5     + median_price  1  181967224       196  1301104015 3162.318
6 + compression.ratio  1  195573632       195  1105530383 3131.578
7       + rear_wheel  1  122660498       194   982869885 3109.939
8           + length  1   64341881       193   918528004 3098.331
```

3.5.2

```
  engine.size         low_price   highway.L.100km      median_price compression.ratio 
     4.901850          5.834145          4.530854          3.569567          1.501441 
   rear_wheel            length 
     1.711962          2.884903 
```

3.6.1

```
Initial Model:
price ~ wheel.base + length + height + engine.size + compression.ratio +
    highway.mpg + highway.L.100km + diesel + gas + front_wheel +
    rear_wheel + four_wheel + low_price + median_price + high_price +
    turbo_aspiration + std_aspiration

Final Model:
price ~ wheel.base + length + engine.size + highway.L.100km +
    diesel + rear_wheel + low_price + median_price


                  Step Df   Deviance Resid. Df Resid. Dev      AIC
1                                          187  896176311 3105.379
2     - std_aspiration  0       0.00       187  896176311 3105.379
3         - high_price  0       0.00       187  896176311 3105.379
4         - four_wheel  0       0.00       187  896176311 3105.379
5                - gas  0       0.00       187  896176311 3105.379
6   - turbo_aspiration  1   24726.06       188  896201037 3103.384
7        - highway.mpg  1  958090.20       189  897159128 3101.599
8        - front_wheel  1 1233361.55       190  898392489 3099.875
9  - compression.ratio  1 2661669.74       191  901054159 3098.470
10            - height  1 2930127.89       192  903984287 3097.123
```

3.6.2

```
 wheel.base          length     engine.size highway.L.100km          diesel      rear_wheel
   4.674337        6.292092        4.930265        4.236303        1.503873        1.705896
  low_price    median_price
   5.835516        3.591458
```

3.7.1

```
Initial Model:
price ~ 1

Final Model:
price ~ engine.size + low_price + highway.L.100km + median_price +
    compression.ratio + rear_wheel + length


                 Step Df   Deviance Resid. Df  Resid. Dev      AIC
1                                         200 12631172689 3611.182
2       + engine.size  1 9611926358       199  3019246331 3325.519
3         + low_price  1 1214029662       198  1805216669 3224.139
4   + highway.L.100km  1  322145429       197  1483071240 3186.629
5      + median_price  1  181967224       196  1301104015 3162.318
6 + compression.ratio  1  195573632       195  1105530383 3131.578
7        + rear_wheel  1  122660498       194   982869885 3109.939
8            + length  1   64341881       193   918528004 3098.331
```

3.7.2

```
     engine.size        low_price   highway.L.100km      median_price compression.ratio
        4.901850         5.834145          4.530854          3.569567          1.501441
      rear_wheel           length
        1.711962         2.884903
```

4.1

```
> AIC(model.new)
[1] 3681.938
> AIC(select.forward.AIC)
[1] 3670.744
> AIC(select.backward.AIC)
[1] 3669.536
> AIC(select.stepwise.AIC)
[1] 3670.744
```

4.2.1

```
> summary(model.new)

Call:
lm(formula = price ~ ., data = partial_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-6196.1 -1151.7  -214.7   965.4  6275.6 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -861.294   2988.668  -0.288 0.773512    
length             104.922     20.550   5.106 7.84e-07 ***
engine.size         32.007      8.369   3.825 0.000177 ***
highway.L.100km    633.464    154.744   4.094 6.23e-05 ***
rear_wheel        2272.002    422.694   5.375 2.18e-07 ***
low_price       -15809.782   1074.895 -14.708  < 2e-16 ***
median_price     -8531.075   1046.864  -8.149 4.43e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2249 on 194 degrees of freedom
Multiple R-squared:  0.9223,	Adjusted R-squared:  0.9199 
F-statistic: 384.1 on 6 and 194 DF,  p-value: < 2.2e-16
```

4.2.2

```
> summary(select.forward.AIC)

Call:
lm(formula = price ~ engine.size + low_price + highway.L.100km + 
    median_price + compression.ratio + rear_wheel + length, data = usedcars_dataset)

Residuals:
    Min      1Q  Median      3Q     Max 
-5869.5 -1239.8  -212.2  1144.0  6663.2 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)          -88.66    2907.50  -0.030 0.975704    
engine.size           27.36       8.22   3.329 0.001045 ** 
low_price         -15746.45    1043.03 -15.097  < 2e-16 ***
highway.L.100km      982.04     178.38   5.505 1.16e-07 ***
median_price       -8787.33    1018.15  -8.631 2.25e-15 ***
compression.ratio    170.78      47.20   3.619 0.000378 ***
rear_wheel          2013.27     416.29   4.836 2.70e-06 ***
length                78.18      21.26   3.677 0.000306 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2182 on 193 degrees of freedom
Multiple R-squared:  0.9273,	Adjusted R-squared:  0.9246 
F-statistic: 351.6 on 7 and 193 DF,  p-value: < 2.2e-16
```

4.2.3

```
> summary(select.backward.AIC)

Call:
lm(formula = price ~ wheel.base + length + engine.size + highway.L.100km + 
    diesel + rear_wheel + low_price + median_price, data = usedcars_dataset)

Residuals:
    Min      1Q  Median      3Q     Max 
-5612.6 -1211.2  -177.2  1132.1  6256.2 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       3996.64    3306.35   1.209 0.228235    
wheel.base         -76.62      54.68  -1.401 0.162757    
length             109.43      31.23   3.504 0.000571 ***
engine.size         26.27       8.20   3.204 0.001589 ** 
highway.L.100km    954.40     171.56   5.563 8.81e-08 ***
diesel            2509.17     627.02   4.002 8.97e-05 ***
rear_wheel        2099.13     413.32   5.079 8.96e-07 ***
low_price       -15904.66    1037.55 -15.329  < 2e-16 ***
median_price     -8890.62    1015.78  -8.752 1.07e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2170 on 192 degrees of freedom
Multiple R-squared:  0.9284,	Adjusted R-squared:  0.9255 
F-statistic: 311.3 on 8 and 192 DF,  p-value: < 2.2e-16
```

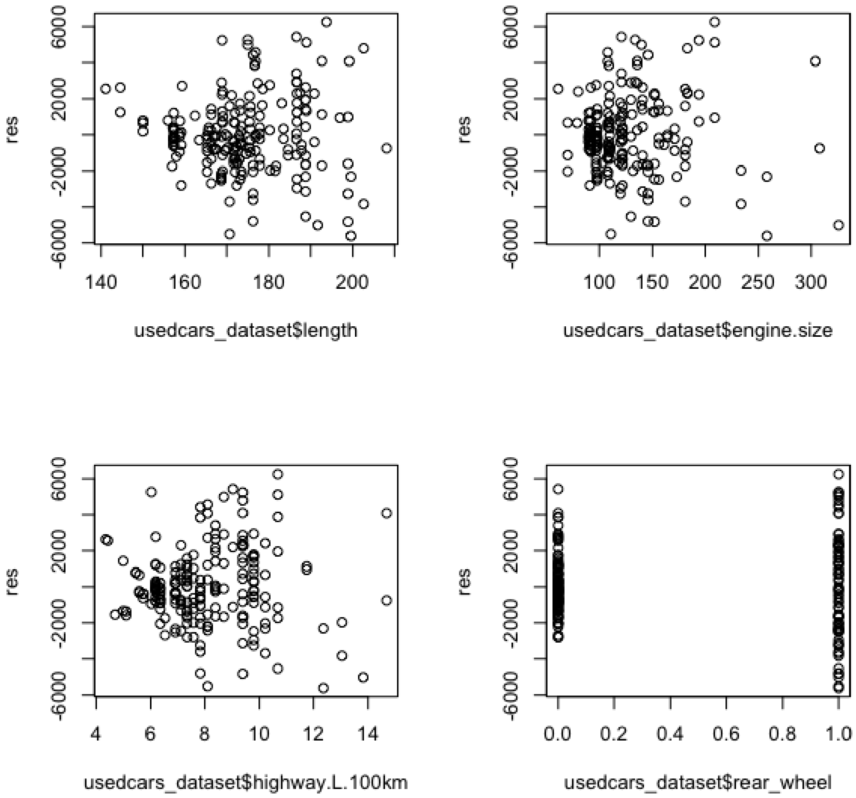
4.2.4

```
> summary(select.stepwise.AIC)

Call:
lm(formula = price ~ engine.size + low_price + highway.L.100km +
    median_price + compression.ratio + rear_wheel + length, data = usedcars_dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-5869.5 -1239.8  -212.2  1144.0  6663.2

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          -88.66    2907.50  -0.030 0.975704
engine.size           27.36       8.22   3.329 0.001045 **
low_price         -15746.45    1043.03 -15.097  < 2e-16 ***
highway.L.100km      982.04     178.38   5.505 1.16e-07 ***
median_price       -8787.33    1018.15  -8.631 2.25e-15 ***
compression.ratio    170.78      47.20   3.619 0.000378 ***
rear_wheel          2013.27     416.29   4.836 2.70e-06 ***
length                78.18      21.26   3.677 0.000306 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2182 on 193 degrees of freedom
Multiple R-squared:  0.9273,    Adjusted R-squared:  0.9246
F-statistic: 351.6 on 7 and 193 DF,  p-value: < 2.2e-16
```

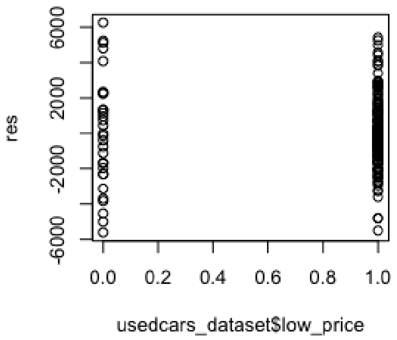
5.1.1

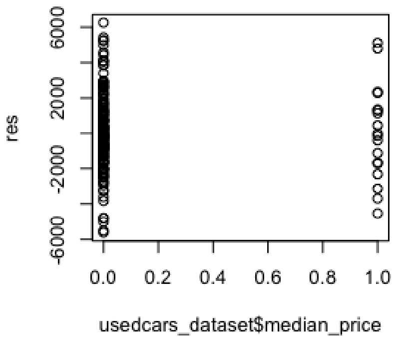
```r
attach(mtcars)
par(mfrow=c(2,2))

res <- resid(select.backward.AIC)
plot(fitted(select.backward.AIC), res,main="residuals vs predicted values")
standard_res <- rstandard(select.backward.AIC)
qqnorm(standard_res)
stud_resids <- studres(select.backward.AIC)
plot(stud_resids,main="stud_residuals vs index")
lev<-hatvalues(select.backward.AIC)
plot(lev,main="leverage vs index")

plot(usedcars_dataset$length,res)
plot(usedcars_dataset$engine.size,res)
plot(usedcars_dataset$highway.L.100km,res)
plot(usedcars_dataset$rear_wheel,res)
plot(usedcars_dataset$low_price,res)
plot(usedcars_dataset$median_price,res)
plot(usedcars_dataset$compression.ratio,res)
```

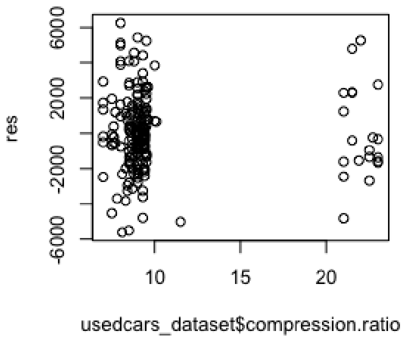
5.1.2


res
usedcars_dataset$length
res
usedcars_dataset$engine.size
res
usedcars_dataset$highway.L.100km
res
usedcars_dataset$rear_wheel

5.1.3


res
6000
2000
-2000
-6000
0.0
0.2
0.4
0.6
0.8
1.0
usedcars_dataset$low_price


res
6000
2000
-2000
-6000
0.0
0.2
0.4
0.6
0.8
1.0
usedcars_dataset$median_price


res
6000
2000
-2000
-6000
10
15
20
usedcars_dataset$compression.ratio