DSE - Data-Driven Economic Analysis
Econometrics Module
Lecture 6 - Logit and Probit Models

Michele De Nadai
michele.denadai@unimi.it

Trimester II, 2023

UNIVERSITÀ
DEGLI STUDI
DI MILANO

Limited Dependent Variables

- So far we focused on modelling dependent variables with support on the entire real line (or supposedly so).
- In real-life applications we often deal with dependent variables with **limited support**:
  - income, $[0, +\infty)$,
  - stock property, $\{0, 1\}$,
  - proportion of preferences during elections, $[0, 1]$.
- These kind of variables are in general **not well suited** to be modelled by means of the linear regression model (or are they?).

Why is that?

- Recall that the first, crucial, assumption for the linear regression model is **linearity** of the conditional expectation:

$$E[Y|\mathbf{X} = \mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}.$$

- However, there is in general no guarantee that, by varying $\mathbf{x}$, the linear combination $\mathbf{x}'\boldsymbol{\beta}$ will produce values **coherent** with the support of Y.
- This is not the only problem arising when applying linear models to limited dependent variables.
- We will focus on models for binary outcomes, i.e. random variables Y with support $\{0, 1\}$.

Example - Labour force participation

- As a running example consider female labour force participation as a binary outcome Y.
- $Y = 1$ if female is employed and $Y = 0$ otherwise.
- We are interested in understanding how labour market attachment is affected by several observed characteristics of the female such as: age, education, county unemployment rate etc...

Binary outcomes

- First recall that, when Y takes only two values $\{0, 1\}$, then Y is a **Bernoulli** random variable with some parameter p

$$Y \sim Be(p)$$

- It follows that, if we are interested in the conditional distribution of Y given a set of controls $\mathbf{X}$, it is:

$$Y|\mathbf{X} = \mathbf{x} \sim Be(p(\mathbf{x})),$$

  for some function $p(\mathbf{x})$.

- Finally note that the **expected value** of a Bernoulli random variable with parameter p is equal to p

- It follows that we can write:

$$E[Y|\mathbf{X} = \mathbf{x}] = Pr(Y = 1|\mathbf{X} = \mathbf{x}) = p(\mathbf{x}).$$

Problem 1 - Support

- Assuming **linearity** of the conditional expectations entails:

  $$\mathrm{E}[Y|\mathbf{X} = \mathbf{x}] = \mathbf{x}'\boldsymbol{\beta} = \Pr(Y = 1|\mathbf{X} = \mathbf{x}),$$

  but $0 \leqslant \Pr(Y = 1|\mathbf{X} = \mathbf{x}) \leqslant 1$.

- If the support of $\mathbf{X}$ is unbounded there will exist values of $\boldsymbol{\beta}$ and $\mathbf{x}$ for which $\mathbf{x}'\boldsymbol{\beta} \notin [0, 1]$.

- In general it is not guaranteed that $\mathbf{x}'\boldsymbol{\beta}$ lies in the **admissible range** for $\mathrm{E}[Y|\mathbf{X} = \mathbf{x}]$.

Problem 2 - Heterosckedasticity

- We also know that $Var(Y) = p(1 - p)$.
- Linearity of the conditional expectation $E[Y|\mathbf{X} = \mathbf{x}] = p(\mathbf{x})$ will then imply:

$$Var(Y|\mathbf{X} = \mathbf{x}) = p(\mathbf{x})\,(1 - p(\mathbf{x})) = \mathbf{x}'\boldsymbol{\beta}(1 - \mathbf{x}'\boldsymbol{\beta}),$$

  which is not constant **by definition**.
- This is in contrast with the usual **omosckedasticity** assumption typical of the linear regression settings:

$$Var(Y|\mathbf{X} = \mathbf{x}) = \sigma^2$$

Linear probability model

- For all these reasons applying a linear model to binary outcomes bears several disadvantages.
- Still, when this occur we talk about **Linear Probability Models** (LPM), since we are modelling the conditional probability of success $(Y = 1)$ through a linear function.
- Coefficients are still interpreted as the marginal effect of the corresponding variable on the **probability** of $Y = 1$.
- Since errors are heteroskedastic by construction, standard errors robust to heteroskedasticity must be considered.
- Let's take a look at more appealing **alternatives** to model binary outcomes.

Latent Index Models

- If problems are originated from limited support of Y, let us assume that there exists a continuous unobserved latent variable, Y^*, whose support is the entire real line, that might be thought as a proxy for Y.
- We could think of Y^* as the **propensity** to observe $Y = 1$. The higher Y^* the more likely it is to observe $Y = 1$.
- If we could observe Y^*, in place of Y, there would be no problems in modelling the conditional distribution $Y^*|\mathbf{X}$ with a linear regression model. For instance:

$$Y^* = \mathbf{x}'\boldsymbol{\beta} - \epsilon$$

Latent Index Models

- Unfortunately, Y^* is unobserved, what we actually observe is:

$$Y = \begin{cases} 1 & \text{if } Y^* > 0 \\ 0 & \text{if } Y^* \leqslant 0 \end{cases}$$

- Hence we will observe $Y = 1$ for higher values of Y^* and $Y = 0$ for lower values of Y^*.
- The choice of the threshold (0) is purely **arbitrary**, since Y^* is unobserved and has no meaningful scale.
- We might then ask what can we say about the conditional expectation of the **observed** Y given $\mathbf{X}$ if the latent variable Y^* has a linear conditional expectation.

Latent Index Models

- If $Y^* = \mathbf{x}'\boldsymbol{\beta} - \epsilon$, it is easy to show that:

$$\begin{aligned} E[Y|\mathbf{X} = \mathbf{x}] &= \Pr(Y = 1|\mathbf{X} = \mathbf{x}), \\ &= \Pr(Y^* > 0|\mathbf{X} = \mathbf{x}), \\ &= \Pr(\mathbf{x}'\boldsymbol{\beta} - \epsilon > 0|\mathbf{X} = \mathbf{x}), \\ &= \Pr(\epsilon < \mathbf{x}'\boldsymbol{\beta}), \\ &= F(\mathbf{x}'\boldsymbol{\beta}), \end{aligned}$$

  where $F(\cdot)$ is the **cumulative distribution function** (c.d.f.) of the (unobserved) random variable ϵ.

Non-linear conditional expectation

- The conditional expectation $E[Y|\mathbf{X} = \mathbf{x}]$ is given by the familiar linear combination $\mathbf{x}'\boldsymbol{\beta}X$ transformed through the cumulative distribution function of ϵ.
- Linearity of $E[Y^*|\mathbf{X} = \mathbf{x}]$ **implies** non-linearity of $E[Y|\mathbf{X} = \mathbf{x}]$, since $F(\cdot)$ is a non-linear function that maps values on the entire real line to values in the range $(0, 1)$.
- Unfortunately, $F(\cdot)$ is unknown. We will then need to make explicit assumptions on the distribution of ϵ.
- Different assumptions on the form of $F(\cdot)$ will translate into different models for the conditional expectation of Y given $\mathbf{X}$.

Probit Model

- A very common choice is $\epsilon \sim N(0,1)$, that is $F(t) = \Phi(t)$, where $\Phi(\cdot)$ is the c.d.f. of the standard normal distribution.
- We will refer to this model as to the **Probit Model**.
- It is:
$$E[Y|\mathbf{X} = \mathbf{x}] = \Phi(\mathbf{x}'\boldsymbol{\beta}),$$
or equivalently:
$$Y|\mathbf{X} = \mathbf{x} \sim Be(\Phi(\mathbf{x}'\boldsymbol{\beta})).$$

Logit Model

- Another very common choice is ϵ distributed according to a **logistic** distribution, that is $F(t) = \Psi(t)$, where $\Psi(\cdot)$ is the c.d.f. of a ligistic distribution given by:

$$\Psi(t) = \frac{e^t}{1 + e^t}.$$

- We will refer to this model as to the **Logit Model**.

- It is:

$$E[Y|\mathbf{X} = \mathbf{x}] = \Psi(\mathbf{x}'\boldsymbol{\beta}),$$

  or equivalently:

$$Y|\mathbf{X} = \mathbf{x} \sim Be(\Psi(\mathbf{x}'\boldsymbol{\beta})).$$

Assumptions

- We still need to make the assumptions typical of the linear regression model, suitably restated in terms of Y^*:
  - **indipendence** - the errors ϵ are independent and identically distributed,
  - **incorrelation** - $E[\epsilon|\mathbf{X}] = 0$,
  - **distribution** - the errors ϵ are distributed according to a random variable with c.d.f. $F(\cdot)$.
- Keep in mind that ϵ here are the residuals of the population regression of Y^* on $\mathbf{X}$.
- Since $Y|\mathbf{X} = \mathbf{x} \sim Be(F(\mathbf{x}'\boldsymbol{\beta}))$ it then follows that:

$$Var(Y|\mathbf{X} = \mathbf{x}) = F(\mathbf{x}'\boldsymbol{\beta})(1 - F(\mathbf{x}'\boldsymbol{\beta})).$$

Estimation

- An estimate for β could be obtained by means of least squares estimation, that is minimizing the sample analogue of the quantity:
$$E[(Y - F(\mathbf{x}'\boldsymbol{\beta}))^2].$$
- This is the so called **non-linear least squares** estimator.
- An estimation method more often adopted in this case is however based on the **likelihood function**, which exploits full knowledge of the distribution of ϵ.
- Both methods provide unbiased estimates for the vector $\boldsymbol{\beta}$. The maximum likelihood estimator is the most efficient one in this case.

Interpreting the Coefficients

- Once we have obtained estimates of $\boldsymbol{\beta}$ how do we interpret the coefficients?
- $\boldsymbol{\beta}$ is the vector of marginal effects of the regressors on Y^*:

$$\frac{\partial E[Y^*|\mathbf{X} = \mathbf{x}]}{\partial \mathbf{x}} = \frac{\partial \mathbf{x}'\boldsymbol{\beta}}{\partial \mathbf{x}} = \boldsymbol{\beta}.$$

- Is this quantity really interesting by itself? No it isn't!
- Since Y^* is a latent random variable without any meaningful scale, the marginal effect of $\mathbf{X}$ on Y^* alone does not tell us anything about the **magnitude** of the effect on Y.
- How can we use $\boldsymbol{\beta}$ to obtain the marginal effect of $\mathbf{X}$ on Y?

Interpreting the Coefficients

- Non-linearity of the conditional expectation of Y given $\mathbf{X}$ implies **non-constant** marginal effects of $\mathbf{X}$ on Y:

$$\begin{aligned} \frac{\partial E[Y|\mathbf{X}=\mathbf{x}]}{\partial \mathbf{x}} &= \frac{\partial F(\mathbf{x}'\boldsymbol{\beta})}{\partial \mathbf{x}}, \\ &= f(\mathbf{x}'\boldsymbol{\beta})\boldsymbol{\beta}, \end{aligned}$$

  where $f(\cdot)$ is the derivative of the function $F(\cdot)$, that is the **density function** of ϵ.

- Marginal effects of $\mathbf{X}$ on Y actually depend on the value of $\mathbf{x}$, i.e. it is not constant for all individuals.
- Still, since $f(\cdot)$ is a strictly positive function in its support, the marginal effect is:
  - of the same sign of $\boldsymbol{\beta}$;
  - equal to zero only when $\boldsymbol{\beta} = 0$.

Interpreting the Coefficients

- When $\boldsymbol{\beta}$ is a vector of coefficients, the marginal effect of one regressor on $\mathrm{E}[Y|\mathbf{X}]$ depends on the entire vector of coefficients $\boldsymbol{\beta}$ and on the full vector $\mathbf{x}$.
- As a consequence, when dealing with non-linear models the marginal effect is almost **individual specific**.
- If the interest is on interpreting and reporting marginal effects we need to summarize this information.
- How can we report estimates about marginal effects?
- Mostly two choices:
  - Average marginal effects over the observed population;
  - Compute marginal effects for a representative individual in the population;

Average partial effect

- The **average partial effect** (APE) of $\mathbf{X}$ on Y is given by the average, over the distribution of $\mathbf{X}$, of the individual marginal effects:

$$APE = E[f(\mathbf{x}'\boldsymbol{\beta})\boldsymbol{\beta}] = \boldsymbol{\beta} E[f(\mathbf{x}'\boldsymbol{\beta})].$$

- This is easily computed from raw data, given an estimate for $\boldsymbol{\beta}$, as:

$$\widehat{APE} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\boldsymbol{\beta}} f(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}) = \widehat{\boldsymbol{\beta}} \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}).$$

- Standard errors for $\widehat{APE}$ could be obtained through Delta method or bootstrap techniques.

Partial effect at the average

- It might be interesting to summarize individual marginal effects by reporting the marginal effect for a **representative individual**.
- A common choice is the (hypothetical) individual with average value of the regressors. This is how we define the **partial effect at the average** (PEA) of $\mathbf{X}$ on Y:

$$PEA = \boldsymbol{\beta} f(E[\mathbf{X}]'\boldsymbol{\beta})$$

- This is again easily computed from raw data, given an estimate for $\boldsymbol{\beta}$, as:

$$\widehat{PEA} = \widehat{\boldsymbol{\beta}} f\left(\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\right)'\boldsymbol{\beta}\right).$$

- Non-linearity of the function $f(\cdot)$ **implies** $\widehat{APE} \neq \widehat{PAE}$.

Evaluating the model

- We should not rely on the R^2 index when comparing different probit/logit models.
- The R^2 index does not vary between 0 and 1 when the support of the dependent variable is $\{0, 1\}$.
- For this purpose there exist pseudo-R^2 indices computed from quantities linked to the likelihood function.
- When interested in comparing the **predictive power** of competing models we could instead compute the proportion of statistical units that we could correctly classify into the two groups.
- This will be our measure of goodness of fit.

Predictive power of the model

- Define $\widehat{p}_i$ as the model predicted probability for $Y = 1$ for individual i. We could then classify individuals according to $\widehat{y}_i$ such that:

$$\widehat{y}_i = \begin{cases} 0 & \widehat{p}_i < 0.5 \\ 1 & \widehat{p}_i \geqslant 0.5 \end{cases}$$

- By comparing y_i and $\widehat{y}_i$ we obtain the two way table:

		y_i	
		0	**1**
$\widehat{y}_i$	**0**	e_{00}	e_{01}
	1	e_{10}	e_{11}

where $e_{00} + e_{11}$ is the proportion of correctly classified statistical units.

Classification errors

- $e_{10} + e_{01}$ will be the proportion of observations incorrectly classified.
- e_{10} will be the proportion of observation incorrectly classified among those with $y_i = 0$ - **false positive**.
- e_{01} will be the proportion of observation incorrectly classified among those with $y_i = 1$ - **false negatives**.
- These proportions are sometimes referred in the literature in terms of:
  - **sensitivity** - $e_{11}/(e_{11} + e_{01})$,
  - **specificity** - $e_{00}/(e_{10} + e_{00})$

Back to the linear probability model

- The LPM is obtained by applying a linear model in the presence of a binary outcome.
- OLS consistently identify the coefficients of the linear regression when assuming:

$$E[Y|X] = \beta_0 + \beta_1 X.$$

- Omosckedasticity of the errors, i.e. $Var(Y|X) = \sigma^2$, is not holding by construction in this context, hence robust standard errors must be considered.
- One major **disadvantage**: predicted probability of "success" might fall outside the range $[0, 1]$.
- One major **advantage**: estimated coefficients can be directly interpreted as marginal effects of the corresponding covariate on the probability of "success".

Marginal effects - Binary covariate

- When X is a dummy variable LPM and Latent Index Models provide the **same results** in terms of estimated probabilities, since we could write:

$$\begin{aligned} E[Y|X] &= F(\beta_0^* + \beta_1^* X), \\ &= F(\beta_0^*) + [F(\beta_0^* + \beta_1^*) - F(\beta_0^*)]\, X, \\ &= \beta_0 + \beta_1 X, \end{aligned}$$

  with $\beta_0 = F(\beta_0^*)$ and $\beta_1 = F(\beta_0^* + \beta_1^*) - F(\beta_0^*)$.

- The conditional expectation of Y given X can then be written as a linear function of the dummy variable X.

Marginal effects - Binary covariate

- We will interpret the coefficients of the linear probability model as:

$$\begin{aligned}\beta_0 &= \text{predicted probability of } Y = 1 \text{ when } X = 0;\\ \beta_1 &= \text{difference in the predicted probabilities of } Y = 1\\ &\quad \text{when } X = 1 \text{ and } X = 0;\end{aligned}$$

- The same argument holds when we consider **saturated** models, that is regression modesl whith only discrete covariates coded as dummies with all interactions.

Marginal effects - Continuous covariate

- When X is a continuous random variable then predicted probabilities for LPM and Latent Index Models differ.
- Still, $\beta_0 + \beta_1 X$ can be shown to be the **best linear approximation** to the true conditional expectation function $E[Y|X]$, while Latent Index Models rely on the correct specification of the function $F(\cdot)$.
- As a consequence we might think of β_1 as the best approximation to the marginal effect for X on the probability that $Y = 1$.

Marginal effects - Continuous covariate

- The advantage here comes from the fact that there is no need for computing individual marginal effects and summarizing the information, since β_1 is already a summary of the marginal effect of interest.
- In practice APE or PEA from Latent Index Models and marginal effects for LPM are very similar.
- If this is not the case you might wonder why...

Bottom line

- Latent Index Models provide a useful alternative to model binary outcomes.
- These models are well suited for making predictions, since they closely approximate the underlying conditional expectation function $\mathrm{E}[Y|\mathbf{X}]$.
- When it comes to reporting marginal effects though there is usually little gain in adopting these more flexible specifications.
- Linear probability models provide an interesting alternative for this task, in that estimated coefficients are directly interpreted as (average) marginal effects.
- An additional advantage is given by the straightforward applicability of Instrumental Variables estimators to account for potential endogeneity of X.