《线性回归》
—线性回归(6)

杨　　瑛

清华大学　数学科学系
Email: yangying@mail.tsinghua.edu.cn
Tel: 62796887

2019.03.28

主要内容：LSE的性质

- σ^2的无偏估计
- 分布理论
- MLE

σ^2的无偏估计

- 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.
  - ✓ 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.

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) $(\hat{\theta} - \theta)^T \mathbf{X}^T \mathbf{X} (\hat{\theta} - \theta)/\sigma^2 \sim \chi^2_p$.
(iii) $\hat{\theta}$与S^2独立.
(iv) $SSE/\sigma^2 = (n-p)S^2/\sigma^2 \sim \chi^2_{n-p}$.

证明参见Seber and Lee, p. 48

定理的用处：

- 可以用于检验假设【后续还要讲】

$$H_0 : \theta_1 = \cdots = \theta_p = 0,$$

或者

$$H_{0j} : \theta_j = 0, j = 1, \cdots, p,$$

等等。

MLE

考虑线性模型

$$\mathbf{Y} = \mathbf{X}\theta + \epsilon, \epsilon \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n),$$

其中$\mathbf{X}$是秩为p的$n \times p$的设计矩阵。参数θ和σ^2可以通过最大似然方法得到。

由于

$$\mathbf{Y} \sim N_n(\mathbf{X}\theta, \sigma^2 \mathbf{I}_n),$$

故似然函数为：

$$L(\theta, \sigma^2 | \mathbf{X}, \mathbf{Y}) = (2\sigma^2)^{-n/2} \exp\left\{-\frac{\|\mathbf{Y} - \mathbf{X}\theta\|^2}{2\sigma^2}\right\}.$$

经常是先做出似然函数，然后再加上一些惩罚

说明：

- θ和σ^2的MLE求法类似《统计推断》课程中正态分布总体参数的MLE的求法. 详见Seber and Lee （2003）, p. 49–50
- 特别注意: $\hat{\theta}_{LSE} = \hat{\theta}_{MLE}$, 但是$\sigma^2$的估计为：

$$\widehat{\sigma}^2 = \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\hat{\theta}\|^2$$

是σ^2的有偏估计。
σ^2的无偏估计是：

$$S^2 = \frac{1}{n-p}\|\mathbf{Y} - \mathbf{X}\hat{\theta}\|^2,$$

其中$p = \dim(\theta)$ (为θ的维数).

思考题：

考虑线性模型

$$\mathbf{Y} = \mathbf{X}\theta + \epsilon, \epsilon \sim N_n(\mathbf{0}, \sigma^2\Sigma),$$

其中$\mathbf{X}$是秩为p的$n \times p$的设计矩阵, θ和σ^2是未知的，Σ是已知的$n \times n$的正定矩阵。

- 试讨论θ和σ^2的MLE以及θ的LSE，以及基于LSE的σ^2的估计
- 试研究这些估计的性质（例如，无偏性，相合性，分布，相对效率等）

作业：

考虑线性模型

$$\mathbf{Y}_i = \mathbf{X}_i^T\theta + \epsilon_i, \epsilon_1, \cdots, \epsilon_n \sim \ iid\ U(-\sigma, \sigma),$$

其中$\mathbf{X}$ 是列满秩, θ 和σ^2 是未知的. 试研究参数θ 和σ 的LSE 和MLE 的性质[包括无偏性，相合性和渐近正态性等]

【第十讲结束】