回归分析 (01714601)
课程主页：http://staff.ustc.edu.cn/~ynyang/lm2020

第十一讲 多重线性回归模型的最小二乘法

2020.3.25

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

多重线性回归模型

<u>多重线性回归模型(总体):</u>

$$y = \mathbf{x}^{\mathsf{T}}\boldsymbol{\beta} + \varepsilon,\quad \varepsilon \sim (0, \sigma^2), \varepsilon\text{与}\mathbf{x}\text{独立。}$$

其中y为响应变量,$\mathbf{x}$为$p \times 1$向量，$\boldsymbol{\beta}_{p\times 1} = (\beta_0, \beta_1, \ldots, \beta_{p-1})^{\mathsf{T}}$为回归系数.

通常$\mathbf{x} = (x_0, x_1, \ldots, x_{p-1})^{\mathsf{T}}$的第一个元素为常数$x_0 = 1$，对应于截距项$\beta_0$。

有时也采用如下模型假设

多重线性模型(总体模型）

均值线性: $\mathrm{E}(y \mid \mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$,　方差齐性: $\mathrm{var}(y \mid \mathbf{x}) = \sigma^2$

数据：假设 $(y_i, \mathbf{x}_i)$，$i=1,2\ldots,n$ 独立，来自于线性回归模型：

$$y_i = \mathbf{x}_i^{\top}\boldsymbol{\beta} + \varepsilon_i,\quad \varepsilon_i \text{ iid} \sim (0,\sigma^2),\quad \varepsilon_i 与 \mathbf{x}_i 独立,\ i=1,2\ldots,n.$$

记 $\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$, $\quad X = \begin{pmatrix} \mathbf{x}_1^{\top} \\ \vdots \\ \mathbf{x}_n^{\top} \end{pmatrix}$, $\quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{pmatrix}$, $\quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$,

其中 X 称为设计阵 (第 i 行为 $\mathbf{x}_i^{\top}$), 上述模型写成矩阵-向量形式：

线性回归模型：

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

预备知识：矩阵/向量函数的导数

矩阵函数的导数：

设$X=(x_{ij})_{1\le i\le n,\ 1\le j\le m}$为$n\times m$矩阵，$y=f(X)$,定义 $\dfrac{\partial y}{\partial X}=\left(\dfrac{\partial y}{\partial x_{ij}}\right)_{1\le i\le n,\ 1\le j\le m}$.

特别地，若$\mathbf{x}=(x_1,\ldots,x_p)^{\top}$为$p\times 1$向量，定义 $\dfrac{\partial y}{\partial \mathbf{x}}=\begin{pmatrix}\dfrac{\partial y}{\partial x_1}\\ \vdots\\ \dfrac{\partial y}{\partial x_p}\end{pmatrix}$

特别地

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

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

最小二乘法

最小二乘法(Least Squares, LS):

假设X列满秩，为了估计参数$\boldsymbol{\beta}$，最小化误差平方和：

$$\min_{\boldsymbol{\beta}\in R^p}\sum \varepsilon_i^{\ 2}=\min_{\boldsymbol{\beta}\in R^p}\sum\left(y_i-\mathbf{x}_i^{\top}\boldsymbol{\beta}\right)^2 \qquad (1)$$

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

$$=\min_{\mathbf{u}\in L(\mathrm{X})}\|\mathbf{y}-\mathbf{u}\|^2 \qquad (3)$$

为了求解，我们可以对(1)或(2)式求导，

也可以利用(3)应用几何投影方法。

解法1: 对(1)或(2)式求导

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

则$\dfrac{\partial Q}{\partial \boldsymbol{\beta}} = -2(\mathbf{y}^\top X)^\top + 2X^\top X\boldsymbol{\beta} = -2X^\top(\mathbf{y} - X\boldsymbol{\beta})$，令$\dfrac{\partial Q}{\partial \boldsymbol{\beta}} = 0,$得

正则方程:

$$X^\top(\mathbf{y} - X\boldsymbol{\beta}) = 0 \quad \Leftrightarrow \quad \sum_{i=1}^{n} \mathbf{x}_i \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta} \right) = 0$$

$$X^\top X\boldsymbol{\beta} = X^\top \mathbf{y} \quad \Leftrightarrow \quad \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top \boldsymbol{\beta} = \sum_{i=1}^{n} \mathbf{x}_i y_i$$

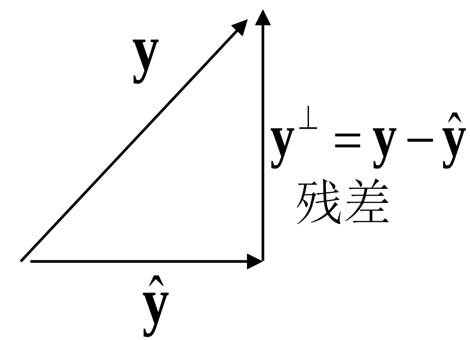
$$\Rightarrow \text{LS估计：} \hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \mathbf{y} = \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \sum_{i=1}^{n} \mathbf{x}_i y_i$$

解法2：投影

由第十讲命题2， $\min_{\mathbf{u}\in L(X)} \|\mathbf{y}-\mathbf{u}\|^2 = \|\mathbf{y}-\hat{\mathbf{y}}\|^2$

其中$\hat{\mathbf{y}} = P_X \mathbf{y}$为$\mathbf{y}$在$L(X)$上的投影，满足正交条件：

$$\boxed{\begin{array}{l}\text{投影的正交条件}\mathbf{y}^{\perp} = \mathbf{y}-\hat{\mathbf{y}} \perp L(X)\text{即为正则方程：}\\ X^{\top}(\mathbf{y}-\hat{\mathbf{y}}) = X^{\top}(\mathbf{y}-X\hat{\boldsymbol{\beta}}) = 0\end{array}}$$



投影/拟合值

投影$\hat{\mathbf{y}}$决定了$\hat{\boldsymbol{\beta}}$：

$$\hat{\mathbf{y}} = P_X\mathbf{y} = X(X^{\top}X)^{-1}X^{\top}\mathbf{y} = X\left\{(X^{\top}X)^{-1}X^{\top}\mathbf{y}\right\} = X\hat{\boldsymbol{\beta}}$$

$\Rightarrow$ LS估计：$\hat{\boldsymbol{\beta}} = (X^{\top}X)^{-1}X^{\top}\mathbf{y}$

关于最小二乘法的几个注解

1. 正则方程为

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

即最优解使得$\boldsymbol{\varepsilon}$与X的各列正交。这也是投影满足的正交条件。

以投影观点看待最小二乘法给出了几何直观。

2. 正则方程 $X^{\top}\boldsymbol{\varepsilon}/n=0$可看作是$\mathrm{E}(X^{\top}\boldsymbol{\varepsilon})=0$的样本版本(矩估计方法)

简单地看，模型$\mathbf{y}=X\boldsymbol{\beta}+\boldsymbol{\varepsilon}$两边左乘$X^{\top}$:

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

并令$X^{\top}\boldsymbol{\varepsilon}=0$即得正则方程。

3. 超定方程(overdetermined system):

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

只有当$\mathbf{y} \in L(X)$时才有解；否则我们如下求近似解：

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

- Pre-conditioning : 两别同乘一个矩阵$A_{p\times n}$ (AX可逆)，得适定方程 :

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

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

经典的线性模型$\mathbf{y}_{n\times 1} = X_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$通常要求$n > p$，但$n < p$的情形近些年开始受到重视.

4. 欠定方程(underdetermined system)

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

有无穷多解，为了求出有意义的解，通常需要施加某些约束，比如

- 数论不定方程问题：限制β为正整数或有理数；
- 压缩感知：限制$\boldsymbol{\beta}$稀疏，即$\boldsymbol{\beta}$的非0的分量个数最少。

拟合值与残差

拟合值/投影: $\hat{\mathbf{y}} = P_X\mathbf{y} = X\hat{\boldsymbol{\beta}} = X(X^\top X)^{-1}X^\top\mathbf{y}$

残差: $\mathbf{e} = \mathbf{y}^\perp = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - X\hat{\boldsymbol{\beta}} = (I_n - P_X)\mathbf{y}$

正则方程: $X^\top\mathbf{e} = 0$, 即$\mathbf{e} \perp L(X)$

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

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