Model/Variable Selection

STAT 757

Most common *model selection* methods are actually *variable selection* methods that assume you are comparing models that are special cases of a larger umbrella model (e.g., comparing MLR models with various predictors omitted from the regressions).

We will discuss four of them: R^2_{adj}, AIC, AIC_c, and BIC.

First, lets generate a toy dataset with a few weakly influential variables, and a few unrelated varibles:

```
set.seed(4916) # to get consistent random numbers

# Coefficients
B = c(200, -5, 0.2, -0.2, 0, 0)
sigma=5

# Design matrix
N = 40*length(B[-1]) # sample size
X = matrix(abs(rnorm(N,mean=20,sd=5)),byrow=TRUE,ncol=length(B[-1]))
X = cbind(1,X)

# Generate a "nice" data set:
Y = rnorm(N, mean=X%*%B, sd=sigma)
mydat = data.frame(Y,X[,-1])
names(mydat) <- c("y",gsub('^','x\\1',1:length(B[-1])))

# check regression:
fit0=lm(y ~ ., data=mydat) # same as fit0=lm(y~x1+x2+x3+x4+x5, data=mydat)
summary(fit0)
```

```
##
## Call:
## lm(formula = y ~ ., data = mydat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2674  -3.2978  -0.2445   2.7801  14.1082
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 192.37072    3.61638  53.194  < 2e-16 ***
## x1           -5.04250    0.07527 -66.990  < 2e-16 ***
## x2            0.25832    0.06457   4.001 8.97e-05 ***
## x3           -0.06304    0.07681  -0.821   0.4128
## x4            0.19086    0.08005   2.384   0.0181 *
## x5            0.05581    0.07908   0.706   0.4812
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.793 on 194 degrees of freedom
## Multiple R-squared:  0.9643, Adjusted R-squared:  0.9634
## F-statistic:  1049 on 5 and 194 DF,  p-value: < 2.2e-16
```

```r
# all submodels: 30+ possibilities!
fit1=lm(y~x2+x3+x4+x5, data=mydat)
fit2=lm(y~x1+x3+x4+x5, data=mydat)
fit3=lm(y~x1+x2+x4+x5, data=mydat)
fit4=lm(y~x1+x2+x3+x5, data=mydat)
fit5=lm(y~x1+x2+x3+x4, data=mydat)

# Or use MuMIn::dredge()! :-)
library(MuMIn)
options(na.action = "na.fail")
fits <- dredge(fit0)
```

```
## Fixed term is "(Intercept)"
```

Criteria

Adjusted R^2 First consider R^2_{adj} (closer to 1 is better):

```r
Rs = data.frame(Rsq.adj = c(Model0=summary(fit0)$adj.r.squared,
                            Model1=summary(fit1)$adj.r.squared,
                            Model2=summary(fit2)$adj.r.squared,
                            Model3=summary(fit3)$adj.r.squared,
                            Model4=summary(fit4)$adj.r.squared,
                            Model5=summary(fit5)$adj.r.squared))
Rs
```

```
##          Rsq.adj
## Model0 0.9633997
## Model1 0.1212890
## Model2 0.9605830
## Model3 0.9634610
## Model4 0.9625204
## Model5 0.9634939
```

AIC Repeat as above using AIC for fit0-fit5.

AICc Repeat as above using AICc for fit0-fit5, or use dredge()

```r
fits ## see above
```

```
## Global model call: lm(formula = y ~ ., data = mydat)
## ---
## Model selection table
##    (Intrc)     x1      x2       x3       x4      x5 df   logLik   AICc
## 12  192.50 -5.067  0.2749           0.18900          5 -594.753 1199.8
## 16  193.80 -5.060  0.2695 -0.06246  0.18280          6 -594.414 1201.3
## 28  191.00 -5.049  0.2639           0.19700 0.05512  6 -594.504 1201.4
## 32  192.40 -5.042  0.2583 -0.06304  0.19090 0.05581  7 -594.157 1202.9
## 4   196.30 -5.059  0.2606                           4 -597.655 1203.5
## 8   197.90 -5.051  0.2543 -0.07960                   5 -597.114 1204.5
```

```
## 20  195.70 -5.050  0.2549                  0.02692  5 -597.596 1205.5
## 24  197.20 -5.041  0.2482 -0.08029         0.02893  6 -597.046 1206.5
## 26  193.50 -4.976                  0.17860 0.13480  5 -602.823 1216.0
## 30  195.50 -4.968         -0.09521 0.16980 0.13340  6 -602.086 1216.6
## 10  197.60 -5.014                  0.15570          4 -604.275 1216.8
## 14  199.60 -5.005         -0.09728 0.14700          5 -603.516 1217.3
## 2   200.60 -5.009                                   3 -606.091 1218.3
## 6   202.60 -5.000         -0.10960                  4 -605.137 1218.5
## 18  197.70 -4.979                           0.10670  4 -605.176 1218.6
## 22  199.70 -4.970         -0.10950          0.10660  5 -604.215 1218.7
## 23   84.79        -0.8208 -0.64150          1.81900  5 -912.594 1835.5
## 19   70.86        -0.7814                   1.82800  4 -914.103 1836.4
## 31   80.75        -0.8126 -0.62740 0.15770 1.84200  6 -912.512 1837.5
## 27   65.66        -0.7713           0.21960 1.85900  5 -913.944 1838.2
## 21   70.34                -0.56680          1.63500  4 -916.218 1840.6
## 17   58.58                                  1.65100  3 -917.363 1840.8
## 29   64.74                -0.54750 0.22660 1.67000  5 -916.053 1842.4
## 25   52.20                         0.27780 1.69300  4 -917.114 1842.4
## 7   117.40        -0.5275 -0.66770                  4 -924.859 1857.9
## 5   105.60                -0.61610                  3 -926.248 1858.6
## 3   103.10        -0.4850                           3 -926.306 1858.7
## 1    93.18                                          2 -927.472 1859.0
## 15  120.80        -0.5384 -0.68050 -0.14660         5 -924.795 1859.9
## 13  107.30                -0.62250 -0.08010         4 -926.229 1860.7
## 11  104.90        -0.4907          -0.08248         4 -926.286 1860.8
## 9    93.71                         -0.02644         3 -927.470 1861.1
##     delta weight
## 12   0.00  0.402
## 16   1.45  0.195
## 28   1.63  0.178
## 32   3.08  0.086
## 4    3.70  0.063
## 8    4.72  0.038
## 20   5.69  0.023
## 24   6.71  0.014
## 26  16.14  0.000
## 30  16.79  0.000
## 10  16.94  0.000
## 14  17.53  0.000
## 2   18.49  0.000
## 6   18.66  0.000
## 18  18.74  0.000
## 22  18.92  0.000
## 23 635.68  0.000
## 19 636.60  0.000
## 31 637.64  0.000
## 27 638.38  0.000
## 21 640.83  0.000
## 17 641.03  0.000
## 29 642.60  0.000
## 25 642.62  0.000
## 7  658.11  0.000
## 5  658.80  0.000
## 3  658.92  0.000
```

```
## 1  659.19  0.000
## 15 660.08  0.000
## 13 660.85  0.000
## 11 660.96  0.000
## 9  661.25  0.000
## Models ranked by AICc(x)
```

BIC Repeat as above using BIC for fit0-fit5.

Summary Which models would you exclude from your final analysis, and why?

Other options? Explore forward and backward stepwise selection (somewhat controversial) and LASSO (least absolute shrinkage and selection operator) and LARS (least angle regression).