Lecture 4

Limited Variables 3 of 5: Binary Response Models II Logit & Probit Models

Michael Curran

Trinity College Dublin

JS Econometrics

Lecture 4 Outline

Introduction

- **Limited dependent variable** (**LDV**): a dependent variable whose range of values is substantively restricted. Examples? LPM.
- Drawbacks of LPM are overcome by logit and probit, though they are more difficult to interpret.
- Alternative limited dep variables: optimizing behaviour often leads to a **corner solution response** for some nontrivial fraction of the population. Examples? Tobit model. Others: count variable (Poisson regression models), sometimes observe LDV due to data censoring (censored and truncated models).
- Finally, general problem of self-selection (nonrandom sample).
- LDV are fine for time series and panel data but most often applied in cross sectional data; sample selection problems are usually confined to cross sectional or panel data. My coverage: cross sectional data. Prof Bénétrix will introduce time series data and panel data.

Link with Lecture 2

- Binary response continued from LPM: could call these three 'binary response limited dep variable models.'
- Three problems with LPM: (i) fitted probabilities can lie outside zero and one, (ii) constant partial effects of any explanatory variable appearing in level form and (iii) heterogeneity.
- First two limitations can be overcome by using more sophisticated **binary response models**. In binary response models, interest lies mainly in **response probability**:

$$P(y = 1|\mathbf{x}) = P(y = 1|x_1, x_2, \ldots, x_k)$$

where $\mathbf{x}$ is full set of explanatory variables.

Specifying logit and probit models

Overcoming first limitation of LPM

LPM: response probability is linear in a set of parameters β_j:

$$P(y = 1|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots \beta_k x_k$$

Now consider a class of binary response models of the form:

$$P(y = 1|\mathbf{x}) = G(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k) = G(\beta_0 + \mathbf{x}\boldsymbol{\beta}) \quad (1)$$

where $0 < G(z) < 1$ for all $z \in \mathbb{R}$. Ensures estimated response probabilities are strictly between zero and one.

$$\mathbf{x}\boldsymbol{\beta} = \beta_1 x_1 + \ldots + \beta_k x_k$$

In **logit model**, G is the logistic function:

$$G(z) = \frac{\exp(z)}{1 + \exp(z)} = \Lambda(z)$$

which is between zero and one for all $z \in \mathbb{R}$.

Specifying logit and probit models

Overcoming first limitation of LPM

In the **probit model**, G is the standard normal CDF, which is expressed as an integral:

$$G(z) = \Phi(z) = \int_{-\infty}^{z} \phi(v)dv$$

where $\phi(z)$ is the standard normal density

$$\phi(z) = \frac{1}{(2\pi)^{-\frac{1}{2}}} \exp\left(-\frac{z^2}{2}\right)$$

G in logit and probit are both increasing functions. Each increases most quickly at $z = 0$; $G(z) \longrightarrow 0$ as $z \longrightarrow -\infty$ and $G(z) \longrightarrow 1$ as $z \longrightarrow \infty$.

Specifying logit and probit models

Derivation

Can derive logit and probit models from underlying **latent variable model**. Let y^* be an unobserved or *latent* variable:

$$y^* = \beta_0 + \mathbf{x}\boldsymbol{\beta} + e, \quad y = 1[y^* > 0] \tag{2}$$

where the *indicator function* $1[\cdot]$ defines a binary outcome:

$$1[\textit{statement}] = \begin{cases} 1 & \text{if statement is true} \\ 0 & \text{otherwise} \end{cases}$$

Assume e is independent of $\mathbf{x}$ and that e either has the standard logistic distribution or the standard normal distribution. Either way, e is symmetrically distributed about zero, so:

$$1 - G(-z) = G(z) \quad \forall z \in \mathbb{R}$$

Probit is more popular since economists like the normality assumption.

Specifying logit and probit models

Derivation & Limitations

From (2) and our assumptions, we can derive response prob. for y:

$$\begin{aligned} P(y = 1|\mathbf{x}) &= P(y^* > 0|\mathbf{x}) = P[e > -(\beta_0 + \mathbf{x}\boldsymbol{\beta})|\mathbf{x}] \\ &= 1 - G[-(\beta_0 + \mathbf{x}\boldsymbol{\beta})] = G(\beta_0 + \mathbf{x}\boldsymbol{\beta}) \end{aligned}$$

which is the same as (1).

Limitations: interpretation. Usual goal in applying binary response models: explain effects of x_j on $P(y = 1|\mathbf{x})$. For logit and probit, the *direction* of the effect of x_j on $E(y^*|\mathbf{x}) = \beta_0 + \mathbf{x}\boldsymbol{\beta}$ and on $E(y|\mathbf{x}) = P(y = 1|\mathbf{x}) = G(\beta_0 + \mathbf{x}\boldsymbol{\beta})$ is always the same. But latent variable y^* rarely has a well-defined unit of measurement, so magnitudes of β_j are not especially useful by themselves. Mostly, we want to estimate effect of x_j on $P(y = 1|\mathbf{x})$, but that's complicated by the nonlinear nature of $G()$.

Specifying logit & probit models

Partial effects

$$\frac{\partial p(\mathbf{x})}{\partial x_j} = g(\beta_0 + \mathbf{x}\boldsymbol{\beta})\beta_j \tag{3}$$

Logit and probit: $G()$ strictly increasing cdf and so $g(z) > 0 \; \forall z$ so partial effect of x_j on $p(\mathbf{x})$ depends on $\mathbf{x}$ through the positive quantity $g(\beta_0 + \mathbf{x}\boldsymbol{\beta})$, which means that the partial effect always has the same sign as β_j. Equation (3) shows that *relative* effects of any two continuous explanatory variables don't depend on $\mathbf{x}$: ratio of partial effects for x_j and x_h is $\frac{\beta_j}{\beta_h}$. x_1: binary explanatory variable, then partial effect from changing x_1 from zero to one holding all other variables fixed is:

$$G(\beta_0 + \beta_1 + \beta_2 x_2 + \ldots + \beta_k x_k) - G(\beta_0 + \beta_2 x_2 + \ldots + \beta_k x_k) \tag{4}$$

This depends on all the values of the other x_j. Sign of β_1 is sufficient to see if the program had a positive or negative effect, but to find the *magnitude*, we must estimate the quantity in (4).

Specifying logit & probit models

Partial effects

Can use (4) for other kinds of discrete variables (e.g. number of children, x_k), so effect on probability of x_k going from c_k to $c_k + 1$:

$$G[\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k (c_k + 1)] - G(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k c_k) \tag{5}$$

Including e.g. z_1^2, $\log(z_2)$ as explanatory variables:

$$P(y = 1|\mathbf{z}) = G(\beta_0 + \beta_1 z_1 + \beta_2 z_1^2 + \beta_3 \log(z_2) + \beta_4 z_3)$$

Partial effect of z_1 on $P(Y = 1|\mathbf{z})$ is
$\frac{\partial P(y=1|\mathbf{z})}{\partial z_1} = g(\beta_0 + \mathbf{x}\boldsymbol{\beta})(\beta_1 + 2\beta_2 z_1)$ and for z_2:
$\frac{\partial P(y=1|\mathbf{z})}{\partial z_2} = g(\beta_0 + \mathbf{x}\boldsymbol{\beta})(\frac{\beta_3}{z_2})$ where
$\mathbf{x}\boldsymbol{\beta} = \beta_1 z_1 + \beta_2 z_1^2 + \beta_3 \log(z_2) + \beta_4 z_3$. So, $g(\beta_0 + \mathbf{x}\boldsymbol{\beta})(\frac{\beta_3}{100})$ is approximate change in response probability when z_2 increases by 1%. Models with interactions among explanatory variables (including those between discrete and continuous variables) are handled similarly. Use (5) to measure effects of discrete variables.

Lecture 4 Outline

Estimating Logit & Probit Models

Unlike LPM, nonlinear nature of $E(y|\mathbf{x})$ render OLS and WLS inapplicable. Can use NLLS / NLWLS / **maximum likelihood estimation** (**MLE**). With MLE, heteroscedasticity is accounted for.

$$f(y|\mathbf{x}_i;\boldsymbol{\beta}) = [G(\mathbf{x}_i\boldsymbol{\beta})]^y[1 - G(\mathbf{x}_i\boldsymbol{\beta})]^{1-y}, \quad y = 0, 1 \qquad (6)$$

Intercept is in vector $\mathbf{x}$. When $y = 1$, we have $G(\mathbf{x}_i\boldsymbol{\beta})$ and when $y = 0$, we have $1 - G(\mathbf{x}_i\boldsymbol{\beta})$. **log-likelihood function** for obs i is a function of the parameters and the data $(\mathbf{x}_i, y_i)$ and is simply the log of (6):

$$\ell_i(\boldsymbol{\beta}) = y_i \log[G(\mathbf{x}_i\boldsymbol{\beta})] + (1 - y_i) \log[1 - G(\mathbf{x}_i\boldsymbol{\beta})]$$

Since $G()$ strictly between 0 and 1 for logit and probit, $\ell_i(\boldsymbol{\beta})$ is well-defined for all values of $\boldsymbol{\beta}$.

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}) \qquad (7)$$

MLE of $\boldsymbol{\beta}$: $\hat{\boldsymbol{\beta}}$ maximises (7).

Estimating Logit & Probit Models

- If $G()$ is standard logit cdf, then $\hat{\boldsymbol{\beta}}$ is *logit estimator*.
- If $G()$ is standard normal cdf, then $\hat{\boldsymbol{\beta}}$ is *probit estimator*.
- Cannot write formulas due to the nonlinear nature of the maximisation problem, which raises issues but MLE is consistent, Asymptotically Normal and Asymptotically Efficient.
- Each $\hat{\beta}_j$ has asymptotic standard errors reported in most stat packages and with these, we can construct asymptotic t tests and confidence intervals.
- To test $H_0 : \beta_j = 0$, form t stat $\frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$ and carry out test normally once decided on one or two sided alternative.

Lecture 4 Outline

Testing Multiple Hypotheses

3 ways to test multiple exclusion restrictions

Focus on exclusion restrictions.

1. LM / score test.
2. Wald test requires estimation of only the unrestricted model. In linear model case, **Wald statistic** after a simple transformation is the F stat so no need to cover Wald statistic separately. Wald statistic is computed by econometrics packages allowing for exclusion restrictions to be tested after unrestricted model has been estimated. It has asymptotic chi-square dist with degrees of freedom equal to number of restrictions being tested.
3. If both restricted and unrestricted models are easy to estimate, then *likelihood ratio* (LR) *test* is attractive.

Testing Multiple Hypotheses

LR Test

- Same concept as F in linear model.
- F measures increase in SSR when variables are dropped from model.
- LR test is based on difference in log-likelihood functions for unrestricted and restricted models.
- Idea: because MLE max log-likelihood, dropping variables generally leads to a *smaller* or at least no larger log-likelihood.

$$LR = 2(\mathcal{L}_{ur} - \mathcal{L}_r) \tag{8}$$

- Since $\mathcal{L}_{ur} \geq \mathcal{L}_r$, LR is nonnegative and usually strictly positive; it has an approximate chi-square distribution under H_0 and if we are testing q exclusion restrictions, $LR \overset{a}{\sim} \chi^2_q$.
- Question 17.1, Wooldridge.

Lecture 4 Outline

Interpreting Estimates

Percent Correctly Predicted

Should report coefficient, SE and value of log-likelihood function. For LPM, goodness-of-fit measure is **percent correctly predicted**. Define binary predictor of y_i to be one if predicted prob is at least .5 and zero otherwise, i.e. $\tilde{y}_i = 1$ if $G(\hat{\beta}_0 + \mathbf{x_i}\boldsymbol{\beta}) \geq .5$ and $\tilde{y}_i = 0$ if $G(\hat{\beta}_0 + \mathbf{x_i}\boldsymbol{\beta}) < .5$. Given $\{\tilde{y}_i : i = 1, \ldots, n\}$, we can see how well $\tilde{y}_i$ predicts y_i across obs. Four possible outcomes on each pair $(y_i, \tilde{y}_i)$ and when both are zero or both are one, we make the correct prediction. In the two cases where one of the pair is zero and the other is one, we make the incorrect prediction. The percent correctly predicted is the percentage of times that $\tilde{y}_i = y_i$. Can be misleading though, e.g. get high percentages correctly predicted even when least likely outcome is very poorly predicted. Also compute the percent correctly predicted for each of the outcomes.

Interpreting Estimates

Percent Correctly Predicted

Criticism: threshold value of .5 especially when one of the outcomes is unlikely, e.g. if $\bar{y} = 0.08$ (only 8% successes in sample) it could be that we *never* predict $y_i = 1$ since estimated probability of success is never greater than .5.

One alternative: use fraction of successes in sample as threshold (e.g. .08 on page 589), i.e. define $\tilde{y}_i = 1$ when $G(\hat{\beta}_0 + \mathbf{x_i}\boldsymbol{\beta}) \geq .08$ and zero otherwise. Using this rule will certainly increase the number of predicted successes, but not without cost: we will necessarily make more mistakes – perhaps many more – in predicting zeros (failures).

Another alternative: choose threshold so fraction of $\tilde{y}_i = 1$ in sample is same / close to $\bar{y}$, i.e. search over threshold values $\tau : 0 < \tau < 1$ so if define $\tilde{y}_i = 1$ when $G(\hat{\beta}_0 + \mathbf{x_i}\boldsymbol{\beta}) \geq \tau$ then $\sum_{i=1}^{n} \tilde{y}_i \approx \sum_{i=1}^{n} y_i$.

Interpreting Estimates

Pseudo R-squared

Various **pseudo R-squared** measures exist for binary response. McFadden (1974): $1 - \frac{\mathcal{L}_{ur}}{\mathcal{L}_0}$ where $\mathcal{L}_{ur}$ is log-likelihood for estimated model and $\mathcal{L}_0$ is that in model with only an intercept. Note $\frac{\mathcal{L}_{ur}}{\mathcal{L}_0} = \frac{|\mathcal{L}_{ur}|}{|\mathcal{L}_0|}$. Also, $|\mathcal{L}_{ur}| \leq |\mathcal{L}_0|$. If covariates have no explanatory power, then $\frac{\mathcal{L}_{ur}}{\mathcal{L}_0} = 1$ and pseudo R-squared is zero. Usually, $|\mathcal{L}_{ur}| < |\mathcal{L}_0|$ so $1 - \frac{\mathcal{L}_{ur}}{\mathcal{L}_0} > 0$. If $\mathcal{L}_{ur} = 0$, pseudo R-squared is one. $\mathcal{L}_{ur}$ can't reach zero in probit or logit model. Alternative pseudo R-squareds for probit and logit are more directly related to usual R^2 from OLS in LPM. Let $\hat{y}_i = G(\hat{\beta}_0 + \mathbf{x_i}\hat{\boldsymbol{\beta}})$. These estimate $E(y_i|\mathbf{x_i})$, so can base R-squared on how close $\hat{y}_i$ are to y_i. Could compute squared correlation between y_i and $\hat{y}_i$. Directly comparable to usual R^2 form of estimation of a LPM. Goodness-of-fit is usually less important than trying to obtain convincing estimates of the *cet par* effects of the explanatory variables.

Interpreting Estimates

Partial Effects

$$\Delta \widehat{P(y = 1|\mathbf{x})} \approx [g(\hat{\beta}_0 + \mathbf{x}\hat{\boldsymbol{\beta}})\hat{\beta}_j]\Delta x_j \quad (9)$$

Cost of logit and probit: partial effects in (9) harder to summarize since scale factor $g(\hat{\beta}_0 + \mathbf{x}\hat{\boldsymbol{\beta}})$ depends on $\mathbf{x}$ (all of the explanatory variables). One fix: plug in values for x_j like means, medians, mins, maxs, lower and upper quartiles and see how $g(\hat{\beta}_0 + \mathbf{x}\hat{\boldsymbol{\beta}})$ changes. Attractive, but tedious and results in too much info even if number of explanatory variables is moderate. Quick way to get magnitudes of partial effects: use single scale factor to multiply each $\hat{\beta}_j$. One method in econometrics packages: replace each explanatory variable with sample average, i.e. adjustment factor is:

$$g(\hat{\beta}_0 + \bar{\mathbf{x}}\hat{\boldsymbol{\beta}}) = g(\hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \hat{\beta}_2\bar{x}_2 + \ldots + \hat{\beta}_k\bar{x}_k) \quad (10)$$

where $g()$ is N(0,1) (probit) and $g(z) = \frac{\exp(z)}{[1+\exp(z)]^2}$ (logit).

Idea: when (10) is multiplied by $\hat{\beta}_j$, we get partial effect of x_j for average person in sample

Interpreting Estimates

Partial Effects

Two potential problems with this motivation:

1. If some explanatory variables are discrete, average of them represents no one in sample (or population). Example?
2. If a continuous explanatory variable appears as a nonlinear function (e.g. natural log or quadratic), it's not clear whether we want to average the nonlinear function or plug the average into the nonlinear function. E.g., should we use $\overline{\log(sales)}$ or $\log \overline{sales}$? Econometrics packages computing scale factor as above default to the former: software is written to compute the average of the regressors included in the probit or logit estimation.

Different approach to computing a scale factor circumvents issue of which values to plug in for explanatory variables. Instead, second scale factor results from averaging the individual partial effects across sample (**average partial effect**) [**APE**].

Interpreting Estimates

Partial Effects

x_j continuous, APE is:

$$\frac{1}{n}\sum_{i=1}^{n}[g(\hat{\beta}_0 + \mathbf{x_i}\hat{\boldsymbol{\beta}})\hat{\beta}_j] = \left[\frac{1}{n}\sum_{i=1}^{n} g(\hat{\beta}_0 + \mathbf{x_i}\hat{\boldsymbol{\beta}})\right]\hat{\beta}_j \qquad (11)$$

Term multiplying $\hat{\beta}_j$ acts as a scale factor. In the probit case $g(\hat{\beta}_0 + \mathbf{x_i}\hat{\boldsymbol{\beta}}) = \phi(\hat{\beta}_0 + \mathbf{x_i}\hat{\boldsymbol{\beta}})$ and $g(\hat{\beta}_0 + \mathbf{x_i}\hat{\boldsymbol{\beta}}) = \frac{\exp(\hat{\beta}_0 + \mathbf{x_i}\hat{\boldsymbol{\beta}})}{[1+\exp(\hat{\beta}_0 + \mathbf{x_i}\hat{\boldsymbol{\beta}})]^2}$ in the logit case. The two scale factors differ since here we are using the average of the nonlinear function (11) rather than the nonlinear function of the average (10). Neither make much sense for discrete explanatory variables; better to use equation (5) to directly estimate the change in the probability. For a change in x_k from c_k to $c_k + 1$, discrete analog of partial effect based on (10) is given by:

$$G[\hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \ldots + \hat{\beta}_{k-1}\bar{x}_{k-1} + \hat{\beta}_k(c_k + 1)] - G(\hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \ldots + \hat{\beta}_{k-1}\bar{x}_{k-1} + \hat{\beta}_k c_k) \qquad (12)$$

For binary x_k, equation (12) is computed by Stata.

Interpreting Estimates

Partial Effects

APE (more comparable to LPM) is given by:

$$\frac{1}{n}\sum_{i=1}^{n} G[\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_{k-1} x_{ik-1} + \hat{\beta}_k (c_k + 1)] - G(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_{k-1} x_{ik-1} + \hat{\beta}_k c_k \qquad (13)$$

Obtaining (13) for probit or logit is simple:

1. For each obs, estimate probability of success for the two chosen values of x_k, plugging in actual outcomes for other explanatory variables (so have n estimated differences).
2. Average the differences in estimated probabilities across all observations.
3. If x_k is binary, plug in one and zero as the only two possible values.

Interpreting Estimates

Partial Effects

- When using logit, probit or LPM, it makes sense to compute scale factors for probit and logit in making comparisons of partial effects.
- But may want a quicker way to compare magnitudes of the different estimates.
- Probit: $g(0) \approx .4$ and logit: $g(0) = .25$ so to make the magnitudes of probit and logit roughly comparable, we can multiply the probit coefficients by $\frac{.4}{.25} = 1.6$ or we can multiply the logit estimates by .625.
- In LPM, $g(0)$ is effectively one, so logit slope estimates can be divided by 4 to make them comparable to LPM estimates and probit slope estimates can be divided by 2.5.
- Still, most cases, want more accurate comparisons obtained by using scale factors (term multiplying β_j in (11)) for logit and probit.

Interpreting Estimates

- Example 17.1. Question 17.2. Figure 17.2.
- Same issues concerning endogenous explanatory variables in linear models arise in logit/probit.
- Not cover them, but possible to test and correct using methods related to 2SLS, (Evans and Schwab, 1995).
- Link to lecture 8 on 2SLS.

Interpreting Estimates

Further Limitations

Further limitations (probit) – misspecification problems in latent variable models:

1. Nonnormality of e in latent variable model: if e doesn't have standard Normal distribution, response probability will not have probit form. Since the response probability is unknown, we couldn't estimate the magnitude of partial effects even if we had consistent estimates of the β_j.
2. Heteroscedasticity in e: if $var(e|\mathbf{x})$ depends on $\mathbf{x}$, response probability no longer has form $G(\beta_0 + \mathbf{x}\boldsymbol{\beta})$; instead, it depends on form of the variance and requires more general estimation. Such models are not often used in practice since logit and probit with flexible functional forms in the independent variables tend to work well.

Interpreting Estimates

Final Comments

- Binary response models apply to time series and panel data (independent but not necessarily identically distributed).
- Logit and probit also used to evaluate impact of certain policies in context of a natural experiment.
- Recently popular: logit and probit with unobserved effects (done in 'Advanced' Wooldridge).

Lecture 4 Outline

Summary

- Specifying binary response: logit & probit overcome LPM drawbacks but difficult to interpret.
- Nonlinear estimation techniques necessary, e.g. MLE.
- Test multiple hypotheses via LM / Wald / LR tests.
- Goodness-of-fit measure: percent correctly predicted & pseudo R^2.
- Can interpret estimates via partial effects, e.g. APE.

References

- Logit & Probit Models: Wooldridge 17.1.