

Modelling the choices on a variable when the zero-value is a relevant alternative (Jordi Jaumandreu^{*}, July 2001)

Many agents' problems have the following structure: the agent can choose either performing or not an action and, if the first alternative is chosen, the agent selects the degree of intensity by which the action is carried out. Examples of this type of problems that have been intensively discussed, both in theoretical and empirical terms, include: the labour market hours of work of women, the consumer expenditures in private health insurances or a durable good, the exports of a firm, the producer's expenditures on advertising or R&D activities, the change of product prices or the hiring/firing of workers by the firm, etc. A variant of this type of problem arises when the choice is partly or totally carried out by a different agent. For example, a previous screening may determine who is entitled to follow a specific educational program, or the firm subsidies corresponding to a public program may be granted among applicants by an agency that also determines the subsidy amount.

A common characteristic of these situations is that "there is an event which at each observation may or may not occur. If it does occur, associated with it will be a continuous, positive random variable. If it does not occur, this variable has a zero value" (Cragg, 1971). It seems then natural to model them in terms of the probability of the event and the expected value of the variable conditioned on the event, that is

$$E(y | x) = P(y > 0 | x) E(y | x, y > 0) \quad (1)$$

where all variables are conditioned on x , the equality follows from the law of iterated expectations (see, for example, Goldberger 1991), and we will refer to the second term of the decomposition as the conditional expectation of y .

This expression has several advantages. Firstly, it stresses the highly non-linear character of the expectation of y . Secondly, it points out that there are –at least- three different functions that can be of interest for the analyst: the expectation of y , the probability of the event, and the conditional expectation of y . Let us write the second and third functions using the notation $P(y > 0 | x) = p(x)$ and $E(y | x, y > 0) = h(x)$. Thirdly, it shows a useful way of modelling y conditional on x : to specify and estimate the functions $p(x)$ and $h(x)$.

Probability is often modelled by assuming an index function $I(x)$ and a random disturbance \mathbf{h} such as, when $I(x) + \mathbf{h} > 0$, y will take a positive value. In this case $p(x)$ is estimated by $F(I(x))$, where F represents the distribution function of \mathbf{h} . But more flexible less parametric ways of modelling $p(x)$ can also be employed. The conditional expectation of y , $h(x)$, can equally be specified in many ways. It is however convenient to note that, while the deviation term $v = y - E(y | x, y > 0)$ has zero mean by definition, $h(x)$ may be highly non-linear, requiring the inclusion of non-linear terms to obtain a good approximation.

^{*} Thanks are due to Manuel Arellano, who first convinced me. Errors, however, are mine.

The analyst, however, may be interested in having a “structural” model that explains how the probability of the event, the positive observations and their values are generated. A popular way of doing this is assuming that the non-zero values of y are the observed values of a variable y^* (sometimes called the latent variable), for which we do not observe the values associated to each x to which corresponds an observed zero-value of y . We will write the expectation of this variable as $E(y^* | x) = g(x)$, what constitutes in principle a fourth function of possible interest. Note, however, the different nature of this function: it is a purely theoretical construct. Then, the probability schedule that determines when y^* will be observed and when not must be specified.

This is mostly done by using the index schedule and assuming that, when $I(x) + \mathbf{h} > 0$, the variable y^* will be observed. When $I(x) + \mathbf{h}$ is specified to be the own y^* , with $I(x)$ being the expected value of y^* , $g(x)$, and assuming $\mathbf{e} = y^* - g(x)$ to be distributed normal, we have the Tobit model. When $I(x) + \mathbf{h}$ is supposed to be a different variable, z^* say, and (\mathbf{h}, \mathbf{e}) are assumed to follow a bivariate normal distribution with correlation parameter \mathbf{r} , we have the “selectivity” model (Heckman, 1979). With $\mathbf{r} = 0$ we have models considered by Cragg (1971).

The popularity of such models has favoured too often a mechanical use of these methods, but structural modelling makes sense when it conforms well with the theoretical structure of the problem at hand. Individuals are generally assumed to base their decisions on the values reached by objective functions (utility, profits, or their duals). It is then natural to suppose that utility or profits should reach a given level (or threshold) to trigger off the action and therefore they will show a range of values for which the action will be discarded.

The variable y^* may be seen constituting either an input for the level of the objective function, $u = u(y^*, x)$, or an optimal choice given other variables, $y^* = y^*(u(x), x)$. In both cases there will be a range of y^* values that correspond to the u values under which the action is discarded. It is in this context that the non-observed values of y can be attributed meaningful shadow values (shadow expenditures, prices, efforts...) when the action is not undertaken: the relevant values of y^* according to the underlying relations.

The assumptions on the observability rule for y^* may lead to a non-coincidence of the expectation of the latent variable y^* with the conditional expectation of y , that is, $E(y | x, y > 0) \neq E(y^* | x)$. This happens when the rule by which the variable is considered to be observed is defined either in terms of y^* or in terms of z^* with \mathbf{h} and \mathbf{e} correlated. The reason is that these cases involve an imperfect observation of censoring that imply a non-null expectation of the random disturbances \mathbf{e} corresponding to the observed values of y . Then we have

$$E(y | x, y > 0) = E(y^* | x) + E(\mathbf{e} | x, y > 0) \quad (2)$$

and hence $h(x) = g(x) + \mathbf{j}(x)$, where $\mathbf{j}(x)$ stands for the conditional expectation of the random disturbances. When this is the relevant model, it is clear that $g(x)$ may not be

obtained from the estimation of $h(x)$. Note, however, that this does not exclude the possibility of having models to explain a latent variable y^* in which $h(x) = g(x)$.

The several methods proposed in the literature as to how approaching the estimation of the functions involved in (1) and (2), must then be seen associated to different final objectives and set-up assumptions. If the focus is the estimation of a model for $E(y|x)$, the natural way seems the specification and separate estimation of $p(x)$ and $h(x)$ (see, for example, Duan, Manning, Morris and Newhouse, 1984). This has often been done using a probit model for $p(x)$ and the best linear predictor of $h(x)$, even if they can be used many alternatives. If the interest lies in a model for $E(y^*|x)$ that makes this expectation different from $h(x)$, a procedure must be found to estimate $g(x)$. The natural way is the use of a FIML procedure that takes into account the stochastic structure of the problem. A popular alternative is the two-step Heckman procedure, by which the estimates obtained from a probit estimation of $p(x)$ are used to construct the \mathbf{I} variable to be included in the second equation to split $h(x)$ in $g(x)$ and $\mathbf{j}(x)$ (Heckman, 1979).

In any case one obtains estimates of the functions $p(x)$ and $h(x)$ (in the second case $h(x)$ is estimated through the estimates of the sum $g(x) + \mathbf{j}(x)$), and hence the two methods can be used to estimate $E(y|x)$. In fact several Monte Carlo studies have focussed on the relative performance of the two methods, comparing their predictive performance, under the headings of “two-part” and “sample selection” models (see Leung and Yu, 1997). These studies reinforce the idea that each method must be used according to the objectives, but also illustrate the importance of the set-up. Sample selection models can reach better estimates than a simple version of two-part estimation when the true model is a selection model, but the alternative methods of estimation of the selection model show a low relative performance under a number of circumstances. In particular they perform poorly when there are few exclusion restrictions among the regressors of the probit and expectation equations, a high degree of censoring, a low variability among the regressors, or a large error variance in the choice equation.

References

- Cragg, J.G. (1971), “Some statistical models for limited dependent variables with application to the demand for durable goods”, *Econometrica*, 39,5, 829-844.
- Duan, N., W.G. Manning, C.N. Morris and J.P. Newhouse (1984), “Choosing between the sample selection model and the multi-part model”, *Journal of Business and Economics Statistics*, 2, 3, 283-289.
- Goldberger, A. (1991), *A course in Econometrics*, Harvard University Press.
- Heckman, J.J. (1979), “Sample selection bias as a specification error”, *Econometrica*, 47,1, 153-161.
- Leung, S.F. and Yu, S. (1996), “On the choice between sample selection and two-part models”, *Journal of Econometrics*, 72, 197-229.

Appendix

Let us illustrate the concepts with the often used specification of $p(x)$ as a probit model and $h(x)$ as an exponential function (y is assumed linear in logarithms). That is,

$$I(x) = x\mathbf{d}, \mathbf{h} \sim N(0,1) \quad (1)$$

$$\ln y = \begin{cases} x\mathbf{b} + \mathbf{e} & \text{if } x\mathbf{d} + \mathbf{h} > 0 \\ \text{undefined} & \text{otherwise} \end{cases} \quad (2)$$

where between \mathbf{d} and \mathbf{b} can be exclusion restrictions. We will make two alternative assumptions on \mathbf{e} , keeping for simplicity the same notation:

$$\mathbf{e} \sim N(0, \mathbf{s}^2) \quad (3a)$$

$$(\mathbf{h}, \mathbf{e}) \sim N(0, \Sigma), \text{ where } \Sigma = \begin{bmatrix} 1 & \mathbf{rs} \\ \mathbf{rs} & \mathbf{s}^2 \end{bmatrix} \quad (3b)$$

In both models, by (1), $p(x) = \Phi(x\mathbf{d})$, where Φ stands for the standard normal cdf. The difference lies in how equation (2) must be interpreted according to assumptions (3a) and (3b). In the first case, we are simply specifying $h(x) = \exp(x\mathbf{b})E(\exp(\mathbf{e})) = \exp(x\mathbf{b} + \frac{1}{2}\mathbf{s}^2)$ according to the assumption that (conditional) $\ln y$ is distributed around $E(\ln y | x, y > 0) = x\mathbf{b}$ with a normal disturbance. In the second we are specifying $h(x) = \exp(x\mathbf{b})E(\exp(\mathbf{e}) | \mathbf{h} > -x\mathbf{d})$, according to the idea that $\ln y^* = x\mathbf{b} + \mathbf{e}$ is a latent variable partially observed. It can be easily shown that $E(\exp(x) | \mathbf{h} > -x\mathbf{d}) = \exp(\frac{1}{2}\mathbf{s}^2)(\Phi(x\mathbf{d} + \mathbf{rs}) / \Phi(x\mathbf{d}))$. Then, defining $g(x) = \exp(x\mathbf{b} + \frac{1}{2}\mathbf{s}^2)$, in this log-linear case we have $h(x) = g(x)\mathbf{j}(x)$ with $\mathbf{j}(x) = \Phi(x\mathbf{d} + \mathbf{rs}) / \Phi(x\mathbf{d})$.

If equation (2) is estimated under assumption (3a) by OLS we obtain an estimate of $h(x)$. If equation (2) is estimated under assumption (3b), for example by a two-step Heckman procedure, we obtain other \mathbf{b} coefficients, this time representing the unconditional assumed relationship $g(x)$. The two sets of \mathbf{b} coefficients can be related by the well known ‘‘short’’ and ‘‘long-regression’’ formulas (see, for example Goldberger 1991). The estimates of $h(x)$ and $g(x)\mathbf{j}(x)$ will be in fact very similar, the difference being only originated in that the first estimate is a more linear approximation to the same expectation. The differences between the two estimations may be evaluated by using the respective predictions of $E(y | x)$

$$E(y | x) = \Phi(x\mathbf{d}) \exp(x\mathbf{b} + \frac{1}{2}\mathbf{s}^2) \quad (4a)$$

$$E(y | x) = \Phi(x\mathbf{d} + \mathbf{rs}) \exp(x\mathbf{b} + \frac{1}{2}\mathbf{s}^2) \quad (4b)$$

Relative performance in this prediction is what Monte Carlo studies have measured.