L09: A biased estimator: Ridge estimator

1. The problem of multicollinearity

   (1) Multicollinearity
   Under the assumption β is estimable, the BLUE for β is

   $$\widehat{\beta} = (X'X)^{-1}X'y \sim N(\beta, \sigma^2(X'X)^{-1}).$$

   But β is estimable $\iff X$ has LI columns $\iff X'X$ is non-singular $\iff |X'X| > 0$.
   We say that there is multicollinearity in X if $|X'X|$ is close to 0. This characterizes the situation where the assumption is true, but almost false.

   (2) Consequence of multicollinearity
   By EVD $X'X = P\Lambda P'$, $|X'X| = |\Lambda| = \lambda_1 \cdots \lambda_p$. So with multicollinearity at least one $\lambda_i > 0$ is close to 0. Consequently, the total variance in $\widehat{\beta}$

   $$\text{tr}[\text{Cov}(\widehat{\beta})] = \text{tr}[\sigma^2(X'X)^{-1}] = \sigma^2\text{tr}\left(P\Lambda^{-1}P'\right) = \frac{\sigma^2}{\lambda_1} + \cdots + \frac{\sigma^2}{\lambda_p}$$

   is very large. Hence the estimator is not stable.

   (3) Common problem for all LUEs
   Let $\widetilde{\beta}$ be a LUE for β. Then $\text{Cov}(\widehat{\beta}) \leq \text{Cov}(\widetilde{\beta})$. So

   $$\text{var}(\widehat{\beta}_i) = e_i'\text{Cov}(\widehat{\beta})e_i \leq e_i'\text{Cov}(\widetilde{\beta})e_i = \text{var}(\widetilde{\beta}_i).$$

   Thus large total variance is a common problem for all LUEs for β.

   (4) A Ridge estimator
   A naive remedy is to increase the value of λ_i to $\lambda_i + k_i$, $k_i > 0$. This changes Λ to $\Lambda + K$. The resulted estimator is called a ridge estimator since it lifted the ridge of matrix Λ. This new estimator is denoted as $\widehat{\beta}(K)$.

2. Initial analysis

   (1) $\widehat{\beta}(K)$ is a linear estimator for β

   With $X'X = P\Lambda P'$, the BLUE for β is $\widehat{\beta} = (X'X)^{-1}X'y = (P\Lambda P')^{-1}X'y$.
   So $\widehat{\beta}(K) = [P(\Lambda + K)P']^{-1}X'y = P(\Lambda + K)^{-1}P'X'y$.
   Thus $\widehat{\beta}(K)$ is still a linear estimator.

   (2) $\widehat{\beta}(K)$ is a biased estimator

   $$\begin{aligned}\widehat{\beta}(K) &= P(\Lambda + K)^{-1}P'X'y = P(\Lambda + K)^{-1}P'(X'X)(X'X)^{-1}X'y \\ &= P(\Lambda + K)^{-1}P'P\Lambda P'\widehat{\beta} = P(\Lambda + K)^{-1}\Lambda P'\widehat{\beta}.\end{aligned}$$

   But $(\Lambda + K)^{-1} = \Lambda^{-1} - \Lambda^{-1}(\Lambda^{-1} + K^{-1})^{-1}\Lambda^{-1}$ since

   $$(\Lambda + K)[\Lambda^{-1} - \Lambda^{-1}(\Lambda^{-1} + K^{-1})^{-1}\Lambda^{-1}] = I.$$

   So $\widehat{\beta}(K) = P[\Lambda^{-1} - \Lambda^{-1}(\Lambda^{-1} + K^{-1})^{-1}\Lambda^{-1}]\Lambda P'\widehat{\beta} = \widehat{\beta} - P\Lambda^{-1}(\Lambda^{-1} + K^{-1})^{-1}P'\widehat{\beta}$.
   Therefore $E[\widehat{\beta}(K)] = \beta - P\Lambda^{-1}(\Lambda^{-1} + K^{-1})^{-1}P'\beta$.
   Thus $\widehat{\beta}(K)$ is a baised estimator with bias $E[\widehat{\beta}(K)] - \beta = -P\Lambda^{-1}(\Lambda^{-1} + K^{-1})^{-1}P'\beta$.

(3) Total variance reduced: $\text{tr}[\text{Cov}(\widehat{\beta}(K))] \leq \text{tr}[\text{Cov}(\widehat{\beta})]$

$$\begin{aligned}
\text{tr}[\text{Cov}(\widehat{\beta}(K))] &= \text{tr}\left[\text{Cov}\left(P(\Lambda+K)^{-1}\Lambda P'\widehat{\beta}\right)\right] \\
&= \text{tr}\left[P(\Lambda+K)^{-1}\Lambda P'(\sigma^2 P\Lambda^{-1}P')P\Lambda(\Lambda+K)^{-1}P'\right] \\
&= \sigma^2\text{tr}\left[(\Lambda+K)^{-1}\Lambda(\Lambda+K)^{-1}\right] = \sum_{i=1}^{p} \frac{\lambda_i\sigma^2}{(\lambda_i+k_i)^2}
\end{aligned}$$

(4) Norm reduced: $\|\widehat{\beta}(K)\| \leq \|\widehat{\beta}\|$

Note that P is an orthogonal matrix that preserves norms. So

$$\begin{aligned}
\|\widehat{\beta}(K)\|^2 &= \|P(\Lambda+K)^{-1}\Lambda P'\widehat{\beta}\|^2 = \|(\Lambda+K)^{-1}\Lambda P'\widehat{\beta}\|^2 \\
&= \|\text{diag}\left(\frac{\lambda_1}{\lambda_1+k_1},..,\frac{\lambda_p}{\lambda_p+k_p}\right)P'\widehat{\beta}\|^2 \leq \|P'\widehat{\beta}\|^2 = \|\widehat{\beta}\|^2
\end{aligned}$$

3. Performance of $\widehat{\beta}(K)$.

(1) Risk $\text{MSE}(\widehat{\gamma})$

For $\widehat{\gamma}$, an estimator for γ, $E[(\widehat{\gamma}-\gamma)(\widehat{\gamma}-\gamma)']$ is a matrix-valued risk. If $\widehat{\gamma}$ is an UE, then this risk is $\text{Cov}(\widehat{\gamma})$. BLUE of β is derived with this risk.
$E[(\widehat{\gamma}-\gamma)'(\widehat{\gamma}-\gamma)] = E\|\widehat{\gamma}-\gamma\|^2$ is a positive valued risk, called the mean squared error denoted as $\text{MSE}(\widehat{\gamma})$. The estimator domination by the matrix-valued risk implies the same domination by the MSE.

$$\begin{aligned}
\text{MSE}(\widehat{\gamma}) &= E[(\widehat{\gamma}-\gamma)'I_p(\widehat{\gamma}-\gamma)] = [E(\widehat{\gamma})-\gamma]'I_p[E(\widehat{\gamma})-\gamma] + \text{tr}[I_p\,\text{Cov}(\widehat{\gamma})] \\
&= \|E(\widehat{\gamma})-\gamma\|^2 + \text{tr}[\text{Cov}(\widehat{\gamma})]
\end{aligned}$$

(2) $\text{MSE}[\widehat{\beta}(K)]$

$$\begin{aligned}
\|E(\widehat{\beta}(K))-\beta\|^2 &= \|-P\Lambda^{-1}(\Lambda^{-1}+K^{-1})^{-1}P'\beta\|^2 = \|\Lambda^{-1}(\Lambda^{-1}+K^{-1})^{-1}P'\beta\|^2 \\
&= \sum_{i=1}^{p}\left(\frac{(1/\lambda_i)}{(1/\lambda_i)+(1/k_i)}\right)^2(P'\beta)_i^2 = \sum_{i=1}^{p}\frac{k_i^2}{(\lambda_i+k_i)^2}(P'\beta)_i^2
\end{aligned}$$

So $\text{MSE}[\widehat{\beta}(K)] = \sum_{i=1}^{p}\frac{k_i^2(P'\beta)_i^2+\lambda_i\sigma^2}{(\lambda_i+k_i)^2}$

(3) Minimizing $\text{MSE}[\widehat{\beta}(K)]$

Let $f(k_i) = \frac{k_i^2(Q'\beta)_i^2+\lambda_i\sigma^2}{(\lambda_i+k_i)^2}$. Then

$$f'(k_i) = \frac{(\lambda_i+k_i)^2 2k_i(Q'\beta)_i^2 - 2(\lambda_i+k_i)[k_i^2(Q'\beta)_i^2+\lambda_i\sigma^2]}{(\lambda_i+k_i)^4} = \cdots = \frac{2\lambda_i(Q'\beta)_i^2}{(\lambda_i+k_i)^3}\left[k_i - \frac{\sigma^2}{(Q'\beta)_i^2}\right].$$

Thus $f(k_i)$ is minimized at $k_i = \frac{\sigma^2}{(Q'\beta)_i^2}$, so is $\text{MSE}[\widehat{\beta}(K)]$.

(4) Better performance

$$\text{MSE}[\widehat{\beta}(K)]_{k_i=\frac{\sigma^2}{(Q'\beta)_i^2}} = \sum_{i=1}^{p}\frac{\frac{\sigma^4}{(Q'\beta)_i^2}+\lambda_i\sigma^2}{\left[\lambda_i+\frac{\sigma^2}{(Q'\beta)_i^2}\right]^2} = \sum_{i=1}^{p}\frac{\sigma^2}{\lambda_i+\frac{\sigma^2}{(Q'\beta)_i^2}} \leq \sum_{i=1}^{p}\frac{\sigma^2}{\lambda_i} = \text{MSE}(\widehat{\beta}).$$

L10 A biased estimator: Principal component estimator

1. Principal component estimator

   (1) An idea
   With EVD $X'X = P\Lambda P'$ where $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_p)$ with $\lambda_1 \geq \cdots \geq \lambda_p > 0$. The total variance in the BLUE for β is

   $$\text{tr}[\text{Cov}(\widehat{\beta})] = \text{tr}[\sigma^2 (X'X)^{-1}] = \left(\frac{\sigma^2}{\lambda_1} + \cdots + \frac{\sigma^2}{\lambda_q}\right) + \left(\frac{\sigma^2}{\lambda_{q+1}} + \cdots + \frac{\sigma^2}{\lambda_p}\right).$$

   In reducing this total variance we wonder if we can have an estimator with reduced total variance $\frac{\sigma^2}{\lambda_1} + \cdots + \frac{\sigma^2}{\lambda_q}$.

   (2) A try
   In $(X'X)^{-1} = \left[(P_I, P_{II}) \begin{pmatrix} \Lambda_I & 0 \\ 0 & \Lambda_{II} \end{pmatrix} (P_I, P_{II})'\right]^{-1} = P_I \Lambda_I^{-1} P_I' + P_{II} \Lambda_{II}^{-1} P_{II}'$,
   $\frac{1}{\lambda_i}$ with $i, = 1, ..., q$ only appear in Λ_I^{-1}. Replacing $(X'X)^{-1}$ in $\widehat{\beta} = (X'X)^{-1} X'y$ by $P_I \Lambda_I^{-1} P_I'$ ends up with $\widehat{\beta}(q) = P_I \Lambda_I^{-1} P_I' X'y$.

   (3) Principal component estimator
   $$\begin{aligned} \text{Cov}(\widehat{\beta}(q)) &= \text{Cov}(P_I \Lambda_I^{-1} P_I' X'y) = \sigma^2 (P_I \Lambda_I^{-1} P_I')(X'X)(P_I \Lambda_I^{-1} P_I') \\ &= \sigma^2 (P_I \Lambda_I^{-1} P_I')(P \Lambda P')(P_I \Lambda_I^{-1} P_I') = \sigma^2 (P_I \Lambda_I^{-1} P_I'). \end{aligned}$$
   Thus $\text{tr}[\text{Cov}(\widehat{\beta}(q))] = \sigma^2 \text{tr}(\Lambda_I^{-1}) = \frac{\sigma^2}{\lambda_1} + \cdots + \frac{\sigma^2}{\lambda_q}$.
   We call this $\widehat{\beta}(q)$ a principal component estimator.

2. Initial analysis

   (1) Linear estimator with reduced total variance
   $\widehat{\beta}(q) = P_I \Lambda_I^{-1} P_I' X'y$ is a linear estimator for β with reduced total variance $\text{tr}[\text{Cov}(\widehat{\beta}(q))] = \frac{\sigma^2}{\lambda_1} + \cdots + \frac{\sigma^2}{\lambda_q}$ from $\text{tr}[\text{Cov}(\widehat{\beta})] = \left(\frac{\sigma^2}{\lambda_1} + \cdots + \frac{\sigma^2}{\lambda_q}\right) + \left(\frac{\sigma^2}{\lambda_{q+1}} + \cdots + \frac{\sigma^2}{\lambda_p}\right)$, the total variance of the BLUE $\widehat{\beta}$.

   (2) Biased estimator
   $$\begin{aligned} \widehat{\beta}(q) &= P_I \Lambda_I^{-} P_I' (X'X)(X'X)^{-1} X'y = P_I \Lambda_I^{-1} P_I' P \Lambda P' \widehat{\beta} \\ &= P_I \Lambda_I^{-1} (I, 0) \begin{pmatrix} \Lambda_I P_I' \\ \Lambda_{II} P_{II}' \end{pmatrix} \widehat{\beta} = P_I P_I' \widehat{\beta}. \end{aligned}$$
   So $\widehat{\beta} = I_p \widehat{\beta} = (P_I P_I' + P_{II} P_{II}') \widehat{\beta} = \widehat{\beta}(q) + P_{II} P_{II}' \widehat{\beta}$. Thus $\beta = E[\widehat{\beta}(q)] + P_{II} P_{II}' \beta$.
   Hence $\widehat{\beta}(q)$ is a biased estimator with bias $E(\widehat{\beta}(q)) - \beta = -P_{II} P_{II}' \beta$.

   (3) Norm reduced
   In $\widehat{\beta} = \widehat{\beta}(q) + P_{II} P_{II}' \widehat{\beta}$, $\left\langle \widehat{\beta}(q), P_{II} P_{II}' \widehat{\beta} \right\rangle = \widehat{\beta}' P_{II} P_{II}' P_I P_I' \widehat{\beta} = 0$.
   So by Pythagorean theorem,

   $$\|\widehat{\beta}\|^2 = \|\widehat{\beta}(q) + P_{II} P_{II}' \widehat{\beta}\|^2 = \|\widehat{\beta}(q)\|^2 + \|P_{II} P_{II}' \widehat{\beta}\|^2 \geq \|\widehat{\beta}(q)\|^2.$$

3. Evaluate the performance

(1) MSEM

For $\widehat{\gamma}$, an estimator for γ, with $E(\widehat{\gamma}) = \mu_{\widehat{\gamma}}$ the matrix-valued risk

$$\begin{aligned} E[(\widehat{\gamma}-\gamma)(\widehat{\gamma}-\gamma)'] &= E[(\widehat{\gamma}-\mu_{\widehat{\gamma}}+\mu_{\widehat{\gamma}}-\gamma)(\widehat{\gamma}-\mu_{\widehat{\gamma}}+\mu_{\widehat{\gamma}}-\gamma)'] \\ &= \text{Cov}(\widehat{\gamma}) + [(\mu_{\widehat{\gamma}}-\gamma)(\mu_{\widehat{\gamma}}-\gamma)']. \end{aligned}$$

is denoted as MSEM($\widehat{\gamma}$). We do comparison of BLUE $\widehat{\beta}$ and principal compnent estimator $\widehat{\beta}(q)$ by MSEM.

(2) MSEM($\widehat{\beta}$ and MSEM($\widehat{\beta}(q)$)

$$\text{MSEM}(\widehat{\beta}) = \text{Cov}(\widehat{\beta}) + 0 = \sigma^2(P_I\Lambda_I^{-1})P_I') + \sigma^2(P_{II}\Lambda_{II}^{-1}P_{II}').$$

$$\begin{aligned} \text{MSEM}[\widehat{\beta}(q)] &= \text{Cov}(\widehat{\beta}(q)) + [E(\widehat{\beta}(q)) - \beta][E(\widehat{\beta}(q)) - \beta]' \\ &= \sigma^2(P_I\Lambda_I^{-1}P_I') + [P_{II}P_{II}'\beta][P_{II}P_{II}'\beta]'. \end{aligned}$$

$$\begin{aligned} \text{So MSEM}[\widehat{\beta}(q)] \le \text{MSEM}(\widehat{\beta}) \quad &\Longleftrightarrow \quad P_{II}P_{II}'\beta\beta'P_{II}P_{II}' \le \sigma^2 P_{II}\Lambda_{II}^{-1}P_{II}' \\ &\overset{*}{\Longleftrightarrow} \quad P_{II}'\beta\beta'P_{II} \le \sigma^2\Lambda_{II}^{-1} \end{aligned}$$

since $A \le B \Longrightarrow CAC' \le CBC'$.

(3) Theorem

If $0 < \lambda_{q+1} \le \frac{\sigma^2}{\|P_{II}'\beta\|^2}$, then MSEM$[\widehat{\beta}(q)] \le$ MSEM$(\widehat{\beta})$

Proof If $0 < \lambda_{q+1} \le \frac{\sigma^2}{\|P_{II}'\beta\|^2}$, then $0 < \|P_{II}'\beta\|^2 \le \frac{\sigma^2}{\lambda_{q+1}}$.

Note that $P_{II}'\beta\beta'P_{II}$ is a symmetric matrix with rank 1, and from

$$(P_{II}'\beta\beta'P_{II})(P_{II}'\beta) = (P_{II}'\beta)\|P_{II}'\beta\|^2$$

we see that $\|P_{II}'\beta\|^2$ is an eigenvalue for $P_{II}'\beta\beta'P_{II}$. By EVD,

$$\begin{aligned} P_{II}'\beta\beta'P_{II} &= Q\,\text{diag}(\|P_{II}'\beta\|^2, 0, ..., 0)\,Q' \le Q\,\text{diag}\left(\frac{\sigma^2}{\lambda_{q+1}}, .., \frac{\sigma^2}{\lambda_{q+1}}\right)Q' \\ &= \sigma^2\text{diag}\left(\frac{1}{\lambda_{q+1}}, ..., \frac{1}{\lambda_{q+1}}\right) \le \sigma^2\text{diag}\left(\frac{1}{\lambda_{q+1}}, ..., \frac{1}{\lambda_p}\right) = \sigma^2\Lambda_{II}^{-1}. \end{aligned}$$

By (2), MSEM$[\widehat{\beta}(q)] \le$ MSEM$(\widehat{\beta})$.

Comment: The cut-off value $\frac{\sigma^2}{\|P_{II}'\beta\|^2}$ depends on unknown parameters β and σ^2.