Chapter 9: Variable Selection

Gunwoong Park

Lecture Note

University of Seoul

Motivations

- We want to explain the data in the simplest way. The simplest is best.
- Unnecessary predictors will add noise.
- Collinearity is caused by having too many variables trying to do the same job.
- Save time and money by not measuring redundant predictors.

Variable Selection

1. Testing-based approaches
   - ▷ Backward elimination
   - ▷ Forward selection
   - ▷ Stepwise regression
2. Criterion-based approaches
   - ▷ AIC and BIC
   - ▷ Adjusted R^2
   - ▷ Mallows' C_p

Testing-based approaches

- General idea: **test significance** of predictors and eliminate in some principled fashion
- Based on individual p-values (t-test)
- Multiple testing is not accounted for, but ranking is more important than the absolute size of p-values
- Different methods use different **rules to add/delete predictors**

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 (e.g., 0.20)

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; Backward Ellimination

```
> 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)
            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
```

Life Expectancy Example Continued; Forward Selection

```
> data(state)
# reassemble the data (add row names)
> 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     69.0   15.1    41.3    20  50708
AK        365   6315        1.5     69.3   11.3    66.7   152 566432
AZ       2212   4530        1.8     70.5    7.8    58.1    15 113417
AR       2110   3378        1.9     70.7   10.1    39.9    65  51945
CA      21198   5114        1.1     71.7   10.3    62.6    20 156361
CO       2541   4884        0.7     72.1    6.8    63.9   166 103766
```

Forward Selection: Step 1

```
> summary(lm(Life.Exp ~ Population, data=statedata))$coef
Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.10e+01   2.65e-01 267.409 7.90e-78
Population  -2.05e-05   4.33e-05  -0.473 6.39e-01

> summary(lm(Life.Exp ~ Income, data=statedata))$coef
Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.76e+01   1.327571   50.91 1.98e-43
Income      7.43e-04   0.000297    2.51 1.56e-02

> summary(lm(Life.Exp ~ Illiteracy, data=statedata))$coef
Estimate Std. Error t value Pr(>|t|)
(Intercept)     72.4      0.338  213.97 3.47e-73
Illiteracy      -1.3      0.257   -5.04 6.97e-06
```

Forward Selection: Step 1

```
> summary(lm(Life.Exp ~ Murder, data=statedata))$coef
Estimate Std. Error t value Pr(>|t|)
(Intercept)   72.974     0.2700  270.30 4.72e-78
Murder        -0.284     0.0328   -8.66 2.26e-11

> summary(lm(Life.Exp ~ HS.Grad, data=statedata))$coef
Estimate Std. Error t value Pr(>|t|)
(Intercept)  65.7397     1.0475   62.76 9.92e-48
HS.Grad       0.0968     0.0195    4.96 9.20e-06

> summary(lm(Life.Exp ~ Frost, data=statedata))$coef
Estimate Std. Error t value Pr(>|t|)
(Intercept) 70.17163     0.4189  167.52 4.33e-68
Frost        0.00677     0.0036    1.88 6.60e-02

> summary(lm(Life.Exp ~ Area, data=statedata))$coef
Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.10e+01   2.49e-01 285.434 3.46e-79
Area        -1.69e-06   2.26e-06  -0.748 4.58e-01
```

Forward Selection: Step 2

```
> summary(lm(Life.Exp ~ Murder + Population, data=statedata))$coef
Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.29e+01   2.58e-01  282.01 1.55e-77
Murder      -3.12e-01   3.32e-02   -9.42 2.15e-12
Population   6.83e-05   2.74e-05    2.49 1.64e-02
> summary(lm(Life.Exp ~ Murder + Income, data=statedata))$coef
Estimate Std. Error t value Pr(>|t|)
(Intercept) 71.22558   0.967395   73.63 3.32e-50
Murder      -0.26976   0.032841   -8.21 1.22e-10
Income       0.00037   0.000197    1.88 6.66e-02
> summary(lm(Life.Exp ~ Murder + Illiteracy, data=statedata))$coef
Estimate Std. Error t value Pr(>|t|)
(Intercept)   73.028     0.2857 255.623 1.56e-75
Murder        -0.264     0.0464  -5.688 7.96e-07
Illiteracy    -0.172     0.2811  -0.613 5.43e-01
```

Forward Selection: Step 2

```
> summary(lm(Life.Exp ~ Murder + HS.Grad, data=statedata))$coef
Estimate Std. Error t value Pr(>|t|)
(Intercept)  70.2971     1.0157   69.21 5.91e-49
Murder       -0.2371     0.0353   -6.72 2.18e-08
HS.Grad       0.0439     0.0161    2.72 9.09e-03
> summary(lm(Life.Exp ~ Murder + Frost, data=statedata))$coef
Estimate Std. Error t value Pr(>|t|)
(Intercept) 73.90032    0.50029  147.71 2.36e-64
Murder      -0.32778    0.03751   -8.74 2.05e-11
Frost       -0.00578    0.00266   -2.17 3.52e-02
> summary(lm(Life.Exp ~ Murder + Area, data=statedata))$coef
Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.29e+01   2.75e-01 265.295 2.73e-76
Murder      -2.90e-01   3.38e-02  -8.584 3.47e-11
Area         1.18e-06   1.46e-06   0.806 4.24e-01
```

Forward Selection: Step 3

```
> summary(lm(Life.Exp ~ Murder + HS.Grad + Population, data=statedata))$coef
Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.04e+01   9.69e-01   72.70 3.95e-49
Murder      -2.66e-01   3.57e-02   -7.45 1.91e-09
HS.Grad      4.07e-02   1.54e-02    2.64 1.12e-02
Population   6.25e-05   2.59e-05    2.41 1.99e-02

> summary(lm(Life.Exp ~ Murder + HS.Grad + Income, data=statedata))$coef
Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.01e+01   1.096336  63.979 1.33e-46
Murder      -2.39e-01   0.035806  -6.664 2.92e-08
HS.Grad      3.91e-02   0.020297   1.924 6.05e-02
Income       9.53e-05   0.000239   0.398 6.92e-01
```

Forward Selection: Step 3

```
> summary(lm(Life.Exp ~ Murder + HS.Grad + Illiteracy, data=statedata))$coef
Estimate Std. Error t value Pr(>|t|)
(Intercept)  69.7354     1.2221  57.063 2.41e-44
Murder       -0.2581     0.0435  -5.934 3.63e-07
HS.Grad       0.0518     0.0188   2.761 8.25e-03
Illiteracy    0.2540     0.3051   0.833 4.09e-01
> summary(lm(Life.Exp ~ Murder + HS.Grad + Frost, data=statedata))$coef
Estimate Std. Error t value Pr(>|t|)
(Intercept) 71.03638    0.98326   72.25 5.25e-49
Murder      -0.28307    0.03673   -7.71 8.04e-10
HS.Grad      0.04995    0.01520    3.29 1.95e-03
Frost       -0.00691    0.00245   -2.82 6.99e-03
> summary(lm(Life.Exp ~ Murder + HS.Grad + Area, data=statedata))$coef
Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.99e+01   1.16e+00  60.093 2.30e-45
Murder      -2.24e-01   4.04e-02  -5.563 1.30e-06
HS.Grad      5.04e-02   1.90e-02   2.649 1.10e-02
Area        -1.06e-06   1.62e-06  -0.658 5.14e-01
```

Forward Selection: Step 4

```
> summary(lm(Life.Exp ~ Murder + HS.Grad + Frost + Population, data=statedata))$coef
Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.10e+01   9.53e-01   74.54 8.61e-49
Murder      -3.00e-01   3.66e-02   -8.20 1.77e-10
HS.Grad      4.66e-02   1.48e-02    3.14 2.97e-03
Frost       -5.94e-03   2.42e-03   -2.46 1.80e-02
Population   5.01e-05   2.51e-05    2.00 5.20e-02

> summary(lm(Life.Exp ~ Murder + HS.Grad + Frost + Income, data=statedata))$coef
Estimate Std. Error t value Pr(>|t|)
(Intercept) 70.836789   1.050471  67.433 7.53e-47
Murder      -0.285558   0.037260  -7.664 1.07e-09
HS.Grad      0.043554   0.018975   2.295 2.64e-02
Frost       -0.006983   0.002469  -2.829 6.96e-03
Income       0.000127   0.000223   0.571 5.71e-01
```

Forward Selection: Step 4

```
> summary(lm(Life.Exp ~ Murder + HS.Grad + Frost + Illiteracy, data=statedata))$coef
Estimate Std. Error t value Pr(>|t|)
(Intercept) 71.51996    1.32049  54.162 1.28e-42
Murder      -0.27312    0.04114  -6.639 3.50e-08
HS.Grad      0.04497    0.01776   2.532 1.49e-02
Frost       -0.00768    0.00283  -2.715 9.36e-03
Illiteracy  -0.18161    0.32785  -0.554 5.82e-01

> summary(lm(Life.Exp ~ Murder + HS.Grad + Frost + Area, data=statedata))$coef
Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.09e+01   1.15e+00  61.705 3.92e-45
Murder      -2.79e-01   4.27e-02  -6.516 5.34e-08
HS.Grad      5.19e-02   1.79e-02   2.906 5.66e-03
Frost       -6.82e-03   2.51e-03  -2.713 9.40e-03
Area        -3.29e-07   1.54e-06  -0.214 8.32e-01
```

- Forward Selection

$$\text{Life Exp} = 71.036 - 0.283\text{Murder} + 0.049\text{HS.Grad} - 0.006\text{Frost}$$

- Backward Elimination

$$\text{Life Exp} = 71.036 - 0.283\text{Murder} + 0.049\text{HS.Grad} - 0.006\text{Frost}$$

In general, the selected models are different.

Remarks on Testing-based approaches

- Greedy. May miss the optimal (true) model.
- Should not take p-values literally.
- Variables not selected can still be correlated with the response, but they do not improve the fit enough to be included.

Criterion-based Model Selection

- General idea: choose the model that optimizes a criterion which balances goodness-of-fit and model size.
- No p-values involved
- Some theoretical guarantees
- Different methods use different goodness-of-fit measures and different penalties for model size

Criterion-based Model Selection

- AIC
- BIC
- adjusted R^2
- Mallow's C_p

- A good model should achieve maximum likelihood with small number of predictors.
    - A likelihood is a function of the parameters (β) of a statistical model given data.
    - A likelihood is often the same as probability of the parameters given data.
    - e.g., $\widehat{\beta} \sim N(\beta, \sigma^2(X^T X)^{-1})$.

- A good model should predict well with small number of predictors.
    - minimize $\sum (Y_i - \widehat{Y}_i)^2$
    - minimize $\sum |Y_i - \widehat{Y}_i|$
    - minimize Residual Sum of Square

AIC

- Akaike information criterion (AIC)

$$\begin{aligned}\text{AIC} &= n\ln(\text{RSS}/n) + 2(p+1) \\ &= -2\text{Likelihood}(\hat{\beta}) + 2(p+1)\end{aligned}$$

- $\hat{\beta}$ can be estimated by LSE
- R function: `step(...,k=2)` (default)
- Pick a model that minimizes AIC

BIC

- Bayes information criterion (BIC)

$$\begin{aligned} \text{BIC} &= n \ln(\text{RSS}/n) + \ln n \times (p+1) \\ &= -2\text{Likelihood}(\hat{\beta}) + \ln n \times (p+1) \end{aligned}$$

- R function: `step(..., k=log(n))`
- Pick a model that minimizes BIC

AIC Backward: Life Expectancy Example

```
> ## AIC Backward
> g = lm(Life.Exp ~ ., data=statedata)
> step(g, direction="backward", k = 2)
Start:  AIC=-22.2
Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad +
Frost + Area

Df Sum of Sq  RSS   AIC
- Area        1      0.00 23.3 -24.2
- Income      1      0.00 23.3 -24.2
- Illiteracy  1      0.00 23.3 -24.2
<none>                    23.3 -22.2
- Population  1      1.75 25.0 -20.6
- Frost       1      1.85 25.1 -20.4
- HS.Grad     1      2.44 25.7 -19.2
- Murder      1     23.14 46.4  10.3
```

```
Step:  AIC=-24.2
Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad +
Frost

Df Sum of Sq  RSS   AIC
- Illiteracy  1      0.00 23.3 -26.2
- Income      1      0.01 23.3 -26.2
<none>                    23.3 -24.2
- Population  1      1.76 25.1 -22.5
- Frost       1      2.05 25.3 -22.0
- HS.Grad     1      2.98 26.3 -20.2
- Murder      1     26.27 49.6  11.6
```

```
Step:  AIC=-26.2
Life.Exp ~ Population + Income + Murder + HS.Grad + Frost

Df Sum of Sq  RSS   AIC
- Income      1       0.0 23.3 -28.2
<none>                    23.3 -26.2
- Population  1       1.9 25.2 -24.3
- Frost       1       3.0 26.3 -22.1
- HS.Grad     1       3.5 26.8 -21.2
- Murder      1      34.7 58.0  17.5
```

```
Step:  AIC=-28.2
Life.Exp ~ Population + Murder + HS.Grad + Frost

Df Sum of Sq  RSS   AIC
<none>                   23.3 -28.2
- Population  1       2.1 25.4 -25.9
- Frost       1       3.1 26.4 -23.9
- HS.Grad     1       5.1 28.4 -20.2
- Murder      1      34.8 58.1  15.5

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

Coefficients:
(Intercept)   Population       Murder      HS.Grad        Frost
7.10e+01     5.01e-05    -3.00e-01     4.66e-02    -5.94e-03
```

BIC Backward: Life Expectancy Example

```
> ## BIC Backward
> g = lm(Life.Exp ~ ., data=statedata)
> step(g, direction="backward", k = log(nrow(statedata)))
Step:  AIC=-18.6
Life.Exp ~ Population + Murder + HS.Grad + Frost

Df Sum of Sq  RSS   AIC
<none>                    23.3 -18.6
- Population  1       2.1 25.4 -18.3
- Frost       1       3.1 26.4 -16.2
- HS.Grad     1       5.1 28.4 -12.6
- Murder      1      34.8 58.1  23.2

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

Coefficients:
(Intercept)   Population       Murder      HS.Grad        Frost
7.10e+01     5.01e-05    -3.00e-01     4.66e-02    -5.94e-03
```

AIC Forward: Life Expectancy Example

```
> g = lm(Life.Exp ~ 1 ,  data=statedata)
> step(g, direction="forward", k= 2,
scope = ( ~ Population + Income + Murder + Illiteracy + HS.Grad + Frost + Area) )

Start:  AIC=30.4
Life.Exp ~ 1

Df Sum of Sq  RSS   AIC
+ Murder      1      53.8 34.5 -14.6
+ Illiteracy  1      30.6 57.7  11.2
+ HS.Grad     1      29.9 58.4  11.7
+ Income      1      10.2 78.1  26.3
+ Frost       1       6.1 82.2  28.9
<none>                    88.3  30.4
+ Area        1       1.0 87.3  31.9
+ Population  1       0.4 87.9  32.2
```

```
Step:  AIC=-14.6
Life.Exp ~ Murder

Df Sum of Sq  RSS   AIC
+ HS.Grad     1      4.69 29.8 -19.9
+ Population  1      4.02 30.4 -18.8
+ Frost       1      3.13 31.3 -17.4
+ Income      1      2.40 32.1 -16.2
<none>                    34.5 -14.6
+ Area        1      0.47 34.0 -13.3
+ Illiteracy  1      0.27 34.2 -13.0
```

```
Step:  AIC=-19.9
Life.Exp ~ Murder + HS.Grad

Df Sum of Sq  RSS   AIC
+ Frost       1      4.40 25.4 -25.9
+ Population  1      3.34 26.4 -23.9
<none>                    29.8 -19.9
+ Illiteracy  1      0.44 29.3 -18.7
+ Area        1      0.28 29.5 -18.4
+ Income      1      0.10 29.7 -18.1
```

```
Step:  AIC=-25.9
Life.Exp ~ Murder + HS.Grad + Frost

Df Sum of Sq  RSS   AIC
+ Population  1     2.064 23.3 -28.2
<none>                    25.4 -25.9
+ Income      1     0.182 25.2 -24.3
+ Illiteracy  1     0.172 25.2 -24.3
+ Area        1     0.026 25.4 -24.0
```

```
Step:  AIC=-28.2
Life.Exp ~ Murder + HS.Grad + Frost + Population

Df Sum of Sq  RSS   AIC
<none>                  23.3 -28.2
+ Income      1   0.00606 23.3 -26.2
+ Illiteracy  1   0.00392 23.3 -26.2
+ Area        1   0.00079 23.3 -26.2

Coefficients:
(Intercept)       Murder      HS.Grad        Frost   Population
7.10e+01    -3.00e-01     4.66e-02    -5.94e-03     5.01e-05
```

BIC Forward: Life Expectancy Example

```
> g = lm(Life.Exp ~ 1 ,  data=statedata)
> step(g, direction="forward", k= log(nrow(statedata)),
scope = ( ~ Population + Income + Murder + Illiteracy + HS.Grad + Frost + Area) ) Step:  AI
Life.Exp ~ Murder + HS.Grad + Frost + Population

Df Sum of Sq  RSS   AIC
<none>                 23.3 -18.6
+ Income      1   0.00606 23.3 -14.7
+ Illiteracy  1   0.00392 23.3 -14.7
+ Area        1   0.00079 23.3 -14.7

Coefficients:
(Intercept)       Murder      HS.Grad        Frost   Population
7.10e+01    -3.00e-01     4.66e-02    -5.94e-03     5.01e-05
```

- Test-based method Forward Selection & Backward Elimination

$$\text{Life Exp} = 71.036 - 0.283\text{Murder} + 0.049\text{HS.Grad} - 0.006\text{Frost}$$

- AIC & BIC: Forward Selection & Backward Elimination

$$\begin{aligned}\text{Life Exp} = 71.036 - 0.300\ \text{Murder} + 0.046\ \text{HS.Grad} \\ - 0.006\ \text{Frost} + 0.0005\ \text{Population}\end{aligned}$$

In general, the selected models are different.

AIC Both Directions: Life Expectancy Example

```
> g = lm(Life.Exp ~ . ,  data=statedata)
> step(g, direction="both", k = 2)
Start:  AIC=-22.2
Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad +
Frost + Area

Df Sum of Sq  RSS   AIC
- Area        1      0.00 23.3 -24.2
- Income      1      0.00 23.3 -24.2
- Illiteracy  1      0.00 23.3 -24.2
<none>                    23.3 -22.2
- Population  1      1.75 25.0 -20.6
- Frost       1      1.85 25.1 -20.4
- HS.Grad     1      2.44 25.7 -19.2
- Murder      1     23.14 46.4  10.3
```

```
Step:  AIC=-24.2
Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad +
Frost

Df Sum of Sq  RSS   AIC
- Illiteracy  1      0.00 23.3 -26.2
- Income      1      0.01 23.3 -26.2
<none>                    23.3 -24.2
- Population  1      1.76 25.1 -22.5
+ Area        1      0.00 23.3 -22.2
- Frost       1      2.05 25.3 -22.0
- HS.Grad     1      2.98 26.3 -20.2
- Murder      1     26.27 49.6  11.6
```

```
Step:  AIC=-26.2
Life.Exp ~ Population + Income + Murder + HS.Grad + Frost

Df Sum of Sq  RSS   AIC
- Income      1       0.0 23.3 -28.2
<none>                    23.3 -26.2
- Population  1       1.9 25.2 -24.3
+ Illiteracy  1       0.0 23.3 -24.2
+ Area        1       0.0 23.3 -24.2
- Frost       1       3.0 26.3 -22.1
- HS.Grad     1       3.5 26.8 -21.2
- Murder      1      34.7 58.0  17.5
```

```
Step:  AIC=-28.2
Life.Exp ~ Population + Murder + HS.Grad + Frost

Df Sum of Sq  RSS   AIC
<none>                    23.3 -28.2
+ Income      1       0.0 23.3 -26.2
+ Illiteracy  1       0.0 23.3 -26.2
+ Area        1       0.0 23.3 -26.2
- Population  1       2.1 25.4 -25.9
- Frost       1       3.1 26.4 -23.9
- HS.Grad     1       5.1 28.4 -20.2
- Murder      1      34.8 58.1  15.5

Coefficients:
(Intercept)   Population       Murder      HS.Grad        Frost
7.10e+01     5.01e-05    -3.00e-01     4.66e-02    -5.94e-03
```

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

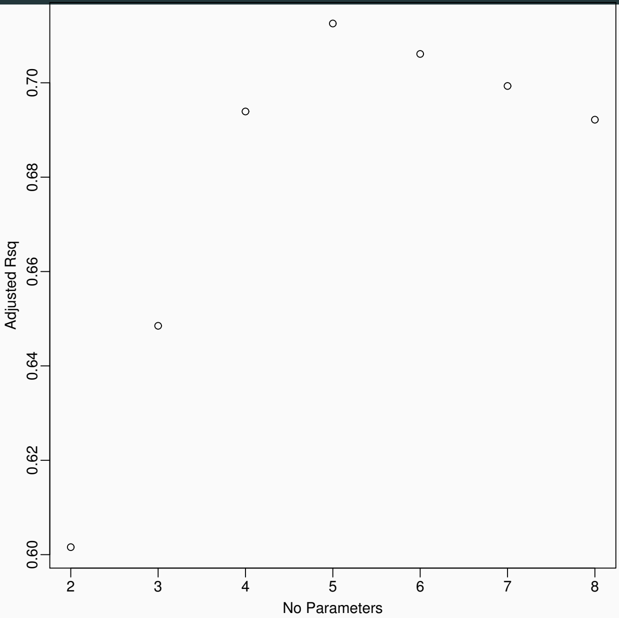
Adjusted R^2 Exhaustive Search: Life Expectancy Example

```
> 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



Adjusted R^2 Backward: Life Expectancy Example

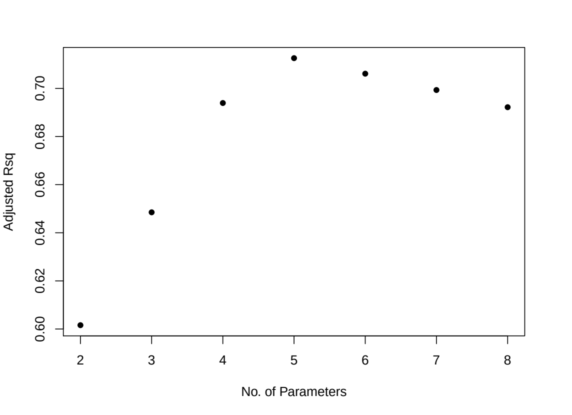
```
> b = regsubsets(Life.Exp ~ ., data=statedata, method="backward")
> summary(b)

Selection Algorithm: backward
         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



Adjusted R^2 Both Direction: Life Expectancy Example

```
> b = regsubsets(Life.Exp ~ ., data=statedata, method="seqrep")
> summary(b)

Selection Algorithm: backward
         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
```

Remarks on Adjusted R^2

- In general, exhaustive search and greedy search have the same result.
- If the number of predicts is large, their results are different. However greedy searches are usually applied because of computational complexity.

Mallows' C_p

Definition:

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

where p is the number of predictors model used.

- $\hat{\sigma}^2$ 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)

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

Mallows' C_p

Definition:

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

where p is the number of predictors model used.

With appropriate predictors,

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

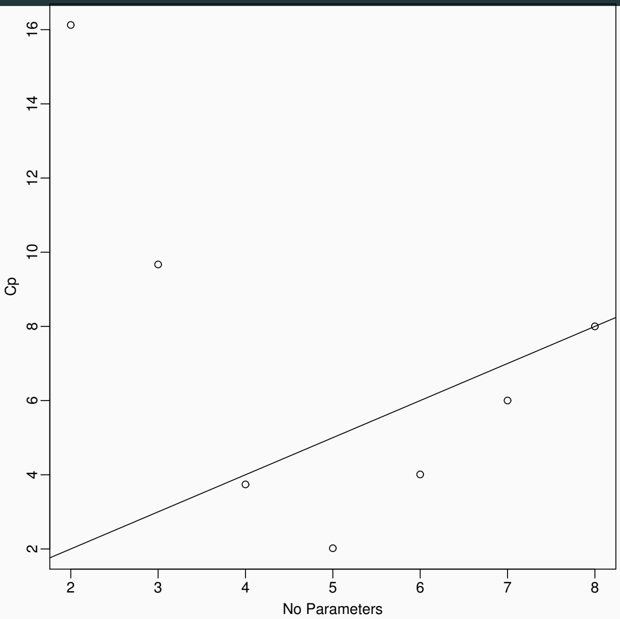
Life Expectancy Example

```
> ## 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

Consistency of Variable Selection

Later...