APPLIED STATISTICS

Variable Selection

Dr Tao Zou

Research School of Finance, Actuarial Studies & Statistics
The Australian National University

Last Updated: Mon Sep 18 18:16:03 2017

Overview

- Motivation
- Sequential Variable Selection
- Variable Selection Among All Subsets
- Cross Validation for Variable Selection Results
- Multicollinearity

References

1. **F.L. Ramsey and D.W. Schafer** (2012)
   Chapter 12 of *The Statistical Sleuth*

2. The slides are made by **R Markdown**.
   http://rmarkdown.rstudio.com

Motivation

There are two prime reasons for variable selection:

1. Simple models with less variables are preferable to complex models with more variables.

2. Including unnecessary variables in a model results in a loss of precision $\Rightarrow$ overfitting.

Variable selection involves choosing a subset of explanatory variables to construct the multiple linear regression model.

Because if the exlanatory variables selected in MLR are determined, then the MLR model with those exlanatory variables is given.

Hence, sometimes we also call model selection.

Different subsets of explanatory variables determine different models. We call those models candidate models.

Motivation Example: Significance Depends on Other Explanatory Variables in the Model (Con'd)

Suppose we are interested in predicting ANU students' 2nd year GPA (Y) given their 1st year GPA (X_1) and UAC score (X_2). The following regression line is fit:

$$\mu\{Y|X_1, X_2\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2. \tag{1}$$

Based on the data, the p-values for the t-tests of whether $\beta_j = 0$ versus $\beta_j \neq 0$ for $j = 1, 2$ are 0.15 and 0.20, respectively.

Does this mean that we **do not need to select both X_1 and X_2** in the model? NO!

The test for β_2 tells us whether X_2 is needed in the model that already contains X_1, i.e., does X_2 offer any information about mean GPA over and above that of X_1?

The meaning of the coefficient of an explanatory variable depends on what other explanatory variables have been included in the regression.

Motivation Example: Significance Depends on Other Explanatory Variables in the Model (Con'd)

If we fit the following two models:

$$\mu\{Y|X_1\} = \alpha_0 + \alpha_1 X_1 \text{ and } \mu\{Y|X_2\} = \gamma_0 + \gamma_2 X_2.$$

For both models, the *p*-values for the *t*-tests of $\alpha_1 = 0$ versus $\alpha_1 \neq 0$ and $\gamma_2 = 0$ versus $\gamma_2 \neq 0$ can be computed. Based on the data, the results of the *p*-values are 0.01 and 0.02, respectively.

Hence at least one of X_1 and X_2 is needed in the model.

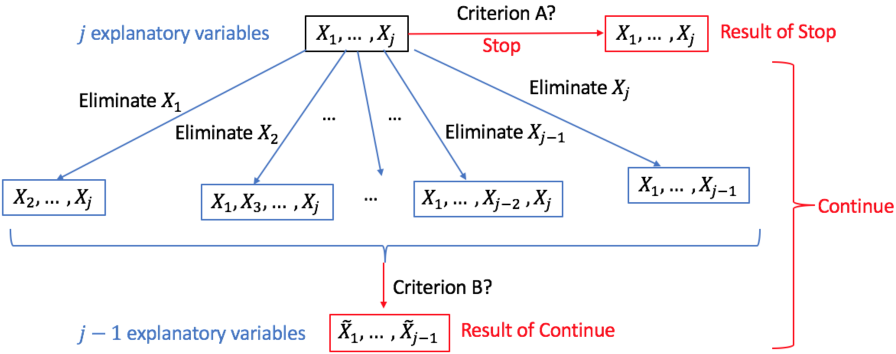
In this example X_1 and X_2 are probably highly correlated so we might expect this to be the case. The following *F*-test of model (1) avoids this problem.

$$\mathrm{H}_0: \text{ none of } X_1 \text{ and } X_2 \text{ is needed in the model} \leftrightarrow$$
$$\mathrm{H}_a: \text{ at least one of } X_1 \text{ and } X_2 \text{ is needed in the model.}$$

However, the *F*-test does not answer: which of X_1 and X_2 should be selected in the model.

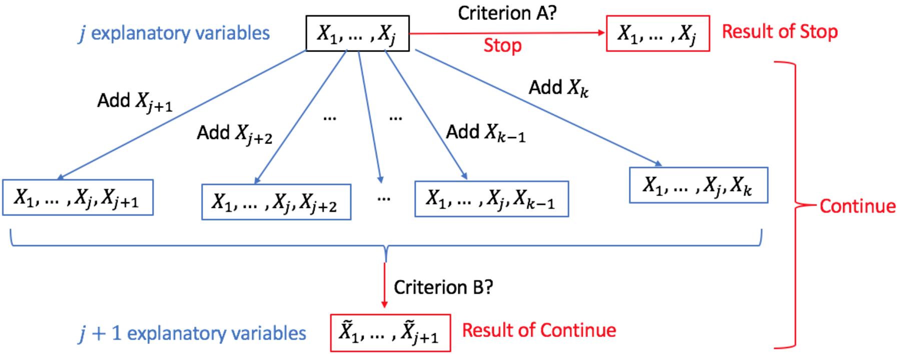
Sequential Variable Selection - Backward Elimination Steps

Backward Elimination Step for j Explanatory Variables



Sequential Variable Selection - Forward Selection Steps

Forward Selection Step for j Explanatory Variables



Suppose we have k explanatory variables $X_1, \dots, X_k$ in total.

Sequential Variable Selection

The idea behind sequential techniques is a sequential search through all possible combinations of variables by either adding or removing a single explanatory variable from the current candidate model at each step.

In order to accomplish the sequential variable selection, we need to determine Criterion A and Criterion B in forward selection steps and backward elimination steps.

Usually the criterion depends on statistical measures.

Sequential Variable Selection by Using *F*-Statistic

We start from using F-statistic as the measure. The F-statistic is

$$F\text{-Stat} = \frac{(\mathrm{SSE}_{\mathrm{reduced}} - \mathrm{SSE}_{\mathrm{full}})/d}{\hat{\sigma}^2_{\mathrm{full}}},$$

which is used to test

$$\mathrm{H}_0 : \text{the } \textbf{reduced model} \text{ is appropriate} \leftrightarrow$$

$$\mathrm{H}_a : \text{the } \textbf{full model} \text{ is appropriate.}$$

The p-value of F-test $< \alpha$ (usually 0.05) $\Leftrightarrow$ F-Stat is too large.

$\Rightarrow$

Reject H_0.

$\Rightarrow$

The **full model** is preferred.

Otherwise, the **reduced model** is preferred

Sequential Variable Selection by Using F-Statistic (Con'd)

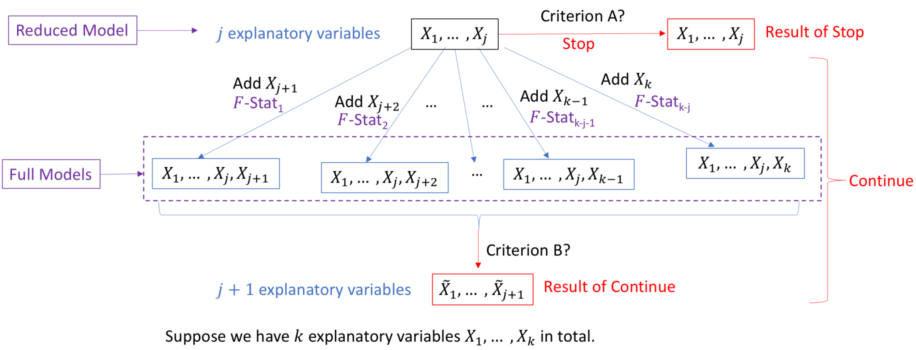
One can verify that the p-value of the F-test $< 0.05 \Leftrightarrow F$-Stat is larger than approximately 4, especially when sample size n is very large.

Hence a predefined "cut-off" for F-Stat is usually 4.

F-stat $> 4 \Rightarrow p$-value $< 0.05 \Rightarrow$ Reject H_0. $\Rightarrow$ The **full model** is preferred.
F-stat $< 4 \Rightarrow p$-value $> 0.05 \Rightarrow$ Not reject H_0. $\Rightarrow$ The **reduced model** is preferred.

We will explain the reduced model and the full model in forward selection steps and backward elimination steps.

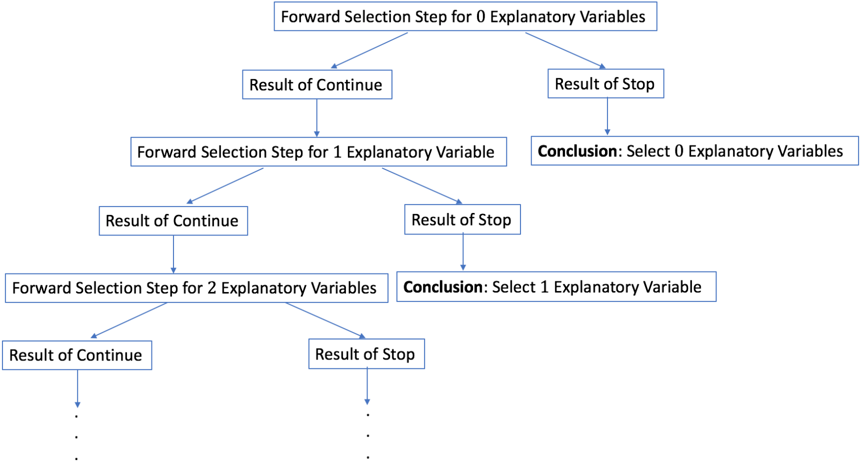
Forward Selection Step by Using F-Statistic


Forward Selection Step for j Explanatory Variables
Reduced Model
j explanatory variables
$X_1, \ldots, X_j$
Criterion A?
Stop
$X_1, \ldots, X_j$
Result of Stop
Add X_{j+1}
F-Stat$_1$
Add X_{j+2}
F-Stat$_2$
...
...
Add X_{k-1}
F-Stat$_{k-j-1}$
Add X_k
F-Stat$_{k-j}$
Full Models
$X_1, \ldots, X_j, X_{j+1}$
$X_1, \ldots, X_j, X_{j+2}$
...
$X_1, \ldots, X_j, X_{k-1}$
$X_1, \ldots, X_j, X_k$
Continue
Criterion B?
$j+1$ explanatory variables
$\tilde{X}_1, \ldots, \tilde{X}_{j+1}$
Result of Continue
Suppose we have k explanatory variables $X_1, \ldots, X_k$ in total.

Criterion A: if $\max\{F\text{-Stat}_1, F\text{-Stat}_2, \cdots, F\text{-Stat}_{k-j-1}, F\text{-Stat}_{k-j}\} < 4$ or $j = k$, then Stop; otherwise Continue.

Criterion B: $\tilde{X}_1, \cdots, \tilde{X}_{j+1}$ are those variables such that the corresponding F-Stat is the largest.

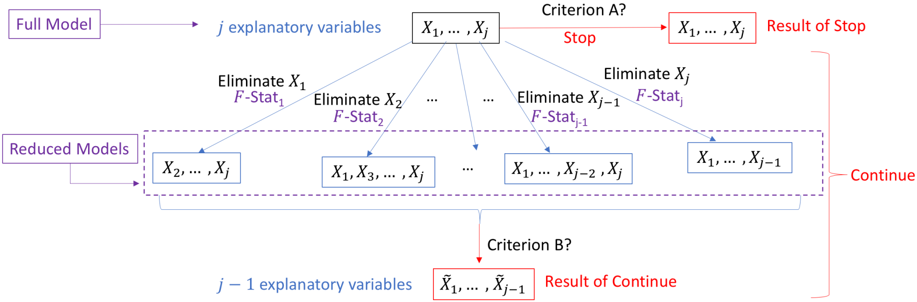
Complete Forward Selection Procedure by Using *F*-Statistic


Forward Selection Step for 0 Explanatory Variables
Result of Continue
Result of Stop
Forward Selection Step for 1 Explanatory Variable
Conclusion: Select 0 Explanatory Variables
Result of Continue
Result of Stop
Forward Selection Step for 2 Explanatory Variables
Conclusion: Select 1 Explanatory Variable
Result of Continue
Result of Stop
.
.
.
.
.
.

Keep doing on the above procedures, until the first time we obtain the **Result of Stop**.

Backward Elimination Step by Using F-Statistic

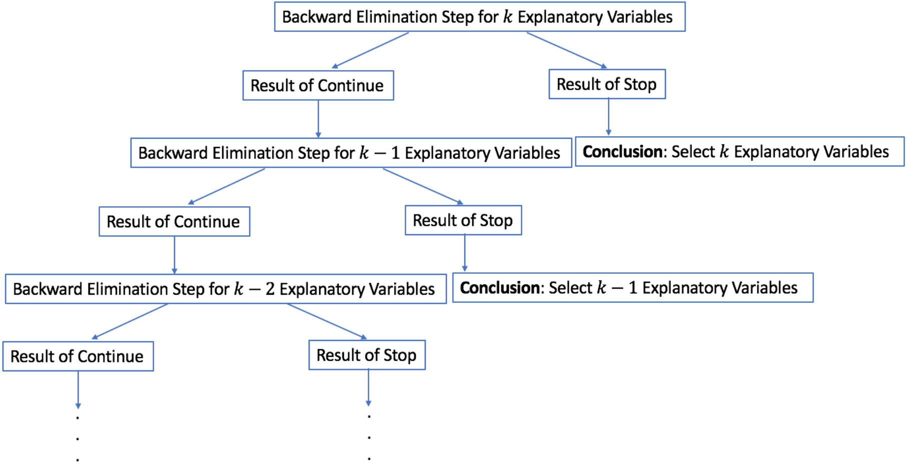
Backward Elimination Step for j Explanatory Variables



Criterion A: if $\min\{F\text{-Stat}_1, F\text{-Stat}_2, \cdots, F\text{-Stat}_{j-1}, F\text{-Stat}_j\} > 4$ or $j = 0$, then Stop; otherwise Continue.

Criterion B: $\tilde{X}_1, \cdots, \tilde{X}_{j-1}$ are those variables such that the corresponding F-Stat is the smallest.

Complete Backward Elimination Procedure by Using F-Statistic


Backward Elimination Step for k Explanatory Variables
Result of Continue
Result of Stop
Backward Elimination Step for $k-1$ Explanatory Variables
Conclusion: Select k Explanatory Variables
Result of Continue
Result of Stop
Backward Elimination Step for $k-2$ Explanatory Variables
Conclusion: Select $k-1$ Explanatory Variables
Result of Continue
Result of Stop

Keep doing on the above procedures, until the first time we obtain the **Result of Stop**.

Stepwise Selection Procedure by Using F-Statistic

This constitutes a combination of backward elimination and forward selection. Each step consists of the following steps. (Any starting model can be chosen, e.g., the model with no explanatory variables, or the model with all the explanatory variables.)

1. Do one Forward Selection Step.

2. Do one Backward Elimination Step.

Repeat Steps 1 and 2 until no explanatory variables can be added or removed.

Remark: The above three methods will sometimes lead to different variable selection results.

Example: SAT Scores

```
rm(list=ls())
library('Sleuth3')
SATdata=case1201
head(SATdata)
```

```
##          State  SAT Takers Income Years Public Expend Rank
## 1         Iowa 1088      3    326 16.79   87.8  25.60 89.7
## 2 SouthDakota 1075      2    264 16.07   86.2  19.95 90.6
## 3 NorthDakota 1068      3    317 16.57   88.3  20.62 89.8
## 4       Kansas 1045      5    338 16.30   83.9  27.14 86.3
## 5     Nebraska 1045      5    293 17.25   83.6  21.05 88.5
## 6      Montana 1033      8    263 15.91   93.7  29.48 86.4
```

```
SATdata=SATdata[-29,]  #removing Alaska
Y<-SATdata[,2]
X<-SATdata[,-c(1,2)]
X<-as.matrix(X)
```

Example: SAT Scores (Con'd)

```r
#install.packages('wle')
library(wle) #need to load this library!
```

```
## Loading required package: circular

## Warning: package 'circular' was built under R version 3.3.2

##
## Attaching package: 'circular'

## The following objects are masked from 'package:stats':
##
##     sd, var
```

```r
mle.stepwise(Y~X,f.in=4,f.out=4,type="Forward")
```

```
##
## Call:
## mle.stepwise(formula = Y ~ X, type = "Forward", f.in = 4, f.out = 4)
##
##
##  Forward  selection procedure
##
## F.in:  4
##
## Last  4  iterations:
##      (Intercept) XTakers XIncome XYears XPublic XExpend XRank
## [1,]           1       0       0      0       0       0     1 162.400
## [2,]           1       0       0      0       0       1     1  22.820
## [3,]           1       0       0      1       0       1     1  16.320
## [4,]           1       0       1      1       0       1     1   5.595
```

Example: SAT Scores (Con’d)

```
full=lm(Y~SATdata$Rank) #full model
reduced = lm(Y ~ 1) #reduced model
anova(reduced,full,test='F')
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ 1
## Model 2: Y ~ SATdata$Rank
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     48 245376
## 2     47  55079  1    190297 162.38 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
full=lm(Y~SATdata$Rank+SATdata$Expend) #full model
reduced = lm(Y~SATdata$Rank) #reduced model
anova(reduced,full,test='F')
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ SATdata$Rank
## Model 2: Y ~ SATdata$Rank + SATdata$Expend
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1     47 55079
## 2     46 36815  1     18265 22.822 1.849e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: SAT Scores (Con'd)

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

```
##
## Call:
## mle.stepwise(formula = Y ~ X, type = "Backward", f.in = 4, f.out = 4)
##
##
##  Backward  selection procedure
##
## F.out:  4
##
## Last  2  iterations:
##       (Intercept) XTakers XIncome XYears XPublic XExpend XRank
## [1,]            1       0       1      1       1       1     1 0.000818
## [2,]            1       0       1      1       0       1     1 0.770000
```

Example: SAT Scores (Con'd)

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

```
##
## Call:
## mle.stepwise(formula = Y ~ X, type = "Stepwise", f.in = 4, f.out = 4)
##
##
##  Stepwise  selection procedure
##
## F.in:  4
## F.out:  4
##
## Last  5  iterations:
##      (Intercept) XTakers XIncome XYears XPublic XExpend XRank
## [1,]           1       0       0      0       0       0     1 1.624e+02
## [2,]           1       0       0      0       0       1     1 2.282e+01
## [3,]           0       0       0      0       0       1     1 1.551e-03
## [4,]           0       0       0      0       1       1     1 1.102e+01
## [5,]           0       0       0      1       1       1     1 6.412e+00
```

Sequential Variable Selection by Using Other Statistical Measures

Recall that in order to accomplish the sequential variable selection, we need to determine Criterion A and Criterion B in forward selection steps and backward elimination steps.

Usually the criterion depends on statistical measures.

We start from using F-statistic as the measure and we adopt the F-test idea to determine Criterion A and Criterion B.

However, other statistical measures can also be used to determine Criterion A and Criterion B. But the idea is different from the F-test.

Idea of Variable Selection by Using Other Statistical Measures

Recall that we pursue good fitting for a MLR model, and SSE (deviance) measures the goodness of fit for MLR.

Based on the definition of SSE (deviance), the smaller the SSE (deviance) is, the better fitting of a model.

Hence, one goal for MLR is to find an appropriate model with smaller SSE.

However, the goal for variable selection is to find a small number of explanatory variables if possible to construct MLR.

The above two goals are contradicted since more explanatory variables results in the decrease in SSE.

Hence, an appropriate variable selection criterion should provide a compromise between how well the model fits the data and the number of explanatory variables $\Rightarrow$ a standard to determine statistical measures.

Idea of Variable Selection by Using Adjusted R-squared

Adjusted R-squared can be considered as one possible statistical measure for variable selection. For the following MLR model

$$\mu\{Y|X_1, \cdots, X_j\} = \beta_0 + \beta_1 X_1 + \cdots + \beta_j X_j, \text{ we have}$$

$$\text{Adjusted } R^2 = 1 - \frac{\text{SSE}/\{n-(j+1)\}}{\text{SST}/(n-1)}.$$

For two candidate models, if their number of the explanatory variables j is the same, then the model with smaller SSE, or equivalently **larger adjusted R-squared**, is preferred.

For two candidate models, if their SSE is the same, then the model with smaller number of explanatory variables j, or equivalently **larger adjusted R-squared**, is preferred.

Hence, the variable selection criterion based on adjusted R-squared is: the model with **larger adjusted R-squared is preferred**.

Idea of Variable Selection by Using AIC and BIC

AIC (Akaike Information Criterion) and BIC (Bayesian) can also be considered. For the following MLR model

$$\mu\{Y|X_1,\cdots,X_j\} = \beta_0 + \beta_1 X_1 + \cdots + \beta_j X_j, \text{ we define}$$

$$\begin{aligned}\text{AIC } &= n\left\{\log\left(\frac{\text{SSE}}{n}\right) + 1 + \log(2\pi)\right\} + 2 \times (j+1) \text{ and}\\ \text{BIC } &= n\left\{\log\left(\frac{\text{SSE}}{n}\right) + 1 + \log(2\pi)\right\} + \log(n) \times (j+1).\end{aligned}$$

For two candidate models, if their number of the explanatory variables j is the same, then the model with smaller SSE, or equivalently **smaller AIC (or BIC)**, is preferred.

For two candidate models, if their SSE is the same, then the model with smaller number of explanatory variables j, or equivalently **smaller AIC (or BIC)**, is preferred.

Hence, the variable selection criterion based on AIC (or BIC) is: the model with **smaller AIC (or BIC) is preferred**.

Idea of Variable Selection by Using Other Statistical Measures

Adjusted R-squared, AIC and BIC all compromise how well the model fits the data (SSE) and the number of explanatory variables (j).

Compared to AIC, BIC assigns a larger weight to the number of explanatory variables k in its expression (usually the sample size n is large such that $\log(n) > 2$).

Hence BIC usually prefers the model with less explanatory variables compared to AIC.

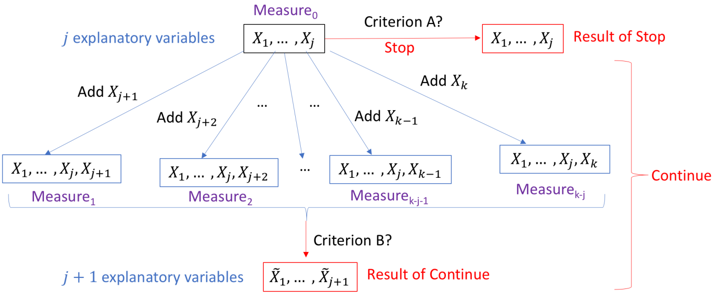
In the following, we will introduce the sequential variable selection procedure based on one of the above three measures.

Let Measure $= -1 \times \mathrm{Adjusted}\ R^2$, or Measure $=$ AIC, or Measure $=$ BIC in the following.

Then, the variable selection criterion based on "Measure" is: the model with **smaller "Measure" is preferred**.

Forward Selection Step by Using Other Statistical Measures

Forward Selection Step for j Explanatory Variables

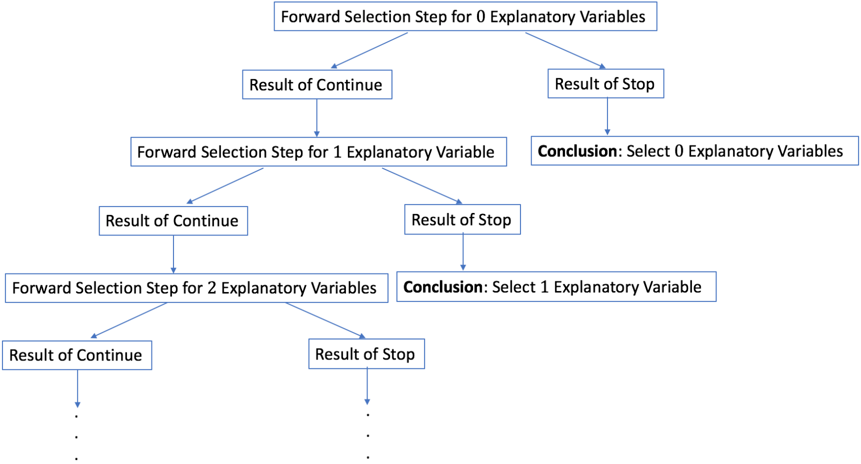


Suppose we have k explanatory variables $X_1, \dots, X_k$ in total.

Criterion A: if $\min\{\text{Measure}_1, \text{Measure}_2, \cdots, \text{Measure}_{k-j-1}, \text{Measure}_{k-j}\} > \text{Measure}_0$ or $j = k$, then Stop; otherwise Continue.

Criterion B: $\tilde{X}_1, \cdots, \tilde{X}_{j+1}$ are those variables such that the corresponding "Measure" is the smallest.

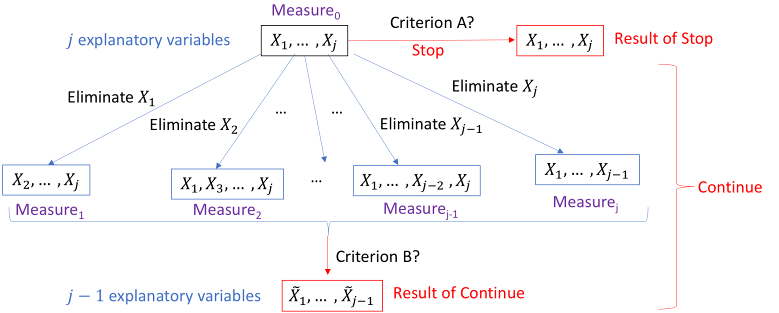
Complete Forward Selection Procedure by Using Other Statistical Measures


Forward Selection Step for 0 Explanatory Variables
Result of Continue
Result of Stop
Forward Selection Step for 1 Explanatory Variable
Conclusion: Select 0 Explanatory Variables
Result of Continue
Result of Stop
Forward Selection Step for 2 Explanatory Variables
Conclusion: Select 1 Explanatory Variable
Result of Continue
Result of Stop

Keep doing on the above procedures, until the first time we obtain the **Result of Stop**.

Backward Elimination Step by Using Other Statistical Measures

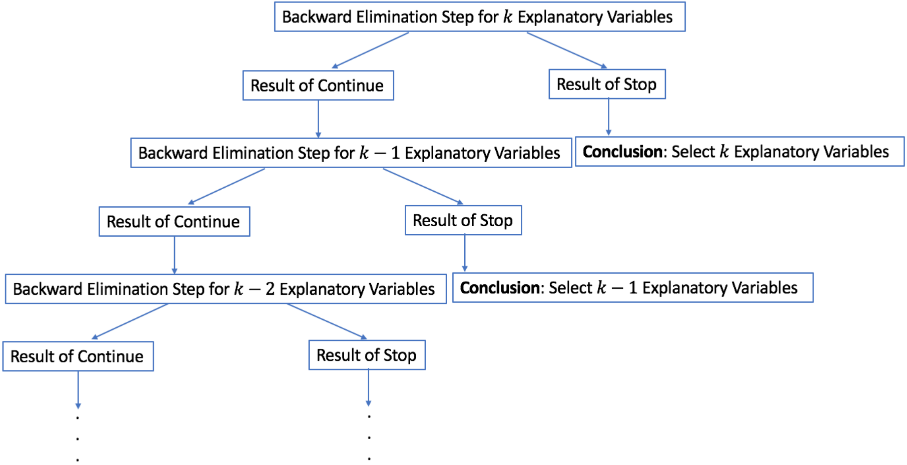
Backward Elimination Step for j Explanatory Variables



Criterion A: if $\min\{\mathrm{Measure}_1, \mathrm{Measure}_2, \cdots, \mathrm{Measure}_{j-1}, \mathrm{Measure}_j\} > \mathrm{Measure}_0$ or $j = 0$, then Stop; otherwise Continue.

Criterion B: $\tilde{X}_1, \cdots, \tilde{X}_{j+1}$ are those variables such that the corresponding "Measure" is the smallest.

Complete Backward Elimination Procedure by Using Other Statistical Measures



Backward Elimination Step for k Explanatory Variables
Result of Continue
Result of Stop
Backward Elimination Step for $k-1$ Explanatory Variables
Conclusion: Select k Explanatory Variables
Result of Continue
Result of Stop
Backward Elimination Step for $k-2$ Explanatory Variables
Conclusion: Select $k-1$ Explanatory Variables
Result of Continue
Result of Stop

Keep doing on the above procedures, until the first time we obtain the **Result of Stop**.

Stepwise Selection Procedure by Other Statistical Measures

This constitutes a combination of backward elimination and forward selection. Each step consists of the following steps. (Any starting model can be chosen, e.g., the model with no explanatory variables, or the model with all the explanatory variables.)

1. Do one Forward Selection Step.

2. Do one Backward Elimination Step.

Repeat Steps 1 and 2 until no explanatory variables can be added or removed.

Remark: The above three methods will sometimes lead to different variable selection results.

Example: SAT Scores (Con’d)

```
X<-data.frame(X)
fit<-lm(Y~.,data=X)
#install.packages('MASS')
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 3.3.2
```

```
#Backward AIC
a=stepAIC(fit,direction="backward",data=X)
```

```
summary(a)
```

```
##
## Call:
## lm(formula = Y ~ Income + Years + Expend + Rank, data = X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.005 -15.548   1.759  13.534  51.808
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -255.2894    88.6074  -2.881 0.006103 **
## Income         0.2412     0.1020   2.365 0.022479 *
## Years         18.9447     5.1563   3.674 0.000644 ***
## Expend         3.3851     0.7775   4.354 7.87e-05 ***
## Rank           9.3764     0.6589  14.229  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.34 on 44 degrees of freedom
## Multiple R-squared:  0.9023, Adjusted R-squared:  0.8934
## F-statistic: 101.6 on 4 and 44 DF,  p-value: < 2.2e-16
```

Example: SAT Scores (Con'd)

```r
#Backward BIC
n=length(Y)
a=stepAIC(fit,direction="backward",data=X, k=log(n))
```

```r
summary(a)
```

```
##
## Call:
## lm(formula = Y ~ Income + Years + Expend + Rank, data = X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.005 -15.548   1.759  13.534  51.808
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -255.2894    88.6074  -2.881 0.006103 **
## Income         0.2412     0.1020   2.365 0.022479 *
## Years         18.9447     5.1563   3.674 0.000644 ***
## Expend         3.3851     0.7775   4.354 7.87e-05 ***
## Rank           9.3764     0.6589  14.229  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.34 on 44 degrees of freedom
## Multiple R-squared:  0.9023, Adjusted R-squared:  0.8934
## F-statistic: 101.6 on 4 and 44 DF,  p-value: < 2.2e-16
```

Example: SAT Scores (Con'd)

```r
#Stepwise AIC
a=stepAIC(fit,direction="both",data=X)
```

```r
summary(a)
```

```
##
## Call:
## lm(formula = Y ~ Income + Years + Expend + Rank, data = X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.005 -15.548   1.759  13.534  51.808
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -255.2894    88.6074  -2.881 0.006103 **
## Income         0.2412     0.1020   2.365 0.022479 *
## Years         18.9447     5.1563   3.674 0.000644 ***
## Expend         3.3851     0.7775   4.354 7.87e-05 ***
## Rank           9.3764     0.6589  14.229  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.34 on 44 degrees of freedom
## Multiple R-squared:  0.9023, Adjusted R-squared:  0.8934
## F-statistic: 101.6 on 4 and 44 DF,  p-value: < 2.2e-16
```

Example: SAT Scores (Con’d)

```
#Forward AIC
fit<-lm(Y~1,data=X)
a=stepAIC(fit,direction="forward",scope=list(lower=~1,upper=~ Takers + Income + Years + Public + Expend
                                             + Rank))
```

```
summary(a)
```

```
##
## Call:
## lm(formula = Y ~ Rank + Expend + Years + Income, data = X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.005 -15.548   1.759  13.534  51.808
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -255.2894    88.6074  -2.881 0.006103 **
## Rank           9.3764     0.6589  14.229  < 2e-16 ***
## Expend         3.3851     0.7775   4.354 7.87e-05 ***
## Years         18.9447     5.1563   3.674 0.000644 ***
## Income         0.2412     0.1020   2.365 0.022479 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.34 on 44 degrees of freedom
## Multiple R-squared:  0.9023, Adjusted R-squared:  0.8934
## F-statistic: 101.6 on 4 and 44 DF,  p-value: < 2.2e-16
```

Variable Selection Among All Subsets

Variable selection involves choosing a subset of explanatory variables to construct the multiple linear regression model.

For instance, consider we have two explanatory variables X_1 and X_2 in total. The subsets are

$$\emptyset, \{X_1\}, \{X_2\}, \text{ and } \{X_1, X_2\}.$$

Different subsets of explanatory variables determine different models. We call those models candidate models.

The methods of variable selection among all subsets fit all possible candidate models (e.g., all the four subsets in the above example).

A value of the statistical measure (e.g., adjusted R-squared, AIC and BIC) is computed for each candidate model.

Let Measure $= -1 \times \text{Adjusted } R^2$, or Measure $=$ AIC, or Measure $=$ BIC.

Then, the model with the **smallest "Measure" is preferred**.

Variable Selection Among All Subsets (Con'd)

The variable selection among all subsets is a search through all possible subsets of variables, in order to obtain the resulting model with the smallest "Measure", which is **an alternative method for variable selection** and is **different from** the sequential variable selection techniques.

The sequential techniques is a sequential search by either adding or removing a single explanatory variable from the current candidate model at each step. The "Measure" is used to determine Criterion A and Criterion B in forward selection steps and backward elimination steps.

Idea of Variable Selection by Using C_p-Statistic

For the following MLR model with **all** the explanatory variables

$$\mu\{Y|X\} = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k, \text{ where } X = (X_1, \cdots, X_k),$$

we can compute $\hat{\sigma}^2_{\text{all}}$.

We consider a new statistical measure C_p-statistic. For the following candidate model

$$\mu\{Y|X_1, \cdots, X_j\} = \beta_0 + \beta_1 X_1 + \cdots + \beta_j X_j,$$

we can compute its SSE. The C_p-statistic is defined as

$$C_p = (j+1) + (n-j-1)\frac{\text{SSE}/(n-j-1) - \hat{\sigma}^2_{\text{all}}}{\hat{\sigma}^2_{\text{all}}} = \frac{\text{SSE}}{\hat{\sigma}^2_{\text{all}}} + 2(j+1) - n.$$

Idea of Variable Selection by Using C_p-Statistic (Con'd)

For two candidate models, if their number of the explanatory variables j is the same, then the model with smaller SSE, or equivalently **smaller** C_p, is preferred.

For two candidate models, if their SSE is the same, then the model with smaller number of explanatory variables j, or equivalently **smaller** C_p, is preferred.

Hence, the variable selection criterion based on C_p-statistic is: the model with **smaller** C_p **is preferred**.

C_p-statistic also compromises how well the model fits the data (SSE) and the number of explanatory variables (j).

The variable selection among all subsets by using C_p-statistic is a search through all possible subsets of variables, in order to obtain the resulting model with the smallest C_p.

Cross Validation for Variable Selection Results

The above variable selection techniques will sometimes lead to different variable selection results.

Cross validation is a process of "checking" the selected model.

The original dataset is split into two parts:

1. a training dataset: for model fitting and variable selection;

2. a test dataset (hold-out sample): for checking selection results and finding the best model.

This procedure allows the performance of a model to be gauged on data that were not used to fit the model.

Mean Squared Prediction Error

We usually use the predictive performance of the test dataset to check the selection results and to find the best model.

A measure of predictive ability is given by the mean squared prediction error (MSPE):

$$\mathrm{MSPE} = \frac{1}{n_{\mathrm{test}}} \sum_{\ell=1}^{n_{\mathrm{test}}} (Y_\ell - \hat{Y}_\ell)^2, \text{ where}$$

- Y_ℓ is the observed ℓ-th response from the test dataset.
- $\hat{Y}_\ell$ is the predicted value of Y_ℓ based on the regression model constructed by the training dataset.
- n_{test} is the number of the observations of the test dataset.

The best model is the model with the smallest MSPE.

Example: SAT Scores (Con'd)

```
#install.packages('leaps')
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 3.3.2
```

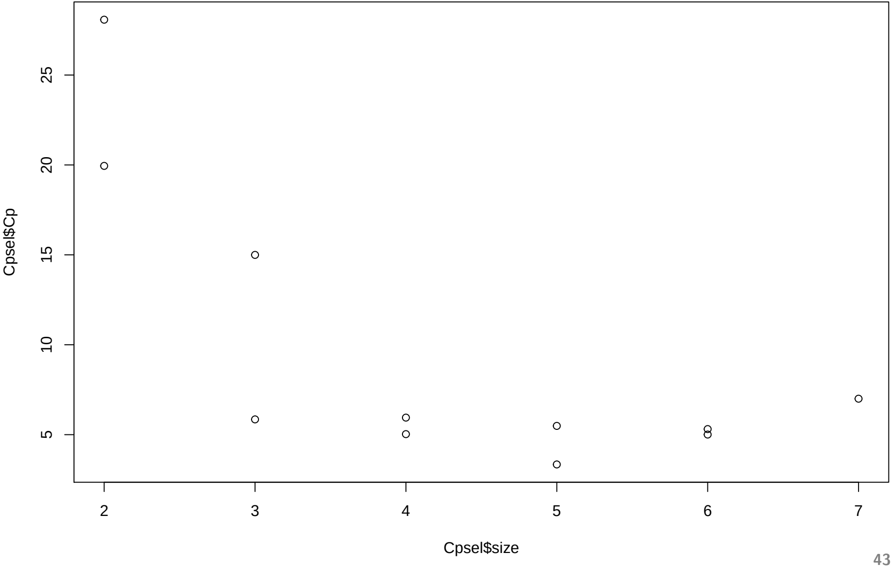
```
#help(leaps) #get some info on the leaps command
X=cbind(log(SATdata[,3]),SATdata[,4:8])
colnames(X)=c('x1','x2','x3','x4','x5','x6')
Y=SATdata[,2]
length(X[,1])
```

```
## [1] 49
```

```
#Training data
Xtraining=X[1:40,]
Ytraining=Y[1:40]
#Test data
Xtest=X[41:49,]
Ytest=Y[41:49]
```

Example: SAT Scores (Con'd)

```
#Training data
Cpsel=leaps(Xtraining,Ytraining,method="Cp",nbest=2)
plot(Cpsel$size,Cpsel$Cp)
```



Example: SAT Scores (Con'd)

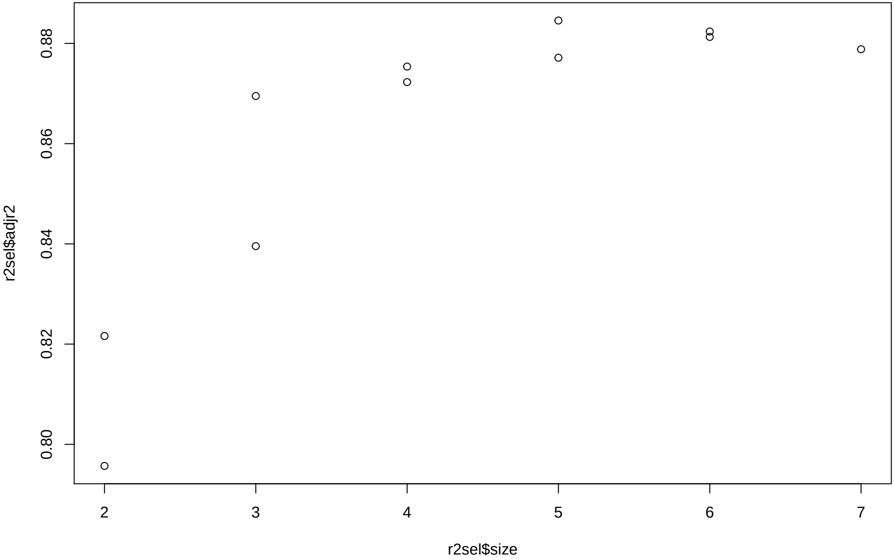
```r
cbind(Cpsel$which, Cpsel$size, Cpsel$Cp)
```

```
##   1 2 3 4 5 6
## 1 1 0 0 0 0 0 2 19.946845
## 1 0 0 0 0 0 1 2 28.076894
## 2 1 0 0 0 1 0 3  5.848913
## 2 1 0 0 1 0 0 3 14.996836
## 3 1 0 0 0 1 1 4  5.029187
## 3 1 0 0 1 1 0 4  5.948021
## 4 1 0 1 0 1 1 5  3.343196
## 4 1 0 1 1 1 0 5  5.487590
## 5 1 0 1 1 1 1 6  5.006153
## 5 1 1 1 0 1 1 6  5.310103
## 6 1 1 1 1 1 1 7  7.000000
```

```r
#Variable Selection Result: size=5, 1 0 1 0 1 1, x1,x3,x5,x6
```

Example: SAT Scores (Con'd)

```
r2sel=leaps(Xtraining,Ytraining,method="adjr2",nbest=2)
plot(r2sel$size,r2sel$adjr2)
```



Example: SAT Scores (Con’d)

```r
cbind(r2sel$which, r2sel$size, r2sel$adjr2)
```

```
##   1 2 3 4 5 6
## 1 1 0 0 0 0 0 2 0.8216154
## 1 0 0 0 0 0 1 2 0.7956931
## 2 1 0 0 0 1 0 3 0.8695092
## 2 1 0 0 1 0 0 3 0.8395530
## 3 1 0 0 0 1 1 4 0.8753745
## 3 1 0 0 1 1 0 4 0.8722821
## 4 1 0 1 0 1 1 5 0.8845738
## 4 1 0 1 1 1 0 5 0.8771504
## 5 1 0 1 1 1 1 6 0.8823800
## 5 1 1 1 0 1 1 6 0.8812968
## 6 1 1 1 1 1 1 7 0.8788383
```

```r
#Variable Selection Result: size=5, 1 0 1 0 1 1, x1,x3,x5,x6
```

Example: SAT Scores (Con’d)

```
#Cross Validation
fitall=lm(Ytraining~.,data=Xtraining)
summary(fitall)
```

```
##
## Call:
## lm(formula = Ytraining ~ ., data = Xtraining)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.963 -12.873   2.109  10.679  44.054
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 484.882137 264.914499   1.830   0.0762 .
## x1          -35.153528  13.983161  -2.514   0.0170 *
## x2            0.009766   0.124502   0.078   0.9380
## x3           11.750673   5.979973   1.965   0.0579 .
## x4            0.278738   0.500545   0.557   0.5814
## x5            2.669543   0.882795   3.024   0.0048 **
## x6            3.637652   2.306533   1.577   0.1243
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.28 on 33 degrees of freedom
## Multiple R-squared:  0.8975, Adjusted R-squared:  0.8788
## F-statistic: 48.15 on 6 and 33 DF,  p-value: 6.3e-15
```

```
#MPSEfull
mean((Ytest-predict(fitall,Xtest))^2)
```

```
## [1] 1558.079
```

Example: SAT Scores (Con’d)

```
fitselect=lm(Ytraining~.,data=Xtraining[,c(1,3,5,6)])
summary(fitselect)
```

```
##
## Call:
## lm(formula = Ytraining ~ ., data = Xtraining[, c(1, 3, 5, 6)])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -48.60 -12.90   2.60  10.56  43.06
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 438.7462   232.7437   1.885  0.06774 .
## x1          -31.3364    11.2500  -2.785  0.00857 **
## x3           11.4040     5.7976   1.967  0.05715 .
## x5            2.8932     0.7569   3.822  0.00052 ***
## x6            4.4030     1.8371   2.397  0.02202 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.79 on 35 degrees of freedom
## Multiple R-squared:  0.8964, Adjusted R-squared:  0.8846
## F-statistic: 75.72 on 4 and 35 DF,  p-value: < 2.2e-16
```

```
#MPSEselect
mean((Ytest-predict(fitselect,Xtest[,c(1,3,5,6)]))^2)
```

```
## [1] 1417.981
```

Multicollinearity – Motivating Example

In a study of primary school reading ability, an educator wished to relate a measure of a student's reading ability (Y) to age (X_1) and gender (female/male).

Gender is a categorical variable with two categories.

To allow for it in MLR, we can consider the following two possible indicator variables.

$$X_2 = 1 \text{ if female; } 0 \text{ otherwise;}$$

$$X_3 = 1 \text{ if male; } 0 \text{ otherwise.}$$

It is worth noting that $X_2 + X_3 = 1$.

Actually we only have the information of X_1 and X_2. The information of X_3 is redundant since $X_3 = 1 - X_2$, which is included in X_2. This phenomenon is called **multicollinearity**.

As a consequence, drop one of the indicator variables (X_2 or X_3) and fit the model.

Multicollinearity

In general, suppose we have explanatory variables $X_1, \cdots, X_k$ in total, if one of the explanatory variables X_j have the situation

$$X_j \approx c_0 + c_1 X_1 + \cdots + c_{j-1} X_{j-1} + c_{j+1} X_{j+1} + \cdots + c_k X_k,$$

for some constants $c_0, c_1, \cdots, c_{j-1}, c_{j+1}, \cdots, c_k$ (which are not all equal to 0), then those explanatory variables are said to be multicollinear.

The MLR based on explanatory variables $X_1, \cdots, X_k$ is said to have a multicollinearity problem.

Multicollinearity can result in

1. For the $n \times (k+1)$ design matrix $\mathbb{X}$, the matrix inverse $(\mathbb{X}^\top \mathbb{X})^{-1}$ may not exist (cannot be computed), and thus the LS estimates may not be obtained.

2. Even if sometimes $(\mathbb{X}^\top \mathbb{X})^{-1}$ still exists, the LS estimates are highly unstable and imprecise $\Rightarrow$ SEs of the estimators are large and a lot of hypothesis testing results are not significant.

Variance Inflation Factors (VIF)

A quantitative measure of the multicollinearity is given by the variance inflation factors (VIF).

The VIF associated with the j-th explanatory variable is:

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the R-squared by regressing X_j on $X_1, \cdots, X_{j-1}, X_{j+1}, \cdots, X_k$.

The "rule of thumb" cut-off for VIF is 10.

If one of VIF_j's is larger than 10, then the MLR based on explanatory variables $X_1, \cdots, X_k$ has a multicollinearity problem.

In this case, all the parameter estimates associated with explanatory variables with large VIF's will have low precision (very large standard errors $\Rightarrow$ insignificant).

How to deal with the multicollinearity problem?

Backward Elimination to Deal with Multicollinearity

An explanatory variable X_j with $\mathrm{VIF}_j > 10$ should be eliminated.

However, usually two or more VIFs are larger than 10.

Can we eliminate all of them in the model? NO!

We can only drop one at each time. This is very similar to the backward elimination for variable selection.

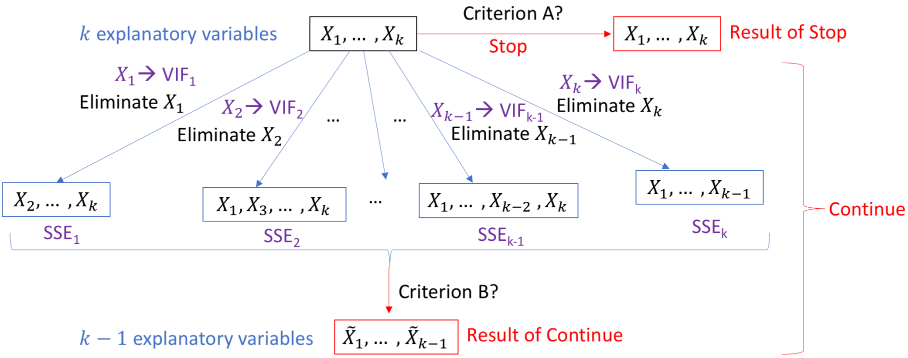
In this case, which one of the explanatory variables should be eliminated?

Intuitively, the criterion is: the resulting model after dropping the explanatory variable **has the best fitting (smallest SSE or deviance).**

Later in this course, we will introduce an alternative approach, principal components analysis (PCA), to deal with multicollinearity.

Backward Elimination Step

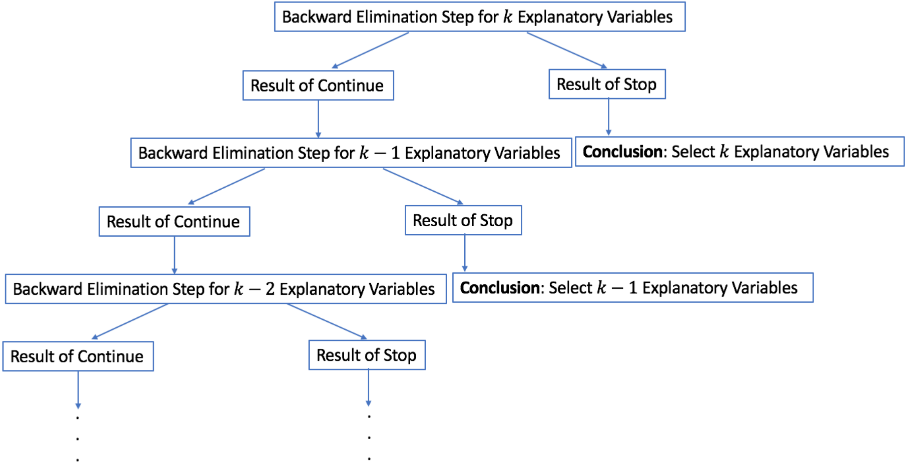
Backward Elimination Step for k Explanatory Variables



Criterion A: if max$\{\text{VIF}_1, \text{VIF}_2, \cdots, \text{VIF}_{k-1}, \text{VIF}_k\} < 10$ or $k = 0$, then Stop; otherwise Continue.

Criterion B: $\tilde{X}_1, \cdots, \tilde{X}_{j-1}$ are those variables such that under the condition $\text{VIF} > 10$, the corresponding SSE is the smallest.

Complete Backward Elimination Procedure to Deal with Multicollinearity


Backward Elimination Step for k Explanatory Variables
Result of Continue
Result of Stop
Backward Elimination Step for $k-1$ Explanatory Variables
Conclusion: Select k Explanatory Variables
Result of Continue
Result of Stop
Backward Elimination Step for $k-2$ Explanatory Variables
Conclusion: Select $k-1$ Explanatory Variables
Result of Continue
Result of Stop

Keep doing on the above procedures, until the first time we obtain the **Result of Stop**.

Example: SAT Scores (Con'd)

```
Y<-SATdata[,2]
X<-SATdata[,-c(1,2)]
fit=lm(Y~.,data=X)
summary(fit)
```

```
## 
## Call:
## lm(formula = Y ~ ., data = X)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.388 -13.268   1.758  13.496  51.024
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -203.92618  192.87664  -1.057  0.29642
## Takers         0.01833    0.64099   0.029  0.97732
## Income         0.18058    0.14807   1.220  0.22944
## Years         16.53592    5.95782   2.775  0.00819 **
## Public        -0.44299    0.52053  -0.851  0.39957
## Expend         3.72998    0.88504   4.214  0.00013 ***
## Rank           9.78937    1.93456   5.060 8.75e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 23.68 on 42 degrees of freedom
## Multiple R-squared:  0.904,  Adjusted R-squared:  0.8903
## F-statistic: 65.95 on 6 and 42 DF,  p-value: < 2.2e-16
```

Example: SAT Scores (Con'd)

```r
#install.packages('car')
library(car)
```

```
## Warning: package 'car' was built under R version 3.3.2
```

```r
vif(fit)
```

```
##    Takers    Income     Years    Public    Expend      Rank
## 17.399186  3.201245  1.469017  2.173352  1.535438 13.917392
```

```r
#Continue
#Try dropping Takers
X1=X[,-1]
fit1=lm(Y~.,data=X1)
deviance(fit1)
```

```
## [1] 23546.74
```

```r
#Try dropping Rank
X2=X[,-6]
fit2=lm(Y~.,data=X2)
deviance(fit2)
```

```
## [1] 37901.82
```

```r
#Based on SSE, we choose to eliminate TAKERS. The resulting VIFs after elimination are:
vif(fit1)
```

```
##   Income    Years   Public   Expend     Rank
## 2.328161 1.446115 2.061058 1.531747 2.311335
```

```r
#Stop
```