STAT 210
Applied Statistics and Data Analysis
Multiple Linear Regression 3
Model Selection

Joaquin Ortega

Model Selection

Model Selection

(This section follows closely Ugarte, Militino and Arnholt, Probability and Statistics with R, Chapman and Hall, 2008)

In this section, we discuss several general methods for model selection.

These methods describe a series of steps that we do not always carry out in the same order.

It is essential to bear in mind that our objective is to increase our understanding of the data and the possible relations between the variables.

The experimenter must have a flexible attitude and be alert for unexpected structures in the data.

Model Selection

When building a model, we usually have a set of explanatory variables, and we seek to select the 'best' subset of regressors.

In this context, the guiding principle is parsimony- also known as Occam's razor. We seek the smallest set of regressors that reasonably explain the data.

Adding too many variables may produce an unnecessarily complicated model and increase the risk of having variables with similar information about the response variable, making the model unreliable.

We consider two approaches for selecting variables:

1. a stepwise testing strategy that compares successive models
2. a criterion approach that attempts to maximize some measure of goodness-of-fit.

Procedures based on testing

Backward Elimination

Backward elimination begins with a model containing all potential regressors and identifies the one with the largest p-value.

This can be done by looking at the p-values for the t tests of the $\hat{\beta}_i, i = 1, \ldots, p$ using the function `summary()` or using the p-values from the `R` function `drop1()`.

If the variable with the largest p-value is above a predetermined value, α_{crit}, that regressor is dropped.

A model with the remaining x-variables is then fitted, and the procedure continues until all the p-values for the remaining variables in the model are below the predetermined α_{crit}.

α_{crit} is sometimes referred to as the 'p-to-remove' and is typically set to 15 or 20%.

Forward Selection

Forward selection starts with no variables in the model and then adds the regressor that produces the smallest p-value below α_{crit} when included in the model.

This procedure is continued until no new predictors can be added.

The user can determine the variable that produces the smallest p-value by regressing the response variable on the x's one at a time using `lm()` and `summary()` or by using the function `add1()`.

Stepwise Regression

Stepwise regression is a combination of backward elimination and forward selection.

This technique allows variables that were either removed or added early to reenter or exit the model later in the process.

At each stage, a variable may be added or removed.

Testing based procedures

Testing-based procedures are relatively straightforward to implement; however, they do have some drawbacks.

One of the chief weaknesses of testing-based procedures is ending up with an overly parsimonious model.

When the analyst has a firm grasp of the subject matter, the analyst may want to include predictors that appear to have no statistical significance.

Although predictors can be added to a model developed from a testing-based perspective, the idea of adding predictors that are not necessarily significant conforms more to a criterion-based procedure.

Criterion-Based Procedures

Criterion-Based Procedures

There are several well-defined optimality criteria used in model building including

- R_a^2 (adjusted R^2),
- Mallows' C_p,
- Bayes Information Criterion (BIC),
- Akaike Information Criterion (AIC).

Adjusted R^2

R_a^2 is used instead of R^2 since R^2 will always increase with the addition of new variables to the model. Recall that

$$R_a^2 = 1 - \frac{SSE/(n-p)}{SST/(n-1)}.$$

Mallows C_p

Mallows C_p statistic is a measure of the total mean square error for the model. Consider a model with p parameters and define

$$\begin{aligned}\Gamma_p &= \frac{1}{\sigma^2}\sum_{i=1}^{n} E(\hat{y}_i - E(y_i))^2 \\ &= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}\left(E(y_i) - E(\hat{y}_i)\right)^2 + \sum_{i=1}^{n} Var(\hat{y}_i)\right) \\ &= \frac{1}{\sigma^2}\left((bias)^2 + variance\right).\end{aligned}$$

This is the mean square prediction error. It will not necessarily get smaller if more terms are added.

Mallows C_p

Using the MSE for the complete model to estimate σ^2, $\hat{\sigma}^2 = MSE$, we get as estimate of Γ_p

$$C_p = \frac{SSE(p)}{\hat{\sigma}^2} - n + 2p$$

which is the C_p statistic. If the model with p terms has a small bias, it can be shown that

$$E(C_p | null\ bias) \simeq p$$

When all p parameters are used in the model, $C_p = p$. A model with a bad fit will produce a C_p much bigger than p. Desirable models have small p and C_p less than or equal to p. It is common practice to plot C_p against p.

AIC and BIC

Recall that $\ln L(\beta, \sigma^2|\mathbf{X})$ is the log-likelihood function. The *BIC* for linear regression models is defined as

$$\begin{aligned} BIC &= -2\max(\ln L(\beta, \sigma^2|\mathbf{X})) + p\ln(n) \\ &= n\ln(SSE/n) + p\ln(n) + constant \end{aligned}$$

while the *AIC* for linear regression models is defined as

$$\begin{aligned} AIC &= -2\max(\ln L(\beta, \sigma^2|\mathbf{X})) + 2p \\ &= n\ln(SSE/n) + 2p + constant \end{aligned}$$

Since the constant is the same for a given data set and error distribution, it can be ignored when comparing models based on the same data. This is what the function `stepAIC()` does.

AIC and BIC

The goal when using *BIC* or *AIC* is to create a model that minimizes either *BIC* or *AIC*. Both *AIC* and *BIC* search for models that have small *SSE*.

However, *BIC* penalizes larger models more so than does *AIC* (assuming $n > e^2 = 7.39$). Consequently, *BIC* will favor smaller models than will *AIC*.

When building a model to be used for predictive purposes, *AIC* will generally be favored over *BIC*.

In `R`, the package `leaps` contains the function `regsubsets()`, which is very useful for computing R_a^2 and Mallows's C_p.

Example

Example

The data frame `HSwrestler` contains information on nine variables for a group of 78 high school wrestlers that was collected by the human performance lab at Appalachian State University. The variables are

- `AGE` (in years),
- `HT` (height in inches),
- `WT` (weight in pounds),
- `ABS` (abdominal skinfold measure),
- `TRICEPS` (tricep skinfold measure),
- `SUBSCAP` (subscapular skinfold measure),
- `HWFAT` (hydrostatic determination of fat),
- `TANFAT` (Tanita determination of fat), and
- `SKFAT` (skinfold determination of fat).

Example

In this example we want to create a model for predicting wrestlers' hydrostatic fat (`HWFAT`).

(a) Use backward elimination with the predictors `AGE`, `HT`, `WT`, `ABS`, `TRICEPS`, and `SUBSCAP` and an α_{crit} of 0.20.

(b) Use forward selection with an α_{crit} of 0.20.

(c) Use the function `regsubsets` in the `R` package `leaps` to select a model using R_a^2 as the criterion.

(d) Use the function `regsubsets` in the `R` package `leaps` to select a model using Mallows's C_p as the criterion.

(e) Use *AIC* as the criterion for selecting a model.

(f) Use *BIC* as the criterion for selecting a model.

Example: Backward elimination

(a) Backward elimination starts with all the variables in the model and eliminates variables with the largest (least significant) p-values:

```
library(PASWR)
attach(HSwrestler)
str(HSwrestler)
```

```
## 'data.frame':    78 obs. of  9 variables:
##  $ AGE    : int  18 15 17 17 17 14 14 17 15 14 ...
##  $ HT     : num  65.8 65.5 64 72 69.5 ...
##  $ WT     : num  134 129 121 145 299 ...
##  $ ABS    : num  8 10 6 11 54 40 6 11 9 19 ...
##  $ TRICEPS: num  6 8 6 10 42 25 8 7 6 13 ...
##  $ SUBSCAP: num  10.5 9 8 10 37 26 7 8 8 11.5 ...
##  $ HWFAT  : num  10.71 8.53 6.78 9.32 41.89 ...
##  $ TANFAT : num  11.9 10 8.3 8.2 41.6 29.9 12.4 11.1 10.1 15.5 ...
##  $ SKFAT  : num  9.8 10.56 8.43 11.77 41.09 ...
```

Example: Backward elimination

We will do the process showing all the steps.

```
reg.all <- lm(HWFAT ~ AGE + HT + WT + ABS + TRICEPS + SUBSCAP)
summary(reg.all)$coefficients
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 13.29369860 9.63026704  1.3804081 1.717917e-01
## AGE         -0.32893403 0.32157778 -1.0228755 3.098393e-01
## HT          -0.06730905 0.16050751 -0.4193514 6.762255e-01
## WT          -0.01365183 0.02590783 -0.5269385 5.998789e-01
## ABS          0.37141976 0.08836595  4.2032001 7.548985e-05
## TRICEPS      0.38742647 0.13761017  2.8153912 6.301113e-03
## SUBSCAP      0.11405213 0.14192779  0.8035927 4.243145e-01
```

Example: Backward elimination

Note that HT has the largest p-value of 6.762255e-01, so it is eliminated from the model:

```
reg.m1 <- lm(HWFAT ~ AGE + WT + ABS + TRICEPS + SUBSCAP)
summary(reg.m1)$coefficients
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  9.69230054 4.33250597  2.2371119 2.837559e-02
## AGE         -0.33352357 0.31954680 -1.0437393 3.000978e-01
## WT          -0.02084061 0.01931391 -1.0790465 2.841686e-01
## ABS          0.38259027 0.08377184  4.5670510 1.996022e-05
## TRICEPS      0.39737898 0.13477014  2.9485685 4.302189e-03
## SUBSCAP      0.11175170 0.14100772  0.7925218 4.306601e-01
```

Example: Backward elimination

Note that SUBSCAP has the largest p-value of 4.306601e-01, so it is eliminated from the model:

```
reg.m2 <- lm(HWFAT ~ AGE + WT + ABS + TRICEPS)
summary(reg.m2)$coefficients
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  9.94025577 4.31017288  2.3062313 2.393916e-02
## AGE         -0.38382444 0.31238134 -1.2287048 2.231289e-01
## WT          -0.01585418 0.01821376 -0.8704507 3.869075e-01
## ABS          0.39968360 0.08074124  4.9501789 4.621329e-06
## TRICEPS      0.46942072 0.09924414  4.7299591 1.068468e-05
```

Example: Backward elimination

Note that WT has the largest p-value of 3.869075e-01, so it is eliminated from the model:

```
reg.m3 <- lm(HWFAT ~ AGE + ABS + TRICEPS)
summary(reg.m3)$coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 10.6160623 4.23272425  2.508092 1.433001e-02
## AGE         -0.5330948 0.26067474 -2.045057 4.440545e-02
## ABS          0.3564311 0.06353588  5.609918 3.323075e-07
## TRICEPS      0.4656071 0.09898493  4.703819 1.158514e-05
```

Example: Backward elimination

The remaining p-values for `AGE`, `ABS`, and `TRICEPS` are all less than 0.20, so the model is composed of these three variables based on backward elimination.

Example: Backward elimination

Alternately, the function `drop1()` can be used in R:

```
drop1(lm(HWFAT ~ AGE + HT +WT + ABS +TRICEPS +SUBSCAP), test="F")
```

```
## Single term deletions
##
## Model:
## HWFAT ~ AGE + HT + WT + ABS + TRICEPS + SUBSCAP
##         Df Sum of Sq    RSS    AIC F value    Pr(>F)
## <none>                651.05 179.51
## AGE      1     9.594 660.64 178.65  1.0463  0.309839
## HT       1     1.613 652.66 177.70  0.1759  0.676225
## WT       1     2.546 653.60 177.81  0.2777  0.599879
## ABS      1   162.000 813.05 194.84 17.6669 7.549e-05 ***
## TRICEPS  1    72.683 723.73 185.76  7.9264  0.006301 **
## SUBSCAP  1     5.921 656.97 178.21  0.6458  0.424315
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: Backward elimination

Note that HT has the largest p-value of 0.676225, so it is eliminated from the model:

```
drop1(lm(HWFAT ~ AGE + WT + ABS +TRICEPS +SUBSCAP), test="F")
```

```
## Single term deletions
##
## Model:
## HWFAT ~ AGE + WT + ABS + TRICEPS + SUBSCAP
##         Df Sum of Sq    RSS    AIC F value    Pr(>F)
## <none>                652.66 177.70
## AGE      1     9.875 662.54 176.87  1.0894  0.300098
## WT       1    10.554 663.22 176.95  1.1643  0.284169
## ABS      1   189.072 841.73 195.54 20.8580 1.996e-05 ***
## TRICEPS  1    78.809 731.47 184.59  8.6941  0.004302 **
## SUBSCAP  1     5.693 658.36 176.38  0.6281  0.430660
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: Backward elimination

Note that SUBSCAP has the largest p-value of 0.430660, so it is eliminated from the model:

```
drop1(lm(HWFAT ~ AGE + WT + ABS +TRICEPS), test="F")
```

```
## Single term deletions
##
## Model:
## HWFAT ~ AGE + WT + ABS + TRICEPS
##         Df Sum of Sq    RSS    AIC F value    Pr(>F)
## <none>                658.36 176.38
## AGE      1    13.615 671.97 175.97  1.5097    0.2231
## WT       1     6.833 665.19 175.18  0.7577    0.3869
## ABS      1   220.994 879.35 196.95 24.5043 4.621e-06 ***
## TRICEPS  1   201.768 860.12 195.23 22.3725 1.068e-05 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: Backward elimination

Note that WT has the largest p-value of 0.3869, so it is eliminated from the model:

```
drop1(lm(HWFAT ~ AGE + ABS +TRICEPS), test="F")
```

```
## Single term deletions
##
## Model:
## HWFAT ~ AGE + ABS + TRICEPS
##         Df Sum of Sq    RSS    AIC F value    Pr(>F)
## <none>                665.19 175.18
## AGE      1    37.595 702.78 177.47  4.1823   0.04441 *
## ABS      1   282.896 948.08 200.82 31.4712 3.323e-07 ***
## TRICEPS  1   198.891 864.08 193.59 22.1259 1.159e-05 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: Backward elimination

The resulting model uses `AGE`, `ABS`, and `TRICEPS` to predict `HWFAT`.

Example: Forward selection

Forward selection assumes a model with an intercept only and adds the most significant (smallest p-values) variables one at a time.

The function `add1()` in `R` is used as the p-values at each stage are shown:

```
add1(lm(HWFAT~1), scope=(~.+ AGE + HT + WT + ABS +
                           TRICEPS + SUBSCAP), test="F")
```

```
## Single term additions
##
## Model:
## HWFAT ~ 1
##           Df Sum of Sq    RSS    AIC  F value    Pr(>F)
## <none>                 6017.8 340.97
## AGE        1     175.0 5842.8 340.67   2.2765    0.1355
## HT         1     117.8 5900.0 341.43   1.5175    0.2218
## WT         1    3237.6 2780.2 282.74  88.5045 2.219e-14 ***
## ABS        1    5072.8  945.0 198.57 407.9929 < 2.2e-16 ***
## TRICEPS    1    5056.3  961.5 199.92 399.6462 < 2.2e-16 ***
## SUBSCAP    1    4939.0 1078.8 208.90 347.9456 < 2.2e-16 ***
## ---
```

Example: Forward selection

The variable ABS has the most significant (smallest) p-value = 2.2e-16 with the largest F value = 407.9929, so it is added to the model:

```
add1(lm(HWFAT~ABS), scope=(~.+AGE +HT +WT +TRICEPS +SUBSCAP), test="F")
```

```
## Single term additions
##
## Model:
## HWFAT ~ ABS
##         Df Sum of Sq    RSS    AIC F value    Pr(>F)
## <none>                944.96 198.57
## AGE      1    80.876 864.08 193.59  7.0199 0.0098255 **
## HT       1    61.598 883.36 195.31  5.2298 0.0250250 *
## WT       1    43.734 901.22 196.87  3.6396 0.0602498 .
## TRICEPS  1   242.173 702.78 177.47 25.8443 2.639e-06 ***
## SUBSCAP  1   132.580 812.38 188.77 12.2400 0.0007904 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: Forward selection

The variable TRICEPS has the most significant (smallest) p-value = 2.639e-06 with the largest F value = 25.8443, so it is added to the model:

```
add1(lm(HWFAT~ABS+TRICEPS),scope=(~.+ AGE + HT + WT + SUBSCAP), test="F
```

```
## Single term additions
##
## Model:
## HWFAT ~ ABS + TRICEPS
##         Df Sum of Sq    RSS    AIC F value  Pr(>F)
## <none>                702.78 177.47
## AGE      1    37.595 665.19 175.18  4.1823 0.04441 *
## HT       1    25.246 677.54 176.62  2.7574 0.10104
## WT       1    30.812 671.97 175.97  3.3932 0.06947 .
## SUBSCAP  1     2.244 700.54 179.22  0.2370 0.62782
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: Forward selection

The variable AGE has the most significant (smallest) p-value = 0.04441 with the largest F value= 4.1823, so it is added to the model:

```
add1(lm(HWFAT~ABS+TRICEPS+AGE), scope=(~.+ HT + WT + SUBSCAP), test="F"
```

```
## Single term additions
##
## Model:
## HWFAT ~ ABS + TRICEPS + AGE
##         Df Sum of Sq    RSS    AIC F value Pr(>F)
## <none>                665.19 175.18
## HT       1    7.0291 658.16 176.35  0.7796 0.3802
## WT       1    6.8332 658.36 176.38  0.7577 0.3869
## SUBSCAP  1    1.9723 663.22 176.95  0.2171 0.6427
```

Example: Forward selection

None of the p-values now meet the α_{crit} level of 0.20, so the model is complete with ABS, TRICEPS, and AGE being used to predict HWFAT.

If a summary is done for the models where ABS, TRICEPS, and AGE are already in the model and HT, WT, or SUBSCAP were added individually, the p-values would match the last column of the last add1() output:

```
summary(lm(HWFAT~ABS+TRICEPS+AGE+HT))$coefficients
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 15.6108355 7.06886737  2.2083928 3.035723e-02
## ABS          0.3701823 0.06550886  5.6508735 2.902965e-07
## TRICEPS      0.4554293 0.09980055  4.5633949 1.990682e-05
## AGE         -0.4236659 0.28898736 -1.4660361 1.469329e-01
## HT          -0.1020099 0.11553071 -0.8829675 3.801523e-01
```

Example: Forward selection

The p-value for HT is 3.801523e-01=0.3802 from the add1() output:

```
summary(lm(HWFAT~ABS+TRICEPS+AGE+WT))$coefficients
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  9.94025577 4.31017288  2.3062313 2.393916e-02
## ABS          0.39968360 0.08074124  4.9501789 4.621329e-06
## TRICEPS      0.46942072 0.09924414  4.7299591 1.068468e-05
## AGE         -0.38382444 0.31238134 -1.2287048 2.231289e-01
## WT          -0.01585418 0.01821376 -0.8704507 3.869075e-01
```

Example: Forward selection

The p-value for WT is 3.869075e-01=0.3869 from the add1() output:

```
summary(lm(HWFAT~ABS+TRICEPS+AGE+SUBSCAP))$coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 10.59636278 4.25550451  2.490037 1.504963e-02
## ABS          0.33934952 0.07364815  4.607713 1.688668e-05
## TRICEPS      0.42485168 0.13249227  3.206615 1.994247e-03
## AGE         -0.53122920 0.26209533 -2.026855 4.633009e-02
## SUBSCAP      0.06218487 0.13346571  0.465924 6.426572e-01
```

Example: Forward selection

The p-value for `SUBSCAP` is `6.426572e-01`$=0.6427$ from the `add1()` output.

Note that the same model results in both the forward and backward selection procedures: (`HWFAT` $\sim$ `ABS + TRICEPS + AGE`). This is not always the case.

Example: Adjusted R^2

The R package `leaps` is needed for the function `regsubsets()`. The arguments have predictors as a matrix first, then the response as a vector. The first six variables of `HSwrestler` are the predictors, while the response, `HWFAT`, is in column 7.

```
str(HSwrestler[,-c(8,9)])
library(leaps)
a <- regsubsets(as.matrix(HSwrestler[,-c(7,8,9)]), HSwrestler[,7])
```

```
## 'data.frame':    78 obs. of  7 variables:
##  $ AGE    : int  18 15 17 17 17 14 14 17 15 14 ...
##  $ HT     : num  65.8 65.5 64 72 69.5 ...
##  $ WT     : num  134 129 121 145 299 ...
##  $ ABS    : num  8 10 6 11 54 40 6 11 9 19 ...
##  $ TRICEPS: num  6 8 6 10 42 25 8 7 6 13 ...
##  $ SUBSCAP: num  10.5 9 8 10 37 26 7 8 8 11.5 ...
##  $ HWFAT  : num  10.71 8.53 6.78 9.32 41.89 ...
```

Example: Adjusted R^2

```
summary(a)
```

```
## Subset selection object
## 6 Variables  (and intercept)
##         Forced in Forced out
## AGE         FALSE      FALSE
## HT          FALSE      FALSE
## WT          FALSE      FALSE
## ABS         FALSE      FALSE
## TRICEPS     FALSE      FALSE
## SUBSCAP     FALSE      FALSE
## 1 subsets of each size up to 6
## Selection Algorithm: exhaustive
##          AGE HT  WT  ABS TRICEPS SUBSCAP
## 1  ( 1 ) " " " " " " "*" " "     " "
## 2  ( 1 ) " " " " " " "*" "*"     " "
## 3  ( 1 ) "*" " " " " "*" "*"     " "
## 4  ( 1 ) "*" "*" " " "*" "*"     " "
## 5  ( 1 ) "*" " " "*" "*" "*"     "*"
## 6  ( 1 ) "*" "*" "*" "*" "*"     "*"
```

Example: Adjusted R^2

```
summary(a)$adjr2
max(summary(a)$adjr2)
which.max(summary(a)$adjr2)
```

```
## [1] 0.8409068 0.8801014 0.8849817 0.8846381 0.8840129
## [6] 0.8826699
## [1] 0.8849817
## [1] 3
```

The largest R_a^2 value is 0.8849817, which corresponds to the model with three predictors.

The row beside the 3 shows "*" symbols for `AGE`, `ABS`, and `TRICEPS`, so these are the appropriate predictor variables.

Example: Mallows's C_p

When using Mallows's C_p, the idea is to select the smallest C_p value less than or equal to p.

In this case, the `R` package leaps and the output from `regsubsets()` gives the optimal value $C_4 = 2.541953$, so the three-predictor (plus an intercept) model using `AGE`, `ABS`, and `TRICEPS` is again selected:

Example: Mallows's C_p

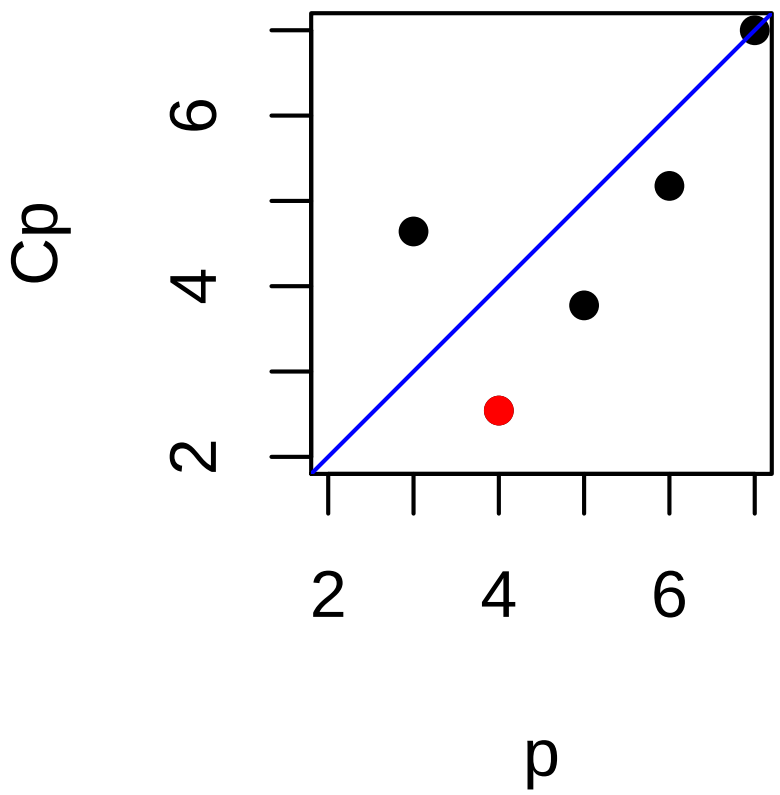
```
## Subset selection object
## 6 Variables  (and intercept)
##         Forced in Forced out
## AGE         FALSE      FALSE
## HT          FALSE      FALSE
## WT          FALSE      FALSE
## ABS         FALSE      FALSE
## TRICEPS     FALSE      FALSE
## SUBSCAP     FALSE      FALSE
## 1 subsets of each size up to 6
## Selection Algorithm: exhaustive
##          AGE HT  WT  ABS TRICEPS SUBSCAP
## 1  ( 1 ) " " " " " " "*" " "     " "
## 2  ( 1 ) " " " " " " "*" "*"     " "
## 3  ( 1 ) "*" " " " " "*" "*"     " "
## 4  ( 1 ) "*" "*" " " "*" "*"     " "
## 5  ( 1 ) "*" " " "*" "*" "*"     "*"
## 6  ( 1 ) "*" "*" "*" "*" "*"     "*"
```

```
summary(a)$cp
```

```
## [1] 29.051861  4.641808  2.541953  3.775400  5.175856
## [6]  7.000000
```

Example: Mallows's C_p

```
par(pty="s")
plot(2:7, summary(a)$cp, ylim=c(2,7), xlab="p",
     ylab="Cp", pch=16); abline(a=0, b=1, col='blue');
points(4,summary(a)$cp[3], col='red', pch=16)
```



Example: AIC

The function `stepAIC()` in the `MASS` package will compute models based on both `AIC` and `BIC` statistics.

The argument `k` of this function will be set equal to 2 for the AIC statistic and `ln(n)` for the BIC statistic.

The user needs to specify the scope of the model with the argument `scope=`. In this case, the scope of the model includes any of the six predictors `AGE, HT, WT, ABS, TRICEPS`, and `SUBSCAP`. For further details, see the `stepAIC()` help file.

The starting AIC value is 179.51. The `stepAIC()` function adds or removes variables until it finds the smallest AIC value.

A - before a variable indicates that the variable will be removed to produce the given AIC, while a '+ indicates the variable will be added to produce the given AIC.

Example: AIC

```
reg.all <- lm(HWFAT ~ AGE + HT + WT + ABS + TRICEPS + SUBSCAP)
mod.aic <- stepAIC(reg.all, direction="both",
                   scope=(~.+SUBSCAP+TRICEPS+ABS+WT+HT+AGE), k=2)
```

```
## Start:  AIC=179.51
## HWFAT ~ AGE + HT + WT + ABS + TRICEPS + SUBSCAP
##
##            Df Sum of Sq    RSS    AIC
## - HT        1     1.613 652.66 177.70
## - WT        1     2.546 653.60 177.81
## - SUBSCAP   1     5.921 656.97 178.21
## - AGE       1     9.594 660.64 178.65
## <none>                  651.05 179.51
## - TRICEPS   1    72.683 723.73 185.76
## - ABS       1   162.000 813.05 194.84
##
## Step:  AIC=177.7
## HWFAT ~ AGE + WT + ABS + TRICEPS + SUBSCAP
##
##            Df Sum of Sq    RSS    AIC
## - SUBSCAP   1     5.693 658.36 176.38
## - AGE       1     9.875 662.54 176.87
## - WT        1    10.554 663.22 176.95
```

Example: AIC

```
mod.aic
```

```
##
## Call:
## lm(formula = HWFAT ~ AGE + ABS + TRICEPS)
##
## Coefficients:
## (Intercept)          AGE          ABS      TRICEPS
##     10.6161      -0.5331       0.3564       0.4656
```

Example: AIC

The final model uses `AGE`, `ABS`, and `TRICEPS` as predictors.

Example: BIC

When BIC is the criterion, the model selected is `HWFAT ~ ABS + TRICEPS`.

```
mod.bic <- stepAIC(reg.all, direction="both",
                   scope=(~.+SUBSCAP+TRICEPS+ABS+WT+HT+AGE),
                   k=log(length(HWFAT)))
```

```
## Start:  AIC=196
## HWFAT ~ AGE + HT + WT + ABS + TRICEPS + SUBSCAP
##
##            Df Sum of Sq    RSS    AIC
## - HT        1     1.613 652.66 191.84
## - WT        1     2.546 653.60 191.95
## - SUBSCAP   1     5.921 656.97 192.35
## - AGE       1     9.594 660.64 192.79
## <none>                  651.05 196.00
## - TRICEPS   1    72.683 723.73 199.90
## - ABS       1   162.000 813.05 208.98
##
## Step:  AIC=191.84
## HWFAT ~ AGE + WT + ABS + TRICEPS + SUBSCAP
##
```

Example: BIC

```
mod.bic
```

```
##
## Call:
## lm(formula = HWFAT ~ ABS + TRICEPS)
##
## Coefficients:
## (Intercept)          ABS      TRICEPS
##      2.0590       0.3371        0.5043
```