Logit models of individual choices

Thierry Magnac
Université de Toulouse

First version: September 2005.
(prepared for the New Palgrave)

The logit function is the reciprocal function to the sigmoid *logistic* function. It maps the interval [0,1] into the real line and is written as:

$$logit(p) = \ln(p/(1-p))$$

Two traditions are involved in the modern theory of logit models of individual choices. The first one concerns *curve fitting* as exposed by Berkson (1944) who coined the term *logit* after its close competitor *probit* which is derived from the normal distribution. Both models are by far the most popular econometric methods used in applied work to estimate models for binary variables, even though the development of semiparametric and non parametric alternatives over the last 30 years has been intensive (Horowitz and Savin, 2001).

In the second strand of literature, models of discrete variables and discrete choices as originally set up by Thurstone (1927) in psychometrics have been known as *random utility models* (RUM) since Marschak (1960) introduced them to economists. As the availability of individual databases and the need for tools to forecast aggregate demands derived from discrete choices were increasing from the 1960s onwards, different waves of innovations, fostered by McFadden (see his Nobel lecture, 2001) elaborated more and more sophisticated and flexible logit models. The use of these models and of simulation methods triggered burgeoning applied research in demand analysis in recent years.

Those who wish to study the subject in greater detail are referred to Gouriéroux (2000), McFadden (2001) or Train (2003) where references to applications in economics and marketing can also be found.

Measurement models

As Berkson (1951) put it, logit (or probit) models may be seen as "merely a convenient way of graphically representing and fitting a function". It is used for any empirical phenomena delivering a binary random variable Y_i, taking values 0 and 1, to be analyzed. In a logit model, it is postulated that its probability distribution conditional on a vector of covariates X_i is given by:

$$\Pr(Y_i = 1 \mid X_i) = \frac{\exp(X_i\beta)}{1+\exp(X_i\beta)}$$

where β is a vector of parameters. This model can also be derived from more general frameworks in statistical mechanics or spatial statistics (Strauss, 1992).

Using cross-sectional samples, the parameter of interest is estimated using maximum likelihood or by GLM methods where the link function is logit (McCullagh and Nelder, 1989). Under the maintained assumption that it is the true model and other standard assumptions, the Maximum Likelihood Estimator (MLE) is consistent, asymptotically normal and efficient (Amemiya, 1985). Nevertheless, the MLE may fail to exist, or more exactly be at the bounds of the parameter space, when the samples are uniformly composed by 0s or 1s for instance (Berkson, 1955).

When repeated observations are available, the method of Berkson delivers an estimator close to MLE since they are asymptotically equivalent. Observe first that the logit function of the true probability obeys the linear equation:

$$logit(\Pr(Y_c = 1 \mid X_c)) = X_c\beta.$$

where the covariates X_c now takes a discrete number of values defining each cell, c. Second use the observed frequencies in each cell, $\hat{p}_c$, and contrast it with theoretical probabilities, p_c, as:

$$\begin{aligned} logit(\hat{p}_c) &= X_c\beta + (logit(\hat{p}_c) - logit(p_c)) \\ &= X_c\beta + \varepsilon_c. \end{aligned}$$

The random term ε_c properly scaled by the square root of the number of observations in cell c is asymptotically normally distributed with variance equal to $1/(p_c(1-p_c))$. The method of Berkson then consists in using minimum chi-square, i.e. a method of moments, to estimate β, an instance of what is know as minimum distance or asymptotic least squares (Gouriéroux, Monfort and Trognon, 1985).

When measurements for a single individual are repeated, Rasch (1960) suspected that individual effects might be important and proposed to write:

$$logit(\Pr(Y_{it} = 1 \mid X_{it})) = X_{it}\beta + \delta_i$$

where t indexes the different items that are measured and δ_i is an individual specific intercept or fixed effect. Items can be different questions in performance tests or different periods. In the original Rasch formulation, parameters were allowed to be different across items, β_t, and there were no covariates.

Given that the number of items is small, it is well known that the estimation of such a model runs into the problem of incidental parameters (see Lancaster, 2000). As the number of parameters δ_i increases with the cross-section dimension, the MLE is inconsistent (Chamberlain, 1984). Nevertheless, the nuisance parameters δ_i can be differenced out using conditional likelihood methods (Andersen, 1971) because:

$$logit(\Pr(Y_{it} = 1 \mid X_{it}, Y_{it} + Y_{it'} = 1)) = (X_{it} - X_{it'})\beta.$$

The conditional likelihood estimator of β is consistent and root n asymptotically normal but it is not efficient although no efficient estimator is known. Furthermore, when binary variables Y_{it} are independent, conditionally on X_i, the only model where a root n consistent estimator exists is a logit model (Chamberlain, 1992). Extensions of Rasch rely on the fact that root n consistent estimators exist if and only if $Y_{it} + Y_{it'}$ is a sufficient statistic for the nuisance parameters δ_i (Magnac, 2004). When the number of items or periods becomes large, profile likelihood methods where individual effects are treated as parameters seem to be accurate in Monte-Carlo experiments as soon as the number of periods is 4 or 5 (Arellano, 2003).

Multinomial logit (or in disuse "conditional logit") is to binary logit what is a multinomial to a binomial distribution (Theil, 1969). Given a vector Y_i consisting of K elements which are binary random variables and lie in the $\mathbb{R}^K-$simplex (their sum is equal to 1), it is postulated that:

$$\Pr(Y_i^{(k)} = 1 \mid X_i) = \frac{\exp(X_i\beta^{(k)})}{1 + \sum_{k=2}^{K} \exp(X_i\beta^{(k)})}$$

where by normalization, $\beta^{(1)} = 0$. Ordered logit has a different flavor since it applies to rank-ordered data such as education levels (Gouriéroux, 2000).

As probits, logit models are very tightly specified parametric models and can be substantially generalized. Much effort has been exerted to relax parametric and conditional independence assumptions starting with Manski (1975). Manski (1988) analyzes the identifying restrictions in binary models and Horowitz (1998) reviews estimation methods. In some cases, Lewbel (2000) and Matzkin (1992) offer alternatives.

Random utility models

The theory of discrete choice is directly set-up in a multiple alternative framework. A choice of an alternative k belonging to a set C is assumed to be probabilistic either because preferences are stochastic, heterogenous or because choices are perturbed in a random way. By definition, choice probability functions map each alternative and choice sets into the simplex of $\mathbb{R}^K$.

A strong restriction on choices is the axiom of Independence of Irrelevant Alternatives (IIA, Luce, 1959). The axiom states that the choice between two alternatives is independent of any other alternative in the choice set. The version that allows for zero probabilities (McFadden, 2001) states that for any pair of choice set C, C' such that $\{k, k'\} \in C$ and $C \subset C'$:

$$\Pr(k \text{ is chosen in } C') = \Pr(k \text{ is chosen in } C). \Pr(\text{An element of } C \text{ is chosen in } C').$$

Under this axiom, choice probabilities take a multinomial generalized logit form.

Moreover, assume that choices are associated with utility functions, $\{u^{(k)}\}_k$ that depend on determinants X_i and random shocks:

$$u^{(k)} = X\beta^{(k)} + \varepsilon^{(k)},$$

and that the actual choice of the decision maker yields maximum utility to her. Then, the IIA axiom is verified if and only if $\varepsilon^{(k)}$ are independent and extreme value distributed (McFadden, 1974). Extensions of decision theory under IIA were proposed in the continuous case (Resnick and Roy, 1991) or in an intertemporal context (Dagsvik, 2002).

The IIA axiom is a strong restriction as in the famous red and blue bus example where if IIA is assumed, the existence of different colours affect choices of transport between bus and other modes while introspection suggests that colours should indeed be irrelevant. Several generalizations which procede from logit were proposed to bypass IIA. Hierarchical or tree structures were the first to be used. At the upper level, the choice set consists in broad groups of alternatives. In each of these groups, there are various alternatives which can consist themselves in subsets of alternatives etc. The most well known model is the two-level nested logit where alternatives are grouped by similarities. For instance, the first level is the choice of the type of the car, the second level being the make of the car. The formula of choice probabilities for nested logit:

$$p^{(k)} = \frac{\exp(X\beta^{(k)}/\lambda_{B_s}) \left(\sum_{j\in B_s} \exp(X\beta^{(j)}/\lambda_{B_s})\right)^{\lambda_{B_s}-1}}{\sum_{t=1}^{T} \left(\sum_{j\in B_t} \exp(X\beta^{(j)}/\lambda_{B_s})\right)^{\lambda_{B_t}}},$$

where alternative k belongs to B_s, is not illuminating but the logic of construction is clear. Choices at each level are modelled as multinomial logit (Train, 2003).

General Extreme Value distributions (McFadden, 1984) provide more extensions although they do not generate all configurations of choice probabilities. In contrast, mixed logit does, as shown by McFadden & Train (2000). Instead of considering that parameters are deterministic, make them random or heterogeneous across agents. The resulting model is a mixture model where individual probabilities of choice are obtained by integrating out the random elements as in

$$p^{(k)} = \int p^{(k)}(\beta) f(\beta) d\beta.$$

Integrals are computed using simulation methods (MacFadden, 2001). The same principle is used by Berry, Levinsohn & Pakes (1995) with a view to generalize the aggregate logit choice models using market data. Logit models are still very much in use in applied settings in demand analysis and marketing and are equivalent to a representative consumer model (Anderson, de Palma and Thisse, 1992). Mixed logits permit much more general patterns of substitution between alternatives and should probably become the standard tool in the near future.

References

Amemiya, T. 1985. *Advanced Econometrics.* Cambridge, Mass.: Harvard University Press.

Andersen, E.B**.,** 1973, *Conditional Inference and Models for Measuring,* Mentalhygiejnisk Forlag:Copenhague.

Anderson, S.P., A. de Palma and J.F. Thisse, 1992, *Discrete choice theory of product differentiation*, Cambridge: MIT Press.

Arellano, M., 2003, "Discrete Choices with Panel Data", *Investigaciones Economicas*, 27(3):423-58.

Berkson, J. 1944, "Application of the logistic function to bioassay", *Journal of the American Statistical Association* 39, 357–65.

Berkson, J., 1951, "Why I prefer Logits to Probits", *Biometrics*, 7(4):327-39.

Berkson, J. 1955. "Maximum likelihood and minimum chi-square estimates of the logistic function", *Journal of the American Statistical Association* 50, 130–62.

Berry, S.T., Levinsohn, J A. and Pakes, A., 1995, "Automobile Prices in Market Equilibrium", *Econometrica,* 63(4):841-890.

Chamberlain, G., 1984, “Panel data”, in eds Z. Griliches and M. Intriligator, *Handbook of Econometrics*, 2(22):1248-1313, North Holland: Amsterdam.

Chamberlain, G., 1992, “Binary Response Models for Panel Data : Identification and Information”, Harvard University, unpublished manuscript.

Dagsvik, J., 2002, "Discrete Choice in Continuous Time: Implications of an Intretemporal Version of IAA", *Econometrica*, 70(2), 817-31.

Gourieroux C., 2000, *Econometrics of Qualitative Dependent Variables*, Cambrige: Cambridge UP.

Gouriéroux, C., A. Monfort and A., Trognon, 1985, « Moindres carrés asymptotiques », *Annales de l’INSEE*, 58, 91-121.

Horowitz, J., 1998, *Semiparametric methods in Econometrics*, Springer: Berlin.

Horowitz, J.L. and N.E., Savin, 2001, "Binary Response Models: Logits, Probits and Semiparametrics", *Journal of Economic Perspectives,* 15(4), 43-56.

Lancaster, T., 2000, “The Incidental Parameter Problem since 1948”, *Journal of Econometrics,* 95:391-413.

Lewbel, A., 2000, “Semiparametric Qualitative Response Model Estimation with Unknown Heteroskedasticity or Instrumental Variables”, *Journal of Econometrics*, 97:145-77.

Luce, R. 1959. *Individual Choice Behavior: a Theoretical Analysis.* New York: Wiley.

Magnac, T., 2004, "Panel Binary Variables and Sufficiency: Generalizing Conditional Logit", *Econometrica*, 72(6), 1859-1877.

Manski, C.F. 1975. The maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics* 3, 205–28.

Manski, C.F., 1988, “Identification of Binary Response Models”, *Journal of the American Statistical Association,* 83:729-738.

Marschak, J., 1960, "Binary Choice Constraints and Random Utility Indicators", in eds K.Arrow, *Mathematical Methods in the Social Sciences*, Stanford: Stanford UP, 312-329.

McCullagh P. and J.A. Nelder, 1989, *Generalized Linear Models*, Chapman and Hall: London.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. In Frontiers in Econometrics, ed. P. Zarembka, New York: Academic Press, 105–42.

McFadden, D. 1984. Econometric analysis of qualitative response models. In *Handbook of Econometrics,* ed. Z. Griliches and M.D. Intriligator, Vol. 2, Amsterdam: North-Holland, 1385–1457.

McFadden, D., 2001, "Economic Choices", *American Economic Review,* 91:351-378.

McFadden, D., and K., Train, 2000, "Mixed MNL models for Discrete Responses", *Journal of Applied Econometrics,* 15(5), 447-70.

Matzkin, R., 1992, “Nonparametric and Distribution-Free Estimation of the Binary Threshold Crossing and The Binary Choice Models”, *Econometrica*, 60:239-270.

Rasch, G., 1960, *Probabilistic Models for Some Intelligence and Attainment Tests*, Denmark Paedagogiske Institut, Copenhaguen.

Resnick, S.I., and R., Roy, 1991, "Random USC functions, max stable process and continuous choice", *The Annals of Applied Probability*, 1(2), 267-92.

Strauss, D., 1992, "The Many Faces of Logistic Regression", *American Statistician*, 46(4), 321-327.

Theil, H. 1969, "A multinomial extension of the linear logit model", *International Economic Review,* 10, 251–9.

Thurstone, L. 1927. A law of comparative judgement. Psychological Review 34, 273–86.

Train, K., 2003, *Discrete Choice Methods with Simulation*, Cambridge: Cambridge UP.