HW 6 Sample Solution

11.12

```
ex1112=data.frame(SPECIES=case0902$Species,BRAIN=case0902$Brain,GESTATION=case0902
$Gestation,LITTER=case0902$Litter)
```

(a) The fitted regression of brain weight on body weight, gestation, and log litter size is given below.

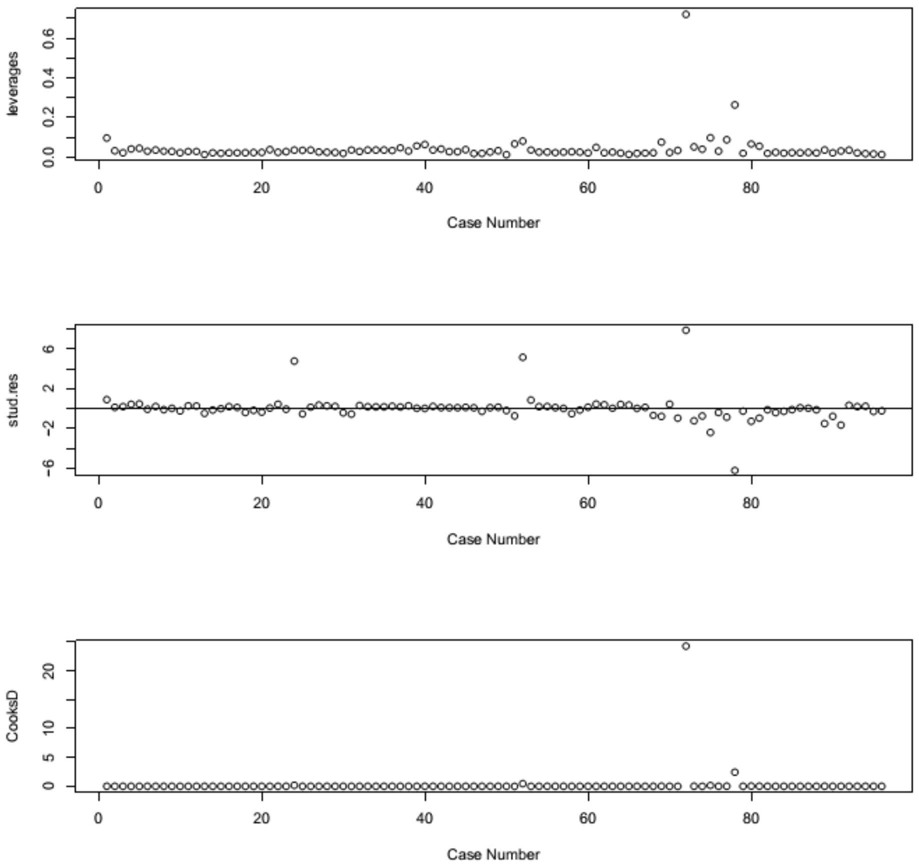
```
> ex1112$loglitter <- with(ex1112,log(LITTER))
> m1 <- lm(BRAIN~BODY+GESTATION+loglitter,data=ex1112)
> summary(m1)

Call:
lm(formula = BRAIN ~ BODY + GESTATION + loglitter, data = ex1112)

Residuals:
     Min       1Q   Median       3Q      Max 
-1004.10   -57.23    19.00    52.39   981.03 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -231.58329   78.05321  -2.967  0.00383 ** 
BODY           0.97318    0.09495  10.249  < 2e-16 ***
GESTATION      1.93012    0.37845   5.100 1.81e-06 ***
loglitter     89.00223   48.78939   1.824  0.07137 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 223.6 on 92 degrees of freedom
Multiple R-squared: 0.8116,Adjusted R-squared: 0.8055 
F-statistic: 132.1 on 3 and 92 DF,  p-value: < 2.2e-16
```

Case-influence statistics are given below. From the plots we see one observation has a very large Cook's distance (24.29). Closer investigation reveals that this observation is case number 72 (African elephant).

Its considerable influence is due to both a very large leverage (0.722) and a high Studentized residual (7.91).

```
> leverages <- hat(model.matrix(m1))
> stud.res <- rstudent(m1)
> CooksD <- cooks.distance(m1)
> par(mfrow=c(3,1))
> plot(leverages,xlab="Case Number")
> plot(stud.res,xlab="Case Number"); abline(0,0)
> plot(CooksD,xlab="Case Number")
> print(cbind(leverages,stud.res,CooksD)[70:76,],digits=2)
   leverages stud.res  CooksD
70     0.023     0.47  0.0013
71     0.033    -0.94  0.0076
72     0.722     7.91 24.2921
73     0.051    -1.21  0.0198
74     0.040    -0.73  0.0056
75     0.097    -2.39  0.1459
76     0.030    -0.38  0.0011
```

```
> ex1112[72,]
            SPECIES BRAIN BODY GESTATION LITTER loglitter
72 African elephant  4480 2800       655      1         0
```



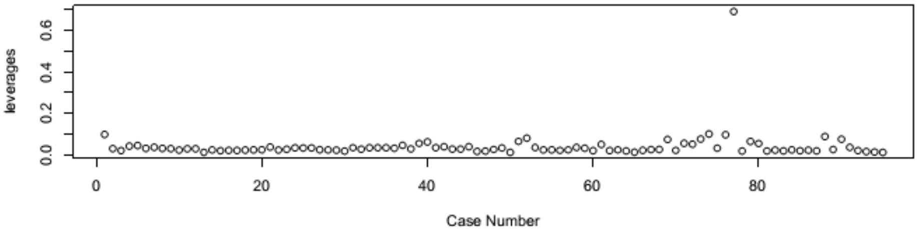
(b) Without the African elephant, we re-fit the model:

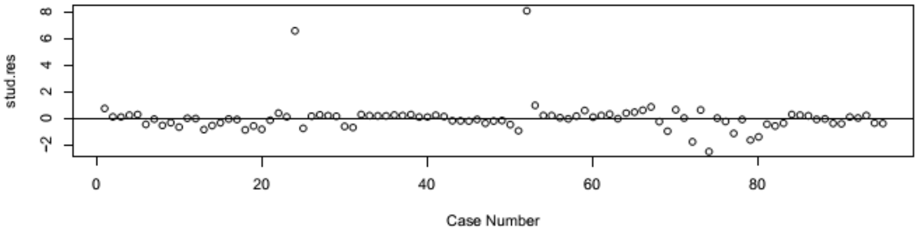
```
> subdat <- ex1112[-72,]
> m2 <- lm(BRAIN~BODY+GESTATION+loglitter,data=subdat)
> summary(m2)

Call:
lm(formula = BRAIN ~ BODY + GESTATION + loglitter, data = subdat)

Residuals:
     Min       1Q   Median       3Q      Max
-400.489  -65.331    1.042   35.591 1026.395

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -155.4376    61.1706  -2.541   0.0127 *
BODY           0.2847     0.1139   2.499   0.0142 *
GESTATION      1.8986     0.2929   6.482 4.59e-09 ***
loglitter     49.1398    38.0948   1.290   0.2003
```

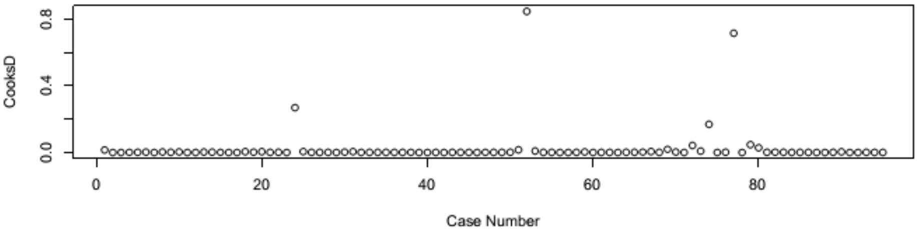
```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 173.1 on 91 degrees of freedom
Multiple R-squared: 0.5509,Adjusted R-squared: 0.5361
F-statistic: 37.21 on 3 and 91 DF,  p-value: 8.696e-16
```

Case-influence statistics are given below. We now notice two relatively influential observations: Cook's distance for dolphin and hippopotamus are (respectively) 0.85 and 0.72. The dolphin has a small leverage but very large Studentized residual (indicating very large brain size), and the hippopotamus has a small studentized residual but very large leverage.

```
> print(cbind(leverages,stud.res,CooksD)[c(52,77),],digits=2)
   leverages stud.res CooksD
52     0.081      8.1   0.85
78     0.692     -1.1   0.72
> subdat[c(52,77),]
        SPECIES BRAIN BODY GESTATION LITTER loglitter
52      Dolphin  1600  160       360      1         0
78 Hippopotamus   590 1400       240      1         0
```







(c) When there is a *true* outlier, its elimination produces an analysis with no other problems. When there is a scale problem requiring transformation, it persists even when apparent "outliers" are eliminated. Since values increase in orders of magnitude, removing the largest observation just re-scales the problem, and the next largest observation could become influential, even if it was not influential before.

11.13. Brain Weights.

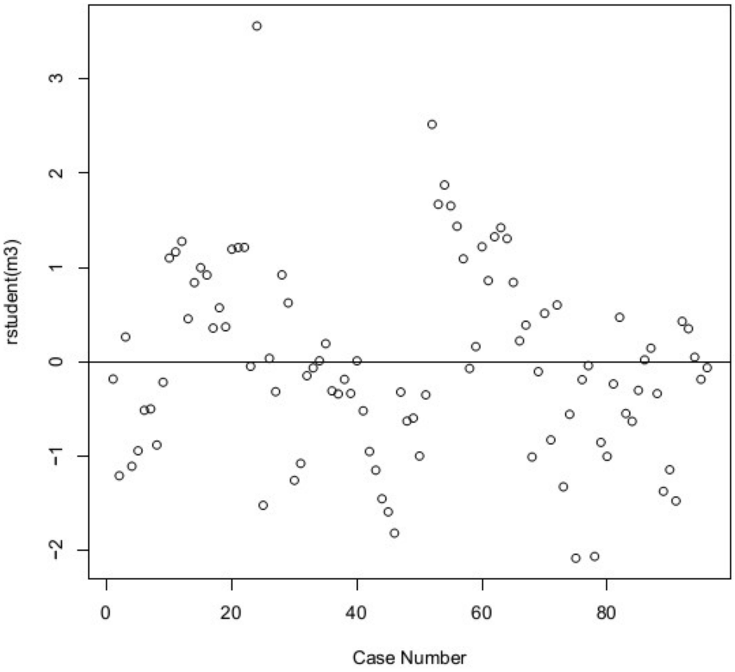
```
> ex1112$logbrain <- with(ex1112,log(BRAIN))
> ex1112$logbody <- with(ex1112,log(BODY))
> ex1112$loggest <- with(ex1112,log(GESTATION))
> m3 <- lm(logbrain~logbody+loggest+loglitter,data=ex1112)
> summary(m3)

Call:
lm(formula = logbrain ~ logbody + loggest + loglitter, data = ex1112)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.95415 -0.29639 -0.03105  0.28111  1.57491 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.85482    0.66167   1.292  0.19962    
logbody      0.57507    0.03259  17.647  < 2e-16 ***
loggest      0.41794    0.14078   2.969  0.00381 ** 
loglitter   -0.31007    0.11593  -2.675  0.00885 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 0.4748 on 92 degrees of freedom
Multiple R-squared: 0.9537,Adjusted R-squared: 0.9522 
F-statistic: 631.6 on 3 and 92 DF,  p-value: < 2.2e-16 
```

The fitted regression line is

$$\hat{y} = 0.855 + 0.575(logbody) + 0.418(loggest) - 0.310(loglitter)$$

Studentized residuals are given below. Human beings have the largest studentized residual (3.562), meaning we have substantially larger brains weight than were predicted by the model.

```
> plot(rstudent(m3),xlab="Case Number"); abline(0,0)
> rstudent(m3)[24]
      24 
3.562002 
> ex1112[24,]
       SPECIES BRAIN BODY GESTATION LITTER loglitter logbrain  logbody  loggest
24 Human being  1300   65       270      1         0  7.17012 4.174387 5.598422
```


rstudent(m3)
3
2
1
0
-1
-2
0
20
40
60
80
Case Number

12.10. ABC Regression.

(a) The estimate of σ^2 for any model is $\frac{\text{RSS}}{\text{d.f.}}$ (see chart below).

(b) The adjusted R^2 for each model is $100\frac{\text{(Total mean square)}-\text{(Residual mean square)}}{\text{Total mean square}}\%$ (see chart below).

(c) The AIC for each model is $n \times \log(\hat{\sigma}^2) + 2p$, where p is the number of regression coefficients (β's) (see chart below).

(d) The BIC for each model is $n \times \log(\hat{\sigma}^2) + p \times \log(n)$ (see chart below).

Model variables	$\hat{\sigma}^2$	adjusted R^2	AIC	BIC
None	300	0.00%	161.71	163.04
A	240	20.00%	157.46	160.12
B	230	23.33%	156.27	158.93
C	260	13.33%	159.70	162.36
AB	220	26.67%	157.02	161.02
AC	210	30.00%	155.72	159.72
BC	230	23.33%	158.27	162.26
ABC	215	28.33%	158.38	163.71

(e) See the chart below for the "best" models, according to different criteria:

Criterion	smallest $\hat{\sigma}^2$	largest adjusted R^2	smallest AIC	smallest BIC
Model variables	AC	AC	AC	B

12.11. ABC Regression.

Step # 1:

The model with the smallest RSS is B. Its extra-sum-of-squares F-statistic is

$$F = \frac{\left(\frac{\text{extra SS}}{\text{extra d.f.}}\right)}{\hat{\sigma}^2_{full}} = \frac{(8,100 - 5,980)}{230} = \mathbf{9.217}$$

with 1 and 26 degrees of freedom. This exceeds $F_{1,26}(.95) = 4.225$, so we choose the model with B and continue.

```
> qf(.95,1,26)
[1] 4.225201
```

Step #2:

Examine models AB and BC. AB has the smaller RSS, and its extra-sum-of-squares F-statistic is

$$F = \frac{5,980 - 5,500}{220} = \mathbf{2.182}$$

with 1 and 25 degrees of freedom. This does **not** exceed $F_{1,25}(.95) = 4.242$, so we do **not** choose the model with AB. Forward selection settles on model B.

```
> qf(.95,1,25)
[1] 4.241699
```

SEX DISCREMENATION STUDY

```
casedat=data.frame(BSAL=case1202$Bsal,SAL77=case1202$Sal77,SEX=case1202$Sex,SENIOR
=case1202$Senior,AGE=case1202$Age,EDUC=case1202$Educ,EXPER=case1202$Exper)
```

Fit a Rich Model.

Since we're modeling the logarithm of beginning salary, we first need to perform the appropriate transformation:

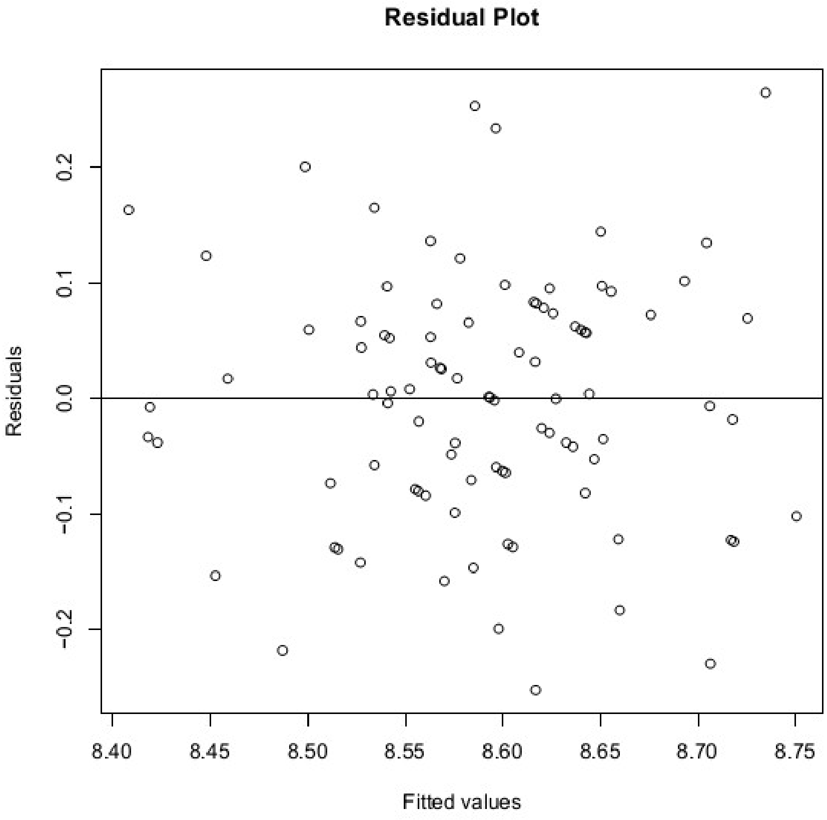
```
casedat$logsal <- with(casedat, log(BSAL))
```

Next we fit a rich model, using all potential explanatory variables *except* "sex":

```
> mrich <- lm(logsal˜SENIOR+AGE+EDUC+EXPER,data=casedat)
```

By looking at a residual plot (on following page), the assumptions of independence and constant variance appear to be valid.

```
> plot(mrich$res˜mrich$fit,xlab="Fitted values",ylab="Residuals",main="Residual Plot")
> abline(0,0)
```



Backward Elimination Using AIC.

We begin with the rich model and remove one variable at a time, until we achieve the model with the smallest possible AIC:

```
> mstep <- step(mrich,direction="backward")
Start:  AIC=-407.8
logsal ~ SENIOR + AGE + EDUC + EXPER

         Df Sum of Sq     RSS     AIC
- AGE     1      0.02    1.06 -408.35
<none>                   1.04 -407.80
- EXPER   1      0.07    1.11 -404.07
- SENIOR  1      0.16    1.20 -396.81
- EDUC    1      0.25    1.30 -389.43

Step:  AIC=-408.35
logsal ~ SENIOR + EDUC + EXPER

         Df Sum of Sq     RSS     AIC
<none>                   1.06 -408.35
- EXPER   1      0.07    1.12 -404.62
- SENIOR  1      0.14    1.20 -398.62
- EDUC    1      0.30    1.36 -386.93
```

The chosen model has *senior*, *education*, and *experience* as the explanatory variables.

Add “sex” to the Chosen Model.

Take the model chosen using backward elimination, and update it by adding the indicator for male employees:

```
> mfinal <- update(mstep,~.+SEX)
> summary(mfinal)

Call:
lm(formula = logsal ~ SENIOR + EDUC + EXPER + SEX, data = casedat)

Residuals:
      Min        1Q    Median        3Q       Max
-0.227738 -0.062870  0.000585  0.054120  0.209065

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.6710332  0.0955077  90.789  < 2e-16 ***
SENIOR      -0.0043456  0.0009435  -4.606 1.38e-05 ***
EDUC         0.0164092  0.0044805   3.662 0.000426 ***
EXPER        0.0002611  0.0001065   2.451 0.016225 *
SEXMALE      0.1295564  0.0213789   6.060 3.30e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09207 on 88 degrees of freedom
Multiple R-squared: 0.5144,Adjusted R-squared: 0.4923
F-statistic:  23.3 on 4 and 88 DF,  p-value: 3.724e-13
```

There is convincing evidence that the median starting salary for males was higher than the median starting salary for females, even after the effects of age, education, previous experience, and seniority are taken into account (two-sided p-value < 0.0001).

The median beginning salary for males was estimated to be 13.8% **greater** than the median salary for females. A 95% confidence interval for the ratio of medians is 109.10% to 118.773%.

```
> exp(0.1295564)
[1] 1.138323
> exp(0.1295564-0.0213789*qt(.975,df=88))
[1] 1.090973
> exp(0.1295564+0.0213789*qt(.975,df=88))
[1] 1.187728
```