

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

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

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

- 现在，将未知参数 θ 视为变量， \mathbf{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)：

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

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

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

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 $x_i \sim \text{Ber}(p_0)$ i.i.d, 那么其联合密度函数为:

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

对数似然函数为:

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

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：

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

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 $x_i \sim N(\mu_0, \sigma_0^2)$ *i.i.d.*, 其中 $\theta = (\mu, \sigma^2)'$, 那么其联合密度函数为

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

对数似然函数为:

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

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：

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

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 σ_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 $x_i \sim U(0, \beta)$ *i.i.d.*，那么密度函数为

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

从而

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

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

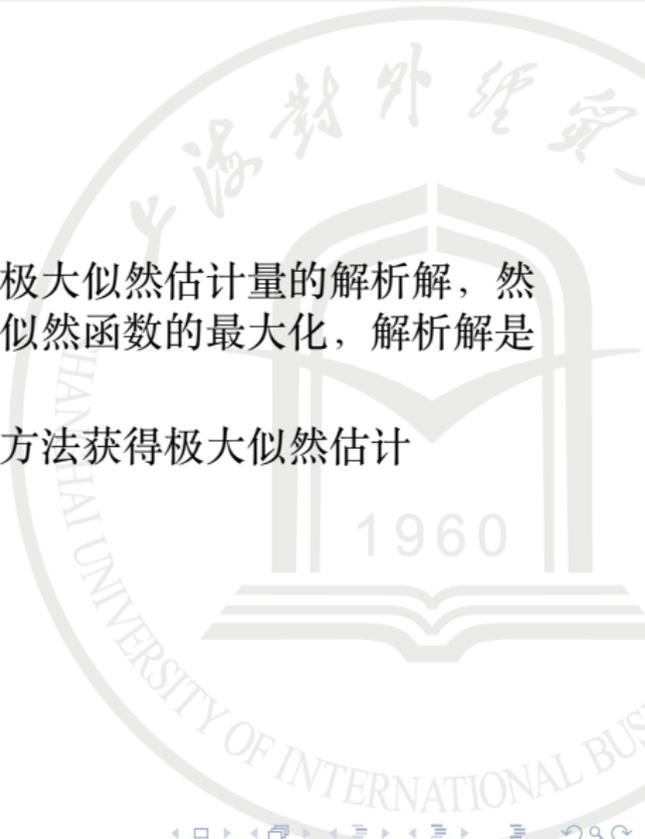
$$L(\beta|\mathbf{x}) = \sum_{i=1}^N \ln f(x_i|\beta) = \begin{cases} -N \ln \beta & \max_i \{x\} \leq \beta \\ -\infty & \text{else} \end{cases}$$

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 $\hat{\beta} \geq \max_i \{x\}$ ；
- 给定 $\hat{\beta} \geq \max_i \{x\}$ ， $-N \ln \beta$ 是一个单调递减的函数
- 从而极大似然估计量即满足 $\hat{\beta} \geq \max_i \{x\}$ 最小的数，即 $\hat{\beta} = \max_i \{x\}$ 。

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 $x_i \sim \text{Beta}(\alpha, \beta)$ *i.i.d.*, 那么

$$\ln f(x_i|\alpha, \beta) = -\ln B(\alpha, \beta) + (\alpha - 1) \ln x_i + (\beta - 1) \ln(1 - x_i)$$

从而对数似然函数为

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

其中

$$\ln B(\alpha, \beta) = \ln \Gamma(\alpha) + \ln \Gamma(\beta) - \ln \Gamma(\alpha + \beta)$$

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

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

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，lnf即为需要返回的变量 $\ln f(x_i|\alpha, \beta)$ ，Stata会根据这一变量自动计算 $L(\alpha, \beta|\mathbf{x}) = \sum_{i=1}^N \ln f(x_i|\alpha, \beta)$
- 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；`lnf`为变量 $\ln f(x_i|\alpha, \beta)$ ，而`g1`、`g2`为对数似然函数对 α 、 β 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 α 、 β 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 α 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 $x_i^* \sim N(\mu, \sigma^2)$ *i.i.d.*，那么我们观察到的数据为：

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

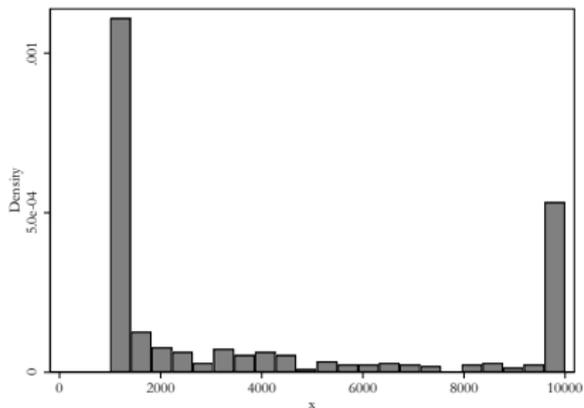
我们称数据存在截尾 (censoring) 现象。

极大似然估计：截尾数据

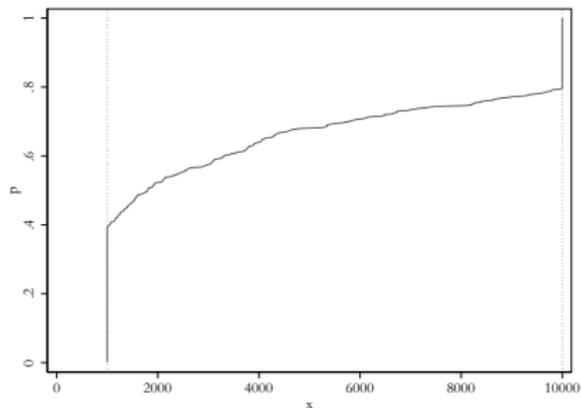
截尾数据

- `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

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

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

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

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

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

极大似然估计：截尾数据

截尾数据

因而其对数似然函数为

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

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

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

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：

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

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 $\frac{1}{N} L(\theta|\mathbf{w})$ 一致收敛到 $\mathcal{L}(\theta)$ ，样本似然函数 $L(\theta|\mathbf{w})$ （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 $L(\theta|\mathbf{w})$ 获得估计，即

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

- 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

- 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

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

- 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 $\arg \max_{\theta} L(\theta | \mathbf{w})$ 。

- 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 $\mathcal{L}(\theta) = \mathbb{E} \ln f(w_i | \theta)$ 呢？

极大似然估计的一致性

伯努利分布的总体似然函数

若 $x_i \sim \text{Ber}(p_0)$ *i.i.d.*, 那么其联合密度函数为:

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

对数似然函数为:

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

根据大数定律, 对于任意的 p , 上述似然函数

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

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 $\frac{\partial}{\partial \theta} \ln f(w_i|\theta)$ 为得分函数 (score function)，记为

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

- 以上结论意味着得分函数在真值处的期望等于0，即 $\mathbb{E}s_i(\theta_0) = 0$ 。
- 至此，如果得分函数 $s_i(\theta)$ 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 θ_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：

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

- 注意到根据Jensen不等式，有

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

以上不等式当且仅当 $f(w|\theta) = f(w|\theta_0)$ 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：

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

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

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

极大似然估计的一致性

极大似然估计的一致性

如果 $\{w_i, i = 1, \dots, N\}$ 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 Θ 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 Θ 上对 θ 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 $\ln f(w_i|\theta)$ 对 θ 是连续的，而并没有要求得分函数 $s_i(\theta)$ 是连续的，所以比基于 $\mathbb{E}s_i(\theta_0) = 0$ 的矩估计假设更宽松。
- 条件(4)结合Kullback-Leiber散度保证了 θ_0 是 $\mathcal{L}(\theta)$ 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 $f(w|\theta_0)$ 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

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

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：

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

- 我们记

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

为对数似然函数的海塞矩阵，我们有

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

极大似然估计的极限分布

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

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

因而根据中心极限定理：

$$\sqrt{N} \frac{1}{N} \sum_{i=1}^N s_i(\theta_0) \stackrel{a}{\approx} N(0, \mathcal{I}_0)$$

- 因而

$$\begin{aligned} \sqrt{N}(\hat{\theta} - \theta_0) &= H_0^{-1} \sqrt{N} \frac{1}{N} \sum_{i=1}^N s_i(\theta_0) + o_p(1) \\ &\stackrel{a}{\approx} N(0, H_0^{-1} \mathcal{I}_0 H_0^{-1}) \end{aligned}$$

极大似然估计的极限分布

注意其中：

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

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：

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

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，

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

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， $\hat{p} = \bar{x} \xrightarrow{D} N\left(p_0, \frac{p_0(1-p_0)}{N}\right)$ 。
- 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：

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

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

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

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

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

。

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： $\lim_{N \rightarrow \infty} \mathbb{E}(\hat{\theta} - \theta) = 0$
- 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 $\mathbb{E}(\tilde{\theta}(x)) = \theta_0$ ，我们有：

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

Cramér-Rao下界

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

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

- 根据Cauchy-Schwarz不等式, 有

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

- 最终我们得到了Cramér-Rao下界:

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

费雪信息矩阵

- 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。

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

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

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：

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

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

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

与直接最大化：

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

条件极大似估计

线性回归

- 如果 $(y_i, x_i)'$, $i = 1, \dots, N$, $x_i \in \mathbb{R}^K$ 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 $y_i|x_i \sim N(x_i'\beta_0, \sigma^2)$, 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 $u_i|x_i \sim N(0, \sigma^2)$ 。

条件极大似估计

线性回归

- 在上述条件下，条件密度函数为：

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

因而似然函数为：

$$L(\beta|y, x) = -\frac{N}{2} \ln(2\pi) - N \ln \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - x_i'\beta)^2$$

如果对 β 求导，可以得到：

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

条件极大似估计

线性回归

- 解得：

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

我们再次得到了最小二乘估计 (Ordinary least squares, OLS)。

- 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 $\mathbb{E}(y_i | x_i) = x_i' \beta_0$ ，而这里不仅仅假设了条件期望，还假设了条件分布： $y_i | x_i \sim N(x_i' \beta_0, \sigma^2)$ ，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。