# Regression Model Selection (continued)

## 4. Sequential F-tests

```
# Forward and backward selection algorithms with partial F-tests at each step.
# This tool is in the package SignifReg (significance testing in regression model building)

> install.packages("SignifReg")
> library(SignifReg)

> data(longley)
> names(longley)
[1] "GNP.deflator" "GNP"          "Unemployed"   "Armed.Forces" "Population"  "Year"   "Employed"
```

# Longley's macroeconomic data set is pre-loaded in R. We'll use it to predict the unemployment rate.

```
> null = lm( Unemployed ~ 1, data=longley )
> SignifReg( null )

Call: lm(formula = Unemployed ~ Population + GNP + Year, data = longley)

Coefficients:
(Intercept)   Population          GNP         Year
 -1.840e+05    3.924e+01   -6.599e+00    9.326e+01
```

**# Of all available variables, R selected only the Population, Gross National Product and Year. By default, each F-test is at the α-to-enter level α = 0.05. With a higher α, more variables will be considered significant and included into the model. With a lower α, it is harder to overcome a significance criterion, and the model will contain fewer variables.**

```
> SignifReg( null, alpha=0.2 )

Call:  lm(formula = Unemployed ~ Population + GNP + Year + Armed.Forces + GNP.deflator, data =
longley)

Coefficients:
 (Intercept)    Population           GNP          Year  Armed.Forces  GNP.deflator
  -1.524e+05     3.510e+01    -6.484e+00     7.686e+01    -2.714e-01     9.649e+00

> SignifReg( null, alpha=0.001 )

Call:  lm(formula = Unemployed ~ 1, data = longley)

Coefficients:
(Intercept)
      319.3
```

**# Backward elimination is similar, and we can set a desired α-to-remove level.**

```
> full = lm( Unemployed ~ ., data=longley )
> SignifReg( full, direction="backward" )

Call:
lm(formula = Unemployed ~ GNP + Population + Year, data = longley)

Coefficients:
(Intercept)          GNP   Population         Year
 -1.840e+05   -6.599e+00    3.924e+01    9.326e+01
```

**# In order to see all the steps of this variable selection, use option trace = TRUE**

```
> SignifReg( full, direction="backward", trace=TRUE )

Call:  lm(formula = Unemployed ~ ., data = longley)

Coefficients:
 (Intercept)  GNP.deflator           GNP  Armed.Forces    Population          Year      Employed
  -1.655e+05     3.815e+00    -3.397e+00    -4.287e-01     1.044e+01     8.582e+01    -3.244e+01
```

```
               RSS         AIC       BIC       adj.rsq max_pvalue alpha_cut-off Bonferroni FDR
<none>         1342.97639  132.28691 138.46762 0.98291 0.25847    FALSE         FALSE      FALSE
- GNP.deflator 1560.0956   132.68465 138.09277 0.98213 0.53545    FALSE         FALSE      FALSE
- GNP          3646.0413   146.26697 151.67509 0.95825 0.39287    FALSE         FALSE      FALSE
- Armed.Forces 5915.34581  154.0095  159.41762 0.93226 0.43263    FALSE         FALSE      FALSE
- Population   1573.97601  132.82638 138.2345  0.98198 0.58404    FALSE         FALSE      FALSE
- Year         11193.49526 164.21402 169.62214 0.87181 0.37876    FALSE         FALSE      FALSE
- Employed     3896.1284   147.32843 152.73655 0.95538 0.06106    FALSE         FALSE      FALSE

Call:
lm(formula = Unemployed ~ GNP.deflator + GNP + Armed.Forces +
    Population + Year, data = longley)

Coefficients:
 (Intercept)  GNP.deflator           GNP  Armed.Forces    Population          Year
  -1.524e+05     9.649e+00    -6.484e+00    -2.714e-01     3.510e+01     7.686e+01



               RSS         AIC       BIC       adj.rsq max_pvalue alpha_cut-off Bonferroni FDR
<none>         3896.1284   147.32843 152.73655 0.95538 0.06106    FALSE         FALSE      FALSE
- GNP.deflator 5630.44594  151.21972 155.85525 0.94138 0.05403    FALSE         FALSE      FALSE
- GNP          36816.1171  181.26367 185.8992  0.61672 0.59076    FALSE         FALSE      FALSE
- Armed.Forces 6310.78085  153.04485 157.68038 0.9343  0.11314    FALSE         FALSE      FALSE
- Population   9166.33965  159.01731 163.65284 0.90457 0.71433    FALSE         FALSE      FALSE
- Year         12142.87902 163.51658 168.15212 0.87358 0.24054    FALSE         FALSE      FALSE

Call:
lm(formula = Unemployed ~ GNP + Armed.Forces + Population + Year,
    data = longley)

Coefficients:
 (Intercept)           GNP  Armed.Forces    Population          Year
  -1.903e+05    -5.799e+00    -2.694e-01     2.625e+01     9.713e+01



               RSS         AIC       BIC       adj.rsq max_pvalue alpha_cut-off Bonferroni FDR
<none>         5630.44594  151.21972 155.85525 0.94138 0.05403    FALSE         FALSE      FALSE
- GNP          39044.27121 180.20384 184.06678 0.62739 0.72154    FALSE         FALSE      FALSE
- Armed.Forces 8010.88252  154.86151 158.72446 0.92355 0.00083    TRUE          TRUE       TRUE
- Population   9283.92097  157.22124 161.08419 0.9114  0.00209    TRUE          TRUE       TRUE
- Year         25320.2248  173.27435 177.1373  0.75836 0.48746    FALSE         FALSE      FALSE

Call:
lm(formula = Unemployed ~ GNP + Population + Year, data = longley)

Coefficients:
(Intercept)          GNP   Population         Year
 -1.840e+05   -6.599e+00    3.924e+01    9.326e+01



Call:
lm(formula = Unemployed ~ GNP + Population + Year, data = longley)

Coefficients:
(Intercept)          GNP   Population         Year
 -1.840e+05   -6.599e+00    3.924e+01    9.326e+01
```

**# Since any stepwise variable selection algorithm involves multiple tests, we can control the familywise error rate by using the Bonferroni correction**

**> SignifReg( full, direction="backward", correction="Bonferroni" )**

**# Compare results of stepwise variable selection without the Bonferroni correction and with it.**
**# Bonferroni reduces the alpha levels and makes it harder for variables to enter the model.**

```
> SignifReg( null, alpha=0.2 )

Call:
lm(formula = Unemployed ~ Population + GNP + Year + Armed.Forces +
    GNP.deflator, data = longley)

Coefficients:
 (Intercept)    Population           GNP          Year  Armed.Forces  GNP.deflator
  -1.524e+05     3.510e+01    -6.484e+00     7.686e+01    -2.714e-01     9.649e+00

> SignifReg( null, alpha=0.2, correction="Bonf" )

Call:
lm(formula = Unemployed ~ Population + GNP + Year, data = longley)

Coefficients:
(Intercept)   Population          GNP         Year
 -1.840e+05    3.924e+01   -6.599e+00    9.326e+01
```

### 5. Visualization – scatterplot matrix

**# Scatterplot matrix – a way to visualize relations between the response and predictor variables.**
**It is used to show (1) whether there is a relation between Y and each $X_j$, (2) whether this**
**relation is linear # or nonlinear, (3) whether there may be strong multicollinearity.**
**This particular data set is known for its strong multicollinearity. Therefore, careful variable**
**selection is really necessary here.**

```
> par(mfrow=c(7,7))
> plot(longley)
```