Regression 9

Model Selection

1 Introduction

Given a possibly large set of potential predictors, which ones do we include in our model? Suppose $[X_1, X_2, \ldots]$ is a “pool” of potential predictors. The model with all predictors,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \epsilon,$$

is the most general model. It holds even if some of the individual β_j's are zero. But if some β_j's zero or close to zero, it is better to omit those X_j's from the model.

The following are the reasons why you should omit variables whose coefficients are close to zero:

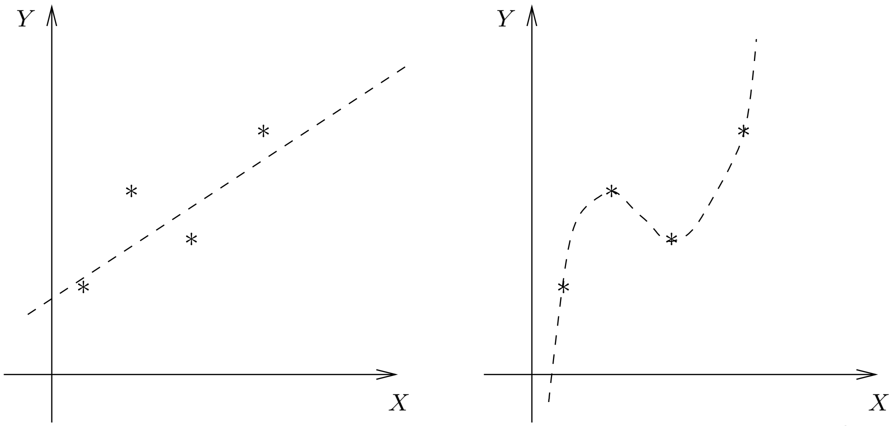
(a) Parsimony principle:

Given two models that perform equally well in terms of prediction, one should choose the model that is more parsimonious (simple).

(b) Prediction principle:

The model should give predictions that are as accurate as possible, *not* just for current observation, *but* for future observations as well. Including unnecessary predictors can apparently improve prediction for the current data, but can harm prediction for future data. Note that SSE never increases as we add more predictors.

It is better to be roughly right than precisely wrong
John Maynard Keynes, British economist (1883 - 1946)



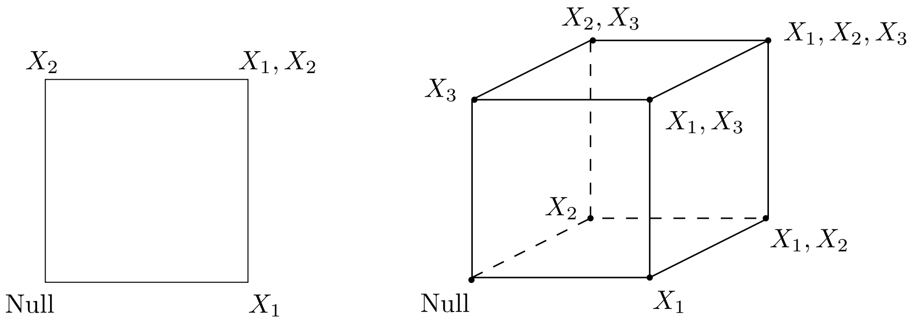
(i) $Y = \beta_0 + \beta_1 X + \epsilon$

(ii) $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$

Model (ii) in the figure gives perfect predictions for the current data ($R^2 = 1$ and $\hat{\epsilon}_i = 0$ for all i), but Model (i) will probably perform better for future data.

2 All possible regressions

The all-possible-regressions procedure calls for considering all possible subsets of the pool of potential predictors and identifying a few good subsets according to some criterion for detailed examination.



Suppose we have two potential predictors X_1 and X_2. This gives four possible models:

$$\underset{\text{Null}}{\binom{2}{0}} \quad + \quad \underset{1\ X}{\binom{2}{1}} \quad + \quad \underset{2\ X's}{\binom{2}{2}} \quad = \quad 4.$$

With $k = 3$ predictors, there are $2^3 = 8$ possible models:

$$\underset{\text{Null}}{\binom{3}{0}} \quad + \quad \underset{1\ X}{\binom{3}{1}} \quad + \quad \underset{2\ X's}{\binom{3}{2}} \quad + \quad \underset{3\ X's}{\binom{3}{3}} \quad = \quad 8.$$

They can be represented by vertices of a 3-dimensional cube. In general with k predictors, there are 2^k possible models. They can be represented by k-dimensional hyper-cube.

The purpose of all-possible-regressions approach is identifying a small group of regression models that are "good" according to a specified criterion (summary statistic) so that a detailed examination can be made of these models leading to the selection of the final regression model to be employed. The main problem of this approach is computationally too expensive. For example, with $k = 10$ predictors, we need to investigate $2^{10} = 1024$ potential regression models. With the aid of modern computing power, this computation is possible. But still the number of 1024 possible models to examine carefully would be an overwhelming task for a data analyst.

Different criteria for comparing the regression models may be used with the all-possible-regressions selection procedure. We discuss five summary statistics:

(i) R^2_p (or SSE_p).

(ii) $R^2_{\text{adj},p}$ (or MSE_p).

(iii) C_p.

(iv) AIC and BIC.

(v) PRESS_p.

We shall denote the number of *all* potential predictors in the pool by $P-1$. Hence including an intercept parameter β_0, we have P potential parameters. The number of predictors in a *subset* will be denoted by $p-1$, as always, so that there are p parameters in the regression function for this subset of predictors. Thus, we have

$$1 \leq p \leq P.$$

2.1 R_p^2 (or SSE_p)

R_p^2 indicates that there are p parameters (*or*, $p-1$ predictors) in the regression model. The coefficient of multiple determination R_p^2 is defined as

$$R_p^2 = 1 - \frac{\mathrm{SSE}_p}{\mathrm{SSTo}}.$$

- It measures the proportion of variance of Y explained by $p-1$ predictors.
- R_p^2 always goes up as we add a predictor. So, it is not appropriate to compare models with different sizes.
- R_p^2 varies inversely with SSE_p because SSTo is always constant for all possible regression models. That is, choosing the model with the largest R_p^2 is equivalent to choosing the model with smallest SSE_p.

2.2 $R_{\mathrm{adj},p}^2$ (or MSE_p)

One often considers models with a large R_p^2 value. However, R_p^2 always increases with the number of predictors. Hence, it can not be used to compare models with *different sizes*. The adjusted coefficient of multiple determination $R_{\mathrm{adj},p}^2$ has been suggested as an alternative criterion:

$$R_{\mathrm{adj},p}^2 = 1 - \frac{\mathrm{SSE}_p/(n-p)}{\mathrm{SSTo}/(n-1)} = 1 - \left(\frac{n-1}{n-p}\right)\frac{\mathrm{SSE}_p}{\mathrm{SSTo}} = 1 - \frac{\mathrm{MSE}_p}{\mathrm{SSTo}/(n-1)}.$$

- It is like R_p^2 but with a penalty for adding unnecessary variables. Thus, $R_{\mathrm{adj},p}^2$ can go down when a useless predictor is added. It can be even negative. If it is negative, `Minitab` sets $R_{\mathrm{adj},p}^2 = 0$.
- $R_{\mathrm{adj},p}^2$ varies inversely with MSE_p because $\mathrm{SSTo}/(n-1)$ is always constant for all possible regression models. That is, choosing the model with the largest $R_{\mathrm{adj},p}^2$ is equivalent to choosing the model with smallest MSE_p.
- R_p^2 is useful when comparing models of the *same size*, while $R_{\mathrm{adj},p}^2$ and C_p are used to compare models with *different sizes*.

2.3 Mallows C_p

The Mallows (1973) C_p is concerned with the *total mean squared error* of the n fitted values for each subset regression model. The mean squared error concept involves the total error in each fitted value:

$$\hat{Y}_i - \mu_i = \underbrace{\hat{Y}_i - E(\hat{Y}_i)}_{\text{random error}} + \underbrace{E(\hat{Y}_i) - \mu_i}_{\text{bias}},$$

where μ_i is the *true mean response* at the ith observation. The mean squared error for $\hat{Y}_i$ is defined as the *expected value* of the square of the total error in the above. It can be shown that

$$\begin{aligned}\text{MSE}(\hat{Y}_i) &= E\{(\hat{Y}_i - \mu_i)^2\} = E\{(\hat{Y}_i - E(\hat{Y}_i))^2\} + (E(\hat{Y}_i) - \mu_i)^2 \\ &= \text{Var}(\hat{Y}_i) + |\text{Bias}(\hat{Y}_i)|^2,\end{aligned}$$

where $\text{Bias}(\hat{Y}_i) = E(\hat{Y}_i) - \mu_i$. The total mean squared error for all n fitted values $\hat{Y}_i$ is the sum over the observation index i:

$$\sum_{i=1}^{n} \text{MSE}(\hat{Y}_i) = \sum_{i=1}^{n} \left\{ \text{Var}(\hat{Y}_i) + |\text{Bias}(\hat{Y}_i)|^2 \right\}.$$

For convenience, we define the row vectors in the data matrix $\mathbf{X}$ by $\mathbf{x}_i{}' = [1 \ X_{i1} \ X_{i2} \ \cdots \ X_{i,p-1}]$ so that we have

$$\underset{n\times p}{\mathbf{X}} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1{}' \\ \mathbf{x}_2{}' \\ \vdots \\ \mathbf{x}_n{}' \end{bmatrix}.$$

Using $\hat{Y}_i = \mathbf{x}_i{}'\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$, we have

$$\sum_{i=1}^{n} \text{Var}(\hat{Y}_i) = \sum_{i=1}^{n} \text{Var}(\mathbf{x}_i{}'\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \mathbf{x}_i{}' \,\text{Var}(\hat{\boldsymbol{\beta}})\mathbf{x}_i = \sigma^2 \sum_{i=1}^{n} \mathbf{x}_i{}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i.$$

The last term in the above equation can be manipulated as follows

$$\begin{aligned}\sum_{i=1}^{n} \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i &= \sum_{i=1}^{n} \mathrm{tr}\big\{\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\big\} = \sum_{i=1}^{n} \mathrm{tr}\big\{\mathbf{x}_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\big\} \\ &= \mathrm{tr}\Big[\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\Big] = \mathrm{tr}\Big[\Big\{\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'\Big\} \cdot (\mathbf{X}'\mathbf{X})^{-1}\Big] \\ &= \mathrm{tr}\big[\mathbf{I}_p\big] = p,\end{aligned}$$

where $\mathbf{I}_p$ is the $p \times p$ identity matrix. Using the above, we have

$$\sum_{i=1}^{n} \mathrm{Var}(\hat{Y}_i) = p\sigma^2.$$

It can also be shown that

$$\sum_{i=1}^{n} |\mathrm{Bias}(\hat{Y}_i)|^2 = (n-p)[E(S_p^2) - \sigma^2],$$

where S_p^2 is the MSE from the current model. Using this, we have

$$\sum_{i=1}^{n} \mathrm{MSE}(\hat{Y}_i) = p\sigma^2 + (n-p)[E(S_p^2) - \sigma^2]. \tag{9.1}$$

Dividing (9.1) by σ^2, we make it scale-free:

$$\sum_{i=1}^{n} \frac{\mathrm{MSE}(\hat{Y}_i)}{\sigma^2} = p + (n-p)\frac{E(S_p^2) - \sigma^2}{\sigma^2}.$$

If the model does not fit well, then S_p^2 is a biased estimate of σ^2. We can estimate $E(S_p^2)$ by MSE_p and estimate σ^2 by the MSE from the maximal model (the largest model we can consider), *i.e.*, $\hat{\sigma}^2 = \mathrm{MSE}_{P-1} = \mathrm{MSE}(X_1, \ldots, X_{P-1})$. Using the estimators for $E(S_p^2)$ and σ^2 gives

$$\begin{aligned}C_p &= p + (n-p)\Big[\frac{\mathrm{MSE}_p - \mathrm{MSE}(X_1, \ldots, X_{P-1})}{\mathrm{MSE}(X_1, \ldots, X_{P-1})}\Big] \\ &= \frac{\mathrm{SSE}_p}{\mathrm{MSE}(X_1, \ldots, X_{P-1})} - (n-2p).\end{aligned}$$

- Small C_p is a good thing. A small value of C_p indicates that the model is relatively precise (has small variance) in estimating the true regression coefficients and predicting future responses. This precision will not improve much by simply adding more predictors. Look for models with small C_p.

- If we have enough predictors in the regression model so that all the significant predictors are included, then $\text{MSE}_p \approx \text{MSE}(X_1, \ldots, X_{P-1})$ and it follows that $C_p \approx p$.
- Thus, C_p being close to p is evidence that the predictors in the pool of potential predictors $(X_1, \ldots, X_{P-1})$ but not in the current model, are not important.
- Models with considerable lack-of-fit have values of C_p larger than p.
- The C_p can be used to compare models with different sizes.
- If we use all the potential predictors $(p = P)$, then $C_p = P$.

2.4 AIC and BIC Criteria

The AIC (Akaike information criterion) and BIC (Bayesian information criterion) are statistical criteria for model selection. This BIC criterion is also called Schwarz' Bayesian criterion (SBC). These criteria are *originally* given by

$$\text{AIC}_p = -2\ln \hat{L} + 2 \cdot (p+1) \tag{9.2}$$

$$\text{BIC}_p = -2\ln \hat{L} + (\ln n) \cdot (p+1), \tag{9.3}$$

where $\hat{L}$ is the likelihood function with parameter estimation. In the likelihood function, the number of all the parameter estimates including $\hat{\sigma}^2$ is $(p+1)$. We search for models having smaller values of AIC or BIC.

Some further derivations result in the following formulas:

$$\begin{aligned}\text{AIC}_p &= -2\ln \hat{L} + 2 \cdot (p+1) \\ &= n\big[\ln(2\pi \text{SSE}_p/n) + 1\big] + 2(p+1) \\ &= \underbrace{n \ln \text{SSE}_p - n\ln n + 2 \cdot p}_{textbook} + n + n\ln(2\pi) + 2\end{aligned}$$

and

$$\begin{aligned}\text{BIC}_p &= -2\ln \hat{L} + (\ln n) \cdot (p+1) \\ &= n\big[\ln(2\pi \text{SSE}_p/n) + 1\big] + (\ln n)(p+1) \\ &= \underbrace{n \ln \text{SSE}_p - n\ln n + (\ln n) \cdot p}_{textbook} + n + n\ln(2\pi) + \ln n.\end{aligned}$$

Our textbook uses the following formulas which are essentially the same as the above.

$$\mathrm{AIC}_p = n \ln \mathrm{SSE}_p - n \ln n + 2 \cdot p$$

$$\mathrm{BIC}_p = n \ln \mathrm{SSE}_p - n \ln n + (\ln n) \cdot p.$$

Note that the values of the last terms in the above equations, $2 \cdot p$ and $(\ln n) \cdot p$, increase as the number of parameters (p) increases, while the value of the SSE_p decreases.

Thus, by analogy with the approach of R^2_{adj} and C_p, these criteria also assign penalties for over-fitting, namely, selecting large numbers of predictors. Notice that $\ln 8 = 2.079442$. Thus, the BIC (or SBC) gives more penalty for over-fitting than AIC when $n \geq 8$. This implies that the BIC tends to favor more simple models (when $n \geq 8$).

Example 9.1. Example on Page 360 of the textbook. (See also Table 9.2, columns 6 and 7, and Table 9.1 for the data set.)

Note that `AIC( )` function in R is the basically the same as the original AIC and BIC formulas in (9.2) and (9.3). The R defines $\mathrm{AIC} = -2 \ln \hat{L} + k \cdot (p + 1)$. Thus, the AIC in R gives AIC_p with $k = 2$ and BIC_p with $k = \ln n$.

R

Ⓡ Read Data

```
> mydata =
    read.table("https://raw.githubusercontent.com/AppliedStat/LM/master/CH09TA01.txt")

> x4 = mydata[,4]
> y1 = mydata[,10]
```

Regression

```
> LM1 = lm (y1 ~  x4)

> n = length(y1)
> p = LM1$rank

> y1.fit = fitted(LM1)
> SSE = sum( (y1-y1.fit)^2 )
>
> AIC(LM1,k=2) - (n + n*log(2*pi) + 2)
[1] -103.2615

> n * log(SSE) - n * log(n) + 2 * p
[1] -103.2615

> AIC(LM1,k=log(n)) - (n + n*log(2*pi) + log(n))
[1] -99.28357

```

```
> n * log(SSE) - n * log(n) + log(n) * p
[1] -99.28357
```

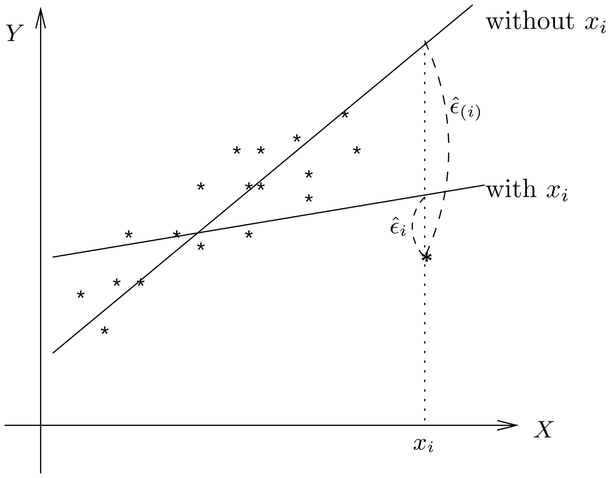
2.5 PRESS$_p$

The PRESS (prediction sum of squares) is defined as

$$\text{PRESS} = \sum_{i=1}^{n} \hat{\epsilon}_{(i)}^2,$$

where $\hat{\epsilon}_{(i)}$ is called PRESS (prediction sum of squares) residual for the ith observation.

Raw residuals $\hat{\epsilon}_i$ and PRESS residuals $\hat{\epsilon}_{(i)}$



The PRESS residual is defined as

$$\hat{\epsilon}_{(i)} = Y_i - \hat{Y}_{(i)},$$

where

$$\hat{Y}_{(i)} = \mathbf{x}_i{}'\hat{\boldsymbol{\beta}}_{(i)}$$

$$\mathbf{x}_i{}' = [X_{i1}, X_{i2}, \ldots, X_{i,p-1}]$$

$$\hat{\boldsymbol{\beta}}_{(i)} = \text{ estimate of } \boldsymbol{\beta} \text{ obtained by leaving out } i\text{th observation.}$$

Models with small PRESS$_p$ fit well in the sense of having small prediction errors. PRESS$_p$ can be calculated without fitting the model n times, each time deleting one of the n cases.

One can show that

$$\hat{\epsilon}_{(i)} = \frac{\hat{\epsilon}_i}{1 - h_{ii}},$$

where h_{ii} is the ith diagonal element of $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. PRESS values are useful for analysis of residuals and influence which will be covered in the next chapter.

Example 9.2. PRESS example on Page 361 of Kutner et al. (2005).

Minitab

Read Data

```
MTB >read C1-C10 ;
SUBC>  file "S:\LM\CH09TA01.txt" .
Entering data from file: S:\LM\CH09TA01.TXT
54 rows read.
```

Regression

```
##
## using regr Minitab command
##
MTB > regr C10 3 C1-C3;
SUBC> resid C21;
SUBC> hi    C22;
SUBC> brief 0.

MTB > let k1 = sum( (C21/(1-C22))**2 )

##
## using PRESS.MAC at https://github.com/AppliedStat/LM
##
MTB > %S:\LM\PRESS  C10  C2-C4  k2

Executing from file: S:\LM\PRESS.MAC

Data Display
K2    4.59693

MTB > print k1 k2.

Data Display
K1    3.91424
K2    4.59693
```

R

Ⓡ Read Data

```
> mydata =
    read.table("https://raw.githubusercontent.com/AppliedStat/LM/master/CH09TA01.txt")
>
> x1 = mydata[,1]
> x2 = mydata[,2]
> x3 = mydata[,3]
> x4 = mydata[,4]
> y  = mydata[,10]
```

Regression

```
##- PRESS function
 PRESS <- function(object ) {
   press.resid = resid(object) / (1-hatvalues(object));
   sum(press.resid^2);
 }

>
> LM = lm(y ~ x1 + x2 + x3)
> PRESS ( LM )
[1] 3.91424
>
> PRESS ( lm(y ~ x2 + x3 + x4) )
[1] 4.596928
```

∥

Best subsets regression

All possible regressions are calculated. However we are given the output of only the best k among $(p-1)$-predictor models. Suppose we have $X_1, \ldots, X_6$. All possible regressions would require a total of $2^6 - 1 = 63$ except the null (intercept-only) model.

$$\underset{6}{\binom{6}{1}} + \underset{15}{\binom{6}{2}} + \underset{20}{\binom{6}{3}} + \underset{15}{\binom{6}{4}} + \underset{6}{\binom{6}{5}} + \underset{1}{\binom{6}{6}} = 63.$$

If we select the best subsets with $k = 3$, we need to look at 16 outputs:

$$3 + 3 + 3 + 3 + 3 + 1 = 16.$$

The `Minitab` command for best subsets regression is `BREG`. `BREG` first looks at all one-predictor regression models and selects the model giving the largest R_p^2. Information on this model and the next best one-predictor model is printed. Then `BREG` looks at all two-predictor modes, finds the one with the largest R_p^2, prints information on this and the next best. This process continues until all $P-1$ predictors are used. Four summary statistics (R_p^2, $R_{\text{adj},p}^2$, C_p and $s = \sqrt{\text{MSE}}$) are printed for each model.

The R command for best subsets regression is `leaps( )`. To use `leaps( )`, "`leaps` R package" should be installed. This package can be downloaded at

`https://cran.r-project.org/web/packages/leaps/`.

In general, we look for models where C_p is small and is also close to p. If the model is adequate (*i.e.*, fits the data well), then the expected value of C_p is approximately equal to p (the number of parameters in the current model).

Example 9.3. Best 2 subsets for each subset size - Surgical Unit data set.

Minitab

Read Data

```
MTB >read C1-C10 ;
SUBC>  file "S:\LM\CH09TA01.txt" .
Entering data from file: S:\LM\CH09TA01.TXT
54 rows read.
```

Best Subsets Regression

```
MTB > bregr C10 8 C1-C8   ## k=2 is default (i.e. Best 2 subsets for each subset size).
Best Subsets Regression: C10 versus C1, C2, C3, C4, C5, C6, C7, C8

Response is C10
                       Mallows           C C C C C C C C
Vars  R-Sq  R-Sq(adj)       Cp        S  1 2 3 4 5 6 7 8
   1  42.8       41.7    117.4  0.37549      X
   1  42.2       41.0    119.2  0.37746        X
   2  66.3       65.0     50.5  0.29079    X X
   2  59.9       58.4     69.1  0.31715      X X
   3  77.8       76.5     18.9  0.23845    X X         X
   3  75.7       74.3     25.0  0.24934  X X X
   4  83.0       81.6      5.8  0.21087  X X X         X
   4  81.4       79.9     10.3  0.22023    X X X       X
   5  83.7       82.1      5.5  0.20827  X X X     X   X
   5  83.6       81.9      6.0  0.20931  X X X   X     X
   6  84.3       82.3      5.8  0.20655  X X X   X X   X
   6  83.9       81.9      7.0  0.20934  X X X     X X X
   7  84.6       82.3      7.0  0.20705  X X X   X X X X
   7  84.4       82.0      7.7  0.20867  X X X X X X   X
   8  84.6       81.9      9.0  0.20927  X X X X X X X X

## If you want best 5 subsets, use the following subcommand.
MTB > bregr C10 8 C1-C8 ;
SUBC>  best 5.

Best Subsets Regression: C10 versus C1, C2, C3, C4, C5, C6, C7, C8
Response is C10
                       Mallows           C C C C C C C C
Vars  R-Sq  R-Sq(adj)       Cp        S  1 2 3 4 5 6 7 8
   1  42.8       41.7    117.4  0.37549      X
   1  42.2       41.0    119.2  0.37746        X
   1  22.1       20.6    177.9  0.43807    X
   1  13.9       12.2    201.8  0.46052                X
   1   6.1        4.3    224.7  0.48101  X
   2  66.3       65.0     50.5  0.29079    X X
   2  59.9       58.4     69.1  0.31715      X X
   2  54.9       53.1     84.0  0.33668  X   X
   2  51.6       49.7     93.4  0.34850      X         X
   2  50.8       48.9     95.9  0.35157        X       X
   3  77.8       76.5     18.9  0.23845    X X         X
   3  75.7       74.3     25.0  0.24934  X X X
   3  71.8       70.1     36.5  0.26885    X X X
   3  68.1       66.2     47.3  0.28587    X X       X
   3  67.6       65.7     48.7  0.28802    X X   X
   4  83.0       81.6      5.8  0.21087  X X X         X
   4  81.4       79.9     10.3  0.22023    X X X       X
   4  78.9       77.2     17.8  0.23498    X X     X   X
   4  78.4       76.6     19.3  0.23785    X X   X     X
   4  78.0       76.2     20.4  0.23982    X X       X X
   5  83.7       82.1      5.5  0.20827  X X X     X   X
   5  83.6       81.9      6.0  0.20931  X X X   X     X
   5  83.3       81.6      6.8  0.21100  X X X X       X
   5  83.2       81.4      7.2  0.21193  X X X       X X
   5  81.8       79.9     11.3  0.22044    X X X   X   X
   6  84.3       82.3      5.8  0.20655  X X X   X X   X
   6  83.9       81.9      7.0  0.20934  X X X     X X X
   6  83.9       81.8      7.2  0.20964  X X X X   X   X
   6  83.8       81.8      7.2  0.20982  X X X   X   X X
   6  83.7       81.6      7.6  0.21066  X X X X X     X
   7  84.6       82.3      7.0  0.20705  X X X   X X X X
   7  84.4       82.0      7.7  0.20867  X X X X X X   X
   7  84.0       81.6      8.7  0.21081  X X X X   X X X
   7  84.0       81.5      8.9  0.21136  X X X X X   X X
   7  82.1       79.4     14.3  0.22306    X X X X X X X
   8  84.6       81.9      9.0  0.20927  X X X X X X X X
```

R

Ⓡ Read Data

```
> mydata =
    read.table("https://raw.githubusercontent.com/AppliedStat/LM/master/CH09TA01.txt")
>
> colnames(mydata) = c("Blood", "Prog", "Enzyme", "Liver", "Age", "Gender",
    "Alc.Mod",  "Alc.Heavy", "Surv", "log-Surv")
>
> # Need to install "leaps" package
> # install.packages("leaps")
> library ("leaps")
>
> xx = mydata[ , c(-9,-10)]
> y  = mydata[ , 10]
>
> best1 = leaps(xx,y,  method = "Cp", nbest=2, names=colnames(xx))  # Cp is default.
>                             ### Use nbest=5 for best 5 subsets
>
> best1
$which
  Blood  Prog Enzyme Liver   Age Gender Alc.Mod Alc.Heavy
1 FALSE FALSE   TRUE FALSE FALSE  FALSE   FALSE     FALSE
1 FALSE FALSE  FALSE  TRUE FALSE  FALSE   FALSE     FALSE
2 FALSE  TRUE   TRUE FALSE FALSE  FALSE   FALSE     FALSE
2 FALSE FALSE   TRUE  TRUE FALSE  FALSE   FALSE     FALSE
3 FALSE  TRUE   TRUE FALSE FALSE  FALSE   FALSE      TRUE
3  TRUE  TRUE   TRUE FALSE FALSE  FALSE   FALSE     FALSE
4  TRUE  TRUE   TRUE FALSE FALSE  FALSE   FALSE      TRUE
4 FALSE  TRUE   TRUE  TRUE FALSE  FALSE   FALSE      TRUE
5  TRUE  TRUE   TRUE FALSE FALSE   TRUE   FALSE      TRUE
5  TRUE  TRUE   TRUE FALSE  TRUE  FALSE   FALSE      TRUE
6  TRUE  TRUE   TRUE FALSE  TRUE   TRUE   FALSE      TRUE
6  TRUE  TRUE   TRUE FALSE FALSE   TRUE    TRUE      TRUE
7  TRUE  TRUE   TRUE FALSE  TRUE   TRUE    TRUE      TRUE
7  TRUE  TRUE   TRUE  TRUE  TRUE   TRUE   FALSE      TRUE
8  TRUE  TRUE   TRUE  TRUE  TRUE   TRUE    TRUE      TRUE

$label
[1] "(Intercept)" "Blood"       "Prog"        "Enzyme"      "Liver"
[6] "Age"         "Gender"      "Alc.Mod"     "Alc.Heavy"

$size
 [1] 2 2 3 3 4 4 5 5 6 6 7 7 8 8 9

$Cp
```

```
 [1] 117.409441 119.171240  50.471575  69.131808  18.914496  24.980500
 [7]   5.750774  10.267014   5.540639   6.018212   5.787389   7.029456
[13]   7.029455   7.735230   9.000000

>
> # The following looks like the Minitab output format.
> cbind( best1$which, best1$Cp)
  Blood Prog Enzyme Liver Age Gender Alc.Mod Alc.Heavy
1     0    0      1     0   0      0       0         0 117.409441
1     0    0      0     1   0      0       0         0 119.171240
2     0    1      1     0   0      0       0         0  50.471575
2     0    0      1     1   0      0       0         0  69.131808
3     0    1      1     0   0      0       0         1  18.914496
3     1    1      1     0   0      0       0         0  24.980500
4     1    1      1     0   0      0       0         1   5.750774
4     0    1      1     1   0      0       0         1  10.267014
5     1    1      1     0   0      1       0         1   5.540639
5     1    1      1     0   1      0       0         1   6.018212
6     1    1      1     0   1      1       0         1   5.787389
6     1    1      1     0   0      1       1         1   7.029456
7     1    1      1     0   1      1       1         1   7.029455
7     1    1      1     1   1      1       0         1   7.735230
8     1    1      1     1   1      1       1         1   9.000000
>
> # Using R-Sq   "adjr2", "r2"
> best2 = leaps(xx,y,  method = "r2", nbest=2)
>
> # Using R-Sq   "adjr2", "r2"
> best3 = leaps(xx,y,  method = "adjr2", nbest=2)
>
> # The following looks like Figure 9.6 on Page 363
> round( cbind(best2$r2*100, best3$adjr2*100, best1$Cp, best1$which), 1)
                  Blood Prog Enzyme Liver Age Gender Alc.Mod Alc.Heavy
1 42.8 41.7 117.4     0    0      1     0   0      0       0         0
1 42.2 41.0 119.2     0    0      0     1   0      0       0         0
2 66.3 65.0  50.5     0    1      1     0   0      0       0         0
2 59.9 58.4  69.1     0    0      1     1   0      0       0         0
3 77.8 76.5  18.9     0    1      1     0   0      0       0         1
3 75.7 74.3  25.0     1    1      1     0   0      0       0         0
4 83.0 81.6   5.8     1    1      1     0   0      0       0         1
4 81.4 79.9  10.3     0    1      1     1   0      0       0         1
5 83.7 82.1   5.5     1    1      1     0   0      1       0         1
5 83.6 81.9   6.0     1    1      1     0   1      0       0         1
6 84.3 82.3   5.8     1    1      1     0   1      1       0         1
6 83.9 81.9   7.0     1    1      1     0   0      1       1         1
7 84.6 82.3   7.0     1    1      1     0   1      1       1         1
7 84.4 82.0   7.7     1    1      1     1   1      1       0         1
8 84.6 81.9   9.0     1    1      1     1   1      1       1         1
```

||

3 Sequential variable selection procedures

If the pool of potential predictors contains 20 to 60 or even more variables, use of a “best” subsets algorithm may not be feasible. An automatic search procedure that develops the best subset of predictors “sequentially” may then be helpful. It was developed to economize on computational efforts, as compared with the all-possible-regressions procedures.

Variable selection algorithms

Two models are said to be nested if one is a special case of the other.

Example 9.4.

$$\begin{aligned} &\text{Model A:} \quad Y = \beta_0 + \beta_1 X_1 + \epsilon \\ &\text{Model B:} \quad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \end{aligned}$$

Model A is a special case of Model B with $\beta_2 = 0$. We say Model A is nested within Model B. ||

Example 9.5.

$$\begin{aligned} &\text{Model A:} \quad Y = \beta_0 + \beta_1 X_1 + \epsilon \\ &\text{Model B:} \quad Y = \beta_0 + \beta_2 X_2 + \epsilon \end{aligned}$$

These are not nested. ||

Nest models can always be compared using a partial F-test or a sequential F-test. Consider the following hypothesis test.

$$\begin{aligned} &H_0: \text{ simpler model holds (nested within } H_1) \\ &H_1: \text{ more complicated model holds} \end{aligned}$$

A large t or F statistic (or small p-value) indicates that H_1 is more plausible than H_0, and we should adopt H_1 rather than H_0. A small t or F statistic (or large p-value) indicates that H_1 is no more plausible than H_0, and thus we should keep H_0.

Suppose we have two potential predictors X_1 and X_2. This gives four possible models. Each line represents a comparison of nested model. Starting from any vertex, we can decide to adopt a more complex model if the p-value for that test is smaller than some cutoff value, say, $\alpha_{\text{enter}} = p_{\text{enter}} = 0.05$, (equivalently, the F test statistic is larger than some cutoff, say, $F_{\text{enter}} \approx 4$, or the t test statistic is larger than some cutoff $t_{\text{enter}} \approx 2$). Or, we can decide to adopt a simpler model if the p-value for that test is above some cutoff, say, $\alpha_{\text{remove}} = p_{\text{remove}} = 0.1$, (equivalently, the F test statistic is smaller than some cutoff, say, $F_{\text{remove}} \approx 3$).

Variable selection algorithms are sets of rules for deciding how to move along the edges of the hypercube. A common problem of these procedures is that where you finally end up often depends on where you start.

1. Forward Selection
   - (i) Start with the *null* model.
   - (ii) Add the most significant variable if p-value is less than $\alpha_{\text{enter}} = p_{\text{enter}}$, (equivalently, F is larger than F_{enter}).
   - (iii) Continue until no more variables enter the model.
2. Backward Elimination
   - (i) Start with the *full* model.
   - (ii) Eliminate the least significant variable whose p-value is larger than $\alpha_{\text{remove}} = p_{\text{remove}}$, (equivalently, F is smaller than F_{remove}).
   - (iii) Continue until no more variables can be discarded from the model.
3. Stepwise Regression
   - (i) Start with any model.
   - (ii) Check each predictor that is currently *in* the model. Suppose the current model contains $X_1, \ldots, X_k$. Then F statistic for X_i is

     $$F = \frac{\text{SSE}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_k) - \text{SSE}(X_1, \ldots, X_k)}{\text{MSE}(X_1, \ldots, X_k)} \sim F(1, n-k-1).$$

     Eliminate the least significant variable whose p-value is larger than $\alpha_{\text{remove}} = p_{\text{remove}}$, (equivalently, F is smaller than F_{remove}).
   - (iii) Continue until no more variables can be discarded from the model.
   - (iv) Add the most significant variable if p-value is less than $\alpha_{\text{enter}} = p_{\text{enter}}$, (equivalently, F is larger than F_{enter}).
   - (v) Go to step (ii)
   - (vi) Repeat until no more predictors can be entered and no more can be discarded.

Remark 9.1.

- Forward selection is a special case of Stepwise regression with $\alpha_{\text{remove}} = p_{\text{remove}} = 1$ (equivalently, $F_{\text{remove}} = 0$). `Minitab` uses `AREMOVE` $= 1$ or `FREMOVE` $= 0$ for the subcommand.
- Backward elimination is a special case of Stepwise regression with $\alpha_{\text{enter}} = p_{\text{enter}} = 0$, (equivalently, $F_{\text{enter}} = \infty$). `Minitab` uses `AENTER` $= 0$ or `FENTER` $= 10000$ for the subcommand.

△

Remark 9.2.

1. Some old `Minitab` versions may ask us to use only F to enter and F to remove, rather than p-values. The default F values are both 4, which correspond roughly to p-values of 0.05. Note that $F(1-\alpha; 1, df) = t(1-\alpha/2; df)^2 \approx z^2_{1-\alpha/2}$.
2. We should satisfy $F_{\text{enter}} \geq F_{\text{remove}}$, (or, $\alpha_{\text{enter}} \leq \alpha_{\text{remove}}$) for the stepwise procedure.
3. A sensible stepwise procedure should allow us to "bundle together" a group of predictors – for example, a set of dummy variables defining a categorical variable – so that they can be entered or removed together.
4. A sensible stepwise procedure should also obey the hierarchy principle. For example, an interaction should not be entered unless both main effects are in. A main effect should not be discarded unless the interaction is out.
5. Stepwise regression procedures may not always find the "best" model (if there exists one). Depending on where you start, many good models may not even be reached by the procedure (it is similar to an initial condition problem in numerical computation).
6. If you can fit all 2^k all-possible-regressions models, then go ahead. (Note: The latest version of `Minitab` can bear up to 31 predictors). Tabulate PRESS, C_p and s^2 and see what you find. Note that there may not be a single "best" model. Stepwise regression procedures may be appropriate when the pool of available predictors is very large and all-possible-regressions method is not feasible.
7. Stepwise regression procedures may not be a good substitute for using your brain. They should not be used automatically or uncritically. The analyst's knowledge of a problem must be used to make conclusions.
8. The final choice of a model should not be made without considering such concepts as analysis of residuals, multi-collinearity diagnostics, transformations, detection of influential observations,

the analyst's knowledge of a problem, etc.

△

4 Stepwise regression example

The data were collected from some college. A random selection of $n = 279$ students was available for research. This data set can be obtained at

`https://raw.githubusercontent.com/appliedstat/LM/master/GPA.txt`

The following information was obtained from the survey:

Z = gender
X_1= SAT verbal
X_2= SAT math
X_3= high school GPA
Y = college school GPA

The cube below shows 8 possible models (gender always in). The t statistics for entering or removing are shown on the edges. Suppose we set $F_{\text{enter}} = 4$ and $F_{\text{remove}} = 4$, (equivalently, $t_{\text{enter}} = 2$ and $t_{\text{remove}} = 2$). Then we have the following paths ending up with the model $(X2, X3)$ regardless of starting model. That does not always happen, though!

`Minitab`

Read Data

```
MTB >read c10 c1-c3 c11 ;
SUBC>  file "S:\LM\GPA.txt" .
Entering data from file: S:\LM\GPA.TXT
279 rows read.
MTB > name c10  'Z'
MTB > name c1   'X1'
MTB > name c2   'X2'
MTB > name c3   'X3'
MTB > name c11  'Y'
```

Forward Selection

```
MTB > stepwise c11 4 c10 c1-c3;
SUBC> force c10;
SUBC> fenter 4 ;
SUBC> fremove 0;
SUBC> steps 5.

Stepwise Regression: Y versus Z, X1, X2, X3

Forward selection.  F-to-Enter: 4

Response is Y on 4 predictors, with N = 279
```

```
Step            1      2        3
Constant    3.154  2.164    1.373

Z           0.228  0.212    0.274
T-Value      4.05   4.30     5.79
P-Value     0.000  0.000    0.000

X3                 0.363    0.290
T-Value             9.12     7.39
P-Value            0.000    0.000

X2                        0.00173
T-Value                      6.21
P-Value                     0.000

S           0.467  0.411    0.385
R-Sq         5.60  27.44    36.37
R-Sq(adj)    5.25  26.91    35.68
Mallows Cp  134.8   42.0      5.2

More? (Yes, No, Subcommand, or Help)
SUBC> yes
No variables entered or removed

More? (Yes, No, Subcommand, or Help)
SUBC> no
```

Backward Elimination

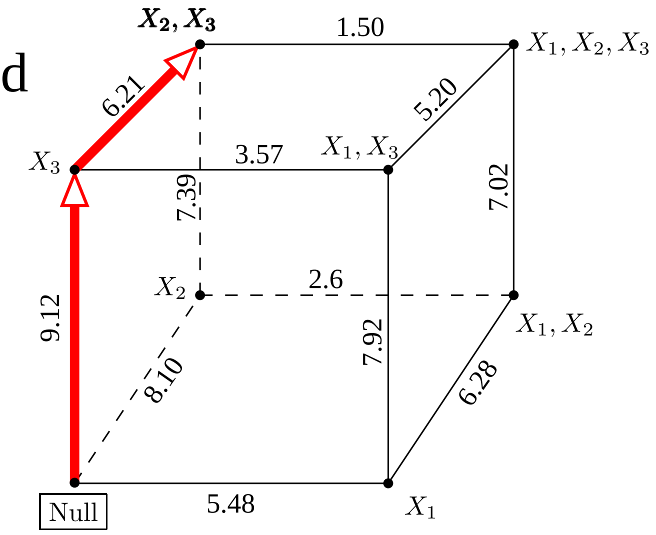
```
MTB > stepwise c11 4 c10 c1-c3;
SUBC> force c10;
SUBC> enter c1-c3;
SUBC> fenter 10000;
SUBC> fremove 4;
SUBC> steps 5.

Stepwise Regression: Y versus Z, X1, X2, X3

  F-to-Enter: 10000  F-to-Remove: 4

Response is Y on 4 predictors, with N = 279

Step              1        2
Constant      1.287    1.373

Z             0.273    0.274
T-Value        5.77     5.79
P-Value       0.000    0.000

X1          0.00042
T-Value        1.50
P-Value       0.135

X2          0.00156  0.00173
T-Value        5.20     6.21
P-Value       0.000    0.000

X3            0.279    0.290
T-Value        7.02     7.39
P-Value       0.000    0.000

S             0.384    0.385
R-Sq          36.89    36.37
R-Sq(adj)     35.97    35.68
Mallows Cp      5.0      5.2

More? (Yes, No, Subcommand, or Help)
SUBC> yes
No variables entered or removed

More? (Yes, No, Subcommand, or Help)
```

```
SUBC> no
```

Stepwise Regression

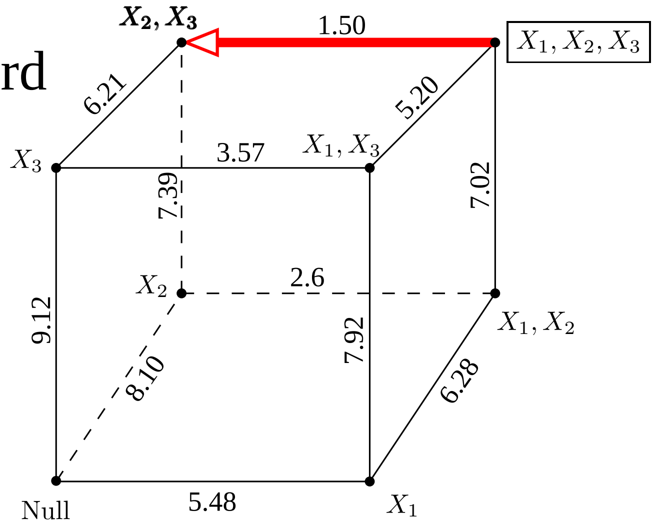
```
MTB > stepwise c11 4 c10 c1-c3;
SUBC> force c10;
SUBC> enter c1;
SUBC> fenter 4 ;
SUBC> fremove 4;
SUBC> steps 5.

Stepwise Regression: Y versus Z, X1, X2, X3

  F-to-Enter: 4  F-to-Remove: 4

Response is Y on 4 predictors, with N = 279

Step              1        2        3        4
Constant      2.358    1.787    1.287    1.373

Z             0.243    0.224    0.273    0.274
T-Value        4.54     4.61     5.77     5.79
P-Value       0.000    0.000    0.000    0.000

X1          0.00159  0.00098  0.00042
T-Value        5.48     3.57     1.50
P-Value       0.000    0.000    0.135

X3                     0.322    0.279    0.290
T-Value                 7.92     7.02     7.39
P-Value                0.000    0.000    0.000

X2                            0.00156  0.00173
T-Value                          5.20     6.21
P-Value                         0.000    0.000

S             0.445    0.402    0.384    0.385
R-Sq          14.86    30.66    36.89    36.37
R-Sq(adj)     14.24    29.90    35.97    35.68
Mallows Cp     96.6     30.0      5.0      5.2

More? (Yes, No, Subcommand, or Help)
SUBC> yes
No variables entered or removed
More? (Yes, No, Subcommand, or Help)
SUBC> no
```

Forward



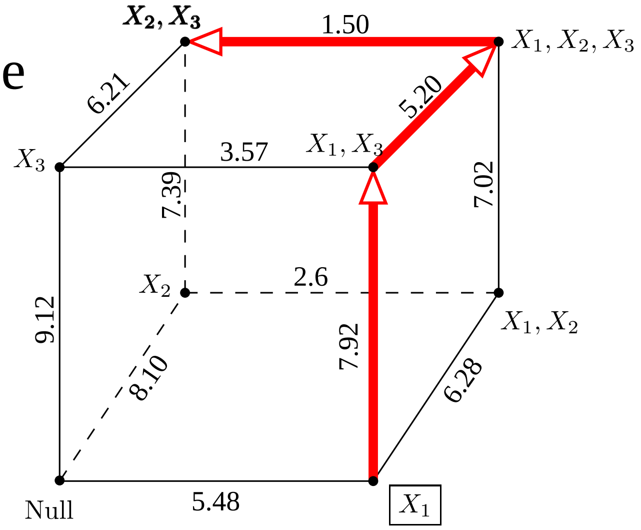
Z and

Backward



Z and

Stepwise



Z and

Example 9.6. Textbook Example. See Figure 9.7 on Page 366.

Minitab

Read Data

```
MTB >read C1-C10 ;
SUBC>  file "S:\LM\CH09TA01.txt" .
Entering data from file: S:\LM\CH09TA01.TXT
54 rows read.
```

Forward Selection

```
MTB > stepwise c10 8 c1-c8;
SUBC> Aenter  0.10 ;
SUBC> Aremove 0.15;
SUBC> steps 5.

Stepwise Regression: C10 versus C1, C2, C3, C4, C5, C6, C7, C8

  Alpha-to-Enter: 0.1  Alpha-to-Remove: 0.15

Response is C10 on 8 predictors, with N = 54

Step             1       2       3       4
Constant     5.264   4.351   4.291   3.852

C3          0.0151  0.0154  0.0145  0.0155
T-Value       6.23    8.19    9.33   11.07
P-Value      0.000   0.000   0.000   0.000

C2                  0.0141  0.0149  0.0142
T-Value               5.98    7.68    8.20
P-Value              0.000   0.000   0.000

C8                           0.429   0.353
T-Value                       5.08    4.57
P-Value                      0.000   0.000

C1                                   0.073
T-Value                               3.86
P-Value                              0.000

S            0.375   0.291   0.238   0.211
R-Sq         42.76   66.33   77.80   82.99
R-Sq(adj)    41.66   65.01   76.47   81.60
Mallows Cp   117.4    50.5    18.9     5.8

More? (Yes, No, Subcommand, or Help)

SUBC>  Yes
No variables entered or removed
More? (Yes, No, Subcommand, or Help)
SUBC> No
```

References

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2005). *Applied Linear Statistical Models*. McGraw-Hill, New York, 5th edition.

Mallows, C. L. (1973). Some comments on c p. *Technometrics*, 15(4):661–675.