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

第十一讲 投影和最小二乘法

2022.11.11

$$(A\mathbf{x})^{\mathrm{T}}\mathbf{y} = \mathbf{x}^{\mathrm{T}}(A^{\mathrm{T}}\mathbf{y})$$

欧氏空间中的正交投影

内积： $\mathbf{u},\mathbf{v}\in R^n,\ (\mathbf{u},\mathbf{v})=\mathbf{u}^{\mathrm{T}}\mathbf{v}=\sum u_i\,v_i$

模长： $\|\mathbf{u}\|=\sqrt{(\mathbf{u},\mathbf{u})}=\sqrt{\sum u_i^2}$

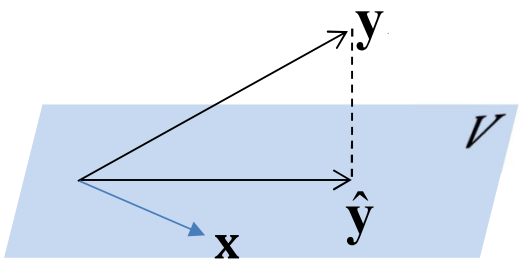
其它形式内积也可以，比如，$(\mathbf{u},\mathbf{v})=\mathbf{u}^{\mathrm{T}}G\mathbf{v},G>0$

投影

设$\mathbf{y}\neq\mathbf{0}\in R^n$,$\mathbf{y}$在线性子空间$V\subset R^n$空间上的正交投影$\hat{\mathbf{y}}$满足：

(1) $\hat{\mathbf{y}}\in V$

(2) $\mathbf{y}-\hat{\mathbf{y}}\perp V$,即对$\forall\mathbf{x}\in V$, $(\mathbf{x},\mathbf{y}-\hat{\mathbf{y}})=0\Leftrightarrow(\mathbf{x},\mathbf{y})=(\mathbf{x},\hat{\mathbf{y}})$.



记$\mathbf{y}^{\perp}=\mathbf{y}-\hat{\mathbf{y}}$,

正交分解： $\mathbf{y}=\hat{\mathbf{y}}+\mathbf{y}^{\perp},\ \hat{\mathbf{y}}\perp\mathbf{y}^{\perp}$;

平方和分解：$\|\mathbf{y}\|^2=\|\hat{\mathbf{y}}\|^2+\|\mathbf{y}^{\perp}\|^2$

投影参考：James H Stapleton (1995) Linear statistical models. Wiley
（下载： /books/5.pdf）

投影唯一

假设$\hat{\mathbf{y}}_1,\hat{\mathbf{y}}_2$都是$\mathbf{y}$的投影，$\mathbf{y}-\hat{\mathbf{y}}_i \perp V$,特别地$\mathbf{y}-\hat{\mathbf{y}}_i \perp \hat{\mathbf{y}}_1-\hat{\mathbf{y}}_2 \in V$

$$\Rightarrow (\hat{\mathbf{y}}_1-\hat{\mathbf{y}}_2,\hat{\mathbf{y}}_1-\hat{\mathbf{y}}_2)=(\mathbf{y}-\hat{\mathbf{y}}_2-(\mathbf{y}-\hat{\mathbf{y}}_1),\hat{\mathbf{y}}_1-\hat{\mathbf{y}}_2)=0 \Rightarrow \hat{\mathbf{y}}_1-\hat{\mathbf{y}}_2=0$$

投影是线性对称幂等矩阵变换

命题4. R^n中的投影是对称幂等线性变换，对应的矩阵是对称幂等矩阵（称为投影矩阵）。

证：假设$\mathbf{y}_1,\mathbf{y}_2$在子空间$V$上的投影为$\hat{\mathbf{y}}_1,\hat{\mathbf{y}}_2$,则容易验证$c_1\hat{\mathbf{y}}_1+c_2\hat{\mathbf{y}}_2$是$c_1\mathbf{y}_1+c_2\mathbf{y}_2$的投影。所以投影是$R^n \to R^n$的线性变换，
一定存在$n\times n$矩阵P_V使得$\hat{\mathbf{y}}=P_V\mathbf{y}$,$P_V$称为$V$的投影矩阵。

幂等：对任何$\mathbf{y}\in R^n$,$\hat{\mathbf{y}}=P\mathbf{y}\in V$,故$P\hat{\mathbf{y}}=\hat{\mathbf{y}}$即

$P\mathbf{y}=P^2\mathbf{y}$,所以$P^2=P$.

对称：对任何$\mathbf{x},\mathbf{y}\in R^n$,因为$P\mathbf{x}\in V$,所以$(P\mathbf{x},\mathbf{y})=(P\mathbf{x},P\mathbf{y})$,

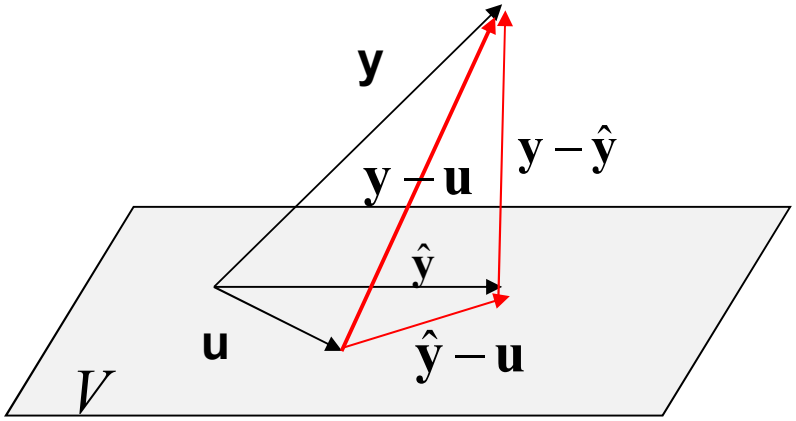
因为$P\mathbf{y}\in V$，所以$(P\mathbf{x},P\mathbf{y})=(\mathbf{x},P\mathbf{y})$,

所以$(P\mathbf{x},\mathbf{y})=(\mathbf{x},P\mathbf{y})$，$P^{\mathsf{T}}=P$。

投影的最优性：最小二乘

定理8（最小二乘）：$\mathbf{y} \in R^n$, 记$\hat{\mathbf{y}} = P(\mathbf{y} \mid V)$为$\mathbf{y}$在子空间$V \subset R^n$上的投影，则$\| \mathbf{y} - \hat{\mathbf{y}} \|^2 = \min_{\mathbf{u} \in V} \| \mathbf{y} - \mathbf{u} \|^2$

证：对任何$\mathbf{u} \in V$，因为$\hat{\mathbf{y}} - \mathbf{u} \in V$，故$\mathbf{y} - \hat{\mathbf{y}} \perp \hat{\mathbf{y}} - \mathbf{u}$.



所以$\| \mathbf{y} - \mathbf{u} \|^2 = \| \mathbf{y} - \hat{\mathbf{y}} + \hat{\mathbf{y}} - \mathbf{u} \|^2 = \| \mathbf{y} - \hat{\mathbf{y}} \|^2 + \| \hat{\mathbf{y}} - \mathbf{u} \|^2 \geq \| \mathbf{y} - \hat{\mathbf{y}} \|^2$.

投影矩阵

投影矩阵

定理9. 设$V=C(A_{n\times m})\subset R^n$，$\mathbf{y}\in R^n$，则$\mathbf{y}$在$V$上的投影为

$$\hat{\mathbf{y}}=P_A\mathbf{y}=A(A^{\mathsf{T}}A)^{-}A^{\mathsf{T}}\mathbf{y},$$

其中A对应的投影矩阵为$P_A=A(A^{\mathsf{T}}A)^{-}A^{\mathsf{T}}$ (n阶方阵).

证明：因为$\hat{\mathbf{y}}\in V$,存在$\boldsymbol{\beta}_{m\times 1}$,使得$\hat{\mathbf{y}}=A\boldsymbol{\beta}$。

因为$\mathbf{y}^{\perp}=\mathbf{y}-\hat{\mathbf{y}}\perp C(A)$,特别地，$\mathbf{y}^{\perp}$与$A$的每列都正交

$$\Rightarrow \quad A^{\mathsf{T}}(\mathbf{y}-\hat{\mathbf{y}})=A^{\mathsf{T}}(\mathbf{y}-A\boldsymbol{\beta})=\mathbf{0}\Leftrightarrow \quad A^{\mathsf{T}}A\boldsymbol{\beta}=A^{\mathsf{T}}\mathbf{y}$$

$$\Rightarrow\hat{\boldsymbol{\beta}}=(A^{\mathsf{T}}A)^{-}A^{\mathsf{T}}\mathbf{y}\Rightarrow\hat{\mathbf{y}}=A(A^{\mathsf{T}}A)^{-}A^{\mathsf{T}}\mathbf{y},P_A=A(A^{\mathsf{T}}A)^{-}A^{\mathsf{T}}$$

证明：因为$\mathbf{y}^{\perp}=\mathbf{y}-\hat{\mathbf{y}}=\mathbf{y}-A\boldsymbol{\beta}\perp C(A)$,所以对任何$\mathbf{x}\in R^m$，$\mathbf{u}=A\mathbf{x}\in C(A)$,

$$0=(\mathbf{u},\mathbf{y}^{\perp})=(A\mathbf{x},\mathbf{y}^{\perp})\overset{\text{adjoint}}{=}\left(\mathbf{x},A^{\mathsf{T}}\mathbf{y}^{\perp}\right)$$

即$A^{\mathsf{T}}\mathbf{y}^{\perp}\in R^m$与任何$\mathbf{x}\in R^m$都正交，所以$A^{\mathsf{T}}\mathbf{y}^{\perp}=\mathbf{0}$，即正则方程：

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

因为$A^{\mathsf{T}}\mathbf{y}\in C(A^{\mathsf{T}})=C(A^{\mathsf{T}}A)$,方程有解,所有解$\hat{\boldsymbol{\beta}}=(A^{\mathsf{T}}A)^{-}A^{\mathsf{T}}\mathbf{y}$

$\Rightarrow\hat{\mathbf{y}}=A\hat{\boldsymbol{\beta}}=A(A^{\mathsf{T}}A)^{-}A^{\mathsf{T}}\mathbf{y}$, 所以$P_A=A(A^{\mathsf{T}}A)^{-}A^{\mathsf{T}}$.

例1. 设$\mathbf{v} \neq \mathbf{0} \in R^n$, $\mathbf{v}$张成的空间$V = C(\mathbf{v}) = \{\mathbf{v}c \mid c \in R\}$，投影矩阵

$$P_{\mathbf{v}} = \mathbf{v}(\mathbf{v}^{\mathsf{T}}\mathbf{v})^{-1}\mathbf{v}^{\mathsf{T}}$$

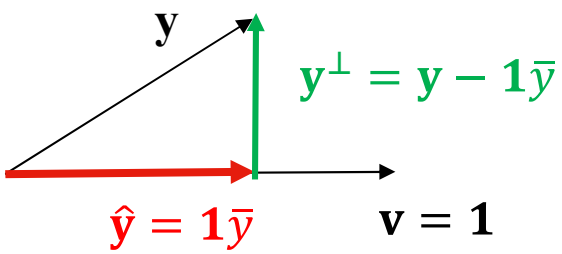
$\mathbf{y} \in R^n$在V上的投影 $\hat{\mathbf{y}} = P_{\mathbf{v}}\mathbf{y} = \mathbf{v}(\mathbf{v}^{\mathsf{T}}\mathbf{v})^{-1}\mathbf{v}^{\mathsf{T}}\mathbf{y} = \mathbf{v}\lambda, \lambda = \mathbf{v}^{\mathsf{T}}\mathbf{y} / \mathbf{v}^{\mathsf{T}}\mathbf{v}$。

特别地，如果$\mathbf{v} = \mathbf{1} = (1,...,1)^{\mathsf{T}} \in R^n$, $P_{\mathbf{1}} = \mathbf{1}(\mathbf{1}^{\mathsf{T}}\mathbf{1})^{-1}\mathbf{1}^{\mathsf{T}} = \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathsf{T}}$,

$$\hat{\mathbf{y}} = P_{\mathbf{1}}\mathbf{y} = \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathsf{T}}\mathbf{y} = \mathbf{1}\bar{y},$$

$$\mathbf{y}^{\perp} = \mathbf{y} - P_{\mathbf{1}}\mathbf{y} = \mathbf{y} - \mathbf{1}\bar{y} = (y_1 - \bar{y},..., y_n - \bar{y})^{\mathsf{T}},$$

$\mathbf{y}^{\perp}$称为$\mathbf{y}$的中心化向量。



中心化向量

$$\mathbf{y}^{\perp} = (I_n - P_{\mathbf{1}})\mathbf{y} = P_{\mathbf{1}^{\perp}}\mathbf{y}$$

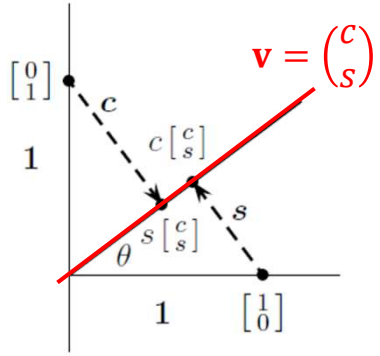
为$\mathbf{y}$在$\mathbf{1}$的正交补空间上的投影

例2.$(n=2)$对任何$\mathbf{v}=(c,s)^{\mathsf{T}}\in R^2,\|\mathbf{v}\|=1$，$C(\mathbf{v})$对应的投影阵

$$P_{\mathbf{v}}=\mathbf{v}(\mathbf{v}^{\mathsf{T}}\mathbf{v})^{-1}\mathbf{v}^{\mathsf{T}}=\mathbf{v}\mathbf{v}^{\mathsf{T}}=\begin{pmatrix}c^2 & cs\\ cs & s^2\end{pmatrix},$$

作为变换，它将$\mathbf{e}_1=\begin{pmatrix}1\\0\end{pmatrix}\to\begin{pmatrix}c^2\\cs\end{pmatrix}=c\mathbf{v}$，$\mathbf{e}_2=\begin{pmatrix}0\\1\end{pmatrix}\to\begin{pmatrix}cs\\s^2\end{pmatrix}=s\mathbf{v}.$

任何$\mathbf{x}=(x_1,x_2)^{\mathsf{T}}$的投影：$P_{\mathbf{v}}\mathbf{x}=\mathbf{v}\mathbf{v}^{\mathsf{T}}\mathbf{x}=\mathbf{v}\left(\mathbf{v}^{\mathsf{T}}\mathbf{x}\right)=\mathbf{v}(cx_1+sx_2).$



当$\mathbf{v}=\mathbf{e}_1=\begin{pmatrix}1\\0\end{pmatrix}$时，$P_{\mathbf{v}}=\begin{pmatrix}1 & 0\\0 & 0\end{pmatrix}$。对任何$\mathbf{x}=\begin{pmatrix}x_1\\x_2\end{pmatrix}$，$P_{\mathbf{v}}\mathbf{x}=\begin{pmatrix}x_1\\0\end{pmatrix}$.

投影矩阵的性质

命题5. 设A是$n\times m$矩阵，则

(1) $P_A = A(AA^{\mathsf{T}})^{-}A^{\mathsf{T}}$唯一，与广义逆$(AA^{\mathsf{T}})^{-}$的具体选择无关。

(2) 投影矩阵等价于对称幂等阵：$P_A = A(A^{\mathsf{T}}A)^{-}A^{\mathsf{T}}$是对称幂等矩阵，
反之，任一对称幂等阵是投影阵。

(3) P_A是$C(A)$对应的投影阵，$I_n - P_A$是$C(A)^{\perp}$对应的投影阵，两者正交：

$$P_A(I_n - P_A) = 0$$

(4) 线性子空间V对应的投影矩阵与空间的基的选取无关。

(5) 按列划分$A=(A_1,A_2)$,若$A_1^{\mathsf{T}}A_2 = 0$,则 $P_A = P_{A_1} + P_{A_2}$.

证明1: (1) 因为投影唯一，投影阵P_A必定唯一,与广义逆$(A^{\mathsf{T}}A)^{-}$的选择无关。

(2) 我们已知投影变换是对称幂等的，所以投影矩阵是对称幂等的。

反之，设P是$n\times n$对称幂等矩阵，$r = rank(P)$,则有谱分解$P = O\begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}O^{\mathsf{T}}$,

其中$OO^{\mathsf{T}} = O^{\mathsf{T}}O = I_n$,划分$O=(O_1,O_2)$,则$O_1^{\mathsf{T}}O_1 = I_r$, $P = O_1O_1^{\mathsf{T}} = O_1\left(O_1^{\mathsf{T}}O_1\right)^{-1}O_1^{\mathsf{T}}$,
所以P具有投影矩阵的形式，所以任何对称幂等矩阵是投影阵。

(4)、(5)：作业。

投影阵$P_A = A(A^{\mathsf{T}}A)^{-}A^{\mathsf{T}}$与广义逆的选择无关这一事实并不显然(除非$A^{\mathsf{T}}A$可逆)，下面我们利用投影阵的表达给出证明。

证明2（1）因为$C(A^{\mathsf{T}}) = C(A^{\mathsf{T}}A)$,所以存在$B$使得$A^{\mathsf{T}} = A^{\mathsf{T}}AB$,

所以$P_A = A(A^{\mathsf{T}}A)^{-}A^{\mathsf{T}} = B^{\mathsf{T}}A^{\mathsf{T}}A(A^{\mathsf{T}}A)^{-}A^{\mathsf{T}}AB = B^{\mathsf{T}}A^{\mathsf{T}}AB$与$(A^{\mathsf{T}}A)^{-}$无关，

且$P_A = B^{\mathsf{T}}A^{\mathsf{T}}AB$是对称的。

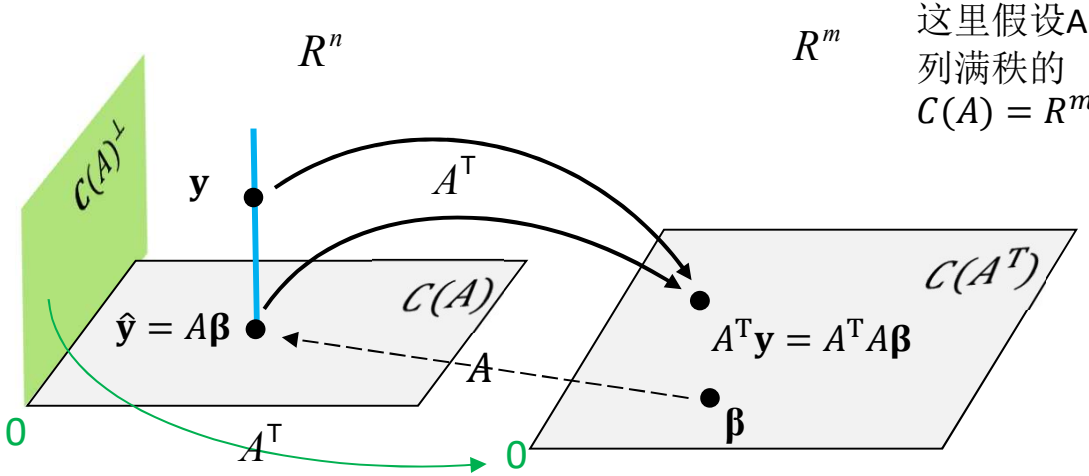
假设P_A, P都是$C(A)$的投影阵，因为$C(P_A) \subset C(A), C(P) \subset C(A)$
所以$PP_A = P_A, P_A = P_A P = P.$

（2）P_A对称幂等：

$$P_A^{\,2} = A(A^{\mathsf{T}}A)^{-}A^{\mathsf{T}}A(A^{\mathsf{T}}A)^{-}A^{\mathsf{T}} = A(A^{\mathsf{T}}A)^{-}A^{\mathsf{T}}A(A^{\mathsf{T}}A)^{-}A^{\mathsf{T}}AB$$

$$= A(A^{\mathsf{T}}A)^{-}A^{\mathsf{T}}AB = A(A^{\mathsf{T}}A)^{-}A^{\mathsf{T}} = P_A。$$

练习：证明 $A(A^{\mathsf{T}}A)^{-}A^{\mathsf{T}}A = A.$



这里假设A是列满秩的 $C(A) = R^m$.

$A^\mathsf{T}\mathbf{y}^\perp = \mathbf{0}$:

蓝线$\mathbf{y} + C(A)^\perp$上的点在$C(A)$上的投影都相同

$\Leftrightarrow$ A^T将它们映射为R^m中的同一点$A^\mathsf{T}\hat{\mathbf{y}}$。

投影矩阵的谱分解表达

命题6. 若$A_{n\times m}$有奇异值分解$A=U_{n\times r}D_{r\times r}(V_{m\times r})^{\mathsf{T}}$,其中$U^{\mathsf{T}}U=V^{\mathsf{T}}V=I_r$,
则$P_A=A(A^{\mathsf{T}}A)^{-}A^{\mathsf{T}}=UU^{\mathsf{T}}$, $P_{A^{\mathsf{T}}}=VV^{\mathsf{T}}$.

证明1: $C(U)=C(A)$,所以$P_A=P_U=U(U^{\mathsf{T}}U)^{-1}U^{\mathsf{T}}=UU^{\mathsf{T}}$.

$A=UDV^{\mathsf{T}}\Rightarrow C(A)\subset C(U)$
$U=AVD^{-1}\Rightarrow C(U)\subset C(A)$

证明2: $A=U_{n\times r}D_{r\times r}(V_{m\times r})^{\mathsf{T}}, U^{\mathsf{T}}U=V^{\mathsf{T}}V=I_r$,

将U,V扩展为正交方阵$\tilde{U}=(U,*)$, $\tilde{V}=(V,*)$, $A=UDV^{\mathsf{T}}=\tilde{U}\begin{pmatrix}D & 0\\ 0 & 0\end{pmatrix}\tilde{V}^{\mathsf{T}}$,

则$A^{\mathsf{T}}A=\tilde{V}\begin{pmatrix}D^2 & 0\\ 0 & 0\end{pmatrix}\tilde{V}^{\mathsf{T}}\Rightarrow (A^{\mathsf{T}}A)^{-}=\tilde{V}\begin{pmatrix}D^{-2} & *\\ * & *\end{pmatrix}\tilde{V}^{\mathsf{T}}$

$$\Rightarrow P_A=A(A^{\mathsf{T}}A)^{-}A^{\mathsf{T}}=\tilde{U}\begin{pmatrix}D & 0\\ 0 & 0\end{pmatrix}\tilde{V}^{\mathsf{T}}\tilde{V}\begin{pmatrix}D^{-2} & *\\ * & *\end{pmatrix}\tilde{V}^{\mathsf{T}}\tilde{V}\begin{pmatrix}D & 0\\ 0 & 0\end{pmatrix}\tilde{U}^{\mathsf{T}}$$

$$=\tilde{U}\begin{pmatrix}D & 0\\ 0 & 0\end{pmatrix}\begin{pmatrix}D^{-2} & *\\ * & *\end{pmatrix}\begin{pmatrix}D & 0\\ 0 & 0\end{pmatrix}\tilde{U}^{\mathsf{T}}=\tilde{U}\begin{pmatrix}I_r & 0\\ 0 & 0\end{pmatrix}\tilde{U}^{\mathsf{T}}=UU^{\mathsf{T}}.$$

注1：命题6提供了投影阵更容易理 解的形式：

设U的各列为$U=(\mathbf{u}_1,...,\mathbf{u}_r)$,则任何$\mathbf{y}\in R^n$在$C(A)$上的投影

$$P_A\mathbf{y}=UU^{\mathsf{T}}\mathbf{y}=\mathbf{u}_1(\mathbf{u}_1{}^{\mathsf{T}}\mathbf{y})+...+\mathbf{u}_r(\mathbf{u}_r{}^{\mathsf{T}}\mathbf{y})$$

正是我们对正交投影的 理解方式： $\mathbf{u}_1,...,\mathbf{u}_r$为基（坐标轴），

$\mathbf{u}_1{}^{\mathsf{T}}\mathbf{y},...,\mathbf{u}_r{}^{\mathsf{T}}\mathbf{y}$ 为$\mathbf{y}$在这些坐标轴上的投影 坐标。

注2：由命题5(3）任何投影矩阵或对称幂等矩阵都可表示成

$$P=O_1O_1{}^{\mathsf{T}},$$

其中$O_1{}^{\mathsf{T}}O_1=I_r$, $r=\mathrm{rank}(A)$ (但O_1未必等于U).

多变量线性回归模型(multiple linear regression)

多变量回归模型

假设 $(y_i, x_{i1}, \ldots, x_{i,p-1})$，$i = 1,2\ldots,n$ 独立，满足多变量/多重线性回归模型：

$$y_i = \beta_0 + x_{i1}\beta_1 + \ldots + x_{i,p-1}\beta_{p-1} + \varepsilon_i ,$$

或

$$y_i = \beta_0 + \mathbf{x}_i^{\mathsf{T}}\mathbf{b} + \varepsilon_i \overset{\Delta}{=} \tilde{\mathbf{x}}_i^{\mathsf{T}}\boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \text{ iid} \sim (0, \sigma^2), \quad \varepsilon_i \text{与} \mathbf{x}_i \text{独立.} \qquad (1)$$

其中 $\mathbf{x}_i = (x_{i1}, \ldots x_{i,p-1})^{\mathsf{T}}$ 为自变量的第 i 此观测，β_0 为截距项，$\mathbf{b} = (\beta_1, \ldots \beta_{p-1})^{\mathsf{T}}$ 为回归系数。$\tilde{\mathbf{x}}_i = \begin{pmatrix} 1 \\ \mathbf{x}_i \end{pmatrix}$，$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \mathbf{b} \end{pmatrix}$，设计阵(design matrix)：

$$X_{n\times p} = \begin{pmatrix} \tilde{\mathbf{x}}_1^{\mathsf{T}} \\ \tilde{\mathbf{x}}_2^{\mathsf{T}} \\ \vdots \\ \tilde{\mathbf{x}}_n^{\mathsf{T}} \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{x}_1^{\mathsf{T}} \\ 1 & \mathbf{x}_2^{\mathsf{T}} \\ \vdots & \vdots \\ 1 & \mathbf{x}_n^{\mathsf{T}} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{pmatrix} = \left(\mathbf{1}, \mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(p-1)}\right)$$

矩阵-向量形式

记 $\mathbf{y}=(y_1,...,y_n)^{\mathsf{T}}$，$\boldsymbol{\varepsilon}=(\varepsilon_1,...,\varepsilon_n)^{\mathsf{T}}$，线性回归模型（1）简记为：

$$\mathbf{y}_{n\times 1}=X_{n\times p}\boldsymbol{\beta}_{p\times 1}+\boldsymbol{\varepsilon}_{n\times 1}=\mathbf{1}\beta_0+\mathbf{x}_{(1)}\beta_1+...+\mathbf{x}_{(p-1)}\beta_{p-1}+\boldsymbol{\varepsilon} \qquad (2)$$

其中 $\boldsymbol{\varepsilon}\sim(\mathbf{0},\sigma^2 I_n)$，$\boldsymbol{\varepsilon}$ 与 X 独立

注：模型也可写作：$\mathbf{y}_{n\times 1}=\boldsymbol{\mu}_{n\times 1}+\boldsymbol{\varepsilon}_{n\times 1}$，$\boldsymbol{\mu}\in C(X)$

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

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

记$\boldsymbol{\beta}=\begin{pmatrix}a\\b\end{pmatrix}$, $\mathbf{y}=(y_1,...,y_n)^{\mathsf{T}}$, 设计阵

$$X=\begin{pmatrix}1 & x_1\\ \vdots & \vdots\\ 1 & x_n\end{pmatrix}=(\mathbf{1},\mathbf{x})$$

模型为

$$\mathbf{y}=X\boldsymbol{\beta}+\boldsymbol{\varepsilon}=\mathbf{1}a+\mathbf{x}b+\boldsymbol{\varepsilon}$$

最小二乘法 (LS: Least Squares)

目标：基于数据$\mathbf{y}$，X，估计参数$\boldsymbol{\beta}$以及σ^2，并研究其性质。

最小二乘法

为了估计参数$\boldsymbol{\beta}$，最小二乘法最小化误差平方和：

$$\min_{\boldsymbol{\beta}\in R^p}\sum \varepsilon_i^{\,2}=\min_{\boldsymbol{\beta}\in R^p}\|\boldsymbol{\varepsilon}\|^2=\min_{\boldsymbol{\beta}\in R^p}\|\mathbf{y}-X\boldsymbol{\beta}\|^2$$

$$=\min_{\boldsymbol{\beta}\in R^p}\sum\left(y_i-\beta_0-\mathbf{x}_i^{\mathsf{T}}\mathbf{b}\right)^2=\min_{\boldsymbol{\beta}\in R^p}\sum\left(y_i-\tilde{\mathbf{x}}_i^{\mathsf{T}}\boldsymbol{\beta}\right)^2$$

使得平方和达到最小值的$\hat{\boldsymbol{\beta}}$称为最小二乘(LS)估计。

定理1. 假设线性回归模型：

$$\mathbf{y}_{n\times 1}=X_{n\times p}\boldsymbol{\beta}_{p\times 1}+\boldsymbol{\varepsilon}_{n\times 1},\quad \boldsymbol{\varepsilon}\sim(\mathbf{0},\sigma^2 I_n),\quad \boldsymbol{\varepsilon}\text{与}X\text{独立。}$$

最小二乘估计$\hat{\boldsymbol{\beta}}$满足正则方程

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

若X列满秩($p\le n$)，正则方程有唯一解

$$\hat{\boldsymbol{\beta}}=(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y}.$$

证法1：投影

证：由第10讲定理8(投影的最小二乘性质，P26)，

$$\min_{\boldsymbol{\beta}\in R^p} \|\mathbf{y}-X\boldsymbol{\beta}\|^2 = \min_{\mathbf{u}\in C(X)} \|\mathbf{y}-\mathbf{u}\|^2 = \|\mathbf{y}-\hat{\mathbf{y}}\|^2$$

其中$\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}}$为$\mathbf{y}$在$C(X)$上的投影，即$\hat{\mathbf{y}} = P_X\mathbf{y} = X\underline{(X^{\mathsf{T}}X)^{-}X^{\mathsf{T}}\mathbf{y}}$,

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

特别地若X列满秩，LS估计$\hat{\boldsymbol{\beta}} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y}$。

$\mathbf{y}-\hat{\mathbf{y}} \perp C(X)$,特别地与$X$的每一列正交，即正则方程：

$$X^{\mathsf{T}}(\mathbf{y}-\hat{\mathbf{y}}) = X^{\mathsf{T}}(\mathbf{y}-X\hat{\boldsymbol{\beta}}) = \mathbf{0}$$

引理. (矩阵向量求导)

(1) 若$\mathbf{a}, \mathbf{x} \in R^n$, 则$\dfrac{\partial(\mathbf{a}^{\mathsf{T}}\mathbf{x})}{\partial \mathbf{x}} = \mathbf{a}$;

(2) 若A为$n \times n$对称矩阵, $\mathbf{x} \in R^n$, 则$\dfrac{\partial(\mathbf{x}^{\mathsf{T}} A\mathbf{x})}{\partial \mathbf{x}} = 2A\mathbf{x}$

更多矩阵微商，参见王松桂《线性模型引论》

证法2：矩阵向量求导

目标函数$Q(\boldsymbol{\beta}) = \| \mathbf{y} - X\boldsymbol{\beta} \|^2 = \mathbf{y}^{\mathsf{T}}\mathbf{y} - 2\mathbf{y}^{\mathsf{T}}X\boldsymbol{\beta} + \boldsymbol{\beta}^{\mathsf{T}}X^{\mathsf{T}}X\boldsymbol{\beta}$,

令$\dfrac{\partial Q}{\partial \boldsymbol{\beta}} = -2(\mathbf{y}^{\mathsf{T}}X)^{\mathsf{T}} + 2X^{\mathsf{T}}X\boldsymbol{\beta} = -2X^{\mathsf{T}}(\mathbf{y} - X\boldsymbol{\beta}) = 0$

得正则方程

$$X^{\mathsf{T}}(\mathbf{y} - X\boldsymbol{\beta}) = 0 \qquad X^{\mathsf{T}}X\boldsymbol{\beta} = X^{\mathsf{T}}\mathbf{y}$$

$$\Rightarrow \text{LS估计}\hat{\boldsymbol{\beta}} = (X^{\mathsf{T}}X)^{-}X^{\mathsf{T}}\mathbf{y} \quad = \left(\sum_{i=1}^{n} \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^{\mathsf{T}}\right)^{-} \left(\sum_{i=1}^{n} \tilde{\mathbf{x}}_i y_i\right)$$

$\Leftrightarrow$

证法3：和函数求导

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

$$\text{令} \frac{\partial Q}{\partial \boldsymbol{\beta}} = -2\sum_{i=1}^{n} \tilde{\mathbf{x}}_i \left(y_i - \tilde{\mathbf{x}}_i^{\mathsf{T}}\boldsymbol{\beta}\right) = 0$$

得正则方程：

$$\sum_{i=1}^{n} \tilde{\mathbf{x}}_i \left(y_i - \tilde{\mathbf{x}}_i^{\mathsf{T}}\boldsymbol{\beta}\right) = 0 \Leftrightarrow \sum_{i=1}^{n} \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^{\mathsf{T}}\boldsymbol{\beta} = \sum_{i=1}^{n} \tilde{\mathbf{x}}_i y_i$$

$$\Rightarrow \text{LS估计}\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^{\mathsf{T}}\right)^{-} \left(\sum_{i=1}^{n} \tilde{\mathbf{x}}_i y_i\right)$$

几个注解

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

LS与矩估计方法

$\mathbf{y}=X\boldsymbol{\beta}+\boldsymbol{\varepsilon}$，$y_i=\tilde{\mathbf{x}}_i^{\mathsf{T}}\boldsymbol{\beta}+\varepsilon_i$，$\varepsilon_i$ 与$\tilde{\mathbf{x}}_i$独立，$E(\tilde{\mathbf{x}}_i\varepsilon_i)=0,$

令相应的样本矩等于0：$0=\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{x}}_i\varepsilon_i=\frac{1}{n}X^{\mathsf{T}}\boldsymbol{\varepsilon}$，即正则方程

$$X^{\mathsf{T}}\boldsymbol{\varepsilon}=X^{\mathsf{T}}\mathbf{y}-X^{\mathsf{T}}X\boldsymbol{\beta}=0\Rightarrow\hat{\boldsymbol{\beta}}=(X^{\mathsf{T}}X)^{-}X^{\mathsf{T}}\mathbf{y}$$

伴随变换

更简单地，矩估计方法实际上在$\mathbf{y}=X\boldsymbol{\beta}+\boldsymbol{\varepsilon}$两边同时左乘$X^{\mathsf{T}}$

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

并令$X^{\mathsf{T}}\boldsymbol{\varepsilon}=0$即可得到正则方程$X^{\mathsf{T}}\mathbf{y}=X^{\mathsf{T}}X\boldsymbol{\beta}$。

工具变量

若$\boldsymbol{\varepsilon}$与X不独立（内生），但存在工具变量$Z_{n\times m}$与自变量相关

但与误差独立，即$E(Z^{\mathsf{T}}\boldsymbol{\varepsilon})=0,$则模型两边同时左乘$Z^{\mathsf{T}}$：

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

并令$Z^{\mathsf{T}}\boldsymbol{\varepsilon}=0$（矩方法）,即得$Z^{\mathsf{T}}\mathbf{y}=Z^{\mathsf{T}}X\hat{\boldsymbol{\beta}}_{\text{IVLS}}\Rightarrow\hat{\boldsymbol{\beta}}_{\text{IVLS}}=\left(Z^{\mathsf{T}}X\right)^{-1}Z^{\mathsf{T}}\mathbf{y}$

超定方程

$\mathbf{y}_{n\times 1} = X_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}$可看作是解超定方程(over-determined system):

$$\mathbf{y}_{n\times 1} = X_{n\times p}\boldsymbol{\beta}_{p\times 1},\quad n > p$$

只有当$\mathbf{y} \in C(X)$时才有解；否则可用如下方式求近似解：

最小二乘

- 求最优近似解使误差 $\|\mathbf{y} - X\boldsymbol{\beta}\|^2$ 最小,即LS。

Pre-conditioning

- 转化为适定方程：两别同乘一个矩阵$A_{p\times n}$ (AX可逆)，得:

$$A_{p\times n}\mathbf{y}_{n\times 1} = A_{p\times n}X_{n\times p}\boldsymbol{\beta}_{p\times 1} \quad \Rightarrow \tilde{\boldsymbol{\beta}} = (AX)^{-1}A\mathbf{y}$$

(当方程个数$n > p$时，适当合并方程以减少方程个数)

特别地，同乘 $A = X^{\mathsf{T}}$得到正则方程：$X^{\mathsf{T}}\mathbf{y} = X^{\mathsf{T}}X\boldsymbol{\beta}$.

同乘一个外生变量矩阵Z(Z与ε独立）得工具变量最小二乘估计。

欠定定程

$n < p$时，$\mathbf{y}_{n\times 1} = X_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}$可看作是解欠定方程(underdetermined system)

$$\mathbf{y}_{n\times 1} = X_{n\times p}\boldsymbol{\beta}_{p\times 1}, \quad n < p,$$

通常有无穷多解，为了求出有意义的解，通常对解施加某些限制，比如

不定方程

- 数论中限制线性方程的解为正整数或有理数；

模长约束

- 主成分回归：极小化欧氏模长 $\|\boldsymbol{\beta}\|_2$，$\tilde{\boldsymbol{\beta}} = X^{+}\mathbf{y}$;
- 压缩感知或lasso：极小化 $\|\boldsymbol{\beta}\|_0$ (非0的个数)或其放松 $\|\boldsymbol{\beta}\|_1$。