《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
   - σ^2的无偏估计
   - 分布理论

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

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

  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: $\widehat{\mathbf{Y}} = \mathbf{X}\hat{\theta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{P}\mathbf{Y}$,
  其中$\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ 称之为投影矩阵。
- 残差: $\widehat{\epsilon} = \mathbf{Y} - \widehat{\mathbf{Y}} = (\mathbf{I}_n - \mathbf{P})\mathbf{Y}$.

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) $rank(\mathbf{I}_n - \mathbf{P}) = tr(\mathbf{I}_n - \mathbf{P}) = n - p$.

(iii) $\mathbf{PX} = \mathbf{X}$

Proof.

【黑板】 □

说明:

若$\mathrm{rank}(\mathbf{X}) = r < p$, 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$\sum_{i=1}^{n}(\mathbf{Y}_i - \widehat{\mathbf{Y}}_i) = 0$.【练习。】
- $\widehat{\mathbf{Y}}^T \times (\mathbf{Y} - \widehat{\mathbf{Y}}) = 0$. 【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$\theta$的无偏估计.

(ii) 若$\mathit{Cov}[\epsilon_i, \epsilon_j] = \delta_{ij}\sigma^2$, 则$\mathit{Var}[\epsilon] = \sigma^2\mathbf{I}_n$, $\mathit{Var}[\mathbf{Y}] = \mathit{Var}[\epsilon]$, 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

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

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$\mathbf{E}[\epsilon] = \mathbf{0}$, $Var[\epsilon] = \sigma^2 \mathbf{I}_n$. 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$\Psi$的所有无偏线性估计类中，$\widehat{\Psi} = \mathbf{c}^T\hat{\theta}$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：$\Psi$的所有无偏估计类中，$\widehat{\Psi} = \mathbf{c}^T\hat{\theta}$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

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

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)

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

但是

$$
\begin{aligned}
\mathrm{Cov}[\mathbf{a}^T\mathbf{Y} - \lambda^T\mathbf{X}^T\mathbf{Y}, \lambda^T\mathbf{X}^T\mathbf{Y}] &= (\mathbf{a}^T - \lambda^T\mathbf{X}^T)\sigma 62\mathbf{I}_n\mathbf{X}\lambda \\
&= (\mathbf{a}^T\mathbf{X} - \lambda\mathbf{X}^T\mathbf{X})\sigma^2\mathbf{I}_n\lambda \\
&= (\mathbf{c}^T - \mathbf{c}^T)\sigma^2\mathbf{I}_n\lambda = 0.
\end{aligned}
$$

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

Gauss-Markov定理的证明(续)

现在由于方差不能为负，我们可以得到：

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

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$\mathrm{Var}[\mathbf{a}^T\mathbf{Y} - \lambda^T\mathbf{X}^T\mathbf{Y}] = 0$, 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$\hat{\theta}$也有相同的性质.
- 在一定的条件下, $\hat{\theta}$在$n \to \infty$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】

- 假定误差满足条件：$\epsilon_i \sim (0, \sigma^2)$, 或者$\epsilon_i$ iid $N(0, \sigma^2)$, 则为推断的缘故，需要估计σ^2.
- 通常利用残差$\hat{\epsilon}_i$来估计

Theorem

假设$\mathbf{E}[\mathbf{Y}] = \mathbf{X}\theta$, $\mathbf{X}$是$n \times p$的秩为r $(r \leq p)$的矩阵，$Var[\mathbf{Y}] = \sigma^2 \mathbf{I}_n$，则

$$S^2 = \frac{(\mathbf{Y} - \widehat{\mathbf{Y}})^T (\mathbf{Y} - \widehat{\mathbf{Y}})}{n - r} = \frac{SSE}{n - r}$$

是σ^2的无偏估计，$\widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\theta}$为残差，$SSE$是残差平方和.

证明：

考虑满秩的表示$V=\mathbf{X}_1\alpha$, 其中$\mathbf{X}_1$是秩为r的$n\times r$矩阵，则

$$\mathbf{Y}-\widehat{\mathbf{Y}}=(\mathbf{I}_n-P)\mathbf{Y},$$

其中α是$r\times 1$的列向量，$P=\mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T$. 从而，

$$\begin{aligned}(n-r)S^2 &= \mathbf{Y}^T(\mathbf{I}_n-P)^T(\mathbf{I}_n-P)\mathbf{Y}\\ &= \mathbf{Y}^T(\mathbf{I}_n-P)\mathbf{Y}\\ &= \mathbf{Y}^T(\mathbf{I}_n-P)\mathbf{Y}.\end{aligned}$$

注意到P是投影矩阵，估计$PV=V$, 由此

$$\begin{aligned}\mathbf{E}[\mathbf{Y}^T(\mathbf{I}_n-P)\mathbf{Y}] &= \sigma^2\mathrm{tr}(\mathbf{I}_n-P)+V^T(\mathbf{I}_n-P)V\\ &= \sigma^2(n-r),\end{aligned}$$

即，$\mathbf{E}[S^2]=\sigma^2$.

说明：

- 当$\mathbf{X}$是满秩时，$S^2 = (\mathbf{Y} - \mathbf{X}\hat{\theta})^T(\mathbf{Y} - \mathbf{X}\hat{\theta})/(n-p)$.
- 而且在略强的条件之下，$(n-p)S^2$是$(n-p)\sigma^2$的唯一的非负无偏方差估计.详见Seber and Lee (2003) Theorem 3.4 (Atiqullah, 1962) p. 45. 研究的是σ^2在一定意义下的最优估计.
- Gauss-Markov定理表明，最小二乘估计$\hat{\theta}$是一个不错的选择，但如果误差相关或方差不等，则可能会有更好的估计.在误差是非正态时，非线性或有偏估计在某种意义上可能会有更好的表现.所以这个定理并没有告诉一个人一直要使用最小二乘法，它只是强烈暗示它，除非有一些强有力的理由不这样做.

说明: (续)

- 考虑使用除普通最小二乘以外的估计的情况是:
  - ✓ 当误差相关或具有不等方差时，应使用广义最小二乘法.
  - ✓ 当错误分布是长尾时，可能会使用稳健估计.稳健的估计通常关于$\mathbf{Y}$是非线性的. 例如中位数估计
  - ✓ 当预测变量高度相关（共线）时，可能更倾向于使用诸如岭回归之类的有偏估计. → 此时 X^TX 未必可逆，可以考虑用 $(\delta I_n + X^TX)$ 代替

当估计是非线性的时候，尽量把非线性估计写出一个独立和部分和余项 此时 $\hat{\beta}_n = \beta + \sum_{i=1}^{n} v_{ni} e_i + \underbrace{R_n}_{\text{余项}}$ $R_n = o\left(\frac{1}{\sqrt{n}}\right)$

分布理论

♠ 在更强的条件$\epsilon \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$之下，我们有

Theorem

如果$\mathbf{Y} \sim N_n(\mathbf{X}\theta, \sigma^2 \mathbf{I}_n)$,$\mathbf{X}$是$n \times p$的秩为$p$的矩阵。则
(i) $\hat{\theta} \sim N_p(\theta, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1})$.
(ii) $(\hat{\theta} - \theta)^T \mathbf{X}^T \mathbf{X} (\hat{\theta} - \theta)/\sigma^2 \sim \chi_p^2$.
(iii) $\hat{\theta}$与S^2独立.
(iv) $SSE/\sigma^2 = (n-p)S^2/\sigma^2 \sim \chi_{n-p}^2$.

证明参见Seber and Lee, p. 48

定理的用处：

- 可以用于检验假设【黑板】

$$H_{0j}: \theta_j = 0$$

$$\theta_1 = \cdots = \theta_p = 0,$$

等等。