LUISS BUSINESS SCHOOL

EFMD EQUIS ACCREDITED

Specialised Master Programmes

Econometrics

Econometrics

Giuseppe Brandi

LUISS

June 2018

Agenda

Limited Dependent Variable Models

- Binary Dependent Variables
  - Linear Probability Model
  - Logit and Probit
- Categorical Dependent Variables
  - Multinomial Logit Model
  - Ordinal Logit Model

Binary Dependent Variable Models - Theory

Binary Dependent Variables

Examples:

- Does the client buy our product?
- Does the bank accept your mortgage request?
- Does a patient heal?
- Do you pass the Econometrics exam?

What are we modelling?

We want to estimate the probability that our dependent variable Y (the response) takes values 0 or 1 given a set of independent variables:

$$Pr(\text{client buying the product} = \text{"true"} | X)$$

where X contains variables such as gender, income...

$Pr(Y = 1)$ is called "probability of success", independently of the context.

Three ways

- Linear Probability Model
- Probit
- Logit

Linear Probability Model

Consider the vector form model:

$$Pr(Y = 1|X) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_K x_K$$

- How do we estimate this model?
- How do we interpret the coefficients?

LPM - Estimation

We use a simple OLS:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

In R :

lm($y \sim x_1 + x_2 + x_3 + ... + x_K$, data $=$ *mydata*)

NB: with binary dependent variables, the model is always heteroskedastic. Remember to use robust estimators for your -standard errors.

LPM - Interpretation

Focusing on x_j while keeping all other variables fixed,

$$\beta_j = \frac{\partial Pr(Y = 1|X)}{\partial x_j}$$

- if x_j is continous : β_j is the change in the probability of success ($Y = 1$) for a 1-unit increment in x_j;
- if x_j is binary : β_j is the change in the probability of success when $x_j = 1$ w.r.t. $x_j = 0$.

Example

Let's use the dataset *affair* from the IDEAS website.

Data on a survey by Psychology Today, 1969, on married individuals. N=601.

Regress the probability of having an affair, **affair**, on

- whether **male** (1) or not;
- whether **kids** (1) or not;
- years of marriage , **yrsmarr**
- how religious, **relig**, 5 very much, 1 not at all. (*)

EXAMPLE CONT.ED

$$Pr(affair = 1|X) = \underset{(0.042)}{0.228} + \underset{(0.035)}{0.037} male_i + \underset{(0.046)}{0.074} kids_i$$
$$+ \underset{(0.004)}{0.010} yrsmarr_i - \underset{(0.015)}{0.063} relig_i + u_i$$

Example cont.ed

Compute the probability of having an affair for a man who has been married for 7 years, two kids, who is an average degree of religiosity (2).

Easy, right?

Now compute the probability of having an affair for a woman who has been married for 1 year, no kids, and who is very religious.

Or compute the probability of having an affair for a man who has been married for 55 years, three kids, and who is not religious at all.

LPM - Pros and Cons

- LPM is easy to estimate/interpret, but...
- The predicted probability

$$\widehat{Pr(y = 1|X)} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_K x_K$$

is not necessarily between 0 and 1!

Counter-argument!

The problem can be solved - not elegant, but effective.

- If $Pr(\widehat{y = 1}|X) > 1$, we set $\hat{y}$ to 1;
- if $Pr(\widehat{y = 1}|X) < 0$, we set $\hat{y}$ to 0;

Counter-counter argument

The effects of the covariates stay constant - they do not depend on the values of the covariates themselves, and this is problematic.

The effect of one more year of marriage on the probability of having an affair is always 1% (50% $\rightarrow$ 51%, but also 99.5% $\rightarrow$ 100.5%)

More than restricting predicted probabilities, we need meaningful effects... (the effect of one more year of marriage should decrease as we approach a probability equal to 1).

CDF

Look into your statistical toolbox...

Is there a function bounded between 0 and 1?

- There are many of them...
- Cumulative Distribution Functions.

Probit

If we assume that the errors in our regression are Normally distributed, we use:

$$G(z) = \int_{-\infty}^{\infty} \phi(\nu) d(\nu) = \Phi(z)$$

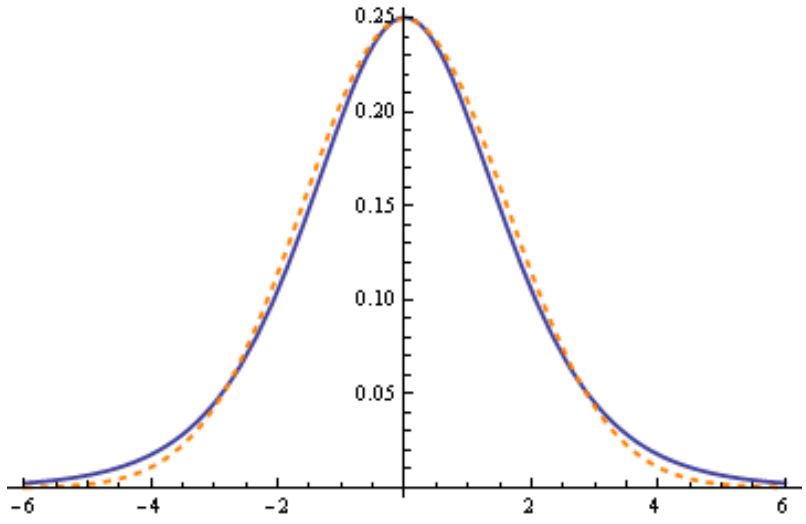
z is called $z - score$, practically our linear model.

$$z = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... + \beta_K x_{K,i} + u_i$$

Logit

If we assume that the errors in our regression are distributed as a logistic, we use:

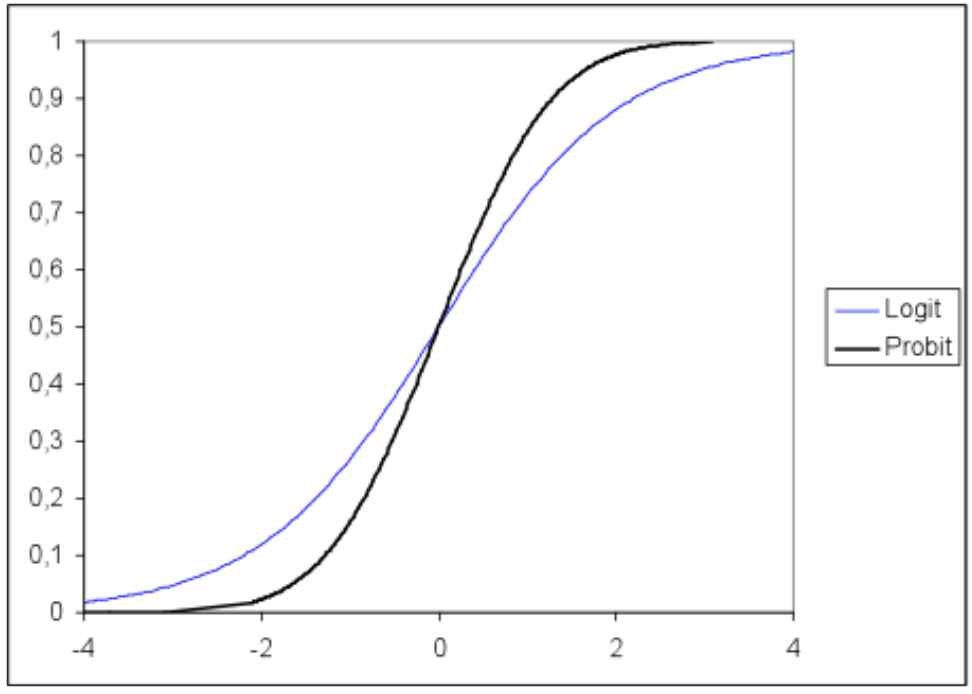
$$G(z) = \frac{e^z}{1 + e^z} = \Lambda(z)$$



Blue: logistic distribution - Orange: Normal distribution.

Probit or Logit?

- Similar results.
- Logistic distribution has fatter tails... but it's basically the same.



Formally

For all Binary Dependent Models:

$$\begin{aligned} Pr(y_i = 1|X_i) &= G(\beta X_i) \\ &= G(\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... + \beta_K x_{K,i} + u_i) \end{aligned}$$

where

- the subscript i identifies a single individual;
- X_i is the row of a matrix containing K regressors we consider useful in order to explain pur dependent variable y_i;
  - the first column of X_i is full of ones, in order to have a constant in the model;
- β is the vector of coefficients;
- $G(.)$ is a function, usually called *link function*.

LINK FUNCTIONS

- Linear Probability Model
  Link function: $G(x) = x$ (identity function $\rightarrow$ linear model):

$$Pr(y_i = 1|X_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... + \beta_K x_{K,i}$$

- Probit
  Link function:

$$G(z) = \int_{-\infty}^{\infty} \phi(\nu) d(\nu) = \Phi(z)$$

  where $\phi(.)$ is the standard Normal density $\rightarrow$ $\Phi(.)$ is the standard Normal cdf.

- Logit
  Link function:

$$G(z) = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}}$$

Logit & Probit - Estimation

- Complicated, non linear link functions. How do we estimate β?
- Maximum Likelihood Estimation.
- MLE tries to find estimates of parameters that make the data actually observed "most likely."
- No need to write a whole new code - pre-existing functions are there for us.

$$glm(y \sim x_1 + x_2 + x_3 + ... + x_K, family = binomial(link = cdf))$$

- *cdf* is 'logit' or 'probit'.
- We obtain $\widehat{Pr(y = 1|X)} \in [0, 1]$.

Logit & Probit - Interpretation

Focusing on x_j while keeping all other variables fixed,

- if x_j is continous:

$$\frac{\partial Pr(Y=1|X)}{\partial x_j} = \frac{dG(z)}{dz}(z)\beta_j$$

  Since $\frac{dG(z)}{dz}(z) > 0$ always, for both P&L, the direction of the effect of a change in x_j is given by the sign of β_j.

- if x_j is binary: the effect of x_j changing from 0 to 1 is:

$$G(\beta_0+\beta_1 x_1+...+\beta_j*1+...+\beta_K x_K)-G(\beta_0+\beta_1 x_1+...+\beta_j*0+...+\beta_K x_K)$$

Example

Probit model: what information can you extract from this output?

$$Pr(affair = 1|X) = \Phi(\underset{(0.146)}{-0.817} + \underset{(0.115)}{0.124} male_i + \underset{(0.167)}{0.285} kids_i + \underset{(0.013)}{0.033} yrsmarr_i - \underset{(0.052)}{0.205} relig_i)$$

Example

Logit model: what information can you extract from this output?

$$Pr(affair = 1|X) = \Lambda(\underset{(0.257)}{-1.350} + \underset{(0.196)}{0.211} male_i + \underset{(0.293)}{0.487} kids_i + \underset{(0.022)}{0.057} yrsmarr_i - \underset{(0.088)}{0.350} relig_i)$$

Logit & Probit - Interpretation cont.ed

Looking at the 'usual' R output, for a given variable we can immediately tell only:

- whether its effect is positive or negative;
- whether its effect is significant or not.

The magnitude of the effect cannot be seen directly.

Logit & Probit - Interpretation cont.ed

It is hard to interpret the coefficient for P&L.

The magnitude of the effect of a variable depends on the values of all the regressor for each individual.

Let's see this on R. Find how gender affects the probability of having an affair for someone without kids and anti-religion, married for a) 10 years or 25) years.

Logit & Probit - Marginal effects

Marginal effects can be an informative means for summarizing how change in the probability of success is related to marginal (very small) change in a (continous) covariate.

- Marginal Effect of Means: the marginal effect is computed for a representative individual whose regressors are set at the average sample level.

$$G'(\beta \bar{X})\beta_j$$

- Average Marginal Effect: the marginal effect is evaluated for each individual and then the average is computed on the whole sample.

$$\frac{\sum G'(\beta X)}{n}\beta_j$$

LOGIT & PROBIT - MARGINAL EFFECTS CONT.ED

For discrete regressors, thinking "marginally" does not make sense.

- It does not make sense to study a, say, 0.1 change in *kids*.

In these cases, R report as marginal effects how the probability of success changes as the binary variable changes from 0 to 1, holding all other variables at their means.

If you have more than two categories, go with factors.

Logit & Probit - Relative effects

We have a direct interpretation only for two regressors' relative effect:

$$\frac{\partial Pr(Y = 1|X)/\partial x_j}{\partial Pr(Y = 1|X)/\partial x_h} = \frac{\beta_j}{\beta_h}$$

- Consider the effect of having kids with respect to the effect of being male, ceteris paribus.
- The first is about twice the latter, no matter the initial characteristics of the individual.
- The result is stable in the three models.

Odds

Odds are an expression of relative probability.
The odds for (against) some event reflect the likelihood that the event will (not) take place.

$$Odds = \frac{Pr(Y = 1|X)}{Pr(Y = 0|X)}$$

Suppose that $Pr(Y = 1|X) = 0.2$: then $Pr(Y = 0|X) = 0.8$ and $OR = \frac{0.2}{0.8} = \frac{1}{4}$ and we say that the odds of success are 1 to 4.

- The transformation from probabilities to OR is a monotonic transformation.
- Odds range from 0 to infinity.

Odds Ratio

The Odds Ratio (OR) is the ratio of the odds in case one regressor is incremented by one unit (ceteris paribus) over the odd in case that regressor is not incremented:

$$OR = \frac{Odd(Y = 1|X + 1)}{Odd(Y = 1|X)}$$

OR are constant: it does not matter what values the other regressors take on.

Odds Ratio example

Suppose we only regress whether a person has an affair or not on just one regressor, gender.

$$Odds(male) = \frac{0.27273}{1 - 0.27273} = 0.375$$

$$Odds(female) = \frac{0.22857}{1 - 0.22857} = 0.296$$

The odds that a man has an affair are about 4 to 10, while for a woman odds are about 3 to 10. The Odds Ratio is:

$$OR = \frac{Odds(male)}{Odds(female)} = \frac{0.375}{0.296} = 1.267$$

The odds of having an affair are about 1.26 times greater for males than for females.

Odds ratio cont.ed

We can also define the log of the OR:

$$lnOR = ln\frac{Odd(Y = 1|X + 1)}{Odd(Y = 1|X)}$$

- The range for the log OR is not restricted.

We like OR and log OR because it is usually difficult to model variables with restricted range.

Odds ratio with logit

Odds and OR are easier to study with a logit model. Notice that:

$$Pr(Y = 1|X) = \frac{1}{1 + e^{-z}}$$

$$Pr(Y = 0|X) = 1 - Pr(Y = 1|X) = \frac{e^{-z}}{1 + e^{-z}}$$

$$Odds = e^{z} = e^{\beta X}$$

$$OR = \frac{e^{z_1}}{e^{z_0}}$$

where z_1 is the z-score with one covariate, j, incremented by one and z_0 is the initial z-score. Then:

$$OR = e^{\beta_j}$$

$$lnOR = \beta_j$$

Odds ratio with logit

β is the marginal effect in terms of the log odds ratio.

If you increase one regressor by one unit,*ceteris paribus*, the log of the odds ratio changes by β units.

Said otherwise, if the regressor increases by one unit, the Odds Ratio increases by e^{β_j}

Goodness of fit

Percentage of correctly classified values.
We build a confusion matrix:

	Actual = 1	Actual = 0
Predicted = 1	True	False
Predicted = 0	False	True

- When *Predicted Probability* $>$ *threshold*, we set $Y = 1$ and viceversa.
- We consider the percentage of correctly predicted values:

$$\frac{\sharp\,(\text{Predicted} = 1 |\ \text{Actual} = 1) + \sharp\,(\text{Predicted} = 0 |\ \text{Actual} = 0)}{\sharp\text{Predictions}}$$

Goodness of fit

- If you are estimating causal effects, prediction is not fundamental.
- With too good in-sample prediction, out-of-sample performance will be poor.
- There will always be an "optimal" threshold to increase accuracy, and this has nothing to do with how good your model is.
- Use training and test datasets...
- You will meet this again in Machine Learning.

Pseudo R^2 - McFadden

- Pseudo $R^2 = 1 - \frac{L_{ur}}{L_r}$, or
- Adjusted pseudo $R^2 = 1 - \frac{L_{ur} - K}{L_r}$ where:
  - L_{ur}: unrestricted log-likelihood (of our model)
  - L_r: restricted log-likelihood (Y regressed only on a constant term)
- If Pseudo $R^2 = 0$ the regressors do not explain our dependent variable.
- It does not tell us the percentage of the variance of the dependent variable explained by the regressors.

Binary Dependent Variable Models - Practice

Dataset

Aspects of daily life, ISTAT

http://www.istat.it/it/archivio/129956

Download data for 2014. In "Leggimi" you find the data description.

We are interested in knowing whether a person has an E-book reader or not depending on some demographic characteristics. We restrict our sample to adults (>=18 years old).

VARIABLES

Identify and use the following variables in your analysis - but pay attention to their description.

- Gender
- Age
- Civil status
- Area of residence
- Education level
- Occupational status

Moreover, build "DENS": number of rooms divided by number of family components, as a proxy for wealth.

VARIABLES REVEALED

- Gender - **SESSO**
- Age - **ETAMi**
- Civil status - **STCIVM**
- Area of residence - **RIPMif**
- Education level - **ISTRMi**
- Occupational status - **CONDMi**
- Proxy for wealth - **DENS**

Questions I

Descriptive analysis

1. What is the age category with the highest percentage of E-book ownership?
2. What about geographical areas?
3. Present some preliminary (descriptive) evidence that the education level and E-book ownership are linked.

Questions II

Let's use a LPM and a Logit model.

1. Do you confirm the previous finding?
2. What is the effect of gender? Is it statistically significant?
3. Predict the probability that a single woman aged 30, living in the Center of Italy, living by herself in a 3 rooms house, working and with a college degree, has an E-book reader?
4. For this individual, find the effect of moving to the Nort-East on the probability of having an E-book reader.
5. What are the odds that this woman has an E-book reader?
6. What can you say about the relative effect of being married and gender?
7. What are the odds ratio of having an E-book reader for a change in the age category?

Prediction

Suppose you are interested in predicting the ownership on an E-book reader. What variables would you include in your analysis?