Chapter 6: Model Comparison and Selection Methods

CZ

Fall, 2014

The Basic Problem

- We have a collection of x-variables for predicting the response variable y. We would like to determine the **best subset** of x-variables for this task, and we do not want to include any unnecessary one in our final model.
- **Basically, model selection procedure is a trade-off between simplicity and accuracy.** When there are two models that give nearly the same fit to the data, then we should choose the one with fewer parameters.

$$R^2 = \frac{SSR_p}{SST} = 1 - \frac{SSE_p}{SST}$$

- We would like R^2 to be as large as possible, **but the difficulty is that as the number of predictors of the model increases, the value of R^2 does too.**
- It does not count for the complexity of the model.

- R^2 may only be appropriate for comparing two models with the same number of predictors. It is not appropriate for comparing models with different number of predictors, especially nested/hierarchical models.

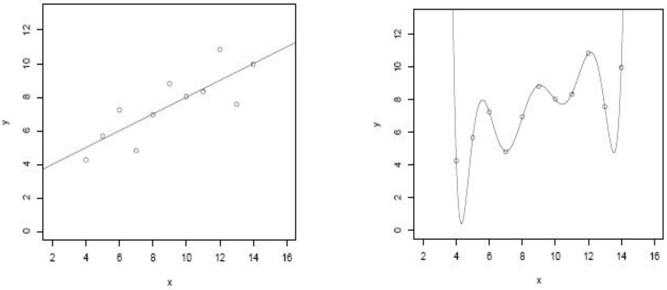

Figure: Left panel: $R^2 = 67\%$; Right panel: $R^2 = 100\%$

$R^2_{adjusted}$

- The interpretation of $R^2_{adjusted}$ is almost the same as for R^2, except that
  - $R^2_{adjusted}$ can actually start decreasing when we are adding unnecessary variables.
- The formula for R^2 is
$$R^2 = 1 - \frac{SSE_p}{SST}$$
- Definition: the formula for $R^2_{adjusted}$ is
$$R^2_{adjusted} = 1 - \frac{SSE_p/(n-p-1)}{SST/(n-1)}$$

$R^2_{adjusted}$

- Notice that
$$R^2_{adjusted} = 1 - \frac{MSE_p}{SST/(n-1)}.$$
- That is, **choosing the model with the largest $R^2_{adjusted}$ is equivalent to selecting the one with the smallest MSE_p.**

Mallow's C_p

- The formula for Mallow's C_p is:

$$C_p = \frac{SSE_p}{MSE(\text{model with all x-variables})} - (n - 2(p+1))$$

- If the model has no bias, the expectation of C_p is $p+1$.
- $C_p - (p+1)$, when positive, is used as a measure of the bias in this p-predictor model
- Pick a model for which C_p is small and close to $p+1$.

Information Criteria for Evaluating Models

- Two information criteria are Akaike's Information Criterion (AIC) and Schwartz's Bayesian Criterion (BIC):

$$
\begin{aligned}
AIC_p &= n \cdot \log(SSE_p) - n \cdot \log(n) + 2(p+1) \\
BIC_p &= n \cdot \log(SSE_p) - n \cdot \log(n) + \log(n) \cdot (p+1)
\end{aligned}
$$

- Pick the model with the smallest AIC_p or BIC_p.
- The difference in the two formulas is the multiplier of $p+1$, the number of parameters. The BIC places a higher penalty on the number of parameters in the model with $n \geq 8$ (because $\log(n) > 2$ for $n \geq 8$), so it will tend to encourage more parsimonious (fewer predictors) models.

Best Subsets/ All Subsets Regression

Best subsets regression consider all the possible models and select the "best" one based on certain criterion, such as largest adjusted R^2, Mallow's Cp close to $p + 1$, lowest AIC/BIC, etc.

Example

- "Swiss" is a data set with 47 observations on 6 variables. y is "Fertility" and x-variables are "Agriculture", "Examination", "Education", "Catholic", and "Infant Mortality".
- In R, to perform a best subset regression:

```
>install.packages("leaps")
>library(leaps)
>data(swiss)
>best=regsubsets(Fertility~Agriculture+Examination+
   Education+Catholic+Infant.Mortality, nbest=1,
   data=swiss)
```

```
> summary(best)
1 subsets of each size up to 5
Selection Algorithm: exhaustive
         Agriculture Examination Education Catholic
1  ( 1 ) " "         " "         "*"       " "
2  ( 1 ) " "         " "         "*"       "*"
3  ( 1 ) " "         " "         "*"       "*"
4  ( 1 ) "*"         " "         "*"       "*"
5  ( 1 ) "*"         "*"         "*"       "*"
         Infant.Mortality
1  ( 1 ) " "
2  ( 1 ) " "
3  ( 1 ) "*"
4  ( 1 ) "*"
5  ( 1 ) "*"
> summary(best)$adjr2
[1] 0.4281849 0.5551665 0.6390004 0.6707140 0.6709710
> summary(best)$cp
[1] 35.204895 18.486158  8.178162  5.032800  6.000000
> summary(best)$bic
[1] -19.60287 -28.61139 -35.65643 -37.23388 -34.55301
```

Example

Based on the $R^2_{Adjusted}$, C_p and BIC_p, the competition is between the four-predictor model and five-predictor model. We fit both models and conclude that the four-predictor model is our final model.

Best Five-Predictors Model

```
> summary(lm(Fertility~., data=swiss))

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      66.91518   10.70604   6.250 1.91e-07
Agriculture      -0.17211    0.07030  -2.448  0.01873
Examination      -0.25801    0.25388  -1.016  0.31546
Education        -0.87094    0.18303  -4.758 2.43e-05
Catholic          0.10412    0.03526   2.953  0.00519
Infant.Mortality  1.07705    0.38172   2.822  0.00734
```

Best Four-Predictors Model

```
>
summary(lm(Fertility~Agriculture+Education+Catholic+Inf
ant.Mortality, data=swiss))

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      62.10131    9.60489   6.466 8.49e-08
Agriculture      -0.15462    0.06819  -2.267  0.02857
Education        -0.98026    0.14814  -6.617 5.14e-08
Catholic          0.12467    0.02889   4.315 9.50e-05
Infant.Mortality  1.07844    0.38187   2.824  0.00722
```

Best Subsets Procedure

- Best subsets consider all the possible models and select the "best" one based on certain criterion, such as largest adjusted R^2, small Mallow's C_p close to $p+1$, lowest AIC, BIC etc.
- When p is above 30 to 40, it is not feasible for us to compare all of the competitive models visually.
- For large p, an alternative search procedure is to develop the best subset of x variables sequentially.

Forward Selection

Two alternative ways to select "best" models are Forward Selection and Backward Elimination

- Forward Selection:
  - We begin with no x-variables in the model, and add variables, **one at a time**, in some optimal way.
  - For example, the variable is added (only) if its associated p-value based on t test is less than a specified standard (α-level). If multiple predictors have p-value less than α-level, we select the one with the smallest p-value (i.e. the most significant predictor).
  - The procedure stops when no available variables have p-value smaller than the α-level.

Backward Elimination

- Backward Elimination:
  - We begin with all potential x-variables in a model, and then remove "weak" variables, **one at a time**, until a desirable stopping point is reached.
  - For example, at each step, we will remove the predictor if its associated p-value is larger than a specified standard (α-level). If multiple choices correspond to p-values bigger than α-level, eliminate the one with the largest p-value (i.e. the most insignificant one).
  - The procedure stops when all p-values are smaller than the specified standard (for example, α level).

Forward Selection/Backward Elimination

- A critical concept is that each step (adding a variable or removing one) is **conditional on** the previous step. For instance, in forward selection we are adding a variable to those already selected.
- The criterion to add in (or eliminate) a predictor in each step for forward selection (or backward elimination) might be considering p-value, AIC value, F-value, etc.

Example

Example: We illustrate the variable selection methods on some data on the 50 states in U.S.A. from the 1970s. We will take the **life expectancy** as the response and the remaining variables as predictors:

`Population`	population estimate of the state
`Income`	per capital income
`Illiteracy`	illiteracy percent of population
`Murder`	murder and non-negligent manslaughter rate
`Hs_Grad`	percent hight-school graduates
`Frost`	mean number of days with min temperature < 32 degrees
`Area`	land area in square miles

Backward Elimination

We start with all potential x-variables in a model.

```
> g<-lm(Life.Exp~., data=statedata)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.094e+01  1.748e+00  40.586  < 2e-16 ***
Population   5.180e-05  2.919e-05   1.775   0.0832 .
Income      -2.180e-05  2.444e-04  -0.089   0.9293
Illiteracy   3.382e-02  3.663e-01   0.092   0.9269
Murder      -3.011e-01  4.662e-02  -6.459 8.68e-08 ***
HS.Grad      4.893e-02  2.332e-02   2.098   0.0420 *
Frost       -5.735e-03  3.143e-03  -1.825   0.0752 .
Area        -7.383e-08  1.668e-06  -0.044   0.9649

Residual standard error: 0.7448 on 42 degrees of freedom
Multiple R-squared: 0.7362, Adjusted R-squared: 0.6922
F-statistic: 16.74 on 7 and 42 DF,  p-value: 2.534e-10
```

Backward Elimination

At each stage we remove the predictor with the largest p-value over 0.05 (α-level=0.05):

```
> g=lm(Life.Exp~Population+Income+Illiteracy+Murder
+HS.Grad+Frost, data=statedata)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.099e+01  1.387e+00  51.165  < 2e-16 ***
Population   5.188e-05  2.879e-05   1.802   0.0785 .
Income      -2.444e-05  2.343e-04  -0.104   0.9174
Illiteracy   2.846e-02  3.416e-01   0.083   0.9340
Murder      -3.018e-01  4.334e-02  -6.963 1.45e-08 ***
HS.Grad      4.847e-02  2.067e-02   2.345   0.0237 *
Frost       -5.776e-03  2.970e-03  -1.945   0.0584 .
```

We remove the predictor “Illiteracy” and refit the model.

```
> g=lm(Life.Exp~Population+Income+Murder+HS.Grad+Frost,
data=statedata)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.107e+01  1.029e+00  69.067  < 2e-16 ***
Population   5.115e-05  2.709e-05   1.888   0.0657 .
Income      -2.477e-05  2.316e-04  -0.107   0.9153
Murder      -3.000e-01  3.704e-02  -8.099 2.91e-10 ***
HS.Grad      4.776e-02  1.859e-02   2.569   0.0137 *
Frost       -5.910e-03  2.468e-03  -2.395   0.0210 *
```

Backward Elimination

We remove the predictor "Income" and refit the model.

```
> g=lm(Life.Exp~Population+Murder+HS.Grad+Frost,
data=statedata)

              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.103e+01  9.529e-01  74.542  < 2e-16 ***
Population   5.014e-05  2.512e-05   1.996  0.05201 .
Murder      -3.001e-01  3.661e-02  -8.199 1.77e-10 ***
HS.Grad      4.658e-02  1.483e-02   3.142  0.00297 **
Frost       -5.943e-03  2.421e-03  -2.455  0.01802 *
```

Backward Elimination

We remove the predictor "Population" and refit the model.

```
> g=lm(Life.Exp~Murder+HS.Grad+Frost, data=statedata)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 71.036379   0.983262  72.246  < 2e-16 ***
Murder      -0.283065   0.036731  -7.706 8.04e-10 ***
HS.Grad      0.049949   0.015201   3.286  0.00195 **
Frost       -0.006912   0.002447  -2.824  0.00699 **

Residual standard error: 0.7427 on 46 degrees of freedom
Multiple R-squared: 0.7127, Adjusted R-squared: 0.6939
F-statistic: 38.03 on 3 and 46 DF,  p-value: 1.634e-12
```

Forward Selection

We start with a null model with no predictors. Add the first variable with the smallest p-value below $\alpha = 0.05$:

```
> g=lm(Life.Exp~Population,data=statedata)
            Pr(>|t|)
Population     0.639
> g=lm(Life.Exp~Income,data=statedata)

            Pr(>|t|)
Income        0.0156 *
> g=lm(Life.Exp~Illiteracy,data=statedata)
           Pr(>|t|)
Illiteracy  6.97e-06 ***
> g=lm(Life.Exp~Murder,data=statedata)
            Pr(>|t|)
Murder      2.26e-11 ***
> g=lm(Life.Exp~HS.Grad,data=statedata)
            Pr(>|t|)
HS.Grad      9.2e-06 ***
> g=lm(Life.Exp~Frost,data=statedata)
            Pr(>|t|)
Frost          0.066 .
> g=lm(Life.Exp~Area,data=statedata)
            Pr(>|t|)
Area           0.458
```

Forward Selection

Add the second variable with the smallest p-value below $\alpha = 0.05$:

```
> g=lm(Life.Exp~Murder+Population,data=statedata)
            Pr(>|t|)
Murder      2.15e-12 ***
Population    0.0164 *
> g=lm(Life.Exp~Murder+Income,data=statedata)
            Pr(>|t|)
Murder      1.22e-10 ***
Income        0.0666 .
> g=lm(Life.Exp~Murder+Illiteracy,data=statedata)
            Pr(>|t|)
Murder      7.96e-07 ***
Illiteracy    0.543
> g=lm(Life.Exp~Murder+HS.Grad,data=statedata)
            Pr(>|t|)
Murder      2.18e-08 ***
HS.Grad     0.00909  **
> g=lm(Life.Exp~Murder+Frost,data=statedata)
            Pr(>|t|)
Murder      2.05e-11 ***
Frost         0.0352 *
> g=lm(Life.Exp~Murder+Area,data=statedata)
            Pr(>|t|)
Murder      3.47e-11 ***
Area           0.424
```

Forward Selection

Add the third variable with the smallest p-value below $\alpha = 0.05$:

```
g=lm(Life.Exp~Murder+HS.Grad+Population,data=statedata)
             Pr(>|t|)
Murder      1.91e-09 ***
HS.Grad      0.0112 *
Population   0.0199 *
>g=lm(Life.Exp~Murder+HS.Grad+Income,data=statedata)
             Pr(>|t|)
Murder      2.92e-08 ***
HS.Grad      0.0605 .
Income       0.6924
>g=lm(Life.Exp~Murder+HS.Grad+Illiteracy,data=statedata
              Pr(>|t|)
Murder       3.63e-07 ***
HS.Grad      0.00825  **
Illiteracy   0.40942
> g=lm(Life.Exp~Murder+HS.Grad+Frost,data=statedata)
Coefficients:
              Pr(>|t|)
Murder       8.04e-10 ***
HS.Grad      0.00195 **
Frost        0.00699 **
> g=lm(Life.Exp~Murder+HS.Grad+Area,data=statedata)
              Pr(>|t|)
Murder       1.3e-06 ***
HS.Grad      0.011 *
Area         0.514
```

Forward Selection

Can not add any more predictors at present significance level $\alpha = 0.05$, stop.

```
g=lm(Life.Exp~Murder+HS.Grad+Frost+Population,data=stat
edata)
            Pr(>|t|)
Murder      1.77e-10 ***
HS.Grad     0.00297 **
Frost       0.01802 *
Population  0.05201 .
g=lm(Life.Exp~Murder+HS.Grad+Frost+Income,data=statedat
a)
            Pr(>|t|)
Murder      1.07e-09 ***
HS.Grad     0.02643 *
Frost       0.00696 **
Income      0.57103
g=lm(Life.Exp~Murder+HS.Grad+Frost+Illiteracy,data=stat
edata)
            Pr(>|t|)
Murder      3.5e-08 ***
HS.Grad     0.01490 *
Frost       0.00936 **
Illiteracy  0.58236
g=lm(Life.Exp~Murder+HS.Grad+Frost+Area,data=statedata)
            Pr(>|t|)
Murder       5.34e-08 ***
HS.Grad      0.00566 **
Frost        0.00940 **
Area         0.83173
```

Comment

- For backward elimination, once a predictor has been eliminated from the model, it will not have the chance to re-enter the model, even if it becomes significant after other predictors being dropped.
- For forward selection, once a predictor entered the model, it remains in the model, even if it becomes non-significant after other predictors have been selected.

Use AIC to Identify Models

- In forward selection (backward elimination), we add in (remove) a predictor based on the p-value of t-test.
- Another selection rule is to look at the AIC value. We know a smaller AIC value is more desirable. Therefore, a model with the smallest AIC value should be selected.

Use AIC to Identify Models

- In forward selection, given a model of size k in the previous step, compare all the candidate models of size $k+1$ by adding in one of the remaining x-variables. Among these size-$(k+1)$ models, select the one with the smallest AIC value.
- In backward elimination, given a model of size k in the previous step, compare all the candidate models of size $k-1$ by eliminating one of the existing predictors. Among these size-$(k-1)$ models, select the one with the smallest AIC value.
- We stop when adding (eliminating) a predictor from the current model can not reduce the AIC value.

Example: Life Expectancy Data

For example, perform a “Forward Selection” based on AIC value.

```
> fullmodel<-lm(Life.Exp~., data=statedata)
> nullmodel<-lm(Life.Exp~1,data=statedata)
> step(nullmodel,scope=list
(upper=fullmodel),direction="forward")
Start:  AIC=30.44
Life.Exp ~ 1

              Df Sum of Sq    RSS     AIC
+ Murder       1    53.838 34.461 -14.609
+ Illiteracy   1    30.578 57.721  11.179
+ HS.Grad      1    29.931 58.368  11.737
+ Income       1    10.223 78.076  26.283
+ Frost        1     6.064 82.235  28.878
<none>                     88.299  30.435
+ Area         1     1.017 87.282  31.856
+ Population   1     0.409 87.890  32.203
```

Example

The second variable to enter:

```
Step:  AIC=-14.61
Life.Exp ~ Murder

             Df Sum of Sq    RSS     AIC
+ HS.Grad     1    4.6910 29.770 -19.925
+ Population  1    4.0161 30.445 -18.805
+ Frost       1    3.1346 31.327 -17.378
+ Income      1    2.4047 32.057 -16.226
<none>                    34.461 -14.609
+ Area        1    0.4697 33.992 -13.295
+ Illiteracy  1    0.2732 34.188 -13.007
```

Example

The third variable to enter:

```
Step:  AIC=-19.93
Life.Exp ~ Murder + HS.Grad

               Df Sum of Sq    RSS     AIC
+ Frost         1    4.3987 25.372 -25.920
+ Population    1    3.3405 26.430 -23.877
<none>                      29.770 -19.925
+ Illiteracy    1    0.4419 29.328 -18.673
+ Area          1    0.2775 29.493 -18.394
+ Income        1    0.1022 29.668 -18.097
```

Example

The fourth variable to enter:

```
Step:  AIC=-25.92
Life.Exp ~ Murder + HS.Grad + Frost

             Df Sum of Sq    RSS     AIC
+ Population  1   2.06358 23.308 -28.161
<none>                    25.372 -25.920
+ Income      1   0.18232 25.189 -24.280
+ Illiteracy  1   0.17184 25.200 -24.259
+ Area        1   0.02573 25.346 -23.970
```

Example

Can not add more variables into the model. Stop.

```
Step:  AIC=-28.16
Life.Exp ~ Murder + HS.Grad + Frost + Population

             Df Sum of Sq    RSS     AIC
<none>                    23.308 -28.161
+ Income      1 0.0060582 23.302 -26.174
+ Illiteracy  1 0.0039221 23.304 -26.170
+ Area        1 0.0007900 23.307 -26.163
```

Use F-statistic to Identify Models

- One other selection rule is to look at the F-statistic. Notice that at each step of either forward selection or backward elimination procedure, the larger model and the smaller model are nested.
- Therefore, we can use the general linear F test to compare them. The F statistic is:

$$F = \frac{SSR(X_j|\text{other predictors in the current model})}{MSE(X_j, \text{other predictors in the current model})}$$

where $SSR(X_j|\text{other predictors in the current model}) =$
$SSR(X_j, \text{other predictors in the current model}) -$
$SSR(\text{other predictors in the current model})$

Sequential Sum of Squares

- Sequential Sum of Squares: the increase in SSR when adding a new predictor to the model.
- Assume there are three predictors: X_1, X_2, X_3.

$$
\begin{aligned}
SSR(X_2|X_1) &= SSR(X_1, X_2) - SSR(X_1) \\
SSR(X_3|X_1, X_2) &= SSR(X_1, X_2, X_3) - SSR(X_1, X_2) \\
SSR(X_1, X_2, X_3) &= SSR(X_1) + SSR(X_2|X_1) + SSR(X_3|X_1, X_2)
\end{aligned}
$$

Sequential Sum of Squares

In R, ANOVA table gives sequential sum of squares. For example,

```
>anova(lm(Life.Exp~Murder+HS.Grad+Frost+Population,
data=statedata))
Analysis of Variance Table

Response: Life.Exp
           Df Sum Sq Mean Sq  F value    Pr(>F)
Murder      1 53.838  53.838 103.9425 2.828e-13 ***
HS.Grad     1  4.691   4.691   9.0567  0.004278 **
Frost       1  4.399   4.399   8.4925  0.005538 **
Population  1  2.064   2.064   3.9841  0.052005 .
Residuals  45 23.308   0.518
```

Sequential Sum of Squares

- SSR(Murder)$= 53.838$
- SSR(HS.Grad$|$Murder)$= 4.691$
- SSR(Frost$|$Murder, HS.Grad)$= 4.399$
- SSR(Population$|$Murder, HS.Grad, Frost)$= 2.064$
- Therefore, SSR(Murder, HS.Grad, Frost, Population)$= 53.838 + 4.691 + 4.399 + 2.064 = 64.992$

Use F-statistic to Identify Models

- In forward selection, we would like to include the predictor corresponds to the **largest** calculated F or the the **smallest** p-value for the corresponding F-test.
- In backward elimination, we would like to discard the predictor corresponds to the **smallest** calculated F or the **largest** p-value for the corresponding F-test.
- One can set a predetermined F value as the stopping point or an alpha-level for the p-value as the stopping point.

Example: Life expectancy data

For example, perform a forward selection based on F-statistic

```
> library(MASS)
> null<-lm(Life.Exp~1, data=statedata)
> fullmodel<-lm(Life.Exp~., data=statedata)
> newmodel<- addterm(null, scope=fullmodel, test="F")
Single term additions

Model:
Life.Exp ~ 1
           Df Sum of Sq    RSS     AIC F Value     Pr(F)
<none>                  88.299  30.435
Population  1     0.409 87.890  32.203   0.223   0.63866
Income      1    10.223 78.076  26.283   6.285   0.01562 *
Illiteracy  1    30.578 57.721  11.179  25.429 6.969e-06 ***
Murder      1    53.838 34.461 -14.609  74.989 2.260e-11 ***
HS.Grad     1    29.931 58.368  11.737  24.615 9.196e-06 ***
Frost       1     6.064 82.235  28.878   3.540   0.06599 .
Area        1     1.017 87.282  31.856   0.559   0.45815
```

Example

The second variable to enter:

```
> newmodel<-lm(Life.Exp~Murder, data=statedata)
> addterm(newmodel, scope=fullmodel, test="F")
Single term additions

Model:
Life.Exp ~ Murder
           Df Sum of Sq    RSS     AIC F Value    Pr(F)
<none>                  34.461 -14.609
Population  1    4.0161 30.445 -18.805  6.1999 0.016369 *
Income      1    2.4047 32.057 -16.226  3.5257 0.066636 .
Illiteracy  1    0.2732 34.188 -13.007  0.3756 0.542910
HS.Grad     1    4.6910 29.770 -19.925  7.4059 0.009088 **
Frost       1    3.1346 31.327 -17.378  4.7029 0.035205 *
Area        1    0.4697 33.992 -13.295  0.6494 0.424375
```

Example

The third variable to enter:

```
> newmodel<-lm(Life.Exp~Murder+HS.Grad, data=statedata)
> addterm(newmodel, scope=fullmodel, test="F")
Single term additions

Model:
Life.Exp ~ Murder + HS.Grad
           Df Sum of Sq    RSS     AIC F Value    Pr(F)
<none>                  29.770 -19.925
Population  1    3.3405 26.430 -23.877  5.8141 0.019949 *
Income      1    0.1022 29.668 -18.097  0.1585 0.692418
Illiteracy  1    0.4419 29.328 -18.673  0.6931 0.409421
Frost       1    4.3987 25.372 -25.920  7.9751 0.006988 **
Area        1    0.2775 29.493 -18.394  0.4329 0.513863
```

Example

Can not add any more predictors at significance level $\alpha = 0.05$, stop.

```
> newmodel<-lm(Life.Exp~Murder+HS.Grad+Frost, data=statedata)
> addterm(newmodel, scope=fullmodel, test="F")
Single term additions

Model:
Life.Exp ~ Murder + HS.Grad + Frost
           Df Sum of Sq    RSS     AIC F Value   Pr(F)
<none>                  25.372 -25.920
Population  1   2.06358 23.308 -28.161  3.9841 0.05201 .
Income      1   0.18232 25.189 -24.280  0.3257 0.57103
Illiteracy  1   0.17184 25.200 -24.259  0.3069 0.58236
Area        1   0.02573 25.346 -23.970  0.0457 0.83173
```

Stepwise Regression

It is a combination of backward and forward method. It addresses the situation where variables are added or removed early in the process and we want to change our mind about them later. The procedure depends on two alphas:

α_1: alpha-to-enter

α_2: alpha-to-drop

At each stage a variable may be added or removed and there are several variations on exactly how this is done.

Stepwise Regression

1. Start as in forward selection using significance lever α_1.
2. At each stage, once a predictor entered the model, check all other predictors previously in the model for their significance. Drop the least significant predictor (the one with the largest *p*-value) if its *p*-value is greater than the significance level α_2.
3. Continue until no predictors can be added and no predictors can be removed.

Stepwise Regression

- Usually, we set α-to-enter$\leq$ α-to-remove. Otherwise, an infinite cycling may occur if one of the predictors has a *p*-value in between α-to-enter and α-to-remove.
- For example, we can set α-to-enter$=0.05$ and α-to-remove$=0.15$.

Example: Life Expectancy Data

Perform a Stepwise Selection based on p-value for the F test.

```
> fullmodel<-lm(Life.Exp~., data=statedata)
> nullmodel<-lm(Life.Exp~1,data=statedata)
> step(nullmodel,scope=list
+ (upper=fullmodel),direction="both",test="F")
Start: AIC=30.44
Life.Exp ~ 1
             Df Sum of Sq    RSS     AIC F value    Pr(>F)
+ Murder      1    53.838 34.461 -14.609 74.9887 2.260e-11 ***
+ Illiteracy  1    30.578 57.721  11.179 25.4289 6.969e-06 ***
+ HS.Grad     1    29.931 58.368  11.737 24.6146 9.196e-06 ***
+ Income      1    10.223 78.076  26.283  6.2847   0.01562 *
+ Frost       1     6.064 82.235  28.878  3.5397   0.06599 .
<none>                    88.299  30.435
+ Area        1     1.017 87.282  31.856  0.5594   0.45815
+ Population  1     0.409 87.890  32.203  0.2233   0.63866
```

Example: Life Expectancy Data

The second variable to enter:

```
Step: AIC=-14.61
Life.Exp ~ Murder
             Df Sum of Sq    RSS     AIC F value   Pr(>F)
+ HS.Grad     1     4.691 29.770 -19.925  7.4059 0.009088 **
+ Population  1     4.016 30.445 -18.805  6.1999 0.016369 *
+ Frost       1     3.135 31.327 -17.378  4.7029 0.035205 *
+ Income      1     2.405 32.057 -16.226  3.5257 0.066636 .
<none>                    34.461 -14.609
+ Area        1     0.470 33.992 -13.295  0.6494 0.424375
+ Illiteracy  1     0.273 34.188 -13.007  0.3756 0.542910
- Murder      1    53.838 88.299  30.435 74.9887 2.26e-11 ***
```

Example: Life Expectancy Data

Check if any variable in the previous model needs to be removed.
The third variable to enter:

```
Life.Exp ~ Murder + HS.Grad
             Df Sum of Sq    RSS     AIC F value    Pr(>F)
+ Frost       1    4.3987 25.372 -25.920  7.9751  0.006988 **
+ Population  1    3.3405 26.430 -23.877  5.8141  0.019949 *
<none>                    29.770 -19.925
+ Illiteracy  1    0.4419 29.328 -18.673  0.6931  0.409421
+ Area        1    0.2775 29.493 -18.394  0.4329  0.513863
+ Income      1    0.1022 29.668 -18.097  0.1585  0.692418
- HS.Grad     1    4.6910 34.461 -14.609  7.4059  0.009088 **
- Murder      1   28.5974 58.368  11.737 45.1482 2.181e-08 ***
```

Example: Life Expectancy Data

No variable needs to be removed from the previous model; Cannot add more variables into the model. Stop.

```
Step: AIC=-25.92
Life.Exp ~ Murder + HS.Grad + Frost
             Df Sum of Sq    RSS     AIC F value    Pr(>F)
+ Population  1     2.064 23.308 -28.161  3.9841  0.052005 .
<none>                    25.372 -25.920
+ Income      1     0.182 25.189 -24.280  0.3257  0.571031
+ Illiteracy  1     0.172 25.200 -24.259  0.3069  0.582361
+ Area        1     0.026 25.346 -23.970  0.0457  0.831727
- Frost       1     4.399 29.770 -19.925  7.9751  0.006988 **
- HS.Grad     1     5.955 31.327 -17.378 10.7968  0.001950 **
- Murder      1    32.756 58.128  13.531 59.3881 8.039e-10 ***
```