第9章 二值选择模型

二值选择模型

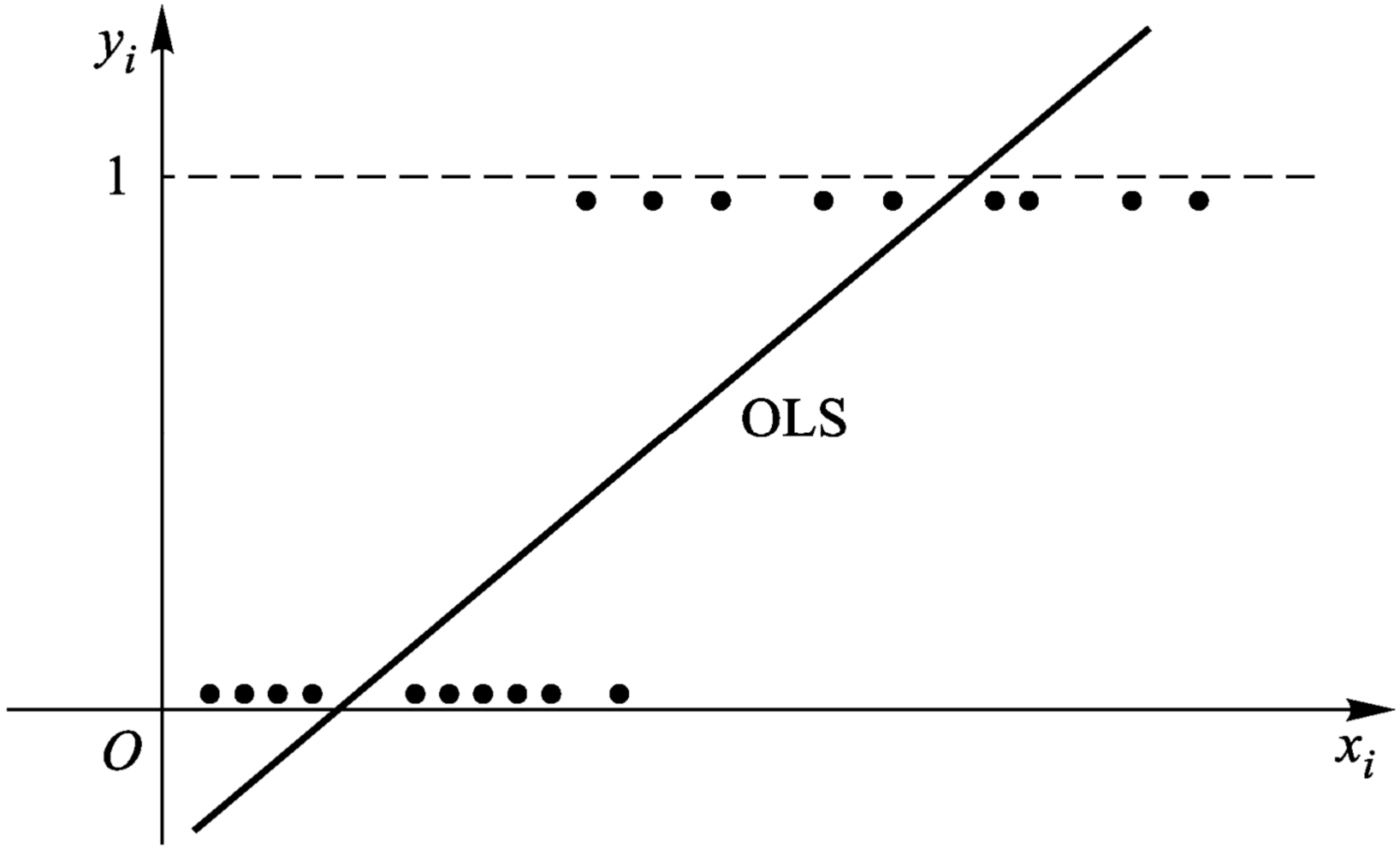
- 如果被解释变量y离散，称为“离散选择模型”(discrete choice model)或“定性反应模型”(qualitative response model)。
- 最常见的离散选择模型是二值选择行为(binary choices)。
- 比如：考研或不考研；就业或待业；买房或不买房；买保险或不买保险；贷款申请被批准或拒绝；出国或不出国；回国或不回国；战争或和平；生或死。

- 假设个体只有两种选择，比如$y = 1$(考研)或$y = 0$(不考研)。
- 最简单的建模方法为**线性概率模型(**Linear Probability Model，LPM)：

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i = \boldsymbol{x}_{\boldsymbol{i}}'\boldsymbol{\beta} + \varepsilon_i \quad (i = 1, \cdots, n)$$

- 其中，解释变量$\boldsymbol{x}_i \equiv (x_{i1} \ \ x_{i2} \ \ \cdots \ \ x_{iK})'$，而参数$\boldsymbol{\beta} \equiv (\beta_1 \ \ \beta_2 \ \ \beta_K)'$
- LPM 的优点是，计算方便，容易得到边际效应(即回归系数)。

- LPM 的缺点是，虽然y的取值非 0即1，但根据线性概率模型所作的预测值却可能出现$\hat{y} > 1$或$\hat{y} < 0$的不现实情形。



- 为使 y的预测值介于[0,1]之间，在给定x的情况下，考虑 y的两点分布概率：

$$\begin{cases} \mathrm{P}(y=1 \mid \boldsymbol{x}) = F(\boldsymbol{x}, \boldsymbol{\beta}) \\ \mathrm{P}(y=0 \mid \boldsymbol{x}) = 1 - F(\boldsymbol{x}, \boldsymbol{\beta}) \end{cases}$$

- 函数$F(\boldsymbol{x}, \boldsymbol{\beta})$称为“连接函数”(link function)，因为它将$\boldsymbol{x}$与y连接起来。
- y的取值要么为0，要么为1，故y肯定服从两点分布。
- 连接函数的选择具有一定灵活性。
- 通过选择合适的连接函数$F(\boldsymbol{x}, \boldsymbol{\beta})$(比如，某随机变量的累积分布函数)，可保证$0 \le \hat{y} \le 1$，并将$\hat{y}$理解为“$y = 1$”发生的概率，因为

$$E(y|\boldsymbol{x}) = 1 \cdot P(y = 1|\boldsymbol{x}) + 0 \cdot P(y = 0|\boldsymbol{x}) = P(y = 1|\boldsymbol{x})$$

- 如果$F(\boldsymbol{x},\boldsymbol{\beta})$为标准正态的累计分布函数，则

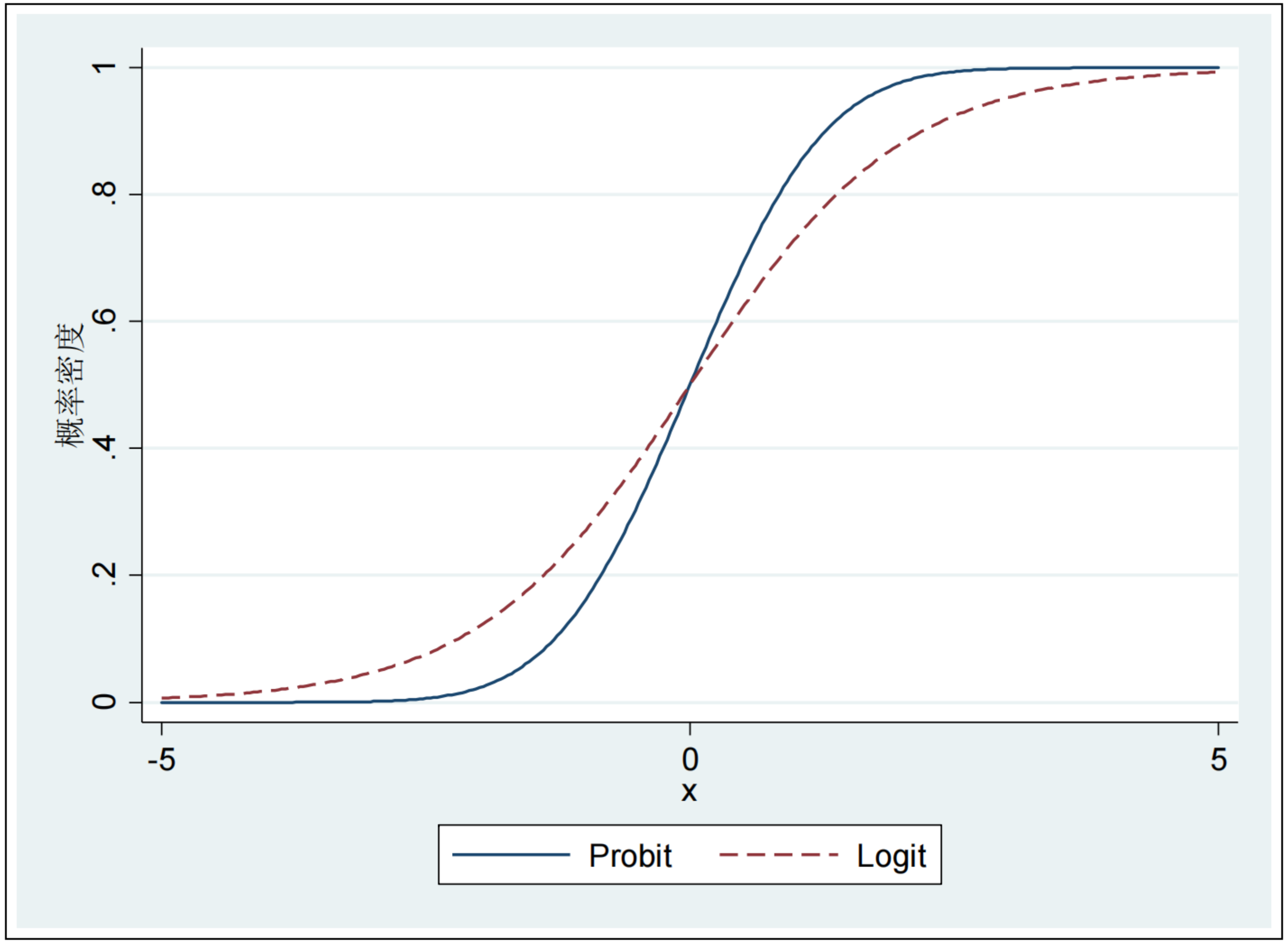
$$\mathrm{P}(y=1\mid\boldsymbol{x})=F(\boldsymbol{x},\boldsymbol{\beta})=\Phi(\boldsymbol{x}'\boldsymbol{\beta})\equiv\int_{-\infty}^{\boldsymbol{x}'\boldsymbol{\beta}}\phi(t)\,dt$$

- $\phi(\cdot)$与$\Phi(\cdot)$ 分别为标准正态的密度与累积分布函数；此模型称为"Probit"。
- 如果$F(\boldsymbol{x},\boldsymbol{\beta})$为"逻辑分布"(logistic distribution)的累积分布函数，则

$$\mathrm{P}(y=1\mid\boldsymbol{x})=F(\boldsymbol{x},\boldsymbol{\beta})=\Lambda(\boldsymbol{x}'\boldsymbol{\beta})\equiv\frac{\exp(\boldsymbol{x}'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}'\boldsymbol{\beta})}$$

- 其中，函数$\Lambda(\cdot)$的定义为$\Lambda(z)\equiv\frac{exp(z)}{1+exp(z)}$；此模型称为"Logit"。

- 逻辑分布的密度函数关于原点对称，期望为 0，方差为$\pi^2/3$（大于标准正态的方差），具有厚尾(fat tails)。
- Probit 与 Logit 都很常用，二者的估计结果(比如边际效应)通常很接近。
- Logit 模型的优势在于：
- （1）逻辑分布的累积分布函数有解析表达式(标准正态没有)，故计算 Logit 更为方便；
- （2）Logit 的回归系数更易解释其经济意义。



概率密度
0
.2
.4
.6
.8
1
-5
0
5
x
Probit
Logit

最大似然估计的原理

- Probit 与 Logit 模型本质上都是非线性模型，无法通过变量转换变为线性模型。
- 对于非线性模型，常使用最大似然估计法(Maximum Likelihood Estimation，MLE 或 ML)。

- 假设随机变量y的概率密度函数$f(y;\theta)$，其中θ为未知参数。
- 为估计θ，从y的总体中抽取样本容量为n的随机样本$\{y_1,\cdots,y_n\}$。
- 假设$\{y_1,\cdots,y_n\}$为iid，样本数据的联合密度函数为

$$f(y_1;\theta)f(y_2;\theta)\cdots f(y_n;\theta)=\prod_{i=1}^{n}f(y_i;\theta)$$

- 其中，$\prod_{i=1}^{n}$ 表示连乘。
- 抽样之前，$\{y_1,\cdots,y_n\}$为随机向量。
- 抽样之后，$\{y_1,\cdots,y_n\}$有了特定的样本值。
- 可将样本的联合密度函数视为在给定$\{y_1,\cdots,y_n\}$情况下，未知参数θ的函数。

- 定义 **似然函数**(likelihood function)为

$$L(\theta; y_1, \cdots, y_n) = \prod_{i=1}^{n} f(y_i; \theta)$$

- 似然函数与联合密度函数完全相等，只是θ与$\{y_1, \cdots, y_n\}$的角色互换，即把θ作为自变量，视$\{y_1, \cdots, y_n\}$为给定。
- 为运算方便，把似然函数取对数：

$$\ln L(\theta; y_1, \cdots, y_n) = \sum_{i=1}^{n} \ln f(y_i; \theta)$$

- 取对数是单调变换，不影响最大化的求解。

- MLE 的思想：给定样本取值后，该样本最可能来自参数θ为何值的总体。
- 寻找$\hat{\theta}_{ML}$，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)：

$$\max_{\theta} \ \ln L(\theta; y_1, \cdots, y_n)$$

- 假设存在唯一内点解，一阶条件为

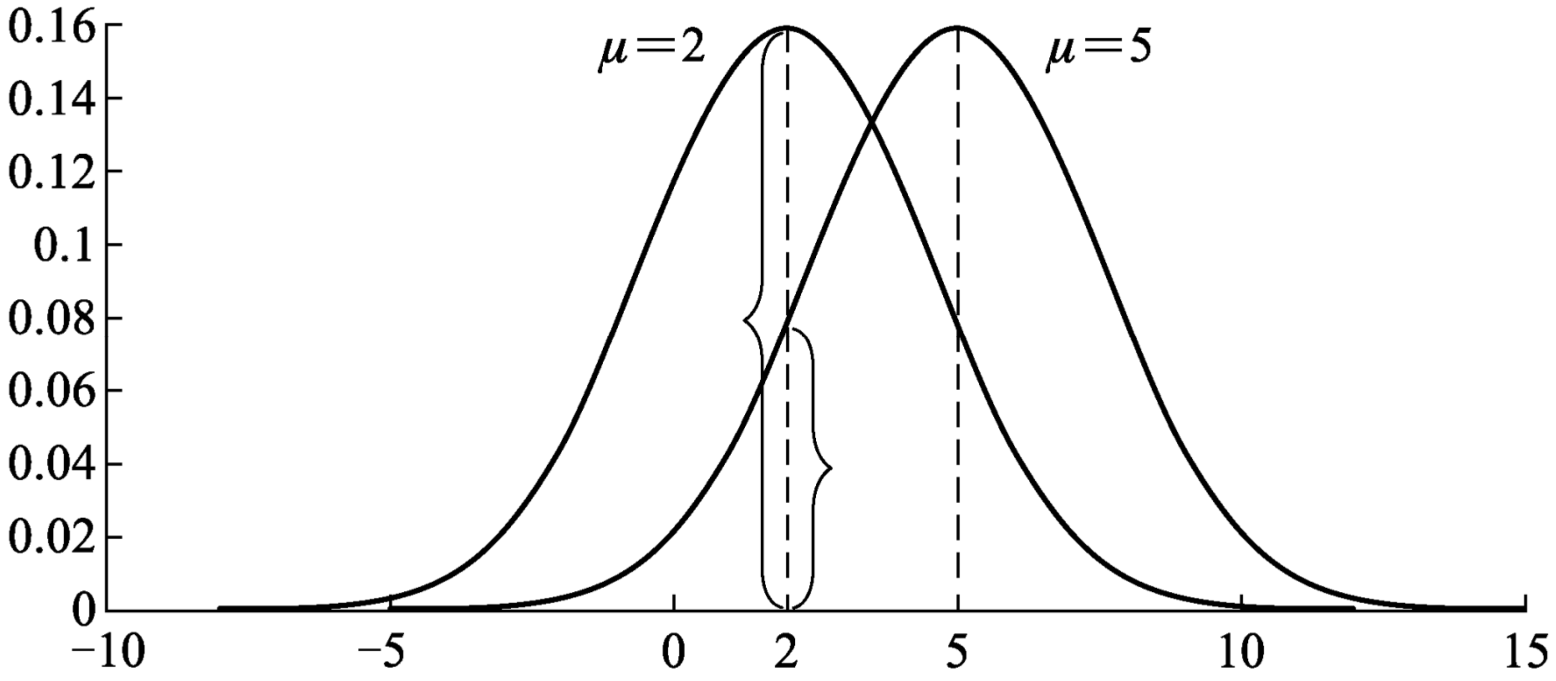
$$\frac{\partial \ln L(\theta; y_1, \cdots, y_n)}{\partial \theta} = 0$$

- 求解一阶条件，可得最大似然估计量$\hat{\theta}_{ML}$。

- **例** 假设$y \sim N(\mu,\sigma^2)$，其中σ^2已知，得到样本容量为 1 的样本$y_1=2$，求对μ的最大似然估计。根据正态分布的密度函数，此样本的似然函数为

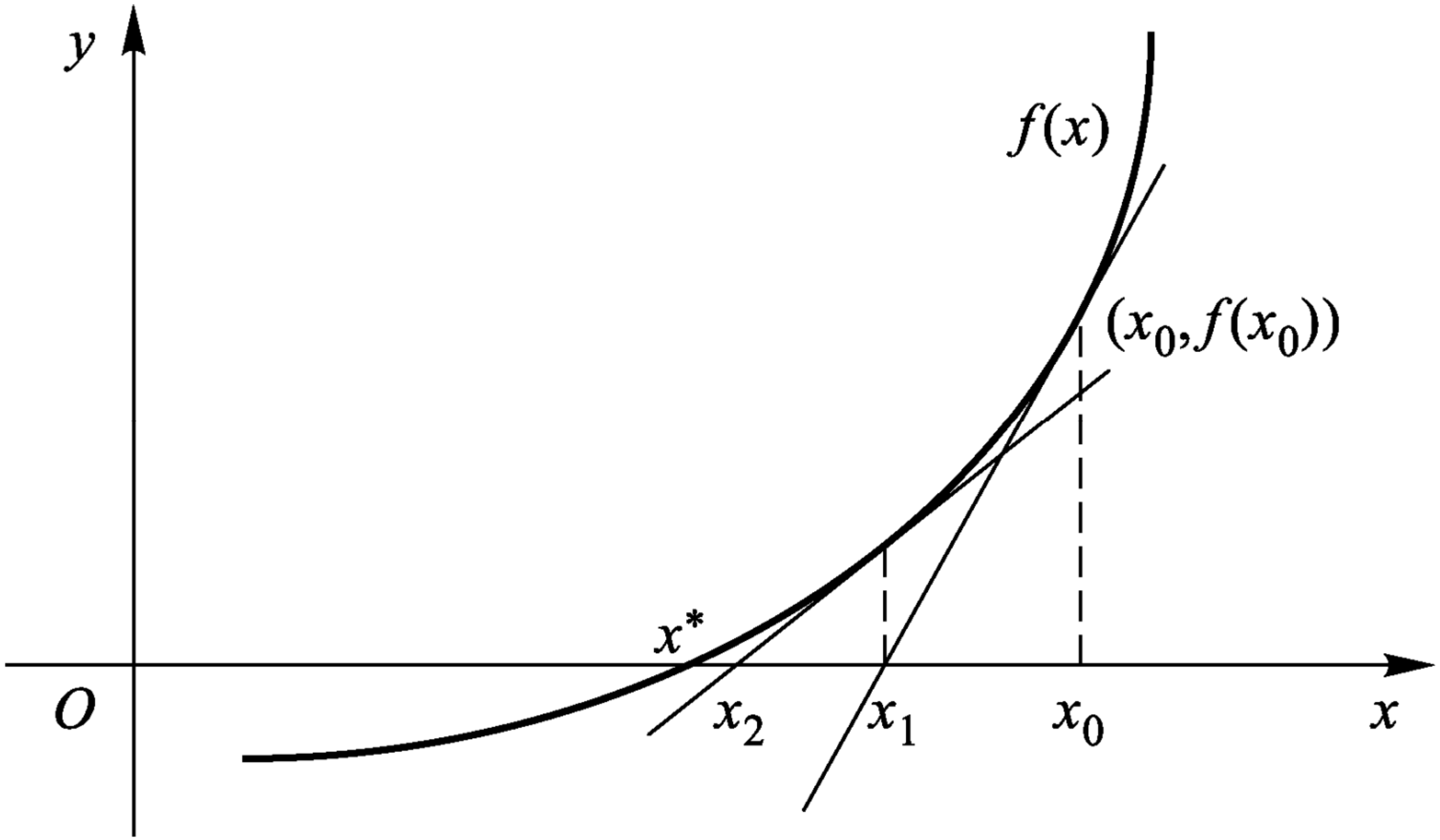
$$L(\mu)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{-(2-\mu)^2}{2\sigma^2}\right\}$$

- 似然函数在$\hat{\mu} = 2$处取最大值



- 在一定的正则条件(regularity conditions)下，MLE 估计量具有良好的大样本性质，可照常进行大样本统计推断。
- （1）$\hat{\theta}_{ML}$为一致估计，即$plim\hat{\theta}_{ML} = \theta$。
- （2）$\hat{\theta}_{ML}$服从渐近正态分布。
- （3）在大样本下，$\hat{\theta}_{ML}$是最有效率的估计（渐近方差最小）。

- 由于模型存在非线性，MLE 通常没有解析解，只能寻找“数值解”(numerical solution)。
- 一般使用“迭代法”(iteration)进行数值求解。
- 常用的迭代法为“高斯-牛顿法”(Gauss-Newton method)。
- MLE 的一阶条件可归结为求非线性方程$f(x) = 0$的解。
- 假设$f(x)$的导数$f'(x)$处处存在。记该方程的解为x^*，满足$f(x^*) = 0$。



- 首先，猜初始值x_0，在点$(x_0, f(x_0))$处作曲线$f(x)$的切线，记切线与横轴的交点为x_1。
- 然后，在点$(x_1, f(x_1))$处再作切线，记切线与横轴的交点为x_2。

- 以此类推，不断迭代，可得序列$\{x_0, x_1, x_2, x_3, \cdots\}$。
- 一般情况下，该序列将收敛至x^*（给定一个精确度，收敛到精确度范围内即停止）。
- 高斯-牛顿法的收敛速度很快，是二次的。如本次迭代的误差为0.1，则下次迭代的误差约为0.1^2，下下次迭代的误差约为0.1^4，等等。
- 如果初始值x_0选择不当，也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)。

- MLE 很容易应用于多参数的情形。
- 假设随机变量y的概率密度函数为$f(y;\boldsymbol{\theta})$，其中$\boldsymbol{\theta}=(\theta_1\ \ \theta_2)'$，则对数似然函数为

$$lnL(\boldsymbol{\theta};y_1,\cdots,y_n)=\sum_{i=1}^{n}lnf(y_i;\boldsymbol{\theta})$$

一阶条件为

$$\begin{cases}\dfrac{\partial lnL(\boldsymbol{\theta};y_1,\cdots,y_n)}{\partial\theta_1}=0\\[2ex]\dfrac{\partial lnL(\boldsymbol{\theta};y_1,\cdots,y_n)}{\partial\theta_2}=0\end{cases}$$

- 求解此联立方程组，可得最大似然估计量$\hat{\theta}_{1,ML}$与$\hat{\theta}_{2,ML}$。
- 高斯-牛顿法也适用于多元函数$f(\boldsymbol{x})=0$的情形，只要在上述迭代的过程中，将切线替换为(超)切平面即可。

二值选择模型的 MLE 估计

- 以 Logit 为例，将 MLE 应用于二值选择模型。
- 对于样本数据$\{\boldsymbol{x}_i, y_i\}_{i=1}^n$，根据逻辑分布的累计分布函数，第i个观测数据的概率密度为

$$f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}) = \begin{cases} \Lambda(\boldsymbol{x}_i'\boldsymbol{\beta}), & \text{若 } y_i = 1 \\ 1 - \Lambda(\boldsymbol{x}_i'\boldsymbol{\beta}), & \text{若 } y_i = 0 \end{cases}$$

- 其中，$\Lambda(z) \equiv \frac{exp(z)}{1+exp(z)}$为逻辑分布的累计分布函数。

- 上式可写为

$$f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}) = \left[\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})\right]^{y_i}\left[1-\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})\right]^{1-y_i}$$

- 取对数可得

$$\ln f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}) = y_i \ln\left[\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})\right] + (1-y_i)\ln\left[1-\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})\right]$$

- 假设样本中的个体相互独立，整个样本的对数似然函数为

$$\ln L(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{x}) = \sum_{i=1}^{n} y_i \ln\left[\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})\right] + \sum_{i=1}^{n}(1-y_i)\ln\left[1-\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})\right]$$

- 把对数似然函数对$\boldsymbol{\beta}$求偏导，可得一阶条件。
- 满足一阶条件的估计量即为MLE估计量，记为$\hat{\boldsymbol{\beta}}_{ML}$。

边际效应

- 对于线性模型，回归系数β_k的经济意义就是变量x_k对y的边际效应(marginal effects)。
- 在非线性模型中，估计量$\hat{\boldsymbol{\beta}}_{ML}$一般并非边际效应。
- 以 Probit 为例，$P(y=1|\boldsymbol{x})=\Phi(\boldsymbol{x}'\boldsymbol{\beta})\equiv\int_{-\infty}^{\boldsymbol{x}'\boldsymbol{\beta}}\phi(t)dt$
- 计算变量x_k的边际效应（使用求倒数的chain rule）

$$\frac{\partial P(y=1|\boldsymbol{x})}{\partial x_k}=\frac{\partial\Phi(\boldsymbol{x}'\boldsymbol{\beta})}{\partial x_k}=\frac{\partial\Phi(\boldsymbol{x}'\boldsymbol{\beta})}{\partial(\boldsymbol{x}'\boldsymbol{\beta})}\cdot\frac{\partial(\boldsymbol{x}'\boldsymbol{\beta})}{\partial x_k}=\phi(\boldsymbol{x}'\boldsymbol{\beta})\beta_k$$

- 由于 Probit 与 Logit 所用分布函数不同，其参数估计值不直接可比。需分别计算二者的边际效应，然后比较。
- 对于非线性模型，边际效应通常不是常数，随着向量$\boldsymbol{x}$而变。
- 非线性模型常用的边际效应概念：
- (1) 平均边际效应 (average marginal effect)：分别计算在每个样本观测值上的边际效应，然后进行简单算术平均。
- (2) 样本均值处的边际效应 (marginal effect at mean)：计算在$\boldsymbol{x} = \bar{\boldsymbol{x}}$处的边际效应。
- (3) 在某代表值处的边际效应 (marginal effect at a representative value)：给定$\boldsymbol{x}^*$，计算$\boldsymbol{x} = \boldsymbol{x}^*$处的边际效应。
- 以上三种边际效应的计算结果可能有较大差异。

- 传统上，常计算样本均值处传统上，常计算样本均值处$\boldsymbol{x} = \bar{\boldsymbol{x}}$的边际效应，因为计算方便。
- 但在非线性模型中，样本均值处的个体行为并不等于样本中个体的平均行为(average behavior of individuals differs from behavior of the average individual)。
- 对于政策分析而言，使用平均边际效应(Stata 的默认方法)，或在某代表值处的边际效应通常更有意义。

回归系数的经济意义

- $\hat{\boldsymbol{\beta}}_{ML}$并非边际效应，它究竟有什么含义?
- 对于 Logit 模型，记事件发生的概率为$p \equiv \mathrm{P}(y=1|\boldsymbol{x})$，则事件不发生的概率为$1-p \equiv \mathrm{P}(y=0|\boldsymbol{x})$。
- 由于$p=\Lambda(\boldsymbol{x}'\boldsymbol{\beta})=\frac{exp(\boldsymbol{x}'\boldsymbol{\beta})}{1+exp(\boldsymbol{x}'\boldsymbol{\beta})}$，$1-p=\frac{1}{1+exp(\boldsymbol{x}'\boldsymbol{\beta})}$，故事件发生与不发生的几率为

$$\frac{p}{1-p}=exp(\boldsymbol{x}'\boldsymbol{\beta})$$

- $\frac{p}{1-p}$称为“几率”(odds)或“相对风险”(relative risk)。

- **例** 在检验药物疗效的随机实验中，“$y=1$”表示“生”，“$y=0$”表示“死”。如几率为 2，意味着存活的概率是死亡概率的两倍。
- 对方程两边取对数

$$ln(\frac{p}{1-p})=\boldsymbol{x}'\boldsymbol{\beta}=\beta_1x_1+\cdots+\beta_Kx_K$$

- $ln(\frac{p}{1-p})$称为“对数几率”(log-odds)。
- 回归系数$\hat{\beta}_j$表示，变量x_j增加一个微小量引起对数几率的边际变化。
- 取对数意味着百分比的变化，故可把$\hat{\beta}_j$视为半弹性(semi-elasticity)，即x_j增加一单位引起几率比($\frac{p}{1-p}$)的变化百分比。

- **例** $\hat{\beta}_j = 0.12$，意味着x_j增加一单位引起几率增加 12%。
- 如x_j为离散变量(比如，性别、子女数)，可使用另一解释法。
- 假设x_j增加一单位，从x_j变为x_j+1，记事件发生概率p的新值为p^*，则新几率与原几率的比率(即几率比，odds ratio)为

$$\frac{\frac{p^*}{1-p^*}}{\frac{p}{1-p}} = \frac{\exp[\beta_1 + \beta_2 x_2 + \cdots + \beta_j (x_j + 1) + \cdots + \beta_K x_K]}{\exp(\beta_1 + \beta_2 x_2 + \cdots + \beta_j x_j + \cdots + \beta_K x_K)} = \exp(\beta_j)$$

- $exp(\hat{\beta}_j)$表示变量x_j增加一单位引起几率的变化倍数。
- Stata称$exp(\hat{\beta}_j)$为几率比(odds ratio)。

- 例　$\hat{\beta}_j = 0.12$，则 $exp(\hat{\beta}_j) = e^{0.12} = 1.13$，故当 x_j 增加一单位时，新几率是原几率的 1.13 倍，或增加 13%，因为$exp(\hat{\beta}_j) - 1 = 1.13 - 1 = 0.13$。
- 如果$\hat{\beta}_j$较小，则$exp(\hat{\beta}_j) - 1 \approx \hat{\beta}_j$（将$exp(\hat{\beta}_j)$泰勒展开），以上两种方法基本等价。
- 对于Probit模型，无法对其系数$\widehat{\boldsymbol{\beta}}_{ML}$进行类似解释。

拟合优度

- 不存在平方和分解公式，无法计算R^2。
- Stata 仍汇报“准R^2”或“伪R^2”(Pseudo R^2)，由 McFadden (1974)提出，定义为

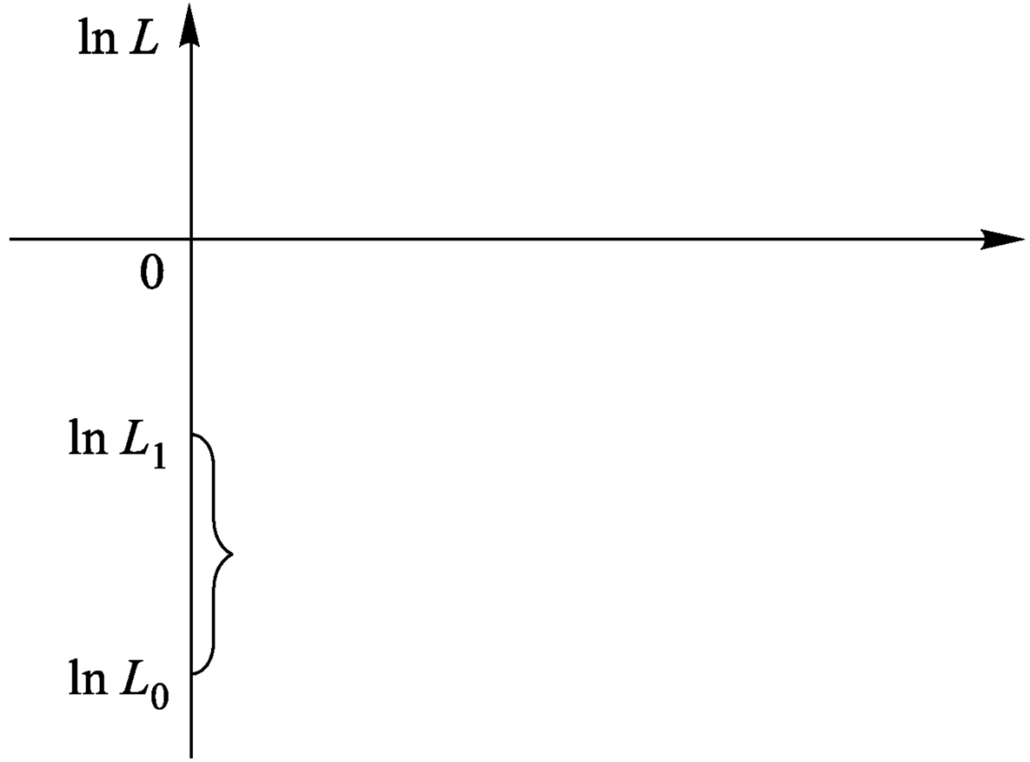
$$准R^2 \equiv \frac{lnL_0 - lnL_1}{lnL_0}$$

- lnL_1为原模型的对数似然函数的最大值
- lnL_0为以常数项为唯一解释变量的对数似然函数之最大值。
- 由于y为两点分布，似然函数的最大可能值为 1(取值概率为 1)，故对数似然函数的最大可能值为 0，记为lnL_{max}。

- 由于$lnL_{max}=0$，可将“准R^2”写为

$$准R^2 \equiv \frac{lnL_1 - lnL_0}{lnL_{max} - lnL_0}$$

- 显然，$0 \ge lnL_1 \ge lnL_0$，而$0 \le 准R^2 \le 1$。
- 分子为对数似然函数的实际增加值($lnL_1 - lnL_0$）；分母为对数似然函数的最大可能增加值($lnL_{max} - lnL_0$）。



- 判断拟合优度的另一方法是计算“正确预测的百分比”(percent correctly predicted)。
- 如发生概率的预测值$\hat{y} \geq 0.5$，则认为其预测值$y = 1$。
- 反之，则认为其预测$y = 0$。
- 将预测值与实际值(样本数据)进行比较，可计算正确预测的百分比。

准最大似然估计

- 使用 MLE 的前提是对总体的分布函数作具体的假定。
- Probit 与 Logit 模型分别假设y的两点分布概率为标准正态或逻辑分布的累积分布函数。
- 此分布函数的设定可能不正确，即存在“设定误差”(specification error)。
- **定义** 使用不正确的分布函数所得到的最大似然估计量，称为“准最大似然估计”(Quasi MLE，简记 QMLE)或“伪最大似然估计”(Pseudo MLE)。
- 准最大似然估计是否一定不一致？不一定！

- **例** 假设线性模型的扰动项服从正态分布，则 MLE 估计量与OLS 估计量完全相同，而 OLS 估计量的一致性并不依赖于关于分布函数的具体假设。
- 关于 QMLE 估计量的标准误，可分两种情况考虑。
- (1) 如果 QMLE 为一致估计量，由于可能存在对分布函数的设定误差，应使用稳健标准误(robust standard errors)，即相对于模型设定稳健的标准误。
- 此稳健标准误与异方差稳健的标准误是一致的，因为扰动项方差是否相同也是一种模型设定。

- (2) 如 QMLE 估计量不一致，即使采用稳健标准误也无济于事。
- QMLE 估计量$\hat{\boldsymbol{\beta}}_{QML} \xrightarrow{p} \boldsymbol{\beta}^* \neq \boldsymbol{\beta}$，应首先担心估计量的一致性。
- 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)。
- 对于二值选择模型(Probit 或 Logit)，只要条件期望函数$E(y|\boldsymbol{x}) = F(\boldsymbol{x}, \boldsymbol{\beta})$设定正确，则MLE估计就一致。
- 由于两点分布的特殊性，在 iid 的情况下，只要$E(y|\boldsymbol{x}) = F(\boldsymbol{x}, \boldsymbol{\beta})$成立，稳健标准误就等于普通标准误。

- 如果认为模型设定正确，就不必使用稳健标准误(使用稳健标准误也没错)。
- 如果模型设定不正确（即$E(y|\boldsymbol{x}) \neq F(\boldsymbol{x}, \boldsymbol{\beta})$），则Probit与Logit模型不能得到对系数$\boldsymbol{\beta}$的一致估计，使用稳健标准误就没有太大意义；首先应解决参数估计的一致性问题。
- 总之，对于二值选择模型，使用普通标准误或稳健标准误都可。

三类渐近等价的大样本检验

- 在计量中，常使用三类在大样本下渐近等价的统计检验。
- 考虑线性回归模型：

$$y_i = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + \varepsilon_i = \boldsymbol{x}_i^{'} \boldsymbol{\beta} + \varepsilon_i \ \ (i = 1, \cdots, n)$$

- 其中，解释变量$\boldsymbol{x} \equiv (x_1 \ \ x_2 \ \ \cdots \ \ x_K)'$，参数$\boldsymbol{\beta} \equiv (\beta_1 \ \ \beta_2 \ \ \cdots \ \ \beta_K)'$
- 检验以下原假设：

$$H_0: \ \boldsymbol{\beta} = \boldsymbol{\beta_0}$$

其中，$\boldsymbol{\beta_0}$已知，共有K个约束。

(1) **沃尔德检验**(Wald Test)

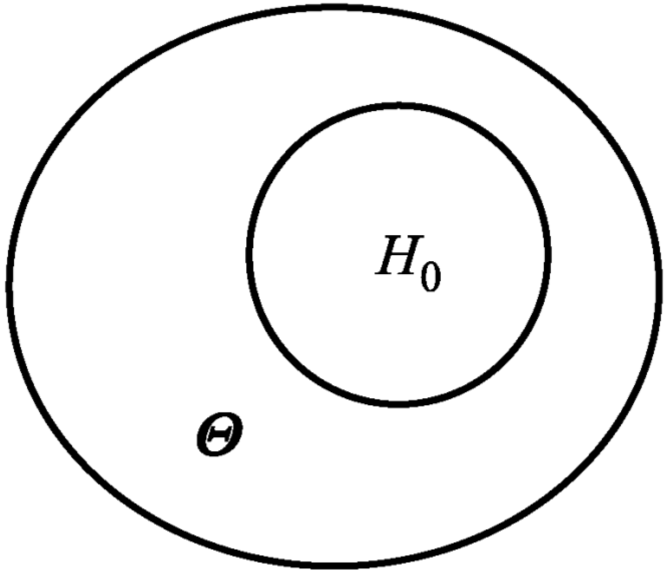
- 沃尔德检验考察$\boldsymbol{\beta}$的无约束估计量$\hat{\boldsymbol{\beta}}$（未施加H_0约束的估计量）与$\boldsymbol{\beta_0}$的距离。
- 基本思想：如果H_0正确，则$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})$不应该很大。
- 由于$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0})$为多维向量，使用二次型：

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

- 其中K为约束条件的个数。
- 单一系数t 检验（大样本下使用标准正态进行检验）、联合线性假设的F 检验（大样本下可使用χ^2 分布进行检验）都是Wald 检验。

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

- 似然比检验比较无约束估计量$\widehat{\boldsymbol{\beta}}$（不施加任何约束）与有约束估计量$\widehat{\boldsymbol{\beta}}^*$（施加$H_0$：$\boldsymbol{\beta} = \boldsymbol{\beta_0}$的约束）的差别。
- 无约束的似然函数最大值$lnL(\widehat{\boldsymbol{\beta}})$比有约束的似然函数最大值$lnL(\widehat{\boldsymbol{\beta}}^*)$更大，因为在无约束条件下的参数空间$\Theta$比有约束条件下(即$H_0$成立时)参数的取值范围更大。
- 基本思想：如果H_0正确，则$[lnL(\widehat{\boldsymbol{\beta}}) - lnL(\widehat{\boldsymbol{\beta}}^*)]$不应该很大。
- 在此例中，有约束的估计量$\widehat{\boldsymbol{\beta}}^* = \boldsymbol{\beta_0}$。



- LR统计量为

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

- 在大样本下，LR统计量也服从渐近$\chi^2(K)$分布。
- F统计量的表达式$F = \frac{(SSR^* - SSR)/(K-1)}{SSR/(n-K)}$，即依据似然比原理而设计。
- 在进行 Probit 或 Logit 回归时，Stata 会汇报一个似然比统计量，检验除常数项外所有参数的联合显著性。

- (3) **拉格朗日乘子检验**(Lagrange Multiplier Test，简记LM)
- Wald 检验只考察无约束估计量$\hat{\boldsymbol{\beta}}$。
- LR检验同时考察无约束估计量$\hat{\boldsymbol{\beta}}$与有约束估计量$\hat{\boldsymbol{\beta}}^*$。
- LM 检验则只考察有约束估计量$\hat{\boldsymbol{\beta}}^*$。
- 有约束条件的对数似然函数最大化问题：

$$\max_{\tilde{\beta}} \quad \ln L(\tilde{\boldsymbol{\beta}})$$

$$s.t. \quad \tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}_0$$

- $\tilde{\boldsymbol{\beta}}$为在最大化过程中假想的参数$\boldsymbol{\beta}$取值(hypothetical value)。

- 对于约束极值问题，引入拉格朗日乘子函数：

$$\max_{\tilde{\boldsymbol{\beta}},\, \boldsymbol{\lambda}} \ \ln L(\tilde{\boldsymbol{\beta}}) - \boldsymbol{\lambda}'(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$

- $\boldsymbol{\lambda}$为拉格朗日乘子向量(Lagrange Multiplier)，其经济含义为约束条件(比如资源约束)的影子价格(shadow price)。
- 如果$\hat{\boldsymbol{\lambda}} = 0$，约束条件完全不起作用(可无偿获取任意数量的资源)。
- 根据一阶条件（对$\tilde{\boldsymbol{\beta}}$求导）可知，最优的拉格朗日乘子向量$\hat{\boldsymbol{\lambda}}$等于对数似然函数在约束估计量$\hat{\boldsymbol{\beta}}^*$处的一阶偏导数（切线的斜率）。

$$\hat{\boldsymbol{\lambda}} = \frac{\partial \ln L(\hat{\boldsymbol{\beta}}^*)}{\partial \tilde{\boldsymbol{\beta}}} \equiv \begin{pmatrix} \dfrac{\partial \ln L(\hat{\boldsymbol{\beta}}^*)}{\partial \tilde{\beta}_1} \\ \vdots \\ \dfrac{\partial \ln L(\hat{\boldsymbol{\beta}}^*)}{\partial \tilde{\beta}_K} \end{pmatrix}$$

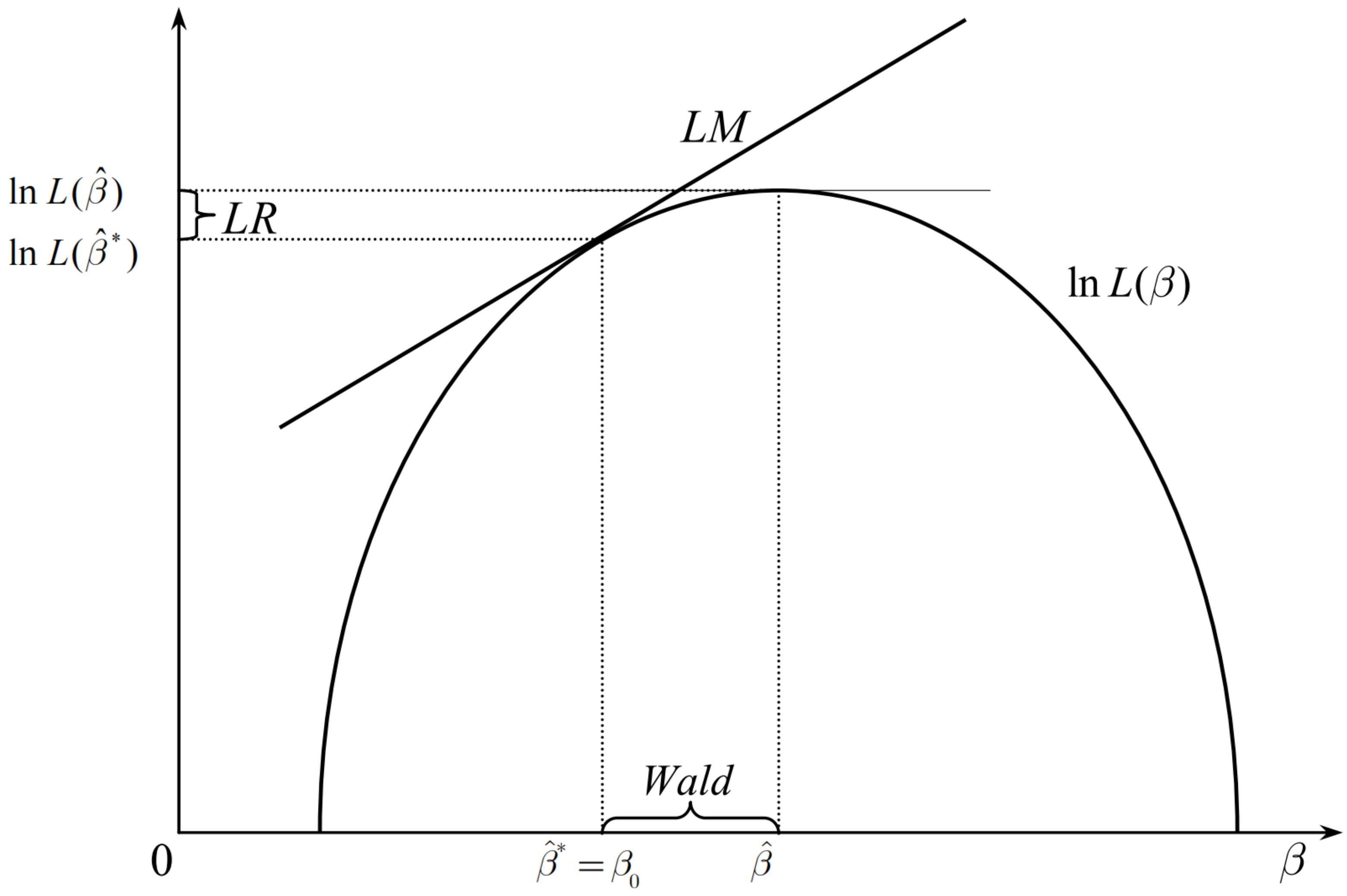
- 如 $\hat{\boldsymbol{\lambda}} \approx \mathbf{0}$，说明约束条件不“紧”(tight)或不是“硬约束”(binding constraint)，加上约束条件不会使似然函数的最大值下降很多，即原假设H_0很可能成立。(基本思想)
- 如果原假设H_0成立，则$(\hat{\boldsymbol{\lambda}} - \mathbf{0})$的绝对值不应很大。
- 以二次型来度量此距离，可得LM 统计量：

$$LM \equiv \hat{\boldsymbol{\lambda}}' \, [\mathrm{Var}(\hat{\boldsymbol{\lambda}})]^{-1} \, \hat{\boldsymbol{\lambda}} \xrightarrow{d} \chi^2(K)$$

- 其中，$Var(\hat{\boldsymbol{\lambda}})$为$\hat{\boldsymbol{\lambda}}$的协方差矩阵。
- 由于$\hat{\boldsymbol{\lambda}} = \frac{\partial lnL(\tilde{\boldsymbol{\beta}})}{\partial \tilde{\boldsymbol{\beta}}}$称为“得分函数”(score function)或“得分向量”(score vector)，此检验也称“得分检验”(score test)。

- 另一直观理解是，由于在无约束估计量$\widehat{\boldsymbol{\beta}}$处，$\frac{\partial lnL(\widehat{\boldsymbol{\beta}})}{\partial \widetilde{\boldsymbol{\beta}}}=\mathbf{0}$ (MLE的一阶条件)，故如原假设H_0成立，则在约束估计量$\widehat{\boldsymbol{\beta}}^*$处，$\frac{\partial lnL(\widehat{\boldsymbol{\beta}}^*)}{\partial \widetilde{\boldsymbol{\beta}}}\approx\mathbf{0}$，而LM 统计量反映的就是此接近程度。
- 对异方差与自相关所进行的nR^2形式的检验都来自LM 检验的推导。

- 这三类检验在大样本下渐近等价，从不同侧面考察同一事物。



- 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。
- 如果无约束估计更方便，常使用 Wald 检验(比如，对线性回归系数的显著性检验)；
- 如果有约束估计更方便，常使用LM 检验(比如，对异方差、自相关的检验)；
- 如果二者都方便，可使用LR检验(比如，对非线性回归方程的显著性检验)。

其他离散选择模型

- (1) 多值选择(multiple choices)：比如，对交通方式的选择(步行、骑车、自驾车、打的、地铁)，对不同职业的选择，对手机品牌的选择。
- (2) 计数数据(count data)：有时被解释变量只能取非负整数。比如，企业在某段时间内获得的专利数；某人在一定时间内去医院看病的次数；某省在一年内发生煤矿事故的次数。
- (3) 排序数据(ordered data)：有些离散数据有着天然的排序。比如，公司债券的评级(AAA, AA, A, B, C 级)，对“春节联欢晚会”的满意度(很满意、满意、不满意、很不满意)。

- 对于以上离散数据，一般也不宜直接进行 OLS 回归，主要估计方法仍为 MLE。
- 由于离散选择模型主要用于微观经济学的实证研究中，故是“微观计量经济学”(Microeconometrics)的重要组成部分。
- 除了离散数据外，微观计量经济学还关注的另一类数据类型为“受限被解释变量”(limited dependent variable)，即被解释变量的取值范围受到限制(包括断尾回归、归并回归与样本选择模型等)。