第 5 章

多元线性回归分析

学习要点

1. 理解经典线性回归模型。
2. 掌握最小二乘估计及其性质。
3. 理解最小二乘估计的几何解释。
4. 理解偏相关系数的概念、性质及计算公式。
5. 掌握线性回归模型的推断及评价。
6. 掌握线性回归分析的 MATLAB 实现。
7. 会用线性回归模型解决实际数据分析问题。

5.1 线性回归模型

我们先来看一个例子。

例 5.1 表 5.1 是来自 4 个不同城市的二手住宅房销售数据 (完整的数据在“二手住宅房销售调查数据.xlsx”文件中)。请分析住宅房成交价格 (y) 与建筑面积 (x_1)、城市人均年收入 (x_2)、是否学区房 (x_3) 之间的关系。

我们想要找到 y 与 $x_i, i = 1, 2, 3$ 之间的依赖关系, 首先想到的是下列线性函数

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \tag{5.1}$$

这就是**多元线性回归模型 (multivariate linear regression model)**, 其中 y 称为**应变量 (dependent variable)** 或**响应变量 (response variable)**, x_i 称为**解释变量 (explanatory variable)** 或**预测变量 (predictor variable)**, ε 称为**残差项**, 代表数据中不能被这个模型解释的部分, 一般假定它是一个随机变量。$\beta_i, i = 0, 1, 2, 3$ 称为回归系数, 是未知的, 需要通过样本数据估计。我们这里所说的“线性”是指式 (5.1) 对回归系数是线性的, 而解释变量 x_i 可以替换成它们的非线性变换, 如 x_i^2、$x_i^2 x_j$ 等, 都不会改变线性模型的本质, 虽然叫法可能不一样。更一般地, 含有 p 个自变量的线性回归模型为

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \tag{5.2}$$

为了估计回归系数, 需要收集样本数据 $\{(x_i, y_i) : i = 1, 2, \cdots, n\}$, 其中 $x_i = (x_{i1}, x_{i2}, \cdots,$

$x_{ip})^{\mathrm{T}} \in \mathbb{R}^p$, $y_i \in \mathbb{R}$。这时, 完整的线性回归模型为

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \qquad i = 1, 2, \cdots, n \tag{5.3}$$

表 5.1 二手住宅房销售调查数据

序 号	成交价 (万元)	建筑面积 (平方米)	城市人均年收入 (万元)	是否学区房
1	81	90	6.9	0
2	115	101	8.2	0
3	102	90	5.2	0
4	89	100	6.9	0
5	71	85	3.85	0
6	117	85	8.2	0
7	130	127	8.2	0
8	100	125	3.85	0
9	101	127	5.2	0
10	139	143	6.9	1
11	80	40	8.2	0
12	61	65	6.9	0
13	70	40	6.9	1
⋮	⋮	⋮	⋮	⋮

这里有 n 个方程, 表示成矩阵的形式就是

$$Y = X\beta + \varepsilon, \qquad Y = (y_1, y_2, \cdots, y_n)^{\mathrm{T}}, \qquad \beta = (\beta_0, \beta_1, \cdots, \beta_p)^{\mathrm{T}} \tag{5.4}$$

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \qquad \varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)^{\mathrm{T}} \tag{5.5}$$

我们假定残差向量 ε 满足条件

$$E(\varepsilon) = 0, \qquad \mathrm{cov}(\varepsilon, \varepsilon) = \sigma^2 I \tag{5.6}$$

其中 σ^2 是一个待估计的参数。我们把式 (5.4)、式 (5.5) 和式 (5.6) 称为**经典线性回归模型 (classical linear regression model)**。

5.2 最小二乘估计

回归分析的一个目的是得到线性方程 (5.2), 并用它来做预测, 即给定一个新的数据 $(x_{i1}, x_{i2}, \cdots, x_{ip})^{\mathrm{T}}$, 利用方程 (5.2) 得到相应的 y_i。要确定方程 (5.2) 就必须用样本数据估计回归系数 $\beta = (\beta_0, \beta_1, \cdots, \beta_p)^{\mathrm{T}}$, 本节介绍一种估计回归系数的方法——**最小二乘估计 (least squares estimation)**。

我们先来考虑这样一个问题: 设 Y 是一个 n 维随机向量, 令

$$\phi(a) = E\left[(Y-a)^{\mathrm{T}}(Y-a)\right], \qquad a \in \mathbb{R}^n \tag{5.7}$$

当 a 为多少时 $\phi(a)$ 最小呢? 设 $EY = \mu$, 则有

$$\begin{aligned}\phi(a) &= E\left[(Y-\mu+\mu-a)^{\mathrm{T}}(Y-\mu+\mu-a)\right] \\ &= E\left[(Y-\mu)^{\mathrm{T}}(Y-\mu)\right] + \|\mu-a\|^2\end{aligned} \tag{5.8}$$

由此可以看出, 当且仅当 $a=\mu$ 时 $\phi(a)$ 取得最小值。

回到经典线性回归模型 (5.4~5.6), 如果 β 是回归系数的准确值, 则有 $E(Y) = E(X\beta+\varepsilon) = X\beta$, 此时 $E[(Y-X\beta)^{\mathrm{T}}(Y-X\beta)]$ 是最小的, 而这个量可以近似地用

$$J = (Y-X\beta)^{\mathrm{T}}(Y-X\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 = \|Y-X\beta\|^2 \tag{5.9}$$

代替, 因此可以通过求解下列优化问题估计回归系数 β:

$$\min_{\beta\in\mathbb{R}^{p+1}} J = \|Y-X\beta\|^2 \tag{5.10}$$

这就是**最小二乘估计法**, 用此方法得到的估计量 $\widehat{\beta}$ 称为 β 的**最小二乘估计（量）(least squares estimation)**。

下面求最小二乘估计 $\widehat{\beta}$。注意到

$$J = Y^{\mathrm{T}}Y - 2Y^{\mathrm{T}}X\beta + \beta^{\mathrm{T}}X^{\mathrm{T}}X\beta \tag{5.11}$$

利用引理 4.1 得

$$\frac{\partial J}{\partial \beta} = 2(X^{\mathrm{T}}X)\beta - 2X^{\mathrm{T}}Y \tag{5.12}$$

因此最小二乘估计 $\widehat{\beta}$ 必须满足

$$(X^{\mathrm{T}}X)\widehat{\beta} = X^{\mathrm{T}}Y \tag{5.13}$$

如果 X 是列满秩的, 则 $X^{\mathrm{T}}X$ 可逆, 于是得到

$$\widehat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y \tag{5.14}$$

这就是**正规方程 (normal equation)**。如果 X 不是列满秩的, 则根据 3.3.2 节的知识, 使得 $\|Y-X\beta\|$ 取最小值的 β 不唯一, 其中范数最小的是 $\widehat{\beta} = X^{+}Y$, 其中 X^{+} 表示 X 的 Moore-Penrose 伪逆。

综合以上讨论, 我们证明了以下定理。

定理 5.1　在经典线性回归模型 (5.4~5.6) 中, 如果 X 是列满秩的, 则回归系数 β 的最小二乘估计由正规方程 (5.14) 给出; 如果 X 不是列满秩的, 则使得 $\|Y-X\beta\|$ 取最小值的 β 不唯一, 其中范数最小的为 $\widehat{\beta}=X^{+}Y$。

从定理 5.1 可立刻得出下列推论:

推论 5.1　在经典线性回归模型 (5.4~5.6) 中, 回归系数 β 的最小二乘估计 $\widehat{\beta}$ 满足

$$E(\widehat{\beta})=\beta, \qquad \operatorname{cov}(\widehat{\beta},\widehat{\beta})=\sigma^2(X^{\mathrm{T}}X)^{-1} \tag{5.15}$$

证明　如果 X 是列满秩的, 则有

$$E(\widehat{\beta})=E\left((X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y\right)=E\left[(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}(X\beta+\varepsilon)\right]=(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X\beta=\beta \tag{5.16}$$

如果 X 不是列满秩的, 则有

$$E(\widehat{\beta})=E(X^{+}Y)=E[X^{+}(X\beta+\varepsilon)]=X^{+}X\beta=\beta \tag{5.17}$$

再注意到

$$\widehat{\beta}-\beta=(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\varepsilon, \qquad E(\varepsilon\varepsilon^{\mathrm{T}})=\sigma^2 I \tag{5.18}$$

因此

$$\operatorname{cov}(\widehat{\beta},\widehat{\beta})=(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}E(\varepsilon\varepsilon^{\mathrm{T}})X(X^{\mathrm{T}}X)^{-1}=\sigma^2(X^{\mathrm{T}}X)^{-1} \tag{5.19}$$ □

接下来考虑回归残差的估计

$$\widehat{\varepsilon}=Y-\widehat{Y}=Y-X\widehat{\beta}=Y-X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y:=(I-H)Y \tag{5.20}$$

其中 $H:=X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ 称为**帽子矩阵 (hat matrix)**。

引理 5.1　设 X 是列满秩矩阵, $H=X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$, 则下列等式成立:

$$(I-H)^2=I-H \tag{5.21}$$

$$X^{\mathrm{T}}(I-H)=0 \tag{5.22}$$

证明　注意到帽子矩阵 H 是幂等矩阵, 即满足 $H^2=H$（本章习题 1）, 因此有

$$(I-H)^2=I-H-H+H^2=I-H \tag{5.23}$$

等式 (5.22) 的证明留作练习（本章习题 2）。 □

定理 5.2　残差估计量 $\widehat{\varepsilon}$ 的均值和协方差阵分别为

$$E(\widehat{\varepsilon})=0, \qquad \operatorname{cov}(\widehat{\varepsilon},\widehat{\varepsilon})=\sigma^2(I-H), \qquad E(\widehat{\varepsilon}^{\mathrm{T}}\widehat{\varepsilon})=(n-p-1)\sigma^2 \tag{5.24}$$

证明　注意到

$$\widehat{\varepsilon}=(I-H)Y=(I-H)(X\beta+\varepsilon) \tag{5.25}$$

再由式 (5.22) 得 $(I-H)X=0$, 因此有

$$E(\widehat{\varepsilon})=(I-H)X\beta=0 \tag{5.26}$$

$$\operatorname{cov}(\widehat{\varepsilon},\widehat{\varepsilon})=(I-H)E(\varepsilon\varepsilon^{\mathrm{T}})(I-H)^{\mathrm{T}}=\sigma^2(I-H)^2=\sigma^2(I-H) \tag{5.27}$$

再注意到

$$\operatorname{tr}(H)=\operatorname{tr}(X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})=\operatorname{tr}((X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X)=\operatorname{tr}(I_{p+1})=p+1 \tag{5.28}$$

因此有

$$\begin{aligned}E(\widehat{\varepsilon}^{\mathrm{T}}\widehat{\varepsilon})&=\sum_{i=1}^{n}\operatorname{var}(\widehat{\varepsilon}_i)=\operatorname{tr}\left(\operatorname{cov}(\widehat{\varepsilon},\widehat{\varepsilon})\right)=\sigma^2\operatorname{tr}(I_n-H)\\&=\sigma^2\left[\operatorname{tr}(I_n)-\operatorname{tr}(H)\right]=(n-p-1)\sigma^2\end{aligned} \tag{5.29}$$ □

注: 之所以关心 $\widehat{\varepsilon}^{\mathrm{T}}\widehat{\varepsilon}$ 的性质, 是因为它是回归残差估计的平方和:

$$\begin{aligned}\widehat{\varepsilon}^{\mathrm{T}}\widehat{\varepsilon}&=(Y-\widehat{Y})^{\mathrm{T}}(Y-\widehat{Y})=\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2\\&=\sum_{i=1}^{n}(y_i-\widehat{\beta}_0-\widehat{\beta}_1x_{i1}-\widehat{\beta}_2x_{i2}-\cdots-\widehat{\beta}_px_{ip})^2\end{aligned} \tag{5.30}$$

定理 5.3 残差估计量 $\widehat{\varepsilon}$ 满足下列性质:

$$X^{\mathrm{T}}\widehat{\varepsilon}=0 \tag{5.31}$$

$$\widehat{Y}^{\mathrm{T}}\widehat{\varepsilon}=0 \tag{5.32}$$

$$\widehat{\varepsilon}^{\mathrm{T}}\widehat{\varepsilon}=Y^{\mathrm{T}}(I-H)Y \tag{5.33}$$

证明 式 (5.31) 可由式 (5.22) 直接得到。至于式 (5.32), 只需要注意到

$$\widehat{Y}=HY,\qquad \widehat{\varepsilon}=(I-H)Y,\qquad H^2=H \tag{5.34}$$

便可得到

$$\widehat{Y}^{\mathrm{T}}\widehat{\varepsilon}=Y^{\mathrm{T}}H^{\mathrm{T}}(I-H)Y=Y^{\mathrm{T}}H(I-H)Y=Y^{\mathrm{T}}(H-H^2)Y=0 \tag{5.35}$$

式 (5.33) 可由式 (5.21) 直接得到。 □

定理 5.4 最小二乘估计量 $\widehat{\beta}$ 和 $\widehat{\varepsilon}$ 是不相关的。

证明 根据推论 5.1 及定理 5.2, 得 $E(\widehat{\beta})=\beta, E(\widehat{\varepsilon})=0$, 因此

$$\operatorname{cov}(\widehat{\beta},\widehat{\varepsilon})=E\left[(\widehat{\beta}-\beta)\widehat{\varepsilon}^{\mathrm{T}}\right]=E\left[\widehat{\beta}\widehat{\varepsilon}^{\mathrm{T}}\right]=E\left[(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y\widehat{\varepsilon}^{\mathrm{T}}\right]$$

$$
\begin{aligned}
&= E\left[(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}YY^{\mathrm{T}}(I-H)\right] \\
&= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}E(YY^{\mathrm{T}})(I-H)
\end{aligned} \tag{5.36}
$$

再注意到

$$
E(YY^{\mathrm{T}}) = E\left[(X\beta+\varepsilon)(X\beta+\varepsilon)^{\mathrm{T}}\right] = X\beta\beta^{\mathrm{T}}X^{\mathrm{T}} + \sigma^2 I \tag{5.37}
$$

将式 (5.37) 代入式 (5.36) 的最后一个表达式, 并利用式 (5.22) 得到 $\operatorname{cov}(\widehat{\beta},\widehat{\varepsilon}) = 0$, 定理得证。 □

对于任意向量 $c \in \mathbb{R}^{p+1}$, 考虑参数 $\theta = c^{\mathrm{T}}\beta$ 的**线性无偏估计量 (linear unbiased estimator)**, 即形如 $a^{\mathrm{T}}Y$ 的无偏统计量, 根据定理 5.1 和推论 5.1, $\widehat{\beta}$ 是 β 的线性无偏估计量, 因此 $c^T\widehat{\beta}$ 是 $c^{\mathrm{T}}\beta$ 的线性无偏估计量, 不仅如此, $c^{\mathrm{T}}\widehat{\beta}$ 还是 $c^{\mathrm{T}}\beta$ 的所有线性无偏估计量中方差最小的, 这就是下面的定理:

定理 5.5 在经典线性回归模型中, 设 X 是列满秩的, $\widehat{\beta}$ 是回归系数 β 的最小二乘估计, $c \in \mathbb{R}^{p+1}$, 则对于 $c^{\mathrm{T}}\beta$ 的任意一个线性无偏估计量 $a^{\mathrm{T}}Y$ 皆有

$$
\operatorname{var}(a^{\mathrm{T}}Y) \geqslant \operatorname{var}(c^{\mathrm{T}}\widehat{\beta}) \tag{5.38}
$$

即 $c^{\mathrm{T}}\widehat{\beta}$ 是 $c^{\mathrm{T}}\beta$ 的**最小方差线性无偏估计量 (minimum-variance linear unbiased estimator)** 或者**最佳线性无偏估计量 (best linear unbiased estimator, BLUE)**。

证明 首先, $a^{\mathrm{T}}Y$ 是 $c^{\mathrm{T}}\beta$ 的线性无偏估计量意味着

$$
c^{\mathrm{T}}\beta = E[a^{\mathrm{T}}Y] = E[a^{\mathrm{T}}(X\beta+\varepsilon)] = a^{\mathrm{T}}X\beta \tag{5.39}
$$

这个等式必须对所有 $\beta \in \mathbb{R}^{p+1}$ 皆成立, 因此必有 $X^{\mathrm{T}}a = c$。于是

$$
\operatorname{var}(a^{\mathrm{T}}Y) = \operatorname{var}(a^{\mathrm{T}}X\beta + a^{\mathrm{T}}\varepsilon) = \operatorname{var}(a^{\mathrm{T}}\varepsilon) = a^{\mathrm{T}}(\sigma^2 I)a = \sigma^2 a^{\mathrm{T}}a \tag{5.40}
$$

$$
\begin{aligned}
\operatorname{var}(c^{\mathrm{T}}\widehat{\beta}) &= \operatorname{var}(c^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y) = \operatorname{var}(c^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\varepsilon) \\
&= \operatorname{var}(a^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\varepsilon) \\
&= \operatorname{var}(a^{\mathrm{T}}H\varepsilon) \\
&= \sigma^2 a^{\mathrm{T}}HH^{\mathrm{T}}a \\
&= \sigma^2 a^{\mathrm{T}}Ha
\end{aligned} \tag{5.41}
$$

$$
\operatorname{var}(a^{\mathrm{T}}Y) - \operatorname{var}(c^{\mathrm{T}}\widehat{\beta}) = \sigma^2 a^{\mathrm{T}}(I-H)a \tag{5.42}
$$

其中 H 是帽子矩阵, 式 (5.41) 推导的最后两个等号用到了 H 的对称性和幂等性（本章习题 1）。既然 H 是幂等的, 其特征值只能是 1 或 0（本章习题 3）, 因此 $I-H$ 是半正定的, 从而有

$$
\operatorname{var}(a^{\mathrm{T}}Y) - \operatorname{var}(c^{\mathrm{T}}\widehat{\beta}) \geqslant 0 \tag{5.43}
$$ □

接下来考虑响应变量 y 的平方和分解问题。记

$$S_T^2=\sum_{i=1}^{n}(y_i-\overline{y})^2,\qquad \overline{y}=\frac{1}{n}\sum_{i=1}^{n}y_i \tag{5.44}$$

称 S_T^2 为**总离差平方和 (total sum of squares)**, 反映了响应变量 y 的变异程度。$\widehat{Y}=X\widehat{\beta}$ 是响应变量 y 的回归估计, 它与 Y 不一致, 其偏差 $\widehat{\varepsilon}=Y-\widehat{Y}$ 就是回归残差的估计。根据定理 5.2 得

$$X^{\mathrm{T}}(Y-\widehat{Y})=X^{\mathrm{T}}\widehat{\varepsilon}=0 \tag{5.45}$$

由于 X 的第一列元素全是 1, 因此有

$$\sum_{i=1}^{n}(y_i-\widehat{y}_i)=0 \tag{5.46}$$

由此推出 $\overline{y}=\overline{\widehat{y}}$。称

$$S_{\mathrm{Reg}}^2:=\sum_{i=1}^{n}(\widehat{y}_i-\overline{\widehat{y}})^2=\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2 \tag{5.47}$$

为**回归平方和 (regression sum of squares)**。称

$$S_{\mathrm{Res}}^2:=\sum_{i=1}^{n}\widehat{\varepsilon}_i^2=\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2 \tag{5.48}$$

为**残差平方和 (residual sum of squares)**。关于这 3 个量之间的关系, 有下列定理:

定理 5.6 在经典线性回归模型中, 总离差平方和 S_T^2、回归平方和 S_{Reg}^2、残差平方和 R_{Res}^2 三者满足下列关系:

$$S_T^2=S_{\mathrm{Reg}}^2+S_{\mathrm{Res}}^2 \tag{5.49}$$

证明 根据定理 5.2 得 $\widehat{Y}^{\mathrm{T}}\widehat{\varepsilon}=0$, 于是有

$$\begin{aligned}S_T^2&=Y^{\mathrm{T}}Y-n(\overline{y})^2=(\widehat{Y}+\widehat{\varepsilon})^{\mathrm{T}}(\widehat{Y}+\widehat{\varepsilon})-n(\overline{y})^2=\widehat{Y}^{\mathrm{T}}\widehat{Y}-n(\overline{\widehat{y}})^2+\widehat{\varepsilon}^{\mathrm{T}}\widehat{\varepsilon}\\&=S_{\mathrm{Reg}}^2+S_{\mathrm{Res}}^2\end{aligned} \tag{5.50}$$

□

回归平方和 S_{Reg}^2 是线性回归模型能够解释的那部分变异性, 它所占的比例为

$$R^2:=\frac{S_{\mathrm{Reg}}^2}{S_T^2} \tag{5.51}$$

通常称之为**决定系数 (coefficient of determination)**, 其大小反映了线性回归模型拟合样本数据的能力。

5.3 几何解释

本节讨论回归分析的几何解释。回归分析本质上是用解释变量的线性组合逼近被解释变量, 在内积空间的框架下讨论这个问题更方便。考虑定义在概率空间 $(\Omega, \mathcal{F}, P)$ 上的随机变量 X, 如果 $E(|X|) < \infty$ 且 $E(|X|^2) < \infty$, 则称 X 是**具有有限二阶矩的**。如果 X 具有有限二阶矩, 则

$$\begin{aligned} DX &= E[(X - EX)^2] = E[X^2 - 2XEX + (EX)^2] \\ &= E(X^2) - (EX)^2 \leqslant E(|X|^2) < \infty \end{aligned} \tag{5.52}$$

因此 X 具有有限方差。定义在 $(\Omega, \mathcal{F}, P)$ 上的所有二阶矩有限的随机变量构成一个线性空间, 记作 $L^2(\Omega)$。在 $L^2(\Omega)$ 上可以按照如下方式定义一个内积:

$$\langle X, Y \rangle_{L^2} = E(XY) \tag{5.53}$$

则 $L^2(\Omega)$ 关于这个内积构成一个内积空间。由这个内积诱导的范数为

$$\|X\|_{L^2} = \sqrt{\langle X, X \rangle_{L^2}} = \sqrt{E(|X|^2)} \tag{5.54}$$

设 $Y, X_1, X_2, \cdots, X_p \in L^2(\Omega)$, 记由 $1, X_1, X_2, \cdots, X_p$ 生成的线性子空间为 $\mathcal{V}_p$, 我们希望找到一个随机变量 $\widehat{Y} \in \mathcal{V}_p$, 使得

$$\|Y - \widehat{Y}\|_{L^2} \leqslant \|Y - Z\|_{L^2}, \qquad \forall\, Z \in \mathcal{V}_p \tag{5.55}$$

称满足此条件的 $\widehat{Y}$ 为 Y 在子空间 $\mathcal{V}_p$ 上的**最佳（线性）逼近元**。寻找最佳逼近元可以通过求解下列优化问题实现:

$$\min_{\beta=(\beta_0,\beta_1,\cdots,\beta_p)\in\mathbb{R}^{p+1}} \|Y - \beta_0 - \beta_1 X_1 - \cdots - \beta_p X_p\|_{L^2} \tag{5.56}$$

这正是最小二乘问题。

最佳逼近元的存在性是有限维赋范线性空间局部紧性的结果, 其证明在泛函分析教材中可以找到, 如文献 [75]。最佳逼近元的唯一性有下列定理作为保障:

定理 5.7 设 $(\Omega, \mathcal{F}, P)$ 是概率空间, $Y, X_1, X_2, \cdots, X_p \in L^2(\Omega)$, $\mathcal{V}_p = \mathrm{Span}\{1, X_1, X_2, \cdots, X_p\}$, 则 Y 在 $\mathcal{V}_p$ 上的最佳逼近元是唯一的。如果 $1, X_1, X_2, \cdots, X_p$ 是线性无关的, 则优化问题 (5.56) 的解是唯一的。

证明 事实上, 设

$$d = \min_{Z \in \mathcal{V}_p} \|Y - Z\|_{L^2} \tag{5.57}$$

如果 $d = 0$, 则 Y 的最佳逼近元 $\widehat{Y}$ 必与 Y 相等, 因此是唯一的; 如果 $d > 0$, 设 Y', Y'' 都是 Y 在 $\mathcal{V}_p$ 上的最佳逼近元, 则有

$$\|Y - Y'\|_{L^2} = \|Y - Y''\|_{L^2} = d \tag{5.58}$$

于是

$$\begin{aligned}\left\|Y-\frac{Y'+Y''}{2}\right\|_{L^2}^2&=\left\|\frac{Y-Y'}{2}+\frac{Y-Y''}{2}\right\|_{L^2}^2\\&=\left\|\frac{Y-Y'}{2}\right\|_{L^2}^2+\left\|\frac{Y-Y''}{2}\right\|_{L^2}^2+2\left\langle\frac{Y-Y'}{2},\frac{Y-Y''}{2}\right\rangle_{L^2}\\&=\frac{d^2}{2}+\frac{1}{2}\langle Y-Y',Y-Y''\rangle_{L^2}\end{aligned}\tag{5.59}$$

如果 $Y'\neq Y''$, 则 $Y-Y'\neq Y-Y''$, 且由于它们长度相等, 因此夹角大于 0, 从而有

$$\langle Y-Y',Y-Y''\rangle_{L^2}<\|Y-Y'\|\cdot\|Y-Y''\|=d^2\tag{5.60}$$

联合式 (5.59) 与式 (5.60) 得到

$$\left\|Y-\frac{Y'+Y''}{2}\right\|_{L^2}<d\tag{5.61}$$

但这与式 (5.57) 矛盾, 因此必然有 $Y'=Y''$。当 $1,X_1,X_2,\cdots,X_p$ 线性无关时, Y 在子空间 $\mathcal{V}_p$ 上的最佳逼近元 $\widehat{Y}$ 有唯一的线性表示, 因此优化问题 (5.56) 的解是唯一的。 □

记 $\varepsilon_Y=Y-\widehat{Y}$, 即 Y 与它在 $\mathcal{V}_p$ 上的最佳逼近元 $\widehat{Y}$ 之差, 称为**最佳逼近误差**或**回归残差**。

命题 5.1 设 $\widehat{Y}$ 是 Y 在 $\mathcal{V}_p$ 上的最佳逼近元, $\varepsilon_Y=Y-\widehat{Y}$ 是回归残差, 则有

$$\langle\varepsilon_Y,Z\rangle_{L^2}=0,\qquad\forall\,Z\in\mathcal{V}_p\tag{5.62}$$

即 $\widehat{Y}$ 是 Y 在 $\mathcal{V}_p$ 上的正交投影。

证明 考虑下列关于实变量 t 的函数

$$\varphi(t)=\|\varepsilon_Y-tZ\|_{L^2}^2-\|\varepsilon_Y\|_{L^2}^2\tag{5.63}$$

由于

$$\|\varepsilon_Y\|_{L^2}=\|Y-\widehat{Y}\|_{L^2}=\min_{Z'\in\mathcal{V}_p}\|Y-Z'\|\leqslant\|Y-\widehat{Y}-tZ\|_{L^2}=\|\varepsilon_Y-tZ\|_{L^2}\tag{5.64}$$

因此 $\varphi(t)$ 是非负的。再注意到 $\varphi(t)$ 是一个二次函数:

$$\varphi(t)=\langle\varepsilon_Y-tZ,\varepsilon_Y-tZ\rangle_{L^2}-\|\varepsilon_Y\|_{L^2}^2=t^2\|Z\|_{L^2}^2-2t\langle\varepsilon_Y,Z\rangle_{L^2}\tag{5.65}$$

要使 $\varphi(t)$ 恒为非负, 必须 $\langle\varepsilon_Y,Z\rangle_{L^2}=0$, 这就证明了式 (5.62)。 □

由式 (5.62) 还可以得到

$$E\varepsilon_Y=\langle\varepsilon_Y,1\rangle_{L^2}=0\tag{5.66}$$

$$\langle\varepsilon_Y,Y\rangle_{L^2}=\langle\varepsilon_Y,Y-\widehat{Y}+\widehat{Y}\rangle_{L^2}=\langle\varepsilon_Y,\varepsilon_Y+\widehat{Y}\rangle_{L^2}=\langle\varepsilon_Y,\varepsilon_Y\rangle_{L^2}=D\varepsilon_Y\tag{5.67}$$

如果另有一个随机变量 $Z \in L^2(\Omega)$, 设 $\widehat{Z}$ 是 Z 在 $\mathcal{V}_p$ 上的最佳逼近, $\varepsilon_Z = Z - \widehat{Z}$ 是回归残差, 则有

$$\langle \varepsilon_Y, Z\rangle_{L^2} = \langle \varepsilon_Y, Z - \widehat{Z} + \widehat{Z}\rangle_{L^2} = \langle \varepsilon_Y, \varepsilon_Z + \widehat{Z}\rangle_{L^2} = \langle \varepsilon_Y, \varepsilon_Z\rangle_{L^2} = \operatorname{cov}(\varepsilon_Y, \varepsilon_Z) \tag{5.68}$$

接下来考虑优化问题 (5.56) 的解。先考虑 $p = 1$ 的情形。令

$$\begin{aligned} J &= \|Y - \beta_0 - \beta_1 X_1\|_{L^2}^2 \\ &= \|Y\|_{L^2}^2 + \beta_0^2 + \beta_1^2 \|X_1\|_{L^2}^2 - 2\beta_0\langle Y, 1\rangle_{L^2} - 2\beta_1\langle Y, X_1\rangle_{L^2} + 2\beta_0\beta_1\langle 1, X_1\rangle_{L^2} \\ &= E(Y^2) + \beta_0^2 + \beta_1^2 E(X_1^2) - 2\beta_0 EY - 2\beta_1 E(X_1 Y) + 2\beta_0\beta_1 EX_1 \end{aligned} \tag{5.69}$$

极小值点应满足下列一阶条件

$$\begin{cases} \dfrac{\partial J}{\partial \beta_0} = 2\beta_0 - 2EY + 2\beta_1 EX_1 = 0 \\ \dfrac{\partial J}{\partial \beta_1} = 2\beta_1 E(X_1^2) - 2E(X_1 Y) + 2\beta_0 EX_1 = 0 \end{cases} \tag{5.70}$$

由此求得极小值点为

$$\widehat{\beta}_0 = EY - \frac{\operatorname{cov}(Y, X_1)}{DX_1} EX_1, \qquad \widehat{\beta}_1 = \frac{\operatorname{cov}(Y, X_1)}{DX_1} \tag{5.71}$$

对于一般的情形, 设 Y 在 $\mathcal{V}_p$ 上的最佳逼近元 $\widehat{Y} := \widehat{\beta}_0 + \sum\limits_{j=1}^{p} \widehat{\beta}_j X_j$, 记 $\mu_Y = EY, \mu_j = EX_j, j = 1, 2, \cdots, p$, 由式 (5.66) 得

$$\mu_Y = EY = E\widehat{Y} = \widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_j \mu_j \tag{5.72}$$

因此有

$$\varepsilon_Y = Y - \widehat{Y} = Y - \mu_Y - (\widehat{Y} - \mu_Y) = Y - \mu_Y - \sum_{j=1}^{p} \widehat{\beta}_j (X_j - \mu_j) \tag{5.73}$$

根据命题 (5.1), 回归残差 ε_Y 与 $\mathcal{V}_p$ 正交, 因此有

$$\begin{aligned} 0 &= \langle \varepsilon_Y, X_k - \mu_k\rangle_{L^2} = \langle Y - \mu_Y, X_k - \mu_k\rangle_{L^2} - \sum_{j=1}^{p} \widehat{\beta}_j \langle X_j - \mu_j, X_k - \mu_k\rangle_{L^2} \\ &= \sigma_{Y,X_k} - \sum_{j=1}^{p} \widehat{\beta}_j \sigma_{jk}, \qquad k = 1, 2, \cdots, p \end{aligned} \tag{5.74}$$

其中 $\sigma_{Y,X_k} = \operatorname{cov}(Y, X_k), \sigma_{kj} = \operatorname{cov}(X_k, X_j), j, k = 1, 2, \cdots, p$。记

$$\Sigma_{XX} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}, \quad \Sigma_{XY} = \begin{pmatrix} \sigma_{Y,X_1} \\ \sigma_{Y,X_2} \\ \vdots \\ \sigma_{Y,X_p} \end{pmatrix}, \quad \widehat{\beta} = \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \vdots \\ \widehat{\beta}_p \end{pmatrix} \tag{5.75}$$