Métodos Automáticos de Selección de Variables

Podemos dividirlos entre aquellos procedimientos de búsqueda que escogen el mejor entre todos los modelos posibles y aquellos que eligen iterativamente, en forma automática.

Búsqueda de todos los subconjuntos posibles

Este método consiste en evaluar todos los modelos posibles que se pueden construir en un conjunto dado de variables independientes.

Es particularmente útil cuando el número de variables no es demasiado grande. En general, uno puede forzar la presencia de ciertas variables y eso reduce el tamaño de la búsqueda. Uno puede imponer el criterio de selección R^2, R^2_a y C_p.

Si bien el C_p parece el más razonable debemos tener en cuenta que asume que el modelo con todas las variables no tiene sesgo. Además, si bien se basa en los errores de predicción no tiene en cuenta que pasaría con futuras observaciones

En R contamos con Leaps.

Consideremos los datos de cemento.

Recordemos que la respuesta y (y.hald) es la temperatura de la mezcla de cemento y las 4 covariables (x.hald) son:
x1: tricalcium aluminate
x2: tricalcium silicate
x3: tetracalcium alumino ferrite
x4: dicalcium silicate.

Recordemos corr(x1,x3)= -0.824 y corr(x2,x4)= -0.975.

```
library(leaps)
library(wle)
data(hald)
hald
```

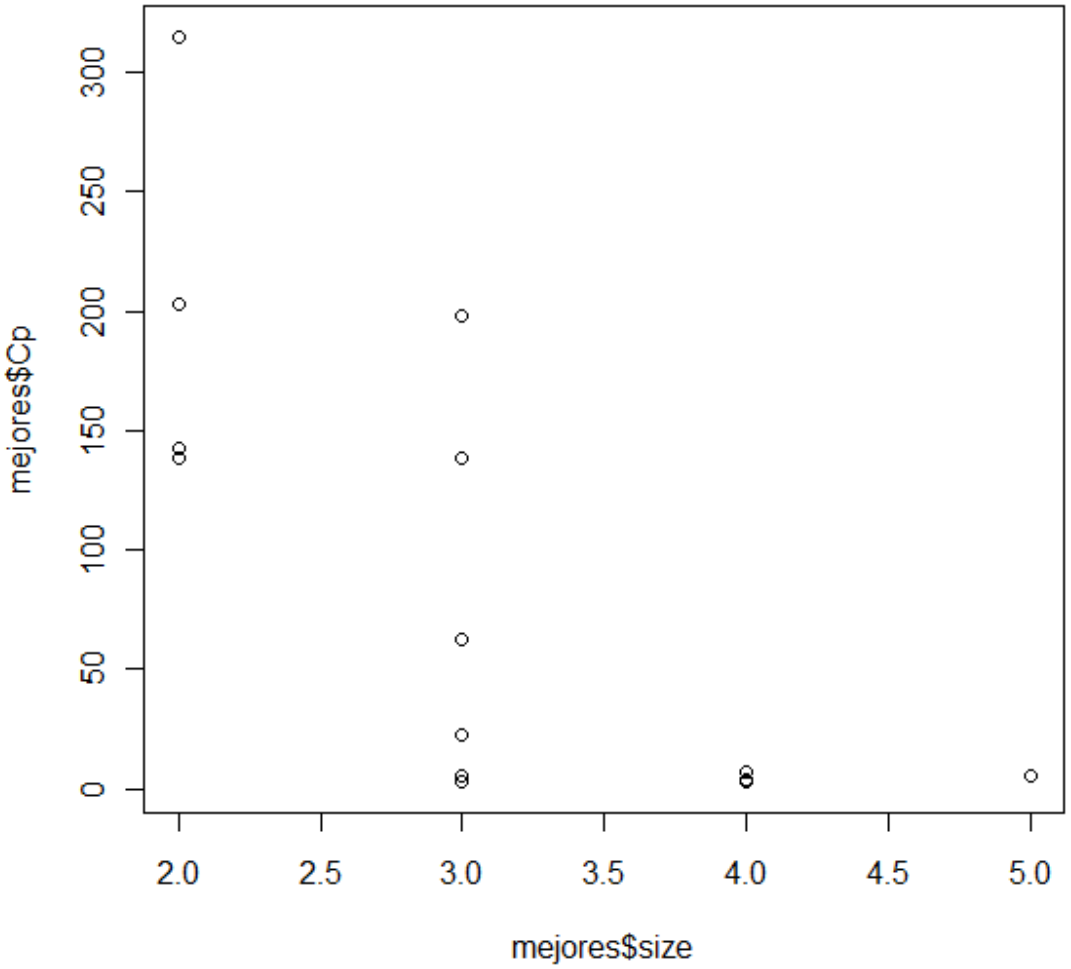
```
> cor(x.hald)
           [,1]       [,2]       [,3]       [,4]
[1,]  1.0000000  0.2285795 -0.8241338 -0.2454451
[2,]  0.2285795  1.0000000 -0.1392424 -0.9729550
[3,] -0.8241338 -0.1392424  1.0000000  0.0295370
[4,] -0.2454451 -0.9729550  0.0295370  1.0000000
```

```
        [,1] [,2] [,3] [,4] [,5]
 [1,]  78.5    7   26    6   60
 [2,]  74.3    1   29   15   52
 [3,] 104.3   11   56    8   20
 [4,]  87.6   11   31    8   47
 [5,]  95.9    7   52    6   33
 [6,] 109.2   11   55    9   22
 [7,] 102.7    3   71   17    6
 [8,]  72.5    1   31   22   44
 [9,]  93.1    2   54   18   22
[10,] 115.9   21   47    4   26
[11,]  83.8    1   40   23   34
[12,] 113.3   11   66    9   12
[13,] 109.4   10   68    8   12
```

all-subsets regression

```
leaps(x=x.hald, y=y.hald, method=c("Cp", "adjr2", "r2"))
```

```
leaps(x=x.hald, y=y.hald, method=c("Cp", "adjr2", "r2"))
$which
      1     2     3     4
1 FALSE FALSE FALSE  TRUE
1 FALSE  TRUE FALSE FALSE
1  TRUE FALSE FALSE FALSE
1 FALSE FALSE  TRUE FALSE
2  TRUE  TRUE FALSE FALSE
2  TRUE FALSE FALSE  TRUE
2 FALSE FALSE  TRUE  TRUE
2 FALSE  TRUE  TRUE FALSE
2 FALSE  TRUE FALSE  TRUE
2  TRUE FALSE  TRUE FALSE
3  TRUE  TRUE FALSE  TRUE
3  TRUE  TRUE  TRUE FALSE
3  TRUE FALSE  TRUE  TRUE
3 FALSE  TRUE  TRUE  TRUE
4  TRUE  TRUE  TRUE  TRUE

$label
[1] "(Intercept)" "1"           "2"           "3"           "4"

$size
 [1] 2 2 2 2 3 3 3 3 3 3 4 4 4 4 5

$Cp
 [1] 138.730833 142.486407 202.548769 315.154284   2.678242   5.495851
 [7]  22.373112  62.437716 138.225920 198.094653   3.018233   3.041280
[13]   3.496824   7.337474   5.000000
```


mejores$Cp
0
50
100
150
200
250
300
2.0
2.5
3.0
3.5
4.0
4.5
5.0
mejores$size

```
> cbind(leap.cem$size,leap.cem$Cp)
```

```
      [,1]       [,2]
 [1,]    2 138.730833
 [2,]    2 142.486407
 [3,]    2 202.548769
 [4,]    2 315.154284
 [5,]    3   2.678242
 [6,]    3   5.495851
 [7,]    3  22.373112
 [8,]    3  62.437716
 [9,]    3 138.225920
[10,]    3 198.094653
[11,]    4   3.018233
[12,]    4   3.041280
[13,]    4   3.496824
[14,]    4   7.337474
[15,]    5   5.000000
```

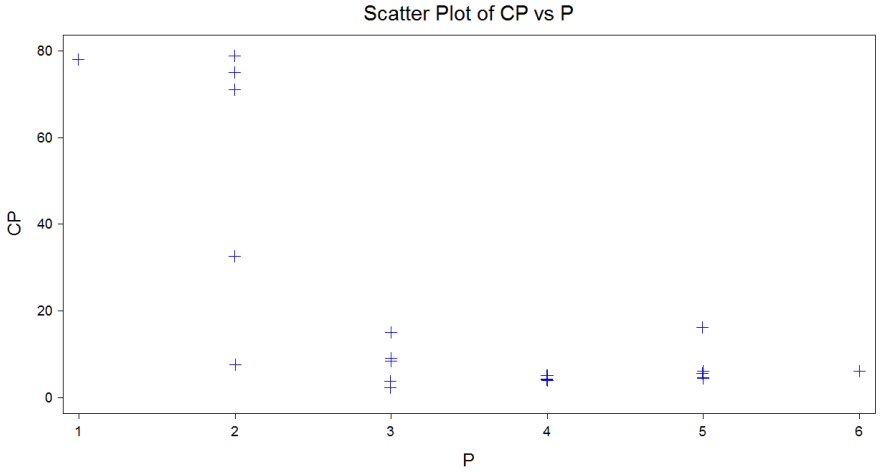
```
leaps(x=x.hald, y=y.hald, method=c("r2"))

$which
      1     2     3     4
1 FALSE FALSE FALSE  TRUE
1 FALSE  TRUE FALSE FALSE
1  TRUE FALSE FALSE FALSE
1 FALSE FALSE  TRUE FALSE
2  TRUE  TRUE FALSE FALSE
2  TRUE FALSE FALSE  TRUE
2 FALSE FALSE  TRUE  TRUE
2 FALSE  TRUE  TRUE FALSE
2 FALSE  TRUE FALSE  TRUE
2  TRUE FALSE  TRUE FALSE
3  TRUE  TRUE FALSE  TRUE
3  TRUE  TRUE  TRUE FALSE
3  TRUE FALSE  TRUE  TRUE
3 FALSE  TRUE  TRUE  TRUE
4  TRUE  TRUE  TRUE  TRUE

$label
[1] "(Intercept)" "1"           "2"           "3"           "4"

$size
 [1] 2 2 2 2 3 3 3 3 3 3 4 4 4 4 5

$r2
 [1] 0.6745420 0.6662683 0.5339480 0.2858727 0.9786784 0.9724710 0.9352896
 [8] 0.8470254 0.6800604 0.5481667 0.9823355 0.9822847 0.9812811 0.9728200
[15] 0.9823756
```

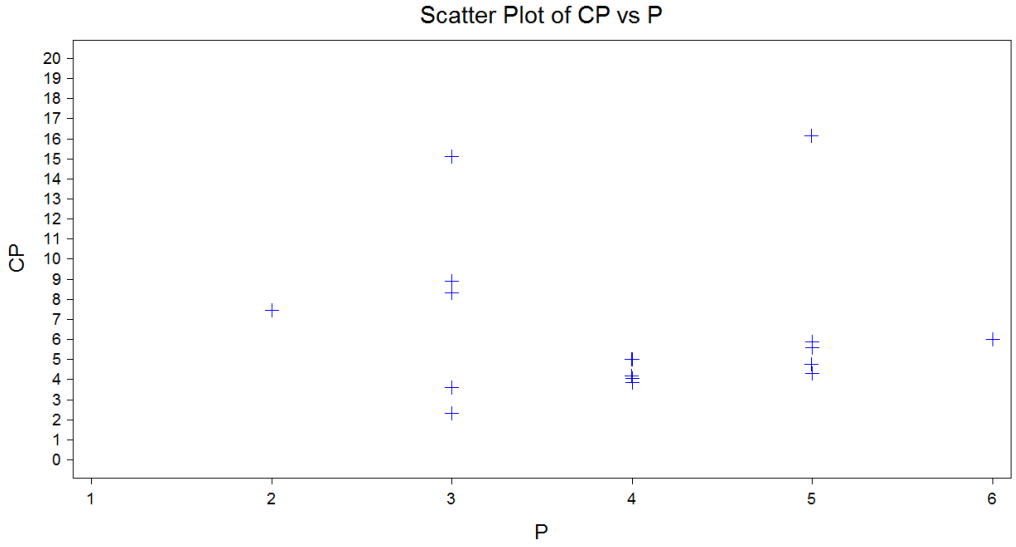
```
leaps(x=x.hald, y=y.hald, method=c("adjr2"))

$which
      1     2     3     4
1 FALSE FALSE FALSE  TRUE
1 FALSE  TRUE FALSE FALSE
1  TRUE FALSE FALSE FALSE
1 FALSE FALSE  TRUE FALSE
2  TRUE  TRUE FALSE FALSE
2  TRUE FALSE FALSE  TRUE
2 FALSE FALSE  TRUE  TRUE
2 FALSE  TRUE  TRUE FALSE
2 FALSE  TRUE FALSE  TRUE
2  TRUE FALSE  TRUE FALSE
3  TRUE  TRUE FALSE  TRUE
3  TRUE  TRUE  TRUE FALSE
3  TRUE FALSE  TRUE  TRUE
3 FALSE  TRUE  TRUE  TRUE
4  TRUE  TRUE  TRUE  TRUE

$label
[1] "(Intercept)" "1"           "2"           "3"           "4"

$size
 [1] 2 2 2 2 3 3 3 3 3 3 4 4 4 4 5

$adjr2
 [1] 0.6449549 0.6359290 0.4915797 0.2209521 0.9744140 0.9669653 0.9223476
 [8] 0.8164305 0.6160725 0.4578001 0.9764473 0.9763796 0.9750415 0.9637599
[15] 0.9735634
```

Datos de Biomasa

UNFORCED INDEPENDENT VARIABLES: (A)K (B)NA (C)PH (D)SAL (E)ZN

P	CP	ADJUSTED R SQUARE	R SQUARE	RESID SS	MODEL VARIABLES
1	77.9	0.0000	0.0000	1.917E+07	INTERCEPT ONLY
2	7.4	0.5900	0.5994	7680575	C
2	32.7	0.3757	0.3899	1.169E+07	E
2	70.9	0.0525	0.0740	1.775E+07	B
2	74.8	0.0198	0.0421	1.836E+07	A
2	78.6	-0.0124	0.0106	1.897E+07	D
3	2.3	0.6422	0.6584	6548174	B C
3	3.6	0.6308	0.6476	6755845	A C
3	8.3	0.5896	0.6083	7509642	C E
3	8.9	0.5845	0.6034	7603247	C D
3	15.1	0.5313	0.5526	8576766	D E
4	3.8	0.6378	0.6625	6471149	B C E
4	4.0	0.6355	0.6604	6511089	A B C
4	4.2	0.6341	0.6590	6536396	B C D
4	5.0	0.6268	0.6522	6667664	A C D
4	5.0	0.6267	0.6521	6669300	A C E
5	4.3	0.6424	0.6749	6232954	A C D E
5	4.7	0.6389	0.6718	6292475	B C D E
5	5.6	0.6306	0.6642	6438038	A B C E
5	5.9	0.6279	0.6617	6485307	A B C D
5	16.1	0.5351	0.5773	8102649	A B D E
6	6.0	0.6360	0.6773	6186048	A B C D E


Scatter Plot of CP vs P
CP
P
0
20
40
60
80
1
2
3
4
5
6


Scatter Plot of CP vs P
CP
P
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
1
2
3
4
5
6

Procedimientos Stepwise

Existen tradicionalmente tres versiones: Forward, Backward y la combinación de ambos que es la Stepwise.

Podríamos decir que hay tantas implementaciones de este método como programas, por lo que es necesario leer detalladamente la descripción del programa que estamos utilizando.

Describiremos la implementación de mle.stepwise de wle.

Forward:
Este procedimiento no incluye inicialmente ninguna covariable, salvo la intercept, y va agregando las variables una a una de acuerdo con la que tiene mayor F parcial en los sucesivos modelos evaluados y superior al valor F.in.

Backard:
Este procedimiento incluye inicialmente todas las covariables y las va eliminando de a una a medida que el valor del F parcial sea inferior al valor F.out.

Stepwise:
Es una combinación de los dos anteriores y tiene en cuenta tanto el valor F.in como el F.out.

Stepwise Regression: veamos un ejemplo de Forward

```
library(wle)
data(hald)
result <- mle.stepwise(y.hald~x.hald)
summary(result)

Forward  selection procedure

F.in:  4

Last  3  iterations:
     (Intercept) x.hald1 x.hald2 x.hald3 x.hald4
[1,]           1       0       0       0       1  22.800
[2,]           1       1       0       0       1 108.200
[3,]           1       1       1       0       1   5.026


> summary(lm(y.hald~x.hald[,1]))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  81.4793     4.9273   16.54 4.07e-09 ***
x.hald[, 1]   1.8687     0.5264    3.55  0.00455 **

Residual standard error: 10.73 on 11 degrees of freedom
Multiple R-squared: 0.5339,     Adjusted R-squared: 0.4916
F-statistic:  12.6 on 1 and 11 DF,  p-value: 0.004552
```

```
> summary(lm(y.hald~x.hald[,2]))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  57.4237     8.4906   6.763  3.1e-05 ***
x.hald[, 2]   0.7891     0.1684   4.686 0.000665 ***

Residual standard error: 9.077 on 11 degrees of freedom
Multiple R-squared: 0.6663,     Adjusted R-squared: 0.6359
F-statistic: 21.96 on 1 and 11 DF,  p-value: 0.0006648

> summary(lm(y.hald~x.hald[,3]))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 110.2027     7.9478  13.866  2.6e-08 ***
x.hald[, 3]  -1.2558     0.5984  -2.098   0.0598 .

Residual standard error: 13.28 on 11 degrees of freedom
Multiple R-squared: 0.2859,     Adjusted R-squared: 0.221
F-statistic: 4.403 on 1 and 11 DF,  p-value: 0.05976

> summary(lm(y.hald~x.hald[,4]))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 117.5679     5.2622  22.342 1.62e-10 ***
x.hald[, 4]  -0.7382     0.1546  -4.775 0.000576 ***

Residual standard error: 8.964 on 11 degrees of freedom
Multiple R-squared: 0.6745,     Adjusted R-squared: 0.645
F-statistic:  22.8 on 1 and 11 DF,  p-value: 0.0005762
```

```
salida.41<-lm(y.hald~ x.hald[,4]+x.hald[,1])
anova(salida.41)
Analysis of Variance Table

Response: y.hald
            Df  Sum Sq Mean Sq F value    Pr(>F)
x.hald[, 4]  1 1831.90 1831.90  245.03 2.319e-08 ***
x.hald[, 1]  1  809.10  809.10  108.22 1.105e-06 ***
Residuals   10   74.76    7.48

salida.43<-lm(y.hald~ x.hald[,4]+x.hald[,3])
anova(salida.43)
Analysis of Variance Table

Response: y.hald
            Df  Sum Sq Mean Sq F value    Pr(>F)
x.hald[, 4]  1 1831.90 1831.90 104.240 1.314e-06 ***
x.hald[, 3]  1  708.13  708.13  40.295 8.375e-05 ***
Residuals   10  175.74   17.57


salida.42<-lm(y.hald~ x.hald[,4]+x.hald[,2])
anova(salida.42)
Analysis of Variance Table

Response: y.hald
            Df  Sum Sq Mean Sq F value    Pr(>F)
x.hald[, 4]  1 1831.90 1831.90 21.0834 0.0009927 ***
x.hald[, 2]  1   14.99   14.99  0.1725 0.6866842
Residuals   10  868.88   86.89
```

```
salida.412<-lm(y.hald~ x.hald[,4]+x.hald[,1]++x.hald[,2])
anova(salida.412)
Response: y.hald
            Df  Sum Sq Mean Sq  F value    Pr(>F)
x.hald[, 4]  1 1831.90 1831.90 343.6758 1.771e-08 ***
x.hald[, 1]  1  809.10  809.10 151.7934 6.150e-07 ***
x.hald[, 2]  1   26.79   26.79   5.0259   0.05169 .
Residuals    9   47.97    5.33

salida.413<-lm(y.hald~ x.hald[,4]+x.hald[,1]++x.hald[,3])
anova(salida.413)
Response: y.hald
            Df  Sum Sq Mean Sq  F value    Pr(>F)
x.hald[, 4]  1 1831.90 1831.90 324.3179 2.285e-08 ***
x.hald[, 1]  1  809.10  809.10 143.2435 7.875e-07 ***
x.hald[, 3]  1   23.93   23.93   4.2358   0.06969 .
Residuals    9   50.84    5.65


> summary(lm(y.hald~ x.hald[,1]+ x.hald[,2]+x.hald[,4]))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  71.6483    14.1424   5.066 0.000675 ***
x.hald[, 1]   1.4519     0.1170  12.410 5.78e-07 ***
x.hald[, 2]   0.4161     0.1856   2.242 0.051687 .
x.hald[, 4]  -0.2365     0.1733  -1.365 0.205395

Residual standard error: 2.309 on 9 degrees of freedom
Multiple R-squared: 0.9823,     Adjusted R-squared: 0.9764
F-statistic: 166.8 on 3 and 9 DF,  p-value: 3.323e-08
```

```
resultb <- mle.stepwise(y.hald~x.hald,type="Backward")
summary(resultb)

Backward  selection procedure

F.out:  4

Last  2  iterations:
     (Intercept) x.hald1 x.hald2 x.hald3 x.hald4
[1,]           1       1       1       0       1 0.01823
[2,]           1       1       1       0       0 1.86300


summary(lm(y.hald~ x.hald[,1]+ x.hald[,2]+ x.hald[,3]+x.hald[,4]))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.4054    70.0710   0.891   0.3991
x.hald[, 1]   1.5511     0.7448   2.083   0.0708 .
x.hald[, 2]   0.5102     0.7238   0.705   0.5009
x.hald[, 3]   0.1019     0.7547   0.135   0.8959
x.hald[, 4]  -0.1441     0.7091  -0.203   0.8441

Residual standard error: 2.446 on 8 degrees of freedom
Multiple R-squared: 0.9824,     Adjusted R-squared: 0.9736
F-statistic: 111.5 on 4 and 8 DF,  p-value: 4.756e-07
```

```
anova(lm(y.hald~ x.hald[,1]+ x.hald[,2]+ x.hald[,4]))

Response: y.hald
            Df  Sum Sq Mean Sq  F value    Pr(>F)
x.hald[, 1]  1 1450.08 1450.08 272.0439 4.934e-08 ***
x.hald[, 2]  1 1207.78 1207.78 226.5879 1.094e-07 ***
x.hald[, 4]  1    9.93    9.93   1.8633    0.2054
Residuals    9   47.97    5.33

anova(lm(y.hald~ x.hald[,1]+ x.hald[,4]+ x.hald[,2]))

Response: y.hald
            Df  Sum Sq Mean Sq  F value    Pr(>F)
x.hald[, 1]  1 1450.08 1450.08 272.0439 4.934e-08 ***
x.hald[, 4]  1 1190.92 1190.92 223.4253 1.163e-07 ***
x.hald[, 2]  1   26.79   26.79   5.0259   0.05169 .
Residuals    9   47.97    5.33

anova(lm(y.hald~ x.hald[,2]+ x.hald[,4]+ x.hald[,1]))

Response: y.hald
            Df  Sum Sq Mean Sq F value    Pr(>F)
x.hald[, 2]  1 1809.43 1809.43 339.460 1.870e-08 ***
x.hald[, 4]  1   37.46   37.46   7.027   0.02644 *
x.hald[, 1]  1  820.91  820.91 154.008 5.781e-07 ***
Residuals    9   47.97    5.33
```

```
anova(lm(y.hald~ x.hald[,1]+ x.hald[,2]))

Response: y.hald
            Df Sum Sq Mean Sq F value    Pr(>F)
x.hald[, 1]  1 1450.1 1450.08  250.43 2.088e-08 ***
x.hald[, 2]  1 1207.8 1207.78  208.58 5.029e-08 ***
Residuals   10   57.9    5.79

anova(lm(y.hald~ x.hald[,2]+ x.hald[,1]))

Response: y.hald
            Df  Sum Sq Mean Sq F value    Pr(>F)
x.hald[, 2]  1 1809.43 1809.43  312.48 7.149e-09 ***
x.hald[, 1]  1  848.43  848.43  146.52 2.692e-07 ***
Residuals   10   57.90    5.79
```

```
results <- mle.stepwise(y.hald~x.hald,type="Stepwise")
summary(results)

mle.stepwise(formula = y.hald ~ x.hald, type = "Stepwise")


 Stepwise  selection procedure

F.in:  4
F.out:  4

Last  4  iterations:
     (Intercept) x.hald1 x.hald2 x.hald3 x.hald4
[1,]           1       0       0       0       1  22.800
[2,]           1       1       0       0       1 108.200
[3,]           1       1       1       0       1   5.026
[4,]           1       1       1       0       0   1.863
```

```
> summary(lm(y.hald~x.hald[,4]))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 117.5679     5.2622  22.342 1.62e-10 ***
x.hald[, 4]  -0.7382     0.1546  -4.775 0.000576 ***

Residual standard error: 8.964 on 11 degrees of freedom
Multiple R-squared: 0.6745,     Adjusted R-squared: 0.645
F-statistic:  22.8 on 1 and 11 DF,  p-value: 0.0005762

> summary(lm(y.hald~ x.hald[,1]+x.hald[,4]))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 103.09738    2.12398   48.54 3.32e-13 ***
x.hald[, 1]   1.43996    0.13842   10.40 1.11e-06 ***
x.hald[, 4]  -0.61395    0.04864  -12.62 1.81e-07 ***

Residual standard error: 2.734 on 10 degrees of freedom
Multiple R-squared: 0.9725,     Adjusted R-squared: 0.967
F-statistic: 176.6 on 2 and 10 DF,  p-value: 1.581e-08

> summary(lm(y.hald~ x.hald[,1]+ x.hald[,2]+x.hald[,4]))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  71.6483    14.1424   5.066 0.000675 ***
x.hald[, 1]   1.4519     0.1170  12.410 5.78e-07 ***
x.hald[, 2]   0.4161     0.1856   2.242 0.051687 .
x.hald[, 4]  -0.2365     0.1733  -1.365 0.205395

Residual standard error: 2.309 on 9 degrees of freedom
Multiple R-squared: 0.9823,     Adjusted R-squared: 0.9764
F-statistic: 166.8 on 3 and 9 DF,  p-value: 3.323e-08
```

```
Forward  selection procedure

F.in:  4

Last  3  iterations:
     (Intercept) x.hald1 x.hald2 x.hald3 x.hald4
[1,]           1       0       0       0       1  22.800
[2,]           1       1       0       0       1 108.200
[3,]           1       1       1       0       1   5.026

############################################################
Backward  selection procedure

F.out:  4

Last  2  iterations:
     (Intercept) x.hald1 x.hald2 x.hald3 x.hald4
[1,]           1       1       1       0       1 0.01823
[2,]           1       1       1       0       0 1.86300

############################################################
Stepwise  selection procedure

F.in:  4
F.out:  4

Last  4  iterations:
     (Intercept) x.hald1 x.hald2 x.hald3 x.hald4
[1,]           1       0       0       0       1  22.800
[2,]           1       1       0       0       1 108.200
[3,]           1       1       1       0       1   5.026
[4,]           1       1       1       0       0   1.863
```