

# Econometrics 1 *Applied Econometrics with R*

## Lecture 11: Binary Dependent Variable

---

黄嘉平

中国经济特区研究中心 讲师

办公室：文科楼2613

E-mail: [huangjp@szu.edu.cn](mailto:huangjp@szu.edu.cn)

Tel: (0755) 2695 0548

Website: <https://huangjp.com>

# Regression with a Binary Dependent Variable

# The HMDA data

---

- HMDA (Home Mortgage Disclosure Act) data are data that related to mortgage applications filed in the Boston area in 1990.
- This data is contained in the AER package.

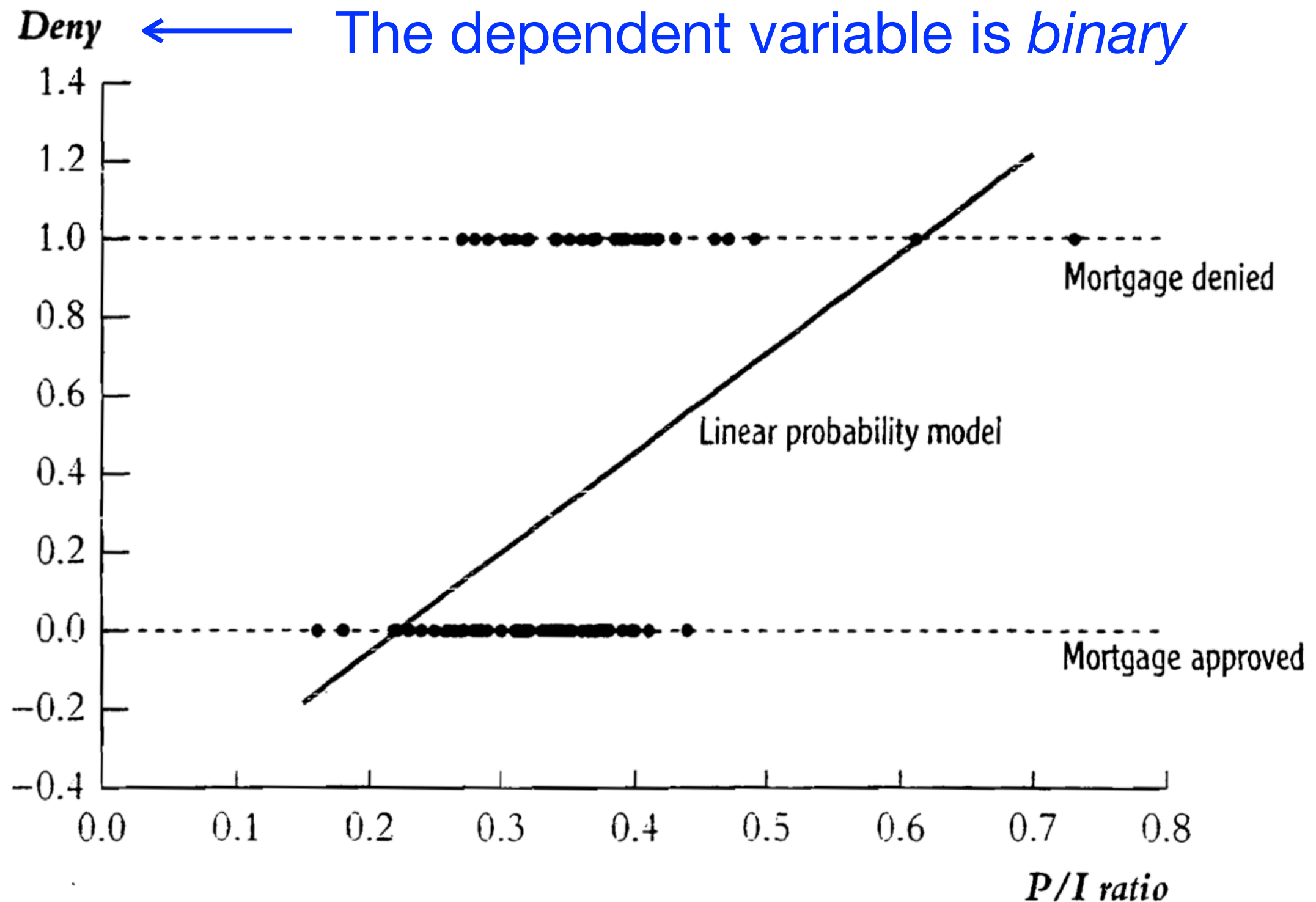
```
> library("AER")  
> data("HMDA")
```

# The HMDA data

---

deny	Factor. Whether the mortgage was denied. Yes or no.
pirat	P/I ratio. Ratio of total monthly debt payments to total monthly income.
hirat	Ratio of monthly housing expenses to total monthly income.
lvrat	Ratio of size of loan to assessed value of property.
chist	Factor. Historical consumer credit score. 6 levels.
mhst	Factor. Historical mortgage credit score. 4 levels.
phist	Factor. Historical public bad credit record. Yes or no.
unemp	1989 Massachusetts unemployment rate in applicant's industry.
selfemp	Factor. Whether the individual is self-employed. Yes or no.
insurance	Factor. Whether the individual was denied mortgage insurance. Yes or no.
condomin	Factor. Whether the unit is a condominium. Yes or no.
afam	Factor. Whether the individual is African-American. Yes or no.
single	Factor. Whether the individual is single. Yes or no.
hschool	Factor. Whether the individual has a high-school diploma. Yes or no.

# What determines whether a mortgage application is denied?



# Regression with a binary dependent variable

---

- The population regression function of a lineal model

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

- Regression with a binary variable

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

$$\Rightarrow E(Y \mid X_1, \dots, X_m) = \Pr(Y = 1 \mid X_1, \dots, X_m)$$

# The linear probability model

---

- The linear probability model

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

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

- The regression coefficient  $\beta_1$  is the change in the probability  $Y = 1$  associated with a unit change in  $X_1$ , holding constant the other regressors, and so forth for  $\beta_2, \dots, \beta_m$
- The regression coefficients can be estimated by OLS.

# A note on the “factor” class in R

---

- Run the following codes

```
> a <- c(0, 1, 1, 0, 0)
> b <- as.factor(a)
```

```
> b
[1] 0 1 1 0 0
Levels: 0 1
```

## Values

a	num [1:5] 0 1 1 0 0
b	Factor w/ 2 levels "0","1": 1 2 2 1 1



# A note on the “factor” class in R

---

- Translating a factor into a numeric vector

A “bad” way:

```
> c <- as.numeric(b)
```

Some “good” ways:

```
> d <- as.numeric(b) - 1
```

```
> f <- as.numeric(levels(b))[b]
```

```
> g <- as.numeric(as.character(b))
```

# Practice

---

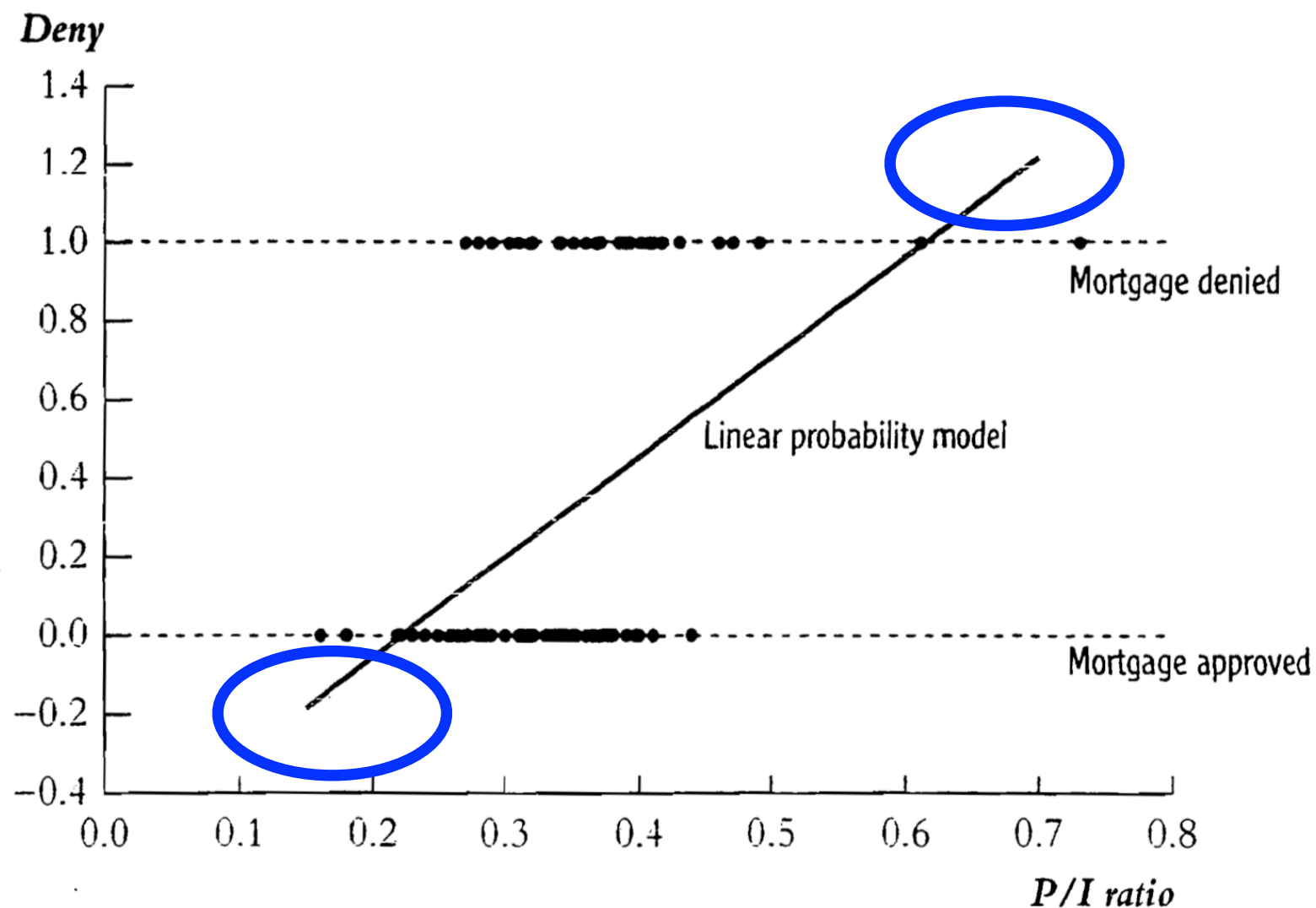
- Import the HMDA data.
- Take `deny` as the dependent variable.
- `deny` is a “factor”. Translate it into a numeric vector with values 0 and 1.
- Run the following regressions and make inference of coefficients.

$$\text{deny}_i = \beta_0 + \beta_1 \text{pirat}_i + u_i$$

$$\text{deny}_i = \beta_0 + \beta_1 \text{pirat}_i + \beta_2 \text{afam}_i + u_i$$

# Shortcomings of the linear probability model

- A probability must be between 0 and 1!



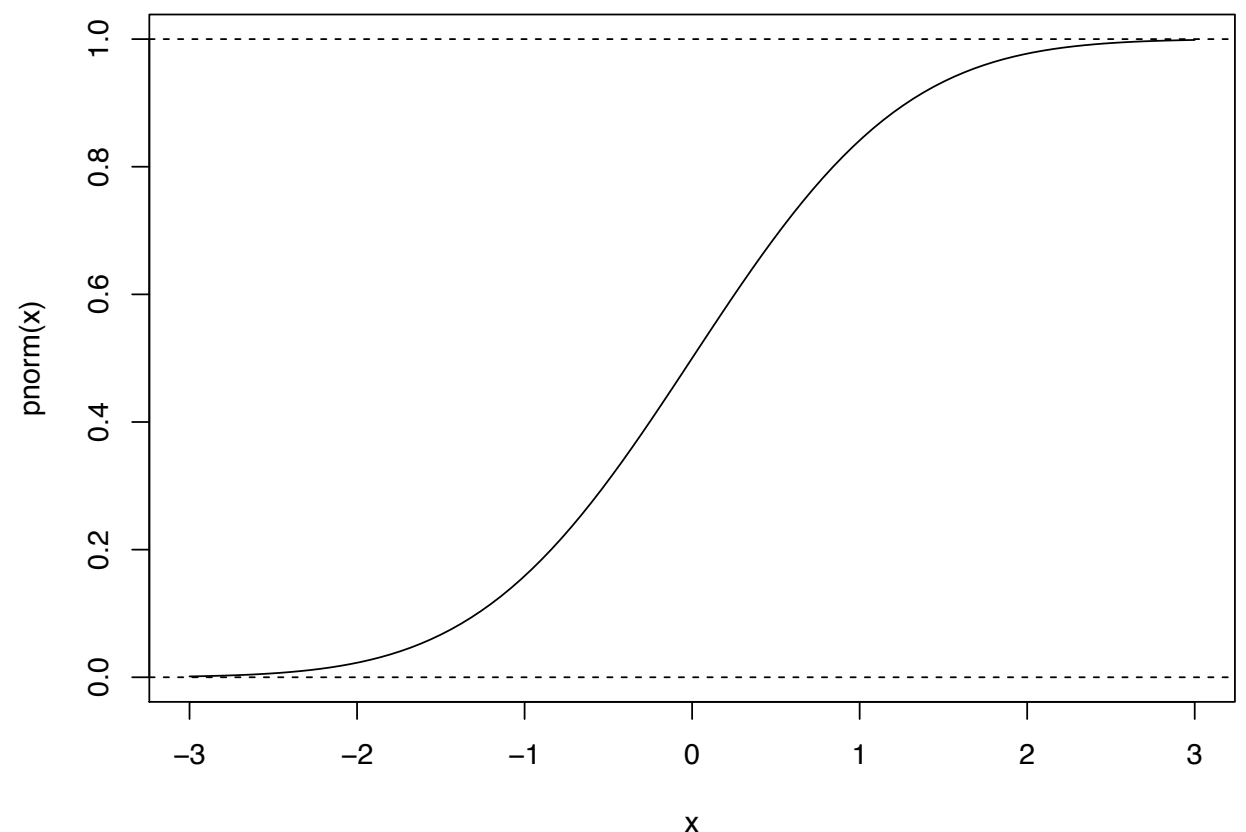
- Nonlinear models are needed.

# The probit regression

---

- Recall the c.d.f. of the standard normal distribution

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



- Probit regression

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

# The probit regression

---

- To predict the probability of  $Y = 1$ 
  1. Calculate the value  $z = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m$
  2. Calculate the cumulative probability at  $z$
- The regression coefficients can be estimated using nonlinear OLS method, or *maximum likelihood* method.
- The **maximum likelihood estimator** has a smaller variance than the nonlinear OLS estimator.

# The `glm()` command in R

---

- GLM refers to “generalized linear model”
- The GLM is a method that allows the distribution of errors being non normal.
- `glm()` uses the “iteratively reweighed least squares” method to find the maximum likelihood estimates of a generalized linear model.
- The probit model is an example of generalized linear model.

# Practice


---

- Take `deny` as the dependent variable and `pirat` (P/I ratio) as the independent variable.
- Use `glm( )` to estimate the coefficients in

$$\Pr(\text{deny} = 1 \mid \text{pirat}) = \Phi(\beta_0 + \beta_1 \text{pirat})$$

e.g.,

```
> glm(deny ~ pirat, family =  
      binomial(link = "probit"))
```

factor

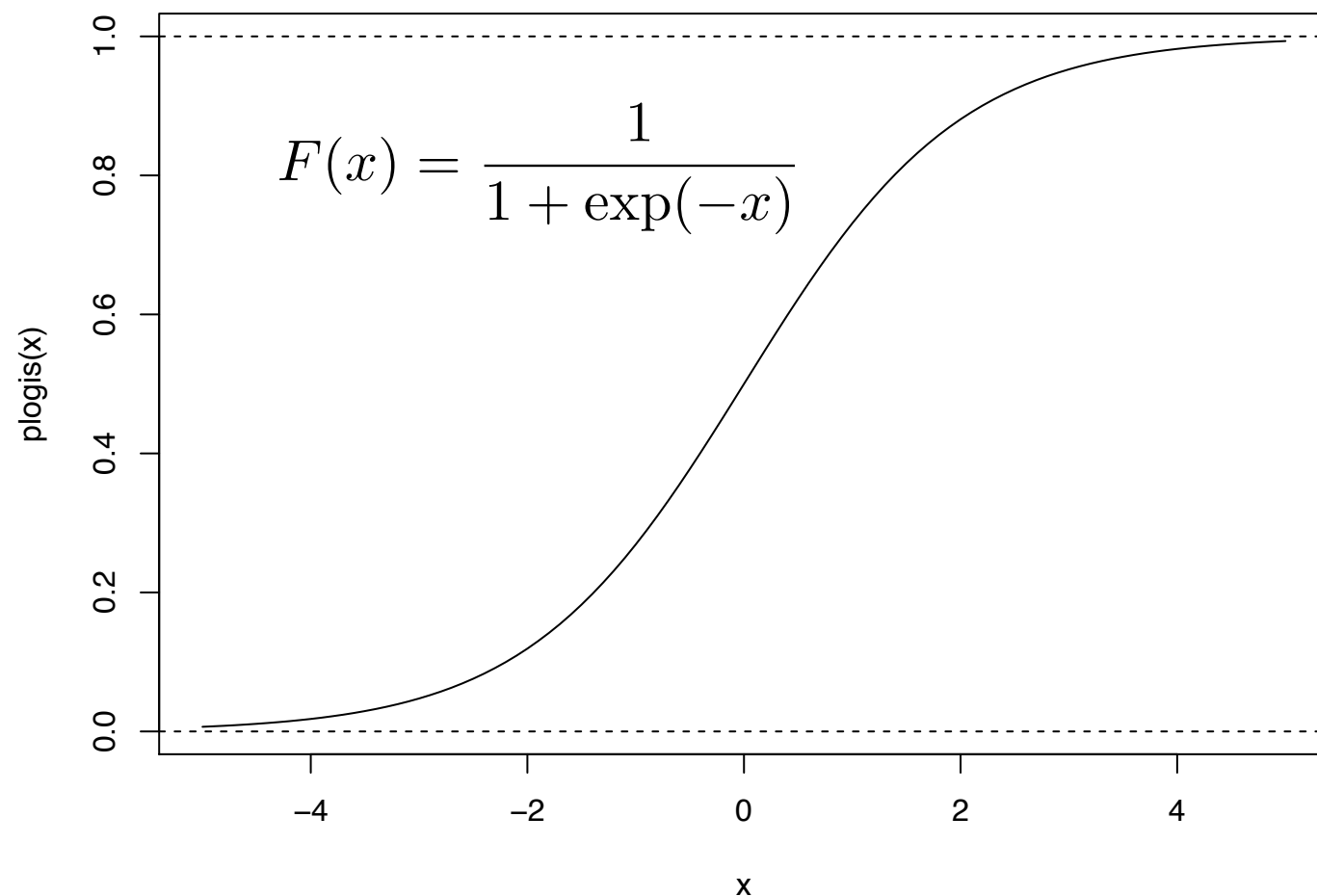
- For more information, read the help of `glm` and `family`

# The logit regression

---

- Logit regression

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



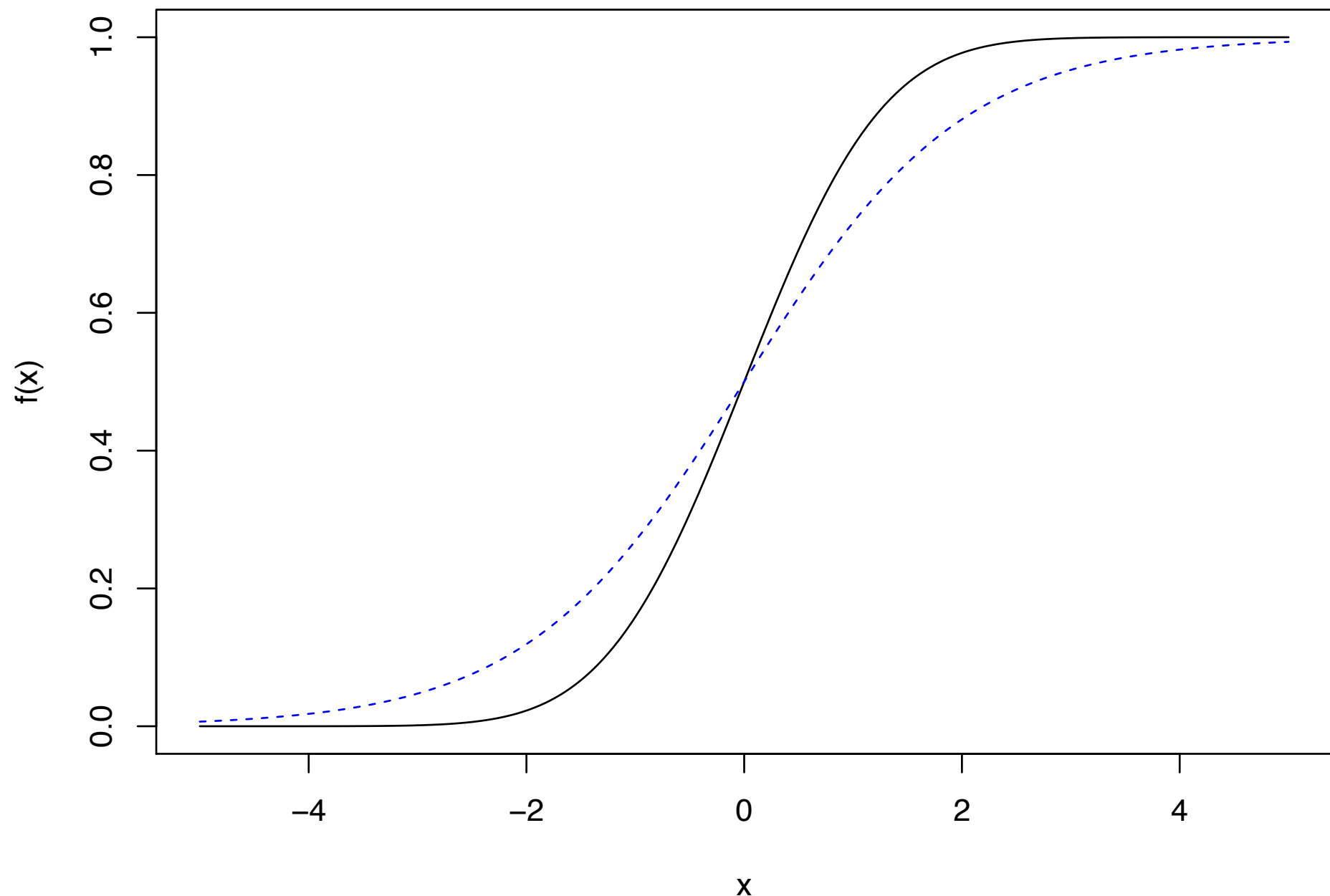
“logit” means the logistic function  $F(x)$



# Practice

---

- Draw a logistic function curve and a standard normal distribution function curve on the same plot from  $x = -5$  to  $x = 5$ .



# Practice

---

- Take `deny` as the dependent variable.
- Try the following regressions
  - > `glm(deny ~ pirat, family = binomial(link = "probit"))`
  - > `glm(deny ~ pirat, family = binomial(link = "logit"))`
- Compare the results.

# Regression results with HC standard errors

---

- The probit model

$$\beta_0 = -2.19, \quad \beta_1 = 2.97$$

$(0.19) \qquad (0.54)$

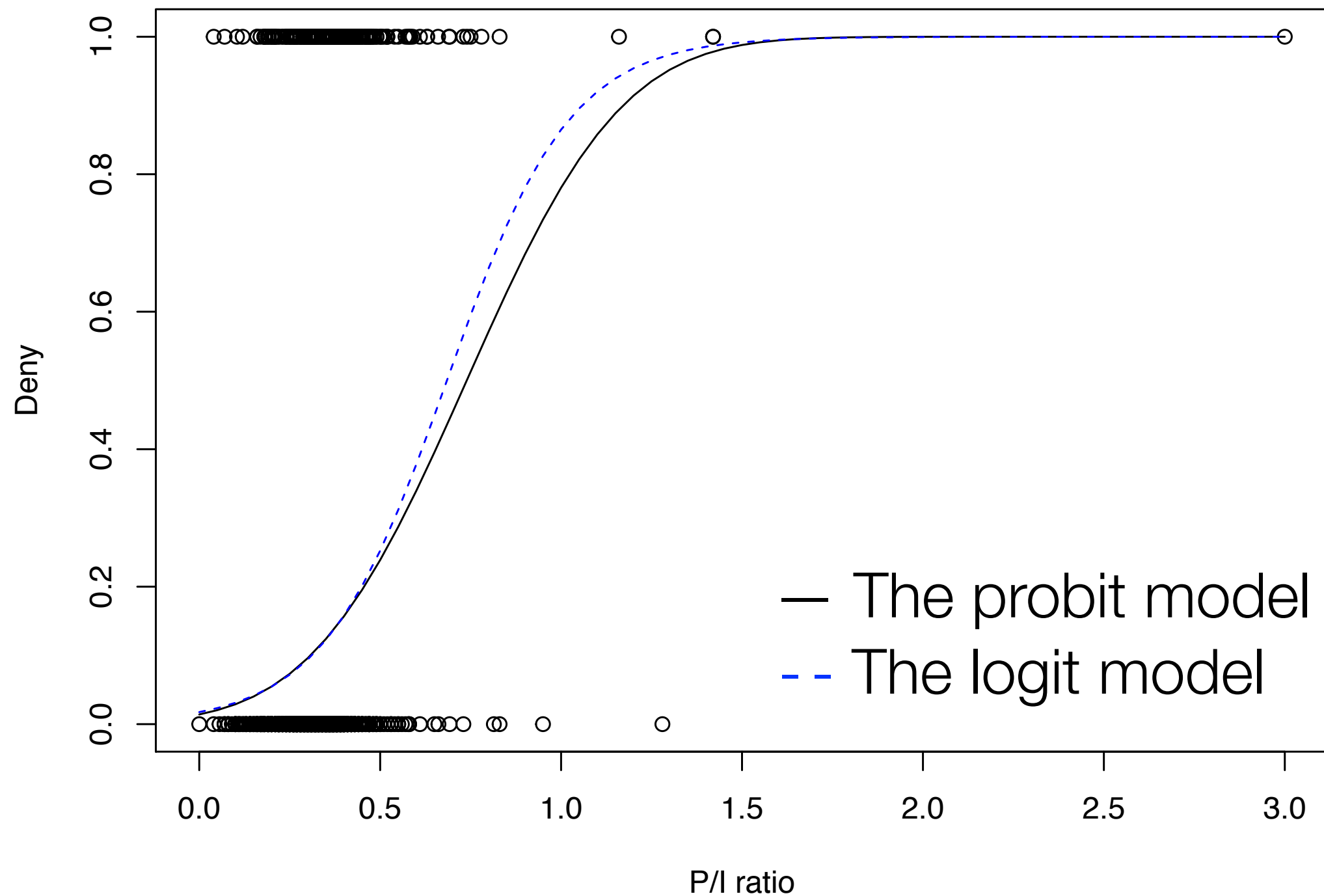
- The logit model

$$\beta_0 = -4.03, \quad \beta_1 = 5.88$$

$(0.36) \qquad (1.00)$

# Practice

- Reproduce the following figure.



# Goodness of fit

---

- McFadden's pseudo- $R^2$

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

where  $\ell(\hat{\beta})$  is the log-likelihood function of the fitted model, and  $\ell(\bar{y})$  is the log-likelihood function of the model containing only a constant term.

- Use `logLik( )` to obtain the log-likelihood function of an `glm` object.

# Practice

---

- Example

```
> fprobit1 <- glm(deny ~ pirat, family =  
  binomial(link = "probit"))  
  
> fprobit0 <- glm(deny ~ 1, family =  
  binomial(link = "probit"))  
  
> pR2probit1 <- 1 - as.numeric(  
  logLik(fprobit1) / logLik(fprobit0))
```

# Practice

---

- Use `pirat` and `afam` as independent variables, i.e.

$$\Pr(\text{deny} = 1 \mid \text{pirat}, \text{afam}) = \Phi(\beta_0 + \beta_1 \text{pirat} + \beta_2 \text{afam})$$

- Run this probit regression. Compare the results with the probit regression with single regressor `pirat`. Does the variable `afam` affects deny probability? Does this model have a better fit than the single regressor model?
- Try the same with logit regression.

# Comparing the linear probability, profit, and logit models

---

- The linear probability model is easy to use and interpret, but it cannot capture the nonlinear nature of the true population regression function.
- Profit and logit models are nonlinear, but their regression coefficients are difficult to interpret.
- The linear probability model uses OLS estimation, while the profit and logit models use ML estimation. When the data is extremely large, ML estimation can be very time consuming.
- Sometimes, the regression results are almost indifferent in practice.



# Assignment 3

---

- Suppose you want to investigate the effect of export on the economic development of a country using econometric methods. You may need to think the following questions:
  1. What variables are you going to use as independent and dependent variables?
  2. Which kind of data as well as regression models you are going to use?
  3. Which kind of estimation biases you need to consider? How can you get rid of them?
- Answer the above questions (either in Chinese or in English) within one page. Submit your report in a PDF file before 2018-12-11 19:00.

# References

---

1. Stock, J. H. and Watson, M. M., *Introduction to Econometrics*, 3rd Edition, Pearson, 2012.
2. Kleiber, C. and Zeileis, A., *Applied Econometrics with R*, Springer, 2008.