Lecture 24: Variable selection in linear models

Consider linear model $X = Z\beta + \varepsilon$, $\beta \in \mathscr{R}^p$ and $\mathrm{Var}(\varepsilon) = \sigma^2 I_n$.

Like the LSE, the ridge regression estimator does not give 0 estimate to a component of β even if that component is 0.

Variable (or model) selection refers to eliminating covariates (columns of Z) corresponding to zero components of β.

Example 1. Linear regression models

- $\mathscr{A}$ = a subset of $\{1,...,p\}$, indices of nonzero components of β
- The dimension of $\mathscr{A}$ is $\dim(\mathscr{A}) = q \le p$
- $\beta_{\mathscr{A}}$: sub-vector of β with indices in $\mathscr{A}$
- $Z_{\mathscr{A}}$: the corresponding sub-matrix of Z
- The number of models could be as large as 2^p
- Approximation to a response surface
    - The ith row of $Z_{\mathscr{A}} = (1, t_i, t_i^2, ..., t_i^h)$, $t_i \in \mathscr{R}$
    - $\mathscr{A} = \{1,...,h\}$: a polynomial of order h
    - $h = 0, 1, ..., p_n$

Example 2. 1-mean vs p-mean

- $n = pr$, $p = p_n$, $r = r_n$
- There are p groups, each has r identically distributed observations
- Select one model from two models
  - 1-mean model: all groups have the same mean μ_1
  - p-mean model: p groups have different means $\mu_1, ..., \mu_p$
- $\mathscr{A} = \mathscr{A}_1$ or $\mathscr{A}_p$

$$Z = \begin{pmatrix} 1_r & 0 & 0 & \cdots & 0 \\ 1_r & 1_r & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1_r & 0 & 0 & \cdots & 1_r \end{pmatrix} \qquad \beta = \begin{pmatrix} \mu_1 \\ \mu_2 - \mu_1 \\ \cdots \\ \mu_p - \mu_1 \end{pmatrix}$$

$$Z_{\mathscr{A}_p} = Z \qquad \beta_{\mathscr{A}_p} = \beta$$
$$Z_{\mathscr{A}_1} = 1_n \qquad \beta_{\mathscr{A}_1} = \mu_1$$

- In traditional studies, p is fixed and n is large, or p/n is small
- In modern applications, both p and n are large, and in some cases $p > n$, $p/n \to \infty$

Methods for variable selection

- Generalized Information Criterion (GIC)
  Put a penalty on the dimension of the parameter: We minimize

$$\|X - Z_{\mathscr{A}}\beta_{\mathscr{A}}\|^2 + \lambda\widehat{\sigma}^2 \mathrm{dim}(\beta_{\mathscr{A}}) \quad \text{over } \mathscr{A},$$

  to obtain a suitable $\mathscr{A}$, and then estimate $\beta_{\mathscr{A}}$.
  $\widehat{\sigma}^2$ is a suitable estimator of the error variance σ^2
  - The term $\|X - Z_{\mathscr{A}}\beta_{\mathscr{A}}\|^2$ measures goodness-of-fit of model $\mathscr{A}$, whereas the term $\lambda\widehat{\sigma}^2 \mathrm{dim}(\beta_{\mathscr{A}})$ controls the "size" of $\mathscr{A}$.
  - If $\lambda_n = 2$, this is the C_p method, and close to the AIC
  - If $\lambda_n = \log n$, this is close to the BIC
- Regularization or penalized optimization simultaneously select variables and estimate θ by minimizing

$$\|X - Z\beta\|^2 + p_\lambda(\beta),$$

  where $p_\lambda(\cdot)$ is a penalty function indexed by the penalty parameter $\lambda \geq 0$, which may depend on n and data.
  Zero components of β are estimated as zeros and automatically eliminated.

- Examples of penalty functions
  - Ridge regression: $p_\lambda(\beta) = \lambda\|\beta\|^2$;
  - LASSO (least absolute shrinkage and selection operator): $p_\lambda(\beta) = \lambda\|\beta\|_1 = \lambda\sum_{j=1}^{p}|\beta_j|$, β_j is the jth component of β;
  - Adaptive LASSO: $p_\lambda(\beta) = \lambda\sum_{j=1}^{p}\tau_j|\beta_j|$, where τ_j's are non-negative leverage factors chosen adaptively such that large penalties are used for unimportant β_j's and small penalties for important ones;
  - Elastic net: $p_\lambda(\beta) = \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|^2$;
  - Minimax concave penalty: $p_\lambda(\beta) = \sum_{j=1}^{p}(a\lambda - \beta_j)_+/a$ for some $a > 0$;
  - SCAD (smoothly clipped absolute deviation): $p_\lambda(\beta) = \sum_{j=1}^{p}\lambda\{I(\beta_j \le \lambda) + \frac{(a\lambda-\beta_j)_+}{(a-1)\lambda}I(\beta_j \ge \lambda)\}$ for some $a > 2$;
  - There are also many modified versions of the previously listed methods.
- Resampling methods

  Cross validation, bootstrap
- Thresholding

  Compare $\widehat{\beta}_j$ with a threshold (may depend on n and data) and eliminate estimates that are smaller than the threshold.

Assessment of variable/model selection procedures

$\mathscr{A}$ = the set containing exactly indices of nonzero components of β

$\widehat{\mathscr{A}}$: a set of variables/model selected based on a selection procedure

The selection procedure is selection consistent if

$$\lim_{n\to\infty} P\left(\widehat{\mathscr{A}} = \mathscr{A}\right) = 1$$

Sometimes the following weaker version of consistency is desired.

Under model $\mathscr{A}$, $\mu = E(X|Z)$ is estimated by $\widehat{\mu}_{\mathscr{A}} = Z_{\mathscr{A}}\widehat{\beta}_{\mathscr{A}}$

We want to minimize the squared error loss

$$L_n(\mathscr{A}) = n^{-1}\|\mu - \widehat{\mu}_{\mathscr{A}}\|^2 \qquad \text{over } \mathscr{A}$$

which is equivalent to minimizing the average prediction error

$$n^{-1}E\left[\|X^* - \widehat{\mu}_{\mathscr{A}}\|^2 \mid X, Z\right] \qquad \text{over } \mathscr{A}$$

X^*: a future independent copy of X

The selection procedure is loss consistent if

$$L_n(\widehat{\mathscr{A}})/L_n(\mathscr{A}) \to_p 1$$

Consistency of the GIC

Let $\mathscr{M}$ denote a set of indices (model).

If $\mathscr{A} \subset \mathscr{M}$, then $\mathscr{M}$ is a correct model; otherwise, $\mathscr{M}$ is a wrong model.
The loss under model $\mathscr{M}$ is equal to

$$L_n(\mathscr{M}) = \Delta_n(\mathscr{M}) + \varepsilon^\tau H_{\mathscr{M}} \varepsilon / n$$

$H_{\mathscr{M}} = Z_{\mathscr{M}}(Z^\tau_{\mathscr{M}} Z_{\mathscr{M}})^{-1} Z^\tau_{\mathscr{M}}$, $\Delta_n(\mathscr{M}) = \|\mu - H_{\mathscr{M}}\mu\|^2/n$ (0 if $\mathscr{M}$ is correct)
Let $\Gamma_{n,\lambda}(\mathscr{M}) = n^{-1}[\|X - Z_{\mathscr{M}}\beta_{\mathscr{M}}\|^2 + \lambda\widehat{\sigma}^2 \mathrm{dim}(\beta_{\mathscr{M}})]$ (to be minimized)

$$\begin{aligned}\|X - Z_{\mathscr{M}}\beta_{\mathscr{M}}\|^2 &= \|X - H_{\mathscr{M}}X\|^2 = \|\mu - H_{\mathscr{M}}\mu + \varepsilon - H_{\mathscr{M}}\varepsilon\|^2 \\ &= n\Delta_n(\mathscr{M}) + \|\varepsilon\|^2 - \varepsilon^\tau H_{\mathscr{M}}\varepsilon + 2\varepsilon^\tau(I - H_{\mathscr{M}})\mu\end{aligned}$$

When $\mathscr{M}$ is a wrong model,

$$\begin{aligned}\Gamma_{n,\lambda}(\mathscr{M}) &= \frac{\|\varepsilon\|^2}{n} + \Delta_n(\mathscr{M}) - \frac{\varepsilon^\tau H_{\mathscr{M}}\varepsilon}{n} + \frac{\lambda\widehat{\sigma}^2 \mathrm{dim}(\mathscr{M})}{n} + O_P\left(\frac{\Delta_n(\mathscr{M})}{n}\right) \\ &= \frac{\|\varepsilon\|^2}{n} + L_n(\mathscr{M}) + O_P\left(\frac{\lambda \mathrm{dim}(\mathscr{M})}{n}\right) + O_P\left(\frac{L_n(\mathscr{M})}{n}\right) \\ &= \frac{\|\varepsilon\|^2}{n} + L_n(\mathscr{M}) + o_P(L_n(\mathscr{M}))\end{aligned}$$

provided that

$$\liminf_{n\to\infty} \min_{\mathscr{M} \text{ is wrong}} \Delta_n(\mathscr{M}) > 0 \quad \text{and} \quad \frac{\lambda p}{n} \to 0$$

(The first condition impies that wrong is always worse than correct)

Among all wrong $\mathscr{M}$, minimizing $\Gamma_{n,\lambda}(\mathscr{M})$ is asymptotically the same as minimizing $L_n(\mathscr{M})$

Hence, the GIC is loss consistent when all models are wrong

The GIC selects the best wrong model, i.e., the best approximation to a correct model in terms of $\Delta_n(\mathscr{M})$, the leading term in the loss $L_n(\mathscr{M})$

For correct models, however, $\Delta_n(\alpha) = 0$ and $L_n(\mathscr{M}) = \varepsilon^\tau H_{\mathscr{M}}\varepsilon/n$

Correct models are nested, and $\mathscr{A}$ has the smallest dimension and

$$\varepsilon^\tau H_{\mathscr{A}}\varepsilon = \min_{\mathscr{M} \text{ is correct}} \varepsilon^\tau H_{\mathscr{M}}\varepsilon$$

$$\begin{aligned}
\Gamma_{n,\lambda}(\mathscr{M}) &= \frac{\|\varepsilon\|^2}{n} - \frac{\varepsilon^\tau H_{\mathscr{M}}\varepsilon}{n} + \frac{\lambda\widehat{\sigma}^2 \mathrm{dim}(\mathscr{M})}{n} \\
&= \frac{\|\varepsilon\|^2}{n} + L_n(\mathscr{M}) + \frac{\lambda\widehat{\sigma}^2 \mathrm{dim}(\mathscr{M})}{n} - \frac{2\varepsilon^\tau H_{\mathscr{M}}\varepsilon}{n}
\end{aligned}$$

If $\lambda \to \infty$, the dominating term in $\Gamma_{n,\lambda}(\mathscr{M})$ is $\lambda\widehat{\sigma}^2\dim(\mathscr{M})/n$.

Among correct models, the GIC selects a model by minimizing $\dim(\mathscr{M})$, i.e., it selects $\mathscr{A}$.

Combining the results, we showed that the GIC is selection consistent.

On the other hand, if $\lambda = 2$ (the C_p method, AIC), the term

$$\frac{2\widehat{\sigma}^2\dim(\mathscr{M})}{n} - \frac{2\varepsilon^\tau H_{\mathscr{M}}\varepsilon}{n}$$

is of the same order as $L_n(\mathscr{M}) = \varepsilon^\tau H_{\mathscr{M}}\varepsilon/n$ unless $\dim(\mathscr{M}) \to \infty$ for all but one correct model.

Under some conditions, the GIC with $\lambda = 2$ is loss consistent if and only if there does not exist two correct models with fixed dimensions.

Conclusion

(1) The GIC with a bounded λ (C_p, AIC) is loss consistent when there is at most one fixed-dimension correct model; otherwise it is inconsistent.

(2) The GIC with $\lambda \to \infty$ and $\lambda p/n \to 0$ (BIC) are selection consistent or loss consistent.

Example 2. 1-mean vs p-mean

$\mathscr{A}_1$ vs $\mathscr{A}_p$ (always correct)

p_n groups, each with r_n observations

$\Delta_n(\mathscr{A}_1) = \sum_{j=1}^{p}(\mu_j - \bar{\mu})^2/p$, $\bar{\mu} = \sum_{j=1}^{p}\mu_j/p$

$n = p_n r_n \to \infty$ means that either $p_n \to \infty$ or $r_n \to \infty$

1. $p_n = p$ is fixed and $r_n \to \infty$
   - The dimensions of correct models are fixed
   - The GIC with $\lambda \to \infty$ and $\lambda/n \to 0$ is selection consistent
   - The GIC with $\lambda = 2$ is inconsistent
2. $p_n \to \infty$ and $r_n = r$ is fixed
   - Only one correct model has a fixed dimension
   - The GIC with $\lambda_n = 2$ is loss consistent
   - The GIC with $\lambda \to \infty$ is inconsistent, because $\lambda p_n/n = \lambda/r \to \infty$
3. $p_n \to \infty$ and $r_n \to \infty$
   - Only one correct model has a fixed dimension
   - The GIC is selection consistent, provided that $\lambda/r_n \to 0$

More on the case where $p_n \to \infty$ and $r_n = r$ is fixed

$\widehat{\sigma}^2 = S(\mathscr{A}_p)/n$, $S(\mathscr{A}) = \|X - Z_{\mathscr{A}}\beta_{\mathscr{A}}\|^2$.

It can be shown that

$$L_n(\mathscr{A}_1) = \Delta_n(\mathscr{A}_1) + \bar{e}^2 \to_p \Delta = \lim_{p\to\infty} \frac{1}{p}\sum_{j=1}^{p}\left(\mu_j - \frac{1}{p}\sum_{i=1}^{p}\mu_i\right)^2$$

$$L_n(\mathscr{A}_p) = \frac{1}{p}\sum_{i=1}^{p}\bar{e}_i^2 \to_p \frac{\sigma^2}{r}$$

where e_{ij}'s are iid, $E(e_{ij}) = 0$, $E(e_{ij}^2) = \sigma^2$, $\bar{e}_i = r^{-1}\sum_{j=1}^{r} e_{ij}$, and $\bar{e} = p^{-1}\sum_{i=1}^{p}\bar{e}_i$.

Then

$$\frac{L_n(\mathscr{A}_1)}{L_n(\mathscr{A}_p)} \to_p \frac{r\Delta}{\sigma^2}$$

The one-mean model is better if and only if $r\Delta < \sigma^2$.

The wrong model may be better!

The GIC with $\lambda_n \to \infty$ minimizes

$$\frac{S(\mathscr{A}_1)}{n} + \frac{\lambda_n}{n}\frac{S(\mathscr{A}_p)}{n-p} \quad \text{and} \quad \frac{S(\mathscr{A}_p)}{n} + \frac{\lambda_n}{r}\frac{S(\mathscr{A}_p)}{n-p}$$

Because

$$\frac{S(\mathscr{A}_1)}{n} = \Delta_n(\mathscr{A}_1) + \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{r}(e_{ij} - \bar{e}^2) \to_p \Delta + \sigma^2$$

$$\frac{S(\mathscr{A}_p)}{n} = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{r}(e_{ij} - \bar{e}_i^2) \to_p \frac{(r-1)\sigma^2}{r}$$

and $\lambda_n/r \to \infty$, $P\{\text{GIC with } \lambda \to \infty \text{ selects } \mathscr{A}_1\} \to 1$

On the other hand, the C_p (GIC with $\lambda_n = 2$) is loss consistent, because the C_p minimizes

$$\frac{S(\mathscr{A}_1)}{n} + \frac{2}{n}\frac{S(\mathscr{A}_p)}{n-p} \quad \text{and} \quad \frac{S(\mathscr{A}_p)}{n} + \frac{2}{r}\frac{S(\mathscr{A}_p)}{n-p}$$

$$\frac{S(\mathscr{A}_1)}{n} + \frac{2}{n}\frac{S(\mathscr{A}_p)}{n-p} \to_p \Delta + \sigma^2,$$

$$\frac{S(\mathscr{A}_p)}{n} + \frac{2}{r}\frac{S(\mathscr{A}_p)}{n-p} \to_p \sigma^2 + \frac{\sigma^2}{r}$$

Asymptotically, the C_p selects $\mathscr{A}_1$ iff $\Delta < \sigma^2/r$, which is the same as the one-mean model is better.

Variable selection by thresholding

Can we do variable selection using p-values?

Or, can we simply select variables by using the values $\widehat{\beta}_j$, $j = 1, ..., p$?

Here $\widehat{\beta}_j$ is the jth component of $\widehat{\beta}$, the least squares estimator of β.

For simplicity, assume that $X|Z \sim N(Z\beta, \sigma^2 I)$.

Then

$$\widehat{\beta}_j - \beta_j = \sum_{i=1}^{n} l_{ij}\varepsilon_i \Big| Z \sim N\left(0, \sigma^2 \sum_{i=1}^{n} l_{ij}^2\right)$$

where ε_i and l_{ij} are the ith components of $\varepsilon = X - Z\beta$ and $(Z^\tau Z)^{-1} z_i$

z_j is the jth row of Z

Because

$$1 - \Phi(t) \le \frac{\sqrt{2\pi}}{t} e^{-t^2/2}, \qquad t > 0$$

where Φ is the standard normal cdf,

$$P\left(|\widehat{\beta}_j - \beta_j| > t\sqrt{\mathrm{var}(\widehat{\beta}|Z)} \Big| Z\right) \le \frac{2\sqrt{2\pi}}{t} e^{-t^2/2}, \quad t > 0$$

Let J_j be the p-vector whose jth component is 1 and other components are 0:

$$l_{ij}^2 = [J_j^\tau (Z^\tau Z)^{-1} z_i]^2 \le J_j^\tau (Z^\tau Z)^{-1} J_j z_i^\tau (Z^\tau Z)^{-1} z_i$$

$$\sum_{i=1}^n l_{ij}^2 \le c_j \sum_{i=1}^n z_i^\tau (Z^\tau Z)^{-1} z_i = pc_j \le p/\eta_n$$

where c_j is the jth diagonal element of $(Z^\tau Z)^{-1}$ and η_n is the smallest eigenvalue of $Z^\tau Z$.

Thus, for any j,

$$P\left(|\widehat{\beta}_j - \beta_j| > t\sigma\sqrt{p/\eta_n} \mid Z\right) \le \frac{2\sqrt{2\pi}}{t} e^{-t^2/2}, \quad t > 0$$

and (letting $t = a_n/(\sigma\sqrt{p/\eta_n})$)

$$P\left(|\widehat{\beta}_j - \beta_j| > a_n \mid Z\right) \le Ce^{-a_n^2\eta_n/(2\sigma^2 p)}$$

for some constant $C > 0$,

$$P\left(\max_{j=1,\ldots,p} |\widehat{\beta}_j - \beta_j| > a_n \mid Z\right) \le pCe^{-a_n^2\eta_n/(2\sigma^2 p)}$$

Suppose that $p/n \to 0$ and $p/(\eta_n \log n) \to 0$ (typically, $\eta_n = O(n)$).
Then, we can choose a_n such that $a_n \to 0$ and $a_n^2(\eta_n \log n/p) \to \infty$ such that

$$P\left(\max_{j=1,\ldots,p}|\widehat{\beta}_j-\beta_j|>ca_n \mid Z\right)=O(n^{-s})$$

for any $c>0$ and some $s\geq 1$; e.g.,

$$a_n=M\left(\frac{p}{\eta_n \log n}\right)^{\alpha}$$

for some constants $M>0$ and $\alpha\in(0,\frac{1}{2})$.

What can we conclude from this?

Let

$$\mathscr{A}=\{j:\beta_j\neq 0\} \quad \text{and} \quad \widehat{\mathscr{A}}=\{j:|\widehat{\beta}_j|>a_n\}$$

That is, $\widehat{\mathscr{A}}$ contains the indices of variables we select by thresholding $|\widehat{\beta}_j|$ at a_n.

Selection consistency:

$$P\left(\widehat{\mathscr{A}}\neq\mathscr{A}\mid Z\right)\leq P\left(|\widehat{\beta}_j|>a_n, j\notin\mathscr{A}\mid Z\right)+P\left(|\widehat{\beta}_j|\leq a_n, j\in\mathscr{A}\mid Z\right)$$

The first term on the right hand side is bounded by

$$P\left(\max_{j=1,\ldots,p}|\widehat{\beta}_j-\beta_j|>a_n \mid Z\right)=O(n^{-s})$$

On the other hand, if we assume that

$$\min_{j\in\mathscr{A}} |\beta_j| \geq c_0 a_n$$

for some $c_0 > 1$, then

$$\begin{aligned} P\left(|\widehat{\beta}_j| \leq a_n, j \in \mathscr{A} \mid Z\right) &\leq P\left(|\beta_j| - |\widehat{\beta}_j - \beta_j| \leq a_n, j \in \mathscr{A} \mid Z\right) \\ &\leq P\left(c_0 a_n - |\widehat{\beta}_j - \beta_j| \leq a_n, j \in \mathscr{A} \mid Z\right) \\ &\leq P\left(\max_{j=1,\dots,p} |\widehat{\beta}_j - \beta_j| \geq (c_0 - 1) a_n \mid Z\right) \\ &= O(n^{-s}) \end{aligned}$$

Hence, we have consistency; in fact, the convergence rate is $O(n^{-s})$.
We can also obtain similar results by thresholding $|\widehat{\beta}_j| / \sqrt{\sum_{i=1}^n l_{ij}^2}$.
This approach may not work if $p/n \not\to 0$.
If $p > n$, then $Z^\tau Z$ is not of full rank.
There exist several other approaches for the case where $p > n$; e.g., we replace $(Z^\tau Z)^{-1}$ by some matrix, or use ridge regression instead of LSE.