# Multiple Hurdle Models in R: The mhurdle Package

**Fabrizio Carlevaro**
Université de Genève

**Yves Croissant**
Université de la Réunion

**Stéphane Hoareau**
Université de la Réunion

**Abstract**

**mhurdle** is a package for R enabling the estimation of a wide set of regression models where the dependent variable is left censored at zero, which is typically the case in household expenditure surveys. These models are of particular interest to explain the presence of a large proportion of zero observations for the dependent variable by means of up to three censoring mechanisms, called hurdles. For the analysis of censored household expenditure data, these hurdles express a good selection mechanism, a desired consumption mechanism and a purchasing mechanism, respectively. **mhurdle** models are specified in a fully parametric form and estimated using the maximum likelihood method for random samples. Model evaluation and selection are tackled by means of goodness of fit measures and Vuong tests. Software rationale and user's guidelines are presented and illustrated with actual examples.

*Keywords*: households' expenditure survey analysis, censored regression models, hurdle models, maximum likelihood estimation, nonlinear goodness of fit measures, Vuong tests for model selection, R.

## 1. Introduction

Data collected by means of households' expenditure survey may present a large proportion of zero expenditures due to many households recording, for one reason or another, no expenditure for some items. Analyzing these data requires to model any expenditure with a large proportion of nil observations as a dependent variable left censored at zero.

Since the seminal paper of Tobin (1958), a large econometric literature has been developed to deal correctly with this problem of zero observations. The problem of censored data has been treated for a long time in the statistics literature dealing with survival models which are implemented in R with the **survival** package of Therneau and Lumley (2008). It has also close links with the problem of selection bias, for which some methods are implemented in the **sampleSelection** package of Toomet and Henningsen (2008b). It is also worth mentioning that a convenient interface to `survreg`, called `tobit`, particularly aimed at econometric applications is available in the **AER** package of Kleiber and Zeileis (2008).

In applied microeconometrics, different decision mechanisms have been put forward to explain the appearance of zero expenditure observations. The original Tobin's model takes only one of these mechanisms into account. With **mhurdle**, up to three mechanisms generating zero expenditure observations may be introduced in the model[1]. More specifically, we consider the

[1]This package has been developed as part of a PhD dissertation carried out by Stéphane Hoareau (2009) at the University of La Réunion under the supervision of Fabrizio Carlevaro and Yves Croissant.

following three zero expenditure generating mechanisms.

**A good selection mechanism (hurdle 1)** . According to this mechanism, the consumer first decides which goods to include in its choice set and, as a consequence, he can discard some marketed goods because he dislikes them (like meat for vegetarians or wine for non-drinkers) or considers them harmful (like alcohol, cigarettes, inorganic food, holidays in dangerous countries), among others.
This censoring mechanism has been introduced in empirical demand analysis by Cragg (1971). It allows to account for the non-consumption of a good as a consequence of a fundamentally non-economic decision motivated by ethical, psychological or social considerations altering the consumer's preferences.

**A desired consumption mechanism (hurdle 2)** . According to this mechanism, once a good has been selected, the consumer decides which amount to consume and, as a consequence of his preferences, resources and selected good prices, its rational decision can turn out to be a negative desired consumption level leading to a nil consumption.
The use of this mechanism to explain the presence of zero observations in family expenditure surveys introduced by Tobin (1958). Its theoretical relevance has been later rationalised by the existence of corner solutions to the microeconomic problem of rational choice of the neoclassical consumer. See section 10.2 of Amemiya (1985), for an elementary presentation of this issue, and chapter 4 of Pudney (1989), for a more comprehensive one.

**A purchasing mechanism (hurdle 3)** . According to this mechanism, once a consumption decision has been taken, the consumer sets up the schedule at which to buy the good and, as a consequence of its purchasing strategy, zero expenditure may be observed if the survey by which these data are collected is carried out over a too short period with respect to the frequency at which the good is bought.
This censoring mechanism has been introduced in empirical demand analysis by Deaton and Irish (1984). It allows to account for the non-purchase of a good not because the good is not consumed but because it is a durable or a storable good infrequently bought. By the same token, this mechanism allows to derive from observed expenditures, the rate of use of a durable good or the rate of consumption of a stored non durable good.

For each of these censoring mechanisms, a continuous latent variable is defined, indicating that censoring is in effect when the latent variable is negative. These latent variables are modelled as the sum of a linear combination of explanatory variables and of a normal random disturbance with a possible correlation between the disturbances of different latent variables. By combining part or the whole set of these censoring mechanisms, we generate a set of non-nested parametric models that can be used to explain censored expenditure data depending on the structural censoring mechanisms that a priori information suggests to be at work.

It is worth mentioning that, although this formal model has been primarily developed to deal with censored household expenditure data, its practical scope is not restricted to empirical demand analysis. A quite natural other area of application is represented by the empirical analysis of labour supply. In this context, hurdle 1 can indeed be reinterpreted as a non-economic mechanism of labour market participation; hurdle 2 as a desired working hours mechanism based on the neoclassical model of labour supply that can generate negative

desired working hours leading to a nil labour supply; hurdle 3 as an unemployment mechanism explaining zero hours worked as a result of spells of unemployment. Note also that even within the realm of demand analysis, the economic interpretation of hurdles 1, 2 and 3 may require to be adapted to the specific features of available data, as we illustrate by an empirical application presented at the end of the paper (see section 5).

Our hurdle models are specified as fully parametric models allowing estimation and inference within an efficient maximum likelihood framework. In order to identify a relevant model specification, goodness of fit measures for model evaluation and selection, as well as Vuong tests for discriminating between nested, strictly non nested and overlapping models have been implemented in **mhurdle** package. Vuong tests remarkably permit to compare two competing models when both, only one, or neither of them contain the true mechanism generating the sample of observations. More precisely, such tests allow to assess which of the two competing models is closest to the true unknown model according to the Kullback-Leibler information criterion. Therefore, such symmetric tests are not intended, as classical Neyman-Pearson tests, to pinpoint the chimeric true model, but to identify a best parametric model specification (with respect to available observations) among a set of competing specifications. As a consequence, they can provide inconclusive results, which prevent from disentangling some competing models, and when they are conclusive, they don't guarantee an identification of the relevant model specification.

The paper is organised as follows: Section 2 presents the rationale of our modelling strategy. Section 3 presents the theoretical framework for model estimation, evaluation and selection. Section 4 discusses the software rationale used in the package. Section 5 illustrates the use of **mhurdle** with several examples. Section 6 concludes.

# 2. Modelling strategy

## 2.1. Model specification

Our modelling strategy is intended to model the level $y$ of expenditures of a household for a given good or service during a given period of observation. To this purpose, we use up to three zero expenditure generating mechanisms, called hurdles, and a demand function.

Each hurdle is represented by a probit model resting on one of the following three latent dependent variables relations:

$$\left\{\begin{array}{l} y_1^* = \beta_1^\top x_1 + \epsilon_1 \\ y_2^* = \beta_2^\top x_2 + \epsilon_2 \\ y_3^* = \beta_3^\top x_3 + \epsilon_3 \end{array}\right. \tag{1}$$

where $x_1$, $x_2$, $x_3$ stand for column-vectors of explanatory variables (called covariates in the followings), $\beta_1$, $\beta_2$, $\beta_3$ for column-vectors of the impact coefficients of the explanatory variables on the continuous latent dependent variables $y_1^*$, $y_2^*$, $y_3^*$ and $\epsilon_1$, $\epsilon_2$, $\epsilon_3$ for normal random disturbances.

- Hurdle 1 models the household decision of selecting or not selecting the good we consider as a relevant consumption good, complying with household's ethical, psychological and

social convictions and habits. This good selection mechanism explains the outcome of a binary choice that can be coded by a binary variable $I_1$ taking value 1 if the household decides to enter the good in its basket of relevant consumption goods and 0 otherwise. The outcome of this binary choice is modelled by associating the decision to select the good to positive values of the latent variable $y_1^*$ and that to reject the good to negative values of $y_1^*$. Therefore, good selection or rejection is modelled as a probability choice where selection occurs with probability $P(I_1 = 1) = P(y_1^* > 0)$ and rejection with probability $P(I_1 = 0) = P(y_1^* \leq 0) = 1 - P(y_1^* > 0)$.
Note that if this mechanism is inoperative, this probit model must be replaced by a singular probability choice model where $P(I_1 = 1) = 1$ and $P(I_1 = 0) = 0$.

- Hurdle 2 models the household decision of consuming or not consuming the selected good, given its actual economic conditions. This desired consumption mechanism explains the outcome of a binary choice coded by a binary variable $I_2$ taking value 1 if the household decides to consume the good and 0 otherwise. The outcome of this binary choice is modelled by associating the decision to consume the selected good to a positive value of its desired consumption level, represented by the latent variable $y_2^*$, and that of not to consume the good to negative values of $y_2^*$. Therefore, when this zero expenditure generating mechanism is operative, it also models the level of desired consumption expenditures by means of a Tobit model identifying the desired consumption expenditures to the value of latent variable $y_2^*$, when it is positive, and to zero, when it is negative.
Conversely, when the desired consumption mechanism is inoperative, implying that the desired consumption cannot be a corner solution of a budget constrained problem of utility minimisation, we must replace not only the probit model explaining the variable $I_2$ by a singular probability choice model where $P(I_2 = 1) = 1$, but also the Tobit demand function by a demand model enforcing non-negative values on the latent variable $y_2^*$. For the time being, two functional forms of this demand model have been programmed in **mhurdle**, namely a log-normal functional form :

$$\ln y_2^* = \beta_2^\top x_2 + \epsilon_2 \tag{2}$$

and a truncated Tobit model, defined by the second of the set of linear relationships (1) with $\epsilon_2$ distributed as a normal random disturbance left-truncated at $\epsilon_2 = -\beta_2^\top x_2$, as suggested by Cragg (1971). Nevertheless, to avoid a cumbersome analytic presentation of our models, in the following we only consider the log-normal model specification.

- Hurdle 3 models the household decision to purchase or not to purchase the good during the survey period over which expenditure data are collected. This purchasing mechanism also explains the outcome of a binary choice, coded by a binary variable $I_3$ taking value 1 if the household decides to buy the good during the period of statistical observation and 0 otherwise. The probit model we use associates the purchasing decision to positive values of latent variable $y_3^*$ and that of not purchasing to negative values of $y_3^*$.
By assuming that consumption and purchases are uniformly distributed over time, but according to different timetables entailing a frequency of consumption higher than that of purchasing, we can also interpret the probability $P(I_3 = 1) = P(y_3^* > 0)$ as measuring the share of purchasing frequency to that of consumption during the observation period.

This allows to relate the observed level of expenditures $y$ to the unobserved level of consumption $y_2^*$ during the observation period, using the following identity:

$$y = \frac{y_2^*}{P(I_3 = 1)} I_1 I_2 I_3. \tag{3}$$

When the purchasing mechanism is inoperative, the previous probit model must be replaced by a singular probability choice model where $P(I_3 = 1) = 1$. In such a case, the observed level of expenditures is identified to the level of consumption, implying $y = y_2^* I_1 I_2$.

A priori information may suggest that one or more of these censoring mechanisms are not in effect. For instance, we know in advance that all households purchase food regularly, implying that the first two censoring mechanisms are inoperative for food. In this case, the relevant model is defined by only two relations: one defining the desired consumption level of food, according to a log-normal specification or a truncated Tobit model, and the other the decision to purchase food during the observation period.

Figure 1 outlines the full set of special models that can be generated by selecting which of these three mechanisms are in effect and which are not. It shows that 8 different models can be dealt with by means of the **mhurdle** package.

Among these models, one is not concerned by censored data, namely model 1. This model is relevant only for modelling uncensored samples. All the other models are potentially able to analyse censored samples by combining up to the three censoring mechanisms described above. With the notable exception of the standard Tobit model 3, that can be estimated also by the **survival** package of Therneau and Lumley (2008) or the **AER** package of Kleiber and Zeileis (2008), these models cannot be found in an other R package.

Some of **mhurdle** models have already been used in applied econometric literature. In particular, model 2 is a single-hurdle good selection model originated by Cragg (1971). The double-hurdle model combining independent good selection (hurdle 1) and desired consumption (hurdle 2) censoring mechanisms is also due to Cragg (1971). An extension of this double-hurdle model to dependent censoring mechanisms has been originated by Blundell and Meghir (1987).

P-Tobit model 7 is due to Deaton and Irish (1984) and explains zero purchases by combining the desired consumption censoring mechanism (hurdle 2) with the purchasing censoring mechanism (hurdle 3). Model 4 is a single-hurdle model not yet used in applied demand analysis, where the censoring mechanism in effect is that of infrequent purchases (hurdle 3).

Among the original models encompassed by **mhurdle**, models 6 is a double-hurdle model combining good selection (hurdle 1) and purchasing (hurdle 3) mechanisms to explain censored samples. Model 8 is an original triple-hurdle model originated in Hoareau (2009). This model explains censored purchases either as the result of good rejection (hurdle 1), negative desired consumption (hurdle 2) or infrequent purchases (hurdle 3).

To derive the form of the probability distribution of the observable dependent variable $y$, we must specify the joint distribution of the random disturbances entering the structural relations of these models.

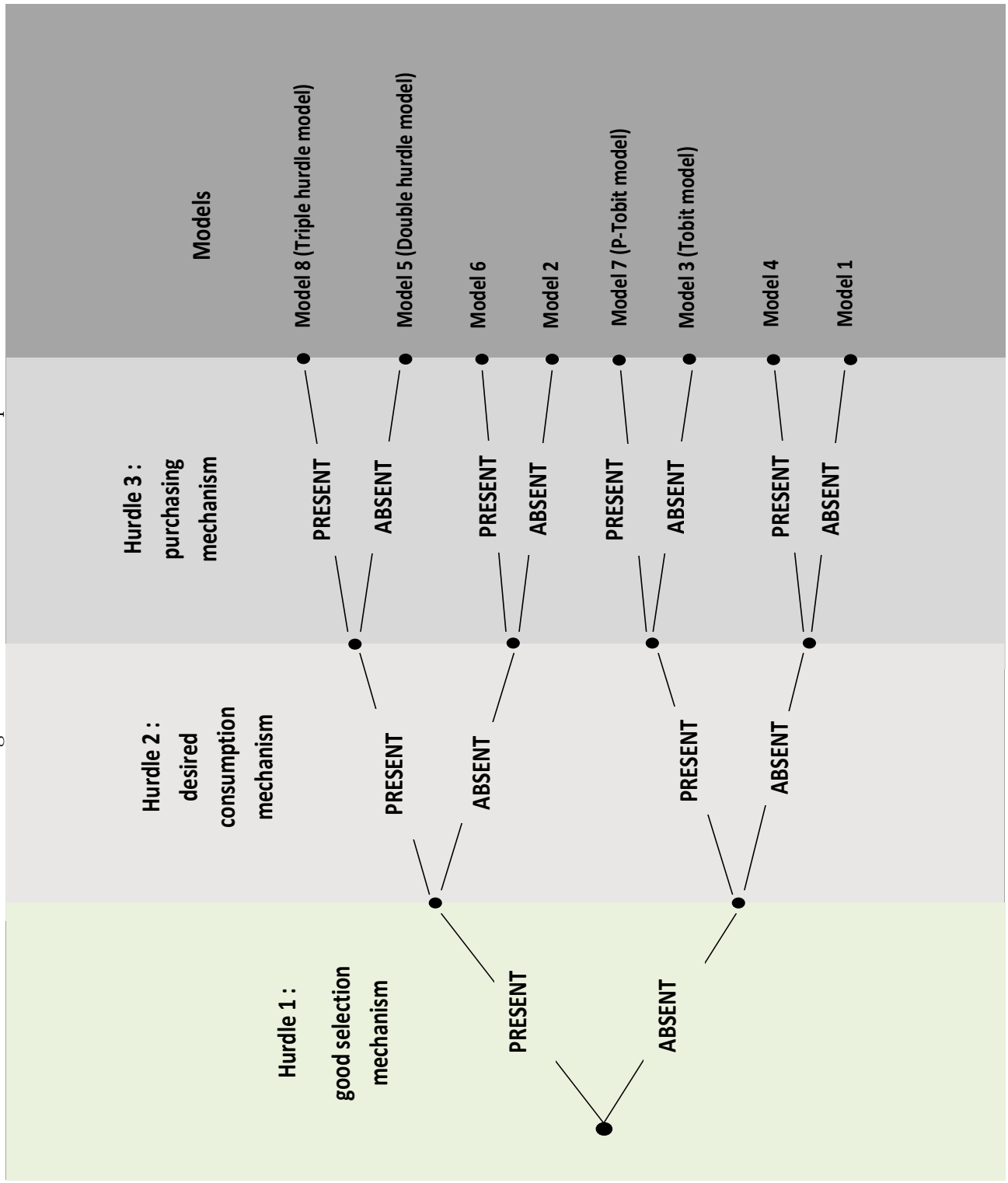


Figure 1: The full set of mhurdle special models.

- Models 8 and 6 are trivariate hurdle models as they involve disturbances $\epsilon_1$, $\epsilon_2$ and $\epsilon_3$, distributed according to the trivariate normal density function:

$$\frac{1}{\sigma}\phi\left(\epsilon_1, \frac{\epsilon_2}{\sigma}, \epsilon_3; \rho_{12}, \rho_{13}, \rho_{23}\right), \tag{4}$$

where

$$\phi(z_1, z_2, z_3; \rho_{12}, \rho_{13}, \rho_{23}) = \frac{\exp\left\{-\frac{\rho^{11}z_1^2+\rho^{22}z_2^2+\rho^{33}z_3^2-2[\rho^{12}z_1z_2+\rho^{13}z_1z_3+\rho^{23}z_2z_3]}{2}\right\}}{\sqrt{(2\pi)^3 \mid R \mid}},$$

with

$$|R| = 1-\rho_{12}^2-\rho_{13}^2-\rho_{23}^2+2\rho_{12}\rho_{13}\rho_{23},$$
$$\rho^{11} = \frac{1-\rho_{23}^2}{\mid R \mid}, \quad \rho^{22} = \frac{1-\rho_{13}^2}{\mid R \mid}, \quad \rho^{33} = \frac{1-\rho_{12}^2}{\mid R \mid},$$
$$\rho^{12} = \frac{\rho_{12}-\rho_{13}\rho_{23}}{\mid R \mid}, \quad \rho^{13} = \frac{\rho_{13}-\rho_{12}\rho_{23}}{\mid R \mid}, \quad \rho^{23} = \frac{(\rho_{23}-\rho_{12}\rho_{13})}{\mid R \mid},$$

denotes the density function of a standard trivariate normal distribution and $\rho_{12}$, $\rho_{13}$, $\rho_{23}$ the correlation coefficients between the couples of normal standard random variables $z_1$ and $z_2$, $z_1$ and $z_3$, $z_2$ and $z_3$, respectively. As the unit of measurement of $\epsilon_1$ and $\epsilon_3$ are not identified, these disturbances are normalised by setting their variances equal to 1.

- Models 7 and 4 are bivariate hurdle models as they involve disturbances $\epsilon_2$ and $\epsilon_3$, distributed according to the bivariate normal density function:

$$\frac{1}{\sigma}\phi\left(\frac{\epsilon_2}{\sigma}, \epsilon_3; \rho_{23}\right), \tag{5}$$

where

$$\phi(z_1, z_2; \rho) = \frac{\exp\left\{-\frac{z_1^2+z_2^2-2\rho z_1 z_2}{2(1-\rho^2)}\right\}}{2\pi\sqrt{1-\rho^2}}$$

denotes the density function of a standard bivariate normal distribution with correlation coefficient $\rho$.

- Models 5 and 2 are also bivariate hurdle models but they involve disturbances $\epsilon_1$ and $\epsilon_2$ which density function is therefore written as:

$$\frac{1}{\sigma}\phi\left(\epsilon_1, \frac{\epsilon_2}{\sigma}; \rho_{12}\right). \tag{6}$$

- Finally, models 3 and 1 are univariate hurdle models involving only disturbance $\epsilon_2$, which density function writes therefore:

$$\frac{1}{\sigma}\phi\left(\frac{\epsilon_2}{\sigma}\right), \tag{7}$$

where

$$\phi(z_1) = \frac{\exp\left\{-\frac{z_1^2}{2}\right\}}{\sqrt{2\pi}}$$

denotes the density function of a standard univariate normal distribution.

A priori information may also suggest to set to zero some or all correlations between the random disturbances entering these models, entailing a partial or total independence between the above defined censoring mechanisms. The use of this a priori information generates, for each trivariate or bivariate hurdle model of Figure 1, a subset of special models all nested within the general model from which they are derived. For a trivariate hurdle model the number of special models so derived is equal to 7, but for a bivariate hurdle model only one special model is generated, namely the model obtained by assuming the independence between the two random disturbances of the model.

In the following, we shall work out the distribution of our hurdle models in their general case, but considering the difficulties of implementing trivariate hurdle models in their full generality, for these models only the special cases of independence or dependence between one of hurdles 1 or 3 and the desired consumption equation, which seems the most relevant for empirical applications, have been programmed in **mhurdle**. The extension of our package to more general model specifications is in progress.

## 2.2. Likelihood function

As for the standard Tobit model, the probability distribution of the observed censored variable $y$ of our hurdle models is a discrete-continuous mixture, which assigns a probability mass $P(y = 0)$ to $y = 0$ and a density function $f_+(y)$ to any $y > 0$, with:

$$P(y = 0) + \int_0^\infty f_+(y)dy = 1. \tag{8}$$

The probability mass $P(y = 0) = 1 - P(y > 0)$ may be computed by integrating the joint density function of the latent variables entering the hurdle model over their positive values.

- For trivariate hurdle model 8, using the change of variables:

$$\begin{cases} z_1 = y_1^* - \beta_1^\top x_1 \\ z_2 = \dfrac{y_2^* - \beta_2^\top x_2}{\sigma} \\ z_3 = y_3^* - \beta_3^\top x_3 \end{cases} \tag{9}$$

this approach leads to:

$$\begin{aligned} P(y = 0) &= 1 - \int_{-\beta_1^\top x_1}^\infty \int_{-\frac{\beta_2^\top x_2}{\sigma}}^\infty \int_{-\beta_3^\top x_3}^\infty \phi(z_1, z_2, z_3; \rho_{12}, \rho_{13}, \rho_{23}) dz_1 dz_2 dz_3 \\ &= 1 - \Phi(\beta_1^\top x_1, \frac{\beta_2^\top x_2}{\sigma}, \beta_3^\top x_3; \rho_{12}, \rho_{13}, \rho_{23}), \end{aligned} \tag{10}$$

where $\Phi(z_1, z_2, z_3; \rho_{12}, \rho_{13}, \rho_{23})$ denotes the distribution function of a standard trivariate normal distribution with correlation coefficients $\rho_{12}$, $\rho_{13}$ and $\rho_{23}$.

- For trivariate hurdle model 6, using the change of variables:

$$\begin{cases} z_1 = y_1^* - \beta_1^\top x_1 \\ z_2 = \dfrac{\ln y_2^* - \beta_2^\top x_2}{\sigma} \\ z_3 = y_3^* - \beta_3^\top x_3 \end{cases} \tag{11}$$

this approach leads to:

$$\begin{aligned} P(y=0) &= 1 - \int_{-\beta_1^\top x_1}^{\infty} \int_{-\infty}^{\infty} \int_{-\beta_3^\top x_3}^{\infty} \phi(z_1, z_2, z_3; \rho_{12}, \rho_{13}, \rho_{23}) dz_1 dz_2 dz_3 \\ &= 1 - \Phi(\beta_1^\top x_1, \beta_3^\top x_3; \rho_{13}), \end{aligned} \tag{12}$$

where $\Phi(z_1, z_2; \rho)$ denotes the distribution function of a standard bivariate normal distribution with correlation coefficient $\rho$.

- The probability mass $P(y = 0)$ for bivariate hurdle models 7 and 5 and univariate hurdle model 3 can be derived from that of trivariate model 8 by eliminating hurdles 1, 3, 1 and 3, respectively. Likewise, this probability for bivariate hurdle models 4 and 2 can be derived from that of trivariate hurdle model 6 by eliminating hurdles 1 and 3, respectively. Corresponding formulas of $P(y = 0)$ for all this special cases implemented in R are presented in Table 1, using the following notations:

$$\Phi_1 = \Phi(\beta_1^\top x_1), \quad \Phi_2 = \Phi\left(\frac{\beta_2^\top x_2}{\sigma}\right), \quad \Phi_3 = \Phi(\beta_3^\top x_3),$$

$$\Phi_{12} = \left(\beta_1^\top x_1, \frac{\beta_2^\top x_2}{\sigma}; \rho_{12}\right), \quad \Phi_{23} = \left(\frac{\beta_2^\top x_2}{\sigma}, \beta_3^\top x_3; \rho_{23}\right),$$

where $\Phi(z)$ denotes the distribution function of a standard univariate normal distribution.

The density function $f_+(y)$ may be computed by performing: first the change of variables $y_2^* = P(I_3 = 1)y = \Phi_3 y$ on the joint density function of the latent variables entering the hurdle model; then by integrating this transformed density function over the positive values of latent variables $y_1^*$ and $y_3^*$.

- For trivariate hurdle model 8 this transformed density function is written as:

$$\frac{\Phi_3}{\sigma} \phi\left(y_1^* - \beta_1^\top x_1, \frac{\Phi_3 y - \beta_2^\top x_2}{\sigma}, y_3^* - \beta_3^\top x_3; \rho_{12}, \rho_{13}, \rho_{23}\right). \tag{13}$$

To perform the analytical integration of this function, it is useful to rewrite it as the product of the marginal distribution of $y$, namely:

$$\frac{\Phi_3}{\sigma} \phi\left(\frac{\Phi_3 y - \beta_2^\top x_2}{\sigma}\right) \tag{14}$$

Table 1: Characteristics of mhurdle special models implemented in R

| id | $h_1$ | $h_2$ | $h_3$ | $\rho_{12}$ | $\rho_{13}$ | $\rho_{23}$ | $P(y=0)$ | $f_+(y)$ | $\mathrm{E}(y \mid y>0)$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | □ | □ | □ | □ | □ | □ | 0 | $\frac{1}{\sigma y}\phi\left(\frac{\ln y-\beta_2^\top x_2}{\sigma}\right)$ | $\exp\left\{\beta_2^\top x_2+\frac{\sigma^2}{2}\right\}$ |
| 2i | ■ | □ | □ | □ | □ | □ | $1-\Phi_1$ | $\frac{1}{\sigma y}\phi\left(\frac{\ln y-\beta_2^\top x_2}{\sigma}\right)\Phi_1$ | $\exp\left\{\beta_2^\top x_2+\frac{\sigma^2}{2}\right\}$ |
| 2d | ■ | □ | □ | ■ | □ | □ | $1-\Phi_1$ | $\frac{1}{\sigma y}\phi\left(\frac{\ln y-\beta_2^\top x_2}{\sigma}\right)\Phi\left(\frac{\beta_1^\top x_1+\rho_{12}\frac{\ln y-\beta_2^\top x_2}{\sigma}}{\sqrt{1-\rho_{12}^2}}\right)$ | $\exp\left\{\beta_2^\top x_2+\frac{\sigma^2}{2}\right\}\frac{\Phi\left(\beta_1^\top x_1+\sigma\rho_{12}\right)}{\Phi_1}$ |
| 3 | □ | ■ | □ | □ | □ | □ | $1-\Phi_2$ | $\frac{1}{\sigma}\phi\left(\frac{y-\beta_2^\top x_2}{\sigma}\right)$ | $\beta_2^\top x_2+\sigma\frac{\phi_2}{\Phi_2}$ |
| 4i | □ | □ | ■ | □ | □ | □ | $1-\Phi_3$ | $\frac{1}{\sigma y}\phi\left(\frac{\ln(\Phi_3 y)-\beta_2^\top x_2}{\sigma}\right)\Phi_3$ | $\exp\left\{\beta_2^\top x_2+\frac{\sigma^2}{2}\right\}\frac{1}{\Phi_3}$ |
| 4d | □ | □ | ■ | □ | □ | ■ | $1-\Phi_3$ | $\frac{1}{\sigma y}\phi\left(\frac{\ln(\Phi_3 y)-\beta_2^\top x_2}{\sigma}\right)\Phi\left(\frac{\beta_3^\top x_3+\rho_{23}\frac{\ln(\Phi_3 y)-\beta_2^\top x_2}{\sigma}}{\sqrt{1-\rho_{23}^2}}\right)$ | $\exp\left\{\beta_2^\top x_2+\frac{\sigma^2}{2}\right\}\frac{\Phi\left(\beta_3^\top x_3+\sigma\rho_{23}\right)}{\Phi_3^2}$ |
| 5i | ■ | ■ | □ | □ | □ | □ | $1-\Phi_1\Phi_2$ | $\frac{1}{\sigma}\phi\left(\frac{y-\beta_2^\top x_2}{\sigma}\right)\Phi_1$ | $\beta_2^\top x_2+\sigma\frac{\phi_2}{\Phi_2}$ |
| 5d | ■ | ■ | □ | ■ | □ | □ | $1-\Phi_{12}$ | $\frac{1}{\sigma}\phi\left(\frac{y-\beta_2^\top x_2}{\sigma}\right)\Phi\left(\frac{\beta_1^\top x_1+\rho_{12}\frac{y-\beta_2^\top x_2}{\sigma}}{\sqrt{1-\rho_{12}^2}}\right)$ | $\beta_2^\top x_2+\sigma\frac{\Psi_{2\mid1}}{\Phi_{12}}$ |
| 6i | ■ | □ | ■ | □ | □ | □ | $1-\Phi_1\Phi_3$ | $\frac{1}{\sigma y}\phi\left(\frac{\ln(\Phi_3 y)-\beta_2^\top x_2}{\sigma}\right)\Phi_1\Phi_3$ | $\exp\left\{\beta_2^\top x_2+\frac{\sigma^2}{2}\right\}\frac{1}{\Phi_3}$ |
| 6d1 | ■ | □ | ■ | ■ | □ | □ | $1-\Phi_1\Phi_3$ | $\frac{1}{\sigma y}\phi\left(\frac{(\ln\Phi_3 y)-\beta_2^\top x_2}{\sigma}\right)\Phi\left(\frac{\beta_1^\top x_1+\rho_{12}\frac{(\ln y\Phi_3)-\beta_2^\top x_2}{\sigma}}{\sqrt{1-\rho_{12}^2}}\right)\Phi_3$ | $\exp\left\{\beta_2^\top x_2+\frac{\sigma^2}{2}\right\}\frac{\Phi\left(\beta_1^\top x_1+\sigma\rho_{12}\right)}{\Phi_1\Phi_3}$ |
| 6d3 | ■ | □ | ■ | □ | □ | ■ | $1-\Phi_1\Phi_3$ | $\frac{1}{\sigma y}\phi\left(\frac{\ln(\Phi_3 y)-\beta_2^\top x_2}{\sigma}\right)\Phi_1\Phi\left(\frac{\beta_3^\top x_3+\rho_{23}\frac{\ln(\Phi_3 y)-\beta_2^\top x_2}{\sigma}}{\sqrt{1-\rho_{23}^2}}\right)$ | $\exp\left\{\beta_2^\top x_2+\frac{\sigma^2}{2}\right\}\frac{\Phi\left(\beta_3^\top x_3+\sigma\rho_{23}\right)}{\Phi_3^2}$ |
| 7i | □ | ■ | ■ | □ | □ | □ | $1-\Phi_2\Phi_3$ | $\frac{1}{\sigma}\phi\left(\frac{\Phi_3 y-\beta_2^\top x_2}{\sigma}\right)\Phi_3^2$ | $\frac{\beta_2^\top x_2}{\Phi_3}+\sigma\frac{\phi_2}{\Phi_2\Phi_3}$ |
| 7d | □ | ■ | ■ | □ | □ | ■ | $1-\Phi_{23}$ | $\frac{1}{\sigma}\phi\left(\frac{\Phi_3 y-\beta_2^\top x_2}{\sigma}\right)\Phi\left(\frac{\beta_3^\top x_3+\rho_{23}\frac{\Phi_3 y-\beta_2^\top x_2}{\sigma}}{\sqrt{1-\rho_{23}^2}}\right)\Phi_3$ | $\frac{\beta_2^\top x_2}{\Phi_3}+\sigma\frac{\Psi_{2\mid3}}{\Phi_{23}\Phi_3}$ |
| 8i | ■ | ■ | ■ | □ | □ | □ | $1-\Phi_1\Phi_2\Phi_3$ | $\frac{1}{\sigma}\phi\left(\frac{\Phi_3 y-\beta_2^\top x_2}{\sigma}\right)\Phi_1\Phi_3^2$ | $\frac{\beta_2^\top x_2}{\Phi_3}+\sigma\frac{\phi_2}{\Phi_2\Phi_3}$ |
| 8d1 | ■ | ■ | ■ | ■ | □ | □ | $1-\Phi_{12}\Phi_3$ | $\frac{1}{\sigma}\phi\left(\frac{\Phi_3 y-\beta_2^\top x_2}{\sigma}\right)\Phi\left(\frac{\beta_1^\top x_1+\rho_{12}\frac{\Phi_3 y-\beta_2^\top x_2}{\sigma}}{\sqrt{1-\rho_{12}^2}}\right)\Phi_3^2$ | $\frac{\beta_2^\top x_2}{\Phi_3}+\sigma\frac{\Psi_{2\mid1}}{\Phi_{12}\Phi_3}$ |
| 8d3 | ■ | ■ | ■ | □ | □ | ■ | $1-\Phi_1\Phi_{23}$ | $\frac{1}{\sigma}\phi\left(\frac{\Phi_3 y-\beta_2^\top x_2}{\sigma}\right)\Phi_1\Phi\left(\frac{\beta_3^\top x_3+\rho_{23}\frac{\Phi_3 y-\beta_2^\top x_2}{\sigma}}{\sqrt{1-\rho_{23}^2}}\right)\Phi_3$ | $\frac{\beta_2^\top x_2}{\Phi_3}+\sigma\frac{\Psi_{2\mid3}}{\Phi_{23}\Phi_3}$ |

A blackened square indicates which hurdle or correlation is assumed to be at work in the model.

and of the joint density function of $y_1^*$ and $y_3^*$ conditioned with respect to $y$, which can be written as follows:

$$\frac{1}{\sigma_{1|2}\sigma_{3|2}}\phi\left(\frac{y_1^*-\mu_{1|2}}{\sigma_{1|2}},\frac{y_3^*-\mu_{3|2}}{\sigma_{3|2}};\rho_{13|2}\right), \tag{15}$$

with:

$$\mu_{1|2}=\beta_1^\top x_1+\rho_{12}\frac{\Phi_3 y-\beta_2^\top x_2}{\sigma},\quad \mu_{3|2}=\beta_3^\top x_3+\rho_{23}\frac{\Phi_3 y-\beta_2^\top x_2}{\sigma},$$
$$\sigma_{1|2}^2=1-\rho_{12}^2,\quad \sigma_{3|2}^2=1-\rho_{23}^2,\quad \rho_{13|2}=\frac{\rho_{13}-\rho_{12}\rho_{23}}{\sqrt{1-\rho_{12}^2}\sqrt{1-\rho_{23}^2}}.$$

Using this factorization of the density function of $y_1^*$, $y$ and $y_3^*$, we obtain:

$$\begin{aligned}
f_+(y)&=\frac{\Phi_3}{\sigma}\phi\left(\frac{\Phi_3 y-\beta_2^\top x_2}{\sigma}\right)\\
&\times\int_0^\infty\int_0^\infty\frac{1}{\sigma_{1|2}\sigma_{3|2}}\phi\left(\frac{y_1^*-\mu_{1|2}}{\sigma_{1|2}},\frac{y_3^*-\mu_{3|2}}{\sigma_{3|2}};\rho_{13|2}\right)dy_1^*dy_3^*\\
&=\frac{\Phi_3}{\sigma}\phi\left(\frac{\Phi_3 y-\beta_2^\top x_2}{\sigma}\right)\int_{-\frac{\mu_{1|2}}{\sigma_{1|2}}}^\infty\int_{-\frac{\mu_{3|2}}{\sigma_{3|2}}}^\infty\phi(z_1,z_3;\rho_{13|2})dz_1dz_3\\
&=\frac{\Phi_3}{\sigma}\phi\left(\frac{\Phi_3 y-\beta_2^\top x_2}{\sigma}\right)\\
&\times\Phi\left(\frac{\beta_1^\top x_1+\rho_{12}\frac{\Phi_3 y-\beta_2^\top x_2}{\sigma}}{\sqrt{1-\rho_{12}^2}},\frac{\beta_3^\top x_3+\rho_{23}\frac{\Phi_3 y-\beta_2^\top x_2}{\sigma}}{\sqrt{1-\rho_{23}^2}};\rho_{13|2}\right).
\end{aligned} \tag{16}$$

- For trivariate hurdle model 6, we proceed as for hurdle model 8 by substituting the joint normal density function (13), by the following joint normal/log-normal density function:

$$\frac{1}{\sigma y}\phi\left(y_1^*-\beta_1^\top x_1,\frac{\ln(\Phi_3 y)-\beta_2^\top x_2}{\sigma},y_3^*-\beta_3^\top x_3;\rho_{12},\rho_{13},\rho_{23}\right). \tag{17}$$

To integrate this density function with respect to the positive values of $y_1^*$ and $y_2^*$, we rewrite it as the product of the marginal distribution of $y$, which is log-normal:

$$\frac{1}{\sigma y}\phi\left(\frac{\ln(\Phi_3 y)-\beta_2^\top x_2}{\sigma}\right) \tag{18}$$

and of the joint density function of $y_1^*|y$ and $y_3^*|y$, which is bivariate normal:

$$\frac{1}{\sigma_{1|2}\sigma_{3|2}}\phi\left(\frac{y_1^*-\mu_{1|2}}{\sigma_{1|2}},\frac{y_3^*-\mu_{3|2}}{\sigma_{3|2}};\rho_{13|2}\right), \tag{19}$$

with:

$$\mu_{1|2}=\beta_1^\top x_1+\rho_{12}\frac{\ln(\Phi_3 y)-\beta_2^\top x_2}{\sigma},\quad \mu_{3|2}=\beta_3^\top x_3+\rho_{23}\frac{\ln(\Phi_3 y)-\beta_2^\top x_2}{\sigma},$$

$$\sigma^2_{1|2} = 1 - \rho^2_{12}, \quad \sigma^2_{3|2} = 1 - \rho^2_{23}, \quad \rho_{13|2} = \frac{\rho_{13} - \rho_{12}\rho_{23}}{\sqrt{1-\rho^2_{12}}\sqrt{1-\rho^2_{23}}}.$$

By integrating this factorisation of the density function of $y_1^*$, $y$ and $y_3^*$, over the positive values of $y_1^*$ and $y_3^*$, we obtain:

$$\begin{aligned} f_+(y) &= \frac{\phi\left(\frac{\ln(\Phi_3 y) - \beta_2^\top x_2}{\sigma}\right)}{\sigma y} \int_{-\frac{\mu_{1|2}}{\sigma_{1|2}}}^{\infty} \int_{-\frac{\mu_{3|2}}{\sigma_{3|2}}}^{\infty} \phi(z_1, z_3; \rho_{13|2}) dz_1 dz_3 \\ &= \frac{\phi\left(\frac{\ln(\Phi_3 y) - \beta_2^\top x_2}{\sigma}\right)}{\sigma y} \\ &\times \Phi\left(\frac{\beta_1^\top x_1 + \rho_{12}\frac{\ln(\Phi_3 y) - \beta_2^\top x_2}{\sigma}}{\sqrt{1-\rho^2_{12}}}, \frac{\beta_3^\top x_3 + \rho_{23}\frac{\ln(\Phi_3 y) - \beta_2^\top x_2}{\sigma}}{\sqrt{1-\rho^2_{23}}}; \rho_{13|2}\right). \end{aligned} \tag{20}$$

- The density function $f_+(y)$ for bivariate hurdle models 7 and 5 and univariate hurdle model 3 can be derived from that of trivariate model 8 by eliminating hurdles 1, 3, 1 and 3, respectively. Likewise, this density function for bivariate hurdle models 4 and 2 can be derived from that of trivariate hurdle model 6 by eliminating hurdles 1 and 3, respectively. Corresponding formulas for $f_+(y)$ for all this special cases implemented in R are presented in Table 1.

From these results it is easy to derive the likelihood function of a random sample of $n$ observations of the censored dependent variable $y$. As these observations are all independently drawn from the same conditional (on covariates $x_1$, $x_2$ and $x_3$) discrete-continuous distribution, which assigns a conditional probability mass $P(y = 0)$ to the observed value $y = 0$ and a conditional density function $f_+(y)$ to the observed values $y > 0$, the log-likelihood function for an observation $y_i$ can be written as :

$$\ln L_i = \begin{cases} \ln P(y_i = 0) & \text{if} \quad y_i = 0 \\ \ln f_+(y_i) & \text{if} \quad y_i > 0 \end{cases} \tag{21}$$

and the log-likelihood for the entire random sample:

$$\ln L = \sum_{i=1}^{n} \ln L_i = \sum_{i|y_i=0} \ln P(y_i = 0) + \sum_{i|y_i>0} \ln f_+(y_i). \tag{22}$$

## 3. Model estimation, evaluation and selection

The econometric framework described in the previous section provides a theoretical background for tackling the problems of model estimation, evaluation and selection within the statistical theory of classical inference.

### 3.1. Model estimation

The full parametric specification of our multiple hurdle models allows to efficiently estimate their parameters by means of the maximum likelihood principle. Indeed, it is well known

from classical estimation theory that, under the assumption of a correct model specification and for a likelihood function sufficiently well behaved, the maximum likelihood estimator is asymptotically efficient within the class of consistent and asymptotically normal estimators [2].

More precisely, the asymptotic distribution of the maximum likelihood estimator $\hat{\theta}$ for the parameter vector $\theta$ of a multiple hurdle model, is written as:

$$\hat{\theta} \overset{A}{\sim} N(\theta, \frac{1}{n} I_A(\theta)^{-1}), \tag{23}$$

where $\overset{A}{\sim}$ stands for "asymptotically distributed as" and

$$I_A(\theta) = \text{plim} \frac{1}{n} \sum_{i=1}^{n} E\left(\frac{\partial^2 \ln L_i(\theta)}{\partial\theta\partial\theta^\top}\right) = \text{plim} \frac{1}{n} \sum_{i=1}^{n} E\left(\frac{\partial \ln L_i(\theta)}{\partial\theta} \frac{\partial \ln L_i(\theta)}{\partial\theta^\top}\right)$$

for the asymptotic Fisher information matrix of a sample of $n$ independent observations.

More generally, any inference about a differentiable vector function of $\theta$, denoted by $\gamma = h(\theta)$, can be based on the asymptotic distribution of its implied maximum likelihood estimator $\hat{\gamma} = h(\hat{\theta})$. This distribution can be derived from the asymptotic distribution of $\hat{\theta}$ according to the so called delta method:

$$\hat{\gamma} \overset{A}{\sim} h(\theta) + \frac{\partial h}{\partial\theta^\top}(\hat{\theta} - \theta) \overset{A}{\sim} N\left(\gamma, \frac{1}{n} \frac{\partial h}{\partial\theta^\top} I_A(\theta)^{-1} \frac{\partial h^\top}{\partial\theta}\right). \tag{24}$$

The practical use of these asymptotic distributions requires to replace the theoretical variance-covariance matrix of these asymptotic distributions with consistent estimators, which can be obtained by using $\frac{\partial h(\hat{\theta})}{\partial\theta^\top}$ as a consistent estimator for $\frac{\partial h(\theta)}{\partial\theta^\top}$ and either $\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \ln L_i(\hat{\theta})}{\partial\theta\partial\theta^\top}$ or $\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ln L_i(\hat{\theta})}{\partial\theta} \frac{\partial \ln L_i(\hat{\theta})}{\partial\theta^\top}$ as a consistent estimator for $I_A(\theta)$. The last two estimators are directly provided by two standard iterative methods used to compute the maximum likelihood parameter's estimate, namely the Newton-Raphson method and the Berndt, Hall, Hall, Hausman or BHHH method, respectively, mentioned in section 4.3.

### 3.2. Model evaluation and selection using goodness of fit measures

Two fundamental principles should be used to appraise the results of a model estimation, namely its economic relevance and its statistical and predictive adequacy. The first principle deals with the issues of accordance of model estimate with the economic rationale underlying the model specification and of its relevance for answering the questions for which the model has been built. These issues are essentially context specific and, therefore, cannot be dealt with by means of generic criteria. The second principle refers to the issues of empirical soundness of model estimate and of its ability to predict sample or out-of-sample observations. These issues can be tackled by means of formal tests of significance, based on the previously presented asymptotic distributions of model estimates, and by measures of goodness of fit/prediction, respectively.

[2] See Amemiya (1985) chapter 4, for a more rigorous statement of this property.

To assess the goodness of fit of **mhurdle** estimates, two pseudo $R^2$ coefficients are provided. The first one is an extension of the classical coefficient of determination, used to explain the fraction of variation of the dependent variable explained by the covariates included in a linear regression model with intercept. The second one is an extension of the likelihood ratio index introduced by McFadden (1974) to measure the relative gain in the maximised log-likelihood function due to the covariates included in a qualitative response model.

To define a pseudo coefficient of determination, we rely on the non linear regression model explaining the dependent variable of a multiple hurdle model. This model is written as:

$$y = E(y) + u, \tag{25}$$

where $u$ stands for a zero expectation, heteroskedastic random disturbance and $E(y)$ for the expectation of the censored dependent variable $y$:

$$E(y) = 0 \times P(y = 0) + \int_0^\infty y f_+(y) dy = \int_0^\infty y f_+(y) dy. \tag{26}$$

To compute this expectation, we reformulate it as a multiple integral of the joint density function of $y_1^*$, $y$ and $y_3^*$ multiplied by $y$, over the positive values of these variables.

- For trivariate hurdle model 8, using the density function (13) and the change of variables:

$$\begin{cases} z_1 = y_1^* - \beta_1^\top x_1 \\ z_2 = \dfrac{\Phi_3 y - \beta_2^\top x_2}{\sigma} \\ z_3 = y_3^* - \beta_3^\top x_3 \end{cases} \tag{27}$$

this reformulation of $E(y)$ is written as:

$$\begin{aligned} E(y) &= \int_{-\beta_1^\top x_1}^\infty \int_{-\frac{\beta_2^\top x_2}{\sigma}}^\infty \int_{-\beta_3^\top x_3}^\infty \frac{\beta_2^\top x_2 + \sigma z_2}{\Phi_3} \phi\left(z_1, z_2, z_3; \rho_{12}, \rho_{13}, \rho_{23}\right) dz_1 dz_2 dz_3 \\ &= \frac{\beta_2^\top x_2}{\Phi_3} \Phi\left(\beta_1^\top x_1, \frac{\beta_2^\top x_2}{\sigma}, \beta_3^\top x_3; \rho_{12}, \rho_{13}, \rho_{23}\right) \\ &+ \frac{\sigma}{\Phi_3} \int_{-\beta_1^\top x_1}^\infty \int_{-\frac{\beta_2^\top x_2}{\sigma}}^\infty \int_{-\beta_3^\top x_3}^\infty z_2 \phi\left(z_1, z_2, z_3; \rho_{12}, \rho_{13}, \rho_{23}\right) dz_1 dz_2 dz_3. \end{aligned} \tag{28}$$

To perform the analytical integration of the second term of the right-hand side of this formula, it is useful to rewrite the density function of $z_1$, $z_2$ and $z_3$ as the product of the marginal density function of $z_1$ and $z_2$, namely $\phi\left(z_1, z_2; \rho_{13}\right)$ and of the density function of $z_2 | z_1, z_3$, which can be written as follows:

$$\frac{\phi\left(\frac{z_2 - \mu_{2|1,3}}{\sigma_{2|1,3}}\right)}{\sigma_{2|1,3}}, \tag{29}$$

where:

$$\mu_{2|1,3} = \varrho_1 z_1 + \varrho_3 z_3, \quad \sigma^2_{2|1,3} = \frac{1 - \rho^2_{12} - \rho^2_{13} - \rho^2_{23} + 2\rho_{12}\rho_{13}\rho_{23}}{1 - \rho^2_{13}},$$

with:

$$\varrho_1 = \frac{\rho_{12} - \rho_{13}\rho_{23}}{1 - \rho^2_{13}}, \quad \varrho_3 = \frac{\rho_{23} - \rho_{12}\rho_{13}}{1 - \rho^2_{13}}.$$

Using this factorisation of the density function of $z_1$, $z_2$ and $z_3$, we obtain:

$$\begin{aligned} &\int_{-\beta_1^\top x_1}^{\infty} \int_{-\frac{\beta_2^\top x_2}{\sigma}}^{\infty} \int_{-\beta_3^\top x_3}^{\infty} z_2 \phi\left(z_1, z_2, z_3; \rho_{12}, \rho_{13}, \rho_{23}\right) dz_1 dz_2 dz_3 \\ &= \int_{-\beta_1^\top x_1}^{\infty} \int_{-\beta_3^\top x_3}^{\infty} \left[ \int_{-\frac{\beta_2^\top x_2}{\sigma}}^{\infty} z_2 \phi\left(\frac{z_2 - \mu_{2|1,3}}{\sigma_{2|1,3}}\right) \frac{dz_2}{\sigma_{2|1,3}} \right] \phi\left(z_1, z_2; \rho_{13}\right) dz_1 dz_3. \end{aligned} \tag{30}$$

By performing the change of variable:

$$z = \frac{z_2 - \mu_{2|1,3}}{\sigma_{2|1,3}}, \tag{31}$$

the integral with respect to $z_2$ simplifies to:

$$\mu_{2|1,3} \Phi\left(\frac{\frac{\beta_2^\top x_2}{\sigma} + \mu_{2|1,3}}{\sigma_{2|1,3}}\right) + \sigma_{2|1,3} \phi\left(\frac{\frac{\beta_2^\top x_2}{\sigma} + \mu_{2|1,3}}{\sigma_{2|1,3}}\right). \tag{32}$$

By inserting this result in formula (30), we finally obtain:

$$\begin{aligned} E(y) = &\frac{\beta_2^\top x_2}{\Phi_3} \Phi\left(\beta_1^\top x_1, \frac{\beta_2^\top x_2}{\sigma}, \beta_3^\top x_3; \rho_{12}, \rho_{13}, \rho_{23}\right) \\ &+ \frac{\sigma}{\Phi_3} \int_{-\beta_1^\top x_1}^{\infty} \int_{-\beta_3^\top x_3}^{\infty} \left[ \left(\varrho_1 z_1 + \varrho_3 z_3\right) \Phi\left(\frac{\frac{\beta_2^\top x_2}{\sigma} + \varrho_1 z_1 + \varrho_3 z_3}{\sigma_{2|1,3}}\right) \right. \\ &\left. + \sigma_{2|1,3} \phi\left(\frac{\frac{\beta_2^\top x_2}{\sigma} + \varrho_1 z_1 + \varrho_3 z_3}{\sigma_{2|1,3}}\right) \right] \phi\left(z_1, z_3; \rho_{13}\right) dz_1 dz_3. \end{aligned} \tag{33}$$

- For trivariate hurdle model 6, we proceed as for hurdle model 8 by first substituting the joint normal density function (13) by the joint normal/log-normal density function (17), then by performing the change of variables:

$$\begin{cases} z_1 = y_1^* - \beta_1^\top x_1 \\ z_2 = \dfrac{\ln\left(\Phi_3 y\right) - \beta_2^\top x_2}{\sigma} \\ z_3 = y_3^* - \beta_3^\top x_3 \end{cases} \tag{34}$$

This leads to the following expression of the expected value of $y$:

$$\begin{aligned} E(y) &= \int_{-\beta_1^\top x_1}^{\infty}\int_{-\infty}^{\infty}\int_{-\beta_3^\top x_3}^{\infty} \frac{\exp\{\beta_2^\top x_2 + \sigma z_2\}}{\Phi_3} \\ &\times\phi\left(z_1, z_2, z_3; \rho_{12}, \rho_{13}, \rho_{23}\right) dz_1 dz_2 dz_3 = \frac{\exp\{\beta_2^\top x_2\}}{\Phi_3} \\ &\times \int_{-\beta_1^\top x_1}^{\infty}\int_{-\beta_3^\top x_3}^{\infty}\left[\int_{-\infty}^{\infty}\exp\{\sigma z_2\}\phi\left(\frac{z_2-\mu_{2|1,3}}{\sigma_{2|1,3}}\right)\frac{dz_2}{\sigma_{2|1,3}}\right]\phi\left(z_1, z_3; \rho_{13}\right) dz_1 dz_3 \end{aligned} \tag{35}$$

obtained by factorising the density function of $z_1$, $z_2$ and $z_3$ as the product of the marginal density function of $z_1$ and $z_3$ times the density function of $z_2|z_1, z_3$.
By performing the change of variable (31), the integral with respect to $z_2$ simplifies to:

$$\int_{-\infty}^{\infty}\exp\{\sigma(\mu_{2|1,3}+\sigma_{2|1,3}z)\}\phi(z)dz = \exp\left\{\sigma\mu_{2|1,3}+\frac{\sigma^2\sigma_{2|1,3}^2}{2}\right\}. \tag{36}$$

By inserting this result in formula (35), we finally obtain:

$$\begin{aligned} E(y) = &\frac{\exp\left\{\beta_2^\top x_2 + \frac{\sigma^2\sigma_{2|1,3}^2}{2}\right\}}{\Phi_3} \\ &\times \int_{-\beta_1^\top x_1}^{\infty}\int_{-\beta_3^\top x_3}^{\infty}\exp\{\sigma\left(\varrho_1 z_1 + \varrho_3 z_3\right)\}\phi\left(z_1, z_3; \rho_{13}\right) dz_1 dz_3. \end{aligned} \tag{37}$$

- $E(y)$ for bivariate hurdle models 7 and 5 and univariate hurdle model 3 can be derived from that of trivariate model 8 by eliminating hurdles 1, 3, 1 and 3, respectively. Likewise, the expectation of $y$ for bivariate hurdle models 4 and 2 can be derived from that of trivariate hurdle model 6 by eliminating hurdles 1 and 3, respectively. Corresponding formulas of $E(y|y>0) = E(y)/P(y>0)$ for all this special cases implemented in R are presented in Table 1, using the following notations:

$$\Psi_{2|1} = \rho_{12}\phi_1\Phi\left(\frac{\frac{\beta_2^\top x_2}{\sigma}-\rho_{12}\beta_1^\top x_1}{\sqrt{1-\rho_{12}^2}}\right)+\phi_2\Phi\left(\frac{\beta_1^\top x_1-\rho_{12}\frac{\beta_2^\top x_2}{\sigma}}{\sqrt{1-\rho_{12}^2}}\right),$$

$$\Psi_{2|3} = \rho_{23}\phi_3\Phi\left(\frac{\frac{\beta_2^\top x_2}{\sigma}-\rho_{23}\beta_3^\top x_3}{\sqrt{1-\rho_{23}^2}}\right)+\phi_2\Phi\left(\frac{\beta_3^\top x_3-\rho_{23}\frac{\beta_2^\top x_2}{\sigma}}{\sqrt{1-\rho_{23}^2}}\right),$$

where $\phi_1 = \phi(\beta_1^\top x_1)$, $\phi_2\left(\frac{\beta_2^\top x_2}{\sigma}\right)$ and $\phi_3 = \phi(\beta_3^\top x_3)$.

Note that formulas of $E(y|y>0)$ for dependent trivariate hurdle models presented in Table 1 are obtained by using closed forms of the following integrals :

$$\int_{-\beta^\top x}^{\infty}\left[\rho z\Phi\left(\frac{\frac{\beta_2^\top x_2}{\sigma}+\rho z}{\sqrt{1-\rho^2}}\right)+\sqrt{1-\rho^2}\phi\left(\frac{\frac{\beta_2^\top x_2}{\sigma}+\rho z}{\sqrt{1-\rho^2}}\right)\right]\phi(z)dz$$

$$= \rho\phi\left(\beta^\top x\right)\Phi\left(\frac{\frac{\beta_2^\top x_2}{\sigma} - \rho\beta^\top x}{\sqrt{1-\rho^2}}\right) + \phi\left(\frac{\beta_2^\top x_2}{\sigma}\right)\Phi\left(\frac{\beta^\top x - \rho\frac{\beta_2^\top x_2}{\sigma}}{\sqrt{1-\rho^2}}\right),$$

$$\int_{-\beta^\top x}^{\infty} \exp\left\{\sigma\rho z\right\}\phi(z)dz = \exp\left\{\frac{\sigma^2\rho^2}{2}\right\}\Phi\left(\beta^\top x + \sigma\rho\right).$$

Denoting by $\hat{y}_i$ the fitted values of $y_i$ obtained by estimating the mean square error predictor $E(y_i)$ for $y_i$ with the maximum likelihood estimate of model parameters, we define a pseudo coefficient of determination for a multiple hurdle model using the following formula:

$$R^2 = 1 - \frac{RSS}{TSS}, \tag{38}$$

with $RSS = \sum(y_i - \hat{y}_i)^2$ the residual sum of squares and $TSS = \sum(y_i - \hat{y}_0)^2$ the total sum of squares, where $\hat{y}_0$ denotes the maximum likelihood estimate of $E(y_i)$ in the multiple hurdle model without covariates (intercept-only model [3]). Note that this goodness of fit measure cannot exceed one but can be negative, as a consequence of the non linearity of $E(y_i)$ with respect to the parameters.

The extension of the McFadden likelihood ratio index for qualitative response models to multiple hurdle models is straightforwardly obtained by substituting in this index formula:

$$\rho^2 = 1 - \frac{\ln L(\hat{\theta})}{\ln L(\hat{\alpha})} = \frac{\ln L(\hat{\alpha}) - \ln L(\hat{\theta})}{\ln L(\hat{\alpha})}, \tag{39}$$

the maximised log-likelihood function of a qualitative response model with covariates and the log-likelihood function of the corresponding model without covariates or intercept-only model, with the maximised log-likelihood functions of a multiple hurdle model with covariates, $\ln L(\hat{\theta})$, and without covariates, $\ln L(\hat{\alpha})$, respectively. This goodness of fit measure takes values within zero and one and, as it can be easily inferred from the above second expression of $\rho^2$, it measures the relative increase of the maximised log-likelihood function due to the use of explanatory variables with respect to the maximised log-likelihood function of a naive intercept-only model.

Model selection deals with the problem of discriminating between alternative model specifications used to explain the same dependent variable, with the purpose of finding the one best suited to explain the sample of observations at hand. This decision problem can be tackled from the point of view of the model specification achieving the best in-sample fit.

This selection criterion is easy to apply as it consists in comparing one of the above defined measures of fit, computed for the competing model specifications, after adjusting them for the loss of sample degrees of freedom due to model parametrisation. Indeed, the value of these measures of fit can be improved by increasing model parametrisation, in particular when the parameter estimates are obtained by optimising a criteria functionally related to the selected measure of fit, as is the case when using the $\rho^2$ fit measure with a maximum likelihood estimate. Consequently, a penalty that increases with the number of model parameters should

[3] For multiple hurdle models involving many intercepts, the estimation of a specification without covariates may face serious numerical problems. If the mhurdle software fails to provide such an estimate, the total sum of squares $TSS$ is computed by substituting the sample average of $y$ for $\hat{y}_0$.

be added to the $R^2$ and $\rho^2$ fit measures to trade off goodness of fit improvements with parameter parsimony losses.

To define an adjusted pseudo coefficient of determination, we rely on Theil (1971)'s correction of $R^2$ in a linear regression model, defined by

$$\bar{R}^2 = 1 - \frac{n - K_0}{n - K}\frac{RSS}{TSS}, \tag{40}$$

where $K$ and $K_0$ stand for the number of parameters of the multiple hurdle model with covariates and without covariates, respectively [4]. Therefore, choosing the model specification with the largest $\bar{R}^2$ is equivalent to choosing the model specification with the smallest model residual variance estimate: $s^2 = \frac{RSS}{n-K}$.

To define an adjusted likelihood ratio index, we replace in this goodness of fit measure $\rho^2$ the log-likelihood criterion with the Akaike information criterion $AIC = -2\ln L(\hat{\theta}) + 2K$. Therefore, choosing the model specification with the largest

$$\bar{\rho}^2 = 1 - \frac{\ln L(\hat{\theta}) - K}{\ln L(\hat{\alpha}) - K_0} \tag{41}$$

is equivalent to choosing the model specification that minimises the Akaike (1973) predictor of the Kullback-Leibler Information Criterion (KLIC). This criterion measures the distance between the conditional density function $f(y|x;\theta)$ of a possibly misspecified parametric model and that of the true unknown model, denoted by $h(y|x)$. It is defined by the following formula:

$$KLIC = E\left[\ln\left(\frac{h(y|x)}{f(y|x;\theta_*)}\right)\right] = \int \ln\left(\frac{h(y|x)}{f(y|x;\theta_*)}\right) dH(y,x), \tag{42}$$

where $H(y,x)$ denotes the distribution function of the true joint distribution of $(y,x)$ and $\theta_*$ the probability limit, with respect to $H(y,x)$, of $\hat{\theta}$ the so called quasi-maximum likelihood estimator obtained by applying the maximum likelihood when $f(y|x;\theta)$ is misspecified.

### 3.3. Model selection using Vuong tests

Model selection can also be tackled from the point of view of the model specification that is favoured in a formal test comparing two model alternatives.

This second model selection criterion relies on the use of a test proposed by Vuong (1989). According to the rationale of this test, the "best" parametric model specification among a collection of competing specifications is the one that minimises the $KLIC$ criterion or, equivalently, the specification for which the quantity:

$$E[\ln f(y|x;\theta_*)] = \int \ln f(y|x;\theta_*) dH(y,x) \tag{43}$$

is the largest. Therefore, given two competing conditional models with density functions $f(y|x;\theta)$ and $g(y|x;\pi)$ and parameter vectors $\theta$ and $\pi$ of size $K$ and $L$, respectively, Vuong suggests to discriminate between these models by testing the null hypothesis:

[4] When the mhurdle software fails to provide the parameter estimates of the intercept-only model and the total sum of squares $TSS$ is computed by substituting the sample average of $y$ for $\hat{y}_0$, $K_0$ is set equal to 1.

$$H_0 : E[\ln f(y|x;\theta_*)] = E[\ln g(y|x;\pi_*)] \Longleftrightarrow E\left[\ln \frac{f(y|x;\theta_*)}{g(y|x;\pi_*)}\right] = 0,$$

meaning that the two models are equivalent, against:

$$H_f : E[\ln f(y|x;\theta_*)] > E[\ln g(y|x;\pi_*)] \Longleftrightarrow E\left[\ln \frac{f(y|x;\theta_*)}{g(y|x;\pi_*)}\right] > 0,$$

meaning that specification $f(y|x;\theta)$ is better than $g(y|x;\pi)$, or against:

$$H_g : E[\ln f(y|x;\theta_*)] < E[\ln g(y|x;\pi_*)] \Longleftrightarrow E\left[\ln \frac{f(y|x;\theta_*)}{g(y|x;\pi_*)}\right] < 0,$$

meaning that specification $g(y|x;\pi)$ is better than $f(y|x;\theta)$.

The quantity $E[\ln f(y|x;\theta_*)]$ is unknown but it can be consistently estimated, under some regularity conditions, by $1/n$ times the log-likelihood evaluated at the quasi-maximum likelihood estimator. Hence $1/n$ times the log-likelihood ratio (LR) statistic

$$LR(\hat{\theta},\hat{\pi}) = \sum_{i=1}^{n} \ln \frac{f(y_i|x_i;\hat{\theta})}{g(y_i|x_i;\hat{\pi})} \tag{44}$$

is a consistent estimator of $E\left[\ln \frac{f(y|x;\theta_*)}{g(y|x;\pi_*)}\right]$. Therefore, an obvious test of $H_0$ consists in verifying whether the LR statistic differs from zero. The distribution of this statistic can be worked out even when the true model is unknown, as the quasi-maximum likelihood estimators $\hat{\theta}$ and $\hat{\pi}$ converge in probability to the pseudo-true values $\theta_*$ and $\pi_*$, respectively, and have asymptotic normal distributions centred on these pseudo-true values.

The resulting distribution of $LR(\hat{\theta},\hat{\pi})$ depends on the relation linking the two competing models. To this purpose, Vuong differentiates among three types of competing models, namely: nested, strictly non nested and overlapping.

A parametric model $G_\pi$ defined by the conditional density function $g(y|x;\pi)$ is said to be nested in parametric model $F_\theta$ with conditional density function $f(y|x;\theta)$, if and only if any conditional density function of $G_\pi$ is equal to a conditional density function of $F_\theta$ almost everywhere (disregarding any zero probability sub-set of $(y,x)$ values, with respect to the true distribution function $H(y,x)$). This means that we can write a parametric constraint in the form $\theta = T(\pi)$, allowing to express model $G_\pi$ as a particular case of model $F_\theta$. Within our multiple hurdle special models this is the case when comparing two specifications differing only with respect to the presence or the absence of correlated disturbances. For these models, it is necessarily the case that $f(y|x;\theta_*) \equiv g(y|x;\pi_*)$. Therefore $H_0$ is tested against $H_f$.

If model $F_\theta$ is misspecified, it has been shown by Vuong that:

- under $H_0$, the quantity $2LR(\hat{\theta},\hat{\pi})$ converges in distribution towards a weighted sum of $K+L$ iid $\chi^2(1)$ random variables, where the weights are the $K+L$ almost surely real and non negative eigenvalues of the following $(K+L)\times(K+L)$ matrix:

$$W = \begin{bmatrix} -B_f A_f^{-1} & -B_{fg} A_g^{-1} \\ B_{fg}^{\top} A_f^{-1} & B_g A_g^{-1} \end{bmatrix},$$

where

$$A_f = E\left(\frac{\partial^2 \ln f(y|x;\theta_*)}{\partial\theta\partial\theta^\top}\right), \quad A_g = E\left(\frac{\partial^2 \ln g(y|x;\pi_*)}{\partial\pi\partial\pi^\top}\right),$$
$$B_f = E\left(\frac{\partial \ln f(y|x;\theta_*)}{\partial\theta}\frac{\partial \ln f(y|x;\theta_*)}{\partial\theta^\top}\right), \quad B_g = E\left(\frac{\partial \ln g(y|x;\pi_*)}{\partial\pi}\frac{\partial \ln g(y|x;\pi_*)}{\partial\pi^\top}\right),$$
$$B_{fg} = E\left(\frac{\partial \ln f(y|x;\theta_*)}{\partial\theta}\frac{\partial \ln g(y|x;\pi_*)}{\partial\pi^\top}\right).$$

To simplify the computation of this limiting distribution, one can alternatively use the weighted sum of $K$ iid $\chi^2(1)$ random variables, where the weights are the $K$ almost surely real and non negative eigenvalues of the following smaller $K \times K$ matrix:

$$\underline{W} = B_f\left[DA_g^{-1}D^\top - A_f^{-1}\right],$$

where $D = \frac{\partial T(\pi_*)}{\partial\pi^\top}$.

- under $H_f$, the same statistic converge almost surely towards $+\infty$.

Performing this standard LR test for nested models, requires to replace the theoretical matrices $W$ and $\underline{W}$ by a consistent estimator. Such an estimator is obtained by substituting matrices $A_f$, $A_g$, $B_f$, $B_g$ and $B_{fg}$ for their sample analogue:

$$\hat{A}_f = \frac{1}{n}\sum_{i=1}^n \frac{\partial^2 \ln f(y_i|x_i;\hat{\theta})}{\partial\theta\partial\theta^\top}, \quad \hat{A}_g = \frac{1}{n}\sum_{i=1}^n \frac{\partial^2 \ln g(y_i|x_i;\hat{\pi})}{\partial\pi\partial\pi^\top},$$
$$\hat{B}_f = \frac{1}{n}\sum_{i=1}^n \frac{\partial \ln f(y_i|x_i;\hat{\theta})}{\partial\theta}\frac{\partial \ln f(y_i|x_i;\hat{\theta})}{\partial\theta^\top}, \quad \hat{B}_g = \frac{1}{n}\sum_{i=1}^n \frac{\partial \ln g(y_i|x_i;\hat{\theta})}{\partial\theta}\frac{\partial \ln g(y_i|x_i;\hat{\theta})}{\partial\theta^\top},$$
$$\hat{B}_{fg} = \frac{1}{n}\sum_{i=1}^n \frac{\partial \ln f(y_i|x_i;\hat{\theta})}{\partial\theta}\frac{\partial \ln g(y_i|x_i;\hat{\theta})}{\partial\theta^\top}$$

and $D$ for $\hat{D} = \partial T(\hat{\pi})/\partial\pi^\top$.

The density function of this asymptotic test statistic has not been worked out analytically. Therefore, we compute it by simulation.

Hence, for a test with critical value $c$, $H_0$ is rejected in favour of $H_f$ if $2LR(\hat{\theta},\hat{\pi}) > c$ or if the p-value associated to the observed value of $2LR(\hat{\theta},\hat{\pi})$ is less than the significance level of the test.

Note that, if model $F_\theta$ is correctly specified, the asymptotic distribution of the LR statistic is, as expected, a $\chi^2$ random variable with $K - L$ degrees of freedom.

Two parametric models $F_\theta$ and $G_\pi$ defined by conditional distribution functions $f(y|x;\theta)$ and $g(y|x;\pi)$ are said to be strictly non-nested, if and only if no conditional distribution function of model $F_\theta$ is equal to a conditional distribution function of $G_\pi$ almost everywhere, and conversely. Within multiple hurdle special models this is the case when comparing two specifications differing with respect either to the censoring mechanisms in effect or to the

functional form of the desired consumption equation. For these models, it is necessarily the case that $f(y|x;\theta_*) \neq g(y|x;\pi_*)$ implying that both models are misspecified under $H_0$.

For such strictly non-nested models, Vuong has shown that:

- under $H_0$, the quantity $n^{-1/2}LR(\hat{\theta},\hat{\pi})$ converges in distribution towards a normal random variable with zero expectation and variance:

$$\omega^2 = V\left(\ln\frac{f(y|x;\theta_*)}{g(y|x;\pi_*)}\right)$$

  computed with respect to the distribution function of the true joint distribution of $(y,x)$.
- under $H_f$, the same statistic converge almost surely towards $+\infty$.
- under $H_g$, the same statistic converge almost surely towards $-\infty$.

Hence, $H_0$ is tested against $H_f$ or $H_g$ using the standardised LR statistic:

$$T_{LR} = \frac{LR(\hat{\theta},\hat{\pi})}{\sqrt{n}\hat{\omega}}, \tag{45}$$

where $\hat{\omega}^2$ denotes the following strongly consistent estimator for $\omega^2$:

$$\hat{\omega}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\ln\frac{f(y_i|x_i;\hat{\theta})}{g(y_i|x_i;\hat{\pi})}\right)^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\ln\frac{f(y_i|x_i;\hat{\theta})}{g(y_i|x_i;\hat{\pi})}\right)^2.$$

As a consequence, for a test with critical value $c$, $H_0$ is rejected in favour of $H_f$ if $T_{LR} > c$ or if the p-value associated to the observed value of $T_{LR}$ in less than the significance level of the test. Conversely, $H_0$ is rejected in favour of $H_g$ if $T_{LR} < -c$ or if the p-value associated to the observed value of $|T_{LR}|$ in less than the significance level of the test.

Note that, if one of models $F_\theta$ or $G_\pi$ is assumed to be correctly specified, the Cox (1961, 1962) LR test of non nested models needs to be used. Because this test is computationally awkward to implement and not really one of model selection, as it can lead to reject both competing models, it has not been programmed in **mhurdle**.

Two parametric models $F_\theta$ and $G_\pi$ defined by conditional distribution functions $f(y|x;\theta)$ and $g(y|x;\pi)$ are said to be overlapping, if and only if part of the conditional distribution function of model $F_\theta$ is equal to the conditional distribution function of $G_\pi$ but none of these models is nested in the other. Within multiple hurdle special models this is the case when comparing two specifications differing only with respect to the covariates taken into consideration, some of them being common to both models and others specific. For these models it is not clear *a priori* as to whether or not $f(y|x;\theta_*) = g(y|x;\pi_*)$ almost everywhere, except if we know *a priori* that at least one of the two competing models is correctly specified. As a consequence, the form of the asymptotic distribution of $LR(\hat{\theta},\hat{\pi})$ under $H_0$ is unknown, which prevents from performing a model selection test based on this statistic.

In the general case where both competing models are wrongly specified, Vuong suggests a sequential procedure which consists in testing first whether or not the variance $\omega^2$ equals zero (since $f(y|x;\theta_*) = g(y|x;\pi_*)$ almost everywhere if and only if $\omega^2 = 0$) and then, according to

the outcome of this test, in using the appropriate asymptotic $LR(\hat{\theta}, \hat{\pi})$ distribution to perform the model selection test.

To test $H_0^\omega : \omega^2 = 0$ against $H_A^\omega : \omega^2 \neq 0$, Vuong suggests to use, as a test statistic, the above defined strongly consistent estimator for $\omega^2$, $\hat{\omega}^2$, and proves that:

- under $H_0^\omega$, the quantity $n\hat{\omega}^2$ converges in distribution towards the same limiting distribution like that of statistic $2LR(\hat{\theta}, \hat{\pi})$ when used for discriminating two misspecified nested models.
- under $H_A^\omega$, the same statistic converge almost surely towards $+\infty$.

Therefore, performing this variance test requires to compute the eigenvalues of a consistent estimate of matrix $W$ or $\underline{W}$, and derive by simulation the density function of the corresponding weighted sum of iid $\chi^2(1)$ random variables.

Hence, for a test with critical value $c$, $H_0^\omega$ is rejected in favour of $H_A^\omega$ if $n\hat{\omega}^2 > c$ or if the p-value associated to the observed value of $n\hat{\omega}^2$ is less than the significance level of the test.

Note, that an asymptotically equivalent test is obtained by replacing in statistics $n\hat{\omega}^2$, $\hat{\omega}^2$ by:

$$\tilde{\omega}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\ln\frac{f(y_i|x_i;\hat{\theta})}{g(y_i|x_i;\hat{\pi})}\right)^2.$$

The second step in discriminating two overlapping models depends on the outcome of the variance test.

- If $H_0^\omega$ is not rejected, one should conclude that the two models cannot be discriminated given the data, since assuming $\omega^2 = 0$ implies that $H_0$ means that the two models are equivalent.
- If $H_0^\omega$ is rejected, the test of $H_0$ against $H_f$ or $H_g$ must be carried out using the standardised LR statistic $T_{LR}$, as for discriminating between two strictly non-nested models. Indeed, $H_0$ is still possible when $\omega^2 \neq 0$. Note, that this sequential procedure of testing $H_0$ against $H_f$ or $H_g$ has a significance level bounded above by the maximum of the significance levels used for performing the variance and the standardised LR tests.

Finally, if one of the two competing models is supposed to be correctly specified, then the two models are equivalent if and only if the other model is correctly specified and if and only if the conditional density functions of the two models are identical almost everywhere. In this case we can bypass the variance test and directly construct a model selection test based on the $2LR(\hat{\theta}, \hat{\pi})$ test statistic used for discriminating between two nested models.

## 4. Software rationale

There are three important issues to be addressed to correctly implement in R the modelling strategy described in the previous sections. The first one is to provide a good interface to describe the model to be estimated. The second one is to find good starting values for

computing model estimates. The third one is to have flexible optimisation tools for likelihood maximisation.

### 4.1. Model syntax

In R, the model to be estimated is usually described using formula objects, the left-hand side denoting the censored dependent variable `y` and the right-hand side the functional relation explaining `y` as a function of covariates. For example, `y ~ x1 + x2 * x3` indicates that `y` linearly depends on variables `x1`, `x2`, `x3` and on the interaction term `x2` times `x3`.

For the models implemented in **mhurdle**, three kinds of covariates should be specified: those of the good selection equation (hurdle 1) denoted $x_1$, those of the desired consumption equation (hurdle 2), denoted $x_2$, and those of the purchasing equation (hurdle 3), denoted $x_3$.

To define a model with three kinds of covariates, a general solution is given by the **Formula** package developed by Zeileis and Croissant (2010), which provides extended formula objects. To define a model where `y` is the censored dependent variable, `x11` and `x12` two covariates for the good selection equation, `x21` and `x22` two covariates for the desired consumption equation, and `x31` and `x32` two covariates for the purchasing equation, we use the following commands :

```
R> library("Formula")
R> f <- Formula(y ~ x11 + x12  | x21 + x22 | x31 + x32)
```

### 4.2. Starting values

For the models we consider, the log-likelihood function will be, in general, not concave. Moreover, this kind of models are highly non linear with respect to parameters, and therefore difficult to estimate. For these reasons, the question of finding good starting values for the iterative computation of parameter estimates is crucial.

As a less computer intensive alternative to maximum likelihood estimation, Heckman (1976) has suggested a two step estimation procedure based on a respecification of the censored variable linear regression model, sometimes called "Heckit" model, avoiding inconsistency of the ordinary least-squares estimator. This two step estimator is consistent but inefficient. It is implemented in package **sampleSelection** (Toomet and Henningsen 2008b).

According to Carlevaro, Croissant, and Hoareau (2008) experience in applying this estimation procedure to double hurdle models, this approach doesn't seem to work well with correlated hurdle models. Indeed, except for the very special case of models 2, 3 and 4, the probability of observing a censored purchase is not that of a simple probit model (see Table 1).

As noted previously, for uncorrelated single hurdle models, the estimation may be performed in a sequence of two simple estimations, namely the maximum likelihood estimation of a standard dichotomous probit model, followed by the ordinary least-squares estimation of a linear, log-linear or linear-truncated regression model. In the last case, package **truncreg** (Croissant 2009) is used.

For correlated single hurdle 1 model 2, the maximum likelihood estimate of the parameters of the corresponding uncorrelated model ($\rho_{12} = 0$) is used as starting values.

For P-Tobit models (4 and 7), the starting values are computed using an Heckman-like two step procedure. In the first step, parameters $\beta_3$ are estimated using a simple probit. In the second step, a linear regression model is estimated by ordinary least squares using the sub-sample of uncensored observations and $y_i\Phi(\hat{\beta}_3^\top x_{3i})$ or $\ln y_i + \ln \Phi(\hat{\beta}_3^\top x_{3i})$ (in the case of a log-normal specification) as dependent variable.

For Tobit model (3), the least squares estimate of the linear regression model is used as starting values.

For double hurdle model (5), the starting values for $\beta_1$ are obtained by estimating a probit model and those for $\beta_2$ using a least squares estimate with the truncated sample of a linear regression model assuming $\rho_{12} = 0$.

Finally, for models involving hurdles 1 and 3 (models 6 and 8), we use two probit models to get starting values for $\beta_1$ and $\beta_2$. Then, we estimate a linear regression model by ordinary least squares with the sub-sample of uncensored observations using $y_i\Phi(\hat{\beta}_3^\top x_{3i})$ or $\ln y_i + \ln \Phi(\hat{\beta}_3^\top x_{3i})$ (in the case of a log-normal specification) as dependent variable and assuming no correlation between the desired consumption equation and these two hurdles.

### 4.3. Optimisation

Two kinds of algorithms are currently used for maximum likelihood estimation. The first kind of algorithms can be called "Newton-like" methods. With these algorithms, at each iteration, the hessian matrix of the log-likelihood is computed, using either the second derivatives of the log-likelihood (Newton-Raphson method) or the outer product of the gradient (Berndt, Hall, Hall, Hausman or BHHH method). This approach is very powerful if the log-likelihood is well-behaved, but it may perform poorly otherwise and fail after a few iterations.

The second algorithm, called Broyden, Fletcher, Goldfarb, Shanno or BFGS method, updates at each iteration an estimate of the hessian matrix of the log-likelihood. It is often more robust and may perform better in cases where the formers don't work.

Two optimisation functions are included in core R: `nlm`, which uses the Newton-Raphson method, and `optim` , which uses the BFGS method (among others). The recently developed **maxLik** package by Toomet and Henningsen (2008a) provides a unified framework. With a unique interface, all the previously described methods are available.

The behaviour of `maxLik` can be controlled by the user using `mhurdle` arguments like `print.level` (from 0-silent to 2-verbal), `iterlim` (the maximum number of iterations), `methods` (the method used, one of `"nr"`, `"bhhh"` or `"bfgs"`) that are passed to `maxLik`.

Some models require the computation of the bivariate normal cumulative density function. We use the **pbivnorm** package (Kenkel 2011) which provides a vectorised (and therefore fast convenient) function to compute the bivariate normal cdf.

## 5. Examples

The package is loaded using:

```
R> library("mhurdle")
```

To illustrate the use of **mhurdle**, we use the `Comics` data frame which contains data about the readings of comics. It is part of a survey conducted by the INSEE (the French national statistical institute) in 2003 about cultural and sportive practises[5]. The explained variable is the number of comics read during the last 12 months by one (randomly chosen) member of the household. There are 5159 observations.

We emphasise that the observed censored variable to be explained is not an expenditure but a service derived from the use of a durable good, namely the comic book library to which the comic book reader has access. Therefore, hurdles 2 and 3 of our modelling paradigm must be reinterpreted as mechanisms describing the process of building up the comic book library and that of planning the intensity of use of the library, respectively. Note also that **mhurdle** treats the dependent variable as a continuous quantitative variable, while it is in fact a discrete count variable. However, the high number of readings during a year by a comic book reader fully justifies this numerical approximation.

```
R> data("Comics", package = "mhurdle")
R> head(Comics, 3)
```

```
  comics  area income cu size age nationality    empl gender couple educ
1      0 paris   80.0  2    2  75       natfr retired   male    yes   18
2      0 paris   41.5  1    1  77       natfr retired female     no   18
3      0 paris   80.0  2    2  43       natfr  interm female     no   22
```

```
R> mean(Comics$comics == 0)
```

```
[1] 0.7828705
```

```
R> max(Comics$comics)
```

```
[1] 520
```

The number of comics read is zero for about 78% of the sample and the maximum value is 520. The covariates of this data frame are :

**area:** one of `rural`, `small`, `medium`, `large` and `paris`

**income:** the income of the household (in thousands of euros per month),

**cu:** the number of consumption units (one for the first two adults, one half for other members of the household),

**size :** the number of persons in the household,

[5]The data is available at http://insee.fr/fr/themes/detail.asp?reg_id=0&ref_id=fd-parcul03. Main results are presented in Muller (2005).

**age :** the age of the person,

**empl :** the kind of occupation, a qualitative factor with 9 levels,

**gender:** one of `male` and `female`,

**couple,** "does the person live in couple ?", a qualitative factor with levels `yes` and `no`,

**educ :** the number of years of education.

### 5.1. Estimation

The estimation is performed using the `mhurdle` function, which has the following arguments:

**formula:** a formula describing the model to estimate. It should have three parts on the right-hand side specifying, in the first part, the good selection equation covariates, in the second part, the desired consumption equation covariates and, in the third part, the purchasing equation covariates.

**data:** a data frame containing the observations of the variables present in the formula.

**subset, weights, na.action:** these are arguments passed on to the model.frame function in order to extract the data suitable for the model. These arguments are present in the `lm` function and in most of the estimation functions.

**start:** the starting values. If `NULL`, the starting values are computed as described in section 4.2.

**dist:** this argument indicates the functional form of the desired consumption equation, which may be either log-normal `"l"` (the default), normal `"n"` or truncated normal `"t"`.

**corr:** this argument indicates whether the disturbance of the good selection equation (hurdle 1) or that of the purchasing equation (hurdle 3) is correlated with that of the desired consumption equation. This argument is in this case respectively equal to `"h1"` or `"h3"`, or `NULL` (the default) in case of no correlation,

**...** further arguments that are passed to the optimisation function `maxLik`.

Different combinations of these arguments lead to a large variety of models. Note that some of them are logically inconsistent and therefore irrelevant. For example, a model with no good selection equation and `corr = "h1"` is logically inconsistent.

To illustrate the use of **mhurdle** package, we first estimate a simple Tobit model, which we call `model3` ; the income is first divided by the number of consumption units and then by its sample mean. Powers up to three for the log of income are introduced.

```
R> Comics$incu <- with(Comics, income / cu)
R> Comics$incum <- with(Comics, incu / mean(incu))
R> model3 <- mhurdle(comics ~ 0 | log(incum) + I(log(incum)^2) +
+                    I(log(incum)^3) + age  + gender + educ +
+                    size| 0, data = Comics, dist = "n", method = 'bfgs')
```

Note that the first and the third part of the formula are 0, as there is no good selection and no purchasing equations.

Consider now that some covariates explain the fact that the good is selected, and not the level of consumption if the good is chosen. In this case, we estimate the following dependent double hurdle model, which we call `model5d`. We keep the income and the size of the household as covariates for the desired consumption equation and move the other covariates to the first part of the formula.

```
R> model5d <- mhurdle(comics ~ gender + educ + age |  log(incum) +
+                     I(log(incum)^2) + I(log(incum)^3) + size | 0,
+                     data = Comics, corr = "h1", dist = "n", method = 'bfgs')
```

The same model without correlation is called `model5i`, and can easily be obtained by updating `model5d` :

```
R> model5i <- update(model5d, corr = NULL)
```

If one wants that zeros only arise from the selection mechanism, one has to switch the `dist` argument to `"l"`, so that a log-normal distribution is introduced. This can be done easily by updating the previous model and this leads to a model called `model2d` :

```
R> model2d <- update(model5d, dist = "l")
```

The independent version is then easily obtained :

```
R> model2i <- update(model2d, corr = NULL)
```

The last model we estimate is a dependent triple hurdle model ; compared to the double hurdle model previously estimated, we move the `age` covariate from the selection to the purchasing equation :

```
R> model8d1 <- mhurdle(comics ~ gender + educ  |  log(incum) +
+                      I(log(incum)^2) + I(log(incum)^3) + size | age,
+                      data = Comics, corr = "h1", dist = "n", method = 'bfgs')
```

## 5.2. Methods

A `summary` method is provided for `mhurdle` objects :

```
R> summary(model8d1)
```

```
Call:
mhurdle(formula = comics ~ gender + educ | log(incum) + I(log(incum)^2) +
    I(log(incum)^3) + size | age, data = Comics, dist = "n",
```

```
    corr = "h1", method = "bfgs")

Frequency of 0:  0.78287

BFGS maximisation method
117 iterations, 0h:0m:12s
g'(-H)^-1g = 4.48E-05


Coefficients :
                     Estimate Std. Error  t-value  Pr(>|t|)
h1.(Intercept)      -1.755587   0.229763  -7.6409 2.154e-14 ***
h1.genderfemale     -0.547607   0.130011  -4.2120 2.531e-05 ***
h1.educ              0.188445   0.014847  12.6927 < 2.2e-16 ***
h2.(Intercept)     -40.533234   3.484658 -11.6319 < 2.2e-16 ***
h2.log(incum)       14.633088   2.896964   5.0512 4.391e-07 ***
h2.I(log(incum)^2)  -3.777712   2.648899  -1.4261  0.153827
h2.I(log(incum)^3)  -4.457532   1.701804  -2.6193  0.008811 **
h2.size              5.253440   0.952567   5.5150 3.487e-08 ***
h3.(Intercept)      11.318180   2.048508   5.5251 3.293e-08 ***
h3.age              -0.149555   0.026846  -5.5709 2.534e-08 ***
sigma               55.212046   1.341107  41.1690 < 2.2e-16 ***
rho                 -0.283790   0.048636  -5.8349 5.381e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -7322.1 on 12 Df

R^2 :
 Coefficient of determination : 0.0048516
 Likelihood ratio index       : 0.032983
```

This method displays the percentage of 0 in the sample, the table of parameter estimates, and two measures of goodness of fit.

`coef`, `vcov`, `logLik`, `fitted` and `predict` methods are provided in order to extract part of the results.

Parameter estimates and the estimated asymptotic variance matrix of maximum likelihood estimators are extracted using the usual `coef` and `vcov` functions. `mhurdle` object methods have a second argument indicating which subset has to be returned (the default is to return all).

```
R> coef(model8d1, "h2")
```

```
    (Intercept)      log(incum) I(log(incum)^2) I(log(incum)^3)            size
     -40.533234       14.633088       -3.777712       -4.457532        5.253440
```

```
R> coef(model5d, "h1")
```

```
 (Intercept) genderfemale         educ          age
  2.34059781  -0.56070370   0.10210778  -0.05598002
```

```
R> coef(model5d, "sigma")
```

```
   sigma
55.48789
```

```
R> coef(summary(model8d1), "h3")
```

```
             Estimate Std. Error   t-value    Pr(>|t|)
(Intercept) 11.318180 2.04850799  5.525085 3.29327e-08
age         -0.149555 0.02684572 -5.570905 2.53419e-08
```

```
R> vcov(model8d1, "h3")
```

```
            (Intercept)           age
(Intercept)  4.19638500 -0.0547908878
age         -0.05479075  0.0007206927
```

Log-likelihood may be obtained for the estimated model or for a "naive" model, defined as a model without covariates :

```
R> logLik(model5d)
```

```
[1] -7274.991
```

```
R> logLik(model5d, naive = TRUE)
```

```
[1] -7571.842
```

Fitted values are obtained using the `fitted` method. The output is a matrix whose two columns are the estimated probability of censoring $\mathrm{P}(y = 0)$ and the estimated expected value of an uncensored dependent variable observation $\mathrm{E}(y|y > 0)$.

```
R> head(fitted(model5d))
```

```
        P(y=0) E(y|y>0)
[1,] 0.8655939 32.94001
[2,] 0.9492517 30.09306
[3,] 0.6639017 37.10684
[4,] 0.9796777 26.43918
[5,] 0.8614599 33.03126
[6,] 0.6387773 37.83426
```

A `predict` function is also provided, which returns the same two columns for given values of the covariates.

```
R> predict(model5d,
+          newdata = data.frame(
+            comics = c(0, 1, 2),
+            gender = c("female", "female", "male"),
+            age = c(20, 18, 32),
+            educ = c(10, 20, 5),
+            incum = c(4, 8, 2),
+            size = c(2, 1, 3)))
```

```
        P(y=0) E(y|y>0)
[1,] 0.6729281 36.35521
[2,] 0.8357142 29.38384
[3,] 0.7754750 35.26525
```

For model evaluation and selection purposes, goodness of fit measures and Vuong tests described in section 3 are provided. These criteria allow to select the most empirically relevant model specification.

Two goodness of fit measures are provided. The first measure is an extension to limited dependent variable models of the classical coefficient of determination for linear regression models. This pseudo coefficient of determination is computed both without (see formula (38)) and with (see formula (40)) adjustment for the loss of sample degrees of freedom due to model parametrisation. The unadjusted coefficient of determination allows to compare the goodness of fit of model specifications having the same number of parameters, whereas the adjusted version of this coefficient is suited for comparing model specifications with a different number of parameters.

```
R> rsq(model5d, type = "coefdet")
```

```
[1] 0.01210285
```

The second measure is an extension to limited dependent variable models of the likelihood ratio index for qualitative response models. This pseudo coefficient of determination is also computed both without (see formula (39)) and with (see formula (41)) adjustment for the loss of sample degrees of freedom due to model parametrisation, in order to allow model comparisons with the same or with a different number of parameters.

```
R> rsq(model5d, type = "lratio", adj = TRUE)
```

```
[1] 0.03825995
```

The Vuong test based on the $T_{LR}$ statistic, as presented in section 3.3 (see formula (45)), is also provided as a criteria for model selection within the family of 8 strictly non-nested models of Figure 1.

```
R> vuongtest(model5d, model8d1)
```

```
        Vuong Test (non-nested)

data:  model5d model8d1
z = 4.7179, p-value = 1.191e-06
```

According to this outcome, the null hypothesis stating the equivalence between the two models is strongly rejected in favour of the alternative hypothesis stating that `model5d` is better than `model8d1`.

Note that Vuong tests for strictly non-nested mhurdle models can also be performed using the `vuong` function of the **pscl** package of Jackman (2011) .

Testing the hypothesis of no correlation between the good selection mechanism and the desired consumption equation can be performed as a Vuong test of selection between two nested models, differing only with respect to the value of the correlation coefficient $\rho_{12}$, namely the test of the hypothesis $H_0 : \rho_{12} = 0$, specifying an independent mhurdle model, against the alternative hypothesis $H_a : \rho_{12} \neq 0$, specifying a corresponding dependent mhurdle model. This test is performed using the log-likelihood ratio (LR) statistic (44). As explained in section 3.3, the critical value or the p-value to be used to perform this test is not the same depending on the model builder believes or not that his unrestricted model, assuming $-1 < \rho_{12} < 1$, is correctly specified. In the first case, the p-value is computed using the standard chi square distribution, whereas in the second case a weighted chi square distribution is used.

```
R> vuongtest(model2d, model2i, type = 'nested', hyp = TRUE)
```

```
        Vuong Test (nested)

data:  model2d model2i
chisq = 3.5204, df = 1, p-value = 0.06062
```

```
R> vuongtest(model2d, model2i, type = 'nested', hyp = FALSE)
```

```
        Vuong Test (nested)

data:  model2d model2i
wchisq = 3.5204, df = 0.919, p-value = 0.066
```

According to these outcomes, the null hypothesis of zero correlation is accepted or rejected at almost the same significance level, which must be set higher than 0.066 for acceptance and lower than 0.061 for rejection.

Testing this hypothesis of no correlation by assuming the unrestricted model correctly specified, can be also performed by means of the classical Wald test, using the t-statistic or the p-value of the correlation coefficient estimate presented in the table of parameter estimates of the dependent model (`model2d`).

```
R> coef(summary(model2d), "rho")
```

```
      Estimate Std. Error   t-value   Pr(>|t|)
rho -0.1513119 0.07783631 -1.943975 0.05189841
```

According to this test outcome, the hypothesis of zero correlation is accepted at a little less stringent significance level than with a Vuong test.

Finally, to illustrate the use of the Vuong test for discriminating between two overlapping models, we consider a slightly different Tobit model obtained by removing the `age` covariate and adding the `empl` and `area` covariates :

```
R> model3bis <- mhurdle(comics ~ 0 | log(incum) + I(log(incum)^2) +
+                       I(log(incum)^3)  + gender + educ + age +
+                       empl+area| 0, data = Comics, dist = "n", method = 'bfgs')
```

In this case, the Vuong test is performed in two steps. Firstly a test of the null hypothesis $\omega^2 = 0$, meaning that the two models are equivalent, is undertaken.

```
R> vuongtest(model3, model3bis, type="overlapping")
```

```
        Vuong Test (overlapping)

data:  model3 model3bis
wchisq = 41.3014, df = 12.54, p-value = 0.001
```

This null hypothesis is here strongly rejected. Therefore, we can test the equivalence of these two models as if they were strictly non-nested.

```
R> vuongtest(model3, model3bis, type="non-nested")
```

```
        Vuong Test (non-nested)

data:  model3 model3bis
z = -0.7078, p-value = 0.2395
```

According to the outcome of this second test, we conclude that these two model specifications cannot be empirically discriminated.

If one of two overlapping models is assumed to be correctly specified, we can bypass the first step of this Voung test (the variance test) and proceed as if we had to discriminate between two nested models.

```
R> vuongtest(model3bis, model3, type="overlapping", hyp=TRUE)

        Vuong Test (overlapping)

data:  model3bis model3
wchisq = 9.0973, df = 10.101, p-value = 0.992
```

Once again, the equivalence of the two models is not rejected.

## 6. Conclusion

**mhurdle** aims at providing a unified framework allowing to estimate and assess a variety of extensions of the standard Tobit model particularly suitable for single-equation demand analysis not currently implemented in R. It explains the presence of a large proportion of zero observations for a dependent variable by means of up to three censoring mechanisms, called hurdles. Inspired by the paradigms used for analysing censored household expenditure data, these hurdles express: (i) a non economic decision mechanism for a good rejection or selection motivated by ethical, psychological or social considerations; (ii) an economic decision mechanism for the desired level of consumption of a previously selected good, which can turn out to be negative leading to a nil consumption; (iii) an economic or non economic decision mechanism for the time frequency at which the desired quantity of a selected good is bought or consumed. Interdependence between these censoring mechanisms is modelled by assuming a possible correlation between the random disturbances in the model relations. Despite the particular area of application from which the above mentioned censoring mechanisms stem, the practical scope of **mhurdle** models doesn't seem to be restricted to empirical demand analysis.

To provide an operational and efficient statistical framework, **mhurdle** models are specified in a fully parametric form allowing statistical estimation and testing within the maximum likelihood inferential framework. Tools for model evaluation and selection are provided, based on the use of goodness of fit measure extensions of the classical coefficient of determination and of the likelihood ratio index of McFadden, as well as on the use of Vuong tests for nested, strictly non-nested and overlapping model comparison when none, one or both of two competing models are misspecified.

Tests of **mhurdle** computing procedures with a wide variety of simulated and observational data have proved the performance and robustness of **mhurdle** package. Still, extensions and improvements of the software are under way, notably the estimation of trivariate hurdle models in their full generality and the design of a consistent Vuong testing strategy for discriminating between a numerous set of competing models. Other desirable extensions, like the use of more general functional forms of the desired consumption relation[6] or of less stringent distributional

[6]See Poirier (1978), Lankford and Wyckoff (1991) and Jones and Yen (2000).

assumptions on which semi-parametric or nonparametric estimation methods are based, will be tackled once the actual scope of our models is established through diversified empirical applications. Research is continuing in this direction.

## References

Akaike H (1973). "Information Theory and an Extension of the Maximum Likelihood Principle." In B Petrov, F Csake (eds.), *Second International Symposion on Information Theory*. Budapest: Akademiai Kiado.

Amemiya T (1985). *Advanced Econometrics*. Harvard University Press, Cambridge (MA).

Blundell R, Meghir C (1987). "Bivariate Alternatives to the Tobit Model." *Journal of Econometrics*, **34**, 179–200.

Carlevaro F, Croissant Y, Hoareau S (2008). "Modélisation Tobit à double obstacle des dépenses de consommation : Estimation en deux étapes et comparaisons avec la méthode du maximum de vraisemblance." In *XXV journées de microéconomie appliqué*. University of la Réunion.

Cox DR (1961). "Tests of Separate Families of Hypotheses." In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 105–123.

Cox DR (1962). "Further Results on Tests of Separate Families of Hypotheses." *Journal of the Royal Statistical Society, Series B*, **24**, 406–424.

Cragg JG (1971). "Some Statistical Models for Limited Dependent Variables with Applications for the Demand for Durable Goods." *Econometrica*, **39**(5), 829–44.

Croissant Y (2009). *truncreg: Truncated Regression Models*. R package version 0.1-1, URL http://CRAN.R-project.org/package=truncreg.

Deaton A, Irish M (1984). "A Statistical Model for Zero Expenditures in Household Budgets." *Journal of Public Economics*, **23**, 59–80.

Heckman J (1976). "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models." *Annals of Economic and Social Measurement*, **5**, 475–92.

Hoareau S (2009). *Modélisation économétrique des dépenses de consommation censurées*. Ph.D. thesis, Faculty of Law and Economics, University of La Réunion.

Jackman S (2011). *pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory, Stanford University*. Department of Political Science, Stanford University, Stanford, California. R package version 1.04.1, URL http://pscl.stanford.edu/,http://CRAN.R-project.org/package=pscl.

Jones A, Yen S (2000). "A Box-Cox double-hurdle model." *Manchester School*, **68(2)**, 203–221.

Kenkel B (2011). *pbivnorm: Vectorized Bivariate Normal CDF*. R package version 0.5-0, URL http://CRAN.R-project.org/package=pbivnorm.

Kleiber C, Zeileis A (2008). *Applied Econometrics with R*. Springer-Verlag, New York. ISBN 978-0-387-77316-2, URL http://CRAN.R-project.org/package=AER.

Lankford R, Wyckoff J (1991). "Modeling charitable giving using a Box-Cox standard Tobit model." *Review of Economics and Statistics*, **73(3)**, 460–470.

McFadden D (1974). "The Measurement of Urban Travel Demand." *Journal of Public Economics*, **3**, 303–328.

Muller L (2005). "Pratique sportive et activités culturelles vont souvent de pair." *INSEE Première*, **1008**.

Poirier D (1978). "The use of the Box-Cox transformation in limited dependent variable models." *Journal of the American Statistical Association*, **73**, 284–287.

Pudney S (1989). *Modelling Individual Choice. The Econometrics of Corners, Kinks and Holes*. Basil Blackwell, Oxford and New York. ISBN 0-631-14589-3.

Theil H (1971). *Principles of Econometrics*. New York: John Wiley and Sons.

Therneau T, Lumley T (2008). *survival: Survival Analysis, Including Penalised Likelihood*. R package version 2.34-1, URL http://CRAN.R-project.org/package=survival.

Tobin J (1958). "Estimation of Relationships for Limited Dependent Variables." *Econometrica*, **26**(1), 24–36.

Toomet O, Henningsen A (2008a). *maxLik: Maximum Likelihood Estimation*. R package version 0.5-8, URL http://CRAN.R-project.org/package=maxLik,http://www.maxLik.org.

Toomet O, Henningsen A (2008b). "Sample Selection Models in R: Package sampleSelection." *Journal of Statistical Software*, **27**(7). URL http://www.jstatsoft.org/v27/i07/,http://CRAN.R-project.org/package=sampleSelection.

Vuong QH (1989). "Likelihood Ratio Tests for Selection and Non-Nested Hypotheses." *Econometrica*, **57**(2), 397–333.

Zeileis A, Croissant Y (2010). "Extended Model Formulas in R: Multiple Parts and Multiple Responses." *Journal of Statistical Software*, **34**(1), 1–13. ISSN 1548-7660. URL http://www.jstatsoft.org/v34/i01,http://CRAN.R-project.org/package=Formula.

## Affiliation:

Fabrizio Carlevaro
Faculté des sciences économiques et sociales
Université de Genève
Uni Mail

40 Bd du Pont d'Arve
CH-1211 Genève 4
Telephone: +41/22/3798914
E-mail:fabrizio.carlevaro@unige.ch

Yves Croissant
Faculté de Droit et d'Economie
Université de la Réunion
15, avenue René Cassin
BP 7151
F-97715 Saint-Denis Messag Cedex 9
Telephone: +33/262/938446
E-mail: yves.croissant@univ-reunion.fr

Stéphane Hoareau
Faculté de Droit et d'Economie
Université de la Réunion
15, avenue René Cassin
BP 7151
F-97715 Saint-Denis Messag Cedex 9
Telephone: +33/262/938446
E-mail: stephane.hoareau@univ-reunion.fr