Specification of Non-Normal Double-Hurdle Models

M. D. Smith

Econometrics and Business Statistics, University of Sydney, Sydney NSW 2006, Australia
(Murray.Smith@econ.usyd.edu.au)

Abstract: In microeconometrics, expenditure data is typically zero-inflated due to many individuals recording, for one reason or another, no consumption expenditure. A two-part model can be appropriate for statistical analysis of such data, with the double-hurdle model (DHM hereafter) one specification that is frequently adopted in econometric practice. Essentially, the DHM is designed to explain individual demand through a joint decision process: a participation decision (first hurdle), and a consumption decision (second hurdle). The statistical model is constructed assuming an underlying bivariate distribution for the decisions, with most empirical studies based on a bivariate normal specification. A number of recent empirical studies do, however, suggest that the assumption of bivariate normality is too restrictive. In turn, these studies are themselves restrictive in the sense that their findings are based on only small departures from bivariate normality (*e.g.* Box-Cox transformations of bivariate normal DHM). It is apparent that practitioners need access to DHM that are based on underlying bivariate non-normality. This paper addresses this need by applying the copula method of construction of non-normal bivariate distributions to the DHM.

Keywords: Double-hurdle model; Bivariate distribution; Non-normal; Copula

1. INTRODUCTION

The double-hurdle model (DHM hereafter) proposed by Cragg [1971] has been used in microeconometrics to analyse a wide range of individual and household commodity demand. Important contributions to the DHM literature include Jones [1989], in which the demand for cigarettes is modelled, and Blundell and Meghir [1987], which was concerned with the labour supply of married women. Other fields in which the DHM has been applied include finance and sociology. The DHM has also been applied to infrequency-of-purchase contexts; *e.g.* Deaton and Irish [1984]. A bibliography of DHM applications appears in the survey by Smith [2001].

The DHM is designed to explain the mechanism of individual demand whereby an individuals decision process is decomposed into separate components that are jointly taken. These are: (i) a market participation decision (whether to buy or not), and (ii) a consumption level decision (how much to buy). Motivating this decomposition is the desire to allow different factors to influence demand; *e.g.* psychological influences may play a prominent role in determining participation, whereas economic considerations are more likely to be important in determining consumption. Pudney [1989, pp.160-162] gives a basis for the DHM in consumer choice theory.

The statistical construction of the DHM is based on assuming an underlying utility structure for both decision components, whose random outcomes are represented by a bivariate distribution. In this respect, the DHM does not differ from many other microeconometric constructions. Section 2 sets out the details. In particular, Cragg's DHM model is presented, for it represents the classical approach to DHM modelling. Foundered on bivariate normality, the popularity of Cragg's DHM permeates the literature. However, recent empirical literature has questioned the restrictiveness of assuming bivariate normality. This evidence originates from a class of DHM model obtained by transforming the observed random variable, termed Transformed DHM (TDHM hereafter). TDHM studies point toward the need for practitioners to access DHM that are based on bivariate non-normality. The remaining sections of the paper address this need by using the copula method of constructing non-normal bivariate distributions, and then demonstrating its application to the DHM.

2. CONSTRUCTION

The DHM is constructed by assuming the existence of a pair of latent variables designed to represent utilities: (i) the utility derived from market participation, denoted by

$$Y_1^{**} = y_1^{**} \in I\!R$$

and (ii) the utility derived from consumption, denoted by

$$Y_2^{**} = y_2^{**} \in I\!R$$

These random variables are then linked to expenditure

$$Y = y \geq 0$$

the latter being observable. The procedure is this - the utility variables are transformed to a pair of latent hurdle decision variables: (i) a participation hurdle

$$Y_1^{*} = 1\{Y_1^{**} > 0\}$$

($1\{A\}$ denotes the indicator function, taking value 1 if event A holds and 0 otherwise) where $Y_1^{*} = y_1^{*} \in \{0, 1\}$, and (ii) a consumption hurdle

$$Y_2^{*} = 1\{Y_2^{**} > 0\}\ Y_2^{**}$$

where $Y_2^{*} = y_2^{*} \geq 0$. The hurdle variables are then linked to expenditure as per

$$Y = Y_1^{*}\ Y_2^{*} \tag{1}$$

Let $F(y_1^{**}, y_2^{**})$ denote the joint cumulative distribution function (cdf) of the utilities Y_1^{**} and Y_2^{**}. The pdf of Y, denoted $f(y)$, is given by the discrete-continuous mixture:

$$\begin{cases} \frac{\partial}{\partial y}\left(F_2(y) - F(0, y)\right) & \text{if } y > 0 \\ F_1(0) + F_2(0) - F(0,0) & \text{if } y = 0 \end{cases} \tag{2}$$

where $F_i(\cdot)$ denotes the marginal cdf of Y_i^{**} $(i = 1, 2)$. With $f(y)$, we can, for example, derive the log-likelihood function.

2.1 Cragg's DHM

One very important illustration of the role of F is given by Cragg's DHM which sets F to

$$\Phi_2\left(y_1^{**} - x'\beta, \frac{y_2^{**} - w'\gamma}{\sigma}; \rho\right) \tag{3}$$

where $\Phi_2(\cdot, \cdot; \rho)$ denotes the cdf of the standard bivariate normal distribution with correlation coefficient ρ, vectors x and w denote regressor variables, and vectors β and γ, and scalar σ denote parameters. This bivariate normal specification has formed the basis of nearly all DHM studies to date. Jones [1992] gives the log-likelihood function when (3) is substituted into (2), for independent observations collected on Y.

2.2 Transformed DHM

Recent empirical studies adopt a more flexible form of DHM - the TDHM. For example, Yen [1993] and Jones and Yen [2000] study a particular TDHM corresponding to a Box-Cox transformation applied to Y assuming Cragg's specification (3) for F. TDHM are obtained by specifying a parametric transformation of Y, one that serves to alter the continuous component of the pdf of Y. The TDHM is obtained by replacing the left hand side of (1) by a parametric function $T(Y)$, that is:

$$T(Y) = Y_1^{*}\ Y_2^{*}$$

The function T is assumed positive-valued, non-decreasing, and differentiable in y for all $Y = y > 0$, and at the origin $T(0) = 0$. Under these conditions on T, only the continuous component of $f(y)$ alters (*i.e.* when $y > 0$), and is given by

$$\frac{\partial T(y)}{\partial y} \times \frac{\partial}{\partial y}\left(F_2(y) - F(0, y)\right)\Big|_{y \to T(y)}$$

This formula shows that the TDHM is obtained by scaling the DHM in which y is replaced by $T(y)$; the scaling factor is simply $\partial T(y)/\partial y$. However, while empirical TDHM studies, such as Yen's, report improved fits when compared to Cragg's DHM, the underlying specification of the joint cdf F was never varied from bivariate normality - the literature is still mired in bivariate normality. What the TDHM literature is clearly pointing toward is the need for practitioners to access DHM that are based on bivariate non-normality. The following section addresses this need.

3. NON-NORMAL DHM

3.1 Bivariate Specifications

The obvious approach when specifying the joint cdf F would be to choose a suitable bivariate distribution from amongst the many discussed in, for example, Kotz et al. [2000]. However, in light of the general functional form of the pdf of Y given in (2), with its interaction between joint and marginal distributions, this conventional approach can suffer from serious mathematical and computational difficulties. In particular, evaluating (mathematically and/or numerically) the following component of the pdf of Y :

$$\frac{\partial}{\partial y} F(0, y) \tag{4}$$

will potentially pose the most problems.

3.2 Independent Specifications

The presence of correlation between the utilities Y_1^{**} and Y_2^{**} is the root cause of the aformentioned difficulties. Consequently, another modelling strategy is to impose independence between Y_1^{**} and Y_2^{**}, leading to the pdf of Y of the form:

$$\begin{cases} (1 - F_1(0))\frac{\partial}{\partial y}F_2(y) & \text{if } y > 0 \\ F_1(0) + F_2(0) - F_1(0)F_2(0) & \text{if } y = 0 \end{cases}$$

which requires knowledge of only the marginal distributions. The justification for imposing independence is argued by Smith [1999], where the conditions under which the correlation coefficient is weakly identified in the DHM are examined. However, this parameter-reduction approach to modelling will struggle to appeal to many practitioners.

3.3 Copula Specifications

A further alternative is to use the theory of copulas; see Joe [1997] and Nelsen [1999]. This method separates the dependency between Y_1^{**} and Y_2^{**} from their respective marginal distributions through the use of a copula function C_θ. Under this approach the bivariate distribution of Y_1^{**} and Y_2^{**} is not specified, rather it is constructed by specifying the marginal distributions F_1 and F_2, along with a copula C_θ. The copula literally joins or couples a multivariate distribution function to its one-dimensional marginal distribution functions. In other words, the copula method builds joint distributions from specific marginals, thus it is the opposite of the conventional method which derives marginals from a specific joint distribution.

For marginals F_1 and F_2, and a copula C_θ, by Sklar's theorem (Nelsen [1999, p.15]) the joint cdf of Y_1^{**} and Y_2^{**} is given by:

$$F(y_1^{**}, y_2^{**}) = C_\theta(F_1(y_1^{**}), F_2(y_2^{**}))$$

where parameter θ represents dependency. Note especially that the marginal cdf's need not be of the same distributional type.

For real-valued $(u, v) \in (0, 1)^2$, Table 1 lists some examples of bivariate copula (attribution is as given by Joe [1997, chp.5]). The first copula - Normal - simply shows that the bivariate normal distribution falls within the class of distributions expressible in copula form (Φ_1^{-1} denotes the inverse of a standard normal cdf). The Morgenstern, Joe, Gumbel, Plackett and Frank copulas provide examples of one-parameter bivariate copula, in these cases dependency is parameterised by θ. For example, positive/negative θ in the Morgenstern copula realises positive/negative dependency, $\theta = 0$ yields independence. The Two-parameter copula allows for asymmetric dependence through the values of $\theta = (\theta_1, \theta_2)$ - in copula theory the ubiquitous Pearsonian definition of linear dependence is extended to cover other concepts of dependence such as concordance, relying on alternate measures of dependence such as Kendall's tau and Spearman's rho, for details see Joe [1997, chp.2] and Nelsen [1999, chp.5].

Table 1. Examples of bivariate copula $C_\theta(u, v)$.

Normal	$\Phi_2\left(\Phi_1^{-1}(u), \Phi_1^{-1}(v); \theta\right)$	for $-1 < \theta < 1$
Morgenstern	$uv(1 + \theta(1-u)(1-v))$	for $-1 < \theta < 1$
Joe	$1 - \left((1-u)^\theta + (1-v)^\theta - (1-u)^\theta(1-v)^\theta\right)^{1/\theta}$	for $1 \leq \theta < \infty$
Gumbel	$\exp\left(-((-\log(u))^\theta + (-\log(v))^\theta)^{1/\theta}\right)$	for $1 \leq \theta < \infty$
Plackett	$\frac{1}{2}\delta^{-1}\left(s - \sqrt{s^2 - 4\theta\delta uv}\right)$	for $0 \leq \theta < \infty$ where $s = 1 + \delta(u+v)$ and $\delta = \theta - 1$
Frank	$-\theta^{-1}\log\left(1 - \delta^{-1}(1 - e^{-\theta u})(1 - e^{-\theta v})\right)$	for $0 \leq \theta < \infty$ where $\delta = 1 - e^{-\theta}$
Two-parameter	$\left(1 + \left(\left(u^{-\theta_1} - 1\right)^{\theta_2} + \left(v^{-\theta_1} - 1\right)^{\theta_2}\right)^{1/\theta_2}\right)^{-1/\theta_1}$	for $\theta_1 > 0$ and $\theta_2 \geq 1$

For given (continuous) random variables R and S, with marginal cdf's $F_1(r)$ and $F_2(s)$, respectively, the chosen copula induces the joint pdf $f(r,s)$ as follows:

$$f(r,s) = \frac{\partial}{\partial r}\frac{\partial}{\partial s} C_\theta(F_1(r), F_2(s))$$

the right hand side of which can be expressed as:

$$f_1(r)f_2(s) \times \frac{\partial}{\partial u}\frac{\partial}{\partial v} C_\theta(u,v)\Big|_{u\to F_1(r), v\to F_2(s)}$$

where f_1 and f_2 are the marginal pdf's. To illustrate, Figure 1 plots the contours (maximum height at the origin) of the bivariate pdf induced by the indicated copula, where the marginals $R \sim N(0,1)$ and $S \sim N(0,1)$. The contour plots indicate that there is a considerable differences in appearance of the bivariate (non-normal) pdf's induced by each copula, even though the marginal distributions are identically standard normal. Depending on the copula, the bivariate pdf can exhibit any type of non-normal feature as may be desired, such as asymmetry (see Nelsen [1999, sec.2.7] for a discussion of concepts of symmetry in bivariate distributions), and short- and long-tails.

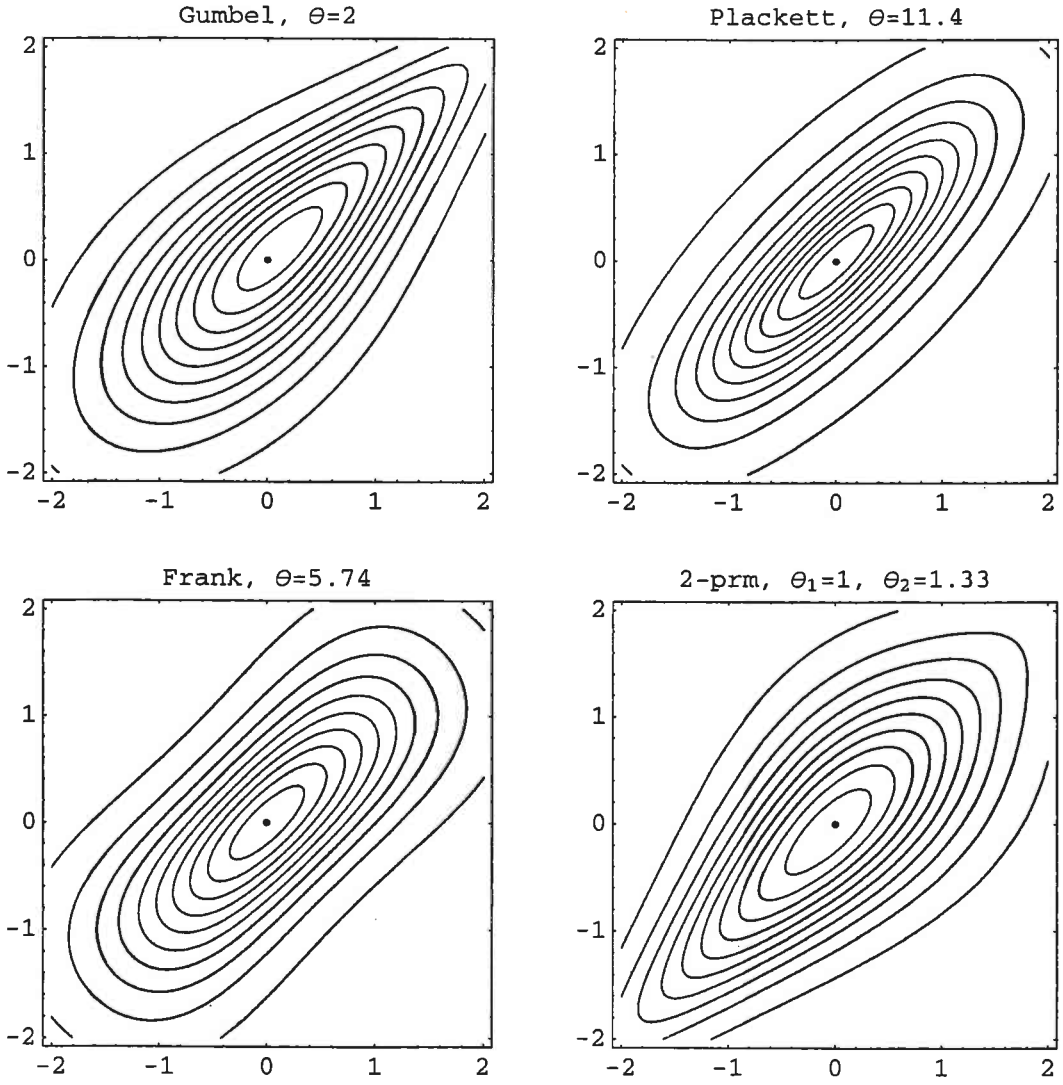

Figure 1. Bivariate pdf contour plots induced by copula, $N(0, 1)$ margins.

Returning to the DHM, the pdf of Y, see (2), under a copula C_θ becomes:

$$\begin{cases} \frac{\partial}{\partial y}\left(F_2(y) - C_\theta(F_1(0), F_2(y))\right) & \text{if } y > 0 \\ F_1(0) + F_2(0) - C_\theta(F_1(0), F_2(0)) & \text{if } y = 0 \end{cases}$$

which, to again stress the point, requires specification of only F_1, F_2 and C_θ. An alternative form for the continuous component ($y > 0$) of the pdf which can sometimes be useful is given by:

$$f_2(y)\left(1 - \frac{\partial}{\partial v} C_\theta(F_1(0), v)\Big|_{v \to F_2(y)}\right) \qquad (5)$$

where f_2 denotes the marginal pdf of Y_2^{**}. (5) will generally be much easier to evaluate than (4), both mathematically and numerically.

Finally, a specific example of a non-normal DHM designed to retain the same marginal normal distributions as Cragg's DHM, but one in which bivariate non-normality is induced through the use of a copula. Let

$$Y_1^{**} \sim N(x'\beta, 1) \quad \text{and} \quad Y_2^{**} \sim N(w'\gamma, \sigma^2)$$

and assign a Morgenstern copula (see Table 1). Under these assumptions, the continuous component of the pdf of Y is given by:

$$\sigma^{-1}\phi_1\left(\frac{y - w'\gamma}{\sigma}\right)\Phi_1(x'\beta) \times (1 - \theta\Phi_1(-x'\beta)(1 - 2\Phi_1\left(\frac{y - w'\gamma}{\sigma}\right)))$$

where ϕ_1 denotes the pdf and Φ_1 the cdf of a $N(0,1)$ variable. The probability mass located at $Y = 0$ is given by:

$$1 - \Phi_1(x'\beta)\Phi_1\left(\frac{w'\gamma}{\sigma}\right) - \theta\Phi_1(x'\beta)\Phi_1\left(\frac{w'\gamma}{\sigma}\right) \times \Phi_1(-x'\beta)\Phi_1\left(\frac{-w'\gamma}{\sigma}\right)$$

For independent observations $Y_1, ..., Y_n$, n_1 of which are strictly positive and n_0 of which are zero-valued ($n_1 + n_0 = n$), and corresponding regressors $(x_1, w_1), ..., (x_n, w_n)$, the log-likelihood function is given by

$$\sum_{i=1}^{n_1} \log f(y_i) + \sum_{i=1}^{n_0} \log \Pr(Y_i = 0)$$

Maximum likelihood estimation of parameters β, γ, σ^2 and θ would then proceed using numerical techniques.

When modelling, information criterion such as minimum AIC can be used to select amongst proposed copulas. Formal hypothesis tests can also be performed, utilising non-nested procedures such as Cox's test.

4. REFERENCES

Blundell, R., and C. Meghir, Bivariate alternatives to the Tobit model, *Journal of Econometrics*, 34, 179-200, 1987.

Cragg, J. G., Some statistical models for limited dependent variables with applications to the demand for durable goods, *Econometrica*, 39, 829-844, 1971.

Deaton, A., and M. Irish, Statistical models for zero expenditures in household budgets, *Journal of Public Economics*, 23, 59-80, 1984.

Joe, H., *Multivariate Models and Dependence Concepts*, Chapman and Hall, London, 1997.

Jones, A. M., A double-hurdle model of cigarette consumption, *Journal of Applied Econometrics*, 4, 23-39, 1989.

Jones, A. M., A note on computation of the double-hurdle model with dependence with an application to tobacco expenditure, *Bulletin of Economic Research*, 44, 67-74, 1992.

Jones, A. M., and S. T. Yen, A Box-Cox double-hurdle model, *The Manchester School*, 68, 203-221, 2000.

Kotz, S., N. Balakrishnan, and N. L. Johnson, *Continuous Multivariate Distributions*, volume 1, 2nd edition, Wiley, New York, 2000.

Nelsen, R. B., *An Introduction to Copulas*, Spinger-Verlag, New York, 1999.

Pudney, S., *Modelling Individual Choice: the Econometrics of Corners, Kinks, and Holes*, Basil Blackwell, London, 1989.

Smith, M. D., Should Dependency be Specified in Double-Hurdle Models?, *Proceedings of the International Congress on Modelling and Simulation*, Vol 2, by L. Oxley, F. Scrimgeour, and M. McAleer (eds.), University of Waikato, Hamilton, New Zealand, 277-282, 1999.

Smith, M. D., On specifying double-hurdle models, forthcoming in *Handbook of Applied Econometrics and Statistical Inference*, by A. Ullah, A. Wan, and A. Chaturvedi (eds.), Marcel-Dekker, 2002.

Yen, S. T., Working wives and food away from home: the Box-Cox double hurdle model, *American Journal of Agricultural Economics*, 75, 884-895, 1993.