课程主页：http://staff.ustc.edu.cn/~ynyang/2022

第十二讲 最小二乘法 II

2022.11.18

误差同分布假设是对所有数据之间的共性(统计意义上相同)，在数学上给出的一种方便的描述，使得我们可以在平均意义下(期望)评价数据分析方法的效果。

Recap

- $\mathbf{y} \in R^n$在$V \subset R^n$上的正交投影 $\hat{\mathbf{y}} = P_V\mathbf{y} \in V$满足

  $\mathbf{y} - P_V\mathbf{y} \perp V \Leftrightarrow$ 对任何$\mathbf{x} \in V$，$(\mathbf{x}, \mathbf{y}) = (\mathbf{x}, P_V\mathbf{y})$

- 若$V = C(A)$, 则$P_V = A(A^{\mathrm{T}}A)^{-}A^{\mathrm{T}}$

- 对于矩阵$B = (\mathbf{b}_1, \ldots, \mathbf{b}_k)$, $P_V B = (P_V\mathbf{b}_1, \ldots, P_V\mathbf{b}_k)$为各列的投影组成的矩阵。

A, B为矩阵或向量，
$A^{\perp} = A - P_C A$，　$B^{\perp} = B - P_C B$，
则 ${A^{\perp}}^{\mathrm{T}} B^{\perp} = {A^{\perp}}^{\mathrm{T}} B$

$$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = \Sigma(x_i - \bar{x})y_i$$

线性回归模型 $\mathbf{y}_{n\times 1} = X_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}, \boldsymbol{\varepsilon} \sim (0, \sigma^2 I_n)$, 正则方程：

$$X^{\mathrm{T}}\boldsymbol{\varepsilon} = X^{\mathrm{T}}(\mathbf{y} - X\boldsymbol{\beta}) = 0,$$

(X各列与$\boldsymbol{\varepsilon}$正交，各个变量与误差不相关)

解得$\boldsymbol{\beta}$的LS估计

$$\hat{\boldsymbol{\beta}} = (X^{\mathrm{T}}X)^{-}X^{\mathrm{T}}\mathbf{y},$$

误差方差的LS估计

拟合值 残差

- 投影$\hat{\mathbf{y}} = P_X\,\mathbf{y} = X\hat{\boldsymbol{\beta}} = X(X^{\mathsf{T}}X)^{-}X^{\mathsf{T}}\mathbf{y}$,称为拟合值向量。
- $\mathbf{e} = \mathbf{y}^{\perp} = \mathbf{y} - \hat{\mathbf{y}} = (I_n - P_X)\mathbf{y}$称为残差向量。

$\hat{\boldsymbol{\beta}}$满足正则方程：$X^{\mathsf{T}}(\mathbf{y} - X\hat{\boldsymbol{\beta}}) = X^{\mathsf{T}}\mathbf{e} = 0$, 即$\mathbf{e} \perp C(X)$,
特别地$\mathbf{e} \perp \mathbf{1}$, 即$\bar{e} = \mathbf{1}^{\mathsf{T}}\mathbf{e}/n = 0$ (样本均值为0)

误差方差估计

残差平方和：$RSS = \|\mathbf{e}\|^2 = \|\mathbf{y} - X\hat{\boldsymbol{\beta}}\|^2 = \mathbf{y}^{\mathsf{T}}(I_n - P_X)\mathbf{y}$

σ^2的LS估计定义为：$\hat{\sigma}^2 = \dfrac{\mathrm{RSS}}{n-p} = \dfrac{1}{n-p}\|\mathbf{e}\|^2$。

拟合优度：决定系数 R^2

模型： $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{1}\beta_0 + Z\mathbf{b} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim (0, \sigma^2 I_n)$

拟合值/投影： $\hat{\mathbf{y}} = P_X\mathbf{y}$; 残差/投影： $\mathbf{e} = \mathbf{y}^{\perp} = \mathbf{y} - \hat{\mathbf{y}} = (I_n - P_X)\mathbf{y}$

注意到$P_X\mathbf{1} = \mathbf{1}$,所以$\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_n)^{\mathrm{T}}$的样本均值:

$$\bar{\hat{y}} = \mathbf{1}^T\hat{\mathbf{y}}/n = \mathbf{1}^T P_X\mathbf{y}/n = \mathbf{1}^T\mathbf{y}/n = \bar{y}$$

所有残差的均值: $\bar{e} = \mathbf{1}^T\mathbf{e}/n = 0$

正交分解及平方和分解

正交分解： $(\mathbf{y} - \mathbf{1}\bar{y}) = (\hat{\mathbf{y}} - \mathbf{1}\bar{y}) + \mathbf{e}, \quad \mathbf{e} \perp \mathbf{1}, \hat{\mathbf{y}}$

$\bar{\hat{y}}$

平方和分解：

$$\|\mathbf{y} - \mathbf{1}\bar{y}\|^2 = \|\hat{\mathbf{y}} - \mathbf{1}\bar{y}\|^2 + \|\mathbf{e}\|^2$$

$$\mathrm{SS}_{\text{总}} = \mathrm{SS}_{\text{回}} + RSS$$

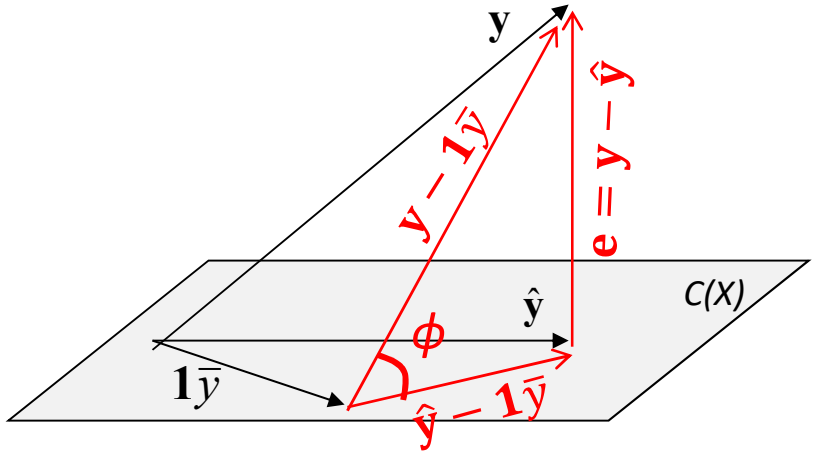
(样本版本的)决定系数

$$R^2 = \frac{\mathrm{SS}_{\text{回}}}{\mathrm{SS}_{\text{总}}} = \frac{VAR(\hat{\mathbf{y}})}{VAR(\mathbf{y})} = \frac{\|\hat{\mathbf{y}} - \mathbf{1}\bar{y}\|^2}{\|\mathbf{y} - \mathbf{1}\bar{y}\|^2}, \quad 0 \le R^2 \le 1.$$

拟合值$\hat{\mathbf{y}}$的样本方差占$\mathbf{y}$的样本方差的比例。

VAR表示样本方差。

图示：总平方和分解 (红色)



中心化后的红色三角形：

$$\mathbf{y}-\mathbf{1}\bar{y}=(\hat{\mathbf{y}}-\mathbf{1}\bar{y})+\mathbf{e},\quad \mathbf{e}\perp\hat{\mathbf{y}}-\mathbf{1}\bar{y}$$

$$\|\mathbf{y}-\mathbf{1}\bar{y}\|^2=\|\hat{\mathbf{y}}-\mathbf{1}\bar{y}\|^2+\|\mathbf{e}\|^2$$

$$SS_{总}=SS_{回}+RSS$$

$$R^2=\frac{\|\hat{\mathbf{y}}-\mathbf{1}\bar{y}\|^2}{\|\mathbf{y}-\mathbf{1}\bar{y}\|^2}=\cos^2(\phi)=\left(r_{y\hat{y}}\right)^2$$

R^2：y 与 ŷ的相关系数的平方

命题1. (1) $R^2=\left(r_{y\hat{y}}\right)^2$, $r_{y\hat{y}}$为$(y_i,\hat{y}_i), i=1,\ldots,n$的样本相关系数。

(2) $R^2=\max\limits_{\mathbf{u}\in L(X)}(r_{\mathbf{u},\mathbf{y}})^2$, 其中$r_{\mathbf{u},\mathbf{y}}$为样本相关系数，

且最大值在$\mathbf{u}=\hat{\mathbf{y}}=X\hat{\boldsymbol{\beta}}$处达到，其中$\hat{\boldsymbol{\beta}}=(X^{\mathsf{T}}X)^{-}X^{\mathsf{T}}\mathbf{y}$。

证明：(1) $[r_{y\hat{y}}]^2=\dfrac{(\mathbf{y}-\mathbf{1}\bar{y},\hat{\mathbf{y}}-\mathbf{1}\bar{y})^2}{\|\mathbf{y}-\mathbf{1}\bar{y}\|^2\|\hat{\mathbf{y}}-\mathbf{1}\bar{y}\|^2}$.

因为$\mathbf{e}\perp\hat{\mathbf{y}}, \mathbf{e}\perp\mathbf{1}$, 所以

$$(\mathbf{y}-\mathbf{1}\bar{y},\hat{\mathbf{y}}-\mathbf{1}\bar{y})=(\hat{\mathbf{y}}-\mathbf{1}\bar{y},\hat{\mathbf{y}}-\mathbf{1}\bar{y})=\|\hat{\mathbf{y}}-\mathbf{1}\bar{y}\|^2$$

$$\Rightarrow[r_{y\hat{y}}]^2=\frac{\|\hat{\mathbf{y}}-\mathbf{1}\bar{y}\|^2}{\|\mathbf{y}-\mathbf{1}\bar{y}\|^2}=R^2$$

虚线框内容可忽略

(2) 记$\mathbf{y}_c\triangleq\mathbf{y}-P_1\mathbf{y}=\mathbf{y}-\mathbf{1}\bar{y}$。对任何$\mathbf{u}=X\boldsymbol{\beta}=\mathbf{1}\beta_0+Z\mathbf{b}\in C(X)$,

令 $\mathbf{u}_c\overset{\Delta}{=}\mathbf{u}-P_1\mathbf{u}=(\mathbf{1}\beta_0+Z\mathbf{b})-P_1(\mathbf{1}\beta_0+Z\mathbf{b})=Z\mathbf{b}-P_1Z\mathbf{b}=Z_c\mathbf{b}$

则 $(r_{\mathbf{u},\mathbf{y}})^2=\dfrac{(\mathbf{u}_c{}^{\mathsf{T}}\mathbf{y}_c)^2}{\|\mathbf{u}_c\|^2\|\mathbf{y}_c\|^2}=\dfrac{(\mathbf{b}^{\mathsf{T}}Z_c^{\mathsf{T}}\mathbf{y}_c)^2}{\mathbf{b}^{\mathsf{T}}Z_c^{\mathsf{T}}Z_c\mathbf{b}\times\|\mathbf{y}_c\|^2}$

令$\mathbf{w}=(Z_c{}^{\mathsf{T}}Z_c)^{1/2}\mathbf{b}$，则 $(r_{\mathbf{u},\mathbf{y}})^2=\dfrac{\left(\mathbf{w}^{\mathsf{T}}\left[(Z_c^{\mathsf{T}}Z_c)^{-1/2}Z_c^{\mathsf{T}}\mathbf{y}_c\right]\right)^2}{\mathbf{w}^{\mathsf{T}}\mathbf{w}\cdot\|\mathbf{y}_c\|^2}$

$\leq\dfrac{\mathbf{w}^{\mathsf{T}}\mathbf{w}\times\ \mathbf{y}_c{}^{\mathsf{T}}Z_c(Z_c^{\mathsf{T}}Z_c)^{-}Z_c^{\mathsf{T}}\mathbf{y}_c}{\mathbf{w}^{\mathsf{T}}\mathbf{w}\times\|\mathbf{y}_c\|^2}$ （Cauchy-Schwartz不等式）

$=\dfrac{\mathbf{w}^{\mathsf{T}}\mathbf{w}\times\|P_{Z_c}\mathbf{y}_c\|^2}{\mathbf{w}^{\mathsf{T}}\mathbf{w}\times\|\mathbf{y}_c\|^2}=\dfrac{\|P_{Z_c}\mathbf{y}\|^2}{\|\mathbf{y}_c\|^2}$ $\leftarrow P_{Z_c}\mathbf{y}=P_{Z_c}\mathbf{y}_c$

$=\dfrac{\|\hat{\mathbf{y}}-\mathbf{1}\bar{y}\|^2}{\|\mathbf{y}-\mathbf{1}\bar{y}\|^2}=R^2$ $\leftarrow \hat{\mathbf{y}}=\mathbf{1}\bar{y}+P_{Z_c}\mathbf{y}$

当$\mathbf{w}\propto(Z_c{}^{\mathsf{T}}Z_c)^{-1/2}Z_c{}^{\mathsf{T}}\mathbf{y}_c\Leftrightarrow\mathbf{b}=(Z_c{}^{\mathsf{T}}Z_c)^{-1/2}\mathbf{w}\propto(Z_c{}^{\mathsf{T}}Z_c)^{-}Z_c{}^{\mathsf{T}}\mathbf{y}_c$时，等号成立.

此时 $\mathbf{u}_c=Z_c\mathbf{b}\propto Z_c(Z_c{}^{\mathsf{T}}Z_c)^{-}Z_c{}^{\mathsf{T}}\mathbf{y}_c=Z_c\hat{\mathbf{b}}$,特别地在$\mathbf{u}=Z_c\hat{\mathbf{b}}+\mathbf{1}\bar{y}=X\hat{\boldsymbol{\beta}}$达到极大。

LS估计的统计性质

即使没有概率模型假设，我们也能操作LS。

$\varepsilon_1, \ldots, \varepsilon_n$ iid $\sim(0, \sigma^2)$的假设反映了我们对误差的认知：

$\varepsilon_1, \ldots, \varepsilon_n$大致相同，$\varepsilon_i \approx 0$且$|\varepsilon_i| \approx \sigma$

该假设可用来评价LS估计是否准确，比如，LS估计$\hat{\boldsymbol{\beta}}$在平均意义下是否等于或接近真正的$\boldsymbol{\beta}$？LS方法的逼近效果如何，逼近误差即误差平方和RSS大概有多大？

例如，$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$中，$\mathrm{RSS} = \|\mathbf{e}\|^2 = \mathbf{y}^{\mathsf{T}}(I_n - P_X)\mathbf{y} = (X\boldsymbol{\beta} + \boldsymbol{\varepsilon})^{\mathsf{T}}(I_n - P_X)(X\boldsymbol{\beta} + \boldsymbol{\varepsilon})$

$= \boldsymbol{\varepsilon}^{\mathsf{T}}(I_n - P_X)\boldsymbol{\varepsilon}$，平均来看（当$X$给定时）：

$$E(\mathrm{RSS}) = E\boldsymbol{\varepsilon}^{\mathsf{T}}(I_n - P_X)\boldsymbol{\varepsilon} = tr\big((I_n - P_X)\,\mathrm{var}(\boldsymbol{\varepsilon})\big) = (n-r)\sigma^2$$

其中$r = tr(P_X) = rank(X)$.

第3讲命题1：若$\mathbf{x} \sim (\boldsymbol{\mu}, \Sigma)$，则

$E(\mathbf{x}^{\mathsf{T}} A\mathbf{x}) = \boldsymbol{\mu}^{\mathsf{T}} A\boldsymbol{\mu} + tr(A\Sigma)$.

LS估计的无偏性和方差

定理1. 假设线性模型$\mathbf{y} = X_{n\times p}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$，$E\boldsymbol{\varepsilon} = 0, \mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$，其中$\boldsymbol{\varepsilon}$与$X$独立。假设$X$是列满秩的（$n \geq p$, 且$X^{\mathsf{T}}X$可逆），则

(1) LS估计的无偏性： $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$，$E(\hat{\sigma}^2) = \sigma^2$.

（2）LS估计的方差：$\mathrm{var}(\hat{\boldsymbol{\beta}} \mid X) = \sigma^2 \left(X^{\mathsf{T}}X\right)^{-1}$.

证明：因为$\boldsymbol{\varepsilon}$与X独立, 所以

- $E(\mathbf{y} \mid X) = E(X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \mid X) = X\boldsymbol{\beta} + E(\boldsymbol{\varepsilon} \mid X) = X\boldsymbol{\beta} + E(\boldsymbol{\varepsilon}) = X\boldsymbol{\beta}$
- $\mathrm{var}(\mathbf{y} \mid X) = \mathrm{var}(X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \mid X) = \mathrm{var}(\boldsymbol{\varepsilon} \mid X) = \mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$

(1) $E(\hat{\boldsymbol{\beta}} \mid X) = E\left(\left(X^{\mathsf{T}}X\right)^{-1}X^{\mathsf{T}}\mathbf{y} \mid X\right) = \left(X^{\mathsf{T}}X\right)^{-1}X^{\mathsf{T}}\left[E(\mathbf{y} \mid X)\right] = \boldsymbol{\beta} \Rightarrow E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}.$

上页我们已经证明了$E(RSS \mid X) = (n-p)\sigma^2 \Rightarrow E(RSS) = (n-p)\sigma^2$

$\Rightarrow E(\hat{\sigma}^2) = E(RSS/(n-p)) = \sigma^2.$

(2) $\mathrm{var}(\hat{\boldsymbol{\beta}} \mid X) = \mathrm{var}(\left(X^{\mathsf{T}}X\right)^{-1}X^{\mathsf{T}}\mathbf{y} \mid X)$

$= \left(X^{\mathsf{T}}X\right)^{-1}X^{\mathsf{T}}\left[\mathrm{var}(\mathbf{y} \mid X)\right]X\left(X^{\mathsf{T}}X\right)^{-1} = \sigma^2\left(X^{\mathsf{T}}X\right)^{-1}$

$\hat{\boldsymbol{\beta}}$的方差估计

在方差公式 $\mathrm{var}(\hat{\boldsymbol{\beta}} \mid X)=\sigma^2\left(X^{\mathsf{T}}X\right)^{-1}$ 中代入（Plug-in）σ^2的估计，即得到$\hat{\boldsymbol{\beta}}$方差的估计：

$$\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}} \mid X)=\hat{\sigma}^2\left(X^{\mathsf{T}}X\right)^{-1}$$

例1. 假设 (y_i, x_i)，$i=1,2...,n$ 独立，满足简单线性回归模型：

$$y_i = a + bx_i + \varepsilon_i ,$$

第6讲命题1、2中我们已经求得LS估计及其方差，这里我们以矩阵向量形式再次计算如下：记 $\boldsymbol{\beta}=\begin{pmatrix} a \\ b \end{pmatrix}, \mathbf{y}=(y_1,...,y_n)^{\mathsf{T}}$，设计阵 $X=\begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}=(\mathbf{1},\mathbf{x})$，$\mathbf{x}\neq\mathbf{1}c$,

模型为 $\mathbf{y}=X\boldsymbol{\beta}+\boldsymbol{\varepsilon}=\mathbf{1}a+\mathbf{x}b+\boldsymbol{\varepsilon}$，LS估计

$$\hat{\boldsymbol{\beta}}=\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix}=(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y}=\begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \Sigma x_i^2 \end{pmatrix}^{-1}\begin{pmatrix} n\bar{y} \\ \Sigma x_i y_i \end{pmatrix}$$

$$=\frac{1}{n(\Sigma x_i^2 - n\bar{x}^2)}\begin{pmatrix} \Sigma x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix}\begin{pmatrix} n\bar{y} \\ \Sigma x_i y_i \end{pmatrix}=\begin{pmatrix} \bar{y}-\bar{x}s_{xy}/s_{xx} \\ s_{xy}/s_{xx} \end{pmatrix}$$

$$\mathrm{var}(\hat{\boldsymbol{\beta}} \mid X)=\sigma^2(X^{\mathsf{T}}X)^{-1}=\frac{\sigma^2}{ns_{xx}}\begin{pmatrix} \Sigma x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix}=\sigma^2\begin{pmatrix} 1/n+\bar{x}^2/s_{xx} & -\bar{x}/s_{xx} \\ -\bar{x}/s_{xx} & 1/s_{xx} \end{pmatrix}$$

LS估计的最优性

定义：对任何两个对称 $n\times n$矩阵A,B, 若$A-B\geq 0$ (非负定), 则称在Loewner偏序意义下 A不小于B, 记作$A\geq B$.

性质：若 $A\geq B$，则对任何 $k\times n$矩阵C，$CAC^{\mathsf{T}}\geq CBC^{\mathsf{T}}$

Gauss-Markov 定理

定理2. 线性模型 $\mathbf{y}=X\boldsymbol{\beta}+\boldsymbol{\varepsilon}$中,假设$X_{n\times p}$列满秩，则 $\boldsymbol{\beta}$的最小二乘估计$\hat{\boldsymbol{\beta}}=(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y}$是最优无偏线性估计 (BLUE, best linear unbiased estimate,), 即对任何 $\boldsymbol{\beta}$的线性无偏估计 $\tilde{\boldsymbol{\beta}}=C\mathbf{y}$ (其中C是仅与X有关的$p\times n$常数矩阵)，$\mathrm{var}(\tilde{\boldsymbol{\beta}}\mid X)\geq \mathrm{var}(\hat{\boldsymbol{\beta}}\mid X)$).

推论：假设$A_{k\times p}$是仅与X有关的常数矩阵，如果A行满秩,则$A\hat{\boldsymbol{\beta}}$ 是$A\boldsymbol{\beta}$的BLUE。

特别地，$\hat{\beta}_k$是β_k的BLUE，$\hat{\beta}_k-\hat{\beta}_j$是$\beta_k-\beta_j$的BLUE，等等。

证明：给定X，设$\tilde{\boldsymbol{\beta}} = C\,\mathbf{y}$是$\boldsymbol{\beta}$的任一线性无偏估计，其中$C$是$p\times n$常数矩阵(可能与$X$有关)，因为$\mathrm{var}(\tilde{\boldsymbol{\beta}}\,|\,X) = C\,\mathrm{var}(\mathbf{y}\,|\,X)C^{\mathsf{T}} = \sigma^2 CC^{\mathsf{T}}$，$\mathrm{var}(\hat{\boldsymbol{\beta}}\,|\,X) = \sigma^2(X^{\mathsf{T}}X)^{-1}$。我们需要证$CC^{\mathsf{T}} \geq (X^{\mathsf{T}}X)^{-1}$.

因为$\tilde{\boldsymbol{\beta}}$线性无偏，所以

$$\boldsymbol{\beta} = \mathrm{E}(\tilde{\boldsymbol{\beta}}\,|\,X) = E(C\mathbf{y}\,|\,X) = CX\boldsymbol{\beta},$$

上式对任何$\boldsymbol{\beta}$成立，故$CX = I_p$。

因为$CX = I_p$，$P_X \leq I_n$,所以

$$(X^{\mathsf{T}}X)^{-1} = CX(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}C^{\mathsf{T}} = CP_XC^{\mathsf{T}} \leq CC^{\mathsf{T}}.$$

不确定性原理

例如，对于简单线性模型$y_i = a + bx_i + \varepsilon_i$的任一无偏估计$\tilde{b}$，由GM定理

$$\mathrm{var}(\tilde{b}\,|\,x) \geq \mathrm{var}(\hat{b}\,|\,x) = \sigma^2 / s_{xx}$$

以$\mathrm{VAR}(x)$代表样本方差$s_x^2 = s_{xx}/(n-1)$，上式等价于

$$\mathrm{var}(\tilde{b}\,|\,x)\times \mathrm{VAR}(x) \geq \sigma^2/(n-1),$$

这说明左端两个方差不可能同时很小，即x和$\tilde{b}$不可能同时精确测量(Uncertainty principle,不确定性原理,测不准原理).

中心化

中心化是数据分析中常见的数据加工方法，它将每一个样本观测值减去样本均值，正等价于所有样本构成的向量或矩阵与向量𝟙的正交化。

一元情形

样本：$y_1,...,y_n$； $\mathbf{y}=(y_1,...,y_n)^{\mathsf{T}}$

样本均值 $\bar{y}=(y_1+...+y_n)/n=\mathbf{1}^{\mathsf{T}}\mathbf{y}/n$, 其中$\mathbf{1}=(1,...,1)^{\mathsf{T}}$

中心化： $y_1-\bar{y},...,y_n-\bar{y}$（样本均值为0）

$$\mathbf{y}_c=\begin{pmatrix} y_1-\bar{y} \\ \vdots \\ y_n-\bar{y} \end{pmatrix}=\mathbf{y}-\mathbf{1}\bar{y}=\mathbf{y}-\mathbf{1}\mathbf{1}^{\mathsf{T}}\mathbf{y}/n=(I_n-P_{\mathbf{1}})\mathbf{y}$$

样本方差： $s^2=\sum_{i=1}^{n}(y_1-\bar{y})^2/(n-1)=\mathbf{y}_c^{\mathsf{T}}\mathbf{y}_c/(n-1)=\mathbf{y}^{\mathsf{T}}(I_n-P_{\mathbf{1}})\mathbf{y}/(n-1).$

中心化基础上再除以标准差，称为标准化 : $(y_1-\bar{y})/s,...,(y_n-\bar{y})/s$，

标准化后样本均值为0，样本方差为1.

多元情形

样本：$\mathbf{x}_1,...,\mathbf{x}_n \in R^m$；样本按行排列成$n\times m$矩阵：$Z=(\mathbf{x}_1,...,\mathbf{x}_n)^{\mathsf{T}}$

样本均值：$\bar{\mathbf{x}}=(\mathbf{x}_1+...+\mathbf{x}_n)/n=Z^{\mathsf{T}}\mathbf{1}/n.$

样本协方差矩阵：$S=\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^{\mathsf{T}}/(n-1)$

中心化：$\mathbf{x}_1-\bar{\mathbf{x}},...,\mathbf{x}_n-\bar{\mathbf{x}}$，按行排列成中心化矩阵：

$$Z_c=\begin{pmatrix}(\mathbf{x}_1-\bar{\mathbf{x}})^{\mathsf{T}}\\ \vdots \\ (\mathbf{x}_n-\bar{\mathbf{x}})^{\mathsf{T}}\end{pmatrix}=Z-\mathbf{1}\bar{\mathbf{x}}^{\mathsf{T}}=Z-\mathbf{1}\mathbf{1}^{\mathsf{T}}Z/n=(I_n-P_1)Z$$

（注意中心化矩阵Z_c的每列均值为0，即$Z_c^{\mathsf{T}}\mathbf{1}=0$）

所以样本协方差矩阵：

$$S=Z_cZ_c^{\mathsf{T}}/(n-1)=Z^{\mathsf{T}}(I_n-P_1)Z/(n-1).$$

如果中心化矩阵Z_c的每一列(变量)除以其标准差，则称为标准化

$$Z_s=Z_cD^{-1/2},D=diag(S)$$

- 如果某个数据矩阵A已经中心化，那么$A^{\mathsf{T}}A/(n-1)=S$,样本协方差矩阵。
- 如果某个数据矩阵A已经标准化，那么$A^{\mathsf{T}}A/(n-1)=R$,样本相关系数矩阵。

线性模型的中心化

线性模型$\mathbf{y}=X\boldsymbol{\beta}+\boldsymbol{\varepsilon}$中我们把截距项单独列 出来，为此划分

$$X_{n\times p}=\begin{pmatrix}1 & \mathbf{x}_1^{\mathsf{T}}\\ \vdots & \vdots\\ 1 & \mathbf{x}_n^{\mathsf{T}}\end{pmatrix}=(\mathbf{1}_{n\times 1},Z_{n\times(p-1)}),\ \boldsymbol{\beta}=\begin{pmatrix}\beta_0\\ \mathbf{b}\end{pmatrix}.$$

- 模型：$\mathbf{y}=X\boldsymbol{\beta}+\boldsymbol{\varepsilon}=\mathbf{1}\beta_0+Z\mathbf{b}+\boldsymbol{\varepsilon}$
- Z中心化：令 中心化$Z_c=Z-\mathbf{1}\bar{\mathbf{x}}^{\mathsf{T}}=(I_n-P_1)Z$，模型等价地表示为

$$\mathbf{y}=\mathbf{1}\beta_0+Z\mathbf{b}+\boldsymbol{\varepsilon}=\mathbf{1}\beta_0+(Z_c+\mathbf{1}\bar{\mathbf{x}}^{\mathsf{T}})\mathbf{b}+\boldsymbol{\varepsilon}=\mathbf{1}(\beta_0+\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{b})+Z_c\mathbf{b}+\boldsymbol{\varepsilon}$$
$$\overset{\Delta}{=}\mathbf{1}\alpha+Z_c\mathbf{b}+\boldsymbol{\varepsilon}=X^*\boldsymbol{\beta}^*+\boldsymbol{\varepsilon}$$

其中$X^*=(\mathbf{1},Z_c)$,回归系数$\boldsymbol{\beta}^*=\begin{pmatrix}\alpha\\ \mathbf{b}\end{pmatrix}$，$\alpha=\beta_0+\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{b}$。

- 所以中心化不改变回归 系数$\mathbf{b}$，但改变了截距项。

下面演示中心化对于计算过程的简化功能：

中心化模型 $\mathbf{y}=\mathbf{1}\alpha+Z_c\mathbf{b}+\boldsymbol{\varepsilon}=X^*\boldsymbol{\beta}^*+\boldsymbol{\varepsilon}$中$Z_c$的列与$\mathbf{1}$正交，这使得

$X^{*\mathsf{T}}X^*=\begin{pmatrix}n & 0\\ 0 & Z_c^{\mathsf{T}}Z_c\end{pmatrix}$为分块对角，对中心化模型应用最小二乘法

$$\hat{\boldsymbol{\beta}}^*=\begin{pmatrix}\hat{\alpha}\\ \hat{\mathbf{b}}\end{pmatrix}=(X^{*\mathsf{T}}X^*)^{-1}X^{*\mathsf{T}}\mathbf{y}=\begin{pmatrix}n & 0\\ 0 & Z_c^{\mathsf{T}}Z_c\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{1}^{\mathsf{T}}\mathbf{y}\\ Z_c^{\mathsf{T}}\mathbf{y}\end{pmatrix}=\begin{pmatrix}\bar{y}\\ (Z_c^{\mathsf{T}}Z_c)^{-1}Z_c^{\mathsf{T}}\mathbf{y}\end{pmatrix}$$

所以LS估计 $\hat{\alpha}=\bar{y}$，$\hat{\mathbf{b}}=(Z_c^{\mathsf{T}}Z_c)^{-1}Z_c^{\mathsf{T}}\mathbf{y}$。由 $\alpha=\beta_0+\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{b}$，得 $\hat{\beta}_0=\bar{y}-\bar{\mathbf{x}}^{\mathsf{T}}\hat{\mathbf{b}}$。注意到

$$Z_c^{\mathsf{T}}Z_c=\sum(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^{\mathsf{T}}=(n-1)S_{\mathbf{xx}},\quad Z_c^{\mathsf{T}}\mathbf{y}=(n-1)\,S_{\mathbf{x}y}$$

$$\Rightarrow \hat{\mathbf{b}}=(Z_cZ_c^{\mathsf{T}})^{-1}Z_c^{\mathsf{T}}\mathbf{y}=S_{\mathbf{xx}}^{-1}S_{\mathbf{x}y}。$$

$$\mathrm{var}(\hat{\mathbf{b}}\mid X)=\sigma^2(Z_c^{\mathsf{T}}Z_c)^{-1}=\sigma^2S_{\mathbf{xx}}^{-1}/(n-1)。$$

另外，$\hat{\mathbf{y}}=P_{\mathbf{1}}\mathbf{y}+P_{Z_c}\mathbf{y}=\mathbf{1}\bar{y}+P_{Z_c}\mathbf{y}$，所以 $\hat{\mathbf{y}}-\mathbf{1}\bar{y}=P_{Z_c}\mathbf{y}\ \ (=Z_c\hat{\mathbf{b}})$,

所以 $\mathrm{SS}_{回}=\|\hat{\mathbf{y}}-\mathbf{1}\bar{y}\|^2=\|P_{Z_c}\mathbf{y}\|^2=\mathbf{y}^{\mathsf{T}}Z_c(Z_c^{\mathsf{T}}Z_c)^{-1}Z_c^{\mathsf{T}}\mathbf{y}=(n-1)S_{y\mathbf{x}}S_{\mathbf{xx}}^{-1}S_{\mathbf{x}y}$

而 $\mathrm{SS}_{总}=\|\mathbf{y}-\mathbf{1}\bar{y}\|^2=(n-1)S_{yy}$

$$\Rightarrow RSS=\mathrm{SS}_{总}-\mathrm{SS}_{回}=(n-1)\left[S_{yy}-S_{y\mathbf{x}}S_{\mathbf{xx}}^{-1}S_{\mathbf{x}y}\right]=(n-1)S_{yy\bullet\mathbf{x}},$$

所以 $R^2=\dfrac{S_{y\mathbf{x}}S_{\mathbf{xx}}^{-1}S_{\mathbf{x}y}}{S_{yy}}$，$\hat{\sigma}^2=\mathrm{RSS}/(n-p)=\dfrac{n-1}{n-p}S_{yy\bullet\mathbf{x}}。$

我们有如下结论（命题1）。

命题2. 假设模型 $y_i = \beta_0 + \mathbf{x}_i^{\mathsf{T}}\mathbf{b} + \varepsilon_i, i = 1,\ldots,n$，记$X = (\mathbf{1}, Z)$列满秩，

$Z = (\mathbf{x}_1,\ldots,\mathbf{x}_n)^{\mathsf{T}}$，$Z_c = Z - \mathbf{1}\bar{\mathbf{x}}^{\mathsf{T}}$，样本方差/协方差矩阵

$$S_{\mathbf{xx}} = \frac{1}{n-1}\sum(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathsf{T}},\ S_{\mathbf{x}y} = \frac{1}{n-1}\sum(\mathbf{x}_i - \bar{\mathbf{x}})(y_i - \bar{y})$$

则

（1）LS估计 $\hat{\mathbf{b}} = (Z_c Z_c^{\mathsf{T}})^{-1} Z_c^{\mathsf{T}}\mathbf{y} = S_{\mathbf{xx}}^{-1} S_{\mathbf{x}y}$，$\hat{\beta}_0 = \bar{y} - \bar{\mathbf{x}}^{\mathsf{T}}\hat{\mathbf{b}}$，$\hat{\sigma}^2 = \frac{n-1}{n-p} S_{yy\bullet\mathbf{x}} \approx S_{yy\bullet\mathbf{x}}$。

（2）$\mathrm{var}(\hat{\mathbf{b}} \mid X) = \sigma^2 (Z_c Z_c^{\mathsf{T}})^{-1} = \sigma^2 S_{\mathbf{xx}}^{-1}/(n-1)$,

（3）$R^2 = \frac{S_{y\mathbf{x}} S_{\mathbf{xx}}^{-1} S_{\mathbf{x}y}}{S_{yy}}$.

对比第5讲命题2(总体版本)

$\mathbf{b} = \Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{x}y}$

$\beta_0 = \mu_y - \mathbf{b}^{\mathsf{T}}\boldsymbol{\mu}_{\mathbf{x}}$

$\sigma^2 = \Sigma_{yy\bullet\mathbf{x}}$

$R^2 = \Sigma_{y\mathbf{x}}\Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{x}y}/\Sigma_{yy}$

命题2的第一部分（**b**的LS估计）可利用投影更简单地证明。

证1（投影）：令$Z_c = Z - P_{\mathbf{1}}Z = Z - \mathbf{1}\bar{\mathbf{x}}^{\mathsf{T}}$，则

$$\hat{\mathbf{y}} = P_X\mathbf{y} = P_{\mathbf{1}}\mathbf{y} + P_{Z_c}\mathbf{y} = \mathbf{1}(\mathbf{1}^{\mathsf{T}}\mathbf{y}/n) + Z_c[(Z_c^{\mathsf{T}}Z_c)^{-1}Z_c^{\mathsf{T}}\mathbf{y}]$$

$$= \mathbf{1}\bar{y} + (Z - \mathbf{1}\bar{\mathbf{x}}^{\mathsf{T}})[S_{\mathbf{xx}}^{-1}S_{\mathbf{x}y}] = \mathbf{1}\left(\bar{y} - \bar{\mathbf{x}}^{\mathsf{T}}[S_{\mathbf{xx}}^{-1}S_{\mathbf{x}y}]\right) + Z[S_{\mathbf{xx}}^{-1}S_{\mathbf{x}y}]$$

1的系数为$\hat{\beta}_0$，Z的系数为$\hat{\mathbf{b}}$，所以$\hat{\mathbf{b}} = S_{\mathbf{xx}}^{-1}S_{\mathbf{x}y}$，$\hat{\beta}_0 = \bar{y} - \bar{\mathbf{x}}^{\mathsf{T}}\hat{\mathbf{b}}$。

其它更直接（但更复杂)的证明：

证2.(应用分块矩阵的逆，求$\hat{\boldsymbol{\beta}}$的分量) $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} = (\mathbf{1}, \mathrm{Z})\begin{pmatrix} \beta_0 \\ \mathbf{b} \end{pmatrix} + \boldsymbol{\varepsilon}$

由定理1，已知*LS*估计$\hat{\boldsymbol{\beta}} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y}$, 下面求分量$\hat{\mathbf{b}}$。

分块：$X^{\mathsf{T}}X = \begin{pmatrix} \mathbf{1}^{\mathsf{T}} \\ Z^{\mathsf{T}} \end{pmatrix}(\mathbf{1}, \mathrm{Z}) = \begin{pmatrix} \mathbf{1}^{\mathsf{T}}\mathbf{1} & \mathbf{1}^{\mathsf{T}}Z \\ Z^{\mathsf{T}}\mathbf{1} & Z^{\mathsf{T}}Z \end{pmatrix} = \begin{pmatrix} n & n\bar{\mathbf{x}}^{\mathsf{T}} \\ n\bar{\mathbf{x}} & \sum \mathbf{x}_i\mathbf{x}_i^{\mathsf{T}} \end{pmatrix}, X^{\mathsf{T}}\mathbf{y} = \begin{pmatrix} \mathbf{1}^{\mathsf{T}} \\ Z^{\mathsf{T}} \end{pmatrix}\mathbf{y} = \begin{pmatrix} n\bar{y} \\ \sum \mathbf{x}_i y_i \end{pmatrix},$

由分块矩阵求逆公式（第3讲命题3）

$$(X^{\mathsf{T}}X)^{-1} = \begin{pmatrix} a & -\bar{\mathbf{x}}^{\mathsf{T}}A^{-1} \\ -A^{-1}\bar{\mathbf{x}} & A^{-1} \end{pmatrix}, \text{其中} A = (n-1)S_{\mathbf{xx}},\ a = 1/n + \bar{\mathbf{x}}^{\mathsf{T}}A^{-1}\bar{\mathbf{x}}$$

$$\Rightarrow \hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\mathbf{b}} \end{pmatrix} = \begin{pmatrix} n & n\bar{\mathbf{x}}^{\mathsf{T}} \\ n\bar{\mathbf{x}} & \sum \mathbf{x}_i\mathbf{x}_i^{\mathsf{T}} \end{pmatrix}^{-1}\begin{pmatrix} n\bar{y} \\ \sum \mathbf{x}_i y_i \end{pmatrix} = \begin{pmatrix} a & -\bar{\mathbf{x}}^{\mathsf{T}}A^{-1} \\ -A^{-1}\bar{\mathbf{x}} & A^{-1} \end{pmatrix}\begin{pmatrix} n\bar{y} \\ \sum \mathbf{x}_i y_i \end{pmatrix}$$

$$= \begin{pmatrix} na\bar{y} - \bar{\mathbf{x}}^{\mathsf{T}}A^{-1}\sum \mathbf{x}_i y_i \\ A^{-1}\sum \mathbf{x}_i y_i - nA^{-1}\bar{\mathbf{x}}\bar{y} \end{pmatrix} = \begin{pmatrix} \bar{y} - \bar{\mathbf{x}}^{\mathsf{T}}S_{\mathbf{xx}}^{-1}S_{\mathbf{x}y} \\ S_{\mathbf{xx}}^{-1}S_{\mathbf{x}y} \end{pmatrix}$$

所以$\hat{\mathbf{b}} = S_{\mathbf{xx}}^{-1}S_{\mathbf{x}y}$，$\hat{\beta}_0 = \bar{y} - \bar{\mathbf{x}}^{\mathsf{T}}\hat{\mathbf{b}}$.

第三讲命题3: $\Sigma > 0$,

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \Sigma_{11\bullet 2}^{-1} & -\Sigma_{11\bullet 2}^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\ -\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11\bullet 2}^{-1} & \Sigma_{22\bullet 1}^{-1} \end{pmatrix}$$

其中各个子块有不同的表达：

$\Sigma_{11\bullet 2}^{-1} = \Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22\bullet 1}^{-1}\Sigma_{21}\Sigma_{11}^{-1}$,

$\Sigma_{11\bullet 2}^{-1}\Sigma_{12}\Sigma_{22}^{-1} = \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22\bullet 2}^{-1}$ 等等

证3（误差平方和分别对 β_0和$\mathbf{b}$求导） $\boldsymbol{\beta}=(\beta_0,\mathbf{b}^{\mathsf{T}})^{\mathsf{T}}$

$$Q(\boldsymbol{\beta})=\|\mathbf{y}-X\boldsymbol{\beta}\|^2=\sum_{i=1}^{n}\left(y_i-\tilde{\mathbf{x}}_i^{\mathsf{T}}\boldsymbol{\beta}\right)^2=\sum_{i=1}^{n}\left(y_i-\beta_0-\mathbf{x}_i^{\mathsf{T}}\mathbf{b}\right)^2$$

$$\frac{\partial Q}{\partial\boldsymbol{\beta}}=0\Leftrightarrow\begin{cases}\dfrac{\partial Q}{\partial\beta_0}=\displaystyle\sum_{i=1}^{n}\left(y_i-\beta_0-\mathbf{x}_i^{\mathsf{T}}\mathbf{b}\right)=0\\ \dfrac{\partial Q}{\partial\mathbf{b}}=\displaystyle\sum_{i=1}^{n}\mathbf{x}_i\left(y_i-\beta_0-\mathbf{x}_i^{\mathsf{T}}\mathbf{b}\right)=0\end{cases}\Leftrightarrow\begin{cases}\sum y_i-n\beta_0-\left(\sum\mathbf{x}_i\right)^{\mathsf{T}}\mathbf{b}=0\\ \sum\mathbf{x}_i y_i-\left(\sum\mathbf{x}_i\right)\beta_0-\left(\sum\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\right)\mathbf{b}=0\end{cases}$$

$$\Leftrightarrow\begin{cases}\bar{y}-\beta_0-\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{b}=0\\ \sum\mathbf{x}_i y_i-n\bar{\mathbf{x}}\beta_0-\left(\sum\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\right)\mathbf{b}=0\end{cases}$$

第一式左乘$n\bar{\mathbf{x}}$减去第二式，消去β_0，得：

消去β_0的过程⇔中心化

$$\left(\sum\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}-n\bar{\mathbf{x}}\bar{\mathbf{x}}^{\mathsf{T}}\right)\mathbf{b}=\sum\mathbf{x}_i y_i-n\bar{\mathbf{x}}\bar{y}$$

$$\Rightarrow\hat{\mathbf{b}}=\left(\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^{\mathsf{T}}\right)^{-1}\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})y_i=S_{\mathbf{xx}}^{-1}S_{\mathbf{x}y},\hat{\beta}_0=\bar{y}-\bar{\mathbf{x}}^{\mathsf{T}}\hat{\mathbf{b}},$$

关于“不妨假设数据已经中心化”

我们已知：

- 如果$\mathbf{y}=X\boldsymbol{\beta}+\boldsymbol{\varepsilon}=\mathbf{1}\beta_0+Z\mathbf{b}+\boldsymbol{\varepsilon}$中$Z$已经中心化，那么$\hat{\mathbf{b}}=(Z^{\mathsf{T}}Z)^{-1}Z^{\mathsf{T}}\mathbf{y}$，$\hat{\beta}_0=\bar{y}$
- 如果$\mathbf{y}$，X都已经中心化，那么$\hat{\mathbf{b}}=(Z^{\mathsf{T}}Z)^{-1}Z^{\mathsf{T}}\mathbf{y}$，$\hat{\beta}_0=0$，此时$\beta_0$无需估计。

在一些问题中，首先假设Z中心化，会简化后续论证过程。

- $\mathbf{y}=Z\mathbf{b}+\boldsymbol{\varepsilon}$（不含截距项），$\mathbf{y}$，$Z$都已经中心化

  LS估计$\hat{\mathbf{b}}=(Z^{\mathsf{T}}Z)^{-1}Z^{\mathsf{T}}\mathbf{y}=S_{\mathbf{xx}}^{-1}S_{\mathbf{xy}}$
- $\mathbf{y}=Z\mathbf{b}+\boldsymbol{\varepsilon}$（不含截距项），$\mathbf{y}$，$Z$都已经标准化

  LS估计$\hat{\mathbf{b}}=(Z^{\mathsf{T}}Z)^{-1}Z^{\mathsf{T}}\mathbf{y}=R_{\mathbf{xx}}^{-1}R_{\mathbf{xy}}$

Recap

我们已知，模型 $\mathbf{y}=X\boldsymbol{\beta}+\boldsymbol{\varepsilon}$两端同时左乘$X^{\mathsf{T}}$，并令$X^{\mathsf{T}}\boldsymbol{\varepsilon}=0$

$$X^{\mathsf{T}}\mathbf{y}=X^{\mathsf{T}}X\boldsymbol{\beta}$$

即得$\boldsymbol{\beta}$的LS估计 $\hat{\boldsymbol{\beta}}=(X^{\mathsf{T}}X)^{-}X^{\mathsf{T}}\mathbf{y}$,

$\boldsymbol{\beta}=\begin{pmatrix}\beta_0\\ \mathbf{b}\end{pmatrix}$，LS估计 $\hat{\mathrm{b}}$ 的直观求解方法

模型$\mathbf{y}=X\boldsymbol{\beta}+\boldsymbol{\varepsilon}=\mathbf{1}\beta_0+Z\mathbf{b}+\boldsymbol{\varepsilon}$两边同时左乘 Z_c (Z_c与$\mathbf{1}$列正交，$Z_c^{\mathsf{T}}\mathbf{1}=0$，但与$Z$最接近)， 并令$Z_c^{\mathsf{T}}\boldsymbol{\varepsilon}=0$:

$$Z_c^{\mathsf{T}}\mathbf{y}=Z_c^{\mathsf{T}}\mathbf{1}\beta_0+Z_c^{\mathsf{T}}Z\mathbf{b}+Z_c^{\mathsf{T}}\boldsymbol{\varepsilon}=Z_c^{\mathsf{T}}Z\mathbf{b}=Z_c^{\mathsf{T}}Z_c\mathbf{b}$$

$$\Rightarrow \hat{\mathbf{b}}=(Z_c^{\mathsf{T}}Z_c)^{-1}Z_c^{\mathsf{T}}\mathbf{y}$$

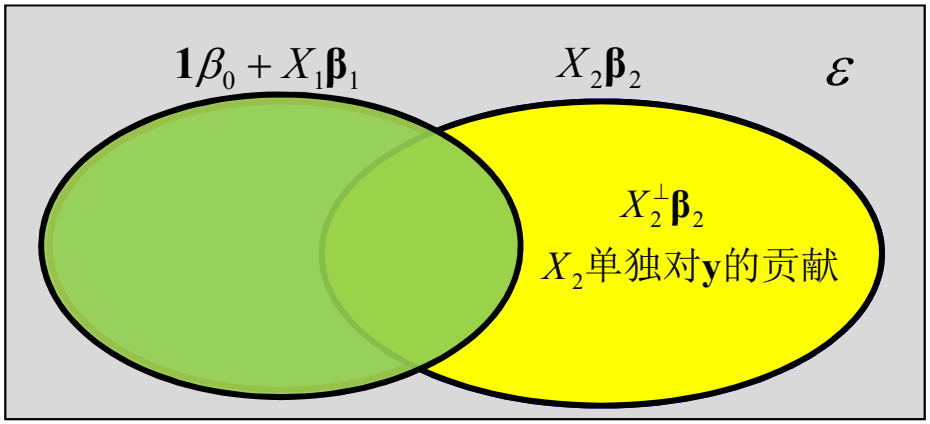
LS估计的分量

假设设计阵X按列划分为$X=(\mathbf{1},X_1,X_2)$，模型为

$$\mathbf{y}=X\boldsymbol{\beta}+\boldsymbol{\varepsilon}=\mathbf{1}\beta_0+X_1\boldsymbol{\beta}_1+X_2\boldsymbol{\beta}_2+\boldsymbol{\varepsilon}$$

其中X_1的列是干扰因素，X_2的列是感兴趣的变量。LS估计$\hat{\boldsymbol{\beta}}_2$是否有效地消除了$X_1$的影响？

$\mathbf{y}=$



$X_2^{\perp}=X_2-P_{\mathbf{1},X_1}X_2$

事实上LS估计 $\hat{\boldsymbol{\beta}}_2=(X_2^{\perp\mathsf{T}}X_2^{\perp})^{-1}X_2^{\perp\mathsf{T}}\mathbf{y}=(X_2^{\perp\mathsf{T}}X_2^{\perp})^{-1}X_2^{\perp\mathsf{T}}\mathbf{y}^{\perp}$，

其中的$X_2^{\perp}=X_2-P_{\mathbf{1},X_1}X_2$，$\mathbf{y}^{\perp}=\mathbf{y}-P_{\mathbf{1},X_1}\mathbf{y}$都消除了$X_1$的影响。

记号：记所有数据$(\mathbf{y}, X_1, X_2)$的样本方差-协方差矩阵为

$$S=\begin{array}{c} \\ y \\ \mathbf{x}_1 \\ \mathbf{x}_2 \end{array}\begin{array}{c} \begin{array}{ccc} y & \mathbf{x}_1 & \mathbf{x}_2 \end{array} \\ \begin{pmatrix} S_{yy} & S_{y1} & S_{y2} \\ S_{1y} & S_{11} & S_{12} \\ S_{2y} & S_{21} & S_{22} \end{pmatrix} \end{array}, \quad \begin{pmatrix} S_{yy\bullet 1} & S_{y2\bullet 1} \\ S_{2y\bullet 1} & S_{22\bullet 1} \end{pmatrix}=\begin{pmatrix} S_{yy} & S_{y2} \\ S_{2y} & S_{22} \end{pmatrix}-\begin{pmatrix} S_{y1} \\ S_{21} \end{pmatrix} S_{11}^{-1}\left(S_{1y}, S_{12}\right)$$

LS估计的自相似性

命题3：假设线性模型：$\mathbf{y}_{n\times 1}=X_{n\times p}\boldsymbol{\beta}_{p\times 1}+\boldsymbol{\varepsilon}_{n\times 1}=\mathbf{1}\beta_0+X_1\boldsymbol{\beta}_1+X_2\boldsymbol{\beta}_2+\boldsymbol{\varepsilon}$

其中$X=(\mathbf{1}, X_1, X_2), \boldsymbol{\beta}=(\beta_0, \boldsymbol{\beta}_1^{\mathsf{T}}, \boldsymbol{\beta}_2^{\mathsf{T}})^{\mathsf{T}}$，记$X_2^{\perp}=X_2-P_{\mathbf{1},X_1}X_2$，则

(1) $\hat{\boldsymbol{\beta}}_2=(X_2^{\perp\mathsf{T}}X_2^{\perp})^{-1}X_2^{\perp\mathsf{T}}\mathbf{y}$,

$\mathrm{var}(\hat{\boldsymbol{\beta}}_2 \mid X)=\sigma^2(X_2^{\perp\mathsf{T}}X_2^{\perp})^{-1}$.

（注意：$\hat{\boldsymbol{\beta}}_2$也等于$(X_2^{\perp\mathsf{T}}X_2^{\perp})^{-1}X_2^{\perp\mathsf{T}}\mathbf{y}^{\perp}$，其中$\mathbf{y}^{\perp}=\mathbf{y}-P_{\mathbf{1},X_1}\mathbf{y}$）

(2) $\hat{\boldsymbol{\beta}}_2$也可表示为$\hat{\boldsymbol{\beta}}_2=S_{22\bullet 1}^{-1}S_{2y\bullet 1}$,

$\mathrm{var}(\hat{\boldsymbol{\beta}}_2 \mid X))=\sigma^2 S_{22\bullet 1}^{-1}/(n-1)$。

对比第5讲命题3(总体版本)

$\boldsymbol{\beta}_2=\Sigma_{22\bullet 1}^{-1}\Sigma_{2y\bullet 1}$

LS估计 $\hat{\boldsymbol{\beta}}_2$ 的直观求解

模型 $\mathbf{y}=X_1\boldsymbol{\beta}_1+X_2\boldsymbol{\beta}_2+\boldsymbol{\varepsilon}$ 两端同时左乘$X_2^{\perp\mathsf{T}}$，并令$X_2^{\perp\mathsf{T}}\boldsymbol{\varepsilon}=\mathbf{0}$（矩估计方法），得

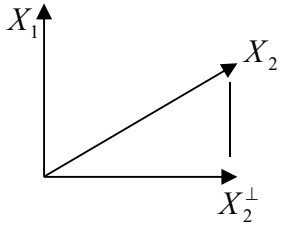
$$X_2^{\perp\mathsf{T}}\mathbf{y}=X_2^{\perp\mathsf{T}}X_2\boldsymbol{\beta}_2\Rightarrow\hat{\boldsymbol{\beta}}_2=\left(X_2^{\perp\mathsf{T}}X_2^{\perp}\right)^{-1}X_2^{\perp\mathsf{T}}\mathbf{y}.$$

为什么方程两端同时左乘$X_2^{\perp\mathsf{T}}$？

(1) $X_2^{\perp}$与X_1列正交（$X_2^{\perp}$：X_2中消除了与X_1有关的部分）。

(2) 在与X_1列正交的矩阵中，$X_2^{\perp}$与X_2最为接近(*LS*,参见右图)。

(3) 如此得到的解是$\boldsymbol{\beta}_2$的LS估计(命题3)，参见下页例子。



两步回归求 $\hat{\boldsymbol{\beta}}_2$

$\hat{\boldsymbol{\beta}}_2=(X_2^{\perp\mathsf{T}}X_2^{\perp})^{-1}X_2^{\perp\mathsf{T}}\mathbf{y}^{\perp}$可以看作是两步回归的结果：

- $X_2\sim X_1$（多元回归）得残差：$X_2^{\perp}=X_2-P_{\mathbf{1},X_1}X_2$，
  $\mathbf{y}\sim X_1$得残差：$\mathbf{y}^{\perp}=\mathbf{y}-P_{\mathbf{1},X_1}\mathbf{y}$，
- $\mathbf{y}^{\perp}\sim X_2^{\perp}$：$\mathbf{y}=X_2^{\perp}\boldsymbol{\beta}_2+\boldsymbol{\delta}$得$\hat{\boldsymbol{\beta}}_2=(X_2^{\perp\mathsf{T}}X_2^{\perp})^{-1}X_2^{\perp\mathsf{T}}\mathbf{y}^{\perp}$。

理解为矩阵$X_2=(\mathbf{x}_1,\ldots,\mathbf{x}_k)$的各列对$X_1$回归：

$\mathbf{x}_i\sim X_1\Rightarrow$ 残差$\hat{\mathbf{e}}_i$，

$X_2^{\perp}=(\hat{\mathbf{e}}_1,\ldots,\hat{\mathbf{e}}_k)$

例2. 模型方程模型 $\mathbf{y}=X_1\boldsymbol{\beta}_1+X_2\boldsymbol{\beta}_2+\boldsymbol{\varepsilon}$, 设$X_2$为$n\times q$矩阵， 设$W$为任何$n\times q$矩阵, $W^{\mathsf{T}}X_1=0, W^{\mathsf{T}}X_2$可逆， 模型两端同时左乘$W^{\mathsf{T}}$,并令$W^{\mathsf{T}}\boldsymbol{\varepsilon}/n=\mathbf{0}$:

$$W^{\mathsf{T}}\mathbf{y}=W^{\mathsf{T}}X_1+W^{\mathsf{T}}X_2\boldsymbol{\beta}_2+W^{\mathsf{T}}\boldsymbol{\varepsilon}=W^{\mathsf{T}}X_2\boldsymbol{\beta}_2$$

$\Rightarrow \widetilde{\boldsymbol{\beta}}_2=(W^{\mathsf{T}}X_2)^{-1}W^{\mathsf{T}}\mathbf{y}$, 则$\widetilde{\boldsymbol{\beta}}_2$是无偏估计,但其方差大于$\hat{\boldsymbol{\beta}}_2$的方差。

证明：

$W^{\mathsf{T}}X_1=0\Rightarrow E(\widetilde{\boldsymbol{\beta}}_2\mid X)=(W^{\mathsf{T}}X_2)^{-1}W^{\mathsf{T}}(X_1\boldsymbol{\beta}_1+X_2\boldsymbol{\beta}_2)=(W^{\mathsf{T}}X_2)^{-1}W^{\mathsf{T}}X_2\boldsymbol{\beta}_2=\boldsymbol{\beta}_2.$

为证 $\mathrm{var}(\widetilde{\boldsymbol{\beta}}_2\mid X)=\sigma^2(W^{\mathsf{T}}X_2)^{-1}W^{\mathsf{T}}W(X_2{}^{\mathsf{T}}W)^{-1}\geq\sigma^2\left(X_2^{\perp\mathsf{T}}X_2^{\perp}\right)^{-1}=\mathrm{var}(\hat{\boldsymbol{\beta}}_2\mid X),$

只需证 $W^{\mathsf{T}}W\geq W^{\mathsf{T}}X_2\left(X_2^{\perp\mathsf{T}}X_2^{\perp}\right)^{-1}X_2{}^{\mathsf{T}}W.$

因为$W^{\mathsf{T}}X_1=0$, 所以 $W^{\mathsf{T}}X_2=W^{\mathsf{T}}X_2^{\perp}$,所以上述不等式等价于

$$W^{\mathsf{T}}W\geq W^{\mathsf{T}}X_2^{\perp}\left(X_2^{\perp\mathsf{T}}X_2^{\perp}\right)^{-1}X_2^{\perp\mathsf{T}}W$$

因为$X_2^{\perp}\left(X_2^{\perp\mathsf{T}}X_2^{\perp}\right)^{-1}X_2^{\perp\mathsf{T}}\leq I_n$,该不等式显然成立。

命题3的证明

证明1（投影）:（1）不妨设 X_1, X_2已经中心化，所以 $\mathbf{1}, X_1, X_2^{\perp}$三部分相互列正交，所以

$$\begin{aligned}\hat{\mathbf{y}} &= P_X\mathbf{y} = P_{(\mathbf{1},X_1,X_2)}\mathbf{y} = P_{(\mathbf{1},X_1,X_2^{\perp})}\mathbf{y} = P_{\mathbf{1}}\,\mathbf{y} + P_{X_1}\mathbf{y} + P_{X_2^{\perp}}\mathbf{y} \\ &= \mathbf{1}\bar{y} + X_1\left((X_1^{\mathsf{T}}X_1)^{-1}X_1^{\mathsf{T}}\mathbf{y}\right) + X_2^{\perp}\left((X_2^{\perp\mathsf{T}}X_2^{\perp})^{-1}X_2^{\perp\mathsf{T}}\mathbf{y}\right) \\ &\overset{\Delta}{=} \mathbf{1}\bar{y} + X_1\mathbf{u} + X_2^{\perp}\mathbf{v} = \mathbf{1}\bar{y} + X_1\mathbf{u} + \left(X_2 - X_1(X_1^{\mathsf{T}}X_1)^{-1}X_1^{\mathsf{T}}X_2\right)\mathbf{v} \\ &= \mathbf{1}\bar{y} + X_1\left(\mathbf{u} - (X_1^{\mathsf{T}}X_1)^{-1}X_1^{\mathsf{T}}X_2\mathbf{v}\right) + X_2\mathbf{v},\end{aligned}$$

X_2的系数$\mathbf{v}$即$\hat{\boldsymbol{\beta}}_2$，$\hat{\boldsymbol{\beta}}_2 = \mathbf{v} = (X_2^{\perp\mathsf{T}}X_2^{\perp})^{-1}X_2^{\perp}\mathbf{y} = \left(X_2^{\perp\mathsf{T}}X_2^{\perp}\right)^{-1}X_2^{\perp\mathsf{T}}\mathbf{y}$

（2）$X_2^{\perp} = X_2 - P_{X_1}X_2 \Rightarrow X_2^{\perp\mathsf{T}}X_2^{\perp} = X_2^{\mathsf{T}}(I_n - P_{X_1})X_2 = X_2^{\mathsf{T}}X_2 - X_2^{\mathsf{T}}X_1(X_1^{\mathsf{T}}X_1)^{-1}X_1^{\mathsf{T}}X_2$

$= (n-1)S_{22} - (n-1)S_{21}S_{11}^{-1}S_{12} = (n-1)S_{22\bullet 1}$，同样$X_2^{\perp\mathsf{T}}\mathbf{y} = (n-1)S_{2y\bullet 1}$

所以$\hat{\boldsymbol{\beta}}_2 = (X_2^{\perp\mathsf{T}}X_2^{\perp})^{-1}X_2^{\perp\mathsf{T}}\mathbf{y} = S_{22\bullet 1}^{-1}S_{2y\bullet 1}$,

$\mathrm{var}(\hat{\boldsymbol{\beta}}_2 \mid X) = \sigma^2(X_2^{\perp\mathsf{T}}X_2^{\perp})^{-1} = \sigma^2 S_{22\bullet 1}^{-1}/(n-1).$

证明2：不妨假设X_1, X_2已经中心化，这不影响$\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$的LS估计.

$X_2^{\perp} = X_2 - P_{\mathbf{1},X_1} X_2 = X_2 - P_{X_1} X_2$， 改写模型：

$$
\begin{aligned}
&\mathbf{y} = \mathbf{1}\beta_0 + X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} = \mathbf{1}\beta_0 + X_1\boldsymbol{\beta}_1 + (X_2^{\perp} + P_{X_1}X_2)\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \\
&= \mathbf{1}\beta_0 + X_1\left[\boldsymbol{\beta}_1 + (X_1^{\mathsf{T}}X_1)^{-1}X_1^{\mathsf{T}}X_2\boldsymbol{\beta}_2\right] + X_2^{\perp}\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \\
&\doteq \mathbf{1}\beta_0 + X_1\boldsymbol{\beta}_1^{*} + X_2^{\perp}\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}, \text{其中} \boldsymbol{\beta}_1^{*} = \boldsymbol{\beta}_1 + (X_1^{\mathsf{T}}X_1)^{-1}X_1^{\mathsf{T}}X_2\boldsymbol{\beta}_2
\end{aligned}
$$

该模型得设计阵的三部分$\mathbf{1}, X_1, X_2^{\perp}$相互正交，$X^{\mathsf{T}}X$分块对角，所以由LS估计公式

易得$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}}_1^{*} \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{1}^{\mathsf{T}}\mathbf{1} & 0 & 0 \\ 0 & X_1^{\mathsf{T}}X_1 & 0 \\ 0 & 0 & X_2^{\perp\mathsf{T}}X_2^{\perp} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{1}^{\mathsf{T}}\mathbf{y} \\ X_1^{\mathsf{T}}\mathbf{y} \\ X_2^{\perp\mathsf{T}}\mathbf{y} \end{pmatrix} = \begin{pmatrix} \bar{y} \\ (X_1^{\mathsf{T}}X_1)^{-1}X_1^{\mathsf{T}}\mathbf{y} \\ (X_2^{\perp\mathsf{T}}X_2^{\perp})^{-1}X_2^{\perp\mathsf{T}}\mathbf{y} \end{pmatrix}$

所以$\hat{\boldsymbol{\beta}}_2 = (X_2^{\perp\mathsf{T}}X_2^{\perp})^{-1}X_2^{\perp\mathsf{T}}\mathbf{y}$。

证明3：（不假设X_1，X_2已经中心化）

$Q = \| \mathbf{y} - \mathbf{1}\beta_0 - X_1\boldsymbol{\beta}_1 - X_2\boldsymbol{\beta}_2 \|^2$ 求导得正则方程$X^{\mathsf{T}}(\mathbf{y} - X\boldsymbol{\beta}) = \mathbf{0}$:

$$\begin{cases} \partial Q / \partial \beta_0 = -\mathbf{1}^{\mathsf{T}}(\mathbf{y} - \mathbf{1}\beta_0 - X_1\boldsymbol{\beta}_1 - X_2\boldsymbol{\beta}_2) = 0 \\ \partial Q / \partial \boldsymbol{\beta}_1 = -X_1^{\mathsf{T}}(\mathbf{y} - \mathbf{1}\beta_0 - X_1\boldsymbol{\beta}_1 - X_2\boldsymbol{\beta}_2) = \mathbf{0} \\ \partial Q / \partial \boldsymbol{\beta}_2 = -X_2^{\mathsf{T}}(\mathbf{y} - \mathbf{1}\beta_0 - X_1\boldsymbol{\beta}_1 - X_2\boldsymbol{\beta}_2) = \mathbf{0} \end{cases}$$

下面应用初等消去法解正则方程(即求$X^{\mathsf{T}}X$的逆）

第一式整理得：$\beta_0 = \bar{y} - \bar{\mathbf{x}}_1^{\mathsf{T}}\boldsymbol{\beta}_1 - \bar{\mathbf{x}}_2^{\mathsf{T}}\boldsymbol{\beta}_2$，代入二、三式

$$\begin{cases} X_1^{\mathsf{T}}\left[\mathbf{y} - \mathbf{1}(\bar{y} - \bar{\mathbf{x}}_1^{\mathsf{T}}\boldsymbol{\beta}_1 - \bar{\mathbf{x}}_2^{\mathsf{T}}\boldsymbol{\beta}_2) - X_1\boldsymbol{\beta}_1 - X_2\boldsymbol{\beta}_2\right] = 0 \\ X_2^{\mathsf{T}}\left[\mathbf{y} - \mathbf{1}(\bar{y} - \bar{\mathbf{x}}_1^{\mathsf{T}}\boldsymbol{\beta}_1 - \bar{\mathbf{x}}_2^{\mathsf{T}}\boldsymbol{\beta}_2) - X_1\boldsymbol{\beta}_1 - X_2\boldsymbol{\beta}_2\right] = 0 \end{cases}$$

整理得：

$$\begin{cases} X_1^{\mathsf{T}}(X_1 - \mathbf{1}\bar{\mathbf{x}}_1^{\mathsf{T}})\boldsymbol{\beta}_1 + X_1^{\mathsf{T}}(X_2 - \mathbf{1}\bar{\mathbf{x}}_2^{\mathsf{T}})\boldsymbol{\beta}_2 = X_1^{\mathsf{T}}(\mathbf{y} - \mathbf{1}\bar{y}) = (X_1 - \mathbf{1}\bar{\mathbf{x}}_1^{\mathsf{T}})^{\mathsf{T}}\mathbf{y} \\ X_2^{\mathsf{T}}(X_1 - \mathbf{1}\bar{\mathbf{x}}_1^{\mathsf{T}})\boldsymbol{\beta}_1 + X_2^{\mathsf{T}}(X_2 - \mathbf{1}\bar{\mathbf{x}}_2^{\mathsf{T}})\boldsymbol{\beta}_2 = X_2^{\mathsf{T}}(\mathbf{y} - \mathbf{1}\bar{y}) \end{cases}$$

记（中心化）$X_{1c}=X_1-\mathbf{1}\bar{\mathbf{x}}_1^{\mathsf{T}},X_{2c}=X_2-\mathbf{1}\bar{\mathbf{x}}_2^{\mathsf{T}},\mathbf{y}_c=\mathbf{y}-\mathbf{1}\bar{y}$,

前述方程组简写为

$$\begin{cases}X_{1c}^{\mathsf{T}}X_{1c}\boldsymbol{\beta}_1+X_{1c}^{\mathsf{T}}X_{2c}\boldsymbol{\beta}_2=X_{1c}^{\mathsf{T}}\mathbf{y}\\ X_{2c}^{\mathsf{T}}X_{1c}\boldsymbol{\beta}_1+X_{2c}^{\mathsf{T}}X_{2c}\boldsymbol{\beta}_2=X_{2c}^{\mathsf{T}}\mathbf{y}\end{cases}$$

第一式左乘$X_{2c}^{\mathsf{T}}X_{1c}\left(X_{1c}^{\mathsf{T}}X_{1c}\right)^{-1}$得：

$$X_{2c}^{\mathsf{T}}X_{1c}\boldsymbol{\beta}_1+X_{2c}^{\mathsf{T}}X_{1c}\left(X_{1c}^{\mathsf{T}}X_{1c}\right)^{-1}X_{1c}^{\mathsf{T}}X_{2c}\boldsymbol{\beta}_2=X_{2c}^{\mathsf{T}}X_{1c}\left(X_{1c}^{\mathsf{T}}X_{1c}\right)^{-1}X_{1c}^{\mathsf{T}}\mathbf{y}$$

即$X_{2c}^{\mathsf{T}}X_{1c}\boldsymbol{\beta}_1+X_{2c}^{\mathsf{T}}P_{X_{1c}}X_{2c}\boldsymbol{\beta}_2=X_{2c}^{\mathsf{T}}P_{X_{1c}}\mathbf{y}$，与第二式相减(消去$\boldsymbol{\beta}_1$)得：

$$X_{2c}^{\mathsf{T}}(I-P_{X_{1c}})X_{2c}\boldsymbol{\beta}_2=X_{2c}^{\mathsf{T}}(I-P_{X_{1c}})\mathbf{y},$$

$$\Leftrightarrow X_2^{\perp\mathsf{T}}X_2^{\perp}\boldsymbol{\beta}_2=X_2^{\perp\mathsf{T}}\mathbf{y}\Rightarrow\hat{\boldsymbol{\beta}}_2=\left(X_2^{\perp\mathsf{T}}X_2^{\perp}\right)^{-1}X_2^{\perp\mathsf{T}}\mathbf{y}$$

证明4：应用分块矩阵的逆（略）