

高级计量经济学

理论经济学博士课程 2023-2024

Lecture 11: Maximum Likelihood Estimation

Davidson, R. & MacKinnon, J. (2009). Econometrics Theory and Methods. Oxford University Press.

黄嘉平

工学博士 经济学博士
深圳大学中国经济特区研究中心 讲师

办公室	粤海校区汇文楼1510
E-mail	huangjp@szu.edu.cn
Website	https://huangjp.com

似然函数和最大似然估计

似然函数

Likelihood function

假设随机向量 \mathbf{y} 服从参数为 $\boldsymbol{\theta}$ 的分布。在 $\boldsymbol{\theta}$ 已知的情况下， \mathbf{y} 的密度函数可以写成 $f(\mathbf{y} | \boldsymbol{\theta})$ 。

例如正态分布 $N(\mu, \sigma^2)$ 有两个参数，其密度函数为

$$f(y | \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y - \mu)^2}{2\sigma^2}\right)$$

而在参数估计中， \mathbf{y} 的观测值是已知的，但 $\boldsymbol{\theta}$ 是未知的。此时，我们可以将上述密度函数理解为 $\boldsymbol{\theta}$ 的函数，即

$$L(\boldsymbol{\theta} | \mathbf{y}) \equiv f(\mathbf{y} | \boldsymbol{\theta})$$

称为似然函数 (likelihood function) 。

之所以称为似然函数，是因为 $L(\boldsymbol{\theta} | \mathbf{y})$ 作为 $\boldsymbol{\theta}$ 的函数不满足概率密度函数的定义

最大似然估计

Maximum likelihood estimation

令似然函数最大的参数估计量

$$\hat{\boldsymbol{\theta}}_{\text{ML}} = \arg \max_{\boldsymbol{\theta}} L(\boldsymbol{\theta} \mid \mathbf{y})$$

称为最大似然估计量 (maximum likelihood estimator, MLE) 。我们也可以用对数似然函数 (log-likelihood function) $\ell(\boldsymbol{\theta} \mid \mathbf{y}) = \log L(\boldsymbol{\theta} \mid \mathbf{y})$ 替代似然函数。

当 y_t 之间相互独立时, \mathbf{y} 的联合密度是 y_t 的密度函数之积, 此时

$$L(\boldsymbol{\theta} \mid \mathbf{y}) = \prod_{t=1}^n f_t(y_t \mid \boldsymbol{\theta}) \Rightarrow \ell(\boldsymbol{\theta} \mid \mathbf{y}) = \sum_{t=1}^n \log f_t(y_t \mid \boldsymbol{\theta})$$

似然函数是多项的乘积, 其取值会变得非常小, 不适合计算机进行数值计算, 这是选择对数似然函数的原因之一。

对指数分布期望值的估计

假设 y_t 是服从指数分布的独立样本，即

$$f(y_t | \theta) = \theta e^{-\theta y_t}, \quad y_t > 0, \theta > 0$$

此时， y_t 的对数似然函数是 $\ell(\theta | y_t) = \log \theta - \theta y_t$ ，则

$$\ell(\theta | \mathbf{y}) = n \log \theta - \theta \sum_{t=1}^n y_t$$

θ 的最大似然估计量可以通过解一阶条件求得：

$$\frac{n}{\theta} - \sum_{t=1}^n y_t = 0 \Rightarrow \hat{\theta}_{\text{ML}} = \frac{n}{\sum_{t=1}^n y_t} = \frac{1}{\bar{y}} \quad (\text{可以确认二阶导为负})$$

在这个例子中， $\hat{\theta}_{\text{ML}}$ 和 MM 估计量是一致的。我们可以计算 $E[y_t] = \theta^{-1}$ （参考期中测验中的第三题），因此 $\hat{\theta}_{\text{MM}} = 1/\bar{y}$ 。ML 估计量比 MM 估计量更容易计算。

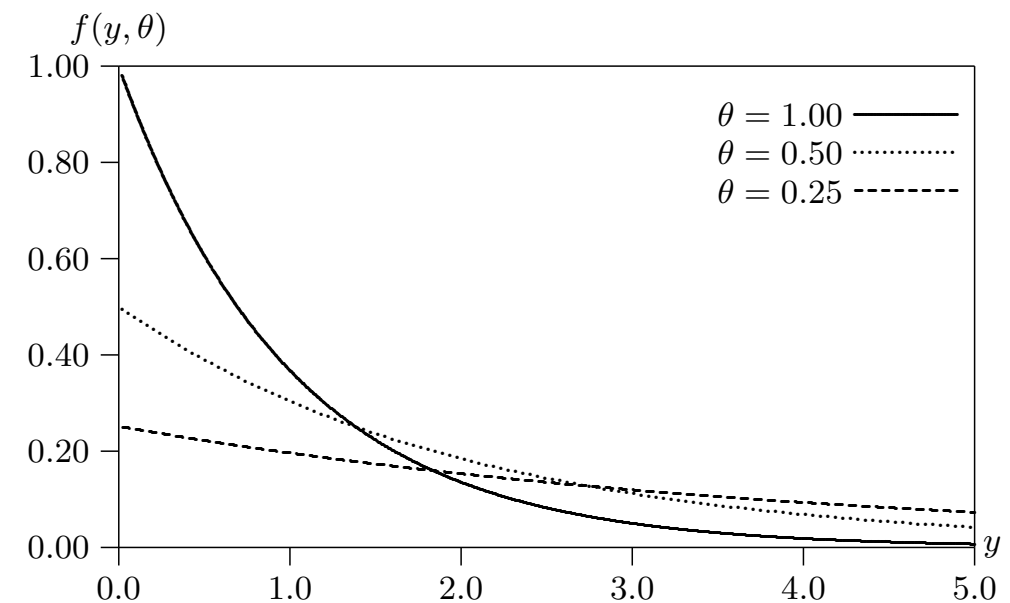


Figure 10.1 The exponential distribution

正态回归模型的最大似然估计

考虑正态回归模型

$$y = X\beta + u, \quad u \sim N(\mathbf{0}, \sigma^2 I)$$

此时 y_t 是独立样本，且服从 $N(X_t\beta, \sigma^2 I)$ ，其密度函数是

$$f_t(y_t | \beta, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_t - X_t\beta)^2}{2\sigma^2}\right)$$

y_t 的对数似然函数是

$$\ell_t(\beta, \sigma | y_t) = -\frac{1}{2} \log 2\pi - \frac{1}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (y_t - X_t\beta)^2$$

因此，

$$\begin{aligned} \ell(\beta, \sigma | y) &= -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=1}^n (y_t - X_t\beta)^2 \\ &= -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (y - X\beta)^\top (y - X\beta) \end{aligned}$$

正态回归模型的最大似然估计

为了求最大似然估计量 $\hat{\beta}_{\text{ML}}$ 和 $\hat{\sigma}_{\text{ML}}$ ，我们首先针对 $\ell(\beta, \sigma | y)$ 求关于 σ 的一阶条件，即

$$\frac{\partial}{\partial \sigma} \ell(\beta, \sigma | y) = -\frac{n}{\sigma} + \frac{1}{\sigma^3} (y - X\beta)^\top (y - X\beta) = 0$$

可得 $\hat{\sigma}^2(\beta) = \frac{1}{n} (y - X\beta)^\top (y - X\beta)$ 。然后将 $\hat{\sigma}^2(\beta)$ 代入 $\ell(\beta, \sigma | y)$ 可得

$$\ell^c(\beta | y) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \left(\frac{1}{n} (y - X\beta)^\top (y - X\beta) \right) - \frac{n}{2}$$

再针对 $\ell^c(\beta | y)$ 解关于 β 的一阶条件可得 $\hat{\beta}_{\text{ML}}$ 。此时 $\hat{\beta}_{\text{ML}} = \hat{\beta}_{\text{OLS}}$ 。

最后， $\hat{\sigma}_{\text{ML}}^2 = \hat{\sigma}^2(\hat{\beta}_{\text{ML}}) = \frac{1}{n} \sum_{t=1}^n \hat{u}_t^2$ ， $\hat{\sigma}_{\text{ML}} = \sqrt{\hat{\sigma}^2(\hat{\beta}_{\text{ML}})}$ 。

β 的 ML 估计量和 OLS 估计量一致是因为我们假设了 $u \sim N(0, \sigma^2 I)$ 。如果换成其他分布，则这个结果不成立，此时 ML 估计量具有渐进有效性。 σ^2 的 ML 估计量和 MM 估计量一致，我们已知它会低估 σ^2 。

ML 估计量的统计学性质

下面我们总结 ML 估计量的统计学性质：

- 如果 ML 估计量即是可识别的又是渐进可识别的，则它是一**一致估计量**。
- 如果 ML 估计量是对数似然函数的一阶条件的解（称为 Type 2 ML 估计量），则它的**极限分布是正态分布**。
- Type 2 ML 估计量在所有渐进非偏的 \sqrt{n} 一致估计量中满足**渐进有效**的。（在小样本中不具有有效性）

基于似然函数的大样本检验方法

这里简单介绍三种大样本下基于最大似然估计的检验方法。设 $H_0: \mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$ 中包含 r 个非线性假设。 $\mathbf{g}(\boldsymbol{\theta}) = [\partial \ell(\boldsymbol{\theta}) / \partial \theta_i]$ 被称为 score 函数, $\mathbf{I}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[\mathbf{g}(\boldsymbol{\theta}) \mathbf{g}^{\top}(\boldsymbol{\theta})]$ 被称为 information matrix。

- Likelihood ratio test (似然比检验)

$$\text{LR} \equiv 2[\ell(\hat{\boldsymbol{\theta}}_{\text{R}}) - \ell(\hat{\boldsymbol{\theta}}_{\text{U}})] = 2 \log[L(\hat{\boldsymbol{\theta}}_{\text{R}})/L(\hat{\boldsymbol{\theta}}_{\text{U}})] \stackrel{a}{\sim} \chi^2(r)$$

原理：在 H_0 下，有约束模型和无约束模型的最大似然函数值应该差不多

- Wald test (沃尔德检验)

$$W \equiv \mathbf{r}^{\top}(\hat{\boldsymbol{\theta}}_{\text{U}}) (\mathbf{R}(\hat{\boldsymbol{\theta}}_{\text{U}}) \widehat{\text{Var}}(\hat{\boldsymbol{\theta}}_{\text{U}}) \mathbf{R}^{\top}(\hat{\boldsymbol{\theta}}_{\text{U}}))^{-1} \mathbf{r}(\hat{\boldsymbol{\theta}}_{\text{U}}) \stackrel{a}{\sim} \chi^2(r), \quad \mathbf{R}(\boldsymbol{\theta}) = [\partial r_i(\boldsymbol{\theta}) / \partial \theta_j]$$

原理：在 H_0 下，无约束模型的估计量应该渐进地满足约束条件

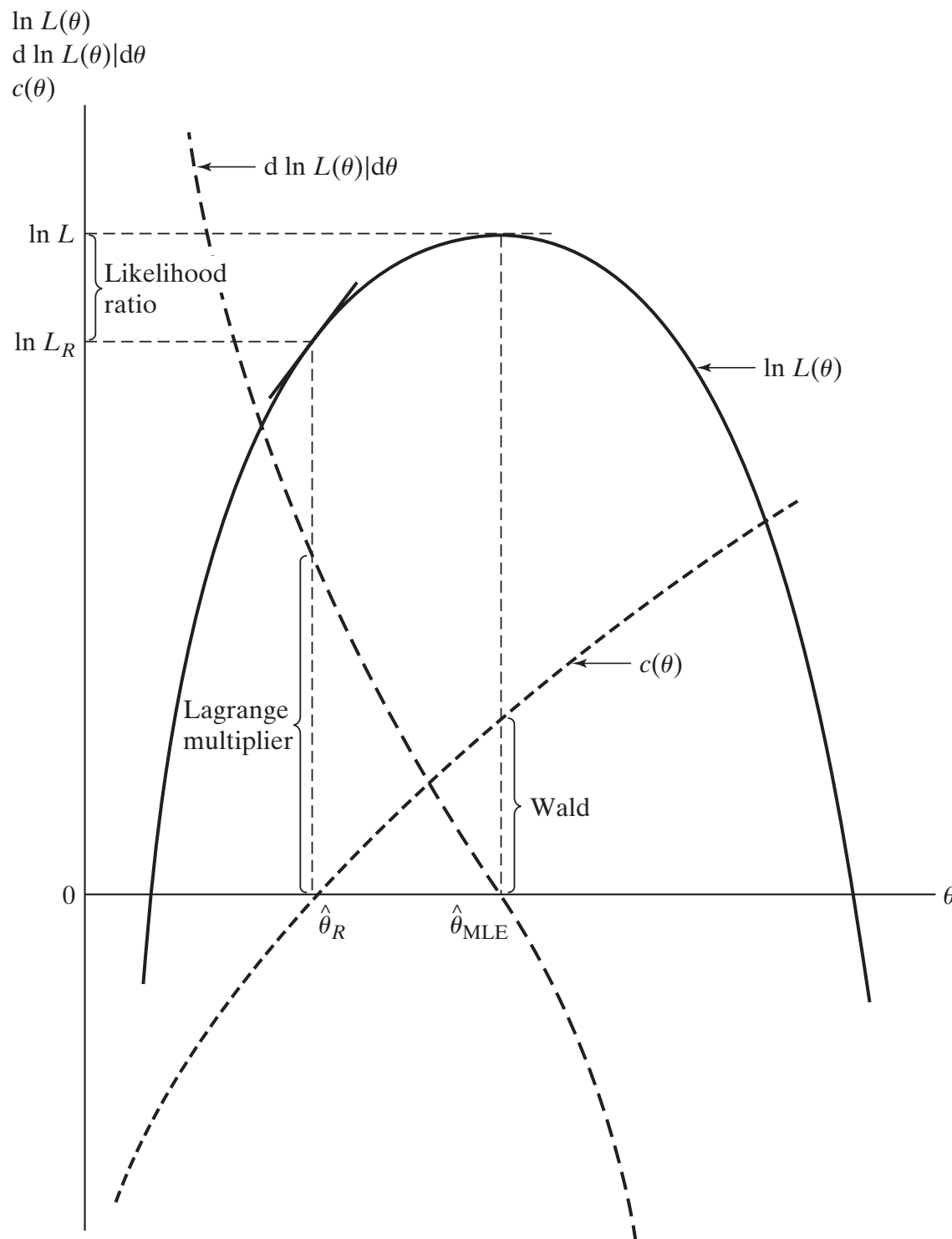
- Lagrange multiplier test (拉格朗日乘数检验) 或 Rao's score test

$$\text{LM} \equiv \mathbf{g}^{\top}(\hat{\boldsymbol{\theta}}_{\text{R}}) \mathbf{I}(\hat{\boldsymbol{\theta}}_{\text{R}})^{-1} \mathbf{g}(\hat{\boldsymbol{\theta}}_{\text{R}}) \stackrel{a}{\sim} \chi^2(r)$$

原理：在 H_0 下，有约束模型的估计量应该很接近真实值，但在 H_1 下应该差异较大

LR, W, and LM Tests

FIGURE 14.2 Three Bases for Hypothesis Tests.



Greene (2020), p.593.

- Wald 检验和 LM 检验可以看作是基于距离函数的检验。
- 三个检验统计量都渐进地服从 $\chi^2(r)$ 。
- 在线性模型和线性约束下，三个统计量都可以用 F 统计量表达，且满足
$$W > LR > LM$$
- 三个检验都是大样本检验，在小样本中的分布未知。
- 在实践中，可以根据计算的简易程度选择统计量。如果有约束模型和无约束模型都很容易估计，就可以选择 LR；如果无约束模型下的统计量相对容易计算（例如线性模型下的非线性约束），就可以选择 W；反之（例如约束条件可以消除模型的非线性）可选择 LM。

R 的典型回归结果：OLS

Kleiber & Zeileis, *Applied Econometrics with R*, Springer.: Section 3.2

```
> library(AER)
> data("CPS1988")
> cps_lm <- lm(log(wage) ~ experience + I(experience^2) + education + ethnicity, data = CPS1988)
> summary(cps_lm)
```

Call:
lm(formula = log(wage) ~ experience + I(experience^2) + education + ethnicity, data = CPS1988)

Residuals:

Min	1Q	Median	3Q	Max
-2.9428	-0.3162	0.0580	0.3756	4.3830

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.321e+00	1.917e-02	225.38	<2e-16 ***
experience	7.747e-02	8.800e-04	88.03	<2e-16 ***
I(experience^2)	-1.316e-03	1.899e-05	-69.31	<2e-16 ***
education	8.567e-02	1.272e-03	67.34	<2e-16 ***
ethnicityafam	-2.434e-01	1.292e-02	-18.84	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5839 on 28150 degrees of freedom
Multiple R-squared: 0.3347, Adjusted R-squared: 0.3346
F-statistic: 3541 on 4 and 28150 DF, p-value: < 2.2e-16

R 的典型回归结果: ML

Kleiber & Zeileis, *Applied Econometrics with R*, Springer.: Section 3.2

```
> library(stats4)
> LL <- function(beta_0, beta_ex, beta_ex2, beta_edu, beta_eth, sigma){
  -sum(dnorm(log(CPS1988$wage), (beta_0 + beta_ex*CPS1988$experience +
    beta_ex2*I(CPS1988$experience^2) + beta_edu*CPS1988$education +
    beta_eth*as.integer(CPS1988$ethnicity)), sigma, log = TRUE))
}
> cps_ML <- mle(LL, start = list(beta_0=0, beta_ex=0, beta_ex2=0, beta_edu=0, beta_eth=0, sigma=10))
> summary(cps_ML)
```

Maximum likelihood estimation

Call:

```
mle(minuslogl = LL, start = list(beta_0 = 0, beta_ex = 0, beta_ex2 = 0,
  beta_edu = 0, beta_eth = 0, sigma = 10))
```

Coefficients:

	Estimate	Std. Error
beta_0	4.565599247	2.436695e-02
beta_ex	0.077455053	8.799206e-04
beta_ex2	-0.001315682	1.898479e-05
beta_edu	0.085634316	1.272004e-03
beta_eth	-0.243661209	1.291627e-02
sigma	0.583852317	2.460192e-03

-2 log L: 49602.68

二值因变量模型

当 y 是二值变量时

二值因变量模型也称为二值响应模型 (binary response model)，是指因变量 $y_t \in \{0,1\}$ 的回归模型。此时，

$$E[y_t | \Omega_t] = \Pr(y_t = 1 | \Omega_t)$$

如果假设线性模型 $E[y_t | \Omega_t] = X_t\beta$ ，则无法满足概率取值在 0 和 1 之间的条件，因此通常不采用线性回归模型。

二值响应模型的一般表达为

$$E[y_t | \Omega_t] = F(X_t\beta) \quad F \text{ 称为变换函数 (transformation function)}$$

其中 F 应满足 $F(-\infty) = 0, F(\infty) = 1, f(x) = \frac{d}{dx}F(x) > 0$ 。

常用的二值响应模型包括 probit 模型和 logit 模型。

Probit 模型

Probit 模型是假设

$$F(x) = \Phi(x) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{1}{2}x^2\right) dx$$

$\Phi(x)$ 是标准正态分布的 CDF

Probit 模型可以从下面的潜在变量 (latent variable) 模型推导出。令

$$y_t^\circ = X_t \boldsymbol{\beta} + u_t, \quad u_t \sim \text{NID}(0, 1),$$

由于无法观测 y_t° 的取值, u_t 的方差也无法识别, 因此假设其为 1

并假设我们只能观测 y_t° 的符号。而 y_t 的取值定义为

$$y_t = \begin{cases} 1 & \text{if } y_t^\circ > 0 \\ 0 & \text{if } y_t^\circ \leq 0 \end{cases}$$

此时,

$$\begin{aligned} \Pr(y_t = 1) &= \Pr(y_t^\circ > 0) = \Pr(X_t \boldsymbol{\beta} + u_t > 0) \\ &= \Pr(u_t > -X_t \boldsymbol{\beta}) = \Pr(u_t \leq X_t \boldsymbol{\beta}) = \Phi(X_t \boldsymbol{\beta}) \end{aligned}$$

Logit 模型

Logit 模型是假设

$$F(x) = \Lambda(x) \equiv \frac{1}{1 + e^{-x}} = \frac{e^x}{1 + e^x} \quad \Lambda(x) \text{ 称为逻辑函数 logistic function}$$

Logit 模型可以利用优势比 (odds ratio, 或称赔率) 推导出。令

$$\log\left(\frac{\Pr(y_t | \Omega_t)}{1 - \Pr(y_t | \Omega_t)}\right) = X_t \beta, \quad \text{优势比就是获胜概率比上失败概率}$$

可得

$$\Pr(y_t | \Omega_t) = \frac{\exp(X_t \beta)}{1 + \exp(X_t \beta)} = \Lambda(X_t \beta)$$

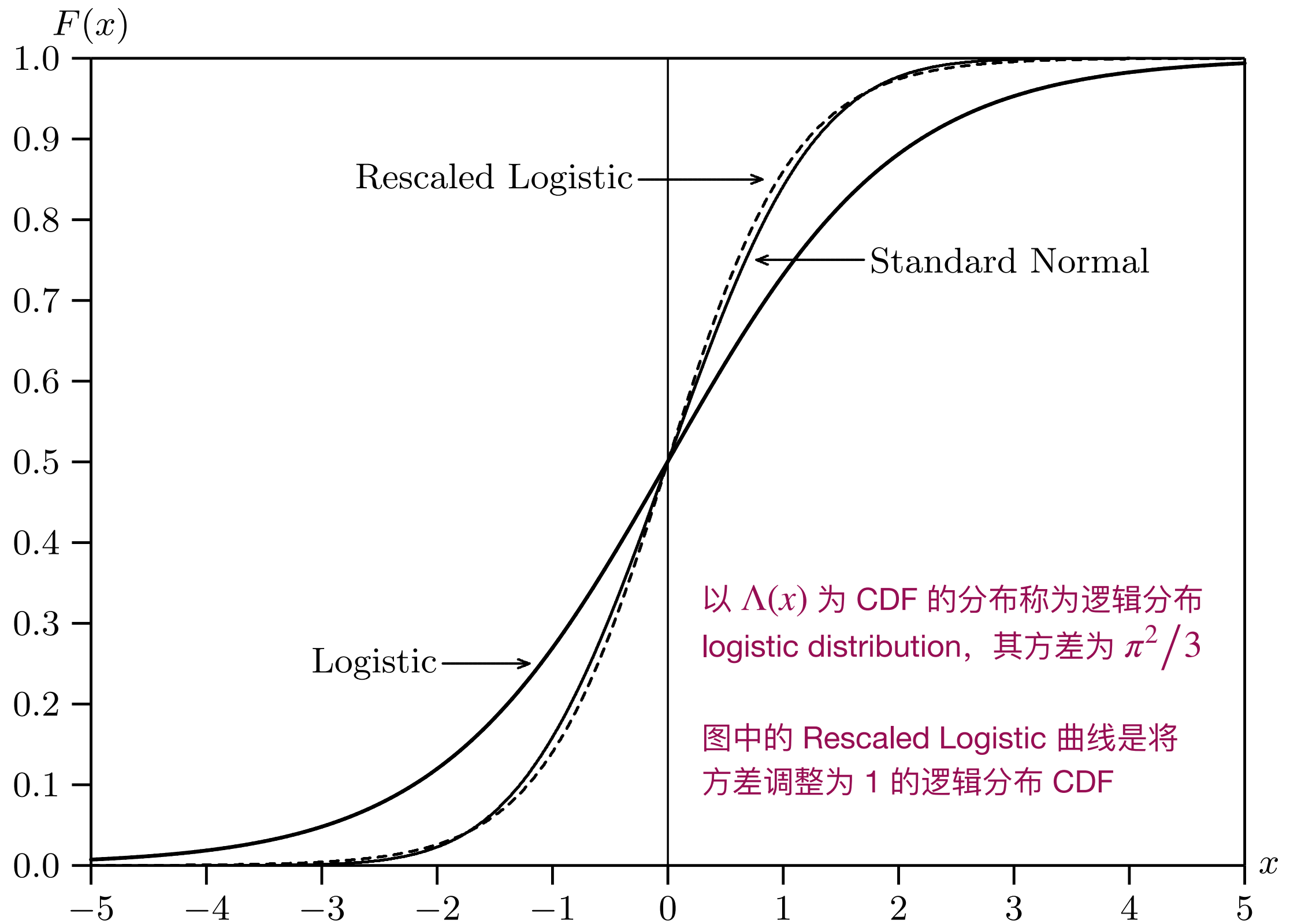


Figure 11.1 Alternative choices for $F(x)$

二值响应模型的估计

线性模型（此时称为线性概率模型 linear probability model）可以用 OLS 估计。

Probit 和 logit 模型一般用 ML 估计。由于 y_t 是离散变量，因此似然函数为概率函数而不是密度函数。根据模型的定义，

$$\ell_t(\boldsymbol{\beta} \mid y_t = 1) = \log \Pr(y_t = 1 \mid \boldsymbol{\beta}) = \log F(X_t \boldsymbol{\beta})$$

$$\ell_t(\boldsymbol{\beta} \mid y_t = 0) = \log \Pr(y_t = 0 \mid \boldsymbol{\beta}) = \log(1 - F(X_t \boldsymbol{\beta}))$$

此时 \mathbf{y} 的对数似然函数为

$$\ell(\boldsymbol{\beta} \mid \mathbf{y}) = \sum_{t=1}^n \left(y_t \log F(X_t \boldsymbol{\beta}) + (1 - y_t) \log(1 - F(X_t \boldsymbol{\beta})) \right)$$

对于 probit 和 logit 模型， $\ell(\boldsymbol{\beta} \mid \mathbf{y})$ 是凹函数，通过解一阶条件可以获得 ML 估计量。

一阶条件和非线性回归模型 $y_t = F(X_t \boldsymbol{\beta}) + v_t$ 的加权最小二乘估计的一阶条件一致，观测值 t 的权重为 $(F(X_t \boldsymbol{\beta})(1 - F(X_t \boldsymbol{\beta})))^{-1/2}$ 。

R 的回归结果: probit

Gerfin, M. (1996). Parametric and semi parametric estimation of the binary response model of labour market participation. *Journal of Applied Econometric*, 11:321-340.

```
> fit_probit = glm(LFP ~ LNNLINC + AGE + I(AGE^2) + EDUC + NYC + NOC + FOREIGN,  
  data = participation, family = binomial(link = "probit"))  
> summary(fit_probit)
```

Call:

```
glm(formula = LFP ~ LNNLINC + AGE + I(AGE^2) + EDUC + NYC + NOC +  
  FOREIGN, family = binomial(link = "probit"), data = participation)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	3.74909	1.40695	2.665	0.00771	**
LNNLINC	-0.66694	0.13196	-5.054	4.33e-07	***
AGE	2.07530	0.40544	5.119	3.08e-07	***
I(AGE^2)	-0.29434	0.04995	-5.893	3.79e-09	***
EDUC	0.01920	0.01793	1.071	0.28428	
NYC	-0.71449	0.10039	-7.117	1.10e-12	***
NOC	-0.14698	0.05089	-2.888	0.00387	**
FOREIGN	0.71437	0.12133	5.888	3.92e-09	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1203.2 on 871 degrees of freedom
Residual deviance: 1017.2 on 864 degrees of freedom
AIC: 1033.2

Number of Fisher Scoring iterations: 4

R 的回归结果: logit

Gerfin, M. (1996). Parametric and semi parametric estimation of the binary response model of labour market participation. *Journal of Applied Econometric*, 11:321-340.

```
> fit_logit = glm(LFP ~ LNNLINC + AGE + I(AGE^2) + EDUC + NYC + NOC + FOREIGN,  
  data = participation, family = binomial(link = "logit"))  
> summary(fit_logit)
```

Call:

```
glm(formula = LFP ~ LNNLINC + AGE + I(AGE^2) + EDUC + NYC + NOC +  
  FOREIGN, family = binomial(link = "logit"), data = participation)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	6.19639	2.38309	2.600	0.00932	**
LNNLINC	-1.10409	0.22571	-4.892	1.00e-06	***
AGE	3.43661	0.68789	4.996	5.86e-07	***
I(AGE^2)	-0.48764	0.08519	-5.724	1.04e-08	***
EDUC	0.03266	0.02999	1.089	0.27611	
NYC	-1.18575	0.17202	-6.893	5.46e-12	***
NOC	-0.24094	0.08446	-2.853	0.00433	**
FOREIGN	1.16834	0.20384	5.732	9.94e-09	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1203.2 on 871 degrees of freedom
Residual deviance: 1017.6 on 864 degrees of freedom
AIC: 1033.6

Number of Fisher Scoring iterations: 4