REGRESSION ANALYISIS WITH BINARY DEPENDENT VARIABLE

14 September 2011

Yogyakarta, Indonesia

Cosimo Beverelli

(World Trade Organization)

NOT CONTINUOUS DEP. VAR.

- In many applications the dependent variable is not continuous but qualitative, discrete or mixed:
  - Qualitative: car ownership (Y/N).
  - Discrete: education degree (Ph.D.,…, no education).
  - Mixed: hours worked per day.
- For such dependent variables, standard linear model are not appropriate.
- In the next slides we will focus on the case of a binary dependent variable

BINARY DEPENDENT VARIABLE

- Let y_i be a DEPENDENT DUMMY VARIABLE explaining whether family *i (i = 1, 2,…, N)* owns a car (y_i = 1) or not (y_i = 0).

- We suppose car ownership to be a function of family income (our exogenous/independent variable) x_i.

- We now want to model car ownership by means of a linear regression model:

$$y_i = \beta x_i + u_i \qquad i=1,2,\dots.n$$

A PROBABILITY

- NOTICE THAT

1. If $E(u_i|x_i) = 0 \;\Rightarrow\; E(y_i|x_i) = \hat{\beta}x_i$
2. Being y_i a binary dependent variable we have that:
   $E(y_i|x_i) = 1 \bullet \text{prob}(y_i = 1|x_i) + 0 \bullet \text{prob}(y_i = 0|x_i)$

- Therefore it is immediate to obtain: $\text{prob}(y_i = 1|x_i) = \hat{\beta}x_i$

- The problem is that there is no guarantee that $0 \leq \widehat{\beta}x_i \leq 1$, while it should be (being a probability)

ERRORS DISTRIBUTION

- In the discussion on the linear regression model, we assumed that errors were normally distributed.
- In case of binary dependent variable, u_i is highly non normal; we have:

$$u_i = \begin{cases} 1 - \hat{\beta} x_i & \text{if } y_i = 1 \\ -\hat{\beta} x_i & \text{if } y_i = 0 \end{cases}$$

- Therefore the distribution of errors for a given independent variable has a 2 mass points instead of a normal distribution!

HETEROSKEDASTICITY

- In the discussion on the linear regression model, we assumed that errors were normally distributed, having a constant variance.
- It is possible to show that in case of binary dependent variable:

$$E({u_i}^2) = (1 - \hat{\beta}x_i)\hat{\beta}x_i$$

- It depends upon the independent variable and/or the coefficient $\Longrightarrow$ there if heteroskedasticity in the model.
- If the model is heteroskedastic, biased standard errors lead to biased inference, so results of hypothesis tests are possibly wrong.

BINARY CHOICE MODELS

- Binary choice model are designed to overcome these issue.
- They postulate that:

$$\text{prob}(y_i = 1|x_i) = G\ (x_i, \beta),$$

  where G is a generic function that takes values in [0,1].
- Thus, the probability of the event depends upon personal attributes (independent variables) and coefficients.
- And usually attention is restricted to functions of the form:

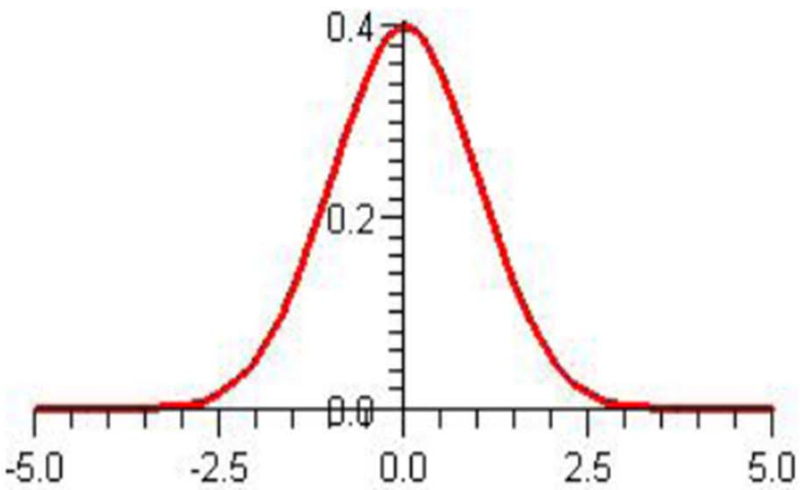
$$G(x_i, \beta) = F(\hat{\beta} x_i) \in (0,1)$$

- Under this restriction, independent variables and coefficients enter the function as linear combinations.

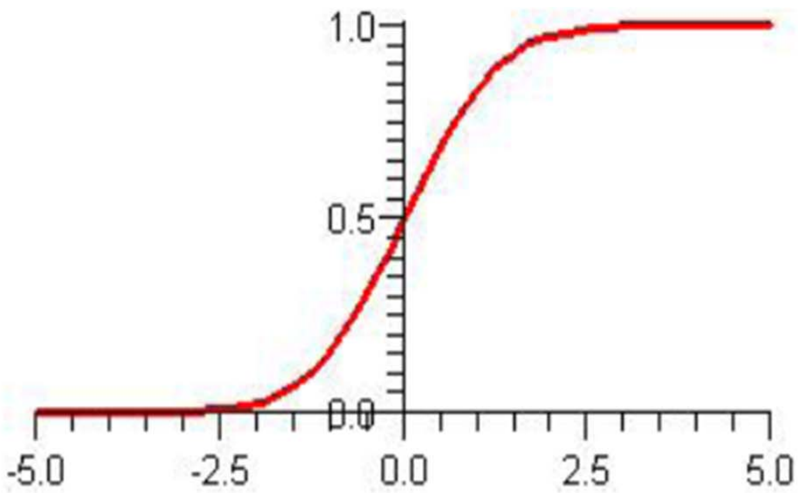
PROBIT MODEL

- The PROBIT MODEL uses the standard normal distribution (with mean 0 and variance 1), whose cumulative distribution function (Φ) is:

$$F(w) = \int_{-\infty}^{w} \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}t^2\right\} dt \equiv \Phi(w)$$



PDF of the Standard Normal Distribution

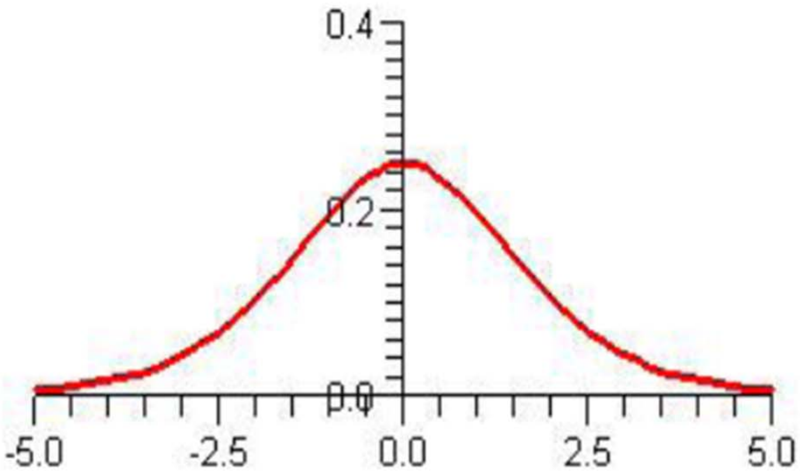


CDF of the Standard Normal Distribution

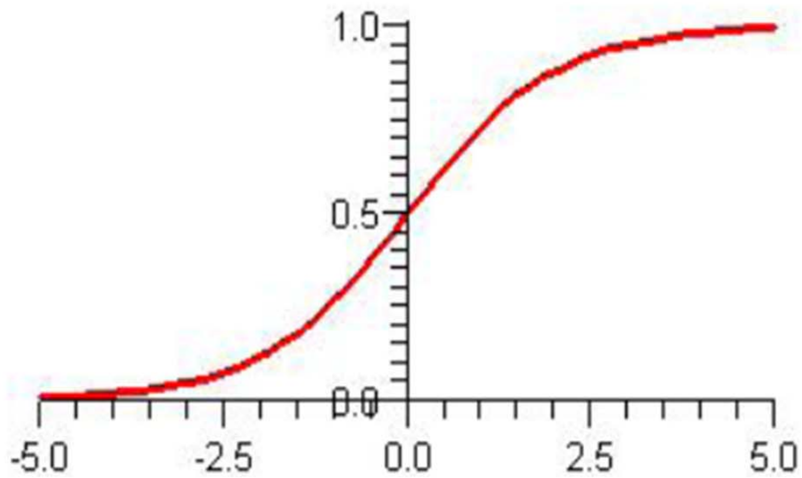
LOGIT MODEL

- The LOGIT MODEL uses the standard logistic distribution (with mean 0 and variance $\pi^2/3$), whose cumulative distribution function *(L)* is:

$$F(w) = \frac{e^w}{1+e^w} \equiv L(w)$$



PDF of the Standard Logistic Distribution



CDF of the Standard Logistic Distribution

LINEAR PROBABILITY MODEL

- The LINEAR PROBABILITY MODEL assumes that there is an uniform distribution on [0,1], therefore:

$$F(w) = \begin{cases} 0 \; if \; w < 0 \\ w \; if \; w \in [0,1] \\ 1 \; if \; w > 1 \end{cases}$$

MARGINAL EFFECT

- Binary dependent variable models yields a set of coefficients $\hat{\beta}$ that parametrize the impact of exogenous variables on the endogenous one, thus providing info on sign and significance.
- Due to the difficulties in the interpretation, it is common practice to evaluate the "marginal effects"; that it, to evaluate the change in the predicted probability induced by a small change in the exogenous variable.
- Given: prob$(y_i = 1|x_i) = F(\hat{\beta}x_i) \Longrightarrow$

$$ME_k = \frac{\delta prob(y_i=1|x_i)}{\delta x_k}$$

Marginal effect of small change in exogenous variable k.

MARGINAL EFFECTS

- PROBIT MODEL

$$ME_k = \phi(\hat{\beta}x_i)\beta_k$$

Standard normal density function

- LOGIT MODEL

$$ME_k = \frac{e^{\hat{\beta}x_i}}{1 + e^{\hat{\beta}x_i}}\beta_k$$

- LINEAR PROBABILITY MODEL

$$ME_k = \begin{cases} \beta_k \; if \; \hat{\beta}x_i \in [0,1] \\ 0 \; otherwise \end{cases}$$

HINT FOR IN-DEPTH EXAMINATIONS

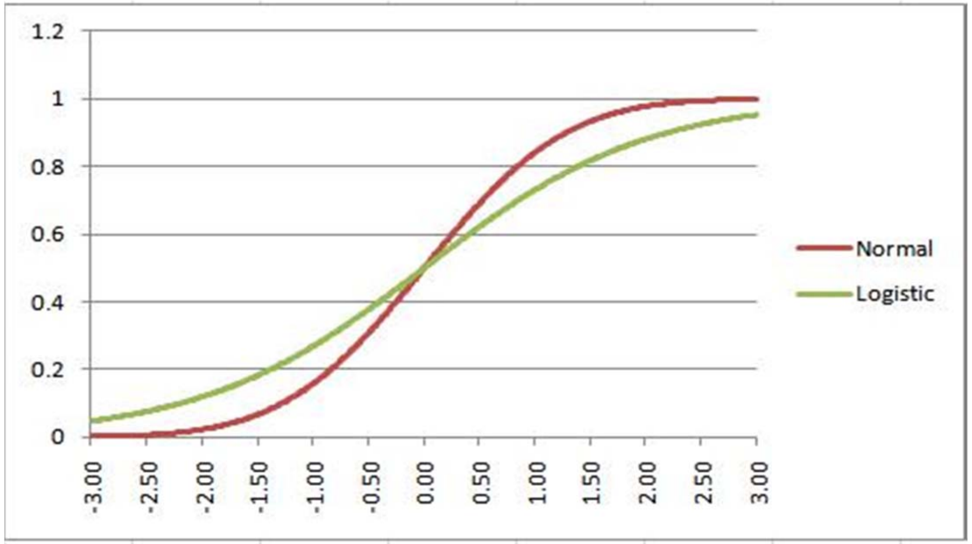
- The estimation of the models is done by Maximum Likelihood Method.
  - Having specified the distribution of errors.
  - Taking logs of the cumulative distribution functions.

$\Longrightarrow$

- Goodness of fit.
  - Pseudo R-squared.

PROBIT /LOGIT: DIFFERENCES

- Cumulative distribution functions of normal and logistic distributions



- Is straightforward to notice that the logistic distribution has lower peak and “fatter” tails with respect to the normal distribution.

PROBIT /LOGIT: DIFFERENCES

- The two models, of course, produce different parameter estimates.
- In binary response models, the estimates of a Logit model are roughly $\pi/\sqrt{3}$ times larger than those of the Probit model.
- These estimators, however, end up with almost the same standardized impacts of independent variables [J. Scott Long, Regression Models for Categorial and Limited Dependent Variables, 1997].
- The choice between Logit and Probit models is more closely related to estimation and familiarity than to theoretical or interpretive aspects.

PROBIT /LOGIT: PANEL DATASET

- Suppose now we are working on a panel dataset.

- Suppose the model is:

$$\text{prob}(y_{it} = 1|x_{it}) = G\ (x_{it}, \beta) \qquad t = 1,2, \dots T$$

- x_{it} can contain a variety of factors: time dummies, their interactions with other variables and lagged dependent variables.

PROBIT (PANEL DATASET)

- Let us consider the unobserved fixed probit model:

$$\text{prob}(y_{it} = 1|x_i, \alpha_i) = \text{prob}(y_{it} = 1|x_{it}, \alpha_i) = \Phi(x_{it}\beta + \alpha_i)$$
$$t = 1,2, \ldots T$$

- Where α_i is the unobserved effect.
- In a fixed effect probit analysis α_i is a parameters to be estimated along with the coefficient, but there is an incidental parameters problem; that is, estimating α_i (N of them) along with β leads to inconsistent estimators of the coefficient itself if T is finite and N→∞ (this problem disappears as T→∞).

LOGIT (PANEL DATASET)

- We can get over this problem by using the fixed effect logit estimator (conditional logit estimator), that makes α_i "vanish" by assuming that the distribution of y_{it} conditional on x_i, α_i does not depend on α_i.
- Notice that α_i as not treated as parameters to be estimated along with β.
- The "escamotge" for the estimate to be consistent is that the identification uses only the individuals who change state.

COMPUTATION

- Stata commands are `logit` (`logistic` with odds ratio and no constant) and `probit`, respectively.
- In Stata it is also possible to highlight the marginal effect of exogenous variables (`dlogit2`, `dprobit2`)
- In the following slides, we illustrate an example [from Park, Hun Myoung, *Regression Models for Binary Dependent Variables Using Stata, SAS, R, LIMDEP, and SPSS*, working paper Indiana University, 2009]
  - y = probability of trusting people
  - x_1 = years of education
  - x_2 = income
  - x_3 = age
  - x_4 = being a male
  - x_5 = internet user

LOGIT – regression (odds ratio)

Pseudo R-square (equivalent of R-square) shows the amount of variance of y explained by x.

Equivalent of P-value of the model. If Prob>chi2 < 0.05 → the model fits the data very well

```
. logistic trust educate income age male www

Logistic regression                               Number of obs   =       1174
                                                  LR chi2(5)      =     128.68
                                                  Prob > chi2     =     0.0000
Log likelihood = -733.97164                       Pseudo R2       =     0.0806

------------------------------------------------------------------------------
       trust | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     educate |   1.163673   .0304619     5.79   0.000     1.105474    1.224935
      income |   1.030814   .0118919     2.63   0.009     1.007768    1.054387
         age |   1.028411   .0050091     5.75   0.000      1.01864    1.038276
        male |   1.292781    .162669     2.04   0.041     1.010228    1.654362
         www |   1.739745   .2885914     3.34   0.001      1.25686    2.408153
------------------------------------------------------------------------------
```

Odd ratio >(<)1 → Positive (negative) effect.

For example, the probability that a male trusts people is larger than the one of a female (1.29 times)

Equivalent to t-values. The higher t value, the higher the significance of the variable

Equivalent to two-tail p-values. In this case, all variables are significant (5%)

LOGIT – marginal effects

The predicted probability of trusting people is 0.4753

for female WWW users at the average age of 41 who graduated in a college (16 years of education) and have average family income of 25,000USD.

```
. mfx, dydx at(mean educate=16 male=0 www=1)

Marginal effects after logit
      y  = Pr(trust) (predict)
         =  .47534926
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
 educate |   .0378032      .0066    5.73   0.000   .024873  .050734       16
  income |   .0075687     .00287    2.63   0.008   .001934  .013203  24.6486
     age |   .0069868     .00121    5.75   0.000   .004606  .009367  41.3075
   male* |   .0640968     .03132    2.05   0.041   .002718  .125475        0
    www* |   .1329051     .03797    3.50   0.000   .058487  .207323        1
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

Marginal Effects

For example, for a year increase in education after college graduation, the predicted probability of trusting people will increase by 3.78%,

holding other independent variables constant at the reference points (column x)

REMINDER

- Stata `probit` estimates the binary probit regression model. If you want to get robust standard errors, add the `robust` option to `logit` and `probit`. The logit and probit models produce almost similar goodness-of-fit measures but their parameter estimates differ.
- The standard normal probability distribution and standard logistic distribution respectively have a unit variance and a variance of $\pi^2/\sqrt{3}$. Therefore, a parameter estimate in a binary logit model is about $\pi/\sqrt{3}$ larger than its corresponding coefficient in its probit counterpart.

PROBIT – regression

- Compare with the result of the logit model (same example).

```
. probit trust educate income age male www

Iteration 0:   log likelihood = -798.31217
Iteration 1:   log likelihood = -734.10951
Iteration 2:   log likelihood = -733.99746
Iteration 3:   log likelihood = -733.99746

Probit regression                                 Number of obs   =       1174
                                                  LR chi2(5)      =     128.63
                                                  Prob > chi2     =     0.0000
Log likelihood = -733.99746                       Pseudo R2       =     0.0806

------------------------------------------------------------------------------
       trust |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     educate |   .0907207   .0154349     5.88   0.000     .0604689    .1209725
      income |   .0185906   .0068681     2.71   0.007     .0051293    .0320519
         age |   .0173105   .0029496     5.87   0.000     .0115293    .0230916
        male |   .1593935   .0768819     2.07   0.038     .0087077    .3100793
         www |   .3417645   .0992156     3.44   0.001     .1473055    .5362235
        cons |  -3.030053   .2786062   -10.88   0.000    -3.576111   -2.483995
------------------------------------------------------------------------------
```

DIFFERENCES WITH ODDS RATIO PROBIT ANALYSIS:
1. Presence of the constant
2. Coefficient instead of odds ratio

SAME AS ODDS RATIO PROBIT ANALYSIS:
1. Same Pseudo R-squared
2. Significance of independent variables

PROBIT – marginal effects

- Compare with the result of the logit model (same example).

The predicted probability of trusting people is 0.4747 (0.4753 in the logit model) for the same female (WWW users, 41, 16 years of education, family income of 25,000USD).

```
. mfx, at(mean educate=16 male=0 www=1)

Marginal effects after probit
      y  = Pr(trust) (predict)
         =  .47469509
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
 educate |   .0361195      .00681    5.30   0.000   .022774  .049465       16
  income |   .0074017      .00264    2.81   0.005   .002234  .012569  24.6486
     age |    .006892      .00118    5.83   0.000   .004574   .00921  41.3075
   male* |   .0635132      .03058    2.08   0.038   .003573  .123453        0
    www* |   .1320435       .0374    3.53   0.000   .058748  .205339        1
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

Marginal Effects
For year increase in education after college graduation, the predicted probability of trusting people will increase by 3.61% (3.78%, in the logit model)
holding other independent variables constant at the reference points (column x),
same values of the logit model

FIRM LEVEL ANALYSIS

- In the previous session, we argued that OLS procedures were not adequate for firm level analysis in case of binary dependent variable, as, for example, export Y/N.
- Here we have discussed the methodologies suitable for binary dependent variable, and in the next slides we will present the results of some published papers.
- These papers study the effects of Foreign Direct Investment, innovations and locations on export propensity (export Y/N) of firms in Mexico [Aitken, Hanson and Harrison, Journal of International Economics, 1997] and in the UK [Wakelin, Research Policy, 1998].

WAKELIN (1/3)

- Uses a probit model in order to discuss the effects of **size, average capital intensity, average wages and unit labour costs** (exogenous variables) on the **probability of exporting** (dependent variable) of UK firms.
- She tests a probit model separating innovating and non-innovating firms finding that the two groups of firms behave differently. What determines such difference? Not the size of firms themselves.
- In order to explore the issue, she introduces two alternative exogenous variables on innovation: the number of innovations in the sector and the level of R&D expenditure in the sector. At this point, differences emerges between the two samples:

WAKELIN (2/3)

1. Skill effect for non-innovating firms only.
2. Effect of unit labour cost: positive for innovating, negative for non-innovating.
3. R&D expenditure of other firms in the sector has a positive impact on non-innovators only.

- These are the result of the empirical analysis. Once we have run the model, we need to give an economic interpretation of the results, paying attention in particular to the policy implications.

WAKELIN (3/3)

- We also need to have a critical point of view: this dataset does not have geographical specific info on the location of firms.
- The effect of firm location on export probability has been shown to matter for Mexican firms [Aitken, Hanson and Harrison, Journal of International Economics, 1997]

AITKEN, HANSON, HARRISON (1/2)

- This paper focuses on Mexican multinational enterprises. The authors estimate a probit model to explore the export propensity of firms.
- The main difference with respect to the previous paper is the greater attention given here to production sectors and firms locations (in terms of site-specific and sector-specific characteristics).
- Beyond the discussion of the general results, we want to highlight the importance of these issues, in order to show that depending on country specific issues (Mexico instead of UK) the choice of the adequate explanatory variables is relevant.

AITKEN, HANSON, HARRISON (2/2)

- Example 1: The measure of overall industry concentration may not adequately control for site-specific characteristics, grouping together industries with dissimilar factor intensities (Shrimps packing is a four-digit industry within food products: it is concentrated in two states but data cannot control for that, being at three-digit level).
- Example 2: Industries intensive in the use of natural resources and/or with high transportation costs have site-specific factors relevant in the export decision. The presence of a relative small set of such industries is determinant for the correlation between local export concentration and export probability.