Model Building with Variable Selection – Part I

Johns Hopkins Engineering

625.461 Statistical Models and Regression

Module 9 – Lecture 9B

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Basics behind Variable Selection

y: response variable

$x_1, \ldots, x_K$: regressors

$$n \geq K + 1$$

Assume: intercept is always in the model

Fit the model containing all K regressors

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \mathbf{X}: \ n \times (K+1)$$

Basics behind Variable Selection

Suppose that we delete r regressors and retain $p = K - r + 1$ regressors

$$\mathbf{y} = \mathbf{X}_p\boldsymbol{\beta}_p + \mathbf{X}_r\boldsymbol{\beta}_r + \boldsymbol{\varepsilon}$$

For the full model, the LS estimator of β

$$\hat{\boldsymbol{\beta}}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Basics behind Variable Selection

An estimator of the residual variance

$$\hat{\sigma}_*^2 = \frac{\mathbf{y}'\mathbf{y} - \hat{\beta}^{*\prime}\mathbf{X}'\mathbf{y}}{n-K-1} = \frac{\mathbf{y}'\left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{y}}{n-K-1}$$

Basics behind Variable Selection

The components of $\hat{\boldsymbol{\beta}}^*$ are $\hat{\boldsymbol{\beta}}_p^*$ and $\hat{\boldsymbol{\beta}}_r^*$.

For the subset model containing *K-r* regressors,

$$\mathbf{y} = \mathbf{X}_p \boldsymbol{\beta}_p + \boldsymbol{\varepsilon}$$

$$\hat{\boldsymbol{\beta}}_p = (\mathbf{X}_p' \mathbf{X}_p)^{-1} \mathbf{X}_p' \mathbf{y}$$

$$\hat{\sigma}^2 = \frac{\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}_p' \mathbf{X}_p' \mathbf{y}}{n-p} = \frac{\mathbf{y}'\left[\mathbf{I} - \mathbf{X}_p (\mathbf{X}_p' \mathbf{X}_p)^{-1} \mathbf{X}_p'\right]\mathbf{y}}{n-p}$$

Basics behind Variable Selection

$$E\left(\hat{\boldsymbol{\beta}}_p\right) = \boldsymbol{\beta}_p + (\mathbf{X}_p'\mathbf{X}_p)^{-1}\mathbf{X}_p'\mathbf{X}_r\boldsymbol{\beta}_r = \boldsymbol{\beta}_p + \mathbf{A}\boldsymbol{\beta}_r$$

Alias matrix: $\mathbf{A} = (\mathbf{X}'_p\mathbf{X}_p)^{-1}\mathbf{X}'_p\mathbf{X}_r$

$\hat{\boldsymbol{\beta}}_p$ is biased for β_p unless $\mathbf{X}'_p\mathbf{X}_r = 0$

$$\operatorname{Var}\left(\hat{\boldsymbol{\beta}}_p\right) = \sigma^2(\mathbf{X}_p'\mathbf{X}_p)^{-1} \qquad \operatorname{Var}\left(\hat{\boldsymbol{\beta}}^*\right) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

$\operatorname{Var}\left(\hat{\boldsymbol{\beta}}_p^*\right) - \operatorname{Var}\left(\hat{\boldsymbol{\beta}}_p\right)$ is positive definite

Basics behind Variable Selection

$$E\left(\hat{\sigma}^2\right) = \sigma^2 + \frac{\boldsymbol{\beta}_r' \mathbf{X}_r' \left[\mathbf{I} - \mathbf{X}_p \left(\mathbf{X}_p' \mathbf{X}_p\right)^{-1} \mathbf{X}_p'\right] \mathbf{X}_r \boldsymbol{\beta}_r}{n - p}$$

The subset-model estimator of σ^2 is biased upward for σ^2

Basics behind Variable Selection

Suppose that we wish to predict the response at $\mathbf{x}' = [\mathbf{x}_p', \mathbf{x}_r']$.

If we use the full model, the predicted value is $\hat{y}^* = \mathbf{x}'\hat{\boldsymbol{\beta}}^*$, with mean $\mathbf{x}'\boldsymbol{\beta}$ and predicted variance

$$\mathrm{Var}(\hat{y}^*) = \sigma^2\left[1 + \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}\right]$$

Basics behind Variable Selection

If the subset model is used, $\hat{y} = \mathbf{x}_p' \hat{\boldsymbol{\beta}}_p$

$$E(\hat{y}) = \mathbf{x}_p' \boldsymbol{\beta}_p + \mathbf{x}_p' \mathbf{A} \boldsymbol{\beta}_r$$

$$\mathrm{MSE}(\hat{y}) = \sigma^2 \left[1 + \mathbf{x}_p' (\mathbf{X}_p' \mathbf{X}_p)^{-1} \mathbf{x}_p \right] + (\mathbf{x}_p' \mathbf{A} \boldsymbol{\beta}_r - \mathbf{x}_r' \boldsymbol{\beta}_r)^2$$

$$Var(\hat{y}^*) \geq Var(\hat{y})$$

JOHNS HOPKINS

WHITING SCHOOL
of ENGINEERING