计量经济学

第十讲：二值因变量回归

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2613
E-mail　huangjp@szu.edu.cn
Website　https://huangjp.com

主要内容

- 线性概率模型
- Probit 和 logit 回归
  - Probit 和 logic 模型
  - 最大似然估计

二值因变量和线性概率模型

HMDA 数据集

- HMDA (Home Mortgage Disclosure Act，房屋抵押公开法) 数据集是波士顿联邦储备银行公布的波士顿地区 1990 年抵押贷款申请的数据。

  数据文件：`hmda_sw1.csv`　说明文件：`hmda.docx`

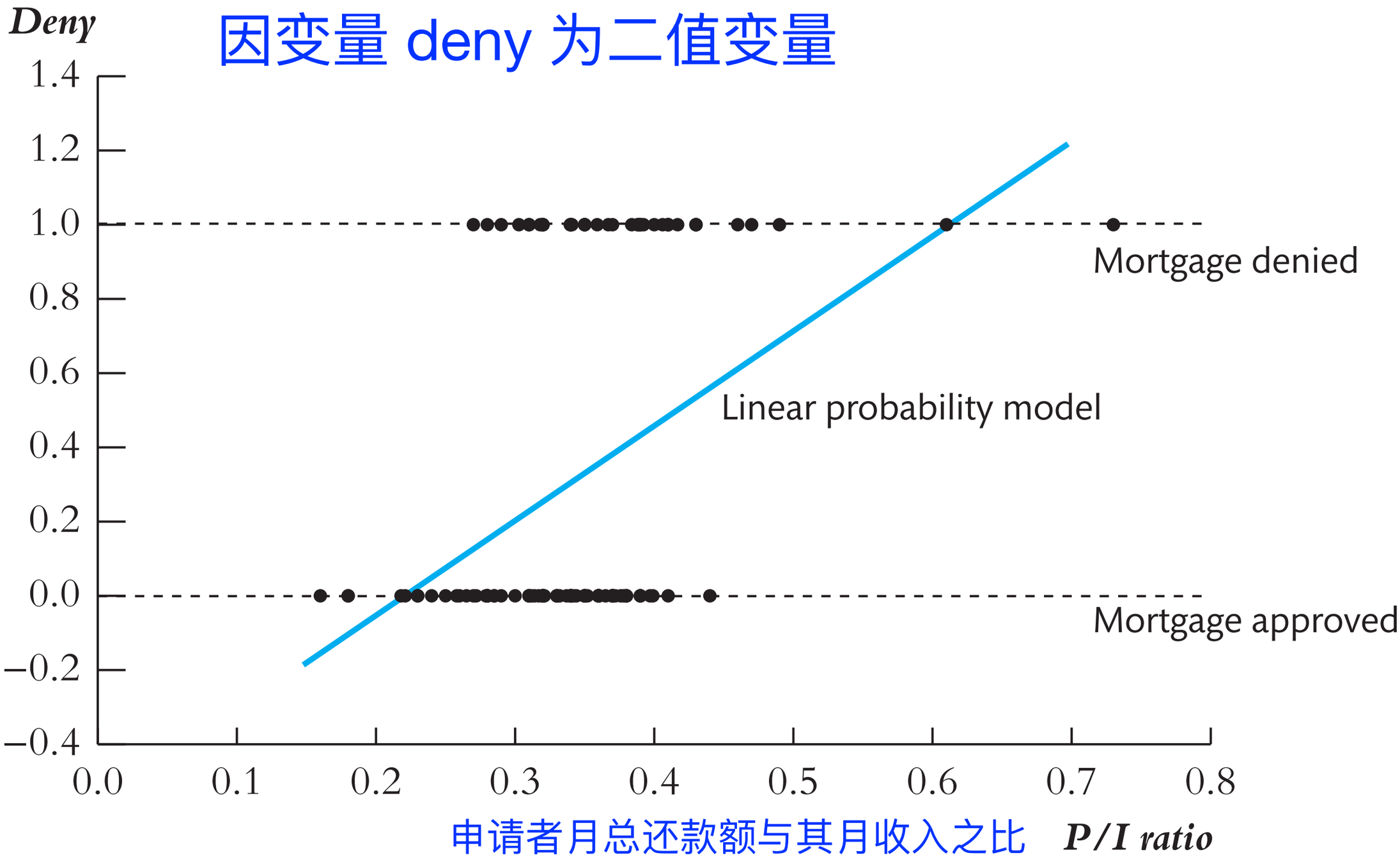
- 数据集共包含 62 个变量，包括数值和字符，同时也包含缺失值（NA值和空白）。

Missing value in the file	
Numeric data:	999999.375
String data:	NA

Missing value after importing	
Numeric:	999999.375
All string:	NA
Partial string:	(blank)

* Learn the `setmiss` command.

什么因素决定贷款申请是否被拒绝?



因变量为二值变量时的总体回归函数

- 当总体回归函数为线性时

$$\begin{aligned}&\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m \\ &= \mathrm{E}(Y \mid X_{1i} = x_1, X_{2i} = x_2, \ldots, X_{mi} = x_m)\end{aligned}$$

- 若因变量 Y 为二值变量，则有

$$\begin{aligned}\mathrm{E}(Y) &= 0 \times \Pr(Y = 0) + 1 \times \Pr(Y = 1) \\ &= \Pr(Y = 1)\end{aligned}$$

$$\Rightarrow \quad \mathrm{E}(Y \mid X_1, \ldots, X_m) = \Pr(Y = 1 \mid X_1, \ldots, X_m)$$

线性概率模型

Linear probability model

- 因变量为二值变量时的多元线性回归模型被称为线性概率模型（linear probability model）

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_m X_{mi} + u_i$$

$$\begin{aligned} \Rightarrow \quad & \Pr(Y = 1 \mid X_1, \ldots, X_m) \\ & = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m \end{aligned}$$

- 系数 β_i 是在其他回归变量保持不变时，X_i 变化一个单位引起的 $Y = 1$ 的概率变化。

- 回归系数依旧可以用 OLS 估计进行估值。

在 gretl 中定义虚拟变量

```
open "@workdir/data/hmda_sw1.csv"
rename s7 deny
dummify deny
```

5		s6	
6		deny	
	65	Ddeny_1	dummy for deny = 1
	66	Ddeny_2	dummy for deny = 2
	67	Ddeny_3	dummy for deny = 3
7		s9	
8		s11	

```
rename 65 Doriginate
rename 66 Dnotaccepted
rename 67 Ddeny
```

自上至下依次为:
同意、同意但申请者未接受、拒绝

5		s6	
6		deny	
	65	Doriginate	dummy for deny = 1
	66	Dnotaccepted	dummy for deny = 2
	67	Ddeny	dummy for deny = 3
7		s9	
8		s11	

练习

- 用 P/I ratio 回归 Ddeny （复制书中 11.1）

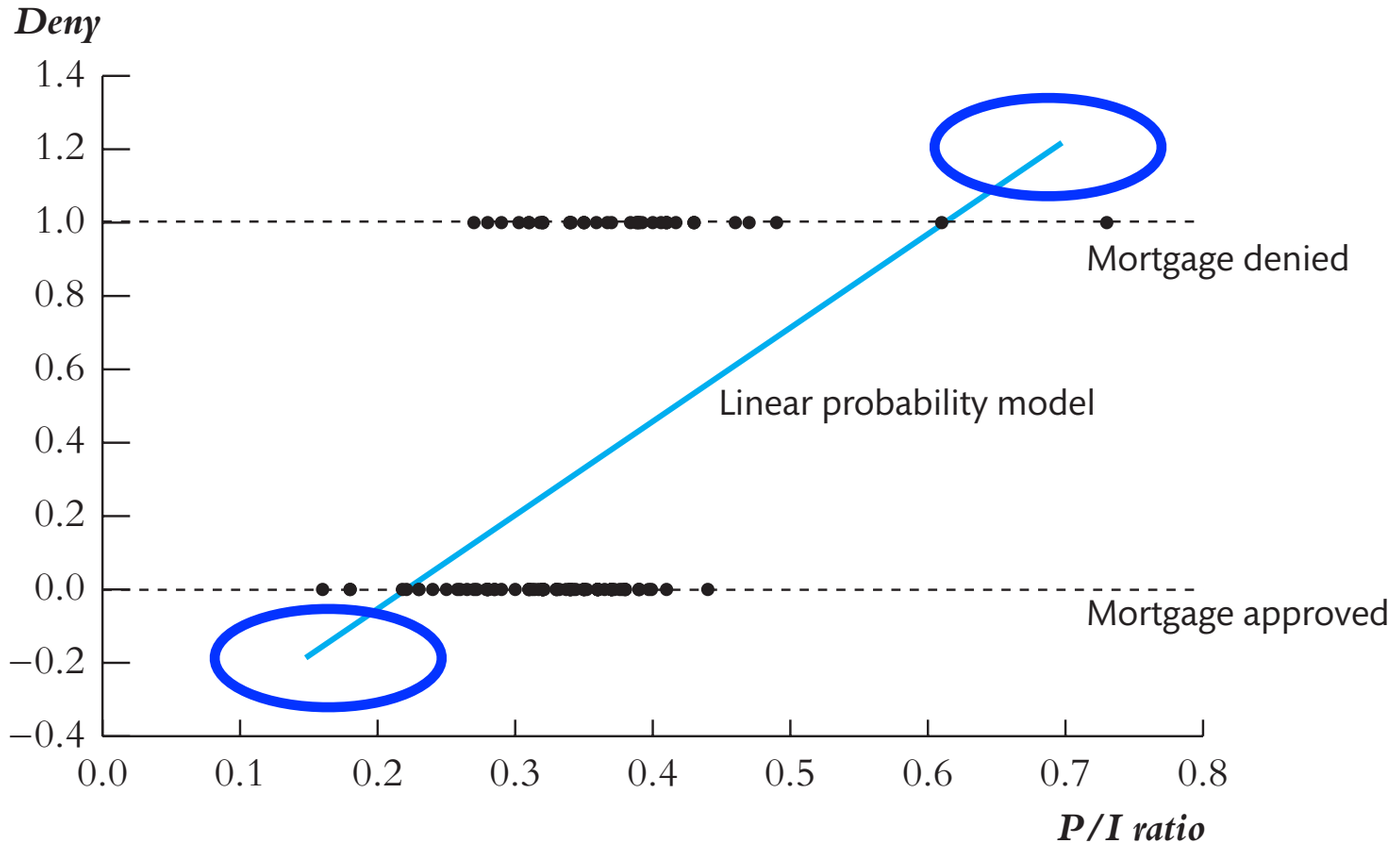
```
rename s46 piratio
genr Npiratio = piratio / 100
ols Ddeny const Npiratio --robust
```

```
Model 1: OLS, using observations 1-2380
Dependent variable: Ddeny
Heteroskedasticity-robust standard errors, variant HC1

            coefficient   std. error     z       p-value
  ---------------------------------------------------------
  const      -0.0799096    0.0319666    -2.500   0.0124   **
  Npiratio    0.603535     0.0984826     6.128   8.88e-10 ***
```

- 尝试用 P/I ratio 和 black 两个变量回归 Ddeny （复制书中 11.3）

线性概率模型的缺点

- 概率应当在 0 和 1 之间取值，而线性概率模型在 X 较大或较小时很容易超出这个范围



- 因此，我们需要考虑非线性模型，即 probit 和 logit 回归模型。

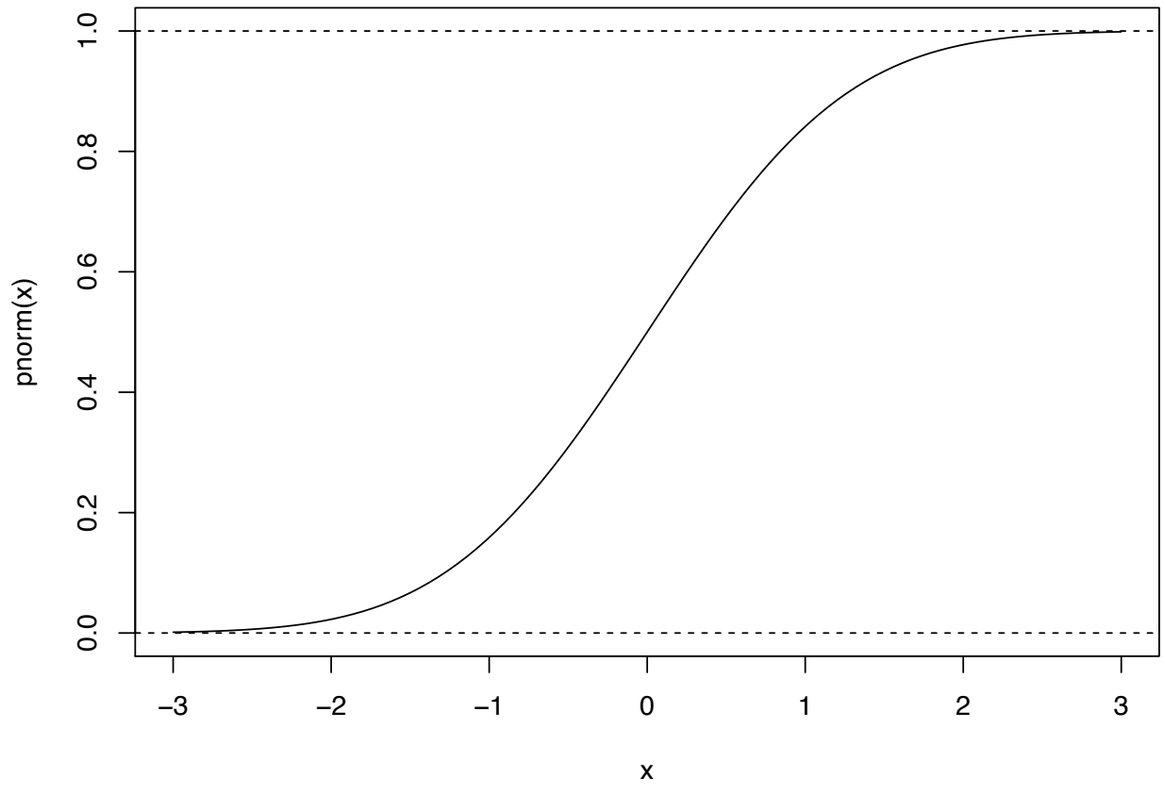
Probit 和 logit 回归

Probit 回归

The probit regression

- Probit 回归利用了标准正态分布的分布函数为 S 型且取值在 0 和 1 之间的特点，即

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



- Probit 回归模型为

$$\Pr(Y = 1 \mid X_1, \ldots, X_m) = \Phi(\beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m)$$

Probit 回归的预测与估计

- 如何预测 $Y = 1$ 的概率
  1. 针对给定的 $X_1, \ldots, X_m$ 和系数估计值 $\hat{\beta}_0, \ldots, \hat{\beta}_m$，计算 $z = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots, + \hat{\beta}_m X_m$
  2. 计算累积概率 $\Phi(z)$
- 系数的估计可以用非线性最小二乘估计，或者**最大似然估计（maximum likelihood estimation）**。

  最大似然估计量是令**似然函数**（即给定总体分布时获得现有样本的概率）最大的参数取值。例如，当总体只包含一个变量并服从伯努利分布，且取值为1的概率为 p 时，p 为未知参数，此时若随机样本中包含两个观测值 0 和 1，则似然函数为 $(p-1)p$。令此似然函数取值最大的 p 的取值为 0.5。详细内容参照 11.3.2 节和附录 11.2。

- 最大似然估计量具有一致性，在大样本下服从正态抽样分布，且比非线性 OLS 估计量更有效（方差更小），因此实践中多采用最大似然估计。

回归系数的解释

- 在 probit 模型中，回归系数无法直接解释成回归变量的变化引起的因变量的概率变化。

- 当 X_k 增加一个单位时，对应的 z 值增加 β_k，进而反映为 $Y = 1$ 的概率的增加。但是概率的增加为非线性的，因此比较直观的解释是，针对一个或多个回归变量的取值计算概率的预测值（或者预测值的变化）。

- 如果只有一个回归变量，也可以用图形表示回归函数。

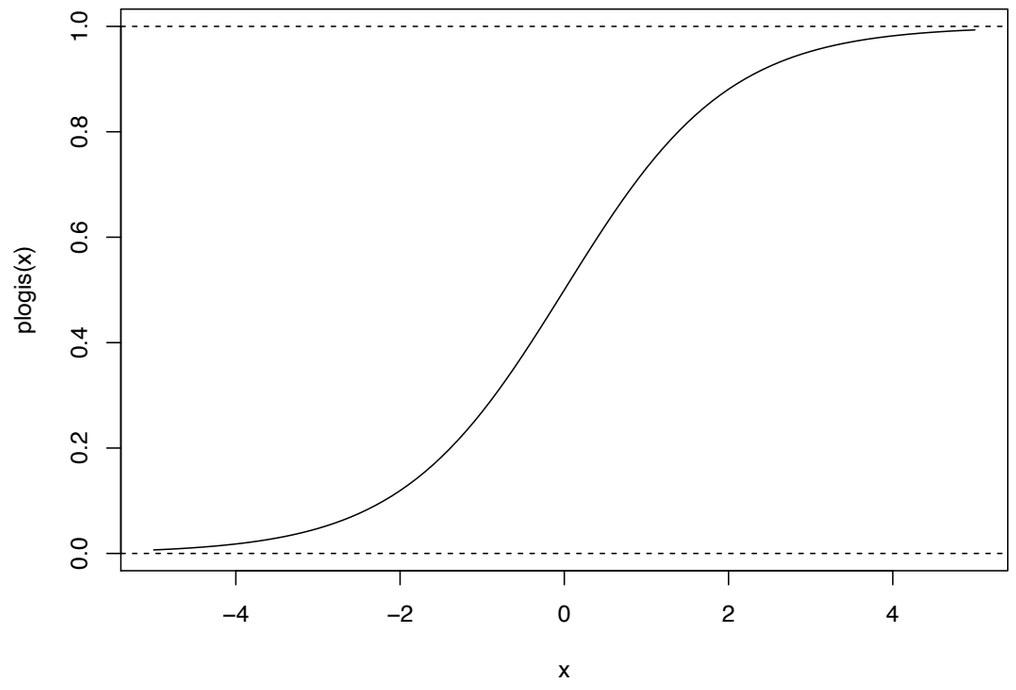
Logit 回归

The logit regression

- Logit 回归利用了 logistic 函数也具备 S 型且取值在 0 和 1 之间的性质

$$\Pr(Y = 1 \mid X_1, \ldots, X_m) = \frac{1}{1 + \exp\left(-\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m\right)\right)}$$

logistic 函数：$\dfrac{1}{1 + e^{-x}}$



- Logit 回归也可用最大似然法估计。

Logit 函数的概率解释

- 当二值变量取值为 1 的概率为 p 时，其优势比（odds ratio）为 $\frac{p}{1-p}$。
- 定义 x 为优势比的自然对数，可得

$$x = \ln(\frac{p}{1-p}) \quad \Rightarrow \quad e^x = \frac{p}{1-p}$$

$$\Rightarrow \quad p = \frac{e^x}{e^x + 1} = \frac{1}{1 + e^{-x}}$$

gretl 中的 `probit` 和 `logit` 命令

- gretl 中的 probit 回归和 logit 回归命令分别为 `probit` 和 `logit` 。
- 需要注意的是，这两个命令默认不输出系数检验的 p 值，而是输出每个回归变量取均值时对应的斜率。如果要输入 p 值，则应添加 **`--p-values`**
- 因变量 Y 的预测值为 0 或 1 ，即

$$\hat{Y}_i = \begin{cases} 1 & \text{if the predicted prob. exceeds } 0.5 \\ 0 & \text{otherwise} \end{cases}$$

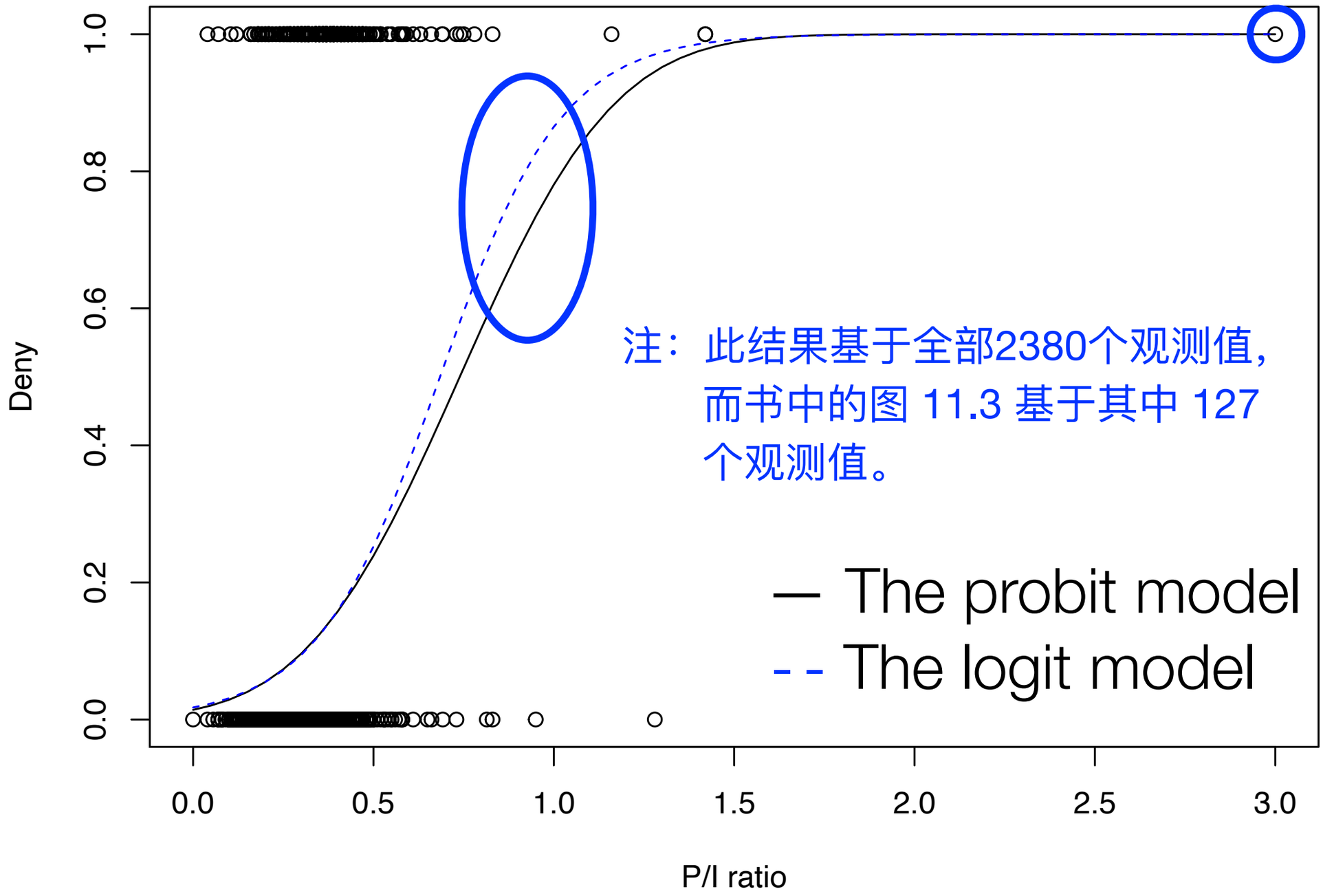
```
probit Ddeny const Npiratio --robust --p-values
```

```
Model 3: Probit, using observations 1-2380
Dependent variable: Ddeny
QML standard errors                                  此结果复制了书中 (11.7)

             coefficient   std. error      z      p-value
  ---------------------------------------------------------
  const       -2.19416      0.164941    -13.30    2.23e-40 ***
  Npiratio     2.96791      0.465224      6.380   1.78e-10 ***

Mean dependent var   0.119748   S.D. dependent var   0.324735
McFadden R-squared   0.046203   Adjusted R-squared   0.043910
Log-likelihood      -831.7923   Akaike criterion     1667.585
Schwarz criterion    1679.134   Hannan-Quinn         1671.788

Number of cases 'correctly predicted' = 2099 (88.2%)
f(beta'x) at mean of independent vars = 0.191
Likelihood ratio test: Chi-square(1) = 80.5859 [0.0000]

            Predicted
                0      1
  Actual 0   2091      4
         1    277      8

Test for normality of residual -
  Null hypothesis: error is normally distributed
  Test statistic: Chi-square(2) = 15.772
  with p-value = 0.000375971
```

练习

- 尝试复制书中 (11.8) 和 (11.10)。
- 分别计算 P/I ratio = 0.3 时白人和黑人申请者的失败概率。

Probit 和 logit 回归结果的差异



拟合优度

- **正确预测比例（fraction correctly predicted）**

  若 $Y_i = 1$ 且其概率预测值超过 50%，或者 $Y_i = 0$ 且其概率预测值小于 50%，则称对 Y_i 的预测正确。

- **伪 R^2 （McFadden's pseudo-R^2）**

$$\text{pseudo-}R^2 = 1 - \frac{\ell(\hat{\beta})}{\ell(\bar{y})}$$

  其中 $\ell(\hat{\beta})$ 是拟合模型的对数似然函数，$\ell(\bar{y})$ 是仅包含常数项的模型的对数似然函数。

```
probit Ddeny const Npiratio --robust --p-values
```

```
Model 3: Probit, using observations 1-2380
Dependent variable: Ddeny
QML standard errors

             coefficient   std. error      z      p-value
  ---------------------------------------------------------
  const       -2.19416      0.164941    -13.30    2.23e-40 ***
  Npiratio     2.96791      0.465224      6.380   1.78e-10 ***

Mean dependent var   0.119748   S.D. dependent var   0.324735
McFadden R-squared   0.046203   Adjusted R-squared   0.043910
Log-likelihood      -831.7923   Akaike criterion     1667.585
Schwarz criterion    1679.134   Hannan-Quinn         1671.788

Number of cases 'correctly predicted' = 2099 (88.2%)
f(beta'x) at mean of independent vars = 0.191
Likelihood ratio test: Chi-square(1) = 80.5859 [0.0000]

            Predicted
                0      1
  Actual 0   2091      4
         1    277      8

Test for normality of residual -
  Null hypothesis: error is normally distributed
  Test statistic: Chi-square(2) = 15.772
  with p-value = 0.000375971
```

线性概率、probit 和 logit 模型的比较

- 线性概率模型简单且容易解释，但无法体现总体回归函数的非线性性质。
- Probit 和 logit 模型为非线性模型，因此更难解释回归系数。
- 线性概率模型用 OLS 估值，而 probit 和 logit 模型用最大似然法估值。如果数据量大，最大似然估计会很花时间（没有简单的公式，需要进行数值计算）。
- Probit 和 logit 回归的结果往往比较相似。在计算中 logit 模型比 probit 略容易。
- 如果不知道如何选择，可以依次尝试。

其他受限因变量模型

Other limited dependent variable models

- **删失（censored）回归模型**

  删失数据是说某个连续变量的实际观测数据被限制在某个固定的范围内，因此超出范围的数据被记录为上限或下限值。针对删失数据有效的模型为 tobit 回归模型。

- **样本选择（sample selection）模型**

  当超出限定范围的数据不是在记录时被修改，而是没有被记录的时候，该数据称为截断（truncated）数据。截断回归模型是样本选择模型的一种。

- **计数数据（count data）**

  计数数据是指因变量为计数数字（发生次数）。针对计数数据的常用模型包括Poisson 回归和负二项（negative binomial）回归。

- **有序因变量（ordered response）**

  有序因变量数据是指互斥的但有自然排序的定性数据，如最终学历为高中、大学、硕士、博士。可用有序 probit 模型（ordered probit model）进行分析。

- **离散选择（discrete choice）数据**

  离散选择数据是无序的定性数据，可用多项 probit（multinomial probit）或多项 logit（multinomial logit）回归模型。

拓展阅读

1. Munnell, A. et al. (1996). Mortgage Lending in Boston: Interpreting HDMA Data, *American Economic Review*, 86:25-53.

2. Ladd, H. (1998). Evidence on Discrimination in Mortgage Lending, *Journal of Economic Perspectives*, 12:41-62.

3. Wooldridge, J. (2010). *Economic Analysis of Cross Section and Panel Data*, 2nd ed. MIT Press.