Regression Analysis Tutoring7

Seung Bong Jung

Seoul National University

November 1, 2021

Motivation of Variable Selection

Assume $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$, where $\mathsf{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\mathsf{var}(\boldsymbol{\epsilon}) = \sigma^2 I$. Let $\hat{\boldsymbol{\beta}}_F = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$ and $\hat{\boldsymbol{\beta}}_{S,1} = (\mathbf{X}_1^t\mathbf{X}_1)^{-1}\mathbf{X}_1^t\mathbf{Y}$, where $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$. Also, define $\mathbf{A} = (\mathbf{X}_1^t\mathbf{X}_1)^{-1}\mathbf{X}_1^t\mathbf{X}_2$ and

- $\hat{\boldsymbol{\beta}}_{F,1} = \hat{\boldsymbol{\beta}}_{S,1} - \mathbf{A}\hat{\boldsymbol{\beta}}_{F,2}$, $\hat{\boldsymbol{\beta}}_{F,2} = (\mathbf{X}_{2,\perp}^t\mathbf{X}_{2,\perp})^{-1}\mathbf{X}_{2,\perp}^t\mathbf{Y}$.
- $\mathsf{var}(\hat{\boldsymbol{\beta}}_{F,1}) = \mathsf{var}(\hat{\boldsymbol{\beta}}_{S,1}) + \mathbf{A}\mathsf{var}(\hat{\boldsymbol{\beta}}_{F,2})\mathbf{A}^t$ since $\mathbf{X}_1 \perp \mathbf{X}_{2,\perp}$, so that $\mathsf{cov}(\hat{\boldsymbol{\beta}}_{S,1}, \hat{\boldsymbol{\beta}}_{F,2}) = \mathbf{O}$. Also, $\mathsf{E}(\hat{\boldsymbol{\beta}}_{S,1}) - \boldsymbol{\beta}_1 = \mathbf{A}\boldsymbol{\beta}_2$.
- Comparison of the mean squared errors of $\hat{\boldsymbol{\beta}}_{F,1}$ and $\hat{\boldsymbol{\beta}}_{S,1}$:

$$\begin{aligned}
&\mathsf{E}(\hat{\boldsymbol{\beta}}_{F,1} - \boldsymbol{\beta}_1)(\hat{\boldsymbol{\beta}}_{F,1} - \boldsymbol{\beta}_1)^t - \mathsf{E}(\hat{\boldsymbol{\beta}}_{S,1} - \boldsymbol{\beta}_1)(\hat{\boldsymbol{\beta}}_{S,1} - \boldsymbol{\beta}_1)^t \\
&\quad = \mathbf{A}[\mathsf{var}(\hat{\boldsymbol{\beta}}_{F,2}) - \boldsymbol{\beta}_2\boldsymbol{\beta}_2^t]\mathbf{A}^t
\end{aligned}$$

Motivation of Variable Selection

- Two predictions of $Y_{\mathbf{z}}$ at $\mathbf{z}$: Decomposing $\mathbf{z}^t$ into $(\mathbf{z}_1^t, \mathbf{z}_2^t)$ in the same way as $\mathbf{X}$ into $(\mathbf{X}_1, \mathbf{X}_2)$,

$$\hat{Y}_{\mathbf{z}}(F) := \mathbf{z}_1^t \hat{\boldsymbol{\beta}}_{F,1} + \mathbf{z}_2^t \hat{\boldsymbol{\beta}}_{F,2}, \ \hat{Y}_{\mathbf{z}}(S) := \mathbf{z}_1^t \boldsymbol{\beta}_{S,1},$$

- Observe that the mean squared prediction error (for both predictions) satisfies

$$\mathsf{E}(\hat{Y}_{\mathbf{z}} - Y_{\mathbf{z}})^2 = \mathsf{E}(\hat{Y}_{\mathbf{z}} - \mathbf{z}^t \boldsymbol{\beta})^2 + \sigma^2,$$

where the first expectation is taken for both the sample $(Y_1, \ldots, Y_n)$ and the new $Y_{\mathbf{z}}$ that is to be predicted.

Motivation of Variable Selection

- MSPE Comparison of the two predictions: Using the identity $\hat{Y}_{\mathbf{z}}(F) = \mathbf{z}_1^t \hat{\boldsymbol{\beta}}_{S,1} + (\mathbf{z}_2 - \mathbf{A}^t \mathbf{z}_1)^t \hat{\boldsymbol{\beta}}_{F,2}$ and the fact that the covariance between $\hat{\boldsymbol{\beta}}_{S,1}$ and $\hat{\boldsymbol{\beta}}_{F,2}$ equals zeros, one gets

$$\mathsf{E}(\hat{Y}_{\mathbf{z}}(F) - \mathbf{z}^t \boldsymbol{\beta})^2 = \mathbf{z}_1^t \mathsf{var}(\hat{\boldsymbol{\beta}}_{S,1}) \mathbf{z}_1 + (\mathbf{z}_2 - \mathbf{A}^t \mathbf{z}_1)^t \mathsf{var}(\hat{\boldsymbol{\beta}}_{F,2})(\mathbf{z}_2 - \mathbf{A}^t \mathbf{z}_1),$$
$$\mathsf{E}(\hat{Y}_{\mathbf{z}}(S) - \mathbf{z}^t \boldsymbol{\beta})^2 = \mathbf{z}_1^t \mathsf{var}(\hat{\boldsymbol{\beta}}_{S,1}) \mathbf{z}_1 + (\mathbf{z}_2 - \mathbf{A}^t \mathbf{z}_1)^t \boldsymbol{\beta}_2 \boldsymbol{\beta}_2^t (\mathbf{z}_2 - \mathbf{A}^t \mathbf{z}_1)$$

- Both the estimation of $\boldsymbol{\beta}_1$ and the prediction of $Y_{\mathbf{z}}$ based on the submodel give better results if $\mathsf{var}(\hat{\boldsymbol{\beta}}_{F,2}) - \boldsymbol{\beta}_2 \boldsymbol{\beta}_2^t$ is positive definite.
- One does variable selection to avoid multicollinearity and also to improve the accuracy of parameter estimation and prediction of the response.

Adjusted R^2

How to select useful predictors? One may consider the coefficient of determination, which turns out to be SSR/SST. But, this is not a good criterion because it is nondecreasing as a new predictor enters the model.

- Suppose that the totality of all predictors at hands are $x_1, \ldots, x_p$. We want to select a subset S of the index set $\{1, 2, \ldots, \}$. Let $|S|$ denote the cardinality of the set S and $q = |S| + 1$.
- Let $R^2(S)$ and SSE(S) denote the coefficient of determination and the residual sum of squares, respectively, when Y is regressed on $\{x_j : j \in S\}$ with an intercept term: $R^2(S) = R(\beta_j : j \in S|\beta_0)/\text{SST}$.
- Adjusted R^2 and the mean squared residual:

$$R_a^2(S) := 1 - (\frac{n-1}{n-q})(1 - R^2(S)) \Leftrightarrow \text{MSE} := \frac{\text{SSE}(S)}{n-q}$$

Mallows's C_p

Now, let $F := \{0, 1, \ldots, p\}$, and think of selecting a subset S of F. Let $\hat{\sigma}^2 = \mathsf{SSE}(F)/(n-p-1)$. The Mallows's C_p statistic is defined by

$$C_p(S) := \frac{\mathsf{SSE}(S)}{\hat{\sigma}^2} - n + 2|S|.$$

- The subset S may not include 0, that is, the C_p statistics is defined in a more general setting where models without the intercept term are also considered to be selected.
- The C_p statistic is an estimate of $\mathsf{E}(\mathsf{ASE}(S))$ where $\mathsf{ASE}(S)$ is the average squared error of the submodel S, defined by

$$\mathsf{ASE}(S) = n^{-1} \sum_{i=1}^{n} (\hat{Y}_i(S) - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2,$$

where $\hat{Y}_i(S)$ is the LS estimate of $\mathsf{E}(Y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ based on $\{(x_{ij}, Y_i) : j \in S\}$.

Derivation of C_p Statistic

- Let $\boldsymbol{\beta}_1 = (\beta_j : j \in S)$ and $\boldsymbol{\beta}_2 = (\beta_j : j \notin S)$. Without loss of generality, assume $\boldsymbol{\beta}^t = (\boldsymbol{\beta}_1^t, \boldsymbol{\beta}_2^t)$. Decompose $\mathbf{X}$ into $(\mathbf{X}_1, \mathbf{X}_2)$ so that $\mathbf{X}\boldsymbol{\beta} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2$.
- Define $\hat{\boldsymbol{\beta}}_{S,1} = (\mathbf{X}_1^t\mathbf{X}_1)^{-1}\mathbf{X}_1^t\mathbf{Y}$, and $\hat{\boldsymbol{\beta}}_S$ by $\hat{\boldsymbol{\beta}}_S^t = (\hat{\boldsymbol{\beta}}_{S,1}^t, \mathbf{0}^t)$.
- $\mathsf{ASE}(S) = n^{-1}(\mathbf{X}\hat{\boldsymbol{\beta}}_S - \mathbf{X}\boldsymbol{\beta})^t(\mathbf{X}\hat{\boldsymbol{\beta}}_S - \mathbf{X}\boldsymbol{\beta})$
- Recalling that $\hat{\boldsymbol{\beta}}_{S,1} = \hat{\boldsymbol{\beta}}_{F,1} + \mathbf{A}\hat{\boldsymbol{\beta}}_{F,2}$, it can be shown that

$$\mathsf{E}(\mathsf{ASE}(S)) = \frac{1}{n}[|S|\sigma^2 + \boldsymbol{\beta}_2^t\mathbf{X}_2^t(I - \Pi_{\mathbf{x}_1})\mathbf{X}_2\boldsymbol{\beta}_2]$$
$$\mathsf{E}(\mathsf{SSE}(S)) = (n - |S|)\sigma^2 + \boldsymbol{\beta}_2^t\mathbf{X}_2^t(I - \Pi_{\mathbf{x}_1})\mathbf{X}_2\boldsymbol{\beta}_2$$

so that

$$\mathsf{E}(\mathsf{ASE}(S)) = \frac{\sigma^2}{n}\mathsf{E}(\frac{\mathsf{SSE}(S)}{\sigma^2} - n + 2|S|).$$

General Framework for Model Selection

Consider a super-model

$$\mathbf{P} \equiv \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$$

that is believed to contatin the true distribution, denoted by $P_{\boldsymbol{\theta}_0}$, where $\boldsymbol{\Theta} \subset \mathbb{R}^k$. For a subset S of the index set $\{1, 2, \ldots, k\}$, define $\boldsymbol{\Theta}_S$, a subset of $\boldsymbol{\Theta}$, and the corresponding submodel $\mathbf{P}_S$ of $\mathbf{P}$ by

$$\begin{aligned} \boldsymbol{\Theta}_S &= \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \theta_j = \theta_{0j} \text{ for all } j \notin S\}, \\ \mathbf{P}_S &= \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}_s\}. \end{aligned}$$

Assume that each $P_{\boldsymbol{\theta}}$ has a density $f(\cdot, \boldsymbol{\theta})$. Define

$$K(\boldsymbol{\theta}) = -2\mathsf{E}_{\boldsymbol{\theta}_0} \log f(Y, \boldsymbol{\theta}),$$

where the expectation is taken with respect to $P_{\boldsymbol{\theta}_0}$, the the true distribution.

General Framework for Model Selection

- It is known that, as $P_{\boldsymbol{\theta}}$ gets away from $P_{\boldsymbol{\theta}_0}$ in a certain sense, the negative expected log-likelihood $k(\boldsymbol{\theta})$ increases. In fact, under certain conditions, $K(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is minimized at $\boldsymbol{\theta}_0$. It may be regarded as a distance between $P_{\boldsymbol{\theta}}$ and $P_{\boldsymbol{\theta}_0}$
- Maximum likelihood estimation of $\boldsymbol{\theta}_0$ based on the submodel $\mathbf{P}_S$:

$$\hat{\boldsymbol{\theta}}_S := \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}_{\boldsymbol{S}}}{\arg\max} \sum_{i=1}^{n} \log f(Y_i, \boldsymbol{\theta}) \ (P_{\hat{\boldsymbol{\theta}}_S} \text{ 'closest' to } P_{\boldsymbol{\theta}_0} \text{ in } \mathbf{P}_S)$$

- If K is available, we may want to select the subset S^* that minimizes $K(\hat{\boldsymbol{\theta}}_S)$ over all subsets S, since $K(\hat{\boldsymbol{\theta}}_S)$ may be regarded as the distance between $P_{\hat{\boldsymbol{\theta}}_S}$ and $P_{\boldsymbol{\theta}_0}$.
- Simply replacing K by $\hat{K} := -2n^{-1} \sum_{i=1}^{n} \log f(Y_i, \cdot)$ is not a proper way since $\hat{K}(\hat{\boldsymbol{\theta}}_S)$ underestimates $K(\hat{\boldsymbol{\theta}}_S)$ and $\hat{K}(\hat{\boldsymbol{\theta}}_{S_1}) > \hat{K}(\hat{\boldsymbol{\theta}}_{S_2})$ if $S_1 \subsetneq S_2$.

Akaike Information Criterion

- Akaike Information Criterion:

$$\mathsf{AIC}(S) := -2n^{-1}\sum_{i=1}^{n}\log f(Y_i, \hat{\boldsymbol{\theta}}_S) + 2\frac{|S|}{n}.$$

- Under certain conditions, $\mathsf{AIC}(S)$ is a good estimate of $K(\hat{\boldsymbol{\theta}}_S)$.
- In our regression setting, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ with $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^t$, $f(Y_i, \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}}\exp[-(Y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2/(2\sigma^2)]$ and $\hat{\sigma}_S^2 = n^{-1}||\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_S||^2$ where S is a subset of $\{1, \ldots, p\}$. Thus,

$$\begin{aligned}\mathsf{AIC}(S) &= \log(2\pi\hat{\sigma}_S^2) + \frac{n^{-1}||\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_S||^2}{\hat{\sigma}_S^2} + \frac{2(|S|+1)}{n} \\ &= \log \mathsf{SSE}(S) + \frac{2|S|}{n},\end{aligned}$$

where the second equation neglects the term $1 + 2n^{-1} + \log(2\pi) - \log n$.

Bayesian Information Criterion

- Bayesian Information Criterion:

$$\mathsf{BIC}(S) := -2n^{-1}\sum_{i=1}^{n}\log f(Y_i, \hat{\boldsymbol{\beta}}_S) + \frac{|S|\log n}{n}$$

- It was derived from a Bayesian framework. In fact, $\mathsf{BIC}(S)$ is a good estimate of the log-posterior probability for the model $\mathbf{P}_S$.
- In our regression setting,

$$\mathsf{BIC}(S) = \log \mathsf{SSE}(S) + \frac{|S|\log n}{n}$$

- The model selection criteria, $C_p(S), \mathsf{AIC}(S)$ and $\mathsf{BIC}(S)$ take the form

$$(\text{Goodness-of-fit}) + (\text{Model Complexity}).$$

- BIC penalizes larger (more complex) models more heavily than AIC when $\log n > 2$, so that it prefers smaller models in comparison with AIC.