Binary Dependent Variable Models

2025/2026, Semester 1

Ralf A. Wilke
Copenhagen Business School

- Binary Dependent Variables/Binary choice models

- Linear Probability Model, Logit, Probit
  - Wooldridge (2025), Chapter 7.5 and 17.1
  - Wooldridge (2010), Chapter 15

Limited Dependent Variable Models

- Limited Dependent Variable (LDV)
- A LDV is defined as a dependent variable whose values are substantially restricted.
- LDV models can be used for time series and cross sectional data but they are more often applied to cross section data.

- Examples:
  - **Binary dependent variable is 0 or 1.**
  - Participation percentage is between 0 and 100.
  - Number of times an individual is arrested (nonnegative integer)

The linear probability model (LPM)

- Binary Dependent Variable
- The dependent variable can just take values 0 and 1.
- What is then the meaning of the model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u \ ?$$

  - β_j can no longer be interpreted as the change in *y* given a one unit change in x_j.
  - Under $E(u|\mathbf{x}) = 0$ we have: $E(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$
  - Since *y* is binary and only takes values 0 and 1, it is always true that:

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

  - The probability of "success" (y=1), $p(x) = P(y = 1|\mathbf{x})$, is a linear function of β. (That's why **LPM**!)

- This implies: $P(y=0|\mathbf{x}) = 1 - P(y=1|\mathbf{x})$
- In the LPM, β_j measures the change in the probability of success when x_j changes other factors fixed:

$$\Delta P(y=1|\mathbf{x}) = \beta_j \Delta x_j$$

- The mechanics of OLS estimation are the same as before.
- Example: In the labour force (MROZ.dta)
  - Success is that a married woman is in the labour force at a given year (*y=1*). We call this variable *inlf*.
  - We have a set of regressors:

    nwifeinc: husband's earnings

    educ: years of education

    exper: labour market experience

    kidslt6: number of kids less than six years

    kidsge6: number of children aged 6-18

- When we estimate the model, we obtain:

$$\begin{aligned}\widehat{inlf} \quad &= \quad 0.586 - 0.0034 nwifeinc + 0.038 educ + 0.039 exper - 0.0006 exper^2 \\ &\quad\ (0.154) \quad (0.0014) \qquad\qquad (0.007) \qquad (0.006) \qquad\qquad (0.00018) \\ &\quad -0.016 age - 0.262 kidslt6 + 0.013 kidsge6 \\ &\quad\ (0.002) \qquad\quad (0.034) \qquad\quad (0.0132) \\ &\quad n = 753, \quad R^2 = 0.264\end{aligned}$$

- Using the *t* statistics, all variables except *kidsge6* are statistically significant and have the expected sign.
- The coefficient educ means that another 10 years of education increases the probability of being in the labour force by 0.038*10=0.38, which is quite a lot.
- Let us now analyse the case for fixed values of ***x*** (except *educ*), *nwifeinc=50, exper=5, age=30, kidslt6=1, kidsge6=0.* In this case the predicted probability is negative until *educ* equals 3.84. This is not good, but since there is nobody in the sample with *educ*<*5*, we shouldn't worry too much.

- When we compute the fitted values for all observations in the sample, we see that for 16 women we obtain a predicted probability less than zero and 17 are greater than one.
- The model predicts that going from zero to four young children implies a predicted drop in the probability by 105p.p..
  - There is no woman in the sample with four young children and just three with three young children.

- The example has illustrated how easy linear probability models are to be interpreted but it also showed some shortcomings.

- How to measure the Goodness of fit in the LPM?
  - The percent correctly predicted: Define $\tilde{y}_i = 1$ if $\hat{y}_i \geq 0.5$ and $\tilde{y}_i = 0$ if $\hat{y}_i < 0.5$ and compare $\tilde{y}_i$ with y_i.
  - The share of $\tilde{y}_i = y_i$ is a measure of goodness of fit.

- There is another problem with the LPM:
  - It does not satisfy one of the Gauss-Markov assumptions.
  - Since *y* is a Bernoulli random variable, its variance is given by:

$$
\begin{aligned}
var(y|\mathbf{x}) &= E(u^2|x) \\
&= (-p(x))^2(1-p(x)) + (1-p(x))^2 p(x) \\
&= p(\mathbf{x})[1-p(\mathbf{x})]
\end{aligned}
$$

with $u = 1 - p(x)$ with probability $p(x)$ y=1

and $u = -p(x)$ with probability $1 - p(x)$ y=0

  - The variance depends on *x*, thus the homoscedasticity assumption is violated.
  - The estimator is still unbiased and consistent but its sample and asymptotic distribution are unknown for us.
  - For this reason t and F statistics are not valid.
  - Use heteroskedasticity robust versions or apply WLS/FGLS.
  - Other work, however, has shown that *t* and *F* statistics are typically not far away from the values obtained with a valid estimator. Therefore, OLS statistics are not completely meaningless.

The Logit and Probit Models

- Also called binary response models.

- Our interest lies primarily in modelling the conditional response probability $P(y = 1|\mathbf{x})$.

- The LPM assumes that it is linear in the parameters β_j which implies several drawbacks.

- To avoid this, we now take a (nonlinear) function G which takes values strictly between zero (>0) and one (<1):
$$P(y = 1|\mathbf{x}) = G(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k) = G(\beta_0 + \mathbf{x}\beta)$$

- Various nonlinear functions have been suggested for G.

- We cover here two which are used in the vast majority of applications (along with the LPM):
  - The logit model: *G* is the logistic function

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

  - The probit model: *G* is the cumulative normal distribution function

$$\begin{aligned} G(z) &= \Phi(z) \\ &= \int_{-\infty}^{z} \phi(\nu) d\nu \\ &= \int_{-\infty}^{z} (2\pi)^{-1/2} exp(-\nu^2/2) d\nu \end{aligned}$$

- Logit and probit models can be derived from an underlying **latent variable model**:

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

where 1[.] is the indicator function:

$$1[y^* > 0] = 1 \text{ if } y* > 0 \text{ and } 1[y^* > 0] = 0 \text{ otherwise.}$$

and *e* is independent of ***x***.

- We assume that *e* either has the standard logistic distribution or the standard normal distribution.
  - Then *e* is symmetrically distributed around zero which means that $G(z) = 1 - G(-z)$.

- Based on the latent variable model and the assumptions we can derive the conditional response probability for y:

$1[y^* > 0] = 1 \text{ if } y* > 0$

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

G is a distribution function of *e*

Symmetry of G

- We are primarily interested in explaining effects of x_j on the conditional response probability $P(y = 1|\mathbf{x})$.
- However, the latent variable model suggests that we are looking at the effects of x_j on y^*.
  - We will see that the direction of the effect on *E(y|**x**)* and *E(y*|**x**)* is the same in logit/probit models.

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

- We want to estimate the effect of x_j on the probability of success $P(y=1|\mathbf{x})$.
- Since G is nonlinear, the magnitudes of the coefficients β_j itself is not useful.
- Instead, the **partial effect** or **marginal effect** of a (roughly) continuous variable x_j is determined by the partial derivative:

$$
\frac{\partial p(\mathbf{x})}{\partial x_j} = g(\beta_0 + \mathbf{x}\beta)\beta_j, \quad where \quad g(z) = \frac{dG}{dz}(z)
$$

  where *g* is a probability density function.
- The magnitude of the partial or marginal effect therefore depends on ***x***.

$$\frac{\partial p(\mathbf{x})}{\partial x_j} = g(\beta_0 + \mathbf{x}\beta)\beta_j, \quad where \quad g(z) = \frac{dG}{dz}(z)$$

- In the logit and probit case: $g(z) > 0$ for all z. This implies that the partial effect has always the same sign as β_j.
  - Logit: $\partial p(x)/\partial x_j = \left(exp(z)/[1 + exp(z)]^2\right)\beta_j$
  - Probit: $\partial p(x)/\partial x_j = \phi(z)\beta_j$
  - …while in the LPM it is simply β_j.
- Moreover, the relative effect of two variables x_j and x_h is the ratio of the two partial effects: β_j/β_h
- When is *g(z)* small and when is it large?
  - One can show that *g(0)=0.4* in the case of probit and *g(0)=0.25* in the case of logit.

- If, say, x_1 is a binary variable switching from 0 to 1, the partial effect is defined by:

$$G(\beta_0 + \beta_1 + \beta_2 x_2 + ...) - G(\beta_0 + \beta_2 x_2 + ...)$$

which also depends on all values of the other x_j.

- Obviously, one can also use squared explanatory variables x_j^2 and logarithmic form $log(x_j)$ among other things. The partial effect then changes.

Maximum Likelihood Estimation of Logit and Probit Models

- As we have now a nonlinear structure, we cannot simply apply OLS.
- We could use non linear least squares estimation techniques but it is no more difficult to use maximum likelihood estimation (MLE).
  - Since MLE is based on the distribution of *y* given $\boldsymbol{x}$, the heteroscedasticity in $var(y|\boldsymbol{x})$ is automatically accounted for.

- We assume that we have a random sample of size n. Then the conditional joint density of two observations *i* and *j* is:

$$f(y_i, y_j|\mathbf{x}) = f(y_i|\mathbf{x})f(y_j|\mathbf{x})$$

- Moreover, the conditional density of y_i given $\mathbf{x}_i$ can be written as

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

  - The contribution to the likelihood is $G(\mathbf{x}_i\beta)$ if *y=1 and* ….

- The log likelihood function of observation *i* is obtained by taking the log:

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

  - It is well defined as G>0 for all values of beta.
- The log likelihood for the sample is obtained by summing across the observations: $L(\beta) = \sum_{i=1}^{n} \ell_i(\beta)$
  - $\hat{\beta}$ is the MLE of β. It maximises $L(\beta)$.
  - Depending on G, $\hat{\beta}$ is called the Logit or Probit Estimator.

- Due to the nonlinear nature of the maximization problem, we don't have closed form solutions for the estimators.
- Computer packages use numerical routines to approximate the derivatives of the ML function.
- Under rather general conditions, the ML estimator is consistent and asymptotically normal and asymptotically efficient and the asymptotic variance is estimated by

$$\widehat{Avar}(\hat{\beta}) = \{\sum_{i=1}^{N} \frac{[g(x_i\hat{\beta})]^2 x_i' x_i}{G(x_i\hat{\beta})[1 - G(x_i\hat{\beta})]}\}^{-1}$$

- Computer packages report asymptotic standard errors and we can construct asymptotic t-statistics and confidence intervals as with OLS.

- Testing multiple hypothesis or (exclusion) restrictions is a bit different as we use different test statistics.
  - Mainly used to test for exclusion restrictions.

- There are three competing approaches:
  - See also tests in general MLE analysis
  - Asymptotically equivalent
  - Likelihood Ratio test (LR)
    - Easy to obtain.
    - Based on a comparison of the restricted and unrestricted model.
  - Lagrange Multiplier (LM) test or core test.
    - Does not require estimation of the unrestricted model.
  - Wald test.
    - Allows for non-linear restrictions to be tested.
    - Based on the unrestricted model.
    - Default in Stata
  - Compare Wooldridge (2010), Sections 15.5.1 and 15.5.2.

Interpretation of Logit and Probit Estimates

- Given the availability of computer packages, the most difficult part of probit and logit estimation is the presentation and interpretation of estimation results.
- The coefficients give the sign of the partial effects of each x_j and we can determine statistical significance by whether we can reject: $H_0 : \beta_j = 0$
- As in the LPM we can use the **percent correctly predicted** as a goodness-of-fit measure.
  - The percent correctly predicted can be misleading as it can be large although the model totally fails to correctly predict one of the outcomes of *y*.
  - For this reason several methods have been suggested to improve this measure (e.g. relate it to the fraction of successes in the sample).

- Alternatively, one can use a **pseudo R-squared** measure.
  - McFadden (1974) suggests the measure: $1 - L_{ur}/L_o$ where L_{ur} is the log likelihood of the model and L_o is the log likelihood of a model with only an intercept.
  - Why does this make sense?
    - $|L_{ur}| \leq |L_o|$ with equality only if the covariates do not have explanatory power at all.
    - If $|L_{ur}|/|L_o| = 1$, then 1-1=0, similar to the original R-squared
    - The measure is one if the estimated probabilities are all unity if y=1 and zero if y=0. But this does not happen in logit and probit models.

- Since the partial or marginal effects depend on the value of ***x***, the same is true for the estimated effect of a continuous regressor on the response probabilities:

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

- So, for $\Delta x_j = 1$, the estimated change in the success probabilities is roughly $g(\hat{\beta}_0 + \mathbf{x}\hat{\beta})\hat{\beta}_j$.
  - This quantity depends on the values of ***x***.
- How shall we choose ***x*** in an application?
  - It depends…
  - Choose interesting values for ***x***, e.g. values of one observation.
  - Define reference group in case of binary variables (=0).
  - Choose quartiles, median or mean of x_j, if it has many values.

- If you choose sample averages ….

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

you are estimating the partial effect for an average individual in the sample.

  - This option is the standard routine in computer packages.

- Note that due to the nonlinearity this is not the same as the sample average effect:

Scale factor

$$n^{-1}\sum_i [g(\hat{\beta}_0 + \mathbf{x_i}\hat{\beta})\hat{\beta}_j] \quad = \quad [n^{-1}\sum_i [g(\hat{\beta}_0 + \mathbf{x_i}\hat{\beta})]]\hat{\beta}_j$$

  - This quantity can be quite easily computed for the logit and probit model and does not rely on the specific choice of ***x*** values.
  - The scale factor is the same for all (continuous) x_j. If you have once the scale factor you can easily compare LPM, Logit and Probit estimates.

- The above approximations are not accurate in case of a discrete or in particular for a binary regressor.
  - How to estimate the change in the response probability in this case?
- Suppose x_k changes from *c* to *c+1*. Then the discrete analogue of the partial effect (at the sample average) is:

$$G(\hat{\beta}_0+\hat{\beta}_1\bar{x}_1+...+\hat{\beta}_{k-1}\bar{x}_{k-1}+\hat{\beta}_k(c+1))-G(\hat{\beta}_0+\hat{\beta}_1\bar{x}_1+...+\hat{\beta}_{k-1}\bar{x}_{k-1}+\hat{\beta}_k c)$$

  - Equivalently you can compute it for other values of ***x***. Moreover, you can also compute the average partial effect in an equivalent way.

- <u>Example</u>: (cont.)
  Married Women's labour force participation
  - Data: MROZ.dta
  - In addition to the LPM we now estimate the model by logit and probit.
  - In addition to the estimated coefficients we report
    - The percent correctly predicted
    - The log likelihood value
    - The Pseudo R-squared as described (R^2 for OLS)
  - LPM standard errors are heteroskedasticity robust

  - Success is that a married woman is in the labour force at a given year (*y=1*). We call this variable *inlf*.
  - We have a set of regressors:

    nwifeinc: husband's earnings

    educ: years of education

    exper: labour market experience

    kidslt6: number of kids less than six years

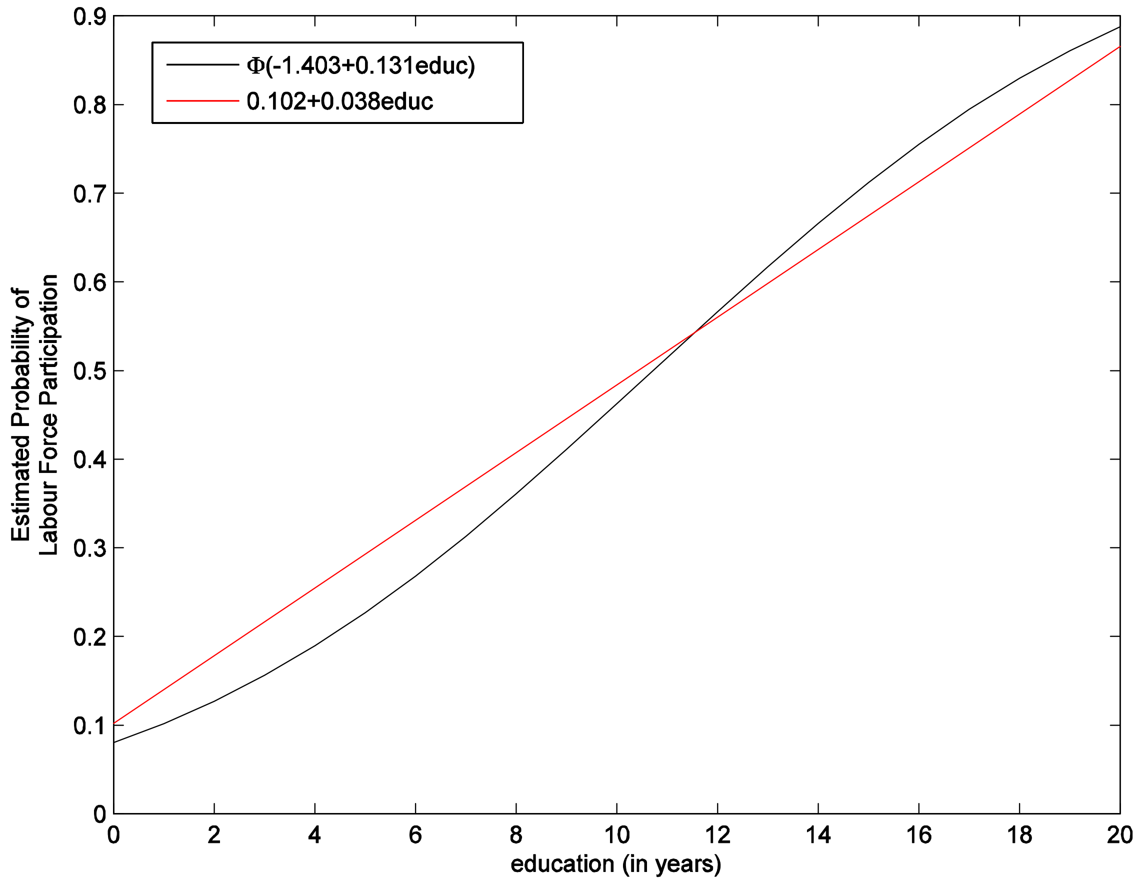
    kidsge6: number of children aged 6-18

Dependent Variable: inlf			
Independent Variables	*(1) LPM (OLS)*	*(2) Logit (MLE)*	*(3) Probit (MLE)*
nwifeinc	*-0.0034 (0.0015)*	*-0.021 (0.008)*	*-0.012 (0.005)*
educ	*0.038 (0.007)*	*0.221 (0.043)*	*0.131 (0.025)*
exper	*0.039 (0.006)*	*0.206 (0.032)*	*0.123 (0.019)*
exper^2	*-0.00060 (0.00018)*	*-0.0032 (0.0010)*	*-0.0019 (0.0006)*
kidslt6	*-0.262 (0.032)*	*-1.443 (0.204)*	*-0.868 (0.119)*
…other controls		*…*	*…*
constant	*0.586 (0.151)*	*0.425 (0.860)*	*0.270 (0.509)*
P C P	*73.4*	*73.6*	*73.4*
Log- Likelihood	*-*	*-401.77*	*-401.30*
Pseudo R^2	*0.264*	*0.220*	*0.221*

- When we compute scale factors for the average partial effect, we obtain:
  - 0.301 in the case of probit
  - 0.179 in the case of logit
  - We can use these scale factors to make estimates comparable.
  - The scaled probit coefficient on *educ* is then 0.301*0.131=0.039.
  - The scaled logit coefficient on *educ* is then 0.179*0.221=0.040.
  - Both are similar and remarkable close to the LPM estimate (0.038).

- The LPM assumes constant marginal effects, while the logit and probit models imply changing magnitudes of the partial effects:
  - In the LPM, one more small child is estimated to reduce the probability of labour force participation by about *0.262* independent of the number of children and regardless of anything else in the model).
  - In contrast, the marginal effect from probit/logit suggest that the drop in the probability decreases with the number of children and moreover it depends on the value of the other regressors.

- The estimated response probabilities from nonlinear binary response models can differ from the LPM.



- At lower levels of education the LPM estimates higher labour force participation probabilities than the probit model.

Summary

- We have introduced several models with a binary variable as dependent variable.
- With the LPM we can easily predict probabilities (with some limitations).
- We have seen nonlinear models for binary responses (logit and probit)
- Interpretation of logit and probit estimates requires some care.
- Use partial or marginal effects or the average partial effect.
- Use scale factors to make LPM, probit and logit estimates comparable.
- The LPM gives often a reasonable estimate of the average partial effect, despite its limitations. OPTIONAL:

  Battey et al. (2019): On the linear in probability model for binary data