极大似然估计

司继春

[1]上海对外经贸大学

2025年11月

概览

极大似然估计的思想

- 与矩估计一样，**极大似然估计量（maximum likelihood estimator，MLE）**也是最常用的估计方法之一。
- 在一定条件下，极大似然估计也可以保证一致性等大样本性质，与矩估计不同的是，极大似然估计通常还额外具有有效性，即极大似然估计量的方差能够达到效率上界
  - 当然，要达到这一性质往往需要比矩估计更强的假设，比如关于数据分布的假设。
- 其思想是，如果我们要对未知参数总体P_θ做推断，估计θ_0，那么我们就寻找一个$\hat{\theta}$，使得这组数据出现的概率最高，则$\hat{\theta}$理应是θ_0的一个合理估计。

极大似然估计的思想

极大似然估计的思想

在抛硬币的例子中，如果我们观察到在100次抛硬币的独立试验中，有55次正面以及45次反面，我们需要对抛一枚硬币得到正面的结果p进行估计。

- 首先猜测$p = 0.1$，那么得到55次正面以及45次反面这一结果的概率为：$0.1^{55} \times 0.9^{45} = 8.73 \times 10^{-58}$，显然该概率过低，即在$p = 0.1$的情况下，得到该结果的可能性很低。
- 如果猜测$p = 0.5$，那么得到该结果的概率为$0.5^{55} \times 0.5^{45} = 7.89 \times 10^{-31}$，可能性大了很多；
- 当猜测$p = 0.55$的时候，得到该结果的概率为$0.55^{55} \times 0.45^{45} = 1.3 \times 10^{-30}$达到最大值，因而很自然的，我们应该更倾向于认为，这枚硬币得到正面的概率最有可能等于0.55。

极大似然估计

- 如果假设一组独立同分布的样本$\boldsymbol{x} = [x_1, ..., x_N]'$来自于参数总体$\mathrm{P}_\theta$，且密度函数为$f(x_i|\theta)$，那么样本的联合分布函数为

$$f(\boldsymbol{x}|\theta) = \prod_{i=1}^{N} f(x_i|\theta)$$

- 现在，将未知参数θ视为变量，$\boldsymbol{x}$为给定的样本，由于对数函数为单调函数，因而可以将联合分布函数取对数，得到**对数似然函数（log-likelihood function）**：

$$L(\theta|\boldsymbol{x}) = \ln f(\boldsymbol{x}|\theta) = \sum_{i=1}^{N} \ln f(x_i|\theta)$$

- 极大似然估计即找到一个$\hat{\theta}$使得对数似然函数最大化

$$\hat{\theta} = \arg\max_{\theta} L(\theta|\boldsymbol{x})$$

从而我们得到了极大似然估计量$\hat{\theta}$。

极大似然估计的简单应用

伯努利分布的极大似然估计

如果$x_i \sim \mathcal{B}er\left(p_0\right)\ i.i.d$，那么其联合密度函数为：

$$f\left(\boldsymbol{x}|p\right)=\prod_{i=1}^{N}p^{x_i}\left(1-p\right)^{1-x_i}$$

对数似然函数为：

$$\begin{aligned}L\left(p|\boldsymbol{x}\right)&=\sum_{i=1}^{N}\left[x_i\ln p+\left(1-x_i\right)\ln\left(1-p\right)\right]\\&=\left(\sum_{i=1}^{N}x_i\right)\ln p+\left(N-\sum_{i=1}^{N}x_i\right)\ln\left(1-p\right)\end{aligned}$$

极大似然估计的简单应用

伯努利分布的极大似然估计

- 现在欲得到p的极大似然估计值，只要对上述对数似然函数求最大值，即:

$$\frac{\partial L\left(p|\boldsymbol{x}\right)}{\partial p}=\frac{\sum_{i=1}^{N}x_i}{p}-\frac{\left(N-\sum_{i=1}^{N}x_i\right)}{1-p}=0$$

  从而得到：$\hat{p}=\frac{1}{N}\sum_{i=1}^{N}x_i$

- 无偏性和一致性讨论略

极大似然估计的简单应用

正态分布的极大似然估计

如果$x_i \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$ $i.i.d$，其中$\theta = \left[\mu, \sigma^2\right]'$，那么其联合密度函数为

$$f\left(\boldsymbol{x}|\theta\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\left(x_i - \mu\right)^2}{2\sigma^2}\right\}$$

对数似然函数为:

$$\begin{aligned} L\left(\theta|\boldsymbol{x}\right) &= \sum_{i=1}^{N}\left[-\frac{1}{2}\ln\left(2\pi\right) - \ln\sigma - \frac{\left(x_i-\mu\right)^2}{2\sigma^2}\right] = -\frac{N}{2}\ln\left(2\pi\right) - N\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(x_i-\mu\right)^2 \\ &= -\frac{N}{2}\ln\left(2\pi\right) - N\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(x_i^2 + \mu^2 - 2\mu x_i\right) \\ &= -\frac{N}{2}\ln\left(2\pi\right) - N\ln\sigma - \frac{N\mu^2}{2\sigma^2} - \frac{N}{2\sigma^2}\overline{x^2} + \frac{N\mu}{\sigma^2}\bar{x} \end{aligned}$$

极大似然估计的简单应用

正态分布的极大似然估计

对其求极大值，得到：

$$\frac{\partial L\left(\theta|\boldsymbol{x}\right)}{\partial\theta}=N\begin{pmatrix}-\frac{\mu}{\sigma^2}+\frac{\bar{x}}{\sigma^2}\\-\frac{1}{2\sigma^2}+\frac{\mu^2+x^2-2\mu\bar{x}}{2\sigma^4}\end{pmatrix}=0$$

解得：

$$\begin{cases}\hat{\mu}=\bar{x}\\\hat{\sigma}^2=\overline{x^2}-\bar{x}^2\end{cases}$$

这与矩估计的估计量是一样的。我们之前已经证明了，$\hat{\mu}$是μ_0的无偏、一致估计量，而$\hat{\sigma}^2$是σ_0^2的一致估计量，$\frac{N}{N-1}\hat{\sigma}^2$是$\sigma_0^2$的无偏估计量。

极大似然估计中的一阶条件

- 上例都使用了一阶条件得到了极大似然估计，但是有时极大似然函数并不可导，比如如下例子：

均匀分布的极大似然估计

如果$x_i \sim \mathcal{U}(0, \beta)$ $i.i.d$，那么密度函数为

$$f(x_i|\beta) = \mathbb{1}\{0 \leq x_i \leq \beta\} \cdot \frac{1}{\beta}$$

从而

$$\ln f(x_i|\beta) = \ln\left[\mathbb{1}\{0 \leq x_i \leq \beta\} \cdot \frac{1}{\beta}\right] = \begin{cases} -\ln \beta & 0 \leq x_i \leq \beta \\ -\infty & else \end{cases}$$

极大似然估计中的一阶条件

均匀分布的极大似然估计

对数似然函数为

$$L(\beta|\boldsymbol{x}) = \sum_{i=1}^{N} \ln f(x_i|\beta) = \begin{cases} -N\ln\beta & \max_i\{\boldsymbol{x}\} \le \beta \\ -\infty & else \end{cases}$$

最大化以上对数似然函数：

- 首先需要保证似然函数取值不为$-\infty$，从而必须有$\hat{\beta} \ge \max_i\{\boldsymbol{x}\}$；
- 给定$\hat{\beta} \ge \max_i\{\boldsymbol{x}\}$，$-N\ln\beta$是一个单调递减的函数
- 从而极大似然估计量即满足$\hat{\beta} \ge \max_i\{\boldsymbol{x}\}$最小的数，即$\hat{\beta} = \max_i\{\boldsymbol{x}\}$。

极大似然估计：无解析解

- 虽然以上例子中我们都能得到极大似然估计量的解析解，然而更多的时候，由于要求对数似然函数的最大化，解析解是无法获得的
- 此时我们需要借助数值最优的方法获得极大似然估计

极大似然估计：无解析解

Beta分布的极大似然估计

如果$x_i \sim \mathcal{B}eta\left(\alpha,\beta\right)\ i.i.d,$ 那么

$$\ln f\left(x_i|\alpha,\beta\right) = -\ln \mathrm{B}\left(\alpha,\beta\right) + \left(\alpha-1\right)\ln x_i + \left(\beta-1\right)\ln\left(1-x_i\right)$$

从而对数似然函数为

$$\begin{aligned} L\left(\alpha,\beta|\boldsymbol{x}\right) &= \sum_{i=1}^{N} \ln f\left(x_i|\alpha,\beta\right) \\ &= \sum_{i=1}^{N} \left[-\ln \mathrm{B}\left(\alpha,\beta\right) + \left(\alpha-1\right)\ln x_i + \left(\beta-1\right)\ln\left(1-x_i\right)\right] \end{aligned}$$

其中

$$\ln \mathrm{B}\left(\alpha,\beta\right) = \ln \Gamma\left(\alpha\right) + \ln \Gamma\left(\beta\right) - \ln \Gamma\left(\alpha+\beta\right)$$

极大似然估计：无解析解

Beta分布的极大似然估计

一阶条件为

$$\begin{aligned}\frac{\partial L(\alpha,\beta|\boldsymbol{x})}{\partial\begin{bmatrix}\alpha\\ \beta\end{bmatrix}} &= \sum_{i=1}^{N}\frac{\partial \ln f(x_i|\alpha,\beta)}{\partial\begin{bmatrix}\alpha\\ \beta\end{bmatrix}}\\ &= \sum_{i=1}^{N}\begin{bmatrix}-\psi(\alpha)+\psi(\alpha+\beta)+\ln x_i\\ -\psi(\beta)+\psi(\alpha+\beta)+\ln(1-x_i)\end{bmatrix}\\ &= 0\end{aligned}$$

其中$\psi(\cdot)$为Digamma函数，即对数Gamma函数$\ln\Gamma(\cdot)$的导数

$$\psi(\alpha)=\frac{d\ln\Gamma(\alpha)}{d\alpha}=\frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$$

极大似然估计：无解析解

Beta分布的极大似然估计

- 由于对数Gamma函数、Digamma函数都是非常复杂的非线性函数，无法直接求解一阶导得到解析解，必须诉诸于数值解。
- 在Stata中，ml命令可以帮助我们得到极大似然估计。
- 在Stata中，需要首先提供极大似然估计的目标函数，而目标函数根据任务的不同以及提供信息的不同有多种写法（Stata附录）
- MLE_beta_nod.do

极大似然估计：无解析解

Beta分布的极大似然估计

- 其中betaobj为计算对数似然函数的程序
- 而“ml model”使用ml命令定义了极大似然估计的模型，其目标函数为betaobj，参数为/alpha和/beta。
- 在betaobj中，lnfj即为需要返回的变量$\ln f\left(x_i|\alpha,\beta\right)$，Stata会根据这一变量自动计算$L\left(\alpha,\beta|\boldsymbol{x}\right)=\sum_{i=1}^{N}\ln f\left(x_i|\alpha,\beta\right)$
- 而theta1对应/alpha，theta2对应/beta。
- ml search为寻找初始点
- ml maximize即让Stata寻找似然函数最大的点。

极大似然估计：无解析解

Beta分布的极大似然估计

- 当然在这里，我们既然已经计算出了导数，就可以综合使用导数信息，即使用ml lf1而非ml lf，这样可能提高计算效率
- MLE_beta.do

极大似然估计：无解析解

Beta分布的极大似然估计

- 注意以上在定义目标函数的时候，todo用于控制是否计算导数；lnfj为变量ln $f(x_i|\alpha,\beta)$，而g1、g2为对数似然函数对\alpha,\beta的导数；
- b为所有的参数。由于我们在定义极大似然问题时使用了lf1，我们将\alpha,\beta分别看做两个“方程”：
  - (alpha: x= , freeparm)为第一个方程，x为数据，在目标函数程序中使用$ML_y1代指，freeparm声明alpha为一个参数；
  - /beta为第二个方程，同样也是一个freeparm，但是没有提供数据，所以直接用/beta即可。
- 由于存在两个方程，所以我们使用了mleval命令区分了两个不同方程的参数。
- 接下来目标函数中只要逐一计算目标函数和导函数即可，其中g1为第一个方程的导数，而g2为第二个方程的导数。

极大似然估计：截尾数据

此外，以上例子要么是离散型随机变量，要么是连续型随机变量。而实际中碰到的数据可能是两种类型分布的混合，比如下例：

极大似然估计：截尾数据

截尾数据

现在正在进行一项调查，其中一项调查为收入（y_i）调查，其中关于收入的问题为：

- 请问您的收入是多少?
  - 小于1000
  - 大于10000
  - 其他_____（请填写具体数值）

如果假设收入的对数（$x_i^* = \log_{10} y_i$）服从正态分布，即$x_i^* \sim \mathcal{N}\left(\mu, \sigma^2\right)$ $i.i.d$，那么我们观察到的数据为：

$$x_i = \begin{cases} 3 & y_i \leq 1000 \\ 4 & y_i \geq 10000 \\ x_i^* & otherwise \end{cases}$$

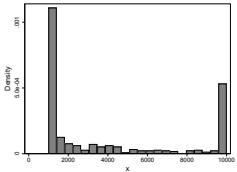
我们称数据存在**截尾（censoring）**现象。

极大似然估计：截尾数据

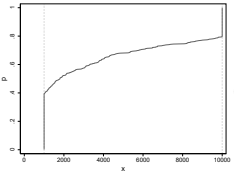
截尾数据

- dgp_mle_censor.ado对以上过程进行了模拟
  - 为了后续程序调用，写成了.ado文件
- 可以观察到很多生成的数据值都是1000或者10000。

极大似然估计：截尾数据



(a)



(b)

极大似然估计：截尾数据

截尾数据

为了估计以上问题，我们可以计算

$$\begin{aligned}\mathrm{P}\left(x_i=3\right)&=\mathrm{P}\left(x_i^*\leq 3\right)\\&=\mathrm{P}\left(\frac{x_i^*-\mu}{\sigma}\leq\frac{3-\mu}{\sigma}\right)=\Phi\left(\frac{3-\mu}{\sigma}\right)\end{aligned}$$

同理$\mathrm{P}\left(x_i=4\right)=1-\Phi\left(\frac{4-\mu}{\sigma}\right)$。因而$x_i$的密度函数为

$$f\left(x_i|\theta\right)=\left[\Phi\left(\frac{3-\mu}{\sigma}\right)\right]^{\mathbb{1}\{x_i=3\}}\left[1-\Phi\left(\frac{4-\mu}{\sigma}\right)\right]^{\mathbb{1}\{x_i=4\}}\left[f\left(x_i\right)\right]^{\mathbb{1}\{3<x_i<4\}}$$

其中$f\left(x_i\right)$为$3<x_i<4$时x_i的密度函数，我们知道$z_i=\frac{x_i-\mu}{\sigma}\sim\mathcal{N}\left(0,1\right)$，从而$F\left(x\right)=\mathrm{P}\left(\frac{x_i-\mu}{\sigma}\leq\frac{x-\mu}{\sigma}\right)=\Phi\left(\frac{x-\mu}{\sigma}\right)$，从而密度函数：

$$f\left(x_i\right)=\frac{1}{\sigma}\phi\left(\frac{x_i-\mu}{\sigma}\right)$$

极大似然估计：截尾数据

截尾数据

因而其对数似然函数为

$$L(\theta|\boldsymbol{x}) = N_3 \ln \Phi\left(\frac{3-\mu}{\sigma}\right) + N_4 \ln\left[1-\Phi\left(\frac{4-\mu}{\sigma}\right)\right] + \sum_{i=1}^{N} \mathbb{1}\{3 < x_i < 4\} \ln\left[\frac{1}{\sigma}\phi\left(\frac{x_i-\mu}{\sigma}\right)\right]$$

最大化以上对数似然函数，我们就得到了正态分布总体的极大似然估计。（MLE_censor.do）

极大似然估计的一致性

- 我们知道，对数似然函数

$$\frac{1}{N}L\left(\theta|\boldsymbol{w}\right)=\frac{1}{N}\sum_{i=1}^{N}\ln f\left(w_i|\theta\right)$$

对于任意一个给定的θ（不一定是真值θ_0），在一定的条件下，根据大数定律，有:

$$\frac{1}{N}L\left(\theta|\boldsymbol{w}\right)\xrightarrow{p}\mathbb{E}\ln f\left(w_i|\theta\right)\triangleq\mathcal{L}\left(\theta\right)$$

即样本似然函数收敛到总体的似然函数。

- 在一定条件下，这个收敛是一致收敛的。

极大似然估计的一致性

- 由于$\frac{1}{N}L(\theta|\boldsymbol{w})$一致收敛到$\mathcal{L}(\theta)$，样本似然函数$L(\theta|\boldsymbol{w})$（红线）随着样本量增大不断逼近总体似然函数$\mathcal{L}(\theta)$（黑线），而极大似然估计的方法是最大化$L(\theta|\boldsymbol{w})$获得估计，即

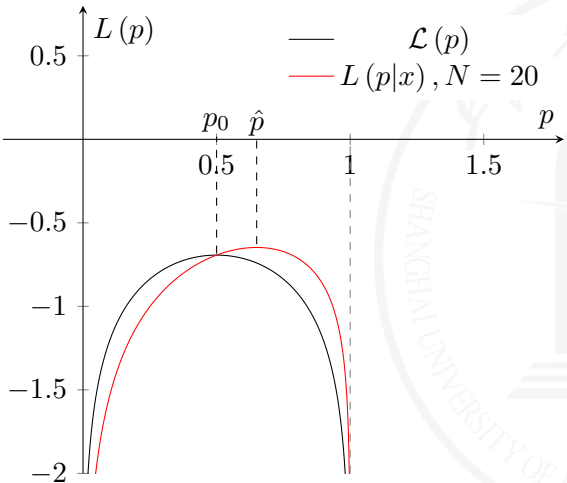
$$\hat{\theta} = \arg\max_{\theta} L(\theta|\boldsymbol{w})$$

- 那么如果真值

$$\theta_0 = \arg\max_{\theta} \mathcal{L}(\theta)$$

可以想象，样本似然函数最大值$\hat{\theta}$应该也会收敛到总体似然函数最大值θ_0。

极大似然估计的一致性



极大似然估计的一致性

- 或者，从矩估计的角度出发，假设$\mathcal{L}(\theta)$连续可微，如果$\theta_0 = \arg\max_\theta \mathcal{L}(\theta)$成立，那么一阶条件为

$$\frac{\partial}{\partial\theta}\mathcal{L}(\theta_0) = \frac{\partial}{\partial\theta}\mathbb{E}\left[\ln f(w_i|\theta_0)\right] = \mathbb{E}\left[\frac{\partial}{\partial\theta}\ln f(w_i|\theta_0)\right] = 0$$

- 以上可以作为总体矩条件方程，而样本矩方程为

$$\frac{1}{N}\sum_{i=1}^{N}\left[\frac{\partial}{\partial\theta}\ln f(w_i|\theta)\right] = 0$$

  求解以上样本矩方程实际上就是得到了$\arg\max_\theta L(\theta|\boldsymbol{w})$。
- 因而如果$\frac{\partial}{\partial\theta}\ln f(w_i|\theta_0)$满足矩估计一致性的条件，那么那么极大似然估计一定是一致的。
- 但是真值θ_0是不是的确能够最大化总体似然函数$\mathscr{L}(\theta) = \mathbb{E}\ln f(w_i|\theta)$呢?

极大似然估计的一致性

伯努利分布的总体似然函数

若$x_i \sim \mathcal{B}er\left(p_0\right)\ i.i.d$，那么其联合密度函数为：

$$f\left(\boldsymbol{x}|p\right)=\prod_{i=1}^{N}p^{x_i}\left(1-p\right)^{1-x_i}$$

对数似然函数为：

$$L\left(p|\boldsymbol{x}\right)=\sum_{i=1}^{N}\left[x_i\ln p+\left(1-x_i\right)\ln\left(1-p\right)\right]$$

根据大数定律，对于任意的p，上述似然函数

$$\begin{aligned}\frac{1}{N}L\left(p|\boldsymbol{x}\right)\xrightarrow{p}\mathcal{L}\left(p\right)&\triangleq\mathbb{E}\left[x_i\ln p+\left(1-x_i\right)\ln\left(1-p\right)\right]\\&=\mathbb{E}\left(x_i\right)\ln p+\mathbb{E}\left(1-x_i\right)\ln\left(1-p\right)\\&=p_0\ln p+\left(1-p_0\right)\ln\left(1-p\right)\end{aligned}$$

极大似然估计的一致性

伯努利分布的总体似然函数

是不是只有当$p = p_0$时，$\mathcal{L}(p)$达到了最大值呢?

- 为求最大值，我们对$\mathcal{L}(p)$求导数并令其等于0得到

$$\frac{\partial \mathcal{L}(p)}{\partial p} = \frac{p_0}{p} - \frac{1-p_0}{1-p} = 0$$

- 从而只有当$p = p_0$时，以上导数等于0。因而真值p_0最大化了总体似然函数$\mathcal{L}(p)$。

极大似然估计的一致性

- 实际上我们可以证明，如果$\mathcal{L}(\theta)$连续可微，那么θ_0一定会使得一阶导数为0。
- 为了证明这一点，我们可以从总体似然函数出发：

$$\begin{aligned}\mathcal{L}(\theta) &= \mathbb{E}\ln f(w_i|\theta) \\ &= \mathbb{E}_{\theta_0}\ln f(w_i|\theta) \\ &= \int_{\mathbb{R}} \ln f(w|\theta)\cdot f(w|\theta_0)\,dw\end{aligned}$$

求其最大值，其一阶条件为

$$\begin{aligned}\frac{\partial \mathcal{L}(\theta)}{\partial\theta} &= \frac{\partial}{\partial\theta}\int_{\mathbb{R}} \ln f(w|\theta)\cdot f(w|\theta_0)\,dw \\ &= \int_{\mathbb{R}} \frac{\partial}{\partial\theta}\ln f(w|\theta)\cdot f(w|\theta_0)\,dw \\ &= \int_{\mathbb{R}} \frac{1}{f(w|\theta)}\frac{\partial f(w|\theta)}{\partial\theta}\cdot f(w|\theta_0)\,dw\end{aligned}$$

极大似然估计的一致性

- 当$\theta = \theta_0$时，有

$$\begin{aligned}\frac{\partial \mathcal{L}(\theta_0)}{\partial \theta} &= \int_{\mathbb{R}} \frac{1}{f(w|\theta_0)} \frac{\partial f(w|\theta_0)}{\partial \theta} \cdot f(w|\theta_0)\, dw \\ &= \int_{\mathbb{R}} \frac{\partial f(w|\theta_0)}{\partial \theta} dw \\ &= \frac{\partial}{\partial \theta} \int_{\mathbb{R}} f(w|\theta_0)\, dw \\ &= 0\end{aligned}$$

其中最后一步由于$\int_{\mathbb{R}} f(w|\theta_0)\, dw = 1$。

极大似然估计的一致性

- 我们通常称对数似然函数的一阶导数$\frac{\partial}{\partial\theta}\ln f\left(w_i|\theta\right)$为**得分函数（score function）**，记为

$$s_i\left(\theta\right)=\frac{\partial}{\partial\theta}\ln f\left(w_i|\theta\right)$$

- 以上结论意味着得分函数在真值处的期望等于0，即$\mathbb{E}\left[s_i\left(\theta_0\right)\right]=0$。
- 至此，如果得分函数$s_i\left(\theta\right)$满足矩估计一致性的要求，那么根据矩条件的一致性结论，必然有极大似然估计量$\hat{\theta}\xrightarrow{p}\theta_0$。

极大似然估计的一致性

- 然而如果使用矩估计的思路证明极大似然的一致性，仍然需要假设似然函数连续可微，且对得分函数的性质有比较高的要求。
- 实际上我们可以通过证明真值θ_0最大化了总体似然函数$\mathcal{L}(\theta)$，从而不依赖导数就可以证明极大似然估计量的一致性。
- 注意前面虽然证明了$\mathbb{E}[s_i(\theta_0)] = 0$，但这只是$\theta_0$最大化了$\mathcal{L}(\theta)$的必要条件。
- 为了更进一步得到真值θ_0是否最大化了总体似然函数，我们引入Kullback-Leiber信息的概念。

Kullback-Leiber信息

Kullback-Leiber信息

令P和Q为同一概率空间中的两个概率函数，f和g分别为其概率密度函数。**Kullback-Leiber信息**被定义为

$$\mathbb{K}(P,Q)=\int_{\mathbb{R}} f(w)\ln\frac{f(w)}{g(w)}dw$$

当P和Q属于参数族P_θ和Q_η时，Kullback-Leiber信息即

$$\mathbb{K}(\theta,\eta)=\int_{\mathbb{R}} f(w|\theta)\ln\frac{f(w|\theta)}{g(w|\eta)}dx=\mathbb{E}_\theta\left[\ln\frac{f(w|\theta)}{g(w|\eta)}\right]$$

其中$f(w|\theta),g(w|\eta)$分别为P_θ和Q_η的密度函数。

Kullback-Leiber信息

- 实际上，Kullback-Leiber信息度量的是两个概率函数的“距离”
- 可以证明，Kullback-Leiber信息$\mathbb{K}(P,Q) \geq 0$，当且仅当$P = Q$时等号成立。
- 然而Kullback-Leiber信息度量并不是一个数学意义上的距离，因为其不具有对称性，即$\mathbb{K}(P,Q) \neq \mathbb{K}(Q,P)$
- 所以我们也将其成为**Kullback-Leiber散度（Kullback-Leiber divergence）**，而非距离。

Kullback-Leiber信息

- 对于极大似然函数，给定任意一个θ，其代表的概率函数与真值θ_0代表的概率函数之间Kullback-Leiber散度为：

$$\mathbb{K}\left(\theta_0,\theta\right)=\mathbb{E}_{\theta_0}\left[\ln\frac{f\left(w|\theta_0\right)}{f\left(w|\theta\right)}\right]=\mathbb{E}\left[\ln\frac{f\left(w|\theta_0\right)}{f\left(w|\theta\right)}\right]$$

- 注意到根据Jensen不等式，有

$$\begin{aligned}-\mathbb{E}\left[\ln\frac{f\left(w|\theta_0\right)}{f\left(w|\theta\right)}\right]&=\mathbb{E}\left[\ln\frac{f\left(w|\theta\right)}{f\left(w|\theta_0\right)}\right]\\&\leq\ln\left(\mathbb{E}\left[\frac{f\left(w|\theta\right)}{f\left(w|\theta_0\right)}\right]\right)=\ln\left(\int_{\mathbb{R}}f\left(w|\theta_0\right)\frac{f\left(w|\theta\right)}{f\left(w|\theta_0\right)}dw\right)\\&=\ln\left(\int_{\mathbb{R}}f\left(w|\theta\right)dw\right)\\&=\ln 1=0\end{aligned}$$

以上不等式当且仅当$f\left(w|\theta\right)=f\left(w|\theta_0\right)$时等号才成立。

Kullback-Leiber信息

- 如果额外假设θ与密度函数一一对应，即不存在$\theta \neq \theta'$使得$f(w|\theta) = f(w|\theta')$，那么不等式

$$\mathbb{E}\left[\ln \frac{f(w|\theta_0)}{f(w|\theta)}\right] \geq 0$$

  只有在$\theta = \theta_0$时等式成立。

- 因而θ_0最小化了：

$$\mathbb{E}\left[\ln \frac{f(w|\theta_0)}{f(w|\theta)}\right] = \mathbb{E}\left[\ln f(w|\theta_0)\right] - \mathbb{E}\left[\ln f(w|\theta)\right]$$

- 由于$\mathbb{E}\left[\ln f(w|\theta_0)\right]$为常数，从而最小化以上目标函数等价于最大化$\mathbb{E}\left[\ln f(w|\theta)\right] = \mathcal{L}(\theta)$，即

$$\theta_0 = \arg\max_{\theta} \mathbb{E}\left[\ln f(w|\theta)\right] = \arg\max_{\theta} \mathcal{L}(\theta)$$

极大似然估计的一致性

极大似然估计的一致性

如果$\{w_i, i = 1, ..., N\}$为一系列独立同分布的随机向量，假设：

1. $\theta_0 \in \Theta \subset \mathbb{R}^K$，其中$\Theta$为紧集；
2. （连续性条件）对于任意的w，函数$\ln f(w_i|\theta) \in \mathbb{R}$在$\Theta$上对$\theta$为连续函数；
3. （收敛性条件）存在一个函数$K(w)$，使得对于任意的$\theta \in \Theta$，$|\ln f(w|\theta)| \leq K(w)$，其中$\mathbb{E}[K(w)] < \infty$；
4. （识别条件）$f(w_i|\theta_0)$为真实的密度函数，且不存在$\theta' \neq \theta_0$使得$f(w_i|\theta_0) = f(w_i|\theta')$。

那么估计量：

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{N} \ln f(w_i|\theta)$$

满足：$\hat{\theta} \xrightarrow{p} \theta_0$。

极大似然估计的一致性

- 注意以上定理中，我们要求$\ln f\left(w_i|\theta\right)$对$\theta$是连续的，而并没有要求得分函数$s_i\left(\theta\right)$是连续的，所以比基于$\mathbb{E}\left[s_i\left(\theta_0\right)\right]=0$的矩估计假设更宽松。
- 条件(4)结合Kullback-Leiber散度保证了θ_0是$\mathcal{L}\left(\theta\right)$的唯一最大值解，此外还要求我们的模型是正确设定的，也就是$f\left(w|\theta_0\right)$是w的真实的密度函数。
- 在这些假设的基础之上我们可以得到结论，极大似然估计是一致估计。

极大似然估计的极限分布

- 由于我们计算极大似然估计量时最大化了极大似然函数，其一阶条件为

$$\frac{\partial L\left(\hat{\theta}|\boldsymbol{x}\right)}{\partial\theta}=\frac{\partial}{\partial\theta}\sum_{i=1}^{N}\ln f\left(x_i|\hat{\theta}\right)=\sum_{i=1}^{N}\frac{\partial}{\partial\theta}\ln f\left(x_i|\hat{\theta}\right)=\sum_{i=1}^{N}s_i\left(\hat{\theta}\right)=0$$

即样本得分函数的均值等于0。

- 我们对上式在$\theta=\theta_0$处进行泰勒展开，得到：

$$0=\sum_{i=1}^{N}s_i\left(\hat{\theta}\right)=\sum_{i=1}^{N}s_i\left(\theta_0\right)+\sum_{i=1}^{N}\frac{\partial}{\partial\theta'}s_i\left(\theta_0\right)\left(\hat{\theta}-\theta_0\right)+O\left(\left(\hat{\theta}-\theta_0\right)^2\right)$$

- 我们记

$$H_i\left(\theta\right)=\frac{\partial}{\partial\theta'}s_i\left(\theta\right)=\frac{\partial}{\partial\theta\partial\theta'}\ln f\left(x_i|\theta\right)$$

为对数似然函数的海塞矩阵，我们有

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\theta'}s_i\left(\theta_0\right)\xrightarrow{p}\mathbb{E}H_i\left(\theta_0\right)\triangleq-H_0$$

极大似然估计的极限分布

- 而由于$\mathbb{E}\left[s_i\left(\theta_0\right)\right]=0$，因而

$$\mathbb{V}\left(s_i\left(\theta_0\right)\right)=\mathbb{E}\left[s_i\left(\theta_0\right)s_i'\left(\theta_0\right)\right]\triangleq\mathcal{I}_0$$

因而根据中心极限定理：

$$\sqrt{N}\frac{1}{N}\sum_{i=1}^{N}s_i\left(\theta_0\right)\overset{a}{\sim}\mathcal{N}\left(0,\mathcal{I}_0\right)$$

- 因而

$$\begin{aligned}\sqrt{N}\left(\hat{\theta}-\theta_0\right)&=H_0^{-1}\sqrt{N}\frac{1}{N}\sum_{i=1}^{N}s_i\left(\theta_0\right)+o_p\left(1\right)\\&\overset{a}{\sim}\mathcal{N}\left(0,H_0^{-1}\mathcal{I}_0H_0^{-1}\right)\end{aligned}$$

极大似然估计的极限分布

注意其中:

$$
\begin{aligned}
-H_0 = \mathbb{E}H_i(\theta_0) &= \int_{\mathbb{R}} \frac{\partial}{\partial\theta\partial\theta'} \ln f(x|\theta_0) \cdot f(x|\theta_0)\, dx \\
&= \int_{\mathbb{R}} \frac{\partial}{\partial\theta'} \left[\frac{1}{f(x|\theta_0)} \frac{\partial f(x|\theta_0)}{\partial\theta} \right] \cdot f(x|\theta_0)\, dx \\
&= \int_{\mathbb{R}} \left[\frac{1}{f(x|\theta_0)} \frac{\partial^2 f(x|\theta_0)}{\partial\theta\partial\theta'} - \frac{1}{f^2(x|\theta_0)} \frac{\partial f(x|\theta_0)}{\partial\theta} \frac{\partial f(x|\theta_0)}{\partial\theta'} \right] \cdot f(x|\theta_0)\, dx \\
&= \int_{\mathbb{R}} \frac{\partial^2 f(x|\theta_0)}{\partial\theta\partial\theta'} - \frac{1}{f(x|\theta_0)} \frac{\partial f(x|\theta_0)}{\partial\theta} \frac{\partial f(x|\theta_0)}{\partial\theta'} dx \\
&= \int_{\mathbb{R}} \frac{\partial^2 f(x|\theta_0)}{\partial\theta\partial\theta'} dx - \int_{\mathbb{R}} \left(\frac{1}{f(x|\theta_0)} \frac{\partial f(x|\theta_0)}{\partial\theta} \frac{\partial f(x|\theta_0)}{\partial\theta'} \right)^2 f(x|\theta_0)\, dx \\
&= \frac{\partial^2}{\partial\theta\partial\theta'} \int_{\mathbb{R}} f(x|\theta_0)\, dx - \mathbb{E}_{\theta_0} \frac{\partial \ln f(x|\theta_0)}{\partial\theta} \frac{\partial \ln f(x|\theta_0)}{\partial\theta'} \\
&= -\mathbb{E}_{\theta_0} \left[s_i(\theta_0)\, s_i(\theta_0)' \right] \\
&= -\mathbb{V}\left[s_i(\theta_0) \right] = -\mathcal{I}_0
\end{aligned}
$$

极大似然估计的极限分布

- 因而我们有：$H_0 = \mathcal{I}_0$，即对数似然函数海塞矩阵的期望等于得分函数的方差。
- 将以上等式带入，可以得到：

$$\sqrt{N}\left(\hat{\theta} - \theta_0\right) \overset{a}{\sim} \mathcal{N}\left(0, \mathcal{I}_0^{-1}\right)$$

即大样本条件下，极大似然估计量的极限分布为正态分布，且其渐近方差为对数似然函数海塞矩阵倒数的逆矩阵。

- 特别的，当θ为标量时，

$$\sqrt{N}\left(\hat{\theta} - \theta_0\right) \overset{a}{\sim} \mathcal{N}\left(0, -\frac{1}{\mathbb{E}\left(\frac{d^2}{d\theta^2} \ln f\left(x|\theta_0\right)\right)}\right)$$

极大似然估计的极限分布

伯努利分布极大似然估计的极限分布

- 在前例中，我们已经得到伯努利分布的极大似然估计为：$\hat{p} = \bar{x}$根据中心极限定理

$$\sqrt{N}\left(\hat{p} - p_0\right) \overset{a}{\sim} \mathcal{N}\left(0, p_0\left(1 - p_0\right)\right)$$

- 另一方面，直接使用以上结论也可以得到$\hat{p}$的极限分布。注意对数似然函数为：

$$L\left(p|\boldsymbol{x}\right) = \sum_{i=1}^{N}\left[x_i \ln p + \left(1 - x_i\right)\ln\left(1 - p\right)\right]$$

因而得分函数为：$s_i\left(p\right) = \frac{x_i}{p} - \frac{1-x_i}{1-p}$进而海塞矩阵（二阶导）：

$$-H_i\left(p\right) = -\frac{x_i}{p^2} - \frac{1 - x_i}{\left(1 - p\right)^2}$$

极大似然估计的极限分布

伯努利分布极大似然估计的极限分布

- 带入真值之后其期望为:

$$\mathcal{I}_0 = H_0 = \mathbb{E}H_i(p_0) = \frac{1}{p_0} + \frac{1}{1-p_0} = \frac{1}{p_0(1-p_0)}$$

从而$H_0^{-1} = p_0(1-p_0)$，从而

$$\sqrt{N}(\hat{p} - p) \overset{a}{\sim} \mathcal{N}(0, p_0(1-p_0))$$

Cramér-Rao下界

- 实际上，可以证明，以上的$\mathcal{I}_0^{-1}$是**渐近无偏估计量（asymptotically unbiased）**所能达到的最小方差
  - 渐近无偏：随着样本量趋向于正无穷，偏差趋向于0：$\lim_{N\to\infty}\mathbb{E}\left(\hat{\theta}-\theta\right)=0$
- 我们称方差为$\mathcal{I}_0^{-1}$的估计量为**渐近有效（asymptotic efficient）**估计量。
- 显然极大似然估计是一个渐近有效估计量。

Cramér-Rao下界

- 为了说明以上结果，现在假设存在一个无偏估计量$\tilde{\theta}(x)$，即$\mathbb{E}\left(\tilde{\theta}(x)\right)=\theta_0$，我们有：

$$\begin{aligned}\int_{\mathbb{R}}\tilde{\theta}(x)\,s(\theta_0)\,f(x|\theta_0)\,dx &= \int_{\mathbb{R}}\tilde{\theta}(x)\,\frac{\partial \ln f(x|\theta_0)}{\partial\theta}f(x|\theta_0)\,dx\\ &= \int_{\mathbb{R}}\tilde{\theta}(x)\,\frac{1}{f(x|\theta_0)}\frac{\partial f(x|\theta_0)}{\partial\theta}\cdot f(x|\theta_0)\,dx\\ &= \int_{\mathbb{R}}\tilde{\theta}(x)\,\frac{\partial f(x|\theta_0)}{\partial\theta}dx\\ &= \frac{\partial}{\partial\theta}\int_{\mathbb{R}}\tilde{\theta}(x)\,f(x|\theta_0)\,dx\\ &= \frac{\partial}{\partial\theta}\mathbb{E}\left(\tilde{\theta}(x)\right)\\ &= \frac{\partial}{\partial\theta}\theta_0=1\end{aligned}$$

Cramér-Rao下界

- 且由于$\int_{\mathbb{R}} s\left(\theta_0\right) \cdot f\left(x|\theta_0\right) dx = 0$从而

$$\begin{aligned}\int_{\mathbb{R}} \tilde{\theta}\left(x\right) s\left(\theta_0\right) \cdot f\left(x|\theta_0\right) dx - \theta_0 \int_{\mathbb{R}} s\left(\theta_0\right) \cdot f\left(x|\theta_0\right) dx \\ = \int_{\mathbb{R}} \left[\tilde{\theta}\left(x\right) - \theta_0\right] s\left(\theta_0\right) \cdot f\left(x|\theta_0\right) dx = 1\end{aligned}$$

- 根据Cauthy-Schwarz不等式，有

$$\begin{aligned}1 &= \left[\int_{\mathbb{R}} \left[\tilde{\theta}\left(x\right) - \theta_0\right] s\left(\theta_0\right) \cdot f\left(x|\theta_0\right) dx\right]^2 \\ &\leq \left[\int_{\mathbb{R}} \left[\tilde{\theta}\left(x\right) - \theta_0\right]^2 f\left(x|\theta_0\right) dx\right] \left[\int_{\mathbb{R}} s\left(\theta_0\right) s\left(\theta_0\right)' \cdot f\left(x|\theta_0\right) dx\right] \\ &= \mathbb{V}\left(\tilde{\theta}\left(x\right)\right) \mathbb{E}\left(s\left(\theta_0\right) s\left(\theta_0\right)'\right) \\ &= \mathbb{V}\left(\tilde{\theta}\left(x\right)\right) \cdot \mathcal{I}_0\end{aligned}$$

Cramér-Rao下界

- 最终我们得到了**Cramér-Rao下界**：

$$\mathbb{V}\left(\tilde{\theta}\left(x\right)\right) \succeq \mathcal{I}_0^{-1}$$

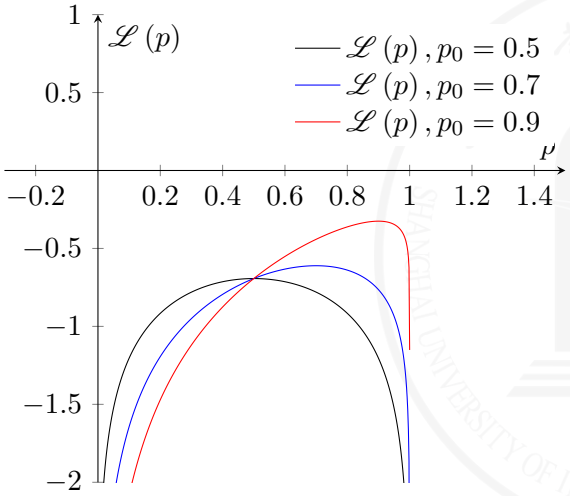
- 基于此，$\mathcal{I}_0$实际上度量了数据中所包含的「信息量」的大小，因而我们称$\mathcal{I}_0$为**费雪信息矩阵（Fisher information matrix）**。
  - $\mathcal{I}_0$越大，意味着所包含的信息越多，极大似然估计所得到的方差也越小，（渐近）无偏估计量所能达到的最小方差也越小。

费雪信息矩阵

伯努利分布的信息矩阵

- 如图所示，当$p_0 = 0.5$时，信息矩阵$\mathcal{I}_0$达到了最小值，而$\hat{p}$的方差达到了最大值，表现在图中即对数似然函数在真值p_0处非常平缓。
- 当真值p_0逐渐接近0或者1时，信息矩阵$\mathcal{I}_0$逐渐变大，$\hat{p}$的方差也逐渐变小，图中对数似然函数在真值p_0处也更加尖锐。
- 因而同样是伯努利分布，真值越接近于0或者1的伯努利分布实际上携带了更多的信息。

费雪信息矩阵


1
0.5
$\mathscr{L}(p)$
$\mathscr{L}(p), p_0 = 0.5$
$\mathscr{L}(p), p_0 = 0.7$
$\mathscr{L}(p), p_0 = 0.9$
p
−0.2
0.2
0.4
0.6
0.8
1
1.2
1.4
−0.5
−1
−1.5
−2

条件极大似估计

- 以上介绍了极大似然估计法，需要设定数据x的完整的分布情况才能得到估计。
- 然而很多时候，我们观察到一系列数据$w_i \in \mathbb{R}^k, i = 1, ..., N$，其中$w_i = [y_i', x_i']', y_i \in \mathbb{R}^{k_1}, x_i \in \mathbb{R}^{k_2}$，很多时候我们仅仅希望研究$x$和$y$之间的关系，而不关心随机向量$x$自身的分布，如果使用极大似然估计，我们就必须设定$x$的联合分布。
- 然而，设定x的联合分布很多时候是多余的，实际上，如果我们能够找到y给定x的条件分布，即$f(y|x, \theta)$，那么基于条件分布的极大似然估计：

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{N} \ln f(y_i|x_i, \theta)$$

仍然能够得到参数θ_0的一致估计。

条件极大似估计

条件极大似然估计的应用

如果$x_i|d_i \sim \mathcal{N}(\mu, 1+d_i), d_i \sim \mathcal{B}er(p), w_i = [x_i, d_i]'$，这个模型中有两个未知参数$\mu, p$，如果我们能够观察到$x_i$和$d_i$，且我们只关心参数$\mu$，那么我们仅仅需要最大化条件对数似然函数

$$\begin{aligned} L(\mu|\boldsymbol{w}) &= \sum_{i=1}^{N} \ln f(x_i|d_i, \mu) \\ &= \sum_{i=1}^{N} \ln \left[\frac{1}{\sqrt{2\pi}\sqrt{1+d_i}} \exp\left\{ -\frac{(x_i-\mu)^2}{1+d_i} \right\} \right] \\ &= -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(1+d_i) - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{1+d_i} \end{aligned}$$

从而得到$\hat{\mu} = \bar{x}$。

条件极大似估计

条件极大似然估计的应用

如果我们对两个未知参数μ, p都感兴趣，那么：

$$f\left(x_i, d_i|\mu, p\right) = f\left(x_i|d_i, \mu, p\right) f\left(d_i|\mu, p\right)$$

注意其中$f\left(x_i|d_i, \mu, p\right) = f\left(x_i|d_i, \mu\right)$与$p$无关，$f\left(d_i|\mu, p\right) = f\left(d_i|p\right)$与$\mu$无关，从而最大化无条件的对数似然函数：

$$\max_{\mu,p} \left[\sum_{i=1}^{N} \ln f\left(x_i|d_i, \mu\right) + \sum_{i=1}^{N} \ln f\left(d_i|p\right)\right]$$

与直接最大化：

$$\max_{\mu} \sum_{i=1}^{N} \ln f\left(x_i|d_i, \mu\right)$$

条件极大似估计

线性回归

- 如果$w_i = [y_i, x_i']', i = 1, ..., N, x_i \in \mathbb{R}^K$为一系列独立同分布的随机向量。为了使用$x_i$预测或者拟合$y_i$，我们可以假设$y_i|x_i \sim \mathcal{N}\left(x_i'\beta_0, \sigma^2\right)$，即给定$x$，$y$服从正态分布。
- 或者等价的，以上模型也可以写成：

$$y_i = x_i'\beta_0 + u_i$$

其中$u_i|x_i \sim \mathcal{N}\left(0, \sigma^2\right)$。

条件极大似估计

线性回归

- 在上述条件下，条件密度函数为：

$$f\left(y_i|x_i,\beta\right)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{\left(y_i-x_i'\beta\right)^2}{2\sigma^2}\right\}$$

因而似然函数为：

$$L\left(\beta|\boldsymbol{w}\right)=-\frac{N}{2}\ln\left(2\pi\right)-N\ln\sigma-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i-x_i'\beta\right)^2$$

如果对β求导，可以得到：

$$-\frac{1}{\sigma^2}\sum_{i=1}^{N}\left(y_i-x_i'\hat{\beta}\right)x_i=0$$

条件极大似估计

线性回归

- 解得:

$$\hat{\beta} = \left[\sum_{i=1}^{N} \left(x_i x_i'\right)\right]^{-1} \sum_{i=1}^{N} \left(x_i y_i\right) = \left(X'X\right)^{-1} X'Y$$

  我们再次得到了**普通最小二乘（ordinary least squares, OLS）**估计。

- 注意在以上设定中，虽然我们得到了与矩估计相同的结果，但是极大似然估计的假设是更强的，因为矩估计中我们只使用了假设$\mathbb{E}\left(y_i|x_i\right) = x_i'\beta_0$，而这里不仅仅假设了条件期望，还假设了条件分布：$y_i|x_i \sim \mathcal{N}\left(x_i'\beta_0, \sigma^2\right)$，因而假设更强。

Stata中的极大似然估计

- 同样，在Stata中，也可以进行极大似然估计。我们只需要向Stata提供极大似然估计的目标函数即可，必要时也可以提供对数似然函数一阶导、二阶导的信息以供Stata能够更好的进行估计计算。
- Stata通过ml命令进行极大似然估计
  - 为了使用该命令，需要首先写一个program以定义对数似然函数，ml命令允许不同的“model”，这其中有的需要提供一阶导数，有的需要提供二阶导数。
  - 此外，还需要指定极大似然估计中需要估计的参数以及数据。

Stata中的极大似然估计

- MLE_censor.do展示了截尾数据例子中的极大似然估计
- 其中：
  - mle_obj_censor为极大似然估计的目标函数
  - 而mle_censor将极大似然估计包装为一个program
  - “ml model”命令定义了极大似然问题的model以及需要的参数和数据，为了方便起见，我们选择“lf”这个model，在这个model中我们需要要计算出$\ln f(w_i|\theta)$并放在变量‘lnfj’中，Stata就可以自动对该变量求和，并求最大化从而找到极大似然估计；
  - “ml search”用于寻找最优化的初始值；
  - “ml maximize”即寻找能使得目标函数的最大化的参数。
- 可以发现最终估计结果与真实值相差无几。

作业

- 11.1
- 11.3
- 11.4
- 11.5