Statistic Models and Estimation

统计模型和估计

Jerry Ling

2024 年 8 月 5 日

概率论由随机变量性质研究其结果，而数理统计由有限观测结果研究随机变量本身性质。为了建立其框架，首先引入统计模型的概念：

> **Definition 1** (Statistic Model)**.** *Suppose our observed outcomes* $\{X_i\}s$ *is generated by r.v. on probability space (E,*$\mathscr{B}$*,*P) described by distribution $\mathbb{P}$, a statistic model of this experiment is a pair
>
> $$(E, (\mathbb{P}_\theta)_{\theta\in\Theta})$$
>
> i.e. a collection of distributions on space E, each described by parameter θ.

注意我们选取的模型未必能代表随机变量。称能代表之的模型（存在 θ 使得 $\mathbb{P}_\theta = \mathbb{P}$）为 **well specified**，正确的参数为**真参数**，此后讨论<u>默认其存在</u>。当参数空间 Θ 有限维（$\in \mathbb{R}^d$）称为**参数化（parametric）模型**。在某些时候我们会直接估计分布函数（及其泛函），因为其属于函数空间，故是非参数估计。线性回归方法则使用与 n 个参数来建立 (n+1) 维数据的关系，因而是参数模型。

另一方面，为了清晰性，我们还要求参数到分布的映射是单射的，即模型的参数是**可分辨的 (identifible)**。这是为了避免多个真参数的情况。指认了模型后，为了找到这个真参数，我们需要利用观测得到的 n 个数据，即 n 个独立等同分布的随机变量X_i：

> **Definition 2** (Statictics and Estimator). *A statistics(统计量) is a function g of n observed r.v.* $\{X_i\}_{i=1,2,\dots,n}$. *An estimator* $\hat{\theta}_n$ *is a statistics that* ***does not*** *contain* θ.
>
> $$\hat{\theta}_n \triangleq g(\{X_i\}_{i=1,2,\dots,n})$$

注意，估计器本身是一个随机变量（序列），且必然由真分布生成。因而我们可以计算其期望、方差（序列）并定义其偏差和标准差：

$$bias \triangleq \mathbb{E}\hat{\theta}_n - \theta \tag{1}$$

$$se(\hat{\theta}_n) \triangleq \sqrt{\mathbb{V}ar\hat{\theta}_n} \tag{2}$$

此后标准差简写为 se。若估计器序列收敛到我们需要的参数，则称之为**一致的 (consistent)**:

Theorem 1 (Consistency Condition).

If

$$bias \to 0, se \to 0$$

then

$$MSE = \mathbb{E}(\hat{\theta} - \theta)^2 \to 0$$

then

$$\hat{\theta} \xrightarrow{\mathbb{P}} \theta$$

其中 MSE 是**均方差 (mean squared error)**，等于偏差和标准差的平方和。

考虑到每个观测量都是独立等同分布，我们容易想到使用中心极限定理来刻画较大观测数量（30+）的估计器的分布。实际上，如果能说明估计器序列本身在真值附近是**渐近正态分布 (asymptotically normal)**，就能直接导出该估计的置信区间：

Confidence Interval

我们的问题是，当模型和估计器已知时，如何（近似）导出置信区间。首先我们考虑 0-1 上的 Ber(p) 变量，p 的估计器 $\hat{p}_n$ 设定为其均值。根据大数定律其必然收敛于其

期望，即偏差趋于 0。记 Ber(p) 的方差为 $\sigma^2 = p(1-p)$，则：

$$se = \frac{\sigma}{\sqrt{n}} \tag{3}$$

显然 se 也趋于 0，则该估计器是一致的。根据中心极限定理 $\frac{\hat{p}_n - p}{se}$ 收敛到正态分布，也就是说此处估计器本身就是渐进正态分布：

$$\sqrt{n}\frac{\hat{p}_n - p}{\sqrt{p(1-p)}} \xrightarrow{d.} N(0,1) \tag{4}$$

此处估计器 $\hat{p}_n$ 是随机的，而真参数 p 是确定值。为估计之，考虑在 $\hat{p}_n$ 周围构造一个区间——其位置参数 p_1,p_2 也是随机变量，构造条件为该区间囊括真参数 p 的概率 $1-\alpha$，也就是**置信度 (level of confidence)**：

> **Definition 3** (Confidence Interval)**.** *At level of confidence 1-α, CI $\triangleq (\hat{p}_n - p_1, \hat{p}_n + p_2)$, such that:*
>
> $$\mathrm{P_p}(\hat{\mathrm{p}}_\mathrm{n} - \mathrm{p_1} < \mathrm{p} < \hat{\mathrm{p}}_\mathrm{n} + \mathrm{p_2}) \geq 1 - \alpha$$

其中 P 的下标表示其中的随机变量均由真参数分布 $\mathbb{P}_p$ 生成。将其同 (4) 比较发现只要分母上的 p 可以去除掉，那么就可以根据高斯分布给出置信区间。其中又可以使用**保守上界 (conservative bound)**、**插入 (plug-in)**、**解方程 (solving equations)** 三种方法得到。此例子中，使用上界是较好的方法。插入法即将方差 p(1-p) 的估计器 $\hat{\sigma}_n^2$ 直接替换之，在 n 较大时有效。

Delta Method for CI Determination

再考虑一个常见的分布 Exp(λ)，因为期望是参数 λ 的倒数，估计器可如下设计：

$$\hat{\lambda}_n = \frac{1}{\frac{1}{n}\sum X_i} \tag{5}$$

由于大数定律，分母依概率收敛于期望，因而这也是一个**无偏的 (unbiased)** 估计器。但中心极限定理却不能直接运用到估计器上：

$$\sqrt{n}\frac{1/\hat{\lambda}_n - 1/\lambda}{\sigma} = \sqrt{n}\frac{1/\hat{\lambda}_n - 1/\lambda}{1/\lambda} \xrightarrow{d.} N(0,1) \tag{6}$$

因而引入一个实用的定理：

Theorem 2 (Delta Method)**.** *For a function g:*

If

$$Y_n \xrightarrow{\mathbb{P}} N(\mu, \frac{\sigma^2}{n}), g'(\mu) \neq 0$$

Then

$$g(Y_n) \xrightarrow{\mathbb{P}} N(g(\mu), g'(\mu)^2 \frac{\sigma^2}{n})$$

则代入反比例函数以导出式 (6) 中估计器 $\hat{\lambda}_n$ 的分布：

$$\hat{\lambda}_n \xrightarrow{d.} N(\lambda, \lambda^4 \frac{\sigma^2}{n}) = N(\lambda, \frac{\lambda^2}{n}) \tag{7}$$

其后置信区间的计算与上一节介绍的流程基本一致，但方程法从二元变为线性，而上界法失效。

Method of Moment