Binary response models

Bernard Lejeune

Supplemental lecture notes VII
Advanced Econometrics
HEC-University of Liège
Academic year 2021-2022

- These lecture notes restate, in matrix form and with more details, the main results of Sections 17-1 of Wooldridge (2016).

1. Logit and Probit models for binary response

- As discussed in Wooldridge (2016), Section 7-5, the usual regression model :

$$y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i$$

$$\Leftrightarrow \quad y_i = X_i\beta + u_i \qquad (1)$$

where $i = 1, ..., n$ indexes the individuals, $X_i = (1, x_{i2}, ..., x_{ik})$ is a $1 \times k$ (row) vector of explanatory variables (including a constant) and $\beta = (\beta_1, \beta_2, ..., \beta_k)$ is $k \times 1$ (column) vector of parameters, may perfectly be used for analyzing data where the dependent variable y_i is a binary variable, i.e., a variable which by definition takes only two values, 0 and 1, and which is used to indicate whether or not an individual has a certain characteristic or a particular event has occurred[1]. The model is then called a linear probability model because in this case the conditional expectation of y_i given X_i is nothing but the probability that y_i is equal to 1 given X_i :

$$E(y_i|X_i) = I\!P(y_i = 1|X_i) = X_i\beta$$

so that, accordingly, the vector of parameters β measures the partial effects of the different explanatory variables on the probability that y_i is equal to 1.

- The linear probability model has some drawbacks :

[1] For example, $y_i = 1$ if an individual i is employed, and $y_i = 0$ otherwise.

  - Due to its linear functional form, the model may easily generate predicted probabilities which are less than zero or greater than one. Also, the model assumes that the partial effect of the different explanatory variables is constant – i.e., that a unit increase in x_{ij} always changes $\mathbb{P}(y_i = 1|X_i)$ by the same amount, regardless of its initial value –, which cannot literally be true[2].

  - When y_i is binary so that $E(y_i|X_i) = \mathbb{P}(y_i = 1|X_i) = X_i\beta$, the conditional variance of y_i given X_i is by definition equal to $Var(y_i|X_i) = X_i\beta(1 - X_i\beta)$. As the homoskedasticity assumption does not hold, the usual OLS estimator of model (1) is not efficient, and the usual inference procedures (confidence interval, hypothesis testing) are not valid[3].

- The logit and probit models overcome the shortcomings of the linear probability model. But this comes at a price: these models are more complicated to interpret.

- The logit model and the probit model are both a special case of the general model:

$$\begin{aligned} \mathbb{P}(y_i = 1|X_i) &= G(\beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik}) \\ &= G(X_i\beta) \qquad i = 1, ..., n \end{aligned} \tag{2}$$

where $G(.)$ is a function whose values are always between 0 and 1: $0 < G(z) < 1$, for all z. This ensures that the probability $\mathbb{P}(y_i = 1|X_i)$ is always between 0 and 1. Several functions are possible for the $G(.)$ function. The logit model specifies for $G(.)$ the logistic function:

$$G(z) = \frac{e^z}{1 + e^z} \tag{3}$$

This logistic function is the cumulative distribution function[4] (cdf) of the standard logistic distribution[5]. On the other hand, the probit model specifies for $G(.)$ the cdf of the standard normal distribution (which can not be written in closed-form):

$$G(z) = \int_{-\infty}^{z} \phi(x)dx\,, \tag{4}$$

where $\phi(x)$ is the standard normal probability distribution function : $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$. Figure 1 below represents the logit and the probit $G(.)$ functions.

[2] because continually increasing one of the explanatory variable would eventually drive $\mathbb{P}(y_i = 1|X_i)$ less than zero or greater than one.

[3] Of course, valid heteroskedasticity robust inference procedures may be used instead, or weighted least squares may be used to obtain an (asymptotically) efficient estimator, but only provided that all predicted probabilities lies between 0 and 1. See Wooldrige (2016), Section 8-5.

[4] As a reminder, the cumulative distribution function $F(x)$ of a random variable X is defined as $F(x) = \mathbb{P}(X \leq x)$.

[5] A random variable X follows a standard logistic distribution if its probability density function (pdf) is $f(x) = \frac{e^x}{(1+e^x)^2}$.

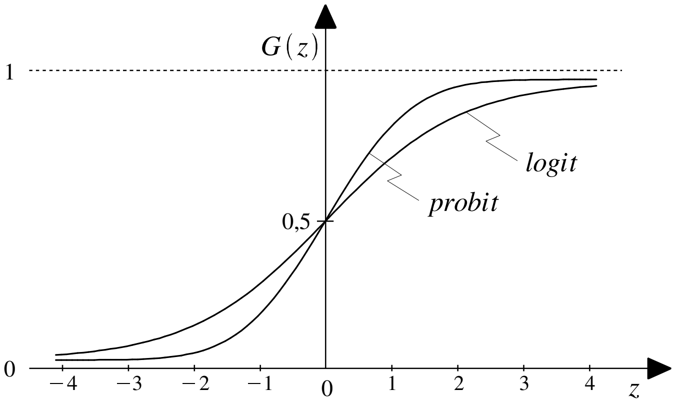

Figure 1: the logit and the probit $G(.)$ functions

Both the logit and the probit $G(.)$ functions are strictly increasing, from zero (for $z \to -\infty$) to one (for $z \to \infty$). Note that model (2) may be derived from an underlying latent variable model such as:

$$\begin{cases} y_i^* = X_i\beta + e_i \\ y_i = 1 \text{ if } y_i^* > 0, \, 0 \text{ otherwise} \end{cases} \tag{5}$$

where y_i^* is an unobserved – i.e., latent – variable, e_i is an error term assumed independent of X_i, and y_i is an observed variable equal to 1 if $y_i^* > 0$, 0 otherwise. Assuming that e_i is distributed according to a standard logistic distribution or a standard normal distribution yields, respectively, the logit model and the probit model. This interpretation of the logit and the probit models is however usually not especially useful. See Wooldridge (2016), Section 17-1a.

- In model (1), interest usually lies in the partial effect of the different explanatory variables x_{ij}. If the variable x_{ij} is (at least roughly) continuous, its partial effect on the probability $I\!P(y_i = 1|X_i)$ is given by:

$$\frac{\partial I\!P(y_i = 1|X_i)}{\partial x_{ij}} = g(X_i\beta)\beta_j \tag{6}$$

where $g(z) = \frac{dG(z)}{dz}$ is given, for the logit model, by:

$$g(z) = \frac{e^z}{(1 + e^z)^2} \tag{7}$$

and for the probit model, by:

$$g(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \,, \tag{8}$$

i.e., by the probability density function (pdf) of, respectively, the standard logistic and the standard normal distribution. Figure 2 below represents the logit and the probit $g(.)$ functions.

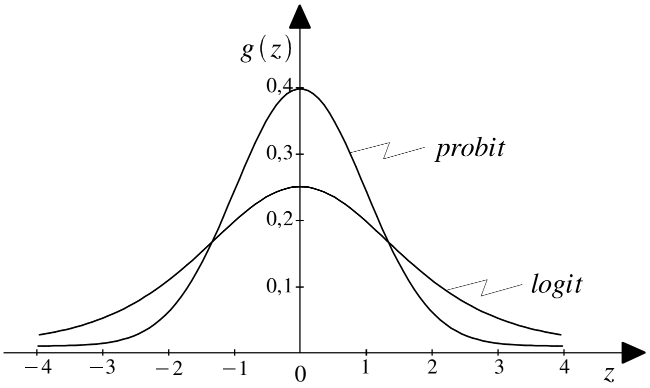

Figure 2 : the logit and the probit $g(.) = \frac{dG(z)}{dz}$ functions

Because the $g(.)$ is always positive, the partial effect $\frac{\partial I\!P(y_i=1|X_i)}{\partial x_{ij}}$ of x_{ij} on the probability $I\!P(y_i = 1|X_i)$ has always the same sign as β_j. Both in the logit and the probit models, for any given β_j, this partial effect is the largest when $X_i\beta = 0$, i.e. when $I\!P(y_i = 1|X_i) = 0.5$.

If the variable x_{ij} is discrete or binary, (6) usually provides only a crude approximation of the actual partial effect. It is preferable to compute the exact partial effect. If, for example, x_{i2} is a binary variable, the exact partial effect from changing x_{i2} from 0 to 1 is simply :

$$\begin{aligned}\frac{\Delta I\!P(y_i = 1|X_i)}{\Delta x_{i2}} &= G\left(\beta_1 + \beta_2 + \beta_3 x_{i3} ... + \beta_k x_{ik}\right) \\ &\quad - G\left(\beta_1 + \beta_3 x_{i3} ... + \beta_k x_{ik}\right)\end{aligned}$$

Also, if there is transformed variables and/or polynomials among the explanatory variables, then the partial effect formula (6) must be adapted. For example, for the model :

$$I\!P(y_i = 1|X_i) = G\left(\beta_1 + \beta_2 x_{i2} + \beta_3 \ln x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i4}^2\right)$$

we have :

$$\begin{aligned}\frac{\partial I\!P(y_i = 1|X_i)}{\partial x_{i2}} &= g(X_i\beta)\beta_2 \\ \frac{\partial I\!P(y_i = 1|X_i)}{\partial x_{i3}} &= g(X_i\beta)\frac{\beta_3}{x_{i3}} \\ \frac{\partial I\!P(y_i = 1|X_i)}{\partial x_{i4}} &= g(X_i\beta)(\beta_4 + 2\beta_5 x_{i4})\end{aligned}$$

See Wooldridge (2016), Section 17-1a, for more examples and some discussion.

2. Maximum likelihood estimation and inference

- Hereafter, we assume that :
  - The available data are realizations of a random sample of size n, $\{(y_i, X_i)$:

$i = 1, ..., n\}$.

- In the sample (and thus in the population), there is no exact linear relationship among the explanatory variables X_i.
- The logit or the probit model of interest:

$$I\!P(y_i = 1|X_i) = G(X_i\beta), \quad i = 1, ..., n$$

where the $G(.)$ function is equal to (3) for the logit model and to (4) for the probit model, is correctly specified, so that the pdf of the conditional distribution of y_i given X_i can be written:

$$f(y_i|X_i;\beta) = G(X_i\beta)^{y_i}(1 - G(X_i\beta))^{1-y_i}, \quad i = 1, ..., n \tag{9}$$

- Under random sampling, the observations are independent across i, so that the joint pdf of $(y_1, ..., y_n)$ given $(X_1, ..., X_n)$ – which is called the likelihood function – is given by:

$$\begin{aligned} & f(y_1, ..., y_n|X_1, ..., X_n;\beta) \\ = \; & f(y_1|X_1;\beta) \times ... \times f(y_n|X_n;\beta) = \prod_{i=1}^{n} f(y_i|X_i;\beta) \end{aligned} \tag{10}$$

The log-likelihood function is obtained taking the natural logarithm of (10):

$$\begin{aligned} L(\beta) &= \log f(y_1, ..., y_n|X_1, ..., X_n;\beta) \\ &= \sum_{i=1}^{n} \log f(y_i|X_i;\beta) \\ &= \sum_{i=1}^{n} [y_i \log G(X_i\beta) + (1 - y_i)\log(1 - G(X_i\beta))] \end{aligned} \tag{11}$$

The maximum likelihood (ML) estimator[6] $\hat{\beta}_{ML}$ of the vector of parameters β is defined as the value of β which maximizes the likelihood function (10), or equivalently[7], the log-likelihood function (11):

$$\begin{aligned} \hat{\beta}_{ML} &= \text{Argmax}_{\beta} \sum_{i=1}^{n} \log f(y_i|X_i;\beta) \\ &= \text{Argmax}_{\beta} \sum_{i=1}^{n} [y_i \log G(X_i\beta) + (1 - y_i)\log(1 - G(X_i\beta))] \end{aligned} \tag{12}$$

The maximization problem (12) has no closed-form solution. The ML estimator $\hat{\beta}_{ML}$ can only be obtained numerically. All econometric software can do it well and fast.

[6] For a discussion of the maximum likelihood approach to estimation, see Wooldridge (2016), Appendix C-4b.

[7] As the natural logarithm is a strictly increasing function, the likelihood and the log-likelihood are necessarily maximum at the same value of β.

- From the general maximum likelihood theory[8], under general regularity conditions, the ML estimator $\hat{\beta}_{ML}$ is consistent and asymptotically normal:

$$\hat{\beta}_{ML} \xrightarrow{p} \beta \tag{13}$$

and

$$\sqrt{n}(\hat{\beta}_{ML} - \beta) \xrightarrow{d} N(0, A^{-1}) \tag{14}$$

where[9]:

$$A = -E\left(\frac{\partial^2 \log f(y_i|X_i;\beta)}{\partial\beta\partial\beta'}\right) = E\left(\frac{g(X_i\beta)^2 X_i'X_i}{G(X_i\beta)(1-G(X_i\beta))}\right) \tag{15}$$

Moreover, the ML estimator $\hat{\beta}_{ML}$ is asymptotically efficient, i.e., it has the smallest (in a matrix sense) asymptotic variance among all consistent asymptotically normal estimators of β[10].

- As for the OLS or the 2SLS estimator, the limiting distribution result (14) provides an approximate finite sample distribution for the ML estimator $\hat{\beta}_{ML}$:

$$\sqrt{n}(\hat{\beta}_{ML} - \beta) \xrightarrow{d} N(0, A^{-1})$$

$$\Leftrightarrow \quad \hat{\beta}_{ML} \approx N(\beta, A^{-1}/n) \tag{16}$$

which can be used – when n is sufficiently large – for performing inference (confidence interval, hypothesis testing).

- For inference based on the limiting distributional result (14), or equivalently on the approximate distributional result (16), we need an estimator of the asymptotic variance $Avar(\hat{\beta}_{ML}) = A^{-1}/n$. This requires a consistent estimator of A. A consistent estimator of A is simply given[11] by the sample counterpart of (15), i.e., $\frac{1}{n}\sum_{i=1}^{n}\frac{g(X_i\hat{\beta}_{ML})^2 X_i'X_i}{G(X_i\hat{\beta}_{ML})(1-G(X_i\hat{\beta}_{ML}))}$, so that an estimator of $Avar(\hat{\beta}_{ML}) = A^{-1}/n$ is given by:

$$\hat{V}_{ML}(\hat{\beta}_{ML}) = \left[\sum_{i=1}^{n}\frac{g(X_i\hat{\beta}_{ML})^2 X_i'X_i}{G(X_i\hat{\beta}_{ML})(1-G(X_i\hat{\beta}_{ML}))}\right]^{-1} \tag{17}$$

As usual, the diagonal elements $V\hat{a}r_{ML}(\hat{\beta}_{ML_j})$ of the $k \times k$ matrix estimator

[8] See Wooldridge (2010), Chapter 13.

[9] The matrix $\frac{\partial^2 \log f(y_i|X_i;\beta)}{\partial\beta\partial\beta'}$ is the hessian matrix of the function $\log f(y_i|X_i;\beta)$, i.e., a square matrix with elements (i,j) equal to the second derivatives $\frac{\partial^2 \log f(y_i|X_i;\beta)}{\partial\beta_i\partial\beta_j}$.

[10] Consistency, asymptotic normality and asymptotic efficiency are general properties of ML estimators. Whenever one wants to estimate the vector of parameters β of a correctly specified model $f(y_i|X_i;\beta)$ for the conditional distribution of y_i given X_i based on a random sample of observations $\{(y_i, X_i)$: $i = 1, ..., n\}$, then the ML estimator $\hat{\beta}_{ML} = \text{Argmax}_{\beta}\sum_{i=1}^{n}\log f(y_i|X_i;\beta)$ always provides a consistent ($\hat{\beta}_{ML} \xrightarrow{p} \beta$), asymptotically normal ($\sqrt{n}(\hat{\beta}_{ML} - \beta) \xrightarrow{d} N(0, A^{-1})$, where $A = -E[\frac{\partial^2 \log f(y_i|X_i;\beta)}{\partial\beta\partial\beta'}]$) and asymptotically efficient estimator of β.

[11] From the law of large numbers (LLN), $\frac{1}{n}\sum_{i=1}^{n}\frac{g(X_i\beta)^2 X_i'X_i}{G(X_i\beta)(1-G(X_i\beta))} \xrightarrow{p} E\left(\frac{g(X_i\beta)^2 X_i'X_i}{G(X_i\beta)(1-G(X_i\beta))}\right)$. As $\hat{\beta}_{ML} \xrightarrow{p} \beta$, we also have $\frac{1}{n}\sum_{i=1}^{n}\frac{g(X_i\hat{\beta}_{ML})^2 X_i'X_i}{G(X_i\hat{\beta}_{ML})(1-G(X_i\hat{\beta}_{ML}))} \xrightarrow{p} E\left(\frac{g(X_i\beta)^2 X_i'X_i}{G(X_i\beta)(1-G(X_i\beta))}\right)$.

$\hat{V}_{ML}(\hat{\beta}_{ML})$[12] being the estimators of the variance $Avar(\hat{\beta}_{ML_j})$ of the estimator $\hat{\beta}_{ML_j}$ of the different parameters β_j ($j = 1, ..., k$), natural estimators of the asymptotic standard error $As.e.(\hat{\beta}_{ML_j}) = \sqrt{Avar(\hat{\beta}_{ML_j})}$ of the estimator $\hat{\beta}_{ML_j}$ of different parameters β_j, as well as a natural estimator of the asymptotic standard error $As.e.(R_0\hat{\beta}_{ML}) = \sqrt{Avar(R_0\hat{\beta}_{ML})} = \sqrt{R_0 Avar(\hat{\beta}_{ML})R_0'}$ of the estimator $R_0\hat{\beta}_{ML}$ of a single linear combination $R_0\beta$ of β, are likewise given by:

$$s.\hat{e}._{ML}(\hat{\beta}_{ML_j}) = \sqrt{V\hat{a}r_{ML}(\hat{\beta}_{ML_j})}, \quad j = 1, ..., k \tag{18}$$

and

$$s.\hat{e}._{ML}(R_0\hat{\beta}_{ML}) = \sqrt{R_0\hat{V}_{ML}(\hat{\beta}_{ML})R_0'} \tag{19}$$

where R_0 is a $1 \times k$ (row) vector of constants.

- The limiting distributional result (16), or equivalently the approximate distributional result (16), and the estimators $\hat{V}_{ML}(\hat{\beta}_{ML})$, $s.\hat{e}._{ML}(\hat{\beta}_{ML_j})$ and $s.\hat{e}._{ML}(R_0\hat{\beta}_{ML})$ given above in respectively (17), (18) and (19), provide all which is needed for performing inference after ML estimation. Following exactly the same reasoning as in Section 4.3 and Section 4.4.2 of SLN-I, it may again readily be checked that if in all the usual OLS inference procedures – confidence interval for β_j or a single linear combination $R_0\beta$, two-sided and one-sided t-tests of β_j or a single linear combination $R_0\beta$, F-test or Wald test of multiple linear restrictions – we replace the usual estimators $\hat{V}(\hat{\beta})$, $s.\hat{e}.(\hat{\beta}_j)$ and $s.\hat{e}.(R_0\hat{\beta})$ by their ML versions $\hat{V}_{ML}(\hat{\beta}_{ML})$, $s.\hat{e}._{ML}(\hat{\beta}_{ML_j})$ and $s.\hat{e}._{ML}(R_0\hat{\beta}_{ML})$, then we obtain asymptotically valid – i.e., approximately valid for n sufficiently large – inference procedures. Note however that, in the present context, for the confidence intervals and the t-tests, it is more common to use quantiles from the standard normal distribution than from the student distribution. Likewise, for testing multiple linear restrictions, it is more common to use the Wald test than the F-test[13]. Another possibility is to use a likelihood ratio test (see Wooldridge (2016), Section 17-1c, for practical details).

3. Remarks

- All modern software provide built-in routines for ML estimation and hypothesis testing of both the logit and the probit models.

- In practice, the logit model and the probit model usually yield very similar results regarding both the estimated probabilities $I\!P(y_i = 1|X_i)$ and the estimated marginal effects $\frac{\partial I\!P(y_i=1|X_i)}{\partial x_{ij}}$ or $\frac{\Delta I\!P(y_i=1|X_i)}{\Delta x_{ij}}$. This basically follows from the fact

[12] As a reminder, the usual estimator of the asymptotic variance of the OLS estimator is $\hat{V}(\hat{\beta}) = \hat{s}^2(X'X)^{-1}$, which may be written $\hat{V}(\hat{\beta}) = \left[\sum_{i=1}^{n} \frac{1}{\hat{s}^2} X_i'X_i\right]^{-1}$. Viewed in this way, $\hat{V}_{ML}(\hat{\beta}_{ML})$ appears a little bit less exotic: the factor $\frac{1}{\hat{s}^2}$ is simply replaced by $\frac{g(X_i\hat{\beta}_{ML})^2}{G(X_i\hat{\beta}_{ML})(1-G(X_i\hat{\beta}_{ML}))}$.

[13] From an asymptotic point of view, it actually does not matter which one is used.

that the $G(.)$ function of the logit model and the probit model are actually much less different than one might think at first glance. With proper scaling, we indeed have:

$$G_{logit}(z) \simeq G_{probit}(\frac{z}{1,6}) \tag{20}$$

where $G_{logit}(.)$ and $G_{probit}(.)$ respectively stand for the logit $G(.)$ function (3) and the probit $G(.)$ function (4). Figure 3 illustrates the approximate equality (20).

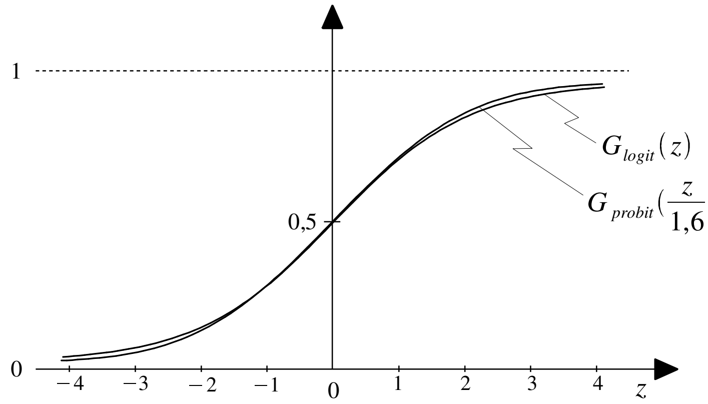

Figure 3: the logit and the probit $G(.)$ functions with proper scaling

As a result, estimation of the logit and the probit models typically yields $\hat{\beta}_{logit} \simeq 1,6\,\hat{\beta}_{probit}$, and we have $G_{logit}(X_i\hat{\beta}_{logit}) \simeq G_{probit}(X_i\frac{\hat{\beta}_{logit}}{1,6}) \simeq G_{probit}(X_i\hat{\beta}_{probit})$. In practice, it thus usually does not really matter which model is used[14].

- Some goodness-of-fit measures may be computed after logit or probit estimation. The most common measure is the percentage of correctly predicted outcomes by the estimated model. Various pseudo R-squared measures may also be computed. See Wooldridge (2016), Section 17-1c, for details.

- A distinctive characteristic of the logit and the probit models is that the partial effect of the different explanatory variables x_{ij} – continuous, discrete or binary – is not constant: it depends on the value of X_i at which it is computed. As a summary of the magnitudes of the partial effect of the different explanatory variables, it is standard either to compute them at some typical value of X_i such as its sample average $\bar{X}$ (these are the so-called partial effects at the average[15]), or to compute them for all observed values of X_i, and then average these computed individual partial effects (these are the so-called average partial effects). For details and some discussion, see Wooldridge (2016), Section 17-1d.

- It is possible to compute standard errors, and thus confidence intervals, for any estimated partial effect, and thus the partial effects at the average (PEA), and further the average partial effects (APE). This is however complicated be-

[14] The probit model tends to be more popular, but the logit model is actually easier to handle because of its closed-form $G(.)$ function.

[15] Note that special attention is needed to properly define $\bar{X}$ when the explanatory variables include dummy variables, transformed variables or polynomials.

cause these effects entail nonlinear functions of the vector of parameters β. See Wooldridge (2010), Section 15.6, for details. Some econometric software provide special options to do it.

Reference

Wooldridge J.M. (2010), *Econometric Analysis of Cross-Section and Panel Data*, Second Edition, MIT Press.

Wooldridge J.M. (2016), *Introductory Econometrics : A Modern Approach*, 6th Edition, Cengage Learning.