MASTERS EPP

ECO 552

LECTURE 5

Probit and logit models

Francis Kramarz and Michael Visser

In many cases, the variable to be explained is a qualitative variable taking the value 1 (when the response is “yes”, for instance) or 0 (when the response is “no”, for instance)

Examples: the questions could be:

- Do you have a personal computer?
- Did you vote in the last election?
- What is your preferred leisure activity among the following choices: going to the movie theater, going to the opera house, reading novels, looking at TV?
- What was your labour market situation in september 2009: employed, self-employed, unemployed, out-of-the labour force?

Difficult to analyze such dependent variables with a linear model, since the corresponding information cannot be naturally *ordered*

Examples :

- *the labour market situation*: some workers may prefer to be employed in a temporary job rather than in a long-term labour contract job; some others may prefer the opposite situation;
- *buying a durable good*: some households want to have a personal computer and can buy it; some others want to have a computer but cannot buy it, while others do not want to have a computer at home (whatever their income is)

The answers correspond to individual choices that are said to be *discrete* since their direct consequence is:

- some *specific action* (example : accepting ot not a job offer)

- but not *the level or the intensity of the corresponding outcome* (example: the number of work hours, the earned income, etc.)

Two possible approaches:

1. assuming that discrete choices result from a rational economic behaviour (maximizing either the individual utility or the firm profit)

2. adopting a *more descriptive approach*

1. The regression approach

The variable to be explained can take only two values, either $Y = 1$ or $Y = 0$

The explanatory variables (which may influence the decision) are denoted X

We represent the relationship between these explanatory variables and the explained variable through a discrete probability model:

$$\Pr(Y = 1) = F(X\beta)$$

$$\Pr(Y = 0) = 1 - F(X\beta)$$

where:

- F is an increasing function from $\mathbb{R}$ to $]0, 1[$
- β is a vector of unknown parameters (to be estimated) associated with the vector X and whose dimension is $(L, 1)$ when the vector X has dimension $(1, L)$

1a. The linear probability model

When F is the identity function:

$$F(X\beta) = X\beta$$

then:

$$E(Y \mid X) = \sum_{j=0}^{1} \Pr(Y = j) \times j = \Pr(Y = 1) = X\beta$$

Now let us assume that Y is generated by the following *linear probability model*:

$$Y = X\beta + u$$

with

$$E(u \mid X) = 0$$

The Ordinary Least Squares (OLS) estimator of β is:

$$\widehat{\beta}_{MCO} = (X'X)^{-1}X'Y$$

Main drawback of this model: $\widehat{E}(Y \mid X) = X\widehat{\beta}_{MCO}$ cannot be easily constrained to belong to the interval $[0,1]$

1b. The probit and logit models

Functions F such as $\widehat{E}(Y \mid X) \in [0,1]$ must verify the following conditions:

1. $F(X\beta)$ should increase with $X\beta$
2. $\lim_{X'\beta\to+\infty} \Pr(Y=1) = 1$
3. $\lim_{X'\beta\to-\infty} \Pr(Y=1) = 0$

Any cumulative density function of a continuous random variable is a good candidate

If we choose the standard normal distribution $N(0,1)$, the corresponding probability model is called the probit model. It is defined as :

$$\Pr(Y = 1) = \int_{-\infty}^{X'\beta} \varphi(t)dt = \Phi(X\beta)$$

$$\Pr(Y = 0) = \int_{X'\beta}^{+\infty} \varphi(t)dt = 1 - \Phi(X\beta)$$

where Φ is the c.d.f. of the standard normal distribution $N(0,1)$, and φ is its density function

The logit model is still easier to implement (no integral):

$$\Pr(Y = 1) = \frac{\exp(X\beta)}{1 + \exp(X\beta)} = \Lambda(X\beta)$$

$$\Pr(Y = 0) = \frac{1}{1 + \exp(X\beta)} = 1 - \Lambda(X\beta)$$

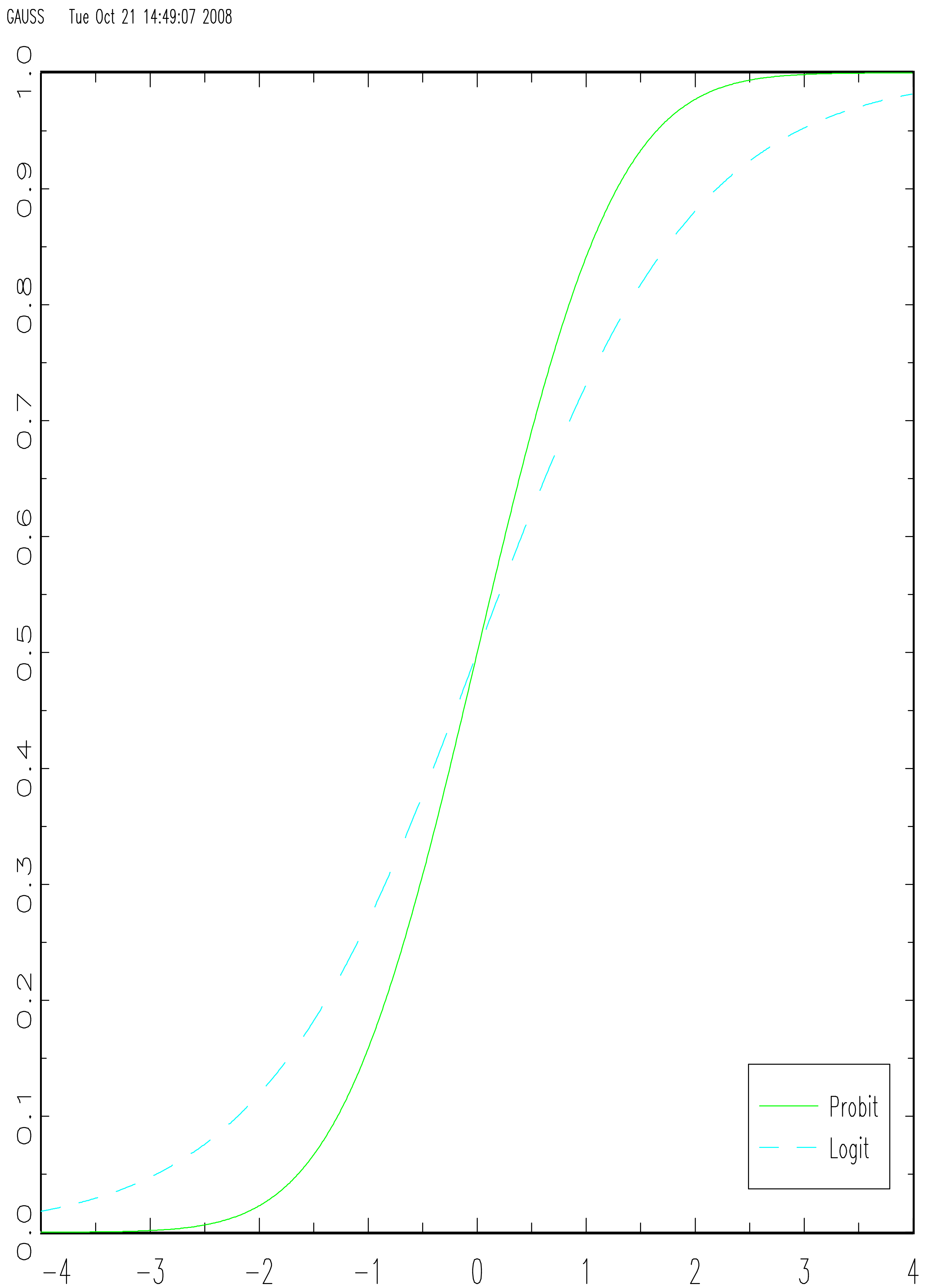
Remark: Probit or Logit?

The logistic function converges less rapidly towards extreme values (0 and 1) than the normal distribution $N(0,1)$: it allows extreme values to be more frequent (see the graph)

The two models give significantly different predictions when the sample contains very few observations such as $Y = 1$ (or such as $Y = 0$)

Generally, the estimated values of parameters β are different (even if predictions $\widehat{E}(Y \mid X)$ are similar)

β is estimated through a maximum likelihood procedure


GAUSS Tue Oct 21 14:49:07 2008
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4
-3
-2
-1
0
1
2
3
4
Probit
Logit

2. The maximum likelihood approach

2.1 The ML estimator

Definition of the likelihood function

- If X is a discrete random variable, the likelihood for the observation x is:

$$L(x,\theta) = \Pr(X = x;\theta)$$

  where θ is the vector of parameters characterizing the distribution of X

- If X is a real random variable, the likelihood for the observation x is:

$$L(x,\theta) = f_X(x;\theta)$$

Let $x = (x_1,\ldots,x_n)$ be a realization of the sample $(X_1,\ldots,X_n)$

The sample likelihood function is:

$$L_n(x;\theta) = \prod_{i=1}^{n} L(x_i;\theta)$$

<u>Definition</u>

A maximum likelihood estimator (MLE) of θ is a solution of the maximization program

$$\max_{\theta\in\Theta} L_n(x;\theta)$$

which is equivalent to

$$\max_{\theta\in\Theta} \ln L_n(x;\theta)$$

<u>Remark</u>: Maximizing the likelihood function with respect to (w.r.t.) β gives the same solution than maximizing the logarithm of this function w.r.t. β, since the logarithm function is an increasing mononotonic transformation

Definition: The *likelihood equations* are derived from the first-order conditions of the program:

$$\frac{\partial L_n\left(x;\widehat{\theta}\right)}{\partial\theta} = \frac{\partial \ln L_n\left(x;\widehat{\theta}\right)}{\partial\theta} = 0$$

In general, these equations are non linear. They can be solved by implementing an iterative algorithm, for instance the *Newton-Raphson algorithm*:

$$\theta^{(l+1)} = \theta^{(l)} - \left[\frac{\partial^2 \ln L(\theta)}{\partial\theta\partial\theta'}\right]^{-1}_{\theta=\theta^{(l)}} \times \left[\frac{\partial \ln L(\theta)}{\partial\theta}\right]_{\theta=\theta^{(l)}}$$

where $\theta^{(l)}$ is the value of the parameter vector θ at iteration l, $\theta^{(0)}$ being an initial value

If the hessian matrix is negative-definite, the log-likelihood function is globally concave. This method converges within a finite number of iterations

The Fisher information matrix

$$I_1(\theta) = E\left(\frac{\partial \ln L(X;\theta)}{\partial \theta} \frac{\partial \ln L(X;\theta)}{\partial \theta'} \right)$$

$$= -E\left(\frac{\partial^2 \ln L(X;\theta)}{\partial \theta \partial \theta'} \right)$$

Proof (simplified, case of a single parameter):

$$E\left(\frac{\partial^2 \ln L(X;\theta)}{\partial \theta^2} \right) = E \frac{\partial}{\partial \theta} \left(\frac{\partial \ln L(X;\theta)}{\partial \theta} \right) = E \frac{\partial}{\partial \theta} \left(\frac{1}{L} \frac{\partial L(X;\theta)}{\partial \theta} \right)$$

$$= -E \frac{1}{L^2} \left(\frac{\partial L(X;\theta)}{\partial \theta} \right)^2 + \underbrace{E \frac{1}{L} \frac{\partial^2 L(X;\theta)}{\partial \theta^2}}_{= \int (\partial^2 L / \partial \theta^2) dx = \partial^2 / \partial \theta^2 \int L dx = 0}$$

$$= -E \frac{1}{L^2} \left(\frac{\partial L(X;\theta)}{\partial \theta} \right)^2 = -E\left(\frac{\partial \ln L(X;\theta)}{\partial \theta} \right)^2$$

2.2 Asymptotic properties of the MLE

Under some (general) regularity assumptions, there exists a sequence $\widehat{\theta}_n$ of local maxima of the log-likelihood function which converges towards θ_0 and which verifies

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \underset{n\to\infty}{\overset{loi}{\to}} \mathrm{N}\left(0, \mathrm{I}_1(\theta_0)^{-1}\right)$$

Properties: The MLE is asymptotically efficient. No other regular estimator has a higher precision.

Remark: $\mathrm{I}_1(\theta_0)$ is unknown since the true value θ_0 of the parameter is unknown, but it can be consistently estimated by $\mathrm{I}_1\left(\widehat{\theta}_n\right)$

Example

Let us consider a sample of n normally distributed random variables:

$$X_i \rightsquigarrow N(m, \sigma^2)$$

The likelihood function of one observation:

$$\frac{1}{\sqrt{2\pi}\,\sigma} \exp(-(\frac{X_i - m}{\sigma})^2/2)$$

The sample log-likelihood function:

$$-\frac{n}{2}\log\pi - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2$$

Likelihood equations:

$$\sum_{i=1}^{n}(x_i - \hat{m}_n) = 0$$

$$-\frac{n}{\hat{\sigma}_n} + \frac{1}{\hat{\sigma}_n^3}\sum_{i=1}^{n}(x_i - \hat{m}_n)^2 = 0$$

MLE:

$$\hat{m}_n = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}_n$$

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = \frac{n-1}{n}s_n^2$$

The Fisher information matrix:

$$I_1(\theta) = \begin{pmatrix} 1 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}$$

2.3 The likelihood ratio test

Testing the null hypothesis: $H_0 \; : \; \widetilde{\theta} = 0$

against the alternative hypothesis: $H_1 \; : \; \widetilde{\theta} \neq 0$

where $\widetilde{\theta}$ is a sub-vector of θ : $\dim\left(\widetilde{\theta}\right) = p \leq \dim(\theta) = k$

Notations: $\widehat{\theta}_n^{(0)}$ MLE of θ under H_0 and $\widehat{\theta}_n^{(1)}$ MLE of θ under H_1

Under some regularity conditions, the test defined by the rejection region

$$W = \left\{ \xi_n^R \geq \chi^2_{1-\alpha}(p) \right\}$$

with

$$\xi_n^R = 2\left[\log L_n\left(x; \widehat{\theta}_n^{(1)}\right) - \log L_n\left(x; \widehat{\theta}_n^{(0)}\right) \right]$$

has an asymptotic level equal to α and it is convergent

3. Estimating the probit and logit models

Each observation may be viewed as a random draw from the Bernouilli distribution with parameter $F(X\beta)$

If the observations are i.i.d., the joint probability of the sample is given by the **likelihood function**:

$$
\begin{aligned}
L(\beta) &= \Pr\left[Y_1 = y_1, \ldots, Y_n = y_n \mid \beta, (X_i)_{i=1,\ldots,n} \right] \\
&= \prod_{i:Y_i=0} [1 - F(X_i\beta)] \times \prod_{i:Y_i=1} F(X_i\beta) \\
&= \prod_{i=1}^{n} [F(X_i\beta)]^{Y_i} \times [1 - F(X_i\beta)]^{1-Y_i}
\end{aligned}
$$

The **log-likelihood function** is thus:

$$\ln L(\beta) = \sum_{i=1}^{n} Y_i \ln F(X_i\beta) + (1 - Y_i) \ln[1 - F(X_i\beta)]$$

3.1 First-order conditions (f.o.c.)

The f.o.c. may be written:

$$\frac{\partial \ln L(\beta)}{\partial \beta} = \sum_{i=1}^{n} \left[Y_i \frac{f_i}{F_i} - (1 - Y_i) \frac{f_i}{1 - F_i} \right] X_i' = 0$$

by setting:

$$F_i = F(X_i\beta) \text{ and } f_i \equiv f(X_i\beta) = \frac{\partial F(X_i\beta)}{\partial (X_i\beta)}$$

1) *For the logit model* :

By setting $\Lambda_i = \dfrac{\exp(X_i\beta)}{1+\exp(X_i\beta)}$

we get:

$$\frac{\partial \ln L(\beta)}{\partial \beta} = \sum_{i=1}^{n} (Y_i - \Lambda_i)X_i' = 0$$

2) *For the probit model* :

By setting $\Phi_i = \Phi(X_i\beta)$ and $\phi_i = \frac{\partial \Phi(X_i\beta)}{\partial (X_i\beta)}$

we get:

$$\frac{\partial \ln L(\beta)}{\partial \beta} = \sum_{i} (Y_i - \Phi_i)\frac{\phi_i}{\Phi_i(1-\Phi_i)}X_i' = 0$$

3.2 Second-order derivatives of the log-likelihood function

1) *For the logit model*:

The hessian matrix of the log-likelihood function may be written as:

$$H = \frac{\partial^2 \ln L(\beta)}{\partial\beta\partial\beta'} = -\sum_i \Lambda_i(1-\Lambda_i)X_i'X_i$$

Since Y_i does not appear in the second-order derivatives, we may write:

$$E_Y\left(-\frac{\partial^2 \ln L(\beta)}{\partial\beta\partial\beta'}\right) = \sum_i \Lambda_i(1-\Lambda_i)X_i'X_i$$

The hessian matrix is always negative-definite: the log-likelihood function is thus globally concave. The Newton-Raphson method converges towards the optimum value within a finite number of iterations

2) *For the probit model*:

We set

$$\lambda_{0i} = \frac{-\phi_i}{1-\Phi_i} \text{ if } Y_i = 0, \text{ and } \lambda_{1i} = \frac{\phi_i}{\Phi_i} \text{ if } Y_i = 1$$

which implies

$$\lambda_i = \lambda_{0i}(1 - Y_i) + \lambda_{1i} Y_i$$

Then the hessian matrix may be written as:

$$H = \frac{\partial^2 \ln L(\beta)}{\partial\beta\partial\beta'} = -\sum_i \lambda_i(\lambda_i + X_i\beta)X_i'X_i$$

Then it may be shown that H is negative-definite for any value of β

3.3 The covariance matrix of the MLE

This matrix is estimated by the inverse of the hessian matrix evaluated at $\widehat{\beta}$:

$$\widehat{V}\left(\widehat{\beta}\right) = \left(-\frac{\partial^2 \ln L(\beta)}{\partial\beta\partial\beta'}\right)^{-1}_{\beta=\widehat{\beta}}$$

It may be also estimated by the inverse of the cross-products of the first-order derivatives of the log-likelihood function evaluated at $\widehat{\beta}$:

$$\widehat{V}\left(\widehat{\beta}\right) = \left(\frac{\partial \ln L(\beta)}{\partial\beta} \times \frac{\partial \ln L(\beta)}{\partial\beta'}\right)^{-1}_{\beta=\widehat{\beta}} = \left(\sum_i g_i^2 X_i' X_i\right)^{-1}$$

with:

- $g_i = Y_i - \widehat{\Lambda}_i$ for the *logit model*
- $g_i = \widehat{\lambda}_{0i}(1 - Y_i) + \widehat{\lambda}_{1i} Y_i$ for the *probit model*

3.4 How can we measure the fit of these two models?

The pseudo-R^2 is defined as:

$$pseudo - R^2 = 1 - \frac{\sum_i [y_i \ln \widehat{p}_i + (1 - y_i) \ln(1 - \widehat{p}_i)]}{N[\bar{y} \ln \bar{y} + (1 - \bar{y}) \ln(1 - \bar{y})]}$$

where $\widehat{p}_i = F\left(x_i \widehat{\beta}\right)$

and $\bar{y} = N^{-1} \sum_i y_i$ is the proportion of observations such as $y_i = 1$

3.5 An example: the probability of a car accident

M. Boyer and G. Dionne (1989): “An Empirical Analysis of Moral Hazard and Experience Rating”, The Review of Economics and Statistics, vol. 71, pp. 128-134

In presence of moral hazard (i.e. when the insurance company cannot observe the behaviour of its prospects), the insurance company has to design a tariff system that incorporates the *ex ante* accident probability of each customer

How does this probability vary with:

1. the individual characteristics of the customer (age, gender, place of residence, number of years with a driving license, class of the driving license, etc.)
2. his/her past driving experience (number of past involvements in accidents and demerit points cumulated in the two last years, number of license suspensions during the last year)

Sample: 19,013 drivers in Quebec, observed between August 1980 and July 1983

Estimation of a *Probit model*

Main results:

- the accident probability of drivers older than 25 is lower by 2 or 3 points than the probability of drivers less than 19 years old
- the accident probability of men is higher by 3.7 points than the accident probability of women
- the number of years with a driving license and the place of residence have no statistically significant effect

- drivers who cumulated five demerit points during the last two years have an accident probability that is higher by 3.4 points (0.6 + 2.8) than the probability of drivers with no demerit points (variable *X*, table 2, last column)
- drivers who were involved in a car accident in the last two years have a probability of accident that is higher by 2.5 points than the probability of drivers who had no accident (variable *Z*, table 2, last column)
- a second accident increases this probability by 3.4 points
- one suspension of license is associated with an accident probability that is higher by 3.9 points (variable *Y*, table 2)

TABLE 1.—DEFINITION OF SIGNIFICANT VARIABLES USED IN THE ECONOMETRIC ANALYSIS

AGE: *A*1619 = 1 if the driver is between 16 and 19 on 1/8/1982 (omitted category); etc.
SEX: *SEXM* = 1 if male.
NUMBER OF YEARS WITH A DRIVING LICENSE: *EXPO* = 1 if permit obtained after 1/8/1982; *EXP*12 = 1 if permit obtained between 1 and 2 years ago (before 1/8/1982) (omitted category); etc.
PLACE OF RESIDENCE: *REG*6 = 1 if the driver lives in the Montreal region (omitted category); *REG*9, Outaouais region; etc.
DRIVING RESTRICTIONS: *RTSA* = 1 if the driver must wear glasses; *RTSJ*, must drive an automatic transmission equipped car; *RTSU*, has a license valid for 6 months only; *RTSY*, cannot drive a taxi or an ambulance; *RTS*0, has no restrictions; etc.
CLASS OF DRIVING LICENCE: *CL*21 = 1 if the driver can drive a vehicle (*CL*22) or a set of vehicles whose weight may exceed 11000 kg; *CL*31, a taxi; etc.
VALIDITY: *VALA* = number of days the individual's license was valid in 1980–81; *VALB*, in 1981–82; *VALC*, in 1982–83.
DEMERIT POINTS: *X* = the number of demerit points cumulated from 8/1980 to 7/1982 for infractions such as not stopping at a stop sign (2 points) or at a red light (3), racing (6), not stopping for a school bus with blinking lights on (9), exceed speed: 1 to 14 km/h over limit (1), 15 to 29 (2), etc.
SUSPENSIONS OF LICENSE: *Y* = the number of license suspensions in 1981–82 for criminal offenses such as negligence causing death or bodily injuries, hit and run, driving under the influence of alcohol, etc.
PAST INVOLVEMENTS IN ACCIDENTS: *Z* = the number of accidents from 8/1980 to 7/1982

TABLE 2.—THE PROBIT ESTIMATES

Variable	Original Coefficient	(t)	Transformed Coeffficient
X	.055	(9.62)[b]	0–1: .006 1–5: .028 5–31: .434
Y	.290	(2.23)[b]	0–1: .039 1–2: .057 2–3: .077
Z	.211	(8.11)[b]	0–1: .025 1–2: .034 2–6: .239
CONSTANT	−2.295	(−10.30)[b]	NC[c]
$A16-$	−.304	(−0.89)	−.025
$A2024$	−.028	(−0.35)	−.003
$A2534$	−.207	(−2.40)[b]	−.020
$A3544$	−.292	(−3.11)[b]	−.027
$A4554$	−.288	(−2.93)[b]	−.026
$A5564$	−.409	(−3.91)[b]	−.033
$A65+$	−.372	(−3.11)[b]	−.030
$SEXM$	.370	(9.80)[b]	.037
$EXPO$	.227	(1.08)	.029
$EXPO1$	.116	(0.72)	.014
$EXP23$	−.145	(−1.10)	−.014
$EXP36$	−.245	(−1.79)[a]	−.022
$EXP611$	−.251	(−1.70)[a]	−.024
$EXP11+$	−.194	(−1.29)	−.021
$REG1$	.107	(1.30)	.012
$REG2$	.035	(0.46)	.004
$REG3$	.073	(1.55)	.008
$REG4$	.106	(1.75)[a]	.012
$REG5$	.025	(0.30)	.003
$REG7$	.073	(1.32)	.008
$REG8$	.008	(0.18)	.001
$REG9$	.245	(3.46)[b]	.031
$REG10$	.128	(1.43)	.015
$REG11$	.158	(1.46)	.019

TABLE 2.—THE PROBIT ESTIMATES

Variable	Original Coefficient	(t)	Transformed Coeffficient
$RTSA$	−.254	(−2.40)[b]	−.025
$RTSB$	.155	(0.60)	.019
$RTSCG$	−.146	(−0.97)	−.014
$RTSD$	.034	(0.28)	.004
$RTSH$	.118	(1.04)	.014
$RTSJ$	.736	(1.73)[a]	.134
$RTSK$	−.535	(−0.99)	−.037
$RTSM$	−.273	(−1.55)	−.023
$RTSO$	.145	(0.67)	.017
$RTSQ$	−.591	(−0.88)	−.039
$RTSU$	.366	(1.86)[a]	.052
$RTSY$	−.977	(−2.31)[b]	−.048
$RTSO$	−.186	(−1.71)[a]	−.021
$CL1112$	.158	(1.41)	.019
$CL13$	−.256	(−0.51)	−.022
$CL21$	.127	(1.88)[a]	.014
$CL22$	.560	(2.81)[b]	.091
$CL31$	.359	(2.74)[b]	.050
$CL42$	.045	(0.77)	.005
$CL54$	.041	(0.52)	.004
$CL55$	−.579	(−1.55)	−.038
$CL56$	−.104	(−0.17)	−.010
$VALA$	.001	(0.97)	NC
$VALB$	−.0002	(−0.42)	NC
$VALC$	.002	(5.70)[b]	NC
Number of observations			19013
Number of variables			53
Likelihood ratio			716.46
Mean estimated probability of accident			0.065
Estimated probability of accident for the average individual			0.052
(Standard error)			(0.002)

[a] Significant at 90%.
[b] Significant at 95%.
[c] NC = not calculated.

4. Random utility models

Models with dependent qualitative variables are often written as *models with an index function*

The outcome resulting from a discrete choice is then assumed to be generated by a latent regression model

Example: the purchase of a durable good:

Economic theory assumes that a consumer compares her utility when she purchases a given durable good with her utility when she does not purchase it

We assume then that the difference between these two utilities is represented by a *latent* variable, which is unobservable :

$$Y^* = X\beta + \varepsilon \quad \text{where} \quad \varepsilon \sim \mathrm{N}(0,1)$$

Indeed we observe the variable Y defined by:

$$Y = 1 \quad \text{if } Y^* > 0 \quad \text{and } Y = 0 \quad \text{otherwise}$$

In this expression, $X\beta$ is called the index function, and thus:

$$Y = \mathbf{1}(X\beta + \varepsilon > 0)$$

where $1(.)$ is a function taking the value 1 when the logical expression within parenthesis is true, 0 otherwise.

Remarks :

1. The variance of ε cannot be identified. To understand that point, we can multiply $X'\beta + \varepsilon$ by σ^2: this does not modify the values of the variable Y ($Y = 0$ or $Y = 1$). Consequently, we assume in general that $\sigma^2 = 1$ (normalization)
2. The assumption of a zero threshold has no consequence as long as the index $X'\beta$ includes an intercept

5. The multinomial logit model

Let us assume that an individual (denoted i) must choose only one item (denoted k) among K possible choices

Example: choosing a place for vacation among three possibilities: mountains, the seaside, the country.

In the sequel, we assume that a given utility level is associated with each of these K possible choices:

$$U_{ik} = \mu_{ik} + \varepsilon_{ik} \quad (k = 1, \ldots, K)$$

where μ_{ik} is deterministic function of some observable variables (for instance, $\mu_{ik} = X_i \beta_k$) and ε_{ik} is an independent random variable

The individual is assumed to choose the item k which gives her the highest utility

Theorem (Mac Fadden, 1973): If the residuals $\{\varepsilon_{ik}\}_{k=1,\ldots,K}$ are i.i.d. random variables which are drawn from the extreme value distribution, whose c.d.f. is :

$$G(x) = \exp[-\exp(-x)]$$

then the probability of choosing item k is:

$$\Pr[Y_i = k] = \frac{\exp(\mu_{ik})}{\sum_{k'=1}^{K} \exp(\mu_{ik'})} = \frac{\exp(X_i\beta_k)}{\sum_{k'=1}^{K} \exp(X_i\beta_{k'})}$$

This model is called the multinomial logit model.

Remarks:

1. These probabilities only depend on the differences:

$$\mu_{ik'} - \mu_{ik} = X_i(\beta_{k'} - \beta_k), \quad k' \neq k$$

They are not modified if we add the same constant term to all parameters β_k

2. Consequently, parameters β_k cannot be separately identified, except if we set $\beta_1 = 0$

3. Estimated parameters are interpreted as differences with respect to the reference parameter β_1. A positive sign means that the explanatory variable increases the probability of choosing a given item (say, item k) relatively to the reference item (say, item 1)

<u>Estimation of the multinomial logit model</u>. We set:

$$P_{ik} = \Pr[Y_i = k] = \frac{\exp(X_i\beta_k)}{\sum_{k'=1}^{K} \exp(X_i\beta_{k'})}$$

with $\beta_1 = 0,\ i = 1,\ldots,n,$ and $k = 1,\ldots,K$

The log-likelihood function of the sample may then be written as:

$$\ln L(\beta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbf{1}(Y_i = k) \times \ln(P_{ik})$$

This log-likelihood function is globally concave

Sketch of the proof: it may be shown that the hessian matrix, whose form is:

$$\frac{\partial^2 \ln L(\beta)}{\partial\beta\partial\beta'} = -\sum_{i=1}^{n}\sum_{k=1}^{K} P_{ik}\left(X_i' - \overline{X}_i'\right)(X_i - \overline{X}_i)$$

$$\text{with } \overline{X}_i' = \frac{\sum_{k'=1}^{K}\exp(X_i\beta_{k'})X_i'}{\sum_{k'=1}^{K}\exp(X_i\beta_{k'})}$$

is negative-definite since $P_{ik} = \Pr[Y_i = k] > 0$

Since the hessian matrix does not depend on Y_i, we show finally that:

$$\widehat{V}\left(\widehat{\beta}\right) = \left[\sum_{i=1}^{n}\sum_{k=1}^{K} P_{ik}\left(X_i' - \overline{X}_i'\right)(X_i - \overline{X}_i)\right]_{\beta=\widehat{\beta}}^{-1}$$