去除模型部分自变量的影响

Remark： 任意维度相同的方阵 M_1, M_2。写法 $M_1 > (\geq) M_2$ 表示 $M_1 - M_2$（半）正定

一、全模型 & 选模型

Full & Reduced Model

1.1 Full & Reduced Model

对于原始线性模型：

$$E(y) = X\beta \qquad cov(y) = \sigma^2 I$$

满足 Gauss-Markov Assumption & $rank(X) = p$，$X \in \mathbb{R}^{n\times p}$

对模型分划：$s+q=p$

$$X = (S, Q), \quad S \in \mathbb{R}^{n\times s} \quad Q \in \mathbb{R}^{n\times q}$$

$$\beta = (\beta_S^T, \beta_Q^T)^T, \quad \beta_S \in \mathbb{R}^s \quad \beta_Q \in \mathbb{R}^q$$

则有原模型的 OLSE 为

$$\hat{\beta} = (\hat{\beta}_S^T, \hat{\beta}_Q^T)^T = (X^TX)^{-1}X^Ty$$

而针对 β_S 的 OLSE 为

$$\tilde{\beta}_S = (S^TS)^{-1}S^Ty$$

$\hat{\beta}$ 为 β 的全模型的 OLSE

$\tilde{\beta}_S$ 为 β_S 的选模型的 OLSE

$$\tilde{\beta}_S \longrightarrow \begin{bmatrix} \beta_S \\ \beta_Q \end{bmatrix} = \beta \longleftarrow \hat{\beta} = \begin{bmatrix} \hat{\beta}_S \\ \hat{\beta}_Q \end{bmatrix}$$

1.2. Full & Reduced：$\hat{\beta}_s$ 与 $\tilde{\beta}_s$ 的关系

由分块矩阵求逆公式：

$$(X^TX)^{-1} = \begin{bmatrix} S^TS & S^TQ \\ Q^TS & Q^TQ \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} (S^TS)^{-1} + ADA^T & -AD \\ -(AD)^T & D \end{bmatrix}$$

其中 $A = (S^TS)^{-1}S^TQ$

$D = \{Q^TQ - Q^TS(S^TS)^{-1}S^TQ\}^{-1}$

$$\hat{\beta} = \begin{bmatrix} (S^TS)^{-1}+ADA^T & -AD \\ -(AD)^T & D \end{bmatrix}\begin{bmatrix} S^Ty \\ Q^Ty \end{bmatrix}$$

又因为 $X^Ty = \begin{bmatrix} S^Ty \\ Q^Ty \end{bmatrix}$，于是可求 $\hat{\beta}$：

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_s \\ \hat{\beta}_Q \end{bmatrix} = \begin{bmatrix} (S^TS)^{-1}S^Ty + AD(SA-Q)^Ty \\ -(SAD)^Ty + DQ^Ty \end{bmatrix}$$

综上可知：$\hat{\beta}_s$ 与 $\tilde{\beta}_s$ 的关系

$$\hat{\beta}_s = \tilde{\beta}_s + AD(SA-Q)^Ty$$

故一般：$\hat{\beta}_s \neq \tilde{\beta}_s$

其中 $\begin{cases} A = (S^TS)^{-1}S^TQ \\ D = \{Q^TQ - Q^TS(S^TS)^{-1}S^TQ\}^{-1} \end{cases}$

1.3 $\hat{\beta}_s$ 与 $\tilde{\beta}_s$ 何时相等

由 $\hat{\beta}_s = \tilde{\beta}_s + AD(SA-Q)^T y$ 可知，$\hat{\beta}_s = \tilde{\beta}_s$ 当：

(a). $AD = 0$，此时有 $\hat{\beta}_s = \tilde{\beta}_s$

Remark：由 $(X^TX)^{-1}$ 的形式，当 $AD=0$ 时，X^TX 为对角阵

$\Rightarrow$ $S^TQ = 0$，即 S 与 Q 正交，此时

$$\hat{\beta} = (X^TX)^{-1}X^Ty = \mathrm{diag}\{(S^TS)^{-1}, (Q^TQ)^{-1}\} \cdot X^Ty$$

$$= \begin{bmatrix} (S^TS)^{-1}S^Ty \\ (Q^TQ)^{-1}Q^Ty \end{bmatrix}$$

说明 $\hat{\beta} = (\hat{\beta}_s^T, \hat{\beta}_Q^T)^T$ 中两部分互不相关

（或：假设 X 每列已中心化，S^TQ 表示两部分自变量的协方差）

(b) $SA - Q = 0$，此时有 $\hat{\beta}_s = \tilde{\beta}_s$

Remark：由 $A = (S^TS)^{-1}S^TQ$ 知，$SA - Q = 0$

$\Rightarrow$ $Q - SA = Q - S(S^TS)^{-1}S^TQ = \{I - S(S^TS)^{-1}S^T\} \cdot Q = 0$

由 Null-Rank Theorem（秩零定理）：

记 $\mathcal{V} = \{ v \in \mathbb{R}^n : \{I - S(S^TS)^{-1}S^T\} \cdot v = 0 \}$ 有：

$$\dim(\mathcal{V}) + \mathrm{rank}(\{I - S(S^TS)^{-1}S^T\}) = n$$

但 $\{I-S(S^TS)^{-1}S^T\}$ 类 $I-H$，为幂等阵 $\Rightarrow$ 它的特征值为 0 或 1

$\therefore$ $\mathrm{rank}(\{I-S(S^TS)^{-1}S^T\}) = \sum \lambda_i = \mathrm{tr}(\{I-S(S^TS)^{-1}S^T\})$

$$= n - \mathrm{tr}(S(S^TS)^{-1}S^T) = n - \mathrm{tr}(SS^T(S^TS)^{-1}) = n - \mathrm{tr}(S^TS(S^TS)^{-1})$$

$= n - tr(I_s) = n - s \quad (S \in \mathbb{R}^s)$

$\Rightarrow \dim(V) = s$

由 $Q \in \mathbb{R}^q$，若 $rank(Q) = q > s = \dim(V)$，则 $Q - SA = 0$ 必矛盾！

二、模型选择对估计的影响

Effect of model selection on estimation

考虑 $\hat{\beta}_s$ 与 $\tilde{\beta}_s$ 的期望，协方差矩阵，均方误差矩阵（mean squared error matrix, MSEM），以及对应的 σ^2 的估计量，比较二者。

2.1 E & Cov

定理 1 β_s 选模型的估计 $\tilde{\beta}_s = (S^TS)^{-1}S^Ty$ 满足：

(a) $E(\tilde{\beta}_s) = \beta_s + A\beta_Q$

(b) $cov(\tilde{\beta}_s) = \sigma^2 (S^TS)^{-1}$

(c) $cov(\hat{\beta}_s) \geq cov(\tilde{\beta}_s)$

其中 $A = (S^TS)^{-1}S^TQ \quad \hat{\beta} = (X^TX)^{-1}X^Ty = (\hat{\beta}_s^T, \hat{\beta}_Q)^T \quad \beta = (\beta_s^T, \beta_Q^T)^T$

证明见习作

(1) 定理 1 (a)： $\tilde{\beta}_s$ 一般是 β_s 的有偏估计

当 $\beta_Q = 0$ 或 $A = 0$ 时，才有无偏性 $E(\tilde{\beta}_s) = \beta_s$

(i) $\beta_Q = 0$ 则 $y = X\beta + \varepsilon = S\beta + \varepsilon$，显然，删去Q合理

(ii) $A = (S^TS)^{-1}S^TQ = 0 \Rightarrow S^TQ = 0$ 说明 $\hat{\beta} = (\hat{\beta}_S^T, \hat{\beta}_Q^T)^T$ 中两部分互不相关，可分别回归

(2) 定理1(c)： $\tilde{\beta}_S$ 相较于 $\hat{\beta}_S$ 的波动更小

Remark： It's also a trade-off between biasedness and variance

2.2 MSEM

定义 MSEM 对于 b 的任意估计量 $\hat{b}$，定义其 MSEM 为

$$MSEM(\hat{b}) = E\{(\hat{b}-b)(\hat{b}-b)^T\}$$
$$= cov(\hat{b}) + \{E(\hat{b})-b\}\{E(\hat{b})-b\}^T$$

相比于 $MSE(\hat{b}) = E\{(\hat{b}-b)^T(\hat{b}-b)\}$ 为标量，$MSEM(\hat{b})$ 为矩阵

Proof： $E\{(\hat{b}-b)(\hat{b}-b)^T\} = E\{(\hat{b}-E(\hat{b})+E(\hat{b})-b)(\hat{b}-E(\hat{b})+E(\hat{b})-b)^T\}$

$= E\{(\hat{b}-E(\hat{b}))(\hat{b}-E(\hat{b}))^T + (E(\hat{b})-b)(E(\hat{b})-b)^T\} + 0 + 0$

$= cov(\hat{b}) + \{E(\hat{b})-b\}\{E(\hat{b})-b\}^T$ □

推论： MSEM 比 MSE 结论更强。已知 $\hat{a}, \hat{b}$ 是对 a, b 的估计。若有 $MSEM(\hat{a}) - MSEM(\hat{b}) \geq 0$ (PSD)

$\Rightarrow$ 则有 $MSE(\hat{a}) - MSE(\hat{b}) \geq 0$。反之不成立

Proof：引理：

$$MSE(\hat{a}) - MSE(\hat{b}) = tr\{MSEM(\hat{a}) - MSEM(\hat{b})\}$$

$$MSE(\hat{a}) - MSE(\hat{b}) = E\{(\hat{a}-a)^T(\hat{a}-a)\} - E\{(\hat{b}-b)^T(\hat{b}-b)\}$$

$$= E\,tr\{(\hat{a}-a)^T(\hat{a}-a)\} - E\,tr\{(\hat{b}-b)^T(\hat{b}-b)\}$$

$$= E\,tr\{(\hat{a}-a)(\hat{a}-a)^T\} - E\,tr\{(\hat{b}-b)(\hat{b}-b)^T\}$$

$$= tr\,E\{(\hat{a}-a)(\hat{a}-a)^T\} - tr\,E\{(\hat{b}-b)(\hat{b}-b)^T\}$$

$$= tr\{MSEM(\hat{a}) - MSEM(\hat{b})\}$$

故，$MSEM(\hat{a}) - MSEM(\hat{b}) \geq 0$ (PSD) 时，必有 $\lambda_i \geq 0$

于是 $tr\{MSEM(\hat{a}) - MSEM(\hat{b})\} = \sum \lambda_i \geq 0$ □

2.3 一定条件下，选模型优化了 MSEM

定理 2

若 $cov(\hat{\beta}_Q) \geq \beta_Q \beta_Q^T$，则 $MSEM(\hat{\beta}_s) \geq MSEM(\tilde{\beta}_s)$

Proof：由定理 1 (a)(b) 知

$$E(\tilde{\beta}_s) = \beta_s + A\beta_Q$$

$$cov(\tilde{\beta}_s) = \sigma^2 (S^TS)^{-1}$$

故 $MSEM(\tilde{\beta}_s) = cov(\tilde{\beta}_s) + \{E(\tilde{\beta}_s) - \beta_s\}\{E(\tilde{\beta}_s) - \beta_s\}^T$

$$= \underline{\sigma^2 (S^TS)^{-1} + A\beta_Q\beta_Q^T A^T}$$

由 $cov(\hat{\beta}) = cov\left(\begin{bmatrix}\hat{\beta}_s \\ \hat{\beta}_Q\end{bmatrix}\right) = \sigma^2 (X^TX)^{-1}$

$$= \sigma^2 \cdot \begin{bmatrix} (S^TS)^{-1} + ADA^T & -AD \\ -(AD)^T & D \end{bmatrix}$$

$$\therefore \quad \mathrm{cov}(\hat{\beta}_S) = \sigma^2 \cdot \{ (S^TS)^{-1} + ADA^T \}$$

$$\mathrm{cov}(\hat{\beta}_Q) = \sigma^2 \cdot D$$

从而 $\mathrm{MSEM}(\hat{\beta}_S) = \mathrm{cov}(\hat{\beta}_S) + 0 = \underline{\sigma^2 \cdot \{ (S^TS)^{-1} + ADA^T \}}$

$$\Rightarrow \mathrm{MSEM}(\hat{\beta}_S) - \mathrm{MSEM}(\tilde{\beta}_S) = \sigma^2 ADA^T - A\beta_Q\beta_Q^T A^T$$

$$= A\,\mathrm{cov}(\hat{\beta}_Q) A^T - A\beta_Q\beta_Q^T A^T$$

$$= A\{ \mathrm{cov}(\hat{\beta}_Q) - \beta_Q\beta_Q^T \} A^T \geq 0$$

当 $\mathrm{cov}(\hat{\beta}_Q) - \beta_Q\beta_Q^T \geq 0$ 时 $\square$

Remark: 定理2 说明 Q 对 y 无影响时，($\beta_Q = 0$)，$\tilde{\beta}_S$ 在 MSEM 标准下优于 $\hat{\beta}_S$

定理2 说明即使 Q 对 y 有影响，当 $\hat{\beta}_Q$ 波动很大 ($\mathrm{cov}(\hat{\beta}_Q) \geq \beta_Q\beta_Q^T$)，$\tilde{\beta}_S$ 在 MSEM 标准下优于 $\hat{\beta}_S$

2.3 对 σ^2 的估计

基于 Reduced Model 对 σ^2 估计：

$$\tilde{\sigma}^2 = \frac{1}{n-s} \| y - S\tilde{\beta}_S \|^2$$

$$= \frac{1}{n-s} \cdot y^T \{ I - S(S^TS)^{-1}S^T \} y$$

下面定理说明 $\tilde{\sigma}^2$ 是 σ 的有偏估计：

定理 3 选模型（Reduced Model）的 σ^2 估计量 $\tilde{\sigma}^2$ 满足：

$$E(\tilde{\sigma}^2) = \sigma^2 + \frac{1}{n-s}\beta_Q^T D^{-1} \beta_Q$$

证明见习作． 回顾：$D = \{Q^TQ - Q^TS(S^TS)^{-1}S^TQ\}^{-1}$

$S(S^TS)^{-1}S^T = H$

$\tilde{\sigma}^2 = y^T(I-H)y \,/\, (n-s)$

$$\Rightarrow E(y^T(I-H)y) = E(y)^T\cdot(I-H)\cdot E(y) + \mathrm{tr}\{\overset{I-H}{\mathrm{cov}}(\cancel{[I-H]}y)\}$$

$$= \beta^TX^T\cdot(I-H)X\beta + \underbrace{\mathrm{tr}\{(I-H)\cdot\sigma^2\cdot I\cdot\cancel{(I-H)}\}}_{(n-s)\cdot\sigma^2}$$

$\beta = \begin{bmatrix}\beta_S\\ \beta_Q\end{bmatrix}$ $X=(S,Q)$

$\beta^TX^T(I-H)X\beta$

$\parallel$

$(\beta_S^TS^T + \beta_Q^TQ^T)(I-H)(S\beta_S + Q\beta_Q)$

而 $\beta_S^TS^T(I-H)S\beta_S = 0$

$\therefore$ 原式 $= \beta_Q^TQ^T(I-H)Q\beta_Q$

$\parallel$

$D^{-1} = Q^TQ - Q^TS(S^TS)^{-1}S^TQ$

$= Q^TQ - Q^THQ$

三、模型选择对预测的影响

Effect of Model Selection on Prediction

假设新观测值 (x_0, y_0) 满足：

$$y_0 \perp \{y_i : i=1,2,\cdots,n\}$$

$$E(y_0) = x_0^T\beta$$

$$var(y_0) = \sigma^2$$

基于全模型和选模型，分别得到对 y_0 的 2 个预测：

$$x_0^T\hat{\beta}$$

$$x_s^T\tilde{\beta}_s$$

其中 $x_0 = (x_s^T, x_Q^T)^T$

记 2 个模型的预测误差为 u 和 $\tilde{u}$：

$$u = y_0 - x_0^T\hat{\beta}$$

$$\tilde{u} = y_0 - x_s^T\tilde{\beta}_s$$

下面定理比较 u，$\tilde{u}$ 的质量：

定理 4 预测误差 u 与 $\tilde{u}$ 满足：

(a). $E(\tilde{u}) = x_Q^T \beta_Q - x_s^T A \beta_Q$ 而 $E(u) = 0$

(b). $var(u) \geq var(\tilde{u})$

(c). 若 $cov(\hat{\beta}_Q) \geq \beta_Q \beta_Q^T$，则 $E(u^2) \geq E(\tilde{u}^2)$

证明见习作

(1) 定理 4 (a)：$x_s^T \tilde{\beta}_s$ 一般是 $E(\beta_0)$ 的有偏估计。

若 $\beta_Q = 0$ 或 $A = 0$，同上讨论 $\Rightarrow$ 此时 $E(x_s^T \tilde{\beta}_s) = E(y_0)$

(2) 定理 4 (b)：相比于 $x_0^T \hat{\beta}$，选模型 $x_s^T \tilde{\beta}_s$ 的误差波动更小

(3) 定理 4 (c)：综合考量