University of Ljubljana

SEB SCHOOL OF ECONOMICS AND BUSINESS

9. Discrete Choice Models

Prof. Dr. Miroslav Verbič

miroslav.verbic@ef.uni-lj.si

www.miroslav-verbic.si

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION OF MBAS ACCREDITED

Ljubljana, October 2025

Motivation

So far, we have focused on modelling continuous dependent variables, but in real life the number of alternatives is often small. This requires a particular modelling approach.

Discrete choice models → the variable to be explained, *y*, is taking a small finite number of outcomes; i.e. we have a discrete dependent variable.

Discrete choice models

The **discrete dependent variable** can be:

- Binary or binomial: dichotomous choice [0, 1].
  Examples: work/not work, buy/not buy etc.
  Approach: **probit model & logit model**.
- Multinomial: multiple choice, e.g. [1, 2, 3, 4], which can be ordered or non-ordered.
  Examples:
  - non-ordered: mode of transport etc;
  - ordered: survey scale etc.

  Approach:
  - non-ordered: **multinomial probit & logit model**;
  - ordered: **ordered probit & logit model**.

Discrete choice models

- Non-negative integer: count data [0, 1, 2, …].
  Examples: number of patents, number of loss events etc.
  Approach: **Poisson model**.

We will only deal here with the **binary** dependent variable.

We are interested in the conditional or response **probability**:

$P(y = 1|X) = P(y = 1|x_1, \dots, x_k)$ for various values of x_j.

Discrete choice models

Why not just use a linear regression model, called the **linear probability model** (LPM), for the binary response variable *y*:

$$P(y_i = 1|x_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} \ ,$$

estimate it by the ordinary least squares estimator (OLS) and obtain the partial effects (regression coefficient estimates):

$$\beta_k = \frac{\partial P(y_i=1|x_i)}{\partial x_k} \ ?$$

Discrete choice models

Bernoulli–type random variables *y* and (thus) *u*:

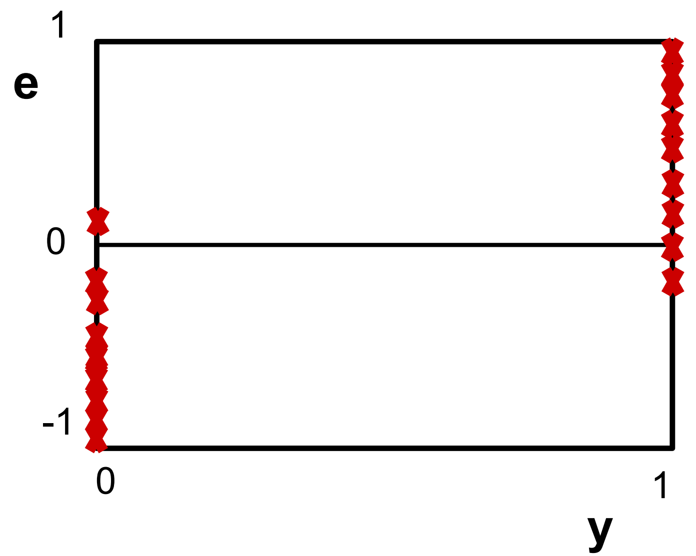
- $P(y = 1|x) = p(x)$
- $P(y = 0|x) = 1 - p(x)$
  - $\mathrm{E}(y|x) = p(x)$
  - $\mathrm{Var}(y|x) = p(x)(1 - p(x))$
- $y \sim \mathrm{Bernoulli}\,(p(x))$

Discrete choice models

Therefore, we have several **reasons**:

1) Non-normality of the random variable *u* (and *y*);
2) Heteroscedasticity of the variance of the disturbance term (variance depends on x_j);
3) Predicted probabilities, i.e. fitted values $\hat{y}$, could lie outside the unit interval $0 \leq E\left(y_i \middle| x_{ji}\right) \leq 1$, which could further imply negative variance (for $\hat{y} < 0$; see previous slide);
4) Questionable power of R^2, since all residuals are concentrated at only two values of *y* (see next slide);
5) Questionable linear relationship between *y* and *x* for such a model; are constant (marginal) effects of *x* on *y* realistic?

Discrete choice models



We usually use a **probit** or a **logit model** instead, which is estimated by the **maximum likelihood estimator** (MLE).

University of Ljubljana

SEB SCHOOL OF ECONOMICS AND BUSINESS

9.1 Maximum Likelihood Estimation

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION OF MBAS ACCREDITED

Basic concepts

- ✓ Maximum likelihood estimation (MLE) is a statistical method (an estimator) to find the most likely density function that would have generated the data.
- ✓ Thus, MLE requires you to make a distributional assumption first.
- ✓ We will provide the intuition behind the MLE using some examples.

Basic concepts

Id	x
1	**1**
2	**4**
3	**5**
4	**6**
5	**9**

- ✓ Let us explain the basic idea of MLE using the data on the left.
- ✓ Let us make an assumption that the variable x follows normal distribution.
- ✓ Remember that the density function of normal distribution with mean μ and variance σ^2 is given by:

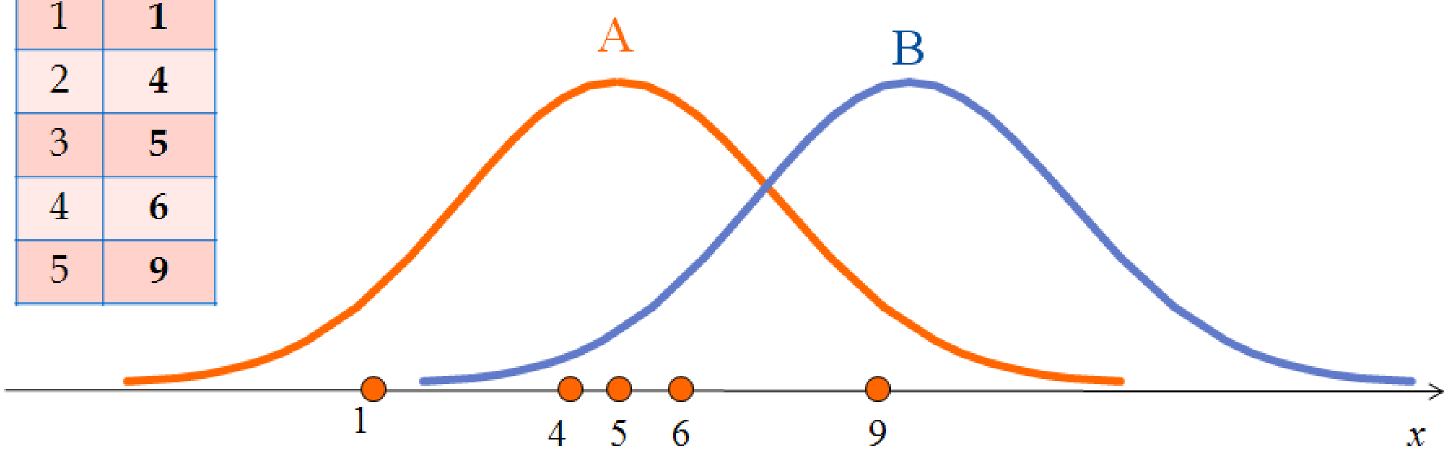
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \quad \text{for} \quad -\infty < x < \infty$$

Basic concepts

- ✓The data are plotted on the horizontal line.
- ✓Now, ask yourself the following question:

"Which distribution, A or B, is more likely to have generated the data?"

Id	x
1	1
2	4
3	5
4	6
5	9



Basic concepts

- ✓ Answer to the question is A, because the data are clustered around the center of the distribution A, but not around the center of the distribution B.
- ✓ This example illustrates that, by looking at the data, it is possible to find the distribution that is most likely to have generated the data.
- ✓ Now, how exactly do we find the distribution in practice?

Estimation procedure in general

- ✓ MLE starts with computing the likelihood contribution of each observation.
- ✓ The likelihood contribution is the height of the density function. We use L_i to denote the likelihood contribution of i^{th} observation.

Estimation procedure in general

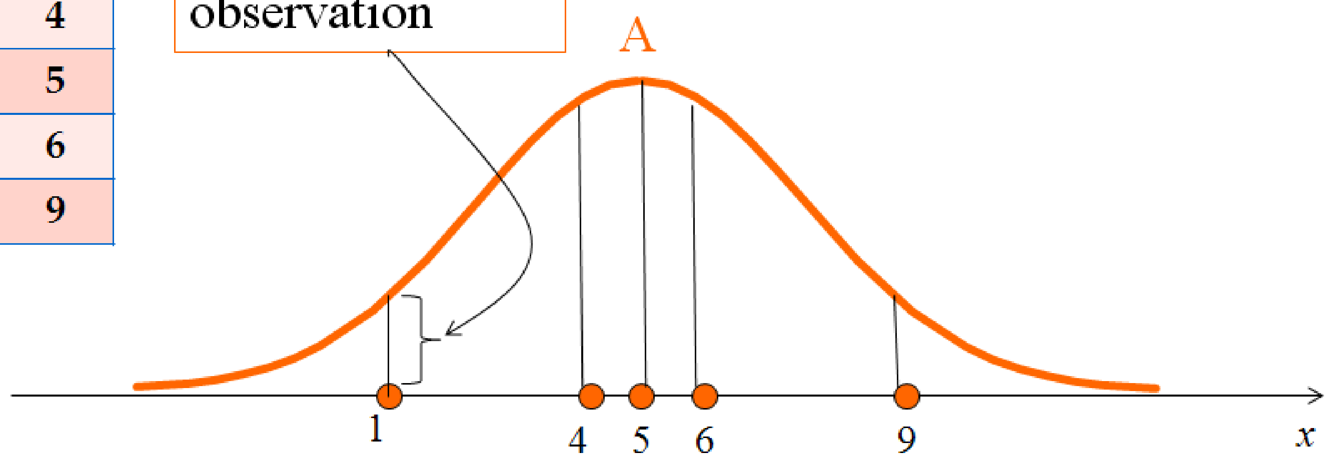
Graphical illustration of the likelihood contribution:

Id	x
1	1
2	4
3	5
4	6
5	9

The likelihood contribution of the first observation

$$= L_1 = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(1-\mu)^2/2\sigma^2}$$

Data value



Estimation procedure in general

- ✓Then, we multiply the likelihood contributions of all the observations. This is called the likelihood function, L:

$$L = \prod_{i=1}^{n} L_i$$

This notation means you multiply from $i = 1$ through n.

- ✓In our example, $n = 5$.

Estimation procedure in general

- In our example, the likelihood function looks like:

Id	x
1	1
2	4
3	5
4	6
5	9

$$L(\mu,\sigma)=\prod_{i=1}^{5}L_i=L_1\times L_2\times L_3\times L_4\times L_5$$

$$=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(1-\mu)^2/2\sigma^2}\times\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(4-\mu)^2/2\sigma^2}$$

$$\times\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(5-\mu)^2/2\sigma^2}\times\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(6-\mu)^2/2\sigma^2}$$

$$\times\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(9-\mu)^2/2\sigma^2}$$

- We write $L(\mu, \sigma)$ to emphasize that the likelihood function depends on these parameters.

Estimation procedure in general

- ✓ Then we find the values of μ and σ that maximize the likelihood function.
- ✓ The values of μ and σ which are obtained this way are called the maximum likelihood estimates (MLEs) of μ and σ.
- ✓ Most of the MLEs cannot be solved 'by hand'. Thus, we need to apply an iterative procedure to solve it on computer.
- ✓ Fortunately, the majority of models that require MLE can be estimated automatically in Stata and R.

Estimation of a regression function

- ✓We are usually interested in estimating a linear regression function.
- ✓We will use a simple bivariate regression model for illustration:

$$y = \beta_0 + \beta_1 x + u.$$

- ✓Estimation of such a model can be done using the MLE.

Estimation of a regression function

Id	y	x
1	2	1
2	6	4
3	7	5
4	9	6
5	15	9

- ✓ Suppose that we have these data, and we are interested in estimating the above model.
- ✓ Let us make an assumption that u follows the normal distribution with mean 0 and variance σ^2.

Estimation of a regression function

- We can rewrite the model as:

$$u = y - (\beta_0 + \beta_1 x).$$

- This means that $y - (\beta_0 + \beta_1 x)$ follows the normal distribution with with mean 0 and variance σ^2.
- The likelihood contribution of each observation is the height of the density function at the data point $y - (\beta_0 + \beta_1 x)$.

Estimation of a regression function

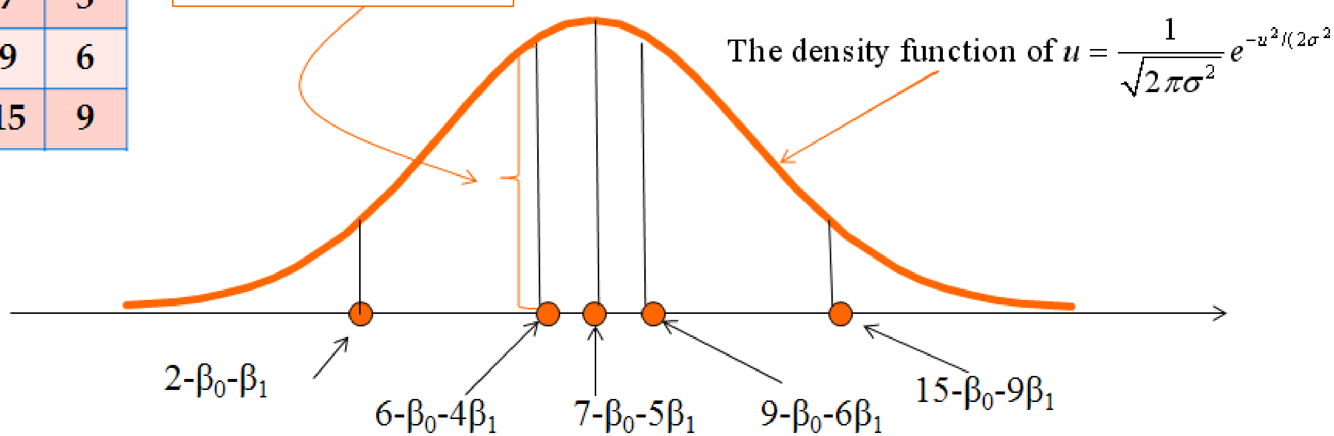
For example, the likelihood contribution of the 2nd observation is given by:

Id	y	x
1	2	1
2	6	4
3	7	5
4	9	6
5	15	9

The likelihood contribution of the 2nd observation

$$=L_2 = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\underbrace{(6-\beta_0-4\beta_1)}_{u_2-\mu=(y_2-\beta_0-\beta_1 x_2)-0}^2/2\sigma^2}$$

Data point



Estimation of a regression function

Then the likelihood function is given by:

Id	y	x
1	2	1
2	6	4
3	7	5
4	9	6
5	15	9

$$L(\beta_0, \beta_1, \sigma) = \prod_{i=1}^{n} L_i = L_1 \times L_2 \times L_3 \times L_4 \times L_5$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(2-\beta_0-\beta_1)^2/2\sigma^2} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(6-\beta_0-4\beta_1)^2/2\sigma^2}$$

$$\times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(7-\beta_0-5\beta_1)^2/2\sigma^2} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(9-\beta_0-6\beta_1)^2/2\sigma^2}$$

$$\times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(15-\beta_0-9\beta_1)^2/2\sigma^2}$$

Estimation of a regression function

- ✓ The likelihood function is thus a function of β_0, β_1, and σ.
- ✓ We choose the values of β_0, β_1, and σ that maximize the likelihood function. These are the maximum likelihood estimates of of β_0, β_1, and σ.
- ✓ Again, maximization can easily be done automatically in Stata and R.

Summary of the MLE procedure

1. Compute the likelihood contribution of each observation, L_i, for i = 1,…,n.
2. Multiply all the likelihood contributions to form the likelihood function, L:

$$L = \prod_{i=1}^{n} L_i$$

3. Maximize L by choosing the values of the parameters. The values of parameters that maximize L are the maximum likelihood estimates of the parameters.

Properties of the ML estimates

1. Consistency
2. Asymptotic normality
3. Asymptotic efficiency
4. Invariance

Invariance means that for any one-to-one transformation of the model parameters, the ML estimates (the maximization solution) remain unchanged.

The log-likelihood function

- It is usually easier to maximize the natural log of the likelihood function than the likelihood function itself:

$$\ln(L) = \ln\left[\prod_{i=1}^{n} L_i\right] = \ln(L_1 \cdot L_2 \cdot \ldots \cdot L_n) =$$

$$= \ln(L_1) + \ln(L_2) + \ldots + \ln(L_n) = \sum_{i=1}^{n} \ln(L_i)$$

- Due to invariance, maximizing the so called log-likelihood function is identical to maximizing the likelihood function.

University of Ljubljana

SEB SCHOOL OF ECONOMICS AND BUSINESS

9.2 Latent Variable Approach

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION OF MBAS ACCREDITED

Example from real life

The university would like to evaluate your **knowledge** at the end of each course.

Unfortunately, the knowledge is not (directly) observed, only your **exams** can be evaluated, which is not always the same thing as obtained knowledge.

At the doctoral/PhD level, often the only two grades awarded are pass and fail (no grades from 1 to 10).

Example from economics

Labour force participation, LFP:

= 1, if an individual participates in the labour market (works) or

= 0, if he/she does not participate in the labour market.

Rational individual maximizes his **direct utility function**, subject to his **budget constraint**:

$$\max\ u(c, j),\quad \text{s.t.}\quad y_N + w(H - j) = c$$

where c stands for consumption of goods, j for consumption of leisure time, y_N for non-labour income, w for wage rate, and H for total available time.

Example from economics

We derive:

- the indirect utility of inactivity: $v(H, y_N)$ and
- the indirect utility of activity: $v(w, H, y_N)$.

Note that apart from w, H *and* y_N, everything else is endogenous.

The following holds for a **rational individual**:

LFP = 1 if and only if $v(w, H, y_N) \geq v(H, y_N)$, and 0 otherwise.

Latent variable approach

We often do not observe the underlying **choice variables** (e.g. indirect utility), but we do observe the **choice itself** (e.g. LFP).

We assume (by rationality) that the option with more favourable choice variable value was chosen.

Latent variable approach

In econometrics, we model this by the so called **latent-variable approach**:

$$y^* = X\beta + u,$$

where *y** is the latent (unobserved) variable (e.g. *v*) and the following observed outcomes (on *y*, e.g. on *LFP*) with assumed relationships (about *y**, e.g. about *v*) hold:

$$y_i = 1, \quad if\ y_i^* \geq 0;$$
$$y_i = 0, \quad if\ y_i^* < 0.$$

Latent variable approach

We model the probability of a choice:

$$P(y=0) = P(y^* < 0) = P(X\beta + u < 0) = P(u < -X\beta) = \\ = \Psi(-X\beta) = 1 - \Psi(X\beta)$$

and

$$P(y=1) = P(y^* \geq 0) = P(X\beta + u \geq 0) = P(u \geq -X\beta) = \\ = 1 - \Psi(-X\beta) = \Psi(X\beta),$$

where $\Psi(\cdot)$ is the cumulative distribution function (CDF) and it holds that $\Psi(X\beta) + \Psi(-X\beta) = 1$ (symmetry).

Latent variable approach

For binary choice, the probability of an observation with outcome either $y_i = 0$ or $y_i = 1$ is:

$$P(y_i|X) = (\Psi(X\beta))^{y_i} \cdot (1 - \Psi(X\beta))^{1-y_i} = L_i(\beta),$$

which is called the **likelihood contribution** of observation *i*, L_i.

Usually, we utilize the **log-likelihood contribution** of observation *i*, denoted by $\ln L_i$:

$$lnL_i(\beta) = y_i ln\Psi(X\beta) + (1 - y_i)ln(1 - \Psi(X\beta)).$$

Latent variable approach

For **maximum likelihood estimation** of **β**, we need to assume a form for the cumulative distribution function, Ψ(Xβ).

For binary choice, we have two possibilities:

1. Standard normal distribution:

$$\Psi(X\beta) = \Phi(X\beta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{X\beta} e^{-\frac{1}{2}t^2} dt, \qquad t \sim N(0,1)$$

leads to the **probit model**;

Latent variable approach

2. Logistic distribution:

$$\Psi(X\beta) = \Lambda(X\beta) = \frac{e^{X\beta}}{1 + e^{X\beta}} = \frac{1}{\frac{1}{e^{X\beta}} + 1} = \frac{1}{1 + e^{-X\beta}}$$

leads to the **logit model**.

Conveniently (as we model probabilities), for both cumulative distribution functions it holds that:

$$0 \leq \Phi(X\beta) \leq 1;$$
$$0 \leq \Lambda(X\beta) \leq 1.$$

Back to the example from real life…

Latent variable: y^* – obtained knowledge

Observed variable – exam result, y:

$$y_i = \begin{cases} 1, if\ y_i^* \geq y_{min}\ (pass) \\ 0, if\ y_i^* < y_{min}\ \ (fail) \end{cases}$$

Implicit assumption: exams were fair in terms of:

a) no cheating and

b) fair grading.

Back to the example from real life…

What are the determinants of exams results?

$y_i = f(age_i, H_i, D_i, E_i, \ldots), \quad \forall$ student *i*

where:

- *y* = 1 if passes, 0 if fails;
- *age* – age of a student;
- *H* – hours of studying the course;
- *D* – finished previous degree abroad (1 if yes, 0 if no);
- *E* – years of work experience
- …

University of Ljubljana

SEB SCHOOL OF ECONOMICS AND BUSINESS

9.3 The Probit Model

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION OF MBAS AMBA ACCREDITED

Probit model and estimation

Let us assume that we have the following model:

$$y_i^* = \beta_0 + \beta_1 x_i + u_i$$

$$\begin{cases} \text{If } y_i = 0, \text{ then } y_i^* < 0 \\ \text{If } y_i = 1, \text{ then } y_i^* \geq 0 \end{cases}$$

Id	y	x
1	0	1
2	0	4
3	1	5
4	1	6
5	1	9

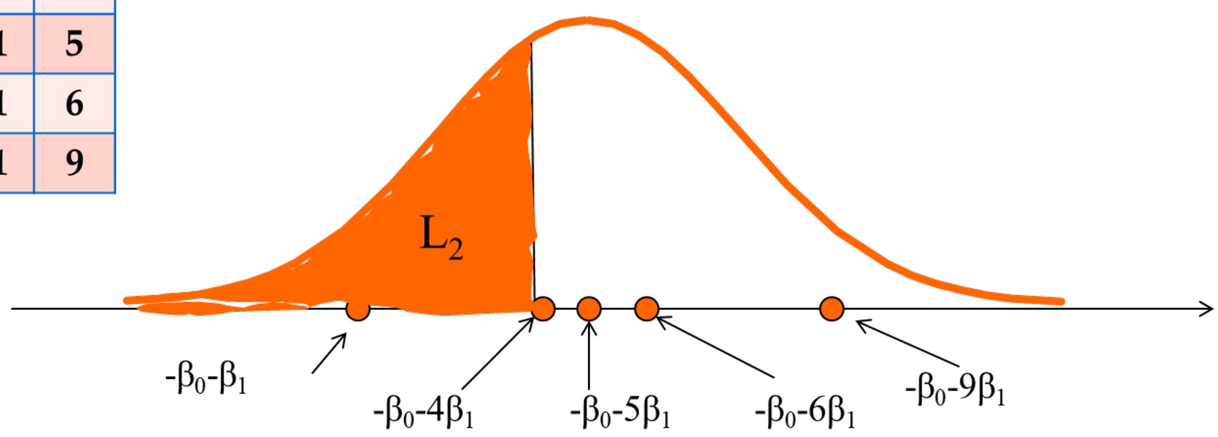
- ✓ Suppose that we have the data on the left.
- ✓ We also assume that $u_i \sim N(0,1)$.

Probit model and estimation

- ✓Take 2nd observation as an example. Since y=0 for this observation, we know y*<0.
- ✓Thus, the likelihood contribution is:

Id	y	x
1	0	1
2	0	4
3	1	5
4	1	6
5	1	9

$$L_2 = P(y_2^* < 0) = P(\beta_0 + 4\beta_1 + u_2 < 0)$$
$$= P(u_2 < -\beta_0 - 4\beta_1) = \underbrace{\Phi(-\beta_0 - 4\beta_1)}_{\text{Cumulative distribution function of standard normal distribution}}$$



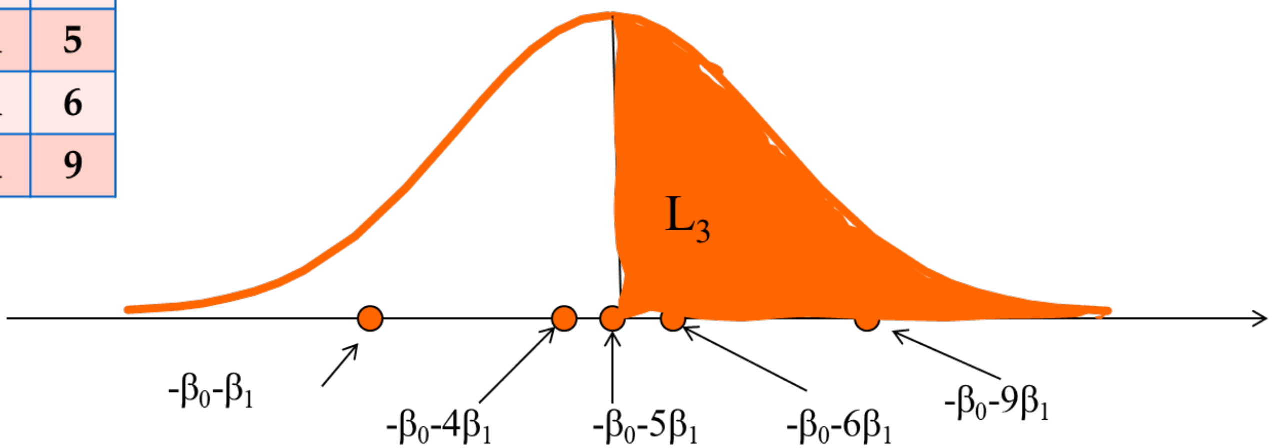
Probit model and estimation

✓ Now, take 3rd observation as an example. Since y=1 for this observation, we know y^*≥0.

✓ Thus, the likelihood contribution is:

Id	y	x
1	0	1
2	0	4
3	1	5
4	1	6
5	1	9

$$L_3 = P(y_3^* \geq 0) = P(\beta_0 + 5\beta_1 + u_3 \geq 0)$$
$$= P(u_3 \geq -\beta_0 - 5\beta_1) = 1 - \Phi(-\beta_0 - 5\beta_1)$$



Probit model and estimation

✓Then the likelihood function is given by:

Id	y	x
1	0	1
2	0	4
3	1	5
4	1	6
5	1	9

$$L(\beta_0, \beta_1) = \prod_{i=1}^{5} L_i = \Phi(-\beta_0 - \beta)\Phi(-\beta_0 - 4\beta)\left[1 - \Phi(-\beta_0 - 5\beta)\right]\times$$
$$\times\left[1 - \Phi(-\beta_0 - 6\beta)\right]\left[1 - \Phi(-\beta_0 - 9\beta)\right]$$

Probit model and estimation

- ✓ Usually, we maximize the $\ln(L)$ instead of the L. Due to invariance, the result is identical.
- ✓ The values of the parameters that maximize $\ln(L)$ are the ML estimators of the (binomial) *probit* model (sometimes it is also being called the *normit* model).
- ✓ The MLE is done automatically in Stata and R.

Generalization of the MLE procedure

The likelihood and the log-likelihood function:

$$L = \prod_{i=1}^{n} L_i(\beta) = \prod_{i=1}^{n} (\Phi(X\beta))^{y_i} \cdot (1 - \Phi(X\beta))^{1-y_i}$$

$$lnL = \sum_{i=1}^{n} l_i(\beta) = \sum_{i=1}^{n} y_i ln\Phi(X\beta) + \sum_{i=1}^{n} (1 - y_i) ln(1 - \Phi(X\beta))$$

Regression coeffcient estimates $\hat{\boldsymbol{\beta}}$ are obtained as the values of $\boldsymbol{\beta}$ that maximize the log-likelihood function $\ln L$ by using numerical methods (iterative procedures). These methods require the gradient, which is the first derivative of the $\ln L$.

Marginal or partial effects

The probit model is non-linear, therefore the estimated coefficients $\hat{\beta}$ do not reflect the "strength" of the effects of a change in *x* on the probability of occurrence of *y*.

Instead, we calculate the **marginal effects**:

Probability density function of the std. normal distr.

- For a continuous x_k: $mfx_k = \frac{\partial p(x)}{\partial x_k} = \phi(X\beta)\beta_k$.

  1. Since $\Phi(\cdot)$ is strictly increasing, sign of β_k is the same as the sign of $\frac{\partial p(x)}{\partial x_k}$;

  2. Relative effects do not depend on *x*: $\frac{\frac{\partial p(x)}{\partial x_k}}{\frac{\partial p(x)}{\partial x_j}} = \frac{\phi(X\beta)\beta_k}{\phi(X\beta)\beta_j} = \frac{\beta_k}{\beta_j}$.

Marginal or partial effects

- For a discrete x_k: a change of x_k from c to c+1 results in:

$$\Delta p(x) = \Phi\big(\beta_0 + \beta_1 x_1 + \cdots + \beta_k (c_k + 1)\big) - \Phi(\beta_0 + \beta_1 x_1 + \cdots + \beta_k c_k).$$

For a dummy explanatory variable x_k, $c = 0$:

$$\Delta p(x) = \Phi(\beta_0 + \beta_1 x_1 + \cdots + \beta_k \cdot 1) - \Phi(\beta_0 + \beta_1 x_1 + \cdots + \beta_k \cdot 0).$$

As you will see later, we usually evaluate marginal effects at mean values of all explanatory variables, $\phi(\overline{\mathrm{X}}\beta)$, but in general, we can choose any values of our explanatory variables.

9.4 The Logit Model

Logit model and estimation

- Again, consider the following model:

$$y_i^* = \beta_0 + \beta_1 x_i + u_i$$

$$\begin{cases} \text{If } y_i = 0, \text{ then } y_i^* < 0 \\ \text{If } y_i = 1, \text{ then } y_i^* \geq 0 \end{cases}$$

- In the logit model, we assume that u_i follows the logistic distribution with mean 0 and variance 1, which has the following density function:

$$f(u) = \frac{e^{-u}}{(1 + e^{-u})^2}$$

Logit model and estimation

Id	y	x
1	0	1
2	0	4
3	1	5
4	1	6
5	1	9

- ✓Now, suppose that you have the data on the left.
- ✓Take the 2nd observation as an example. Since y=0, it must have been the case that y*<0.
- ✓Thus, the likelihood contribution is:

$$L_2 = P(y_2^* < 0) = P(\beta_0 + 4\beta_1 + u_2 < 0)$$

$$= P(u_2 < -\beta_0 - 4\beta_1) = \underbrace{\Lambda(-\beta_0 - 4\beta_1)}_{\text{Cumulative distribution function of logistic distribution}}$$

$$= \frac{1}{1+e^{-(-\beta_0 - 4\beta_1)}} = \frac{1}{1+e^{\beta_0 + 4\beta_1}}$$

Logit model and estimation

Id	y	x
1	0	1
2	0	4
3	1	5
4	1	6
5	1	9

- ✓Now, take the 3rd observation as an example. Since y=1, it must have been the case that $y^*\geq 0$.
- ✓Thus, the likelihood contribution is:

$$L_3 = P(y_3^* \geq 0) = P(\beta_0 + 5\beta_1 + u_3 \geq 0)$$
$$= P(u_3 \geq -\beta_0 - 5\beta_1) = 1 - \Lambda(-\beta_0 - 5\beta_1)$$
$$= 1 - \frac{1}{1+e^{-(-\beta_0 - 5\beta_1)}} = 1 - \frac{1}{1+e^{\beta_0 + 5\beta_1}} = \frac{e^{\beta_0 + 5\beta_1}}{1+e^{\beta_0 + 5\beta_1}}$$

Logit model and estimation

✓Thus the likelihood function for the data set is given by:

Id	y	x
1	0	1
2	0	4
3	1	5
4	1	6
5	1	9

$$L=\prod_{i=1}^{5} L_i =$$

$$=\frac{1}{1+e^{\beta_0+\beta_1}}\times\frac{1}{1+e^{\beta_0+4\beta_1}}\times\frac{e^{\beta_0+5\beta_1}}{1+e^{\beta_0+5\beta_1}}\times\frac{e^{\beta_0+6\beta_1}}{1+e^{\beta_0+6\beta_1}}\times\frac{e^{\beta_0+9\beta_1}}{1+e^{\beta_0+9\beta_1}}$$

Logit model and estimation

- ✓ Again, we usually maximize the $\ln(L)$ instead of the L. Due to invariance, the result is identical.
- ✓ The values of the parameters that maximize $\ln(L)$ are the ML estimators of the (binomial) *logit* model.
- ✓ The MLE is done automatically in Stata and R.

Generalization of the MLE procedure

The likelihood and the log-likelihood function:

$$L = \prod_{i=1}^{n} L_i(\beta) = \prod_{i=1}^{n} (\Lambda(X\beta))^{y_i} \cdot (1 - \Lambda(X\beta))^{1-y_i}$$

$$lnL = \sum_{i=1}^{n} l_i(\beta) = \sum_{i=1}^{n} y_i ln\Lambda(X\beta) + \sum_{i=1}^{n} (1 - y_i) ln(1 - \Lambda(X\beta))$$

Again, the regression coeffcient estimates $\hat{\boldsymbol{\beta}}$ are obtained as the values of $\boldsymbol{\beta}$ that maximize the log-likelihood function $\ln L$ by using numerical methods (iterative procedures). These methods require the gradient, which is the first derivative of the $\ln L$.

Marginal or partial effects

The logit model is also non-linear, therefore the estimated coefficients $\hat{\boldsymbol{\beta}}$ do not reflect the "strength" of the effects of a change in *x* on the probability of occurrence of *y*.

Instead, we calculate the **marginal effects**:

Probability density function of the logistic distr.

- For a continuous x_k: $mfx_k = \frac{\partial p(x)}{\partial x_k} = \lambda(X\beta)\beta_k$.

1. Since $\Lambda(\cdot)$ is strictly increasing, sign of β_k is the same as the sign of $\frac{\partial p(x)}{\partial x_k}$;

2. Relative effects do not depend on *x*: $\frac{\frac{\partial p(x)}{\partial x_k}}{\frac{\partial p(x)}{\partial x_j}} = \frac{\lambda(X\beta)\beta_k}{\lambda(X\beta)\beta_j} = \frac{\beta_k}{\beta_j}$.

Marginal or partial effects

- For a discrete x_k: a change of x_k from c to c+1 results in:

$$\Delta p(x) = \Lambda\big(\beta_0 + \beta_1 x_1 + \cdots + \beta_k (c_k + 1)\big) - \Lambda(\beta_0 + \beta_1 x_1 + \cdots + \beta_k c_k).$$

For a dummy explanatory variable x_k, c = 0:

$$\Delta p(x) = \Lambda(\beta_0 + \beta_1 x_1 + \cdots + \beta_k \cdot 1) - \Lambda(\beta_0 + \beta_1 x_1 + \cdots + \beta_k \cdot 0).$$

In case of the logistic distribution: $\lambda(X\beta) = \Lambda(X\beta)(1 - \Lambda(X\beta))$.

Again, most often we evaluate marginal effects at mean values of all explanatory variables, $\lambda(\bar{X}\beta)$, but in general, we can choose any values of our explanatory variables.

Odds, odds ratio and the logit

Logit model can also be analyzed in terms of **odds**, i.e. the ratio between the probability of a “positive” outcome (1) and the probabilty of a “negative” outcome (0).

Example for an unspecified course:

- Probability of passing: 3/5
- Probability of failing: 2/5

- Odds of passing (if 1 – pass): $\frac{3/5}{2/5} = \frac{3}{2}$
- Odds of failing (if 1 – fail): $\frac{2/5}{3/5} = \frac{2}{3}$

Odds, odds ratio and the logit

In our case, the **odds** Ω are defined as:

$$\Omega = \frac{P(y=1|x)}{P(y=0|x)} = \frac{P(y=1|x)}{1-P(y=1|x)} =$$

$$= \frac{\frac{e^{X\beta}}{1+e^{X\beta}}}{1-\frac{e^{X\beta}}{1+e^{X\beta}}} = \frac{e^{X\beta}}{1+e^{X\beta}} \cdot \frac{1+e^{X\beta}}{1+e^{X\beta}-e^{X\beta}} = e^{X\beta}$$

The log of odds, also called the **logit**, is then defined as:

$$ln\Omega = lne^{X\beta} = X\beta = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

The logit model is thus *simply* a linear OLS regression of log of odds on explanatory variables x_j. Unfortunately, we do ***not*** **observe the odds** in practice.

Odds, odds ratio and the logit

We also have the **odds ratio**, *OR*:

$$OR_k = \frac{\Omega(X; x_k + 1)}{\Omega(X;\ x_k)} = e^{\beta_k}$$

Interpretation: If x_k increases by 1 unit of measurement, then the odds that $y = 1$ *change*, ceteris paribus, by a factor of e^{β_k}.

We thus have three possibilities:

- *OR* = 1: odds unchanged;
- *OR* > 1: odds increase;
- *OR* < 1: odds decrease.

University of Ljubljana

SEB SCHOOL OF ECONOMICS AND BUSINESS

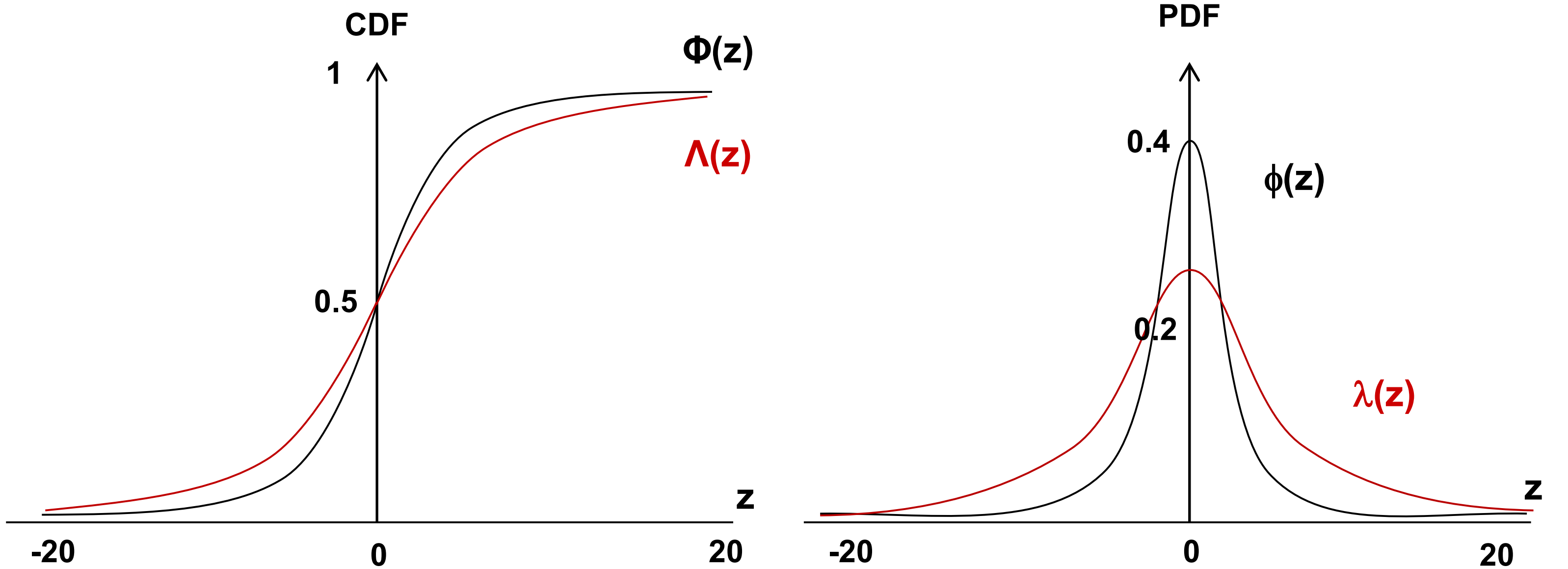
9.5 Model Comparison and Interpretation

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION AMBA ACCREDITED

Model comparison



1. Estimated coefficients differ: $\beta_{\text{logit}} \approx 1.7 \cdot \beta_{\text{probit}}$.
2. Logistic distribution has "fatter tails", which matters only if the distribution of sample values *y* is extreme (e.g. 95% of 1s).

Model interpretation

- ✓ The statistical significance of the regression coefficients can tell you whether the effect of x on the probability of y exists or not.
- ✓ The sign of the regression coefficients can tell you if the probability of y will increase or decrease, when x increases.
- ✓ What if we also want to know "by how much" the probability increases or decreases (the strength of the effect)?
- ✓ This can be done by computing the **partial effects**, also called the **marginal effects**.
- ✓ An alternative approach is to compute the **odds ratios**, done solely based on the *logit model*.

Model interpretation: marginal effects

- ✓ As we already observed, the partial or marginal effect depends on the value of the explanatory variables. Therefore, it is different for every observation in the data.
- ✓ However, we want to know the overall effect of x on the probability of y.
- ✓ For this purpose, we calculate the **partial effect at average**, also called the **marginal effect at average**.

Model interpretation: marginal effects

This is also most commonly done in practice, i.e. the marginal effects are evaluated at mean values of explanatory variables:

$$\overline{mfx}_k = PEA_k = \phi(\bar{\mathrm{X}}\beta)\beta_k \text{ for the probit model;}$$
$$\overline{mfx}_k = PEA_k = \lambda(\bar{\mathrm{X}}\beta)\beta_k \text{ for the logit model.}$$

Interpretation of marginal effects at average involves three relativizations: 1) on average, 2) ceteris paribus, and 3) given the mean (average) values of explanatory variables.

Model interpretation: marginal effects

Three **most common cases**:

a) Numerical explanatory variable **in levels, $\boldsymbol{x_k}$**:

If x_k increases by 1 unit of measurement, then the probability that y = 1, on average, ceteris paribus, given the means of explanatory variables, increases/decreases by $100 \cdot \overline{mfx_k}$ percentage points.

b) Numerical explanatory variable **in logs, ln $\boldsymbol{x_k}$**:

If x_k increases by 1 percent, then the probability that y = 1, on average, ceteris paribus, given the means of explanatory variables, increases/decreases by $\overline{mfx_k}$ percentage points.

Model interpretation: marginal effects

c) Explanatory variable is a **dummy variable, *D***:

If *D* = 1, then the probability that *y* = 1, on average, ceteris paribus, given the means of explanatory variables, increases/decreases by $100 \cdot \overline{mfx_k}$ percentage points, compared to *D* = 0.

Of course, we can evaluate marginal effects at any feasible values of explanatory variables. In that case, we need to use the chosen values of explanatory variables in the calculation and **adjust the third relativization** accordingly.

Alternatively, we could also have calculated the **average partial effect** or **average marginal effect**, which is the average of the partial effects calculated separately by observations.

Model interpretation: odds ratios

Three **most common cases**:

a) Numerical explanatory variable **in levels, x_k**:

$$OR_k = e^{\beta_k}$$

If x_k increases by 1 unit of measurement, then the odds that $y = 1$, ceteris paribus, increase by $100 \cdot (OR_k - 1)$ percent (if OR > 1) or decrease by $100 \cdot (1 - OR_k)$ percent (if OR < 1).

b) Explanatory variable is a **dummy variable, D**:

If $D = 1$, then the odds that $y = 1$, ceteris paribus, increase by $100 \cdot (OR_k - 1)$ percent (if OR > 1) or decrease by $100 \cdot (1 - OR_k)$ percent (if OR < 1), compared to $D = 0$.

Model interpretation: odds ratios

c) Numerical explanatory variable **in logs, ln x_k**:

$$e^{\beta_k \cdot \ln 1.01}$$

If x_k increases by 1 percent, then the odds that y = 1, ceteris paribus, increase by 100 · ($e^{\beta_k \cdot \ln 1.01}$ – 1) percent (if $e^{\beta_k \cdot \ln 1.01}$ > 1) or decrease by 100 · (1 – $e^{\beta_k \cdot \ln 1.01}$) percent (if $e^{\beta_k \cdot \ln 1.01}$ < 1).

Of course, we can evaluate odds ratios for larger changes in explanatory variables. In that case, we need to insert above the chosen change of explanatory variable appropriately.

An example in Stata and R

We have data on several variables concerning the annual holidays:

- *abroad*: dichotomous variable for a person spending holidays abroad: 1 – abroad, 0 – in the home country;
- *log_income*: logarithm of annual family net income per household member;
- *age*: person's age;
- *pet*: dichotomous variable for the presence of pets in the family: 1 – yes, 0 – no.

Estimate the logit model for the *abroad* variable as the dependent variable and all the other variables as explanatory variables.

Calculate the marginal effects at the means of explanatory variables (centroid). Interpret the calculated marginal effects.

An example in Stata

```
. logit abroad log_income age i.pet

Iteration 0:   log likelihood = -27.525553
Iteration 1:   log likelihood = -17.032774
Iteration 2:   log likelihood = -16.960203
Iteration 3:   log likelihood = -16.959864
Iteration 4:   log likelihood = -16.959864

Logistic regression                             Number of obs     =         40
                                                LR chi2(3)        =      21.13
                                                Prob > chi2       =     0.0001
Log likelihood = -16.959864                     Pseudo R2         =     0.3839
```

$\ln L_{\text{const}}$ (Iteration 0)

$\ln L_{\text{model}}$ (Iteration 4)

$$1 - \frac{lnL_{model}}{lnL_{const}}$$

abroad	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
log_income	4.250643	1.449276	2.93	0.003	1.410114	7.091172
age	-.1004235	.0442162	-2.27	0.023	-.1870858	-.0137613
1.pet	-2.278437	1.075051	-2.12	0.034	-4.385498	-.171377
_cons	-33.26217	12.14418	-2.74	0.006	-57.06433	-9.460013

An example in Stata

```
. margins, dydx(log_income age pet) atmeans

Conditional marginal effects                    Number of obs     =         40
Model VCE    : OIM

Expression   : Pr(abroad), predict()
dy/dx w.r.t. : log_income age 1.pet
at           : log_income      =    9.193383 (mean)
               age             =       47.85 (mean)
               0.pet           =          .7 (mean)
               1.pet           =          .3 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  log_income |   1.034782    .350304     2.95   0.003     .3481992    1.721366
         age |  -.0244472   .0108509    -2.25   0.024    -.0457146   -.0031799
       1.pet |  -.5135038   .1927674    -2.66   0.008     -.891321   -.1356866
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

University of Ljubljana
SEB School of Economics and Business

An example in R

```
> mod_logit = glm(abroad ~ log_income + age + factor(pet), family=binomial(link="logit"),
 data=holiday)
> summary(mod_logit)

Call:
glm(formula = abroad ~ log_income + age + factor(pet), family = binomial(link = "logit"),
    data = holiday)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7062  -0.6569   0.1760   0.5858   1.9286

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -33.26217   12.14386  -2.739  0.00616 **
log_income     4.25064    1.44923   2.933  0.00336 **
age           -0.10042    0.04422  -2.271  0.02313 *
factor(pet)1  -2.27844    1.07503  -2.119  0.03405 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 55.051  on 39  degrees of freedom
Residual deviance: 33.920  on 36  degrees of freedom
AIC: 41.92

Number of Fisher Scoring iterations: 5
```

An example in R

$$1 - \frac{lnL_{model}}{lnL_{const}}$$

```
> PseudoR2(mod_logit, which="all")
       McFadden     McFaddenAdj        CoxSnell      Nagelkerke   AldrichNelson
      0.3838502       0.2385307       0.4103844       0.5490215       0.3456715
VeallZimmermann           Efron McKelveyZavoina            Tjur             AIC
      0.5968356       0.4203134       0.6137739       0.4331495      41.9197276
            BIC          logLik         logLik0              G2
     48.6752455     -16.9598638     -27.5255525      21.1313775
> mfx_logit = logitmfx(abroad ~ log_income + age + factor(pet), data=holiday)
> mfx_logit
Call:
logitmfx(formula = abroad ~ log_income + age + factor(pet), data = holiday)

Marginal Effects:
                  dF/dx Std. Err.       z    P>|z|
log_income     1.034783  0.350294  2.9540 0.003136 **
age           -0.024447  0.010851 -2.2531 0.024254 *
factor(pet)1 -0.513504  0.192763 -2.6639 0.007724 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

dF/dx is for discrete change for the following variables:

[1] "factor(pet)1"
```

LR

University of Ljubljana

SEB SCHOOL OF ECONOMICS AND BUSINESS

9. Discrete Choice Models

Prof. Dr. Miroslav Verbič

miroslav.verbic@ef.uni-lj.si

www.miroslav-verbic.si

EFMD EQUIS ACCREDITED

AACSB ACCREDITED

ASSOCIATION OF MBAS ACCREDITED

Ljubljana, October 2025