Should Dependency be Specified in Double-Hurdle Models?

Murray D. Smith, Department of Econometrics, University of Sydney

Abstract: In microeconometrics, consumption data is typically zero-inflated due to many individuals recording no consumption. A mixture model can be appropriate for statistical analysis of such data, with the Dependent Double-Hurdle model (DDH hereafter) one specification that is frequently adopted in econometric practice. Essentially, the DDH model is designed to explain individual demand through a sequential two-step process: a market participation decision (first hurdle), followed by a consumption level decision (second hurdle) - specification of a non-zero correlation/covariance parameter in the underlying bivariate utility structure allows for dependency between the hurdles. A significant feature of the majority of empirical DDH studies has been the lack of support for the existence of dependency. This empirical phenomenon is studied from a theoretical perspective using examples based on the bivariate normal, bivariate logistic, and bivariate Poisson distributions. The Fisher Information matrix for the parameters of the model is considered, especially the component corresponding to the dependency parameter. The main finding is that the DDH model contains too little statistical information to support estimation of dependency, even when dependency is truly present. Consequently, the paper calls for the elimination of attempts to estimate dependency using the DDH framework. The advantage of this strategy is that it permits modelling based on flexible distributional structures, for in the absence of dependency the underlying variables are independent. Two approaches to model construction are explored: (i) models resulting from specifications for the underlying utility variables, and (ii) models resulting from specifications for the hurdle variables.

1 Introduction

The Double-Hurdle model (DH hereafter) has been used in economics to analyse a wide range of individual commodity demand and labour supply behaviour. In terms of commodity demand, the DH model is designed to explain the mechanism of individual demand through a sequential two-step process: a market participation decision (first hurdle), followed by a consumption level decision (second hurdle). The statistical origins of the model are due to Cragg [3], and its basis in consumer choice theory is due to Pudney [9, pp.160-162].

The generalisation of the DH model to allow for dependence between the participation and consumption decisions - the Dependent Double-Hurdle model (DDH hereafter) - has recently been the subject of empirical attention. Importantly, the arguments mounted for this generalisation have not been based on economic theory. Rather, justification has been based on intuitive behavioural grounds, and on statistical grounds. In statistical terms, a parameter θ, representing dependency, is incorporated into the DH model. Typically, a DDH model nests its DH counterpart through the restriction $\theta = 0$.

A summary of a number of published DDH studies appears in Table I. The entries in the last column - "DDH vs DH" - indicate whether the fitted DDH is either insignificant from (insig), or significantly different to (sig), its nested DH version. The relevant hypothesis test showed that the data in the majority of the studies did not support the DDH model over the DH model at any conventional level of significance. The persistent finding against the DDH model provides the motivation for this paper - an explanation is sought for why DDH models appear to be statistically indistinguishable from their nested DH counterparts.

Table I: Dependent Double-Hurdle Studies

	Application: Demand for	Sample size	% of 0's	DDH vs DH
Blaylock and Blisard [1]	Cigarettes (USA)	2962	60.7	insig
Burton *et al.* [2]	Meat (UK)	2144	6.3	insig
Gao *et al.* [4]	Rice (USA)	4273	67.0	insig & sig
Garcia and Labeaga [5]	Cigarettes (Spain)	23669	41.2	insig
Gould [6]	Cheese (USA)	5017	59.0	sig
Jones [7]	Cigarettes (UK)	1573	na	insig
Jones [8]	Cigarettes (UK)	2321	48.5	insig
Yen and Jones [10]	Cheese (USA)	4245	18.1	insig

The DH and DDH models are members of the class of hierarchical limited and qualitative dependent variable models. This class of model often suffers parameter identification problems, although detection of this typically surfaces only when attempting to compute parameter estimates. However, by focusing on the distribution of the DDH model (as opposed to secondary issues such as the properties of estimators and test statistics used in the model), the problem of *weakly identified parameters* is shown to be present (section 3).

In the paper's remaining sections, the extent of the identification problem is quantified using *Fisher's information* to measure the amount of statistical information on the parameters of the model. Section 4 focuses on Fisher's information on θ, while in section 5, Fisher information matrices are inspected.

The main suggestion of the paper, is that the introduction of dependency into DH models (thereby yielding a DDH model) is a statistically spurious generalisation - the DDH model adds little to the informational content of its nested DH counterpart. This seemingly negative outcome is, however, of considerable benefit to practitioners. In the absence of dependency, there is greater opportunity to explore more flexible distributional forms. Section 6 concludes with a proposal on this theme.

2 Statistical Construction

Begin by defining Y_1^{**} as the utility derived by an individual from market participation, and Y_2^{**} the utility derived by an individual from consumption. Assume (for now) that these variables are continuous, and real-valued. Next, assume a parametric bivariate model for (Y_1^{**}, Y_2^{**}) is specified by assigning a joint cumulative distribution function (cdf), denoted by $F(y_1^{**}, y_2^{**})$, for real-valued pairs (y_1^{**}, y_2^{**}). The cdf depends upon unknown parameters, one in particular being the dependency parameter θ. Importantly, variables Y_1^{**} and Y_2^{**} are not observed. The observed variable is individual consumption $Y \geq 0$. The relationship between (Y_1^{**}, Y_2^{**}) and Y is established by defining the hurdle variables:

$$Y_1^* = 1\{Y_1^{**} > 0\}, \quad Y_2^* = 1\{Y_2^{**} > 0\}\, Y_2^{**},$$

where $1\{A\}$ is the indicator function, taking value 1 if event A holds and 0 otherwise. Y_1^* represents the first hurdle decision, and Y_2^* represents the second hurdle consumption. In general, Y_1^* and Y_2^* are latent. Finally, to complete the construction of the DDH model, individual consumption

$$Y = Y_1^*\, Y_2^*.$$

Due to the sequential decomposition of the decision, a zero observation on Y can occur in two ways: (i) if the first hurdle is not passed ($Y_1^* = 0$), or (ii) if the first hurdle is passed but the second hurdle is not ($Y_1^* = 1$ *and* $Y_2^* = 0$). Any positive-valued observation occurs only when both hurdles are passed ($Y_1^* = 1$ *and* $Y_2^* > 0$).

Under continuity, the probability density function (pdf) of Y is a continuous-discrete mixture, with functional form depending upon the specification assumed for F. Denote it by

$$f(y) = \begin{cases} f_+(y) & \text{if } y > 0 \\ f_0 & \text{if } y = 0. \end{cases}$$

When $y > 0$, the $f_+(y)$ component may be derived as follows:

$$\begin{aligned} f_+(y) &= \frac{\partial}{\partial y} \Pr(Y \leq y) \\ &= \frac{\partial}{\partial y} \left(F_2(y) - F(0, y)\right), \end{aligned}$$

where $F_i(\cdot)$ denotes the marginal cdf of Y_i^{**} $(i = 1, 2)$. When $y = 0$, the f_0 component is the probability mass at the origin:

$$\begin{aligned} f_0 &= P(Y = 0) \\ &= F_1(0) + F_2(0) - F(0, 0). \end{aligned}$$

3 Bivariate Normal DDH

In this first example, assume (Y_1^{**}, Y_2^{**}) is distributed according to the following bivariate normal:

$$\begin{bmatrix} Y_1^{**} \\ Y_2^{**} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} 1 & \sigma\theta \\ \sigma\theta & \sigma^2 \end{bmatrix} \right).$$

Without loss of generality, $\mathrm{Var}(Y_1^{**})$ is normalised to unity because in the construction of the DDH model all scale information on Y_1^{**} is lost due to the transformation of Y_1^{**} to Y_1^*. As is well known, the dependency parameter θ is equivalent to the correlation coefficient between Y_1^{**} and Y_2^{**}. There are four parameters in the model $(\mu_1, \mu_2, \sigma^2, \theta)$.

For the bivariate normal DDH model, the joint cdf of (Y_1^{**}, Y_2^{**}) is given by:

$$F(y_1^{**}, y_2^{**}) = \Omega\left(y_1^{**} - \mu_1, \frac{y_2^{**} - \mu_2}{\sigma}; \theta\right),$$

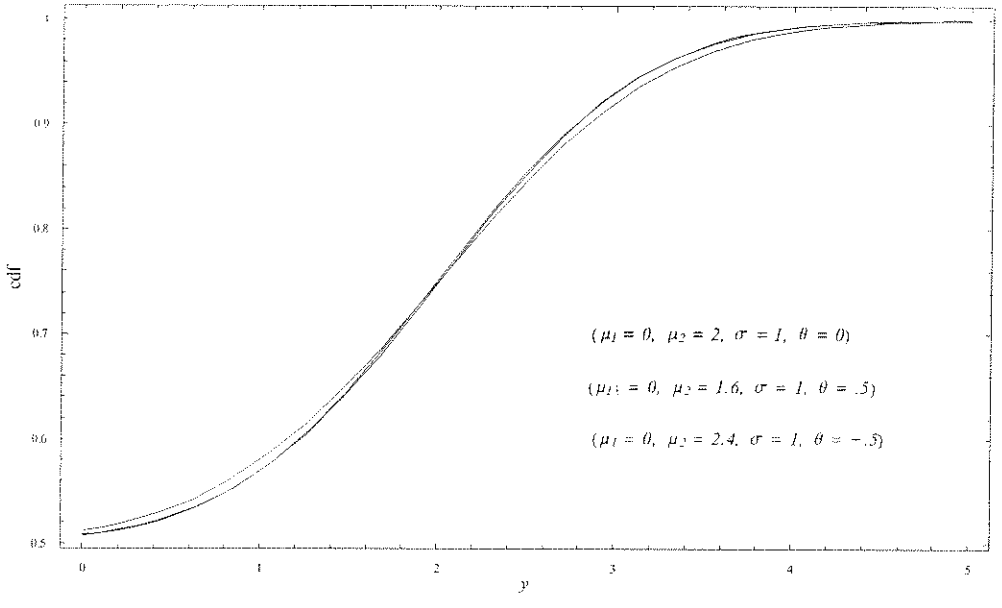
where $\Omega(\cdot, \cdot; \theta)$ denotes the cdf of a standardised bivariate normal distribution with correlation coefficient θ. The cdf of Y, $\Pr(Y \leq y)$, is given by: $\Phi(-\mu_1) + \Phi(z) - \Omega(-\mu_1, z; \theta)$ if $y > 0$,

and $\Phi(-\mu_1)+\Phi(-\sigma^{-1}\mu_2)-\Omega(-\mu_1,-\sigma^{-1}\mu_2;\theta)$ if $y=0$. Here $\Phi(\cdot)$ denotes the cdf of a standardised univariate normal distribution, and $z=(y-\mu_2)/\sigma$.

In Figure I, distributions of Y are plotted for three far-apart points in the parameter space; namely, at $(\mu_1,\mu_2,\sigma^2,\theta)=(0,2,1,0)$, $(0,1.6,1,0.5)$, $(0,2.4,1,-0.5)$. The significant feature to notice, is that the three distributions are almost indistinguishable across the support of Y. Certainly, the fact that the distributions are not identical, is sufficient to identify the parameters of the bivariate normal DDH model; however, what Figure I reveals is that identification is weak in the selected neighbourhoods of the parameter space. Estimation of demonstrably weakly-identified parameters must be of considerable concern, for it can lead to computational problems such as lack of convergence - this in fact occurred in the Burton *et al.* DDH study, see [2, p.205].

In the following examples, attempts to quantify the implications of weak identification in the DDH model are undertaken using *Fisher's information*, a well-known measure of statistical information.

Figure I: Distributions of Y (bivariate normal DDH)



4 Bivariate Logistic DDH

For this example, assume (Y_1^{**},Y_2^{**}) is distributed according to Gumbel's Type II bivariate logistic distribution with cdf $F(y_1^{**},y_2^{**})$ equal to:

$$F(y_1^{**})F(y_2^{**})(1+\theta(1-F(y_1^{**}))(1-F(y_2^{**}))),$$

for real-valued pairs (y_1^{**},y_2^{**}). The notation

$$F(y_i^{**})=(1+\exp(-(y_i^{**}-\mu_i)))^{-1}$$

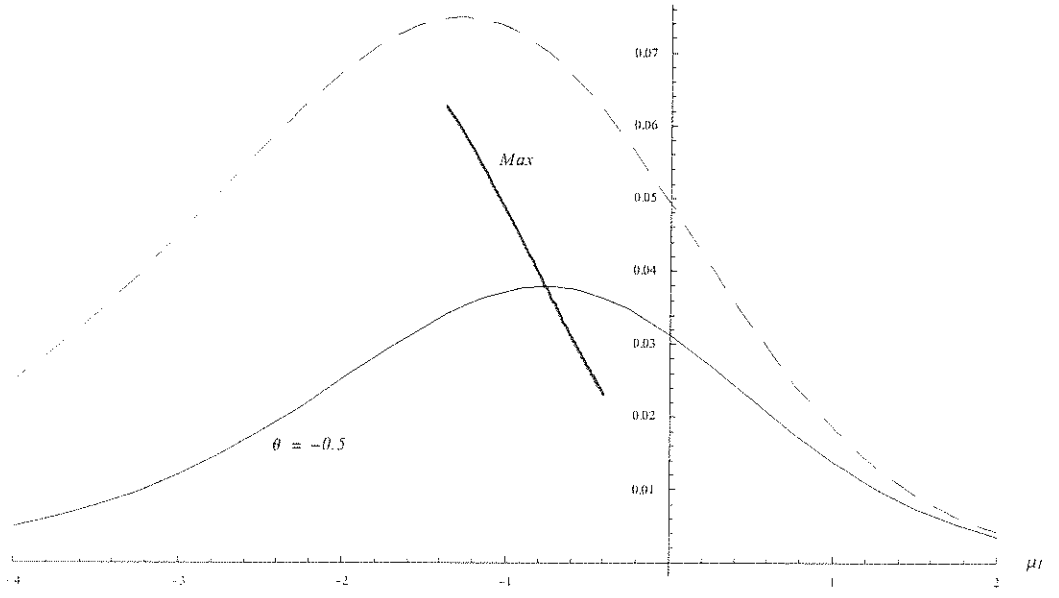
$(i=1,2)$, corresponds to the cdf of a logistic random variable with mean μ_i and variance $\pi^2/3$. Also, the dependency parameter θ is such that $-1<\theta<1$, moreover, in this model it is equivalent to the covariance between Y_1^{**} and Y_2^{**}. The parameters of the model are (μ_1,μ_2,θ).

For this specification, Fisher's information on θ, $i=E\left(\frac{\partial}{\partial\theta}\log f(Y)\right)^2$, may be derived in closed form. Figure II plots i against values of μ_1 in the interval $[-4,2]$. The dashed curve bounds Fisher's information on θ, for all μ_2 and θ in the parameter space. The solid curve depicts Fisher's information on θ, setting $\mu_2=1$ and $\theta=-0.5$. Significantly, the plot provides evidence for the *claim*:

$$\left.\frac{\partial i}{\partial\mu_1}\right|_{\mu_1=0}<0 \quad \text{for all } \theta \text{ and } \mu_2,$$

implying that i is maximised for some $\mu_1<0$, whatever the value of θ and μ_2.

Figure II: Fisher's information on θ (bivariate logistic DDH)



The claim is further evidenced by the thicker line, labelled *Max*, which traverses through all maximums of i for every θ such that $-1 < \theta < 1$, where μ_2 is fixed at unity; this line is situated entirely over negative values of μ_1. If the claim is true, then statistical information on θ will be maximised only when more than half of the population do not participate in the market. In other words, when sampling from a population, if more than 50% of respondents announce zero consumption, then it is under these conditions that we are best-placed to perform inference on θ, for this is precisely the situation when the amount of statistical information present on θ can be at its greatest. This feature is seen empirically, with Gao *et al.* [4] reporting 67% of their sample with zero consumption, and for Gould [6] the proportion was 59% - both of these DDH studies report significant dependency parameter estimates (see Table I).

5 Bivariate Poisson DDH

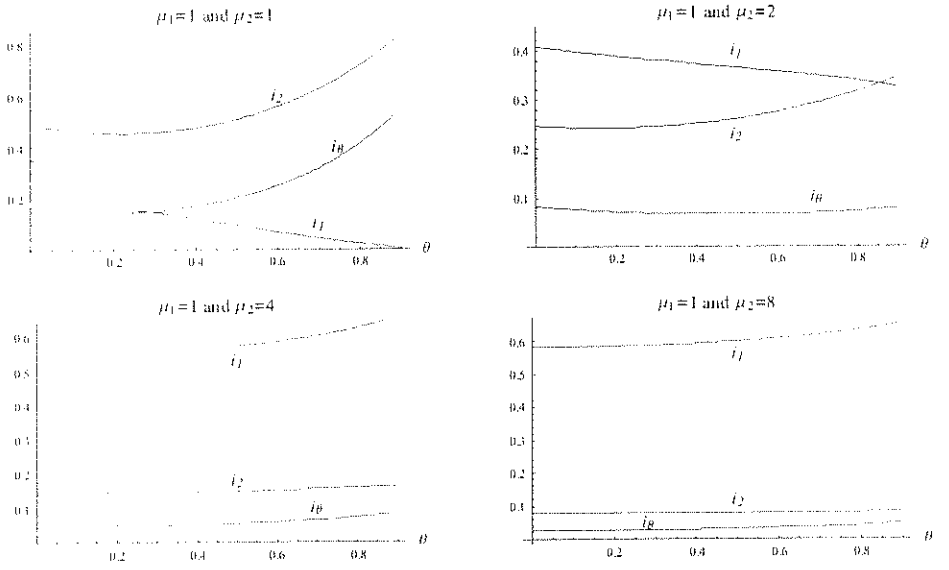
In the previous example, attention focused solely on Fisher's information on the dependency parameter. Of course, DDH models will, in general, contain parameters in addition to the dependency parameter. Accordingly, in this example the impact of dependency on all DDH parameters is examined by inspecting elements of Fisher's information matrix, where (Y_1^{**}, Y_2^{**}) is assumed distributed according to Holgate's bivariate Poisson distribution. The marginal pdf is such that $Y_i^{**} \sim \text{Poisson}(\mu_i)$, while θ $(0 \leq \theta < \min(\mu_1, \mu_2))$ is the covariance between Y_1^{**} and Y_2^{**}.

Figure III gives four plots, against values of θ, of Fisher's information on parameters μ_1, μ_2, and θ, denoted respectively by i_1, i_2, and i_θ. These measures correspond to the elements on the leading diagonal of Fisher's information matrix: $E\left(\frac{\partial}{\partial\alpha}\log f(Y) \times \frac{\partial}{\partial\alpha'}\log f(Y)\right)$ where the column vector $\alpha = (\mu_1, \mu_2, \theta)'$. For the three plots, the true value of μ_1 is fixed at unity, whereas μ_2 is assigned value 1, 2, 4, and finally 8. Note the differing vertical scales.

In the first plot ($\mu_1 = \mu_2 = 1$), it is apparent that the statistical information associated with θ increases with the true value of θ. This is a positive finding, and one that accords with intuition. However, also evident in the plot is the trade-off between Fisher information on μ_1 and θ, the former virtually disappearing as θ increases. There appears here to be a "competition" amongst the parameters for statistical information.

In the second plot ($\mu_1 = 1$ and $\mu_2 = 2$), the trade-off in Fisher information is still in evidence, however, the magnitudes of i_1 and i_2 are such that neither vanishes as the true value of θ increases. Nevertheless, the situation in respect of Fisher Information on θ has worsened considerably: i_θ remains fairly constant and fairly small. There is little statistical information on

Figure III: Fisher's information on μ_1, μ_2, and θ (bivariate Poisson DDH)



θ present in the model, irrespective of its true value.

In the remaining plots there is a vast disparity between the Fisher information on μ_1, and that of μ_2 and θ. Both plots clearly demonstrate that there is barely no Fisher information on θ, irrespective of the true value of θ. Given the scarcity of statistical information on θ, there seems little chance of data (except perhaps if it is collected in very large quantity) being able to reliably estimate θ, much less it being able to support the dependency hypothesis *even when θ truly is non-zero*.

The trade-off in statistical information evidenced in this example, suggests that DDH models are over-parameterised. Incorporating a dependency parameter into a DH model (yielding a DDH model), while perhaps justified on behavioural grounds, manages only to expose a statistical weakness in the DDH model.

There is an alternative viewpoint here, one which may be seen in the approach of Gao *et al.* [4]. In that study, estimation of a bivariate normal DDH model returned an insignificant estimate of θ. The authors then proceeded to specify a second DDH model, based on an inverse hyperbolic sine transformation of Y_i which ultimately returned a significant estimate of θ. Electing, in light of the results of this paper, to ascribe insignificance in the first DDH model to the difficulties caused by weak identification, then this can be "overcome" by inducing sufficient non-linearity into the likelihood function. Unfortunately, non-linear transformations (such as the one used by Gao *et al.*) may manage to hide weak identification in a shower of parameters, but it typically comes at a cost of violating the principle of parsimony.

6 Remarks

6.1 Summary

Taken as a whole, the results of this paper demonstrate that the DDH model represents a *spurious statistical generalisation* of the DH model. The economic underpinnings of the model are not affected by this conclusion, nor does it invalidate the behavioural arguments mounted to justify the DDH model over the DH model. It is the statistical nature of the DDH model which is deficient. This has manifested itself in the empirical literature, with most studies being unable to support the existence of the dependency parameter, and it has been studied in this paper under ideal theoretical circumstances through means of Fisher's information.

In practice, knowing the true DDH model is no longer the luxury it has been here. The indicator - an excessive proportion of zeros in the data - may provide favourable evidence to justify fitting a DDH model, but taken in the broader perspective of all parameters in the model, it may be a costly strategy. To

the extent that mean/regression parameters are usually of greater importance in estimation, it would appear safer to ignore dependency altogether and specify a DH model, the statistical information in the data can then reveal as much about these parameters as is possible.

6.2 Other Modelling Strategies

The seemingly negative conclusion to the paper is, however, a boon to practitioners for it allows far more flexible distributional structures to be employed for the models' random variables. To see this, suppose a DH model is to be fitted. By construction, the underlying decision utility variables are independent, in which case $F(y_1^{**}, y_2^{**}) = F_1(y_1^{**})F_2(y_2^{**})$, and the pdf of Y becomes:

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

Now only univariate distributions F_1 and F_2 are required. This approach to DH modelling is well-known.

Now the previous construction focused on the relationship between (Y_1^{**}, Y_2^{**}) and Y. Evidently there is a second, less-explored possibility - one that emphasizes specification of distributions for the hurdle variables Y_1^* and Y_2^*. Of course, specifying a distribution for Y_1^* is easy, it must be Bernoulli distributed:

$$\Pr(Y_1^* = y_1^*) = (1 - r)^{1 - y_1^*} r^{y_1^*},$$

where y_1^* takes values 0 and 1, and real-valued r is such that $0 \leq r \leq 1$. The success probability r may depend on parameters and covariates, and can be parameterised with any function whose range is $(0, 1)$; *e.g.*, the cdf of the normal distribution yields the familiar probit, but possibly more flexible would be the pdf of a beta distribution. For the second hurdle variable Y_2^*, assume, for the moment, that it is observable. Those observations would give Y_2^* the appearance of being zero-inflated, hence it would be natural to specify a distribution for it from amongst this class.

To illustrate the construction of the second approach, suppose that the pdf of Y_2^* is given by:

$$g(y_2^*) = \begin{cases} g_+(y_2^*) & \text{if } y_2^* > 0 \\ g_0 & \text{if } y_2^* = 0, \end{cases}$$

for suitable functions g_+ and g_0, both of which may depend on parameters and covariates. Following the steps outlined in section 2, the pdf of observed consumption Y is:

$$f(y) = \begin{cases} r g_+(y) & \text{if } y > 0 \\ 1 - r + r g_0 & \text{if } y = 0. \end{cases}$$

It is a subject of future research to contrast the performance of DH models based on these two approaches to modelling.

References

[1] Blaylock, J. R., and Blisard, W. N. (1992), "U.S. cigarette consumption: the case of low-income women", *American Journal of Agricultural Economics*, **74**, 698-705.

[2] Burton, M., Tomlinson, M., and Young, T. (1994), "Consumers' decisions whether or not to purchase meat: a double hurdle analysis of single adult households", *Journal of Agricultural Economics*, **45**, 202-212.

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

[4] Gao, X. M., Wailes, E. J., and Cramer, G. L. (1995), "Double-hurdle model with bivariate normal errors: an application to U.S. rice demand", *Journal of Agricultural and Applied Economics*, **27**, 363-376.

[5] Garcia, J., and Labeaga, J. M. (1996), "Alternative approaches to modelling zero expenditure: an application to Spanish demand for tobacco", *Oxford Bulletin of Economics and Statistics*, **58**, 489-506.

[6] Gould, B. W. (1992), "At-home consumption of cheese: a purchase-infrequency model", *American Journal of Agricultural Economics*, **72**, 453-459.

[7] Jones, A. M. (1989), "A double-hurdle model of cigarette consumption", *Journal of Applied Econometrics*, **4**, 23-39.

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

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

[10] Yen, S. T., and Jones, A. M. (1997), "Household consumption of cheese: an inverse hyperbolic sine double-hurdle model with dependent errors", *American Journal of Agricultural Economics*, **79**, 246-251.