Chapter 11: Variable Selection

Copyright: Prof Legesse Kassa Debusho, UNISA, South Africa
Prof Ziv Shkedy, University of Hasselt, Belgium

The course material development sponsored by the Flemish Interuniversity Council (Developmental Cooperation VLIR-UOS) through the Cross cutting project on biostatistics

October 20, 2016

Table of contents

1. Introduction
2. Hierarchical Models
3. Stepwise Procedures
   - Backward Elimination
   - Forward Selection
   - Stepwise Regression
4. Criterion-based procedures

Introduction

- This chapter deals with model building problems and the selection of the regression model where such topics as
  - miss-specification of predictor variables,
  - selection criteria like Mallows C_p, $PRESS_p$, AIC_P, BIC_P, and
- variable selection methods such as
  - stepwise regression,
  - forward and
  - backward selection procedures are considered.
- By the use of these selection methods, the possible 'best' model will be determined among the alternative predictor variables.
- That is, variable selection is intended to select the 'best' subset of predictors.

Introduction

- If we have, say, p predictor variables, we will be able to have a total of 2^{P-1} possible candidate models with $P-1$ predictor variables from which the best one could be selected.
- And such problems, related with the method of identifying such group of regressors, are known as variable selection problems.

Why do we bother

1. We want to explain the data in the simplest way - redundant predictors should be removed. The smallest model that fits the data is best.
2. Unnecessary predictors will add noise to the estimation of other quantities that we are interested in. Degrees of freedom will be wasted.
3. Collinearity is caused by having too many variables trying to do the same job.
4. Cost: if the model is to be used for prediction, we can save time and/or money by not measuring redundant predictors.

Why do we bother

- Prior to variable selection:
  1. Identify outliers and influential points - maybe exclude them at least temporarily.
  2. Add in any transformations of the variables that seem appropriate.

Hierarchical models

- Some models have a natural hierarchy. For example, in polynomial models, x^2 is a higher order term than x.
- When selecting variables, it is important to respect the hierarchy.
- Lower order terms should not be removed from the model before higher order terms in the same variable.
- There two common situations where this situation arises:
  1. Polynomials models: Consider the model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

Hierarchical models

- Suppose we fit this model and find that the regression summary shows that the term in x is not significant but the term in x^2 is.
- If we then removed the x term, our reduced model would then become

$$y = \beta_0 + \beta_2\, x^2 + \varepsilon$$

but suppose we then made a scale change $x \to x + 1$, then the model would become

$$y = \beta_0 + \beta_2\, a^2 + 2\, \beta_2\, a\, x + \beta_2\, x^2 + \varepsilon.$$

- The first order x term has now reappeared.
- Scale changes should not make any important change to the model but in this case an additional term has been added.

Hierarchical models

- This is not good. This illustrates why we should not remove lower order terms in the presence of higher order terms.
- We would not want interpretation to depend on the choice of scale.
- Removal of the first order term here corresponds to the hypothesis that the predicted response is symmetric about and has an optimum at $x = 0$.
- Often this hypothesis is not meaningful and should not be considered.
- Only when this hypothesis makes sense in the context of the particular problem could we justify the removal of the lower order term.

Hierarchical models

2. Models with interactions: Consider the second order response surface model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 a x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon.$$

- We would not normally consider removing the $x_1 x_2$ interaction term without simultaneously considering the removal of the x_1^2 and x_2^2 terms.
- A joint removal would correspond to the clearly meaningful comparison of a quadratic surface and linear one.
- Just removing the $x_1 x_2$ term would correspond to a surface that is aligned with the coordinate axes.
- This is hard to interpret and should not be considered unless some particular meaning can be attached.

Hierarchical models

- Any rotation of the predictor space would reintroduce the interaction term and, as with the polynomials, we would not ordinarily want our model interpretation to depend on the particular basis for the predictors.

Backward elimination

- This is the simplest of all variable selection procedures and can be easily implemented without special software.
- In situations where there is a complex hierarchy, backward elimination can be run manually while taking account of what variables are eligible for removal.
- The steps are
  1. Start with all the predictors in the model
  2. Remove the predictor with highest p-value greater than α_{crit}
  3. Refit the model and goto 2
  4. Stop when all p-values are less than α_{crit}.
- The α_{crit} is sometimes called the "*p*-to-remove" and does not have to be 5%.
- If prediction performance is the goal, then a 15-20% cut-off may work best, although methods designed more directly for optimal prediction should be preferred.

Forward selection

- This just reverses the backward method and the steps are
  1. Start with no variables in the model.
  2. For all predictors not in the model, check their *p*-value if they are added to the model. Choose the one with lowest *p*-value less than α_{crit}.
  3. Continue until no new predictors can be added.

Stepwise regression

- This is a combination of backward elimination and forward selection.
- It addresses the situation where variables are added or removed early in the process and we want to change our mind about them later.
- At each stage a variable may be added or removed and there are several variations on exactly how this is done.
- Therefore a stepwise method adds/delets regressors one at a time.
- Stepwise procedures are relatively cheap computationally but they do have some drawbacks.
    1. Because of the "one-at-a-time" nature of adding/dropping variables, its possible to miss the "optimal" model.

Stepwise regression

2. The p-values used should not be treated too literally.
    - There is so much multiple testing occurring that the validity is dubious.
    - The removal of less significant predictors tends to increase the significance of the remaining predictors.
    - This effect leads one to overstate the importance of the remaining predictors.
3. The procedures are not directly linked to final objectives of prediction or explanation and so may not really help solve the problem of interest.
    - With any variable selection method, it is important to keep in mind that model selection cannot be divorced from the underlying purpose of the investigation.
    - Variable selection tends to amplify the statistical significance of the variables that stay in the model.

Stepwise regression

- Variables that are dropped can still be correlated with the response.
- It would be wrong to say these variables are unrelated to the response, its just that they provide no additional explanatory effect beyond those variables already included in the model.

4. Stepwise variable selection tends to pick models that are smaller than desirable for prediction purposes.
   To give a simple example, consider the simple regression with just one predictor variable. Suppose that the slope for this predictor is not quite statistically significant. We might not have enough evidence to say that it is related to y but it still might be better to use it for predictive purposes.

Example: state data - US State Facts and Figures

Data sets related to the 50 states of the United States of America. The variables are `population estimate` as of July 1, 1975; `per capita income` (1974); `illiteracy` (1970, percent of population); `life expectancy` in years (1969-71); `murder and non-negligent manslaughter` rate per 100,000 population (1976); `percent high-school graduates` (1970); `mean number of days with min temperature` $<$ 32 degrees (1931-1960) in capital or large city; and `land area` in square miles. The data was collected from US Bureau of the Census. We will take life expectancy as the response and the remaining variables as predictors - a fix is necessary to remove spaces in some of the variable names.

Example: state data - US State Facts and Figures

```
data(state)
statedata <- data.frame(state.x77,
     row.names = state.abb, check.names = T)
MLR.fit.state <- lm(Life.Exp ˜ ., data=statedata)
summary(MLR.fit.state)
Coefficients:
              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 .
```

Example: state data - US State Facts and Figures

```
Area        -7.383e-08  1.668e-06  -0.044   0.9649
---
Signif. codes: 0 '***'0.001 '**'0.01 '*'0.05 '.'0.1 ' '1

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
```

- Which predictors should be included - can you tell from the p-values?
- Looking at the coefficients, can you see what operation would be helpful?
- Does the murder rate decrease life expectancy - that's obvious a priori, but how should these results be interpreted?

Example: state data: Backward method

- At each stage we remove the predictor with the largest $p-value$ over 0.05:

```
MLR.fit.state.1 <- update(MLR.fit.state, . ~ . - Area)
summary(MLR.fit.state.1)

Coefficients:
              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 .
```

Example: state data: Backward method

```
Signif. codes: 0 ’***’0.001 ’**’0.01 ’*’0.05 ’.’0.1 ’ ’1

Residual standard error: 0.7361 on 43 degrees of freedom
Multiple R-squared:  0.7361,    Adjusted R-squared:  0.699
F-statistic: 19.99 on 6 and 43 DF,  p-value: 5.362e-11
```

Example: state data: Backward method

```
MLR.fit.state.2 <- update(MLR.fit.state.1, . ~ . - Illitera
summary(MLR.fit.state.2)

Coefficients:
              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 *
```

Example: state data: Backward method

```
MLR.fit.state.3 <- update(MLR.fit.state.2, . ~ . - Income)
summary(MLR.fit.state.3)

Coefficients:
              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 *
```

Example: state data: Backward method

```
MLR.fit.state.4 <- update(MLR.fit.state.3, . ~ . - Populat
summary(MLR.fit.state.4)

Coefficients:
             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 **
---
Signif. codes: 0 ’***’0.001 ’**’0.01 ’*’0.05 ’.’0.1 ’ ’1
```

Example: state data: Backward method

```
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
```

- The final removal of the Population variable is a close call.
- We may want to consider including this variable if interpretation is aided.
- Notice that the R^2 for the full model of 0.736 is reduced only slightly to 0.713 in the final model.
- Thus the removal of four predictors causes only a minor reduction in fit.

Criterion-based procedures

- If there are p potential predictors, then there are 2 p possible models.
- We fit all these models and choose the best one according to some criterion.
- Clever algorithms such as the "branch-and-bound" method can avoid actually fitting all the models - only likely candidates are evaluated.
- Some criteria are are discussed next.

Akaike Information Criterion and Bayes Information Criterion

- Information criteria is a measure of goodness of fit or uncertainty for the range of values of the data.
- In the context of multiple linear regression, information criteria measures the difference between a given model and the "true" underlying model.
- Akaike (1973) introduced the concept of information criteria as a tool for optimal model selection.
- The Akaike's Information Criteria (AIC) and the Bayesian information criteria (BIC) are functions of the number of observations n, the SSE and the number of parameters p.
- AIC and BIC are based on the maximum likelihood estimates of the model parameters.

Akaike Information Criterion and Bayes Information Criterion

- In maximum likelihood, the idea is to estimate parameters so that, under the model, the probability of the observed data would be as large as possible.
- In general,

  $$AIC = -2\ \log(likelihood) + 2\,p$$

  and

  $$BIC = -2\ \log(likelihood) + p\,log(n)$$

- For linear regression models, the $-2\ \log(likelihood) = n\ \log(SSE/n)$, this is known as the *deviance*.
- We want to minimize AIC or BIC.

Akaike Information Criterion and Bayes Information Criterion

- Larger models will fit better and so have smaller *SSE* but use more parameters. Thus the best choice of model will balance fit with model size.
- *BIC* penalizes larger models more heavily and so will tend to prefer smaller models in comparison to *AIC*.
- *AIC* and *BIC* can be used as selection criteria for other types of model too.

Example: state data - AIC

- We can apply the *AIC* (and optionally the *BIC*) to the state data.
- The function does not evaluate the *AIC* for all possible models but uses a search method that compares models sequentially.
- Thus it bears some comparison to the stepwise method described above but with the advantage that no dubious p-values are used.
- The R code:

```
MLR.fit.state <- lm(Life.Exp ~ ., data=statedata)
step(MLR.fit.state)
```

Example: state data - AIC

```
Start:  AIC=-22.18
Life.Exp ~ Population + Income + Illiteracy + Murder +
           HS.Grad +  Frost + Area

             Df Sum of Sq    RSS     AIC
- Area        1    0.0011 23.298 -24.182
- Income      1    0.0044 23.302 -24.175
- Illiteracy  1    0.0047 23.302 -24.174
<none>                    23.297 -22.185
- Population  1    1.7472 25.044 -20.569
- Frost       1    1.8466 25.144 -20.371
- HS.Grad     1    2.4413 25.738 -19.202
- Murder      1   23.1411 46.438  10.305
```

Example: state data - AIC

```
Step:  AIC=-24.18
Life.Exp ~ Population + Income + Illiteracy + Murder +
           HS.Grad + Frost

             Df Sum of Sq    RSS     AIC
- Illiteracy  1    0.0038 23.302 -26.174
- Income      1    0.0059 23.304 -26.170
<none>                    23.298 -24.182
- Population  1    1.7599 25.058 -22.541
- Frost       1    2.0488 25.347 -21.968
- HS.Grad     1    2.9804 26.279 -20.163
- Murder      1   26.2721 49.570  11.569
```

Example: state data - AIC

```
Step:  AIC=-26.17
Life.Exp ~ Population + Income + Murder + HS.Grad + Frost

             Df Sum of Sq    RSS     AIC
- Income      1     0.006 23.308 -28.161
<none>                    23.302 -26.174
- Population  1     1.887 25.189 -24.280
- Frost       1     3.037 26.339 -22.048
- HS.Grad     1     3.495 26.797 -21.187
- Murder      1    34.739 58.041  17.456
```

Example: state data - AIC

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

             Df Sum of Sq    RSS     AIC
<none>                    23.308 -28.161
- Population  1     2.064 25.372 -25.920
- Frost       1     3.122 26.430 -23.877
- HS.Grad     1     5.112 28.420 -20.246
- Murder      1    34.816 58.124  15.528
```

Example: state data - AIC

```
Call:
lm(formula = Life.Exp ~ Population + Murder + HS.Grad +
       Frost, data = statedata)

Coefficients:
(Intercept) Population   Murder    HS.Grad     Frost
  7.103e+01  5.014e-05  -3.001e-01  4.658e-02 -5.943e-03
```

- The sequence of variable removal is the same as with backward elimination.
- The only difference is the the Population variable is retained.

Adjusted R^2

- Adjusted R^2 called R_a^2. Recall that $R^2 = 1 - SSE/SS_{Total}$.
- Adding a variable to a model can only decrease the SSE and so only increase the R^2 so R^2 by itself is not a good criterion because it would always choose the largest possible model.

$$R_a^2 = 1 - \frac{SSE/(n-p)}{SS_{Total}/(n-1)} = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2) = 1 - \frac{\widehat{\sigma}^2_{model}}{\widehat{\sigma}^2_{null}}.$$

- Adding a predictor will only increase R_a^2 if it has some value.
- Do you see the connection to $\widehat{\sigma}^2$?
- Minimizing the standard error for prediction means minimizing $\widehat{\sigma}^2$ which in term means maximizing R_a^2.

Predicted Residual Sum of Squares (PRESS)

- The Prediction Error Sum of Squares (PRESS) is a statistic which is based on the deleted residuals, say d_i such that, $d_i = y_i - \widehat{y}_{i(i)}$, where $\hat{y}_{i(i)}$ is the predicted value for the ith case when the regression function is fitted without the ith case.
- There are n deleted residuals associated with each model.
- Then the PRESS criterion is the sum of the squares of these deleted residuals:

$$PRESS_p = \sum_{i=1}^{n} (d_i)^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_{(i)})^2$$

- PRESS values could be calculated using the equivalent expression for d_i,

$$d_i = \left(\frac{\widehat{\varepsilon}_i}{1 - h_{ii}} \right)$$

Predicted Residual Sum of Squares (PRESS)

- Then $PRESS_p$ becomes

$$PRESS_p = \sum_{i=1}^{n} \left(\frac{\widehat{\varepsilon}_i}{1 - h_{ii}} \right)^2 ,$$

  where $\widehat{\varepsilon}_i$ is the ordinary residual and h_{ii} is the leverage value such that both are based on all n cases.
- The models with small PRESS values are considered to be 'good' candidate models because the deleted residual d_i is the prediction error when a regression model is fitted without the ith case and $\widehat{y}_{i(i)}$ is used as the predicted value.
- Then small prediction errors involve small d_i values and hence small sum of d_i^2 values.

Predicted Residual Sum of Squares (PRESS)

- Therefore, such models with small PRESS values fit well due to the fact that they have small prediction errors, which may be desirable if prediction is the objective.

Mallow's C_p Statistic

- A good model should predict well so average MSE of prediction might be a good criterion:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} E(\widehat{y}_i - E(y_i))^2$$

which can be estimated by the C_p statistic

$$C_p = \frac{SSE_p}{\widehat{\sigma}^2} + 2\,p - n$$

where $\widehat{\sigma}^2$ is from the model with all predictors and SSE_p indicates the SSE from a model with p parameters.

Mallow's C_p Statistic

- C_p has the following properties:
  (a) C_p is easy to compute
  (b) It is closely related to R_a^2 and the AIC.
  (c) For the full model $C_p = p$ exactly.
  (d) If a p predictor model fits then $E(SSE_p) = (n - p)\,\sigma^2$ and then $E(C_p) \approx p$. A model with a bad fit will have C_p much bigger than p.
- It is usual to plot C_p against p. We desire models with small p and C_p around or less than p.

Example: state data - C_p and R_a^2

- We default for the leaps() function is the Mallows Cp criterion:

  ```
  library(leaps)
  MLR.fit.state.Cp <- regsubsets(Life.Exp ~ .,
          nbest=3, data=statedata)
  par(mfrow=c(1,1))
  plot(MLR.fit.state.Cp)
  ```

- Now let's see which model the adjusted R^2 criterion selects. The R code:

```
x <- model.matrix(MLR.fit.state)[, -1]
y <- statedata$Life
leaps(x, y, wt=rep(1, NROW(x)), int=TRUE,
  method=c("adjr2"))
```

Example: state data - C_p and R_a^2

- Observe from the output of the above R code that the Population, Frost, HS graduation and Murder model has the largest R_a^2.

Effects of outliers and influential points

- Variable selection methods are sensitive to outliers and influential points.
- Let's check for high leverage points for state data:

```
h <- hat(x)
names(h) <- state.abb
Lev <- round(rev(sort(h)), 3)
Lev[1:8]
   AK    CA    HI    NV    NM    TX    NY    WA
0.810 0.409 0.379 0.365 0.325 0.284 0.257 0.223
```

- Which state sticks out?

Effects of outliers and influential points

- Let's try excluding it (Alaska is the second state in the data).

```
MLR.exc.Alaska <- leaps(x[-2,],y[-2],
        method="adjr2")
mod.sum <- cbind(MLR.exc.Alaska$which,
           MLR.exc.Alaska$adjr2)
mod.sum[order(MLR.exc.Alaska$adjr2),]
```

```
  1 2 3 4 5 6 7
6 1 0 1 1 1 1 1  0.70383148
5 0 1 0 1 1 1 1  0.70709124
6 1 1 0 1 1 1 1  0.70730267
4 1 0 0 1 1 1 0  0.70867029
5 1 0 0 1 1 1 1  0.71044047
```

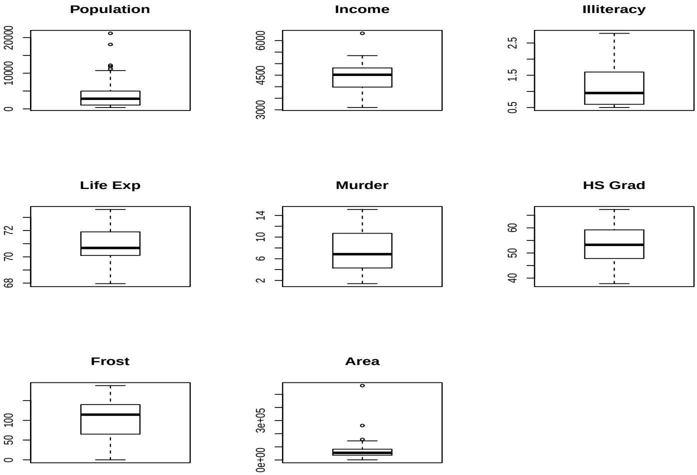
- We see that area now makes it into the model.

Effects of transformation

- Transforming the predictors can also have an effect.
- Take a look at the variables using boxplot. The R code

```
par(mfrow=c(3,3))
for(i in 1:8) boxplot(state.x77[,i],
      main=dimnames(state.x77)[[2]][i])
```

Boxplots of the State data



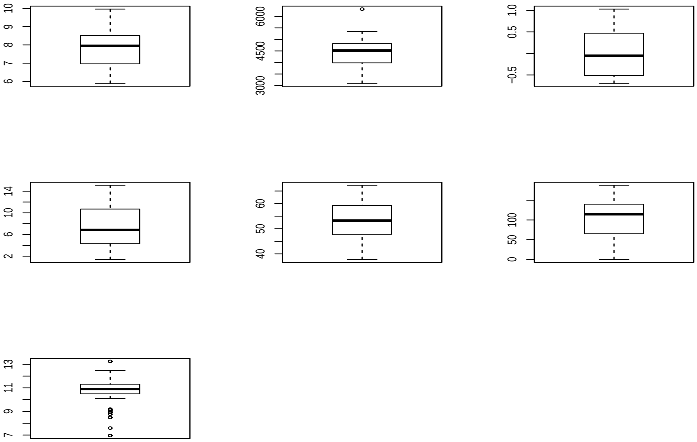
Boxplots of the State data

- In the above figure, we see that Population, Illiteracy and Area are skewed - we try transforming them:

```
x.tra <- cbind(log(x[,1]),x[,2],log(x[,3]),
               x[,4:6],log(x[,7]))

par(mfrow=c(3,3))
apply(x.tra,2, boxplot)
```

Boxplots of the State data



Boxplots of the State data

- which shows the appropriately transformed data.
- Now try the adjusted R^2 method again.

```
MLR.trans <- leaps(x.tra, y, method="adjr2")
mod.sum <- cbind(MLR.trans$which,MLR.trans$adjr2)
mod.sum[order(MLR.trans$adjr2, decreasing = TRUE),][1:5,]
```

```
  1 2 3 4 5 6 7
4 1 0 0 1 1 1 0 0.7173392
5 1 0 0 1 1 1 1 0.7136360
5 1 0 1 1 1 1 0 0.7117179
5 1 1 0 1 1 1 0 0.7110497
6 1 0 1 1 1 1 1 0.7083894
```

Boxplots of the State data

- This changes the "best" model again to log(Population), Frost, HS graduation and Murder.
- The adjusted R^2 is the highest models we have seen so far.

Summary

- Variable selection is a means to an end and not an end itself.
- The aim is to construct a model that predicts well or explains the relationships in the data.
- Automatic variable selections are not guaranteed to be consistent with these goals.
- Use these methods as a guide only.
- Stepwise methods use a restricted search through the space of potential models and use a dubious hypothesis testing based method for choosing between models.
- Criterion-based methods typically involve a wider search and compare models in a preferable manner.
- For this reason, the recommendation is use a criterion-based method.

Summary

- Accept the possibility that several models may be suggested which fit about as well as each other. If this happens, consider:
    1. Do the models have similar qualitative consequences?
    2. Do they make similar predictions?
    3. What is the cost of measuring the predictors?
    4. Which has the best diagnostics?
- If you find models that seem roughly equally as good but lead to quite different conclusions then it is clear that the data cannot answer the question of interest unambiguously.
- Be alert to the danger that a model contradictory to the tentative conclusions might be out there.