PSTAT 126

Lab 8

Roupen Khanjian

Spring 2021

```
library(faraway) # Functions and Datasets for Books by Julian Faraway
library(alr4) # Data to Accompany Applied Linear Regression 4th Edition
library(tidyverse) # Easily Install and Load the 'Tidyverse'
library(leaps) # Regression Subset Selection
library(patchwork) # The Composer of Plots
```

Model selection

Data from Faraway book (Chapter 10)

- Suppose the intercept is included in the model. For the remaining p - 1 covariates (predictors) , they could be in the model or out. Then in total we have 2^{p-1} choices. When p = 8, we have 128 potential models (not counting interaction or polynomial terms!).

```
data(state)
statedata <- data.frame(state.x77, row.names = state.abb)
head(statedata)
```

```
##    Population Income Illiteracy Life.Exp Murder HS.Grad Frost   Area
## AL       3615   3624        2.1    69.05   15.1    41.3    20  50708
## AK        365   6315        1.5    69.31   11.3    66.7   152 566432
## AZ       2212   4530        1.8    70.55    7.8    58.1    15 113417
## AR       2110   3378        1.9    70.66   10.1    39.9    65  51945
## CA      21198   5114        1.1    71.71   10.3    62.6    20 156361
## CO       2541   4884        0.7    72.06    6.8    63.9   166 103766
```

```
lmod <- lm(Life.Exp ~ ., statedata)
```

```r
summary(lmod)
```

```
## 
## Call:
## lm(formula = Life.Exp ~ ., data = statedata)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.48895 -0.51232 -0.02747  0.57002  1.49447 
## 
## 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 .  
## 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
```

```r
b <- regsubsets(formula(lmod),
                data=statedata)
rs <- summary(b)
rs$which # for each model of size p+1, chooses the model with the lowest RSS value.
```

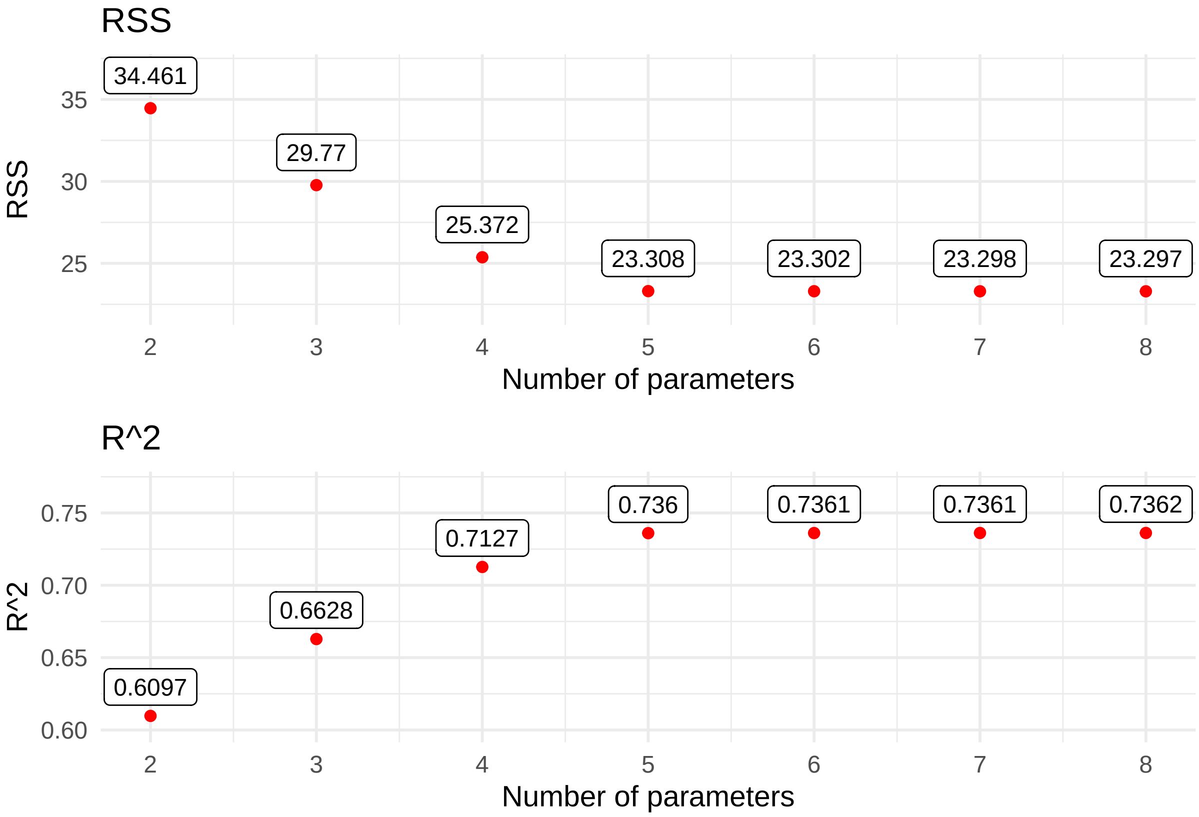
```
##   (Intercept) Population Income Illiteracy Murder HS.Grad Frost  Area
## 1        TRUE      FALSE  FALSE      FALSE   TRUE   FALSE FALSE FALSE
## 2        TRUE      FALSE  FALSE      FALSE   TRUE    TRUE FALSE FALSE
## 3        TRUE      FALSE  FALSE      FALSE   TRUE    TRUE  TRUE FALSE
## 4        TRUE       TRUE  FALSE      FALSE   TRUE    TRUE  TRUE FALSE
## 5        TRUE       TRUE   TRUE      FALSE   TRUE    TRUE  TRUE FALSE
## 6        TRUE       TRUE   TRUE       TRUE   TRUE    TRUE  TRUE FALSE
## 7        TRUE       TRUE   TRUE       TRUE   TRUE    TRUE  TRUE  TRUE
```

```r
# plot(rs$rss ~ I(1:7), ylab="RSS",
#      xlab="Number of Predictors", main = "RSS vs # of Predictors" )
```

```r
r1 <- ggplot(data = data.frame(rs$rss), aes(x = 2:8, y = rs$rss)) +
  geom_point(colour = "red", size = 1.5) +
  geom_label(aes(label= round(rs$rss, 3)), size = 3, nudge_y = 2 ) +
  scale_x_continuous(breaks = seq(2,8,1)) +
  ylim(22, 37) +
  labs(x = "Number of parameters", y = "RSS",
       title = "RSS") +
  theme_minimal()

r2 <- ggplot(data = data.frame(rs$rsq), aes(x = 2:8, y = rs$rsq)) +
  geom_point(colour = "red", size = 1.5) +
  geom_label(aes(label = round(rs$rsq, 4)), size = 3, nudge_y = 0.02) +
  scale_x_continuous(breaks = seq(2, 8, 1)) +
  ylim(0.6, 0.77) +
  labs(x = "Number of parameters", y = "R^2",
       title = "R^2") +
  theme_minimal()

r1 / r2
```



Now we introduce information criteria for model selection.

- **Akaike's Information Criterion (AIC)**

$$AIC = nlog(RSS) - nlog(n) + 2p = nlog(RSS/n) + 2p$$

- In AIC $k = 2$
- **Bayesian Information Criterion (BIC)**

$$BIC = nlog(RSS) - nlog(n) + p(log(n)) = nlog(RSS/n) + p(log(n))$$

- In BIC $k = log(n)$

Notes on AIC/BIC

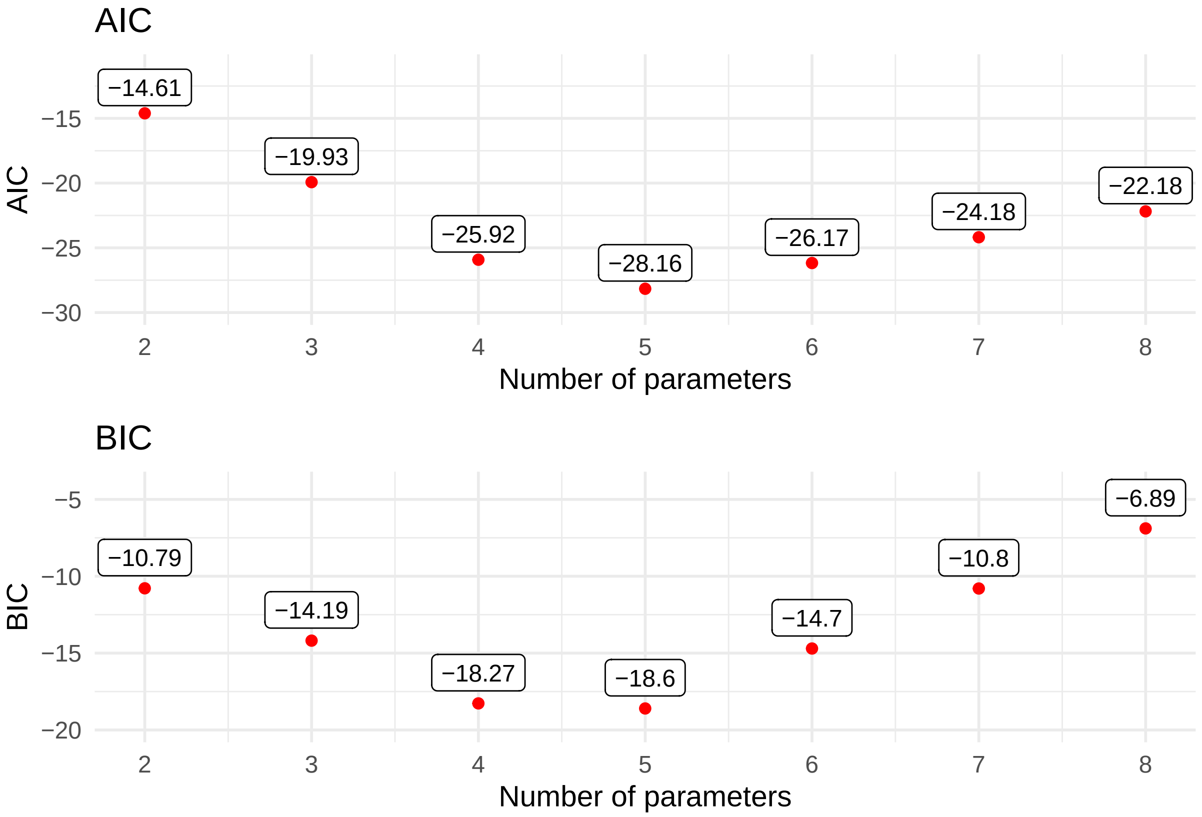
- BIC penalizes larger models more heavily and so will tend to prefer smaller models in comparison to AIC.
- The goal is to identify a subset of predictors such that AIC or BIC are minimized.

```
n <- nrow(statedata)
AIC <- n*log(rs$rss/n) + (2:8)*2
BIC <- n*log(rs$rss/n) + (2:8)*(log(n))
# plot(BIC~ I(1:7), ylab="BIC", xlab="Number of Predictors")
# plot(AIC ~ I(1:7), ylab="AIC", xlab="Number of Predictors")
```

```
a1 <- ggplot(data = data.frame(AIC), aes(x = 2:8, y = AIC)) +
  geom_point(colour = "red", size = 1.5) +
  geom_label(aes(label= round(AIC, 2)), size = 3, nudge_y = 2 ) +
  scale_x_continuous(breaks = seq(2,8,1)) +
  ylim(-30, -11) +
  labs(x = "Number of parameters", y = "AIC") +
  ggtitle("AIC") +
  theme_minimal()

b1 <- ggplot(data = data.frame(BIC), aes(x = 2:8, y = BIC)) +
  geom_point(colour = "red", size = 1.5) +
  geom_label(aes(label= round(BIC, 2)), size = 3, nudge_y = 2) +
  scale_x_continuous(breaks = seq(2,8,1)) +
  ylim(-20, -4) +
  labs(x = "Number of parameters", y = "BIC") +
  ggtitle("BIC") +
  theme_minimal()

a1 / b1
```



Model Selection

- Forward selection
  - Start with no variables (just intercept)
  - Add one variable at a time according to some criterion
  - Stop when no more variables should be added
- Backward selection
  - Start with a Full model with all possible predictors
  - Remove one variable at a time according to some criterion
  - Stop when no more variables should be dropped

Forward selection using p-values

- Let $\alpha = 0.10$ be our stopping criteria.

```r
mod0 <- lm(Life.Exp ~ 1, statedata)
add1(mod0, ~.+Population+Income+Illiteracy+Murder+HS.Grad+Frost+Area, 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  0.2233   0.63866
## Income      1    10.223 78.076  26.283  6.2847   0.01562 *
## Illiteracy  1    30.578 57.721  11.179 25.4289 6.969e-06 ***
## Murder      1    53.838 34.461 -14.609 74.9887 2.260e-11 ***
## HS.Grad     1    29.931 58.368  11.737 24.6146 9.196e-06 ***
## Frost       1     6.064 82.235  28.878  3.5397   0.06599 .
## Area        1     1.017 87.282  31.856  0.5594   0.45815
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
mod1 <- update(mod0, ~.+Murder)
add1(mod1, ~.+Population+Income+Illiteracy+HS.Grad+Frost+Area, 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
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
mod2 <- update(mod1, ~.+HS.Grad)
add1(mod2, ~.+Population+Income+Illiteracy+Frost+Area, 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   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
mod3 <- update(mod2, ~.+Frost)
add1(mod3, ~.+Population+Income+Illiteracy+Area, 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  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
mod4 <- update(mod3, ~.+Population)
add1(mod4, ~.+Income+Illiteracy+Area, test = "F")
```

```
## Single term additions
## 
## Model:
## Life.Exp ~ Murder + HS.Grad + Frost + Population
##            Df Sum of Sq    RSS     AIC F value Pr(>F)
## <none>                  23.308 -28.161               
## Income      1 0.0060582 23.302 -26.174  0.0114 0.9153
## Illiteracy  1 0.0039221 23.304 -26.170  0.0074 0.9318
## Area        1 0.0007900 23.307 -26.163  0.0015 0.9694
```

```r
summary(mod4)
```

```
## 
## Call:
## lm(formula = Life.Exp ~ Murder + HS.Grad + Frost + Population, 
##     data = statedata)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.47095 -0.53464 -0.03701  0.57621  1.50683 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  7.103e+01  9.529e-01  74.542  < 2e-16 ***
## 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 *
## Population   5.014e-05  2.512e-05   1.996  0.05201 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7197 on 45 degrees of freedom
## Multiple R-squared:  0.736,  Adjusted R-squared:  0.7126
## F-statistic: 31.37 on 4 and 45 DF,  p-value: 1.696e-12
```

`step` function allows you to choose a model using AIC as the information criteria. Can use forward or backward selection.

```
mod0 <- lm(Life.Exp ~ 1, statedata)
mod.upper <- lm(Life.Exp ~ ., statedata)
step(mod0,
     scope = list(lower = mod0, upper = mod.upper),
     direction = "forward") # Forward (start with intercept)
```

Forward selection using AIC values

```
## 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
##
## 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
##
## 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
##
## 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
##
## 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
```

```
##
## Call:
## lm(formula = Life.Exp ~ Murder + HS.Grad + Frost + Population,
##     data = statedata)
##
## Coefficients:
## (Intercept)       Murder      HS.Grad        Frost   Population
##   7.103e+01   -3.001e-01    4.658e-02   -5.943e-03    5.014e-05
```

```r
lmod <- lm(Life.Exp ~ ., statedata)
step(lmod, direction = "backward") # backward is the default direction in R
```

Backward selection using AIC values

```
## 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
##
## 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
##
```

```
## 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
##
## 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
```

```
##
## 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
```