PhD in Economics and Statistics - University of Milano-Bicocca

Lecture 5: The variable selection problem

May 9, 2019

Lecturer: Aldo Solari

1 Linear model

Consider a random response $\underset{n\times 1}{y}$ and a fixed design matrix $\underset{n\times p}{X}$ of full column rank, i.e. $\mathrm{rk}(X) = p$, whose columns correspond to predictors, with $n > p$. We assume a Gaussian model

$$y \sim N_n(X\beta, \sigma^2 I_n)$$

where $\underset{n\times 1}{\mu} = X\beta$ is the mean vector and $\underset{p\times 1}{\beta}$ and σ^2 are unknown parameters.

Let $P = X(X^\mathsf{T} X)^{-1}X^\mathsf{T}$ be the orthogonal projector onto $\mathrm{Sp}(X)$, the column space of X. Recall that P is symmetric (i.e. $P = P^\mathsf{T}$) and idempotent (i.e. $P = PP$) with rank $\mathrm{rk}(P) = p$ and trace $\mathrm{tr}(P) = p$.

The least squares estimator for μ is given by

$$\hat{\mu} = Py \sim N_n(\mu, \sigma^2 P)$$

To see this, note that $P\mu = \mu$ and recall that if $z \sim N_p(\mu, \Sigma)$ and $\underset{q\times p}{B}$ is a $q \times p$ matrix, then $Bz \sim N_q(B\mu, B\Sigma B^\mathsf{T})$.

An unbiased estimator for σ^2 is given by

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n-p} = \frac{\|y - \hat{\mu}\|^2}{n-p} \sim \sigma^2 \frac{\chi^2_{n-p}}{n-p}$$

where $\|a\|^2 = a^\mathsf{T} a = \sum_{i=1}^q a_i$ denotes the squared Euclidean norm for a vector $\underset{q\times 1}{a} = (a_1, \ldots, a_q)^\mathsf{T}$. Recall that if $z \sim N_p(0, I_p)$ and $\underset{p\times p}{B}$ is a symmetric semidefinite matrix with $\mathrm{rk}(B) = r$, then the quadratic forms $z^\mathsf{T} Bz \sim \chi^2_r$ and $z^\mathsf{T}(I_p - B)z \sim \chi^2_{p-r}$ are independent.

Finally, the estimator for β is given by

$$\hat{\beta} = (X^\mathsf{T} X)^{-1}X^\mathsf{T} y \sim N_p(\beta, \sigma^2(X^\mathsf{T} X)^{-1})$$

and it is independent from $\hat{\sigma}^2$.

Confidence intervals for components of β are based on the distributions of $\hat{\beta}$ and $\hat{\sigma}^2$. Under the normal linear model the ith element of β

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2 v_i)$$

where v_i is the ith diagonal element of $V = (X^\mathsf{T} X)^{-1}$, therefore

$$t_i = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}\sqrt{v_i}} \sim t_{n-p}$$

and a $1-\alpha$ confidence interval for β_i is

$$\mathrm{CI}_i = \hat{\beta}_i \pm t_\alpha \hat{\sigma}\sqrt{v_i}$$

where $t_\alpha = t_{1-\alpha/2,n-p}$ is the $1-\alpha/2$ quantile of the Student t distribution with $n-p$ degrees of freedom. Then the interval is marginally valid with a $1-\alpha$ coverage guarantee

$$\mathrm{P}(\beta_i \in \mathrm{CI}_i) \geq 1-\alpha$$

This is useful if the predictor of interest is the ith predictor.
If all predictors are of interest, we may consider constructing simultaneous confidence intervals such that

$$\mathrm{P}(\beta_i \in \widetilde{\mathrm{CI}}_i \ \forall i \in \{1,\ldots,p\}) \geq 1-\alpha$$

A simple but conservative solution is given by Bonferroni which uses

$$\widetilde{\mathrm{CI}}_i = \hat{\beta}_i \pm t_{\alpha/m} \hat{\sigma}\sqrt{v_i}$$

but in general better solutions can be found.
Suppose we want to predict a future observation

$$y^* \sim N_n(X\beta, \sigma^2 I_n)$$

by $\hat{y}^* = X\hat{\beta} = \hat{\mu}$. Then the prediction error is

$$\begin{aligned}
\mathrm{PE} &= \mathrm{E}\|y^* - \hat{y}^*\|^2 = \mathrm{E}\|(\varepsilon^*) + (\mu - \hat{\mu})\|^2 \\
&= \mathrm{E}\|\varepsilon^*\|^2 + \mathrm{E}\|\mu - \hat{\mu}\|^2 + 2\mathrm{E}[(\varepsilon^*)^\mathsf{T}(\mu - \hat{\mu})] \\
&= n\sigma^2 + \mathrm{E}\|P(\mu - y)\|^2 + 0 = n\sigma^2 + \mathrm{E}\|P\varepsilon\|^2 \\
&= n\sigma^2 + p\sigma^2
\end{aligned}$$

where ε and ε^* are independent and identically distributed as $N_n(0, \sigma^2 I_n)$, $\|\varepsilon\|^2 \sim \sigma^2\chi^2_n$.

2 The variable selection problem

A *variable selection problem* arises when the researcher suspects that some regressors in the full model are not necessary for explaining or predicting y, but does not know which.

We denote a (sub)model by the index set $M \subseteq F = \{1,\ldots,p\}$ of regressors it includes, with size $m = |M|$. Let X_M be the design matrix of model M, i.e. the submatrix of X with columns indexed by M. For the particular case of the full model $M = F$, we have $X_F = X$ and size $|F| = p$. Then the candidate models are all submodels of the full model F, i.e.

$$\mathcal{M} = \{M \cup \{1\} : M \subseteq F\} \tag{1}$$

where we required that each model contains the intercept term, which by convention is the first column of X, i.e. $X_{\{1\}} = 1_n$. The number of candidate models is $|\mathcal{M}| = 2^{p-1}$. The set-up just described is often termed *all subset selection*.

Let $P_M = X_M(X_M^\mathsf{T} X_M)^{-1} X_M^\mathsf{T}$ is the orthogonal projector onto $\mathrm{Sp}(X_M)$, the column space of X_M, with $\mathrm{Sp}(X_M) \subseteq \mathrm{Sp}(X)$. Each candidate model M has mean parameter

$$\mu_M = P_M \mu$$

and coefficients

$$\beta_M = (X_M^\mathsf{T} X_M)^{-1} X_M^\mathsf{T} \mu$$

We will use the notation

$$\beta_{i \cdot M}, \quad i \in M$$

for the components of β_M.

What is the relationship to full model coefficients? A little algebra shows that

$$\beta_M = (\beta_{i \cdot M}, i \in M)^\mathsf{T} = (\beta_{i \cdot F}, i \in M)^\mathsf{T}$$

if and only if

$$X_M^\mathsf{T} X_{F \setminus M} (\beta_{i \cdot F}, i \in F \setminus M)^\mathsf{T} = \underset{m \times 1}{0}$$

This happens if

$$(\beta_{i \cdot F}, i \in F \setminus M)^\mathsf{T} = \underset{p-m \times 1}{0}$$

and if the column space of X_M is orthogonal to that of $X_{F \setminus M}$, i.e.

$$\mathrm{Sp}(X_M) \perp \mathrm{Sp}(X_{F \setminus M}).$$

How many parameters do we have? We have 2^{p-1} candidate models. For each candidate model, we have m coefficient parameters $\beta_{i \cdot M}$. The intercept appears in all 2^{p-1} submodels, and each regressor appears in 2^{p-2} submodels. This implies that the overall number of parameters is

$$\sum_{M \in \mathcal{M}} |M| = 2^{p-1} + (p-1)2^{p-2}$$

2.1 Submodels

Based on a model M, the estimator of μ_M is given by

$$\hat{\mu}_M = P_M y \sim N_n(\mu_M, \sigma^2 P_M)$$

and the estimator for β_M is given by

$$\hat{\beta}_M = (X_M^\mathsf{T} X_M)^{-1} X_M^\mathsf{T} y \sim N_m(\beta_M, \sigma^2 (X_M^\mathsf{T} X_M)^{-1})$$

We have

$$\hat{\beta}_{i\cdot M} \sim N(\beta_{i\cdot M}, \sigma^2 v_{i\cdot M})$$

where $v_{i\cdot M}$ is the ith diagonal element of $V_M = (X_M^\mathsf{T} X_M)^{-1}$. If we estimate σ^2 by the full model estimator $\hat{\sigma}^2 = \|y - \hat{\mu}\|^2/(n-p)$, which is independent from $\hat{\beta}_M$ for all $M \in \mathcal{M}$, then

$$t_{i\cdot M} = \frac{\hat{\beta}_{i\cdot M} - \beta_{i\cdot M}}{\hat{\sigma}\sqrt{v_{i\cdot M}}} \sim t_{n-p}$$

and a $1-\alpha$ confidence interval for $\beta_{i\cdot M}$ is

$$\mathrm{CI}_{i\cdot M} = \hat{\beta}_{i\cdot M} \pm t_\alpha \hat{\sigma}\sqrt{v_{i\cdot M}}$$

where $t_\alpha = t_{1-\alpha/2, n-p}$ is the $1-\alpha/2$ quantile of the Student t distribution with $n-p$ degrees of freedom. Then the interval is marginally valid with a $1-\alpha$ coverage guarantee

$$\mathrm{P}(\beta_{i\cdot M} \in \mathrm{CI}_{i\cdot M}) \geq 1 - \alpha$$

This holds if the submodel M is specified a priori, that is, it is not the result of a variable selection algorithm.
Suppose we want to predict a future observation $y^* \sim N_n(X\beta, \sigma^2 I_n)$ by $\hat{y}^*_M = X\hat{\beta}_M = \hat{\mu}_M$. Then the prediction error is

$$\begin{aligned}
\mathrm{PE}_M &= \mathrm{E}\|y^* - \hat{y}^*_M\|^2 = n\sigma^2 + \mathrm{E}\|\mu - P_M y\|^2 \\
&= n\sigma^2 + \mathrm{E}\|(\mu - \mu_M) + (\mu_M - P_M y)\|^2 \\
&= n\sigma^2 + \|\mu - \mu_M\|^2 + \mathrm{E}\|\mu_M - P_M y\|^2 + 2\mathrm{E}[(\mu - \mu_M)^\mathsf{T}(\mu_M - P_M y)] \\
&= n\sigma^2 + \|\mu - \mu_M\|^2 + \mathrm{E}\|P_M(\mu - y)\|^2 + 0 \\
&= n\sigma^2 + \|\mu - \mu_M\|^2 + m\sigma^2
\end{aligned}$$

which decomposes into *irreducible error* $n\sigma^2$, *squared bias* $\|\mu - \mu_M\|^2$ and *variance* $m\sigma^2$.

2.2 Variable selection procedures

In practice, the model M tends to be the result of some variable selection procedure that makes use of the stochastic component of the data y (X being fixed). For example, *best subset selection*:

- Set B_1 as the null model (only intercept)
- For $m = 2, \ldots, p$:

1. Fit all $\binom{p-1}{m-1}$ models of size m that contain exactly $m-1$ regressors and the intercept
2. Pick the "best" among these $\binom{p-1}{m-1}$ models, and call it B_m, where "best" is defined having the smallest residual sum of squares

$$\text{RSS}_M = \|y - \hat{\mu}_M\|^2$$

- Select a single best model from among $B_1, B_2, \ldots, B_p$ using $\mathrm{C_p}$, BIC, etc.

Note that $B_1 = \{1\}$ and $B_p = F$.
The selected model should be expressed as

$$\hat{M} = \hat{M}(y)$$

Data dependence of the selected model $\hat{M}$ has strong consequences, because the selected model $\hat{M}$ is random.

3 Post-Selection Inference

Let X be an $n \times 3$ matrix of rank 3, where the 1st column of X is the intercept term, i.e. $X_{\{1\}} = 1_n$, which is always included in the model. Suppose that we want inference for the 2nd predictor but we don't know whether or not include the 3rd, that is

$$\mathcal{M} = \{\{1,2\}, \{1,2,3\}\}$$

For the model selector, we set $\hat{M} = \{1,2,3\}$ if $\hat{\beta}_{3\cdot\{1,2,3\}}/\hat{\sigma}\sqrt{v_{3\cdot\{1,2,3\}}}$ is larger than $t_{1-0.05/2,n-3}$, and $\hat{M} = \{1,2\}$ otherwise. We are interested in the coverage probability of the interval (that ignores the selection)

$$\text{CI}_2 = \hat{\beta}_{2\cdot\hat{M}} \pm t_{1-0.05/2,n-3}\hat{\sigma}\sqrt{v_{2\cdot\hat{M}}}$$

in two scenarios:

- the target is fixed $\beta_{2\cdot\{1,2,3\}}$, i.e.
$$\text{P}(\beta_{2\cdot\{1,2,3\}} \in \text{CI}_2)$$
- the target is random $\beta_{2\cdot\hat{M}}$, i.e.
$$\text{P}(\beta_{2\cdot\hat{M}} \in \text{CI}_2)$$

4 PoSI

The PoSI procedure proposed by Berk et al. (2013) produces a constant K_{PoSI} that provides universally valid post-selection inference when the target is random.

Theorem 4.1. *Let* K_{PoSI} *such that*

$$\mathrm{P}(\max_{M \in \mathcal{M}} \max_{i \in M} |t_{i \cdot M}| \leq K_{\mathrm{PoSI}}) \geq 1 - \alpha$$

Then with

$$\mathrm{CI}_{i \cdot \hat{M}} = \hat{\beta}_{i \cdot \hat{M}} \pm K_{\mathrm{PoSI}} \hat{\sigma} \sqrt{v_{i \cdot \hat{M}}}$$

we have

$$\mathrm{P}(\beta_{i \cdot \hat{M}} \in \mathrm{CI}_{i \cdot \hat{M}} \ \forall i \in \hat{M}) \geq 1 - \alpha \quad \forall \ \hat{M}$$

Proof. For any $\hat{M}$, the following inequality holds

$$\max_{i \in \hat{M}} |t_{i \cdot \hat{M}}| \leq \max_{M \in \mathcal{M}} \max_{i \in M} |t_{i \cdot M}|$$

By definition, K_{PoSI} is equal or greater than the $1-\alpha$ quantile of the distribution of $\max\limits_{M \in \mathcal{M}} \max\limits_{i \in M} |t_{i \cdot M}|$ Then

$$\mathrm{P}(\max_{i \in \hat{M}} |t_{i \cdot \hat{M}}| \leq K_{\mathrm{PoSI}}) \geq 1 - \alpha$$

$\square$

The PoSI constant K_{PoSI} depends on the design matrix X, the collection of candidate models $\mathcal{M}$, the desired coverage $1 - \alpha$ and the degrees of freedom $r = n - p$ in $\hat{\sigma}^2$, hence

$$K_{\mathrm{PoSI}} = K_{\mathrm{PoSI}}(X, \mathcal{M}, \alpha, r)$$

It turns out the Scheffe constant

$$K_{\mathrm{Scheffe}} = \sqrt{p f_{1-\alpha, p, n-p}}$$

provides an upper bound for the PoSI constant

$$K_{\mathrm{PoSI}} \leq K_{\mathrm{Scheffe}}$$

4.1 PoSI1

Sometimes the interest is on the ith predictor only. Here variable selection is limited to the models that contain this predictor (and the intercept term)

$$\mathcal{M} = \{M \cup \{1, i\} : M \subseteq F \setminus \{1, i\}\}$$

Let K_{PoSI1} be such that

$$\mathrm{P}(\max_{M \in \mathcal{M}} |t_{i \cdot M}| \leq K_{\mathrm{PoSI1}}) \geq 1 - \alpha$$

then

$$\mathrm{P}(\beta_{i \cdot \hat{M}} \in \mathrm{CI}_{i \cdot \hat{M}}) \geq 1 - \alpha \quad \forall\ \hat{M}$$

with

$$\mathrm{CI}_{i \cdot \hat{M}} = \hat{\beta}_{i \cdot \hat{M}} \pm K_{\mathrm{PoSI1}} \hat{\sigma} \sqrt{v_{i \cdot \hat{M}}}$$

and

$$K_{\mathrm{PoSI1}} \leq K_{\mathrm{PoSI}}$$