Linear Statistical Models Assignment 2

Kim Seang CHY

Question 1: Maximum likelihood of σ^2

Recall that in the general full model $\boldsymbol{\varepsilon} \sim MVN(0, \sigma^2 I)$, hence the likelihood function is given by:

$$\begin{aligned} L(\beta, \sigma) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-\varepsilon_i^2}{2\sigma^2}\right) \\ &= (2\pi)^{-\frac{n}{2}} \sigma^{-n} \exp\left(\sum_{i=1}^{n} \frac{-\varepsilon_i^2}{2\sigma^2}\right) \end{aligned}$$

Let $\ell(\beta, \sigma)$ be the log likelihood function of $L(\beta, \sigma)$.

$$\ell(\beta, \sigma) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} \varepsilon_i^2\right)$$

Since $\sum_{i=1}^{n} \varepsilon_i^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = SS_{Res}$:

$$\begin{aligned} \ell(\beta, \sigma) &= -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{SS_{Res}}{2\sigma^2} \\ \frac{\partial \ell}{\partial \sigma} &= -\frac{1}{\sigma} + \frac{SS_{Res}}{\sigma^3} \end{aligned}$$

Setting $\frac{\partial \ell}{\partial \sigma} = 0$ and solve for σ we get:

$$\begin{aligned} -\frac{1}{\sigma} + \frac{SS_{Res}}{\sigma^3} &= 0 \\ \implies \sigma &= \frac{SS_{Res}}{n} \end{aligned}$$

Question 3: Show that the SS_{Res} for the first model is at least the SS_{Res} for the second model.

For the second model our sum of residual is given by $SS_{Res} = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}\mathbf{y} = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta}$. Since $\mathbf{H}\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$, then $SS_{Res} = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta}$

Let $SS_{Res_{\gamma_1}} = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_1\mathbf{y}$ be the sum of residual of the first model.
We can partition $\mathbf{X}$ and $\boldsymbol{\beta}$ as follow:

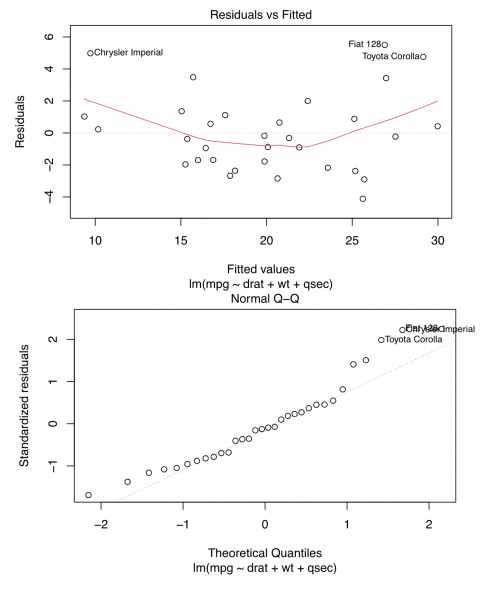
$$\mathbf{X} = \left[\begin{array}{c|c} \mathbf{X}_1 & \mathbf{X}_2 \end{array}\right] \text{ and } \boldsymbol{\beta} = \left[\begin{array}{c} \boldsymbol{\gamma}_1 \\ \hline \boldsymbol{\gamma}_2 \end{array}\right]$$

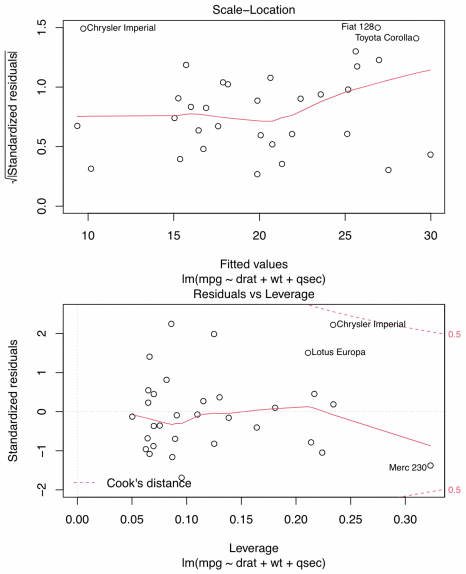
Hence, $\mathbf{X}\boldsymbol{\beta} = \mathbf{X}_1\boldsymbol{\gamma}_1 + \mathbf{X}_1\boldsymbol{\gamma}_2 = \mathbf{H}_1\mathbf{y} + \mathbf{H}_2\mathbf{y}$, where $\mathbf{H}_1$ is the hat matrix the first model and $\mathbf{H}_2$ is hat matrix for the rest of the predictor that was in the second model but not in first model. Thus SS_{Res} of the full model is given by:

$$\begin{aligned} SS_{Res} &= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T(\mathbf{H}_1 + \mathbf{H}_2)\mathbf{y} \\ &= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{H}_1\mathbf{y} - \mathbf{y}^T\mathbf{H}_2\mathbf{y} \\ &= SS_{Res_{\gamma_1}} - \mathbf{y}^T\mathbf{H}_2\mathbf{y} \end{aligned}$$

Since $\mathbf{H}_2$ is an idempotent and symmetric matrix, this implied its eigenvalue are either 0 or 1, hence it is positive semi-define, implies $\mathbf{y}^T\mathbf{H}_2\mathbf{y} \geq 0$. Hence $SS_{Res_{\gamma_1}} \geq SS_{Res}$.

e. Produce diagnostic plots for your final model from stepwise selection and comment.





The residuals vs fitted graph trend line is around 0. However, the two end which trend downward(upward) at the start points(end points). The start points does not seem to be a problem as indicated by the scale-location graph. However for the end point there seem to be a increasing variance which indicate we may not have a constant variances.

From the QQplots, also reflect this as while the starting point not be perfectly normal they are close. On the other hand, the end points is not normally distributed .
The residuals vs leverage graph, nothing really stand out, as point that may considered to be troublesome like Chrysler Imperial and Merc 230 are still with 0.5 crooks distance.

Question 5: Ridge Regression

a. Show the estimator:

Since $\sum_{i=1}^{n} e_i^2 = (y - X\beta)^T(y - X\beta)$ and $\sum_{i=1}^{n} b_i^2 = \beta^T\beta$. The parameter above is give by:

$$\begin{aligned} L(\beta) &= (y - X\beta)^T(y - X\beta) + \lambda(\beta^T\beta) \\ &= y^Ty - 2y^TX\beta + \beta^T(X^TX)\beta + \lambda(\beta^T\beta) \end{aligned}$$

To find the maximum likelihood estimator we find $\frac{\partial L}{\partial \beta}$ setting it to 0 then solve for β.

$$\begin{aligned} &\frac{\partial L}{\partial \beta} = -2(X^Ty) + 2(X^TX)\beta + 2\lambda\beta = 0 \\ &(X^TX + \lambda I)\beta = X^Ty \\ &\beta = (X^TX + \lambda I)^{-1}X^Ty \end{aligned}$$

Hence the estimator is give by $b = (X^TX + \lambda I)^{-1}X^Ty$.

b. Show that b is unbiased if $\lambda \neq 0$.

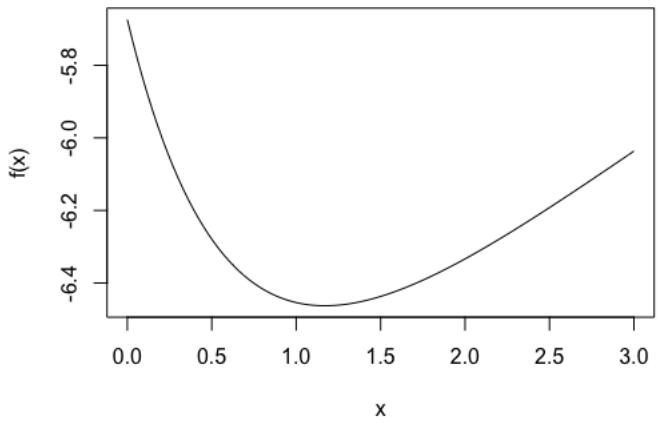
$$\begin{aligned} E(b) &= E(X^TX + \lambda I)^{-1}X^Ty) \\ &= (X^TX)^{-1}X^TE(y) + (\lambda I)^{-1}X^TE(y) \\ &= (X^TX)^{-1}X^TXb + (\lambda I)^{-1}X^TXb \qquad \text{Since E(y)=Xb} \\ &= b + (\lambda I)^{-1}X^TXb \end{aligned}$$

Since $(\lambda I)^{-1}X^TXb = 0$ if and only if $\lambda = 0$ then $E(b) \neq b$ if $\lambda \neq 0$ hence it is biased when $\lambda \neq 0$.

c. Optimal λ value.

```
f1 <-function(lambda){
  #Calling Data Frame
  q2data <- data.matrix(read.csv(file = "Q2_Data.csv"))
  q2frame <- read.csv(file="Q2_Data.csv")
  #Converting Data frame to matrix
  y <- matrix(q2data[,1],7,1)
  x <- matrix(c(rep(1,7),q2data[,-1]),7,4)
  #Scaling data
  x <- scale(x[,-1],center=T,scale=T)
  y <- scale(y,center=T,scale=T)
  p <- 3
  #Definting Degree of freedom
  lambda = matrix(c(lambda,0,0,0,lambda,0,0,0,lambda),3,3)
  H <- x%*%solve(t(x)%*%x+lambda)%*%t(x)
  df <- sum(diag(H))
  #Computing Beta
  b <- solve((t(x)%*%x)+lambda)%*%t(x)%*%y
  #Computing Sum-squared
  e <- (y-x%*%b)
  SSres <- sum(e^2)
  n <- dim(y)[1]
  gof <- n*log(SSres/n)+2*df
  return(gof)
}
```

```
> f <- Vectorize(f1); curve(f, 0 ,3)
> f(1.17)<f(1.171)
[1] TRUE
> f(1.17)<f(1.169)
[1] TRUE
```



Hence, the optimal value for λ is approximately 1.17.