# Multiple hurdle models in R: The mhurdle Package

**Fabrizio Carlevaro**
Université de Genève
Yves Croissant
Université de la Réunion
Stéphane Hoareau
Université de la Réunion

**Abstract**

**mhurdle** is a package for R enabling the estimation of a wide set of models for which the response is zero left-censored. This kind of models are called limited dependent or *Tobit* models in the econometric literature and are of particular interest to analyze households' consumption data provided by family expenditure surveys.

*Keywords*:˜limited dependent variable, maximum likelihood estimation, R.

## 1. Introduction

In applied econometric studies, the dependent variable often exhibits a large proportion of fixed values, *e.g.*:

- the number of hours of work supplied is zero for all unemployed or inactive persons;
- the expenditure for particular goods are nil for all households not consuming these goods;
- the attendance of a show is always equal to the capacity of the room each time the show is performed at "position closed".

In these circumstances, ordinary least-squares estimation is biased and inconsistent. However, the model can be estimated consistently using maximum likelihood methods by taking into account the censored nature of the dependent variable.

This problem has been treated for a long time in the statistics literature dealing with survival models which are implemented in R with the **survival** package of Therneau and Lumley (2008).

It has also close links with the problem of selection bias, for which some methods are implemented in the **sampleSelection** package of Toomet and Henningsen (2008b).

**mhurdle** deals specifically with models where the dependent variable is zero-left censored and the observations' sample consequently may present a large proportion of zeros, which is typically the case in household expenditure surveys[1].

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

Since Tobin (1958) seminal paper, a large econometric literature has been developed to deal correctly with this problem of zero observations. More specifically, zero observations may appear for the following three reasons:

**lack of resources** : the household would like to consume the good, but cannot afford it with its present budget;

**good rejection** : the good is not selected by the household, because it is harmful or can be replaced by some substitute good ;

**purchase infrequency** : the good is bought by the household, but with a low frequency so that zero expenditure may be observed if the survey is carried out over a too short period (see Deaton and Irish 1984).

The original Tobin's model takes only the first source of zeros into account. With **mhurdle**, the three sources of zero may be introduced in the model.

For each of the three sources of zeros, a continuous latent variable is defined, with a zero observed if the latent variable is negative. These latent variables are defined as the sum of a linear combination of covariates and a random disturbance with a possible correlation between the disturbances of different latent variables.

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

# 2. Econometric framework

## 2.1. Model specification

Our modeling strategy rests on the following three equations:

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

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

- The first equation defines the *good selection mechanism* : if $y_1^* < 0$, the good is not consumed because it is not identified by the household as a relevant consumption good.
- The second equation defines the *desired consumption level* of the good ; therefore, if $y_2^* < 0$, the good is not consumed, as a negative consumption level implied by the budget constraint cannot be realized.

- The third equation defines the *frequency of purchase mechanism*: if $y_3^* < 0$ the good is not purchased during the survey period, while it is purchased at least one time when $y_3^* > 0$. Assuming that the survey period is a fraction $P$ of the purchase period, a purchase $y = y_2^*/P$ is observed with probability $P = Prob\{y_3^* > 0\}$ during the survey while no purchase is observed with probability $(1 - P)$.

As $y_1^*$ and $y_3^*$ are unobservable indicators of dichotomous variables, $\epsilon_1$ and $\epsilon_3$ stand for $N(0, 1)$ random disturbances, while $\epsilon_2 \sim N(0, \sigma^2)$ with unknown $\sigma^2$, since $y_2^*$ is an observable variable when uncensored.

A priori information may suggest that one or more of these censoring mechanisms is ineffective. 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 equations: one defining the desired consumption level of food and the other the decision of food purchasing during the survey period. Besides, the desired consumption equation explaining dependent variable $y_2^*$ must be specified as a non negative parametric function of covariates $x_2$ and random disturbance $\epsilon_2$. For the time being, two functional forms of this equation have been programmed in **mhurdle**, namely a log-normal functional form :

$$\ln y_2^* = \beta_2^\top x_2 + \epsilon_2$$

and a truncated normal functional form, defined by a linear desired consumption equation with $\epsilon_2$ distributed according to a $N(0, \sigma^2)$ left-truncated at $\epsilon_2 = -\beta_2^\top x_2$, as suggested by Cragg (1971).

A priori information may also suggest to set to zero some or all correlations between random disturbances $\epsilon_1$, $\epsilon_2$, $\epsilon_3$, entailing a partial or total independence between the above defined censoring mechanisms. In particular, it seems appropriate to a priori suppose zero correlation between $\epsilon_1$ and $\epsilon_3$ as well as between $\epsilon_2$ and $\epsilon_3$, as a consequence of the different nature in the determinants responsible, on one hand, of the good selection and desired consumption level decisions and, on the other hand, of those responsible of the frequency of purchase decision.

Figure 1 outlines the full set of special models that can be generated from this general econometric framework by enforcing ineffective censoring mechanisms and by selecting an appropriate functional form of the desired consumption equation.

This figure shows that 12 different consistent models can be estimated by the **mhurdle** package, leading to 23 different parametric specifications when no correlation between random disturbances $\epsilon_1$ and $\epsilon_2$ is considered as a different specification assumption from that of correlated disturbances.

Note that among these models, two are not concerned by censored data, namely models 1 and 2. These two specifications are relevant only for modeling uncensored samples.

All the other models are potentially able to analyze censored samples by combining up to the three censoring mechanisms described above. With the notable exception of the standard Tobit model, that can be estimated also by the **survival** package of Therneau and Lumley (2008), these models cannot be found in an other R-library.

Some of **mhurdle** models have already been used in the applied econometric literature. In particular, models 3 and 4 are single hurdle good selection models originated by Cragg (1971).

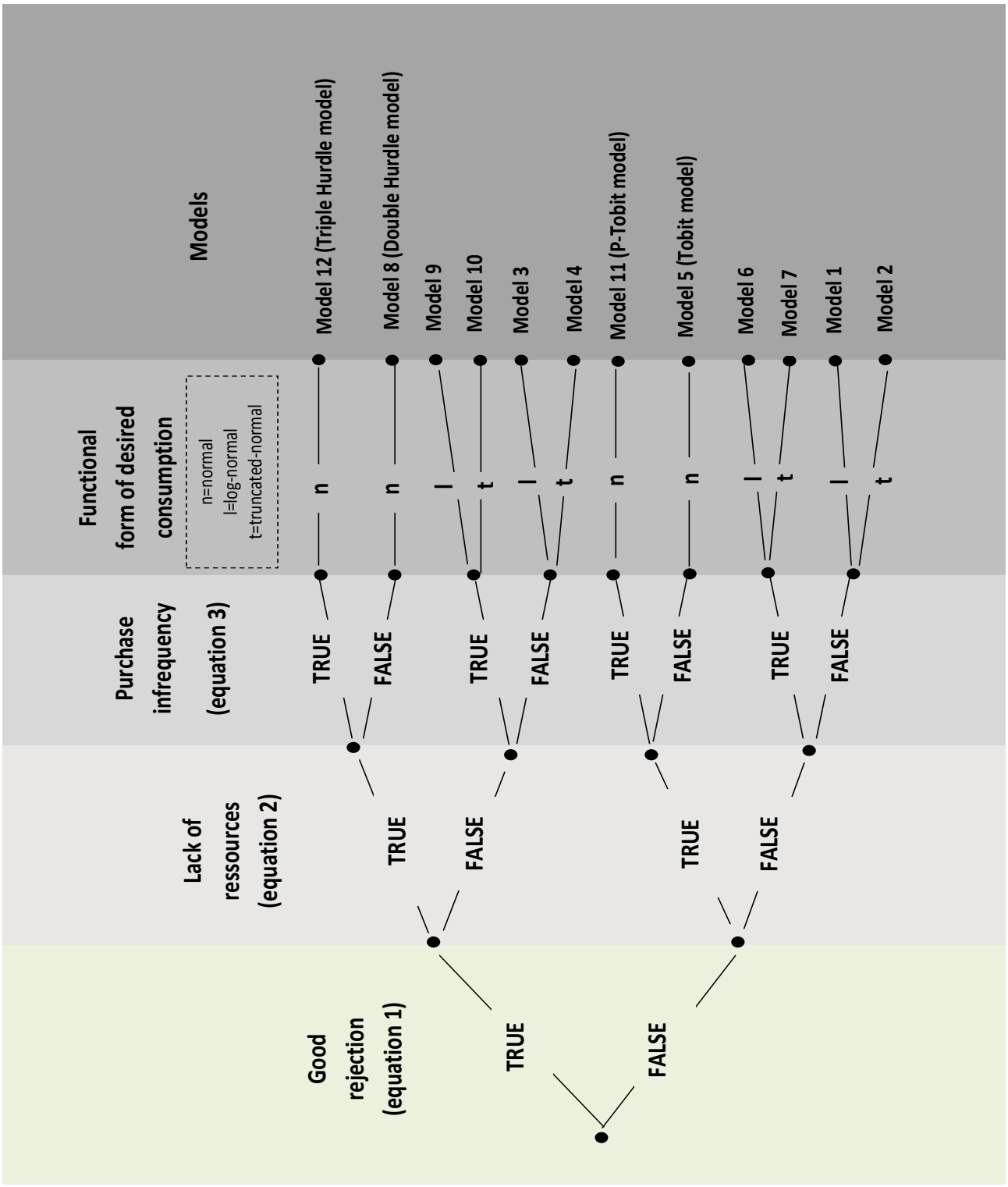


Figure 1: The full set of mhurdle special models.

The double hurdle model combining uncorrelated good selection and lack of resources cen-

soring mechanisms is also due to Cragg (1971); the correlated version of this double hurdle model has been originated by Blundell and Meghir (1987).

P-Tobit model is due to Deaton and Irish (1984) and explains zero purchases as the result of lack of resources and/or infrequent purchases. Models 6 and 7 are single hurdle models not yet used in applied demand analysis, where the operating censoring mechanism is due to infrequent purchases.

Among the original models encompassed by **mhurdle**, models 9 and 10 are double hurdle models combining good selection and frequency of purchase mechanisms to explain censored samples.

Model 12 is an original three hurdle model originated in Hoareau (2009). This model explains censored purchases either as the result of good rejection, lack of resources or infrequent purchases.

## 2.2. Likelihood function

As for the standard Tobit model, the likelihood of our censored models have two components: the first one is the probability of a binary choice (purchasing or not), the second one is the density function of the chosen expenditure level of consumption for the households that consume.

The contribution of a zero observation to the sample log-likelihood function can be written as follow:

$$\ln L_i^- = \begin{cases} \ln\left(1-\Phi\left(\beta_1^\top x_{1i}, \frac{\beta_2^\top x_{2i}}{\sigma}; \rho\right)\Phi(\beta_3^\top x_{3i})\right) & \text{for normal models} \\ \ln\left(1-\Phi(\beta_1^\top x_{1i})\Phi(\beta_3^\top x_{3i})\right) & \text{for log-normal models} \\ \ln\left(1-\frac{\Phi\left(\beta_1^\top x_{1i}, \frac{\beta_2^\top x_{2i}}{\sigma}; \rho\right)}{\Phi\left(\frac{\beta_2^\top x_{2i}}{\sigma}\right)}\Phi(\beta_3^\top x_{3i})\right) & \text{for truncated-normal models} \end{cases}$$

where $\Phi(z)$ denotes the distribution function of a $N(0,1)$ random variable. We remind that log-normal and truncated normal models assume that the lack of resources mechanism is inoperative, whereas it is operative in normal models.

The second and the third expression correspond to the case where a zero purchase is observed only for good rejection or for purchase infrequency reasons. In the second expression, the desired consumption equation is specified according to a log-normal functional form whereas, in the third expression, it is specified according to a truncated-normal functional form. The first expression corresponds to the case where a zero purchase is observed either for good rejection, for lack of resources or for purchase infrequency reasons.

These expressions become simpler in the following special cases:

- when the good selection mechanism is inoperative, implying:

$$P\{y_{1i}^* > 0\} = \Phi(\beta_1^\top x_{1i}) = 1$$

and consequently:

$$\ln L_i^- = \begin{cases} \ln\left(1-\Phi\left(\frac{\beta_2^\top x_{2i}}{\sigma}\right)\Phi(\beta_3^\top x_{3i})\right) & \text{for normal models} \\ \ln\left(1-\Phi(\beta_3^\top x_{3i})\right) & \text{otherwise} \end{cases}$$

- when the purchase frequency mechanism is inoperative, implying:

$$P\{y_{3i}^* > 0\} = \Phi(\beta_3^\top x_{3i}) = 1$$

and consequently:

$$\ln L_i^- = \begin{cases} \ln\left(1 - \Phi\left(\beta_1^\top x_{1i}, \frac{\beta_2^\top x_{2i}}{\sigma}; \rho\right)\right) & \text{for normal models} \\ \ln\left(1 - \Phi(\beta_1^\top x_{1i})\right) & \text{for log-normal models} \\ \ln\left(1 - \frac{\Phi\left(\beta_1^\top x_{1i}, \frac{\beta_2^\top x_{2i}}{\sigma}; \rho\right)}{\Phi\left(\frac{\beta_2^\top x_{2i}}{\sigma}\right)}\right) & \text{for truncated-normal models} \end{cases}$$

- when the good selection mechanism and the desired consumption equation are uncorrelated ($\rho = 0$), implying:

$$\Phi\left(\beta_1^\top x_{1i}, \frac{\beta_2^\top x_{2i}}{\sigma}; 0\right) = \Phi(\beta_1^\top x_{1i})\Phi\left(\frac{\beta_2^\top x_{2i}}{\sigma}\right)$$

and consequently:

$$\ln L_i^- = \begin{cases} \ln\left(1 - \Phi(\beta_1^\top x_{1i})\Phi\left(\frac{\beta_2^\top x_{2i}}{\sigma}\right)\Phi(\beta_3^\top x_{3i})\right) & \text{for normal models} \\ \ln\left(1 - \Phi(\beta_1^\top x_{1i})\Phi(\beta_3^\top x_{3i})\right) & \text{otherwise} \end{cases}$$

- when both the good selection and the frequency of purchase mechanisms are inoperative, implying:

$$\ln L_i^- = \begin{cases} \ln\left(1 - \Phi\left(\frac{\beta_2^\top x_{2i}}{\sigma}\right)\right) & \text{for normal models} \\ -\infty & \text{for log-normal and truncated-normal models} \end{cases}$$

Consequently, in this very special case, log-normal and truncated-normal model specifications can only be used to analyze uncensored samples.

The contribution of a positive observation to the log-likelihood function is best presented by defining a "residual" of the fit as :

$$e_i = \begin{cases} \ln y_i + \ln\Phi(\beta_3^\top x_{3i}) - \beta_2^\top x_{2i} & \text{for log-normal models} \\ y_i\Phi(\beta_3^\top x_{3i}) - \beta_2^\top x_{2i} & \text{otherwise} \end{cases}$$

One observes that the parameters and the covariates of the frequency of purchase equation enter the definition of this "residual", because this residual is defined for the average consumption, which depends on the probability of purchasing, as described previously.

The contribution of a positive observation to the log-likelihood function is then written as :

$$\begin{aligned} \ln L_i^+ &= -\ln\sigma + \ln\phi\left(\tfrac{e_i}{\sigma}\right) + \ln\Phi\left(\frac{\beta_1^\top x_{1i} + \frac{\rho}{\sigma}e_i}{\sqrt{(1-\rho^2)}}\right) + \ln\Phi(\beta_3^\top x_{3i}) \\ &+ \begin{cases} \ln\Phi(\beta_3^\top x_{3i}) & \text{for normal models} \\ -\ln y_i & \text{for log-normal models} \\ -\ln\Phi\left(\frac{\beta_2^\top x_{2i}}{\sigma}\right) + \ln\Phi(\beta_3^\top x_{3i}) & \text{for truncated-normal models} \end{cases} \end{aligned}$$

where $\phi(z)$ denotes the density function of a $N(0,1)$ random variable.

As for the log-likelihood function of a censored observation, the expression of the log-likelihood function of an uncensored observation become simpler in the following special cases:

- when the good selection mechanism is inoperative, implying:

$$\begin{aligned}\ln L_i^+ &= -\ln\sigma + \ln\phi\left(\tfrac{e_i}{\sigma}\right) + \ln\Phi(\beta_3^\top x_{3i}) \\ &+ \begin{cases} \ln\Phi(\beta_3^\top x_{3i}\ ) & \text{for normal models} \\ -\ln y_i & \text{for log-normal models} \\ -\ln\Phi\left(\frac{\beta_2^\top x_{2i}}{\sigma}\right) + \ln\Phi(\beta_3^\top x_{3i}\ ) & \text{for truncated-normal models} \end{cases}\end{aligned}$$

- when the purchase frequency mechanism is inoperative, implying:

$$e_i = \begin{cases} \ln y_i - \beta_2^\top x_{2i} & \text{for log-normal models} \\ y_i - \beta_2^\top x_{2i} & \text{otherwise} \end{cases}$$

and

$$\begin{aligned}\ln L_i^+ &= -\ln\sigma + \ln\phi\left(\tfrac{e_i}{\sigma}\right) + \ln\Phi\left(\frac{\beta_1^\top x_{1i} + \frac{\rho}{\sigma} e_i}{\sqrt{(1-\rho^2)}}\right) \\ &+ \begin{cases} -\ln y_i & \text{for log-normal models} \\ -\ln\Phi\left(\frac{\beta_2^\top x_{2i}}{\sigma}\right) & \text{for truncated-normal models} \end{cases}\end{aligned}$$

- when the good selection mechanism and the desired consumption equation are uncorrelated ($\rho = 0$), implying:

$$\begin{aligned}\ln L_i^+ &= -\ln\sigma + \ln\phi\left(\tfrac{e_i}{\sigma}\right) + \ln\Phi(\beta_1^\top x_{1i}) + \ln\Phi(\beta_3^\top x_{3i}) \\ &+ \begin{cases} \ln\Phi(\beta_3^\top x_{3i}\ ) & \text{for normal models} \\ -\ln y_i & \text{for log-normal models} \\ -\ln\Phi\left(\frac{\beta_2^\top x_{2i}}{\sigma}\right) + \ln\Phi(\beta_3^\top x_{3i}\ ) & \text{for truncated-normal models} \end{cases}\end{aligned}$$

- when both the good selection and the frequency of purchase mechanisms are inoperative, implying:

$$e_i = \begin{cases} \ln y_i - \beta_2^\top x_{2i} & \text{for log-normal models} \\ y_i - \beta_2^\top x_{2i} & \text{otherwise} \end{cases}$$

and

$$\begin{aligned}\ln L_i^+ &= -\ln\sigma + \ln\phi\left(\tfrac{e_i}{\sigma}\right) \\ &+ \begin{cases} -\ln y_i & \text{for log-normal models} \\ -\ln\Phi\left(\frac{\beta_2^\top x_{2i}}{\sigma}\right) & \text{for truncated-normal models} \end{cases}\end{aligned}$$

Combining these log-likelihood function for zero and positive expenditure observations, the sample log-likelihood function is written as :

$$\ln L = \sum_{i|y_i=0} \ln L_i^- + \sum_{i|y_i>0} \ln L_i^+$$

Note that for uncorrelated single-hurdle good selection models, $\ln L_i^-$ depends only on $\beta_1$ and $\ln L_i^+$ depends only on $\beta_2$ and $\sigma$, allowing to separate the model estimation according to two independent models :

- a binary probit model allowing to estimate $\beta_1$ independently of $\beta_2$ and $\sigma$;
- a linear, log-linear or truncated regression model allowing to estimate $\beta_2$ and $\sigma$ independently of $\beta_1$.

## 3. Model estimation, evaluation and selection

The econometric framework described in the previous section provides a theoretical framework 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 correct model specification and for a likelihood function sufficiently well behaved, the maximum likelihood estimator is asymptotically efficient within the class of consistent and asymptotically normal estimators[2].

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

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

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

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

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

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

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

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

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

Two fundamental principles should be used to appraise the results of a model estimation, namely its economic relevance and its statistical and predictive adequacy. The first principle deals with the issues of accordance of model estimate with the economic rationale underlying the model specification and of its relevance for answering the questions for which the model has been built. These issues are essentially context specific and, therefore, cannot be dealt with 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 McFadden (1974) to measure the relative gain in the maximized 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 mhurdle model. This model is written as:

$$y_i = E(y_i) + \varepsilon_i, i = 1, ..., n$$

where $\varepsilon_i$ stands for a zero expectation, heteroskedastic random disturbance and $E(y_i) = Prob\{y_i > 0\}E(y_i|y_i > 0)$, with $Prob\{y_i > 0\} = 1 - L_i^-$ and

$$E(y_i|y_i > 0) = \begin{cases} \frac{\beta_2^\top x_{2i}}{\Phi(\beta_3^\top x_{3i})} + \sigma \frac{\psi_n(\beta_1^\top x_{1i}, \frac{\beta_2^\top x_{2i}}{\sigma}; \rho)}{\Phi(\beta_1^\top x_{1i}, \frac{\beta_2^\top x_{2i}}{\sigma}; \rho)\Phi(\beta_3^\top x_{3i})} & \text{for normal and truncated-normal models} \\ \frac{\exp\left\{\beta_2^\top x_{2i} + 0.5\sigma^2\left(1-\rho^2\right)\right\}\psi_l(\beta_1^\top x_{1i}; \rho\sigma)}{\Phi(\beta_1^\top x_{1i})\Phi(\beta_3^\top x_{3i})} & \text{for log-normal models} \end{cases}$$

where

$$\psi_n(\beta_1^\top x_{1i}, \frac{\beta_2^\top x_{2i}}{\sigma}; \rho) = \int_{-\beta_1^\top x_{1i}}^{\infty} \left[\rho\varepsilon_1 \Phi\left(\frac{\frac{\beta_2^\top x_{2i}}{\sigma} + \rho\epsilon_1}{\sqrt{1-\rho^2}}\right) + \sqrt{1-\rho^2}\phi\left(\frac{\frac{\beta_2^\top x_{2i}}{\sigma} + \rho\epsilon_1}{\sqrt{1-\rho^2}}\right)\right] \phi(\varepsilon_1) d\epsilon_1$$

and

$$\psi_l(\beta_1^\top x_{1i};\rho\sigma)=\int_{-\beta_1^\top x_{1i}}^{\infty}\exp\{\rho\sigma\varepsilon_1\}\phi(\varepsilon_1)d\epsilon_1$$

Notice that these two last integrals can be computed using the first terms of a Taylor series expansion around $\rho=0$ and $\rho\sigma=0$, respectively, as detailed for the first integral in Carlevaro, Croissant, and Hoareau (2008). Moreover, the above general expressions of $E(y_i|y_i>0)$ become simpler in the following special cases:

- when the good selection mechanism is inoperative ($\Phi(\beta_1^\top x_{1i})=1$), leading to:

$$E(y_i|y_i>0)=\begin{cases}\frac{\beta_2^\top x_{2i}}{\Phi(\beta_3^\top x_{3i})}+\sigma\frac{\phi(\frac{\beta_2^\top x_{2i}}{\sigma})}{\Phi(\frac{\beta_2^\top x_{2i}}{\sigma})\Phi(\beta_3^\top x_{3i})} & \text{for normal and truncated-normal models}\\ \frac{\exp\left\{\beta_2^\top x_{2i}+0.5\sigma^2\right\}}{\Phi(\beta_3^\top x_{3i})} & \text{for log-normal models}\end{cases}$$

- when the purchase frequency mechanism is inoperative ($\Phi(\beta_3^\top x_{3i})=1$), leading to:

$$E(y_i|y_i>0)=\begin{cases}\beta_2^\top x_{2i}+\sigma\frac{\psi_n(\beta_1^\top x_{1i},\frac{\beta_2^\top x_{2i}}{\sigma};\rho)}{\Phi(\beta_1^\top x_{1i},\frac{\beta_2^\top x_{2i}}{\sigma};\rho)} & \text{for normal and truncated-normal models}\\ \exp\left\{\beta_2^\top x_{2i}+0.5\sigma^2\left(1-\rho^2\right)\right\}\frac{\psi_l(\beta_1^\top x_{1i};\rho\sigma)}{\Phi(\beta_1^\top x_{1i})} & \text{for log-normal models}\end{cases}$$

- when the good selection mechanism and the desired consumption equation are uncorrelated ($\rho=0$), implying:

$$\Phi\left(\beta_1^\top x_{1i},\frac{\beta_2^\top x_{2i}}{\sigma};0\right)=\Phi(\beta_1^\top x_{1i})\Phi\left(\frac{\beta_2^\top x_{2i}}{\sigma}\right)$$

as well as:

$$\psi_n\left(\beta_1^\top x_{1i},\frac{\beta_2^\top x_{2i}}{\sigma};0\right)=\psi_l\left(\beta_1^\top x_{1i};0\right)=\Phi(\beta_1^\top x_{1i})$$

and consequently:

$$E(y_i|y_i>0)=\begin{cases}\frac{\beta_2^\top x_{2i}}{\Phi(\beta_3^\top x_{3i})}+\sigma\frac{\phi\left(\frac{\beta_2^\top x_{2i}}{\sigma}\right)}{\Phi(\frac{\beta_2^\top x_{2i}}{\sigma})\Phi(\beta_3^\top x_{3i})} & \text{for normal and truncated-normal models}\\ \frac{\exp\left\{\beta_2^\top x_{2i}+0.5\sigma^2\right\}}{\Phi(\beta_3^\top x_{3i})} & \text{for log-normal models}\end{cases}$$

namely the same formulas of $E(y_i|y_i>0)$ as when the good selection mechanism is inoperative;

- when both the good selection and the frequency of purchase mechanisms are inoperative, leading to:

$$E(y_i|y_i>0)=\begin{cases}\beta_2^\top x_{2i}+\sigma\dfrac{\phi\left(\frac{\beta_2^\top x_{2i}}{\sigma}\right)}{\Phi(\frac{\beta_2^\top x_{2i}}{\sigma})} & \text{for normal and truncated-normal models}\\ \exp\left\{\beta_2^\top x_{2i}+0.5\sigma^2\right\} & \text{for log-normal models}\end{cases}$$

Denoting by $\hat{y}_i$ the fitted values of $y_i$ obtained by computing predictor $E(y_i)$ for $y_i$ with the maximum likelihood estimate of model parameters, we define a pseudo coefficient of determination for a mhurdle model according to the following formula:

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

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 mhurdle model without covariates (intercept-only model). Notice 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.

Two other formulas, which are equivalent to compute $R^2$ in the linear regression model with intercept, could have been used to define a pseudo coefficient of determination, namely: the ratio of the explained sum of square to the total sum of squares or the squared correlation between actual and fitted values. We disregarded these alternatives because the former measure can exceed one in a non linear regression model, while the latter, although providing values always within zero and one, cannot be adjusted for degrees of freedom for a use as a model selection criterion. A more promising approach consists in computing $RSS$ and $TSS$ with standardized residuals, to correct for the heteroskedasticity of row residuals. This Pearson goodness of fit measure, requiring to write down analytically the variance of $\varepsilon_i$, is not currently implemented.

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

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

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

### 3.3. Model selection

Model selection deals with the problem of discriminating between alternative model specifications used to explain the same dependent variable, with the view of finding the one best

suited to explain the sample of observations at hand. This decision problem can be tackled from two point of view, namely that of the model specification achieving the best in-sample fit, on one hand, and that of the model specification that is favored in a formal test comparing two model alternatives, on the other hand.

The first 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 parametrization. Indeed, the value of these measures of fit can be improved by increasing model parametrization, in particular when the parameter estimates are obtained by optimizing a criteria functionally related to the selected measure of fit, as it is the case when using the $\rho^2$ fit measure with a maximum likelihood estimate. Consequently, a penalty that increases with the number of model parameters should be added to the $R^2$ and $\rho^2$ fit measures to trade off goodness of fit improvements with parameter parsimony losses.

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

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

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

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

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

is equivalent to choosing the model specification that minimizes the Akaike (1973) predictor of the Kullback-Liebler 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:

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

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

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

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

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

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

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:

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

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

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

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 work 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 $\hat{\theta_*}$ and $\hat{\pi_*}$, respectively, and have asymptotic normal distributions centered 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. However, for model comparisons within the set of mhurdle special models presented in FIG. 1, only the first two cases are really relevant, at least as long as we compare model specifications with identical covariates.

A parametric model $G_\pi$ defined by the conditional density function (cdf) $g(y|x;\pi)$ is said to be nested in parametric model $F_\theta$ with cdf $f(y|x;\theta)$, if and only if any cdf of $G_\pi$ is equal to a cdf of $F_\theta$, for almost all $x$. Within our mhurdle special models this is the case when comparing two specifications differing only with respect to the presence or the absence of correlated disturbances. For these models, it is necessarily the case that $f(y|x;\theta_*) \equiv g(y|x;\pi_*)$. Therefore $H_0$ is tested against $H_f$.

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

- under $H_0$, the quantity $2LR(\hat{\theta},\hat{\pi})$ converges in distribution towards a weighted sum of $K+L$ iid $\chi^2(1)$ random variables, where the weights are the $K+L$ possibly negative eigenvalues of a theoretical symmetric matrix, that can be consistently estimated by a

sample analogue. Notice that the density function of this random variable has not been worked out analytically. Therefore, we compute it by simulation.

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

As a consequence, for a test with critical value $c$, $H_0$ is rejected in favor 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. Notice that, if model $F_\theta$ is correctly specified, the asymptotic distribution of the LR statistic is, as expected, a $\chi^2$ random variable with $K - L$ degrees of freedom.

Two parametric models $F_\theta$ and $G_\pi$ defined by cdf $f(y|x;\theta)$ and $g(y|x;\pi)$ are said to be strictly non-nested, if and only if no cdf of model $F_\theta$ is equal to a cdf of $G_\pi$, for almost all $x$, and conversely. Within mhurdle special models this is the case when comparing two specifications differing with respect either to the effective censoring mechanisms 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 standardized LR statistic:

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

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

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

As a consequence, for a test with critical value $c$, $H_0$ is rejected in favor 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 favor 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. Notice that, if one of models $F_\theta$ or $G_\pi$ is assumed to be correctly specified, the Cox (1961, 1962) LR test of non nested models needs to be used.

## 4. Software rationale

There are three important issues to be addressed to correctly implement in R the econometric framework described in the previous section. The first one is to provide a good interface to describe the model to be estimated. The second one is the problem of finding good starting values for computing model estimates. The third one is to offer flexible optimization tools for likelihood maximization.

### 4.1. Model syntax

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

For the models implemented in **mhurdle**, three kinds of covariates should be specified: the ones of the consumption equation (denoted $x_2$), the ones of the selection equation (denoted $x_1$) and those of the infrequency equation (denoted $x_3$). To define a model with three kinds of covariates, a general solution is given by the **Formula** package developed by Zeileis and Croissant (2010), which provides extended formula objects. To define a model where `y` is the censored dependent variable, `x21` and `x22` two covariates for the desired consumption equation, `x11` and `x12` two covariates for the selection and `x31` and `x32` two covariates for the infrequency of purchase equation, we use the following commands :

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

To illustrate the use of **Formula**, let's use the `tobin` data.frame from the **survival** package. This data.frame is a sub-sample of 20 observations of the original data used by Tobin (1958) in his seminal paper.

```
R> data("tobin", package = "survival")
R> head(tobin, 3)
```

```
  durable  age quant
1     0.0 57.7   236
2     0.7 50.9   283
3     0.0 48.5   207
```

The variables of this data.frame are :

**durable:** the durable good expenditures in thousands of US$;

**age:** the age of the head of the family in years;

**quant:** the liquidity ratio in per thousands.

To estimate a model for durable good expenditures using `age` and `quant` as covariates for the desired consumption equation, `age` for the selection equation, and `quant` for the purchase infrequency equation, we use the following syntax:

```
R> f <- Formula(durable ~ age | age + quant | quant)
```

Several methods are provided to deal with these extended formulas. In particular, the model covariate matrices for the three equations are easily computed using:

```
R> S <- model.matrix(f, data = tobin, rhs = 1)
R> X <- model.matrix(f, data = tobin, rhs = 2)
R> P <- model.matrix(f, data = tobin, rhs = 3)
R> head(X, 3)

  (Intercept)  age quant
1           1 57.7   236
2           1 50.9   283
3           1 48.5   207

R> head(S, 3)

  (Intercept)  age
1           1 57.7
2           1 50.9
3           1 48.5

R> head(P, 3)

  (Intercept) quant
1           1   236
2           1   283
3           1   207
```

For the end user, all these manipulations are internal to `mhurdle` function. All he should do is entering a formula of the type `y ~ x11 + x12 | x21 + x22 | x31 + x32` as first argument of the function.

## 4.2. Starting values

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

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

According to Carlevaro *et˜al.* (2008) experience in applying this estimation procedure to two-hurdle models, this approach doesn't seem to work well with our correlated hurdle-models.

Indeed, except for the very special case of model 3 (log-normal correlated single-hurdle selection model), the probability of observing a censored purchase is not that of a simple probit model (see the formula of $\ln L_i^-$).

As noted previously, for uncorrelated single-hurdle good selection 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.

In case of correlated single-hurdle good selection models, the coefficient maximum likelihood estimate of the corresponding uncorrelated model ($\rho = 0$) is used as starting values.

For purchase infrequency models (P-Tobit models), the starting values are computed using an Heckman-like two step procedure. In the first step, parameters $\beta_3$ are estimated using a simple probit. In the second step, a linear, log-linear or linear-truncated model is estimated on the sub-sample of uncensored observations using $y_i \Phi(\beta_3' x_3)$ or $\ln y_i + \ln \Phi(\beta_3' x_3)$ (in the case of a log-normal specification) as the dependent variable of the regression model estimated by ordinary least squares.

### 4.3. Optimisation

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

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

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

The behavior 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`.

## 5. Examples

The package is loaded using:

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

### 5.1. Estimation

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

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

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

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

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

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

**corr:** a logical argument indicating whether the disturbances of the selection equation and the consumption equation are correlated or not. The default is `FALSE`.

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

Different combinations of these arguments lead to a large variety of models. Note that some of them are logically inconsistent and therefore irrelevant. For example, a model with no good selection equation and `corr = TRUE` is logically inconsistent because only good selection and desired consumption equations can be correlated.

To illustrate the use of `mhurdle` package, we first estimate an independent triple-hurdle model, which we call `model12i` :

```
R> model12i <- mhurdle(durable ~ age + quant | age + quant | age +
+     quant, tobin, dist = "n", method = "bfgs")
```

In applied work, the issue may be to select the relevant hurdles. As an alternative to the previously estimated three hurdle model we can now estimate more a priori restricted models where only one or two hurdles are relevant.

To estimate a model where only lack of resources is relevant to explain censored durable good expenditures, we use :

```
R> model5 <- mhurdle(durable ~ 0 | age + quant | 0, tobin, dist = "n",
+     method = "nr")
```

To estimate an independent log-normal single-hurdle good rejection model, we use:

```
R> model3i <- mhurdle(durable ~ age + quant | age + quant | 0, tobin,
+     dist = "l")
```

To estimate a log-normal single-hurdle purchase infrequency model, we use:

```
R> model6 <- mhurdle(durable ~ 0 | age + quant | age + quant, tobin,
+     dist = "l")
```

To estimate an independent model where censured durable good expenditures may be explained by lack of resources or good rejection, we use :

```
R> model8i <- mhurdle(durable ~ age + quant | age + quant | 0, tobin,
+     dist = "n")
```

We then update this model in order to estimate a dependent double-hurdle (lack of resources or good rejection) model:

```
R> model8d <- update(model8i, corr = TRUE)
```

### 5.2. Methods

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

```
R> summary(model8i)

Call:
mhurdle(formula = durable ~ age + quant | age + quant | 0, data = tobin,
    dist = "n")

Frequency of 0:  0.65

Newton-Raphson maximisation
gradient close to zero
5 iterations, 0h:0m:0s
g'(-H)^-1g = 6.12E-22

Coefficients :
                 Estimate Std. Error t-value  Pr(>|t|)
sel.(Intercept)  1.461792   3.710805  0.3939 0.6936338
sel.age         -0.122834   0.072312 -1.6987 0.0893830 .
sel.quant        0.017997   0.015802  1.1389 0.2547578
reg.(Intercept) 12.841869   5.321390  2.4133 0.0158108 *
reg.age          0.404577   0.098441  4.1098 3.959e-05 ***
reg.quant       -0.113719   0.019904 -5.7132 1.108e-08 ***
sigma            1.434599   0.397077  3.6129 0.0003028 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -22.186 on 7 Df
rho: score test : z = 0.016 (p.value = 0.494)
```

```
R^2 :
 McFadden   : 0.25143
 Regression : 0.37556
```

This method displays the percentage of 0 in the sample, the coefficient table, several measures of goodness of fit and, for independent models, a score test of correlation.

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

Coefficients 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(model12i, "reg")

 (Intercept)          age        quant
 0.200499276  0.013152493 -0.002808693

R> coef(model12i, "sel")

(Intercept)         age       quant
 9.64594522  0.27385000 -0.07744716

R> coef(model12i, "sigma")

    sigma
0.7949458

R> coef(summary(model12i), "ifr")

              Estimate Std. Error      t-value  Pr(>|t|)
(Intercept)  2.0520340  696.83632  0.002944786 0.9976504
age         -2.5905681   37.01866 -0.069980062 0.9442095
quant        0.5774693   10.34643  0.055813405 0.9554905

R> vcov(model12i, "reg")

            (Intercept)          age         quant
(Intercept) 13.86563162 -0.113600675 -0.0350976222
age         -0.11360067  0.008281271 -0.0011242924
quant       -0.03509762 -0.001124292  0.0003623083
```

Log-likelihood may be obtained for the estimated model or for a "naive" model, *i.e.* a model without covariates. Moreover, the component of the likelihood for null and for positive observations may be obtained separately :

```
R> logLik(model12i)
```

```
[1] -18.1697

R> logLik(model12i, which = "positive")

[1] -13.47851

R> logLik(model12i, naive = TRUE, which = "zero")

     zero
-5.600195
```

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

```
R> head(fitted(model12i))

          zero      positive
[1,] 1.0000000  1.809788e+29
[2,] 0.4795935  1.703160e+00
[3,] 0.9999849  2.743720e+05
[4,] 0.4478696  4.027116e+00
[5,] 0.4366752  4.276167e+00
[6,] 1.0000000 4.486644e+174
```

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

```
R> pr <- predict(model12i, newdata = data.frame(durable = c(0, 1,
+     0), age = c(50, 32, 48), quant = c(206, 232, 245)))
R> head(pr)

          zero     positive
[1,] 1.0000000 9.074063e+17
[2,] 0.6532498 8.239173e-01
[3,] 0.4428734 3.816195e+00
```

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 appropriate model specification.

Two goodness of fit measures are provided. The first measure is an extention to limited dependent variable models of the classical coefficient of determination for linear regression models. This pseudo coefficient of determination is computed both without ($R^2$) and with ($\bar{R}^2$) adjustment for the loss of sample degrees of freedom due to model parametrization. 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> r.squared(model12i, which = "all", type = "regression")
```

```
[1] 0.4608687
```

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 ($\rho^2$) and with ($\bar{\rho}^2$) adjustment for the loss of sample degrees of freedom due to model parametrization, in order to allow model comparisons with the same or with a different number of parameters.

```
R> r.squared(model12i, type = "mcfadden", which = "all", dfcor = TRUE)
```

```
[1] 0.1922852
```

The Vuong test based on the $T_{LR}$ statistic, as presented in section 3.3, is also provided as a criteria for model selection within the family of 12 strictly non nested models of FIG. 1.

```
R> vuongtest(model6, model8d)
```

```
        Vuong Test (non-nested)

data:  model6 model8d
z = -1.0376, p-value = 0.1497
alternative hypothesis: The second model is better
```

Testing the hypothesis of no correlation between the good selection mechanism and the desired consumption equation can be performed by means of a Wald test, a Lagrange multiplier (LM) test or a log-likelihood ratio (LR) statistic.

Likelihood ratio tests are performed using a Vuong test, and more precisely the nested version of this test. 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 model is correctly specified. In the first case, the p-value is computed using the standard chi square distribution, in the second case a weighted chis square distribution is used.

```
R> vuongtest(model8d, model8i, type = "nested", hyp = TRUE)
```

```
        Vuong Test (nested)

data:  model8d model8i
chisq = 0.002, df = 1, p-value = 0.9646
alternative hypothesis: The larger model is better
```

```
R> vuongtest(model8d, model8i, type = "nested", hyp = FALSE)
```

```
        Vuong Test (nested)

data:  model8d model8i
wchisq = 0.002, df = 1, p-value = 0.661
alternative hypothesis: The larger model is better
```

The LM test is performed using the independent model (`model8i`). The `summary` performs the test, as seen previously, the p-value for this test is 0.494. The Wald test is simply obtained in the coefficient table of the dependent model (`model8d`)

```
R> coef(summary(model8d), "rho")

      Estimate Std. Error    t-value  Pr(>|t|)
rho 0.05469814   1.218839 0.04487724 0.9642052
```

In the previous example, all the tests don't reject the hypothesis of no correlation.

## 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 .

## References

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

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

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

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

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

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

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

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

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

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

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

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

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

Therneau T, Lumley T (2008). *survival: Survival Analysis, Including Penalised Likelihood*. R package version 2.34-1.

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

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

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

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

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

## 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