Model selection [F. 3.4]

L12
24.02.2017

SPSE = Expected squared prediction error

$E(Y) = \mu$ is the truth, but we model μ as $\mu = X\beta_M$ and we assume

$$Y = X\beta_M + \varepsilon \quad , \quad E(\varepsilon) = 0, \ Cov(\varepsilon) = \sigma^2 I$$

$M \subset \{0, 1, 2, .., k\}$ and $|M|$ = size of model = # parameters

- M: our model is based on a subset of all the available covariates
- $\{0, 1, 2, .., k\}$: available covariates $x_1, .., x_k$

TRAINING: $i = 1, .., n$ our observations, available Y_i, X_i^T and $\hat{\beta}_M$ is the estimator from the training set for our model M.

VALIDATION: $j = 1, .., J$ new observations, available as Y_j and X_j^T.

$$\sum_{j=1}^{J} E\left((Y_j - \hat{Y}_{jM})^2 \right) = SPSE$$

- Y_j: new obs
- $\hat{Y}_{jM}$: predicted value based on $\hat{\beta}_M$ and x_j^T

lots of work

$$= \cdots =$$

$\mathrm{Cov}(\varepsilon) = \sigma^2 I$

$E(x_j^T \hat{\beta}_M)$ — expected

μ_j — true value of Y_j

$$n\sigma^2 + |M| \cdot \sigma^2 + \sum_{j=1}^{J} (\mu_{jM} - \mu_j)^2$$

- $n\sigma^2$: variance of errors = irreducible prediction error
- $|M| \cdot \sigma^2$: variance of model = can be made smaller by choosing fewer variables
- $\sum_{j=1}^{J} (\mu_{jM} - \mu_j)^2$: squared bias = this can be made smaller by including more variables

bias-variance trade-off

Want to minimize SPSE, but difficult since σ^2 and μ_j is unknown.

PLAN: estimate SPSE and choose the M that minimizes this estimate.

Finding the best model ← all subsets method

smallest SSE

1) Have k covariates that might be used

$$Y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

How many possible models can I make (want to have intercept)?

$2 \cdot 2 \cdot 2 \cdots 2 = 2^k$ possible models

Fit all possible 2^k models.

2) For $M = \{0, 1, 2, \ldots, k\}$ choose the model with the smallest SSE.

Ex: Add rain: $k = 7 \Rightarrow$ total $2^7 = 128$ possible models

complexity	searched	best model
~~$\lvert M\rvert = 0$~~ 1	1	—
$\lvert M\rvert =$ ~~1~~ 2:	7	x_4 (Al)
$\lvert M\rvert =$ ~~2~~ 3:	$\binom{7}{2} = 21$	x_1, x_3
$\lvert M\rvert =$ ~~3~~ 4:	$\binom{7}{3} = 35$	x_1, x_2, x_3
$\lvert M\rvert =$ ~~4~~ 5:	$\binom{7}{4} = 35$	x_1, x_2, x_3, x_5
$\lvert M\rvert =$ ~~5~~ 6	$\binom{7}{5}$	x_1, x_2, x_3, x_5, x_7
$\lvert M\rvert =$ ~~6~~ 7	$\binom{7}{6}$	$x_1, x_2, x_3, x_4, x_5, x_7$
$\lvert M\rvert =$ ~~7~~ 8	$\binom{7}{7} = 1$	$x_1, x_2, x_3, x_4, x_5, x_6, x_7$
	2^7	

3) Now we need to choose between these k+1 models found in 2). Which criterion should I use?

$|M| = p = k+1$

i) $R^2_{adj} = 1 - \dfrac{\frac{SSE}{(n-p)}}{\frac{SST}{n-1}}$

Ex: Happiness:

$|M|=1$: only intercept

$|M|=2$: love (3) : 60.5

$|M|=3$: love + work (10) : 66.3

$|M|=4$: love + work + money (13) : 68.4

$|M|=5$: love + work + sex + money (15) : 67.6

ii) Mallows' $C_p = \dfrac{SSE}{\hat{\sigma}^2_{FULL}} - n + 2|M|$

vs $\hat{SPSE} = SSE + 2|M| \cdot \hat{\sigma}^2_{FULL}$ ← give same result.

iii) $AIC = n \cdot \ln(\hat{\sigma}^2) + 2(|M|+1)$

$\hat{\sigma}^2 = \dfrac{SSE}{n - p_{k+1}}$

iv) $BIC = n \cdot \ln(\hat{\sigma}^2) + \ln(n)(|M|+1)$

BIC gives more penalty than AIC to large models.

Ex: Happy: BIC best model = love + work

Homework: slide 24 Acid rain

Transformation of response and predictors might improve the fit of the regression model.

The BoxCox transform

$$g_\lambda(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln(y) & \lambda = 0 \end{cases}$$

↑ Class of function

For $Y = X\beta + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 I)$ the best value of λ on based on maximizing the (profile) likelihood

$$l(\lambda) = -\frac{n}{2} \ln\left(\frac{SSE_\lambda}{n}\right) - (\lambda - 1) \sum_{i=1}^{n} \ln y_i$$

SSE_λ is the SSE when $g_\lambda(Y)$ is the response

R: boxcox(fit), see plot.