

第 6 章 最大似然估计

如果回归模型存在非线性, 常使用最大似然估计法(MLE)。

6.1 最大似然估计法的定义

假设随机向量 \mathbf{y} 的概率密度函数为 $f(\mathbf{y}; \boldsymbol{\theta})$, 其中 $\boldsymbol{\theta}$ 为 K 维未知参数向量, $\boldsymbol{\theta} \in \Theta$ 。

Θ 为参数空间, 即参数 $\boldsymbol{\theta}$ 所有可能取值所构成的集合。

通过抽取随机样本 $\{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ 来估计 $\boldsymbol{\theta}$ 。

假设 $\{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ 为 iid，则样本数据的联合密度函数为 $f(\mathbf{y}_1; \boldsymbol{\theta}) f(\mathbf{y}_2; \boldsymbol{\theta}) \cdots f(\mathbf{y}_n; \boldsymbol{\theta})$ 。

在抽样前， $\{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ 为随机向量。

抽样后， $\{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ 有了特定的样本值，可将样本联合密度函数视为在 $\{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ 给定情况下，未知参数 $\boldsymbol{\theta}$ 的函数。

定义似然函数(likelihood function)为

$$L(\boldsymbol{\theta}; \mathbf{y}_1, \cdots, \mathbf{y}_n) = \prod_{i=1}^n f(\mathbf{y}_i; \boldsymbol{\theta})$$

似然函数与联合密度函数完全相等，只是 $\boldsymbol{\theta}$ 与 $\{\mathbf{y}_1, \cdots, \mathbf{y}_n\}$ 的角色互换，即把 $\boldsymbol{\theta}$ 作为自变量，而视 $\{\mathbf{y}_1, \cdots, \mathbf{y}_n\}$ 为给定。

为了运算方便，常把似然函数取对数：

$$\ln L(\boldsymbol{\theta}; \mathbf{y}_1, \cdots, \mathbf{y}_n) = \sum_{i=1}^n \ln f(\mathbf{y}_i; \boldsymbol{\theta})$$

“最大似然估计法”(Maximum Likelihood Estimation, 简记 MLE 或 ML)的思想: 给定样本取值后, 该样本最有可能来自参数 θ 为何值的总体。

寻找 $\hat{\theta}_{\text{ML}}$, 使得观测到样本数据的可能性最大, 即最大化对数似然函数(loglikelihood function):

$$\max_{\theta \in \Theta} \ln L(\theta; y)$$

最大似然估计量 $\hat{\theta}_{\text{ML}}$ 可写为,

$$\hat{\theta}_{\text{ML}} \equiv \arg \max \ln L(\theta; y)$$

其中，“argmax” (argument of the maximum)表示能使 $\ln L(\boldsymbol{\theta}; \mathbf{y})$ 最大化的 $\boldsymbol{\theta}$ 取值。

假设存在唯一内点解，一阶条件：

$$\mathbf{s}(\boldsymbol{\theta}; \mathbf{y}) \equiv \frac{\partial \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}} \equiv \begin{pmatrix} \frac{\partial \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \theta_K} \end{pmatrix} = \mathbf{0}$$

一阶条件要求，对数似然函数的梯度向量 (gradient, 偏导数、斜率) $s(\boldsymbol{\theta}; \mathbf{y})$ 为 $\mathbf{0}$ ，实际上是 K 个未知参数 $(\theta_1 \theta_2 \cdots \theta_K)$ ， K 个方程的方程组。

该向量也称“得分函数” (score function) 或“得分向量” (score vector)。

得分函数 $s(\boldsymbol{\theta}; \mathbf{y})$ 是 \mathbf{y} 的函数，也是随机向量。

在下面，记真实参数为 $\boldsymbol{\theta}_0$ ，而 $\boldsymbol{\theta}$ 为该参数的任何可能取值。

命题(得分函数的期望为 0)

如果似然函数正确(correctly specified), 则

$$E[s(\boldsymbol{\theta}_0; \boldsymbol{y})] = 0$$

其中, $s(\boldsymbol{\theta}_0; \boldsymbol{y})$ 表示得分函数 $s(\boldsymbol{\theta}; \boldsymbol{y})$ 在 $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ 处的取值。

例 假设随机样本 $y_i \sim N(\theta_0, 1)$, $i = 1, \dots, n$ 。则样本数据的对数似然函数为,

$$L(\theta) = -\frac{n}{2} \ln 2\pi - \frac{1}{2} \sum_{i=1}^n (y_i - \theta)^2$$

其得分函数为,

$$s(\theta) = \frac{\partial L(\theta)}{\partial \theta} = \sum_{i=1}^n (y_i - \theta)$$

故得分函数的期望值为,

$$\mathbb{E}[s(\theta)] = \sum_{i=1}^n [\mathbb{E}(y_i) - \theta] = \sum_{i=1}^n [\theta_0 - \theta]$$

在 $\theta = \theta_0$ 处,

$$\mathbb{E}[s(\theta)] \Big|_{\theta=\theta_0} = \sum_{i=1}^n [\theta_0 - \theta] \Big|_{\theta=\theta_0} = 0$$

此结果与上述命题一致。

得分函数可分解为

$$\begin{aligned} s(\boldsymbol{\theta}; \mathbf{y}) &\equiv \frac{\partial \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}} = \frac{\partial \sum_{i=1}^n \ln f(\mathbf{y}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ &= \sum_{i=1}^n \frac{\partial \ln f(\mathbf{y}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \equiv \sum_{i=1}^n \mathbf{s}_i(\boldsymbol{\theta}; \mathbf{y}_i) \end{aligned}$$

其中, $s_i(\boldsymbol{\theta}; y_i) \equiv \frac{\partial \ln f(y_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ 为第 i 个观测值对得分函数的贡献。

在上例中, $s(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i=1}^n (y_i - \theta)$, 而 $s_i(\boldsymbol{\theta}, y_i) = (y_i - \theta)$ 。

二阶条件要求, 对数似然函数的黑赛矩阵(Hessian matrix)

$$\frac{\partial^2 \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \equiv \frac{\partial \left(\frac{\partial \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}} \right)}{\partial \boldsymbol{\theta}'}$$

为负定矩阵，即对数似然函数为严格凹函数(strictly concave)。黑赛矩阵可分解为

$$\mathbf{H}(\boldsymbol{\theta}; \mathbf{y}) \equiv \frac{\partial^2 \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \sum_{i=1}^n \frac{\partial^2 \ln L(\boldsymbol{\theta}; \mathbf{y}_i)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \equiv \sum_{i=1}^n \mathbf{H}_i(\boldsymbol{\theta}; \mathbf{y}_i)$$

其中， $\mathbf{H}_i(\boldsymbol{\theta}; \mathbf{y}_i)$ 为第 i 个观测值对黑赛矩阵的贡献。

在上例中， $\mathbf{H}(\boldsymbol{\theta}; \mathbf{y}) = \frac{\partial s(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial \sum_{i=1}^n (y_i - \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -n$ ，而

$$\mathbf{H}_i(\boldsymbol{\theta}; \mathbf{y}_i) = \frac{\partial s_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial (y_i - \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -1。$$

6.2 线性回归模型的最大似然估计

例 假设 $X \sim N(\mu, \sigma^2)$ ，其中 σ^2 已知，得到一个样本容量为 1 的样本 $x_1 = 2$ ，求对 μ 的最大似然估计。

$$\text{似然函数为 } L(\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(2-\mu)^2}{2\sigma^2}\right\}。$$

似然函数在 $\hat{\mu} = 2$ 处取最大值，见图 6.1。

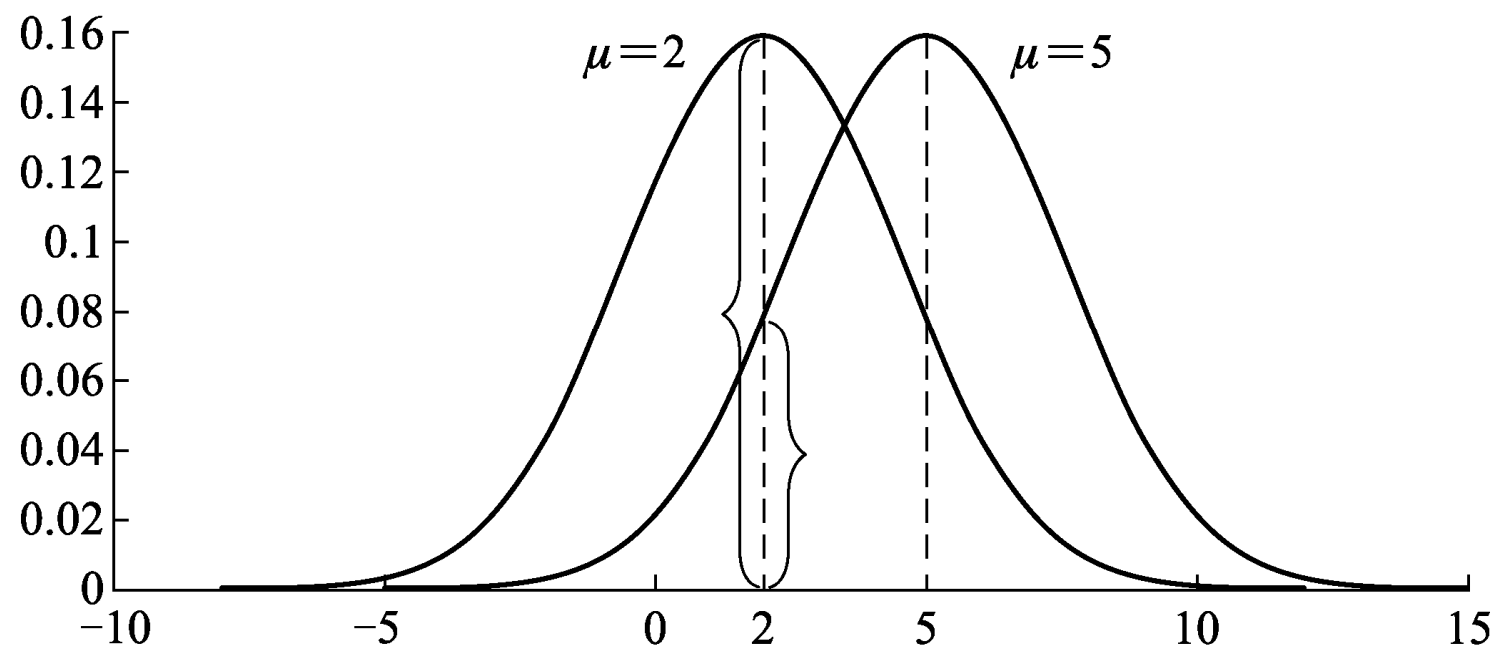


图 6.1 选择参数使观测到样本的可能性最大

例(非正式) 某人操一口浓重的四川口音，则判断他最有可能来自四川。

考虑线性回归模型：

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

假设 $\boldsymbol{\varepsilon} | \mathbf{X} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ ，则 $\mathbf{y} | \mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$ ，条件密度函数为：

$$f(\mathbf{y} | \mathbf{X}) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}$$

用假想 $\tilde{\beta}, \tilde{\sigma}^2$ 代替真实 β, σ^2 ，取对数可得

$$\ln L(\tilde{\beta}, \tilde{\sigma}^2) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \tilde{\sigma}^2 - \frac{1}{2\tilde{\sigma}^2} (\mathbf{y} - \mathbf{X}\tilde{\beta})'(\mathbf{y} - \mathbf{X}\tilde{\beta})$$

分两步最大化。第一步，给定 $\tilde{\sigma}^2$ ，选择最优 $\tilde{\beta}$ 。第二步，代入第一步的最优 $\tilde{\beta}$ ，选择最优 $\tilde{\sigma}^2$ 。

第一步。由于 $\tilde{\beta}$ 只出现在第三项中，故等价于使 $(\mathbf{y} - \mathbf{X}\tilde{\beta})'(\mathbf{y} - \mathbf{X}\tilde{\beta})$ 最小，正是 OLS 的目标函数 $\mathbf{e}'\mathbf{e}$ 。

$$\hat{\boldsymbol{\beta}}_{\text{ML}} = \hat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$$

第二步。对数似然函数变为 $-\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \tilde{\sigma}^2 - \frac{1}{2\tilde{\sigma}^2} \mathbf{e}'\mathbf{e}$ ，称为“集中对数似然函数”(concentrated log likelihood function)，因为 $\tilde{\boldsymbol{\beta}}$ 的取值已在第一步固定，称为“concentrated with respect to $\tilde{\boldsymbol{\beta}}$ ”。对 $\tilde{\sigma}^2$ 求导可得

$$-\frac{n}{2} \frac{1}{\tilde{\sigma}^2} + \frac{1}{2\tilde{\sigma}^4} \mathbf{e}'\mathbf{e} = 0$$

求解 σ^2 的 MLE 估计量为

$$\hat{\sigma}_{\text{ML}}^2 = \frac{\mathbf{e}'\mathbf{e}}{n} \neq \hat{\sigma}_{\text{OLS}}^2 = \frac{\mathbf{e}'\mathbf{e}}{n-K} \equiv s^2$$

MLE 对 $\boldsymbol{\beta}$ 的估计与 OLS 一样，但对 σ^2 的估计略有不同，此差别在大样本下消失。

由于 OLS 估计量 s^2 是对 σ^2 的无偏估计，故 MLE 估计量 $\hat{\sigma}_{\text{ML}}^2$ 是有偏的(小样本性质)。

MLE 的主要优点是大样本性质良好，比如一致性、最小渐近方差。

6.3 最大似然估计的数值解

最大似然估计通常没有解析解，只能寻找“数值解”(numerical solution)。

方法一为“网格搜索”(grid search)：

如果待估参数 θ 为一维，且大致知道取值范围，比如 $\theta \in (0, 1)$ 。

如果待估参数 θ 为多维，或对 θ 的取值范围所知不多，网格搜索不现实。

方法二为“高斯-牛顿法” (Gauss-Newton method)。

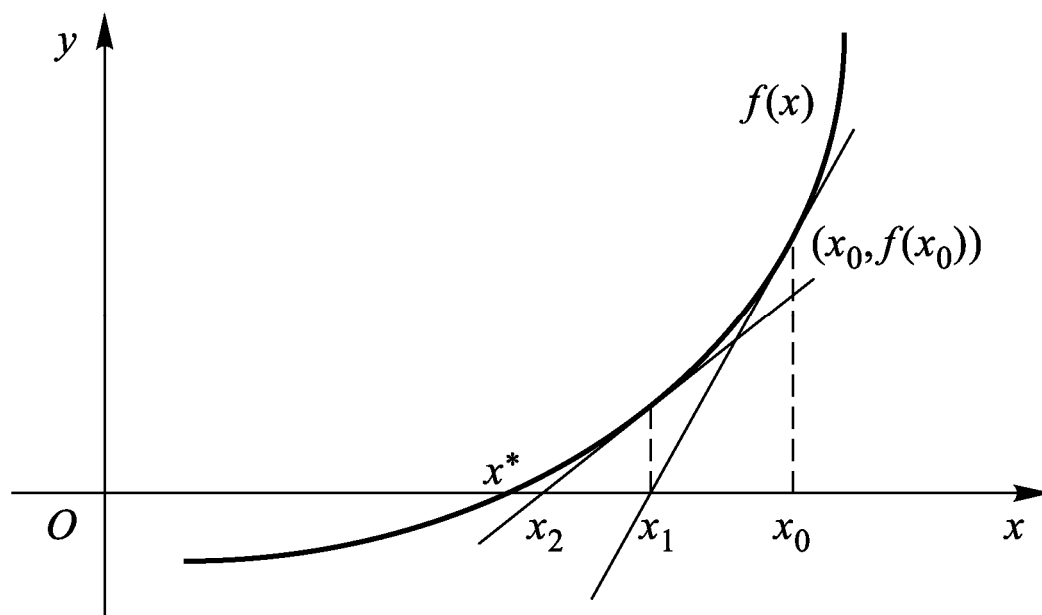


图 6.2 牛顿法

牛顿法收敛很快，是二次的。比如，如果本次迭代的误差为 0.1，则下次迭代的误差约为 0.1^2 。

如果初始值 x_0 选择不当，可能出现迭代不收敛的情形。

使用牛顿法得到的可能只是“局部最大值”(local maximum)，而非“整体最大值”(global maximum)。

牛顿法也适用于多元函数的情形 $f(\mathbf{x}) = 0$ ，将切线替换为(超)切平面即可。

如对原函数 $f(x)$ 作二阶近似(二阶泰勒展开)，称为“牛顿-拉夫森法”(Newton-Raphson method)。

6.4 信息矩阵与无偏估计的最小方差

定义信息矩阵(information matrix)为对数似然函数的黑塞矩阵之期望值(对 \mathbf{y} 求期望)的负数,

$$\mathbf{I}(\boldsymbol{\theta}) \equiv -\mathbf{E} \left[\frac{\partial^2 \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right]$$

一维情形下, $-\frac{\partial^2 \ln L}{\partial \theta^2}$ 为对数似然函数的二阶导数之负数。

对数似然函数为凹函数, 故二阶导数为负数, 加负号为正数。

$-\frac{\partial^2 \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$ 表示对数似然函数在 $\boldsymbol{\theta}$ 空间中的曲率 (curvature)，取期望值之后为平均曲率(对 \mathbf{y} 进行平均)。

如果曲率大，对数似然函数陡峭，较易根据样本分辨真实 $\boldsymbol{\theta}$ 的位置；反之，如果曲率小，对数似然函数平坦，不易根据样本判断真实 $\boldsymbol{\theta}$ 的位置。

如果似然函数完全平坦，则似然函数不存在唯一最大值，MLE 没有唯一解；则无法根据样本数据来判断 $\boldsymbol{\theta}$ 的位置。

$I(\boldsymbol{\theta})$ 包含了 $\boldsymbol{\theta}$ 是否容易估计的信息，故称“信息矩阵”。

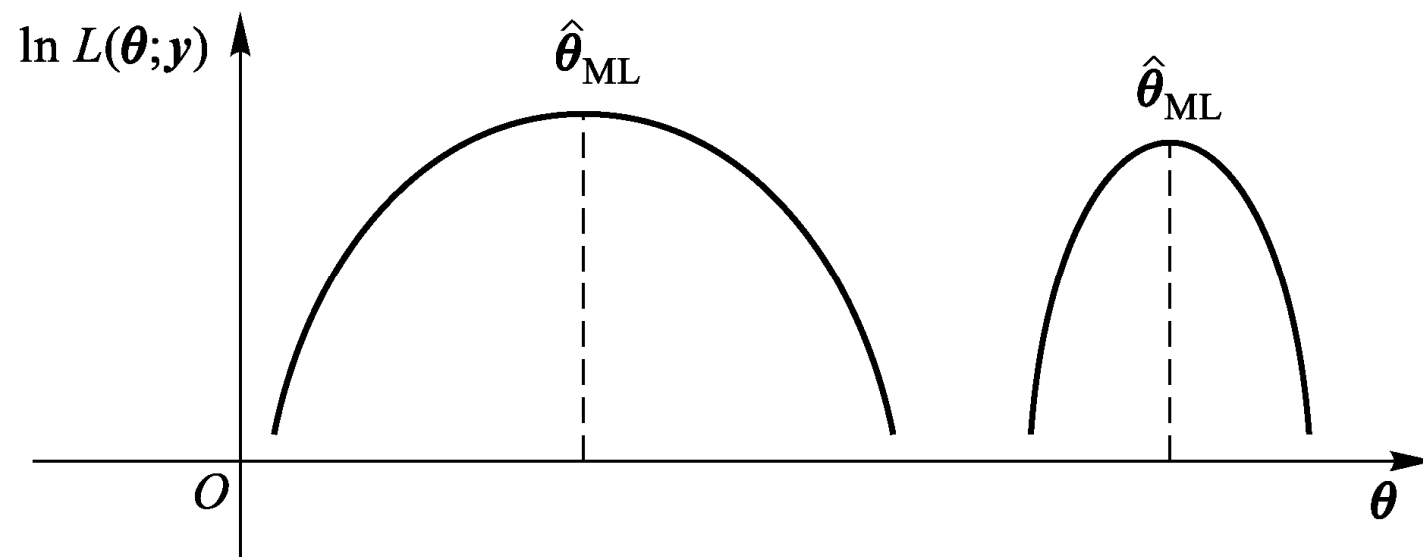


图 6.3 平坦(左)与陡峭(右)的对数似然函数

命题 (信息矩阵等式)

在 $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ 处，以下“信息矩阵等式”(information matrix equality)成立，

$$\begin{aligned} I(\boldsymbol{\theta}_0) &\equiv -\mathbf{E}\left[\frac{\partial^2 \ln L(\boldsymbol{\theta}_0; \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right] \\ &= \mathbf{E}\left[\frac{\partial \ln L(\boldsymbol{\theta}_0; \mathbf{y})}{\partial \boldsymbol{\theta}} \frac{\partial \ln L(\boldsymbol{\theta}_0; \mathbf{y})}{\partial \boldsymbol{\theta}'}\right] = \mathbf{E}[s(\boldsymbol{\theta}_0; \mathbf{y})s(\boldsymbol{\theta}_0; \mathbf{y})'] \end{aligned}$$

证明参见附录。

命题 (得分函数的方差为信息矩阵)

在 $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ 处，信息矩阵 $\mathbf{I}(\boldsymbol{\theta}_0)$ 就是得分函数的协方差矩阵 $\text{Var}[s(\boldsymbol{\theta}_0; \mathbf{y})]$ 。

证明：

$$\begin{aligned}\text{Var}[s(\boldsymbol{\theta}_0; \mathbf{y})] &= \text{E}[s(\boldsymbol{\theta}_0; \mathbf{y})s(\boldsymbol{\theta}_0; \mathbf{y})'] - \underbrace{\text{E}[s(\boldsymbol{\theta}_0; \mathbf{y})]}_{=0} \underbrace{\text{E}[s(\boldsymbol{\theta}_0; \mathbf{y})]'}_{=0} \\ &= \text{E}[s(\boldsymbol{\theta}_0; \mathbf{y})s(\boldsymbol{\theta}_0; \mathbf{y})'] \\ &= \mathbf{I}(\boldsymbol{\theta}_0)\end{aligned}$$

最后一步用到了信息矩阵等式。

假设 $\hat{\boldsymbol{\theta}}$ 是对真实参数 $\boldsymbol{\theta}_0$ 的任意无偏估计，则在一定的正则条件(regularity conditions)下， $\hat{\boldsymbol{\theta}}$ 的方差不会小于 $[\mathbf{I}(\boldsymbol{\theta}_0)]^{-1}$ ，即 $\text{Var}(\hat{\boldsymbol{\theta}}) \geq [\mathbf{I}(\boldsymbol{\theta}_0)]^{-1}$ 。

称 $[\mathbf{I}(\boldsymbol{\theta}_0)]^{-1}$ 为“克莱默-劳下限”(Cramer-Rao Lower Bound)。

无偏估计所能达到的最小方差与信息矩阵有关。曲率 $\mathbf{I}(\boldsymbol{\theta}_0)$ 越大，则 $[\mathbf{I}(\boldsymbol{\theta}_0)]^{-1}$ 越小，无偏估计可能达到的最小方差越小。

在古典线性回归模型中，可证明(参见附录)

$$[I(\boldsymbol{\theta}_0)]^{-1} = \begin{pmatrix} \sigma^2(\mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0} & 2\sigma^4/n \end{pmatrix}$$

其中， $\boldsymbol{\theta}_0 = (\boldsymbol{\beta} \ \sigma^2)'$ 。由于 $\text{Var}(\hat{\boldsymbol{\beta}}_{\text{OLS}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ ，故 $\hat{\boldsymbol{\beta}}_{\text{ML}} = \hat{\boldsymbol{\beta}}_{\text{OLS}}$ 均达到了无偏估计的最小方差。

命题 在高斯-马尔可夫定理中，如果加上扰动项为正态分布的假定，则 OLS 是“最佳无偏估计”(Best Unbiased Estimator, 简记 BUE)，而不仅仅是 BLUE。

克莱默-劳下限的结论可推广到渐近分布的情形。

在一定的正则条件下，对于真实参数 θ_0 的渐近正态一致估计 (Consistent and Asymptotically Normally distributed estimators, 简记 CAN)所能达到的最小方差为 $[I(\theta_0)]^{-1}$ ，即克莱默-劳下限。

6.5 最大似然法的大样本性质

定理(MLE 的大样本性质) 在一定的正则条件下，MLE 估计量拥有以下良好的大样本性质。

(1) 一致性, 即 $\text{plim}_{n \rightarrow \infty} \hat{\boldsymbol{\theta}}_{\text{ML}} = \boldsymbol{\theta}_0$ 。

(2) 渐近有效性, 即渐近协方差矩阵 $\text{Avar}(\hat{\boldsymbol{\theta}}_{\text{ML}}) = n[\boldsymbol{I}(\boldsymbol{\theta}_0)]^{-1}$, 在大样本下达到了克莱默-劳下限。

(3) 渐近正态, 即 $\sqrt{n}(\hat{\boldsymbol{\theta}}_{\text{ML}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\boldsymbol{0}, n[\boldsymbol{I}(\boldsymbol{\theta}_0)]^{-1})$, 可近似地认为 $\hat{\boldsymbol{\theta}}_{\text{ML}} \xrightarrow{d} N(\boldsymbol{\theta}, [\boldsymbol{I}(\boldsymbol{\theta}_0)]^{-1})$ 。

证明 (选读)

定理 (不变性) 如果将参数 θ “参数变换” (reparameterize) 为 $\alpha \equiv g(\theta)$, 则对 α 的最大似然估计就是 $\hat{\alpha}_{\text{ML}} = g(\hat{\theta}_{\text{ML}})$ 。

其中, $g(\cdot)$ 可以是多维函数, 也不要求 α 与 θ 有一一对应的函数关系。

利用最大似然估计的不变性, 有时可以大大简化计算。

【例】对 $(\mu^2 + \sigma^2)$ 的最大似然估计就是 $(\hat{\mu}_{\text{ML}}^2 + \hat{\sigma}_{\text{ML}}^2)$ 。

6.6 MLE 估计量的渐近协方差矩阵

在大样本下，最大似然估计量的渐近协方差矩阵为

$$\text{Avar}(\hat{\boldsymbol{\theta}}_{\text{ML}}) = n[\mathbf{I}(\boldsymbol{\theta}_0)]^{-1} = n \left\{ -\text{E} \left[\frac{\partial^2 \ln L(\boldsymbol{\theta}_0; \mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right] \right\}^{-1}$$

此表达式依赖于未知参数 $\boldsymbol{\theta}_0$ ，有三种估计方法。

1. 期望值法。如果知道黑塞矩阵期望值的具体函数形式，则直接以 $\hat{\boldsymbol{\theta}}_{\text{ML}}$ 替代 $\boldsymbol{\theta}_0$ 可得，

$$\widehat{\text{Avar}}(\hat{\boldsymbol{\theta}}_{\text{ML}}) = n \left\{ -\text{E} \left[\frac{\partial^2 \ln L(\hat{\boldsymbol{\theta}}_{\text{ML}}; \mathbf{y})}{\partial \hat{\boldsymbol{\theta}} \partial \hat{\boldsymbol{\theta}}'} \right] \right\}^{-1}$$

黑塞矩阵通常包含复杂的非线性函数,期望值可能无解析解。

2. 观测信息矩阵法。以 $\hat{\boldsymbol{\theta}}_{\text{ML}}$ 替代 $\boldsymbol{\theta}_0$ 后,将期望算子忽略掉:

$$\widehat{\text{Avar}}(\hat{\boldsymbol{\theta}}_{\text{ML}}) = n \left[-\frac{\partial^2 \ln L(\hat{\boldsymbol{\theta}}_{\text{ML}}; \mathbf{y})}{\partial \hat{\boldsymbol{\theta}} \partial \hat{\boldsymbol{\theta}}'} \right]^{-1}$$

此法称为“观测信息矩阵”(Observed Information Matrix,

简记 OIM)法。但二阶偏导数可能不易计算。

3. 梯度向量外积或 **BHHH** 法。利用信息矩阵等式，用 $\sum_{i=1}^n \hat{s}_i \hat{s}_i'$ 来估计 $I(\theta_0)$ ：

$$\text{即 } \widehat{\text{Avar}}(\hat{\theta}_{\text{ML}}) = n \left(\sum_{i=1}^n \hat{s}_i \hat{s}_i' \right)^{-1}$$

其中， $\hat{s}_i \equiv \frac{\partial \ln f(y_i; \hat{\theta}_{\text{ML}})}{\partial \theta}$ 为第 i 个观测值对得分函数的贡献之估计值。此法称为“梯度向量外积” (Outer Product of Gradients, 简记 OPG) 或 BHHH 法，只需计算一阶偏导数；

且协方差估计量总是非负定的(nonnegative definite)，而 OIM 法的协方差估计量无此保证。

这三种方法在大样本下渐近等价 (asymptotically equivalent)。

在有限样本中，计算结果可能差别较大，甚至导致统计推断作出不同的结论，参见 Greene (2012, p.522)。

这三种方法都建立在似然函数正确的前提下。如果似然函数不正确，则三种方法都失效，应使用稳健标准误。

6.7 三类渐近等价的统计检验

对于线性回归模型，检验原假设 $H_0: \boldsymbol{\beta} = \boldsymbol{\beta}_0$ ，其中 $\boldsymbol{\beta}_{K \times 1}$ 为未知参数， $\boldsymbol{\beta}_0$ 已知，共有 K 个约束。

(1) 沃尔德检验(Wald Test)

通过 $\boldsymbol{\beta}$ 的无约束估计量 $\hat{\boldsymbol{\beta}}_U$ 与 $\boldsymbol{\beta}_0$ 的距离来进行检验。如果 H_0 正确，则 $(\hat{\boldsymbol{\beta}}_U - \boldsymbol{\beta}_0)$ 的绝对值不应该很大。沃尔德统计量：

$$W \equiv (\hat{\boldsymbol{\beta}}_U - \boldsymbol{\beta}_0)' [\text{Var}(\hat{\boldsymbol{\beta}}_U)]^{-1} (\hat{\boldsymbol{\beta}}_U - \boldsymbol{\beta}_0) \xrightarrow{d} \chi^2(K)$$

K 为约束条件的个数，证明类似于第 5 章对于线性假设“ $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ ”的大样本检验。 t 检验、 F 检验都是 Wald 检验。

(2) 似然比检验 (Likelihood Ratio Test, 简记 LR)

无约束的似然函数最大值 $\ln L(\hat{\boldsymbol{\beta}}_U)$ 比有约束的似然函数最大值 $\ln L(\hat{\boldsymbol{\beta}}_R)$ 更大，因为在无约束条件下的参数空间 Θ 比有约束条件下(即 H_0 成立时)参数的取值范围更大。

在此例中，有约束的估计量 $\hat{\boldsymbol{\beta}}_R = \boldsymbol{\beta}_0$ 。

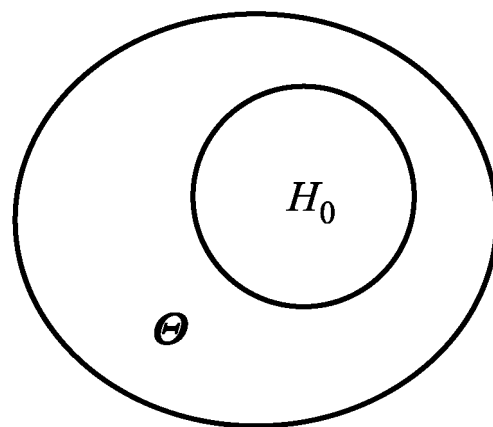


图 6.5 无约束与有约束的参数空间

如果 H_0 正确, 则 $\ln L(\hat{\beta}_U) - \ln L(\hat{\beta}_R)$ 不应该很大。LR 统计量:

$$\text{LR} \equiv -2 \ln \left[\frac{L(\hat{\beta}_R)}{L(\hat{\beta}_U)} \right] = 2 \left[\ln L(\hat{\beta}_U) - \ln L(\hat{\beta}_R) \right] \xrightarrow{d} \chi^2(K)$$

证明方法是，将对数似然函数作二阶泰勒展开(根据 MLE 一阶条件，一阶项为 0)。

F 统计量的另一表达式 $F = \frac{(\mathbf{e}^{*'} \mathbf{e}^* - \mathbf{e}' \mathbf{e}) / (K - 1)}{\mathbf{e}' \mathbf{e} / (n - K)}$ ，可视为 LR 统计量。

(3) 拉格朗日乘子检验(Lagrange Multiplier Test, 简记 LM):

考虑有约束条件的对数似然函数最大化问题:

$$\begin{aligned} \max_{\tilde{\boldsymbol{\beta}}} \quad & \ln L(\tilde{\boldsymbol{\beta}}) \\ \text{s.t.} \quad & \boldsymbol{\beta} = \boldsymbol{\beta}_0 \end{aligned}$$

引入拉格朗日乘子函数,

$$\max_{\tilde{\beta}, \lambda} \ln L(\tilde{\beta}) - \lambda'(\tilde{\beta} - \beta_0)$$

其中, λ 为拉格朗日乘子向量。

如果 $\hat{\lambda} \approx \mathbf{0}$, 则说明此约束条件不“紧”(tight)或不是“硬约束”(binding constraint), 加上此约束条件不会使似然函数的最大值下降很多, 即原假设 H_0 很可能成立。

根据一阶条件(对 $\tilde{\beta}$ 求导)可知, $\hat{\lambda} = \frac{\partial \ln L(\hat{\beta}_R)}{\partial \tilde{\beta}}$ 。LM 统计量:

$$\text{LM} \equiv \left(\frac{\partial \ln L(\hat{\boldsymbol{\beta}}_R)}{\partial \tilde{\boldsymbol{\beta}}} \right)' \left[\mathbf{I}(\hat{\boldsymbol{\beta}}_R) \right]^{-1} \left(\frac{\partial \ln L(\hat{\boldsymbol{\beta}}_R)}{\partial \tilde{\boldsymbol{\beta}}} \right) \xrightarrow{d} \chi^2(K)$$

其中, $\mathbf{I}(\hat{\boldsymbol{\beta}}_R)$ 为信息矩阵在 $\hat{\boldsymbol{\beta}}_R$ 处的取值。由于 $\frac{\partial \ln L(\tilde{\boldsymbol{\beta}})}{\partial \tilde{\boldsymbol{\beta}}}$ 为“得分函数” (score function), 故也称“得分检验” (score test); 而 $\mathbf{I}(\hat{\boldsymbol{\beta}}_R)$ 为得分函数的协方差矩阵。

在 $\hat{\boldsymbol{\beta}}_U$ 处, $\frac{\partial \ln L(\hat{\boldsymbol{\beta}}_U)}{\partial \tilde{\boldsymbol{\beta}}} = \mathbf{0}$ 。如 H_0 成立, 则在 $\hat{\boldsymbol{\beta}}_R$ 处, 也应有 $\frac{\partial \ln L(\hat{\boldsymbol{\beta}}_R)}{\partial \tilde{\boldsymbol{\beta}}} \approx \mathbf{0}$, 而 LM 统计量反映此接近程度。

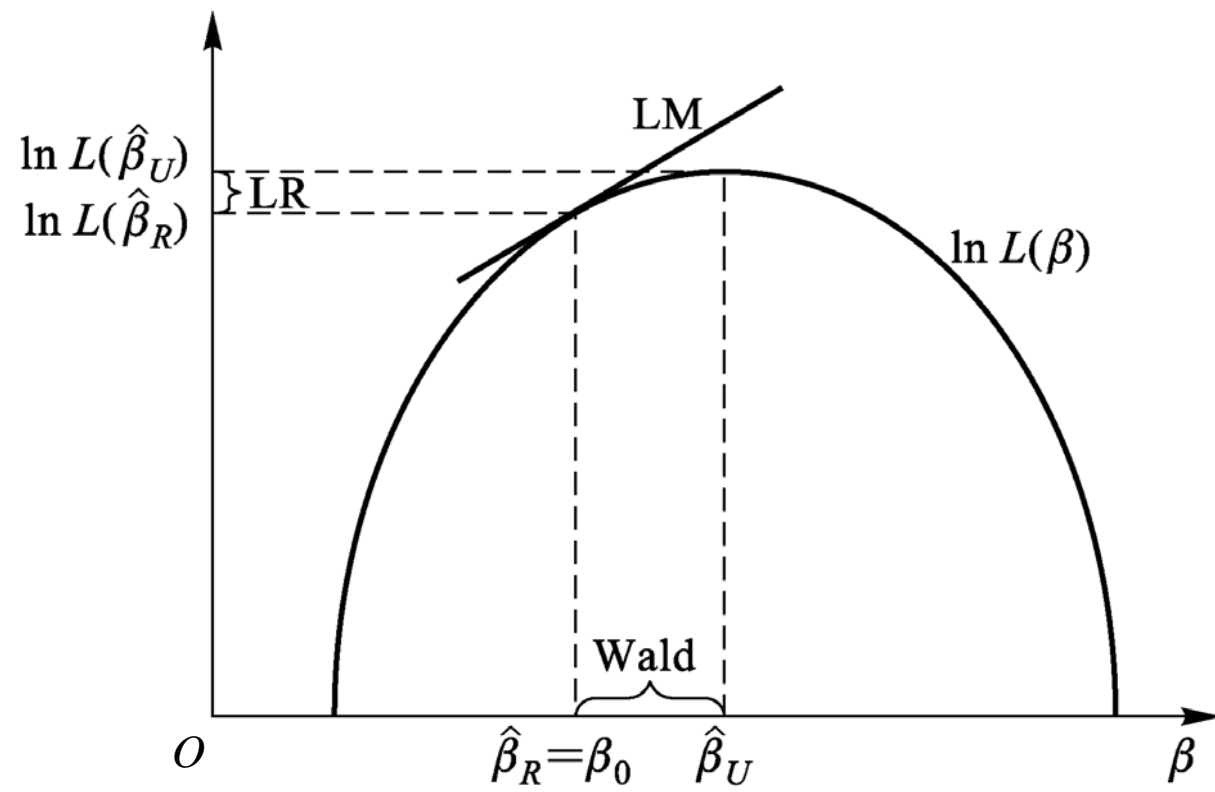


图 6.6 三类渐近等价的统计检验

Wald 检验仅利用无约束估计的信息，LM 检验仅利用有约束估计的信息，而 LR 检验同时利用二者的信息。

这三类检验在大样本下渐近等价，但小样本性质不同。

实际应用中究竟采取哪种检验常取决于“无约束估计”与“有约束估计”哪种更方便。

6.8 准最大似然估计法

如果随机变量不服从正态分布，却使用了以正态分布为前提的 MLE，该估计量是否一致？

对于线性模型，MLE 估计量等价于 OLS 估计量，而 OLS 估计量的一致性不依赖于正态分布的假定。

定义 使用不正确的似然函数而得到的最大似然估计，被称为**准最大似然估计** (Quasi MLE, 简记 QMLE)或“伪最大似然估计” (Pseudo MLE)。

虽然 MLE 常要求随机变量服从正态，此假定可能并不强。

如果 QMLE 估计量满足以下两个条件，则为一致估计量。

(i) 模型设定的概率密度函数属于“线性指数分布

族” (linear exponential family), 即概率密度函数可以写为

$$f(y; \boldsymbol{\theta}) = \frac{p(y)e^{r(\boldsymbol{\theta})}}{q(\boldsymbol{\theta})} \text{ 的形式。}$$

线性指数分布族包括正态分布, 二项分布, 泊松分布, 负二项分布, Γ 分布, 以及逆高斯分布(inverse Gaussian)等。

(ii) 条件期望 $E(y | \mathbf{x})$ 的函数形式设定正确。

一般情况下, QMLE 并不一致, 譬如第 14 章的 Tobit 回归。即使 QMLE 碰巧一致, $\hat{\theta}_{\text{QML}}$ 的渐近方差也不再是 $n[\mathbf{I}(\theta_0)]^{-1}$ 。

假设正确的对数似然函数为 $\ln L(\theta; \mathbf{y})$, 被误设为 $\ln L^*(\theta; \mathbf{y})$, 称为“准对数似然函数” (pseudo log likelihood function)。最大化 $\ln L^*(\theta; \mathbf{y})$ 的结果即 QMLE 估计量,

$$\hat{\theta}_{\text{QML}} \equiv \arg \max \ln L^*(\theta; \mathbf{y})$$

遵循类似于 MLE 一致性的证明步骤, 可证明 $\hat{\theta}_{\text{QML}} \xrightarrow{p} \theta^*$, 其中 θ^* 称为“准真实值” (pseudo-true value), 通常 $\theta^* \neq \theta_0$ 。

对于 $\hat{\boldsymbol{\theta}}_{\text{QML}}$ 的大样本分布，可证明：

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\text{ML}} - \boldsymbol{\theta}^*) \xrightarrow{d} N(0, \boldsymbol{A}_0^{*-1} \boldsymbol{B}_0^* \boldsymbol{A}_0^{*-1})$$

由于 $\ln L^*(\boldsymbol{\theta}; \boldsymbol{y})$ 并非真正的对数似然函数，信息矩阵等式不成立，故一般 $\boldsymbol{A}_0^* \neq \boldsymbol{B}_0^*$ ，夹心估计量 $\boldsymbol{A}_0^{*-1} \boldsymbol{B}_0^* \boldsymbol{A}_0^{*-1}$ 无法进简化。

基于 $\boldsymbol{A}_0^{*-1} \boldsymbol{B}_0^* \boldsymbol{A}_0^{*-1}$ 的标准误差被称为“胡贝尔-怀特稳健标准误” (Huber-White robust standard errors)。

胡贝尔-怀特稳健标准误也简称为“稳健标准误”，因为它

与异方差稳健标准误是一致的。

假设用 MLE 来估计古典线性回归模型，真实模型存在异方差，但在同方差的错误设定下来求 MLE 估计量，即得到的就是 QMLE 估计量。

此 $\hat{\beta}_{\text{QML}}$ 依然是真实参数 β 的一致估计，而胡贝尔-怀特稳健标准误就是异方差稳健的标准误。

在使用 MLE 估计非线性模型时，如果对模型的正确设定无把握，而 QMLE 估计量依然一致，应使用(胡贝尔-怀特)稳健标准误。Stata 的选择项为 “r” 或 “vce(robust)”。

如果对于模型设定很有信心，可直接使用 OIM 或 OPG 方法来估计渐近方差，没有必要使用稳健标准误。

当 QMLE 估计量不一致时，即使采用(胡贝尔-怀特)稳健标准误也无济于事，应担心估计量的一致性。

(胡贝尔-怀特)稳健标准误只是一致地估计了一个不一致估计量的方差(a consistent estimator of the variance of an inconsistent estimator)。

无论 OIM、OPG 法，还是(胡贝尔-怀特)稳健标准误都假设样本数据为 iid。如果样本数据可分为若干组，而同一组内的

观测值存在自相关，则应使用“聚类稳健标准误”(cluster-robust standard errors)，在 Stata 中由选择项“`vce(cluster clustvar)`”来实现。

总结：对于线性回归模型，建议总是使用稳健标准误。对于非线性模型，可分四种情况考虑。

(i) 如果对模型设定较有信心或模型拟合得较好，可不用稳健标准误。

(ii) 如果对模型设定缺乏信心，且 QMLE 为一致估计，应使用稳健标准误。

(iii) 如果对模型设定缺乏信心，但 QMLE 也不一致，应首先担心 QMLE 估计量的一致性，仅使用稳健标准误进行校正无济于事。

(iv) 对于聚类样本，应使用聚类稳健的标准误。

6.9 对正态分布假设的检验

对于线性回归模型，即使扰动项不服从正态分布，OLS 依然一致，且服从渐近正态，可用进行大样本推断；故检验扰动项是否服从正态分布意义不大。

对非线性模型，由于正态分布假定是推导 MLE 的前提，故检验扰动项是否服从正态分布可能较重要。

在某些情况下，“准最大似然估计”也一致，但毕竟不如真正的 MLE 有效率。

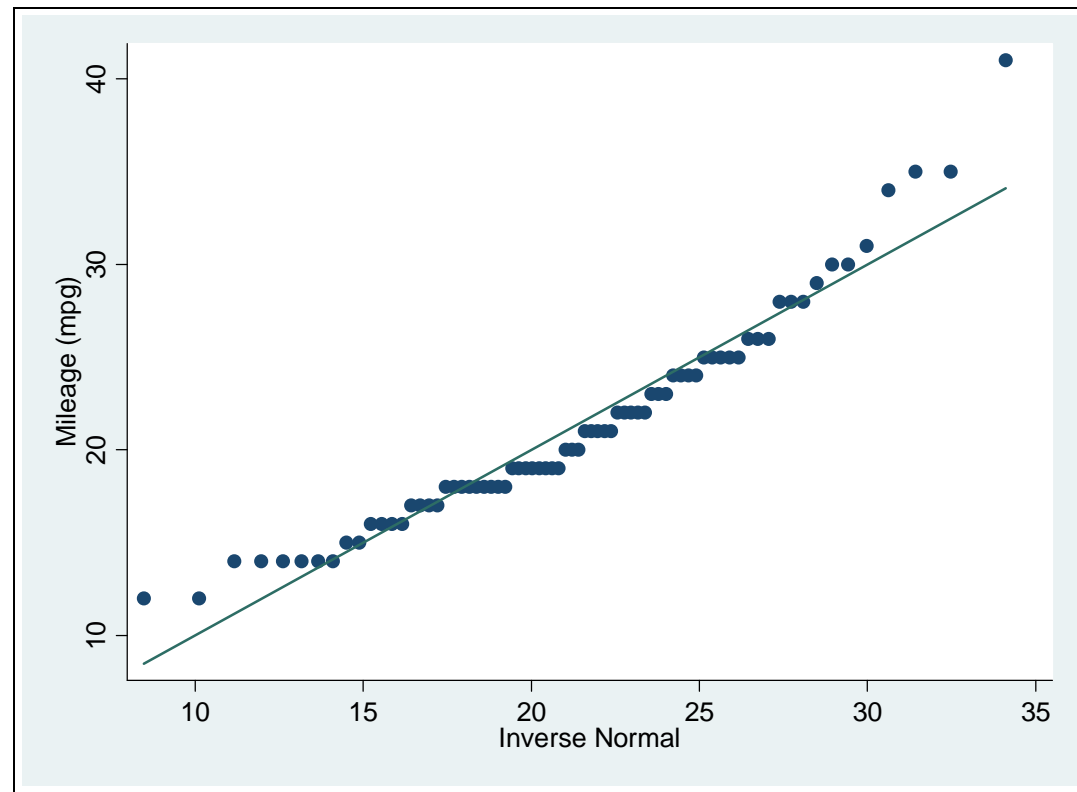
为了考察扰动项是否为正态，最直观的方法是画图。可把残差画成直方图(histogram)，并与正态分布的密度函数比较。

但直方图是不连续的。为得到光滑估计，可使用“核密度估计法”(kernel density estimation)，并与正态密度相比。

另一画图方法是，将正态分布的分位数(quantiles)与残差的分位数画成散点图(scatter plot)。

如果残差来自正态分布，则该图上的散点应该集中在 45° 线附近。

称这种图为“分位数-分位数图”(Quantile-Quantile plot, 简记 QQ plot)。



QQ 图

常用的检验方法利用了正态分布的偏度与峰度性质。

随机变量 X 的偏度为 $E[(X - \mu) / \sigma]^3$, 峰度为 $E[(X - \mu) / \sigma]^4$, 超额峰度为 $E[(X - \mu) / \sigma]^4 - 3$ 。

对于残差 $\{e_1, \dots, e_n\}$, 其偏度与超额峰度的样本估计值分别为

$$\frac{1}{n\hat{\sigma}^3} \sum_{i=1}^n e_i^3 \quad \text{与} \quad \left(\frac{1}{n\hat{\sigma}^4} \sum_{i=1}^n e_i^4 \right) - 3$$

其中, $\bar{e} = 0$ 。在扰动项服从正态分布的原假设下, 这两个统计量服从正态分布。

“雅克-贝拉检验” (Jarque and Bera, 1987, 简记 JB)使用它们的平方之加权平均作为检验统计量:

$$JB \equiv \frac{n}{6} \left[\left(\frac{1}{n\hat{\sigma}^3} \sum_{i=1}^n e_i^3 \right)^2 + \frac{1}{4} \left(\frac{1}{n\hat{\sigma}^4} \sum_{i=1}^n e_i^4 - 3 \right)^2 \right] \xrightarrow{d} \chi^2(2)$$

JB 检验虽常用, 但收敛速度较慢, 对样本容量要求较高。Stata 官方案程序中提供了 D'Agostino et al (1990)的改进方法, 基于偏度与峰度设计了更复杂的检验统计量。

如果发现某变量不服从正态分布, 有时可以通过取对数, 使之变得更接近于正态分布。