제11장 Variable selection procedures

◆ Formulation of the problem

$Y$ : reponse

$X_1$, ..., $X_q$ : full set of predictors

• moodel(11.1) : $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_q x_{iq} + \epsilon_i$ $\Rightarrow$ $\hat{\beta}_j{*}$, $\hat{y}{*}$

• moodel(11.2) : $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$ $p < q$ $\Rightarrow$ $\hat{\beta}_j$, $\hat{y}$

▷ Condition

1. $\beta_0$, $\beta_1$, ..., $\beta_q$ : non-zero

model(11.1) : true model

model(11.2) : underspecified model

$\Rightarrow$ Under fitting problem : bias $E(\hat{\beta}_j) \neq \beta_j$, $E(\hat{y}|\boldsymbol{x}_0) \neq \boldsymbol{x}_0{}'\boldsymbol{\beta}$

2. $\beta_0$, $\beta_1$, ..., $\beta_p$ : non-zero, $\beta_{p+1}$, ..., $\beta_q$ : zero

model(11.2) : true model

model(11.1) : overspecified model

$\Rightarrow$ over fitting problem : large variance $Var(\hat{\beta}_j{*}) \geq Var(\hat{\beta}_j)$, $Var(\hat{y}{*}) \geq Var(\hat{y})$

$\Rightarrow$ Need to compare $MSE = Var + bias^2$

▷ Use of regression equations

- Description and model building : two conflict requirement

  (1) to account for as much of the variation as possible

  $\Rightarrow$ tend to include more variables

  (2) to adhere to the principle of parsimony

  $\Rightarrow$ for easy of understanding .. with as few variables as possible

- Estimation and prediction

  $\Rightarrow$ minimizing the $MSE$ of prediction

- Control :

  to determine the magnitude by which the value of a predictor variable must be altered to obtain a specified value of response.

  $\Rightarrow$ need $s.e.(\hat{\beta}_j)$ : small

◆ Criteria for evaluation equations

▷ Residual Mean Squares

$$RMS_p = \frac{SSE_p}{(n-p)} = MSE \quad : (p-1) \text{ variables, } p \text{ parameters}$$

cf. $R_p^2 = 1 - \frac{SSE}{SST} = 1 - \frac{RMS_p}{SST}(n-p) \quad : R^2$

$$R_{ap}^2 = 1 - \frac{RMS_p}{SST}(n-1) \quad : \quad adj - R^2$$

▷ Mallow's $C_p$ : $(p-1)$ variables, $p$ parameters

$$Variance + bias^2 \quad : \quad J_p = \frac{1}{\sigma^2}\sum MSE(\hat{y}_i)$$

⇒ To estimate $J_p$, Mallow(1973) uses the following statistic :

$$C_p = \frac{SSE_p}{\hat{\sigma}^2} + (2p-n) = p + \frac{(s_p^2 - \hat{\sigma}^2)(n-p)}{\hat{\sigma}^2} \quad ; \quad \hat{\sigma}^2 : \text{estimate of } \sigma^2 \text{ of full model}$$

⇒ the small, the better & $C_p \approx p$

eg. $C_1 = 1.9$, $C_2 = 2.1$, $C_3 = 2.6$, $C_4 = 3.9$, $C_5 = 5$

⇒ model with $p = 2$ is the best!

[reg213f] 4.다음 질문에 답하라.

(2) 가능한 모든 독립변수가 5개일 때 $C_p$를 기준으로 가장 좋은 모델을 결정하고자 한다.
독립변수의 개수가 $p-1$인 회귀모델 중 가장 작은 $C_p$값을 갖는 경우만 기록하니 다음과 같았다.
즉, 독립변수가 2인 회귀모델 중 가장 작은 $C_p$를 갖는 모델의 $C_p$는 5.7이다.
$C_p$를 기준으로 가장 좋은 모델은 독립변수가 몇 개인 경우인가?

| 변수의 수 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $C_p$ | 8.7 | 5.7 | 5.1 | 5.2 | 6 |

▷ Information criteria

- Akaike IC : $AIC_p = n\log\left(\frac{SSE_p}{n}\right) + 2p$
- Bayesian IC :$BIC_p = n\log\left(\frac{SSE_p}{n}\right) + p\log(n)$ : to avoid over fitting

⋮

◆ Evaluating all possible equations

moderate size of $p$ : $2^p$ possible equations

$R^2$, $R_a^2$, $C_p$, $IC$, $PRESS$

eg. Supervisor performance data(Table3.3)

result : Table 11.4, Fig 11.1

◆ Variable Selection procedures (regressors are not collinear )

6 predictors : $x_1$, $x_2$, $x_3$, $x_4$, $x_5$, $x_6$

▷ Forward Selection

- fit 6 equations :

$$y = \beta_0 + \beta_j x_j + \epsilon, \quad j = 1, \ldots, 6$$

- choose $x_j$ with biggest $|t_j|$
- let $x_1$ be the first in : fit 5 equations :

$$y = \beta_0 + \beta_1 x_1 + \beta_j x_j + \epsilon, \quad j = 2, \ldots, 6$$

- choose $x_j$ with biggest $|t_j|$ : OK if significant
- let $x_2$ be the second in : fit 4 equations :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_j x_j + \epsilon, \quad j = 3, \ldots, 6$$

- choose $x_j$ with biggest $|t_j|$ : OK if significant

⋮

until no significant variable left

eg Supervisor performance data : see Table 11.2

▷ Backward elimination

- fit : $y = \beta_0 + \beta_1 x_1 + \cdots \beta_6 x_6 + \epsilon$
- delete $x_j$ with smallest $|t_j|$ & not significant
- let $x_6$ be the first out : fit : $y = \beta_0 + \beta_1 x_1 + \cdots \beta_5 x_5 + \epsilon$
- delete $x_j$ with smallest $|t_j|$ & not significant

$\vdots$

until all variables in the model are significant..

eg Supervisor performance data : see Table 11.3

▷ Stepwise procedure

modified version of Forward selection

In each step, after adding one variable, perform backward elimination...

eg.
- let $x_1$ be the first in OK
- let $x_2$ be the second in : $x_1, x_2$

  delete if any of $x_1$, $x_2$ is insignificant.. both are significant..
- let $x_3$ be the third in : $x_1, x_2, x_3$

  delete if any of $x_1$, $x_2$, $x_3$ is insignificant... delete $x_1$ : $x_2, x_3$

$\vdots$

until no significant variable left, all variables in the model are significant..

[reg211f] 2. 종속변수 $Y$에 대하여 4개의 독립변수 $X_1$, $X_2$, $X_3$, $X_4$로 가능한 회귀모형을 적합시키고 다음과 같은 결과를 얻었다. 이 결과를 이용하여 다음 물음에 답하라. (2번째~5번째 칸의 값은 회귀모수 추정치의 $p$-값이다.)

| Model | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $R^2_{adj}$ | $C_p$ |
|---|---|---|---|---|---|---|
| 1 | 3.12e-07 | - | - | - | 0.9075 | 29.3524 |
| 2 | - | 0.0533 | - | - | 0.2348 | 308.1845 |
| 3 | - | - | 2.61e-07 | - | 0.9104 | 28.1400 |
| 4 | - | - | - | 9.33e-07 | 0.8873 | 37.7306 |
| 5 | 7.56e-09 | 0.000326 | - | - | 0.9736 | 2.9470 |
| 6 | 0.374 | - | 0.295 | - | 0.9093 | 27.1788 |
| 7 | 0.0386 | - | - | 0.1220 | 0.9208 | 22.8362 |
| 8 | - | 0.0191 | 3.13e-07 | - | 0.9446 | 13.8801 |
| 9 | - | 0.000322 | - | 2.01e-08 | 0.9679 | 5.0867 |
| 10 | - | - | 0.137 | 0.868 | 0.9017 | 30.0330 |
| 11 | 0.00677 | 0.000669 | 0.319856 | - | 0.9739 | 3.8561 |
| 12 | 0.19616 | 0.00216 | - | 0.91981 | 0.9707 | 4.9351 |
| 13 | 0.0115 | - | 0.0333 | 0.0170 | 0.9482 | 12.5541 |
| 14 | - | 0.00155 | 0.51261 | 0.02395 | 0.9661 | 6.4933 |
| 15 | 0.0986 | 0.0149 | 0.2017 | 0.3819 | 0.9735 | 5.0000 |

(1) 유의수준 5%로 Forward selection, Backward elimination, Stepwise selection 방법에 의하여 변수를 선택하고 결과를 비교하라.

(2) 모든 가능한 모델을 대상으로 $R^2_{adj}$ 또는 $C_p$을 고려하여 가장 좋은 모델을 선택한다면 각 경우에 어느 모델이 좋겠는가?

(3) 위 (1)과 (2)의 결과로부터 가장 바람직한 모델을 선택한다면 어느 모델을 선택하겠는지 밝히고 그 이유를 설명하라.

◆ Variable selection with collinear data

Perform principal component procedure...

◎ Example ( the Homicide data ) on p.314

to investigate the role of firearms in accounting for the rising homicide rate in Detroit..

data were collected for the years 1961-1973

- Variable decription : Table 11.6 on p.315
- Data : Table 11.7-8 on pp.315-316

  ⇒ use these data to illustrate the danger of mechanical variable selection procedures
- model : $H = \beta_0 + \beta_1 G_1 + \beta_2 M + \beta_3 W + \epsilon$
- centered and scaled model : $\widetilde{H} = \theta_1 \widetilde{G_1} + \theta_2 \widetilde{M} + \theta_3 \widetilde{W} + \epsilon'$
- OLS result : Table 11.9 on p.316

  $VIF_1 = 42$, $VIF_3 = 51$ : multicollinearity, $G$ is not significant... but..
- Variable selection procedure :

  Forward selection : G-M-W ⇒ (f) is the final model

  Backward elimination : delete G, ⇒ (g) is the final model

  Stepwise procedure : G-M-W- delete G ⇒ (g) is the final model

⇒ the first variable eliminated by the BE is the first variable selected by the FS... $G$

⇒ this example shows clearly that automatic applications of variable selection procedure in multicollinear data can lead to the selection of a wrong model...

◆ A possible strategy for fitting regression model

1. Examine variables : $Y, X_1, \ldots, X_p$ : one at a time

   try to make them not be too skewed $\Rightarrow$ make transformation!

2. Construct pairwise scatter plots :

   point out obvious collinearity $\Rightarrow$ delete redundant variables..

3. Fit the full regression model

   delete variables with no significant explanatory power.

   - check linearity
   - check heteroscedasticity
   - look for high leverage pt, outlier, influential pt.

4. add or delete variables and repeat 3

   monitor the fitting process by examining $C_p, AIC, BIC, \ldots$

5. For the final model, check $VIF$'s, residual plots

6. validate the fitted model : cross validation...