



统计与机器学习

第一章：线性回归 - Part I

倪 蓓

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

2020 年 10 月 2 日



目录

- ① 线性回归的模型与假设
- ② 线性回归模型的参数估计
 - 最小二乘估计
 - 极大似然估计
 - 参数估计的性质
- ③ 中心化和标准化
 - 中性化
 - 标准化
- ④ 显著性检验
 - F 检验
 - t 检验
 - 复相关系数
- ⑤ 置信区间与预测

线性回归的模型

- 线性回归模型为

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \quad (1)$$

- 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

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

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

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

线性回归的模型

- 令

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix},$$

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \cdots, \beta_p)',$$

$$\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)'.$$

- 线性回归模型的矩阵形式为

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (4)$$

- 问题：如何估计回归系数 $\boldsymbol{\beta}$?

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 \mathbf{X}

- 是确定性变量，不是随机变量；
- 要求 $\text{rank}(\mathbf{X}) = p + 1 < n$ ，这表明了自变量之间不相关，样本量应大于自变量的个数， \mathbf{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

- $E(\varepsilon_i) = 0, \quad i = 1, 2, \dots, n$ ，表示没有系统误差；
- $\text{Cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2, & i = j \\ 0, & i \neq j \end{cases} \quad i, j = 1, 2, \dots, n$ ，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

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

在假设 (3) 下，随机误差向量服从

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

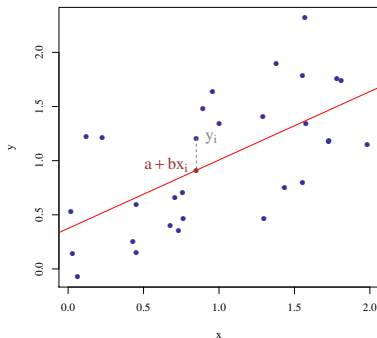
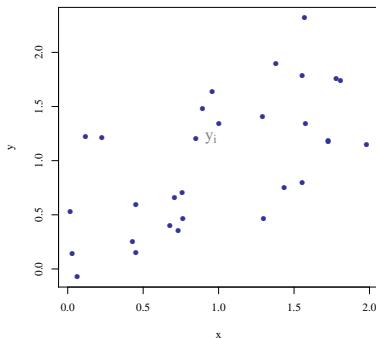
等价于假定因变量 \mathbf{y} 服从 n 维正态分布，其期望向量和协方差矩阵分别为

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

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 $x_i'\beta$.



- 离差（实际观测值和估计值的差）定义为 $y_i - x_i'\beta$.

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

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

最小二乘估计

具体计算方法

离差平方和 $Q(\beta)$ 可写为

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

最小二乘估计

常见的矩阵计算（补充）

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

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

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

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

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

特别地, 如果 \mathbf{B} 是一个对称矩阵, 那么

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

最小二乘估计

具体计算方法

- 对 β 求导, 可得

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

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

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

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

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

最小二乘估计

最小二乘估计为

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

说明

- 在求最小二乘估计 $\hat{\beta}$ 时, 需要 $(\mathbf{X}'\mathbf{X})^{-1}$ **必须存在**. 也就是说, $(\mathbf{X}'\mathbf{X})$ 是一非奇异矩阵, 即 $|\mathbf{X}'\mathbf{X}| \neq 0$.
- 由线性代数可知, $\text{rank}(\mathbf{X}) \geq \text{rank}(\mathbf{X}'\mathbf{X})$. 如果 $\mathbf{X}'\mathbf{X}$ 为 $p+1$ 阶满秩矩阵, 也就是说 $\text{rank}(\mathbf{X}'\mathbf{X}) = p+1$, 那么 $\text{rank}(\mathbf{X}) \geq p+1$.
- 另一方面, 设计矩阵 \mathbf{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

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

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

- $\hat{\mathbf{y}}$ 也可以写为

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

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

最小二乘估计

定理 1-1 (帽子矩阵的性质)

帽子矩阵 $H = X(X'X)^{-1}X'$ 具有以下的一些性质。

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

证明:

- H 的转置矩阵为

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

由于 H 的转置矩阵等于 H , 所以 H 是对称的。

最小二乘估计

- 我们可知,

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

因此, H 是幂等矩阵.

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

$$\begin{aligned} \text{tr}(H) &= \text{tr}(X(X'X)^{-1}X') \\ &= \text{tr}((X'X)^{-1}X'X) \\ &= \text{tr}(I_{p+1}) = p+1 \end{aligned}$$

最小二乘估计

- 残差定义为

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

- 也可写为

$$e = y - Hy = (I - H)y$$

- 几何上的关系：回归值 \hat{y} 与残差 e 垂直，即

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

.

最小二乘估计

- 残差的协方差矩阵为

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

- 由此，我们可以构造误差项方差 σ^2 的估计，即

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

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, \mathbf{y} 的分布为

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

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

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

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

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

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

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

极大似然估计

具体计算方法

- 对数似然函数为

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

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

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

极大似然估计

具体计算方法

- 极大似然估计为

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

说明

- $\hat{\beta}_{\text{ML}} = \hat{\beta}_{\text{LS}}$, 一般记为 $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$;
- $\hat{\sigma}_{\text{ML}}^2$ 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 $\mathbf{x} = (x_1, x_2, \dots, x_n)'$ 和 $\mathbf{y} = (y_1, y_2, \dots, y_n)'$ 均是 n 维随机变量。对于任意一个 $m \times n$ 维常矩阵 $\mathbf{A} = \{a_{ij}\}_{m \times n}$ 和一个 $m' \times n$ 维常矩阵 $\mathbf{B} = \{b_{ij}\}_{m' \times n}$ ，以及一个 m 维的常向量 $\mathbf{c} = (c_1, c_2, \dots, c_m)'$ ，我们有

- $E(\mathbf{Ax} + \mathbf{c}) = \mathbf{A}E(\mathbf{x}) + \mathbf{c}$;
- $\text{Var}(\mathbf{Ax} + \mathbf{c}) = \mathbf{A}\text{Var}(\mathbf{x})\mathbf{A}'$
- $\text{Cov}(\mathbf{Ax}, \mathbf{By}) = \mathbf{A}\text{Cov}(\mathbf{x}, \mathbf{y})\mathbf{B}'$

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 $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, 那么 $\hat{\beta}$ 满足

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

证明:

- 我们计算 $\hat{\beta}$ 的期望, 即

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

因此, $\hat{\beta}$ 是 β 的无偏估计。

参数估计的性质

- 我们计算 $\hat{\beta}$ 的方差，即

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

参数估计的性质

- β 的最小二乘估计 $\hat{\beta} = (X'X)^{-1}X'y$ 是一个随机变量;
- 残差 $e = y - X\hat{\beta} = (I_n - H)y$ 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 $\hat{\beta}$ 与残差 e 线性不相关，即

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

说明：

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

参数估计的性质

- β 的最小二乘估计 $\hat{\beta} = (X'X)^{-1}X'y$ 是一个随机变量;
- 残差 $e = y - X\hat{\beta} = (I_n - H)y$ 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 $\hat{\beta}$ 与残差 e 线性不相关，即

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

说明：

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

参数估计的性质

证明：由于

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

因此，最小二乘估计 $\hat{\beta}$ 和残差 e 线性不相关。

中心化

矩阵的知识（补充）

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

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

其中,

$$E_{11} = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1}$$

$$E_{12} = -A^{-1}B(D - CA^{-1}B)^{-1}$$

$$E_{21} = -(D - CA^{-1}B)^{-1}CA^{-1}$$

$$E_{22} = (D - CA^{-1}B)^{-1}$$

中心化

回顾

- 原本的模型:

$$y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon$$

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

$$\hat{y} = \hat{\beta}_{\text{intercept}} + \mathbf{x}'\hat{\boldsymbol{\beta}}_{\text{slope}}$$

中心化

- 原始数据集为

$$\begin{cases} \mathbf{y} = (y_1, \dots, y_n)' \\ \mathbf{X} = (\mathbf{1}_n, \mathbf{X}_o), \quad \mathbf{X}_o = (\mathbf{x}_1, \dots, \mathbf{x}_p) \end{cases}$$

- 最小二乘估计为

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

其中, $\mathbf{A}_o = (\mathbf{X}_o'\mathbf{X}_o - n^{-1}\mathbf{X}_o'\mathbf{1}_n\mathbf{1}_n'\mathbf{X}_o)^{-1}$

中心化

- 中心化:

$$x_{ij}^* = x_{ij} - \bar{x}_j, \quad \bar{x}_j = n^{-1} \sum_{i=1}^n x_{ij}$$

$$y_i^* = y_i - \bar{y}, \quad \bar{y} = n^{-1} \sum_{i=1}^n y_i$$

令

$$\begin{cases} \mathbf{y}^* = (y_1^*, \dots, y_n^*)' \\ \mathbf{X}_c = (\mathbf{x}_1^*, \dots, \mathbf{x}_p^*) \\ \mathbf{X}^* = (\mathbf{1}_n, \mathbf{X}_c) \end{cases}$$

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

中心化

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

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

其中，

$$\mathbf{A}_c = (\mathbf{X}'_c \mathbf{X}_c - n^{-1} \mathbf{X}'_c \mathbf{1}_n \mathbf{1}'_n \mathbf{X}_c)^{-1}$$

中心化

- 中心化的因变量与未中心化的因变量之间的关系：

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

其中 $\mathbf{H}_{\mathbf{1}_n} = \mathbf{1}_n(\mathbf{1}_n'\mathbf{1}_n)^{-1}\mathbf{1}_n'$ 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：

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

而且

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

中心化

- 最小二乘估计分为回归常数和回归系数，即

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

- 一方面，回归常数为

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

中心化

- 由于

$$\begin{aligned} \mathbf{A}_c &= (\mathbf{X}'_c \mathbf{X}_c - n^{-1} \mathbf{X}'_c \mathbf{1}_n \mathbf{1}'_n \mathbf{X}_c)^{-1} \\ &= (\mathbf{X}'_o (\mathbf{I}_n - \mathbf{H}_{1_n}) \mathbf{X}_o \\ &\quad - n^{-1} \mathbf{X}'_c (\mathbf{I}_n - \mathbf{H}_{1_n}) \mathbf{1}_n \mathbf{1}'_n (\mathbf{I}_n - \mathbf{H}_{1_n}) \mathbf{X}_o)^{-1} \\ &= (\mathbf{X}'_o (\mathbf{I}_n - \mathbf{H}_{1_n}) \mathbf{X}_o)^{-1} = \mathbf{A}_o \end{aligned}$$

- 另一方面，回归系数为

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

中心化

- 而

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

- 于是,

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

- 采用中心化的数据得到的经验回归方程为

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

标准化

- 标准化:

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

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

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

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

其中,

$$\mathbf{L} = \text{diag} \left\{ \frac{1}{\sqrt{L_{11}}}, \dots, \frac{1}{\sqrt{L_{pp}}} \right\}$$

标准化

- 最小二乘估计为

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

- 回归系数为

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

标准化

- 注意到

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

- 因此,

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

其中每一个分量为

$$\hat{\beta}_{sj} = \frac{\sqrt{L_{jj}}}{\sqrt{L_{yy}}}\hat{\beta}_{cj} = \frac{\sqrt{L_{jj}}}{\sqrt{L_{yy}}}\hat{\beta}_j, \quad j = 1, 2, \dots, p.$$



统计与机器学习

第一章：线性回归 - Part II

倪 蓓

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

2020 年 10 月 2 日



目录

- ① 线性回归的模型与假设
- ② 线性回归模型的参数估计
 - 最小二乘估计
 - 极大似然估计
 - 参数估计的性质
- ③ 中心化和标准化
 - 中性化
 - 标准化
- ④ 显著性检验
 - F 检验
 - t 检验
 - 复相关系数
- ⑤ 置信区间与预测

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, \dots, x_p 从整体上对因变量 y 是否有明显的影响.

- 原假设为

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

- 备择假设为

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

- 如果 H_0 为真, 则表明因变量 y 与 x_1, x_2, \dots, 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 $\hat{y}_i = \mathbf{x}_i' \hat{\boldsymbol{\beta}}$;
- 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 $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$, 有

- $SS_E/\sigma^2 \sim \chi^2(n-p-1)$, 其中 $SS_E = \mathbf{e}'\mathbf{e}$;
- 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 $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n)$.

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

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

其中

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

当且仅当

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

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,

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

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

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

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

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

那么残差平方和为

$$SS_E = \mathbf{e}'\mathbf{e} = \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}$$

由于 $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, 而且 $(\mathbf{I} - \mathbf{H})$ 是对称幂等矩阵, 且其秩为 $\text{rank}(\mathbf{I} - \mathbf{H}) = n - (p + 1)$. 我们可以计算

$$\frac{1}{\sigma^2}E(\mathbf{y})'(\mathbf{I} - \mathbf{H})E(\mathbf{y}) = \frac{1}{\sigma^2}\boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = 0.$$

因此, $SS_E/\sigma^2 \sim \chi^2(n - p - 1)$.

F 检验

- 由于

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

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

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

那么, SS_R 和 SS_E 相互独立.

F 检验

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

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

所以,

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

F 检验

由于 $\mathbf{X} = (\mathbf{1}_n, \mathbf{X}_o)$, 因此, 注意到

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

于是,

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

F 检验

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

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

根据多元统计的知识可知,

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

其中

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

F 检验

结论

- 那么，检验统计量

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

- 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

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

- 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

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

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

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

参数估计的性质

定理 1-5

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

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

证明:

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

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

t 检验

多元统计的知识（补充）

假设随机向量 $\mathbf{x} = (x_1, x_2, \dots, x_p)'$ 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

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

- $(\mathbf{X}'\mathbf{X})^{-1}$ 是一个 $(p+1) \times (p+1)$ 矩阵;
- 这里 c_{ij} 表示 $(\mathbf{X}'\mathbf{X})^{-1}$ 中第 $(i+1)$ 行第 $(j+1)$ 列元素, $i, j = 0, 1, \dots, p$;
- 根据多元统计的知识, 我们有

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

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, \dots, 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, \dots, 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, \dots, 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 $\mathbf{x}_0 = (1, x_{01}, x_{02}, \dots, x_{0p})'$, 我们关心的是

$$y_0 = \mathbf{x}_0' \boldsymbol{\beta} + \varepsilon_0$$

- 基本假定: $E(\varepsilon_0) = 0$ 和 $\text{Var}(\varepsilon_0) = \sigma^2$.
- 也就是说, $E(y_0) = \mathbf{x}_0' \boldsymbol{\beta}$ 和 $\text{Var}(y_0) = \sigma^2$.
- 基于数据集 $\{(y_i, \mathbf{x}_i'), i = 1, 2, \dots, n\}$, 我们可以得到 $\boldsymbol{\beta}$ 的最小二乘估计为 $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)'$.

置信区间与预测

点预测

- y_0 的预测值为

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

- 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

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

- 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

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

那么,

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

置信区间与预测

点预测

- 考虑 $\text{Bias}^2(\hat{y}_0) = 0$, 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 $\hat{\beta}$ 的无偏性, 我们有

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

置信区间与预测

点预测

- 我们可以将 \hat{y}_0 写为

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

即 \hat{y}_0 是 $\mathbf{y} = (y_1, y_2, \dots, y_n)'$ 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 \mathbf{c} , $\mathbf{c}'\hat{\boldsymbol{\beta}}$ 是 $\mathbf{c}'\boldsymbol{\beta}$ 的最小方差线性无偏估计。

置信区间与预测

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

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

于是， $d'X = c'$. 由此，

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

和

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

置信区间与预测

从而

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

所以, $\mathbf{c}'\hat{\boldsymbol{\beta}}$ 是 $\mathbf{c}'\boldsymbol{\beta}$ 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

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

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

- 根据定理 1-5 可知,

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

- 预测值为

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

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

- 预测值的分布为

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

置信区间与预测

预测值的分布

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

- 而 y_0 的分布为

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

- 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

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

- $\hat{y}_0 - y_0$ 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,

$$\frac{\hat{y}_0 - y_0}{\hat{\sigma} \sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}} \sim t_{n-p-1}$$

- y_0 的置信水平为 $1 - \alpha$ 的预测区间为

$$\hat{y}_0 \pm t_{1-\frac{\alpha}{2}}(n - p - 1) \hat{\sigma} \sqrt{1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}.$$

- $E(y_0)$ 的置信水平为 $1 - \alpha$ 的置信区间为

$$\hat{y}_0 \pm t_{1-\frac{\alpha}{2}}(n - p - 1) \hat{\sigma} \sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}.$$

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