

《线性回归》 —线性回归(5)

杨 瑛

清华大学 数学科学系

Email: yangying@mail.tsinghua.edu.cn

Tel: 62796887

2019.03.26

主要内容：LSE的性质

1 线性模型LSE的性质

- 拟合值和残差
- 投影矩阵的性质
- LSE的性质
- 为什么LSE是一个好的估计？
- 可估函数
- Gauss-Markov定理
- Gauss-Markov定理的证明

拟合值和残差

♠ 对于标准线性模型：

$$\mathbf{Y} = \mathbf{X}\theta + \epsilon, \epsilon \sim N(\mathbf{0}, \Sigma)$$

系数 θ 的LSE是

$$\hat{\theta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

这个估计有时也称之为普通最小二乘估计(**Ordinary Least Squares Estimate**).

♠ 拟合值: $\hat{\mathbf{Y}} = \mathbf{X}\hat{\theta} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = \mathbf{P}\mathbf{Y}$,
其中 $\mathbf{P} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ 称之为投影矩阵。

♠ 残差: $\hat{\epsilon} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I}_n - \mathbf{P})\mathbf{Y}$.

投影矩阵的性质

定理.

设 \mathbf{X} 是秩为 p 的 $n \times p$ 的矩阵, $\mathbf{P} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$. 则

- (i) \mathbf{P} 和 $\mathbf{I}_n - \mathbf{P}$ 是对称的幂等矩阵.
- (ii) $\text{rank}(\mathbf{I}_n - \mathbf{P}) = \text{tr}(\mathbf{I}_n - \mathbf{P}) = n - p$.
- (iii) $\mathbf{P}\mathbf{X} = \mathbf{X}$

Proof.

【黑板】



说明:

若 $\text{rank}(\mathbf{X}) = r < p$, 则上面的定理仍然成立, 只须将 p 替换为 r .

【黑板 (学生), 利用线性代数的知识证明之! 】

进一步的性质

- ♠ 若 \mathbf{X} 满秩, 则 $\sum_{i=1}^n (\mathbf{Y}_i - \hat{\mathbf{Y}}_i) = 0$. 【练习。】
- ♠ $\hat{\mathbf{Y}}^T \times (\mathbf{Y} - \hat{\mathbf{Y}}) = 0$. 【试给出几何解释.】

性质.

- (i) 若 \mathbf{X} 满秩, $\mathbf{E}[\epsilon] = \mathbf{0}$, 则 $\mathbf{E}[\hat{\theta}] = \theta$. 即, $\hat{\theta}$ 是 θ 的无偏估计.
- (ii) 若 $\text{Cov}[\epsilon_i, \epsilon_j] = \delta_{ij}\sigma^2$, 则 $\text{Var}[\epsilon] = \sigma^2\mathbf{I}_n$, $\text{Var}[\mathbf{Y}] = \text{Var}[\epsilon]$, 其中, 若 $i \neq j$, $\delta_{ij} = 0$, 若 $i = j$, $\delta_{ij} = 1$.
- (iii) 若 \mathbf{X} 是列满秩且非随机的设计矩阵, 则

$$\text{Var}[\hat{\theta}] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.$$

为什么LSE是一个好的估计？

- ♠ LSE $\hat{\theta}$ 是到 \mathbf{X} 的列向量张成空间上的正交投影，有几何的直观解释和意义.
- ♠ 如果随机误差是iid 正态的，则 θ 的LSE和MLE是相同的. 粗略来说， θ 的最大似然估计可使得到观察到数据的概率最大.
- ♠ Gauss-Markov定理：LSE是BLUE (best linear unbiased estimate, 最佳线性无偏估计)

可估函数

- ♠ 首先我们解释可估函数(**estimable function**)的概念. 参数 θ 的线性组合 $\Psi = \mathbf{c}^T \theta$ 是可估的, 当且仅当存在组合 $\mathbf{a}^T \mathbf{Y}$ 使得

$$\mathbf{E}[\mathbf{a}^T \mathbf{Y}] = \mathbf{c}^T \theta, \forall \theta.$$

- ♠ 可估函数包括对未来观测的预测, 这解释了为什么要考虑可估函数的原因.
- ♠ 如果 \mathbf{X} 是满秩 (对观测数据而言), 那么所有线性组合都是可估计的.

Theorem (Gauss-Markov 定理)

假设 $\mathbf{E}[\epsilon] = \mathbf{0}$, $\text{Var}[\epsilon] = \sigma^2 \mathbf{I}_n$. 假设模型的结构部分 $\mathbf{E}[\mathbf{Y}] = \mathbf{X}\theta$ 是正确的. 设 $\Psi = \mathbf{c}^T \theta$ 是可估函数, 则在 Ψ 的**所有无偏线性估计类**中, $\hat{\Psi} = \mathbf{c}^T \hat{\theta}$ 是方差最小的估计, 且唯一.

说明一:

如果不对估计的类型加以限制, 很难研究最优的估计! 回顾《统计推断》中对于估计最优性的讨论,

说明二:

如果将定理中的条件加强为 $\epsilon \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, 则定理的结论可以加强为: Ψ 的**所有无偏估计类**中, $\hat{\Psi} = \mathbf{c}^T \hat{\theta}$ 是方差最小的估计, 且唯一.

Gauss-Markov定理的证明

我们从初步计算开始：假设 $\mathbf{a}^T \mathbf{Y}$ 是 $\mathbf{c}^T \theta$ 的某一个无偏估计，那么

$$\begin{aligned}\mathbf{E}[\mathbf{a}^T \mathbf{Y}] &= \mathbf{c}^T \theta, \forall \theta \\ \implies \mathbf{a}^T \mathbf{X} \theta &= \mathbf{c}^T \theta, \forall \theta\end{aligned}$$

上面的式子蕴含着： $\mathbf{a}^T \mathbf{X} = \mathbf{c}^T$ ， \mathbf{c} 必须在 \mathbf{X}^T 的空间范围中，同时也蕴含着 \mathbf{c} 亦在 $\mathbf{X}^T \mathbf{X}$ 的空间范围中，从而，存在 λ 使得

$$\begin{aligned}\mathbf{c} &= \mathbf{X}^T \mathbf{X} \lambda \\ \mathbf{c}^T \hat{\theta} &= \lambda^T \mathbf{X}^T \mathbf{X} \hat{\theta} = \lambda^T \mathbf{X}^T \mathbf{Y}.\end{aligned}$$

现在我们可以证明最小二乘估计具有最小方差—选择任意可估计函数 $\mathbf{a}^T \mathbf{Y}$ ，并计算其方差：

Gauss-Markov定理的证明(续)

$$\begin{aligned}
 \text{Var}[\mathbf{a}^T \mathbf{Y}] &= \text{Var}[\mathbf{a}^T \mathbf{Y} - \mathbf{c}^T \hat{\theta} + \mathbf{c}^T \hat{\theta}] \\
 &= \text{Var}[\mathbf{a}^T \mathbf{Y} - \lambda^T \mathbf{X}^T \mathbf{Y} + \mathbf{c}^T \hat{\theta}] \\
 &= \text{Var}[\mathbf{a}^T \mathbf{Y} - \lambda^T \mathbf{X}^T \mathbf{Y}] + \text{Var}[\mathbf{c}^T \hat{\theta}] \\
 &\quad + 2\text{Cov}[\mathbf{a}^T \mathbf{Y} - \lambda^T \mathbf{X}^T \mathbf{Y}, \lambda^T \mathbf{X}^T \mathbf{Y}]
 \end{aligned}$$

但是

$$\begin{aligned}
 \text{Cov}[\mathbf{a}^T \mathbf{Y} - \lambda^T \mathbf{X}^T \mathbf{Y}, \lambda^T \mathbf{X}^T \mathbf{Y}] &= (\mathbf{a}^T - \lambda^T \mathbf{X}^T) \sigma^2 \mathbf{I}_n \mathbf{X} \lambda \\
 &= (\mathbf{a}^T \mathbf{X} - \lambda^T \mathbf{X}^T \mathbf{X}) \sigma^2 \mathbf{I}_n \lambda \\
 &= (\mathbf{c}^T - \mathbf{c}^T) \sigma^2 \mathbf{I}_n \lambda = 0.
 \end{aligned}$$

这样以来, $\text{Var}(\mathbf{a}^T \mathbf{Y}) = \text{Var}[\mathbf{a}^T \mathbf{Y} - \lambda^T \mathbf{X}^T \mathbf{Y}] + \text{Var}[\mathbf{c}^T \hat{\theta}]$.

Gauss-Markov定理的证明(续)

现在由于方差不能为负，我们可以得到：

$$\text{Var}[\mathbf{a}^T \mathbf{Y}] \geq \text{Var}[\mathbf{c}^T \hat{\theta}].$$

换言之， $\mathbf{c}^T \hat{\theta}$ 具有最小方差.

剩下的就是证明唯一性. 如果 $\text{Var}[\mathbf{a}^T \mathbf{Y} - \lambda^T \mathbf{X}^T \mathbf{Y}] = 0$, 则要求 $\mathbf{a}^T - \lambda^T \mathbf{X}^T = 0$, 即 $\mathbf{a}^T \mathbf{Y} = \lambda^T \mathbf{X}^T \mathbf{Y} = \mathbf{c}^T \hat{\theta}$, 当等号成立时, 只有 $\mathbf{a}^T \mathbf{Y} = \mathbf{c}^T \hat{\theta}$. 因此估计量是唯一的.

说明二中结论的证明:

回想《统计推断》中的有效估计是如何证明的. 【先回顾什么有效估计的定义.】

几点说明:

- ♠ 若 \mathbf{X} 是满秩的, 则 $\mathbf{a}^T \hat{\theta}$ 是 $\mathbf{a}^T \theta$ 的BLUE.
- ♠ 在研究BLUE时, 对 ϵ 的分布没有做假设.
- ♠ 而当 $\epsilon \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ 时, $\mathbf{a}^T \hat{\theta}$ 是 $\mathbf{a}^T \theta$ 的所有无偏估计中方差最小的.
- ♠ 特别的, θ 的每一个分量 θ_i 的估计 $\hat{\theta}_i$ 也有相同的性质.
- ♠ 在一定的条件下, $\hat{\theta}$ 在 $n \rightarrow \infty$ 时, 有相合性和渐近正态性
【黑板: 简单线性模型参数的LSE的相合性和渐近正态性】

随机阵列的CLT

详见R. J. Serfling (1980), Approximation Theorems of Mathematical Statistics. p. 31-32

【第九讲结束】