Multiple Hurdle Tobit 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 a real-world example.

Keywords: Households' expenditure survey analysis, censored regression models, hurdle models, Tobit models, maximum likelihood estimation, goodness of fit measures, Vuong tests, 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 also been treated for a long time in the statistics literature dealing with survival models.

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 following three zero expenditure generating mechanisms.

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

A good selection mechanism (hurdle 1) . According to this mechanism, the consumer[2] 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, was 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 function of explanatory variables and of a normal random disturbance, with a possible correlation between the disturbances of different latent variables in order to account for a possible simultaneity of the decisions modelled by censoring mechanisms. To model possible departures of the observed dependent variable to normality, we use flexible transformations allowing to rescale skewed or leptokurtic random variables to normality. 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.

These formal models have been primarily developed to deal with censored household expenditure data, and numerous applications have been carried out in this field. Complementing a

[2]The consumer we are reffering to is that of the microeconomic theory, an abstract economic agent responsible of the decisions of a consumption unit that may be an individual, a family, a household. According to the economic litterature, we term this concept "the consumer" by convenience.

previous survey by Smith (2002), Table 1 gives an updated overview of these studies. We note a late popularity of Cragg's approach, as the first applications of hurdle models are published in the late of 1980s, namely almost two decades after the publication of Cragg's paper in Econometrica. Since then, a large variety of demand models including one or two among the previous three censoring mechanisms are estimated. However, none of these studies use the three censoring mechanisms we consider, jointly. From the 1990s on, many studies account for deviations of the desired consumption relation to homoscedasticity and normality by modelling the standard error of this variable as a non negative parametric function of some explanatory variables and by transforming its distribution to normality using either the Box-Cox or the inverse hyperbolic sine transformations. The estimation of a correlation coefficient between disturbances is also performed in several of these studies, with an increasing succes over time, in terms of statistical significance of estimates.

Table 1: Litterature overview of applied papers on demand analysis using multiple hurdle Tobit models

Authors	Goods	Hurdles			Distribution of y_2^* [a]	Skedasticity	Sample size	% of zeros	Correlation coefficient	
		h1	h2	h3					estimate [b]	signif. [c]
Cheng and Capps (1988)	finfish and shellfish	■	□	□	log-normal	□	9422	72-83	-0.69, 0.04	**, ns
Haines, Guilkey, and Popkin (1988)	milk, bread	■	□	□	tr. normal	□	5406	43-71		
Jones (1989)	cigarettes	■	□	□	normal	□	1565	40		ns
Jones (1989)	cigarettes	■	■	□	normal	□	1565	40		ns
Popkin, Guilkey, and Haines (1989)	milk, bread	■	□	□	tr. normal	□	1068-5406	na		
Reynolds (1990)	fresh vegetables	■	□	□	tr. normal	□	3368	32		
Blaylock and Blisard (1992)	cigarettes	■	■	□	normal	□	2962	61	na	ns
Gould (1992)	cheese	□	■	■	normal	□	5017	59	0.22	***
Blaylock and Blisard (1993)	wine	■	■	□	normal	□	916	80	na	ns
Blaylock and Blisard (1993)	wine	■	□	□	normal	□	916	80	na	ns
Yen (1993)	food away from home	■	□	□	Box-Cox	■	4199	12		
Blisard and Blaylock (1993)	butter	■	■	□	normal	□	5000	na		
Blisard and Blaylock (1993)	butter	□	■	■	normal	□	5000	na		
Yen (1994)	alcohol	■	□	□	Box-Cox	■	4245	72		
Gao, Wailes, and Cramer (1995)	rice	■	■	□	IHS	■	4273	67	0.24	***
Yen, Dellenbarger, and Schupp (1995)	crawfish	■	■	□	Box-Cox	□	915	78		
Yen and Su (1995)	butter	■	■	□	normal	■	8083	81		
Yen and Su (1995)	butter	□	■	■	normal	■	8083	81		
Yen (1995)	alcohol	■	□	□	Box-Cox	□	4538	88		
Yen (1995)	alcohol	■	□	□	IHS	□	4538	88		
Wang, Jensen, and Yen (1996b)	eggs	■	■	□	normal	□	1492	na		
Garcia and Labeaga (1996)	cigarettes	■	■	□	normal	□	23669	41	na	ns
Garcia and Labeaga (1996)	cigarettes	■	□	□	normal	□	23669	41		
Yen and Huang (1996)	finfish	■	■	□	IHS	■	4066	74		
Yen and Jones (1996)	cigarettes	■	■	□	Box-Cox	□	3801	57	-2.63*	***
Su and Yen (1996)	pork	■	■	□	IHS	■	4198	30		
Su and Yen (1996)	pork	□	■	■	IHS	■	4198	30		
Burton, Dorsett, and Young (1996)	meat	■	■	□	Box-Cox	□	na	na	na	ns
Yen and Jensen (1996)	alcohol	■	■	□	IHS	■	9552	54		
Wang, Gao, Wailes, and Cramer (1996a)	alcohol	■	■	□	normal	□	ns	ns	0.77	***

`na` means that the information was not in the article. A black box in columns `h1-h2-h3` indicates that the relevant hurdle is taken into account, and in column `skedasticity` that heteroscedasticity is taken into account. (a) `IHS` stands for Inverse Hyperbolic Sine transformation and `tr.normal` for truncated normal distribution. (b) A starred value indicates a covariance estimate. (c) (*), (**) and (***) indicate that the correlation coefficient or the covariance estimate is significant at the 10%, 5% or 1% level, respectively ; `ns` that it is non significant at the 10% level.

Continued on next page

Table 1 – *Continued from previous page*

Authors	Goods	Hurdles			Distribution of y_2^* [(a)]	Skedasticity	Sample size	% of zeros	Correlation coefficient	
		h1	h2	h3					estimate[(b)]	signif.[(c)]
Yen, Jensen, and Wang (1996)	egg	■	■	□	IHS	□	4230	58	4.72*	***
Yen and Jones (1997)	cheese	■	■	□	IHS	□	4245	18	na	ns
Yen (1999)	cigarettes	■	□	□	tr. normal	□	5814	75		
Bertail, Caillavet, and Nichele (1999)	home produced food	■	■	□	normal	□	6121	76	0.20	ns
Burton, Dorsett, and Young (2000)	meat	■	■	□	Box-Cox	□	na	na	na	ns
Jones and Yen (2000)	beef	■	■	□	Box-Cox	□	4150	15	-0.45	***
Angulo, Gil, and Gracia (2001)	alcohol	■	■	□	normal	■	15087	na		
Angulo *et al.* (2001)	alcohol	■	□	■	normal	■	15087	na		
Newman, Henchion, and Matthews (2001)	meat	■	■	□	IHS	□	7112-7332	15-68		
Newman, Henchion, and Matthews (2003)	prepared meals	■	■	□	IHS	■	7112-7332	51-57		
Newman *et al.* (2003)	prepared meals	□	■	■	IHS	■	7112-7332	51-57		
Mutlu and Gracia (2004)	food away from home	■	■	□	normal	■	12430	61		
Chaze (2005)	health services	■	■	□	Box-Cox	□	9295	23	-0.15	ns
Yen (2005)	cigarettes	■	□	□	tr. normal	□	4921-4672	75-78	-0.88, -0.59	***
Yen (2005)	cigarettes	■	■	□	normal	□	4921-4672	75-78	-0.71, -0.88	***
Zhang, Huang, and Lin (2008)	fresh organic product	■	□	□	log-normal	□	6916	42	0.90	***
Fuller, Beghin, and Rozelle (2007)	dairy products	■	■	□	normal	□	314	na	0.35, 0.56	***
Aristei and Pieroni (2008)	cigarettes	■	■	□	Box-Cox	■	27499	67	na	ns
Aristei, Perali, and Pieroni (2008)	alcohol	■	■	□	normal	□	134515	41	0.12	***
Aristei *et al.* (2008)	alcohol	■	□	□	normal	□	134515	41	0.14	***
Keelan, Henchion, and Newman (2009)	food service	■	■	□	Box-Cox	□	7526-7877	47-56		
Aristei and Pieroni (2009)	cigarettes	■	■	□	Box-Cox	□	47777	40	0.23	***
Okunade, Suraratdecha, and Benson (2010)	healthcare	■	■	□	normal	□	98632	28	-0.27	***
Brouhle and Khanna (2012)	eco-labeled market	■	■	□	normal	□	2483-2933	na		
Brouhle and Khanna (2012)	eco-labeled market	■	□	□	normal	□	2483-2933	na		
Crowley, Eakins, and Jordan (2012)	lottery	■	■	□	normal	■	3082	55		

`na` means that the information was not in the article. A black box in columns `h1-h2-h3` indicates that the relevant hurdle is taken into account, and in column `skedasticity` that heteroscedasticity is taken into account. (a) `IHS` stands for Inverse Hyperbolic Sine transformation and `tr.normal` for truncated normal distribution. (b) A starred value indicates a covariance estimate. (c) (*), (**) and (***) indicate that the correlation coefficient or the covariance estimate is significant at the 10%, 5% or 1% level, respectively ; `ns` that it is non significant at the 10% level.

The practical scope of multiple hurdle Tobit models is not restricted to empirical demand analysis but has been fruitfully used in other fields of economics. This includes labor economics (Elek, Köllö, Reizer, and Szabó 2011), contingent valuation studies (Saz-Salazar and Rausell-Köster 2008; Martínez-Espineira 2006), finance (Moffatt 2005), sport activities (Humphreys and Ruseski 2010), internet use (Wodjao 2020), gambling (Humphreys, Lee, and Soebbing 2009), production (Akpan, Nkanta, and Essien 2012; Mal, Anik, Bauer, and Schmitz 2012; Okello, Kirui, and Gitonga 2012; Teklewold, Dadi, Yami, and Dana 2006)

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.

Survival models are implemented in R with the **survival** package of Therneau (2013). 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 (2008). 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). More enhanced censored regression models (left and right censoring, random effect models) are available in the **censReg** package (Henningsen 2013). Some flavor of hurdle models have also been developed for count data and are implemented by the `hurdle` of the **pscl** package (Zeileis, Kleiber, and Jackman 2008).

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 a real-world example. 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:

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

where x_1, x_2, x_3 stand for column-vectors of explanatory variables (called covariates in the followings), β_1, β_2, β_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 ϵ_1, ϵ_2, ϵ_3 for normal random disturbances. Since variables y_1^* and y_3^* are never observed, contrary to y_2^*, the units of measurement of ϵ_1 and ϵ_3 are not identified. Hence, these disturbances are normalized by setting their variances equal to 1, i.e. by identifying them to standard normal random variables.

- 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^*. Cragg (1971) suggested two types of functional forms for this demand model, namely a log-normal functional form :

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

  and a truncated normal functional form where y_2^* is generated by a linear relationship $y_2^* = \beta_2^\top x_2 + \epsilon_2$ with ϵ_2 distributed according to a normal distribution left-truncated at

$\epsilon_2 = -\beta_2^\top x_2$. Nevertheless, to avoid a cumbersome analytic presentation of our models, in the following we only consider the log-normal model specification. More flexible generalizations of these functional forms will be discussed in section 2.3.

- 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 (theoretical or real-world knowledge) 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 or a truncated normal specification, 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 **mhurdle** package. To use a mnemonically rule, we number the models by 3 binary digits, each of which indicates if a censoring mechanism is or is not in effect, using figures `1` and `0`, respectively. For example, `011` indicates the model for which hurdle 1 (good selection mechanism) is not in effect while hurdles 2 (desired consumption mechanism) and 3 (purchasing mechanism) are.

Among these models, one is not concerned by censored data, namely model `000`. 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 `010`, that can be estimated also by the **survival** package of Therneau (2013) 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 `100` is a single-hurdle good selection model originated by Cragg (1971) by assuming independence between disturbances ϵ_1 and ϵ_2. The dependent version of this model

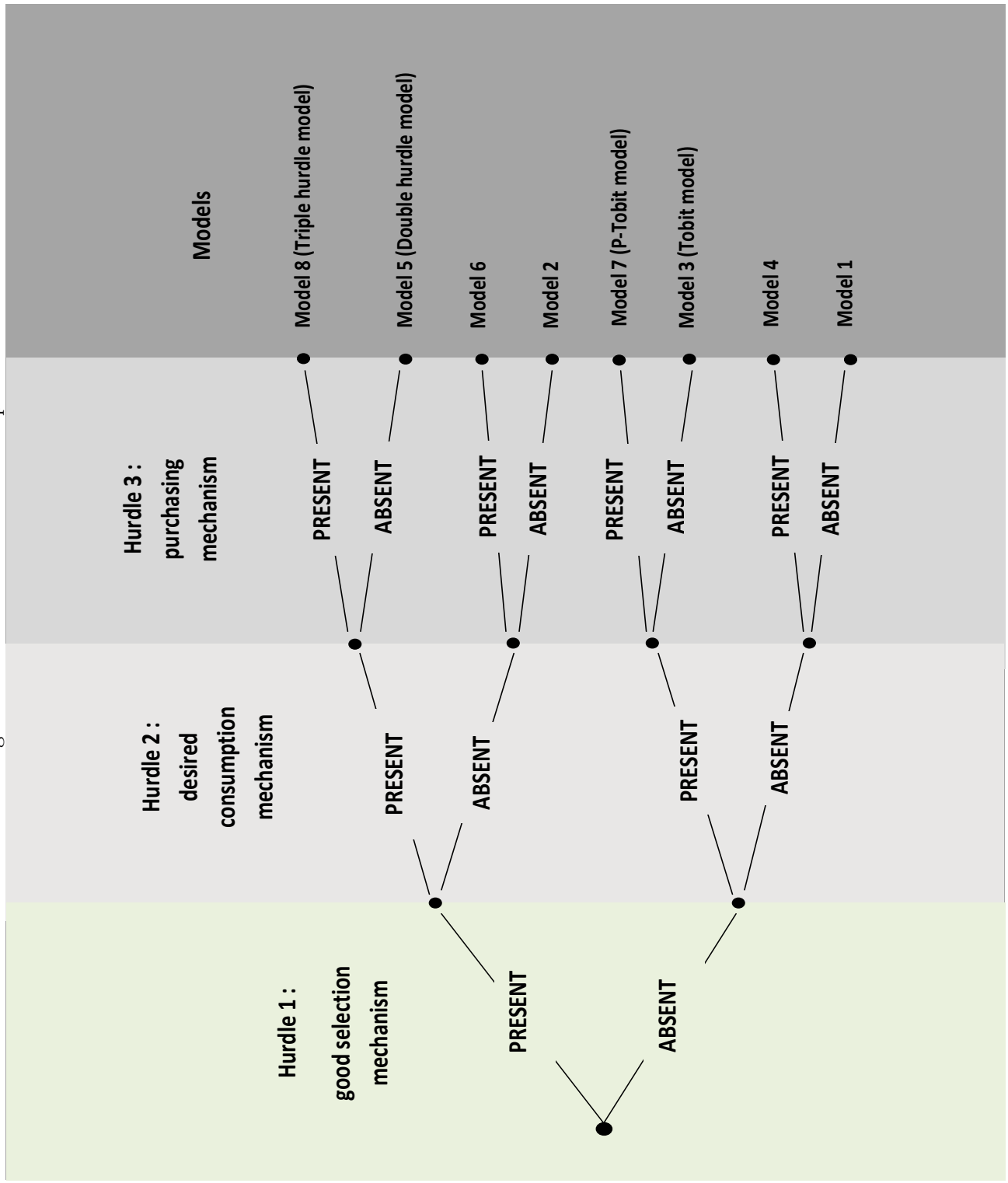

Figure 1: The full set of mhurdle special models.

may be viewed as a sample selection model in which only the desired consumption is observed, but it diffears from this model populazied by Heckman (1979) to illustrate linear regression model estimation given sample selection, in that desired consumption is generated by a demand model enforcing non negative values on latent variable y_2^*. Still, in many applications presented in Table 1, Heckman's sample selection model is used as a dependent single-hurdle good selection model in which good selection decision is assumed to dominate good consumption decision. From our point of view, this model is theoretically misspecified to analyse a latent dependent variable that could take negative values while assuming that hurdle 2 is not in effect.

The double-hurdle model `110` 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 `011` 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 `001` 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 `101` is a double-hurdle model combining good selection (hurdle 1) and purchasing (hurdle 3) mechanisms to explain censored samples. Model `111` 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.

- Models `111` and `101` are trivariate hurdle models as they involve disturbances ϵ_1, ϵ_2 and ϵ_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 ρ_{12}, ρ_{13}, ρ_{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.

- Models `011` and `001` are bivariate hurdle models as they involve disturbances ϵ_2 and ϵ_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 ρ.

- Models `110` and `100` are also bivariate hurdle models but they involve disturbances ϵ_1 and ϵ_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 `010` and `000` are univariate hurdle models involving only disturbance ϵ_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.

While the assumption of correlated disturbances is intended to account for the interdependence between latent variables y_1^*, y_2^* and y_3^* unexplained by covariates x_1, x_2 and x_3, a priori information (theoretical or real-world knowledge) 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 model relations. 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 the present version of **mhurdle**. To

identify the presence or absence of assumed dependence between couples of disturbances of a given **mhurdle** special model, we add to the 3 binary digit number of the model a letter `i` , if independence is assumed, and a letter `d` , otherwise. For exemple, `101dii` indicates a trivariate model `101` for which the couple of disturbances (ϵ_1, ϵ_2) are assumed to be correlated, while (ϵ_1, ϵ_3) and (ϵ_2, ϵ_3) are not.

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 `111`, using the change of variables:

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

  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 ρ_{12}, ρ_{13} and ρ_{23}.

- For trivariate hurdle model `101`, using the change of variables:

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

  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 ρ.

- The probability mass $P(y = 0)$ for bivariate hurdle models `011` and `110` and univariate hurdle model `010` can be derived from that of trivariate model `111` by eliminating hurdles 1, 3, 1 and 3, respectively. Likewise, this probability for bivariate hurdle models `100` and `001` can be derived from that of trivariate hurdle model `101` 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 2, 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 variable $y_2^* = P(I_3 = 1)y = \Phi_3 y$ 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 `111` 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:

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

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:

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

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

Table 2: Characteristics of mhurdle special models implemented in R

id	h_1	h_2	h_3	ρ_{12}	ρ_{13}	ρ_{23}	$P(y=0)$	$f_+(y)$	$\mathrm{E}(y \mid y>0)$
000	□	□	□				0	$\frac{1}{\sigma y}\phi\left(\frac{\ln y-\beta_2^\top x_2}{\sigma}\right)$	$\exp\left\{\beta_2^\top x_2+\frac{\sigma^2}{2}\right\}$
100i	■	□	□	□			$1-\Phi_1$	$\frac{1}{\sigma y}\phi\left(\frac{\ln y-\beta_2^\top x_2}{\sigma}\right)\Phi_1$	$\exp\left\{\beta_2^\top x_2+\frac{\sigma^2}{2}\right\}$
100d	■	□	□	■			$1-\Phi_1$	$\frac{1}{\sigma y}\phi\left(\frac{\ln y-\beta_2^\top x_2}{\sigma}\right)\Phi\left(\frac{\beta_1^\top x_1+\rho_{12}\frac{\ln y-\beta_2^\top x_2}{\sigma}}{\sqrt{1-\rho_{12}^2}}\right)$	$\exp\left\{\beta_2^\top x_2+\frac{\sigma^2}{2}\right\}\frac{\Phi\left(\beta_1^\top x_1+\sigma\rho_{12}\right)}{\Phi_1}$
010	□	■	□				$1-\Phi_2$	$\frac{1}{\sigma}\phi\left(\frac{y-\beta_2^\top x_2}{\sigma}\right)$	$\beta_2^\top x_2+\sigma\frac{\phi_2}{\Phi_2}$
001i	□	□	■			□	$1-\Phi_3$	$\frac{1}{\sigma y}\phi\left(\frac{\ln(\Phi_3 y)-\beta_2^\top x_2}{\sigma}\right)\Phi_3$	$\exp\left\{\beta_2^\top x_2+\frac{\sigma^2}{2}\right\}\frac{1}{\Phi_3}$
001d	□	□	■			■	$1-\Phi_3$	$\frac{1}{\sigma y}\phi\left(\frac{\ln(\Phi_3 y)-\beta_2^\top x_2}{\sigma}\right)\Phi\left(\frac{\beta_3^\top x_3+\rho_{23}\frac{\ln(\Phi_3 y)-\beta_2^\top x_2}{\sigma}}{\sqrt{1-\rho_{23}^2}}\right)$	$\exp\left\{\beta_2^\top x_2+\frac{\sigma^2}{2}\right\}\frac{\Phi\left(\beta_3^\top x_3+\sigma\rho_{23}\right)}{\Phi_3^2}$
110i	■	■	□	□			$1-\Phi_1\Phi_2$	$\frac{1}{\sigma}\phi\left(\frac{y-\beta_2^\top x_2}{\sigma}\right)\Phi_1$	$\beta_2^\top x_2+\sigma\frac{\phi_2}{\Phi_2}$
110d	■	■	□	■			$1-\Phi_{12}$	$\frac{1}{\sigma}\phi\left(\frac{y-\beta_2^\top x_2}{\sigma}\right)\Phi\left(\frac{\beta_1^\top x_1+\rho_{12}\frac{y-\beta_2^\top x_2}{\sigma}}{\sqrt{1-\rho_{12}^2}}\right)$	$\beta_2^\top x_2+\sigma\frac{\Psi_{2\mid 1}}{\Phi_{12}}$
101iii	■	□	■	□	□	□	$1-\Phi_1\Phi_3$	$\frac{1}{\sigma y}\phi\left(\frac{\ln(\Phi_3 y)-\beta_2^\top x_2}{\sigma}\right)\Phi_1\Phi_3$	$\exp\left\{\beta_2^\top x_2+\frac{\sigma^2}{2}\right\}\frac{1}{\Phi_3}$
101dii	■	□	■	■	□	□	$1-\Phi_1\Phi_3$	$\frac{1}{\sigma y}\phi\left(\frac{(\ln\Phi_3 y)-\beta_2^\top x_2}{\sigma}\right)\Phi\left(\frac{\beta_1^\top x_1+\rho_{12}\frac{(\ln y\Phi_3)-\beta_2^\top x_2}{\sigma}}{\sqrt{1-\rho_{12}^2}}\right)\Phi_3$	$\exp\left\{\beta_2^\top x_2+\frac{\sigma^2}{2}\right\}\frac{\Phi\left(\beta_1^\top x_1+\sigma\rho_{12}\right)}{\Phi_1\Phi_3}$
101iid	■	□	■	□	□	■	$1-\Phi_1\Phi_3$	$\frac{1}{\sigma y}\phi\left(\frac{\ln(\Phi_3 y)-\beta_2^\top x_2}{\sigma}\right)\Phi_1\Phi\left(\frac{\beta_3^\top x_3+\rho_{23}\frac{\ln(\Phi_3 y)-\beta_2^\top x_2}{\sigma}}{\sqrt{1-\rho_{23}^2}}\right)$	$\exp\left\{\beta_2^\top x_2+\frac{\sigma^2}{2}\right\}\frac{\Phi\left(\beta_3^\top x_3+\sigma\rho_{23}\right)}{\Phi_3^2}$
011i	□	■	■			□	$1-\Phi_2\Phi_3$	$\frac{1}{\sigma}\phi\left(\frac{\Phi_3 y-\beta_2^\top x_2}{\sigma}\right)\Phi_3^2$	$\frac{\beta_2^\top x_2}{\Phi_3}+\sigma\frac{\phi_2}{\Phi_2\Phi_3}$
011d	□	■	■			■	$1-\Phi_{23}$	$\frac{1}{\sigma}\phi\left(\frac{\Phi_3 y-\beta_2^\top x_2}{\sigma}\right)\Phi\left(\frac{\beta_3^\top x_3+\rho_{23}\frac{\Phi_3 y-\beta_2^\top x_2}{\sigma}}{\sqrt{1-\rho_{23}^2}}\right)\Phi_3$	$\frac{\beta_2^\top x_2}{\Phi_3}+\sigma\frac{\Psi_{2\mid 3}}{\Phi_{23}\Phi_3}$
111iii	■	■	■	□	□	□	$1-\Phi_1\Phi_2\Phi_3$	$\frac{1}{\sigma}\phi\left(\frac{\Phi_3 y-\beta_2^\top x_2}{\sigma}\right)\Phi_1\Phi_3^2$	$\frac{\beta_2^\top x_2}{\Phi_3}+\sigma\frac{\phi_2}{\Phi_2\Phi_3}$
111dii	■	■	■	■	□	□	$1-\Phi_{12}\Phi_3$	$\frac{1}{\sigma}\phi\left(\frac{\Phi_3 y-\beta_2^\top x_2}{\sigma}\right)\Phi\left(\frac{\beta_1^\top x_1+\rho_{12}\frac{\Phi_3 y-\beta_2^\top x_2}{\sigma}}{\sqrt{1-\rho_{12}^2}}\right)\Phi_3^2$	$\frac{\beta_2^\top x_2}{\Phi_3}+\sigma\frac{\Psi_{2\mid 1}}{\Phi_{12}\Phi_3}$
111iid	■	■	■	□	□	■	$1-\Phi_1\Phi_{23}$	$\frac{1}{\sigma}\phi\left(\frac{\Phi_3 y-\beta_2^\top x_2}{\sigma}\right)\Phi_1\Phi\left(\frac{\beta_3^\top x_3+\rho_{23}\frac{\Phi_3 y-\beta_2^\top x_2}{\sigma}}{\sqrt{1-\rho_{23}^2}}\right)\Phi_3$	$\frac{\beta_2^\top x_2}{\Phi_3}+\sigma\frac{\Psi_{2\mid 3}}{\Phi_{23}\Phi_3}$

A blackened square indicates which hurdle or correlation is assumed to be at work in the model ; an empty square indicates a hurdle which is not in effect or a zero correlation.

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

$$\begin{aligned} f_+(y) &= \frac{\Phi_3}{\sigma}\phi\left(\frac{\Phi_3 y - \beta_2^\top x_2}{\sigma}\right) \\ &\times \int_0^\infty \int_0^\infty \frac{1}{\sigma_{1|2}\sigma_{3|2}}\phi\left(\frac{y_1^* - \mu_{1|2}}{\sigma_{1|2}}, \frac{y_3^* - \mu_{3|2}}{\sigma_{3|2}}; \rho_{13|2}\right) dy_1^* dy_3^* \\ &= \frac{\Phi_3}{\sigma}\phi\left(\frac{\Phi_3 y - \beta_2^\top x_2}{\sigma}\right) \int_{-\frac{\mu_{1|2}}{\sigma_{1|2}}}^\infty \int_{-\frac{\mu_{3|2}}{\sigma_{3|2}}}^\infty \phi(z_1, z_3; \rho_{13|2}) dz_1 dz_3 \\ &= \frac{\Phi_3}{\sigma}\phi\left(\frac{\Phi_3 y - \beta_2^\top x_2}{\sigma}\right) \\ &\times \Phi\left(\frac{\beta_1^\top x_1 + \rho_{12}\frac{\Phi_3 y - \beta_2^\top x_2}{\sigma}}{\sqrt{1-\rho_{12}^2}}, \frac{\beta_3^\top x_3 + \rho_{23}\frac{\Phi_3 y - \beta_2^\top x_2}{\sigma}}{\sqrt{1-\rho_{23}^2}}; \rho_{13|2}\right). \end{aligned} \tag{16}$$

- For trivariate hurdle model `101`, we proceed as for hurdle model `111` 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:

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

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

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:

$$\begin{aligned} f_+(y) &= \frac{\phi\left(\frac{\ln(\Phi_3 y)-\beta_2^\top x_2}{\sigma}\right)}{\sigma y}\int_{-\frac{\mu_{1|2}}{\sigma_{1|2}}}^{\infty}\int_{-\frac{\mu_{3|2}}{\sigma_{3|2}}}^{\infty}\phi(z_1,z_3;\rho_{13|2})dz_1dz_3 \\ &= \frac{\phi\left(\frac{\ln(\Phi_3 y)-\beta_2^\top x_2}{\sigma}\right)}{\sigma y} \\ &\times \Phi\left(\frac{\beta_1^\top x_1+\rho_{12}\frac{\ln(\Phi_3 y)-\beta_2^\top x_2}{\sigma}}{\sqrt{1-\rho_{12}^2}},\frac{\beta_3^\top x_3+\rho_{23}\frac{\ln(\Phi_3 y)-\beta_2^\top x_2}{\sigma}}{\sqrt{1-\rho_{23}^2}};\rho_{13|2}\right). \end{aligned} \tag{20}$$

- The density function $f_+(y)$ for bivariate hurdle models `011` and `110` and univariate hurdle model `010` can be derived from that of trivariate model `111` by eliminating hurdles 1, 3, 1 and 3, respectively. Likewise, this density function for bivariate hurdle models `100` and `001` can be derived from that of trivariate hurdle model `101` by eliminating hurdles 1 and 3, respectively. Corresponding formulas for $f_+(y)$ for all this special cases implemented in R are presented in Table 2.

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 :

$$\ln L_i = \left\{\begin{array}{ll} \ln P(y_i=0) & \text{if} \quad y_i=0 \\ \ln f_+(y_i) & \text{if} \quad y_i>0 \end{array}\right. \tag{21}$$

and the log-likelihood for the entire random sample:

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

2.3. Heteroscedasticity and nonnormality

Contrary to the classical linear regression model which estimation is robust with respect to deviations from the assumptions of homoscedasticity and normality of disturbances, the maximum likelihood estimation of a Tobit model become inconsistent under heteroscedasticity and nonnormality of disturbances. Therefore, it is important to have methods allowing to test whether these assumptions are acceptable, on grounds of empirical evidence provided by a sample of observations, and to suggest how to respecify the model in case where a misspecification is brought out. In this section, we shall tackle this problem by using more flexible specifications of the desired consumption relation, where homoscedasticity and normality assumptions appear as special cases. Our choices are inspired by model specifications identified

in our survey of hurdle Tobit model applications reported in Table 1. Stated in general terms, our generalizations of models presented in section 2.1 are written as:

$$\begin{cases} y_1^* = \beta_1^\top x_1 + z_1 \\ T(y_2^*) = \beta_2^\top x_2 + \sigma(\beta_0^\top x_0) z_2 \\ y_3^* = \beta_3^\top x_3 + z_3 \end{cases} \tag{23}$$

where $T(y_2^*)$ denotes a monotonic transformation of desired consumption variable y_2^*, $\sigma(\beta_0^\top x_0)$ a positive monotonic transformation of a linear function of a vector of covariates x_0, selected to explain the heteroscedasticity of desired consumption, z_1, z_2 and z_3 standard normal random variables, with z_2 possibly truncated at the bounds of an interval $]B_1, B_2[$ ensuring that the range of values of $\beta_2^\top x_2 + \sigma(\beta_0^\top x_0) z_2$ corresponds to the domain of definition of the inverse transformation $y_2^* = T^{-1}(\beta_2^\top x_2 + \sigma(\beta_0^\top x_0) z_2)$.

With these assumptions in mind, we can derive the form of the observable dependent variable

$$y = \frac{T^{-1}(\beta_2^\top x_2 + \sigma(\beta_0^\top x_0) z_2)}{P(I_3 = 1)} I_1 I_2 I_3. \tag{24}$$

from the joint distribution of random variables z_1, z_2 and z_3, which is written as:

$$\frac{\phi\left(z_1, z_2, z_3; \rho_{12}, \rho_{13}, \rho_{23}\right)}{\Phi(B_2) - \Phi(B_1)}. \tag{25}$$

Using the change of variables:

$$\begin{cases} z_1 = y_1^* - \beta_1^\top x_1 \\ z_2 = \dfrac{T(y_2^*) - \beta_2^\top x_2}{\sigma(\beta_0^\top x_0)} \\ z_3 = y_3^* - \beta_3^\top x_3 \end{cases} \tag{26}$$

we derive the joint density function of latent variables y_1^*, y_2^* and y_3^*:

$$f(y_1^*, y_2^*, y_3^*) = \frac{T^{'}(y_2^*)}{\sigma(\beta_0^\top x_0)} \frac{\phi\left(y_1^* - \beta_1^\top x_1, \frac{T(y_2^*) - \beta_2^\top x_2}{\sigma(\beta_0^\top x_0)}, y_3^* - \beta_3^\top x_3\right)}{\Phi(B_2) - \Phi(B_1)}, \tag{27}$$

where $T^{'}(y_2^*)$ stands for the derivative of $T(y_2^*)$. Using this density function we compute the probability mass $P(y = 0) = 1 - P(y > 0)$ as follows:

$$\begin{aligned} P(y=0) &= 1 - \int_0^\infty \int_0^\infty \int_0^\infty \frac{\phi(y_1^*, y_2^*, y_3^*; \rho_{12}, \rho_{13}, \rho_{23})}{\Phi(B_2) - \Phi(B_1)} dy_1^* dy_2^* dy_3^* \\ &= 1 - \int_{-\beta_1^\top x_1}^{\infty} \int_{\frac{T(0) - \beta_2^\top x_2}{\sigma(\beta_0^\top x_0)}}^{B_2} \int_{-\beta_3^\top x_3}^{\infty} \frac{\phi(z_1, z_2, z_3; \rho_{12}, \rho_{13}, \rho_{23})}{\Phi(B_2) - \Phi(B_1)} dz_1 dz_2 dz_3 \\ &= 1 - \frac{\Phi(\beta_1^\top x_1, \frac{\beta_2^\top x_2 - T(0)}{\sigma(\beta_0^\top x_0)}, \beta_3^\top x_3; \rho_{12}, \rho_{13}, \rho_{23}) - \Phi(\beta_1^\top x_1, -B_2, \beta_3^\top x_3; \rho_{12}, \rho_{13}, \rho_{23})}{\Phi(B_2) - \Phi(B_1)}. \end{aligned} \tag{28}$$

To compute the density function $f_+(y)$, we first perform the change of variable $y_2^* = \Phi_3 y$ on the joint density function , leading to the joint density function of variables y_1^*, y and y_3^*:

$$f(y_1^*, y, y_3^*) = \frac{\Phi_3 T'(\Phi_3 y)}{\sigma(\beta_0^\top x_0)} \frac{\phi\left(y_1^* - \beta_1^\top x_1, \frac{T(\Phi_3 y) - \beta_2^\top x_2}{\sigma(\beta_0^\top x_0)}, y_3^* - \beta_3^\top x_3\right)}{\Phi(B_2) - \Phi(B_1)}. \tag{29}$$

Then we integrate this transformed density function, rewritten as the product of the marginal density function of y, namely:

$$\frac{\Phi_3 T'(\Phi_3 y)}{\sigma(\beta_0^\top x_0)} \frac{\phi\left(\frac{T(\Phi_3 y) - \beta_2^\top x_2}{\sigma(\beta_0^\top x_0)}\right)}{\Phi(B_2) - \Phi(B_1)}, \tag{30}$$

and of the joint density function of y_1^* and y_2^* conditioned with respect to y, stated by formula , over the positive values of latent variables y_1^* and y_3^*. By this way we get:

$$\begin{aligned} f_+(y) = & \frac{\Phi_3 T'(\Phi_3 y)}{\sigma(\beta_0^\top x_0)} \frac{\phi\left(\frac{T(\Phi_3 y) - \beta_2^\top x_2}{\sigma(\beta_0^\top x_0)}\right)}{\Phi(B_2) - \Phi(B_1)} \\ & \times \Phi\left(\frac{\beta_1^\top x_1 + \rho_{12} \frac{T(\Phi_3 y) - \beta_2^\top x_2}{\sigma(\beta_0^\top x_0)}}{\sqrt{1-\rho_{12}^2}}, \frac{\beta_3^\top x_3 + \rho_{23} \frac{T(\Phi_3 y) - \beta_2^\top x_2}{\sigma(\beta_0^\top x_0)}}{\sqrt{1-\rho_{23}^2}}; \rho_{13|2}\right). \end{aligned} \tag{31}$$

A natural choice for the heteroscedastic model $\sigma(\beta_0^\top x_0)$ is given by the exponential functional form $\sigma(\beta_0^\top x_0) = \exp\{\beta_0^\top x_0\}$, as exponential is a strictly increasing function mapping the set of real numbers $]-\infty, +\infty[$ onto the the set of positive real numbers $]0, +\infty[$, and leads to homoscedasticity when the linear function $\beta_0^\top x_0$ is a constant. This allows to test the assumption of homoscedasticity by inserting an intercept among parameters β_0 and assessing whether the other parameters are statistically significant or not.

The main drawbacks of this functional form lies in its implicit assumption of unboundedness of the variance of disturbance ϵ_2 with respect to covariates x_0. When a priori information suggests that this conditional variance remains bounded with respect to any possible values of these covariates, it is best to consider a functional model of the form $\sigma(\beta_0^\top x_0) = \exp\{\alpha F(\beta_0^\top x_0)\}$, where $F(.)$ is a distribution function mapping the set of real numbers $]-\infty, +\infty[$ onto the unit interval $]0, 1[$, and α a parameter playing the role of the intercept in the exponential functional form. Therefore, testing the assumption of homoscedasticity amounts to assessing the statistical non significance of parameter vector β_0 without an intercept, and setting the standard deviation of ϵ_2 equal to $\sigma = \exp\{\alpha F(0)\}$. For a practical application of this parametric model of heteroscedasticity, a menu of functional forms of $F(.)$ have been programmed in **mhurdle**, including the distribution functions of standard normal, logistic, Cauchy and Gompertz random variables.

As far as the choice of transformation $T(y_2^*)$ is concerned, two families of parametric functions have been considered, in order to generate departures from normality towards skewed and leptokurtic (more sharply peaked) distributions, of the kind encountered in collected economic data.

To generate skewed distributions of y_2^* we use the two parameters Box and Cox (1964) transformation, as suggested by Chaze (2005). This transformation is written as:

$$T(y_2^*) = \begin{cases} \frac{(y_2^*+\gamma)^\lambda - 1}{\lambda} & \text{if} \quad \lambda \neq 0 \\ \ln(y_2^* + \gamma) & \text{if} \quad \lambda = 0 \end{cases} \tag{32}$$

with λ a parameter characterizing the non linearity of the transformation. As shown in Figure 1, this transformation is linear for $\lambda = 1$, convex for $\lambda > 1$, and concave for $\lambda < 1$, with a ceiling asymptote at $-1/\lambda$ when $\lambda < 0$. γ is a location parameter, restricting the domain where the transformation holds, to the interval $]-\lambda, +\infty[$. Hence, the image of this interval by the Box-Cox transformation is given by:

$$T(]-\gamma, +\infty[) = \begin{cases}]-1/\lambda, +\infty[& \text{if} \quad \lambda > 0 \\ ]-\infty, +\infty[& \text{if} \quad \lambda = 0 \\ ]-\infty, -1/\lambda[& \text{if} \quad \lambda < 0 \end{cases} \tag{33}$$

As the inverse Box-Cox transformation is written as:

$$y_2^* = \begin{cases} \{\lambda(\beta_2^\top x_2 + \sigma(\beta_0^\top x_0) z_2) + 1\}^{1/\lambda} - \gamma & \text{if} \quad \lambda \neq 0 \\ \exp\{\beta_2^\top x_2 + \sigma(\beta_0^\top x_0) z_2\} - \gamma & \text{if} \quad \lambda = 0 \end{cases} \tag{34}$$

it turns out that z_2 is truncated at the bounds of the interval $]B_1, B_2[$, with:

$$B_1 = \begin{cases} -\frac{\beta_1^\top x_1 + (1/\lambda)}{\sigma(\beta_0^\top x_0)} & \text{if} \quad \lambda > 0 \\ -\infty & \text{if} \quad \lambda \leq 0 \end{cases} \tag{35}$$

and

$$B_2 = \begin{cases} +\infty & \text{if} \quad \lambda \geq 0 \\ -\frac{\beta_1^\top x_1 + (1/\lambda)}{\sigma(\beta_0^\top x_0)} & \text{if} \quad \lambda < 0 \end{cases} \tag{36}$$

Hence:

$$\Phi(B_2) - \Phi(B_1) = \begin{cases} \Phi(sign(\lambda) \frac{\beta_1^\top x_1 + (1/\lambda)}{\sigma(\beta_0^\top x_0)}) & \text{if} \quad \lambda \neq 0 \\ 1 & \text{if} \quad \lambda = 0 \end{cases} \tag{37}$$

with

$$sign(\lambda) = \begin{cases} + & \text{if} \quad \lambda > 0 \\ - & \text{if} \quad \lambda < 0 \end{cases} \tag{38}$$

Finally, the distribution of the observed variable y is obtained by inserting these results in formulas and , jointly with the derivative of the Box-Cox transformation, $T'(y_2^*) = (y_2^* + \gamma)^{\lambda - 1}$.

Note that when $\gamma = 0$, the Box-Cox transformation holds only for $y_2^* > 0$, corresponding to the restriction enforced to the desired consumption relation, when hurdle 2 is not in effect, by means of Cragg's log-normal or truncated normal functional forms. These functional forms arise as special cases of the Box-Cox transformation, when $\lambda = 0$ and $\lambda = 1$, respectively.

Thus, testing the statistical significance of parameter γ, against the alternative $\gamma > 0$, amounts to testing the assumption that hurdle 2 is not in effect within the framework of a nested model. Conversely, when $\gamma < 0$ the Box-Cox transformation holds for $y_2^* > -\gamma$ meaning that $-\gamma$ stands for a "committed" consumption of a basic necessity, while $\gamma > 0$ typify a luxury good whose consumption occurs only above a given income threshold.

The way the Box-Cox transformation generates the profile of the density function of y_2^* from that of z_2 can be illustrated by considering the special case of $\beta_2^\top x_2 = 0$ and $\sigma(\beta_0^\top x_0) = 1$. In this special case, the marginal density function of y_2^* is defined for $y_2^* \geq 0$,and takes the form of the product of the density function of z_2, written as a function of y_2^*:

$$\frac{\phi(\frac{y_2^{*\lambda}-1}{\lambda})}{\Phi(sign(\lambda)\frac{1}{\lambda})} \tag{39}$$

times the value of the Jacobian of the Box-Cox transformation:

$$J = y_2^{*(\lambda-1)}. \tag{40}$$

This Jacobian magnifies or reduces the value of the density function of z_2 depending on whether $J > 1$ or $J < 1$, respectively. According to the values of λ, three cases must be considered.

- When $\lambda = 1$, $J = 1$ whatever the value of y_2^*. Hence, the shape of the density function of y_2^* is the same as that of z_2, namely a normal random variable left truncated at $y_2^* = 0$ and with a mode at $y_2^* = 1$.

- When $\lambda > 1$, J is an unbounded increasing function of y_2^*, taking value 0 at $y_2^* = 0$ and 1 at $y_2^* = 1$, while the density function of z_2 at $y_2^* = 0$ holds finite. Hence, the density function of y_2^* starts from 0 and shifts the mode beyond $y_2^* = 1$. Furthermore, the Jacobian being a linear function of y_2^* when $\lambda = 2$, a concave function when $1 < \lambda < 2$, and a convex function when $\lambda > 2$, the concentration of the density function around its mode increases with the value of λ, until the collapse of the entire probability mass at $y_2^* = 1$, when $\lambda \to \infty$.

- When $\lambda < 1$, J is a decreasing function of y_2^*, from $+\infty$ at $y_2^* = 0$, to 0 when $y_2^* \to \infty$, taking value 1 at $y_2^* = 1$, while the density function of z_2 at $y_2^* = 0$ holds finite or takes value 0 depending on whether $0 < \lambda < 1$ or $\lambda \leq 0$. Hence, contrary to the previous case, the mode of the density function of y_2^* shifts short of $y_2^* = 1$, while, at the origin, the density function of y_2^* tends to $+\infty$ and turns out to be an undetermined product depending on whether $0 < \lambda < 1$ or $\lambda \leq 0$. The application of Hospital's rule allows to remove the indetermination of this limit, which turns out to be equal to 0. Note that the same tendency towards a concentration of the density function of y_2^* around its degenerated limit at $y_2^* = 1$ appears when $\lambda \to -\infty$. Note also that the J-shaped profile of this density function, when $0 < \lambda < 1$, has a mode preceded by an antimode that can coalesce into a point of inflexion.

All these profiles are illustrated in Figures 2.

To generate leptokurtic distributions of y_2^* we use the transformation popularized by Johnson (1949), based on the inverse hyperbolic sine function, namely:

$$T(y_2^*) = \frac{1}{\theta}\sinh^{-1}(\theta y_2^*) = \frac{1}{\theta}\ln\{\theta y_2^* + \sqrt{(\theta y_2^*)^2 + 1}\} \tag{41}$$

with θ a parameter characterizing the non linearity of the transformation. Indeed, as illustrated by Figure 3, while, for $\theta > 0$ and finite, the general shape of this transformation is that of an odd [3] increasing function, from $-\infty$ to $+\infty$, with an inflexion point at $y_2^* = 0$, concave for $y_2^* > 0$ and convex for $y_2^* < 0$, the transformation becomes linear, when $\theta \to 0$, and constant (equal to 0), when $\theta \to +\infty$ [4], by progressively loosing its sigmoidal profile. As these profiles are invariant with respect to a change of sign of θ, this parameter can be assumed to be non negative, without loss of generality.

To determine the distribution of the observed variable y, we need to insert in formulas and the expressions of $\Phi(B_2) - \Phi(B_1)$ and $T^{'}(y_2^*)$ for Johnson's transformation. We first note that the inverse function of this transformation, which is written:

$$\begin{aligned} y_2^* &= \frac{\sinh(\theta(\beta_2^\top x_2 + \sigma(\beta_0^\top x_0)z_2))}{\theta} \\ &= \frac{\exp\{\theta(\beta_2^\top x_2 + \sigma(\beta_0^\top x_0)z_2)\} - \exp\{-\theta(\beta_2^\top x_2 + \sigma(\beta_0^\top x_0)z_2)\}}{2\theta} \end{aligned} \tag{42}$$

does not require a truncation of the support of z_2, implying that $B_1 = -\infty$, $B_2 = +\infty$, and hence $\Phi(B_2) - \Phi(B_1) = 1$.

Secondly, the derivative of Johnson's transformation, namely:

$$T^{'}(y_2^*) = \frac{1}{\sqrt{(\theta y_2^*)^2 + 1}} \tag{43}$$

is a bell-shaped positive pair function [5] taking constant values 1 and 0 when $\theta \to 0$ and $\theta \to +\infty$, respectively.

To analyze the profile of the density function of y_2^* generated by this transformation, we consider, as we did before for the Box-Cox transformation, the special case of $\beta_2^\top x_2 = 0$ and $\sigma(\beta_0^\top x_0) = 1$. In this special case, the marginal density function of y_2^* takes the form of the product of the density function of $\phi(T(y_2^*))$ times the value of the Jacobian $T^{'}(y_2^*)$. As a function of θ, this formula generates a family of bell-shaped density functions which kurticity increases with the value of θ, from metakurticity (that of the normal distribution) when $\theta = 0$, to extreme leptokurticity resulting in a concentration of all the density function at point $y_2^* = 0$, when $\theta \to +\infty$, as shown in Figure 4. Therefore, the inverse hyperbolic sine transformation must be used only when hurdle 2 is in effect.

3. Model estimation, evaluation and selection

[3] A function $f(x)$ is said odd if $f(-x) = -f(x)$ whatever x, implying that $f(0) = 0$.
[4] These limits can be obtained easily by using Hospital's rule.
[5] A function $f(x)$ is said pair if $f(-x) = f(x)$ whatever x.

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 [6].

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

$$\hat{\theta} \overset{A}{\sim} N(\theta, \frac{1}{n} I_A(\theta)^{-1}), \tag{44}$$

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 θ, 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:

$$\hat{\gamma} \overset{A}{\sim} h(\theta) + \frac{\partial h}{\partial\theta^\top}(\hat{\theta} - \theta) \overset{A}{\sim} N\left(\gamma, \frac{1}{n} \frac{\partial h}{\partial\theta^\top} I_A(\theta)^{-1} \frac{\partial h^\top}{\partial\theta}\right). \tag{45}$$

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

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

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.

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:

$$y = E(y) + u, \tag{46}$$

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

$$E(y) = 0 \times P(y = 0) + \int_0^\infty y f_+(y) dy = \int_0^\infty y f_+(y) dy. \tag{47}$$

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 `111`, using the density function (13) and the change of variables:

$$\begin{cases} z_1 = y_1^* - \beta_1^\top x_1 \\ z_2 = \dfrac{\Phi_3 y - \beta_2^\top x_2}{\sigma} \\ z_3 = y_3^* - \beta_3^\top x_3 \end{cases} \tag{48}$$

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

$$\begin{aligned} E(y) &= \int_{-\beta_1^\top x_1}^\infty \int_{-\frac{\beta_2^\top x_2}{\sigma}}^\infty \int_{-\beta_3^\top x_3}^\infty \frac{\beta_2^\top x_2 + \sigma z_2}{\Phi_3} \phi\left(z_1, z_2, z_3; \rho_{12}, \rho_{13}, \rho_{23}\right) dz_1 dz_2 dz_3 \\ &= \frac{\beta_2^\top x_2}{\Phi_3} \Phi\left(\beta_1^\top x_1, \frac{\beta_2^\top x_2}{\sigma}, \beta_3^\top x_3; \rho_{12}, \rho_{13}, \rho_{23}\right) \\ &+ \frac{\sigma}{\Phi_3} \int_{-\beta_1^\top x_1}^\infty \int_{-\frac{\beta_2^\top x_2}{\sigma}}^\infty \int_{-\beta_3^\top x_3}^\infty z_2 \phi\left(z_1, z_2, z_3; \rho_{12}, \rho_{13}, \rho_{23}\right) dz_1 dz_2 dz_3. \end{aligned} \tag{49}$$

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:

$$\frac{\phi\left(\frac{z_2-\mu_{2|1,3}}{\sigma_{2|1,3}}\right)}{\sigma_{2|1,3}}, \tag{50}$$

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:

$$\begin{aligned}
&\int_{-\beta_1^\top x_1}^{\infty}\int_{-\frac{\beta_2^\top x_2}{\sigma}}^{\infty}\int_{-\beta_3^\top x_3}^{\infty} z_2\phi\left(z_1, z_2, z_3; \rho_{12}, \rho_{13}, \rho_{23}\right) dz_1 dz_2 dz_3 \\
&= \int_{-\beta_1^\top x_1}^{\infty}\int_{-\beta_3^\top x_3}^{\infty}\left[\int_{-\frac{\beta_2^\top x_2}{\sigma}}^{\infty} z_2\phi\left(\frac{z_2-\mu_{2|1,3}}{\sigma_{2|1,3}}\right)\frac{dz_2}{\sigma_{2|1,3}}\right]\phi\left(z_1, z_2; \rho_{13}\right) dz_1 dz_3.
\end{aligned} \tag{51}$$

By performing the change of variable:

$$z = \frac{z_2-\mu_{2|1,3}}{\sigma_{2|1,3}}, \tag{52}$$

the integral with respect to z_2 simplifies to:

$$\mu_{2|1,3}\Phi\left(\frac{\frac{\beta_2^\top x_2}{\sigma}+\mu_{2|1,3}}{\sigma_{2|1,3}}\right) + \sigma_{2|1,3}\phi\left(\frac{\frac{\beta_2^\top x_2}{\sigma}+\mu_{2|1,3}}{\sigma_{2|1,3}}\right). \tag{53}$$

By inserting this result in formula (51), we finally obtain:

$$\begin{aligned}
E(y) = &\frac{\beta_2^\top x_2}{\Phi_3}\Phi\left(\beta_1^\top x_1, \frac{\beta_2^\top x_2}{\sigma}, \beta_3^\top x_3; \rho_{12}, \rho_{13}, \rho_{23}\right) \\
&+ \frac{\sigma}{\Phi_3}\int_{-\beta_1^\top x_1}^{\infty}\int_{-\beta_3^\top x_3}^{\infty}\left[\left(\varrho_1 z_1 + \varrho_3 z_3\right)\Phi\left(\frac{\frac{\beta_2^\top x_2}{\sigma}+\varrho_1 z_1+\varrho_3 z_3}{\sigma_{2|1,3}}\right)\right. \\
&\left.+ \sigma_{2|1,3}\phi\left(\frac{\frac{\beta_2^\top x_2}{\sigma}+\varrho_1 z_1+\varrho_3 z_3}{\sigma_{2|1,3}}\right)\right]\phi\left(z_1, z_3; \rho_{13}\right) dz_1 dz_3.
\end{aligned} \tag{54}$$

- For trivariate hurdle model `101`, we proceed as for hurdle model `111` 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:

$$\begin{cases} z_1 = y_1^* - \beta_1^\top x_1 \\ z_2 = \dfrac{\ln(\Phi_3 y) - \beta_2^\top x_2}{\sigma} \\ z_3 = y_3^* - \beta_3^\top x_3 \end{cases} \tag{55}$$

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

$$\begin{aligned} E(y) = &\int_{-\beta_1^\top x_1}^{\infty}\int_{-\infty}^{\infty}\int_{-\beta_3^\top x_3}^{\infty} \frac{\exp\{\beta_2^\top x_2 + \sigma z_2\}}{\Phi_3} \\ &\times\phi(z_1, z_2, z_3; \rho_{12}, \rho_{13}, \rho_{23})\, dz_1 dz_2 dz_3 = \frac{\exp\{\beta_2^\top x_2\}}{\Phi_3} \\ &\times\int_{-\beta_1^\top x_1}^{\infty}\int_{-\beta_3^\top x_3}^{\infty}\left[\int_{-\infty}^{\infty}\exp\{\sigma z_2\}\phi\left(\frac{z_2-\mu_{2|1,3}}{\sigma_{2|1,3}}\right)\frac{dz_2}{\sigma_{2|1,3}}\right]\phi(z_1, z_3; \rho_{13})\, dz_1 dz_3 \end{aligned} \tag{56}$$

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 (52), the integral with respect to z_2 simplifies to:

$$\int_{-\infty}^{\infty}\exp\{\sigma(\mu_{2|1,3} + \sigma_{2|1,3}z)\}\phi(z)dz = \exp\left\{\sigma\mu_{2|1,3} + \frac{\sigma^2\sigma_{2|1,3}^2}{2}\right\}. \tag{57}$$

By inserting this result in formula (56), we finally obtain:

$$\begin{aligned} E(y) = &\frac{\exp\left\{\beta_2^\top x_2 + \frac{\sigma^2\sigma_{2|1,3}^2}{2}\right\}}{\Phi_3} \\ &\times\int_{-\beta_1^\top x_1}^{\infty}\int_{-\beta_3^\top x_3}^{\infty}\exp\{\sigma(\varrho_1 z_1 + \varrho_3 z_3)\}\phi(z_1, z_3; \rho_{13})\, dz_1 dz_3. \end{aligned} \tag{58}$$

- $E(y)$ for bivariate hurdle models `011` and `110` and univariate hurdle model `010` can be derived from that of trivariate model `111` by eliminating hurdles 1, 3, 1 and 3, respectively. Likewise, the expectation of y for bivariate hurdle models `100` and `001` can be derived from that of trivariate hurdle model `101` 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 2, 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 2 are obtained by using closed forms of the following integrals :

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

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

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

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

$$R^2 = 1 - \frac{RSS}{TSS}, \tag{59}$$

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 [7]). 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 computing this index formula:

$$\rho^2 = 1 - \frac{\ln L(\hat{\theta})}{\ln L(\hat{\alpha})} = \frac{\ln L(\hat{\alpha}) - \ln L(\hat{\theta})}{\ln L(\hat{\alpha})}, \tag{60}$$

using 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 ρ^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.

[7] 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 $\hat{y}_0$.

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 ρ^2 fit measure with a maximum likelihood estimate. Consequently, a penalty that increases with the number of model parameters should be added to the R^2 and ρ^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

$$\bar{R}^2 = 1 - \frac{n - K_0}{n - K}\frac{RSS}{TSS}, \tag{61}$$

where K and K_0 stand for the number of parameters of the multiple hurdle model with covariates and without covariates, respectively [8]. 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 ρ^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

$$\bar{\rho}^2 = 1 - \frac{\ln L(\hat{\theta}) - K}{\ln L(\hat{\alpha}) - K_0} \tag{62}$$

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:

$$KLIC = E\left[\ln\left(\frac{h(y|x)}{f(y|x;\theta_*)}\right)\right] = \int \ln\left(\frac{h(y|x)}{f(y|x;\theta_*)}\right) dH(y,x), \tag{63}$$

where $H(y,x)$ denotes the distribution function of the true joint distribution of (y,x) and θ_* 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

[8] 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.

collection of competing specifications is the one that minimises the $KLIC$ criterion or, equivalently, the specification for which the quantity:

$$E[\ln f(y|x;\theta_*)] = \int \ln f(y|x;\theta_*) dH(y,x) \tag{64}$$

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

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

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

$$LR(\hat{\theta},\hat{\pi}) = \sum_{i=1}^{n} \ln \frac{f(y_i|x_i;\hat{\theta})}{g(y_i|x_i;\hat{\pi})} \tag{65}$$

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 θ_* and π_*, 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_π defined by the conditional density function $g(y|x;\pi)$ is said to be nested in parametric model F_θ with conditional density function $f(y|x;\theta)$, if and only if any conditional density function of G_π is equal to a conditional density function of F_θ 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_π as a particular case of model F_θ. 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_θ 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_θ is correctly specified, the asymptotic distribution of the LR statistic is, as expected, a χ^2 random variable with $K - L$ degrees of freedom.

Two parametric models F_θ and G_π 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_θ is equal to a conditional distribution function of G_π 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:

$$T_{LR} = \frac{LR(\hat{\theta},\hat{\pi})}{\sqrt{n}\hat{\omega}}, \tag{66}$$

where $\hat{\omega}^2$ denotes the following strongly consistent estimator for ω^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_θ or G_π 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_θ and G_π 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_θ is equal to the conditional distribution function of G_π 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 ω^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 ω^2, $\hat{\omega}^2$, and proves that:

- under H_0^ω, 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^ω, 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^ω is rejected in favour of H_A^ω 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^ω 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^ω 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**, four kinds of covariates should be specified: those of

- the good selection equation (hurdle 1) denoted x_1,
- the desired consumption equation (hurdle 2), denoted x_2,
- the purchasing equation (hurdle 3), denoted x_3,
- the variance equation, denoted x_4.

To define a model with several 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 and optimisation

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 2008).

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 `100`, `010` and `001`, the probability of observing a censored purchase is not that of a simple probit model (see Table 2).

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.

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 (2012) provides a unified framework. With a unified 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 (code by Alan Genz. R code by Brenton Kenkel and based on Adelchi Azzalini's 'mnormt' package. 2012) which provides a vectorised (and therefore fast and 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 one surveys conducted by the Bureau of Labour Statistics of the U.S. Department of Labour, called the "Interview Survey". Data from 25813 households on all expenditures are collected, on a quarterly basis. The micro-data files are publicly available on the website of the Bureau of Labour Statistics, and may be downloaded and used without permission. We use a small subset of 1000 randomly selected house-

holds. The total expenditure is divided in 14 main chapters (food, alcool, housing, apparel, transport, health, entertainment, personal care, reading, eduction, tobacco, miscellanous expenditures, cash contributions and insurance) and we also report expenditures on fees and admissions (called `shows`) which is part of the entertainment chapter and trip expenditures (called `vacations` on the data set).

All the expenditures are in tousands of USD, measured on an annual basis and are divided by the number of consumption units (obtained by counting for one the first adult of the household, 0.7 the subsequent adults and 0.5 every other person aged under 18).

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

```
  month size  cu   income       linc      linc2 smsa    sex  race hispanic educ
1     5    1 1.0 13.37900 -1.2753120 1.62642081  yes female white      yes    7
2     4    4 2.5 72.40440  0.4132687 0.17079103  yes female white       no   13
3     8    2 1.7 55.80412  0.1528493 0.02336291  yes female white       no   12
  age age2 car     food    alcool  housing   apparel transport   health
1  37 1369   0 1.733333 0.0000000 2.548000 0.0000000  0.000000 0.000000
2 -11  121   1 5.216000 0.0000000 8.857600 0.2080000  2.667200 2.600000
3  31  961   3 3.568628 0.5835294 6.316863 0.5976471  2.286275 2.257882
  entertainment  perscare   reading education tobacco    miscexp  cashcont
1         0.000 0.0000000 0.0000000         0       0 0.00000000 0.0000000
2         1.344 0.0640000 0.0000000         0       0 0.00000000 0.0000000
3         0.680 0.4941176 0.2988235         0       0 0.09411765 0.8117647
  insurance     shows foodaway vacations
1    0.0000 0.0000000 0.000000         0
2    8.8912 0.0800000 3.136000         0
3    0.0000 0.1176471 1.223529         0
```

```
R> mean(Interview$shows == 0)
```

```
[1] 0.694
```

```
R> max(Interview$shows)
```

```
[1] 6.844706
```

The covariates are :

income the anual net income by consumption unit,

smsa does the household live in a SMSA (`yes` or `no`),

age the age of the reference person of the household,

educ the number of year of education of the reference person of the household,

sex the sex of the reference person of the household (`male` and `female`),

size the number of persons in the household,

month the month of the interview (between 1 and 12),

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 between two and four 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, in the third part, the purchasing equation covariates and in the fourth part, the covariates of the variance equation.

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 `"ln"` (the default), two-parameters log-normal `"ln2"`, normal `"n"`, truncated normal `"tn"`, Box-Cox `"bc"`, two-parameters Box-Cox `"bc2"` or inverse Hyperbolic Sine `"ihs"`,

scaled: if `TRUE`, the dependent variable is divided by the geometric mean of the positive values,

corr: this boolean argument indicates whether the disturbance of the different equations are correlated, the default value is `FALSE`,

robust: if `TRUE`, transformations of some parameters are used, so that they lie in the required range (positive values for the standard deviation and for the position parameter, between -1 and +1 for the coefficients of correlation),

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

One equation models

We start with models that only contain the consumption equation. In this case, the only source of null consumption is the lack of ressources. In this case, the distribution of y^* must admit negative values, this is the case for the two-parameters box-cox normal distribution, the log-normal distribution with a position parameter and the normal distribution (this later case corresponds to the standard tobit model). We use the expenditures on vacation as an exemple and the results are presented in table 3, using the **texreg** library.

	normal tobit	log-normal tobit	box-cox tobit
h2.(Intercept)	-12.99^{***}	-2.08	-2.30^{**}
	(1.54)	(2.44)	(0.72)
h2.I(car > 0)TRUE	1.25	0.24	0.26
	(1.04)	(0.23)	(0.18)
h2.size	1.05^{**}	0.20	0.21^{**}
	(0.40)	(0.16)	(0.08)
h2.linc	5.35^{***}	0.91	0.97^{***}
	(0.74)	(0.69)	(0.21)
h2.linc2	0.23	-0.10	-0.11
	(0.68)	(0.16)	(0.15)
h2.age	-0.01	-0.00	-0.00
	(0.03)	(0.01)	(0.01)
h2.age2	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)
sd.sd	8.65^{***}	1.50	1.60^{***}
	(0.20)	(1.12)	(0.29)
tr		-0.03	
		(0.32)	
pos		1.05	0.98^{**}
		(0.92)	(0.31)
Num. obs.	1000	1000	1000
Log Likelihood	-758.51	-658.61	-658.62
R^2	0.03	1.00	0.02
McFadden R^2			
$R^2(y = 0)$			
$R^2(y > 0)$			

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3: Estimation of one-equation models

```
R> Sn <- mhurdle(vacations ~ 0 | I(car > 0) + size + linc + linc2 + age + age2,
+                Interview, dist = "n", h2 = TRUE, corr = FALSE,
+                method = "bhhh", print.level = 0)
R> Sbc <- update(Sn, dist = "bc")
R> Sln <- update(Sn, dist = "ln")
R> library(texreg)

R> texreg(list(Sn, Sbc, Sln),
+         custom.model.names = c("normal tobit", "log-normal tobit", "box-cox tobit"),
+         caption = "Estimation of one-equation models",
+         label = "tab:oneq", pos = "ht")
```

Simple selection models

The original independent single hurdle models proposed by Cragg (1971) required the distinct

estimation of two models : a probit to explain the null *vs* the positive observations and an estimation on the truncated sample in order to explain the level of consumption of those who consume the good. This latter estimation can be realised with the hypothesis that the distribution of the errors are truncated normal, log-normal or box-cox normal. We use the expenditure on food-away as an exemple and the results are presented in table 4.

```
R> Stn <- mhurdle(foodaway ~ size + smsa + age + age2 | linc + linc2, Interview,
+                 dist = "n", h2 = FALSE, corr = FALSE, method = "bhhh", print.level = 0)
R> Sbc <- update(Stn, dist = "bc")
R> Sln <- update(Stn, dist = "ln")
```

The dependent model is easily obtained by setting the `corr` argument to `TRUE`. For the log-normal model, we obtain :

```
R> Slnd <- update(Sln, corr = TRUE)
R> coef(summary(Slnd), "corr")

         Estimate Std. Error   t-value Pr(>|t|)
corr12 -0.8735068 0.03420676 -25.53609        0
```

The coefficient of correlation between the selection and the consumption is in this case highly significant.

```
R> texreg(list(Stn, Sln, Sbc),
+        custom.model.names = c("truncated-normal", "log-normal", "box-cox"),
+        caption = "Estimation of single hurdle selection models",
+        label = "tab:sep", pos = "ht")
```

P-tobit model

Apparel is a good candidate to illustrate the estimation of a P-tobit model as the infrequency of purchase is clearly the only source of null expenditure in this case. To explain the probability of purchasing during the quarter of the survey, we use the month (as expenditures in apparels can be concentrated on certain periods during the year because of the sales) and the fact that the household lives in a SMSA. No corner solution is allowed as the good is consumed even for households with low ressources, so we choose a log-normal distribution.

```
R> ptobit <- mhurdle(apparel ~ 0 | linc + linc2 | factor(month) + smsa,
+                    Interview, corr = TRUE, dist = "ln", h2 = FALSE,
+                    method = "bhhh")
```

To illustrate the use of `mhurdle`, we estimate several models explaining the expenditure on fees and admission. The expenditure is supposed to depends on income and its square, age and its square, education, the size of the household and the fact that the household lives in a smsa. We first estimate a triple hurdle model, using `educ` and `size` as covariates for the selection equation and `age` and `smsa` as covariates for the purchasing equation. We use a two-parameter log-normal distribution and a general structure of correlation between the errors of the three equations is estimated.

	truncated-normal	log-normal	box-cox
h1.(Intercept)	0.64^{***}	0.64^{***}	0.64^{***}
	(0.16)	(0.16)	(0.16)
h1.size	0.02	0.02	0.02
	(0.04)	(0.04)	(0.04)
h1.smsayes	0.35^{**}	0.35^{**}	0.35^{**}
	(0.13)	(0.13)	(0.13)
h1.age	-0.01^{*}	-0.01^{*}	-0.01^{*}
	(0.00)	(0.00)	(0.00)
h1.age2	−0.00	−0.00	−0.00
	(0.00)	(0.00)	(0.00)
h2.(Intercept)	-19.12^{*}	0.12^{***}	0.16^{***}
	(9.33)	(0.04)	(0.04)
h2.linc	20.30^{**}	0.70^{***}	0.72^{***}
	(7.88)	(0.05)	(0.04)
h2.linc2	−4.99	−0.01	0.01
	(2.59)	(0.05)	(0.05)
sd.sd	5.86^{***}	0.86^{***}	0.86^{***}
	(1.23)	(0.02)	(0.02)
tr			0.08^{**}
			(0.03)
Num. obs.	1000	1000	1000
Log Likelihood	-1550.64	-1505.17	-1500.78
R^2	-0.34	-0.69	1.00
McFadden R^2			
$R^2(y = 0)$			
$R^2(y > 0)$			

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4: Estimation of single hurdle selection models

```
R> H3D <- mhurdle(shows ~ educ + size | linc + linc2 | age + age2 + smsa,
+                 Interview, dist = "ln", h2 = TRUE, corr = TRUE, method = "bhhh")
```

The independent version of the same model is obtained by setting the `corr` argument to `FALSE`.

```
R> H3I <- update(H3D, corr = FALSE)
```

A three equation - two hurdles model is obtained by making the second hurdle ineffective,

```
R> H2D <- update(H3D, h2 = FALSE)
```

We then estimate a double-hurdle selection model by using a two-part formula with `educ`, `size`, `age` and `smsa` in the first part which describe the selection process:

```
R> S2D <- mhurdle(shows ~ educ + size + age + age2 + smsa | linc + linc2,
+                 Interview, dist = "ln", h2 = TRUE, corr = TRUE, method = "bhhh")
```

Finaly, we estimate a double-hurdle p-tobit model, all the covariates which were in the first part of the formula being moved to the third part:

```
R> P2D <- mhurdle(shows ~ 0 | linc + linc2 | educ + size + age + age2 + smsa,
+                 Interview, dist = "ln", h2 = TRUE, corr = TRUE, method = "bhhh")
```

5.2. Methods

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

```
R> summary(H3D)

Call:
mhurdle(formula = shows ~ educ + size | linc + linc2 | age +
    age2 + smsa, data = Interview, dist = "ln", h2 = TRUE, corr = TRUE,
    method = "bhhh")

Frequency of 0:  0.694

BHHH maximisation method
45 iterations, 0h:0m:1s
g'(-H)^-1g = 0.000121


Coefficients :
                  Estimate  Std. Error t-value  Pr(>|t|)
h1.(Intercept) -1.51710780  0.61786864 -2.4554  0.014073 *
```

```
h1.educ          0.12270773  0.04100103  2.9928  0.002764 **
h1.size          0.21831543  0.09235265  2.3639  0.018082 *
h2.(Intercept)   0.18925615  0.23700720  0.7985  0.424566
h2.linc          0.70949799  0.08172746  8.6813 < 2.2e-16 ***
h2.linc2         0.09630606  0.05295770  1.8185  0.068981 .
h3.(Intercept)   0.16209936  0.29148871  0.5561  0.578137
h3.age          -0.04029869  0.01914732 -2.1047  0.035320 *
h3.age2          0.00088737  0.00062155  1.4277  0.153388
h3.smsayes       0.66856877  0.27613351  2.4212  0.015470 *
sd.sd            0.91141074  0.15295485  5.9587 2.543e-09 ***
corr12          -0.59929329  0.22205211 -2.6989  0.006957 **
corr13          -0.75827415  0.39347672 -1.9271  0.053966 .
corr23           0.30786399  0.32667919  0.9424  0.345986
pos              0.78683766  0.18780995  4.1895 2.795e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -993.03 on 15 Df

R^2 :
 Coefficient of determination : 0.12309
 Likelihood ratio index       : 0.10135
```

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(H3D, "h2")

(Intercept)        linc       linc2
 0.18925615  0.70949799  0.09630606

R> coef(H3D, "h1")

(Intercept)        educ        size
 -1.5171078   0.1227077   0.2183154

R> coef(H3D, "sd")

       sd
0.9114107
```

```
R> coef(summary(H3D), "h3")
```

```
                Estimate   Std. Error    t-value   Pr(>|t|)
(Intercept)  0.1620993584 0.2914887149  0.5561085 0.57813667
age         -0.0402986918 0.0191473160 -2.1046653 0.03532045
age2         0.0008873679 0.0006215521  1.4276645 0.15338845
smsayes      0.6685687685 0.2761335069  2.4211794 0.01547024
```

```
R> vcov(H3D, "h3")
```

```
              (Intercept)           age          age2       smsayes
(Intercept)  8.496567e-02 -1.671992e-03  3.135901e-05 -3.546625e-02
age         -1.671992e-03  3.666197e-04 -1.038823e-05 -2.233757e-03
age2         3.135901e-05 -1.038823e-05  3.863270e-07  6.150671e-05
smsayes     -3.546625e-02 -2.233757e-03  6.150671e-05  7.624971e-02
```

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

```
R> logLik(H3D)
```

```
'log Lik.' -993.0319 (df=15)
```

```
R> logLik(H3D, naive = TRUE)
```

```
'log Lik.' -1105.029 (df=4)
```

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(H3D))
```

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

```
R> predict(H3D,
+         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)))
```

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 59) and with (see formula 61) 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(H3D, type = "coefdet")
```

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 60) and with (see formula 62) 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(H3D, type = "lratio", adj = TRUE)
```

```
[1] 0.09106802
```

The Vuong test based on the T_{LR} statistic, as presented in section 3.3 (see formula 66), is also provided as a criteria for model selection within the family of 8 strictly non-nested models of Figure 1[9] :

```
R> vuongtest(S2D, P2D)
```

```
        Vuong Test (non-nested)

data:  S2D-P2D
z = -2.4301, p-value = 0.007546
```

According to this outcome, the null hypothesis stating the equivalence between the two models is rejected in favour of the alternative hypothesis stating that `P2D` is better than `S2D`.

Testing the hypothesis of no correlation between the good selection mechanism, the purchasing 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 coefficients, namely the test of the hypothesis $H_0 : \rho_{12} = \rho_{13} = \rho_{23} = 0$, specifying an independent mhurdle model, against the alternative hypothesis specifying a corresponding dependent mhurdle model. This test is performed using the log-likelihood ratio (LR) statistic

[9]Note that Vuong tests for strictly non-nested models can be performed using the `vuong` function of the **pscl** package of Jackman (2012) for `glm` models and some specific count data models.

(65). 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 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(H3D, H3I, type = 'nested', hyp = TRUE)

        Vuong Test (nested)

data:  H3D-H3I
chisq = 8.7817, df = 3, p-value = 0.03234

R> vuongtest(H3D, H3I, type = 'nested', hyp = FALSE)

        Vuong Test (nested)

data:  H3D-H3I
wchisq = 8.7817, sev = 11.798, p-value = 0.453
```

According to these outcomes, the null hypothesis of zero correlation is rejected if the unrestricted model is assumed to be correctly specified and accepted otherwise.

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

```
R> S2Db <- mhurdle(shows ~ educ + size + sex + smsa | linc + linc2,
+                  Interview, dist = "ln", h2 = TRUE, corr = TRUE, method = "bhhh")
```

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(S2D, S2Db, type="overlapping")

        Vuong Test (overlapping)

data:  S2D-S2Db
wchisq = 14.654, sev = 182.36, p-value = 0.823
```

The null hypothesis is not rejected.

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(S2D, S2Db, type="overlapping", hyp=TRUE)
```

```
        Vuong Test (overlapping)

data:  S2D-S2Db
wchisq = 12.534, sev = -11.564, p-value < 2.2e-16
```

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. Unexplained interdependence between latent variables 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 continuing.

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.

Akpan S, Nkanta VS, Essien U (2012). "A double-hurdle model of fertilizer adoption and optimum use among farmers in southern Nigeria." *Tropicultura*, **30**(4), 249–253.

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

Angulo AM, Gil JM, Gracia A (2001). "The demand for alcoholic beverages in Spain." *Agricultural economics*, **26**, 71–83.

Aristei D, Perali F, Pieroni L (2008). "Cohort, age and time effects in alcohol consumption by Italian households: a double-hurdle approach." *Empirical Economics*, **35**(1), 29–61.

Aristei D, Pieroni L (2008). "A double-hurdle approach to modelling tobacco consumption in Italy." *Applied Economics*, **40**(19), 2463–2476.

Aristei D, Pieroni L (2009). "Addiction, social interactions and gender differences in cigarette consumption." *Empirica*, **36**(3), 245–272.

Bertail P, Caillavet F, Nichele V (1999). "A bootstrapped double hurdle analysis: consumption of home-produced food." *Applied Economics*, **31**(12), 1631–1639.

Blaylock JR, Blisard WN (1992). "Self-evaluated health status and smoking behaviour." *Applied Economics*, **24**(4), 429–435.

Blaylock JR, Blisard WN (1993). "Wine consumption by US men." *Applied Economics*, **25**(5), 645–651.

Blisard N, Blaylock J (1993). "Distinguishing between Market Participation and Infrequency of Purchase Models of Butter Demand." *American Journal of Agricultural Economics*, **75**(2), pp. 314–320.

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

Box GEP, Cox DR (1964). "An analysis of transformations." *Journal of the Royal Statistical Society. Series B (Methodological)*, **26**(2), 211–252.

Brouhle K, Khanna M (2012). "Determinants of participation versus consumption in the Nordic Swan eco-labeled market." *Ecological Economics*, **73**(0), 142 – 151.

Burton M, Dorsett R, Young T (1996). "Changing preferences for meat: Evidence from UK household data, 1973-93." *European review of agricultural ecoomics*, **23**, 357–370.

Burton M, Dorsett R, Young T (2000). "An investigation of the increasing prevalence of nonpurchase of meat by British households." *Applied Economics*, **32**(15), 1985–1991.

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.

Chaze JP (2005). "Assessing household health expenditure with Box-Cox censoring models." *Health economics*, **14**, 893–907.

Cheng H, Capps OJ (1988). "Demand analysis of fresh and frozen finfish and shellfish in the United States." *American Journal of Agricultural Economics*, **70**(3), 533–542.

code by Alan Genz R code by Brenton Kenkel F, based on Adelchi Azzalini's 'mnormt' package (2012). *pbivnorm: Vectorized Bivariate Normal CDF*. R package version 0.5-1, URL http://CRAN.R-project.org/package=pbivnorm.

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.

Crowley F, Eakins J, Jordan D (2012). "Participation, expenditure and regressivity in the Irish lottery: evidence from Irish household budget survey 2004-05." *Economic and social review*, **43**(2), 199–225.

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

Elek P, Köllö J, Reizer B, Szabó PA (2011). "Detecting wage under-reporting using a double hurdle model." Discussion paper no 6224, IZA discussion paper series.

Fuller FH, Beghin JC, Rozelle S (2007). "Consumption of dairy products in urban China: results from Beijing, Shangai and Guangzhou." *Australian Journal of Agricultural and Resource Economics*, **51**(4).

Gao X, Wailes EJ, Cramer GL (1995). "Double-hurdle model with bivariate normal errors: an application to U.S. rice demand." *Journal of Agricultural and applied economics*, **27**(2), 363–376.

Garcia J, Labeaga JM (1996). "Alternative approaches to modelling zero expenditure: an application to spanish demand for tobacco." *Oxford Bulletin of Economics and Statistics*, **58**(3), 489–506.

Gould BW (1992). "At-Home Consumption of Cheese: A Purchase-Infrequency Model." *American Journal of Agricultural Economics*, **74**(2), pp. 453–459.

Haines PS, Guilkey DK, Popkin BM (1988). "Modeling Food Consumption Decisions as a Two-Step Process." *American Journal of Agricultural Economics*, **70**(3), pp. 543–552.

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.

Henningsen A (2013). *censReg: Censored Regression (Tobit) Models*. R package version 0.5-20, URL http://CRAN.R-project.org/package=censReg.

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

Humphreys BR, Lee YS, Soebbing BP (2009). "Consumer behaviour in lotto markets: the double hurdle approach and zeros in gambling survey data." WP no 2009-27, University of Alberta.

Humphreys BR, Ruseski JE (2010). "The economic choice of participation and time spent in physical activity and sport in Canada." WP no 2010-14, University of Alberta.

Jackman S (2012). *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 1.04.4, URL http://pscl.stanford.edu/,http://CRAN.R-project.org/package=pscl.

Johnson N (1949). "Systems of frequency curves generated by methods of translation." *Biometrika*, **36**(1-2), 149–176.

Jones AM (1989). "A double-hurdle model of cigarette consumption." *Journal of applied econometrics*, **4**(1), 23–39.

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

Keelan CD, Henchion MM, Newman CF (2009). "A Double-Hurdle Model of Irish Households' Food Service Expenditure Patterns." *Journal of International Food & Agribusiness Marketing*, **21**(4), 269–285.

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.

Mal P, Anik AR, Bauer S, Schmitz PM (2012). "Bt cotton adoption: a double-hurdle approach for north Indian farmers." *AgBioForum*, **15**(3), 294–302.

Martínez-Espineira (2006). "A Box-Cox double-hurdle model of wildlife valuation: the citizen's perspective." *Ecological economics*, **58**, 192–208.

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

Moffatt P (2005). "Hurdle models of loan default." *Journal of the operational research society*, **56**(9), 1063–1071.

Mutlu S, Gracia A (2004). "Food Consumption Away from Home in Spain." *Journal of Food Products Marketing*, **10**(2), 1–16.

Newman C, Henchion M, Matthews A (2001). "Infrequency of purchase and doubleâĂŘhurdle models of Irish households' meat expenditure." *European Review of Agricultural Economics*, **28**(4), 393–419.

Newman C, Henchion M, Matthews A (2003). "A double-hurdle model of Irish household expenditure on prepared meals." *Applied Economics*, **35**(9), 1053–1061.

Okello JJ, Kirui OK, Gitonga Z (2012). "A triple hurdle analysis of the use of electronic-based agricultural market information services: the case of smallholder farmers in Kenya." Internatinal association of agricultural economists triennal conference, Foz de Iguacu, Brazil.

Okunade AA, Suraratdecha C, Benson DA (2010). "Determinants of Thailand household healthcare expenditure: the relevance of permanent resources and other correlates." *Health Economics*, **19**(3), 365–376. ISSN 1099-1050. doi:10.1002/hec.1471.

Popkin BM, Guilkey DK, Haines PS (1989). "Food Consumption Changes of Adult Women between 1977 and 1985." *American Journal of Agricultural Economics*, **71**(4), pp. 949–959.

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

Reynolds A (1990). "Analyzing Fresh Vegetable Consumption From Household Survey Data." *Southern Journal of Agricultural Economics*, **22**(02), 31–38.

Saz-Salazar Sd, Rausell-Köster P (2008). "A double-hurdle model of urban green areas valuation: dealing with zero responses." *Landscape and urban planning*, **84**, 241–251.

Smith M (2002). *Handbook of applied econometrics and statistical inference*, chapter On specifying double hurdle models, p. chapter 25. Marcel Dekker, New-York.

Su SJB, Yen ST (1996). "Microeconometric models of infrequently purchased goods: an application to household pork consumption." *Empirical economics*, **21**, 513–533.

Teklewold H, Dadi L, Yami A, Dana N (2006). "Determinants of adoption of poultry technology: a double-hurdle approach." *Livestock research for rural development*, **18**(3).

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

Therneau T (2013). *survival: Survival Analysis, Including Penalised Likelihood*. R package version 2.37-4, 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 (2008). "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.

Toomet O, Henningsen A (2012). *maxLik: Maximum Likelihood Estimation*. R package version 1.1-2, URL http://CRAN.R-project.org/package=maxLik,http://www.maxLik.org.

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

Wang J, Gao XM, Wailes EJ, Cramer GL (1996a). "U.S. Consumer Demand for Alcoholic Beverages: Cross-Section Estimation of Demographic and Economic Effects." *Review of Agricultural Economics*, **18**(3), pp. 477–489.

Wang Q, Jensen HH, Yen ST (1996b). "Impact of cholesterol information on US egg consumption: evidence from consumer survey data." *Applied Economics Letters*, **3**(3), 189–191.

Wodjao TB (2020). "A double-hurdle model of computer and internet use in american households." Departement of Economics, Western Michigan University.

Yen ST (1993). "Working Wives and Food Away from Home: The Box-Cox Double Hurdle Model." *American Journal of Agricultural Economics*, **75**(4), pp. 884–895.

Yen ST (1994). "Cross-section estimation of US demand for alcoholic beverage." *Applied Economics*, **26**(4), 381–392.

Yen ST (1995). "Alternative transformations in a class of limited dependent variable models: alcohol consumption by US women." *Applied Economics Letters*, **2**(8), 258–262.

Yen ST (1999). "Gaussian versus count-data hurdle models: cigarette consumption by women in the US." *Applied Economics Letters*, **6**(2), 73–76.

Yen ST (2005). "Zero observations and gender differences in cigarette consumption." *Applied Economics*, **37**(16), 1839–1849.

Yen ST, Dellenbarger LE, Schupp AR (1995). "Determinants Of Participation And Consumption: The Case Of Crawfish In South Louisiana." *Journal of Agricultural and Applied Economics*, **27**(01).

Yen ST, Huang CL (1996). "Household demand for Finfish: a generalized double-hurdle model." *Journal of agricultural and resource economics*, **21**(2), 220–234.

Yen ST, Jensen HH (1996). "Determinants of Household Expenditures on Alcohol." *Journal of Consumer Affairs*, **30**(1), 48–67. ISSN 1745-6606.

Yen ST, Jensen HH, Wang O (1996). "Cholesterol information and egg consumption in the US: A nonnormal and heteroscedastic double-hurdle model." *European Review of Agricultural Economics*, **23**(3), 343–356.

Yen ST, Jones AM (1996). "Individual cigarette consumption and addiction: a flexible limited dependent variable approach." *Health economics*, **5**, 105–117.

Yen ST, Jones AM (1997). "Household consumption of cheese: an inverse hyperbolic sine double-hurdle model with dependent errors." *American journal of agricultural economics*, **79**(1), 246–251.

Yen ST, Su SJ (1995). "Modeling U.S. Butter Consumption With Zero Observations." *Agricultural and Resource Economics Review*, **24**(1).

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.

Zeileis A, Kleiber C, Jackman S (2008). "Regression Models for Count Data in R." *Journal of Statistical Software*, **27**(8), 1–25. ISSN 1548-7660. URL http://www.jstatsoft.org/v27/i08.

Zhang F, Huang CL, Lin BH (2008). "Modeling fresh organic produce consumption: a generalized double-hurdle model approach." *Agribusiness*, **24**(4), 510–522.

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