Econometrics 1

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

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.

- Data file: `hmda_sw1.csv`. Description: `hmda.docx`

- 62 variables, numeric and string data, with missing values.

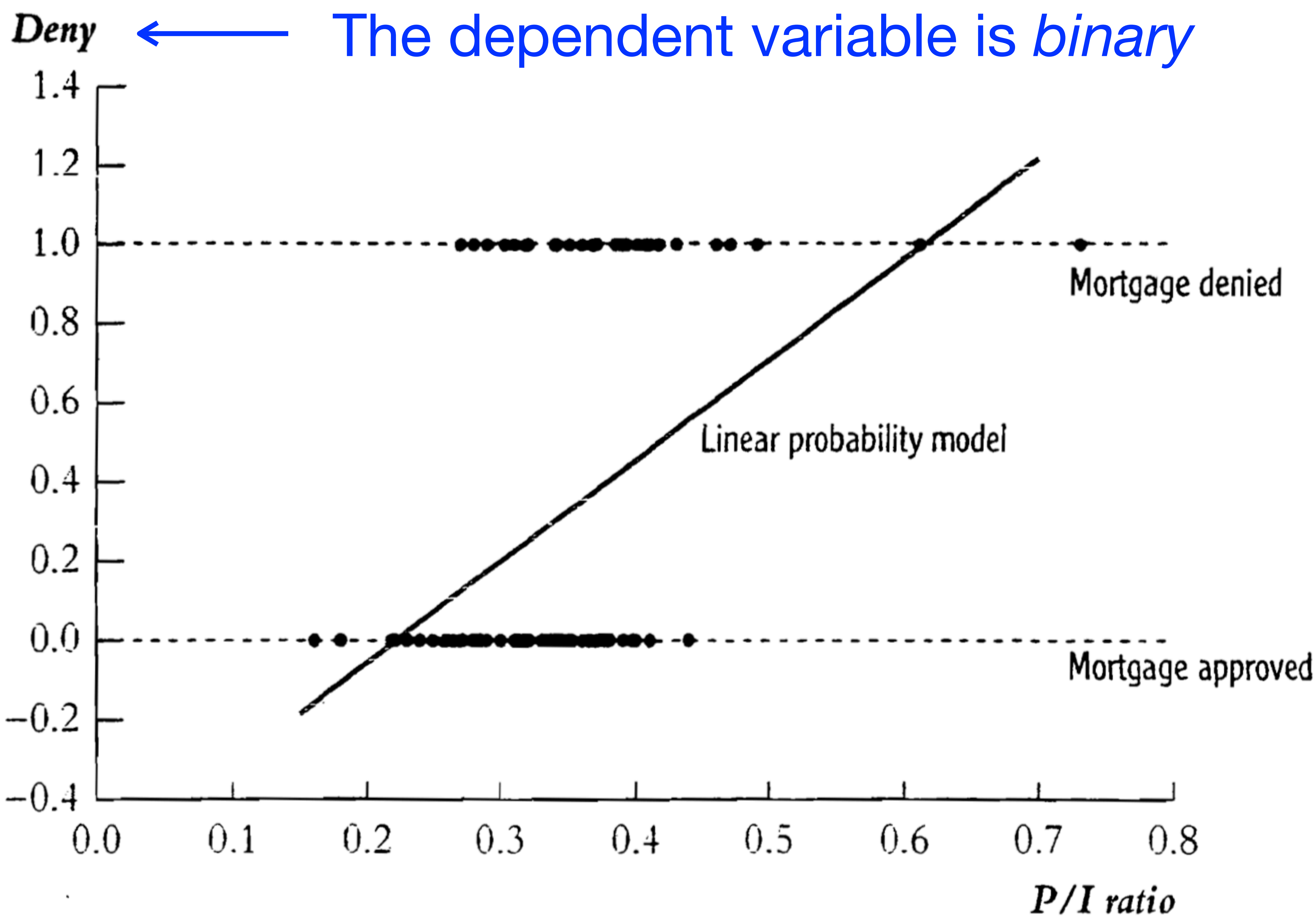
Missing value in the file	
Numeric data:	999999.375
String data:	NA

→

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

* Learn the `setmiss` command.

What determines whether a mortgage application is denied?



Regression with a binary dependent variable

- The population regression function of a lineal model

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

- Regression with a binary variable

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

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

The linear probability model

- The linear probability model

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

$$\Rightarrow \quad \Pr(Y = 1 \mid X_1, \ldots, X_m)$$
$$= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m$$

- The regression coefficient β_1 is the change in the probability that $Y = 1$ associated with a unit change in X_1, holding constant the other regressors, and so forth for $\beta_2, \ldots, \beta_m$

- The regression coefficients can be estimated by OLS.

Dummifying in gretl

```
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
```

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	

Practice

- Regress Ddeny with P/I ratio (Eq. (11.1))

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

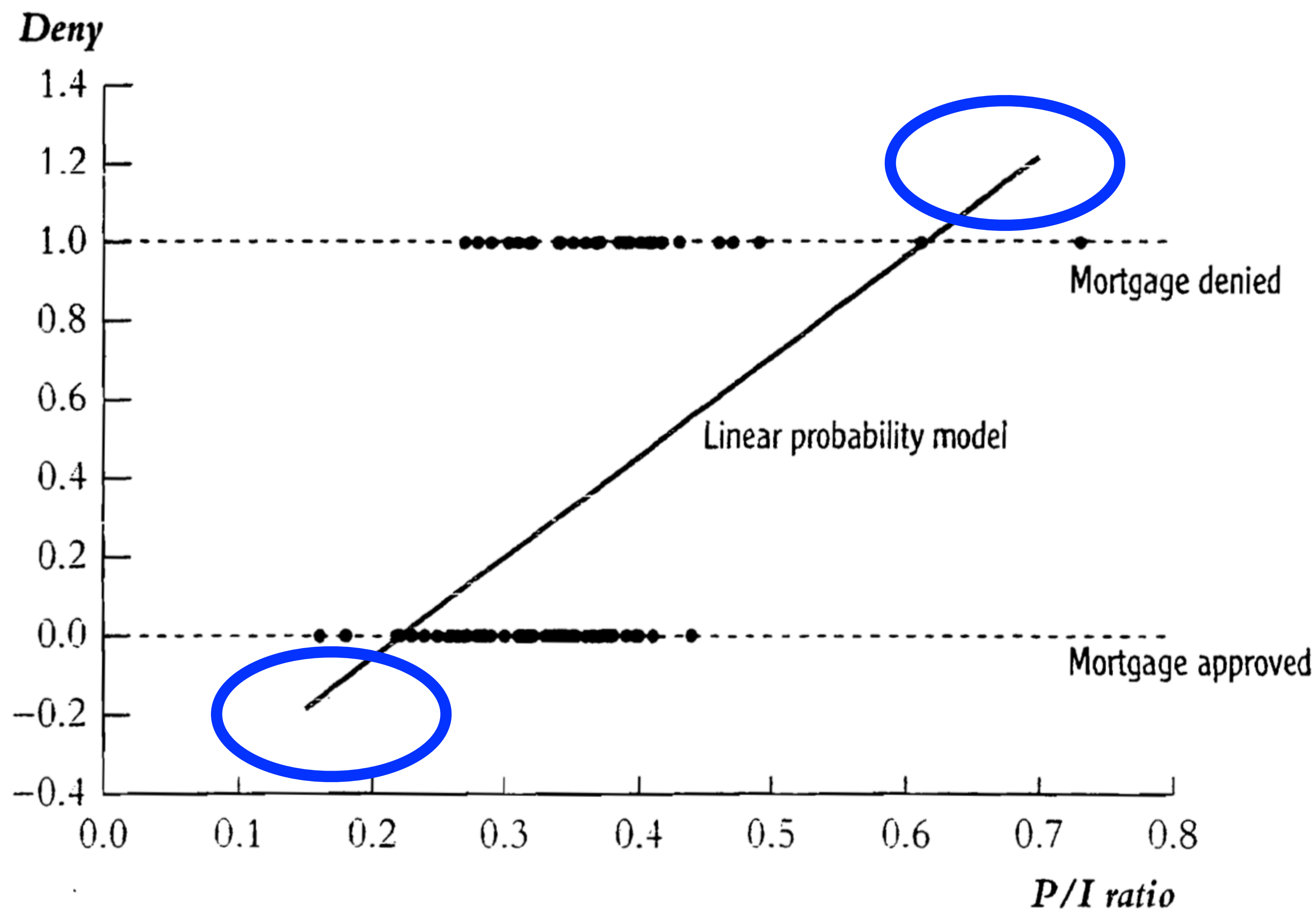
```
Model 1: OLS, using observations 1-2380
Dependent variable: Ddeny
Heteroskedasticity-robust standard errors, variant HC1

             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 ***
```

- Reproduce Eq. (11.2).

Shortcomings of the linear probability model

- A probability must be between 0 and 1!

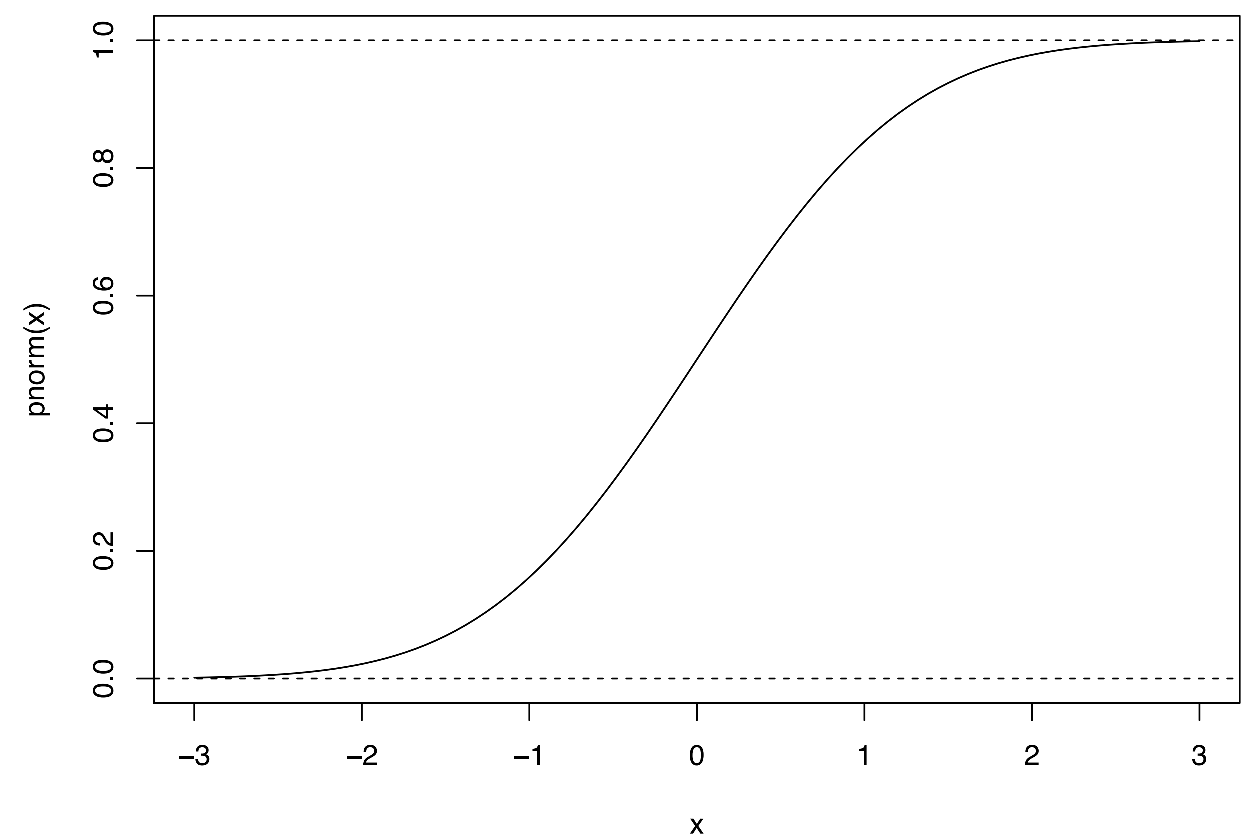


- Nonlinear models are needed.

The probit regression

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

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



- Probit regression

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

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

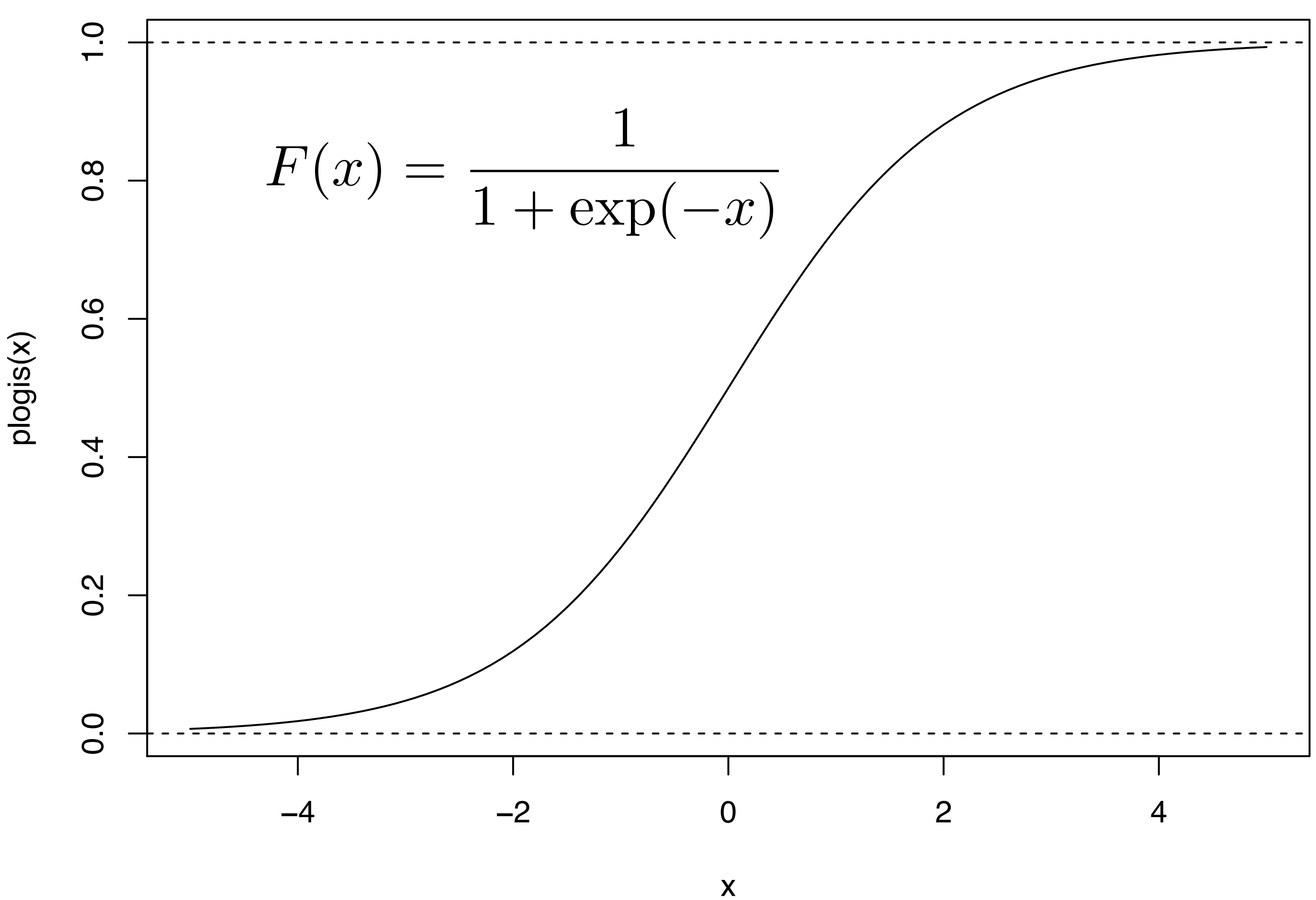
- The regression coefficients can be estimated using nonlinear OLS method, or *maximum likelihood* method.

- The **maximum likelihood estimator** has a smaller variance than the nonlinear OLS estimator.

The logit regression

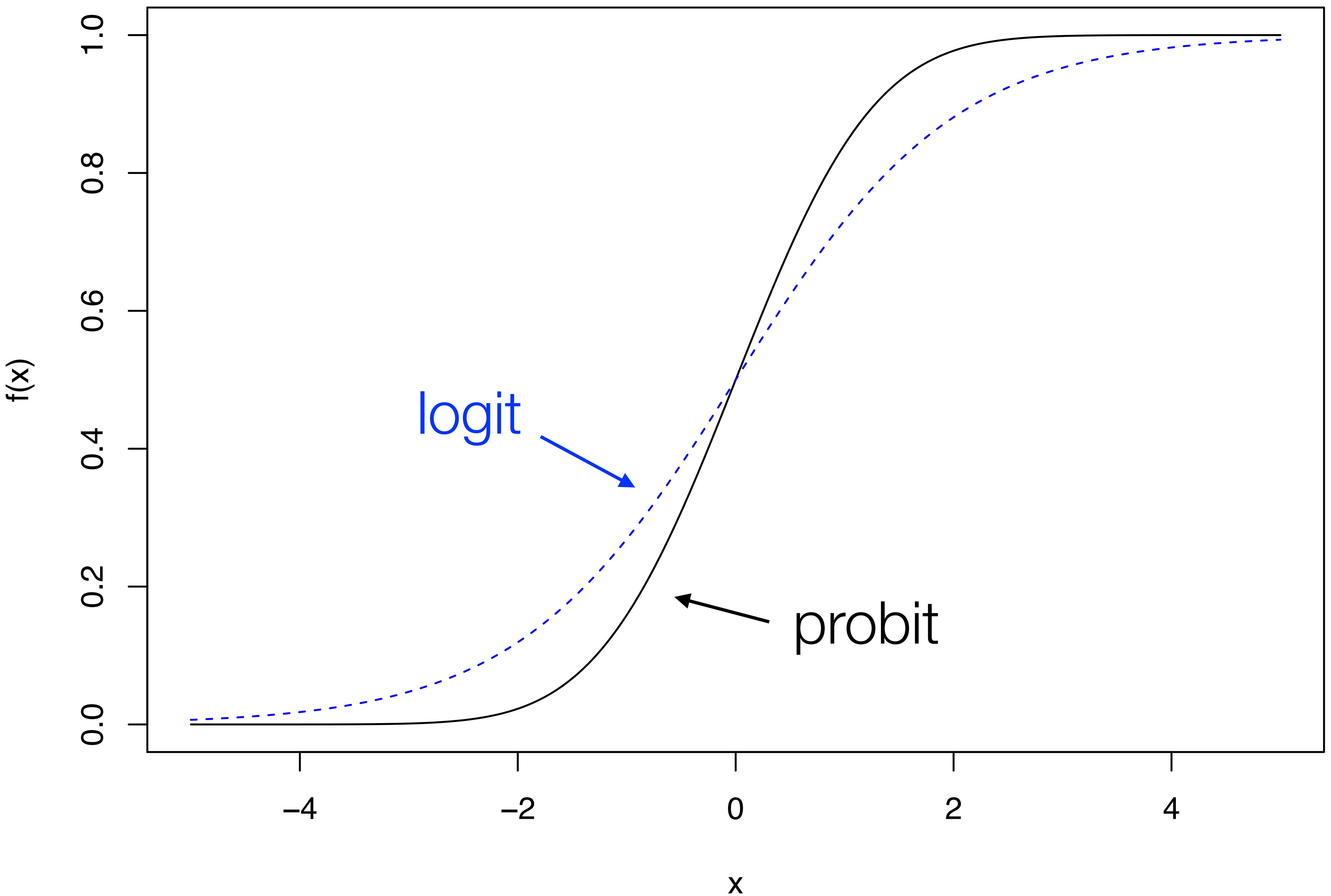
- Logit regression

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



"logit" means the logistic function $F(x)$

Difference between probit and logit function



The `probit` and `logit` command in gretl

- Run probit/logit regression using maximum likelihood estimation.

- Slopes at the mean of the independent variables are reported instead of p-values. If p-values are preferred, the **`--p-values`** option is needed.

- Predictions of the dependent variable with estimated parameters are also reported.

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

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

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

             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
```

Practice

- Reproduce Eq. (11.8) and (11.10) with and without the **`--p-values`** option.

- Calculate the predicted probability for a White/Black applicant with a P/I ratio = 0.3 for each model.

  Hint: you may use `$coeff` in your calculation.

- Compare your results with those given in the textbook.

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.

- The **fraction correctly predicted**.

```
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
```

Comparing the linear probability, probit, 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.

- Probit and logit models are nonlinear, but their regression coefficients are difficult to interpret.

- The linear probability model uses OLS estimation, while the probit and logit models use ML estimation (or nonlinear least squares estimation). When the data is extremely large, ML estimation can be very time consuming.

- Sometimes, the regression results are almost indifferent in practice. (The logit model was easier in calculation than the probit model.)

Take-home practice

- Learn Section 11.4 and try to reproduce Table 11.2.

References

1. Stock, J. H. and Watson, M. M., *Introduction to Econometrics*, 3rd Edition, Pearson, 2012.