统计与机器学习

第一章：线性回归 - Part I

倪 葎

DaSE@ECNU
(lni@dase.ecnu.edu.cn)

2020 年 10 月 2 日

目录

线性回归的模型

- 线性回归模型为

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \tag{1}$$

  - y 为响应变量/因变量，为一个随机变量；
  - x 为协变量/自变量，通常假定是确定性的变量；
  - $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ 是 $p+1$ 个未知参数；
  - ε 为随机误差，并假定

$$\begin{cases} E(\varepsilon) = 0, \\ \mathrm{Var}(\varepsilon) = \sigma^2 \end{cases} \tag{2}$$

线性回归的模型

在实际问题中，

- n 组观测数据 $(x_{i1}, x_{i2}, \cdots, x_{ip}, y_i)$；
- 基于观测数据，线性回归方程模型可写为

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_1 x_{12} + \cdots + \beta_p x_{1p} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_1 x_{22} + \cdots + \beta_p x_{2p} + \varepsilon_2 \\ \quad\vdots \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_1 x_{n2} + \cdots + \beta_p x_{np} + \varepsilon_n \end{cases} \tag{3}$$

线性回归的模型

- 令

$$
\begin{aligned}
\boldsymbol{y} &= \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \\
\boldsymbol{\beta} &= (\beta_0, \beta_1, \beta_2, \cdots, \beta_p)', \\
\boldsymbol{\varepsilon} &= (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)'.
\end{aligned}
$$

- 线性回归模型的矩阵形式为

$$
\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{4}
$$

- 问题：如何估计回归系数 β?

线性回归的基本假定

为了便于进行参数估计，需要对回归方程进行一些假设：

(1) 设计矩阵 $\boldsymbol{X}$

- 是**确定性变量**，不是随机变量；
- 要求 $\operatorname{rank}(\boldsymbol{X}) = p + 1 < n$，这表明了自变量之间不相关，样本量应大于自变量的个数，$\boldsymbol{X}$ 是满秩矩阵.

线性回归的基本假定

为了便于进行参数估计，需要对回归方程进行一些假设:

(2) 随机误差是零均值和等方差的，即

- $E(\varepsilon_i) = 0, \quad i = 1, 2, \cdots, n$，表示没有系统误差;
- $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2, & i = j \\ 0, & i \neq j \end{cases} \quad i, j = 1, 2, \cdots, n$，表明随机在不同的样本点之间是不相关的（在正态假定下即为独立的），不存在序列相关，并且有相同的精度;

这个条件常称为**高斯-马尔可夫条件**.

线性回归的基本假定

为了便于进行参数估计，需要对回归方程进行一些假设:

(3) 假定随机误差项服从正态分布，即

$$\begin{cases} \varepsilon_i \sim N(0, \sigma^2), & i = 1, 2, \cdots, n \\ \varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n \text{ 相互独立} \end{cases}$$

在假设（3）下，随机误差向量服从

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n).$$

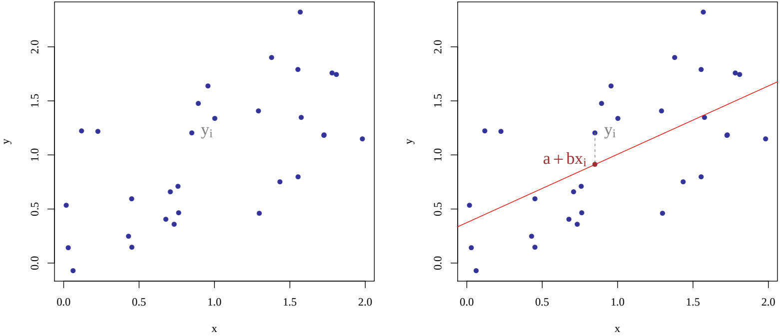
等价于假定因变量 $\boldsymbol{y}$ 服从 n **维正态分布**，其期望向量和协方差矩阵分别为

$$\begin{aligned} E(\boldsymbol{y}) &= \boldsymbol{X\beta} \\ \mathrm{Var}(\boldsymbol{y}) &= \sigma^2 \boldsymbol{I}_n \end{aligned}$$

最小二乘估计

基本思想

- 对于第 i 个样本，实际观测值为 y_i，估计值为 $\boldsymbol{x}_i'\boldsymbol{\beta}$.



- 离差（实际观测值和估计值的差）定义为 $y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$.

最小二乘估计

基本思想

最小二乘估计：通过最小化离差平方和而得到的估计方法.

- 对于线性模型，离差平方和定义为

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2$$

- 最小二乘估计为

$$\hat{\boldsymbol{\beta}}_{\mathrm{LS}} = \arg\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta})$$

最小二乘估计

具体计算方法

离差平方和 $Q(\boldsymbol{\beta})$ 可写为

$$\begin{aligned}
Q(\boldsymbol{\beta}) &= \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 \\
&= \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 \\
&= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\
&= \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{y}'\boldsymbol{y}
\end{aligned}$$

最小二乘估计

常见的矩阵计算（补充）

对于一个 p 维列向量 $\boldsymbol{x} = (x_1, x_2, \cdots, x_p)'$,

- 线性函数求导：对于任意常向量 $\boldsymbol{a} = (a_1, a_2, \cdots, a_p)'$, 我们有

$$\frac{\partial(\boldsymbol{x}'\boldsymbol{a})}{\partial\boldsymbol{x}} = \frac{\partial(\boldsymbol{a}'\boldsymbol{x})}{\partial\boldsymbol{x}} = \boldsymbol{a}.$$

- 二次型求导：对于任意 $p \times p$ 常值矩阵 $\boldsymbol{B}$，我们有

$$\frac{\partial(\boldsymbol{x}'\boldsymbol{B}\boldsymbol{x})}{\partial\boldsymbol{x}} = (\boldsymbol{B} + \boldsymbol{B}')\boldsymbol{x}.$$

特别地，如果 $\boldsymbol{B}$ 是一个对称矩阵，那么

$$\frac{\partial(\boldsymbol{x}'\boldsymbol{B}\boldsymbol{x})}{\partial\boldsymbol{x}} = 2\boldsymbol{B}\boldsymbol{x}.$$

最小二乘估计

具体计算方法

- 对 $\boldsymbol{\beta}$ 求导，可得

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{X}'\boldsymbol{y}$$

- 令 $\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0$，可得

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{y}$$

- 基于假设（1），$\boldsymbol{X}'\boldsymbol{X}$ 是满秩的，因此，$(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 存在.
- 由此，最小二乘估计为

$$\hat{\boldsymbol{\beta}}_{\mathrm{LS}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

最小二乘估计

最小二乘估计为

$$\hat{\boldsymbol{\beta}}_{\mathrm{LS}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

说明

- 在求最小二乘估计 $\hat{\boldsymbol{\beta}}$ 时，需要 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$必须存在. 也就是说，$(\boldsymbol{X}'\boldsymbol{X})$ 是一非奇异矩阵，即 $|\boldsymbol{X}'\boldsymbol{X}| \neq 0$.
- 由线性代数可知，$\mathrm{rank}(\boldsymbol{X}) \geq \mathrm{rank}(\boldsymbol{X}'\boldsymbol{X})$. 如果 $\boldsymbol{X}'\boldsymbol{X}$ 为 $p+1$ 阶满秩矩阵，也就是说 $\mathrm{rank}(\boldsymbol{X}'\boldsymbol{X}) = p+1$，那么 $\mathrm{rank}(\boldsymbol{X}) \geq p+1$.
- 另一方面，设计矩阵 $\boldsymbol{X}$ 为 $n \times (p+1)$ 阶矩阵，于是应有 $n \geq p+1$. 这表明了采用最小二乘法估计方法求解线性回归的未知参数，样本量必须不少于模型中的参数个数.

最小二乘估计

- 回归值或拟合值定义为

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

其中，$\hat{y}_i = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}, i = 1, 2, \cdots, n.$

- $\hat{\boldsymbol{y}}$ 也可以写为

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

- 矩阵 $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$：将观测值 $\boldsymbol{y}$ 变换为 $\hat{\boldsymbol{y}}$. 从形式上来看，就是给 $\boldsymbol{y}$ 戴上了一顶帽子“ ^ ”，因而形象地称矩阵 $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ 为帽子矩阵，记为 $\boldsymbol{H}$.

- 于是，$\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$。

最小二乘估计

定理 1-1 (帽子矩阵的性质)

帽子矩阵 $\boldsymbol{H}=\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ 具有以下的一些性质。

- $\boldsymbol{H}$ 是 n 阶对称矩阵;
- $\boldsymbol{H}$ 是幂等矩阵，即 $\boldsymbol{H}=\boldsymbol{H}^2$;
- $\boldsymbol{H}$ 的迹为 $p+1$，即 $\mathrm{tr}(\boldsymbol{H})=p+1$.

证明:

- $\boldsymbol{H}$ 的转置矩阵为

$$\begin{aligned}\boldsymbol{H}' &= \left(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right)'=(\boldsymbol{X}')'\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\right)'\boldsymbol{X}' \\ &= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'=\boldsymbol{H}\end{aligned}$$

由于 $\boldsymbol{H}$ 的转置矩阵等于 $\boldsymbol{H}$，所以 $\boldsymbol{H}$ 是对称的。

最小二乘估计

- 我们可知，

$$
\begin{aligned}
\boldsymbol{H}^2 &= \left(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right)^2 \\
&= \left(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right)\left(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right) \\
&= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \\
&= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \boldsymbol{H}
\end{aligned}
$$

  因此，$\boldsymbol{H}$ 是幂等矩阵.

- 易知 $\boldsymbol{X}'\boldsymbol{X}$ 是一个 $(p+1)\times(p+1)$ 的满秩矩阵。于是，我们计算 $\boldsymbol{H}$ 的迹，即

$$
\begin{aligned}
\mathrm{tr}(\boldsymbol{H}) &= \mathrm{tr}(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') \\
&= \mathrm{tr}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}) \\
&= \mathrm{tr}(\boldsymbol{I}_{p+1}) = p+1
\end{aligned}
$$

最小二乘估计

- 残差定义为

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$$

- 也可写为

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{H}\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$$

- 几何上的关系：回归值 $\hat{\boldsymbol{y}}$ 与残差 $\boldsymbol{e}$ 垂直，即

$$\hat{\boldsymbol{y}}'\boldsymbol{e} = (\boldsymbol{H}\boldsymbol{y})'((\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}) = \boldsymbol{y}'\boldsymbol{H}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} = 0$$

.

最小二乘估计

- 残差的协方差矩阵为

$$
\begin{aligned}
\mathrm{Var}(\boldsymbol{e}) &= \mathrm{Cov}(\boldsymbol{e}, \boldsymbol{e}) \\
&= \mathrm{Cov}((\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}, (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}) \\
&= (\boldsymbol{I}-\boldsymbol{H})\mathrm{Cov}(\boldsymbol{y}, \boldsymbol{y})(\boldsymbol{I}-\boldsymbol{H})' \\
&= \sigma^2(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{I}_n(\boldsymbol{I}-\boldsymbol{H})' \\
&= \sigma^2(\boldsymbol{I}-\boldsymbol{H})
\end{aligned}
$$

- 由此，我们可以构造误差项方差 σ^2 的估计，即

$$
\hat{\sigma}^2 = \frac{1}{n-p-1}(\boldsymbol{e}'\boldsymbol{e}) = \frac{1}{n-p-1}\sum_{i=1}^{n} e_i^2
$$

极大似然估计

基本思想

- 极大似然估计依赖于误差向量的正态分布假定. 于是, $\boldsymbol{y}$ 的分布为

$$\boldsymbol{y} \sim N_n(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}_n)$$

- $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)'$ 的联合密度函数为

$$\begin{aligned} f(\boldsymbol{y}; \boldsymbol{\beta}, \sigma^2) &= \frac{1}{(2\pi)^{n/2} \left|\sigma^2 \boldsymbol{I}_n\right|^{1/2}} \cdot \\ &\quad \exp\left\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\beta})' \left(\sigma^2 \boldsymbol{I}_n\right)^{-1} (\boldsymbol{y} - \boldsymbol{X\beta})\right\}. \end{aligned}$$

- 参数 $(\boldsymbol{\beta}, \sigma^2)$ 的似然函数为

$$L(\boldsymbol{\beta}, \sigma^2) = (2\pi)^{-n/2}(\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})\right\}.$$

极大似然估计

基本思想

极大似然估计：通过最大化似然函数而得到的估计方法。

- 极大似然估计

$$
\begin{aligned}
(\hat{\boldsymbol{\beta}}_{\mathrm{ML}}, \hat{\sigma}^2_{\mathrm{ML}}) &= \arg\max_{(\boldsymbol{\beta},\sigma^2)} L(\boldsymbol{\beta}, \sigma^2) \\
&= \arg\max_{(\boldsymbol{\beta},\sigma^2)} \ln\big(L(\boldsymbol{\beta}, \sigma^2)\big)
\end{aligned}
$$

极大似然估计

具体计算方法

- 对数似然函数为

$$\ln L(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

- 对数似然函数分布关于 $\boldsymbol{\beta}$ 和 σ^2 求偏导，即

$$\begin{cases} \dfrac{\partial \ln L(\boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta}} = -\dfrac{1}{\sigma^2}(\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}'\boldsymbol{y}) = 0 \\ \dfrac{\partial \ln L(\boldsymbol{\beta}, \sigma^2)}{\partial \sigma^2} = -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = 0 \end{cases}$$

极大似然估计

具体计算方法

- 极大似然估计为

$$\begin{cases} \hat{\boldsymbol{\beta}}_{\mathrm{ML}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\ \hat{\sigma}^2_{\mathrm{ML}} = \dfrac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{ML}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{ML}}) = \dfrac{1}{n}\boldsymbol{e}'\boldsymbol{e} \end{cases}$$

说明

- $\hat{\boldsymbol{\beta}}_{\mathrm{ML}} = \hat{\boldsymbol{\beta}}_{\mathrm{LS}}$，一般记为 $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$;
- $\hat{\sigma}^2_{\mathrm{ML}}$ 不是一个无偏估计，但是相合估计。

参数估计的性质

概率论复习

假设 $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)'$ 和 $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)'$ 均是 n 维随机变量。对于任意一个 $m \times n$ 维常矩阵 $\boldsymbol{A} = \{a_{ij}\}_{m\times n}$ 和一个 $m' \times n$ 维常矩阵 $\boldsymbol{B} = \{b_{ij}\}_{m'\times n}$，以及一个 m 维的常向量 $\boldsymbol{c} = (c_1, c_2, \cdots, c_m)'$，我们有

- $E(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{c}) = \boldsymbol{A}E(\boldsymbol{x}) + \boldsymbol{c}$;
- $\mathrm{Var}(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{c}) = \boldsymbol{A}\mathrm{Var}(\boldsymbol{x})\boldsymbol{A}'$
- $\mathrm{Cov}(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{B}\boldsymbol{y}) = \boldsymbol{A}\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y})\boldsymbol{B}'$

参数估计的性质

定理 1-2 (最小二乘估计的性质)

最小二乘估计 $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$，那么 $\hat{\boldsymbol{\beta}}$ 满足

- $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$，即 $\hat{\boldsymbol{\beta}}$ 是 $\boldsymbol{\beta}$ 的无偏估计；
- $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

证明：

- 我们计算 $\hat{\boldsymbol{\beta}}$ 的期望，即

$$\begin{aligned}
E(\hat{\boldsymbol{\beta}}) &= E((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{y}) \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + E(\boldsymbol{\varepsilon})) \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}
\end{aligned}$$

  因此，$\hat{\boldsymbol{\beta}}$ 是 $\boldsymbol{\beta}$ 的无偏估计。

参数估计的性质

- 我们计算 $\hat{\boldsymbol{\beta}}$ 的方差，即

$$
\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}) &= \mathrm{Var}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}) \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{Var}(\boldsymbol{y})((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')' \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{Var}(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')' \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\sigma^2\boldsymbol{I}_n)\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}
\end{aligned}
$$

参数估计的性质

- $\boldsymbol{\beta}$ 的最小二乘估计 $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ 是一个随机变量；
- 残差 $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{y}$ 是一个随机变量；

问题：这两个随机变量是否有关?

定理 1-3

最小二乘估计 $\hat{\boldsymbol{\beta}}$ 与残差 $\boldsymbol{e}$ 线性不相关，即

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}, \boldsymbol{e}) = \boldsymbol{0}.$$

说明：

- 特别地，在正态分布的假定下，最小二乘估计 $\hat{\boldsymbol{\beta}}$ 与残差 $\boldsymbol{e}$ 独立。基于此，最小二乘估计 $\hat{\boldsymbol{\beta}}$ 与残差平方和 $SS_E = \boldsymbol{e}'\boldsymbol{e}$ 独立。

参数估计的性质

- $\boldsymbol{\beta}$ 的最小二乘估计 $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ 是一个随机变量；
- 残差 $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{y}$ 是一个随机变量；

问题：这两个随机变量是否有关?

定理 1-3

最小二乘估计 $\hat{\boldsymbol{\beta}}$ 与残差 $\boldsymbol{e}$ 线性不相关，即

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}, \boldsymbol{e}) = \mathbf{0}.$$

说明：

- 特别地，在正态分布的假定下，最小二乘估计 $\hat{\boldsymbol{\beta}}$ 与残差 $\boldsymbol{e}$ 独立。基于此，最小二乘估计 $\hat{\boldsymbol{\beta}}$ 与残差平方和 $SS_E = \boldsymbol{e}'\boldsymbol{e}$ 独立。

参数估计的性质

证明： 由于

$$
\begin{aligned}
\mathrm{Cov}(\hat{\boldsymbol{\beta}}, \boldsymbol{e}) &= \mathrm{Cov}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}, (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}) \\
&= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') \\
&= \sigma^2 \cdot 0 \\
&= 0
\end{aligned}
$$

因此，最小二乘估计 $\hat{\boldsymbol{\beta}}$ 和残差 $\boldsymbol{e}$ 线性不相关。

中心化

矩阵的知识（补充）

假定 $\boldsymbol{A}$ 是 $m \times m$ 可逆矩阵，$\boldsymbol{B}$ 是 $m \times n$ 矩阵，$\boldsymbol{C}$ 是 $n \times m$ 矩阵，$\boldsymbol{D}$ 是 $n \times n$ 矩阵。如果 $\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}$ 是 $n \times n$ 可逆矩阵，那么

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{E}_{11} & \boldsymbol{E}_{12} \\ \boldsymbol{E}_{21} & \boldsymbol{E}_{22} \end{pmatrix}$$

其中，

$$\begin{aligned}
\boldsymbol{E}_{11} &= \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} \\
\boldsymbol{E}_{12} &= -\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1} \\
\boldsymbol{E}_{21} &= -(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} \\
\boldsymbol{E}_{22} &= (\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}
\end{aligned}$$

中心化

回顾

- 原本的模型:
$$y = \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon$$
即
$$y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \varepsilon.$$
- $\boldsymbol{\beta}$ 的估计记为 $\hat{\boldsymbol{\beta}} = (\hat{\beta}_{\text{intercep}}, \hat{\boldsymbol{\beta}}_{\text{slope}})'$.
- 经验回归方程为
$$\hat{y} = \hat{\beta}_{\text{intercept}} + \boldsymbol{x}'\hat{\boldsymbol{\beta}}_{\text{slope}}$$

中心化

- 原始数据集为

$$\begin{cases} \boldsymbol{y} = (y_1, \cdots, y_n)' \\ \boldsymbol{X} = (\boldsymbol{1}_n, \boldsymbol{X}_{\mathrm{o}}), \quad \boldsymbol{X}_{\mathrm{o}} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p) \end{cases}$$

- 最小二乘估计为

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\ &= \begin{pmatrix} n & \boldsymbol{1}_n'\boldsymbol{X}_{\mathrm{o}} \\ \boldsymbol{X}_{\mathrm{o}}'\boldsymbol{1}_n & \boldsymbol{X}_{\mathrm{o}}'\boldsymbol{X}_{\mathrm{o}} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{1}_n' \\ \boldsymbol{X}_{\mathrm{o}}' \end{pmatrix} \boldsymbol{y} \\ &= \begin{pmatrix} n^{-1} + n^{-2}\boldsymbol{1}_n'\boldsymbol{X}_{\mathrm{o}}\boldsymbol{A}_{\mathrm{o}}\boldsymbol{X}_{\mathrm{o}}'\boldsymbol{1}_n & -n^{-1}\boldsymbol{1}_n'\boldsymbol{X}_{\mathrm{o}}\boldsymbol{A}_{\mathrm{o}} \\ -n^{-1}\boldsymbol{A}_{\mathrm{o}}\boldsymbol{X}_{\mathrm{o}}'\boldsymbol{1}_n & \boldsymbol{A}_{\mathrm{o}} \end{pmatrix} \begin{pmatrix} \boldsymbol{1}_n' \\ \boldsymbol{X}_{\mathrm{o}}' \end{pmatrix} \boldsymbol{y} \\ &= \begin{pmatrix} n^{-1}\boldsymbol{1}_n' + n^{-2}\boldsymbol{1}_n'\boldsymbol{X}_{\mathrm{o}}\boldsymbol{A}_{\mathrm{o}}\boldsymbol{X}_{\mathrm{o}}'\boldsymbol{1}_n\boldsymbol{1}_n' - n^{-1}\boldsymbol{1}_n'\boldsymbol{X}_{\mathrm{o}}\boldsymbol{A}_{\mathrm{o}}\boldsymbol{X}_{\mathrm{o}}' \\ -n^{-1}\boldsymbol{A}_{\mathrm{o}}\boldsymbol{X}_{\mathrm{o}}'\boldsymbol{1}_n\boldsymbol{1}_n' + \boldsymbol{A}_{\mathrm{o}}\boldsymbol{X}_{\mathrm{o}}' \end{pmatrix} \boldsymbol{y} \end{aligned}$$

其中，$\boldsymbol{A}_{\mathrm{o}} = (\boldsymbol{X}_{\mathrm{o}}'\boldsymbol{X}_{\mathrm{o}} - n^{-1}\boldsymbol{X}_{\mathrm{o}}'\boldsymbol{1}_n\boldsymbol{1}_n'\boldsymbol{X}_{\mathrm{o}})^{-1}$

中心化

- 中心化:

$$
\begin{aligned}
x_{ij}^* &= x_{ij} - \bar{x}_j, \quad \bar{x}_j = n^{-1} \sum_{i=1}^n x_{ij} \\
y_i^* &= y_i - \bar{y}, \quad \bar{y} = n^{-1} \sum_{i=1}^n y_i
\end{aligned}
$$

令

$$
\begin{cases}
\boldsymbol{y}^* = (y_1^*, \cdots, y_n^*)' \\
\boldsymbol{X}_{\mathrm{c}} = (\boldsymbol{x}_1^*, \cdots, \boldsymbol{x}_p^*) \\
\boldsymbol{X}^* = (\mathbf{1}_n, \boldsymbol{X}_{\mathrm{c}})
\end{cases}
$$

其中，$\boldsymbol{x}_j^* = (x_{1j}^*, \cdots, x_{nj}^*)'$

中心化

- 那么，基于 $\boldsymbol{y}^*$ 和 $\boldsymbol{X}^*$，最小二乘估计为

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_{\mathrm{c}} &= ((\boldsymbol{X}^*)'\boldsymbol{X}^*)^{-1}(\boldsymbol{X}^*)'\boldsymbol{y}^* \\
&= \begin{pmatrix} n & \mathbf{1}_n'\boldsymbol{X}_{\mathrm{c}} \\ \boldsymbol{X}_{\mathrm{c}}'\mathbf{1}_n & \boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{1}_n' \\ \boldsymbol{X}_{\mathrm{c}}' \end{pmatrix} \boldsymbol{y}^* \\
&= \begin{pmatrix} n^{-1} + n^{-2}\mathbf{1}_n'\boldsymbol{X}_{\mathrm{c}}\boldsymbol{A}_{\mathrm{c}}\boldsymbol{X}_{\mathrm{c}}'\mathbf{1}_n & -n^{-1}\mathbf{1}_n'\boldsymbol{X}_{\mathrm{c}}\boldsymbol{A}_{\mathrm{c}} \\ -n^{-1}\boldsymbol{A}_{\mathrm{c}}\boldsymbol{X}_{\mathrm{c}}'\mathbf{1}_n & \boldsymbol{A}_{\mathrm{c}} \end{pmatrix} \begin{pmatrix} \mathbf{1}_n' \\ \boldsymbol{X}_{\mathrm{c}}' \end{pmatrix} \boldsymbol{y}^* \\
&= \begin{pmatrix} n^{-1}\mathbf{1}_n' + n^{-2}\mathbf{1}_n'\boldsymbol{X}_{\mathrm{c}}\boldsymbol{A}_{\mathrm{c}}\boldsymbol{X}_{\mathrm{c}}'\mathbf{1}_n\mathbf{1}_n' - n^{-1}\mathbf{1}_n'\boldsymbol{X}_{\mathrm{c}}\boldsymbol{A}_{\mathrm{c}}\boldsymbol{X}_{\mathrm{c}}' \\ -n^{-1}\boldsymbol{A}_{\mathrm{c}}\boldsymbol{X}_{\mathrm{c}}'\mathbf{1}_n\mathbf{1}_n' + \boldsymbol{A}_{\mathrm{c}}\boldsymbol{X}_{\mathrm{c}}' \end{pmatrix} \boldsymbol{y}^*
\end{aligned}
$$

其中，

$$
\boldsymbol{A}_{\mathrm{c}} = (\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}} - n^{-1}\boldsymbol{X}_{\mathrm{c}}'\mathbf{1}_n\mathbf{1}_n'\boldsymbol{X}_{\mathrm{c}})^{-1}
$$

中心化

- 中心化的因变量与未中心化的因变量之间的关系:

$$\boldsymbol{y}^* = \boldsymbol{y} - \mathbf{1}_n(\mathbf{1}_n'\mathbf{1}_n)^{-1}\mathbf{1}_n'\boldsymbol{y} = (\boldsymbol{I}_n - \boldsymbol{H}_{\mathbf{1}_n})\boldsymbol{y},$$

  其中 $\boldsymbol{H}_{\mathbf{1}_n} = \mathbf{1}_n(\mathbf{1}_n'\mathbf{1}_n)^{-1}\mathbf{1}_n'$ 是对称幂等矩阵。

- 中心化的自变量与未中心化的自变量之间的关系:

$$\begin{aligned}\boldsymbol{X}_{\mathrm{c}} &= \boldsymbol{X}_{\mathrm{o}} - \mathbf{1}_n(\mathbf{1}_n'\mathbf{1}_n)^{-1}\mathbf{1}_n'\boldsymbol{X}_{\mathrm{o}} = (\boldsymbol{I}_n - \mathbf{1}_n(\mathbf{1}_n'\mathbf{1}_n)^{-1}\mathbf{1}_n')\boldsymbol{X}_{\mathrm{o}} \\ &= (\boldsymbol{I}_n - \boldsymbol{H}_{\mathbf{1}_n})\boldsymbol{X}_{\mathrm{o}}\end{aligned}$$

  而且

$$\begin{aligned}\mathbf{1}_n'(\boldsymbol{I}_n - \boldsymbol{H}_{\mathbf{1}_n}) &= \mathbf{1}_n' - \mathbf{1}_n'\boldsymbol{H}_{\mathbf{1}_n} \\ &= \mathbf{1}_n' - \mathbf{1}_n'\mathbf{1}_n(\mathbf{1}_n'\mathbf{1}_n)^{-1}\mathbf{1}_n' = 0\end{aligned}$$

中心化

- 最小二乘估计分为回归常数和回归系数，即

$$\hat{\boldsymbol{\beta}}_{\mathrm{c}} = (\hat{\beta}_{\mathrm{c,intercept}}, \hat{\boldsymbol{\beta}}_{\mathrm{c,slope}})'.$$

- 一方面，回归常数为

$$\begin{aligned}
\hat{\beta}_{\mathrm{c,intercept}} &= (n^{-1}\mathbf{1}_n' + n^{-2}\mathbf{1}_n'\boldsymbol{X}_{\mathrm{c}}\boldsymbol{A}_{\mathrm{c}}\boldsymbol{X}_{\mathrm{c}}'\mathbf{1}_n\mathbf{1}_n' \\
&\quad -n^{-1}\mathbf{1}_n'\boldsymbol{X}_{\mathrm{c}}\boldsymbol{A}_{\mathrm{c}}\boldsymbol{X}_{\mathrm{c}}')\boldsymbol{y}^* \\
&= n^{-1}\mathbf{1}_n'\boldsymbol{y}^* \\
&= n^{-1}\mathbf{1}_n'(\boldsymbol{I}_n - \boldsymbol{H}_{\mathbf{1}_n})\boldsymbol{y} = 0
\end{aligned}$$

中心化

- 由于

$$
\begin{aligned}
\boldsymbol{A}_{\mathrm{c}} &= (\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}} - n^{-1}\boldsymbol{X}_{\mathrm{c}}'\mathbf{1}_n\mathbf{1}_n'\boldsymbol{X}_{\mathrm{c}})^{-1} \\
&= \Big(\boldsymbol{X}_{\mathrm{o}}'(\boldsymbol{I}_n - \boldsymbol{H}_{\mathbf{1}_n})\boldsymbol{X}_{\mathrm{o}} \\
&\quad -n^{-1}\boldsymbol{X}_{\mathrm{c}}'(\boldsymbol{I}_n - \boldsymbol{H}_{\mathbf{1}_n})\mathbf{1}_n\mathbf{1}_n'(\boldsymbol{I}_n - \boldsymbol{H}_{\mathbf{1}_n})\boldsymbol{X}_{\mathrm{o}}\Big)^{-1} \\
&= (\boldsymbol{X}_{\mathrm{o}}'(\boldsymbol{I}_n - \boldsymbol{H}_{\mathbf{1}_n})\boldsymbol{X}_{\mathrm{o}})^{-1} = \boldsymbol{A}_{\mathrm{o}}
\end{aligned}
$$

- 另一方面，回归系数为

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_{\mathrm{c,slope}} &= (-n^{-1}\boldsymbol{A}_{\mathrm{c}}\boldsymbol{X}_{\mathrm{c}}'\mathbf{1}_n\mathbf{1}_n' + \boldsymbol{A}_{\mathrm{c}}\boldsymbol{X}_{\mathrm{c}}')\boldsymbol{y}^* \\
&= \boldsymbol{A}_{\mathrm{c}}\boldsymbol{X}_{\mathrm{c}}'(\boldsymbol{I}_n - \boldsymbol{H}_{\mathbf{1}_n})\boldsymbol{y} \\
&= \boldsymbol{A}_{\mathrm{o}}\boldsymbol{X}_{\mathrm{o}}'(\boldsymbol{I}_n - \boldsymbol{H}_{\mathbf{1}_n})\boldsymbol{y}
\end{aligned}
$$

中心化

- 而

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_{\text{slope}} &= (-n^{-1}\boldsymbol{A}_{\text{o}}\boldsymbol{X}_{\text{o}}'\mathbf{1}_n\mathbf{1}_n' + \boldsymbol{A}_{\text{o}}\boldsymbol{X}_{\text{o}}')\boldsymbol{y} \\
&= \boldsymbol{A}_{\text{o}}\boldsymbol{X}_{\text{o}}'(\boldsymbol{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n')\boldsymbol{y} \\
&= \boldsymbol{A}_{\text{o}}\boldsymbol{X}_{\text{o}}'(\boldsymbol{I}_n - \boldsymbol{H}_{\mathbf{1}_n})\boldsymbol{y}
\end{aligned}
$$

- 于是，

$$\hat{\boldsymbol{\beta}}_{\text{c,slope}} = \hat{\boldsymbol{\beta}}_{\text{slope}}$$

- 采用中心化的数据得到的经验回归方程为

$$\hat{y}^* = (\boldsymbol{x}^*)'\hat{\boldsymbol{\beta}}_{\text{slope}}$$

标准化

- 标准化:

$$x_{ij}^{**} = \frac{x_{ij}^*}{\sqrt{L_{jj}}} = \frac{x_{ij} - \bar{x}_j}{\sqrt{L_{jj}}}, \quad i = 1, 2, \cdots, n; j = 1, 2, \cdots, p.$$

$$y_i^{**} = \frac{y_i^*}{\sqrt{L_{yy}}}, \quad i = 1, 2, \cdots, n.$$

其中，L_{jj} 是自变量 x_j 的离差平方和，即

$$L_{jj} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

而 L_{yy} 是因变量 y 的离差平方和，即

$$L_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

标准化

- 令

$$
\begin{aligned}
\boldsymbol{y}^{**} &= \left(\frac{y_1 - \bar{y}}{\sqrt{L_{yy}}}, \cdots, \frac{y_n - \bar{y}}{\sqrt{L_{yy}}}\right)' = \frac{1}{\sqrt{L_{yy}}}\boldsymbol{y}^* \\
\boldsymbol{X}_\mathrm{s} &= \left(\frac{1}{\sqrt{L_{11}}}\boldsymbol{x}_1^*, \cdots, \frac{1}{\sqrt{L_{pp}}}\boldsymbol{x}_p^*\right) \\
&= \boldsymbol{X}_c\boldsymbol{L}.
\end{aligned}
$$

其中，

$$
\boldsymbol{L} = \mathrm{diag}\left\{\frac{1}{\sqrt{L_{11}}}, \cdots, \frac{1}{\sqrt{L_{pp}}}\right\}
$$

标准化

- 最小二乘估计为

$$\hat{\boldsymbol{\beta}}_{\mathrm{s}} = (\hat{\boldsymbol{\beta}}_{\mathrm{s,intercept}}, \hat{\boldsymbol{\beta}}_{\mathrm{s,slope}})' = (0, \hat{\boldsymbol{\beta}}_{\mathrm{s,slope}})'.$$

- 回归系数为

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\mathrm{s,slope}} &= (\boldsymbol{X}_{\mathrm{s}}'\boldsymbol{X}_{\mathrm{s}})^{-1}\boldsymbol{X}_{\mathrm{s}}'\boldsymbol{y}^{**} \\
&= (\boldsymbol{L}\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}}\boldsymbol{L})^{-1}\boldsymbol{L}\boldsymbol{X}_{\mathrm{c}}'\frac{1}{\sqrt{L_{yy}}}\boldsymbol{y}^{*} \\
&= \boldsymbol{L}^{-1}(\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}})^{-1}\boldsymbol{L}^{-1}\boldsymbol{L}\boldsymbol{X}_{\mathrm{c}}'\frac{1}{\sqrt{L_{yy}}}\boldsymbol{y}^{*} \\
&= \frac{1}{\sqrt{L_{yy}}}\boldsymbol{L}^{-1}(\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}})^{-1}\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{y}^{*}
\end{aligned}$$

标准化

- 注意到

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\mathrm{c,slope}} &= (-n^{-1}\boldsymbol{A}_{\mathrm{c}}\boldsymbol{X}_{\mathrm{c}}'\mathbf{1}_n\mathbf{1}_n' + \boldsymbol{A}_{\mathrm{c}}\boldsymbol{X}_{\mathrm{c}}')\boldsymbol{y}^* \\ &= \boldsymbol{A}_{\mathrm{c}}\boldsymbol{X}_{\mathrm{c}}'(\boldsymbol{I}_n - \boldsymbol{H}_{\mathbf{1}_n})\boldsymbol{y}^* \\ &= (\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}})^{-1}\boldsymbol{X}_{\mathrm{c}}'(\boldsymbol{I}_n - \boldsymbol{H}_{\mathbf{1}_n})(\boldsymbol{I}_n - \boldsymbol{H}_{\mathbf{1}_n})\boldsymbol{y} \\ &= (\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{X}_{\mathrm{c}})^{-1}\boldsymbol{X}_{\mathrm{c}}'\boldsymbol{y}^*\end{aligned}$$

- 因此，

$$\hat{\boldsymbol{\beta}}_{\mathrm{s,slope}} = \frac{1}{\sqrt{L_{yy}}}\boldsymbol{L}^{-1}\hat{\boldsymbol{\beta}}_{\mathrm{c,slope}}$$

其中每一个分量为

$$\hat{\beta}_{\mathrm{s}j} = \frac{\sqrt{L_{jj}}}{\sqrt{L_{yy}}}\hat{\beta}_{\mathrm{c}j} = \frac{\sqrt{L_{jj}}}{\sqrt{L_{yy}}}\hat{\beta}_j, \quad j = 1, 2, \cdots, p.$$

统计与机器学习

第一章：线性回归 - Part II

倪 葎

DaSE@ECNU
(lni@dase.ecnu.edu.cn)

2020 年 10 月 2 日

目录

显著性检验

概述

- 考虑线性回归模型为

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

- 在实际问题中，我们进一步判断因变量 y 和自变量 $x_1, x_2, \cdots, x_p$ 之间是否存在显著的线性关系.
- 在统计学中，我们可以采用**显著性检验**来解决这一问题. 在多元线性回归模型中，有两种统计检验方法:
  - F 检验：用于检验**回归方程**的显著性;
  - t 检验：用于检验**回归系数**的显著性;
- 除了显著性检验，我们将介绍常用的指标用于衡量线性回归的拟合优度.

F 检验

原假设与备择假设

- 对多元线性回归方程的显著性检验是要看自变量 $x_1, x_2, \cdots, x_p$ 从整体上对因变量 y 是否有明显的影响.
- 原假设为

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0,$$

- 备择假设为

$$H_1 : \text{存在 } \beta_j \text{ 不为零，} \ j = 1, 2, \cdots, p.$$

- 如果 H_0 为真，则表明因变量 y 与 $x_1, x_2, \cdots, x_p$ 之间的关系用线性回归模型来刻画是不合适的.

F 检验

检验过程

- 离差平方和为

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

  记为

$$SS_T = SS_R + SS_E.$$

- 拟合值 $\hat{y}_i = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$;
- 偏差 $e_i = y_i - \hat{y}_i$;

F 检验

- 检验统计量

$$F_0 = \frac{SS_R/p}{SS_E/(n-p-1)}$$

定理 1-4

在正态假设下，即 $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$，有

- $SS_E/\sigma^2 \sim \chi^2(n-p-1)$，其中 $SS_E = \boldsymbol{e}'\boldsymbol{e}$；
- SS_E 和 SS_R 独立；
- 在 H_0 下，$SS_R \sim \chi^2(p)$；

F 检验

多元统计的知识（补充）

假设 n 维随机变量 $\boldsymbol{x} \sim N_n(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_n)$.

- 如果 $\boldsymbol{C}$ 是一个对称矩阵，且 $\mathrm{rank}(\boldsymbol{C}) = r$，那么二次型

  $$\boldsymbol{x}'\boldsymbol{C}\boldsymbol{x}/\sigma^2 \sim \chi^2(r, \delta)$$

  其中

  $$\delta = \frac{1}{\sigma^2}\boldsymbol{\mu}'\boldsymbol{C}\boldsymbol{\mu}$$

  当且仅当

  $$\boldsymbol{C}^2 = \boldsymbol{C} \quad 且 \quad \mathrm{rank}(\boldsymbol{C}) = r (r \leq n)$$

F 检验

多元统计的知识（补充）

假设 n 维随机变量 $\boldsymbol{x} \sim N_n(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_n)$.

- $\boldsymbol{A}$ 为 n 阶对称矩阵，$\boldsymbol{B}$ 为 $m \times n$ 矩阵，那么，

$$\boldsymbol{B}\boldsymbol{A} = \boldsymbol{0}_{m \times n}$$

  当且仅当 $\boldsymbol{B}\boldsymbol{x}$ 和 $\boldsymbol{x}\boldsymbol{A}\boldsymbol{x}$ 相互独立.

- $\boldsymbol{A}, \boldsymbol{B}$ 为 n 阶对称矩阵，则

$$\boldsymbol{A}\boldsymbol{B} = \boldsymbol{0}_{n \times n}$$

  当且仅当 $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$ 和 $\boldsymbol{x}'\boldsymbol{B}\boldsymbol{x}$ 相互独立.

F 检验

证明：

- 由于残差
$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$$
那么残差平方和为
$$SS_E = \boldsymbol{e}'\boldsymbol{e} = \boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$$
由于 $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I}_n)$，而且 $(\boldsymbol{I} - \boldsymbol{H})$ 是对称幂等矩阵，且其秩为 $\mathrm{rank}(\boldsymbol{I} - \boldsymbol{H}) = n - (p+1)$. 我们可以计算
$$\frac{1}{\sigma^2}E(\boldsymbol{y})'(\boldsymbol{I} - \boldsymbol{H})E(\boldsymbol{y}) = \frac{1}{\sigma^2}\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X\beta} = 0.$$
因此，$SS_E/\sigma^2 \sim \chi^2(n-p-1)$.

F 检验

- 由于

$$\begin{aligned} SS_R &= \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 = \sum_{i=1}^{n}(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}-\bar{\boldsymbol{x}}'\hat{\boldsymbol{\beta}})^2 \\ &= \hat{\boldsymbol{\beta}}'\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})'\hat{\boldsymbol{\beta}} \\ &= \boldsymbol{y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \end{aligned}$$

其中，$\boldsymbol{A}=\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})'$，而 $SS_E=\boldsymbol{e}'\boldsymbol{e}=\boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}$. 易知，

$$\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{H})=0$$

那么，SS_R 和 SS_E 相互独立.

F 检验

- 因为 $\bar{\boldsymbol{x}}' = (\mathbf{1}_n'\mathbf{1}_n)^{-1}\mathbf{1}_n'\boldsymbol{X}$，我们有

$$\begin{aligned}
\boldsymbol{A} &= \sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})' \\
&= \sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i' - n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}' \\
&= \boldsymbol{X}'\boldsymbol{X} - (\mathbf{1}_n'\mathbf{1}_n)\boldsymbol{X}'\mathbf{1}_n(\mathbf{1}_n'\mathbf{1}_n)^{-1}(\mathbf{1}_n'\mathbf{1}_n)^{-1}\mathbf{1}_n'\boldsymbol{X} \\
&= \boldsymbol{X}'\boldsymbol{X} - \boldsymbol{X}'\mathbf{1}_n(\mathbf{1}_n'\mathbf{1}_n)^{-1}\mathbf{1}_n'\boldsymbol{X} \\
&= \boldsymbol{X}'(\boldsymbol{I}_n - \boldsymbol{H}_{\mathbf{1}_n})\boldsymbol{X}
\end{aligned}$$

所以，

$$\begin{aligned}
SS_R &= \boldsymbol{y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{I}_n - \boldsymbol{H}_{\mathbf{1}_n})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\
&= \boldsymbol{y}'\boldsymbol{H}(\boldsymbol{I}_n - \boldsymbol{H}_{\mathbf{1}_n})\boldsymbol{H}\boldsymbol{y}.
\end{aligned}$$

F 检验

由于 $\boldsymbol{X} = (\mathbf{1}_n, \boldsymbol{X}_\text{o})$，因此，注意到

$$\begin{aligned}
\boldsymbol{H}\boldsymbol{H}_{\mathbf{1}_n} &= (\mathbf{1}_n \quad \boldsymbol{X}_\text{o}) \begin{pmatrix} n^{-1} + n^{-2}\mathbf{1}_n'\boldsymbol{X}_\text{o}\boldsymbol{A}_\text{o}\boldsymbol{X}_\text{o}'\mathbf{1}_n & -n^{-1}\mathbf{1}_n'\boldsymbol{X}_\text{o}\boldsymbol{A}_\text{o} \\ -n^{-1}\boldsymbol{A}_\text{o}\boldsymbol{X}_\text{o}'\mathbf{1}_n & \boldsymbol{A}_\text{o} \end{pmatrix} \begin{pmatrix} \mathbf{1}_n' \\ \boldsymbol{X}_\text{o}' \end{pmatrix} (n^{-1}\mathbf{1}_n\mathbf{1}_n') \\
&= (n^{-1}\mathbf{1}_n\mathbf{1}_n' + n^{-2}\mathbf{1}_n\mathbf{1}_n'\boldsymbol{X}_\text{o}\boldsymbol{A}_\text{o}\boldsymbol{X}_\text{o}'\mathbf{1}_n\mathbf{1}_n' - n^{-1}\boldsymbol{X}_\text{o}\boldsymbol{A}_\text{o}\boldsymbol{X}_\text{o}'\mathbf{1}_n\mathbf{1}_n' \\
&\quad -n^{-1}\mathbf{1}_n\mathbf{1}_n'\boldsymbol{X}_\text{o}\boldsymbol{A}_\text{o}\boldsymbol{X}_\text{o}' + \boldsymbol{X}_\text{o}\boldsymbol{A}_\text{o}\boldsymbol{X}_\text{o}')(n^{-1}\mathbf{1}_n\mathbf{1}_n') \\
&= (-n^{-1}\mathbf{1}_n\mathbf{1}_n')^2 \\
&= \boldsymbol{H}_{\mathbf{1}_n}
\end{aligned}$$

于是，

$$\begin{aligned}
SS_R &= \boldsymbol{y}'\boldsymbol{H}(\boldsymbol{I}_n - \boldsymbol{H}_{\mathbf{1}_n})\boldsymbol{H}\boldsymbol{y} \\
&= \boldsymbol{y}'(\boldsymbol{H} - \boldsymbol{H}_{\mathbf{1}_n})\boldsymbol{y}.
\end{aligned}$$

F 检验

令 $\boldsymbol{B} = (\boldsymbol{H} - \boldsymbol{H}_{\mathbf{1}_n})$. 易证 $\boldsymbol{B}$ 是对称幂等矩阵，且 $\text{rank}(\boldsymbol{B}) = \text{tr}(\boldsymbol{B}) = \text{tr}(\boldsymbol{H}) - \text{tr}(\boldsymbol{H}_{\mathbf{1}_n}) = (p+1) - 1 = p$.
在原假设成立时，$\beta_1 = \beta_2 = \cdots = \beta_p = 0$. 于是，

$$\boldsymbol{y} \sim N(\beta_0 \mathbf{1}_n, \sigma^2 \boldsymbol{I}).$$

根据多元统计的知识可知，

$$\frac{SS_R}{\sigma^2} = \frac{\boldsymbol{y}'\boldsymbol{B}\boldsymbol{y}}{\sigma^2} \overset{H_0}{\sim} \chi^2(p, \delta)$$

其中

$$\begin{aligned}
\delta &= \frac{1}{\sigma^2}(\beta_0 \mathbf{1}_n)'\boldsymbol{B}(\beta_0 \mathbf{1}_n) = \frac{\beta_0^2}{\sigma^2}\mathbf{1}_n'(\boldsymbol{H} - \boldsymbol{H}_{\mathbf{1}_n})\mathbf{1}_n \\
&= \frac{\beta_0^2}{\sigma^2}\text{tr}(\mathbf{1}_n'(\boldsymbol{H} - \boldsymbol{H}_{\mathbf{1}_n})\mathbf{1}_n) = \frac{\beta_0^2}{\sigma^2}\text{tr}((\boldsymbol{H} - \boldsymbol{H}_{\mathbf{1}_n})\mathbf{1}_n\mathbf{1}_n') \\
&= \frac{\beta_0^2}{\sigma^2}\text{tr}(n(\boldsymbol{H} - \boldsymbol{H}_{\mathbf{1}_n})\boldsymbol{H}_{\mathbf{1}_n}) = 0
\end{aligned}$$

F 检验

结论

- 那么，检验统计量

$$F_0 = \frac{SS_R/p}{SS_E/(n-p-1)} \overset{H_0}{\sim} F(p, n-p-1)$$

- 给定显著性水平 α，
  - 当 $F_0 > F_{1-\alpha}(p, n-p-1)$，拒绝原假设 H_0，或
  - 当 $p_0 = P(F \geq F_0) < \alpha$，拒绝原假设 H_0.

- 通常我们可以借助方差分析表来展示这一分析结果.

来源	平方和	自由度	均方	F 值	p 值
回归	SS_R	p	$\frac{SS_R}{p}$	$F_0 = \frac{\frac{SS_R}{p}}{\frac{SS_E}{(n-p-1)}}$	$p_0 = P(F \geq F_0)$
误差	SS_E	$n-p-1$	$\frac{SS_E}{(n-p-1)}$		
总和	SS_T	$n-1$			

F 检验

说明

- $SS_T = SS_R + SS_E$，即

$$\boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{H}_{\boldsymbol{1}_n})\boldsymbol{y} = \boldsymbol{y}'(\boldsymbol{H} - \boldsymbol{H}_{\boldsymbol{1}_n})\boldsymbol{y} + \boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{y}$$

- SS_R 与 SS_E 独立；

t 检验

动机

- 在多元线性回归中，回归方程显著并不意味着每个自变量对因变量的影响都显著.
- 我们希望从回归方程中剔除那些次要的、可有可无的自变量，建立更为简化的回归方程.
- 所以，我们需要对每个自变量进行显著性检验.
- 如果某个自变量 x_j 对 y 的作用不显著，那么在回归模型中，其对应的回归系数 β_j 为零.

t 检验

原假设与备择假设

- 因此，我们想要检验变量 x_j 是否显著?
- 原假设为

$$H_{0j}: \beta_j = 0, \quad j = 1, 2, \cdots, p$$

- 备择假设为

$$H_{1j}: \text{至少存在 } j \text{ 使得 } \beta_j \neq 0.$$

- 如果我们拒绝原假设 H_{0j}，那么我们认为自变量 x_j 对因变量 y 显著.

多元统计知识（补充）

- 假设 n 维随机变量 $\boldsymbol{x} \sim N_n(\boldsymbol{\mu}, \Sigma)$。
  对于任何 $m \times n$ 常数矩阵 $\boldsymbol{A}$ 和 m 维常向量 $\boldsymbol{b}$，我们有

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} \sim N_m(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\Sigma\boldsymbol{A}').$$

参数估计的性质

定理 1-5

在正态假设下，即 $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}_n)$，有

- $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$；

证明:

- 由于 $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$，其为 $\boldsymbol{y}$ 的线性组合，而且 $\boldsymbol{y}$ 服从多元正态分布。根据定理 1-2 可知，$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ 且 $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$. 因此，

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}).$$

多元统计的知识（补充）

假设随机向量 $\boldsymbol{x} = (x_1, x_2, \cdots, x_p)'$ 是一个 p 维正态分布 $N(\boldsymbol{\mu}, \Sigma)$，其中

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} = \begin{pmatrix} E(x_1) \\ E(x_2) \\ \vdots \\ E(x_p) \end{pmatrix} \quad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}$$

- 每一个分量 x_j 也服从正态分布，即

$$x_j \sim N(\mu_j, \sigma_{jj}).$$

t 检验

检验统计量

- 在正态假设下，根据定理 1-5 可知

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}).$$

- $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 是一个 $(p+1) \times (p+1)$ 矩阵;

- 这里 c_{ij} 表示 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 中第 $(i+1)$ 行第 $(j+1)$ 列元素，$i, j = 0, 1, \cdots, p$;

- 根据多元统计的知识，我们有

$$\hat{\beta}_j \sim N(\beta_j, c_{jj}\sigma^2), j = 0, 1, \cdots, p$$

t 检验

检验统计量

- 检验统计量为

$$t_j = \frac{\hat{\beta}_j}{\sqrt{c_{jj}}\hat{\sigma}}$$

其中,

$$\hat{\sigma}^2 = \frac{1}{n-p-1} SS_E$$

- 当原假设 $H_{0j}: \beta_j = 0$ 为真时,

$$t_j \sim t(n-p-1)$$

- 给定显著性水平 α，当 $|t_j| \geq t_{1-\alpha/2}(n-p-1)$ 时，拒绝原假设 $H_{0j}: \beta_j = 0$，并认为 β_j 显著不为零.

t 检验

说明

- 在一元线性回归中，回归系数显著性的 t 检验与回归方程显著性的 F 检验是等价的。

- 在多元线性回归中，这两种检验是不等价的。
  - F 检验显著，说明因变量 y 对自变量 $x_1, x_2, \cdots, x_p$ 整体的线性回归效果是显著的，但不等价于因变量 y 对每个自变量 x_j 的回归效果都显著。
  - 反之，某个或某几个 x_j 的系数不显著，回归方程显著性的 F 检验仍可能是显著的。

复相关系数

定义

- **拟合优度**可以用来度量回归方程对样本观测值的拟合程度。
- 在多元线性回归中，定义**样本决定系数**为

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

  - R^2 的取值在 $[0, 1]$ 区间内。
  - R^2 越接近 1，表明回归拟合的效果越好；
  - R^2 越接近 0，表明回归拟合的效果越差。

复相关系数

定义

- 定义 y 关于 $x_1, x_2, \cdots, x_p$ 的样本复相关系数为

$$R = \sqrt{R^2} = \sqrt{\frac{SS_R}{SS_T}}$$

- 与样本相关系数的区别:
  - 复相关系数的符号恒为正；相关系数的符号有正有负；
  - 复相关系数衡量作为一个整体的 $x_1, x_2, \cdots, x_p$ 与 y 的线性关系；
  - 相关系数衡量单个随机变量 x_j 与 y 的线性关系；

置信区间与预测

概述

- 给定 $\boldsymbol{x}_0 = (1, x_{01}, x_{02}, \cdots, x_{0p})'$，我们关心的是
$$y_0 = \boldsymbol{x}_0'\boldsymbol{\beta} + \varepsilon_0$$
- 基本假定：$E(\varepsilon_0) = 0$ 和 $\mathrm{Var}(\varepsilon_0) = \sigma^2$.
- 也就是说，$E(y_0) = \boldsymbol{x}_0'\boldsymbol{\beta}$ 和 $\mathrm{Var}(y_0) = \sigma^2$.
- 基于数据集 $\{(y_i, \boldsymbol{x}_i'), i = 1, 2, \cdots, n\}$，我们可以得到 $\boldsymbol{\beta}$ 的最小二乘估计为 $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p)'$.

置信区间与预测

点预测

- y_0 的预测值为

$$\hat{y}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + x_{01}\hat{\beta}_1 + \cdots + x_{0p}\hat{\beta}_p.$$

- 如何判断 $\hat{y}_0$ 的好坏?
- 我们可以比较 $\hat{y}_0$ 与 y_0 的差距，即考虑 $(\hat{y}_0 - y_0)^2$ 的大小。
- 由于 $\hat{y}_0$ 和 y_0 均是随机期望，我们一般考虑 $(\hat{y}_0 - y_0)^2$ 的均值，即 $E(\hat{y}_0 - y_0)^2$.

置信区间与预测

点预测

- y_0 的预测值为

$$\hat{y}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + x_{01}\hat{\beta}_1 + \cdots + x_{0p}\hat{\beta}_p.$$

- 如何判断 $\hat{y}_0$ 的好坏?
- 我们可以比较 $\hat{y}_0$ 与 y_0 的差距，即考虑 $(\hat{y}_0 - y_0)^2$ 的大小。
- 由于 $\hat{y}_0$ 和 y_0 均是随机期望，我们一般考虑 $(\hat{y}_0 - y_0)^2$ 的均值，即 $E(\hat{y}_0 - y_0)^2$.

置信区间与预测

点预测

由于

$$
\begin{aligned}
& (\hat{y}_0 - y_0)^2 \\
= \; & (\hat{y}_0 - E(\hat{y}_0) + E(\hat{y}_0) - E(y_0) + E(y_0) - y_0)^2 \\
= \; & (\hat{y}_0 - E(\hat{y}_0))^2 + 2(\hat{y}_0 - E(\hat{y}_0))(E(\hat{y}_0) - E(y_0)) \\
& + (E(\hat{y}_0) - E(y_0))^2 + 2(E(\hat{y}_0) - E(y_0))(E(y_0) - y_0) \\
& + (E(y_0) - y_0)^2 + 2(\hat{y}_0 - E(\hat{y}_0))(E(y_0) - y_0) \\
= \; & (\hat{y}_0 - E(\hat{y}_0))^2 + 2(\hat{y}_0 - E(\hat{y}_0))(E(\hat{y}_0) - E(y_0)) \\
& + (E(\hat{y}_0) - E(y_0))^2 + 2(E(\hat{y}_0) - E(y_0))(-\varepsilon_0) \\
& + \varepsilon_0^2 + 2(\hat{y}_0 - E(\hat{y}_0))(-\varepsilon_0)
\end{aligned}
$$

那么，

$$
E(\hat{y}_0 - y_0)^2 = \mathrm{Var}(\hat{y}_0) + \mathrm{Bias}^2(\hat{y}_0) + \mathrm{Var}(\varepsilon_0)
$$

置信区间与预测

点预测

- 考虑 $\mathrm{Bias}^2(\hat{y}_0) = 0$，即 $E(\hat{y}_0) = E(y_0)$?

定理 1-6

$\hat{y}_0$ 是 y_0 的无偏预测，即 $E(\hat{y}_0) = E(y_0)$.

证明： 根据最小二乘估计 $\hat{\boldsymbol{\beta}}$ 的无偏性，我们有

$$E(\hat{y}_0) = E(\boldsymbol{x}_0'\hat{\boldsymbol{\beta}}) = \boldsymbol{x}_0'E(\hat{\boldsymbol{\beta}}) = \boldsymbol{x}_0'\boldsymbol{\beta} = E(y_0).$$

置信区间与预测

点预测

- 我们可以将 $\hat{y}_0$ 写为

$$\hat{y}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}} = \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

  即 $\hat{y}_0$ 是 $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)'$ 的线性函数，因此，$\hat{y}_0$ 是 y_0 的线性预测。

- 在 y_0 所有的线性无偏预测中，$\hat{y}_0$ 是否是最好的?

定理 1-7

假定 $\hat{\boldsymbol{\beta}}$ 是 $\boldsymbol{\beta}$ 的最小二乘估计。对于任意一个 $(p+1)$ 维常数向量 $\boldsymbol{c}$，$\boldsymbol{c}'\hat{\boldsymbol{\beta}}$ 是 $\boldsymbol{c}'\boldsymbol{\beta}$ 的最小方差线性无偏估计。

置信区间与预测

证明：假设 $\boldsymbol{d}'\boldsymbol{y}$ 是 $\boldsymbol{c}'\boldsymbol{\beta}$ 的任意一个线性无偏估计，即，对于一切 $\boldsymbol{\beta}$，有

$$\boldsymbol{c}'\boldsymbol{\beta} = E(\boldsymbol{d}'\boldsymbol{y}) = E(\boldsymbol{d}'(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})) = \boldsymbol{d}'\boldsymbol{X}\boldsymbol{\beta}.$$

于是，$\boldsymbol{d}'\boldsymbol{X} = \boldsymbol{c}'$. 由此，

$$\mathrm{Var}(\boldsymbol{d}'\boldsymbol{y}) = \boldsymbol{d}'\mathrm{Var}(\boldsymbol{y})\boldsymbol{d} = \sigma^2\boldsymbol{d}'\boldsymbol{d}$$

和

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{c}'\hat{\boldsymbol{\beta}}) &= \mathrm{Var}(\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}) \\
&= \sigma^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c} \\
&= \sigma^2\boldsymbol{d}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{d}.
\end{aligned}$$

置信区间与预测

从而

$$\begin{aligned}\mathrm{Var}(\boldsymbol{d}'\boldsymbol{y}) - \mathrm{Var}(\boldsymbol{c}\hat{\boldsymbol{\beta}}) &= \sigma^2\boldsymbol{d}'\boldsymbol{d} - \sigma^2\boldsymbol{d}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{d} \\ &= \sigma^2\boldsymbol{d}'(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{d} \\ &= \sigma^2\boldsymbol{d}'(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{d} \\ &\geq 0\end{aligned}$$

所以，$\boldsymbol{c}'\hat{\boldsymbol{\beta}}$ 是 $\boldsymbol{c}'\boldsymbol{\beta}$ 的最小方差线性无偏估计。

推论 1-8

在 y_0 的一切线性无偏预测中，$\hat{y}_0$ 的方差最小。

置信区间与预测

从而

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{d}'\boldsymbol{y}) - \mathrm{Var}(\boldsymbol{c}\hat{\boldsymbol{\beta}}) &= \sigma^2\boldsymbol{d}'\boldsymbol{d} - \sigma^2\boldsymbol{d}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{d} \\
&= \sigma^2\boldsymbol{d}'(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{d} \\
&= \sigma^2\boldsymbol{d}'(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{d} \\
&\geq 0
\end{aligned}$$

所以，$\boldsymbol{c}'\hat{\boldsymbol{\beta}}$ 是 $\boldsymbol{c}'\boldsymbol{\beta}$ 的最小方差线性无偏估计。

推论 1-8

在 y_0 的一切线性无偏预测中，$\hat{y}_0$ 的方差最小。

置信区间与预测

预测值的分布

假设 $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)' \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$，$\varepsilon_0 \sim N(0, \sigma^2)$，而且 ε_0 与 ε_i 相互独立，$i = 1, 2, \cdots, n$.

- 根据定理 1-5 可知，

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}).$$

- 预测值为

$$\hat{y}_0 = \boldsymbol{x}_0' \hat{\boldsymbol{\beta}}$$

  其期望为 $\boldsymbol{x}_0'\boldsymbol{\beta}$，和方差为 $\sigma^2 \boldsymbol{x}_0' (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{x}_0$.

- 预测值的分布为

$$\hat{y}_0 \sim N(\boldsymbol{x}_0'\boldsymbol{\beta}, \sigma^2 \boldsymbol{x}_0' (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{x}_0)$$

置信区间与预测

预测值的分布

假设 $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)' \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$，$\varepsilon_0 \sim N(0, \sigma^2)$，而且 ε_0 与 ε_i 相互独立，$i = 1, 2, \cdots, n$.

- 而 y_0 的分布为

$$y_0 \sim N(\boldsymbol{x}_0' \boldsymbol{\beta}, \sigma^2).$$

- 由于 ε_0 与 ε_i 相互独立，$\hat{y}_0$ 与 y_0 相互独立。
- $\hat{y}_0 - y_0$ 的分布为

$$\hat{y}_0 - y_0 \sim N(0, \sigma^2(1 + \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0))$$

- $\hat{y}_0 - y_0$ 与 $\hat{\sigma}^2$ 相互独立。

置信区间与预测

预测值的分布

- 根据定理 1-4 可知，$\frac{SS_E}{\sigma^2} \sim \chi^2(n-p-1)$.
- 于是，
$$\frac{\hat{y}_0 - y_0}{\hat{\sigma}\sqrt{1 + \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}} \sim t_{n-p-1}$$
- y_0 的置信水平为 $1-\alpha$ 的预测区间为
$$\hat{y}_0 \pm t_{1-\frac{\alpha}{2}}(n-p-1)\hat{\sigma}\sqrt{\mathbf{1} + \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}.$$
- $E(y_0)$ 的置信水平为 $1-\alpha$ 的置信区间为
$$\hat{y}_0 \pm t_{1-\frac{\alpha}{2}}(n-p-1)\hat{\sigma}\sqrt{\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}.$$

其中，$t_{1-\frac{\alpha}{2}}(n-p-1)$ 表示相应的分位数。