第30讲 基于子模型的预测

2020.6.5

Essentially, all models are wrong, but some are useful

- Box

线性模型中的预测问题

训练数据：$(\mathbf{x}_i, y_i), i = 1,...,n, \mathbf{x}_i$为$p\times 1$自变量，$y_i$为响应变量。

模型：$y_i = \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta} + \varepsilon_i,\ \ \varepsilon_i \sim (0,\sigma^2)$

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \ \boldsymbol{\varepsilon} \sim (0,\sigma^2 I_n),$$

对$\boldsymbol{\beta}$的任一估计$\tilde{\boldsymbol{\beta}}, \mathbf{y}$的拟合值向量$\tilde{\mathbf{y}} = X\tilde{\boldsymbol{\beta}}.$

待预测数据：$(\mathbf{x}_0, y_0)$，$\mathbf{x}_0$所对应的y_0待预测。y_0与$y_1,..., y_n$独立。

假设同样的模型：$y_0 = \mathbf{x}_0^{\mathsf{T}}\boldsymbol{\beta} + \varepsilon_0, \varepsilon_0 \sim (0,\sigma^2)$

若$\mathbf{x}_0$等于某个$\mathbf{x}_i, i \in \{1,2,...,n\}$, 称为in-sample问题；
若$\mathbf{x}_0$不等于任何$\mathbf{x}_i$，称为out-of-sample问题。

预测统计量：设$\tilde{\boldsymbol{\beta}}$为$\boldsymbol{\beta}$的一个估计，以$\tilde{y}_0 = \mathbf{x}_0^{\mathsf{T}}\tilde{\boldsymbol{\beta}}$预测$y_0$.

命题1. 假设$\tilde{\boldsymbol{\beta}}$为$\boldsymbol{\beta}$的任一估计，记其MSE矩阵$M(\tilde{\boldsymbol{\beta}})=E(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})^{\mathsf{T}}$。
我们以$\tilde{y}_0=\mathbf{x}_0^{\mathsf{T}}\tilde{\boldsymbol{\beta}}$ 预测$y_0=\mathbf{x}_0^{\mathsf{T}}\boldsymbol{\beta}+\varepsilon_0$，则$\tilde{y}_0$预测误差

$$e(\tilde{y}_0)=E(\tilde{y}_0-y_0)^2=m(\tilde{y}_0)+\sigma^2=\mathbf{x}_0^{\mathsf{T}}M(\tilde{\boldsymbol{\beta}})\mathbf{x}_0+\sigma^2,$$

其中均方误差 $m(\tilde{y}_0)=E(\tilde{y}_0-\mathbf{x}_0^{\mathsf{T}}\boldsymbol{\beta})^2=\mathbf{x}_0^{\mathsf{T}}M(\tilde{\boldsymbol{\beta}})\mathbf{x}_0.$

证：被预测变量$y_0=\mathbf{x}_0^{\mathsf{T}}\boldsymbol{\beta}+\varepsilon_0$的期望$\theta=E(y_0)=\mathbf{x}_0^{\mathsf{T}}\boldsymbol{\beta}$, 则

$$m(\tilde{y}_0)=E(\tilde{y}_0-\theta)^2=E\left(\mathbf{x}_0^{\mathsf{T}}\tilde{\boldsymbol{\beta}}-\mathbf{x}_0^{\mathsf{T}}\boldsymbol{\beta}\right)^2=E\left(\mathbf{x}_0^{\mathsf{T}}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})^{\mathsf{T}}\mathbf{x}_0\right)$$
$$=\mathbf{x}_0^{\mathsf{T}}E\left((\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})^{\mathsf{T}}\right)\mathbf{x}_0=\mathbf{x}_0^{\mathsf{T}}M(\tilde{\boldsymbol{\beta}})\mathbf{x}_0.$$

推论1. 设$\hat{\boldsymbol{\beta}}=(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y}$为LS估计, y_0的预测量取为$\hat{y}_0=\mathbf{x}_0^{\mathsf{T}}\hat{\boldsymbol{\beta}}$,
则预测误差$e(\hat{y}_0)=(1+\mathbf{x}_0^{\mathsf{T}}(X^{\mathsf{T}}X)^{-1}\mathbf{x}_0)\sigma^2$,
均方误差$m(\hat{y}_0)=\mathbf{x}_0^{\mathsf{T}}M(\hat{\boldsymbol{\beta}})\mathbf{x}_0=\mathbf{x}_0^{\mathsf{T}}(X^{\mathsf{T}}X)^{-1}\mathbf{x}_0\sigma^2.$

证：$bias(\hat{\boldsymbol{\beta}})=0,\ M(\hat{\boldsymbol{\beta}})=\mathrm{var}(\hat{\boldsymbol{\beta}})=\sigma^2(X^{\mathsf{T}}X)^{-1}.$

令$\tilde{\mathbf{y}} = X\tilde{\boldsymbol{\beta}}$为原数据$\mathbf{y}$的拟合$(\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon})$, 其$MSE$矩阵为:

$$M(\tilde{\mathbf{y}}) = E(\tilde{\mathbf{y}} - X\boldsymbol{\beta})(\tilde{\mathbf{y}} - X\boldsymbol{\beta})^{\mathsf{T}} = E(X\tilde{\boldsymbol{\beta}} - X\boldsymbol{\beta})(X\tilde{\boldsymbol{\beta}} - X\boldsymbol{\beta})^{\mathsf{T}} = XM(\tilde{\boldsymbol{\beta}})X^{\mathsf{T}}.$$

推论2. 在关于设计阵X通常的假设下$(n > p$及X列满秩$)$, $\tilde{y}_0 = \mathbf{x}_0^{\mathsf{T}}\tilde{\boldsymbol{\beta}}$的均方误差和预测误差与均方误差矩阵$M(\tilde{\mathbf{y}}) = XM(\tilde{\boldsymbol{\beta}})X^{\mathsf{T}}$有关,。

证明: X 列满秩$Rank(X) = p \Rightarrow L(X^{\mathsf{T}}) = R^p$, 对任何$\mathbf{x_0} \in R^p$, 存在$\mathbf{a} \in R^n$,

使得$\mathbf{x_0} = X^{\mathsf{T}}\mathbf{a}$, 此时$\tilde{y}_0 = \mathbf{x}_0^{\mathsf{T}}\tilde{\boldsymbol{\beta}} = \mathbf{a}^{\mathsf{T}}X\tilde{\boldsymbol{\beta}} = \mathbf{a}^{\mathsf{T}}\tilde{\mathbf{y}}$的$MSE$:

$$m(\tilde{y}_0) = \mathbf{x}_0^{\mathsf{T}}\mathrm{M}(\tilde{\boldsymbol{\beta}})\mathbf{x}_0 = \mathbf{a}^{\mathsf{T}}XM(\tilde{\boldsymbol{\beta}})X^{\mathsf{T}}\mathbf{a} = \mathbf{a}^{\mathsf{T}}M(\tilde{\mathbf{y}})\mathbf{a}.$$

$\tilde{y}_0 = \mathbf{x}_0^{\mathsf{T}}\tilde{\boldsymbol{\beta}}$的预测误差: $e(\tilde{y}_0) = \mathbf{a}^{\mathsf{T}}M(\tilde{\mathbf{y}})\mathbf{a} + \sigma^2$

推论2说明,在新数据上的预测效果主要由训练数据拟合值向量的均方误差$M(\tilde{\mathbf{y}})$决定，所以后续内容我们将主要考察$M(\tilde{\mathbf{y}})$.

常用的有偏（压缩）预测方法有：

- 变量选择：选取部分变量(其它变量的回归系数压缩为0);
- 压缩估计方法(shrinkage)：比如：$\tilde{\boldsymbol{\beta}} = \Lambda\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$，$\Lambda < I_p$
- 规则化方法 /惩罚最小二乘：对回归系数大小进行约束
- Bayes方法：通过假设参数的随机性，实现压缩、减少参数的效果。

注：这些方法界限不一定分明，比如（1）如果压缩估计把某些估计压缩为0, 则达到了选择变量的效果;（2）规则化方法可理解为Bayes方法。

基于子模型的预测

全模型（真模型）： $\boldsymbol{\beta}_1 : k\times 1,\ \boldsymbol{\beta}_2 : (p-k)\times 1$

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon},\quad \boldsymbol{\varepsilon} \sim (0, \sigma^2 I_n)$$

LS估计 $\hat{\boldsymbol{\beta}} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y}$，拟合值： $\hat{\mathbf{y}} = P_X\mathbf{y}$

子模型（选变量模型,错误模型）：

$$\mathbf{y} = X_1\boldsymbol{\beta}_1 + \boldsymbol{\delta},\quad \boldsymbol{\delta} \sim (0, \tau^2 I_n),$$

LS估计 $\tilde{\boldsymbol{\beta}}_1 = (X_1{}^{\mathsf{T}}X_1)^{-1}X_1{}^{\mathsf{T}}\mathbf{Y}$，全模型中 $\boldsymbol{\beta}$ 的估计 取为 $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1{}^{\mathsf{T}}, 0)^{\mathsf{T}}$，

拟合值： $\tilde{\mathbf{y}} = X\tilde{\boldsymbol{\beta}} = X_1\tilde{\boldsymbol{\beta}}_1 = P_{X_1}\mathbf{y}$

我们假设全模型是真模型，而子模型是错误的模型。在一定条件下，基于简单的较少变量的子模型比基于更多变量的真模型预测效果更好。

为了衡量子模型的预测效果，我们需要计算 $M(\tilde{\boldsymbol{\beta}})$ 或 $M(\tilde{\mathbf{y}})$。

引理1. 设$A=\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}>0$, 记$A_{11\bullet 2}=A_{11}-A_{12}A_{22}^{-1}A_{21}$, $A_{22\bullet 1}=A_{22}-A_{21}A_{11}^{-1}A_{12}$, 则

(1) $A_{11\bullet 2}^{-1}=A_{11}^{-1}+A_{11}^{-1}A_{12}A_{22\bullet 1}^{-1}A_{21}A_{11}^{-1}$, 同样$A_{22\bullet 1}^{-1}$有类似表达.

(2) $A_{11\bullet 2}^{-1}A_{12}A_{22}^{-1}=\left(A_{22\bullet 1}^{-1}A_{21}A_{11}^{-1}\right)^{\mathsf{T}}$.

(3) $A^{-1}=\begin{pmatrix} A_{11\bullet 2}^{-1} & -A_{11\bullet 2}^{-1}A_{12}A_{22}^{-1} \\ -A_{22\bullet 1}^{-1}A_{21}A_{11}^{-1} & A_{22\bullet 1}^{-1} \end{pmatrix}\geq\begin{pmatrix} A_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix}$

证:(1),(2)容易验证。由第4讲命题6, $A^{-1}=\begin{pmatrix} A_{11\bullet 2}^{-1} & -A_{11\bullet 2}^{-1}A_{12}A_{22}^{-1} \\ -A_{22\bullet 1}^{-1}A_{21}A_{11}^{-1} & A_{22\bullet 1}^{-1} \end{pmatrix}$

$$=\begin{pmatrix} A_{11}^{-1}+A_{11}^{-1}A_{12}A_{22\bullet 1}^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}A_{22\bullet 1}^{-1} \\ -A_{22\bullet 1}^{-1}A_{21}A_{11}^{-1} & A_{22\bullet 1}^{-1} \end{pmatrix}$$

$$=\begin{pmatrix} A_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix}+\begin{pmatrix} A_{11}^{-1}A_{12}A_{22\bullet 1}^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}A_{22\bullet 1}^{-1} \\ -A_{22\bullet 1}^{-1}A_{21}A_{11}^{-1} & A_{22\bullet 1}^{-1} \end{pmatrix}$$

$$=\begin{pmatrix} A_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix}+\begin{pmatrix} A_{11}^{-1}A_{12} \\ -I \end{pmatrix}A_{22\bullet 1}^{-1}\begin{pmatrix} A_{21}A_{11}^{-1} & -I \end{pmatrix}\geq\begin{pmatrix} A_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix}$$

引理2：对任何$\mathbf{x} \in R^n$，$\mathbf{x}^{\mathsf{T}}\mathbf{x} \le \lambda \Leftrightarrow \mathbf{x}\mathbf{x}^{\mathsf{T}} \le \lambda I_n$

证明：注意到$\mathbf{x}\mathbf{x}^{\mathsf{T}}$是秩为1的矩阵，其唯一非0特征根为$\mathbf{x}^{\mathsf{T}}\mathbf{x}$，

$\mathbf{x}\mathbf{x}^{\mathsf{T}} \le \lambda I_n \Leftrightarrow \mathbf{x}\mathbf{x}^{\mathsf{T}}$的所有特征根$\le \lambda \Leftrightarrow \mathbf{x}^{\mathsf{T}}\mathbf{x} \le \lambda$

引理3：实数$\lambda > 0$, 矩阵$A_{n\times n} > 0$，向量$\mathbf{x} \in R^n$, 则

$$\mathbf{x}^{\mathsf{T}}A^{-1}\mathbf{x} \le \lambda \Leftrightarrow \mathbf{x}\mathbf{x}^{\mathsf{T}} \le \lambda A$$

证明：若$\mathbf{x}^{\mathsf{T}}A^{-1}\mathbf{x} \le \lambda$，即$(A^{-1/2}\mathbf{x})^{\mathsf{T}}(A^{-1/2}\mathbf{x}) \le \lambda$

$\Rightarrow (A^{-1/2}\mathbf{x})(A^{-1/2}\mathbf{x})^{\mathsf{T}} \le \lambda I_n$, 即$A^{-1/2}\mathbf{x}\mathbf{x}^{\mathsf{T}}A^{-1/2} \le \lambda I_n$

$\Rightarrow \mathbf{x}\mathbf{x}^{\mathsf{T}} \le \lambda A$

反之，若$\lambda A \ge \mathbf{x}\mathbf{x}^{\mathsf{T}}$，则$\lambda I_n \ge A^{-1/2}\mathbf{x}\mathbf{x}^{\mathsf{T}}A^{-1/2} = (A^{-1/2}\mathbf{x})(A^{-1/2}\mathbf{x})^{\mathsf{T}}$，

$\Rightarrow \lambda \ge (A^{-1/2}\mathbf{x})^{\mathsf{T}}(A^{-1/2}\mathbf{x}) = \mathbf{x}^{\mathsf{T}}A^{-1}\mathbf{x}$

引理4：假设模型 $\mathbf{y}_{n\times 1}=X_{n\times p}\boldsymbol{\beta}_{p\times 1}+\boldsymbol{\varepsilon}_{n\times 1}=X_1\boldsymbol{\beta}_1+X_2\boldsymbol{\beta}_2+\boldsymbol{\varepsilon}$，$\boldsymbol{\varepsilon}\sim(0,\sigma^2 I_n)$,

其中$\boldsymbol{\beta}_1$和$\boldsymbol{\beta}_2$的长度分别为k和$p-k$，记$\tilde{\boldsymbol{\beta}}=\begin{pmatrix}\tilde{\boldsymbol{\beta}}_1\\ \mathbf{0}\end{pmatrix}$，其中$\tilde{\boldsymbol{\beta}}_1=(X_1{}^{\mathsf{T}}X_1)^{-1}X_1{}^{\mathsf{T}}\mathbf{y}$.

记$\tilde{\mathbf{y}}=X\tilde{\boldsymbol{\beta}}=X_1\tilde{\boldsymbol{\beta}}_1, A=(X_1{}^{\mathsf{T}}X_1)^{-1}X_1{}^{\mathsf{T}}X_2$，则

(1) $M(\tilde{\boldsymbol{\beta}})=\begin{pmatrix}\sigma^2(X_1{}^{\mathsf{T}}X_1)^{-1} & 0\\ 0 & 0^{\mathsf{T}}\end{pmatrix}+\begin{pmatrix}A\\ -I\end{pmatrix}\boldsymbol{\beta}_2\,\boldsymbol{\beta}_2{}^{\mathsf{T}}(A^{\mathsf{T}},-I).$

(2) $M(\tilde{\mathbf{y}})=\sigma^2 P_{X_1}+X_2^{\perp}\boldsymbol{\beta}_2\boldsymbol{\beta}_2{}^{\mathsf{T}}X_2^{\perp\mathsf{T}}$，$m(\tilde{\mathbf{y}})=k\sigma^2+\|X_2^{\perp}\boldsymbol{\beta}_2\|^2$.

$k=p$时，$\hat{\mathbf{y}}=X\hat{\boldsymbol{\beta}}$, $M(\hat{\mathbf{y}})=\sigma^2 P_X$, $m(\hat{\mathbf{y}})=p\sigma^2$

$M(\hat{\mathbf{y}})=\mathrm{var}(\hat{\mathbf{y}})=X\,\mathrm{var}(\hat{\boldsymbol{\beta}})X^{\mathsf{T}}=\sigma^2X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}$

$m(\hat{\mathbf{y}})=\mathrm{tr}M(\hat{\mathbf{y}})=p\sigma^2$

证明：(1) $E(\tilde{\boldsymbol{\beta}}_1) = E(X_1^{\mathsf{T}}X_1)^{-1}X_1^{\mathsf{T}}\mathbf{y} = (X_1^{\mathsf{T}}X_1)^{-1}X_1^{\mathsf{T}}(X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2) = \boldsymbol{\beta}_1 + A\boldsymbol{\beta}_2$

$\Rightarrow \text{Bias}(\tilde{\boldsymbol{\beta}}_1) = E(\tilde{\boldsymbol{\beta}}_1) - \boldsymbol{\beta}_1 = A\boldsymbol{\beta}_2,$

$\tilde{\boldsymbol{\beta}}$的偏差 $\mathbf{b} \triangleq E(\tilde{\boldsymbol{\beta}}) - \boldsymbol{\beta} = \mathrm{E}\begin{pmatrix}\tilde{\boldsymbol{\beta}}_1 \\ 0\end{pmatrix} - \begin{pmatrix}\boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2\end{pmatrix} = \begin{pmatrix}A\boldsymbol{\beta}_2 \\ -\boldsymbol{\beta}_2\end{pmatrix} = \begin{pmatrix}A \\ -I\end{pmatrix}\boldsymbol{\beta}_2$

另外，$\text{var}(\tilde{\boldsymbol{\beta}}_1) = \text{var}\left((X_1^{\mathsf{T}}X_1)^{-1}X_1^{\mathsf{T}}\mathbf{y}\right) = \sigma^2(X_1^{\mathsf{T}}X_1)^{-1}$，所以MSE矩阵

$$M(\tilde{\boldsymbol{\beta}}) = \text{var}(\tilde{\boldsymbol{\beta}}) + \mathbf{b}\mathbf{b}^{\mathsf{T}} = \begin{pmatrix}\sigma^2(X_1^{\mathsf{T}}X_1)^{-1} & 0 \\ 0 & 0\end{pmatrix} + \begin{pmatrix}A \\ -I\end{pmatrix}\boldsymbol{\beta}_2\boldsymbol{\beta}_2^{\mathsf{T}}(A^{\mathsf{T}}, -I)$$

(2) 因为 $\tilde{\mathbf{y}} = X_1\tilde{\boldsymbol{\beta}}_1 = X_1(X_1^{\mathsf{T}}X_1)^{-1}X_1^{\mathsf{T}}\mathbf{y} = P_{X_1}\mathbf{y},$ 所以，

偏差 $\mathrm{b}(\tilde{\mathbf{y}}) = E(\tilde{\mathbf{y}}) - X\boldsymbol{\beta} = P_{X_1}X\boldsymbol{\beta} - X\boldsymbol{\beta} = -X_2^{\perp}\boldsymbol{\beta}_2,$

$$\text{var}(\tilde{\mathbf{y}}) = \text{var}(P_{X_1}\mathbf{y}) = \sigma^2 P_{X_1},$$

$$\Rightarrow M(\tilde{\mathbf{y}}) = \text{var}(\tilde{\mathbf{y}}) + \mathrm{b}(\tilde{\mathbf{y}})\mathrm{b}(\tilde{\mathbf{y}})^{\mathsf{T}} = \sigma^2 P_{X_1} + X_2^{\perp}\boldsymbol{\beta}_2(X_2^{\perp}\boldsymbol{\beta}_2)^{\mathsf{T}}.$$

(2) 也可由(1)得到：

由 $A=(X_1{}^{\mathsf{T}}X_1)^{-1}X_1{}^{\mathsf{T}}X_2$ 及(1)的结果

$$M(\tilde{\mathbf{y}})=XM(\tilde{\boldsymbol{\beta}})X^{\mathsf{T}}=(X_1,X_2)\begin{pmatrix}\sigma^2\left(X_1{}^{\mathsf{T}}X_1\right)^{-1} & 0\\ 0 & 0'\end{pmatrix}\begin{pmatrix}X_1{}^{\mathsf{T}}\\ X_2{}^{\mathsf{T}}\end{pmatrix}$$

$$+(X_1,X_2)\begin{pmatrix}A\\ -I\end{pmatrix}\boldsymbol{\beta}_2\boldsymbol{\beta}_2{}^{\mathsf{T}}(A^{\mathsf{T}},-I)\begin{pmatrix}X_1{}^{\mathsf{T}}\\ X_2{}^{\mathsf{T}}\end{pmatrix}=\sigma^2P_{X_1}+X_2^{\perp}\boldsymbol{\beta}_2\boldsymbol{\beta}_2{}^{\mathsf{T}}X_2^{\perp\mathsf{T}}$$

$$\Rightarrow\ m(\tilde{\mathbf{y}})=tr(M(\tilde{\mathbf{y}}))=\sigma^2trP_{X_1}+tr(X_2^{\perp}\boldsymbol{\beta}_2\boldsymbol{\beta}_2{}^{\mathsf{T}}X_2^{\perp\mathsf{T}})=k\sigma^2+\|X_2^{\perp}\boldsymbol{\beta}_2\|^2\ .$$

命题2：假设模型 $\mathbf{y}_{n\times 1}=X_{n\times p}\boldsymbol{\beta}_{p\times 1}+\boldsymbol{\varepsilon}_{n\times 1}=X_1\boldsymbol{\beta}_1+X_2\boldsymbol{\beta}_2+\boldsymbol{\varepsilon}$，$\boldsymbol{\varepsilon}\sim(0,\sigma^2 I_n)$，

其中 X_1和X_2分别为$n\times k$和$n\times(p-k)$。令$\tilde{\mathbf{y}}=X\tilde{\boldsymbol{\beta}}=X_1\tilde{\boldsymbol{\beta}}_1$，其中$\tilde{\boldsymbol{\beta}}=\begin{pmatrix}\tilde{\boldsymbol{\beta}}_1\\ \mathbf{0}\end{pmatrix}$，

$\tilde{\boldsymbol{\beta}}_1=(X_1{}^{\mathsf{T}}X_1)^{-1}X_1{}^{\mathsf{T}}\mathbf{y}$。令$\hat{\boldsymbol{\beta}}$为LS估计，$\hat{\mathbf{y}}=X\hat{\boldsymbol{\beta}}$。则

(1) $\mathrm{var}(\tilde{\boldsymbol{\beta}}_1)\le\mathrm{var}(\hat{\boldsymbol{\beta}}_1),\ \mathrm{var}(\tilde{\boldsymbol{\beta}})\le\mathrm{var}(\hat{\boldsymbol{\beta}}).$

(2) 若$\|X_2^{\perp}\boldsymbol{\beta}_2\|\le\sqrt{p-k}\sigma$，则$m(\tilde{\mathbf{y}})\le m(\hat{\mathbf{y}})=p\sigma^2.$

(3) 若$\|X_2^{\perp}\boldsymbol{\beta}_2\|\le\sigma$，则$M(\tilde{\mathbf{y}})\le M(\hat{\mathbf{y}})=\sigma^2P_X$，且$M(\tilde{\boldsymbol{\beta}})\le M(\hat{\boldsymbol{\beta}})$。

注1：(1)与GM定理并不矛盾，因为$\tilde{\boldsymbol{\beta}}_1,\tilde{\boldsymbol{\beta}}$是线性有偏估计。

注2：$H_0:\boldsymbol{\beta}_2=0$的检验$F=\dfrac{\|X_2^{\perp}\hat{\boldsymbol{\beta}}_2\|^2}{(p-k)\hat{\sigma}^2}.$

条件(2)：$\|X_2^{\perp}\boldsymbol{\beta}_2\|\le\sqrt{p-k}\sigma$大致等价于$F\le 1$;

条件(3)：$\|X_2^{\perp}\boldsymbol{\beta}_2\|\le\sigma$ 大致等价于$F\le\dfrac{1}{p-k}.$

证明： (1) 所有期望计算都是 给定自变量条件下，

由引理1, $\mathrm{var}(\tilde{\boldsymbol{\beta}}) = \begin{pmatrix} \sigma^2 (X_1^{\mathsf{T}} X_1)^{-1} & 0 \\ 0 & 0 \end{pmatrix} \le \sigma^2 (X^{\mathsf{T}} X)^{-1} = \mathrm{var}(\hat{\boldsymbol{\beta}}) \Rightarrow \mathrm{var}(\tilde{\boldsymbol{\beta}}_1) = \mathrm{var}(\hat{\boldsymbol{\beta}}_1).$

或由 $\mathrm{var}(\tilde{\boldsymbol{\beta}}_1) = \mathrm{var}\left((X_1^{\mathsf{T}} X_1)^{-1} X_1^{\mathsf{T}} \mathbf{y}\right) = \sigma^2 (X_1^{\mathsf{T}} X_1)^{-1} \le \sigma^2 (X_1^{\perp\mathsf{T}} X_1^{\perp})^{-1} = \mathrm{var}(\hat{\boldsymbol{\beta}}_1).$

(2) 若 $\| X_2^{\perp} \boldsymbol{\beta}_2 \| \le \sqrt{p-k}\,\sigma$，则

$$m(\tilde{\mathbf{y}}) = k\sigma^2 + \| X_2^{\perp} \boldsymbol{\beta}_2 \|^2 \le k\sigma^2 + (p-k)\sigma^2 = p\sigma^2 = m(\hat{\mathbf{y}}).$$

(3) 因为 $M(\tilde{\mathbf{y}}) = E(X\tilde{\boldsymbol{\beta}} - X\boldsymbol{\beta})(X\tilde{\boldsymbol{\beta}} - X\boldsymbol{\beta})^{\mathsf{T}} = XM(\tilde{\boldsymbol{\beta}})X^{\mathsf{T}}$,

$M(\hat{\mathbf{y}}) = E(X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta})(X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta})^{\mathsf{T}} = XM(\hat{\boldsymbol{\beta}})X^{\mathsf{T}}$,

所以 $M(\tilde{\boldsymbol{\beta}}) \le M(\hat{\boldsymbol{\beta}}) \Leftrightarrow XM(\tilde{\boldsymbol{\beta}})X^{\mathsf{T}} \le XM(\hat{\boldsymbol{\beta}})X^{\mathsf{T}} \Leftrightarrow M(\tilde{\mathbf{y}}) \le M(\hat{\mathbf{y}}).$

下面我们仅需证明给定条件下，$M(\tilde{\mathbf{y}}) \le M(\hat{\mathbf{y}})$

由引理4(2), $M(\tilde{\mathbf{y}}) = \sigma^2 P_{X_1} + X_2^{\perp}\boldsymbol{\beta}_2\boldsymbol{\beta}_2^{\mathsf{T}} X_2^{\perp\mathsf{T}}$, $M(\hat{\mathbf{y}}) = \sigma^2 P_X$

$$M(\tilde{\mathbf{y}}) \le M(\hat{\mathbf{y}}) \Leftrightarrow X_2^{\perp}\boldsymbol{\beta}_2\boldsymbol{\beta}_2^{\mathsf{T}} X_2^{\perp\mathsf{T}} \le \sigma^2(P_X - P_{X_1}) = \sigma^2 P_{X_2^{\perp}} = \sigma^2 X_2^{\perp}(X_2^{\perp\mathsf{T}} X_2^{\perp})^{-1} X_2^{\perp\mathsf{T}}$$

$$\Leftrightarrow \boldsymbol{\beta}_2\boldsymbol{\beta}_2^{\mathsf{T}} \le \sigma^2(X_2^{\perp\mathsf{T}} X_2^{\perp})^{-1}$$

(两边同时左乘$X_2^{\perp\mathsf{T}}$,右乘$X_2^{\perp}$)

$$\overset{\text{引理3}}{\Leftrightarrow} \boldsymbol{\beta}_2^{\mathsf{T}} X_2^{\perp\mathsf{T}} X_2^{\perp}\boldsymbol{\beta}_2 \le \sigma^2$$

注意到最后一式即条件 $\| X_2^{\perp}\boldsymbol{\beta}_2 \| \le \sigma$,证毕。

$M(\tilde{\boldsymbol{\beta}}) \le M(\hat{\boldsymbol{\beta}})$的另外一个直接证明(利用命题1给出的$M(\tilde{\boldsymbol{\beta}})$表达式):

由于 $M(\tilde{\boldsymbol{\beta}}) = \begin{pmatrix} \sigma^2 \left(X_1{}^{\mathsf{T}} X_1\right)^{-1} & 0 \\ 0 & 0' \end{pmatrix} + \begin{pmatrix} A \\ -I \end{pmatrix} \boldsymbol{\beta}_2 \boldsymbol{\beta}_2{}^{\mathsf{T}} (A^{\mathsf{T}}, -I)$, 其中 $A = (X_1{}^{\mathsf{T}} X_1)^{-1} X_1{}^{\mathsf{T}} X_2$.

记 $B = (X_2^{\perp\mathsf{T}} X_2^{\perp})^{-1}$, $\hat{\boldsymbol{\beta}}$的MSE矩阵: $M(\hat{\boldsymbol{\beta}}) = \sigma^2 (X^{\mathsf{T}} X)^{-1} = \sigma^2 \begin{pmatrix} \left(X_1{}^{\mathsf{T}} X_1\right)^{-1} + ABA^{\mathsf{T}} & -AB \\ -BA^{\mathsf{T}} & B \end{pmatrix}$

$$= \sigma^2 \begin{pmatrix} \left(X_1{}^{\mathsf{T}} X_1\right)^{-1} & 0 \\ 0 & 0' \end{pmatrix} + \sigma^2 \begin{pmatrix} ABA^{\mathsf{T}} & -AB \\ -BA^{\mathsf{T}} & B \end{pmatrix}$$

所以 $M(\hat{\boldsymbol{\beta}}) - M(\tilde{\boldsymbol{\beta}}) = \sigma^2 \begin{pmatrix} ABA^{\mathsf{T}} & -AB \\ -BA' & B \end{pmatrix} - \begin{pmatrix} A \\ -I \end{pmatrix} \boldsymbol{\beta}_2 \boldsymbol{\beta}_2{}^{\mathsf{T}} (A^{\mathsf{T}}, -I)$

$= \sigma^2 \begin{pmatrix} A \\ -I \end{pmatrix} B (A^{\mathsf{T}}, -I) - \begin{pmatrix} A \\ -I \end{pmatrix} \boldsymbol{\beta}_2 \boldsymbol{\beta}_2{}^{\mathsf{T}} (A^{\mathsf{T}}, -I).$

由引理3知，条件 $\| X_2^{\perp} \boldsymbol{\beta}_2 \|^2 = \boldsymbol{\beta}_2{}^{\mathsf{T}} X_2^{\perp\mathsf{T}} X_2^{\perp} \boldsymbol{\beta}_2 \le \sigma^2 \Leftrightarrow \boldsymbol{\beta}_2 \boldsymbol{\beta}_2{}^{\mathsf{T}} \le \sigma^2 (X_2^{\perp\mathsf{T}} X_2^{\perp})^{-1} = \sigma^2 B$,

所以 $M(\hat{\boldsymbol{\beta}}) - M(\tilde{\boldsymbol{\beta}}) \ge 0.$