Chapter 10: Model Selection

Stats 500, Fall 2015
Brian Thelen, University of Michigan
443 West Hall, bjthelen@umich.edu

Variable Selection

- Testing-based approaches
- Criterion-based approaches

Testing-based Model Selection

- Backward elimination
- Forward selection
- Stepwise regression

Backward Elimination

1. Start with all the predictors in the model
2. **Remove** the predictor with the **highest** p**-value** greater than α
3. Refit the model and go to step 2
4. Stop when all p-values are less than α

$\alpha > 0.05$ may be better if **prediction is the goal** .

Forward Selection

1. Start with no predictor variables
2. For all predictors not in the model, check the p-value **if** they are added to the model
3. **Add** the one with the **smallest p-value** less than α
4. Refit the model and go to step 2
5. Stop when no new predictors can be added

Stepwise regression is a combination of backward elimination and forward selection (allows to add variables back after they have been removed).

Life Expectancy Example

- Census data from 50 states
- Response: life expectancy in years (1969-71)
- Predictors:

```
'Population': population estimate as of July 1, 1975
'Income': per capita income (1974)
'Illiteracy': illiteracy (1970, percent of population)
'Murder': murder and non-negligent manslaughter rate
   per 100,000 population (1976)
'HS Grad': percent high-school graduates (1970)
'Frost': mean number of days with minimum temperature
   below freezing (1931-1960) in capital or large city
'Area': land area in square miles
```

Life Expectancy Example Continued

```
> data(state)
# reassemble the data (add row names)
> statedata = data.frame(state.x77, row.names=state.abb)
> g = lm(Life.Exp ~ ., data=statedata)
> summary(g)
```

```
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

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
```

```
## Backward elimination - drop largest p-value
> g = update(g, . ~ . - Area)
> summary(g)
             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
Residual standard error:   0.7361 on 43 degrees of freedom
Multiple R-Squared: 0.7361      Adjusted R-squared: 0.6993
F-statistic: 19.99 on 6 and 43 DF      p-value: 5.362e-11
```

```
## Continue dropping
> g = update(g, . ~ . - Illiteracy)
> summary(g)
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
Residual standard error:   0.7277 on 44 degrees of freedom
Multiple R-Squared: 0.7361      Adjusted R-squared: 0.7061
F-statistic: 24.55 on 5 and 44 DF      p-value: 1.019e-11
```

```
## Continue dropping
> g = update(g, . ~ . - Income)
> summary(g)
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
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
```

```
## Borderline case... would keep for prediction,
## but try dropping
> g = update(g, . ~ . - Population)
> summary(g)
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
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
```

```
## Cannot conclude other predictors have no effect
## on response: e.g., Illiteracy
> summary(lm(Life.Exp ~ Illiteracy + Murder
     + Frost, statedata))
Coefficients:
           Estimate Std.Error t value Pr(>|t|)
Intercept 74.556717  0.584251 127.611  < 2e-16
Illiteracy-0.601761  0.298927  -2.013  0.04998
Murder    -0.280047  0.043394  -6.454 6.03e-08
Frost     -0.008691  0.002959  -2.937  0.00517
Residual standard error:   0.7911 on 46 degrees of freedom
Multiple R-Squared: 0.6739      Adjusted R-squared: 0.6527
F-statistic: 31.69 on 3 and 46 DF       p-value: 2.915e-11
```

Remarks on Testing-based approaches

- **Greedy** . May miss the optimal model.
- Do not take p-values at face value (multiple testing).
- Variables not selected can still be correlated with the response, but they do not improve the fit enough to be included.
- Tend to pick **smaller models** than desirable for prediction purposes.

Criterion-based Model Selection

General idea: choose the model that optimizes a criterion which **balances goodness-of-fit and model size** .

- **AIC** and **BIC**
- Adjusted R^2
- Mallows' $\mathbf{C_p}$

AIC and BIC

- Akaike information criterion (**AIC**)

$$\text{AIC} = n \ln(\text{RSS}/n) + 2(p+1)$$

  R function: `step(...,k=2)` (default)

- Bayes information criterion (**BIC**)

$$\text{BIC} = n \ln(\text{RSS}/n) + (p+1) \ln n$$

  R function: `step(..., k=log(n))`

Pick a model that **minimizes AIC or BIC**

Life Expectancy Example

```
> ## AIC
> g = lm(Life.Exp ~ ., data=statedata)
> step(g)
Start:  AIC= -22.18
 Life.Exp ~ Population + Income + Illiteracy +
   Murder + HS.Grad + Frost + Area
             Df Sum of Sq     RSS     AIC
- Area        1     0.001  23.298 -24.182
- Income      1     0.004  23.302 -24.175
- Illiteracy  1     0.005  23.302 -24.174
<none>                     23.297 -22.185
- Population  1     1.747  25.044 -20.569
- Frost       1     1.847  25.144 -20.371
- HS.Grad     1     2.441  25.738 -19.202
- Murder      1    23.141  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     0.004 23.302 -26.174
- Income      1     0.006 23.304 -26.170
<none>                    23.298 -24.182
- Population  1     1.760 25.058 -22.541
- Frost       1     2.049 25.347 -21.968
- HS.Grad     1     2.980 26.279 -20.163
- Murder      1    26.272 49.570  11.568
```

```
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  17.457
```

```
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 -23.876
- HS.Grad     1     5.112 28.420 -20.246
- Murder      1    34.816 58.124  15.528

Coefficients:
(Intercept   Population   Murder    HS.Grad      Frost
  71.03      5.014e-05   -0.3001   4.658e-02  -5.943e-03
```

- BIC picked the same model.

Adjusted R^2

Recall

$$R^2 = 1 - \frac{RSS}{TSS}$$

Definition of adjusted R^2:

$$\begin{aligned} R_a^2 &= 1 - \frac{RSS/(n-(p+1))}{TSS/(n-1)} \\ &= 1 - \left(\frac{n-1}{n-(p+1)}\right)(1-R^2) \end{aligned}$$

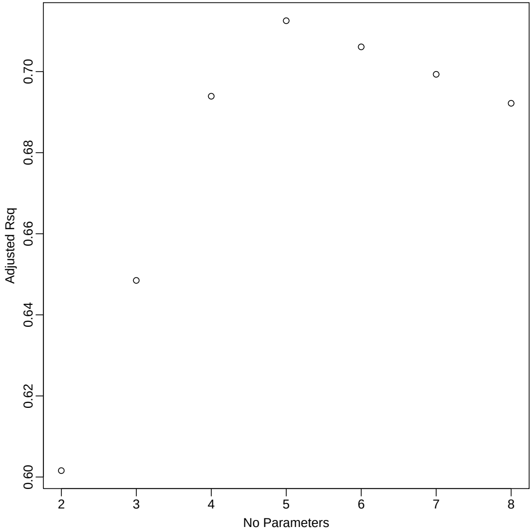
- Adding a predictor will not necessarily increase R_a^2
- Maximizing R_a^2 is equivalent to minimizing RSE $\hat{\sigma}$.

Life Expectancy Example

```
> ## Adjusted R^2
> library(leaps)
> b = regsubsets(Life.Exp ~ ., data=statedata)
> summary(b)
Selection Algorithm: exhaustive
         Population Income Illiteracy Murder HS.Grad Frost Area
1  ( 1 ) " "        " "    " "        "*"    " "     " "   " "
2  ( 1 ) " "        " "    " "        "*"    "*"     " "   " "
3  ( 1 ) " "        " "    " "        "*"    "*"     "*"   " "
4  ( 1 ) "*"        " "    " "        "*"    "*"     "*"   " "
5  ( 1 ) "*"        "*"    " "        "*"    "*"     "*"   " "
6  ( 1 ) "*"        "*"    "*"        "*"    "*"     "*"   " "
7  ( 1 ) "*"        "*"    "*"        "*"    "*"     "*"   "*"

# plot adjusted R2 against p+1
> rs = summary(b)
> plot(2:8, rs$adjr2, xlab="No. of Parameters",
  ylab="Adjusted Rsq")
# select model with largest adjusted R2
> which.max(rs$adjr2)
[1] 4
```

Adjusted R^2 for the Life Expectancy Data



Mallows' C_p

Definition:

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} + 2(p+1) - n$$

- $\hat{\sigma}^2$ is estimated from the model with all predictors
- RSS_p is from the model with p predictors
- Goal: minimize C_p.
- C_p around or less than $p+1$ indicates good fit.
- C_p estimates the mean squared error (**MSE**)

$$\frac{1}{\sigma^2}\sum_i E(\hat{y}_i - Ey_i)^2$$

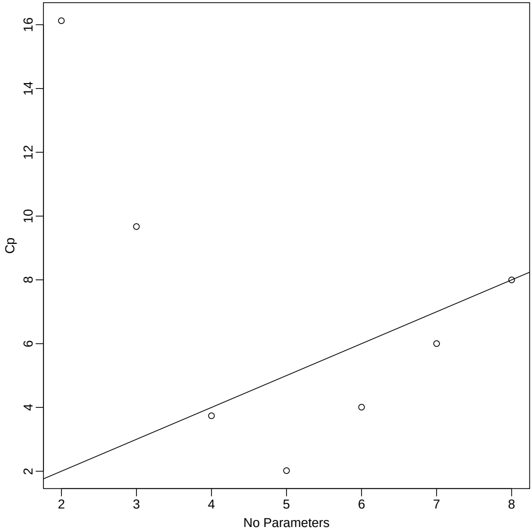
Life Expectancy Example

```
> ## Mallows Cp
> library(leaps)
> b = regsubsets(Life.Exp ~ ., data=statedata)
> rs = summary(b)

> which.min(rs$cp)
[1] 4

> plot(2:8, rs$cp, xlab="No. Parameters",
       ylab="Cp")
> abline(0, 1)
```

C_p Plot for the Life Expectancy Data



Variable Selection Summary

- Variable selection methods are sensitive to outliers
- Generally, criterion-based methods are preferred
- It may happen that several models provide very similar fit
- If models with similar fit lead to very different conclusions, the data are ambiguous
- If conclusions are similar, choose a simpler model and/or predictors that are easier to measure