Homework 3

Yuanyou Yao

November 17, 2019

Problem 1

(a) $\Sigma_{j=0}^{p-1} c_j \hat{\beta}_j$ can be denoted as $c^T\hat{\beta}$,where c and $\hat{\beta}$ are vectors. In least squares estimate, $\hat{\beta} = (X^TX)^{-1}X^TY$, so

$$\text{Mean: } E(\Sigma_{j=0}^{p-1} c_j \hat{\beta}_j) = c^T\beta$$
$$\text{Variance: } Var(c^T\hat{\beta}) = c^T\sigma^2(X^TX)^{-1}c$$
$$\text{Distribution: } c^T\hat{\beta}\text{\textasciitilde}N(c^T\beta, c^T\sigma^2(X^TX)^{-1}c)$$

(b) Consider that $c^T\hat{\beta}$~$N(c^T\beta, c^T\sigma^2(X^TX)^{-1}c)$. Then we know that $c^T\hat{\beta} - c^T\beta$~$N(0, c^T\sigma^2(X^TX)^{-1}c)$.

Under H_0, we use T test, and the test statistics is

$$T = \frac{c^T\hat{\beta} - h}{c^T\sigma^2(X^TX)^{-1}c}$$

Under significant level α, the reject region is

$$|T| > t_{n-p,\alpha/2}$$

In this case, $\hat{\beta}$ is independent.

(c)

$$E(Y_{n+1} - \hat{Y}_{n+1}) = 0$$
$$Var(Y_{n+1} - \hat{Y}_{n+1}) = \sigma^2(1 + Z^T(X^TX)^{-1}Z)$$
$$Y_{n+1} - \hat{Y}_{n+1}\text{\textasciitilde}N(0, \sigma^2(1 + Z^T(X^TX)^{-1}Z))$$

(d) Proof:

$$MSE(\hat{Y}_{n+1}) = E(\hat{Y}_{n+1} - Y_{n+1})^2 = \sigma^2(1 + Z^T(X^TX)^{-1}Z)$$
$$Z^T(X^TX)^{-1}Z = Z^T(X^TX)^{-1}X^TX(X^TX)^{-1}Z > 0$$
$$MSE(\hat{Y}_{n+1}) = \sigma^2(1 + Z^T(X^TX)^{-1}Z) > \sigma^2$$

(e) We consider the T test.

$$T = \frac{Y_{n+1} - \hat{Y}_{n+1}}{\sqrt{\sigma^2(1 + Z^T(X^TX)^{-1}Z)}} \sim t_{n-p}$$

So, the interval I

$$(\hat{Y}_{n+1} - t_{n-p,\alpha/2}\sqrt{\sigma^2(1 + Z^T(X^TX)^{-1}Z)}, \hat{Y}_{n+1} + t_{n-p,\alpha/2}\sqrt{\sigma^2(1 + Z^T(X^TX)^{-1}Z)}$$

Problem 2

(a)

The first model can be described as a single line because it is a SLR.

The second model can be described as a parallel lines because it contains indicator variables thus it can change the intercept when $lmass = 0$, but cannot change the slope.

The third model can be described as a separate lines because it contains cross variables so it has both different intercepts and slopes.

(b)

F-test: whether there is a difference between the in-flight energy and body mass among birds, echolocating and non-echolocation bats. The second model is the reduced model and the first model is the full model.

Other test: whether there is a difference for birds, echolocating bats and non-echolocating bats after body mass is accounted for. The second model is the full model and the first model is the reduced model.

(c)

```
> anova(lm1,lm2,lm3)
Analysis of Variance Table

Model 1: lenergy ~ lmass
Model 2: lenergy ~ lmass + Type
Model 3: lenergy ~ Type * lmass
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     18 0.58289
2     16 0.55332  2  0.029574 0.4100 0.6713
3     14 0.50487  2  0.048450 0.6718 0.5265
```

Comparing the first, the second and third model, the p-values are 0.6713, 0.5265 seperately. Hence, the first model is the best, wich means that the full model is reduced to

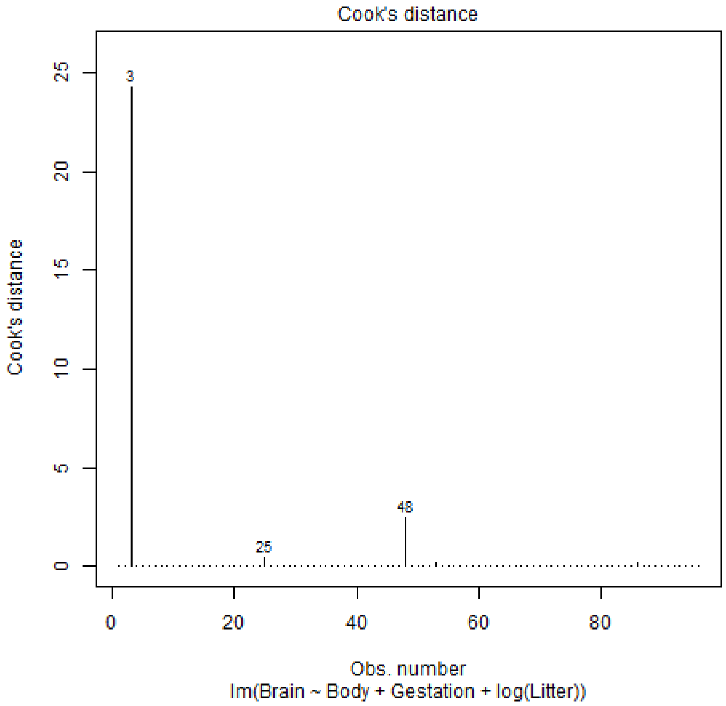
$$lenergy{\sim}lmass$$

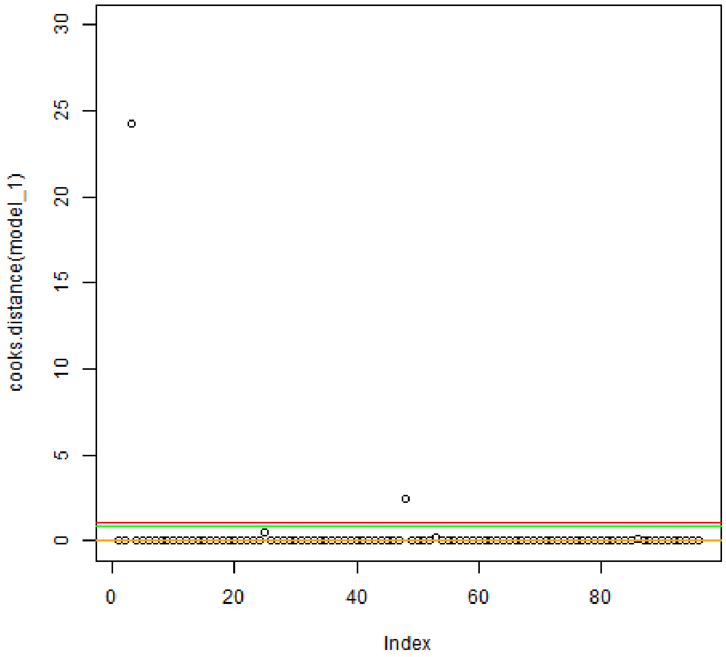
Problem 3

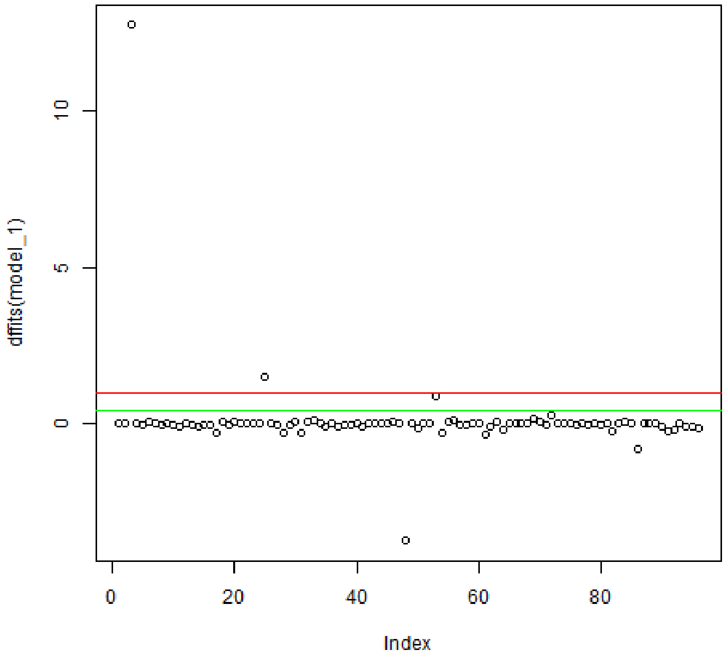
(a) By using R, we do the regression:

$$Y = -231.58 + 0.97X_1 + 1.93X_2 + 89ln(X_3)$$

To find influential points, we compute the cook's distance and fiffits value and the graphs are shown below:


Cook's distance
Cook's distance
0
5
10
15
20
25
3
25
48
0
20
40
60
80
Obs. number
lm(Brain ~ Body + Gestation + log(Litter))


cooks.distance(model_1)
0
5
10
15
20
25
30
0
20
40
60
80
Index

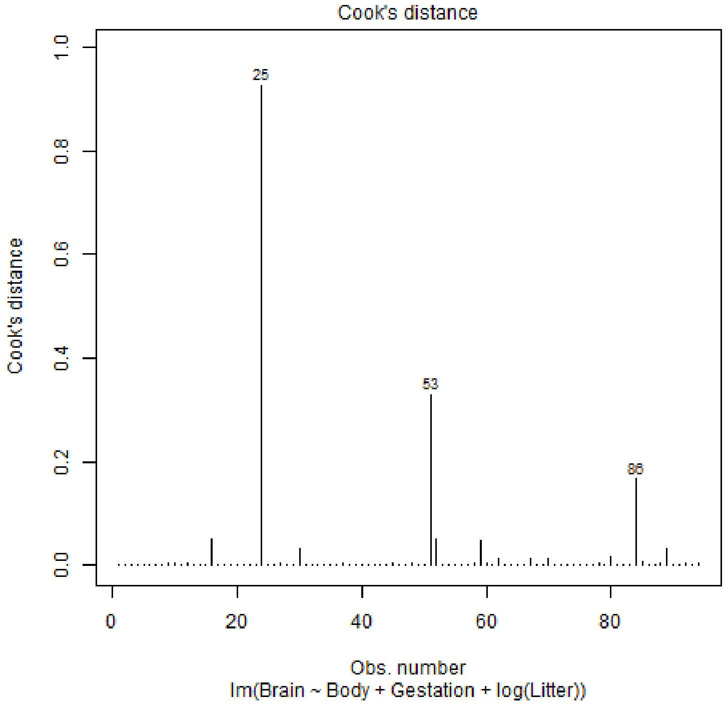


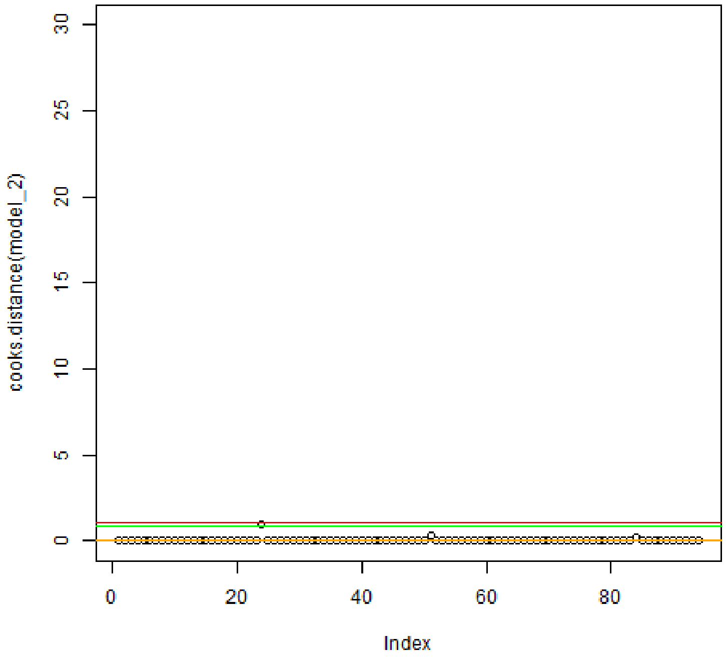
Based on the graphs, we can see that there are two influential points: 3, 25, 48, 53 and 86, which corresponding to African elephant, dolpin, hippopotamus, human being and Tapir respectively.

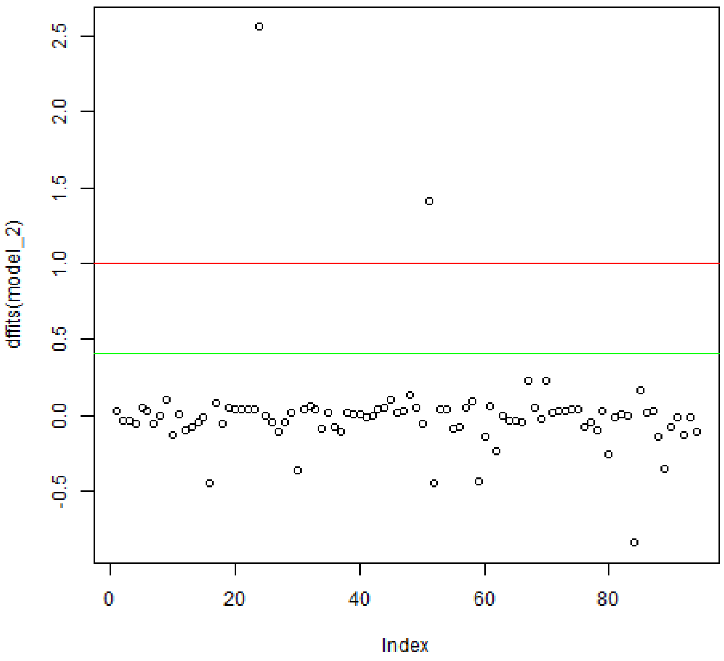
(b) After deleting the influential points, we do the regression again and get

$$Y = -135.44 + 0.47X_1 + 1.972X_2 + 39.38ln(X_3)$$

To find influential points, we compute the cook's distance and diffits value and the graphs are shown below:


Cook's distance
Cook's distance
0.0
0.2
0.4
0.6
0.8
1.0
25
53
86
0
20
40
60
80
Obs. number
lm(Brain ~ Body + Gestation + log(Litter))


cooks.distance(model_2)
0
5
10
15
20
25
30
0
20
40
60
80
Index



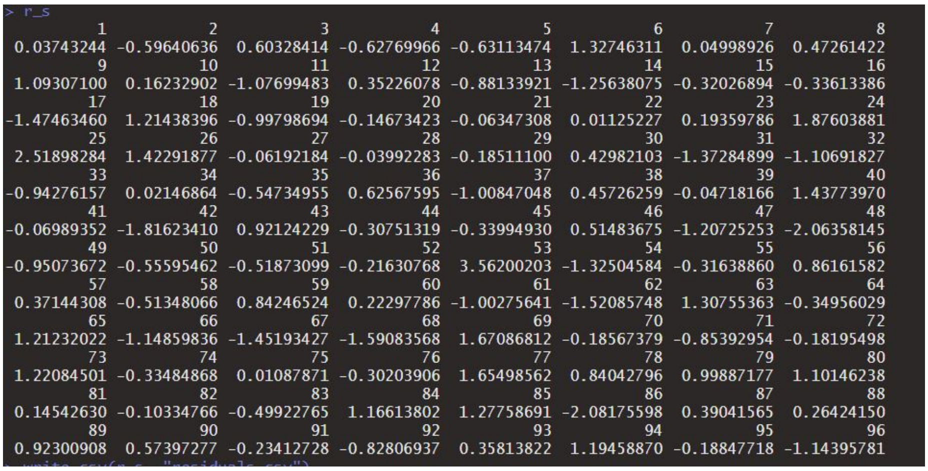
Based on the graphs, there are still two influential points; 25 and 53.

Main differences: The coefficient of body weight and litter size and interception changed, F-value and t-value are decreasing.

(c) Consider all the data and fit the regression of log brain weight on log body weight, log gestation, and log litter size:

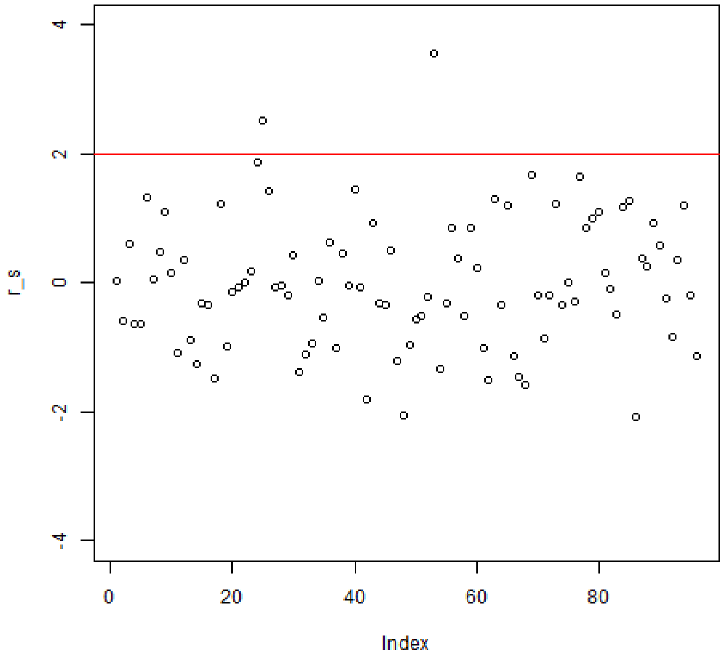
$$ln(Y) = 0.85482 + 0.57507ln(X_1) + 0.41794ln(X_2) - 0.31007ln(X_3)$$

Then, we can calculate the studentized residuals:



```
> r_s
          1           2           3           4           5           6           7           8 
 0.03743244 -0.59640636  0.60328414 -0.62769966 -0.63113474  1.32746311  0.04998926  0.47261422 
          9          10          11          12          13          14          15          16 
 1.09307100  0.16232902 -1.07699483  0.35226078 -0.88133921 -1.25638075 -0.32026894 -0.33613386 
         17          18          19          20          21          22          23          24 
-1.47463460  1.21438396 -0.99798694 -0.14673423 -0.06347308  0.01125227  0.19359786  1.87603881 
         25          26          27          28          29          30          31          32 
 2.51898284  1.42291877 -0.06192184 -0.03992283 -0.18511100  0.42982103 -1.37284899 -1.10691827 
         33          34          35          36          37          38          39          40 
-0.94276157  0.02146864 -0.54734955  0.62567595 -1.00847048  0.45726259 -0.04718166  1.43773970 
         41          42          43          44          45          46          47          48 
-0.06989352 -1.81623410  0.92124229 -0.30751319 -0.33994930  0.51483675 -1.20725253 -2.06358145 
         49          50          51          52          53          54          55          56 
-0.95073672 -0.55595462 -0.51873099 -0.21630768  3.56200203 -1.32504584 -0.31638860  0.86161582 
         57          58          59          60          61          62          63          64 
 0.37144308 -0.51348066  0.84246524  0.22297786 -1.00275641 -1.52085748  1.30755363 -0.34956029 
         65          66          67          68          69          70          71          72 
 1.21232022 -1.14859836 -1.45193427 -1.59083568  1.67086812 -0.18567379 -0.85392954 -0.18195498 
         73          74          75          76          77          78          79          80 
 1.22084501 -0.33484868  0.01087871 -0.30203906  1.65498562  0.84042796  0.99887177  1.10146238 
         81          82          83          84          85          86          87          88 
 0.14542630 -0.10334766 -0.49922765  1.16613802  1.27758691 -2.08175598  0.39041565  0.26424150 
         89          90          91          92          93          94          95          96 
 0.92300908  0.57397277 -0.23412728 -0.82806937  0.35813822  1.19458870 -0.18847718 -1.14395781 
> write.csv(r_s, "residuals.csv")
```

(d) To find influential points, we compute the cook's distance and fiffits value and the graphs are shown below:



we can conclude that at the points 25 and 53, which corresponding to dolphin and human, have substantially larger brain weights than were predicted by the model.

(e) According to the graphics shown in (d), the conclusion is — there is no mammals have substantially smaller brain weights than were predicted by the model.

(f)

1. After log transformation, F-value increased, which means the model fits better.
2. Removing influential points and refit the regression may not be a good way to make thing right, even make it worse.
3. The assumption of normality can hold by log transformation.

Problem 4

(a) SSTO = 8100 and $\hat{\sigma^2} = \frac{SSE}{df}$. The estimates of σ^2 are shown below:

Model variables	$\hat{\sigma}^2$
None	300
A	240
B	230
C	260
AB	220
AC	210
BC	230
ABC	215

(b) Adjusted $R^2 = 1 - \frac{\frac{SSE_p}{n-p}}{\frac{SSTO}{N-1}}$. The adjusted R^2 are shown below:

Model variables	Adjusted R^2
None	0
A	0.2
B	0.233
C	0.133
AB	0.267
AC	0.3
BC	0.233
ABC	0.283

(c) $C_p(M) = \frac{SSE(M)}{\hat{\sigma}^2} - n + 2p(M)$, the C_p statistics are shown below:

Model variables	C_p
None	11.674
A	5.023
B	3.814
C	7.442
AB	3.851
AC	2.419
BC	4.744
ABC	4

(d) $BIC_p = nln(SSE_p) - nln(n) + pln(n)$. The BIC are shown below:

Model variables	BIC_p
None	162.0198
A	158.0473
B	156.8556
C	160.2885
AB	157.8450
AC	156.5424
BC	159.0896
ABC	159.3905

(e)

(i)The last model with variables A, B and C has the smallest $\hat{\sigma}^2$, 198.46.

(ii)The model with variables A and C has the largest adjusted R^2, 0.3.

(iii)The model with variables A and C has the smallest C_p statistic, 2.419.

(iv)The model with variables A and C has the smallest BIC, 156.54.

(f)

Starting with none variable, we get the model contains Bhas the smallest SSE. Then we perform an extra-sum-of-squares F -test to see whether that variable is significant.

Under $\alpha = 0.05$, $F = 7.75$. Thus, B is significant.

Then, we get AB has the smaller SSE. Samely, we perform an extra-sum-of-squares F -test to see whether the additional variable is significant.

Under $\alpha = 0.05$, $F = 2.182 < 4.24 = F_{1,n-4,0.05}$. So Ais insignificant.

Finally, by forward selecion, we have the model form:

$$Y\beta_0 + \beta_2 B + \epsilon$$

(g) To find the posterior probability for these 8 models, we first subtract the smallest BIC from the other ones, and then calculate the difference of each model as below:

$$BIC_j^* = BIC_j - BIC_{min}$$

Then,we have the posterior probability:

$$p(M_j) = \frac{e^{-\frac{1}{2}BIC_j^*}}{\sum_{j=1}^{k} e^{-\frac{1}{2}BIC_j^*}}$$

The posterior distributions are shown below:

Model variables	Posterioir distribution
None	0.01802701
A	0.13138647
B	0.23840663
C	0.04284313
AB	0.14537115
AC	0.27882110
BC	0.07802002
ABC	0.06712450

A Appendix: Problem 2

```
bats = read.csv( file = ”bats.csv” , header = T)
bats = transform(bats , lenergy = log(Energy))
bats = transform(bats , lmass = log(Mass))
lm1 = lm(lenergy ~ lmass , data = bats)
lm2 = lm(lenergy ~ lmass + Type, data = bats)
lm3 = lm(lenergy ~ Type*lmass , data = bats)
anova(lm1,lm2,lm3)
```

B Appendix: Problem 3

```
data = read.csv( ”brain.csv”)
model_1=lm(Brain ~ Body+Gestation+log(Litter) ,data = data)
summary(model_1)
cooks.distance(model_1)
n = dim(model.matrix(model_1))[1]
p = dim(model.matrix(model_1))[2]
png(”3_a_1.png”)
plot(model_1, which = 4)
dev.off ()
png( ”3_a_2.png”)
plot(cooks.distance(model_1) ,ylim = c (0,30))
abline(h=1, col=”red”)
abline(h=qf (0.5 , p, n-p) , col=”green”)
abline(h=4/n, col=”blue”)
abline(h=4/(n-p-1-1), col=”orange”)
dev.off ()
dffits (model_1)
png(”3_a_3.png”)
plot( dffits (model_1) )
abline(h=1, col=”red”)
abline(h=2*sqrt(p/n) , col=”green”)
```

```
dev.off ()
summary(influence.measures(model_1))
data_new = data[-48,][-3,]
model_2=lm(Brain ~ Body+Gestation+log(Litter) ,data = data_new)
summary(model_2)
cooks.distance(model_2)
n = dim(model.matrix(model_2))[1]
p = dim(model.matrix(model_2))[2]
png("3_b_1.png")
plot(model_2, which = 4)
dev.off ()
png( "3_b_2.png")
plot(cooks.distance(model_2) ,ylim = c (0,30))
abline(h=1, col="red")
abline(h=qf (0.5 , p, n-p) , col="green")
abline(h=4/n, col="blue")
abline(h=4/(n-p-1-1), col="orange")
dev.off ()
dffits (model_2)
png("3_b_3.png")
plot( dffits (model_2) )
abline(h=1, col="red")
abline(h=2*sqrt(p/n) , col="green")
dev.off ()
summary(influence.measures(model_2))
```