Stat 331 Applied Linear Models – Assignment 3
Solution

1(a) 2 points

$$\begin{aligned} E(\hat{\beta}) &= (X^TX)^{-1}X^TE(Y) \\ &= (X^TX)^{-1}X^T(X\beta + Z\gamma) \\ &= \beta + (X^TX)^{-1}X^TZ\gamma. \end{aligned}$$

$$\begin{aligned} E(r) &= E(Y - \hat{Y}) \\ &= (X\beta + Z\gamma) - \{X\beta + X(X^TX)^{-1}X^TZ\gamma\} \\ &= Z\gamma - HZ\gamma. \end{aligned}$$

1(b) 2 points It follows from

$$\begin{aligned} Var(r) &= E\{r - E(r)\}\{r - E(r)\}^T \\ &= E\{rr^T\} - E\{rE(r)^T\} - E\{E(r)r^T\} + E(r)E(r)^T \\ &= E\{rr^T\} - E(r)E(r)^T. \end{aligned}$$

1(c) 4 points

$$\begin{aligned} E(r^Tr) &= E\{tr(rr^T)\} = tr\{E(rr^T)\} \\ &= tr\{Var(r) + E(r)E(r)^T\} \\ &= tr\{Var(r)\} + tr\{(I-H)Z\gamma\gamma^TZ^T(I-H)^T\}. \end{aligned}$$

We have shown that $tr\{Var(r)\} = (n - p_1 - 1)\sigma^2$ in class. So

$$\begin{aligned} E(r^Tr) &= (n - p_1 - 1)\sigma^2 + tr\{(I-H)^T(I-H)Z\gamma\gamma^TZ^T\} \\ &= (n - p_1 - 1)\sigma^2 + tr\{(I-H)Z\gamma\gamma^TZ^T\} \\ &= (n - p_1 - 1)\sigma^2 + tr\{Z^T(I-H)Z\gamma\gamma^T\} \\ &= (n - p_1 - 1)\sigma^2 + tr\{\gamma^TZ^T(I-H)Z\gamma\} \\ &= (n - p_1 - 1)\sigma^2 + \gamma^TZ^T(I-H)Z\gamma. \end{aligned}$$

2(a) 4 points

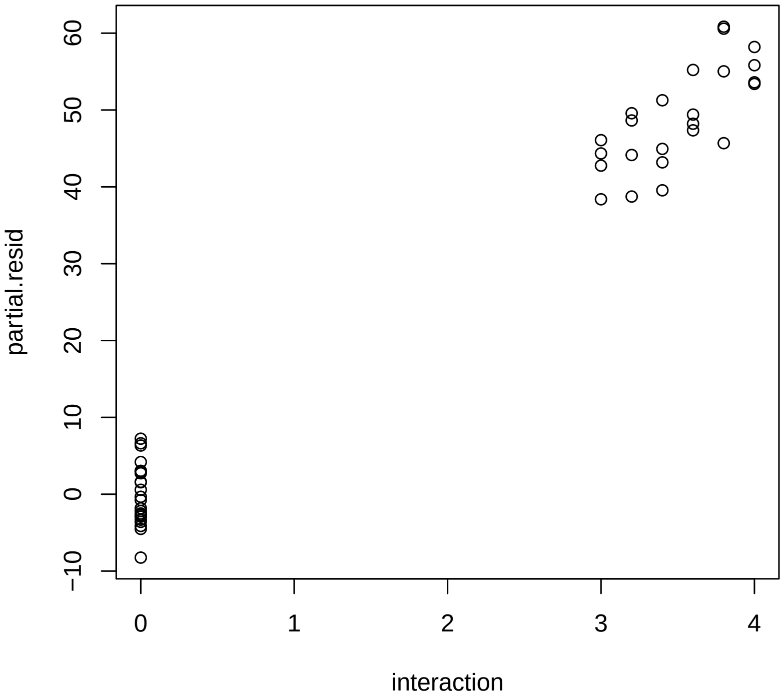
Null hypothesis: H_0 : there is no lack of fit in the model.

Alternative hypothesis: H_1 : there is lack of fit in the model.

The calculated F statistic is 2.59, while $F_{(0.05,9,36)} = 2.15$ and $p - value = 0.02$. Hence we reject the null hypothesis at 5% level and conclude that there is lack of fit in the model with only GPA and region as the predictors.

2(b) 6 points

From the partial residual plot, we can see a strong linear relationship. Hence, we need to consider including the interaction term in the model.



The calculated F statistic is 1.03, while $F_{(0.05,8,36)} = 2.21$ and $p - value = 0.43$. Hence we fail to reject the null hypothesis at 5% level and conclude that there is no lack of fit in the model with GPA, region and their interaction as the predictors.

2(c) 2 points

It is the estimated difference in the effect of GPA on income, between job in US and jobs in Europe.

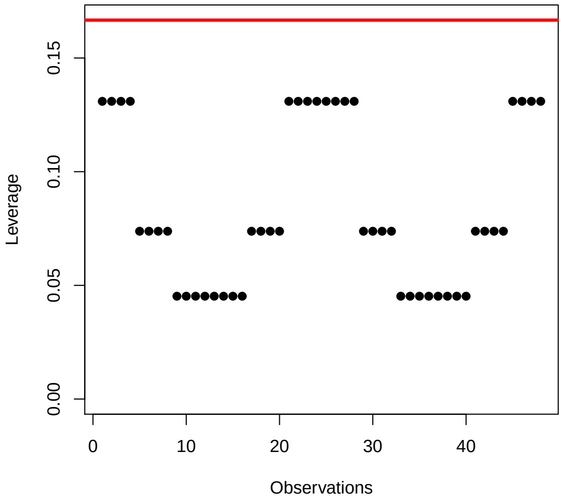
Or it is the estimated difference in the slope of GPA between the two models when $x_2 = 1$ (jobs in US) versus $x_2 = 0$ (jobs in Europe) .

2(d) 2 points

Although the p-value is larger than 0.05, we cannot drop GPA, as the interaction term is significant (Rule of Hierarchy).

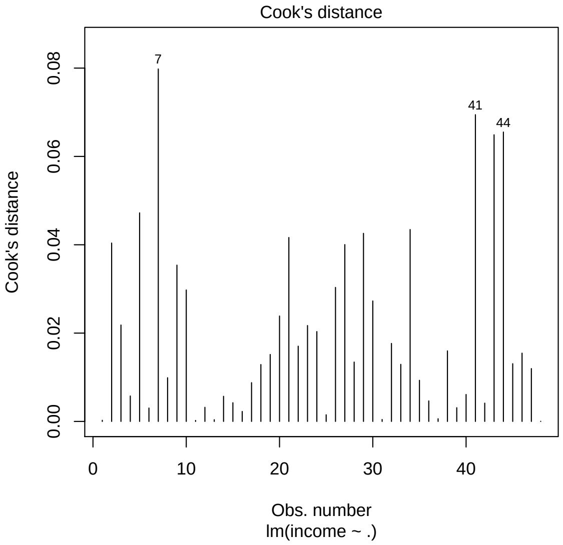
2(e) 2 points

Based on the plot, no leverage exceeds the threshold, and no observation is considered as a point of high leverage.



2(f) 2 points

The threshold value is 0.85, and no observation has a Cook's Distance larger than that. Hence, no observation is considered as an influential point.



```
math=read.delim("math.txt")

#### (a)

attach(math)

GPA.temp = GPA[1]
region.temp = region[1]
group=1

for (i in 2:length(income))
{
if (GPA.temp!= GPA[i] | region.temp != region[i])
{
GPA.temp = GPA[i]
region.temp = region[i]
group=c(group,group[length(group)]+1)
} else
{
group=c(group,group[length(group)])
}

}

k=max(group)

group.mean=numeric(0)

for (j in 1:k)
{
group.mean=c(group.mean,mean(income[group==j]))
}

SSE.beta = 0

for (j in 1:k)
{
SSE.beta=SSE.beta + sum((income[group==j]-group.mean[j])^2)
}


fm1=lm(income~GPA+region,data=math)

SSE.betaA = anova(fm1)[3,2]

F = ( (SSE.betaA-SSE.beta)/(k-2-1) ) / (SSE.beta / (nrow(math) - k))

F
```

```
qf(0.95,df1=(k-2-1),df2=(nrow(math) - k))

#### (b)

fm2=lm(income~.,data=math)

partial.resid = resid(fm2) + interaction*coef(fm2)[4]

plot(interaction,partial.resid)


SSE.betaA = anova(fm2)[4,2]

F = ( (SSE.betaA-SSE.beta)/(k-3-1) ) / (SSE.beta / (nrow(math) - k))

F

qf(0.95,df1=(k-3-1),df2=(nrow(math) - k))

#### (d)

summary(fm2)

#### (e)

threshold = 2*(3+1)/nrow(math)

plot(1:nrow(math),hatvalues(fm2),pch=19,ylim=c(0,threshold),
xlab="Observations",ylab="Leverage")

abline(threshold,0,lwd=3,col=2)

#### (f)

threshold = qf(0.5,4,nrow(math)-4)

threshold

plot(fm2, which=4)

detach(math)
```

3(a) 3 points

Based on the R output, the final model is chosen to be

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_5 x_{i,5} + \epsilon_i.$$

In other words, we select Stay, Cultures, and Beds as the predictors in the final model.

The reason is the corresponding adjusted R^2 is the largest, the corresponding Mallow's Cp is small, and the corresponding BIC is the smallest.

3(b) 3 points

In the first step, the p-value for Cultures is the smallest and smaller than 0.05. Hence we select x_2 in the model.

In the second step, the p-value for Stay is the smallest and smaller than 0.05. Hence we select x_1 in the model, and now the model includes x_1 and x_2.

In the third step, the p-value for Beds is the smallest and smaller than 0.05. Hence we select x_5 in the model, and now the model includes x_1, x_2 and x_5.

In the fourth step, the p-values for both Age and Census are larger than 0.05. Hence we stop and claim that the model in the third step is the best one.

The final model is

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_5 x_{i,5} + \epsilon_i.$$

In other words, we select Stay, Cultures, and Beds as the predictors in the final model.

3(c) 3 points

Based on the R output, the final model from backward elimination using AIC is

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_5 x_{i,5} + \epsilon_i.$$

In other words, we select Stay, Cultures, and Beds as the predictors in the final model.

```
Senic=read.delim("Senic.txt")

##### (a)

library(leaps)

best=regsubsets(InfctRsk~Stay+Cultures+Age+Census+Beds,data=Senic,
        nbest=2)

summary(best)

summary(best)$adjr2

summary(best)$cp

summary(best)$bic

##### (b)

## First step

g=lm(InfctRsk~Stay,data=Senic)

anova(g)

g=lm(InfctRsk~Cultures,data=Senic)

anova(g)

g=lm(InfctRsk~Age,data=Senic)

anova(g)

g=lm(InfctRsk~Census,data=Senic)

anova(g)

g=lm(InfctRsk~Beds,data=Senic)

anova(g)

## Second step

g=lm(InfctRsk~Cultures+Stay,data=Senic)

anova(g)

g=lm(InfctRsk~Cultures+Age,data=Senic)
```

```
anova(g)

g=lm(InfctRsk~Cultures+Census,data=Senic)

anova(g)

g=lm(InfctRsk~Cultures+Beds,data=Senic)

anova(g)

## Third step

g=lm(InfctRsk~Cultures+Stay+Age,data=Senic)

anova(g)

g=lm(InfctRsk~Cultures+Stay+Census,data=Senic)

anova(g)

g=lm(InfctRsk~Cultures+Stay+Beds,data=Senic)

anova(g)

## Fourth step

g=lm(InfctRsk~Cultures+Stay+Beds+Age,data=Senic)

anova(g)

g=lm(InfctRsk~Cultures+Stay+Beds+Census,data=Senic)

anova(g)

##### (c)

fullmodel<-lm(InfctRsk~Stay+Cultures+Age+Census+Beds,data=Senic)
nullmodel<-lm(InfctRsk~1,data=Senic)

step(fullmodel,scope=list(lower=nullmodel),direction="backward")
```