

Econometrics 1

Lecture 22

Dummy Dependent Variables Models

Probability models/Dummy dependent variables

Alternative names: dichotomous dependent variables, discrete dependent random variable, binary variable, either or choice variables

$$Y_i = \begin{cases} \beta_0 + \beta_1 X_t + u_t & \text{if event occurs} \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

- the labour force participation (1 if a person participates in the labour force, 0 otherwise)
- yes or no vote in particular issue
- to marry or not to marry
- to study further or to start a job
- to buy or not to buy a particular stock
- choice of transportation mode to work (1 if a person drives to work, 0 otherwise)
- Union membership (1 if one is a member of the union, 0 otherwise)
- Owning a house (1 if one owns 0 otherwise)
- Multinomial choices: work as a teacher, or as a clerk, or as a self employed or professional or as a factory worker
- Multinomial ordered choices: strongly agree, agree, neutral, disagree

Four types of Probability models/Dummy dependent variable Models

There are mainly four types of limited dependent variable models

- Linear probability models
- Probit model
- Logit model
- Tobit

Linear Probability Model

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad (9)$$

where Y_i if person owns a house, 0 otherwise; X_i is family income.

$E(Y_i = 1/X_i)$ probability that the event y will occur given x

$$E(Y_i = 1/X_i) = 0(1 - P_i) + 1(P_i) = P_i \quad (10)$$

$$E(Y_i = 1/X_i) = \beta_1 + \beta_2 X_i = P_i \Rightarrow$$

$$0 \leq E(Y_i = 1/X_i) \leq 1$$

Linear Probability Model

Problems: Error term is heteroscedastic

$$u_i = 1 - \beta_1 - \beta_2 X_i \text{ with probability } (1 - P_i) \text{ and}$$
$$u_i = -\beta_1 - \beta_2 X_i \text{ with probability } P_i$$

Variance of the error term is not constant

$$\text{var}(u_i) = (-\beta_1 - \beta_2 X_i)^2 (1 - P_i) + (1 - \beta_1 - \beta_2 X_i)^2 P_i$$
$$\text{var}(u_i) = (-\beta_1 - \beta_2 X_i)^2 (1 - \beta_1 - \beta_2 X_i) + (1 - \beta_1 - \beta_2 X_i)^2 (\beta_1 + \beta_2 X_i)$$
$$\text{var}(u_i) = (\beta_1 + \beta_2 X_i)(1 - \beta_1 - \beta_2 X_i) = P_i(1 - P_i) \quad \mathbf{(11)}$$

thus the variance depends on x.

Limitations of a linear probability model

It is possible to transform this model to make it homoscedastic by dividing the original variables by

$$\sqrt{(\beta_1 + \beta_2 X_i)(1 - \beta_1 - \beta_2 X_i)} = \sqrt{P_i(1 - P_i)} = \sqrt{w_i}$$

$$\frac{Y_i}{\sqrt{w_i}} = \frac{\beta_1}{\sqrt{w_i}} + \beta_2 \frac{X_i}{\sqrt{w_i}} + \frac{u_i}{\sqrt{w_i}} \quad (12)$$

- a. It does not guarantee that the probability lies inside (0,1) bands
- b. Probability in non-linear phenomenon: at very low level of income a family does not own a house; at very high level of income **every** one owns a house ; marginal effect of income is very negligible. The linear probability model does not explain this fact well.

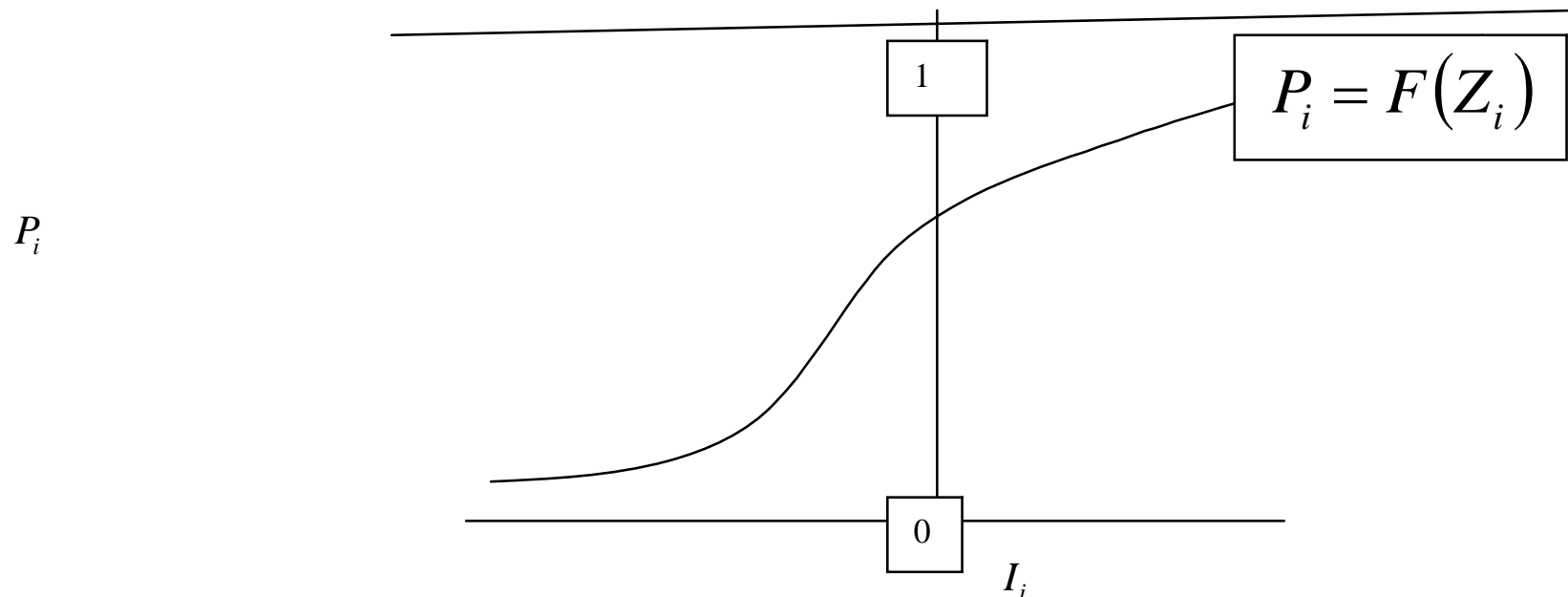
Probit model

$$\begin{aligned}\Pr(Y_i = 1) &= \Pr(Z_i^* \leq Z_i) = F(Z_i) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z_i} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\beta_i + \beta_2 X} e^{-\frac{t^2}{2}} dt \quad (13)\end{aligned}$$

Here t is standardised normal variable,
 $t \sim N(0,1)$

Steps for a probit model

- 1. probability depends upon unobserved utility index Z_i** which depends upon observable variables such as income. There is a thresh-hold of this index when after which family starts owning a house, $Z_i \geq Z_i^*$.



Logit model

variable Y_i which takes value 1 ($Y_i = 1$) if a student gets a first class mark, value 0 ($Y_i = 0$) otherwise. Probability of getting a first class mark in an exam is a function of student effort index denoted by Z_i . $P_i = \frac{1}{1 + e^{-Z_i}}$; where $Z_i = \beta_1 + \beta_2 X_i + u_i$

An example of a logit model: what determines that a student gets the first class degree?

$$Z_i = \beta_1 + \beta_2 H + \beta_3 E + \beta_4 A + \beta_5 P + e_t$$

H = hours of study, E = exercises, A = attendance in lectures and classes; P = papers written for assignment.

Ratio of odds: $\left[\frac{P_i}{1 - P_i} = \frac{1 + e^{Z_i}}{1 + e^{-Z_i}} = e^{Z_i} \right]$; taking log of the odds

$$\ln\left(\frac{P_i}{1 - P_i}\right) = Z_i$$

Features of a logit model

- a. probability goes from 0 to 1 as the index variable z_i goes from $-\infty$ to $+\infty$. Probability lies between 0 and 1.
- b. Log of the odds is linear in x , characteristic variables but probabilities themselves are not linear but non linear function of the parameters. Probabilities are estimated using the maximum likelihood method.
- c. Any explanatory variable that determines the value of z_i , measures how the log of odds of an event (i.e. owning a house) changes as a result of change in explanatory variable such as income.
- d. We can calculate P_i for given estimates of $\hat{\beta}_1$ and $\hat{\beta}_2$ or all other $\hat{\beta}_i$.
- e. Limiting case when $P_i=1$; $\ln\left(\frac{P_i}{1-P_i}\right)$ or when $P_i=0$ $\ln\left(\frac{0}{1-0}\right)$; OLS cannot be applied in such case but the maximum likelihood method may be used to estimate the parameters.

Other Choice Models

- Multinomial choice models: choice of different brands; different subjects (economics, finance, accounting, management)
- Ordered probit or ordered logit for choice of bonds such as AAA BBB; orders are used to rank the outcome; survey questions with ranking
- Count data models: How many trips to hospital by a patient?

Poission random variable $P(Y = y) = \frac{e^{-\lambda} \lambda^y}{y!}$

Tobit Model

It is an extension of the probit model, named after Tobin. We observe variables if the event occurs: ie if some one buys a house. We do not observe explanatory variables for people who have not bought a house. The observed sample is censored, contains observations for only those who buy the house.

$$Y_i = \begin{cases} \beta_0 + \beta_1 X_t + u_t & \text{if event occurs} \\ 0 & \text{otherwise} \end{cases}$$

Y_i is equal to $\beta_0 + \beta_1 X_t + u_t$ if the event is observed equal to zero if the event is not observed.

It is unscientific to estimate probability only with observed sample without worrying about the remaining observations in the truncated distribution. The Tobit model tries to correct this bias.

- **Inverse Mill's ratio:** Example first estimate probability of work then estimate the hourly wage as a function of socio-economic background variables

Summary on Probability Models

The effect of observed variables on probability

$$\frac{\partial P_i}{\partial x_{i,j}} = \begin{cases} \beta_j & \text{for the linear probability model} \\ \beta_j P_j (1 - P_j) & \text{for the logit probability model} \\ \beta_j \phi(Z_i) & \text{for the probit probability model} \end{cases}$$

where $Z_i = \beta_0 + \sum_{i=1}^k \beta_i x_{i,j}$ **and** $\phi(\cdot)$ **is the standard normal density function.**