STAT 350

Assignment4

Yuhan Zhang
301345627

Assignment 4

1. $$D_i = \frac{(\hat{\beta}_i-\hat{\beta})'(X'X)(\hat{\beta}_i-\hat{\beta})}{p\,MS_{Res}}$$

$\left[\ \hat{\beta}-\hat{\beta}_i = \dfrac{(X'X)^{-1}x_i e_i}{1-h_{ii}}\ \text{From (8.1)}\right]$

$$= \frac{[-(X'X)^{-1}x_i e_i]'(X'X)[-(X'X)^{-1}x_i e_i]}{(1-h_{ii})^2\,p\cdot MS_{Res}}$$

$$= \frac{e_i' x_i'(X'X)^{-1}\cancel{(X'X)(X'X)^{-1}}x_i e_i}{(1-h_{ii})^2\,p\cdot MS_{Res}}$$

$$= \frac{e_i' x_i'(X'X)^{-1}x_i e_i}{(1-h_{ii})^2\,p\cdot MS_{Res}}$$

$$= \frac{x_i'(X'X)^{-1}e_i^2 x_i}{(1-h_{ii})^2\,p\cdot MS_{Res}}$$

$$= \frac{h_{ii}e_i^2}{(1-h_{ii})^2\,p\cdot MS_{Res}}$$

$$= \frac{e_i^2}{(1-h_{ii})MS_{Res}}\cdot\frac{h_{ii}}{(1-h_{ii})\cdot p}$$

$\left[\ r_i = \dfrac{e_i}{\sqrt{MS_{Res}(1-h_{ii})}}\ \right]$

$$= \frac{r_i^2\cdot h_{ii}}{p\cdot(1-h_{ii})}$$

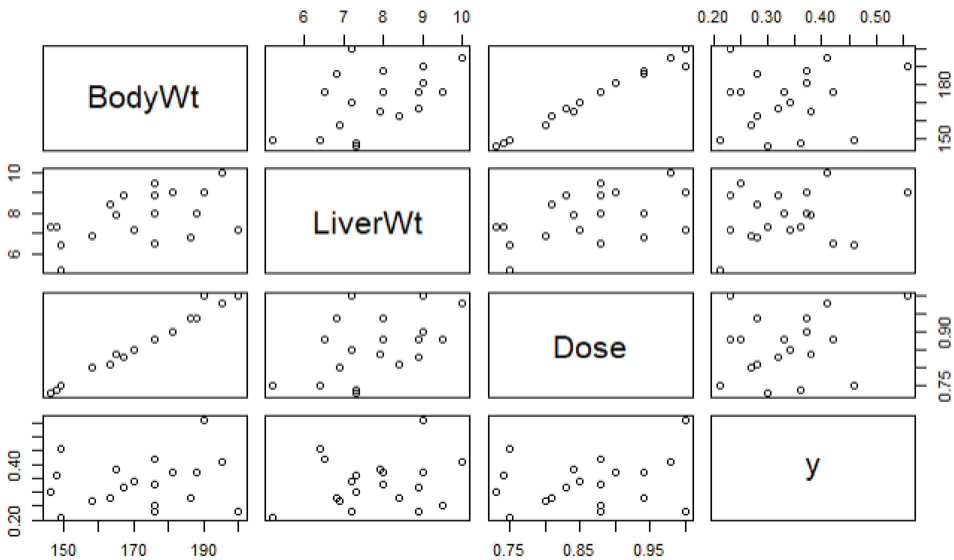
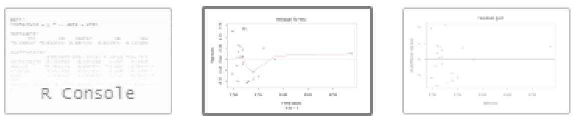
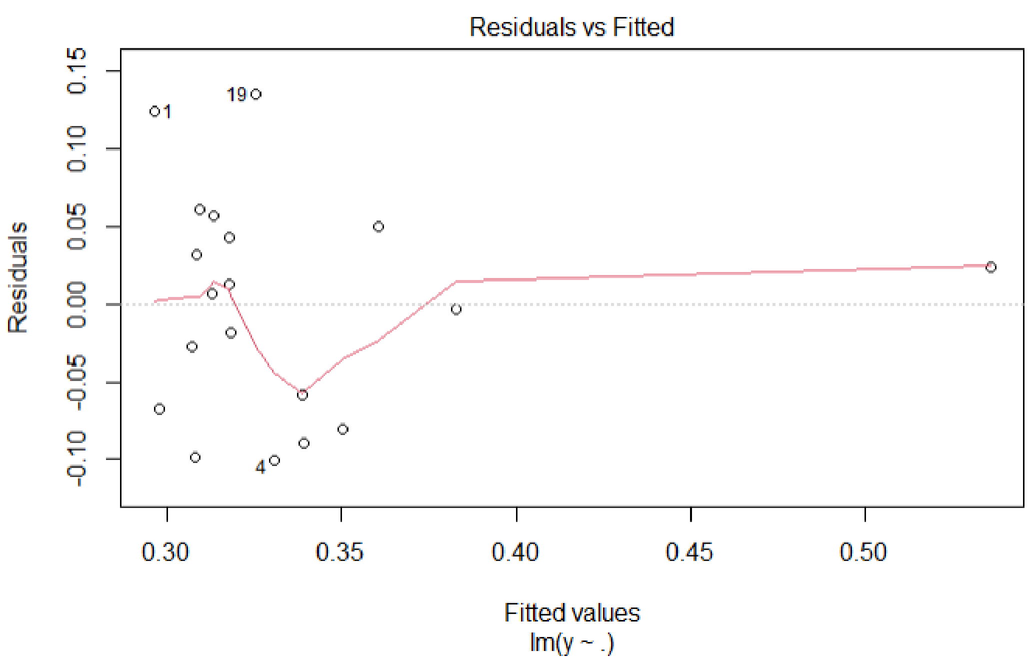
Q2 a

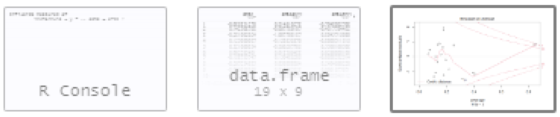
```
Coefficients:
(Intercept)           z1           z2
 -5.959e-17    9.137e-01    3.068e-01
```

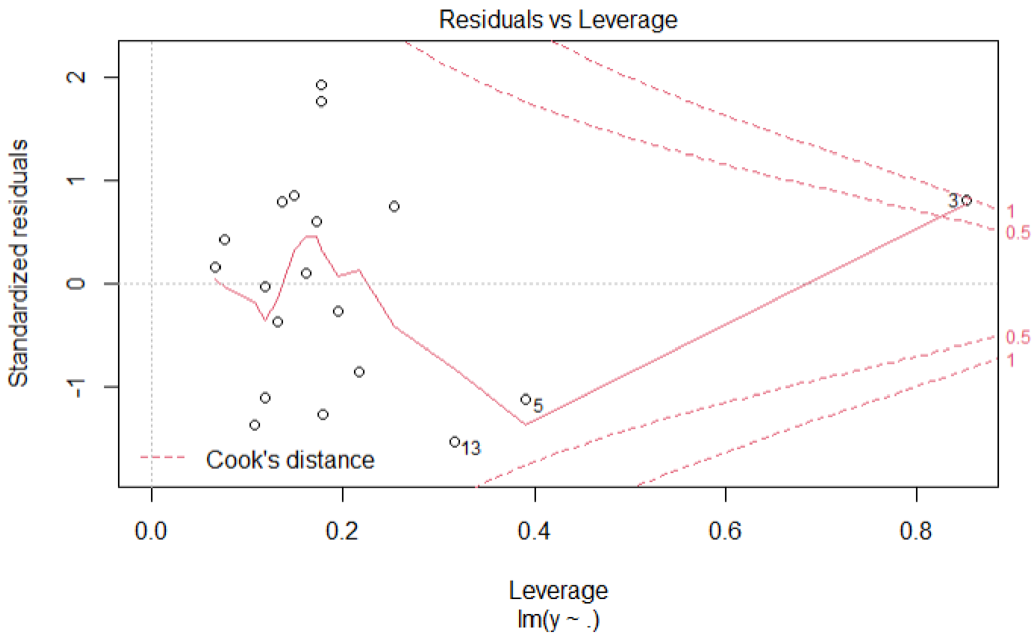
Q2 b

```
##b.
```{r}
round(vcov(fit1),5)
```
```

```
            (Intercept)      z1      z2
(Intercept)     0.00132 0.00000 0.00000
z1              0.00000 0.01185 0.00000
z2              0.00000 0.00000 0.01185
```

As the covariance are all zero, They are independent.

Q2 c

```
##c
```{r}
vif(fit1)
```
```

```
z1 z2
 1  1
```

Q2 d

```
Coefficients:
(Intercept)          z1           z2
 -5.508e-17   8.941e-01   -1.155e-02
```

Q2 e

```
## e
```{r}
round(vcov(fit2),5)
```
```

```
            (Intercept)       z1       z2
(Intercept)     0.00395  0.00000  0.00000
z1              0.00000  0.05908 -0.03727
z2              0.00000 -0.03727  0.05908
```

They are not independent because Var(Beta1(hat),Beta2(hat))is not equal to 0.

Q2 f

```
##f
```{r}
vif(fit2)
```

      z1   z2
1.660661 1.660661
```

Q2 g

```
## g.
The first is preferable as it does not have multicollinearity.
Besides, the variance the beta(hat) is smaller.
```

Q4 a

````
## a.

```{r}
setwd("C:/Users/carol/Desktop/stat350")
df3 = read.csv("guinea_pig.csv")[,-1]
tail(df3)
pairs(df3)
```
````

R Console

data.frame
6 x 4




```
##i
There is a positive linear relation between body weight and dose.There is a weak positive linear
relationship between body weight and liver weight, and a weak positive linear relationship between
liverweight and dose.
##ii
There is no obvioudd relationship between predictors and response variables.
```

Q4 bc

````
## b.
body weight and dose
```{r}
fit3 = lm(y~., data=df3)
summary(fit3)

##body and dose, there is a significant linear relationship.

## c.
plot(fit3, 1)
plot(fitted(fit3),rstudent(fit3),main="residual plot", ylab="studentaized residual")
abline(h=0)
```
````






```
##c
There is no linear relationship.
```

Q4 d

d.

```{r}
nrow(df3)
X<-cbind(rep(1,nrow(df3)), df3[,1],df3[,2],df3[,3])

hii<-diag(X%*%solve(t(X)%*%X)%*%t(X))
hii
# Identify points of high Leverage
p<-ncol(X) # number of betas in the model (beta0,beta1,beta2)
n<-nrow(X) # number of observations
which(hii>2*p/n) #points with high leverage
```

```
[1] 19
 [1] 0.17798270 0.17934099 0.85091457 0.10761585 0.39153825 0.16115958 0.13688107 0.25367448
 [9] 0.06701578 0.11968672 0.11950583 0.17239599 0.31618336 0.13140699 0.07617481 0.21661460
[17] 0.19522441 0.14872221 0.17796183
[1] 3
```

Yes.The third one has large hii.

Q4 e

````
## e.
```{r}
influence.measures(fit3)
plot(fit3, 5)
```
````






```
There is no cook's distance larger than one.

The 3rd obs has large leverage and cook's distance, it has relative large impact on the regression coef.
```

Q4 f

f.

```{r}
vif(fit3)
```

```
   BodyWt    LiverWt       Dose 
52.101917  1.335679 51.427154 
```

Yes there is evidence of collinearity, as the vif of bodywt and does is 52.1 and 51.43 which are larger than 10.

Q4 g

g.

```{r}
fit4 = lm(y~., data=df3[-3,])
summary(fit4)
vif(fit4)
```

```
Call:
lm(formula = y ~ ., data = df3[-3, ])

Residuals:
      Min        1Q    Median        3Q       Max 
-0.102154 -0.056486  0.002838  0.046519  0.137059 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.311427   0.205094   1.518    0.151
BodyWt      -0.007783   0.018717  -0.416    0.684
LiverWt      0.008989   0.018659   0.482    0.637
Dose         1.484877   3.713064   0.400    0.695

Residual standard error: 0.07825 on 14 degrees of freedom
Multiple R-squared:  0.02106,	Adjusted R-squared:  -0.1887 
F-statistic: 0.1004 on 3 and 14 DF,  p-value: 0.9585

    BodyWt    LiverWt       Dose 
259.449422   1.445674 253.199751 
```

As all the p-value is larger than 0.05, no independent variable have significant linear relationship with the response variable, which is as a result of the multicollinearity increse the standard error of the beata hats. Therefore reduce the t-test and increase the p-value leading to non-significant result.