



Regression Analysis

课程笔记

作者: Fir1247

组织: USTC

时间: December 13, 2023

版本: 1.1

联系方式: fa1247@mail.ustc.edu.cn 或者 QQ: 3105292483



回归分析

前言

本文档为杨亚宁老师 2023 年秋季学期回归分析课程的笔记。

December 13, 2023

Fir1247

目录

第 1 章 预备知识	1
1.1 随机向量	1
1.2 相关系数	2
1.2.1 Pearson 相关系数	2
1.2.2 独立性假设检验	2
1.2.3 去相关化	4
1.2.4 决定系数和典则相关系数	4
1.2.5 偏相关系数	6
1.3 多元正态分布	9
1.4 线性代数知识补充	10
1.4.1 向量空间	10
1.4.2 矩阵	13
1.4.3 行列空间的特征刻画	15
1.4.4 广义逆矩阵	16
1.4.5 欧氏空间中的正交投影	17
第 2 章 线性回归	18
2.1 引言	18
2.2 简单线性回归模型	21
2.2.1 定义	21
2.2.2 最小二乘法	21
2.2.3 误差方差的估计	24
2.2.4 “LS 估计的方差”的估计	24
2.2.5 正态模型下的统计推断	24
2.2.6 应用：幂次律	26
2.2.7 回归模型结果的解释：因果还是关联？	29
2.3 多重线性回归模型	29
2.3.1 定义	29
2.3.2 最小二乘法	30
2.3.3 误差方差的估计	30



2.3.4	LS 估计的统计性质.....	30
2.3.5	工具变量.....	31
2.4	中心化.....	31
2.4.1	基本概念.....	32
2.4.2	部分回归系数.....	34
2.5	回归模型的 F 检验	35
2.5.1	F 分布.....	35
2.5.2	一般线性假设的 F 检验.....	36
2.5.3	约束最小二乘.....	36
2.5.4	全模型与零模型的差别——F 检验的另一种观点	37
2.5.5	部分回归系数的 F 检验.....	38
2.5.6	应用：回归方程的显著性检验.....	39

第 1 章 预备知识

1.1 随机向量

定义 1.1.1

随机向量 $\mathbf{x} = (x_1, \dots, x_n)^T$ 的均值定义为:

$$\mathbb{E}(\mathbf{x}) = \boldsymbol{\mu}_x \stackrel{\text{def}}{=} (\mathbb{E}x_1, \dots, \mathbb{E}x_n)^T$$

随机向量 $\mathbf{x}_{n \times 1}, \mathbf{y}_{m \times 1}$ 的协方差矩阵定义为:

$$\text{cov}(\mathbf{x}, \mathbf{y}) \stackrel{\text{def}}{=} \mathbb{E}(\mathbf{x} - \boldsymbol{\mu}_x)(\mathbf{y} - \boldsymbol{\mu}_y)^T = (\text{cov}(x_i, y_j))_{n \times m}$$

特别地, 随机向量 $\mathbf{x}$ 的方差-协方差矩阵:

$$\text{var}(\mathbf{x}) = \text{cov}(\mathbf{x}) = \text{cov}(\mathbf{x}, \mathbf{x}) = \mathbb{E}\mathbf{x}\mathbf{x}^T - \boldsymbol{\mu}_x\boldsymbol{\mu}_x^T$$

用 $\mathbf{x} \sim (\boldsymbol{\mu}, \Sigma)$ 表示 $\mathbb{E}(\mathbf{x}) = \boldsymbol{\mu}, \text{var}(\mathbf{x}) = \Sigma$.

命题 1.1.1

设 $\mathbf{x}, \mathbf{y}$ 为 $n \times 1$ 的随机向量, $\mathbf{z}, \mathbf{w}$ 为 $m \times 1$ 的随机向量, 则

1. $\mathbb{E}(\mathbf{x} + \mathbf{y}) = \mathbb{E}(\mathbf{x}) + \mathbb{E}(\mathbf{y})$, 对于任意常数矩阵 A 和常数向量 $\mathbf{b}$, $\mathbb{E}(A\mathbf{x} + \mathbf{b}) = A\mathbb{E}(\mathbf{x}) + \mathbf{b}$.
2. $\text{cov}(\mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{w}) = \text{cov}(\mathbf{x} + \mathbf{y}) + \text{cov}(\mathbf{x} + \mathbf{w}) + \text{cov}(\mathbf{y} + \mathbf{z}) + \text{cov}(\mathbf{y} + \mathbf{w})$.
3. $\text{var}(\mathbf{x} + \mathbf{y}) = \text{var}(\mathbf{x}) + \text{var}(\mathbf{y}) + \text{cov}(\mathbf{x}, \mathbf{y}) + \text{cov}(\mathbf{y}, \mathbf{x})$.
4. $\text{cov}(A\mathbf{x} + \mathbf{b}, B\mathbf{y} + \mathbf{c}) = A\text{cov}(\mathbf{x}, \mathbf{y})B^T$, 其中 A, B 为常数矩阵, $\mathbf{b}, \mathbf{c}$ 为常数向量。特别地, $\text{var}(A\mathbf{x} + \mathbf{b}) = A\text{var}(\mathbf{x})A^T$.

证明 只说明 4:

$$\begin{aligned}\text{cov}(A\mathbf{x} + \mathbf{b}, B\mathbf{y} + \mathbf{c}) &\stackrel{\text{def}}{=} \mathbb{E}(A\mathbf{x} + \mathbf{b} - A\boldsymbol{\mu}_x - \mathbf{b})(B\mathbf{y} + \mathbf{c} - B\boldsymbol{\mu}_y - \mathbf{c})^T \\ &= A[\mathbb{E}(\mathbf{x} - \boldsymbol{\mu}_x)(\mathbf{y} - \boldsymbol{\mu}_y)^T]B^T \\ &= A\text{cov}(\mathbf{x}, \mathbf{y})B^T\end{aligned}$$

定义 1.1.2

随机向量 $\mathbf{x}$ 的相关系数矩阵定义为 $R = (\rho_{ij})$, 其中 ρ_{ij} 为 x_i, x_j 的相关系数。

推论 1.1.1

$\Sigma = (\sigma_{ij})$ 为随机向量 $\mathbf{x}$ 的协方差矩阵, 则

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

设 $D = \text{diag}(\Sigma) = \text{diag}(\sigma_{11}, \dots, \sigma_{nn})$,

$$R = D^{-\frac{1}{2}} \Sigma D^{-\frac{1}{2}}$$



推论 1.1.2

随机向量 $\mathbf{x}_{n \times 1} \sim (\boldsymbol{\mu}, \Sigma)$, $D = \text{diag}(\Sigma)$, R 为 $\mathbf{x}$ 的相关系数矩阵, 则

1. 中心化: $\mathbf{y} = \mathbf{x} - \boldsymbol{\mu} \sim (\mathbf{0}, \Sigma)$.

2. 标准化:

$$\mathbf{y} = \Sigma^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\mu}) \sim (\mathbf{0}, I_n)$$

$$\mathbf{y} = D^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\mu}) \sim (\mathbf{0}, R)$$



1.2 相关系数

1.2.1 Pearson 相关系数

定义 1.2.1

随机变量 x, y 的 Pearson 相关系数:

$$\rho_{xy} \stackrel{\text{def}}{=} \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)\text{var}(y)}}$$

样本 $(x_1, y_1), \dots, (x_n, y_n)$ 的 Pearson 相关系数:

$$r_{xy} \stackrel{\text{def}}{=} \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$



直观上来看, 样本的 Pearson 相关系数度量了中心化样本 (向量) 之间的相似性 (夹角的余弦值)。

1.2.2 独立性假设检验

考察样本 $(x_1, y_1), \dots, (x_n, y_n)$, 其中 x_i 的总体为 x , y_i 的总体为 y , 对于检验问题 $H_0: x \perp y$, 原假设成立时 $\rho_{xy} = 0$, 在此条件下研究 Pearson 相关系数 r_{xy} 的分布, 就可以用来进行独立性检验。

1.2.2.1 正态总体下的精确检验

引理 1.2.1

若随机向量 $\mathbf{y} \sim N_n(0, I_n)$, 假设 O 是任一正交矩阵, $O\mathbf{y} \sim N_n(0, I_n)$.



证明 写出 $O\mathbf{y}$ 的概率密度函数即可。

引理 1.2.2

假设随机向量 $\mathbf{x} \sim N_n(0, I_n)$, 则

1. 对于任何常数向量 $\mathbf{a} \in \mathbb{R}^n$, 满足 $\|\mathbf{a}\| = 1$, 则 $\mathbf{a}^T \mathbf{x} \sim N(0, 1)$, $\|\mathbf{x}\|^2 - (\mathbf{a}^T \mathbf{x})^2 \sim \chi_{n-1}^2$, 且二者独立, 因

此

$$\sqrt{n-1} \frac{\mathbf{a}^T \mathbf{x}}{\sqrt{\|\mathbf{x}\|^2 - (\mathbf{a}^T \mathbf{x})^2}} \sim t_{n-1}$$

2. 对于任何常数向量 $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, 满足 $\|\mathbf{a}\| = \|\mathbf{b}\| = 1$, $\mathbf{a}^T \mathbf{b} = 0$, 则 $\mathbf{a}^T \mathbf{x} \sim N(0, 1)$, $\|\mathbf{x}\|^2 - (\mathbf{a}^T \mathbf{x})^2 - (\mathbf{b}^T \mathbf{x})^2 \sim \chi_{n-2}^2$, 且二者独立, 因此

$$\sqrt{n-2} \frac{\mathbf{a}^T \mathbf{x}}{\sqrt{\|\mathbf{x}\|^2 - (\mathbf{a}^T \mathbf{x})^2 - (\mathbf{b}^T \mathbf{x})^2}} \sim t_{n-2}$$

♥

证明 见作业。

命题 1.2.1 (正态总体)

假设样本 $(x_1, y_1), \dots, (x_n, y_n)$ 服从二元正态总体, 则当 $\rho_{xy} = 0$ 时,

$$t = \sqrt{n-2} \frac{r}{\sqrt{1-r^2}} \sim t_{n-2}$$

这里 $r = r_{xy}$.

♣

证明 当 $\rho_{xy} = 0$ 时, y_i 与 x_i 独立, 所以分别服从某个一元正态, 不妨 $y_i \sim N(0, 1)$, $\mathbf{y} = (y_1, \dots, y_n)^T \sim N_n(0, I_n)$, 所以

$$\frac{r}{\sqrt{1-r^2}} = \frac{s_{xy}/\sqrt{s_{xx}}}{\sqrt{s_{yy} - s_{xy}^2/s_{xx}}} = \frac{s_{xy}/\sqrt{s_{xx}}}{\sqrt{\sum_{i=1}^n y_i^2 - n\bar{y}^2 - (s_{xy}/\sqrt{s_{xx}})^2}}$$

下面我们在给定 $\mathbf{x} = (x_1, \dots, x_n)^T$ 的条件下计算 t 的分布, 记

$$\mathbf{a} = (x_1 - \bar{x}, \dots, x_n - \bar{x})^T / \sqrt{s_{xx}}, \quad \mathbf{b} = (1, \dots, 1)^T / \sqrt{n}$$

则 $\|\mathbf{a}\| = \|\mathbf{b}\| = 1$, $\mathbf{a} \perp \mathbf{b}$, 因此

$$t = \sqrt{n-2} \frac{\mathbf{a}^T \mathbf{y}}{\sqrt{\sum_{i=1}^n y_i^2 - (\mathbf{b}^T \mathbf{y})^2 - (\mathbf{a}^T \mathbf{y})^2}} \sim t_{n-2}$$

该分布与 $\mathbf{x}$ 无关, 因此无条件成立, 命题得证。

根据上述命题, 可得正态总体下独立性假设的精确检验: $|t| > t_{n-2}(\alpha/2)$ 时否定原假设, p 值为 $P(|T| \geq |t|)$, $T \sim t_{n-2}$.

推论 1.2.1

两样本 t 检验等价于相关性检验。

♥

证明 见作业。

1.2.2.2 大样本检验

命题 1.2.2

样本 $(x_1, y_1), \dots, (x_n, y_n)$, 存在有限二阶矩, 样本相关系数为 r , $\rho_{xy} = 0$ 时,

$$\sqrt{nr} \xrightarrow{d} N(0, 1)$$

♣

证明比较麻烦, 这里不抄录了。由上述命题可见, 独立性假设的大样本检验: $|\sqrt{nr}| > u_{\alpha/2}$ 时拒绝原假设。

推论 1.2.2

独立性假设的大样本检验等价于 Pearson 卡方检验。

♥

证明 见作业。

1.2.2.3 置换检验

本节内容在 Lab1 中有详细介绍和 R 语言应用。

1.2.2.4 相关系数的置信区间

我甚至还没学置信区间，本节内容待补充。（课件 Week2 P22-P26）

1.2.3 去相关化

假设任意 $q \times 1$ 随机向量 $\mathbf{y}$ ， $p \times 1$ 随机向量 $\mathbf{x}$ ，方差-协方差矩阵：

$$\Sigma = \text{cov} \begin{pmatrix} \mathbf{y} \\ \mathbf{x} \end{pmatrix} = \begin{pmatrix} \text{var}(\mathbf{y}) & \text{cov}(\mathbf{y}, \mathbf{x}) \\ \text{cov}(\mathbf{x}, \mathbf{y}) & \text{var}(\mathbf{x}) \end{pmatrix}$$

$\mathbf{y}$ 和 $\mathbf{x}$ 相关，协方差为 $q \times p$ 矩阵 Σ_{yx} 。我们希望从 $\mathbf{y}$ 中消去与 $\mathbf{x}$ 相关的线性成分，即求 $A_{q \times p}$ ，使得 $\mathbf{y} - A\mathbf{x}$ 与 $\mathbf{x}$ 不相关。

$$0 = \text{cov}(\mathbf{y} - A\mathbf{x}, \mathbf{x}) = \text{cov}(\mathbf{y}, \mathbf{x}) - A(\text{cov})(\mathbf{x}, \mathbf{x}) = \Sigma_{yx} - A\Sigma_{xx}$$

得到 $A = \Sigma_{yx}\Sigma_{xx}^{-1}$ 。

定义 1.2.2

我们称 $\mathbf{y}^\perp = \mathbf{y} - \Sigma_{yx}\Sigma_{xx}^{-1}\mathbf{x}$ 为 $\mathbf{y}$ 的去相关化（与 $\mathbf{x}$ 不相关）。

容易验证：

$$\begin{aligned} \text{var}(\mathbf{y}^\perp) &= \text{cov}(\mathbf{y} - \Sigma_{yx}\Sigma_{xx}^{-1}\mathbf{x}, \mathbf{y} - \Sigma_{yx}\Sigma_{xx}^{-1}\mathbf{x}) \\ &= \text{cov}(\mathbf{y}, \mathbf{y}) - \Sigma_{yx}\Sigma_{xx}^{-1}\text{cov}(\mathbf{x}, \mathbf{y}) - \text{cov}(\mathbf{y}, \mathbf{x})\Sigma_{xx}^{-1}\Sigma_{xy} + \text{cov}(\Sigma_{yx}\Sigma_{xx}^{-1}\mathbf{x}, \Sigma_{yx}\Sigma_{xx}^{-1}\mathbf{x}) \\ &= \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy} \\ &\stackrel{\text{def}}{=} \Sigma_{yy \cdot x} \end{aligned}$$

于是

$$\text{cov} \begin{pmatrix} \mathbf{y}^\perp \\ \mathbf{x} \end{pmatrix} = \begin{pmatrix} \Sigma_{yy \cdot x} & 0 \\ 0 & \Sigma_{xx} \end{pmatrix}$$

去相关化对应于协方差矩阵相合分块的对角化：

$$\begin{pmatrix} I_q & -\Sigma_{yx}\Sigma_{xx}^{-1} \\ 0 & I_p \end{pmatrix} \begin{pmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix} \begin{pmatrix} I_q & 0 \\ -\Sigma_{xx}^{-1}\Sigma_{xy} & I_p \end{pmatrix} = \begin{pmatrix} \Sigma_{yy \cdot x} & 0 \\ 0 & \Sigma_{xx} \end{pmatrix}$$

1.2.4 决定系数和典则相关系数

$\mathbf{x}$ 和 $\mathbf{y}$ 的协方差矩阵 $\Sigma_{yx} = \text{cov}(\mathbf{y}, \mathbf{x})$ 或相关系数矩阵度量了 $\mathbf{x}$ 和 $\mathbf{y}$ 各个分量之间的相关程度，这是一个 $q \times p$ 矩阵，我们希望汇总矩阵所有元素，得到向量 $\mathbf{x}$ 和 $\mathbf{y}$ 的一个相关性的数字度量。

利用前述去相关化技术，我们可以定义典则相关系数或决定系数用于度量两个随机向量之间的相关性。

定义 1.2.3

$\mathbf{y}^\perp = \mathbf{y} - \Sigma_{yx}\Sigma_{xx}^{-1}\mathbf{x}$ 与 $\mathbf{x}$ 不相关，也与 $\hat{\mathbf{y}} \stackrel{\text{def}}{=} \Sigma_{yx}\Sigma_{xx}^{-1}\mathbf{x}$ 不相关，所以有正交分解： $\mathbf{y} = \hat{\mathbf{y}} + \mathbf{y}^\perp$ 。

定理 1.2.1 (方差分解勾股定理)

$\text{var}(\mathbf{y}) = \text{var}(\hat{\mathbf{y}}) + \text{var}(\mathbf{y}^\perp)$, 即

$$\text{var}(\hat{\mathbf{y}}) = \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy}$$

定义 1.2.4

定义:

$$\Phi = \Sigma_{yy}^{-\frac{1}{2}} (\Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy}) \Sigma_{yy}^{-\frac{1}{2}}$$

称为 $\mathbf{x}$ 与 $\mathbf{y}$ 的关系密切程度。也可以定义为 $\Psi = (\Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy}) \Sigma_{yy}$, 但该矩阵不对称。

在多元分析中, Φ 的最大特征根的平方根 $\sqrt{\lambda_{\max}(\Phi)}$ 称为第一典则相关系数, 第二大特征根的平方根称为第二点则相关系数, 等等。

命题 1.2.3

典则相关系数关于 $\mathbf{x}$ 和 $\mathbf{y}$ 对称, 介于 0 和 1 之间。

引理 1.2.3

A, B 是 n 阶对称矩阵,

1. 若 $A \leq B$, 即 $A - B$ 半负定, 则对于任何 $m \times n$ 矩阵, $CAC^T \leq CBC^T$.
2. 若 $A \leq I$, 则 $\lambda(A) \leq 1$.

证明

1. 对于任何 $\mathbf{x} \in \mathbb{R}^m$, $\mathbf{x}^T (CAC^T - CBC^T) \mathbf{x} = \mathbf{y}^T (A - B) \mathbf{y} \leq 0$, 其中 $\mathbf{y} = C^T \mathbf{x}$.
2. 假设正交矩阵 O 使得 $OAO^T = \Lambda$ 对角阵, $A \leq I \Rightarrow OAO^T \leq OO^T = I$, 即 $\Lambda \leq I$, 于是 Λ 对角元 ≤ 0 .

证明 方差分解公式中 $\Sigma_{yy} \geq 0$, 所以 $0 \leq \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy} \leq \Sigma_{yy}$, 于是

$$0 \leq \Phi = \Sigma_{yy}^{-\frac{1}{2}} \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy} \Sigma_{yy}^{-\frac{1}{2}} \leq I_q$$

所以 $0 \leq \lambda_{\max}(\Phi) \leq 1$.

由于 $\Psi = \Sigma_{xx}^{-\frac{1}{2}} \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma_{yx} \Sigma_{xx}^{-\frac{1}{2}}$ 与 Φ 有着相同的非 0 特征根, 所以二者的最大特征根相同。

定义 1.2.5

在回归分析的讨论范围中 $q = 1$, 即 $\mathbf{y}$ 只是随机变量, 此时

$$\phi = \frac{\Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy}}{\Sigma_{yy}}$$

为 $\mathbf{x}$ 能解释的部分在 $\mathbf{y}$ 的方差中所占比例, 我们称之为 (回归方程的) 决定系数。它度量了随机变量 $\mathbf{y}$ 和随机向量 $\mathbf{x}$ 之间的相关程度。

注

1. ϕ 的最优性: $\phi = \max_{\beta \in \mathbb{R}^p} (\rho(\mathbf{y}, \beta^T \mathbf{x}))^2$, 见本节命题 3.
2. 如果 x, y 都是随机变量, 则 $\phi = \rho_{xy}^2$. 故决定/典则系数是随机变量相关系数的推广。

命题 1.2.4

$\mathbf{x}$ 和 $\mathbf{y}$ 分别是 $p \times 1, q \times 1$ 的随机向量, $\Phi = \Sigma_{yy}^{-\frac{1}{2}} \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy} \Sigma_{yy}^{-\frac{1}{2}}$, 则

$$\max_{0 \neq \mathbf{a} \in \mathbb{R}^p, 0 \neq \mathbf{b} \in \mathbb{R}^q} \rho(\mathbf{a}^T \mathbf{x}, \mathbf{b}^T \mathbf{y}) = \lambda_{\max}^{\frac{1}{2}}(\Phi)$$

等号在 $\mathbf{a} \propto \Sigma_{xx}^{-1} \Sigma_{xy} \Sigma_{yy}^{-\frac{1}{2}} \mathbf{v}_{\max}$, $\mathbf{b} \propto \Sigma_{yy}^{-\frac{1}{2}} \mathbf{v}_{\max}$ 时达到, 其中 $\mathbf{v}_{\max}$ 为 Φ 的特征根 $\lambda_{\max}(\Phi)$ 对应的特征向量。(这其实是典则相关系数通常的定义方式)

证明

1.2.5 偏相关系数

1.2.5.1 定义

x, y 为随机变量, 干扰因素 z 为随机向量或变量, 我们希望在控制 z 的条件下计算 x, y 之间的相关性, 即偏相关系数, 理解为控制 z 的条件下, x, y 不相关。

定义 1.2.6

令 $y^\perp = y - \Sigma_{yz}\Sigma_{zz}^{-1}z$, $x^\perp = x - \Sigma_{xz}\Sigma_{zz}^{-1}z$, 它们都与 z 不相关, 偏相关系数定义为 $y^\perp$ 和 $x^\perp$ 的相关系数:

$$\rho_{yx \cdot z} \stackrel{\text{def}}{=} \rho_{y^\perp x^\perp} = \frac{\text{cov}(y^\perp, x^\perp)}{\sqrt{\text{var}(x^\perp)\text{var}(y^\perp)}}$$

约定记号:

$$\Sigma_{xy \cdot z} = \Sigma_{xy} - \Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zy}$$

推论 1.2.3

1. 由定义, $|\rho_{yx \cdot z}| \leq 1$.
2. 后面将会证明, 多元正态假设下, 偏相关系数实际上是给定 z 的条件下 x, y 的条件相关系数。
3. $\rho_{yz \cdot z} = 0$ 意味着: 控制 z 的条件下, x 和 y 不相关 (正态假设下条件独立)。
4. $\rho_{yx \cdot z}$ 不依赖于 z 的具体值, 所以只有在各个组内 x, y 的分布形状类似的情况下, 偏相关系数才有意义。

命题 1.2.5

$$\rho_{yx \cdot z} = \frac{\Sigma_{xy \cdot z}}{\sqrt{\Sigma_{xx \cdot z}}\sqrt{\Sigma_{yy \cdot z}}}$$

证明

当 z 也是随机变量时, 相关系数矩阵

$$R = \begin{pmatrix} 1 & \rho_{xy} & \rho_{xz} \\ \rho_{xy} & 1 & \rho_{yz} \\ \rho_{xz} & \rho_{yz} & 1 \end{pmatrix}$$

则偏相关系数 $\rho_{xy \cdot z}$ 具有如下形式:

$$\rho_{xy \cdot z} = \frac{\rho_{xy} - \rho_{xz}\rho_{yz}}{\sqrt{1 - \rho_{xz}^2}\sqrt{1 - \rho_{yz}^2}}$$

其中 ρ_{ab} 代表随机变量 a, b 的 Pearson 相关系数。

注 $|\rho_{xy \cdot z}| \leq 1 \Leftrightarrow 1 - \rho_{xy}^2 - \rho_{yz}^2 - \rho_{zx}^2 + 2\rho_{xy}\rho_{xz}\rho_{yz} \geq 0 \Leftrightarrow \det(R) \geq 0$.

例 1.2.1.

假设三个变量的相关系数矩阵如下:

$$R = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}$$

即两两相关系数都是 ρ , 则必定 $\rho \geq -\frac{1}{2}$ (正定性约束), 且 $\rho_{xy \cdot z} = \frac{\rho}{1+\rho}$, 这表明偏相关系数比相关系数更靠近 0:

1. 当 $\rho \geq 0$ 时, 且 $0 \leq \rho_{xy \cdot z} \leq \rho$.
2. 当 $-\frac{1}{2} \leq \rho \leq 0$ 时, $\rho \leq \rho_{xy \cdot z} \leq 0$.

思考: 为什么约束 $\rho \geq -\frac{1}{2}$? 稳定性的要求?

例 1.2.2.

高中男、女生的数学成绩 x 与语文成绩 y 的相关系数都为 0.6, 假设男女生数学成绩没有差异, 但女生的语文成绩要好于男生。如果男女生混合在一起, 得到的相关系数是大于、等于还是小于 0.6?

解答: 以 $z = 1, 0$ 分别代表男、女, 已知 $\rho_{xz} = 0$ (数学与性别不相关), 已知 $\rho_{xy \cdot z} = 0.6$ (在各个 z 组内 x, y 的分布/相关性相同)。故

$$0.6 = \rho_{xy \cdot z} = \frac{\rho_{xy}}{\sqrt{1 - \rho_{yz}^2}} \geq \rho_{xy}$$

1.2.5.2 偏相关系数的检验

类似于相关系数的检验, 等价于多重回归模型中回归系数的检验。

精确检验: 数据 $(x_i, y_i, z_i), i = 1, \dots, n$. $H_0: \rho_{xy \cdot z} = 0$. 假设 $y_i | x_i, z_i, i = 1, \dots, n$ iid. 服从正态分布, 记样本偏相关系数为 $r_{xy \cdot z}$, 变量总个数为 p . 检验统计量在原假设下

$$T = \sqrt{n-p} \frac{r_{xy \cdot z}}{\sqrt{1 - r_{xy \cdot z}^2}} \sim t_{n-p}$$

大样本检验: 对于一般总体, H_0 下, $z = \sqrt{n-p} \times r_{xy \cdot z} \xrightarrow{d} N(0, 1), n \rightarrow \infty$.

例 1.2.3.

身高 x 、阅读能力 y 和年龄段 z 的相关系数矩阵为

$$R = \begin{pmatrix} 1 & 0.6 & 0.8 \\ 0.6 & 1 & 0.7 \\ 0.8 & 0.7 & 1 \end{pmatrix}$$

$n = 100$, 检验阅读与身高是否相关。

解答: 计算可得 $r_{xy \cdot z} = 0.089$, $z = \sqrt{100} \times 0.089 = 0.89$, 得 $p = 0.37$, 不显著相关。

1.2.5.3 偏相关系数的一般情形

定义 1.2.7

三个随机向量: $\mathbf{x}_{p \times 1}, \mathbf{y}_{q \times 1}, \mathbf{z}_{r \times 1}$, 控制 $\mathbf{z}$ 条件下研究 $\mathbf{x}, \mathbf{y}$ 的相关性:

$$\text{var} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \\ \mathbf{z} \end{pmatrix} = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} & \Sigma_{xz} \\ \Sigma_{yx} & \Sigma_{yy} & \Sigma_{yz} \\ \Sigma_{zx} & \Sigma_{zy} & \Sigma_{zz} \end{pmatrix}$$

去相关化: $\mathbf{x}^\perp = \mathbf{x} - \Sigma_{xz}\Sigma_{zz}^{-1}\mathbf{z}$, $\mathbf{y}^\perp = \mathbf{y} - \Sigma_{yz}\Sigma_{zz}^{-1}\mathbf{z}$

$$\text{var} \begin{pmatrix} \mathbf{x}^\perp \\ \mathbf{y}^\perp \\ \mathbf{z} \end{pmatrix} = \begin{pmatrix} \Sigma_{xx \cdot z} & \Sigma_{xy \cdot z} & 0 \\ \Sigma_{yx \cdot z} & \Sigma_{yy \cdot z} & 0 \\ 0 & 0 & \Sigma_{zz} \end{pmatrix}$$

令 $\Phi_{xy \cdot z} = \Sigma_{yy \cdot z}^{-\frac{1}{2}}(\Sigma_{yx \cdot z}\Sigma_{xx \cdot z}^{-1}\Sigma_{xy \cdot z})\Sigma_{yy \cdot z}^{-\frac{1}{2}}$, 代表控制 $\mathbf{z}$ 之后, 随机向量 $\mathbf{x}, \mathbf{y}$ 的相关性。当 $q > 1$ 时, $\Phi_{xy \cdot z}$ 是 $q \times q$ 矩阵, 是偏相关系数的推广。

第一典则偏相关系数: $\sqrt{\lambda_{\max}(\Phi_{xy \cdot z})}$, 度量了控制 $\mathbf{z}$ 条件下, 随机向量 $\mathbf{x}, \mathbf{y}$ 之间相关性大小 (介于 0, 1 之间)。

$q = 1$ 时, $\Phi_{xy \cdot z} = \Sigma_{yx \cdot z}\Sigma_{xx \cdot z}^{-1}\Sigma_{xy \cdot z}/\Sigma_{yy \cdot z} \in [0, 1]$, 称为偏决定系数。特别地, 若 $p = q = 1$, 即 x, y 都是一元随机变量, 则 $\Phi_{xy \cdot z} = \rho_{xy \cdot z}^2$ 。

如何检验? “大概”可以这么检验:

$$nr^2 > \chi_{|p-q|+1}^2(\alpha), \text{ 否定原假设}$$

这里 r^2 代表典则相关系数的平方、决定系数、“典则偏相关系数的平方或偏决定系数”。

例 1.2.4.

1.2.5.4 偏相关系数矩阵

命题 1.2.6 (分块矩阵的逆)

设正定矩阵

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

则

$$\Sigma^{-1} = \begin{pmatrix} \Sigma_{11 \cdot 2}^{-1} & -\Sigma_{11 \cdot 2}^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\ -\Sigma_{22 \cdot 1}^{-1}\Sigma_{21}\Sigma_{11}^{-1} & \Sigma_{22 \cdot 1}^{-1} \end{pmatrix}$$

其中 $\Sigma_{11 \cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, $\Sigma_{22 \cdot 1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ 。

证明

命题 1.2.7

随机向量 $\mathbf{x}_{p \times 1}$ 的协方差矩阵为 $\Sigma = (\sigma_{ij})$, $\Omega = \Sigma^{-1} = (\omega_{ij})$ 称为精度矩阵 (precision matrix), 记 $C = \text{diag}(\Omega)$, 则 x_i, x_j 的偏相关系数

$$\rho_{ij \cdot \text{other}} = \begin{cases} -\omega_{ij}/\sqrt{\omega_{ii}\omega_{jj}}, & i \neq j \\ 1, & i = j \end{cases}$$

且偏相关系数矩阵

$$R_{\text{partial}} = (\rho_{ij \cdot \text{other}}) = 2I_p - C^{-\frac{1}{2}}\Omega C^{-\frac{1}{2}}$$

注

1. Σ 元素包含了相关性信息; Σ^{-1} 元素包含了片相关性信息。
2. 基于 Σ^{-1} 的偏相关系数计算法则类似于基于 Σ 的相关系数计算法则, 但相差一个符号。
3. $\rho_{ij \cdot \text{other}} = 0 \Leftrightarrow \omega_{ij} = 0$ 。

1.3 多元正态分布

1. 多元正态假设下，偏相关系数为条件相关系数。
2. 条件分布满足（启发）所有线性模型的假设。

定义 1.3.1

p 维随机向量 $\mathbf{x}$ 的密度为

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right), \mathbf{x} \in \mathbb{R}^p$$

其中参数 $\boldsymbol{\mu} \in \mathbb{R}^p$, Σ 为 $p \times p$ 正定矩阵，则 $\mathbf{x}$ 服从多元正态分布，记为 $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$ 。

当 $\boldsymbol{\mu} = \mathbf{0}, \Sigma = I_p$ 时，为多元标准正态分布，即每个分量都服从标准正态分布且相互独立。

引理 1.3.1

若 $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$ ，则对任意可逆常数矩阵 A ，常数向量 $\mathbf{b}_{p \times 1}$ ， $A\mathbf{x} + \mathbf{b} \sim N_p(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma A^T)$ 。

推论 1.3.1

1. $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma) \Rightarrow \mathbf{y} = \Sigma^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\mu}) \sim N_p(\mathbf{0}, I_p)$ 。
2. $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma) \Leftrightarrow \mathbf{x}$ 可以表示为 $\mathbf{x} = B\mathbf{y} + \boldsymbol{\mu}$ ，其中 $\mathbf{y} \sim N_p(\mathbf{0}, I_p)$ ，方阵 B 使得 $\Sigma = BB^T$ （通常 $B = \Sigma^{-\frac{1}{2}}$ ）。

定理 1.3.1

1. $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma) \Rightarrow \mathbb{E}(\mathbf{x}) = \boldsymbol{\mu}, \text{var}(\mathbf{x}) = \Sigma$ 。
2. $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma) \Rightarrow$ 矩母函数 $\mathbb{E}(\exp(\mathbf{t}^T \mathbf{x})) = \exp(\mathbf{t}^T \boldsymbol{\mu} + \frac{1}{2} \mathbf{t}^T \Sigma \mathbf{t}), \forall \mathbf{t} \in \mathbb{R}^p$ 。

定义 1.3.2 (奇异多元正态分布, P.L.Hsu)

假设 $\mathbf{y} \sim N_p(\mathbf{0}, I_p)$ ， B 是任一 $q \times p$ 常数矩阵， $\boldsymbol{\mu}$ 是 $q \times 1$ 常数向量，则称 $\mathbf{x} = B\mathbf{y} + \boldsymbol{\mu}$ 的分布为奇异 (singular) 多元正态分布，记作 $\mathbf{x} \sim N_q(\boldsymbol{\mu}, BB^T)$ 。

下面讨论多元正态 $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$ 的分量的分布，为此划分

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

其中 $\mathbf{x}_1, \boldsymbol{\mu}_1$ 为 $q \times 1$ 向量， Σ_{11} 为 $q \times q$ 矩阵。

引理 1.3.2

若 $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$ ，且如上划分，则 $\mathbf{x}_1$ 与 $\mathbf{x}_2$ 独立 $\Leftrightarrow \Sigma_{12} = \mathbf{0}$ ，此时 $\mathbf{x}_1 \sim N_q(\boldsymbol{\mu}_1, \Sigma_{11}), \mathbf{x}_2 \sim N_{p-q}(\boldsymbol{\mu}_2, \Sigma_{22})$ 。

定理 1.3.2 (边际与条件分布)

若 $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$ ，且如前划分，则

1. $\mathbf{x}_1 \sim N_q(\boldsymbol{\mu}_1, \Sigma_{11}), \mathbf{x}_2 \sim N_{p-q}(\boldsymbol{\mu}_2, \Sigma_{22})$ 。
2. $\mathbf{x}_1 | \mathbf{x}_2 \sim N_q(\boldsymbol{\mu}_1 + \Sigma_{12} \Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \Sigma_{11.2})$ 。

证明

条件分布 $\mathbf{x}_1 | \mathbf{x}_2 \sim N_q(\boldsymbol{\mu}_1 + \Sigma_{12} \Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \Sigma_{11.2})$ 蕴含了如下事实：

1. 条件方差常数： $\text{var}(\mathbf{x}_1 | \mathbf{x}_2) = \Sigma_{11.2}$ 是常数矩阵，与 $\mathbf{x}_2$ 无关。
2. 条件期望线性： $\mathbb{E}(\mathbf{x}_1 | \mathbf{x}_2) = \boldsymbol{\mu}_1 + \Sigma_{12} \Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$ 是 $\mathbf{x}_2$ 的线性函数。
3. 去相关化： $\boldsymbol{\varepsilon} = \mathbf{x}_1 - \mathbb{E}(\mathbf{x}_1 | \mathbf{x}_2) = \mathbf{x}_1 - [\boldsymbol{\mu}_1 + \Sigma_{12} \Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)]$ ， $\boldsymbol{\varepsilon} | \mathbf{x}_2 \sim N_q(\mathbf{0}, \Sigma_{11.2})$ 与 $\mathbf{x}_2$ 无关，所以 $\boldsymbol{\varepsilon} \perp \mathbf{x}_2, \boldsymbol{\varepsilon} \sim$

$N_q(0, \Sigma_{11 \cdot 2})$.

4. 多元正态线性模型（改写去相关化）：

$$\mathbf{x}_1 = (\boldsymbol{\mu}_1 - \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{\mu}_2) + \Sigma_{12}\Sigma_{22}^{-1}\mathbf{x}_2 + \varepsilon \stackrel{\text{def}}{=} \mathbf{a} + \mathbf{b}^T \mathbf{x}_2 + \varepsilon$$

其中 $\varepsilon \perp \mathbf{x}_2, \varepsilon \sim N_q(0, \Sigma_{11 \cdot 2})$.

一般的多元线性回归模型放松误差正态性：

$$\mathbf{x}_1 = \mathbf{a} + \mathbf{b}^T \mathbf{x}_2 + \varepsilon, \varepsilon \perp \mathbf{x}_2, \varepsilon \sim (0, \Sigma_{11 \cdot 2})$$

定理 1.3.3

若 $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$ ，则对任意行满秩 $q \times p$ 常数矩阵 A ，常数向量 $\mathbf{b}_{p \times 1}$ ，有 $A\mathbf{x} + \mathbf{b} \sim N_q(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma A^T)$.

证明

定理 1.3.4

1. 多元正态假设下，偏相关系数等于条件相关系数。
2. 设 $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \Sigma)$ ， $\Sigma^{-1} = (\omega_{ij})$. 若 $\omega_{ij} = 0$ ，则给定其他变量时， x_i 与 x_j 条件独立。

1.4 线性代数知识补充

去相关化是随机变量的投影，条件期望是投影，回归模型的最小二乘法是欧式向量空间的投影。

1.4.1 向量空间

定义 1.4.1

集合 S 上所有实函数组成的集合： $\mathbb{R}^S = \{f: S \rightarrow \mathbb{R}\}$ ，定义 $\mathbb{R}^S$ 上的运算： $\forall f, g \in \mathbb{R}^S, s \in S, \lambda \in \mathbb{R}$,

1. 加法： $(f+g)(s) = f(s) + g(s)$;
2. 数乘： $(\lambda f)(s) = \lambda f(s)$.

于是 $\mathbb{R}^S$ 构成向量空间，称之为函数空间。

特别地， $\mathbb{R}^{\mathbb{R}}$ 称为实函数空间，定义内积：

$$\langle f, g \rangle = \int f(x)g(x)dx$$

诱导出模长：

$$\|f\| = \left(\int f(x)^2 dx \right)^{\frac{1}{2}}$$

所有平方可积函数组成空间 $L^2 = \{f: \int f(x)^2 dx < \infty\} \subset \mathbb{R}^{\mathbb{R}}$.

例 1.4.1.

实向量空间 $\mathbb{R}^n = \{\mathbf{x} = (x_1, \dots, x_n)^T : x_i \in \mathbb{R}\}$ 实际上是 $\mathbb{R}^{\{1, 2, \dots, n\}}$ ：

$$\mathbf{x}(i) = x_i, i = 1, 2, \dots, n$$

例 1.4.2.

考虑概率空间 $(\Omega, \mathcal{F}, P)$, 样本空间 Ω 上的实函数集合 $\mathbb{R}^\Omega$, 只考虑 $\mathbb{R}^\Omega$ 中 $\{w : x(w) < t\} \in \mathcal{F}$ 的那些函数 (即可计算概率的), 称之为随机变量。定义内积

$$\langle x, y \rangle = \mathbb{E}(xy)$$

诱导出模长:

$$\|x\| = \sqrt{\mathbb{E}x^2}$$

所有二阶矩存在的随机变量组成内积向量空间:

$$L^2(\mathcal{F}) = \{x : x \text{ 关于 } \mathcal{F} \text{ 可测}, \mathbb{E}x^2 < \infty\} \subset \mathbb{R}^\Omega$$

虽然随机变量叫“变量”, 实际上它是函数, 是向量空间中的向量。

定理 1.4.1

Hilbert 空间之间的连续线性变换 $T \in L(V, W)$ 具有唯一的伴随变换 $T^* \in L(W, V)$ 使得

$$(Tv, w) = (v, T^*w), \quad \forall v \in V, w \in W$$

若 $T = T^*$, 称其自伴随; 欧氏空间中, 矩阵 A 的伴随矩阵就是其转置矩阵 A^T , 自伴随即对称阵。

定义 1.4.2

假设 V 是内积向量空间, $U \subset V$ 是其子空间, 则任何 $v \in V$ 在 U 上的正交投影 $\hat{v} \in U$ 满足 $v - \hat{v} \perp U$, 即

$$\forall u \in U, \langle v - \hat{v}, u \rangle = 0 \Leftrightarrow \langle v, u \rangle = \langle \hat{v}, u \rangle$$

容易验证投影是一个线性变换, 记 $\hat{v} = P_U v$.

正交投影是极小化误差平方和问题的最优解。

定理 1.4.2

$$P_U(v) = \operatorname{argmin}_{u \in U} \|u - v\|^2.$$

证明 $u \in U$, $\hat{v} = P_U v \in U$, 所以 $u - \hat{v} \in U$. 故 $v - \hat{v} \perp u - \hat{v}$, 由毕达哥拉斯定理:

$$\|u - v\|^2 = \|v - \hat{v} + u - \hat{v}\|^2 = \|v - \hat{v}\|^2 + \|u - \hat{v}\|^2 \geq \|\hat{v} - v\|^2$$

命题 1.4.1

V 是内积空间, $u \in V$, $U = \{\lambda u : \lambda \in \mathbb{R}\}$, 则

$$P_U(v) = \frac{\langle v, u \rangle}{\langle u, u \rangle} u$$

证明 $\hat{v} \in U \Rightarrow \hat{v} = \lambda u$, 由

$$\langle v - \hat{v}, u \rangle = 0 \Rightarrow \lambda = \frac{\langle v, u \rangle}{\langle u, u \rangle}$$

命题 1.4.2 (Cauchy-Schwarz)

V 是内积空间, $\forall u, v \in V$ 有 $|\langle u, v \rangle| \leq \|u\| \cdot \|v\|$.

证明 利用 $\|v\|^2 \geq \|\hat{v}\|^2$, 再由命题 1.6.1 得证。

命题 1.4.3

V 是内积空间, $\{u_n\}_{n=1}^m \subset V$ 相互正交, $U = \text{span}\{u_n\}_{n=1}^m$, 则

$$P_U(v) = \sum_{n=1}^m \frac{\langle v, u_n \rangle}{\langle u_n, u_n \rangle} u_n$$

命题 1.4.4 (Schmidt 正交化)

内积空间中任何一组线性无关向量 $\{u_n\}_{n=1}^m$ 可以将其正交化:

$$e_1 = u_1, e_2 = u_2 - \frac{\langle u_2, e_1 \rangle}{\langle e_1, e_1 \rangle} e_1, e_3 = u_3 - \frac{\langle u_3, e_1 \rangle}{\langle e_1, e_1 \rangle} e_1 - \frac{\langle u_3, e_2 \rangle}{\langle e_2, e_2 \rangle} e_2, \dots$$

命题 1.4.5

U 是内积空间 V 的有限维子空间, 则任何 $v \in V$ 都存在 U 上的正交投影 $P_U(v)$.

泛函中的结论是: Hilbert 空间的闭线性子空间上必有正交投影。

定义 1.4.3 (条件期望)

概率空间 $(\Omega, \mathcal{F}, P)$, 子 σ 域 $\mathcal{G} \in \mathcal{F}$, 考虑二阶矩存在的随机变量空间:

$$V = L^2(\mathcal{F}) = \{y: y \text{ 关于 } \mathcal{F} \text{ 可测}, \mathbb{E}y^2 < \infty\}$$

限制 r.v. 的复杂性, 取其子空间

$$U = L^2(\mathcal{G}) = \{u: u \text{ 关于 } \mathcal{G} \text{ 可测}, \mathbb{E}u^2 < \infty\}$$

那么, 随机变量 $y \in V$ 在 U 上的投影 $P_U y$ 就称为条件期望 $\mathbb{E}(y|\mathcal{G})$, 其满足:

$$\langle y - \mathbb{E}(y|\mathcal{G}), u \rangle = \mathbb{E}(y - \mathbb{E}(y|\mathcal{G}))u = 0, \forall u \in U$$

随机向量 $\mathbf{x}$ 的所有函数

$$U = \{f(\mathbf{x}): f \text{ 可测}, \mathbb{E}f(\mathbf{x})^2 < \infty\}$$

即 $U = L^2(\sigma(\mathbf{x}))$, 此时条件期望是 y 在所有 $\mathbf{x}$ 的可测函数构成的空间上的投影, 是下述最小二乘问题的最优解:

$$\mathbb{E}(y|\mathbf{x}) = \underset{f}{\operatorname{argmin}} \mathbb{E}(y - f(\mathbf{x}))^2$$

例 1.4.3.

$V = L^2(\mathcal{F})$, 对于任何 r.v. x 和常数 $1 \in V$,

$$P_1(x) = \frac{\langle x, 1 \rangle}{\langle 1, 1 \rangle} 1 = \mathbb{E}x$$

令 $x^\perp = x - P_1 x = x - \mathbb{E}x$, 则

$$\|x^\perp\|^2 = \mathbb{E}(x^\perp)^2 = \mathbb{E}(x - \mathbb{E}x)^2 = \operatorname{var}(x)$$

这表明方差和随机变量长度的平方有某种联系。

例 1.4.4.

$V = L^2(\mathcal{F})$, r.v. $x_1, \dots, x_p$ 和 $1 \in V$ 张成的子空间

$$U = \{a + \mathbf{b}^T \mathbf{x} : a \in \mathbb{R}, \mathbf{b} \in \mathbb{R}^p\}$$

假设 r.v. y 在 U 上的投影为 $P_U y = a + \mathbf{b}^T \mathbf{x}$, 则

$$\begin{aligned} \langle y - (a + \mathbf{b}^T \mathbf{x}), 1 \rangle &= \mathbb{E}(y - (a + \mathbf{b}^T \mathbf{x})) = 0 \\ \langle y - (a + \mathbf{b}^T \mathbf{x}), x_i \rangle &= \mathbb{E}x_i(y - (a + \mathbf{b}^T \mathbf{x})) = 0, i = 1, 2, \dots, p \end{aligned}$$

解得 $\mathbf{b}^T = \Sigma_{yx} \Sigma_{xx}^{-1}$, $a = \mathbb{E}y - \mathbf{b}^T \mathbb{E}\mathbf{x}$, 即

$$P_U y = \mathbb{E}y + \Sigma_{yx} \Sigma_{xx}^{-1} (\mathbf{x} - \mathbb{E}\mathbf{x})$$

这是如下最小二乘问题的最优解:

$$\operatorname{argmin}_{a, \mathbf{b}} \mathbb{E}(y - (a + \mathbf{b}^T \mathbf{x}))^2$$

这表明, y 关于 $\mathbf{x}$ 去相关化 $\Leftrightarrow y$ 向 $\mathbf{x}$ 正交补空间投影, 令

$$\varepsilon := y - P_U y = y - \Sigma_{yx} \Sigma_{xx}^{-1} \mathbf{x} - (\mathbb{E}y - \mathbf{b}^T \mathbb{E}\mathbf{x})$$

其与 $\mathbf{x}$ 不相关, 改写即线性模型:

$$y = (\mathbb{E}y - \mathbf{b}^T \mathbb{E}\mathbf{x}) + \Sigma_{yx} \Sigma_{xx}^{-1} \mathbf{x} + \varepsilon = a + \mathbf{b}^T \mathbf{x} + \varepsilon$$

命题 1.4.6 (投影算子相关性质)

内积空间 V 的子空间为 U , 即 $P_U = P$, 则

1. P 若存在必唯一;
2. 对于任何 $\mathbf{u} \in U, P\mathbf{u} = \mathbf{u}$;
3. $P^2 = P$;
4. P 自共轭 (自伴随);
5. 若算子 $T \neq 0 \in L(V, V)$ 满足幂等且自伴随, 则 T 是投影算子。

1.4.2 矩阵

定义 1.4.4

约定记号: 矩阵 $A = (a_{ij})_{m \times n} \in \mathbb{R}^{m \times n}$,

$$A = \begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_m \end{pmatrix} = \begin{pmatrix} \alpha_1^T \\ \vdots \\ \alpha_n^T \end{pmatrix}$$

记

$$C(A) = C(\mathbf{a}_1, \dots, \mathbf{a}_m) = \operatorname{span}\{\mathbf{a}_1, \dots, \mathbf{a}_m\} = \{A\mathbf{x} : \mathbf{x} \in \mathbb{R}^m\}$$

$$N(A) = \ker(A) = \{\mathbf{x} \in \mathbb{R}^m : A\mathbf{x} = 0\}$$

定义

$$\text{rank}(A) = \dim(C(A))$$

$$\text{nullity}(A) = \dim(N(A))$$

定理 1.4.3

$$\text{rank}(A) = \text{rank}(A^T).$$

定理 1.4.4

$$C(A) = C(AA^T).$$

定理 1.4.5

$N(A) = C(A^T)^\perp$, 即 $C(A^T) \oplus N(A) = \mathbb{R}^m$, 因此

$$\text{rank}(A_{n \times m}) = m - \text{nullity}(A) = n - \text{nullity}(A^\perp)$$

例 1.4.5. 常用线性变换方阵

取单位向量

$$\boldsymbol{v} = \begin{pmatrix} c \\ s \end{pmatrix} = \begin{pmatrix} \cos \theta \\ \sin \theta \end{pmatrix} \in \mathbb{R}^2$$

1. 拉伸: 等同于矩阵数乘 $cA = (cI) \cdot A = (c \cdot a_{ij})$, 但这会引起混乱 (因为我们把常量视为 1×1 的矩阵);
2. 旋转: $\mathbb{R}^2$ 上的旋转阵

$$O = \begin{pmatrix} c & -s \\ s & c \end{pmatrix}$$

3. 投影: $\mathbb{R}^2$ 上的投影阵

$$P = \boldsymbol{v}\boldsymbol{v}^T = \begin{pmatrix} c^2 & cs \\ cs & s^2 \end{pmatrix}$$

一般的 n 阶投影阵满足对称幂等: $P^2 = P, P^T = P$. P 的特征根为 0, 1, 所以存在正交矩阵 O 使得

$$P = O \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} O^T, r = \text{rank}(P)$$

二次型可简化为平方和:

$$\boldsymbol{x}^T P \boldsymbol{x} = \|\boldsymbol{P}\boldsymbol{x}\|^2 = \boldsymbol{x}^T O \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} O^T \boldsymbol{x} = \|\boldsymbol{y}\|^2$$

其中 $\boldsymbol{y} = O^T \boldsymbol{x}$.

4. 反射: $\mathbb{R}^2$ 上的反射阵:

$$R = 2\boldsymbol{v}\boldsymbol{v}^T - I_n = \begin{pmatrix} 2c^2 - 1 & 2cs \\ 2cs & 2s^2 - 1 \end{pmatrix} = \begin{pmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{pmatrix}$$

一般的 n 阶反射阵 $R = I_n - 2P$ 或者 $2P - I_n$, 其中 P 为投影矩阵. $R^2 = I_n$, 可称之为“平方根”矩阵。

5. 若 n 阶方阵 O 满足 $O^T O = I_n$, 称之为正交矩阵, 正交 = 旋转 + 反射。

命题 1.4.7

若 n 阶方阵 O 满足 $O^T O = I_n$, 则 $OO^T = I_n$.

1.4.3 行列空间的特征刻画

特征值相关理论就不再赘述了, 后面用到结论的 gap 会标明。。

定理 1.4.6 (对称矩阵的谱分解)

任一 n 阶对称方阵 A 可表示为 $A = V\Lambda V^T$, 其中 $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ 为特征值对角矩阵, $V = (v_1, \dots, v_n)$ 是正交矩阵。

定理 1.4.7 (复特征根)

A 是 n 阶实矩阵, 下列事实等价: ($a, b \in \mathbb{R}, u, v \in \mathbb{R}^n$)

1. $A(u + vi) = (u + vi)(a + bi)$, 即 $a + bi$ 是 A 的复特征根, 特征向量 $u + vi$.
2. $A(u - vi) = (u - vi)(a + bi)$, 即 $a - bi$ 是 A 的复特征根, 特征向量 $u - vi$.
- 3.

$$A(u, v) = (u, v) \begin{pmatrix} a & b \\ -b & a \end{pmatrix}$$

对于非方阵矩阵, 其行和列如何用类似于特征向量的不变量进行刻画? 我们将利用对称方阵 AA^T 以及 $A^T A$ 的特征向量, $A^T A$ 的正特征值的平方根称为 A 的奇异值。奇异值分解 SVD 的发现来自于以下事实: 若 v 是 $A^T A$ 的特征向量, 则 Av 是 AA^T 的特征向量。

定理 1.4.8 (奇异值分解, SVD)

任一秩为 r 的 $n \times m$ 矩阵 A 可以表示为

$$A = U_{n \times r} D_{r \times r} V_{r \times m}^T$$

其中 $U^T U = V^T V = I_r$, $D = \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_r})$, 其中 $\lambda_1 \geq \dots \geq \lambda_r > 0$ 为 $A^T A$ 或者 AA^T 的 r 个正特征根, $U_{n \times r} = (u_1, \dots, u_r)$ 的第 i 列 u_i 为 AA^T 的对应于特征根 λ_i 的特征向量, $V_{m \times r} = (v_1, \dots, v_r)$ 的第 i 列 v_i 为 $A^T A$ 的对应于特征根 λ_i 的特征向量。

也可以将 U, V 扩展成正交方阵 $\tilde{U}, \tilde{V}$, 奇异值分解也表述为

$$A_{n \times m} = \tilde{U}_{n \times n} \begin{pmatrix} D_{r \times r} & 0 \\ 0 & 0 \end{pmatrix} \tilde{V}_{m \times m}^T$$

小结 1.

$$A = UDV^T,$$

1. 行特征: V 是行空间 $C(A^T)$ 的正交基, $A^T A V = V D^2$.
2. 列特征: U 是行空间 $C(A)$ 的正交基, $A A^T U = U D^2$.
3. 对偶: $AV = UD, A^T U = VD$.
4. 若 A 各列是中心化的, 则 UD 的各列称为主成分 (第 j 列称为第 j 主成分, 重要性、方差贡献为 λ_j), V 的各列称为主成分方向。

5.

$$A = UDV^T = \sum_{j=1}^r \sqrt{\lambda_j} \mathbf{u}_j \mathbf{v}_j^T$$

前 k 项是 A 的 k 阶 (秩 k) 最佳逼近。

1.4.4 广义逆矩阵

定义 1.4.5

任何 $A_{n \times m}$, 若 $X_{m \times n}$ 满足 $AXA = A$, 则称 X 为 A 的广义逆, 记作 A^- .

A^- 是一定存在的, 例如 $VD^{-1}U^T$, 那唯一性如何?

命题 1.4.8

当 A 是满秩方阵时广义逆唯一, 此时 $A^- = A^{-1}$, 否则不唯一。

证明 当 $r = \text{rank}(A) < \max(m, n)$ 时, 若 A 有奇异值分解:

$$A = \tilde{U} \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} \tilde{V}^T$$

其中 $\tilde{U}, \tilde{V}$ 是正交阵, 设 X 是任一广义逆, 则

$$\begin{aligned} AXA = A &\Leftrightarrow \tilde{U} \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} \tilde{V}^T X \tilde{U} \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} \tilde{V}^T = \tilde{U} \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} \tilde{V}^T \\ &\Leftrightarrow \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{pmatrix} \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} \end{aligned}$$

其中

$$Y = \tilde{V}^T X \tilde{U} = \begin{pmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{pmatrix}$$

则 $Y_{11} = D^{-1}$, 其他任意, 广义逆可取:

$$X = \tilde{V} Y \tilde{U}^T = \tilde{V} \begin{pmatrix} D^{-1} & * \\ * & * \end{pmatrix} \tilde{U}^T$$

定义 1.4.6

任何 $A_{n \times m}$, 若 $X_{n \times m}$ 满足:

$$AXA = A, XAX = X, (AX)^T = AX, (XA)^T = XA$$

则称 X 为 A 的 Moore-Penrose 广义逆, 记作 A^+ .

命题 1.4.9

$A^+ = VD^{-1}U^T$, 存在唯一。

定理 1.4.9

若方程 $Ax = b$ 有解 ($b \neq 0$), 则任何一个解都具有形式 $x = A^-b$.

命题 1.4.10

若方程 $Ax = b$ 有解 ($b \neq 0$), 则 $x = A^+b$ 是所有解中长度最小的解。

1.4.5 欧氏空间中的正交投影

定理 1.4.10 (最小二乘的欧式情形)

$y \in \mathbb{R}^n$, 记 $\hat{y} = P(y|V)$ 为 y 在子空间 $V \subset \mathbb{R}^n$ 上的投影, 则 $\|y - \hat{y}\|^2 = \min_{u \in V} \|y - u\|^2$.

命题 1.4.11

$V = C(A_{n \times m}) \subset \mathbb{R}^n$, $y \in \mathbb{R}^n$, 则 y 在 V 上的投影为

$$\hat{y} = P_A y = A(A^T A)^{-1} A^T y$$

其中 A 对应的投影矩阵为 $P_A = A(A^T A)^{-1} A^T$.

命题 1.4.12

A 是 $n \times m$ 矩阵,

1. $P_A = A(A^T A)^{-1} A^T$ 唯一, 与广义逆的具体选取无关。
2. 投影矩阵 = 对称幂等阵: $P_A = A(A^T A)^{-1} A^T$ 是对称幂等阵; 反之, 任一对称幂等阵是某个投影矩阵。
3. 投影矩阵特征值只有 1, 0, 而且 1 的代数重数为 r .
4. $P_A A = A$.
5. P_A 是 $C(A)$ 对应的投影阵, 则 $I_n - P_A$ 是 $C(A)^\perp$ 对应的投影阵, 二者正交。
6. 线性子空间 V 对应的投影矩阵与空间基的选取无关。
7. 若 $A^T A$ 为对角阵, 则 $P_A = \sum P_{a_k}$.
8. 按列划分 $A = (A_1 \ A_2)$, 若 $A_1^T A_2 = 0$, 则 $P_A = P_{A_1} + P_{A_2}$.

命题 1.4.13

若 $A_{n \times m}$ 有奇异值分解 $A = U_{n \times r} D_{r \times r} V_{r \times m}^T$, 那么 $P_A = U U^T$.

第 2 章 线性回归

2.1 引言

定义 2.1.1

(一元) 线性回归模型假设响应变量 y 和自变量向量 $\mathbf{x}$ 满足方程:

$$y = a + \mathbf{b}^T \mathbf{x} + \varepsilon$$

其中误差项 ε 是不可观测的随机变量, 满足 Gauss-Markov 假设:

- (1) 截距 a 可识别: $\mathbb{E}\varepsilon = 0$.
- (2) 方差齐性 (Homoscedasticity): $\text{var}(\varepsilon) = \sigma^2$.
- (3) 外生性 (Exogeneity): ε 与 $\mathbf{x}$ 独立。

条件 (1)(2) 简记为 $\varepsilon \sim (0, \sigma^2)$. 特殊地, 当 $(y, \mathbf{x})$ 服从多元正态时, GM 假设都成立。

命题 2.1.1

若 $(y, \mathbf{x})$ 满足线性模型的 GM 假设, 则

- (1) 回归函数线性: $\mathbb{E}(y|\mathbf{x}) = a + \mathbf{b}^T \mathbf{x}$.
- (2) 条件方差常数: $\text{var}(y|\mathbf{x}) = \sigma^2$.

由此可见, 线性回归模型蕴含了条件期望 $\mathbb{E}(y|\mathbf{x})$ 关于 $\mathbf{x}$ 线性, 条件期望 $\mathbb{E}(y|\mathbf{x})$ 也被称为回归函数, 是平方损失下所有 $\mathbf{x}$ 的函数中对 y 的最佳逼近。所以线性回归模型实际上假设了“ $\mathbf{x}$ 的线性函数是所有 y 的函数 $g(\mathbf{x})$ 逼近中最佳的逼近”。

命题 2.1.1 可以推出 GM 假设的弱化版本:

- (1) $\mathbb{E}\varepsilon = \mathbb{E}(y - \mathbb{E}(y|\mathbf{x})) = 0$.
- (2) $\text{var}(\varepsilon) = \text{var}(y - a - \mathbf{b}^T \mathbf{x}) = \text{var}(\mathbb{E}(y - a - \mathbf{b}^T \mathbf{x}|\mathbf{x})) + \mathbb{E}(\text{var}(y - a - \mathbf{b}^T \mathbf{x}|\mathbf{x})) = 0 + \mathbb{E}(\text{var}(y|\mathbf{x})) = \sigma^2$.
- (3*) 由条件期望的性质, $\varepsilon = y - \mathbb{E}(y|\mathbf{x})$ 与 $\mathbf{x}$ 不相关 (但不一定独立)。

定义 2.1.2 (模型的等价定义)

假设 y 是响应变量, $\mathbf{x}$ 是自变量, $\mathbb{E}(y|\mathbf{x})$ 称为回归函数, 则线性回归模型假设:

- (1) 线性回归函数: $\mathbb{E}(y|\mathbf{x}) = a + \mathbf{b}^T \mathbf{x}$.
- (2) 方差常数/齐性: $\text{var}(y|\mathbf{x}) = \sigma^2$.

(3) $y - \mathbb{E}(y|\mathbf{x})$ 与 $\mathbf{x}$ 独立。

其他（非线性）回归模型通常对回归函数做假设，例如：

1. logistic 回归：响应变量 y 是伯努利变量，它与自变量 $\mathbf{x}$ 有关，假设 $y|\mathbf{x} \sim B(1, \pi(\mathbf{x}))$,

$$\pi(\mathbf{x}) = P(y = 1|\mathbf{x}) = 1 - P(y = 0|\mathbf{x}) = \mathbb{E}(y|\mathbf{x}) = \frac{\exp(a + \mathbf{b}^T \mathbf{x})}{1 + \exp(a + \mathbf{b}^T \mathbf{x})}$$

2. Poisson 回归：假设 y 是计数变量，服从 Poisson 分布 $\text{Pois}(\lambda(\mathbf{x}))$, $\lambda(\mathbf{x}) = \mathbb{E}(y|\mathbf{x}) = \exp(a + \mathbf{b}^T \mathbf{x})$.

定义 2.1.3 (向量空间中的投影)

$U \subset V$ 是向量空间，任何 $v \in V$ 在 U 上的正交投影（记作 $\hat{v} \in U$ ）满足

$$\langle v - \hat{v}, u \rangle = 0, \forall u \in U$$

定义 2.1.4 (从投影的角度定义条件期望/回归函数)

若对于任何可测的 $g(\mathbf{x})$,

$$\langle y - \varphi(\mathbf{x}), g(\mathbf{x}) \rangle = \mathbb{E}(y - \varphi(\mathbf{x}))g(\mathbf{x}) = 0$$

$\varphi(\mathbf{x})$ 称为 y 给定 $\mathbf{x}$ 的条件期望/回归函数，记作 $\mathbb{E}(y|\mathbf{x})$,

定义表明 $y - \varphi(\mathbf{x}) \perp L^2(\sigma(\mathbf{x}))$ ，其中

$$L^2(\sigma(\mathbf{x})) = \{g(\mathbf{x}) : g \text{ 可测}, \mathbb{E}g(\mathbf{x})^2 < \infty\}$$

称为所有 $\sigma(\mathbf{x})$ 可测函数组成的随机变量空间。因此， $\mathbb{E}(y|\mathbf{x})$ 是 y 在 $\mathbf{x}$ 的函数组成的空间上的正交投影，从而是 $\mathbf{x}$ 的线性函数是所有可测函数中对 y 的最佳逼近：

$$\mathbb{E}(y|\mathbf{x}) = \underset{g}{\operatorname{argmin}} \mathbb{E}(y - g(\mathbf{x}))^2$$

推论 2.1.1

假设 y 是任一随机变量， $\mathbf{x}$ 是随机向量，

1. $\mathbb{E}(\mathbb{E}(y|\mathbf{x})) = \mathbb{E}(y)$.

2. $\operatorname{var}(y) = \operatorname{var}[\mathbb{E}(y|\mathbf{x})] + \mathbb{E}[\operatorname{var}(y|\mathbf{x})]$

这也是决定系数定义的由来之一：

$$\Phi \stackrel{\text{def}}{=} \frac{\operatorname{var}(\mathbb{E}(y|\mathbf{x}))}{\operatorname{var}(y)}$$

证明

1. 条件期望定义中取 $g(\mathbf{x}) = 1$ 即可。

2. 令 $\varepsilon = y - \mathbb{E}(y|\mathbf{x})$ ，定义中取 $g(\mathbf{x}) = x_i$ ，则 $\mathbb{E}[y - \mathbb{E}(y|\mathbf{x})]x_i = 0$ ，即 ε 与 x_i 不相关， ε 与 $\mathbf{x}$ 不相关，正交分解 $y = \mathbb{E}(y|\mathbf{x}) + \varepsilon$ 两边同时求方差：

$$\operatorname{var}(y) = \operatorname{var}[\mathbb{E}(y|\mathbf{x})] + \operatorname{var}(\varepsilon)$$

而 $\operatorname{var}(\varepsilon) = \operatorname{var}[y - \mathbb{E}(y|\mathbf{x})] = \mathbb{E}[\operatorname{var}(y|\mathbf{x})]$.

命题 2.1.2

假设线性回归模型

$$y = a + \mathbf{b}^T \mathbf{x} + \varepsilon, \mathbb{E}\varepsilon = 0, \operatorname{var}(\varepsilon) = \sigma^2, \varepsilon \text{ 与 } \mathbf{x} \text{ 独立}$$

则参数 a, b, σ^2 均由 y, x 的均值和方差决定:

$$b = \Sigma_{xx}^{-1} \Sigma_{xy}, a = \mu_y - b^T \mu_x, \sigma^2 = \Sigma_{yy} - \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy} = \Sigma_{yy} (1 - \Phi)$$

其中

$$\Phi = \frac{\text{var}(b^T x)}{\text{var}(y)} = \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy} / \Sigma_{yy}$$

为决定系数。

由命题 2.1.2, 线性回归模型可以表示为:

$$y = a + \Sigma_{yx} \Sigma_{xx}^{-1} x + \varepsilon, \varepsilon \text{ 与 } x \text{ 独立}$$

这与去相关化非常相似:

$$y = \Sigma_{yx} \Sigma_{xx}^{-1} x + y^\perp, y^\perp \text{ 与 } x \text{ 不相关}$$

$y^\perp$ 可看作“误差项”。因为要求 $\mathbb{E}(\varepsilon) = 0$, 回归模型多出了常数项 a , 这并不影响相关性和独立性。

定义 2.1.5 (多元线性回归模型)

响应变量 y 是向量的情形: 假设随机向量 $y_{q \times 1}, x_{p \times 1}$ 满足:

$$y_{q \times 1} = a_{q \times 1} + B(q \times p) x_{p \times 1} + \varepsilon_{q \times 1}, \varepsilon \sim (0, \Sigma), \varepsilon \text{ 与 } x \text{ 独立}$$

称为多元线性回归模型, 其中 y 称为响应变量, x 称为自变量, 可观测; ε 不可观测, 一般称为误差。 a, B, Σ 为参数, 由 y, x 的均值和方差决定:

$$B = \Sigma_{yx} \Sigma_{xx}^{-1}, a = \mu_y - B \mu_x, \Sigma = \Sigma_{yy} - \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy}$$

但本课程范围内一般不讨论多元回归模型。

考虑 $y = a + b^T x + \varepsilon$ 中部分变量的回归系数, 假设 $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, 相应的 $b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$, 其中 x_1 是感兴趣的变量, x_2 是控制变量/干扰因素。

命题 2.1.3

假设模型:

$$y = a + b^T x + \varepsilon = a + b_1^T x_1 + b_2^T x_2 + \varepsilon, \varepsilon \sim (0, \sigma^2) \text{ 与 } x \text{ 独立}$$

其中 $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ 独立, $b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$. 令

$$x_1^\perp = x_1 - \Sigma_{12} \Sigma_{22}^{-1} x_2, y^\perp = y - \Sigma_{yx} \Sigma_{22}^{-1} x_2$$

则

- (1) $y^\perp = a + b_1^T x_1^\perp + \varepsilon, \mathbb{E}\varepsilon = 0, \text{var}(\varepsilon) = \sigma^2, \varepsilon \text{ 与 } x_1^\perp \text{ 独立}.$
- (2) $b_1 = \Sigma_{11 \cdot 2}^{-1} \Sigma_{1y \cdot 2}.$
- (3) 当 x_1 是随机变量时, $b_1 \propto \rho_{1y \cdot 2}$ 与偏相关系数成正比。
- (4) 部分/偏决定系数

$$R_{1y \cdot 2}^2 = \frac{\text{var}(b_1^T x_1^\perp)}{\text{var}(y^\perp)} = \frac{\Sigma_{y1 \cdot 2} \Sigma_{11 \cdot 2}^{-1} \Sigma_{1y \cdot 2}}{\Sigma_{yy \cdot 2}}$$

推论 2.1.2 (参数的矩估计)

假设样本 $(y_i, x_i), i = 1, 2, \dots, n$ 来自于线性模型

$$y = a + b^T x + \varepsilon$$

假设样本方差矩阵和样本均值为

$$S = \begin{pmatrix} S_{yy} & S_{yx} \\ S_{xy} & S_{xx} \end{pmatrix}, \begin{pmatrix} \bar{y} \\ \bar{x} \end{pmatrix}$$

参数 a, b, σ^2 的矩估计为:

$$\begin{cases} \hat{b} = S_{xx}^{-1} S_{xy} \\ \hat{a} = \bar{y} - \hat{b}^T \bar{x} \\ \hat{\sigma}^2 = S_{yy \cdot x} = S_{yy} - S_{yx} S_{xx}^{-1} S_{xy} \end{cases}$$

另外, $R^2 = \hat{\Phi} = S_{yx} S_{xx}^{-1} S_{xy} / S_{yy}$.



后面将会看到, 除了误差方差之外, 所有参数的最小二乘估计等于上述矩估计, 误差方差的最小二乘估计:

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

2.2 简单线性回归模型

2.2.1 定义

定义 2.2.1 (简单线性模型)

二元随机向量 (x, y) 满足模型 $y = a + bx + \varepsilon, \varepsilon \sim (0, \sigma^2)$, 其中 ε 与 x 独立。



我们知道, 简单模型的参数由 x, y 的均值、方差和协方差决定:

$$b = \frac{\Sigma_{xy}}{\Sigma_{xx}} = \rho \frac{\sigma_y}{\sigma_x}, a = \mu_y - b\mu_x, \sigma^2 = (1 - \rho^2)\sigma_y^2$$

所以回归函数或均值函数:

$$\mathbb{E}(y|x) = a + bx = \mu_y + \rho \frac{\sigma_y}{\sigma_x} (x - \mu_x)$$

其中的 $\rho, |\rho| < 1$ 导致回归效应: 当 x 比其均值 μ_x 大 k 个标准差, 即 $x = \mu_x + k\sigma_x$ 时 $\mathbb{E}(y|x) = \mu_y + \rho k \sigma_y$ 比其均值 μ_y 大 $\rho k < k$ 个标准差。而 x, y 等比例增加的情况对应于上述方程中 $\rho = 1$ 的情况。

2.2.2 最小二乘法

定义 2.2.2 (简单线性模型 (样本模型))

假设独立样本 $(x_i, y_i), i = 1, 2, \dots, n$ 来自于总体模型

$$y = a + bx + \varepsilon, \varepsilon \sim (0, \sigma^2), \varepsilon \perp x$$

即 $(x_i, y_i), i = 1, 2, \dots, n$ 满足模型:

$$y_i = a + bx_i + \varepsilon_i, \varepsilon_i \text{ iid. } \sim (0, \sigma^2), \varepsilon_i \perp x_i$$



以 x_i 预测 y_i 的误差: $\varepsilon_i = y_i - a - bx_i = y_i - \mathbb{E}(y_i|x_i)$. 为了求解回归直线, 即求解 a, b , 利用最小二乘法 (least squares, LS) 极小化误差平方和:

$$\min_{a, b} \sum_{i=1}^n \varepsilon_i^2 = \min_{a, b} \sum_{i=1}^n (y_i - a - bx_i)^2$$

注 为什么我们只考虑极小化竖直方向 y_i 的误差, 而不是水平方向 x_i 的误差, 或者样本点到直线的垂直距离? 这是因为 x 和 y 的地位不同, 模型以 x 的函数描述 y . 如果认为地位对称, 我们希望求解一条直线能最好地表示二

者的关系, 那么该直线应该写成对称形式:

$$ax + by = c$$

此时, 应当极小化样本点到直线的垂直距离, 这被称为对称回归 (total least squares).

命题 2.2.1

最小二乘问题

$$\min_{a,b} \sum_{i=1}^n \varepsilon_i^2 = \min_{a,b} \sum_{i=1}^n (y_i - a - bx_i)^2$$

的最优解, 即最小二乘估计 LS 为

$$\hat{b} = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} \stackrel{\text{def}}{=} \frac{S_{xy}}{S_{xx}}, \quad \hat{a} = \bar{y} - \hat{b} \bar{x}$$

可以发现, 其与矩估计完全相同。

证明 误差平方和

$$\sum_{i=1}^n (y_i - a - bx_i)^2$$

对 a, b 分别求导并令其等于零, 得到正则方程:

$$\begin{cases} \sum \varepsilon_i = \sum (y_i - a - bx_i) = 0 \\ \sum x_i \varepsilon_i = \sum x_i (y_i - a - bx_i) = 0 \end{cases}$$

第一个方程解得 $a = \bar{y} - b\bar{x}$, 带入到第二个方程可得

$$\sum x_i (y_i - \bar{y} - b(x_i - \bar{x})) = 0$$

然后就可以得出结论了。

LS 估计是否符合直观? 只考虑 $\hat{b}$, 形式上

$$\hat{b} = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x}) (y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x})^2 \cdot \frac{y_i - \bar{y}}{x_i - \bar{x}}}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x})^2 \frac{y_i - \bar{y}}{x_i - \bar{x}}}{\sum (x_i - \bar{x})^2} \stackrel{\text{def}}{=} \sum w_i z_i$$

其中

$$w_i = \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}, \quad \sum w_i = 1, \quad z_i = \frac{y_i - \bar{y}}{x_i - \bar{x}} \text{ or } \frac{y_i}{x_i - \bar{x}}$$

所以 $\hat{b}$ 是每个样本点处得局部斜率估计 z_i 的加权平均, 斜率 z_i 是十分直观的, 该如何理解权重 w_i ?

引理 2.2.1

$z_1, \dots, z_n$ 独立, $\sigma_i^2 = \text{var}(z_i)$ 已知但不相等, 假设 $w_i \geq 0, \sum w_i = 1$, 则

$$\frac{\sum z_i / \sigma_i^2}{\sum 1 / \sigma_i^2}$$

是所有加权平均 $\sum w_i z_i$ 中方差最小的, 即

$$\text{var}(\sum w_i z_i) = \sum w_i^2 \sigma_i^2 \geq \frac{1}{\sum 1 / \sigma_i^2}$$

证明 由 Cauchy-Schwarz 不等式,

$$\text{var}(\sum w_i z_i) = \sum w_i^2 \sigma_i^2 \geq \frac{1}{\sum 1 / \sigma_i^2}, \quad w_i \propto 1 / \sigma_i^2 \text{ 时等号成立}$$

考虑 b 的加权平均估计 $\tilde{b} = \sum w_i z_i, w_i \geq 0, \sum w_i = 1$, 因为

$$\text{var}(z_i | x_i) = \text{var}\left(\frac{y_i}{x_i - \bar{x}} | x_i\right) = \frac{\sigma^2}{(x_i - \bar{x})^2}$$

因此最优权重 $w_i \propto (x_i - \bar{x})^2 / \sigma^2$, 此时 $\hat{b} = \sum w_i z_i$ 就是 LS 估计。

注 也存在许多其他权重选择, 这依赖于具体问题背景, 例如:

Error-in-variable 模型的广义矩估计 (GMM)

$$\tilde{b}_{EV} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})^2}{\sum (x_i - \bar{x})^2 (y_i - \bar{y})}$$

GM 假设第三条不成立时的工具变量估计

$$\tilde{b}_{IVLS} = \frac{\sum (u_i - \bar{u})(y_i - \bar{y})}{\sum (u_i - \bar{u})(x_i - \bar{x})}$$

其中 u_i 是外生变量/工具变量。

我们甚至可以考虑其他局部斜率, 比如 $z_{ij} = (y_i - y_j)/(x_i - x_j)$, 并构造 b 的估计 $\tilde{b} = \sum w_{ij} z_{ij}$, 这被称为 U 统计量。那么该如何选择权重?

命题 2.2.2

- $\hat{a}, \hat{b}$ 是无偏估计。
- 给定所有自变量 $\mathbf{x} = (x_1, \dots, x_n)^T$ 条件下,

$$\text{var} \left(\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} \middle| \mathbf{x} \right) = \begin{pmatrix} \sigma^2/n + \bar{x}^2 \sigma^2/s_{xx} & -\bar{x} \sigma^2/s_{xx} \\ -\bar{x} \sigma^2/s_{xx} & \sigma^2/s_{xx} \end{pmatrix}$$

这说明自变量方差越大, LS 估计的方差越小, 估计精度就越高。

证明 待补充。

定义 2.2.3 (拟合值与残差)

求出 LS 估计值后, 得到拟合回归直线: $y = \hat{a} + \hat{b}x$, 定义拟合值 $\hat{y}_i = \hat{a} + \hat{b}x_i$, 残差 $e_i = y_i - \hat{y}_i$ 。

所以正则方程能写为:

$$\sum e_i = 0, \sum x_i e_i = 0$$

可以记作 $\mathbf{e} \perp \mathbf{1}, \mathbf{e} \perp \mathbf{x}$

- 残差与拟合值不相关: $\sum \hat{y}_i e_i = 0, \mathbf{e} \perp \hat{\mathbf{y}}$
- 残差的均值: $\bar{e} = \sum e_i/n = 0$
- 拟合值的样本均值: $\bar{\hat{y}} = \sum \hat{y}_i/n = \bar{y}$

定义 2.2.4

记 $a_1, \dots, a_n$ 的平方和

$$s_{aa} = \sum (a_i - \bar{a})^2 = \|\mathbf{a} - \bar{a}\mathbf{1}\|^2$$

那么

- 总平方和 (响应的平方和)

$$SS_{\text{总}} = s_{yy} = \sum (y_i - \bar{y})^2 = \|\mathbf{y} - \bar{y}\mathbf{1}\|^2$$

- 回归平方和 (拟合值的平方和)

$$SS_{\text{回}} = s_{\hat{y}\hat{y}} = \sum (\hat{y}_i - \bar{y})^2 = \|\hat{\mathbf{y}} - \bar{y}\mathbf{1}\|^2$$

- 残差平方和:

$$RSS = s_{ee} = \sum e_i^2$$

除以 $(n-1)$ 就是残差的样本方差。

定义 2.2.5 ((样本版本的) 决定系数)

定义为自变量所能解释的响应变量总平方和的百分比

$$R^2 = \frac{SS_{\text{回}}}{SS_{\text{总}}} = \frac{s_{\hat{y}\hat{y}}}{s_{yy}} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

命题 2.2.3

$SS_{\text{回}} = s_{xy}^2/s_{xx}$, $RSS = s_{yy} - s_{xy}^2/s_{xx}$, $R^2 = r_{xy}^2$, 因此 $SS_{\text{总}} = s_{yy}$

2.2.3 误差方差的估计

实际上, $\sigma^2 = \text{var}(\varepsilon_i) = \mathbb{E}(\varepsilon_i^2)$, 而 e_i 可以看作对随机变量 ε_i 的预测, 我们可以尝试基于 $\{e_i\}$ 估计 σ^2 :

$$\hat{\sigma}^2 = \frac{1}{n-2} RSS = \frac{1}{n-2} \sum_{i=1}^n e_i^2$$

虽然不是由最小二乘法直接得到的, 但通常还是称之为 LS 估计。为什么除以 $n-2$ 而不是 $n-1$? 因为估计了两个参数 a, b 。

引理 2.2.2

1. $e_i = (\varepsilon_i - \bar{\varepsilon}) - (x_i - \bar{x})s_{x\varepsilon}/s_{xx}$.
2. $RSS = s_{yy} - s_{xy}^2/s_{xx} = s_{\varepsilon\varepsilon} - s_{x\varepsilon}^2/s_{xx}$.

证明 待补充。

命题 2.2.4

$\hat{\sigma}^2$ 是 σ^2 的无偏估计, 即 $\mathbb{E}(\hat{\sigma}^2) = \sigma^2$.

证明 待补充。

2.2.4 “LS 估计的方差” 的估计

$\text{var}(\hat{b}|\mathbf{x}) = \sigma^2/s_{xx}$ 中将 σ^2 的估计代入可得:

$$\widehat{\text{var}}(\hat{b}|\mathbf{x}) = \frac{\hat{\sigma}^2}{s_{xx}}$$

标准差: $\text{se}(\hat{b}) = \sqrt{\widehat{\text{var}}(\hat{b}|\mathbf{x})} = \hat{\sigma}/\sqrt{s_{xx}}$

Wald 检验方法是构造检验统计量常用的方法之一 (还有似然比检验、Score 检验)。一般地, 若参数 θ 的估计为 $\hat{\theta}$, 其标准差为 $\text{se}(\hat{\theta})$, 则 $H_0: \theta = \theta_0$ 的 Wald 检验统计量定义为

$$W = (\hat{\theta} - \theta_0)/\text{se}(\hat{\theta})$$

对于简单模型的 $H_0: b = b_0$, b_0 已知, Wald 检验统计量

$$W = \frac{\hat{b} - b_0}{\text{se}(\hat{b})} = \frac{\sqrt{s_{xx}}(\hat{b} - b_0)}{\hat{\sigma}}$$

2.2.5 正态模型下的统计推断

对于截距项一般没必要做统计推断, 只有斜率 b (x 带来的效应) 才是我们关心的, 下面只考虑正态假设下 b 的统计推断 (非正态的情形有大样本检验或置信区间)。

命题 2.2.5

假设模型 $y_i = a + bx_i + \varepsilon_i, \varepsilon_1, \dots, \varepsilon_n \text{ iid} \sim N(0, \sigma^2)$, 则

1. $\sqrt{s_{xx}}(\hat{b} - b)/\sigma \sim N(0, 1)$;
2. $(n-2)\hat{\sigma}^2/\sigma^2 \sim \chi_{n-2}^2$, 且 $\hat{\sigma}^2$ 与 $(\hat{a}, \hat{b})$ 独立;
3. $\sqrt{s_{xx}}(\hat{b} - b)/\hat{\sigma} \sim t_{n-2}$.

证明 待补充

原假设 $H_0: b = b_0$, b_0 已知, Wald 检验统计量定义为

$$t = (\hat{b} - b_0)/\text{se}(\hat{b}) = \sqrt{s_{xx}}(\hat{b} - b_0)/\hat{\sigma}$$

$t \sim t_{n-2}$, 当 $|t| \geq t_{n-2}(\alpha/2)$ 时拒绝原假设。

绝大多数情况下我们只关系 $H_0: b = 0$, Wald 检验统计量定义为

$$t = \hat{b}/\text{se}(\hat{b}) = \sqrt{s_{xx}}\hat{b}/\hat{\sigma} \sim t_{n-2}$$

$|t| \geq t_{n-2}(\alpha/2)$ 时拒绝原假设。

命题 2.2.6

$H_0: b = 0$ 的检验统计量 $t = \sqrt{s_{xx}}\hat{b}/\hat{\sigma} = \sqrt{n-2} \frac{r}{\sqrt{1-r^2}}$. 这说明 Wald 检验就是相关性检验。

证明

$$t = \frac{\sqrt{s_{xx}}\hat{b}}{\hat{\sigma}} = \frac{s_{xy}/\sqrt{s_{xx}}}{\sqrt{\frac{1}{n-2}(s_{yy} - s_{xy}^2/s_{xx})}} = \sqrt{n-2} \frac{r}{\sqrt{1-r^2}}$$

推论 2.2.1 (两样本 t-检验)

两样本 t-检验是回归系数显著性检验的特殊情形:

$$y_1, \dots, y_{n_1} \text{ iid} \sim N(\mu_1, \sigma^2) \leftarrow x_1, \dots, x_{n_1} = 1$$

$$y_{n_1+1}, \dots, y_{n_1+n_2} \text{ iid} \sim N(\mu_2, \sigma^2) \leftarrow x_{n_1+1}, \dots, x_{n_1+n_2} = 1$$

给两组样本分别赋予标号 $x_i = 0, 1$, 此时两样本问题写成线性模型:

$$y_i = a + bx_i + \varepsilon_i \quad (a = \mu_2, b = \mu_1 - \mu_2), \varepsilon_i \text{ iid} \sim N(0, \sigma^2)$$

$$H_0: b = 0 \Leftrightarrow \mu_1 = \mu_2$$

该模型 $H_0: b = 0$ 的检验统计量等于两样本 t-检验统计量:

$$t = \frac{\hat{b}}{\sqrt{\hat{\sigma}^2/s_{xx}}} = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{(n_1^{-1} + n_2^{-1})s^2}}$$

推论 2.2.2 (斜率的置信区间)

b 的 $(1-\alpha)100\%$ 置信区间可以取为:

$$\left[\hat{b} \mp \frac{\hat{\sigma}}{\sqrt{s_{xx}}} t_{n-2}(\alpha/2) \right]$$

推论 2.2.3 (均值函数的置信带)

均值函数/回归函数 $m(x_0) = \mathbb{E}(y|x = x_0) = a + bx_0, x_0 \in \mathbb{R}$, 对于给定的 x_0 , $m(x_0)$ 的 LS 估计 $\hat{m}(x_0) = \hat{a} + \hat{b}x_0$, 可以证明:

$$\hat{m}(x_0) \sim N(m(x_0), \sigma^2 \left(\frac{1}{\sigma^2} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right))$$

与 $\hat{\sigma}^2$ 独立，类似可得

$$\frac{\hat{m}(x_0) - m(x_0)}{\hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}}} \sim t_{n-2}$$

基于该分布，我们得到 $m(x_0)$ 的置信水平为 $1 - \alpha$ 的置信区间：

$$\left[\hat{m}(x_0) \pm t_{n-2}(\alpha) \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}} \right]$$



2.2.6 应用：幂次律

定义 2.2.6

如果变量 x, y 满足

$$y = cx^k$$

则称它们满足幂次律。幂次律的发现一般通过 $\log$ 尺度上的线性回归模型得到：

$$\log(y) = \log(c) + k \cdot \log(x)$$

无标度：改变 x 的单位，幂次律依然成立。



2.2.6.1 Benford 定律（首位数字）

Newcomb(1881), Benford(1938) 发现“自然的”正数的首位非 0 数字并不是均匀分布，首位数字是 1, 2, ..., 9 的概率依次下降，1 的概率最大，即所谓 Benford 定律。“自然的数字”包括人口数、河流面积、财务报表、新闻中出现的数字等等。

Benford 定律是一个严格的概率分布，并不是数据分析得到的经验，而是在一定条件下经过严格的数学论证得到的分布：

定理 2.2.1 (Benford 定律)

在一定的假设下，自然数字的首位数字为 d 的概率为

$$p(d) = \log_{10}(1 + d^{-1}), d = 1, 2, \dots, 9$$

假设有多种，其中一个为：随机变量 $x > 0$, $\log(x) \sim N(0, 1)$.



Benford 定律可以近似为幂次律：

$$p(d) = \log_{10}(1 + d^{-1}) \approx 0.31d^{-0.86}$$

例 2.2.1.

下表数据是 2020 年美国 3144 个县的人口数的首位数字 (d) 和频率 (p).

首位数字	1	2	3	4	5	6	7	8	9
频数	957	581	381	298	237	204	172	161	153
频率	0.304	0.185	0.121	0.095	0.075	0.065	0.055	0.051	0.049

拟合简单线性模型： $\log(p) = a + b \log(d) + e$ ，LS 估计得到

$$\hat{a} = -1.153, \hat{b} = -0.874$$

拟合得到的回归直线：

$$\log(p) = -1.153 - 0.874 \times \log(d)$$

即

$$p = 0.316 \cdot d^{-0.874}$$

得到下表：

首位数字	1	2	3	4	5	6	7	8	9
频率	0.304	0.185	0.121	0.095	0.075	0.065	0.055	0.051	0.049
幂次律拟合值	0.316	0.172	0.121	0.094	0.077	0.066	0.058	0.051	0.046
Benford 理论值	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

2.2.6.2 齐夫定律（体量与排名）

齐夫定律 (Zipf's law) 是由哈佛大学语言学家 G.K.Zipf 于 1949 年发现的单词使用频率与排名关系的经验幂次定律：在自然语言的语料库里，一个单词出现的频率 (p_k) 与它在频率表里的排名 (k) 的幂次成反比：

$$p_k \propto 1/k^\alpha, \text{ specially } \alpha = 1$$

不同类型、不同作者的词频分布可能不同，其中指数 α 刻画了文本风格，不一定等于 1。

一般的齐夫定律描述体量 (size, 可以是概率、体量等) 与排名的幂次关系，刻画的是这样一种现象：排名越高，差距越大，比如第一名是第二名的 2 倍，第二名是第三名的 1.25 倍，排名靠后的差异很小。例如 size 可以是财富榜的财富、金牌榜的金牌数目、河流面积等等。

例 2.2.2.

例如美国 Hamilton, Madison 所作的联邦文献的前十个高频词分布如下，显然两人写作风格有差异。

Words	the	of	to	and	in	a	be	that	it	is
Rank(k)	1	2	3	4	5	6	7	8	9	10
Freq(q)	91.27	64.65	40.71	24.5	24.37	22.85	20.06	14.98	13.82	11.7

表 2.1: Hamilton

Words	the	of	to	and	in	a	be	that	it	is
Rank(k)	1	2	3	4	5	6	7	8	9	10
Freq(q)	93.65	57.8	35.25	27.55	23.05	20.22	16.45	14.37	13.34	12.76

表 2.2: Madison

两人数据分别拟合线性模型：

$$\log(q) = a - \alpha \times \log(k)$$

得到二人的 α 分别等于 0.900, 0.902.

这也差别不大啊？

例 2.2.3.

下表列出了中文报刊中前十个高频汉字（前 42 个高频词占 25%）：

单字	的	一	是	不	了	在	有	人	这	大
排名 (k)	1	2	3	4	5	6	7	8	9	10
频率%(q)	4.87	1.41	1.32	1.07	0.95	0.93	0.91	0.78	0.76	0.58

拟合线性模型得到 $\log(q) = 1.23 - 0.75\log(k)$, $\alpha = 0.75$ 与偏离较大, 可能是因为统计的是单字频率而非词汇频率。

2.2.6.3 Kleiber 定律与异速生长

Square-cube law（同速生长）：生物器官或物体的体积 V （或重量）正比于长度的 3 次方，而表面积 S 正比于长度的 2 次方，所以面积正比于体积的 $2/3$ 次方：

$$S \propto V^{2/3}$$

生物生长过程中, 如果器官随着身体的增长而等比例地线性增长, 满足 square-cube law, 称为同速生长 (isometric scaling). 但多数器官的生长不呈线性关系, 比如眼睛生长较慢, 而腿部生长较快, 这称为异速生长 (allometry).

Kleiber's law 指出, 动物代谢速率 (R) 与体重 (M) 存在如下幂次律关系

$$R = 70 \times M^{3/4}$$

其中代谢速率单位为卡路里/秒, 与表面积有关。

2.2.6.4 BMI 指数

指数或指标 (index) 是反映复杂系统整体表现的度量, 比如物价指数、消费指数、普尔指数。指数需要有普适性。为了衡量人的体重是否超标, 单纯用体重 W 作为指标不具有普适性, 因为体重与身高 H 等因素有关 (你可以对每个身高段定义体重标准, 但这不够简洁)。BMI 被认为是适用于所有身高的成年人的一个体重 (特别是脂肪) 指数。

我们在对数尺度上建立线性模型：

$$\log(W) = a + b\log(H) + \varepsilon, \varepsilon \sim N(0, \sigma^2)$$

等价地, $\log(W) \sim N(a + b\log(H), \sigma^2)$, 标准化：

$$z = \frac{\log(W) - (a + b\log(H))}{\sigma} \sim N(0, 1)$$

那么 z 的分布与 H 无关, 具有普适性。例如一个人的体重指标 $z > 1.645 = \alpha_{0.95}$, 这意味着他的体重指标超过了 95% 的人, 可以认为是体重超标。所以 W/H^b 的分布与身高无关, 可以作为体重指数。经验数据表明 $b \approx 2$ 。

显然, $BMI = W/H^2$ 在某种意义上校正了身高因素 (消除了身高的影响), 人们普遍认为该指标适用于所有身高的人的体重量, 应用广泛。

2.2.6.5 无标度社交网络

无标度社交网络 (scale-free social network) 中大多数成员有较少的连结 (hub), 而少数点有较多的连结。每个节点的连结个数称为度数 (degree), 节点度数 k 服从 Pareto 分布/幂次律：

$$P(k) \propto k^{-r}$$

2.2.7 回归模型结果的解释：因果还是关联？

随机化控制试验或天然试验： x 是外生的，即 x 与 ε 独立，则 LS 估计 $\hat{b}$ 是无偏的，结果可表述为因果关系：对同一个研究对象， x 每增加一个单位， y 的期望增加 b 个单位。

观察研究：自变量一般是内生的，LS 估计 $\hat{b}$ 有偏，结果只能表述为关联关系：如果一个研究对象的 x 比另外一个研究对象大 1 个单位，则相应的 y 的期望大 b 个单位。

例 2.2.4.

分析 2001 年人口抽样调查数据，得到妻子教育水平（上学的年数）与丈夫教育水平的回归方程如下：

$$\text{WifeEdLevel} = 5.60 + 0.57\text{HusbandEdLevel} + \text{residual}$$

如果公司送王先生到大学在职培养一年，你是否预期王太太的教育水平会上升 0.57 年？若不是，0.57 的含义是什么？

这是观察研究而非试验，没有证据表明误差与自变量独立，结果是关联而不是因果。 $b = 0.57$ 的含义是：如果该研究中某人比另外一个人多上一年学，那么他的妻子比另外一人的妻子期望多上 0.57 年学。

2.3 多重线性回归模型

2.3.1 定义

定义 2.3.1 (多重线性回归模型)

总体：

$$y = \beta_0 + x_1\beta_1 + \cdots + x_{p-1}\beta_{p-1} + \varepsilon, \varepsilon \sim (0, \sigma^2), \varepsilon \perp \mathbf{x}$$

记 $\mathbf{x} = (x_1, \dots, x_{p-1})^T$, $\mathbf{b} = (\beta_1, \dots, \beta_{p-1})^T$, 则

$$y = \beta_0 + \mathbf{x}^T \mathbf{b} + \varepsilon$$

样本 $(y_i, x_{i1}, \dots, x_{i,p-1}), i = 1, 2, \dots, n$ 来自上述总体，那么误差

$$\varepsilon_i = y_i - \beta_0 - \mathbf{x}_i^T \mathbf{b}$$

独立同分布 $\sim (0, \sigma^2)$ ，且与 $\mathbf{x}_i$ 独立， $i = 1, 2, \dots, n$ 。

记所有响应变量 $\mathbf{y} = (y_1, \dots, y_n)^T$ ，所有误差 $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$ ，所有回归系数 $\boldsymbol{\beta} = (\beta_0, \dots, \beta_{p-1})^T$ ，所有自变量组成的矩阵

$$X_{n \times p} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{pmatrix}$$

称为设计阵，则模型的矩阵-向量表达为：

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

注 模型也可以写作：

$$\mathbf{y} = \mathbf{u} + \boldsymbol{\varepsilon}, \mathbf{u} \in C(X)$$

2.3.2 最小二乘法

为了估计参数 β , 最小化误差平方和:

$$\min_{\beta \in \mathbb{R}^p} \sum \varepsilon_i^2 = \min_{\beta \in \mathbb{R}^p} \sum (y_i - \beta_0 - \mathbf{x}_i^T \mathbf{b})^2 = \min_{\beta \in \mathbb{R}^p} \|\mathbf{y} - X\beta\|^2$$

定理 2.3.1

假设线性回归模型

$$\mathbf{y}_{n \times 1} = X_{n \times p} \beta_{p \times 1} + \varepsilon_{n \times 1}, \varepsilon \sim (0, \sigma^2 I_n), \varepsilon \perp X$$

满足 GM 假设, 则其 LS 估计 $\hat{\beta}$ 应当满足估计方程 (也叫正则方程):

$$X^T \varepsilon = X^T (\mathbf{y} - X\beta) = 0$$

则得到

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

如果 X 列满秩, 则 LS 估计唯一, 且无偏。

证明

推论 2.3.1

回顾命题 1.5.11,

$$P_X(\mathbf{y}) = X(X^T X)^{-1} X^T \mathbf{y} = X\hat{\beta}$$

矩方法的观点: 略。

解方程的观点: 略。

2.3.3 误差方差的估计

定义 2.3.2

投影 $\hat{\mathbf{y}} = P_X \mathbf{y} = X\hat{\beta} = X(X^T X)^{-1} X^T \mathbf{y}$ 称为拟合值向量, $\mathbf{e} = \mathbf{y}^\perp = \mathbf{y} - \hat{\mathbf{y}}$ 称为残差向量。

那么估计方程 $X^T \varepsilon = 0$ 就是 $X^T \mathbf{e} = 0$, 即 $\mathbf{e} \perp C(X)$.

残差平方和 $\text{RSS} = \|\mathbf{e}\|^2 = \mathbf{y}^T (I_n - P_X) \mathbf{y}$, σ^2 的 LS 估计定义为

$$\hat{\sigma}^2 = \frac{\text{RSS}}{n-p} = \frac{1}{n-p} \|\mathbf{e}\|^2$$

2.3.4 LS 估计的统计性质

2.3.4.1 无偏性

引理 2.3.1

若 $\mathbf{x} \sim (\mu, \Sigma)$, 则 $\mathbb{E}(\mathbf{x}^T A \mathbf{x}) = \mu^T A \mu + \text{tr}(A \Sigma)$.

定理 2.3.2

线性模型

$$\mathbf{y} = X_{n \times p} \beta + \varepsilon, \varepsilon \sim (0, \sigma^2 I_n), \varepsilon \perp X$$

其中 $n \geq p$, X 列满秩, 则

1. LS 估计的无偏性: $\mathbb{E}\hat{\beta} = \beta$.

2. LS 估计的方差: $\text{var}(\hat{\beta}|X) = \Sigma^2(X^T X)^{-1}$.

3. 误差方差估计的无偏性: $\mathbb{E}\hat{\sigma}^2 = \sigma^2$.



$\text{var}(\hat{\beta}|X) = \Sigma^2(X^T X)^{-1}$ 中代入 σ^2 的估计, 即得到 $\hat{\beta}$ 方差的估计:

$$\hat{\text{var}}(\hat{\beta}|X) = \frac{1}{n-p} \|e\|^2 (X^T X)^{-1}$$

2.3.4.2 最优性

定理 2.3.3 (Gauss-Markov)

线性模型

$$\mathbf{y} = X_{n \times p} \boldsymbol{\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim (0, \sigma^2 I_n), \boldsymbol{\varepsilon} \perp X$$

其中 $n \geq p$, X 列满秩, 则 $\boldsymbol{\beta}$ 的 LS 估计 $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}$ 是最优无偏线性估计 (BLUE, best linear unbiased estimate), 即对于任何 $\boldsymbol{\beta}$ 的线性无偏估计 $\tilde{\boldsymbol{\beta}} = C\mathbf{y}$, 有

$$\text{var}(\tilde{\boldsymbol{\beta}}) \geq \text{var}(\hat{\boldsymbol{\beta}})$$

其中 C 是仅与 X 有关的 $p \times n$ 的常数矩阵。



2.3.4.3 测不准原理

对于简单线性模型 $y_i = a + bx_i + \varepsilon_i$ 的任意无偏估计 $\tilde{b}$,

$$\text{var}(\tilde{b}|x) \geq \text{var}(\hat{b}|x) = \sigma^2 / s_{xx}$$

记 $s_x^2 = s_{xx}/(n-1)$, 上式等价于:

$$\text{var}(\tilde{b}|x) \cdot s_x^2 \geq \frac{\sigma^2}{n-1}$$

这说明左端两个方差不可能同时很小, 即 x 和 $\tilde{b}$ 不可能同时精准测量。

2.3.5 工具变量

工具变量法 (Instrumental Variable method) 试图在线性模型中误差与自变量不独立的情况下, 求解回归系数的无偏估计, 发现因果。

定义 2.3.3

如果线性模型 $y = a + bx + \varepsilon$ 中 y 不是研究对象本身固有而是外界随机赋予的, 称 Y 是外生的 (exogenous), 此时 x 与 ε 独立。外生性是推断 x 和 y 之间因果关系的关键条件。



2.4 中心化

先明确一下各种记号, 对于线性模型:

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

其中,

$$\begin{aligned} \mathbf{y} &= \mathbf{y}_{n \times 1} = (y_1, y_2, \dots, y_n)^T \\ X &= X_{n \times p} = \begin{pmatrix} \mathbf{1}_{n \times 1} & \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}_{n \times (p-1)} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{pmatrix} \\ \boldsymbol{\beta} &= \boldsymbol{\beta}_{p \times 1} = (\beta_0, \beta_1, \dots, \beta_{p-1}) \\ \boldsymbol{\varepsilon} &= \boldsymbol{\varepsilon}_{n \times 1} = (\varepsilon_1, \dots, \varepsilon_n) \end{aligned}$$

注意每个 $\mathbf{x}_i = (x_{i1}, \dots, x_{i,p-1})^T$ 是 $p-1$ 维列向量。同时:

$$\begin{aligned} \mathbf{b} &= \mathbf{b}_{(p-1) \times 1} = (\beta_1, \dots, \beta_{p-1})^T \\ Z &= Z_{n \times (p-1)} = \begin{pmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_n \end{pmatrix}^T = \begin{pmatrix} x_{11} & \cdots & x_{1,p-1} \\ x_{21} & \cdots & x_{2,p-1} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{n,p-1} \end{pmatrix} \end{aligned}$$

在没有特殊说明时, 用 $\mathbf{1}$ 代表 $\mathbf{1}_{n \times 1}$, 于是

$$X = \begin{pmatrix} \mathbf{1} & Z \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \mathbf{b} \end{pmatrix}$$

最后, 记

$$\bar{\mathbf{x}} = \frac{1}{n}(\mathbf{x}_1 + \cdots + \mathbf{x}_n) = \begin{pmatrix} \bar{x}_1 & \cdots & \bar{x}_{p-1} \end{pmatrix} = \begin{pmatrix} \frac{1}{n} \sum_{k=1}^n x_{k1} & \cdots & \frac{1}{n} \sum_{k=1}^n x_{k,p-1} \end{pmatrix} = \frac{1}{n} Z^T \mathbf{1}$$

以及

$$\bar{y} = \frac{1}{n}(y_1 + \cdots + y_n)$$

2.4.1 基本概念

定义 2.4.1 (中心化)

Z 的中心化矩阵:

$$Z_c = Z - P_1 Z = Z - \frac{1}{n} \mathbf{1} \mathbf{1}^T Z = Z - \mathbf{1} \bar{\mathbf{x}}^T$$

$\mathbf{y}$ 的中心化:

$$\mathbf{y}_c = \mathbf{y} - P_1 \mathbf{y} = \mathbf{y} - \mathbf{1} \bar{y}$$

如果令 $\tilde{X} = \begin{pmatrix} \mathbf{1} & Z_c \end{pmatrix}$, 那么 $C(X) = C(\tilde{X})$, 且 $\mathbf{1} \perp Z_c$, 由命题 1.5.12 的 8 可得

$$P_X = P_{\tilde{X}} = P_1 + P_{Z_c} = \frac{1}{n} \mathbf{1} \mathbf{1}^T + Z_c (Z_c^T Z_c)^- Z_c^T$$

定义 2.4.2 (标准化)

样本协方差矩阵:

$$S = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T = \frac{1}{n-1} Z_c^T Z_c = \frac{1}{n-1} Z^T (I_n - P_1) Z$$

Z 的标准化矩阵: 设 $D = \text{diag}(S)$,

$$Z_s = D^{-1/2} Z_c$$

相关系数矩阵:

$$R = \frac{1}{n-1} Z_s^T Z_s$$

$\mathbf{y}$ 的标准化:

$$\mathbf{y}_s = (y_1 - \bar{y}, \dots, y_n - \bar{y})^T / s = \mathbf{y}_c / s$$

这里 s 是 $\mathbf{y}$ 的样本方差。

线性模型的中心化: Z 中心化, 模型改写为

$$\begin{aligned} \mathbf{y} &= \mathbf{1}(\beta_0 + \bar{\mathbf{x}}^T \mathbf{b}) + Z_c \mathbf{b} + \varepsilon \\ &\stackrel{\text{def}}{=} \mathbf{1}\beta_0^* + Z_c \mathbf{b} + \varepsilon \end{aligned}$$

$\mathbf{y}$ 和 Z 都中心化, 模型改写为

$$\begin{aligned} \mathbf{y}_c &= \mathbf{1}(\beta_0 + \bar{\mathbf{x}}^T \mathbf{b} - \bar{y}) + Z_c \mathbf{b} + \varepsilon \\ &\stackrel{\text{def}}{=} \mathbf{1}\beta_0^{**} + Z_c \mathbf{b} + \varepsilon \end{aligned}$$

可以发现, 中心化不影响回归系数 $\mathbf{b}$, 只改变截距项。

定义 2.4.3 (拟合优度)

总平方和:

$$SS_{\text{总}} = s_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \|\mathbf{y} - \mathbf{1}\bar{y}\|^2$$

回归平方和:

$$SS_{\text{回}} = s_{\hat{y}\hat{y}} = \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2 = \|\hat{\mathbf{y}} - \mathbf{1}\bar{\hat{y}}\|^2 = \|\hat{\mathbf{y}} - \mathbf{1}\bar{y}\|^2$$

这里用到了 $\bar{\hat{y}} = \bar{y}$, 这是 $\sum e_i = 0$ 的直接推论。

决定系数:

$$R^2 = \frac{SS_{\text{回}}}{SS_{\text{总}}} = \frac{\|\hat{\mathbf{y}} - \mathbf{1}\bar{y}\|^2}{\|\mathbf{y} - \mathbf{1}\bar{y}\|^2} = \frac{\text{var}(\hat{\mathbf{y}})}{\text{var}(\mathbf{y})}$$

决定系数度量了模型拟合的程度 (拟合优度) / 自变量对响应的解释能力。

根据 $\mathbf{e} \perp \hat{\mathbf{y}}$, $\mathbf{e} \perp \mathbf{1}$, 所以

$$\mathbf{e} \perp (\hat{\mathbf{y}} - \mathbf{1}\bar{y}) \Rightarrow \mathbf{y} - \mathbf{1}\bar{y} = (\hat{\mathbf{y}} - \mathbf{1}\bar{y}) \oplus \mathbf{e}$$

因此有平方分解:

$$\|\mathbf{y} - \mathbf{1}\bar{y}\|^2 = \|\hat{\mathbf{y}} - \mathbf{1}\bar{y}\|^2 + \|\mathbf{e}\|^2$$

$$SS_{\text{总}} = SS_{\text{回}} + RSS$$

所以 $SS_{\text{回}} \leq SS_{\text{总}}$, $0 \leq R^2 \leq 1$.

命题 2.4.1

$$R^2 = (r_{y\hat{y}})^2 = S_{yx} S_{xx}^{-1} S_{xy} / S_{yy}$$

其中

$$S_{xx} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

$$S_{xy} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(y_i - \bar{y}), S_{yx} = S_{xy}^T$$

证明

命题 2.4.2

$$R^2 = \max_{\mathbf{u} \in L(X)} (r_{\mathbf{u}, y})^2$$

其中 $r_{\mathbf{u}, y}$ 为样本相关系数, 且最大值在 $\mathbf{u} = \hat{\mathbf{y}} = X\hat{\beta}$ 处达到。

证明

命题 2.4.3 (回归系数的估计)

假设 X 列满秩, 则

1. LS 估计:

$$\hat{\mathbf{b}} = (Z_c Z_c^T)^{-1} Z_c^T \mathbf{y} = S_{xx}^{-1} S_{xy}, \hat{\beta}_0 = \bar{y} - \bar{\mathbf{x}}^T \hat{\mathbf{b}}$$

2. LS 估计的方差:

$$\text{var}(\hat{\mathbf{b}}|X) = \sigma^2 (Z_c Z_c^T)^{-1} = \frac{1}{n-1} \sigma^2 S_{xx}^{-1}$$

3. 方差的估计:

$$\hat{\sigma}^2 = \frac{n-1}{n-p} S_{yy \cdot \mathbf{x}} \approx S_{yy \cdot \mathbf{x}} = S_{yy} - S_{yx} S_{xx}^{-1} S_{xy} = S_{yy} (1 - R^2)$$

可以看出, 自变量的回归系数 $\mathbf{b}$ 的 LS 估计仅依赖于样本协方差矩阵, 截距项 LS 估计只与样本均值有关。如果 $\mathbf{y}, Z$ 都已经中心标准化, 那么 $\hat{\mathbf{b}} = R_{xx}^{-1} R_{xy}$, 其中 R 是样本相关系数矩阵。

2.4.2 部分回归系数

将设计阵 X 拆分为 $X = (1, X_1, X_2)$, 模型为

$$\mathbf{y} = X\beta + \varepsilon = \mathbf{1}\beta_0 + X_1\beta_1 + X_2\beta_2 + \varepsilon$$

其中 X_1 的列认为是干扰因素, X_2 的列是感兴趣的变量, LS 估计 $\hat{\beta}_2$ 是否有效地消除了 X_1 的影响? 约定记号:

$$X_2^\perp = X_2 - P_{1, X_1} X_2, \hat{\beta}_2 = (X_2^{\perp T} X_2^\perp)^{-1} X_2^{\perp T} \mathbf{y}$$

所有数据 $(\mathbf{y}, X_1, X_2)$ 的样本方差-协方差矩阵为:

$$S = \begin{matrix} & \begin{matrix} \mathbf{x}_1 & \mathbf{x}_2 & y \end{matrix} \\ \begin{matrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ y \end{matrix} & \begin{pmatrix} S_{11} & S_{12} & S_{1y} \\ S_{21} & S_{22} & S_{2y} \\ S_{y1} & S_{y2} & S_{yy} \end{pmatrix} \end{matrix}$$

以及

$$\begin{pmatrix} S_{22 \cdot 1} & S_{2y \cdot 1} \\ S_{y2 \cdot 1} & S_{yy \cdot 1} \end{pmatrix} = \begin{pmatrix} S_{22} & S_{2y} \\ S_{y2} & S_{yy} \end{pmatrix} - \begin{pmatrix} S_{21} \\ S_{y1} \end{pmatrix} S_{11}^{-1} (S_{12} \quad S_{1y})$$

命题 2.4.4

线性模型:

$$\mathbf{y}_{n \times 1} = X_{n \times p} \beta_{p \times 1} + \varepsilon_{n \times 1} = \mathbf{1}\beta_0 + X_1\beta_1 + X_2\beta_2 + \varepsilon, \varepsilon \sim (0, \sigma^2 I_n), \varepsilon \perp X$$

其中 $X = (\mathbf{1}, X_1, X_2)$ 列满秩, $\beta = (\beta_0, \beta_1^T, \beta_2^T)^T$, 令 $X_1^\perp = X_2 - P_{\mathbf{1}, X_1} X_2$, 则

$$\begin{aligned}\hat{\beta}_2 &= (X_2^{\perp T} X_2^\perp)^{-1} X_2^{\perp T} \mathbf{y} = S_{22 \cdot 1}^{-1} S_{2y \cdot 1} \\ \text{var}(\hat{\beta}_2 | X) &= \sigma^2 (X_2^{\perp T} X_2^\perp)^{-1} = \frac{1}{n-1} \sigma^2 S_{22 \cdot 1}^{-1}\end{aligned}$$

例 2.4.1. 命题 2.4.4 的 $p=3$ 特殊情形

假设数据 $(y_i, x_i, z_i), i=1, 2, \dots, n$ 满足模型:

$$y_i = a + bx_i + cz_i + \varepsilon_i, \varepsilon_i \text{ iid } \sim (0, \sigma^2), \varepsilon_i \perp x_i, z_i$$

那么

$$\begin{aligned}\hat{b} &= \frac{S_{xy \cdot z}}{S_{xx \cdot z}} = \frac{S_{xy} - S_{xz} S_{yz} / S_{zz}}{S_{xx} - S_{xz}^2 / S_{zz}} = \sqrt{\frac{S_{yy}}{S_{xx}}} \left(\frac{r_{yx} - r_{yz} r_{xz}}{1 - r_{xz}^2} \right) \propto r_{yx \cdot z} \\ \text{var}(\hat{b} | x, z) &= \frac{\sigma^2}{(n-1)(S_{xx} - S_{xz}^2 / S_{zz})} = \frac{\sigma^2}{(n-1)S_{xx}} \times \frac{1}{1 - r_{xz}^2}\end{aligned}$$

注意 $r_{xz} = 0$ 时, 方差 $\text{var}(\hat{b} | x, z) = \frac{\sigma^2}{(n-1)S_{xx}}$, 记

$$VIF = \frac{1}{1 - r_{xz}^2}$$

称为方差膨胀因子 (variance inflation factor).

应用: 路径模型。

2.5 回归模型的 F 检验

2.5.1 F 分布

定义 2.5.1 (F 分布)

随机变量 $U \sim \chi_{d_1}^2, V \sim \chi_{d_2}^2$, 且相互独立, 则

$$F = \frac{U/d_1}{V/d_2} \sim F_{d_1, d_2}$$

在大多数检验中, d_1 与参数的个数相关, d_2 则是样本量减去一个数, 所以 d_2 往往 $\rightarrow \infty$, 此时 $V/d_2 \rightarrow 1$. 当 d_1 也较大时,

$$F \approx \frac{1}{\sqrt{d_1}} \cdot \frac{U}{\sqrt{d_1}} = \frac{1}{\sqrt{d_1}} \cdot \frac{\frac{d_1}{2} y_1^2}{\sqrt{d_1}} \rightarrow N(1, \frac{2}{d_1})$$

这是一个常用的近似。

引理 2.5.1

假设 $\mathbf{x} \sim N(0, I_n)$, 假设 B 是投影矩阵, 则

1. $\mathbf{x}^T B \mathbf{x} = \|B \mathbf{x}\|^2 \sim \chi_r^2$, $r = \text{rank}(B)$.
2. $AB = 0 \Rightarrow A \mathbf{x} \perp B \mathbf{x}$, 特别地 $A \mathbf{x} \perp \mathbf{x}^T B \mathbf{x}$.

命题 2.5.1

假设 $P = P_V$ 是子空间 V 的 $n \times n$ 投影矩阵, $r = \text{rank}(P) = \dim(V)$, 若 $\mathbf{x} \sim N(0, I_n)$, 则

$$\frac{\|P\mathbf{x}\|^2/r}{\|(I_n - P)\mathbf{x}\|^2/(n-r)} \sim F_{r, n-r}$$

2.5.2 一般线性假设的 F 检验

一般的 Wald 检验是这样的: θ 是参数向量, $\hat{\theta}$ 是其估计, 则原假设 $H_0: \varphi(\theta) = \mathbf{c}_0$ 成立时, Wald 检验统计量

$$W \stackrel{\text{def}}{=} (\varphi(\hat{\theta}) - \mathbf{c}_0)^T \hat{\Sigma}^{-1} (\varphi(\hat{\theta}) - \mathbf{c}_0) \xrightarrow{d} \chi_q^2$$

其中 $\hat{\Sigma}$ 是对

$$\Sigma = \text{var}(\varphi(\hat{\theta}))$$

的估计。

现假设正态线性回归模型: $\mathbf{y}_{n \times 1} = X_{n \times p} \boldsymbol{\beta}_{p \times 1} + \boldsymbol{\varepsilon}_{n \times 1}, \boldsymbol{\varepsilon} \sim N_n(0, \sigma^2 I_n)$, 原假设 $H_0: A\boldsymbol{\beta} = \mathbf{c}_0$ 称为一般线性假设, 其中 $A_{q \times p}, \mathbf{c}_0$ 已知, A 行满秩。那相应的 Wald 检验如何?

定理 2.5.1

正态线性回归模型: $\mathbf{y}_{n \times 1} = X_{n \times p} \boldsymbol{\beta}_{p \times 1} + \boldsymbol{\varepsilon}_{n \times 1}, \boldsymbol{\varepsilon} \sim N_n(0, \sigma^2 I_n)$, 其中 X 列满秩, 则

1. $\hat{\boldsymbol{\beta}} | X \sim N(\boldsymbol{\beta}, \sigma^2 (X^T X)^{-1})$.
2. $(n-p)\hat{\sigma}^2/\sigma^2 \sim \chi_{n-p}^2$, 且 $\hat{\boldsymbol{\beta}}, \hat{\sigma}^2$ 独立。

证明

定理 2.5.2

正态线性回归模型: $\mathbf{y}_{n \times 1} = X_{n \times p} \boldsymbol{\beta}_{p \times 1} + \boldsymbol{\varepsilon}_{n \times 1}, \boldsymbol{\varepsilon} \sim N_n(0, \sigma^2 I_n)$, 考虑零假设 $H_0: A\boldsymbol{\beta} = \mathbf{c}_0$, 则

$$W \stackrel{\text{def}}{=} (A\hat{\boldsymbol{\beta}} - \mathbf{c}_0)^T (\hat{\sigma}^2 A(X^T X)^{-1} A^T)^{-1} (A\hat{\boldsymbol{\beta}} - \mathbf{c}_0) \sim qF_{q, n-p}$$

或者 $F \stackrel{\text{def}}{=} W/q \sim F_{q, n-p}$. 这就是回归模型一般线性假设下的 F 检验。

证明

注 有关 $\boldsymbol{\beta} = (\beta_0, \dots, \beta_{p-1})^T$ 的检验问题一般都可以写成上述线性假设的形式, 例如 $p = 3$ 时,

$$\beta_1 = \beta_2 = 0 \Leftrightarrow H_0: A\boldsymbol{\beta} = \mathbf{0}, A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

2.5.3 约束最小二乘

考虑使得零假设 $H_0: A\boldsymbol{\beta} = \mathbf{c}_0$ 成立的 $\boldsymbol{\beta}$ 的最小二乘估计, 即约束最小二乘问题:

$$\tilde{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p, A\boldsymbol{\beta} = \mathbf{0}}{\text{argmin}} \|\mathbf{y} - X\boldsymbol{\beta}\|^2$$

我们不妨假设 $\mathbf{c}_0 = \mathbf{0}$, 这是因为可以取

$$\boldsymbol{\beta}^* = \boldsymbol{\beta} - A^T(AA^T)^{-1}\mathbf{c}_0$$

此时模型转化为

$$\mathbf{y}^* \stackrel{\text{def}}{=} \mathbf{y} - XA^T(AA^T)^{-1}\mathbf{c}_0 = X\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$$

原假设 $H_0: A\boldsymbol{\beta} = \mathbf{c}_0 \Leftrightarrow A\boldsymbol{\beta}^* = \mathbf{0}$.

命题 2.5.2

约束最小二乘问题的最优解为

$$\tilde{\beta} = \hat{\beta} - (X^T X)^{-1} A^T (A (X^T X)^{-1} A^T)^{-1} A \hat{\beta}$$

证明

由上述命题可得拟合值 $\hat{y}_0 = X \tilde{\beta} = \hat{y} - (X^T X)^{-1} A^T (A (X^T X)^{-1} A^T)^{-1} A \hat{\beta}$, 我们还可以从投影的角度得到这个结果。

命题 2.5.3

约束 $A_{q \times p} \beta = 0$ 时, X 张成的空间 (称为自变量空间) 为

$$V_0 = \{X\beta : A\beta = 0, \beta \in \mathbb{R}^p\} = C(X(X^T X)^{-1} A^T)^{\perp} \cap C(X)$$

且该空间的投影阵为

$$P_{V_0} = P_X - P_{X(X^T X)^{-1} A^T}$$

证明

y 在 V_0 上的投影就是使得约束最小二乘达到最小的拟合值, 即

$$\hat{y}_0 = P_{V_0} y = (P_V - P_{X(X^T X)^{-1} A^T}) y = X \hat{\beta} - (X^T X)^{-1} A^T (A (X^T X)^{-1} A^T)^{-1} A \hat{\beta}$$

实际应用时, 可能不需要计算上述公式, 而可以在具体的线性假设检验问题中将约束代入模型进行简化, 见下例。

例 2.5.1.

假设模型

$$y = \mathbf{1}\beta_0 + x_1\beta_1 + \cdots + x_{p-1}\beta_{p-1} + \varepsilon$$

约束 $\beta_1 = \cdots = \beta_{p-1}$, 带入模型得

$$y = \mathbf{1}\beta_0 + (x_1 + \cdots + x_{p-1})\beta_1 + \varepsilon$$

此时设计阵就是 $X_0 = (\mathbf{1} \quad x_1 + \cdots + x_{p-1})$, 这是一个更简单的线性模型, 直接计算其 LS 估计即可。

2.5.4 全模型与零模型的差别——F 检验的另一种观点

我们将约束条件下的回归模型称为零模型, 原来的回归模型称为全模型。由命题 2.5.2 可知

$$\hat{y} - \hat{y}_0 = (X^T X)^{-1} A^T (A (X^T X)^{-1} A^T)^{-1} A \hat{\beta}$$

于是

$$\|\hat{y} - \hat{y}_0\|^2 = (\hat{y} - \hat{y}_0)^T (\hat{y} - \hat{y}_0) = \hat{\beta}^T A^T [A (X^T X)^{-1} A^T]^{-1} A \hat{\beta}$$

回顾定理 2.5.2, 得到

$$F = \frac{1}{\hat{\sigma}^2 q} \|\hat{y} - \hat{y}_0\|^2 = \frac{n-p}{q} \times \frac{\|\hat{y} - \hat{y}_0\|^2}{\|y - \hat{y}\|^2}$$

记零模型下的残差平方和与决定系数为

$$RSS_0 = \|y - \hat{y}_0\|^2, R_0^2 = \frac{\|\hat{y}_0 - \mathbf{1}\bar{y}\|^2}{\|y - \mathbf{1}\bar{y}\|^2}$$

因为

$$\|\hat{y} - \hat{y}_0\|^2 = \|y - \hat{y}_0\|^2 - \|y - \hat{y}\|^2 = RSS_0 - RSS$$

所以

$$F = \frac{n-p}{q} \times \frac{RSS_0 - RSS}{RSS} \sim F_{q, n-p}$$

可以发现, 检验统计量 F 的关键部分是两个模型的残差。对此结果我个人的直观理解是: 当原假设成立时, 我们上面的“基于约束条件下的最小二乘估计”是一个合理的估计¹, 所以零模型下的残差满足正态分布, 于是残差的平方和就是卡方分布, RSS 服从 $\chi^2_{n-\dim(V)}$, 自由度为 $n-p$, RSS_0 服从 $\chi^2_{n-\dim(V_0)}$, 自由度为 $n-p+q$, 所以就有

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p)} \sim F_{q, n-p}$$

这一观点非常直观地展现出了回归模型 F 检验的本质: 残差的对比。

命题 2.5.4

若 $A = (\mathbf{0}, *)$, 即 A 的第一列为 0, 则

$$F = \frac{n-p}{q} \times \frac{R^2 - R_0^2}{1 - R^2} = \frac{n-p}{q} \times \frac{R_p^2}{1 - R_p^2}$$

其中

$$R_p^2 = \frac{R^2 - R_0^2}{1 - R_0^2} = \frac{\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_0\|^2}{\|\mathbf{y} - \hat{\mathbf{y}}_0\|^2}$$

证明

2.5.5 部分回归系数的 F 检验

对于线性回归模型 $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, 记 $V = C(X)$, 将设计阵 X 拆分为 $X = (\mathbf{1} \quad X_1 \quad X_2)$, $\mathbf{y} = \mathbf{1}\beta_0 + X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$, 其中 X_2 为 $n \times q$ 阶矩阵, $\boldsymbol{\beta}_2$ 是我们关心的部分: 原假设 $H_0: \boldsymbol{\beta}_2 = \mathbf{0}_{q \times 1}$. 那么 H_0 成立时的子模型为 $\mathbf{y} = \mathbf{1}\beta_0 + X_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$, $V_0 = C(\mathbf{1}, X_1)$.

命题 2.5.5

令

$$X_2^\perp = X_2 - P_{V_0}X_2 = X_2 - P_{\mathbf{1}, X_1}X_2$$

则 H_0 的 F 检验统计量为

$$F = \frac{\|X_2^\perp \hat{\boldsymbol{\beta}}_2\|^2}{q\hat{\sigma}^2} \sim F_{q, n-p}$$

其中

$$\hat{\boldsymbol{\beta}}_2 = (X_2^{\perp T} X_2^\perp)^{-1} X_2^{\perp T} \mathbf{y}$$

为 $\boldsymbol{\beta}_2$ 的 LS 估计。(命题 2.4.4)

证明 只需注意到

$$\hat{\mathbf{y}} = P_X \mathbf{y} = P_{\mathbf{1}, X_1, X_2} \mathbf{y} = P_{\mathbf{1}, X_1, X_2^\perp} \mathbf{y} = \hat{\mathbf{y}}_0 + P_{X_2^\perp} \mathbf{y} = \hat{\mathbf{y}}_0 + X_2^\perp \hat{\boldsymbol{\beta}}_2$$

本节的原假设 $\boldsymbol{\beta}_2 = \mathbf{0}$, 也是前面所介绍的一般线性假设, 当时我们定义了

$$R_p^2 = \frac{R^2 - R_0^2}{1 - R_0^2}$$

这代表 $\mathbf{y}$ 中 X_1 不能解释的部分中, X_2 能解释得比例。

¹没有所谓约束条件, 事实就是满足 $A\boldsymbol{\beta} = \mathbf{0}$ 的, 这是一个真正的回归模型, 我们得到的结果也将满足回归模型的性质。

命题 2.5.6

命题 2.5.5 中的线性模型在正态假设下的 F 检验统计量为

$$F = \frac{n-p}{q} \times \frac{R_p^2}{1-R_p^2} \sim F_{q, n-p}$$

其中

$$R_p^2 = \frac{S_{y2 \cdot 1} S_{22 \cdot 1}^{-1} S_{2y \cdot 1}}{S_{yy \cdot 1}} \stackrel{\text{def}}{=} R_{y x_2 \cdot x_1}^2$$

为控制 X_1 时 y 与 X_2 的偏决定系数。(也叫部分决定系数)

证明 由 $\hat{y} - \hat{y}_0 = X_2^\perp \hat{\beta}_2$,

$$\|\hat{y} - \hat{y}_0\|^2 = \mathbf{y}^T X_2^\perp (X_2^\perp{}^T X_2^\perp)^{-1} X_2^\perp{}^T \mathbf{y} = (n-1) S_{y2 \cdot 1} S_{22 \cdot 1}^{-1} S_{2y \cdot 1}$$

另外,

$$\|\mathbf{y} - \hat{y}_0\|^2 = \|\mathbf{y} - \mathbf{1}\bar{y}\|^2 - \|\hat{y}_0 - \mathbf{1}\bar{y}\|^2 = (n-1) S_{yy} - (n-1) S_{y1} S_{11}^{-1} S_{1y} = (n-1) S_{yy \cdot 1}$$

然后由命题 2.5.4 中的

$$R_p^2 = \frac{\|\hat{y} - \hat{y}_0\|^2}{\|\mathbf{y} - \hat{y}_0\|^2}$$

可知得证。

2.5.6 应用：回归方程的显著性检验

F 检验统计量统一形式:

$$F = \frac{n-p}{q} \times \frac{R_p^2}{1-R_p^2}$$

当 $q = p-1$ 时, 为显著性检验: 用来判断回归模型选取的自变量 X 对响应变量 y 的影响, 拆分 $X = (\mathbf{1}, Z)$, $\beta = (\beta_0, \mathbf{b})^T$, 原假设 $H_0: \mathbf{b} = 0$ 成立时

$$F = \frac{\|Z_c \hat{\mathbf{b}}\|^2}{(p-1) \hat{\sigma}^2} = \frac{n-p}{p-1} \times \frac{R^2}{1-R^2} \sim F_{p-1, n-p}$$

当 $q = 1$ 时, 为单个变量的显著性检验: 原假设 $H_0: \beta_k = 0, k$ 为 $1 \sim p-1$ 中任一值, 记 $\mathbf{x}_k^\perp = \mathbf{x}_k - P_{X_{(-k)}} \mathbf{x}_k$, 其中 $X_{(-k)}$ 是 X 除去第 k 列之外构成的矩阵, 记 $r_p = r_{y x_k \cdot \text{others}}$ 为 y 与 x_k 的偏相关系数, 那么 H_0 的 F 检验为

$$F = \frac{\|\mathbf{x}_k^\perp \hat{\beta}_k\|^2}{\hat{\sigma}^2} = (n-p) \frac{r_p^2}{1-r_p^2} \sim F_{1, n-p}$$

这是特殊的 t 检验:

$$t = \pm \sqrt{F} = \sqrt{n-p} \frac{r_p}{\sqrt{1-r_p^2}} \sim t_{n-p}$$

特别地, $p = 2, q = 1$ 时为一维简单线性模型, 检验的是 x 和 y 的相关性。 $p > 2$ 时检验的是单个变量 x_k 与 y 的偏相关性。