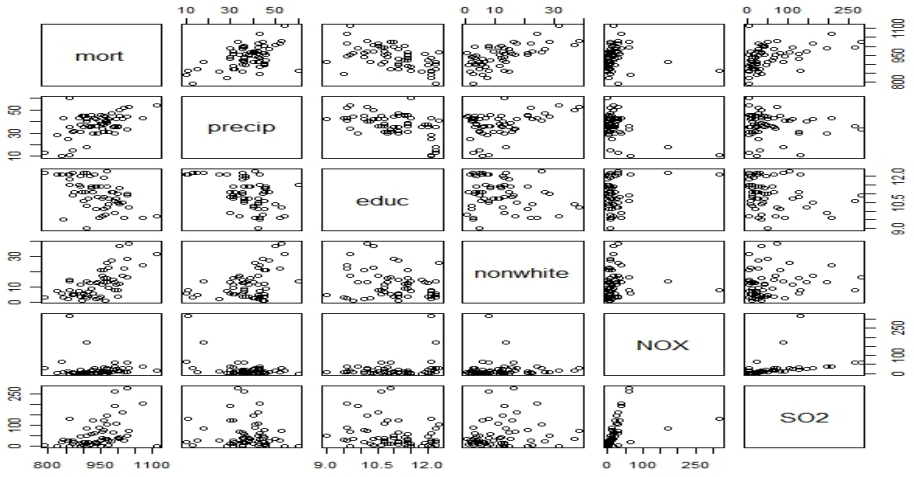
Tutorial 6 Solution

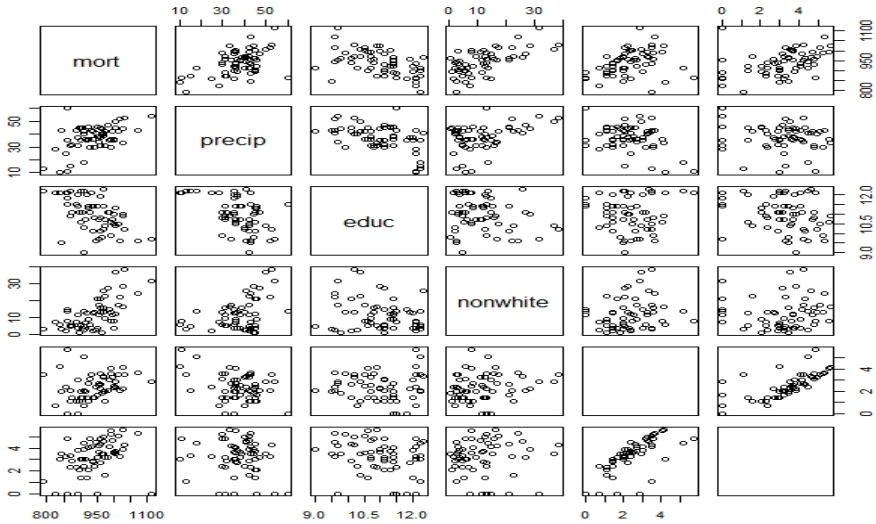
QUESTION 1 (revised based on the exercise in Chapter 11 from "The Statistical Sleuth")

Does pollution kill people? Data in one early study designed to explore this issue came from 5 Standard Metropolitan Statistical Areas (SMSA) in the United States, obtained for the years 1959-1961. Total age-adjusted mortality from all causes (mortality), in deaths per 100,000 population, is the response variable. The explanatory variables listed in the display below include mean annual precipitation (in inches); median number of school years completed, for persons of age 25 years or older; percentage of 1960 population that is nonwhite; relative pollution potential of oxides of nitrogen, NO_x; and relative pollution potential of sulphur dioxide, SO_2. "Relative pollution potential" is the product of the tons emitted per day per square kilometre and a factor correcting for SMSA dimension and exposure. The data is contained in "pollution.txt". [data from: G.C McDonald and J.A. Ayers, "Some applications of the 'Chernoff Faces': A technique for graphically representing multivariate data", in Graphical Representation of Multivariate Data, New York, 1978]

a) Is there evidence that mortality is associated with either of the pollution variables, after the effects of the climate and socioeconomic variables are accounted for? To answer this question you must first decide upon an appropriate model. **[Do not worry about higher order terms or interactions in your modelling process.]** After finding an appropriate model make sure to look at the case-influence statistics.

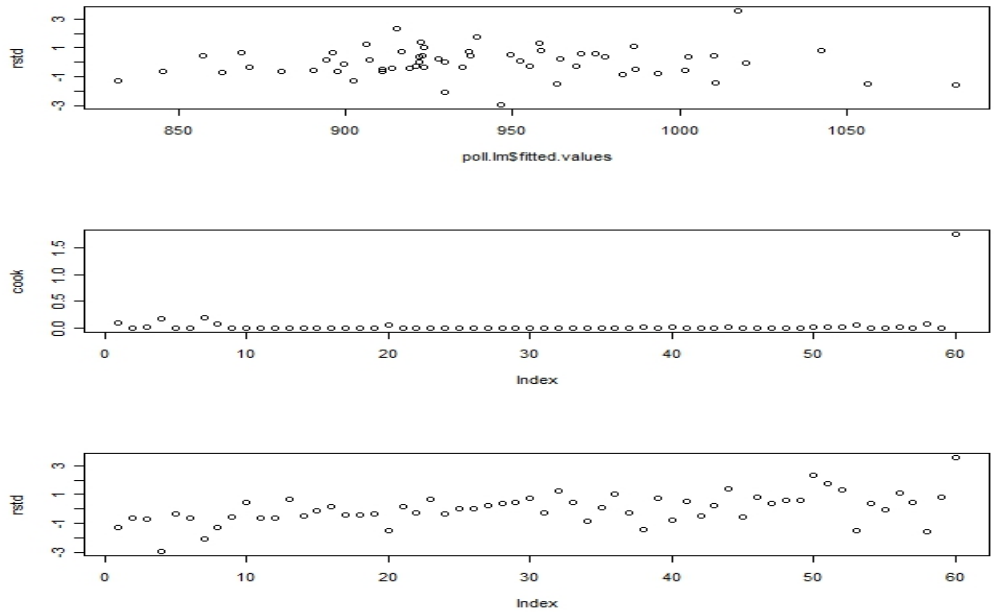
```
> pollution=read.table("pollution.csv",header=T,sep=",")
> city=pollution$CITY
> mort=pollution$MORT
> precip=pollution$PRECIP
> educ=pollution$EDUC
> nonwhite=pollution$NONWHITE
> NOX=pollution$NOX
> SO2=pollution$SO2
> par(mfrow=c(2,1))
> pairs(cbind(mort,precip,educ,nonwhite,NOX,SO2))
> pairs(cbind(mort,precip,educ,nonwhite,log(NOX),log(SO2)))
```





These plots suggest that taking the log transformation of the two pollutant variables is appropriate.

```
> poll.lm=lm(mort~precip+educ+nonwhite+log(NOX)+log(SO2))
> cook<-cooks.distance(poll.lm)
> rstd<-rstandard(poll.lm)
> par(mfrow=c(3,1))
> plot(poll.lm$fitted.values,rstd)
> plot(cook)
> plot(rstd)
```



These plots clearly indicate that point 60 is a problem (the large residual also corresponds to point 60). We will refit the regression with this point removed.

```
>poll.lm=lm(mort[-60]~precip[-60]+educ[-60]+nonwhite[-60]+log(NOX[-60])+log(SO2[-
60]))
> summary(poll.lm)

Call:
lm(formula = mort[-60] ~ precip[-60] + educ[-60] + nonwhite[-60] +
    log(NOX[-60]) + log(SO2[-60]))

Residuals:
    Min      1Q  Median      3Q     Max
-92.886 -20.610  -1.425  20.586  76.900
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   852.3761    85.9328   9.919 1.12e-13 ***
precip[-60]     1.3633     0.6357   2.145   0.0366 *
educ[-60]      -5.6669     6.5238  -0.869   0.3889
nonwhite[-60]   3.0397     0.5906   5.147 3.95e-06 ***
log(NOX[-60])  -9.8984     7.7306  -1.280   0.2060
log(SO2[-60])  26.0326     5.9311   4.389 5.46e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 32.07 on 53 degrees of freedom
Multiple R-squared: 0.7247,     Adjusted R-squared: 0.6987
F-statistic:  27.9 on 5 and 53 DF,  p-value: 9.928e-14
```

To test whether SO_2 and NOX are important we can use an F-test.

```
> poll.r<-lm(mort[-60]~precip[-60]+educ[-60]+nonwhite[-60])
> anova(poll.r,poll.lm,test='F') #the partial F test anova
Analysis of Variance Table

Model 1: mort[-60] ~ precip[-60] + educ[-60] + nonwhite[-60]
Model 2: mort[-60] ~ precip[-60] + educ[-60] + nonwhite[-60] + logNox[-60] +
    logSo2[-60]
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1     55 93695
2     53 54501  2     39193 19.057 5.812e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The very small p-value suggests that at least one of SO_2 and NOX is important. The small t-value of -1.28 for NOX suggests that only SO_2 is needed in the model.

b) Find a 95% confidence interval for the mean mortality for a city with the following characteristics precipitation=30, education=10, nonwhite=20, NO_x=1, and SO_2=1.

```
> x1<-precip[-60]
> x2<-educ[-60]
> x3<-nonwhite[-60]
> x4<-log(NOX[-60])
> x5<-log(SO2[-60])
> y<-mort[-60]
> fit<-lm(y~x1+x2+x3+x4+x5)
> xnew<-data.frame(x1=3,x2=10,x3=20,x4=0,x5=0)
> predict(fit,xnew,interval="confidence")
       fit      lwr      upr
1 897.4001 858.8557 935.9445
```

The confidence interval runs from 858.8557 to 935.9445.

c) Conduct an appropriate hypothesis test(s) to determine whether the variables education and nonwhite provide important information about mortality, over and above what is explained by precipitation, NO_x and SO_2.

To answer this question formally we need to perform an F-test of the null: $\beta_2 = \beta_3 = 0$.

```
> fitnew<-lm(y~x1+x4+x5+x2+x3)
> fitnewr<-lm(y~x1+x4+x5)
> anova(fitnewr,fitnew,test='F') #the partial F test anova
Analysis of Variance Table

Model 1: y ~ x1 + x4 + x5
Model 2: y ~ x1 + x4 + x5 + x2 + x3
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
```

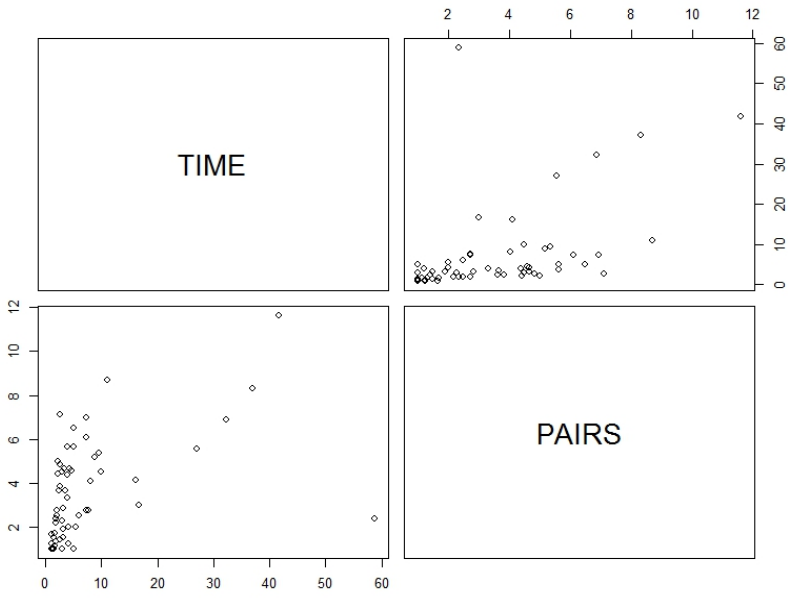
```
1     55 82737
2     53 54501  2     28236 13.729 1.569e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

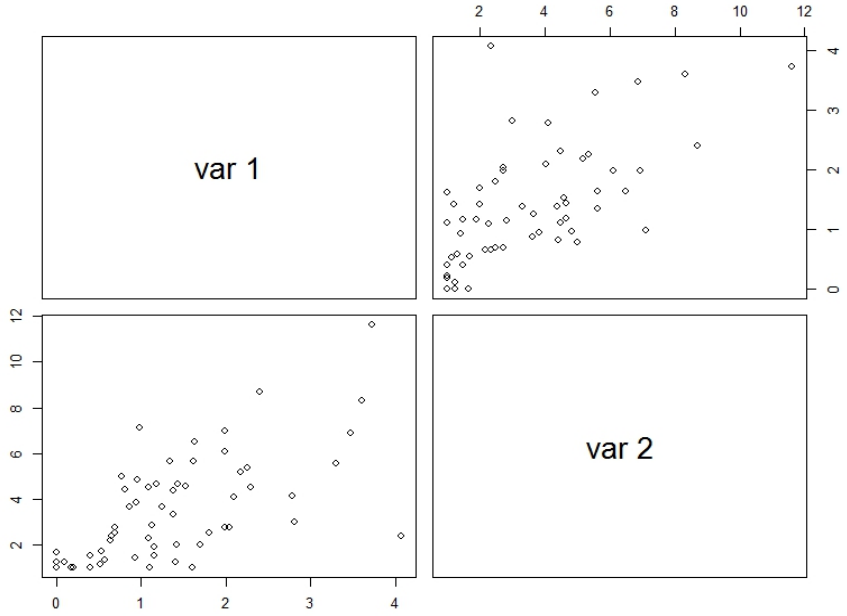
The p-value for our F-test is <0.05. This means we reject the null and conclude education and nonwhite provide important information about the age-adjusted mortality rate. Looking at the t-value for education we would likely conclude that it is only non-white that is important.

QUESTION 2 (revised based on the exercise in Chapter 10 from "The Statistical Sleuth")

The file "birds" contains measurements on breeding pairs of land bird species collected from 16 islands around Britain over the course of several decades. For each species, the data set contains average time of extinction on those island where it appeared; the average number of nesting pairs; the size of the species (large or small); and the migratory status of the species (migrant or resident). It is expected that species with larger numbers of nesting pairs will tend to remain longer before becoming extinct. Of interest is wether, after adjusting for number of nesting pairs, size or migratory status has any effect. There is also some interest in whether the effect of size differs depending on the number of nesting pairs. Analyse the data to answer these two questions of interest. Your answer must include appropriate documentation.
[data from: S.L. Pitman, H.L. Jones, and J. Diamond, "On the risk of extinction", American Naturalist 132 (1988): 757-85).

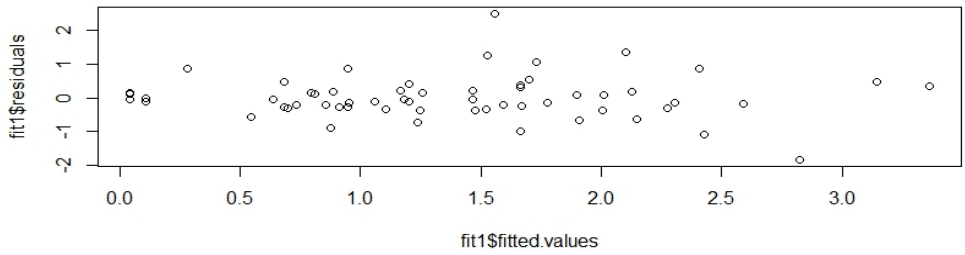
```
> birds<-read.table("birds.csv",header=T,sep=",")
> pairs(birds[,2:3])
> pairs(cbind(log(birds[,2]),birds[,3]))
```

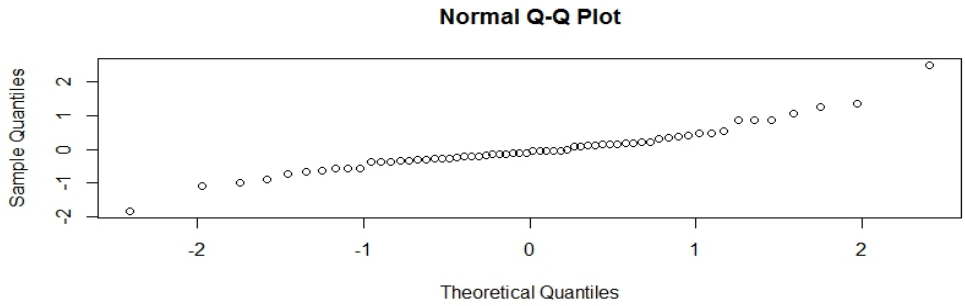




These plots suggest that taking the log transformation of the response is reasonable. Other transformations were also possible. To answer the first part of the question we need to fit the following model:

```
> fit1<-lm(log(birds$TIME)~birds$PAIRS+as.factor(birds$SIZE)+as.factor(birds$STATUS))
> summary(fit1)

Call:
lm(formula = log(birds$TIME) ~ birds$PAIRS + as.factor(birds$SIZE) +
    as.factor(birds$STATUS))

Residuals:
     Min       1Q   Median       3Q      Max
-1.83997 -0.29458 -0.07187  0.21712  2.51691

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              0.43056    0.20706   2.079 0.042011 *
birds$PAIRS              0.26509    0.03679   7.206 1.32e-09 ***
as.factor(birds$SIZE)S  -0.65237    0.16665  -3.915 0.000241 ***
as.factor(birds$STATUS)R  0.50406    0.18261   2.760 0.007717 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6523 on 58 degrees of freedom
Multiple R-squared: 0.5984,     Adjusted R-squared: 0.5776
F-statistic: 28.81 on 3 and 58 DF,  p-value: 1.557e-11

> par(mfrow=c(2,1))
> plot(fit1$fitted.values,fit1$residuals)
> qqnorm(fit1$residuals)
```





Based on these two plots the fitted model appears adequate. To answer the question of interest we need to test the null hypothesis that $\beta_2 = \beta_3 = 0$. This test can be performed using an F-test. The test statistic can be computed as follows:

```
> fitr<-lm(log(birds$TIME)~birds$PAIRS)
> anova(fitr,fit1,test='F') #the partial F test anova
Analysis of Variance Table

Model 1: log(birds$TIME) ~ birds$PAIRS
Model 2: log(birds$TIME) ~ birds$PAIRS + as.factor(birds$SIZE) + as.factor(birds$STATUS)
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     60 34.795
2     58 24.682  2    10.113 11.883 4.73e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This suggests that at least one of size or status is an important variable. Further to this test, the fact that the individual p-values for each variable are less than 0.05 suggests that both variables are needed in the model.

To answer the second part of the question we need to include an interaction term between pairs and size. This can be done in the following way:

```
> #note need to create and indicator for SIZE here to compute the product term.
> size<-ifelse(birds$SIZE=="S",1,0)
> fit2<-lm(log(birds$TIME)~birds$PAIRS+size+as.factor(birds$STATUS)+I(birds$PAIRS*size))
> summary(fit2)

Call:
lm(formula = log(birds$TIME) ~ birds$PAIRS + size + as.factor(birds$STATUS) +
    I(birds$PAIRS * size))

Residuals:
     Min       1Q   Median       3Q      Max
-1.59931 -0.37147 -0.06669  0.22977  2.45041

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            0.67184    0.27764   2.420  0.01874 *
birds$PAIRS            0.20069    0.06172   3.252  0.00193 **
```

```
size                      -0.99378    0.31135  -3.192  0.00230 **
as.factor(birds$STATUS)R   0.48060    0.18245   2.634  0.01084 *
I(birds$PAIRS * size)      0.09892    0.07638   1.295  0.20048
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6486 on 57 degrees of freedom
Multiple R-squared: 0.6099,     Adjusted R-squared: 0.5825
F-statistic: 22.28 on 4 and 57 DF,  p-value: 4.105e-11
```

The p-value for the interaction term is 0.20, suggesting that it is not needed in our model. Based on the data we observed, there is no evidence to suggest that the impact of size depends on the number of nesting pairs.

QUESTION 3 (revised based on the exercise in Chapter 12 from class text "The Statistical Sleuth")

Blood-Brain Barrier. Please use "install.packages('Sleuth3')" and "library(Sleuth3)" to call the dataset in this problem. Using the data stored in the object "case1102" of the R library "Sleuth3", perform the following variable-selection techniques to find a subset of the covariates-days after inoculation (Days), tumor weight (Tumor), weight loss (Loss), initial weight (Weight), and sex (Sex)-for explaining log of the ratio of brain tumor antibody count (Brain) to liver antibody count (Liver). (a) Forward selection by using F-statistic; (b) backward elimination by using F-statistic; (c) stepwise selection by using F-statistic. Are the above results the same?

```
> library(Sleuth3)
> head(case1102)
   Brain   Liver Time Treatment Days    Sex Weight Loss Tumor
1  41081 1456164  0.5        BD   10 Female    239  5.9   221
2  44286 1602171  0.5        BD   10 Female    225  4.0   246
3 102926 1601936  0.5        BD   10 Female    224 -4.9    61
4  25927 1776411  0.5        BD   10 Female    184  9.8   168
5  42643 1351184  0.5        BD   10 Female    250  6.0   164
6  31342 1790863  0.5        NS   10 Female    196  7.7   260
> attach(case1102)
> IndSex=ifelse(Sex=='Female',1,0)
> Variable=c("Days","Tumor","Loss", "Weight", "IndSex")
> p=length(Variable)
> X=data.frame(Days,Tumor,Loss, Weight, IndSex)
> Y=log(Brain/Liver)
> detach(case1102)
>
> library(wle) #need to load this library!
> X=as.matrix(X)
>
> #(a)
> mle.stepwise(Y~X,f.in=4,f.out=4,type="Forward")

Call:
mle.stepwise(formula = Y ~ X, type = "Forward", f.in = 4, f.out = 4)


 Forward  selection procedure

F.in:  4

Last  2  iterations:
     (Intercept) XDays XTumor XLoss XWeight XIndSex
[1,]           1     0      0     0       0       1 14.67
[2,]           1     1      0     0       0       1 10.56

> result=mle.stepwise(Y~X,f.in=4,f.out=4,type="Forward")
> result=as.vector(result$step[length(result$step[,1]),2:(p+1)])
> #Selected Variables
```

```
> Variable[as.logical(result)]
[1] "Days"   "IndSex"
```

Variables selected are "Days" , "IndSex".

```
> #(b)
> mle.stepwise(Y~X,f.in=4,f.out=4,type="Backward")

Call:
mle.stepwise(formula = Y ~ X, type = "Backward", f.in = 4, f.out = 4)


 Backward  selection procedure

F.out:  4

Last  3  iterations:
     (Intercept) XDays XTumor XLoss XWeight XIndSex
[1,]           1     1      1     0       1       1 0.000835
[2,]           1     1      0     0       1       1 0.005907
[3,]           1     1      0     0       0       1 1.525000

> result=mle.stepwise(Y~X,f.in=4,f.out=4,type="Backward")
> result=as.vector(result$step[length(result$step[,1]),2:(p+1)])
> #Selected Variables
> Variable[as.logical(result)]
[1] "Days"   "IndSex"
```

Variables selected are "Days" , "IndSex".

```
> #(c)
> mle.stepwise(Y~X,f.in=4,f.out=4,type="Stepwise")

Call:
mle.stepwise(formula = Y ~ X, type = "Stepwise", f.in = 4, f.out = 4)


 Stepwise  selection procedure

F.in:  4
F.out:  4

Last  2  iterations:
     (Intercept) XDays XTumor XLoss XWeight XIndSex
[1,]           1     0      0     0       0       1 14.670
[2,]           0     0      0     0       0       1  1.616

> result=mle.stepwise(Y~X,f.in=4,f.out=4,type="Stepwise")
> result=as.vector(result$step[length(result$step[,1]),2:(p+1)])
> #Selected Variables
> Variable[as.logical(result)]
[1] "IndSex"
```

Variable selected is "IndSex". It is worth noting that the intercept is eliminated in this case. Hence, the results for the forward selection and backward elimination are the same, but they are different from the result for stepwise selection.