pstat-126-extra

Zejie Gao

2023-03-23

Our goal is to model the response mpg in terms of the rest of the variables (except name).

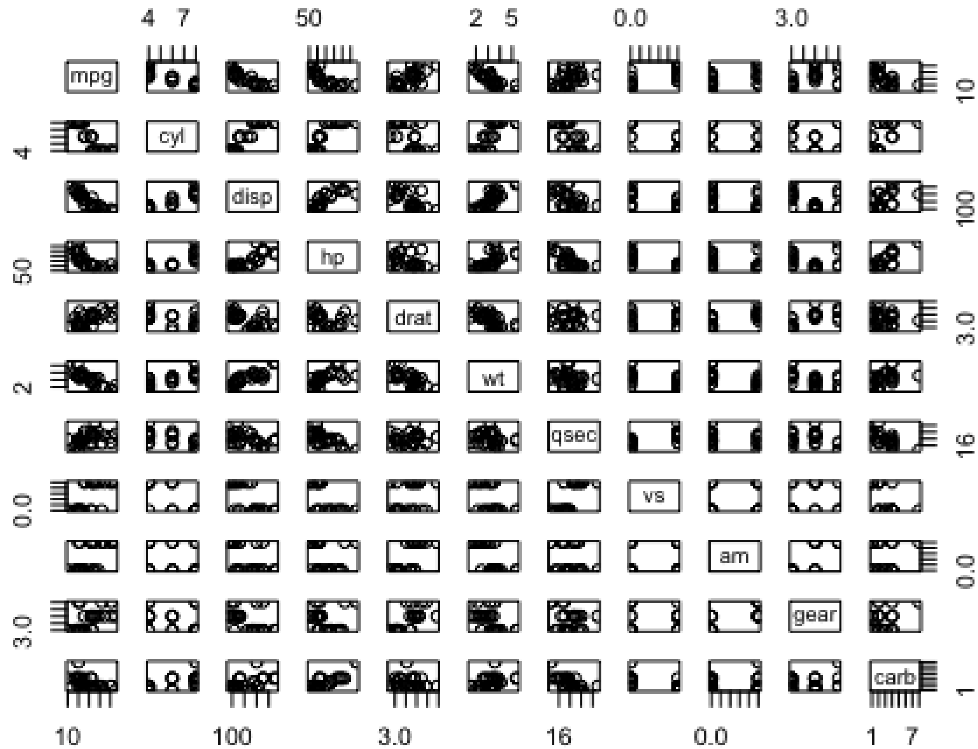
Partition the data set into two sets a training data and a test data. Remove every fifth observation from the data for use as a test sample. Perform an exploratory analysis. Comment on your findings. Perform a regression analysis and come up with the best multiple linear regression model that explains the response mpg in terms of the rest (except name). Comment on your findings and explain the methods and strategies that you employed in order to select the model you picked. Things you have to include in this part: - Model diagnostics - Justification on whether it is necessary or not to do any transformation on the response or the predictors - Variable selection Assess the prediction performance by using the test sample.

```
Car <- read.table("cars (1).txt",header=T)
str(Car)

## 'data.frame':    32 obs. of  12 variables:
##  $ name: chr  "Mazda RX4" "Mazda RX4 Wag" "Datsun 710" "Hornet 4 Dri
ve" ...
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : int  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : int  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : int  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : int  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: int  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: int  4 4 1 1 2 1 4 2 2 4 ...

Car <- as.data.frame(Car)

test_indices <- seq(5, nrow(Car), by=5)
test_data <- Car[test_indices,]
train_data <- Car[-test_indices, ]
```

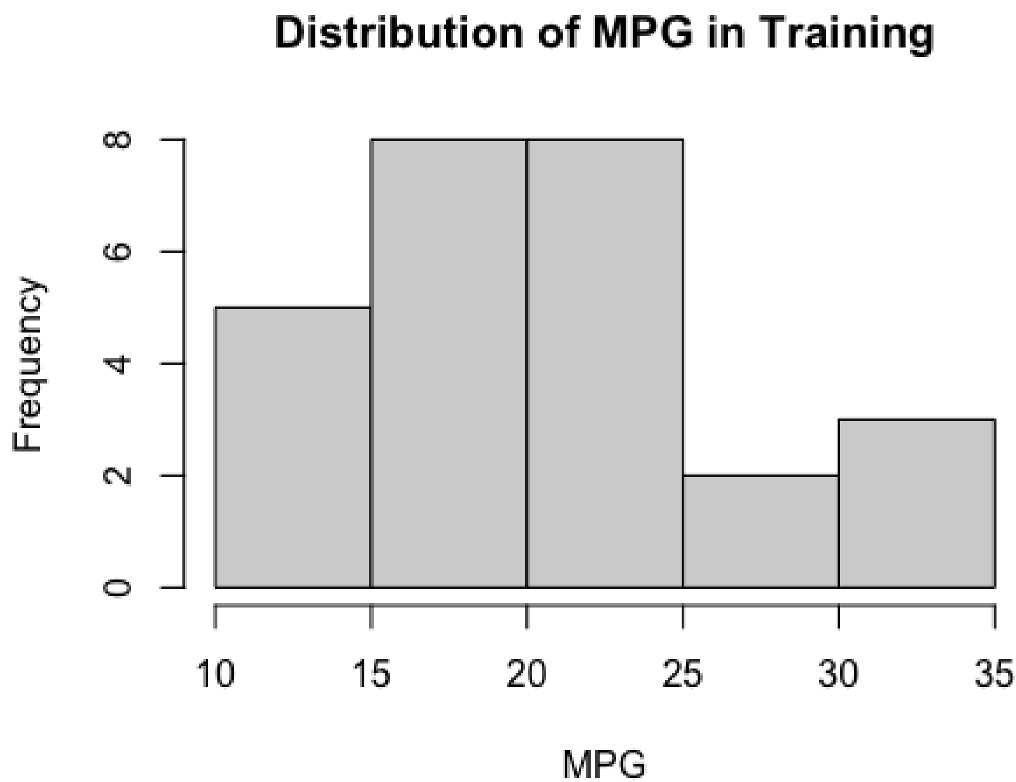
1. To perform some exploratory analysis on data car, I create a scatterplot matrix to visualize the relationships between all the variables, a correlation matrix to examine the pairwise correlations between variables, and histograms, density plots, and boxplots to explore the distribution of the response variable “mpg”. From the correlation matrix, there are 13.36577%

correlation between variables higher than 0.9 or lower than -0.9. This data indicate possible high pairwise collinearity that may impact our data analysis. Based on the histogram and density plot, most of the mpg value fall bettween 15 and 25 and the distribution is right-skewed.

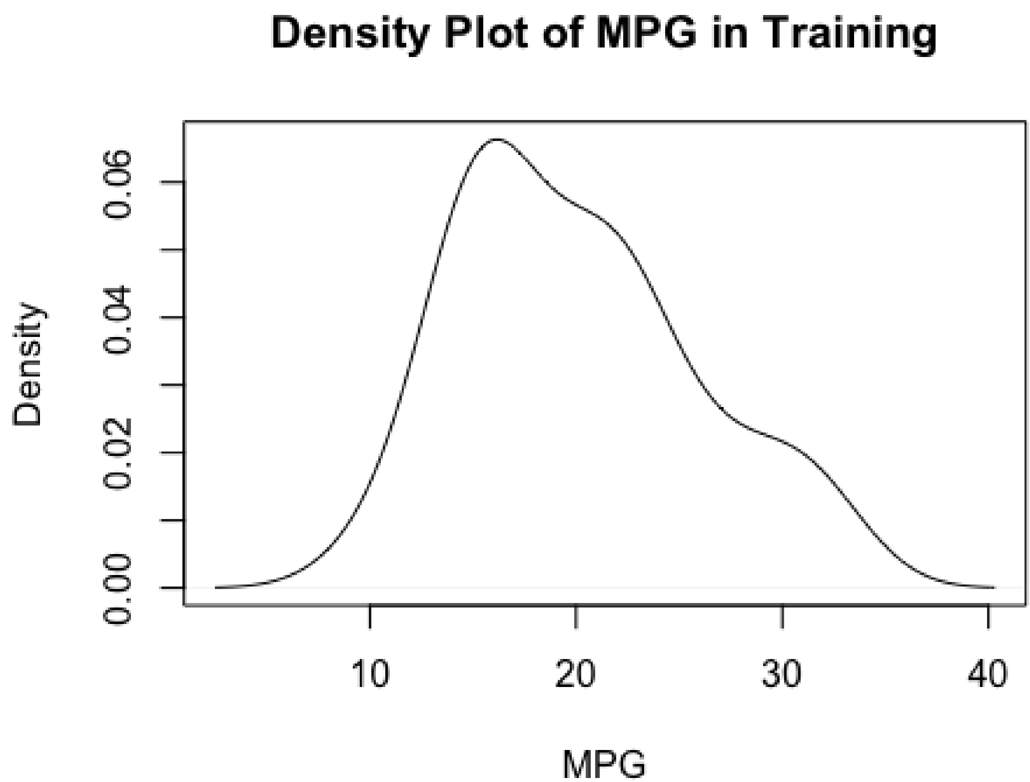
```
summary(train_data)

##      name                mpg             cyl             disp      
##  Length:26          Min.   :10.40   Min.   :4.000   Min.   : 75.7  
##  Class :character   1st Qu.:15.28   1st Qu.:4.000   1st Qu.:120.5  
##  Mode  :character   Median :19.55   Median :6.000   Median :196.3  
##                     Mean   :20.07   Mean   :6.077   Mean   :221.8  
##                     3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:303.2  
##                     Max.   :32.40   Max.   :8.000   Max.   :460.0  
##        hp             drat             wt             qsec      
##  Min.   : 52.0   Min.   :2.760   Min.   :1.513   Min.   :14.50  
##  1st Qu.: 95.5   1st Qu.:3.098   1st Qu.:2.504   1st Qu.:16.88  
##  Median :111.5   Median :3.715   Median :3.203   Median :17.71  
##  Mean   :145.2   Mean   :3.622   Mean   :3.168   Mean   :17.90  
##  3rd Qu.:180.0   3rd Qu.:3.920   3rd Qu.:3.570   3rd Qu.:18.90  
##  Max.   :335.0   Max.   :4.930   Max.   :5.424   Max.   :22.90  
##        vs               am              gear            carb      
##  Min.   :0.0000   Min.   :0.0000   Min.   :3.000   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
##  Median :0.0000   Median :0.0000   Median :4.000   Median :2.000  
##  Mean   :0.4615   Mean   :0.4231   Mean   :3.692   Mean   :2.731  
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :1.0000   Max.   :1.0000   Max.   :5.000   Max.   :8.000

sum(is.na(train_data))

## [1] 0

pairs(train_data[, -1])
```




```
corr_matrix <- cor(train_data[,c(-1,-2)])
high_cor <- sum(corr_matrix> 0.9 | corr_matrix < 0.9) / sum(corr_matrix
)
high_cor

## [1] 13.36577

hist(train_data$mpg,
     main="Distribution of MPG in Training", xlab="MPG")
```

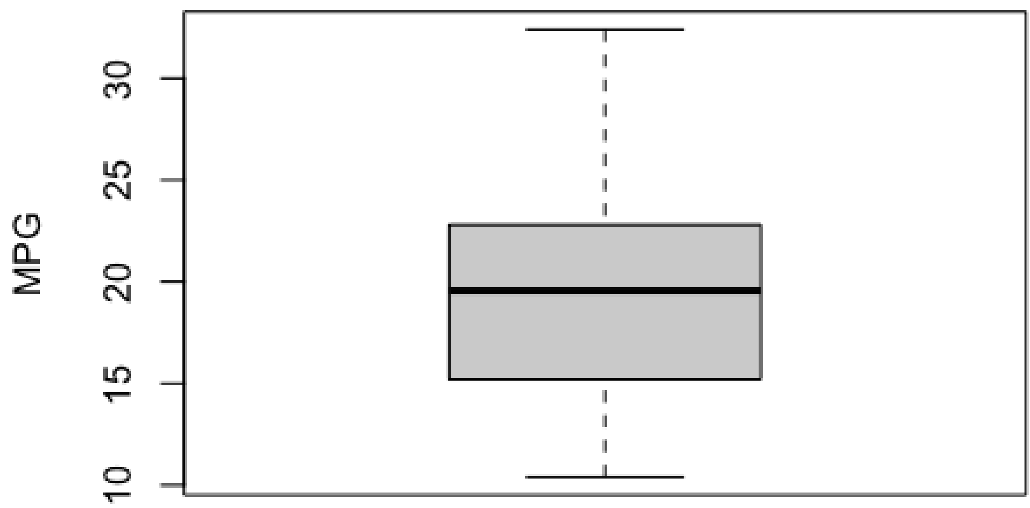



```
plot(density(train_data$mpg),
     main="Density Plot of MPG in Training",
     xlab="MPG",
     ylab="Density")
```




```
boxplot(train_data$mpg,
        main="Boxplot of MPG in Training", ylab="MPG")
```

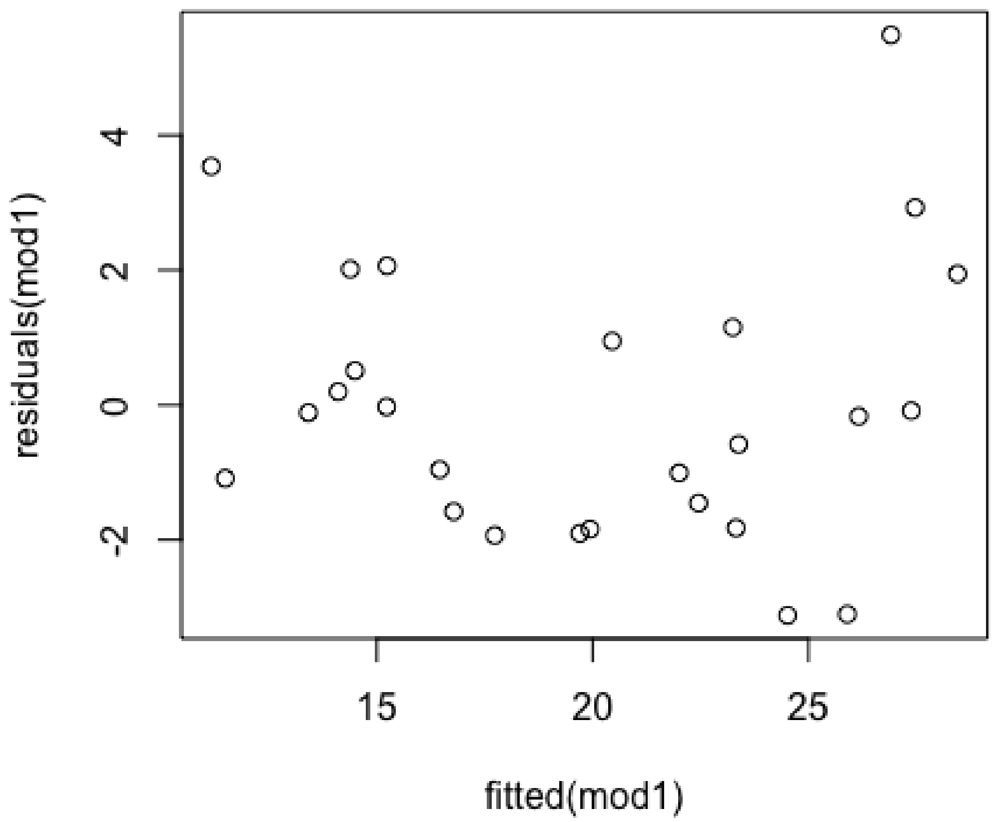
Boxplot of MPG in Training



2.

Model diagnostics on error (a) constant variance No clear trend on this graph represent the residual could have a constant variance. In addition, ncvTest help prove the contstant variance.

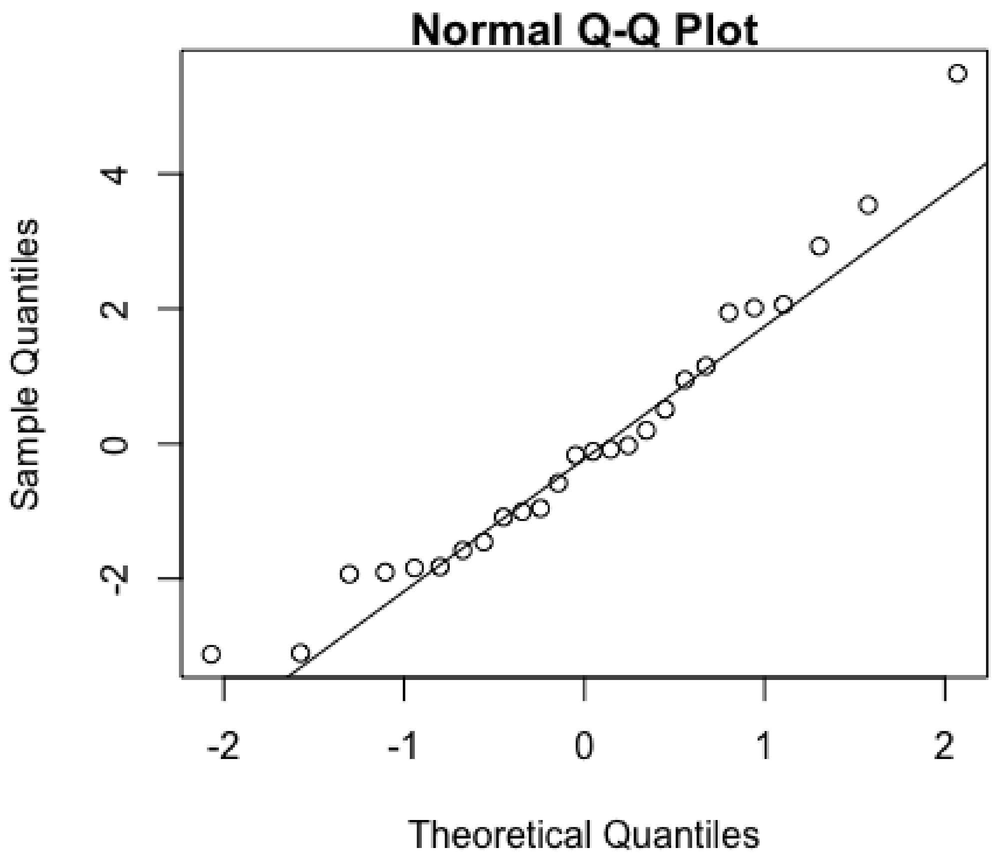
```
mod1 <- lm(mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb,
           train_data)
par(mar = c(5,5,1,2))
plot(fitted(mod1), residuals(mod1))
```




```
car::ncvTest(mod1)

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 2.222176, Df = 1, p = 0.13604
```

(b) normality Due to small p-value, we could not reject null hypothesis of the normality. Thus it is normal.

```
par(mar = c(5,5,1,2))
qqnorm(residuals((mod1),
                 ylab = "Residuals",
                 main = 'Residual vs Theoretical quantiles',
                 pch = 18))
qqline(residuals(mod1))
```

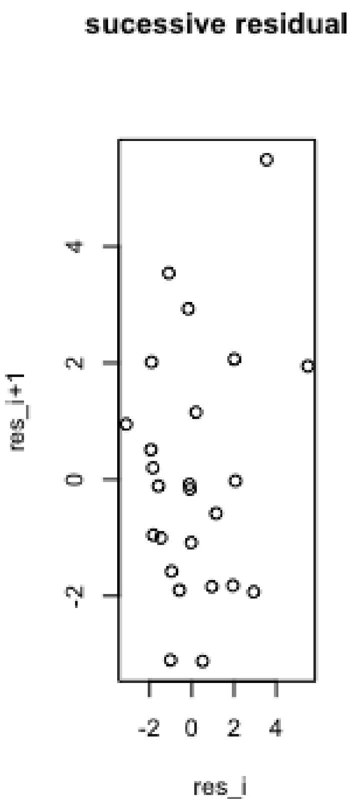



```
shapiro.test(residuals(mod1))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(mod1)
## W = 0.95012, p-value = 0.2334
```

(c) Independence Due to small value 0.03571, we could accept the alternative hyposis that the true autocorrealtion is greater than 0.

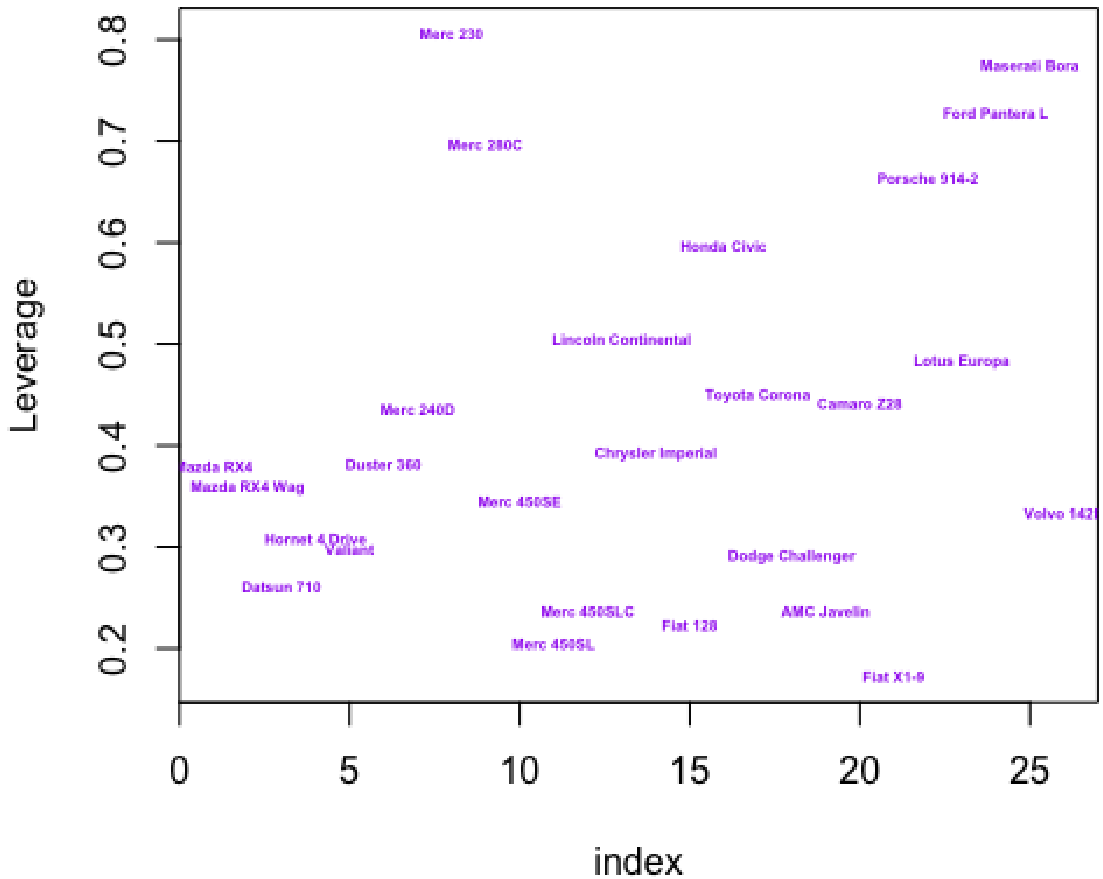
```
dim(train_data)[1]

## [1] 26

y_hat <- mod1$fitted.values
e_hat <- mod1$residuals
par(mfrow = c(1, 3), mar = c(4,4,8,2))
n <- dim(train_data)[1]
plot(mod1$residuals[1:(n-1)], mod1$residuals[2:n],
     xlab = " res_i",
     ylab = "res_i+1",
     main = "sucessive residual")
dwtest(mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
, data = train_data)
```

```
##
##  Durbin-Watson test
##
## data:  mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + c
arb
## DW = 1.5555, p-value = 0.03571
## alternative hypothesis: true autocorrelation is greater than 0
```



3. Model diagnosis on unusual observation (a) high leverage No high leverage point exist in this data.

```
hatv <- hatvalues(mod1)
Car_lev <- data.frame(index = seq(length(hatv)),
                           Leverage = hatv, namesC = train_data$name)
par(mar = c(4,4,0.5,0.5))
plot(Leverage ~ index, data = Car_lev, col = "white", pch = NULL)
text(Leverage ~index, labels = namesC, data = Car_lev , cex = 0.4, font
 = 2, col = "purple")
abline(h =2*sum(hatv)/dim(Car_lev)[1], col = "orange", lty = 2)
```




```r
sum(hatv > 2*sum(hatv)/dim(Car_lev)[1])

## [1] 0

high_lev <- train_data|>
  filter(hatv > 2*sum(hatv)/dim(Car_lev)[1])
high_lev

##  [1] name mpg  cyl  disp hp   drat wt   qsec vs   am   gear carb
## <0 行> (或0-长度的row.names)
```

(b) outliers In this case, we do not have outlier.

```r
r <- rstandard(mod1)
outliers <- sum(r > 3 | r< -3)
outliers

## [1] 0
```

(c) influential observations There are five influential observations exists in our train_data.

```r
X <- model.matrix(mod1)
H <- X %*% solve(t(X) %*% X) %*% t(X)
print(H[1:5, 1:5])
```

```
##              1            2            3           4           6
## 1  0.37939459  0.352379043 -0.029983249 -0.02826913 -0.04532256
## 2  0.35237904  0.357957531 -0.006488445 -0.06933762 -0.03610984
## 3 -0.02998325 -0.006488445  0.261212766  0.03569016  0.10706049
## 4 -0.02826913 -0.069337616  0.035690155  0.30826984  0.22559781
## 6 -0.04532256 -0.036109843  0.107060489  0.22559781  0.29798286
```

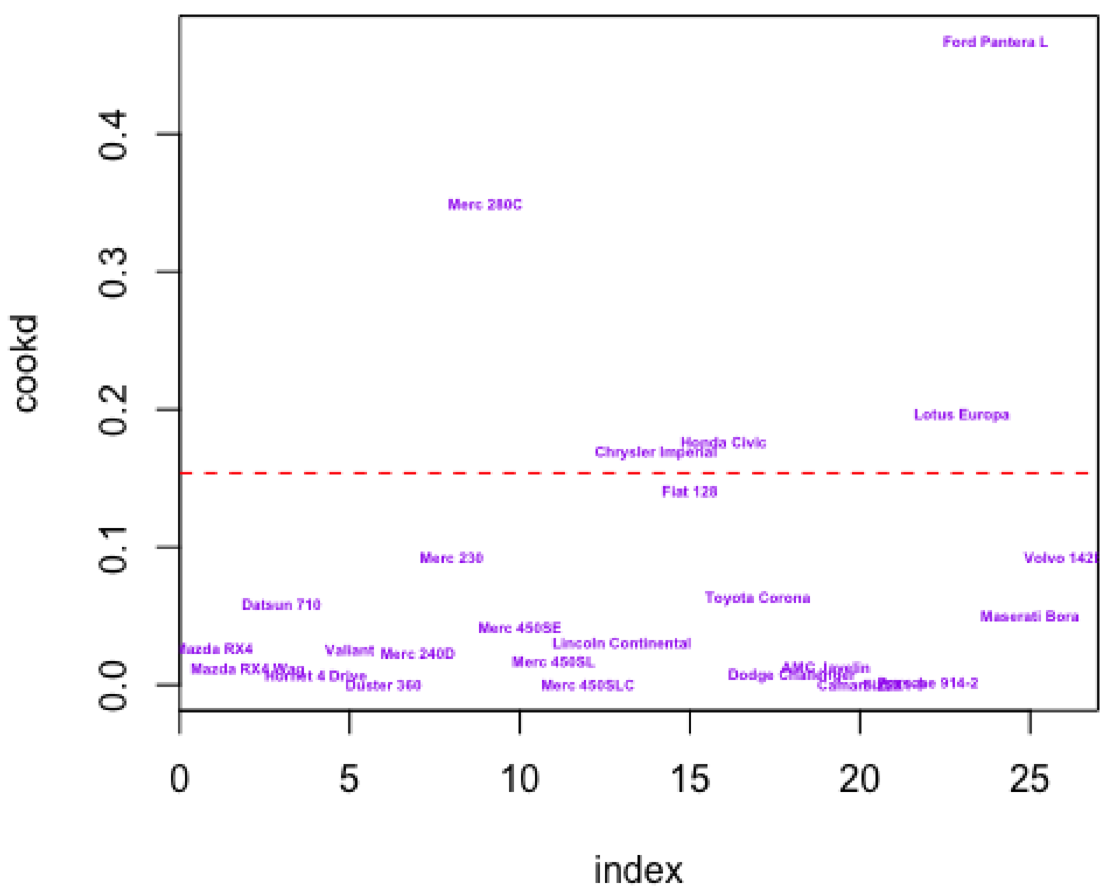
```
sum_diag <-sum(diag(H)); sum_diag
```

```
## [1] 11
```

```
p_star <- ncol(X); p_star
```

```
## [1] 11
```

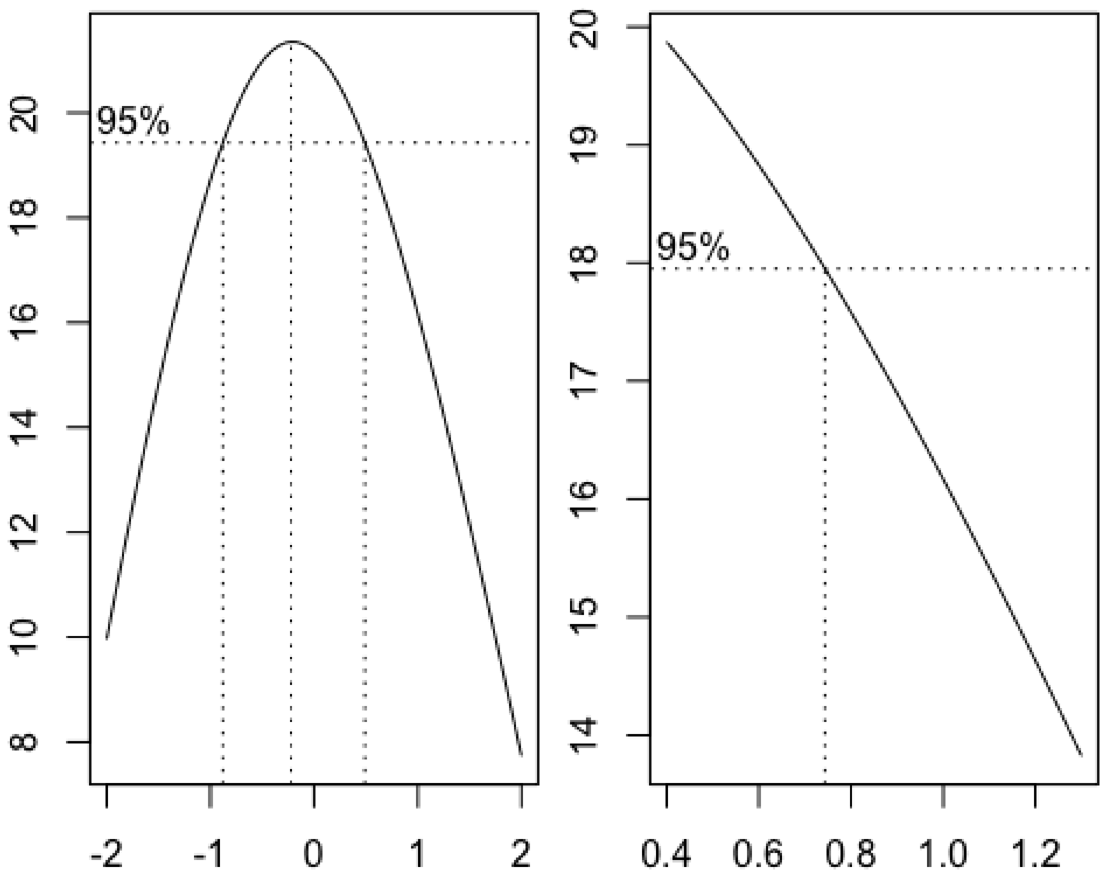
```
cook <- cooks.distance(mod1)
Car_cook <- data.frame(index = seq(length(cook)),
                            cookd = abs(cook), namesC = train_data$name
)
par(mar = c(4,4,0.5,0.5))
plot(cookd ~ index, data = Car_cook, col = "white", pch = NULL)
text(cookd ~index, labels = namesC, data = Car_cook , cex = 0.4,
     font = 2, col = "purple")
abline(h = 4/dim(X)[1], col = "red", lty = 2)
```



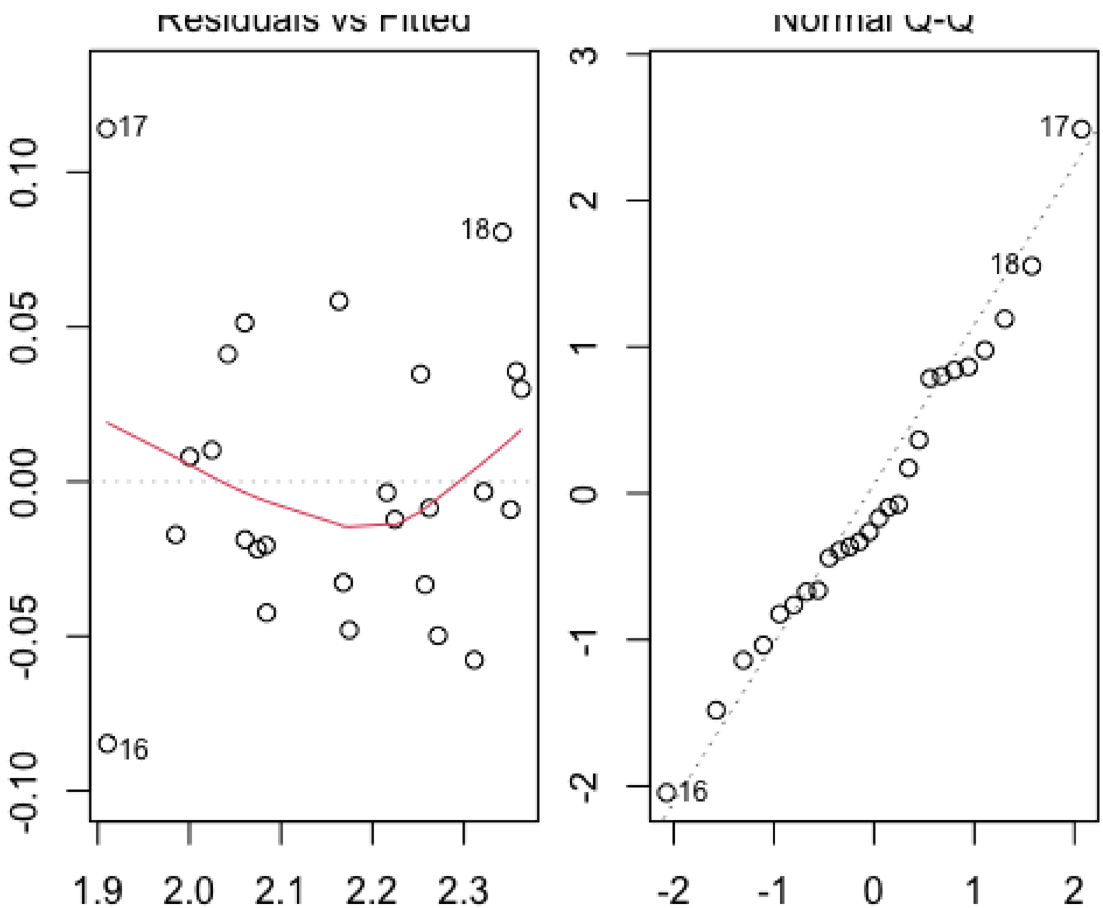
```
sum(cook >= 4/dim(X)[1])

## [1] 5
```

3. Transformation Since the confidence interval do not contains lambda = 1, transformation is necessary.

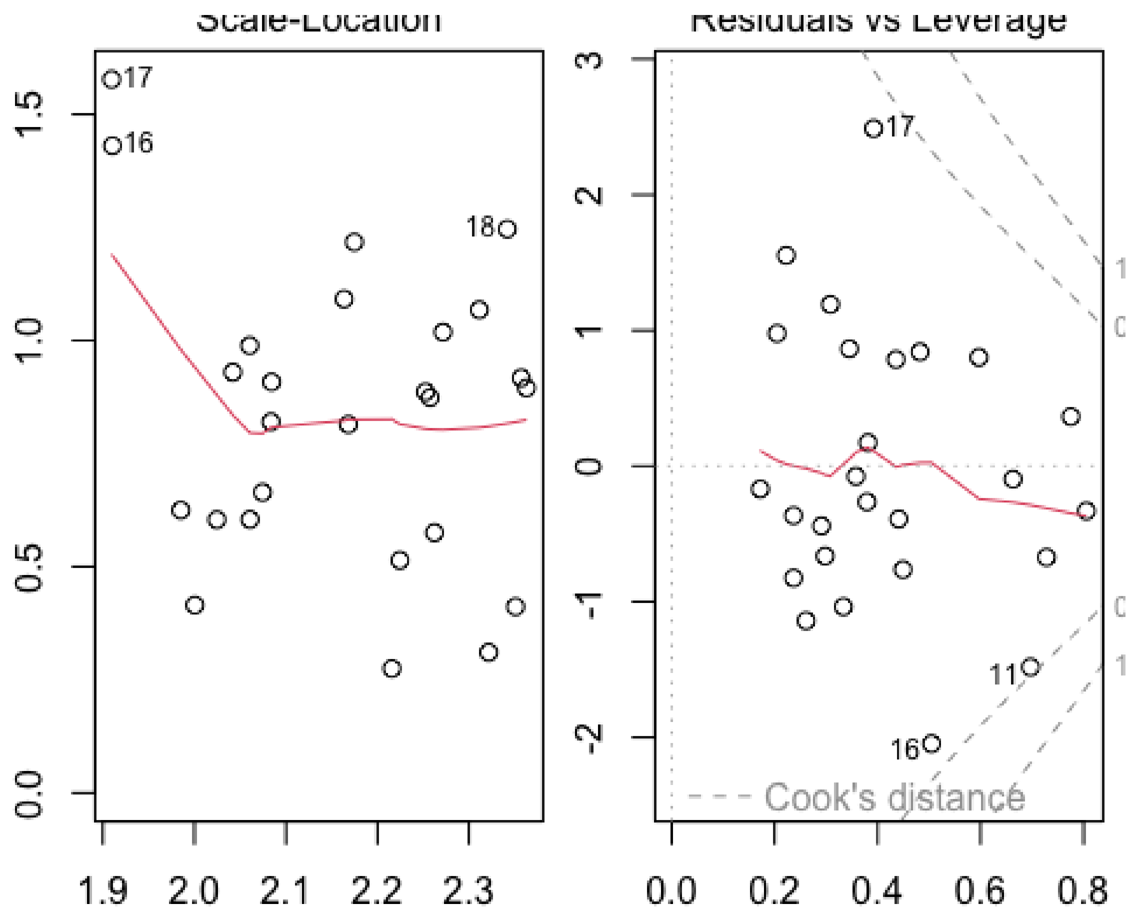
```
par(mfrow = c(1, 2), mar = c(2, 2, 0.8, 0.5))
bc <- boxcox(mod1, plotit = TRUE)
boxcox(mod1, plotit = TRUE, lambda = seq(0.4, 1.3, by = 0.1))
```




```
lambda <- bc$x[which.max(bc$y)]; lambda

## [1] -0.2222222

train_data_new <- train_data |>
  mutate(mpg = (mpg^(lambda)-1)/lambda)

test_data_new <- train_data |>
  mutate(mpg = (mpg^(lambda)-1)/lambda)
# change both train and test
mod2 <- lm(mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear +
carb,
           train_data_new)
plot(mod2)
```

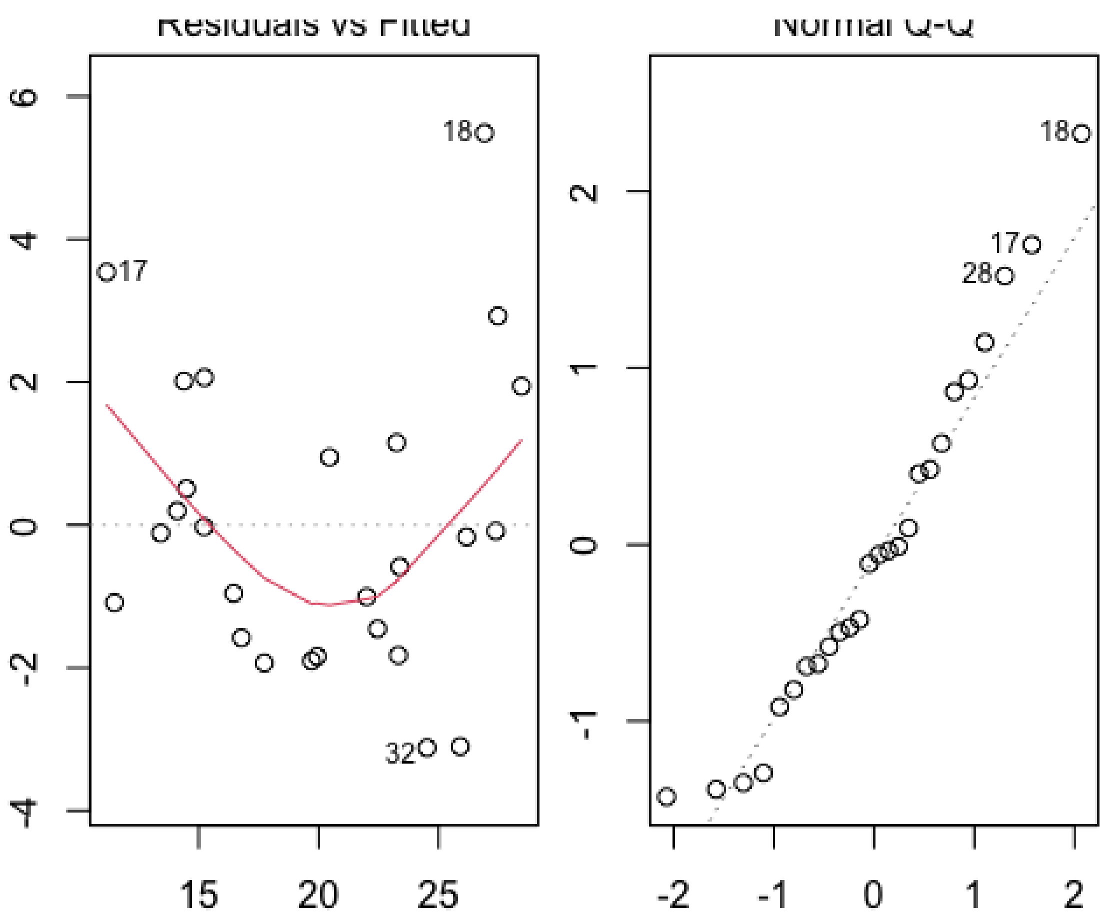


Residuals vs Fitted
Normal Q-Q
17
18
16
0.10
0.05
0.00
-0.05
-0.10
1.9
2.0
2.1
2.2
2.3
3
2
1
0
-1
-2

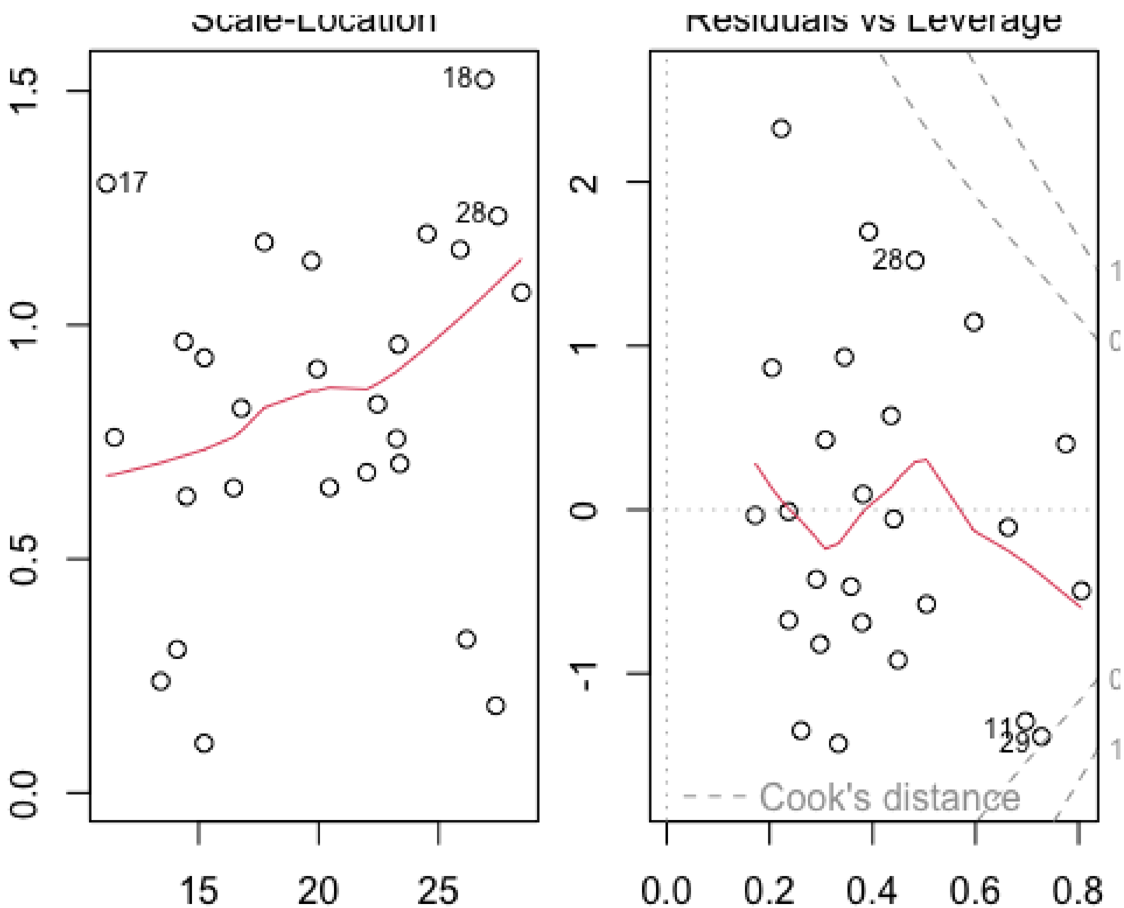


Scale-Location
Residuals vs Leverage
17
16
18
11
Cook's distance
1.5
1.0
0.5
0.0
3
2
1
0
-1
-2
1.9
2.0
2.1
2.2
2.3
0.0
0.2
0.4
0.6
0.8


```
plot(mod1)
```



Residuals vs Fitted
Normal Q-Q
18
17
32
28
6
4
2
0
-2
-4
15
20
25
-2
-1
1



4. model selection After performing the necessary analyses, it was found that the mod3 model (mpg ~ hp + wt + qsec + gear) has the lowest AIC and MSE compared to the other models tested using ridge and lasso regression. Based on these findings, it is suggested that lasso regression favors the inclusion of only the four predictors in mod3.

Furthermore, ridge regression resulted in a higher MSE compared to mod3, indicating that mod3 provides a better fit to the data. However, the difference in MSE between ridge regression and mod3 was not very large. Therefore, if researchers want to include more variables in the model, ridge regression may be a better choice.

```
step(mod2, direction = "backward")

## Start:  AIC=-139.62
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##          Df Sum of Sq       RSS     AIC
## - drat  1 0.0000127 0.051926 -141.62
## - vs    1 0.0000235 0.051937 -141.61
## - cyl   1 0.0002562 0.052170 -141.50
## - carb  1 0.0003009 0.052215 -141.47
## - qsec  1 0.0005078 0.052422 -141.37
```

```
## - am    1 0.0007037 0.052617 -141.27
## - disp  1 0.0008407 0.052754 -141.21
## - hp    1 0.0021348 0.054049 -140.57
## - gear  1 0.0024701 0.054384 -140.41
## - wt    1 0.0036118 0.055526 -139.87
## <none>              0.051914 -139.62
##
## Step:  AIC=-141.62
## mpg ~ cyl + disp + hp + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq      RSS     AIC
## - vs    1 0.0000276 0.051954 -143.60
## - cyl   1 0.0002508 0.052177 -143.49
## - carb  1 0.0003728 0.052299 -143.43
## - qsec  1 0.0005087 0.052435 -143.36
## - am    1 0.0006920 0.052618 -143.27
## - disp  1 0.0008982 0.052825 -143.17
## - hp    1 0.0021266 0.054053 -142.57
## - gear  1 0.0024577 0.054384 -142.41
## - wt    1 0.0036715 0.055598 -141.84
## <none>              0.051926 -141.62
##
## Step:  AIC=-143.6
## mpg ~ cyl + disp + hp + wt + qsec + am + gear + carb
##
##        Df Sum of Sq      RSS     AIC
## - cyl   1 0.0003680 0.052322 -145.42
## - carb  1 0.0003935 0.052348 -145.41
## - am    1 0.0006805 0.052634 -145.26
## - qsec  1 0.0007615 0.052715 -145.22
## - disp  1 0.0008903 0.052844 -145.16
## - hp    1 0.0021799 0.054134 -144.53
## - gear  1 0.0024402 0.054394 -144.41
## - wt    1 0.0039513 0.055905 -143.70
## <none>              0.051954 -143.60
##
## Step:  AIC=-145.42
## mpg ~ disp + hp + wt + qsec + am + gear + carb
##
##        Df Sum of Sq      RSS     AIC
## - carb  1 0.0006382 0.052960 -147.10
## - am    1 0.0009574 0.053279 -146.95
## - disp  1 0.0013411 0.053663 -146.76
## - qsec  1 0.0015226 0.053845 -146.67
## - hp    1 0.0023320 0.054654 -146.29
## - wt    1 0.0039120 0.056234 -145.54
## - gear  1 0.0039819 0.056304 -145.51
## <none>              0.052322 -145.42
##
## Step:  AIC=-147.1
```

```
## mpg ~ disp + hp + wt + qsec + am + gear
##
##        Df Sum of Sq      RSS     AIC
## - disp  1 0.0007279 0.053688 -148.75
## - am    1 0.0010949 0.054055 -148.57
## - qsec  1 0.0032692 0.056229 -147.55
## - gear  1 0.0033627 0.056323 -147.50
## <none>              0.052960 -147.10
## - hp    1 0.0064932 0.059453 -146.10
## - wt    1 0.0104369 0.063397 -144.43
##
## Step:  AIC=-148.75
## mpg ~ hp + wt + qsec + am + gear
##
##        Df Sum of Sq      RSS     AIC
## - am    1 0.0015707 0.055259 -150.00
## <none>              0.053688 -148.75
## - gear  1 0.0044069 0.058095 -148.70
## - qsec  1 0.0066794 0.060368 -147.70
## - hp    1 0.0094394 0.063128 -146.54
## - wt    1 0.0308971 0.084585 -138.93
##
## Step:  AIC=-150
## mpg ~ hp + wt + qsec + gear
##
##        Df Sum of Sq      RSS     AIC
## <none>              0.055259 -150.00
## - qsec  1  0.005123 0.060382 -149.69
## - hp    1  0.012225 0.067484 -146.80
## - gear  1  0.014106 0.069365 -146.09
## - wt    1  0.036715 0.091974 -138.75

##
## Call:
## lm(formula = mpg ~ hp + wt + qsec + gear, data = train_data_new)
##
## Coefficients:
## (Intercept)           hp           wt         qsec         gear
##   2.1198251   -0.0006875   -0.0788298    0.0128869    0.0442731
```

```
mod3 <- lm(mpg ~ hp + wt + qsec + gear, data = train_data_new)
summary(mod3)
```

```
##
## Call:
## lm(formula = mpg ~ hp + wt + qsec + gear, data = train_data_new)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.081176 -0.030843 -0.009427  0.025218  0.126001
```

```
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.1198251  0.2045819  10.362 1.04e-09 ***
## hp          -0.0006875  0.0003190  -2.155  0.04288 *
## wt          -0.0788298  0.0211038  -3.735  0.00122 **
## qsec         0.0128869  0.0092357   1.395  0.17750
## gear         0.0442731  0.0191219   2.315  0.03080 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0513 on 21 degrees of freedom
## Multiple R-squared:  0.8997, Adjusted R-squared:  0.8806
## F-statistic: 47.08 on 4 and 21 DF,  p-value: 3.42e-10
```

```r
X_test <- test_data_new[,c("hp", "wt", "qsec", "gear")]
y_pred <- predict(mod3, newdata = X_test)
mse1 <- mean((test_data_new$mpg - y_pred)^2); mse1
```

```
## [1] 0.002125338
```

```r
library(glmnet)
```

```
## 载入需要的程辑包：Matrix
```

```
##
## 载入程辑包：'Matrix'
```

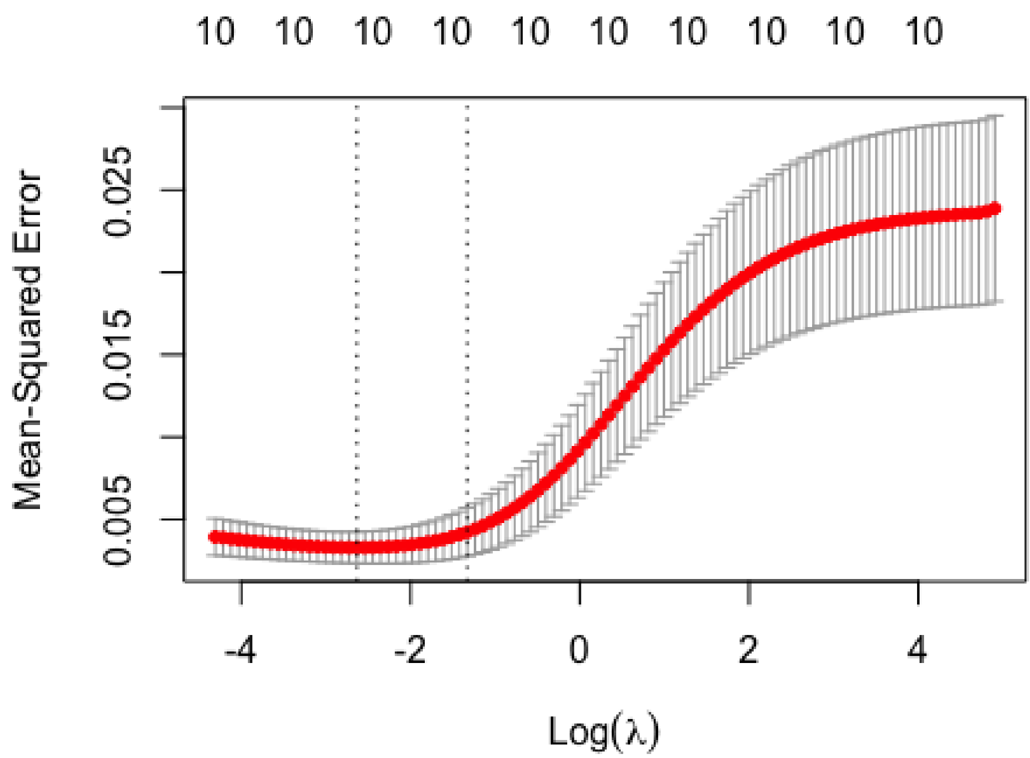
```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loaded glmnet 4.1-6
```

```r
x <- scale(data.matrix(train_data_new[, c(-1,-2)]))
y <- train_data_new$mpg

ridge_model <- cv.glmnet(x, y, alpha = 0)
```

```
## Warning: Option grouped=FALSE enforced in cv.glmnet, since < 3 obser
vations per
## fold
```

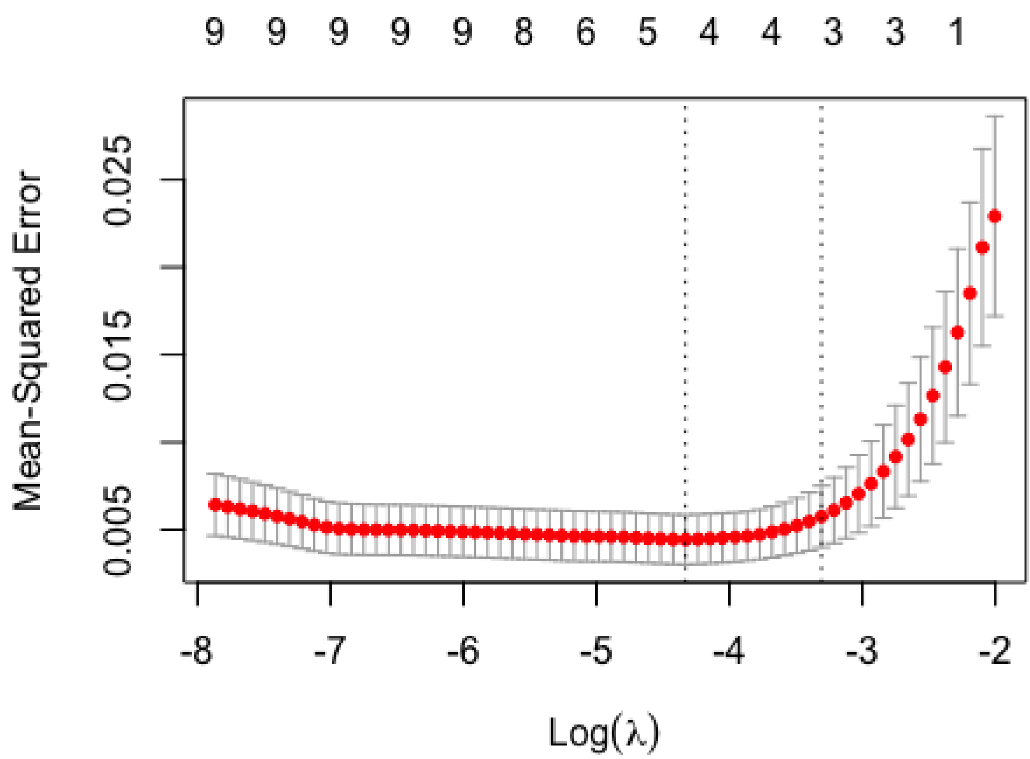
```r
best_lambda <- ridge_model$lambda.min
best_model <- glmnet(x, y, alpha = 0, lambda = best_lambda)

ridge_coef <- coef(best_model, s = "lambda.min")

plot(ridge_model)
```




```
X_test <- scale(data.matrix(test_data_new[, c(-1,-2)]))
y_pred <- predict(best_model, newx = X_test)

mse2 <- mean((test_data_new$mpg - y_pred)^2); mse2

## [1] 0.002268497

x <- scale(data.matrix(train_data_new[, c(-1,-2)]))
y <- train_data_new$mpg

lasso_model <- cv.glmnet(x, y, alpha = 1)

## Warning: Option grouped=FALSE enforced in cv.glmnet, since < 3 obser
vations per
## fold

best_lambda <- lasso_model$lambda.min
best_model <- glmnet(x, y, alpha = 1, lambda = best_lambda)

lasso_coef <- coef(best_model, s = "lambda.min")
lasso_coef

## 11 x 1 sparse Matrix of class "dgCMatrix"
##                       s1
```

```
## (Intercept)  2.16438687
## cyl         -0.03522733
## disp        -0.04211118
## hp          -0.01315853
## drat          .
## wt          -0.04386359
## qsec          .
## vs            .
## am            .
## gear          .
## carb          .
```

```
plot(lasso_model)
```




```
X_test <- scale(data.matrix(test_data_new[, c(-1,-2)]))

y_pred <- predict(best_model, newx = X_test)

mse3 <- mean((test_data_new$mpg - y_pred)^2); mse3

## [1] 0.002406127

mse_combined <- c(mse1, mse2, mse3)
which.min(mse_combined)
```

```
## [1] 1
```