Introducción a la Economía Cuantitativa y al análisis de datos con STATA 18

Doctorado en Economía, Empresa, Finanzas y Computación

Universidad de Huelva

Junio-Julio 2024

Mónica Carmona (UHU y CCTH)

Emilio Congregado (UHU y CCTH)

Concepción Román (UHU y CCTH)

[md]ecofin] smart people

uhu.es

Universidad Internacional de Andalucía

AEET ASOCIACIÓN ESPAÑOLA ECONOMÍA DEL TRABAJO

Introducción a la Economía Cuantitativa y al análisis de datos con STATA 18

PROGRAMA

Día 27 de junio	Día 2 de julio
El análisis de datos en la investigación económica El modelo de regresión: un refresco de conceptos	Regresión con Stata Modelos de elección discreta con Stata
Día 28 de junio	**Día 4 de julio**
Introducción a Stata 18: una guía rápida Ventanas y ficheros Preparando Stata para trabajar Estructura básica de comandos y sintaxis	Lecturas de bases de datos Data Management

Contacto

Mónica **Carmona** monica@uhu.es
Emilio **Congregado** congregado@uhu.es
Concepción **Román** concepcion.roman@dege.uhu.es

[m d] ecofin] smart people

uhu.es Universidad Internacional de Andalucía AEET ASOCIACIÓN ESPAÑOLA ECONOMÍA DEL TRABAJO

Introducción a la Economía Cuantitativa y al análisis de datos con STATA 18

PROGRAMA

Día 27 de junio	Día 2 de julio
El análisis de datos en la investigación económica El modelo de regresión: un refresco de conceptos	Regresión con Stata Modelos de elección discreta con Stata
Día 28 de junio	**Día 4 de julio**
Introducción a Stata 18: una guía rápida Ventanas y ficheros Preparando Stata para trabajar Estructura básica de comandos y sintaxis	Lecturas de bases de datos Data Management

Contacto

Mónica **Carmona** monica@uhu.es
Emilio **Congregado** congregado@uhu.es
Concepción **Román** concepcion.roman@dege.uhu.es

Previous considerations

Classification of econometric models depending on characteristics of the dependent variable

- Discrete choice models
- Censored and truncated models
- Panel data models
- Duration (survival) models

Discrete choice models

- Discrete choice models are those in which the dependent variable (y) is discrete
  - Discrete decisions
  - Discrete event
  - Discrete way of collecting data

Discrete choice models

Classification of discrete choice models:

- Binary (y={0,1})
  - Linear probability model
  - Probit
  - Logit

- Multiple discrete choice models:
  - Non ordered:
    - Multinomial logit
    - Conditional logit
  - Ordered:
    - Ordered probit
    - Ordered logit

Binary discrete choice models

Binary discrete choice models

The simplest discrete choice model is that one in which the dependent variable takes on 2 values (usually y={0,1})

Examples:

- Individuals in the labour market
- To invest or not to invest
- To buy a house or not to buy
- …

Binary discrete choice models

Dependent variable y may represent a decision making or the occurrence of an event

Dependent variable y only takes two values:

- Value 1 represents taking the decision (or the occurrence of the event)
- Value 0 represents that the decision has not been taken (or the event has not occurred)

Binary discrete choice models

x represents a vector of characteristics that have an influence over the decision (or occurrence)

- Which is the probability that the variable y, conditioned on x, takes value 0 or 1?

- Which is the probability that an individual, given the characteristics in x, takes decision …?

The statistical model – A latent-variable model

- The model can be written as

$$y_i^* = \boldsymbol{x}_{\boldsymbol{i}}'\boldsymbol{\beta} + \varepsilon_i$$

Structural model

where y^*_i is a latent variable (non observed) that can be interpreted as the increase or differential utility or benefit between taking (doing) an option or another

For a single independent variable, we can simplify the notation to

$$y_i^* = \alpha + \beta x_i + \varepsilon_i$$

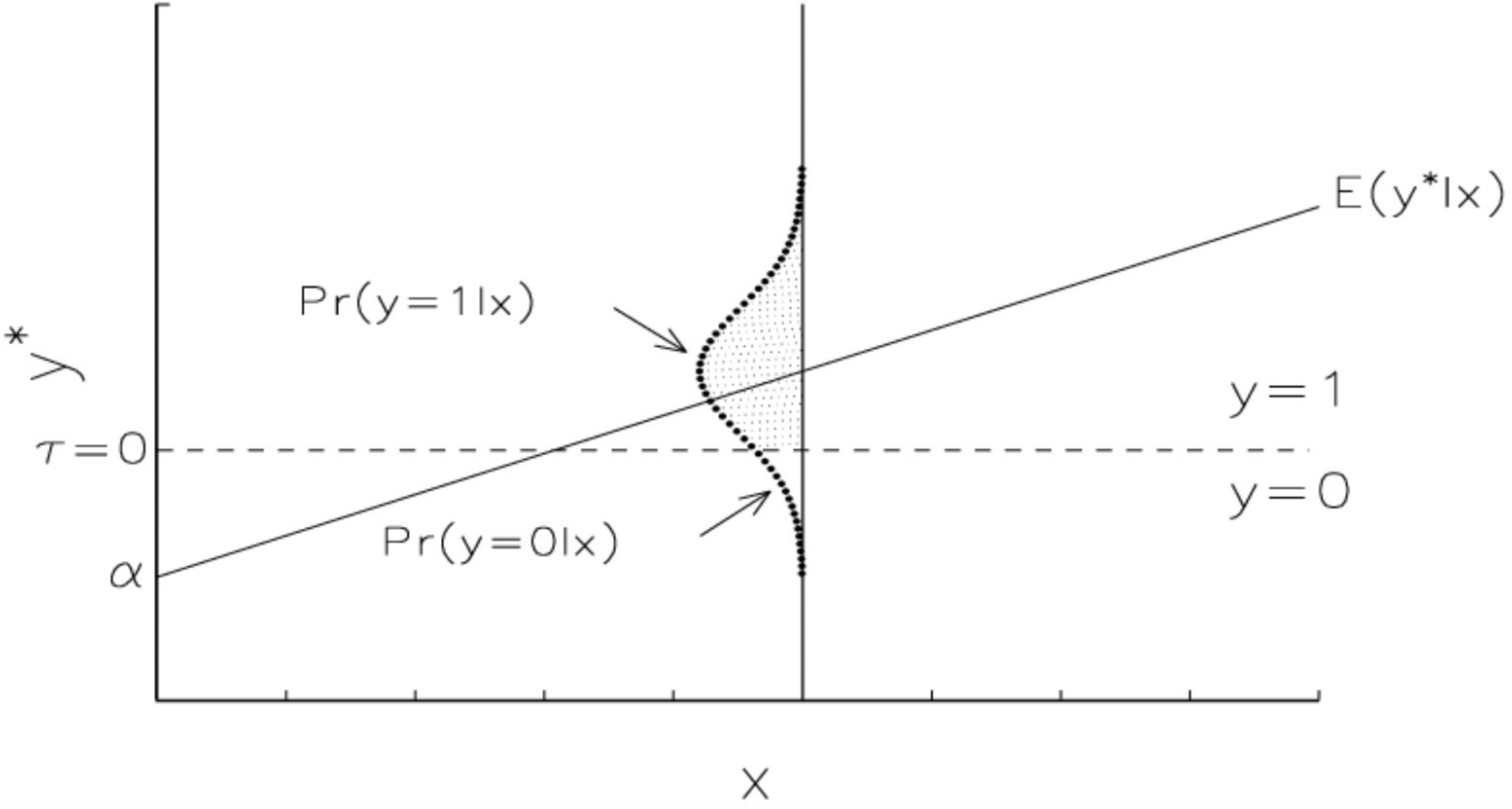
- The latent variable relates to the observed variable in the following way

$$y_i = \begin{cases} 1\ if\ y_i^* > 0 \\ 0\ if\ y_i^* \le 0 \end{cases}$$

Measurement equation

The statistical model – A latent-variable model

Relationship between latent variable y* and Pr(y=1) for the binary model



The statistical model – A latent-variable model

- For a given value of x,

$$\Pr(y = 1|x) = \Pr(y^* > 0|x)$$

- Substituting the structural model and rearranging terms,

$$\Pr(y = 1|x) = \Pr(\varepsilon > -[\alpha + \beta x]|x)$$

which shows how the probability depends on the distribution of the error term

$$\Pr(\varepsilon > -[\alpha + \beta x]|x) = \Pr(\varepsilon < [\alpha + \beta x]|x) = F(\alpha + \beta x)$$

- Two distributions of ε are commonly used, both with an assumed mean of 0:
  - If ε is assumed to be distributed logistically with Var(ε)=$^{\pi^2}/_{3}$, this lead to the **binary logit model**
  - If ε is assumed to be normal with Var(ε)=1, this lead to the **binary probit model**

The statistical model – A latent-variable model

- **Binary logit model**

$$\Pr(y = 1|x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$

- **Binary probit model**

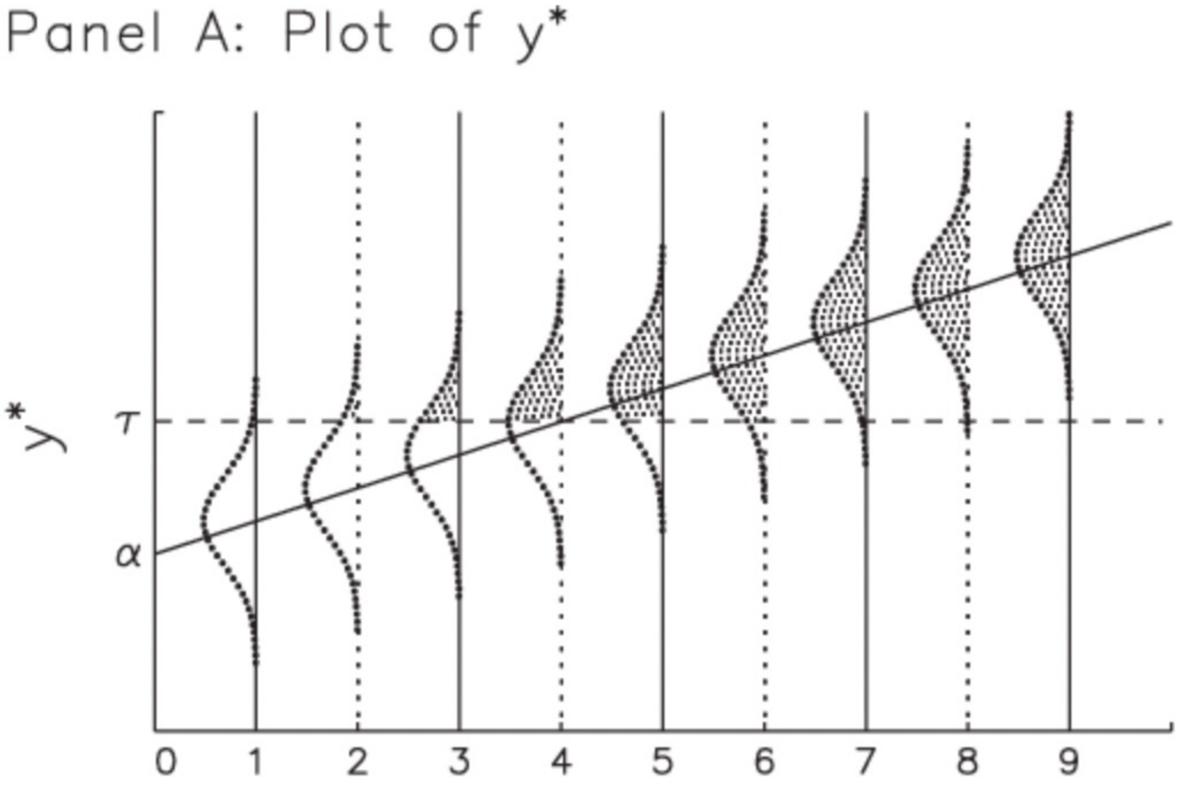
$$\Pr(y = 1|x) = \int_{-\infty}^{\alpha+\beta x} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{t^2}{2}\right) dt$$

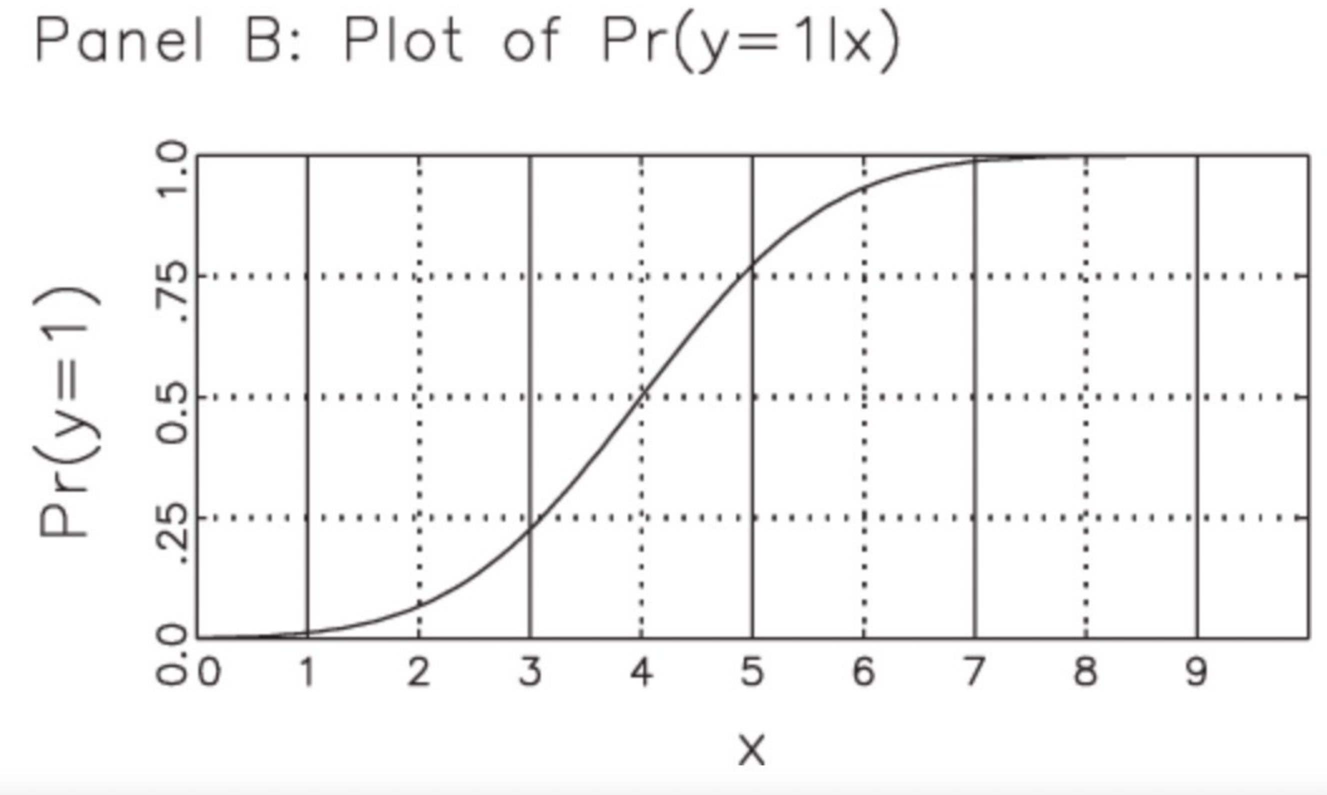
- For both logit and probit, the probability of the event conditional on x is the cumulative distribution function (CDF) of ε evaluated at $\mathbf{x}\boldsymbol{\beta}$, where F is the logistic CDF Λ for the logit model and the normal CDF Φ for the probit model

$$Pr(y = 1|x) = F(\boldsymbol{x\beta}$$

The statistical model – A latent-variable model

Relationship between the linear model $y_i^* = \alpha + \beta x_i + \varepsilon_i$ and the nonlinear probability model $\Pr(\mathrm{y} = 1|\mathrm{x}) = F(\alpha + \beta x)$





Linear probability model

The first approximation to the estimation of these models was to use OLS

The dominance in econometrics of OLS induced that F was selected as an identity function, and then

$$\text{Prob}(y_i = 1) = F(x_i'\beta) = x_i'\beta$$

So,

$$E(y_i \mid x_i) = x_i'\beta \quad \Rightarrow \quad y_i = x_i'\beta + u_i \,, \qquad E(u_i) = 0$$

Linear probability model

The linear probability model yields the usual linear regression model, with binary data for the dependent variable

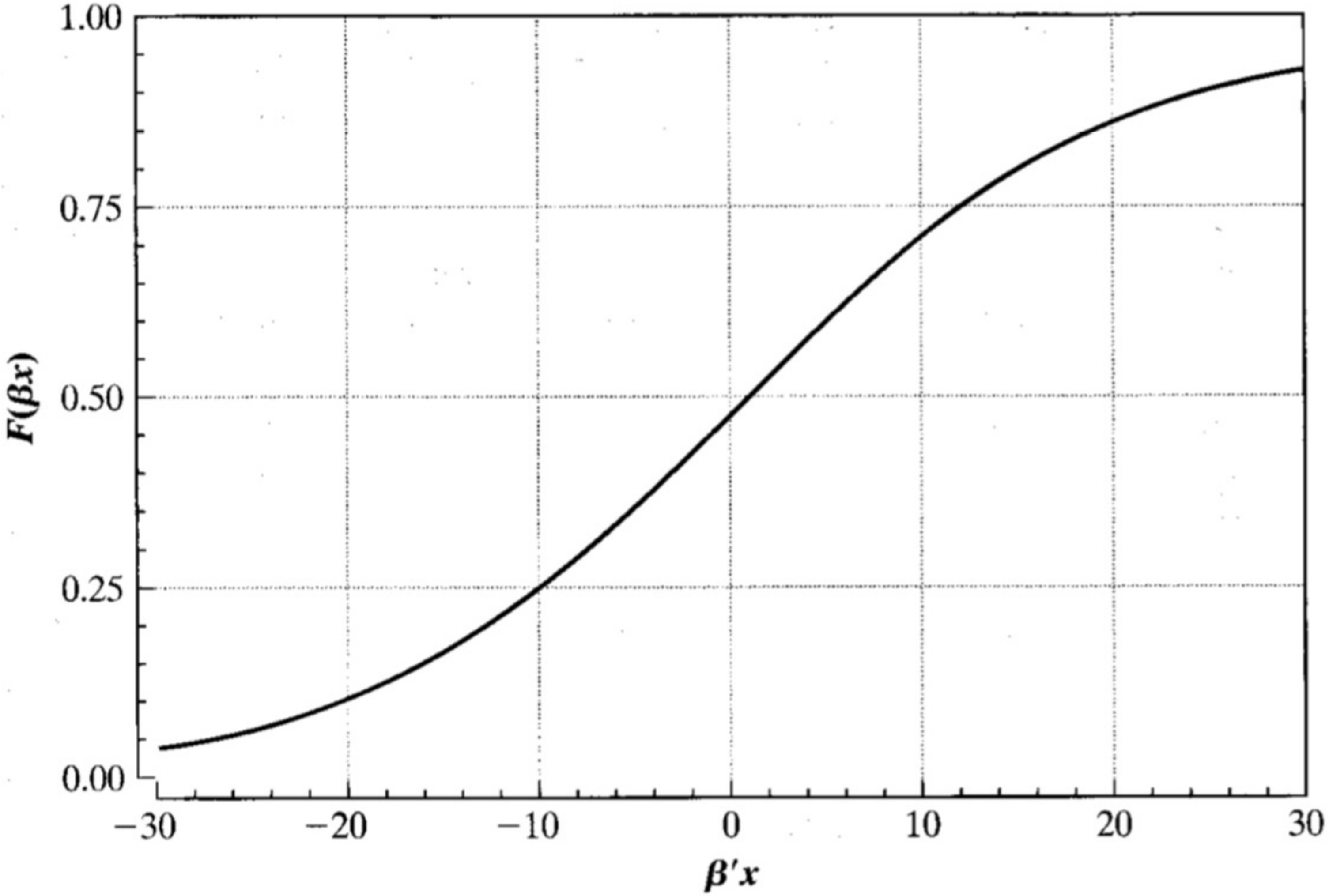
However, this approximation has some **problems**:

- The distribution for the error terms u_i does not have a normal distribution
- The errors are heteroscedastic, so OLS estimators are not efficient
- Applying OLS nothing guaranties that the probabilities fall in the [0,1] interval, which makes difficult to interpret the results
- Constant marginal effect is problematic

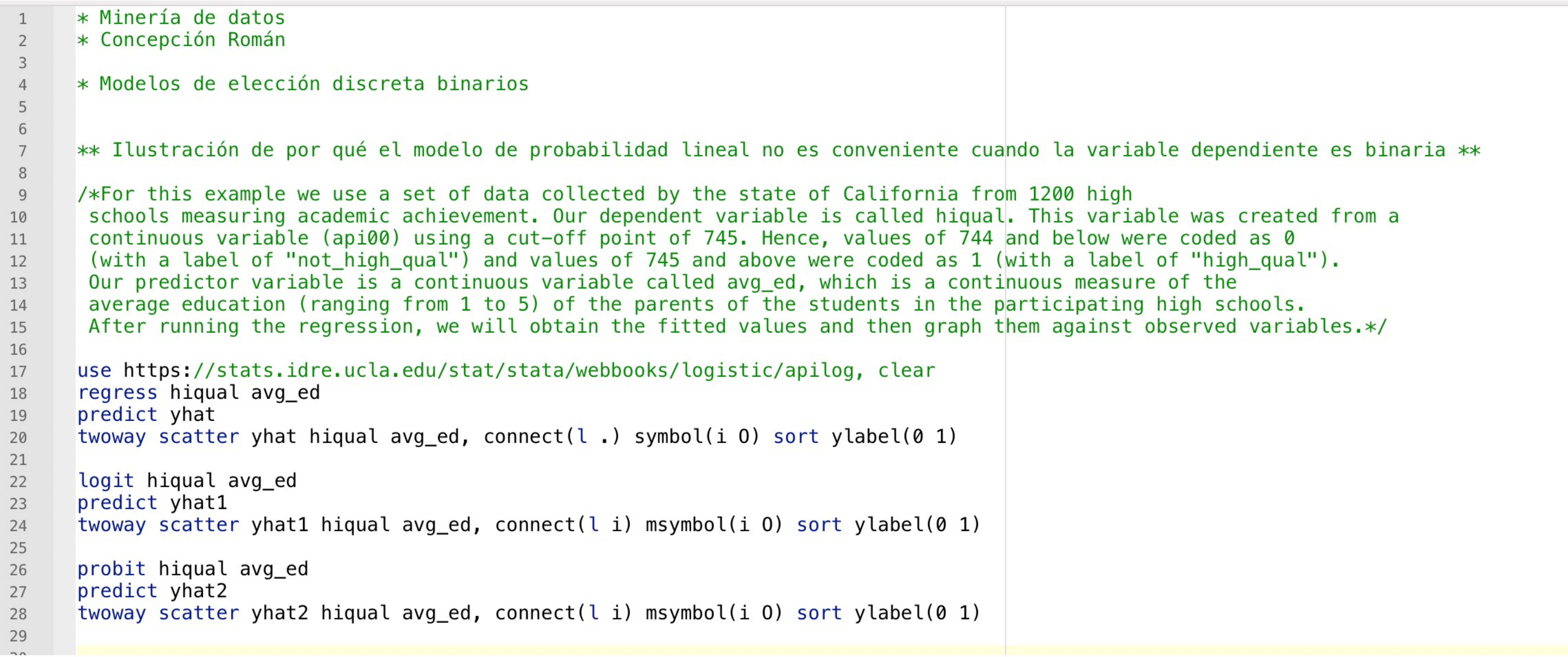
Linear probability model

- The relationship is non linear
- The dependent variable yields between 0 and 1

- Alternative models
  - Logit model (based on logistic CDF)
  - Probit model (based on normal CDF)



Linear probability model



```
* Minería de datos
* Concepción Román

* Modelos de elección discreta binarios


** Ilustración de por qué el modelo de probabilidad lineal no es conveniente cuando la variable dependiente es binaria **

/*For this example we use a set of data collected by the state of California from 1200 high
 schools measuring academic achievement. Our dependent variable is called hiqual. This variable was created from a
 continuous variable (api00) using a cut-off point of 745. Hence, values of 744 and below were coded as 0
 (with a label of "not_high_qual") and values of 745 and above were coded as 1 (with a label of "high_qual").
 Our predictor variable is a continuous variable called avg_ed, which is a continuous measure of the
 average education (ranging from 1 to 5) of the parents of the students in the participating high schools.
 After running the regression, we will obtain the fitted values and then graph them against observed variables.*/

use https://stats.idre.ucla.edu/stat/stata/webbooks/logistic/apilog, clear
regress hiqual avg_ed
predict yhat
twoway scatter yhat hiqual avg_ed, connect(l .) symbol(i O) sort ylabel(0 1)

logit hiqual avg_ed
predict yhat1
twoway scatter yhat1 hiqual avg_ed, connect(l i) msymbol(i O) sort ylabel(0 1)

probit hiqual avg_ed
predict yhat2
twoway scatter yhat2 hiqual avg_ed, connect(l i) msymbol(i O) sort ylabel(0 1)

```

Comparing linear and non-linear models

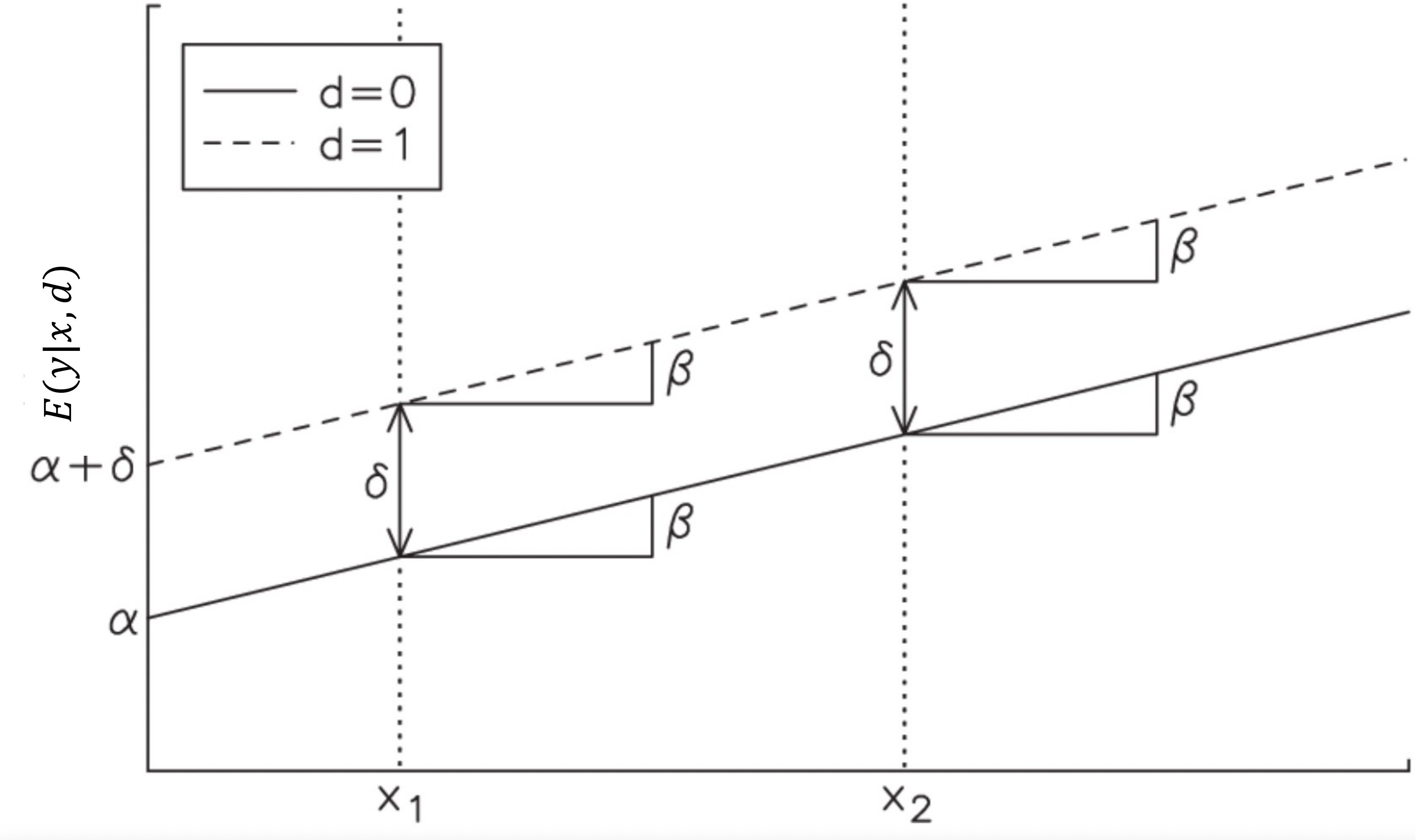
A simple linear model

Consider a linear regression model

$$y = \alpha + \beta x + \delta d + \varepsilon$$

where y is the dependent variable, x is a continuous independent variable and d is a binary independent variable

The effect of a given change in an independent variable is the same regardless of the value of that variable at the start of its change and regardless of the level of the other variables in the model



Comparing linear and non-linear models

Non linear models

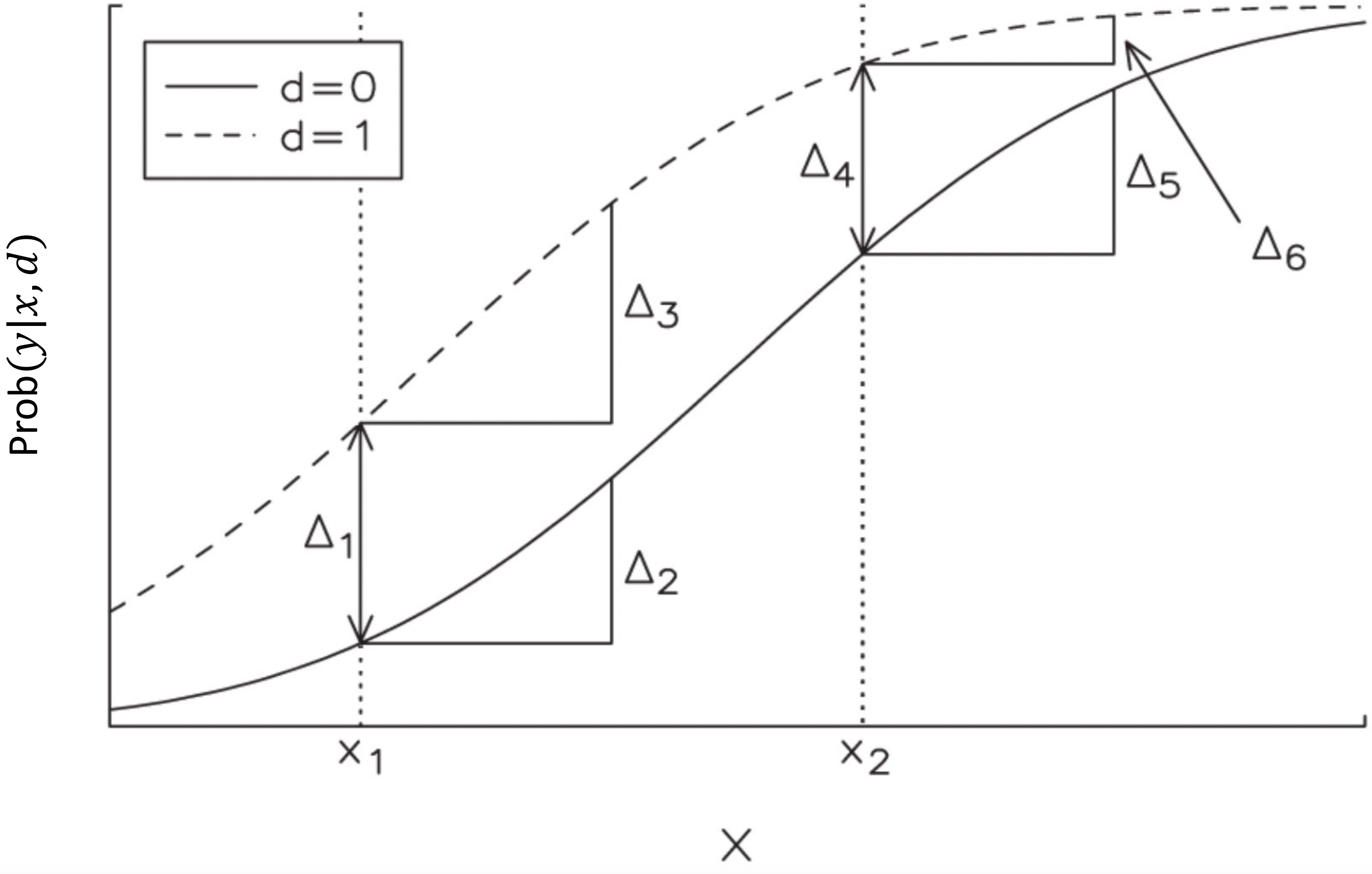
Using a logit model to illustrate the idea of nonlinearity

$$\Pr(y = 1|x, d) = \frac{\exp(\alpha + \beta x + \delta d)}{1 + \exp(\alpha + \beta x + \delta d)}$$

where x is a continuous independent variable and d is a binary independent variable

The nonlinearity of the model makes more difficult to interpret the effects of x and d on the probability of y occurring

In nonlinear models, the effect of a change in a variable depends on the values of all variables in the model and is no longer simply equal to a parameter of the model.



The *logit* and *probit* models

- ***Logit* model**

$$\Pr(y_i = 1|x) = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})}$$

- ***Probit* model**

$$\Pr(y_i = 1|x) = \int_{-\infty}^{\boldsymbol{x}_i'\boldsymbol{\beta}} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{t^2}{2}\right) dt$$

- These models are non linear in the parameters
- The normal distribution and the logistic distributions are similar, and only differ in the extreme values in the tails

The *logit* and *probit* models

- Estimation of the logit and probit models is carried out using the **maximum likelihood method**

- What is a likelihood function? A likelihood function or joint distribution function for the variables y_i given the x_i is:
  - the joint density function for the variables y_i given the variables x_i if the variable y_i/x_i is a continuous variable
  - the joint probability function for the y_i given the variables x_i if the variable y_i/x_i is discrete

The *logit* and *probit* models

- Estimation of the logit and probit models is carried out using the **maximum likelihood method**

  (Caution with small -lower than 100 obs.- sample sizes!!!)

- What is a likelihood function? A likelihood function or joint distribution function for the variables y_i given the x_i is:
  - the joint density function for the variables y_i given the variables x_i if the variable y_i/x_i is a continuous variable
  - the joint probability function for the y_i given the variables x_i if the variable y_i/x_i is discrete

The *logit* and *probit* models

- The joint probability function or likelihood function, assuming independent observations, is:

$$L = \text{Prob}(Y_1 = y_1, Y_2 = y_2, \dots, Y_N = y_N) =$$
$$\text{Prob}(Y_1 = y_1)\cdot\text{Prob}(Y_2 = y_2)\cdot\ldots\cdot\text{Prob}(Y_N = y_N) =$$
$$\prod_{y_i=0}\left[1 - F(x_i'\beta)\right]\cdot\prod_{y_i=1}\left[F(x_i'\beta)\right]$$

- Taking *logs* from the above expression:

$$\log L = \sum_{y_i}\left\{y_i \log\left[F(x_i'\beta)\right] + (1 - y_i)\log\left[1 - F(x_i'\beta)\right]\right\}$$

The *logit* and *probit* models

- The criterion for estimation: β that maximises the likelihood function, i.e., that better reproduces our sample:

$$\underset{\beta}{Max} \quad \log L \rightarrow \beta_{ML} \rightarrow \frac{\partial \log L}{\partial \beta} = 0$$

- β_{ML} maximises the probability of observing our particular sample, assuming that a specific distribution has generated our data

Interpretation of the coefficients

- In the *probit* and *logit* models the **marginal effects** (that measure the effect of a change in an explanatory variable on the dependent variable) do not coincide with the estimation coefficients

- To calculate the marginal effects we have to calculate:

$$\frac{\partial P(y_i = 1)}{\partial x_j} = \frac{\partial F(x_i'\beta)}{\partial x_j} = \frac{\partial F(x_i'\beta)}{\partial x_i'\beta}\frac{\partial x_i'\beta}{\partial x_j} = f(x_i'\beta)\beta_j$$

- In the case of the *probit* and *logit* models we can only interpret the sign of the coefficient as this coincides with the sign of the marginal effect

Methods of interpretation

- Based on predictions
  - Predictions for each observation
  - Predictions at specified values
  - Marginal effects
    - Marginal effect at the mean (MEM)
    - Marginal effect at representative values (MER)
    - Average marginal effect (AME)
  - Graphs of predictions
- Using parameters
  - Odd ratios

The *logit* and *probit* models

Comparing logit and probit

- The coefficients from logit and probit models are not directly comparable as the underlying distributions for the error terms are different

- Following Amemiya (1981) we see that the relationship between coefficient from logit and the same coefficient from probit is around 1.7

- The estimated logit coefficients are about 1.7 times larger than the probit estimates. This illustrates how the magnitudes of the coefficients are affected by the assumed

- Values of the z tests for logit and probit are quite similar because they are not affected by the assumed Var(ε), but they are not exactly the same because the models assume different distributions of the errors

Goodness of fit

- Information criteria
  - AIC (Akaike's information criterion)

$$AIC = -2logL + 2P$$

where *logL* is the maximum value for the likelihood function and P is the number of parameters in the model

  - BIC (Bayesian Information criterion)

$$\text{BIC}= -2logL + dflogN$$

where *logL* is the maximum value for the likelihood function and df is the number of parameters in the model

The smaller AIC or BIC, the better the fit

Goodness of fit

- Pseudo R^2 = $1 - \frac{\log L}{\log L_{\text{Restricted}}}$

Where *logL* is the maximum value for the likelihood function and *logL*$_{Restricted}$ is the likelihood function calculated with a constant only

Higher values implied better fit

- 2x2 hits and misses table:

	Observed values for *y*		
Predicted values for *y*	1	0	
1	# Hits	# Misses	# 1 predicted
0	# Misses	# Hits	# 0 predicted
Total	# 1 observed	# 0 observed	

Stata commands

```
logit depvar [indepvars] [if] [in] [weight] [, options]

probit depvar [indepvars] [if] [in] [weight] [, options]

margins [marginlist] [if] [in] [weight] [, response_options options]

marginsplot [, options]
```

Multivariate discrete choice models

Multivariate discrete choice models

Discrete choice model in which the dependent variable take more than 2 values

Examples:

- The selection of a mode of transport (train, car, bus)
- Wage survey. The salary of an individual can be collected in intervals (observed thresholds)
- Marketing survey about a new product (unobserved thresholds)
- …

Multivariate non-ordered discrete choice models

The type of explanatory variables we have to explain the probabilities will eventually determine the type of multinomial model we can estimate:

- If we have both attributes of the choices and characteristics of the individuals as explanatory variables, then we can estimate a *conditional logit*
- If we only have characteristics of the individuals making the choices (which are constant across alternatives), then the only model we can estimate is the *multinomial logit*

Multinomial non-ordered models

x represents a vector of characteristics that have an influence over the decision (or occurrence)

The explanatory variables are individual characteristics

Unordered-choice models can be motivated by a random utility model

For the i_{th} individual faced with J choices (J >2), let's suppose that the unobserved utility of choice j, y_{ij}^* is:

Multinomial non-ordered models

- The model can be written as

$$y_{ij}^{*} = x_i'\beta_j + u_{ij}$$

where y_{ij}^{*} is a latent variable (non observed) that can be interpreted as the increase or differential utility or benefit between taking (doing) an option or another

- If the individual makes choice j in particular, we assume that is the maximum among the J utilities

Multinomial non-ordered models

- The latent variable relates to the observed variable in the following way:

$$y_i = j \quad if \quad y_{ij}^* > y_{ik}^* \quad \forall j \neq k$$

- Hence, the statistical model is driven by the probability that choice j is made, which is:

$$Prob(y_i = j) = Prob(y_{ij}^* > y_{ik}^*) \quad \forall j \neq k$$

The model is made operational by a particular choice of distribution for the disturbances

Multinomial non-ordered models

- The model is made operational by a particular choice of distribution for the disturbances

- As in the binary case, two models could be considered, *logit* or *probit*

- However, because of the need to evaluate multiple integrals of the normal distribution, the probit model is computationally more intensive

- Then, in general, it is used the *logit* specification

Multinomial *logit* model

- Following McFadden (1973), if (and only if) the J disturbances are independent and identically distributed with Weibull distribution, we have:

$$Prob(y_i = j) = Prob\left(y_{ij}^* > y_{ik}^*\right) = \frac{e^{x'\beta_j}}{\sum_{k=1}^{J} e^{x'\beta_k}} \quad j = 1, \dots, J$$

- The estimated equations provide a set of probabilities for the J choices

Multinomial *logit* model

- Indeterminacy problem

$$\frac{e^{x'(\beta_j+q)}}{\sum_{k=1}^{J} e^{x'(\beta_k+q)}} = \frac{e^{x'q} \cdot e^{x'\beta_j}}{e^{x'q} \sum_{k=1}^{J} e^{x'\beta_k}} = \frac{e^{x'\beta_j}}{\sum_{k=1}^{J} e^{x'\beta_k}}$$

- A convenient normalization that solves the problem of indeterminacy is to assume that one set of parameters equals zero:

$$Prob(y_i = j) = \frac{e^{x'\beta_j}}{\sum_{k=1}^{J} e^{x'\beta_k}} \quad j = 2, \ldots, J$$

$$Prob(y_i = 1) = \frac{1}{1 + \sum_{k=2}^{J} e^{x'\beta_k}} \quad j = 1$$

Multinomial *logit* model

- Estimation of the multinomial logit model is carried out using the **maximum likelihood method**

- Assuming that the alternative are independent, the likelihood function is:

$$L = Prob(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n) = \prod_{i} \prod_{j=1}^{J} [Prob(y_i = j)]^{d_{ij}}$$

where d_{ij}=1 if the alternative chosen by individual *i* is *j*, and 0 otherwise

Multinomial *logit* model

- Taking *logs* from the above expression:

$$\log L = \sum_{i} \sum_{j=1}^{J} d_{ij} \cdot Prob(y_i = j)$$

- The criterion for estimation this function is finding $\beta=(\beta_1, \beta_2, \ldots, \beta_J)$ that maximises the likelihood function, i.e., that better reproduces our sample.

$$\underset{\beta}{Max} \quad \log L \rightarrow \beta_{ML} \rightarrow \frac{\partial \log L}{\partial \beta} = 0$$

- β_{ML} maximises the probability of observing our particular sample, assuming that a specific distribution has generated our data

Interpretation of the coefficients

- Depending on the alternative we take for the normalization we will get different coefficients estimated but irrespectively of this, the marginal effects, probabilities estimated and the likelihood function maximum will be the same.

- We cannot even interpret the sign of the coefficients as they do not usually coincide with the sign of the **marginal effects**. So we need to calculate the marginal effects to interpret the results of the parameters estimated. The marginal effects of the multinomial logit are:

$$\frac{\partial Prob(y_i = j)}{\partial x_i} = P_j \left[\beta_j - \sum_{k=1}^{J} P_k \, \beta_j \right]$$

The independence of irrelevant alternatives assumption (IIA)

- The IIA comes from the assumption of alternative independent disturbances
- This is a convenient assumption for estimation but it is not a particularly appealing restriction in some real life examples
- This property makes that the ratio of the probabilities of two alternatives are independent of all the other alternatives

Introducción a la Economía Cuantitativa y al análisis de datos con STATA 18

Doctorado en Economía, Empresa, Finanzas y Computación

Universidad de Huelva

Junio-Julio 2024

Mónica Carmona (UHU y CCTH)

Emilio Congregado (UHU y CCTH)

Concepción Román (UHU y CCTH)