Statistics 206

Homework 9

Not Due

Problems 1 and 2. Model validation and model diagnostic case study in R. *Diabetes data (Cont'd from homework 8). This data consist of* 19 *variables on* 403 *subjects from* 1046 *subjects who were interviewed in a study to understand the prevalence of obesity, diabetes, and other cardiovascular risk factors in central Virginia for African Americans. We will consider building regression models with* `glyhb` *as the response variable as Glycosolated Hemoglobin* > 70 *is often taken as a positive diagnostics of diabetes. The data set and description are under Files/Homework. Please attach your R codes and plots.*

1. **Model validation.** We now consider validation of the two models fs1 and fs2 selected by the forward stepwise procedure.

   (a) **Internal validation of Models fs1 and fs2**. For this purpose, we need to compute C_p and $Press_p$ for these models. For C_p, we need an unbiased estimator of the error variance σ^2. The largest model we have considered so far is Model 2. However, this model has a very large number of regression coefficients (relative to the sample size), making its parameter estimation unreliable due to large sampling variability. Therefore, we decided to use a smaller model consisting of all predictors identified by Model fs1 (the forward stepwise selected first-order model), as well all the 2-way interaction terms among these predictors. Denote this model by Model 3. Note that, Model fs2 is also a sub-model of Model 3. How many regression coefficients are there in Model 3 ? What is MSE from Model 3? Calculate SSE_p, MSE_p, C_p and $Press_p$ for Models fs1 and fs2 and briefly comment on the results, e.g., does it appear to be substantial model bias in these two models? Should overfitting be a concern?

```
> data.cc=data.c[, c("glyhb", "stab.glu", "age", "waist", "ratio")]
> fit3=lm(glyhb~.^2, data=data.cc)
> length(fit3$coefficients) #number of coefficents in Model 3
[1] 11
> mse3= anova(fit3)["Residuals",3] #MSE for Model 3
> mse3
[1] 0.001346963
> sse.fs1=anova(step.f)["Residuals",2] #first order selected
> sse.fs1
[1] 0.2401432
> sse.fs2=anova(step.f2)["Residuals",2] #second order selected
> sse.fs2
[1] 0.2339843
> mse.fs1=anova(step.f)["Residuals",3] #MSE for Model fs1
> mse.fs1
```

```
[1] 0.001349119
> mse.fs2=anova(step.f2)["Residuals",3] #MSE for Model fs2
> mse.fs2
[1] 0.001329456
> p.fs1=length(step.f$coefficients) #5
> p.fs1
[1] 5
> p.fs2=length(step.f2$coefficients) #7
> p.fs2
[1] 7
> cp.fs1=sse.fs1/mse3-(n-2*p.fs1) #C_p for Model fs1
> cp.fs1
[1] 5.284958
> cp.fs2=sse.fs2/mse3-(n-2*p.fs2) #C_p for Model fs2
> cp.fs2
[1] 4.712495
> press.fs1=sum(step.f$residuals^2/(1-influence(step.f)$hat)^2)
> press.fs1
[1] 0.2535404
> press.fs2=sum(step.f2$residuals^2/(1-influence(step.f2)$hat)^2)
> press.fs2
[1] 0.2534834
```

For both Model fs1 and Model fs2, $C_p \approx p$ and $Press_p$ and SSE_p are reasonably close, supporting their validity: little bias and not much overfitting.

(b) **External validation using the validation set**. We now fit Models fs1 and fs2 on the validation data set. Compare the fitted regression coefficients from the training data and those from the validation data. Are the two sets of estimated regression coefficients having the same sign? Are their values similar? How about the two sets of standard errors? Does it appear that Models fs1 and fs2 have consistent estimates on the training data and validation data? Calculate the mean squared prediction error (MSPE) using the validation data for each of the two models. How do these $MSPE_v$ compare with the respective $Press_p/n$ and SSE_p/n (Note here n is the sample size of the training data, i.e., 183)? Which model among the two has a smaller $MSPE_v$?

```
### Model fs1
> fit.fs1.v=lm(step.f,data=data.v) #Model fs1 on validation data
> summary(step.f) #summary on training data

Call:
lm(formula = glyhb ~ stab.glu + age + waist + ratio, data = data.c)

Residuals:
Min        1Q    Median        3Q       Max
```

```
-0.091989 -0.022720 -0.001251  0.020707  0.144356

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.490e-01  1.843e-02  18.932  < 2e-16 ***
stab.glu    -5.368e-04  5.219e-05 -10.287  < 2e-16 ***
age         -6.412e-04  1.698e-04  -3.776 0.000217 ***
waist       -1.398e-03  5.075e-04  -2.756 0.006465 **
ratio       -2.848e-03  1.997e-03  -1.427 0.155439
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.03673 on 178 degrees of freedom
Multiple R-squared: 0.5345,Adjusted R-squared: 0.524
F-statistic: 51.09 on 4 and 178 DF,  p-value: < 2.2e-16

> summary(fit.fs1.v) #summary on validation data

Call:
lm(formula = step.f, data = data.v)

Residuals:
Min        1Q    Median        3Q       Max
-0.151518 -0.018954  0.000226  0.017982  0.133835

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.3287126  0.0187828  17.501  < 2e-16 ***
stab.glu    -0.0004436  0.0000575  -7.715 8.31e-13 ***
age         -0.0006694  0.0001815  -3.687 0.000301 ***
waist       -0.0008451  0.0004945  -1.709 0.089243 .
ratio       -0.0042812  0.0014718  -2.909 0.004089 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.03633 on 178 degrees of freedom
Multiple R-squared:  0.47,Adjusted R-squared: 0.4581
F-statistic: 39.46 on 4 and 178 DF,  p-value: < 2.2e-16

#percent change in parameter estimation
> round(abs(coef(step.f)-coef(fit.fs1.v))/abs(coef(step.f))*100,3)
(Intercept)    stab.glu         age       waist       ratio
5.813      17.359       4.402      39.573      50.310
> sd.fs1= summary(step.f)$coefficients[,"Std. Error"]
```

```
> sd.fs1.v= summary(fit.fs1.v)$coefficients[,"Std. Error"]
#percent change in standard errors
> round(abs(sd.fs1-sd.fs1.v)/sd.fs1*100,3)
(Intercept)    stab.glu         age       waist       ratio
1.889      10.190       6.925       2.545      26.283
```

ANS. Consistency for Model fs1: reasonable. Signs for parameter estimates are all the same, but percent change can be as big as 50%.

```
##mean squared prediction error
> newdata=data.v[,-5]
> pred.fs1=predict.lm(step.f, newdata)
> mspe.fs1=mean((pred.fs1-data.v[,5])^2)
> mspe.fs1
[1] 0.001329283
> press.fs1/n
[1] 0.001385467
> mse.fs1
[1] 0.001349119
```

```
### Model fs2
> fit.fs2.v=lm(step.f2,data=data.v) #Model fs1 on validation data
> summary(step.f2) #summary on training data

Call:
lm(formula = glyhb ~ stab.glu + age + waist + ratio + stab.glu:ratio +
age:ratio, data = data.c)

Residuals:
Min        1Q    Median        3Q       Max
-0.089202 -0.022258 -0.003599  0.021182  0.145324

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.527e-01  3.162e-02  11.152  < 2e-16 ***
stab.glu       -9.522e-04  2.186e-04  -4.355 2.25e-05 ***
age             7.247e-05  5.277e-04   0.137   0.8909
waist          -1.305e-03  5.079e-04  -2.570   0.0110 *
ratio          -2.158e-03  6.565e-03  -0.329   0.7427
stab.glu:ratio  7.507e-05  3.775e-05   1.988   0.0483 *
age:ratio      -1.724e-04  1.231e-04  -1.401   0.1631
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.03646 on 176 degrees of freedom
```

```
Multiple R-squared: 0.5464,Adjusted R-squared: 0.531
F-statistic: 35.34 on 6 and 176 DF,  p-value: < 2.2e-16

> summary(fit.fs2.v) #summary on validation data

Call:
lm(formula = step.f2, data = data.v)

Residuals:
Min        1Q    Median        3Q       Max
-0.148870 -0.017817 -0.000883  0.018924  0.125159

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.122e-01  3.031e-02  10.302   <2e-16 ***
stab.glu       -2.435e-04  1.413e-04  -1.724   0.0865 .
age            -8.409e-04  5.465e-04  -1.539   0.1257
waist          -9.390e-04  4.991e-04  -1.881   0.0616 .
ratio           7.797e-05  6.220e-03   0.013   0.9900
stab.glu:ratio -3.984e-05  2.581e-05  -1.544   0.1245
age:ratio       3.366e-05  1.201e-04   0.280   0.7796
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.03629 on 176 degrees of freedom
Multiple R-squared: 0.4772,Adjusted R-squared: 0.4594
F-statistic: 26.78 on 6 and 176 DF,  p-value: < 2.2e-16

> #percent change in parameter estimation
> round(abs(coef(step.f2)-coef(fit.fs2.v))/abs(coef(step.f2))*100,3)
(Intercept)       stab.glu            age          waist          ratio
11.465         74.424       1260.423         28.060        103.612
stab.glu:ratio      age:ratio
153.071        119.518
> sd.fs2= summary(step.f2)$coefficients[,"Std. Error"]
> sd.fs2.v= summary(fit.fs2.v)$coefficients[,"Std. Error"]
> #percent change in standard errors
> round(abs(sd.fs2-sd.fs2.v)/sd.fs2*100,3)
(Intercept)       stab.glu            age          waist          ratio
4.157         35.370          3.571          1.731          5.248
stab.glu:ratio      age:ratio
31.637          2.436
```

ANS. Consistency for Model fs2: Both sign and magnitude changed.

```
#mean squared prediction error
> newdata=data.v[,-5]
> pred.fs2=predict.lm(step.f2, newdata)
> mspe.fs2=mean((pred.fs2-data.v[,5])^2)
> mspe.fs2 #larger than mspe.fs1
[1] 0.00152642
> press.fs2/n
[1] 0.001385155
> mse.fs2
[1] 0.001329456
```

ANS. For both models, $MSPE_v$ is not much bigger than $Press_p/n$ and SSE_p/n, though $MSPE_v$ is closer to $Press_p/n$ and SSE_p/n in Model fs1. Moreover, Model fs1 has smaller $MSPE_v$.

(c) Based on both internal and external validation, which model you would choose as the final model? Fit the final model using the entire data set (training and validation combined). Write down the fitted regression function and report the R summary() and anova() output.

ANS. Model fs1 is preferred based on smaller $MSPE_v$ and more consistent parameter estimation in training and validation data sets.

```
> fit.fs1.final=lm(step.f, data=data.s) #fit Model fs1 on whole data
> summary(fit.fs1.final)

Call:
lm(formula = step.f, data = data.s)

Residuals:
Min        1Q    Median        3Q       Max
-0.152555 -0.020528 -0.000382  0.019560  0.148412

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.380e-01  1.306e-02  25.881  < 2e-16 ***
stab.glu    -4.922e-04  3.838e-05 -12.825  < 2e-16 ***
age         -6.561e-04  1.229e-04  -5.338 1.67e-07 ***
waist       -1.080e-03  3.516e-04  -3.071  0.00229 **
ratio       -3.661e-03  1.181e-03  -3.100  0.00209 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.03643 on 361 degrees of freedom
Multiple R-squared: 0.5005,Adjusted R-squared: 0.495
F-statistic: 90.45 on 4 and 361 DF,  p-value: < 2.2e-16
```

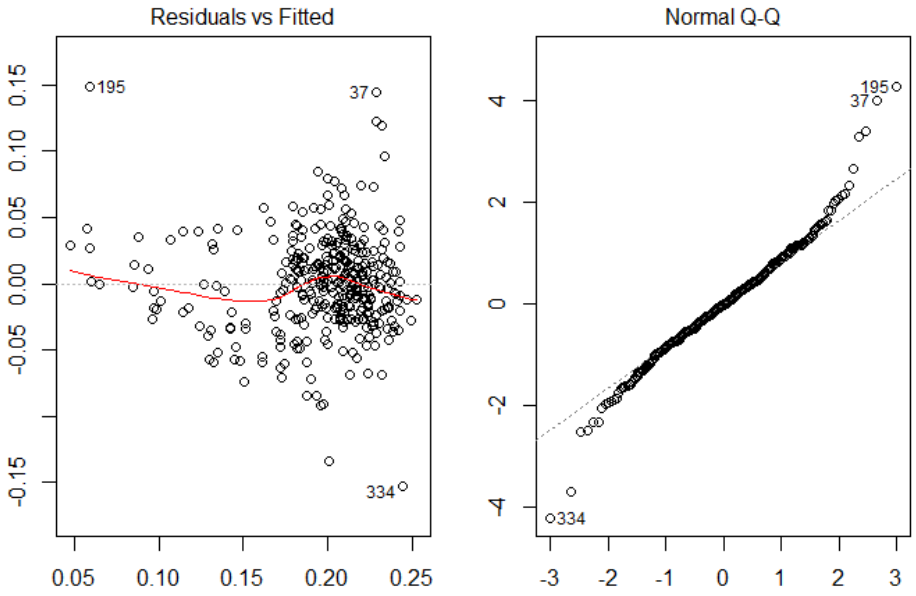
```
> anova(fit.fs1.final)
Analysis of Variance Table

Response: glyhb
Df  Sum Sq Mean Sq  F value    Pr(>F)
stab.glu    1 0.39753 0.39753 299.5043 < 2.2e-16 ***
age         1 0.04867 0.04867  36.6682 3.515e-09 ***
waist       1 0.02125 0.02125  16.0081 7.655e-05 ***
ratio       1 0.01276 0.01276   9.6103  0.002087 **
Residuals 361 0.47915 0.00133
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

2. **Model diagnostic: Outlying and influential cases.** Conduct model diagnostic for the final model from the previous problem.

   (a) Draw residual vs. fitted value plot and residual Q-Q plot and comment on these plots.

Figure 1: Model Diagnostics for final model



   ANS. The residual plot shows non-constancy in error variance. The Normal QQ plot shows heavy tails probably due to outliers.

   (b) Obtain the studentized deleted residuals and identify any outlying Y observations. Use the Bonferroni outlier test procedure at $\alpha = 0.1$.

The studentized deleted residuals are calculated through this equation:

$$t_i = e_i \sqrt{\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2}}$$

To identify the outlying Y observations, we use the Bonferroni outlier test procedure at $\alpha = 0.1$. The Bonferronis threshold is

$$t(1 - \frac{\alpha}{2n}; n-p-1) = 3.676928$$

The Y observations corresponding to those studentized deleted residuals which are greater than the Bonferroni's threshold can be deemed as significant outlying observations. They are as follows:

```
> idx.Y ##  outliers
[1]  34 176 303 330
```

The code is as follows:

```
## check outliers in Y
res=residuals(fit.fs1.full)# residuals of the final model
n = nrow(data.s)
p = ncol(data.s)
h1 = influence(fit.fs1.full)$hat
d.res.std=studres(fit.fs1.full) #studentized deleted residuals

max(abs(d.res.std))
sort(abs(d.res.std),decreasing=T)
qt(1-0.1/(2*n),n-p-1) # bonferronis thresh hold
idx.Y = as.vector(which(abs(d.res.std)>=qt(1-0.1/(2*n),n-p-1)))
idx.Y ## outliers
```

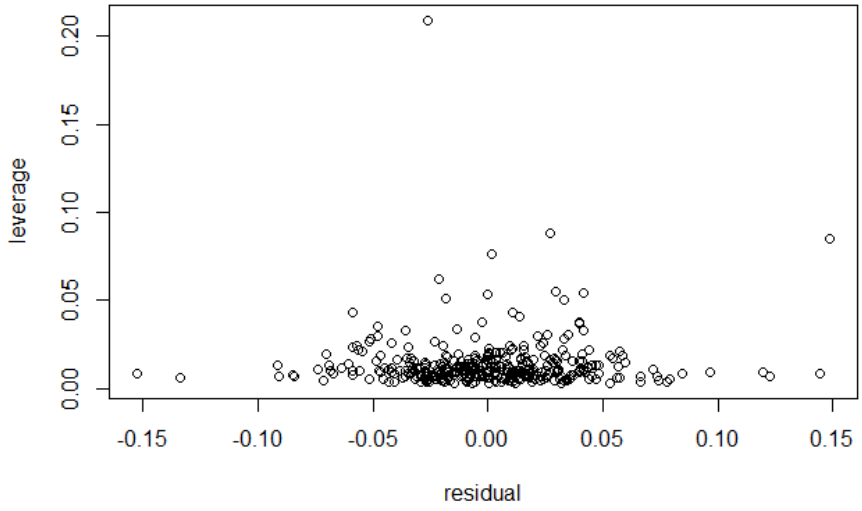
(c) Obtain the leverage and identify any outlying X observations. Draw residual vs. leverage plot.

```
idx.X = as.vector(which(h1>(2*p/n)))
idx.X ## two outliers
plot(h1,res,xlab="leverage",ylab="residuals")
```

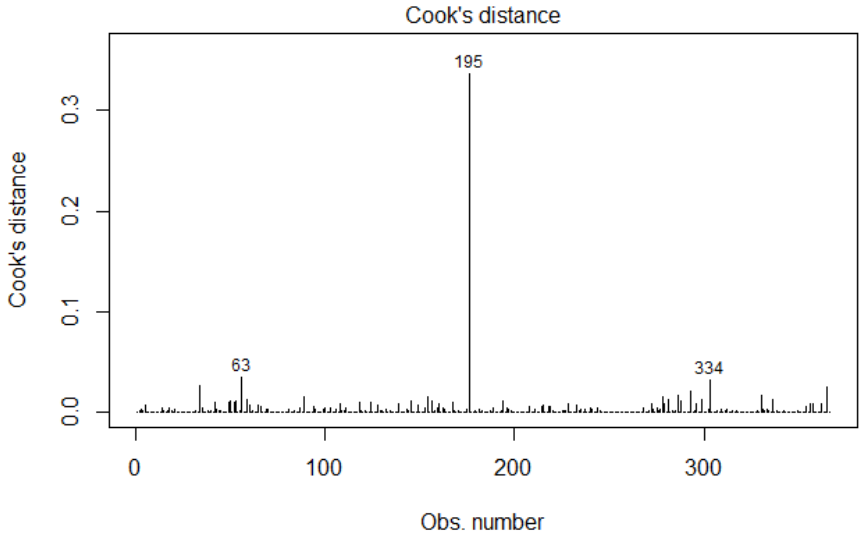
```
> idx.X ## outliers
[1]  56 156
```

The leverages are obtained and compared with the value of $\frac{2p}{n} = 0.08743169$. The cases with $h_{ii} > \frac{2p}{n}$ are defined as outlying X observations. There are 2 cases defined as outlying X observations, their indexes are shown above.

Figure 2: Residuals vs. Leverage Plot



(d) Draw an influence index plot using Cook's distance. Are there any influential cases according to this measure?




```
cook.d = res^2*h1/(p*1.293*(1-h1)^2)
cook.max = cook.d[which(cook.d==max(cook.d))]
```

```
pf(cook.max,p,n-p)

idx = c(idx.X,idx.Y)
cook.d[idx]
pf(cook.d[idx],p,n-p)
```

According to the Cook's distance plot, case 195 has the biggest Cook's distance.

$$D_{195} = \frac{e_i^2}{p * MSE} \frac{h_{ii}}{(1-h_{ii})^2} = 0.0001079778$$

$$p_{195} = P(F_{16,350} < 0.0001079778) = 8.99587e-30$$

Therefore, even case 195 has little aggregated influence on all the fitted values. Hence there is no influential cases according to this measure.

(e) Calculate the average absolute percent difference in the fitted values with and without the most influential case identified from the previous question. What does this measure indicate the influence of this case?

The potential influential case identified previously is the 195th case,
we fit the model without 195th case and calculate the average absolute percent difference in the fitted values as 0.03163563. For 195th case, the percentage change on the fitted value with or without the case is very small. Therefore, no case have an unduly large influence on prediction and thus all cases may be retained.

```
fit.fs1.full2=lm(fit.fs1, data=data.s[-195,])
f1=fitted(fit.fs1.full)
f2=fitted(fit.fs1.full2)
f1=f1[-195]
f=f1-f2
sum=0
for(i in 1:length(f1))
{
sum=sum+abs(f[i]/f1[i]);
}
yhat_195=fitted(fit.fs1.full)[195]
beta_new=as.vector(fit.fs1.full2$coefficients)
x_195=c(1,data.s$stab.glu[195],data.s$age[195],
data.s$waist[195],data.s$ratio[195])
y_195=t(beta_new)%*%x_195
sum=sum+abs((yhat_195-y_195)/yhat_195)
per.average=sum*100/366
```

3. **(Optional Problem). Studentized deleted residuals.** In the following, no assumption is made on the data or the model unless it is explicitly stated.

(a) Assume the observed response vector $\mathbf{Y} \in \mathbb{R}^n$ has $Var(\mathbf{Y}) = \sigma^2 \mathbf{I}_n$. Show that, the ith deleted residual $d_i = Y_i - \widehat{Y}_{i(i)}$ has

$$Var(d_i) = \frac{\sigma^2}{1 - h_{ii}}.$$

ANS For the ith deleted residual:

$$d_i = Y_i - \hat{Y}_{i,(i)} = \frac{e_i}{1 - h_{ii}}$$

Therefore:

$$Var(d_i) = Var\left(\frac{e_i}{1 - h_{ii}}\right) = \frac{1}{(1 - h_{ii})^2} Var(e_i) = \frac{1}{(1 - h_{ii})^2} \times \sigma^2 \times (1 - h_{ii}) = \frac{\sigma^2}{(1 - h_{ii})}$$

(b) Let

$$SSE_{(i)} = \sum_{j:j \neq i} (Y_j - \widehat{Y}_{j(i)})^2, \quad MSE_{(i)} = \frac{SSE_{(i)}}{n - p - 1},$$

i.e., $SSE_{(i)}$ and $MSE_{(i)}$ are the SSE and MSE of the regression fit excluding case i, respectively. Show that

$$SSE_{(i)} = SSE - \frac{e_i^2}{1 - h_{ii}}.$$

Hints: Recall that

$$SSE_{(i)} = \tilde{\mathbf{Y}}^T (\mathbf{I} - \mathbf{H}) \tilde{\mathbf{Y}},$$

where

$$\tilde{\mathbf{Y}} = \mathbf{Y} - \mathbf{d}_{(i)}, \quad where, \quad \mathbf{d}_{(i)} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ d_i \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

i.e., $\tilde{\mathbf{Y}}$ *is the same as* $\mathbf{Y}$ *except for the* i*th element, where it is* $\widehat{Y}_{i(i)}$.

ANS Denote

$$\mathbf{D}' = \begin{bmatrix} 0 & ... & 0 & d_i & 0 & ... & 0 \end{bmatrix}$$

Where $d_i = \frac{e_i}{1-h_{ii}}$. Therefore:

$$\begin{aligned}
SSE_{(i)} &= \tilde{\mathbf{Y}}'(\mathbf{I_n} - \mathbf{H})\tilde{\mathbf{Y}} = (\mathbf{Y} - \mathbf{D})'(\mathbf{I_n} - \mathbf{H})(\mathbf{Y} - \mathbf{D})' \\
&= \mathbf{Y}'(\mathbf{I_n} - \mathbf{H})\mathbf{Y} - \mathbf{D}'(\mathbf{I_n} - \mathbf{H})\mathbf{Y} - \mathbf{Y}'(\mathbf{I_n} - \mathbf{H})\mathbf{D} + \mathbf{D}'(\mathbf{I_n} - \mathbf{H})\mathbf{D} \\
&= SSE - \mathbf{D}'\mathbf{e} - \mathbf{e}'\mathbf{D} + \mathbf{D}'(\mathbf{I_n} - \mathbf{H})\mathbf{D} \text{ since } (\mathbf{I_n} - \mathbf{H})\mathbf{Y} = \mathbf{e} \\
&= SSE - \frac{e_i^2}{1-h_{ii}} - \frac{e_i^2}{1-h_{ii}} + (1-h_{ii})d_i^2 \\
&= SSE - \frac{e_i^2}{1-h_{ii}} - \frac{e_i^2}{1-h_{ii}} + (1-h_{ii})\frac{e_i^2}{(1-h_{ii})^2} \\
&= SSE - \frac{e_i^2}{1-h_{ii}} - \frac{e_i^2}{1-h_{ii}} + \frac{e_i^2}{1-h_{ii}} = SSE - \frac{e_i^2}{1-h_{ii}}
\end{aligned}$$

(c) Show that the studentized deleted residual

$$t_i = \frac{d_i}{s\{d_i\}} = \frac{d_i}{\sqrt{MSE_{(i)}/(1-h_{ii})}}$$

can be computed by:

$$t_i = e_i\sqrt{\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2}}.$$

ANS

$$\begin{aligned}
t_i &= \frac{d_i}{s\{d_i\}} = \sqrt{1-h_{ii}}\frac{d_i}{\sqrt{MSE_{(i)}}} = \sqrt{1-h_{ii}}\frac{1}{\sqrt{MSE_{(i)}}} \times \frac{e_i}{1-h_{ii}} \\
&= \frac{e_i}{\sqrt{MSE_{(i)} \times (1-h_{ii})}} = e_i\sqrt{\frac{n-p-1}{SSE_{(i)} \times (1-h_{ii})}} \\
&= e_i\sqrt{\frac{n-p-1}{(SSE - \frac{e_i^2}{1-h_{ii}}) \times (1-h_{ii})}} \\
&= e_i\sqrt{\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2}}
\end{aligned}$$

(d) Under the Normality assumption, i.e., $\mathbf{Y}$ is an n-dimensional Normal random vector with $Var(\mathbf{Y}) = \sigma^2\mathbf{I}_n$, show that $SSE_{(i)}$ is independent with Y_i and $\widehat{Y}_{i(i)}$. Therefore, $SSE_{(i)}$ is independent with d_i. If we further assume that the model is correct, then the deleted residual d_i has mean zero and the studentized deleted residual t_i follows a $t_{(n-p-1)}$ distribution.

ANS For any given regression model, its SSE is always independent with a fitted value based on this model, so $SSE_{(i)}$ is independent with $\widehat{Y}_{i(i)}$. Also $SSE_{(i)}$ only involve Y_{-i} which are all independent with Y_i, therefore $SSE_{(i)}$ is also independent with Y_i.

4. **(Optional Problem). Cook's distance.** Show that the Cook's distance

$$D_i := \frac{\sum_{j=1}^{n}(\widehat{Y}_j - \widehat{Y}_{j(i)})^2}{p \times MSE}, \quad i = 1, \cdots, n$$

can be computed by:

$$D_i = \frac{e_i^2}{p \times MSE} \frac{h_{ii}}{(1-h_{ii})^2}.$$

Hints: Note that

$$\sum_{j=1}^{n}(\widehat{Y}_j - \widehat{Y}_{j(i)})^2 = (\mathbf{Y} - \tilde{\mathbf{Y}})^T \mathbf{H} (\mathbf{Y} - \tilde{\mathbf{Y}}).$$

ANS

$$\sum_{j=1}^{n}\left(\hat{Y}_j - \hat{Y}_{j,(i)}\right)^2 = (\mathbf{Y} - \tilde{\mathbf{Y}})' \mathbf{H} (\mathbf{Y} - \tilde{\mathbf{Y}}) = \mathbf{D}'\mathbf{H}\mathbf{D} = d_i^2 h_{ii}$$

Therefore:

$$Di = \frac{\sum_{j=1}^{n}\left(\hat{Y}_j - \hat{Y}_{j,(i)}\right)^2}{p \times MSE} = \frac{d_i^2 h_{ii}}{p \times MSE} = \frac{e_i^2}{p \times MSE} \frac{h_{ii}}{(1-h_{ii})^2}$$

5. **Regression formulations of one-way ANOVA model**.

   (a) Consider the *cell means formulation* discussed in class:

$$Y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, \cdots, I, j = 1, \cdots, n_i.$$

   Express this model as a linear regression model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

   Specify $\mathbf{Y}$, $\mathbf{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$. What is $\mathbf{X}^T\mathbf{X}$ and what is $\mathbf{X}^T\mathbf{Y}$? What is the LS estimator of $\boldsymbol{\beta}$? Derive the fitted values and residuals.

ANS

$$\underset{n_T \times 1}{\mathbf{Y}} = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{I1} \\ \vdots \\ Y_{In_I} \end{pmatrix}; \quad \underset{n_T \times I}{\mathbf{X}} = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \\ 0 & \cdots & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix}; \quad \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_I \end{pmatrix}; \quad \underset{n_T \times 1}{\boldsymbol{\epsilon}} = \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2n_2} \\ \vdots \\ \epsilon_{I1} \\ \vdots \\ \epsilon_{In_I} \end{pmatrix}$$

where for $\mathbf{X}$, column X_i has n_i 1's and the other entries are 0.

$$\mathbf{X}^T\mathbf{X} = \text{diag}(n_1, n_2, ..., n_I)$$

$$\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} n_1\bar{Y}_{1.} \\ n_2\bar{Y}_{2.} \\ \vdots \\ n_I\bar{Y}_{I.} \end{pmatrix}$$

The LS estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \bar{Y}_{1.} \\ \bar{Y}_{2.} \\ \vdots \\ \bar{Y}_{I.} \end{pmatrix}$$

The fitted values

$$\underset{n_T\times 1}{\hat{\mathbf{Y}}} = \begin{pmatrix} \bar{Y}_{1.} \\ \vdots \\ \bar{Y}_{1.} \\ \bar{Y}_{2.} \\ \vdots \\ \bar{Y}_{2.} \\ \vdots \\ \bar{Y}_{I.} \\ \vdots \\ \bar{Y}_{I.} \end{pmatrix}$$

where each $\bar{Y}_{i.}$ appears n_i times.
The residuals are

$$\underset{n_T\times 1}{\mathbf{Y}} - \underset{n_T\times 1}{\hat{\mathbf{Y}}} = \begin{pmatrix} Y_{11} - \bar{Y}_{1.} \\ \vdots \\ Y_{1n_1} - \bar{Y}_{1.} \\ Y_{21} - \bar{Y}_{2.} \\ \vdots \\ Y_{2n_2} - \bar{Y}_{2.} \\ \vdots \\ Y_{I1} - \bar{Y}_{I.} \\ \vdots \\ Y_{In_I} - \bar{Y}_{I.} \end{pmatrix}$$

(b) Consider the alternative formulation used by `R` function `lm`:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, \cdots, I, j = 1, \cdots, n_i, \quad \alpha_1 = 0.$$

Express this model as a linear regression model by specifying $\mathbf{Y}$, $\mathbf{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$. Compare it with the linear regression model with $I-1$ indicator variables for factor levels (with level 1 as the reference class). What do you find?

ANS

$$\underset{n_T\times 1}{\mathbf{Y}} = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{I1} \\ \vdots \\ Y_{In_I} \end{pmatrix}; \quad \underset{n_T\times I}{\mathbf{X}} = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & 0 & \cdots & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & \cdots & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & \cdots & \cdots & 0 & 1 \end{pmatrix}; \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_2 \\ \vdots \\ \alpha_I \end{pmatrix}; \quad \underset{n_T\times 1}{\epsilon} = \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2n_2} \\ \vdots \\ \epsilon_{I1} \\ \vdots \\ \epsilon_{In_I} \end{pmatrix}.$$

It is the same as the linear regression model which sets factor level 1 as baseline, and define $I-1$ dummy variables $X_2, \cdots, X_I$:

$$X_2 = \begin{cases} 1 & \text{if } i = 2 \\ 0 & \text{otherwise} \end{cases}, \cdots, X_I = \begin{cases} 1 & \text{if } i = I \\ 0 & \text{otherwise} \end{cases}$$

6. **One-way ANOVA case study in R.**

A company uses six filling machines of the same make and model to place detergent into cartons that show a label weight of 32 ounces. The production manger has complained that the six machines do not place the same amount of fill into the cartons. A consultant requested that 20 filled cartons be selected randomly from each of the six machines and the content of each carton weighted. The observations were recorded in terms of the deviations of weights from 32 ounces. The data is under Files/Homework/filling.txt: The first column is the observation, the second column is the index for the filling machine and the third column is the index for the carton. Consider fitting the one-way ANOVA model to this data.

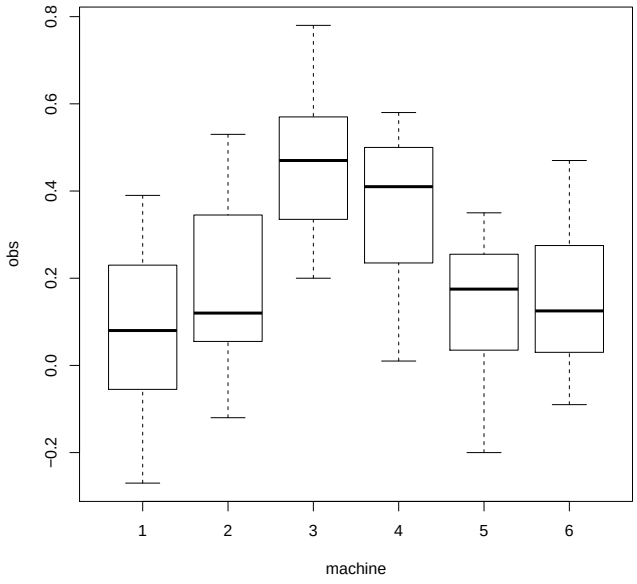
(a) What is the response variable? What is the factor? How many levels are there for this factor? Name the design of this study.

ANS The response variable is the the deviation of weight from 32 ounces.
The factor is the filling machine. It has 6 levels.
This is a balanced complete randomized design.

(b) Draw side-by-side box plots of the response for the factor levels. Do the factor level means appear to differ? Does the variability of the observations within each factor level appear to be approximately the same for all factor levels?

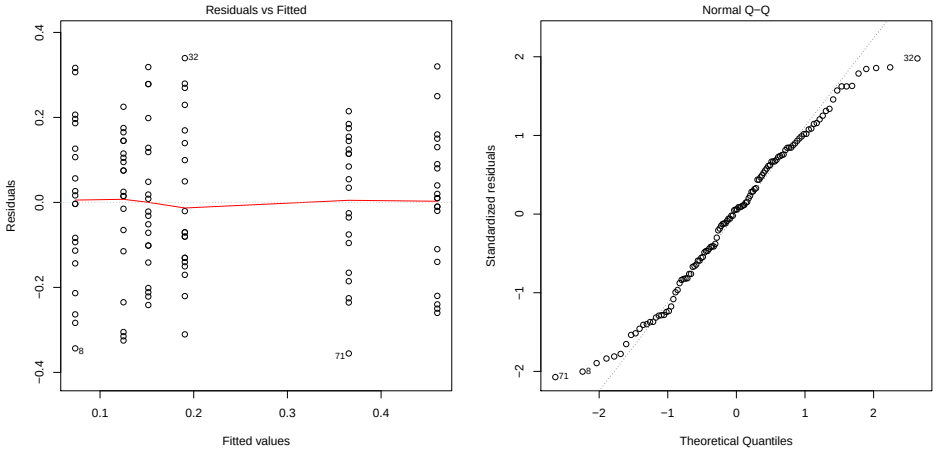
ANS The box plots of the response for the factor levels are shown below. The factor level means appear to differ. The sizes of the boxes are roughly equal so there is no

obvious sign of unequal variance



(c) Fit the one-way ANOVA model. Draw residual versus fitted value plot and residual Q-Q plot. Comment on model assumptions. Are remedial measures needed? *(Hint: When using the* `lm` *function, remember to declare the factor as* `factor`*.)*

ANS The normal QQ plot and residuals plots are shown below. The normal QQ plot indicates the slightly light tailed of the residuals' distribution; while the residuals versus fitted values plots shows no sign for unequal variance. Hence we could conclude that the model assumptions hold.



(d) Obtain the estimated factor levels means and obtain the ANOVA table.

ANS We fit the model under the restriction $\alpha_1 = 0$, i.e. we set level 1 as our baseline

```
> data=read.table("filling.txt")
> obs = data[,1]
> machine = as.factor(data[,2])
> n = nrow(data)

> fit=lm(obs~machine)
> summary(fit)

Call:
lm(formula = obs ~ machine)

Residuals:
Min      1Q  Median      3Q     Max
-0.3555 -0.1305  0.0100  0.1289  0.3395

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.07350    0.03935   1.868   0.0644 .
machine2     0.11700    0.05565   2.102   0.0377 *
machine3     0.38650    0.05565   6.945 2.47e-10 ***
machine4     0.29200    0.05565   5.247 7.22e-07 ***
machine5     0.05150    0.05565   0.925   0.3567
machine6     0.07800    0.05565   1.402   0.1638
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.176 on 114 degrees of freedom
Multiple R-squared:  0.3934,    Adjusted R-squared:  0.3668
F-statistic: 14.78 on 5 and 114 DF,  p-value: 3.636e-11
```

The estimated group means can be obtained from the summary output:

$$\hat{\mu}_1 = 0.0735 \qquad \hat{\mu}_2 = \hat{\mu}_1 + 0.1170 = 0.1905 \qquad \hat{\mu}_3 = \hat{\mu}_1 + 0.38650 = 0.4600$$

$$\hat{\mu}_4 = \hat{\mu}_1 + 0.29200 = 0.3655 \qquad \hat{\mu}_5 = \hat{\mu}_1 + 0.05150 = 0.1250 \qquad \hat{\mu}_6 = \hat{\mu}_1 + 0.07800 = 0.1515$$

ANS ANOVA output from R is shown as bellow:

```
> anova(fit)
Analysis of Variance Table

Response: obs
```

```
            Df Sum Sq Mean Sq F value    Pr(>F)
machine      5 2.2893 0.45787  14.784 3.636e-11 ***
Residuals  114 3.5306 0.03097
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The SSR is 2.2893, the SSE is 3.5306 from the output. The degrees of freedom for SSR is $I-1=5$, the degrees of freedom for SSE is $n-I=114$. Hence the ANOVA table is:

Source of variation	Sum of Squares (SS)	Degree of Freedom (df)	MS
Between treatments	$SSTR = 2.2893$	$I-1=5$	$MSTR = 0.45787$
Within treatments	$SSE = 3.5306$	$n_T - I = 114$	$MSE = 0.03097$
Total	$SSTO = 5.8200$	$n_T - 1 = 119$	

(e) Test whether or not the mean fill differs among the six machines at level 0.05. State the null and alternative hypotheses, the decision rule and the conclusion.

ANS

$$H_0 : \mu_1 = \cdots = \mu_6 \qquad \text{versus} \qquad H_a : \text{not all factor level means are the same}$$

$$F^* = \frac{MSTR}{MSE} = 14.784$$

Under the null hypothesis:$F^* \overset{iid}{\sim} F_{(5,114)}$
We can reject the null hypothesis if $F^* > F(1-\alpha, I-1, n_T-1)$ or if $p = P(F_{(5,114)} > 14.784) < \alpha$

$$p = P(F_{(5,114)} > 14.784) = 3.636 \times 10^{-11} < \alpha$$

Therefore, we can reject the null hypothesis at 5% level.

(f) Construct a 99% confidence interval for μ_2. If we are interested in all factor level means, what multiple comparison procedure shall we use? What would be the corresponding 99% confidence interval for μ_2?

ANS The $(1-\alpha)$-confidnce interval of μ_i is:

$$\widehat{\mu}_2 \pm s\left(\widehat{\mu}_2\right) t\left(1-\frac{\alpha}{2}; n_T - I\right)$$

$$\widehat{\mu}_2 = 0.1905 \qquad s\left(\widehat{\mu}_2\right) = \sqrt{\frac{MSE}{n_i}} = \frac{0.176}{\sqrt{20}} = 0.0393511 \qquad t\left(1-\frac{\alpha}{2}; n_T - I\right) = 2.6189$$

Therefore, the 99% confidence interval for μ_2 is:

$$[0.0874, 0.2936]$$

If we are interested in all factor level means, we should use Bonferroni procedure (here $g = 6$). The corresponding Bonferroni's multiplier is:

$$B = t\left(1 - \frac{\alpha}{2g}; n_T - I\right) = t(1 - \frac{0.01}{2 \times 6}; 114) = 3.2207$$

The corresponding confidence interval for μ_2 is:

$$\widehat{\mu}_2 \pm s\left(\widehat{\mu}_2\right) \times B = [0.0638, 0.3172]$$

(g) How many pairwise comparisons of factor levels means are there? If we are interested in all these pairwise comparisons, what multiple comparison procedure shall we use and what is its corresponding multiplier for $\alpha = 0.05$? Construct the corresponding 95% confidence intervals for all pairwise comparisons. At familywise significance level 0.05, which pairs of factor level means should be declared as being different?

ANS There are 15 pairwise comparisons of factor levels means. If we are interested in all these pairwise comparisons, we should use Tukey's procedure. The corresponding Tukey's multiplier is:

$$T = \frac{1}{\sqrt{2}} q\left(1 - \alpha; I, n_T - I\right) = \frac{1}{\sqrt{2}} q\left(0.95; 6, 114\right) = 2.8988$$

The Tukey's 95% confidence intervals for all pairwise comparisons were obtained by using the function `TukeyHSD`, and the output from R is shown as below:

```
> TukeyHSD(aov(fit))
Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = fit)

$mach
diff        lwr        upr     p adj
2-1  0.1170 -0.0443194  0.2783194 0.2934937
3-1  0.3865  0.2251806  0.5478194 0.0000000*
4-1  0.2920  0.1306806  0.4533194 0.0000106*
5-1  0.0515 -0.1098194  0.2128194 0.9392011
6-1  0.0780 -0.0833194  0.2393194 0.7260015
3-2  0.2695  0.1081806  0.4308194 0.0000588*
4-2  0.1750  0.0136806  0.3363194 0.0252432*
5-2 -0.0655 -0.2268194  0.0958194 0.8469184
6-2 -0.0390 -0.2003194  0.1223194 0.9815028
4-3 -0.0945 -0.2558194  0.0668194 0.5359056
5-3 -0.3350 -0.4963194 -0.1736806 0.0000003*
6-3 -0.3085 -0.4698194 -0.1471806 0.0000029*
5-4 -0.2405 -0.4018194 -0.0791806 0.0004684*
```

```
6-4 -0.2140 -0.3753194 -0.0526806 0.0026737*
6-5  0.0265 -0.1348194  0.1878194 0.9968910
```

At familywise significance level 0.05, there are eight pairs should be declared as being different, which are:
1 versus 3, 1 versus 4;
2 versus 3, 2 versus 4;
3 versus 5, 3 versus 6;
4 versus 5, 4 versus 6

(h) What if we are only interested in 6 **pre-specified** pairwise comparisons, which multiple comparison procedure shall we use and what is its corresponding multiplier for $\alpha = 0.05$? What if we are only interested in the 6 pairwise comparisons that show the most differences in the data (i.e corresponding to the six largest $|\widehat{D}|$), which procedure shall we use and what are the corresponding C.Is?

ANS If we are only interested in 6 pre-specified pairwise comparison, Bonferroni's procedure should be applied and the corresponding multiplier at $\alpha = 0.05$ is (note Bonferroni's multiplier in this case is smaller than Tukey's multiplier):

$$B = t\left(1 - \frac{\alpha}{2g}; n_T - I\right) = t(1 - \frac{0.05}{2 \times 6}; 114) = 2.6851$$

If we are only interested in the 6 pairwise comparisons that show the most differences in the data,we should use Tukey's multiplier as Bonferron's procedure is not applicable anymore: It is only applicable to pre-specified comparisons.

(i) If we are interested in all possible contrasts, which multiple comparison procedure shall we use? Construct the corresponding 95% confidence interval for the contrast:

$$L = \frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2}.$$

Test whether or not $L = 0$ with family-wise-error-rate controlled at 0.05. What if we are interested in 20 pre-specified contrasts (including L), which multiple comparison procedure shall we use and what is its corresponding multiplier for $\alpha = 0.05$? Construct the corresponding 95% confidence interval for L (assuming L is in this prespecified set of contrasts). What do you find?

ANS If we are interested in all possible contrasts, we should use Scheffe's procedure. Consider the constrast:

$$L = \frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2}$$
$$c_1 = c_2 = 0.5, c_3 = c_4 = -0.5$$

An unbiased estimator of L:

$$\hat{L} = \frac{\overline{Y}_{1.} + \overline{Y}_{2.}}{2} - \frac{\overline{Y}_{3.} + \overline{Y}_{4.}}{2} = \frac{0.0735 + 0.1905}{2} - \frac{0.4600 + 0.3655}{2} = -0.28075$$

$$s(\hat{L}) = \sqrt{MSE \times \sum_{i=1}^{4} \frac{c_i^2}{n_i}} = 0.03935$$

Since we are interested in all possible constrasts, we choose Scheffe's multipler at $\alpha = 0.05$:

$$S = \sqrt{(I-1)F(1-\alpha; I-1, n_T - I)} = \sqrt{5 \times F(0.95; 5, 114)} = 3.3867$$

So the 95% confidence interval for L is:

$$\hat{L} \pm s(\hat{L}) \times S = [-0.4140, -0.1475]$$

Since 0 is not contained in the above interval, we can reject the null hypothesis that $L = 0$.
If we are interested in 20 pre-specified contrasts, we should use Bonferroni's multiplier. Since att $\alpha = 0.05$, the corresponding Bonferroni's multiplier is:

$$B = t\left(1 - \frac{\alpha}{2g}; n_T - I\right) = t(1 - \frac{0.05}{2 \times 20}; 114) = 3.0919 < S = 3.39.$$

The corresponding 95% confidence interval for L is: $[-0.4024, -0.1591]$. The resulting confidence interval is narrower compared with the one obtain through Scheffe's procedure.