1 高斯－马尔可夫定理

假设样本数据以如下形式给出：

$$\boldsymbol{Y}_{n\times 1}=\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},\quad \boldsymbol{X}_{n\times k}=\begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{pmatrix}=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}.$$

假设有如下的一个简单线性关系，变量 y_i 是 k 个变量（$x_{i1},x_{i2},\cdots,x_{ik}$）和不可观察随机扰动项 u_i 的线性函数，

$$y_i=\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_k x_{ik}+u_i,\ i=1,\ 2,\ \cdots,\ n.$$

用矩阵形式

$$\boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{u},\tag{1}$$

其中，$\boldsymbol{\beta}=(\beta_1,\beta_2,\cdots,\beta_k)^{\mathrm{T}}$ 是 $k\times 1$ 维常量。

回归的目的是利用观察到的 y 和 $\boldsymbol{X}$ 对式（1）中的未知参数进行推断，包括回归系数 $\boldsymbol{\beta}$. 为了使回归得以进行，必须对数据生成过程进行若干假定。除了对模型（1）的线性假设之外，还有如下假定：

1) $E(\boldsymbol{u}\mid \boldsymbol{X})=0$；

2) $E(\boldsymbol{uu}^{\mathrm{T}}\mid \boldsymbol{X})=\sigma^2\boldsymbol{I}_{n\times n}$，其中，$\boldsymbol{I}_{n\times n}$为 $n\times n$ 单位矩阵；

3) rank（$\boldsymbol{X}$）$=k$.

在假定 1）和 2）下，基本线性回归模型（1）有 $k+1$ 个未知参数，即 k 个系数β 以及方差 σ^2.

回归分析就是通过观察到的 y 和 $\boldsymbol{X}$ 推断出这些参数。事实上有多种方法估计参数 $\boldsymbol{\beta}$ 和 σ^2. 给定 $\boldsymbol{\beta}$ 值，在假定 1）下，y_i 在给定 $k\times 1$ 向量 $\boldsymbol{x}_i$ 下的条件均值 $E(y_i)=\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}$，相应的与真实值的离差 $u_i=y_i-E(y_i)$. 最小二乘法[2]就是选择 $\boldsymbol{\beta}$ 的值使得 y_i 的真实值与预测的条件均值 $E(y_i)$ 的离差平方和最小。更正规地，最小二乘法就是用沿 y 轴估计的残差平方和最小值 S 作为 $\boldsymbol{\beta}$ 的估计值，S 定义为

$$S=\sum_{i=1}^{n}(y_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2=\sum_{i=1}^{n}u_i^2=(\boldsymbol{Y}-\boldsymbol{X\beta})^{\mathrm{T}}(\boldsymbol{Y}-\boldsymbol{X\beta})=\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}-2\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}+\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X\beta}.\tag{2}$$

为得到使 S 最小化的 $\boldsymbol{\beta}$ 值，将式（2）中的 S 对 $\boldsymbol{\beta}$ 求偏导得

$$\frac{\partial S}{\partial \boldsymbol{\beta}}=-2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}+2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X\beta}.$$

令$\dfrac{\partial S}{\partial \boldsymbol{\beta}}=0$，并用 $\hat{\boldsymbol{\beta}}$ 表示 $\boldsymbol{\beta}$ 的估计值，得到最小二乘估计量 $\hat{\boldsymbol{\beta}}$ 的方程为 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\hat{\boldsymbol{\beta}}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$.

如果 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 满秩，$\boldsymbol{\beta}$ 的最小二乘估计量可由下面 k 个线性方程求得：

$$\hat{\boldsymbol{\beta}}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}.\tag{3}$$

2 矩阵的分解

可以通过式（3）得到最小二乘估计量 $\hat{\boldsymbol{\beta}}$，假定 $\hat{\boldsymbol{\beta}}$ 可以分拆为 $\hat{\boldsymbol{\beta}}=(\hat{\boldsymbol{\beta}}_1,\hat{\boldsymbol{\beta}}_2)$，其中 $\hat{\boldsymbol{\beta}}_1$，$\hat{\boldsymbol{\beta}}_2$ 分别为$k_1\times 1$ 和 $k_2\times 1$ 向量，且 $k_1+k_2=k$. 相应地，将 $\boldsymbol{X}$ 分拆为 $\boldsymbol{X}=(\boldsymbol{X}_1,\boldsymbol{X}_2)$，那么方程 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\hat{\boldsymbol{\beta}}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$ 可以表示为

$\begin{pmatrix} \boldsymbol{X}_1^{\mathrm{T}} \\ \boldsymbol{X}_2^{\mathrm{T}} \end{pmatrix}(\boldsymbol{X}_1,\ \boldsymbol{X}_2)\begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix}=\begin{pmatrix} \boldsymbol{X}_1^{\mathrm{T}} \\ \boldsymbol{X}_2^{\mathrm{T}} \end{pmatrix}\boldsymbol{Y}$，化简得

$$\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1+\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2=\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{Y},\tag{4}$$

$$\boldsymbol{X}_2^{\mathrm{T}}\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1+\boldsymbol{X}_2^{\mathrm{T}}\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2=\boldsymbol{X}_2^{\mathrm{T}}\boldsymbol{Y}.\tag{5}$$

对式（4）两边左乘（$\boldsymbol{X}_2^{\mathrm{T}}\boldsymbol{X}_1)^{-1}$，得

$$\hat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{Y} - (\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2. \tag{6}$$

将式（6）代入式（5），得

$$\hat{\boldsymbol{\beta}}_2 = \{\boldsymbol{X}_2^{\mathrm{T}}[\boldsymbol{I} - \boldsymbol{X}_1(\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{X}_1)^{-1}]\boldsymbol{X}_2\}^{-1}\boldsymbol{X}_2^{\mathrm{T}}[\boldsymbol{I} - \boldsymbol{X}_1(\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^{\mathrm{T}}]\boldsymbol{Y}\,. \tag{7}$$

类似地，可以对式（5）两边左乘（$\boldsymbol{X}_2^{\mathrm{T}}\boldsymbol{X}_2)^{-1}$，得

$$\hat{\boldsymbol{\beta}}_1 = \{\boldsymbol{X}_1^{\mathrm{T}}[\boldsymbol{I} - \boldsymbol{X}_2(\boldsymbol{X}_2^{\mathrm{T}}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^{\mathrm{T}}]\boldsymbol{X}_1\}^{-1}\boldsymbol{X}_1^{\mathrm{T}}[\boldsymbol{I} - \boldsymbol{X}_2(\boldsymbol{X}_2^{\mathrm{T}}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^{\mathrm{T}}]\boldsymbol{Y}\,. \tag{8}$$

由于多元线性回归常数项的存在，对于多元线性回归的设计矩阵 $\boldsymbol{X}_{n\times k}$ 来说，有一列都为 1. 令 $\boldsymbol{X}_{n\times k}$ 的最后一列（第 k 列）为 $\boldsymbol{e}$，$\boldsymbol{e}=\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}_{n\times 1}$，将 $\boldsymbol{X}_2=\boldsymbol{e}$ 代入式（5），得

$$(n\bar{x}_1,\cdots,n\bar{x}_{k-1})\begin{pmatrix}\hat{\beta}_1\\ \vdots\\ \hat{\beta}_{k-1}\end{pmatrix} + n\hat{\beta}_k = n\bar{\boldsymbol{Y}}，即\ \hat{\beta}_k = \bar{Y} - \bar{x}_1\hat{\beta}_1 - \cdots - \bar{x}_{k-1}\hat{\beta}_{k-1}, \tag{9}$$

其中，$\bar{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij}$，$j=1$，…，$k-1$；$\bar{Y}=\frac{1}{n}\sum_{i=1}^{n}y_i$.

将 $\boldsymbol{X}_2=\boldsymbol{e}$ 代入式（8），再令 $\boldsymbol{M}=\boldsymbol{I}-\boldsymbol{X}_2(\boldsymbol{X}_2^{\mathrm{T}}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^{\mathrm{T}}$，那么 $\boldsymbol{M}=\boldsymbol{I}-\boldsymbol{e}(\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e})^{-1}\boldsymbol{e}^{\mathrm{T}}$.

已知 $\boldsymbol{M}=\boldsymbol{M}^{\mathrm{T}}$，$\boldsymbol{M}\boldsymbol{M}^{\mathrm{T}}=[\boldsymbol{I}-\boldsymbol{e}(\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e})^{-1}\boldsymbol{e}^{\mathrm{T}}][\boldsymbol{I}-\boldsymbol{e}(\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e})^{-1}\boldsymbol{e}^{\mathrm{T}}]=\boldsymbol{I}-\frac{1}{n}\boldsymbol{E}_{n\times n}=\boldsymbol{M}$. 那么式（8）可以改写为

$$\hat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{M}\boldsymbol{M}^{\mathrm{T}}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{M}\boldsymbol{M}^{\mathrm{T}}\boldsymbol{Y} = [\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{M}(\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{M})^{\mathrm{T}}]^{-1}\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{M}\boldsymbol{M}^{\mathrm{T}}\boldsymbol{Y}, \tag{10}$$

其中，$\boldsymbol{X}_1=\begin{pmatrix}x_{11} & x_{12} & \cdots & x_{1,k-1}\\ x_{21} & x_{22} & \cdots & x_{2,k-1}\\ \vdots & \vdots & & \vdots\\ x_{n1} & x_{n2} & \cdots & x_{n,k-1}\end{pmatrix}$，代入式（10），得

$$\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{M}=\begin{pmatrix}x_{11}-\bar{x}_1 & x_{21}-\bar{x}_1 & \cdots & x_{n1}-\bar{x}_1\\ x_{12}-\bar{x}_2 & x_{22}-\bar{x}_2 & \cdots & x_{n2}-\bar{x}_2\\ \vdots & \vdots & & \vdots\\ x_{1,k-1}-\bar{x}_{k-1} & x_{2,k-1}-\bar{x}_{k-1} & \cdots & x_{n,k-1}-\bar{x}_{k-1}\end{pmatrix}_{(k-1)\times n},\quad \boldsymbol{M}^{\mathrm{T}}\boldsymbol{Y}=\begin{pmatrix}y_1-\bar{Y}\\ y_2-\bar{Y}\\ \vdots\\ y_n-\bar{Y}\end{pmatrix}_{n\times 1},$$

$$\hat{\boldsymbol{\beta}}_1=\begin{pmatrix}\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2 & \cdots & \sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i,k-1}-\bar{x}_{k-1})\\ \vdots & & \vdots\\ \sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i,k-1}-\bar{x}_{k-1}) & \cdots & \sum_{i=1}^{n}(x_{i,k-1}-\bar{x}_{k-1})^2\end{pmatrix}^{-1}\begin{pmatrix}\sum_{i=1}^{n}(x_{i,1}-\bar{x}_1)(y_i-\bar{Y})\\ \vdots\\ \sum_{i=1}^{n}(x_{i,k-1}-\bar{x}_{k-1})(y_i-\bar{Y})\end{pmatrix}.$$

从 $\hat{\boldsymbol{\beta}}_1$ 的形式可以得出另外一种求最小二乘估计量[3]的方法。首先，将每个变量变化为与各自均值的离差形式，即 $\dot{y}_i=y_i-\bar{y}$，$\dot{x}_{ij}=x_{ij}-\bar{x}_j$，$j=1$，…，$k-1$.

然后作 $\dot{y}_i$ 对 $\dot{x}_{i1}$，…，$\dot{x}_{i,k-1}$ 的回归求得最小二乘估计量 $\hat{\beta}_1$，…，$\hat{\beta}_{k-1}$，最后将它们代入式（9）中得到 $\hat{\beta}_k$.

3 估计量的方差最小化

最小二乘估计是从残差的平方和最小的角度出发而得出的估计量。那么能不能从估计量的方差最小这个角度出发而得出估计量呢？回答是肯定的。

首先考虑简单线性回归情形，从式(3)可知参数的估计量是因变量的线性组合。那么不妨设 $\hat{\beta}_1=\sum c_i y_i=\sum(\beta_1 x_i+\beta_2+u_i)$.

要满足的条件是 $E(\hat{\beta}_1)=\beta_1\sum c_i x_i+\beta_2\sum c_i=\beta_1$，即 $\sum c_i x_i=1$，$\sum c_i=0$.

在这两个约束条件下，求方差最小化

$$\mathrm{var}(\hat{\beta}_1)=\sum c_i^2\mathrm{var}(y_i)=\sigma^2\sum c_i^2 .$$

这是一个条件极值问题，因此可以利用构造拉格朗日函数[4~5]如下：

$$\mathrm{L}(c_i,\lambda_1,\lambda_2)=\sigma_2\sum c_i^2+\lambda_1(\sum c_i x_i-1)+\lambda_2\sum c_i .$$

分别对 c_i，λ_1，λ_2 求偏导得

$$\frac{\partial \mathrm{L}}{\partial c_i}=2\sigma^2 c_i+\lambda_1 x_i+\lambda_2,\ \frac{\partial \mathrm{L}}{\partial \lambda_1}=\sum c_i x_i-1,\ \frac{\partial \mathrm{L}}{\partial \lambda_2}=\sum c_i.$$

从 $\frac{\partial \mathrm{L}}{\partial c_i}=0$，得出 $c_i=-\frac{1}{2\sigma^2}(\lambda_1 x_i+\lambda_2)$. (11)

从 $\frac{\partial \mathrm{L}}{\partial \lambda_1}=0$，$\frac{\partial \mathrm{L}}{\partial \lambda_2}=0$，得出 $\sum c_i x_i=1$，$\sum c_i=0$.

对式（11）两边同时加上 $\sum$ ，得

$$\lambda_1\sum x_i+n\lambda_2=0. \tag{12}$$

对式（11）两边同时乘以 x_i，再加上 $\sum$ 得

$$\lambda_1\sum x_i^2+\lambda_2\sum x_i=-2\sigma^2. \tag{13}$$

联立式（12）和式（13），可得

$$\lambda_1=\frac{2n\sigma^2}{(\sum x_i)^2-n\sum x_i^2},\ \lambda_2=-\frac{2\sigma^2\sum x_i}{(\sum x_i)^2-n\sum x_i^2}.$$

再将 λ_1，λ_2 代入式（11），得出

$$c_i=\frac{\sum x_i-nx_i}{(\sum x_i)^2-n\sum x_i^2}.$$

那么可以得出 $\hat{\beta}_1$ 的参数估计值为

$$\hat{\beta}_1=\frac{(\sum x_i)(\sum y_i)-n\sum x_i y_i}{(\sum x_i)^2-n\sum x_i^2}.$$

同样也可以通过此方法得出 $\hat{\beta}_2=\overline{y}-\hat{\beta}_1\overline{x}$，这与最小二乘估计出的估计量是一样的。现在要做的是把一元情况推广到多元，考虑估计量为 $\hat{\boldsymbol{\beta}}_{k\times1}=\boldsymbol{c}_{k\times n}\boldsymbol{y}_{n\times1}$，满足 $E(\hat{\boldsymbol{\beta}}_{k\times1})=\boldsymbol{c}_{k\times n}E(\boldsymbol{y}_{n\times1})=\boldsymbol{c}E(\boldsymbol{X\beta}+\boldsymbol{u})=\boldsymbol{\beta}$，得出 $\boldsymbol{cX}=\boldsymbol{I}_{n\times n}$. 很容易看出 $\boldsymbol{c}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$ 满足无偏性的约束条件。因此考虑估计量 $\hat{\boldsymbol{\beta}}_{k\times1}$ 的方差是否达到最小。不妨设最小二乘估计量为 $\boldsymbol{\beta}$，那么

$$\boldsymbol{\beta} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y},$$
$$\hat{\boldsymbol{\beta}}_{k\times 1} = \boldsymbol{\beta} - (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} + \boldsymbol{c}\boldsymbol{y} = \boldsymbol{\beta} + [\boldsymbol{c} - (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}]\boldsymbol{u},$$
$$\mathrm{cov}(\hat{\boldsymbol{\beta}}_{k\times 1}) = \sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1} + \sigma^2[\boldsymbol{c} - (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}][\boldsymbol{c} - (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}]^{\mathrm{T}}. \tag{14}$$

从式（14）中的第二项是非负半正定矩阵[6]可以知道最小二乘估计量的方差在所有的线性无偏估计量中是最小的。

4 结论

本文主要是围绕线性回归模型的最小二乘估计方法进行阐述，利用矩阵的分解原理对最小二乘估计量进行分解，得出另外一种与最小二乘估计等价的估计方法，另外从估计量的方差出发得出的估计量也与最小二乘估计量等价，因此最小二乘估计量的方差在所有的线性估计量中是最小的。

［参考文献］(References)

［1］ 魏宗舒. 概率论与数理统计教程[M]. 北京：高等教育出版社，1983.
WEI Z S. Probability and statistics course[M]. Beijing：Higher Education Press，1983.（in Chinese）

［2］ 因特里格特，博德金，萧政. 经济计量模型、技术和应用[M]. 李双杰，张涛. 北京：中国社会科学出版社，2004.
YINTELIGETE，BO D J，XIAO Z. Econometric model，technology and application[M]. LI S J，ZHANG T. Beijing：China Social Sciences Press，2004.（in Chinese）

［3］ 于秀林，任雪松. 多元统计分析[M]. 北京：中国统计出版社，1999.
YU X L，REN X S. Multivariate statistical analysis[M]. Beijing：China Statistics Press，1999.（in Chinese）

［4］ 许国祥. 统计学[M]. 北京：高等教育出版社，2004.
XU G X. Statistics[M]. Beijing：Higher Education Press，2004.（in Chinese）

［5］ 王梓坤. 概率论基础及其应用[M]. 北京：科学出版社，1976.
WANG Z K. Probability theory and its application[M]. Beijing：Science Press，1976.（in Chinese）

［6］ 陈上珠. 应用统计[M]. 北京：经济科学出版社，1987.
CHEN S Z. Application of statistics[M]. Beijing：Economic Science Press，1987.（in Chinese）