Variable Selection

PYJ

Variable Selection

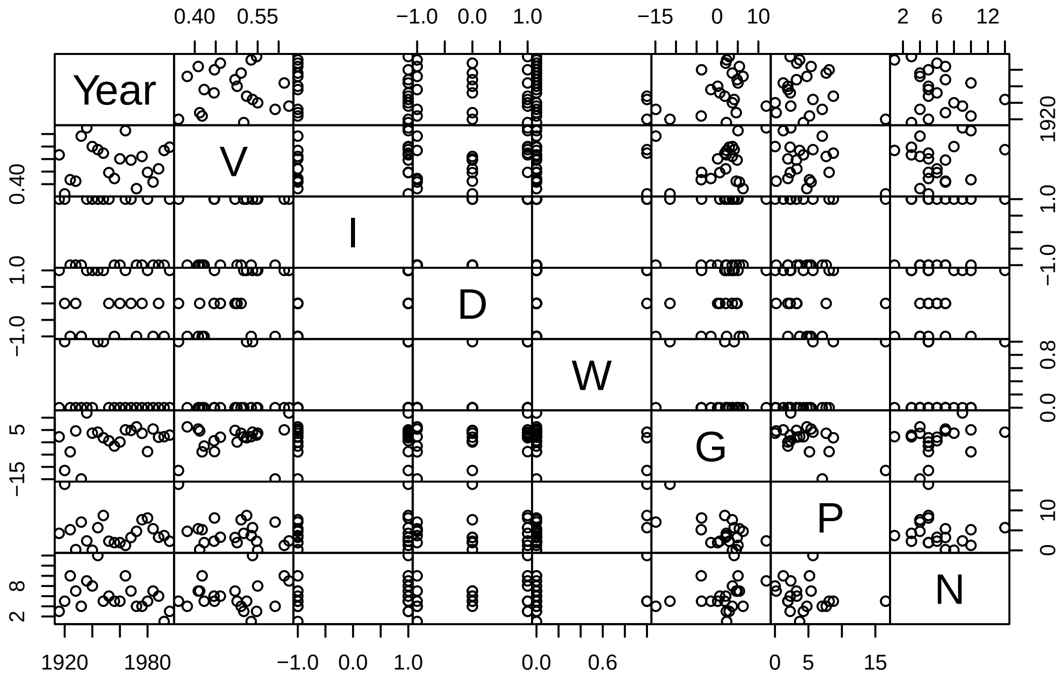
- Core: Mean Squared Error: $\text{MSE}(\hat{\beta}) = E[(\hat{\beta} - \beta)^2]$
  - MSE = Variance + $(Bias)^2$
  - When we miss necessary predictors: gain a smaller variance (intuitively, deleting variables cannot increase the variance); also get a biased estimates (how large is the bias depends on what's the X_j
  - Variance-Bias Trade-off: compare the increment in $(Bias)^2$ and the reduction in variance, vice versa.
- Model Comparison Methods
  - Nested model: F test -> they follow F distribution
    * Example:
    $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$
  - Any two models w/ the same response
    * Similar to nested, but not nested
      · MSE (SSE/(n-p-1))
      · AIC (Akaike) $= nlog_e(SSE_p/n) + 2p$
      · BIC (Bayesian) $= n\log_e(SSE_p/n) + p\log_e(n)$
  - Any two models w/ the same response but possibly differently transformed
    * Example: $log(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, $\sqrt{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
- Forward Selection (FS): 2^q subsets
- Backward Elimination (BE): 2^q subsets

- Stepwise Selection (SW): each iteration choose the lowest AIC or BIC; stop when no candidate can lower the score

Coding Example

```r
p160 = read.table("P160.txt", h = T)
```

```r
pairs(p160, gap = 0, oma = c(2, 2, 2, 2))
```




```r
# quick visualization of how variables correlated
```

```r
# BE, include everything at the beginning
step(lm(V ~ I + D + W + G + P + N, data = p160), test = "F")
```

```
Start:  AIC=-104.98
V ~ I + D + W + G + P + N

       Df Sum of Sq      RSS     AIC F value Pr(>F)
- N     1 0.0000079 0.072712 -106.98  0.0015 0.9694
- I     1 0.0000400 0.072744 -106.97  0.0077 0.9313
- W     1 0.0000894 0.072793 -106.96  0.0172 0.8975
- G     1 0.0016214 0.074325 -106.52  0.3122 0.5851
```

```
- P     1 0.0044157 0.077119 -105.75  0.8503 0.3721
<none>              0.072704 -104.98
- D     1 0.0101039 0.082808 -104.25  1.9456 0.1848

Step:  AIC=-106.98
V ~ I + D + W + G + P

       Df Sum of Sq      RSS     AIC F value Pr(>F)
- I     1 0.0000436 0.072755 -108.97  0.0090 0.9257
- W     1 0.0001396 0.072851 -108.94  0.0288 0.8675
- G     1 0.0016497 0.074361 -108.51  0.3403 0.5683
- P     1 0.0048827 0.077594 -107.62  1.0073 0.3315
<none>              0.072712 -106.98
- D     1 0.0101469 0.082859 -106.24  2.0933 0.1685

Step:  AIC=-108.97
V ~ D + W + G + P

       Df Sum of Sq      RSS     AIC F value  Pr(>F)
- W     1 0.0001571 0.072912 -110.92  0.0346 0.85488
- G     1 0.0016185 0.074374 -110.51  0.3559 0.55912
- P     1 0.0050355 0.077791 -109.56  1.1074 0.30829
<none>              0.072755 -108.97
- D     1 0.0245242 0.097280 -104.87  5.3932 0.03373 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step:  AIC=-110.92
V ~ D + G + P

       Df Sum of Sq      RSS     AIC F value  Pr(>F)
- G     1 0.0017808 0.074693 -112.42  0.4152 0.52794
<none>              0.072912 -110.92
- P     1 0.0110706 0.083983 -109.95  2.5812 0.12655
- D     1 0.0270882 0.100001 -106.29  6.3158 0.02234 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step:  AIC=-112.42
V ~ D + P

       Df Sum of Sq      RSS     AIC F value  Pr(>F)
<none>              0.074693 -112.42
```

```
- P     1 0.0099223 0.084616 -111.80  2.3911 0.13943
- D     1 0.0255565 0.100250 -108.24  6.1588 0.02317 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Call:
lm(formula = V ~ D + P, data = p160)

Coefficients:
(Intercept)            D            P
   0.514022     0.043134    -0.006017
```

```
# we are using F test here to compare current model with the potential model, the score is s
```

BE stops when there's only D and P included in this model.

```
summary(lm(V ~ D + P, data = p160))
```

```

Call:
lm(formula = V ~ D + P, data = p160)

Residuals:
      Min        1Q    Median        3Q       Max
-0.101121 -0.036838 -0.006987  0.019029  0.163250

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.514022   0.022793  22.552  1.2e-14 ***
D            0.043134   0.017381   2.482   0.0232 *
P           -0.006017   0.003891  -1.546   0.1394
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06442 on 18 degrees of freedom
Multiple R-squared:  0.3372,    Adjusted R-squared:  0.2636
F-statistic: 4.579 on 2 and 18 DF,  p-value: 0.02468
```

```r
# Forward selection
step(lm(V ~ 1, data = p160), scope = V ~ I + D + W + G + P + N, direction = "forward", test =
```

```
Start:  AIC=-107.78
V ~ 1

       Df Sum of Sq      RSS     AIC F value  Pr(>F)
+ D     1 0.0280805 0.084616 -111.80  6.3054 0.02124 *
+ I     1 0.0135288 0.099167 -108.47  2.5921 0.12389
+ P     1 0.0124463 0.100250 -108.24  2.3589 0.14106
<none>              0.112696 -107.78
+ G     1 0.0060738 0.106622 -106.94  1.0824 0.31123
+ N     1 0.0024246 0.110271 -106.24  0.4178 0.52579
+ W     1 0.0009518 0.111744 -105.96  0.1618 0.69197
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step:  AIC=-111.8
V ~ D

       Df Sum of Sq      RSS     AIC F value Pr(>F)
+ P     1 0.0099223 0.074693 -112.42  2.3911 0.1394
<none>              0.084616 -111.80
+ W     1 0.0068141 0.077801 -111.56  1.5765 0.2253
+ I     1 0.0012874 0.083328 -110.12  0.2781 0.6044
+ G     1 0.0006325 0.083983 -109.95  0.1356 0.7170
+ N     1 0.0000033 0.084612 -109.80  0.0007 0.9793

Step:  AIC=-112.42
V ~ D + P

       Df  Sum of Sq      RSS     AIC F value Pr(>F)
<none>               0.074693 -112.42
+ G     1 0.00178078 0.072912 -110.92  0.4152 0.5279
+ W     1 0.00031940 0.074374 -110.51  0.0730 0.7903
+ N     1 0.00018496 0.074508 -110.47  0.0422 0.8397
+ I     1 0.00002633 0.074667 -110.42  0.0060 0.9392


Call:
lm(formula = V ~ D + P, data = p160)
```

```
Coefficients:
(Intercept)            D            P
   0.514022     0.043134    -0.006017
```

Only D and P included in the model.

```
# stepwise
step(lm(V ~ D + W, data = p160), scope = V~ I + D + W + G + P + N, direction = "both", test =
```

```
Start:  AIC=-111.56
V ~ D + W

       Df Sum of Sq      RSS     AIC F value  Pr(>F)
- W     1  0.006814 0.084616 -111.80  1.5765 0.22532
<none>              0.077801 -111.56
+ P     1  0.003428 0.074374 -110.51  0.7835 0.38843
+ N     1  0.000374 0.077428 -109.66  0.0820 0.77802
+ I     1  0.000178 0.077623 -109.61  0.0391 0.84567
+ G     1  0.000011 0.077791 -109.56  0.0023 0.96213
- D     1  0.033943 0.111744 -105.96  7.8529 0.01178 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step:  AIC=-111.8
V ~ D

       Df Sum of Sq      RSS     AIC F value  Pr(>F)
+ P     1 0.0099223 0.074693 -112.42  2.3911 0.13943
<none>              0.084616 -111.80
+ W     1 0.0068141 0.077801 -111.56  1.5765 0.22532
+ I     1 0.0012874 0.083328 -110.12  0.2781 0.60439
+ G     1 0.0006325 0.083983 -109.95  0.1356 0.71703
+ N     1 0.0000033 0.084612 -109.80  0.0007 0.97928
- D     1 0.0280805 0.112696 -107.78  6.3054 0.02124 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step:  AIC=-112.42
V ~ D + P

       Df Sum of Sq      RSS     AIC F value  Pr(>F)
```

```
<none>               0.074693 -112.42
- P     1 0.0099223 0.084616 -111.80  2.3911 0.13943
+ G     1 0.0017808 0.072912 -110.92  0.4152 0.52794
+ W     1 0.0003194 0.074374 -110.51  0.0730 0.79026
+ N     1 0.0001850 0.074508 -110.47  0.0422 0.83968
+ I     1 0.0000263 0.074667 -110.42  0.0060 0.93919
- D     1 0.0255565 0.100250 -108.24  6.1588 0.02317 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Call:
lm(formula = V ~ D + P, data = p160)

Coefficients:
(Intercept)            D            P
   0.514022     0.043134    -0.006017
```

Difference between using AIC and BIC in programming: substitute 2 to $\log_e(n)$, therefore, need to specify k = log(n) as last command.

FS, BE, SW may not choose the same model: Highly correlated predictors can cause variables to appear significant in one method but not in another; order dependency exists.