The Probit Model

Alexander Spermann
University of Freiburg

SoSe 2009

Course outline

1. Notation and statistical foundations
2. Introduction to the Probit model
3. Application
4. Coefficients and marginal effects
5. Goodness-of-fit
6. Hypothesis tests

Notation and statistical foundations

1. $y_i = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$ Gujarati

$y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + u_t$ Wooldridge

2. Matrix

$$Y = X\beta + \varepsilon$$

$$Y = x'\beta + \varepsilon$$

$$Y = X\hat{\beta} + \hat{u}$$

$$y_i = x_i'\beta + \varepsilon_i$$

$$\begin{array}{cc} x_i'\beta & x_i'\beta \\ \begin{pmatrix} 1 & x_1 & x_2 \end{pmatrix}\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} & \begin{pmatrix} 1 & x_2 & x_3 \end{pmatrix}\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \end{array}$$

Notation and statistical foundations – Vectors

- Column vector:

$$\underset{nx1}{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

- Transposed (row vector): $\underset{1xn}{a}' = \begin{bmatrix} a_1 & a_2 & \dots & a_n \end{bmatrix}$

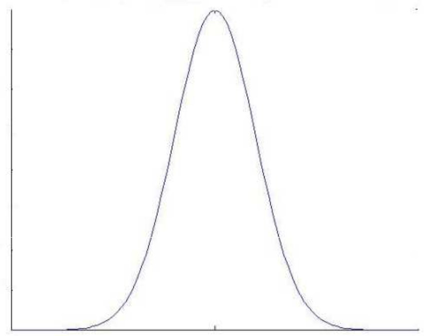
- Inner product:

$$a'b = \begin{bmatrix} a_1 & a_2 & \dots & a_n \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = \sum a_i b_i$$

Notation and statistical foundations – density function

- PDF: probability density function f(x)
- Example: Normal distribution:

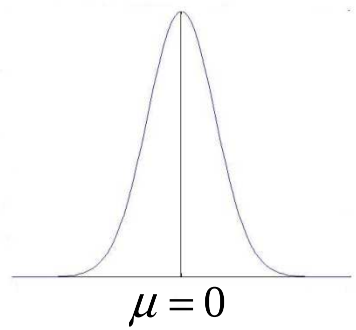
$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{(x-\mu)^2}{\sigma^2}\right]}$$



- Example: Standard normal distribution:

N(0,1), $\mu = 0$, $\sigma = 1$

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$



Notation and statistical foundations – distibutions

- Standard logistic distribution:

$$f(x) = \frac{e^x}{\left(1+e^x\right)^2}, \mu = 0, \sigma^2 = \frac{\pi^2}{3}$$

Exponential distribution:

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}}, x \geq 0 \\ 0, x \leq 0 \end{cases}, \theta > 0, \mu = \theta, \sigma^2 = \theta^2$$

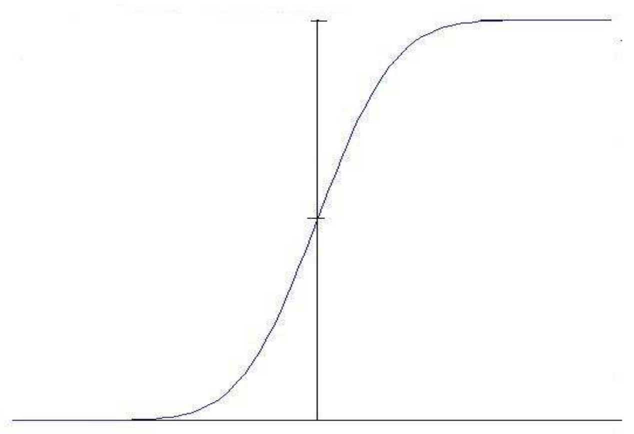
- Poisson distribution:

$$f(x) = \frac{e^{-\theta}\theta^x}{x!}, \mu = \theta, \sigma^2 = \theta$$

Notation and statistical foundations – CDF

- CDF: cumulative distribution function F(x)
- Example: Standard normal distribution:

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$



- The cdf is the integral of the pdf.

Notation and statistical foundations – logarithms

- Rule I:

$$y = x\,z$$

$$\log y = \log x + \log z$$

- Rule II:

$$y = x^{n}$$

$$\log y = n \log x$$

- Rule III:

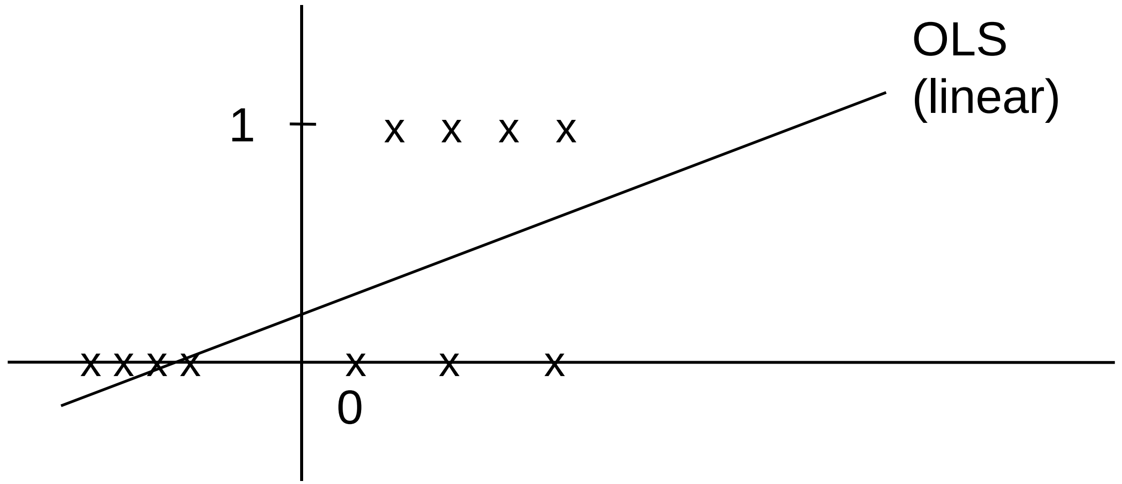
$$y = a\,x^{b}$$

$$\log y = \log a + b \log x$$

Introduction to the Probit model – binary variables

- Why not use OLS instead?

$$y = \begin{cases} 1 \\ 0 \end{cases}$$



- Nonlinear estimation, for example by maximum likelihood.

Introduction to the Probit model – latent variables

- Latent variable: Unobservable variable y* which can take all values in (-∞, +∞).
- Example: y* = Utility(Labour income) - Utility(Non labour income)
- Underlying latent model:

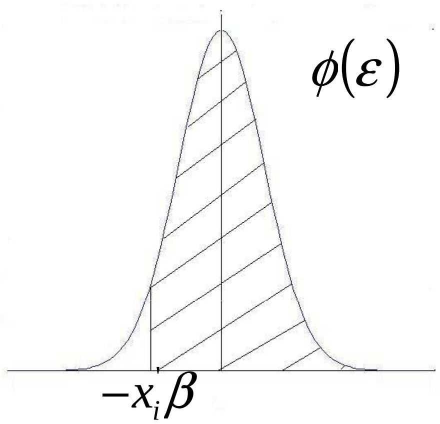
$$y_i = \begin{cases} 1, & y_i^* > 0 \\ 0, & y_i^* \leq 0 \end{cases}$$

$$y_i^* = x_i^{'}\beta + \varepsilon_i$$

Introduction to the Probit model – latent variables

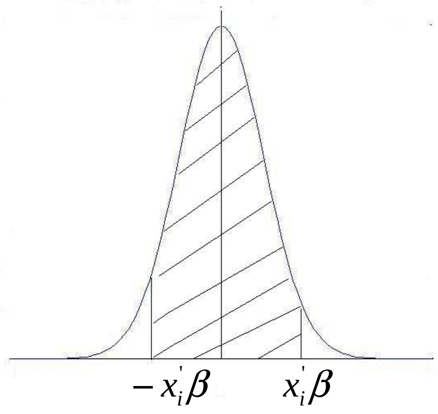
➤ Probit is based on a latent model:

$$P(y_i = 1 \mid x) = P(y_i^* > 0 \mid x)$$
$$= P(x_i'\beta + \varepsilon_i > 0 \mid x)$$
$$= P(\varepsilon_i > -x_i'\beta \mid x)$$
$$= 1 - F(-x_i'\beta)$$



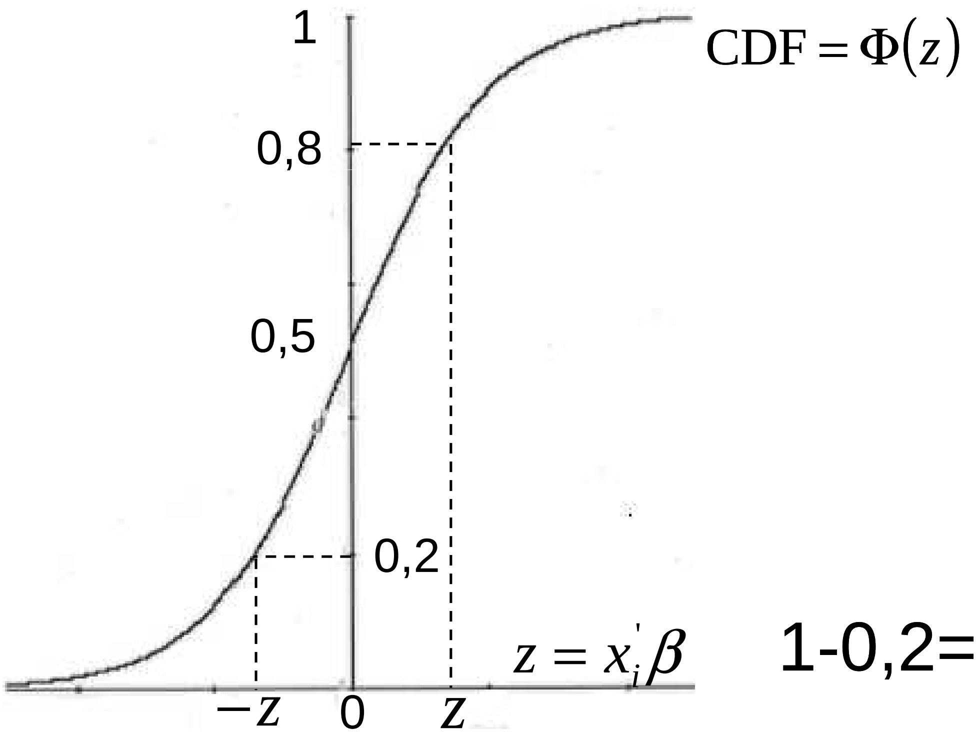
Assumption: Error terms are independent and normally distributed:

$$P(y_i = 1 \mid x) = 1 - \Phi(-\frac{x_i'\beta}{\sigma}), \sigma \equiv 1$$
$$= \Phi(x_i'\beta) \quad \text{because of symmetry}$$



Introduction to the Probit model – CDF

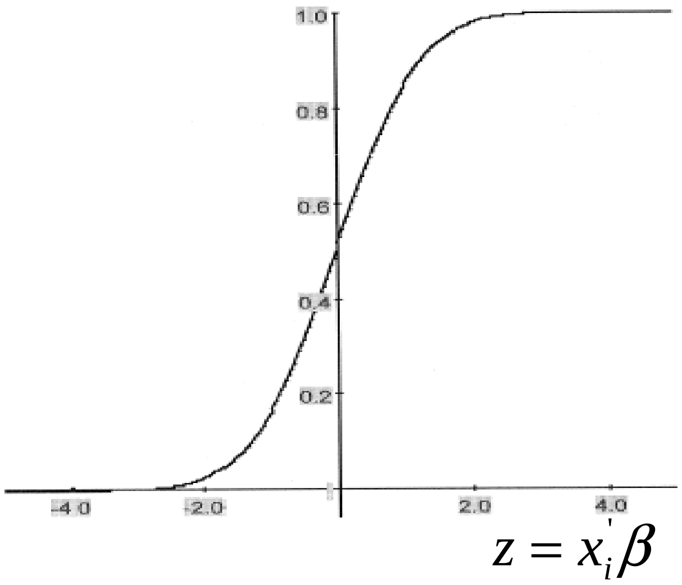
➢ Example:

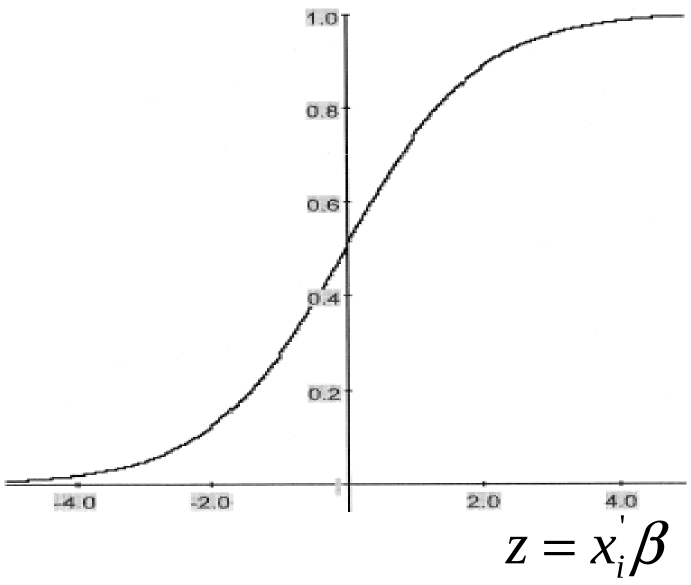


1-0,2=0,8

Introduction to the Probit model – CDF Probit vs. Logit

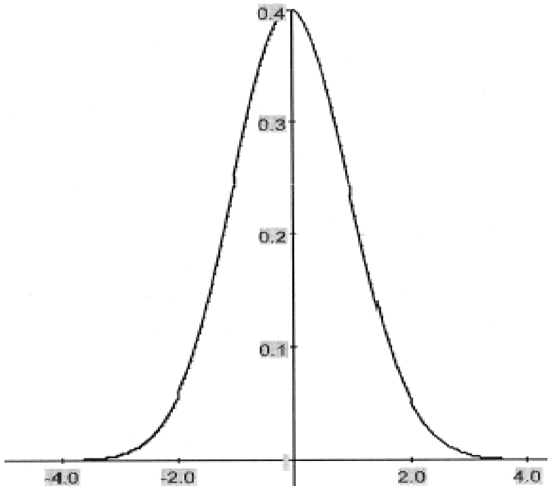
- F(z) lies between zero and one
- CDF of Probit: CDF of Logit:



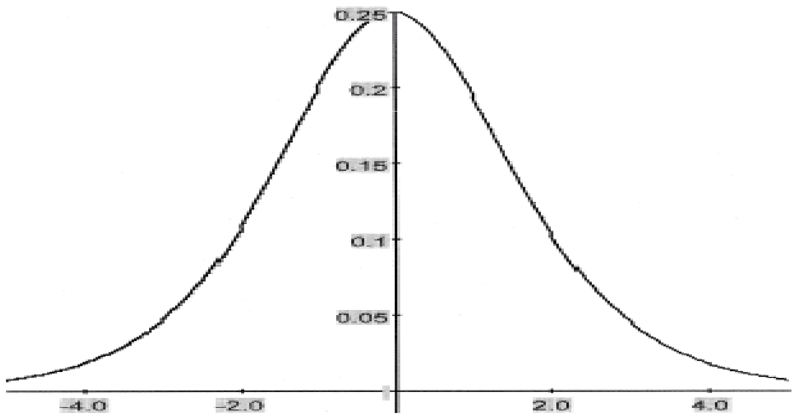


Introduction to the Probit model – PDF Probit vs. Logit

- PDF of Probit:



PDF of Logit:



Introduction to the Probit model – The ML principle

- Joint density:

$$f(y \mid x, \beta) = \prod_i F(x_i^{'}\beta)^{y_i} \left[1 - F(x_i^{'}\beta)\right]^{(1-y_i)}$$

$$= \prod_i F_i^{y_i} (1 - F_i)^{1-y_i}$$

- Log likelihood function:

$$\ln L = \sum_i y_i \ln F_i + (1 - y_i) \ln(1 - F_i)$$

Introduction to the Probit model – The ML principle

- The principle of ML: Which value of β maximizes the probability of observing the given sample?

$$\frac{\partial \ln L}{\partial \beta} = \sum_i \left[\frac{y_i f_i}{F_i} + \frac{(1-y_i)(-f_i)}{1-F_i} \right] x_i$$

$$= \sum_i \left[\frac{y_i - F_i}{F_i(1-F_i)} f_i \right] x_i$$

$$= 0$$

Introduction to the Probit model – Example

- Example taken from Greene, Econometric Analysis, 5. ed. 2003, ch. 17.3.
- 10 observations of a discrete distribution
- Random sample: 5, 0, 1, 1, 0, 3, 2, 3, 4, 1
- PDF:

$$f(x_i,\theta)=\frac{e^{-\theta}\theta^{x_i}}{x_i!}$$

- Joint density :

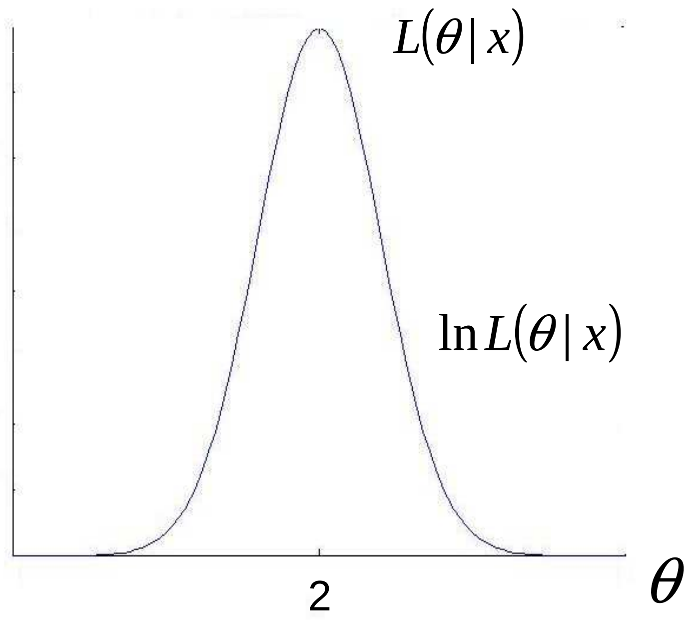
$$f(x_1,x_2,\ldots,x_{10}\,|\,\theta)=\prod_{i=1}^{10} f(x_i,\theta)=\frac{e^{-10\theta}\cdot\theta^{\sum_i x_i}}{\prod_{i=1}^{10} x_i!}=\frac{e^{-10\theta}\cdot\theta^{20}}{207{,}36}$$

- Which value of θ makes occurance of the observed sample most probable?

Introduction to the Probit model – Example

$$\ln L(\theta) = -10\theta + 20\ln\theta - 12{,}242$$

$$\frac{d\ln L(\theta)}{d\theta} = -10 + \frac{20}{\theta} = 0$$



$$\frac{d^2 \ln L(\theta)}{d\theta^2} = -\frac{20}{\theta^2}$$

$$\Rightarrow Maximum$$

Application

- Analysis of the effect of a new teaching method in economic sciences
- Data:

Beobachtung	GPA	TUCE	PSI	Grade	Beobachtung	GPA	TUCE	PSI	Grade
1	2,66	20	0	0	17	2,75	25	0	0
2	2,89	22	0	0	18	2,83	19	0	0
3	3,28	24	0	0	19	3,12	23	1	0
4	2,92	12	0	0	20	3,16	25	1	1
5	4	21	0	1	21	2,06	22	1	0
6	2,86	17	0	0	22	3,62	28	1	1
7	2,76	17	0	0	23	2,89	14	1	0
8	2,87	21	0	0	24	3,51	26	1	0
9	3,03	25	0	0	25	3,54	24	1	1
10	3,92	29	0	1	26	2,83	27	1	1
11	2,63	20	0	0	27	3,39	17	1	1
12	3,32	23	0	0	28	2,67	24	1	0
13	3,57	23	0	0	29	3,65	21	1	1
14	3,26	25	0	1	30	4	23	1	1
15	3,53	26	0	0	31	3,1	21	1	0
16	2,74	19	0	0	32	2,39	19	1	1

Source: Spector, L. and M. Mazzeo, Probit Analysis and Economic Education. In: Journal of Economic Education, 11, 1980, pp.37-44

Application – Variables

- **Grade**
  Dependent variable. Indicates whether a student improved his grades after the new teaching method PSI had been introduced (0 = no, 1 = yes).

- **PSI**
  Indicates if a student attended courses that used the new method (0 = no, 1 = yes).

- **GPA**
  Average grade of the student

- **TUCE**
  Score of an intermediate test which shows previous knowledge of a topic.

Application – Estimation

- Estimation results of the model (output from Stata):

```
. probit  grade psi tuce gpa

Iteration 0:   log likelihood =  -20.59173
Iteration 1:   log likelihood = -13.315851
Iteration 2:   log likelihood = -12.832843
Iteration 3:   log likelihood = -12.818826
Iteration 4:   log likelihood = -12.818803

Probit estimates                                  Number of obs   =         32
                                                  LR chi2(3)      =      15.55
                                                  Prob > chi2     =     0.0014
Log likelihood = -12.818803                       Pseudo R2       =     0.3775

------------------------------------------------------------------------------
       grade |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         psi |   1.426332    .595037     2.40   0.017      .2600814    2.592583
        tuce |   .0517289   .0838901     0.62   0.537     -.1126927    .2161506
         gpa |    1.62581   .6938818     2.34   0.019      .2658269    2.985794
       _cons |   -7.45232   2.542467    -2.93   0.003     -12.43546   -2.469177
------------------------------------------------------------------------------
```

Application – Discussion

- ML estimator: Parameters were obtained by maximization of the log likelihood function.
  Here: 5 iterations were necessary to find the maximum of the log likelihood function (-12.818803)

- Interpretation of the estimated coefficients:
  - Estimated coefficients do not quantify the influence of the rhs variables on the probability that the lhs variable takes on the value one.
  - Estimated coefficients are parameters of the latent model.

Coefficients and marginal effects

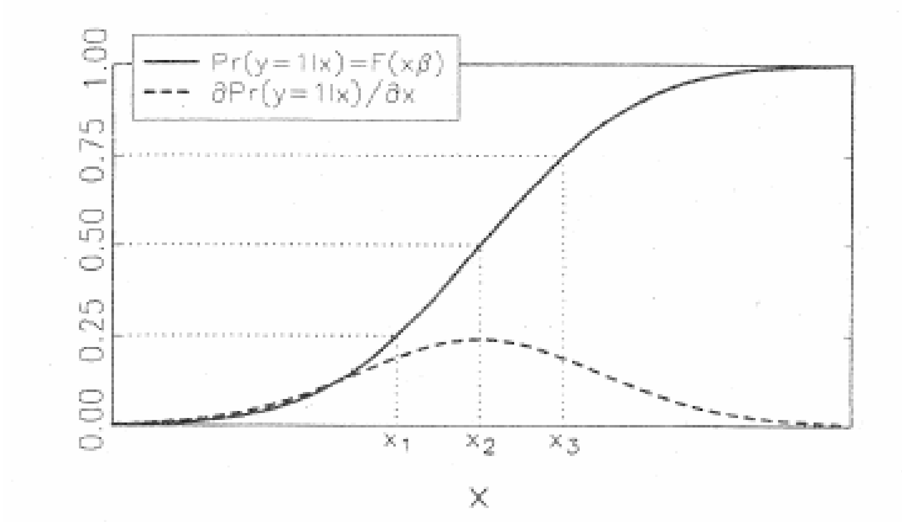
- The marginal effect of a rhs variable is the effect of an unit change of this variable on the probability P(Y = 1|X = x), given that all other rhs variables are constant:

$$\frac{\partial P(y_i = 1 \mid x_i)}{\partial x_i} = \frac{\partial E(y_i \mid x_i)}{\partial x_i} = \varphi(x_i' \beta)\beta$$

- Recap: The slope parameter of the linear regression model measures directly the marginal effect of the rhs variable on the lhs variable.

Coefficients and marginal effects

- The marginal effect depends on the value of the rhs variable.
- Therefore, there exists an individual marginal effect for each person of the sample:



Coefficients and marginal effects – Computation

- Two different types of marginal effects can be calculated:
  - Average marginal effect
    Stata command: `margin`

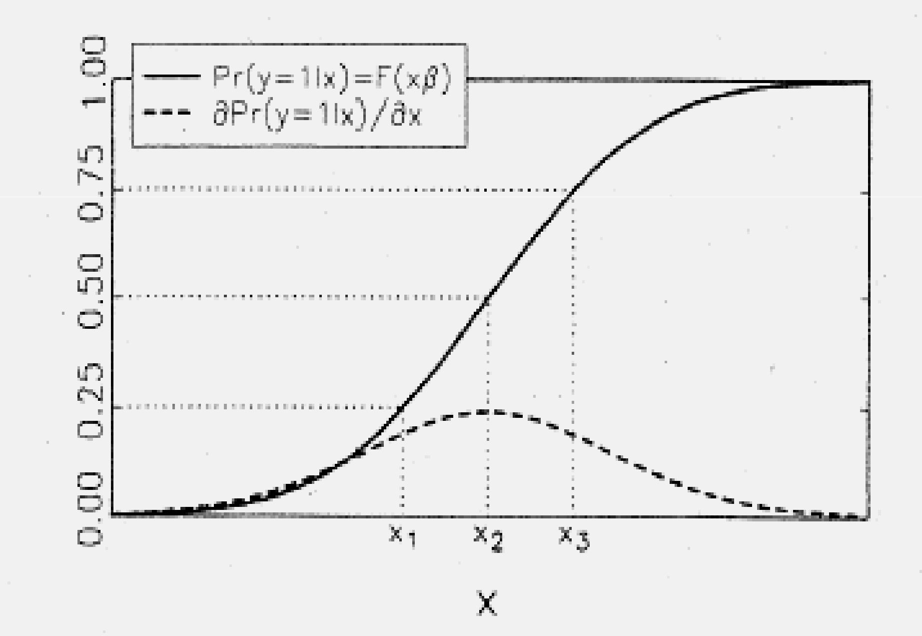
```
Marginal effects on Prob(grade==1) after probit
```

grade	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
gpa	.3637883	.1129461	3.22	0.001	.1424181	.5851586
tuce	.011476	.0184085	0.62	0.533	-.024604	.047556
psi	.3737518	.1399912	2.67	0.008	.0993741	.6481295

  - Marginal effect at the mean:
    Stata command: `mfx compute`

Coefficients and marginal effects – Computation

- Principle of the computation of the average marginal effects:



- Average of individual marginal effects

Coefficients and marginal effects – Computation

- Computation of average marginal effects depends on type of rhs variable:
  - Continuous variables like TUCE and GPA:

$$AME = \frac{1}{n}\sum_{i=1}^{n}\varphi(x_i^{'}\beta)\beta$$

  - Dummy variable like PSI:

$$AME = \frac{1}{n}\sum_{i=1}^{n}\left[\Phi(x_i^{'}\beta \mid x_i^{k}=1) - \Phi(x_i^{'}\beta \mid x_i^{k}=0)\right]$$

Coefficients and marginal effects – Interpretation

- Interpretation of average marginal effects:
  - Continuous variables like TUCE and GPA:
    An infinitesimal change of TUCE or GPA changes the probability that the lhs variable takes the value one by X%.
  - Dummy variable like PSI:
    A change of PSI from zero to one changes the probability that the lhs variable takes the value one by X percentage points.

Coefficients and marginal effects – Interpretation

Variable	Estimated marginal effect	Interpretation
GPA	0.364	If the average grade of a student goes up by an infinitesimal amount, the probability for the variable grade taking the value one rises by 36.4 %.
TUCE	0.011	Analog to GPA,with an increase of 1.1%.
PSI	0.374	If the dummy variable changes from zero to one, the probability for the variable grade taking the value one rises by 37.4 ppts.

Coefficients and marginal effects – Significance

- Significance of a coefficient: test of the hypothesis whether a parameter is significantly different from zero.
- The decision problem is similar to the t-test, wheras the probit test statistic follows a standard normal distribution. The z-value is equal to the estimated parameter divided by its standard error.
- Stata computes a p-value which shows directly the significance of a parameter:

	z-value	p-value	Interpretation
GPA :	3.22	0.001	***significant***
TUCE:	0,62	0,533	***insignificant***
PSI:	2,67	0,008	***significant***

Coefficients and marginal effects

- Only the average of the marginal effects is displayed.
- The individual marginal effects show large variation:

Descriptive statistics for individual marginal effects

	Mean	SD	Min	Max
gpa	0.36379	0.21358	0.06783	0.64807
tuce	0.01148	0.00687	0.00209	0.02063
psi	0.37375	0.12878	0.06042	0.51959

Stata command: `margin, table`

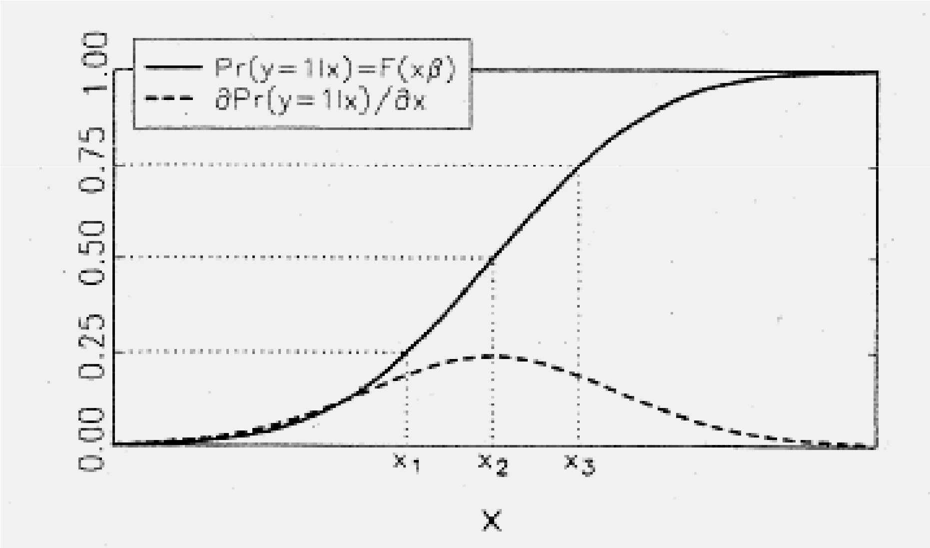
Coefficients and marginal effects

- Variation of marginal effects may be quantified by the confidence intervals of the marginal effects.
- In which range one can expect a coefficient of the population?
- In our example:

	Estimated coefficient	Confidence interval (95%)
GPA:	0,364	- 0,055 - 0,782
TUCE:	0,011	- 0,002 - 0,025
PSI:	0,374	0,121 - 0,626

Coefficients and marginal effects

- What is calculated by `mfx`?
- Estimation of the marginal effect at the sample mean.



Sample mean

Goodness of fit

- Goodness of fit may be judged by McFaddens Pseudo R^2.
- Measure for proximity of the model to the observed data.
- Comparison of the estimated model with a model which only contains a constant as rhs variable.
  - $\ln \hat{L}(M_{Full})$: Likelihood of model of interest.
  - $\ln \hat{L}(M_{Intercept})$: Likelihood with all coefficients except that of the intercept restricted to zero.
  - It always holds that $\ln \hat{L}(M_{Full}) \geq \ln \hat{L}(M_{Intercept})$

Goodness of fit

- The Pseudo R^2 is defined as:

$$Pseudo R^2 = R^2_{McF} = 1 - \frac{\ln \hat{L}(M_{Full})}{\ln \hat{L}(M_{Intercept})}$$

- Similar to the R^2 of the linear regression model, it holds that $0 \leq R^2_{McF} \leq 1$

- An increasing Pseudo R^2 may indicate a better fit of the model, whereas no simple interpretation like for the R^2 of the linear regression model is possible.

Goodness of fit

- A high value of R^2_{McF} does not necessarily indicate a good fit, however, as $R^2_{McF} = 1$ if $\ln \hat{L}(M_{Full}) = 0$.
- R^2_{McF} increases with additional rhs variables. Therefore, an adjusted measure may be appropriate:

$$PseudoR^2_{adjusted} = \overline{R}^2_{McF} = 1 - \frac{\ln \hat{L}(M_{Full}) - K}{\ln \hat{L}(M_{Intercept})}$$

- Further goodness of fit measures: R^2 of McKelvey and Zavoinas, Akaike Information Criterion (AIC), etc. See also the Stata command `fitstat`.

Hypothesis tests

- Likelihood ratio test: possibility for hypothesis testing, for example for variable relevance.
- Basic principle: Comparison of the log likelihood functions of the unrestricted model ($\ln L_U$) and that of the restricted model ($\ln L_R$)
- Test statistic: $LR = -2\ln\lambda = -2(\ln L_R - \ln L_U) \sim \chi^2(K)$

$$\lambda = \frac{L_R}{L_U} \qquad 0 \leq \lambda \leq 1$$

- The test statistic follows a χ^2 distribution with degrees of freedom equal to the number of restrictions.

Hypothesis tests

- Null hypothesis: All coefficients except that of the intercept are equal to zero.
- In the example: LR $\chi^2(3) = 15,55$
- Prob > chi2 = 0.0014
- Interpretation: The hypothesis that all coefficients are equal to zero can be rejected at the 1 percent significance level.