9.4. Variable Selection

- Goal is to develop a model with the best set of independent variables
  - Easier to interpret if unimportant variables are removed
  - Lower probability of collinearity

- Stepwise regression procedure
  - Provide evaluation of alternative models as variables are added (or withdrawn)

- Best-subset approach
  - Try all combinations and select the best using various criteria, such as the highest adjusted R^2

$x_1, \cdots, x_k$. $2^k - 1$

```
> install.packages("olsrr")
> library(olsrr)
```

Data: Statedata

Dependent Variable: Life.Exp - the life expectancy in years of residents of the state in 1970

Independent Variables

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

加油可以跑多少公里

$AIC = 2K - 2\log(\hat{L})$

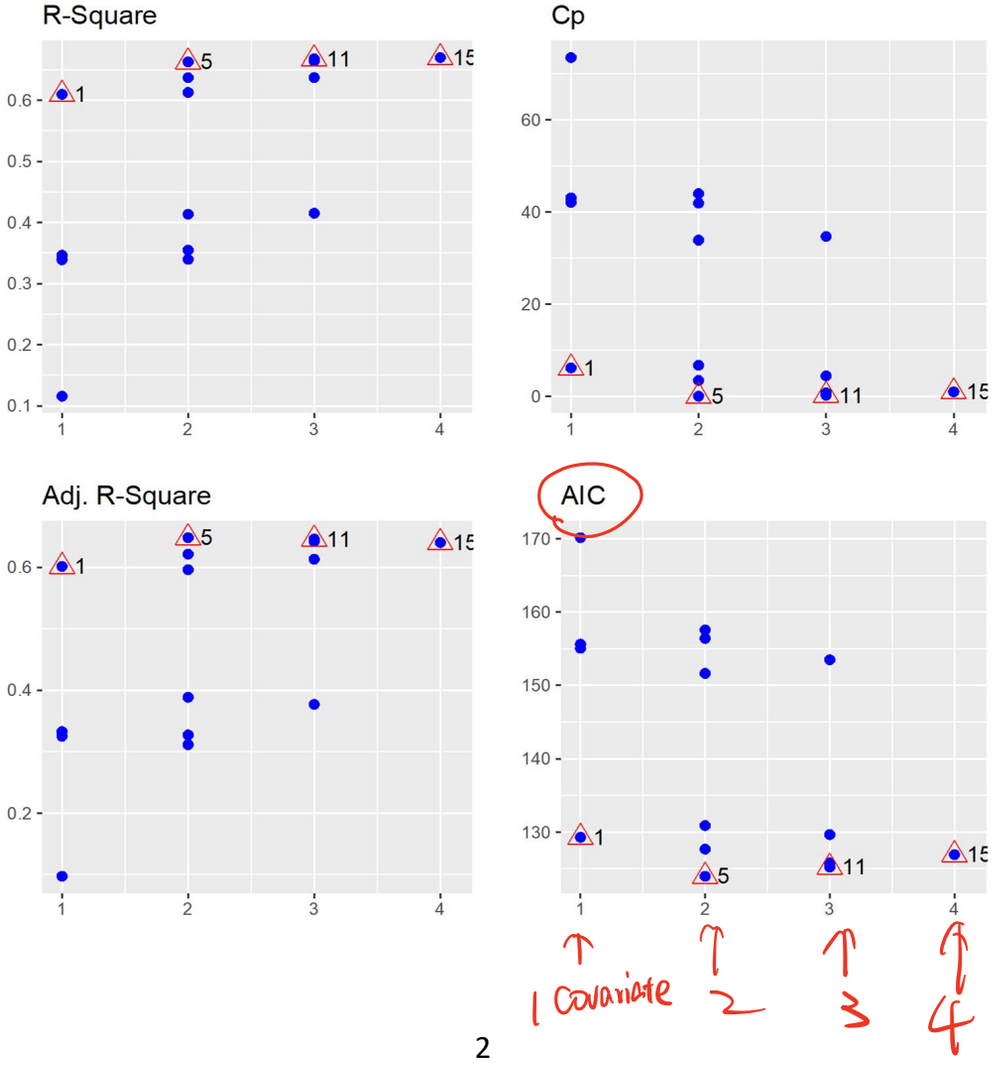
↑ # of unknown parameter

All-possible subset selection

Assume we have independent variables: Murder, Income, Illiteracy, and HS.Grad

```
> reg <- lm(Life.Exp ~ Murder+Income+Illiteracy+HS.Grad, data=statedata)
> ols_step_all_possible(reg)
# A tibble: 15 x 6
   Index     N Predictors   `R-Square` `Adj. R-Square` `Mallow's Cp`
   <int> <int> <chr>             <dbl>           <dbl>         <dbl>
 1     1     1 Murder            0.610           0.602          7.19
 2     2     1 Illiteracy        0.346           0.333         43.1
 3     3     1 HS.Grad           0.339           0.325         44.1
 4     4     1 Income            0.116           0.0974        74.5
 5     5     2 Murder HS.G…      0.663           0.648          1.95
 6     6     2 Murder Inco…      0.637           0.622          5.48
 7     7     2 Murder Illi…      0.613           0.596          8.77
 8     8     2 Illiteracy …      0.414           0.389         35.9
 9     9     2 Income Illi…      0.355           0.327         43.9
10    10     2 Income HS.G…      0.340           0.312         46.0
11    11     3 Murder Illi…      0.668           0.646          3.27
12    12     3 Murder Inco…      0.664           0.642          3.79
13    13     3 Murder Inco…      0.637           0.613          7.46
14    14     3 Income Illi…      0.415           0.377         37.7
15    15     4 Murder Inco…      0.670           0.640          5

> p <- ols_step_all_possible(reg)
> plot(p)
```



Stepwise regression (Use all of the available independent variables)

```
> reg <- lm(Life.Exp ~ ., data=statedata)
> summary(reg)

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
```

```
> start = lm(Life.Exp~1, data=statedata)
> fitALL = lm(Life.Exp~., data=statedata)

> stepwise <- step(start, direction="both",scope=formula(fitALL))
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     4.691 29.770 -19.925
+ Population  1     4.016 30.445 -18.805
+ Frost       1     3.135 31.327 -17.378
+ Income      1     2.405 32.057 -16.226
<none>                   34.461 -14.609
+ Area        1     0.470 33.992 -13.295
+ Illiteracy  1     0.273 34.188 -13.007
- Murder      1    53.838 88.299  30.435

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
- HS.Grad     1    4.6910 34.461 -14.609
- Murder      1   28.5974 58.368  11.737

Step:  AIC=-25.92
Life.Exp ~ Murder + HS.Grad + Frost

             Df Sum of Sq    RSS     AIC
+ Population  1     2.064 23.308 -28.161
<none>                   25.372 -25.920
+ Income      1     0.182 25.189 -24.280
+ Illiteracy  1     0.172 25.200 -24.259
+ Area        1     0.026 25.346 -23.970
- Frost       1     4.399 29.770 -19.925
- HS.Grad     1     5.955 31.327 -17.378
- Murder      1    32.756 58.128  13.531
```

$y_i = \beta_0 + \varepsilon_i$

$y_i = \beta_0 + \beta_1 \text{Murder}_i + \varepsilon_i$

$y_i = \beta_0 + \beta_1 \text{Murder}_i + \beta_2 \text{HS. Grad}_i + \beta_3 \text{Forest}_i + \varepsilon_i$

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

             Df Sum of Sq    RSS     AIC
<none>                     23.308 -28.161
+ Income      1     0.006 23.302 -26.174
+ Illiteracy  1     0.004 23.304 -26.170
+ Area        1     0.001 23.307 -26.163
- 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
```

final model.

n: # sample size.
k: # variable

Remark 9.4: Modern techniques for variable selection.

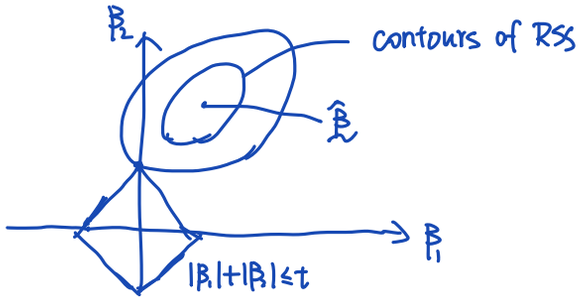
* LASSO:

for model: $\underset{\sim}{y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$ $\quad n \ll k$.

LSE: $\min (\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\beta})'(\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\beta}) \Rightarrow \hat{\underset{\sim}{\beta}} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{y}$

LASSO: $\min_{\underset{\sim}{\beta}} L_{LASSO} = (\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\beta})'(\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\beta}) + \lambda \sum_{j=1}^{k} |\beta_j|$

$\Leftrightarrow \min_{\underset{\sim}{\beta}} (\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\beta})'(\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\beta})$ s.t. $\sum_j |\beta_j| \le t$

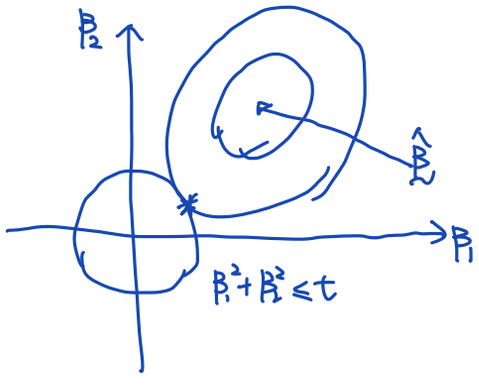


* Ridge regression.

$\min_{\underset{\sim}{\beta}} L_{ridge} = (\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\beta})'(\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\beta}) + \lambda \sum_{j=1}^{k} \beta_j^2$

$\Leftrightarrow \min_{\underset{\sim}{\beta}} (\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\beta})'(\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\beta})$ s.t. $\sum_{j=1}^{k} \beta_j^2 \le t$

$\Rightarrow \hat{\underset{\sim}{\beta}} = (\underset{\sim}{X}^T\underset{\sim}{X} + \lambda \underset{\sim}{I})^{-1}\underset{\sim}{X}^T\underset{\sim}{y}$



* Elastic Net.

$\min L_{Elr} = (\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\beta})^T(\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\beta}) + \lambda_2 \sum_j \beta_j^2 + \lambda_1 \sum_j |\beta_j|$

$\Rightarrow$ Variables with high correlation

R package: glmnet. LASSO, Ridge, Elastic

Final model

```
> finalmodel <- lm(Life.Exp ~ Murder + HS.Grad + Frost + Population,data=statedata)
> summary(finalmodel)

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

> library(faraway)
> x <- model.matrix(finalmodel) [,-1]

> vif(x)
    Murder    HS.Grad      Frost Population
  1.727844   1.356791   1.498077   1.189835
```