# Limited dependent variable. Models for binary and multinomial outcome variable. Panel data and limited dependent variable.

Jakub Mućk
SGH Warsaw School of Economics

# Introduction

- In previous meetings, we have dealt with models in which the range of dependent variable is unbounded.
- The common cases when the response (dependent) variable is restricted:
  - binary: $y \in \{0, 1\}$,
  - multinomial: $y \in \{0, 1, 2, \ldots, k\}$,
  - integer: $y \in \{0, 1, 2, \ldots\}$,
  - censored: $y \in \{y^* \text{ if } y > y^*\}$.

# Binary dependent variable

- For binary outcome data the dependent variable $y$ takes one of two values:

$$y = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } (1-p) \end{cases} . \tag{1}$$

- The binary choice variable $y$ is restricted and the binary outcome is Bernoulli distributed.
- The probability ($p$) is not observed (latent variable).
- Examples:
  - dummy variables indicating whether some loan application is accepted ($y = 1$) or not ($y = 0$),
  - dummy variables indicating whether individual decided to work ($y = 1$) or not ($y = 0$),
  - binary variable indicating whether individual takes the second or third job ($y = 1$) or not ($y = 0$),
  - dummy variable indicating whether the birthweight was low, i.e., below 2500 g, ($y = 1$) or not ($y = 0$).
- Models for binary dependent variable
  - linear probability model (LMP);
  - logistic regression (logit);
  - probit regression (ptobit).

# Linear Probability Model (LMP) I

SGH

- Linear Probability Model (LMP) is the OLS regression of $y$ on $X$ that ignores the discreteness of the dependent variable. Moreover, the LMP does not constrain predicted probabilities to be between zero and one.
- In general, it is assumed that the (conditional to a set of covariates) probability is as follows:

$$\begin{aligned} Prob(y=1|X) &= F(X,\beta), \quad (2) \\ Prob(y=0|X) &= 1-F(X,\beta). \quad (3) \end{aligned}$$

- If the function $F(X,\beta)$ is assumed to linear, i.e., $F(X,\beta)=X'\beta$, then

$$y = \underbrace{\mathbb{E}(y|X,\beta)}_{Prob(y=1|X)} + \underbrace{(1-\mathbb{E}(y|X,\beta))}_{Prob(y=0|X)} = F(X,\beta) = X'\beta + \varepsilon. \quad (4)$$

- Finally, the LMP can be estimated by OLS:

$$y = X'\beta + \varepsilon. \quad (5)$$

where $\varepsilon$ is the error term.
- Shortcomings of the LMP:
  1. The predicted values of the dependent variable are not constrained to be between zero and one.

2. It is assumed that the probability is linearly related to some continuous explanatory variable.
3. The problem of the error heteroskedasticity. By construction, errors vary with the explanatory variables:

$$\begin{aligned} Var(\varepsilon|x) &= Prob(y=1|X)\left(1-X'\beta\right)^2 + Prob(y=0|X)\left(-X'\beta\right)^2 \\ &= X'\beta\left(1-X'\beta\right)^2 + \left(1-X'\beta\right)\left(-X'\beta\right)^2 \\ &= X'\beta\left(1-X'\beta\right). \end{aligned}$$

   As a consequence, the estimated variance-covariance matrix are biased (also standard errors, $t$ statistics, $\mathcal{F}$ statistic, etc.). To challenge this issue one might apply:
   - robust standard errors;
   - feasible GLS estimation that accounts for heteroskedastic residuals.
4. By construction, error term is also not normally distributed.
   - The statistical inference in small samples is not reliable.

- In the logit model, the conditional (to $X$) probability is described by the cumulative logistic distribution (conditional to some explanatory variables $X$):

$$p = Prob(y = 1|X) = \frac{\exp(X'\beta)}{1 + \exp(X'\beta)}. \tag{6}$$

- The predicted probabilities are always between zero and one.
- It can be shown that the logit (log of odds):

$$\ln\left(\frac{p}{1-p}\right) = X'\beta. \tag{7}$$

- The parameters of $\beta$ are estimated using the maximum likelihood (ML) method. In general, the log likelihood for the logit model can be written as:

$$\ln \mathcal{L} = \sum_{i=1}^{N} \left[y_i \ln(F(X,\beta)) + (1 - y_i)\ln(1 - F(X,\beta))\right], \tag{8}$$

where $\mathcal{L}$ is the likelihood function, the index $i$ stands for observation and $F(X,\beta) = \exp(X'\beta)/(1 + \exp(X'\beta))$.

- The Log Likelihood, given by (8), is maximized using some optimization methods.
- The Likelihood test for the significance of parameters. The null hypothesis:

$$\mathcal{H}_0: \quad \beta_1 = \beta_2 = \ldots = \beta_k = 0, \tag{9}$$

and the test statistics ($LR$) bases on the log-likelihood difference between the considered model ($\mathcal{L}$) and the model with only intercept ($\mathcal{L}_0$)):

$$LR = 2(\mathcal{L} - \mathcal{L}_0), \tag{10}$$

where $LR$ is $\chi^2$ distributed with $k$ (number of explanatory variables) degrees of freedom.

- The logit model is nonlinear. The sign of the estimates informs only about the direction of the relationship between explanatory variable and probability.
- To interpret the logit estimates it is useful to introduce the odds ratio. The odds is an exponential function of fitted $F(X, \beta)$. For instance, the odds ratio for $x_1$ variable can be described as:

$$OR = \frac{\exp\left(\beta_0 + \beta_1(x_1 + 1) + \ldots + \beta_k x_k\right)}{\exp\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k\right)} = \exp(\beta_1), \tag{11}$$

so an 1-unit increase in $x_1$ multiplies the odds ratio by $\exp(\beta_1)$.

- In nonlinear models, more common approach is to use marginal effects. In the logit model, the marginal effect for the $k$-th explanatory variable can be written:

$$MFX(x_k) = \frac{\partial p}{\partial x_k} = \beta_k p(1-p) = \frac{\exp\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k\right)}{\left[\exp\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k\right)\right]^2}\beta_k. \quad (12)$$

- Some remarks about the marginal effects:
  - The marginal effects vary for different values of explanatory variables.
    - The usual approach is to calculate the marginal effects for the average explanatory variables, i.e., $\bar{x}_1, \ldots, \bar{x}_k$.
    - However, the marginal effects for the mean explanatory variables are not reasonable when we are interested in the effect of some dummy variable on the probability. In such cases, one should calculated the marginal effect when this indicator variable is set to 0.
  - Apart from the point estimates, it is essentially to analyze confidence intervals or standard errors for the estimated marginal effects are useful.

- In the probit model, the conditional (to $X$) probability is described by the cumulative standard normal distribution (conditional to some explanatory variables $X$):
$$p = Prob(y = 1|X) = \Phi(X'\beta), \tag{13}$$
where $\Phi(X'\beta)$ is the cumulative distribution function for the standard normal.
- Alternatively,
$$p = Prob(y = 1|X) = \int_{-\infty}^{X'\beta} (2\pi)^{-\frac{1}{2}} \exp\left(-z^2/2\right) dz, \tag{14}$$
- The marginal effects for the $k$-th explanatory variable:
$$MFX(x_k) = \frac{\partial p}{\partial x_k} = \phi\left(\beta_0 + \beta_1 x_1 + \dots \beta_k x_k\right)\beta_k, \tag{15}$$
where $\phi(\cdots)$ denotes the density for the standard normal distribution.

# Multinomial Logit

- **Key assumption:** several ($K$) possible outcomes/exclusive alternatives:

$$p_{i1} + p_{2i} + \ldots + p_{Ki} = 1, \tag{16}$$

where $p_{ij}$ is the probability of the $j$-th alternative for $i$-th observation/unit.

- Assuming the logistic distribution:

$$p_{ij} = Prob(y = j|X) = \frac{\exp(X'\beta_j)}{\sum_{k=1}^{K} \exp(X'\beta_j)}, \tag{17}$$

where $\beta_j$ are alternative specific parameters while $X$ is a set of explanatory variables that are not varying between alternatives.

- Since the probabilities of exclusive alternatives sum to one the normalization of parameters is required. Taking the first possible outcome ($j = 1$) it is restricted that $\beta_0$ is set at zero. This implies that

$$p_{ij} = \begin{cases} \frac{1}{\sum_{k=2}^{K} \exp(X'\beta_j)} & \text{if} \quad j = 1, \\ \frac{\exp(X'\beta_j)}{\sum_{k=2}^{K} \exp(X'\beta_j)} & \text{if} \quad j. \neq 1 \end{cases} \tag{18}$$

- The parameters $\beta_2, \ldots, \beta_K$ can be estimated with the standard MLE.

- **Interpretation is relative**. Note that:

$$\frac{Prob(y=j|X)}{Prob(y=1|X)} = \exp\left(X'\beta_j\right), \tag{19}$$

  is the probability ratio.
- the exponentiated value of a single estimate is the the relative-risk ratio for a unit change in a given explanatory variables.
- The cross-equation restriction can be imposed and test with the LR tests.
- Further issues:
  - **Conditional logit**.
  - **Mixed multinomial models**.
  - **Ordered mulinomial models**.

# Count data

- Count data is a special data when observations take only non-negative integer.
- The dependent variable is a count of the number of occurrences of an event, i.e., $y \in \{0, 1, 2, \ldots\}$.
- In many empirical applications, the sample of such dependent variable is concentrated on a few small discrete values, i.e., 0, 1, 2.
- Examples:
  - The number of children in a household.
  - The number of alcoholic drinks a college student takes in a week.
  - The number of patents.
  - The number of new products introduced in market.
  - The number of doctor visits.
- Models:
  - Poisson regression model.
  - Negative binomial model.

# The Poisson regression model I

- The natural stochastic environment for counted variable is a Poisson point process for occurrence of the event of interest. The probability function for a Poisson distribution:

$$Prob(Y = y) = \frac{\exp(-\mu)\mu^y}{y!}, \qquad y = 0, 1, 2, \ldots \tag{20}$$

where $\mu$ denotes the intensity parameter.

- It can be shown that the Poisson distribution has the equidispersion property:

$$\begin{aligned} \mathbb{E}(y) &= \mu, \\ Var(y) &= \mu. \end{aligned}$$

- In the Poisson regression, the intensity parameter captures the relationship between the dependent variable and explanatory variables. Usually, the exponential mean parameterization is assumed:

$$\mu = \exp(x'\beta) \tag{21}$$

- Estimation: the pseudo maximum likelihood (PML) or quasi maximum likelihood (QML) estimation.
- **Interpretation:** marginal effects:

$$MFX(x_j) = \frac{\partial \mathbb{E}(y|X)}{\partial x_j} = \beta_j \exp(X'\beta). \tag{22}$$

- The Poisson regression bases on the very restrictive assumption that $\mathbb{E}(y) = Var(y)$.
  - Very often, the variance is far from the mean.

- In many empirical application, a Poisson density underpredicts the zero count.

- There are several test statistics designed to verify the hypothesis that mean equals variance. General idea bases on the following relationship:

$$Var(y) = \mu + \alpha g(\mu), \tag{23}$$

where $g(\cdot)$ is some known function ($g(\mu) = \mu$ or $g(\mu) = \mu^2$).

- Having fitted values ($\hat{\mu} = \exp(X'\beta)$)from the Poisson model the following OLS regression can be run to test null ($\alpha = 0$):

$$\frac{(y - \hat{\mu})^2 - y}{\hat{\mu}} = \alpha \frac{g(\hat{\mu})}{\hat{\mu}} + u \tag{24}$$

where $u$ is the error term.

# The negative binomial regression I

- The negative binomial regression is an extension of the Poisson regression that accounts for overdispersion, i.e., extra variation that is not included in the standard Poisson process.
- In the negative binomial regression, the following moments can be assumed

$$\mathbb{E}(y) = \mu,$$
$$Var(y) = \mu(1 + \alpha\mu).$$

Where $\alpha > 0$. Note that if $\alpha = 0$ then it is standard Poisson regression. $\alpha$ is the overdispersion parameter.

- Estimation: PML or QML.
- **Interpretation:** marginal effects:

# Censored data and tobit regression

- Usual causes of incompletely observed data are truncation and censoring.
  - truncated data $\Longrightarrow$ some observations on both dependent and explanatory variables are missing;
  - censored data $\Longrightarrow$ some observations on dependent variable are missing but information on explanatory variables are complete.
- Censoring can be perceived as a feature of data-gathering process. For instance, for confidentiality reasons the income of high-income workers may be top-coded (higher than 200k USD).
- Examples:
  - Ticket sales to some event. We want to explain (latent, unobservable) demand for tickets to some sport events. Sometimes all tickets are sold out and (unobservable) demand can be higher than the total number of available tickets but we observe only the number of tickets that were sold out.
- Model:
  - Tobit model.

# Censoring Mechanism

- When data are censored we always observe the explanatory variables.
- Our dependent variable is the latent variable $y^*$ for which we have incomplete observations ($y$).
- $y^*$ may be censored from below/left. Then we observe:

$$y = \begin{cases} y^* & \text{if} \quad y^* > L \\ L & \text{if} \quad y^* \leq L \end{cases} .$$

- $y^*$ may be censored from above/right. Then we observe:

$$y = \begin{cases} y^* & \text{if} \quad y^* < U \\ U & \text{if} \quad y^* \geq U \end{cases} .$$

- It it possible to consider more sophisticated censoring mechanisms.
- The OLS estimates in such cases are not consistent.

- In the censored regression, information on censoring is included.
- Consider the following example:

$$y^* = X'\beta + \varepsilon, \tag{25}$$
$$y = 0 \quad \text{if } y^* \leq 0, \tag{26}$$
$$y = y^* \quad \text{if } y^* > 0. \tag{27}$$

Then the conditional expected value of $y$:

$$\mathbb{E}(y|x) = \Phi\left(\frac{X'\beta}{\sigma}\right)\left(X'\beta + \sigma\lambda\right), \tag{28}$$

where $\Phi(\cdot)$ is the probability density function of normal distribution and $\lambda$ stands for the Mills ratio, ($\lambda = \phi(X'\beta/\sigma)/\Phi(X'\beta/\sigma)$).

- Estimation: MLE.
- Interpretation: marginal effects:
  - For a latent variable ($y^*$), marginal effects are constant:

$$MFX(x_j) = \frac{\partial\mathbb{E}(y^*|x)}{\partial x_j} = \beta_j. \tag{29}$$

  but $y^*$ in unobserved.
  - For the observed variable $y$, the marginal effects become more sophisticated:

$$MFX(x_j) = \frac{\partial\mathbb{E}(y|x)}{\partial x_j} = \beta_j\Phi\left(\frac{X'\beta}{\sigma}\right). \tag{30}$$

# The Binary Outcomes Models & Panel Data

- We can use logit or probit binary response function:

$$Pr(y_{it} = 1|x_{it}) = G\left(x_{it}'\beta\right), \tag{31}$$

where $G(\cdot)$ is a known function taking on values in the open unit interval.

- Note that in the (31) we assume that our model is dynamically complete. In other words, we don't assume that the scores (latent variable $\implies$ probabilities) is serially correlated or contains the individual-specific component. For instance,

$$Pr(y_{it} = 1|x_{it}) = Pr(y_{it} = 1|x_{it}, x_{it-1}, x_{it-2}) \tag{32}$$

- Some useful procedures to test dynamic completeness:
  - Add the lagged dependent variable/ independent variables to considered model and test their significance.
  - Make a pooled probit/logit regression. Based on the pooled estimates make prediction and include lagged fitted values (scores) in basic model. Then, test its significance.
- To account for an unobserved heterogeneity it is useful to apply robust standard errors.

- The fixed effect logit model:

$$Pr(y_{it} = 1) = Pr(y^*_{it} > 0) = F(x'_{it}\beta), \tag{33}$$

where $F$ is the logistic cumulative distribution function and

$$y^*_{it} = x'_{it}\beta + \mu_i + \varepsilon_{it}, \tag{34}$$

where $\mu_i$ is the individual-specific intercept and $\varepsilon_{it}$ denotes the idiosyncratic error.

- Natural way to estimate the parameters of the FE logit model is to include dummy variables and perform ML estimation but ...
- Incidental parameters problem. As $N \to \infty$ for the fixed $T$, the number of parameters capturing fixed effecst increases with T. As a result, $\mu_i$ cannot be consistently estimated for a fixed $T$.
- The above problem can be overcome by using conditional likelihood function. It is assumed that the fixed effects and explantory variables are not correlated with error term.

- Conditional likelihood function:

$$\mathcal{L}_C = \prod_{i=1}^{N} Pr\left(y_{i1}, \ldots, y_{iT} / \sum_{t=1}^{T} y_{it}\right). \tag{35}$$

- Let's illustrate for $T = 2$.
  - For $T = 2$, the sum $\sum_{t=1}^{T} y_{it}$ can be 0, 1 or 2.
  - But if the sum $\sum_{t=1}^{T} y_{it}$ is 0 (or 2) then both $y_{i1}$ and $y_{i2}$ are 0 (or 1). These cases are irrelevant for the $\ln \mathcal{L}_C$ because $\ln(1) = 0$.
  - Two remaining cases (when sum equals 1). Let's start with the sum that equals unity:

$$Pr(y_{i1} + y_{i2} = 1) = Pr(y_{i1} = 0, y_{i2} = 1) + Pr(y_{i1} = 1, y_{i2} = 0) \tag{36}$$

  - General probability:

$$\begin{aligned} Pr(y_{i1} = 1) &= \exp(\mu_i + x_{it}'\beta) / \left[1 + \exp(\mu_i + x_{it}'\beta)\right] \\ Pr(y_{i1} = 0) &= 1 - \exp(\mu_i + x_{it}'\beta) / \left[1 + \exp(\mu_i + x_{it}'\beta)\right] \\ &= 1 / \left[1 + \exp(\mu_i + x_{it}'\beta)\right]. \end{aligned}$$

- Conditional probability in the second period:

$$Pr(y_{i1}=1, y_{i2}=0) = \frac{\exp(\mu_i + x_{i1}'\beta)}{1+\exp(\mu_i + x_{i1}'\beta)} \cdot \frac{1}{1+\exp(\mu_i + x_{i2}'\beta)}$$

$$Pr(y_{i1}=0, y_{i2}=1) = \frac{1}{1+\exp(\mu_i + x_{i1}'\beta)} \cdot \frac{\exp(\mu_i + x_{i2}'\beta)}{1+\exp(\mu_i + x_{i2}'\beta)}$$

As a result:

$$\begin{aligned} Pr(y_{i1}+y_{i2}=1) &= Pr(y_{i1}=0, y_{i2}=1) + Pr(y_{i1}=1, y_{i2}=0) \\ &= \frac{\exp(\mu_i + x_{i1}'\beta) + \exp(\mu_i + x_{i2}'\beta)}{\left(1+\exp(\mu_i + x_{i1}'\beta)\right)\left(1+\exp(\mu_i + x_{i2}'\beta)\right)}, \end{aligned}$$

and

$$Pr(y_{i1}=1, y_{i2}=0 | y_{i1}+y_{i2}=1) = \frac{Pr(y_{i1}=1, y_{i2}=0)}{Pr(y_{i1}+y_{i2}=1)} =$$

$$\frac{\exp(\mu_i + x_{i1}'\beta)}{\exp(\mu_i + x_{i1}'\beta) + \exp(\mu_i + x_{i2}'\beta)} = \frac{\exp(x_{i1}'\beta)}{\exp(x_{i1}'\beta) + \exp(x_{i2}'\beta)}$$

Finally,

$$Pr(y_{i1}=1, y_{i2}=0 | y_{i1}+y_{i2}=1) = \frac{1}{1+\exp(x_{i2}-x_{i1})'\beta} \tag{37}$$

and analogously for the remaining case:

$$Pr(y_{i1} = 1, y_{i2} = 0 | y_{i1} + y_{i2} = 1) = \frac{\exp(x_{i2} - x_{i1})'\beta}{1 + \exp(x_{i2} - x_{i1})'\beta}. \tag{38}$$

Using the conditional likelihood we eliminate the fixed effect (individual-specific intercept). Apart from that, all time invariant explanatory variables are also wiped out from the estimation.

- **The random effects binary outcomes models** assume that the individual effects are normally distributed, i.e. $\mu_i \sim \mathcal{N}(0, \sigma_\mu^2)$.
- This yields:

$$Pr(y_{it} = 1 | x_{it}, \beta, \mu_i) = \begin{cases} \Lambda(\mu_i + x_{it}'\beta) & \text{for logit model,} \\ \Phi(\mu_i + x_{it}'\beta) & \text{for probit model,} \end{cases} \tag{39}$$

  where $\Lambda(\cdot)$ and $\Phi(\cdot)$ is the logistic and standard normal cumulative distribution, respectively.
- The underlying parameters ($\beta$) and the variance of random unit-specific effects ($\sigma_\mu^2$) can be estimated with the Maximum Likelihood estimation (MLE). The MLE of $\beta$ and $\sigma_\mu^2$ maximizes the log-likelihood, i.e., $\sum_{i=1}^{N} \ln f(y_i | X_i, \beta, \sigma_\mu^2)$, where

$$f(y_i | X_i, \beta, \sigma_\mu^2) = \int f(y_i | X_i, \beta) \frac{1}{\sqrt{2\pi\sigma_\mu^2}} \exp\left(\frac{-\mu_i}{2\sigma_\mu^2}\right) d\mu_i, \tag{40}$$

  where $f(y_i | X_i, \beta)$ is the considered probability distribution function (logistic or standard normal).
- The MLE estimates are calculated numerically using quadrature method.

- We can test the presence of cross-sectional heterogeneity. The standard likelihood test are designed to verify the following null hypothesis:

$$\mathcal{H}_0 : \quad \sigma_\mu^2 = 0. \tag{41}$$

- Unlike the FE logit the random effects logit/probit models use variables that are constant over time.
- In analogous fashion to the linear FE models, it is assumed that individual effects are independent of the explanatory variables.