Introduction to mhurdle

Contents

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]his package is a new version of a package first developed as part of a PhD dissertation carried out by St'ephane Hoareau (2009) at the University of Reunion 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 rationalized by the existence of corner solutions to the microeconomic problem of rational choice of the neoclassical consumer.[3] Cragg (1971)'s combination of this desired consumption mechanism and the selection mechanism leads a double hurdle model, also called two-part model, which has been extended by Blundell and Meghir (1987) to the case where the two equations are correlated.

- 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, a 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 an infrequently bought good or service. This is typically the case for durable or storable non durable goods, but also for many infrequently bought services like health, leisure, education and training services. Remarkably, this mechanism allows to derive from observed expenditures the rate of use of a durable good, the rate of consumption of a stored non durable good, and the rate of purchase of an infrequently consumed non storable non durable good or service.

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 modeled as functions of explanatory variables and of a random disturbance, with a possible correlation between the disturbances of different latent variables in order to account for a possible simultaneity of the decisions modeled by censoring mechanisms. To model possible departures of the observed dependent variable from normality, as it is common with economic variables, we use transformations of this variable allowing to rescale non normal 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. Smith (2002) presented an overview of these studies. Using an updated overview of these studies carried out by ourselves, 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 (1971) 's paper. 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 modeling

[2]The consumer we are referring 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 literature, we term this concept "the consumer" by convenience.

[3]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.

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 logarithmic, 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 success over time, in terms of statistical significance of estimates.

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, contingent valuation studies, finance, sport activities, internet use, gambling, production.

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 as well as Vuong tests for discriminating between nested and non nested models have been implemented in `mhurdle` package. Vuong tests remarkably permit to compare two competing models when 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 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 in the `hurdle` function 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 modeling 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 qualitative binary response model resting on one of the following three latent dependent variables relations:

$$\left\{\begin{array}{l} y_1^* = \beta_1^\top x_1 + z_1 \\ T(y_2^*) = \beta_2^\top x_2 + \sigma z_2 \\ y_3^* = \beta_3^\top x_3 + z_3 \end{array}\right. \tag{1}$$

where y_1^*, y_2^*, y_3^* stand for continuous latent dependent variables x_1, x_2, x_3 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 latent dependent variables, z_1, z_2, z_3 for standard normal random disturbances. Since variables y_1^* and y_3^* are never observed, contrary to y_2^*, the units of measurement of y_1^* and y_3^* are not

identified. Hence, the disturbances in the corresponding binary response models are normalized by setting their variances equal to 1.

Relying on a Johnson (1949)'s proposal, $T(\cdot)$ denotes a differentiable one-to-one transformation intended to rescale a potentially not normally distributed random variable y_2^* into a normally distributed ones, with σ standard deviation. As the domain of definition of the inverse transformation $T^{-1}(\cdot)$ my be restrained to a subset of real numbers, the standard normal random disturbance z_2 is potentially truncated at the bounds of an interval $[B_1, B_2]$ with B_1 and/or B_2 finite. To account for heteroscedasticity of z_2, in `mhurdle` the standard deviation σ can be specified as a positive monotonic transformation of a linear function of a vector of covariates x_4, namely $\beta_4^\top x_4$. However, for notational simplicity, we will adopt in what follows the notation of an homoscedastic standard deviation σ rather than the more general one of an heteroscedastic standard deviation $\sigma(\beta_4^\top x_4)$.[4]

These equations model the hurdles, namely the sequence of binary decisions a consumer faces when buying a good. Each hurdle decision is coded by a binary variable I_j taking value 1 if $y_j^* > 0$ and 0 otherwise. Therefore, the outcome of each of such decisions is modeled as a probit model with outcome probabilities: $\mathrm{P}(I_j = 1) = \mathrm{P}(y_j^* > 0)$ and $\mathrm{P}(I_j = 0) = 1 - \mathrm{P}(y_j^* > 0)$.

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 $\mathrm{P}(I_3 = 1) = \mathrm{P}(y_3^* > 0)$ 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:

$$y = \frac{T^{-1}(\beta_2^\top x_2 + \sigma z_2)}{\mathrm{P}(I_3 = 1)} I_1 I_2 I_3. \tag{2}$$

This modeling strategy provides an explanatory framework where censoring of the observed dependent variable results from the effect of up to three different censoring mechanisms, namely: $I_1 = 0$ and/or $I_2 = 0$ and/or $I_3 = 0$.

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 and the other the decision to purchase food during the observation period.

When the first and/or the third hurdles are supposed not to be in effect, the previous binary response models for I_1 and/or I_3 must be replaced by singular probability response models where $\mathrm{P}(I_1) = 1$ and/or $\mathrm{P}(I_3) = 1$. When the second hurdle is supposed to be inoperative, we must not only replace the binary response model explaining I_2 by a singular probability response model where $\mathrm{P}(I_2 = 1) = 1$ but also select a monotonic transformation $T(y_2^*)$ enforcing non-negative values to y_2^*.

Cragg (1971) suggested two types of functional forms of the distribution of y_2^* enforcing such constraint, namely the (inverse) log-normal distribution:

$$y_2^* = \exp\{\beta_2^\top x_2 + \sigma z_2\} \tag{3}$$

and the normal distribution left-truncated at 0:

$$y_2^* = \max\{0; \beta_2^\top x_2 + \sigma z_2\} = \beta_2^\top x_2 + \sigma \max\{B_1; z_2\} \tag{4}$$

[4]For a practical application of this parametric model of heteroscedasticity, a functional form of $\sigma(\beta_4^\top x_4)$ written as $\exp\{\beta_{40} F(\beta_4^\top x_4)\}$, where $F(.)$ is the distribution function of the standard normal random variable, has been programmed in `mhurdle`. This functional form remains bounded with respect to any possible values of covariates x_4 and allows testing the assumption of homoscedasticity, $\sigma = \exp\{\beta_{40}/2\}$, by assessing the statistical non significance of parameter vector β_4 without an intercept.

where $\max\{B_1; z_2\}$ is a standard normal random variable left-truncated at $B_1 = -\beta_2^\top x_2/\sigma$.

To account for these two functional forms within our general formal framework (1) and illustrate in section 5 their empirical use with `mhurdle`, in this paper we consider two transformations $T(y_2^*)$.[5]

The first transformation is the identity transformation which generates an un-truncated normal distribution of y_2^* when hurdle 2 is in effect, and the left-truncated normal distribution (4) when hurdle 2 is not in effect.

The second transformation is a slight generalization of the logarithmic transformation leading to the (inverse) log-normal distribution (3), namely:

$$T(y_2^*) = \ln(y_2^* + \alpha) \tag{5}$$

with α a location parameter which shifts the left-truncation bound of y_2^* towards negative or positive values according to the sign of α. As the inverse of this transformation is written as:

$$y_2^* = \exp\{\beta_2^\top x_2 + \sigma z_2\} - \alpha \tag{6}$$

it turns out that this transformation allows to model skewed distributions of y_2^* according to a translated (inverse) log-normal functional form where the support of z_2 is un-truncated. By the same token, testing the statistical significance of parameter α, against the alternative $\alpha > 0$, amounts to testing the assumption that hurdle 2 is not in effect. Conversely, when $\alpha < 0$ the transformation (5) holds for $y_2^* > -\alpha$ meaning that $-\alpha$ stands for a "committed" consumption of a basic necessity, while $\alpha > 0$ typify a luxury good whose consumption occurs only above a given income threshold.[6]

To derive the form of the probability distribution of the observable dependent variable y, we must still specify the joint distribution of the random disturbances entering the hurdle relations assumed to be in effect and that of the demand function, which in turn depends on the domain of definition of transformation $T^{-1}(\cdot)$.

For trivariate hurdle models, where hurdles 1 and 3 are in effect (but not necessarily hurdle 2), the joint density function of z_1, z_2 and z_3 is a possibly truncated standard trivariate normal density function written as:

$$\phi\left(z_1, z_2, z_3; \rho\right)/\Pi \tag{7}$$

where $\Pi = \Phi(B_2) - \Phi(B_1)$ with $\Phi(z)$ the distribution function of a standard univariate normal distribution, $\phi(z_1, z_2, z_3; \rho)$ the density function of a standard trivariate normal distribution with ρ the vector of the three symmetric correlation coefficients ρ_{12}, ρ_{13}, ρ_{23} between the couples of normal standard random variables z_1 and z_2, z_1 and z_3, z_2 and z_3, respectively.

For bivariate hurdle models, where either hurdle 1 or hurdle 3 is not in effect, the joint density function of z_i, with $i = 1$ or 3, and z_2 is a possibly truncated standard bivariate normal density function written as:

$$\phi\left(z_i, z_2; \rho_{i2}\right)/\Pi \tag{8}$$

where $\phi\left(z_i, z_2; \rho_{i2}\right)$ denotes the density function of a standard bivariate normal distribution.

Finally, for univariate hurdle models, where both hurdles 1 and 3 are not in effect, the density function of z_2 is a possibly truncated standard univariate normal density function:

[5] To model possible departures of the observed dependent variable y from normality towards distributions of the kind encountered with real-world economic data, two families of flexible parametric transformations $T(.)$ were programmed in `mhurdle`, namely the two parameter Box and Cox (1964) transformation, as suggested by Lankford and Wyckoff (1991) and Chaze (2005), and Johnson (1949)'s one parameter inverse hyperbolic sine transformation. The first family, which includes transformations (4) and (5) as special cases, generates a broad range of skewed distributions, while the second family generates a broad range of symmetric leptokurtic (more sharply peaked than the normal) distributions.

[6] This economic interpretation of parameter α my justify to model this parameter as a function of some covariates, like household demographic characteristics.

$$\phi(z_2)/\Pi \tag{9}$$

where $\phi(z_2)$ 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, 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 seven, 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.

2.2 Probability distribution of censored dependent variable

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 $\mathrm{P}(y=0)$ to $y=0$ and a density function $f_+(y)$ to any $y>0$, with:

$$\mathrm{P}(y=0)+\int_0^\infty f_+(y)dy=1. \tag{10}$$

To compute the probability mass $\mathrm{P}(y=0)=1-\mathrm{P}(y>0)$ and the density function $f_+(y)$ we must first establish the joint density function of the latent variables entering the hurdle model.

For trivariate hurdle models, the joint density function of latent variables y_1^*, y_2^* and y_3^* can be derived from the density function (7) of variables z_1, z_2 and z_3 by means of the change of variables:

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

which leads to:

$$f(y_1^*, y_2^*, y_3^*) = \frac{T^{'}(y_2^*)}{\sigma} \frac{\phi\left(y_1^* - \beta_1^\top x_1, (T(y_2^*) - \beta_2^\top x_2)/\sigma, y_3^* - \beta_3^\top x_3; \rho\right)}{\Pi}, \tag{12}$$

where $T^{'}(y_2^*)$ stands for the derivative of $T(y_2^*)$.

To compute $\mathrm{P}(y>0)$, we integrate this density function over the three-dimensional positive octant using the change of variables (11) and the symmetry property of density function (7). Performing this integration for a strictly increasing transformation $T(\cdot)$[7], which is the case for the identity and logarithmic transformations we focus in this paper, and z_2 truncated over the interval $[B_1, B_2]$ leads to:

[7]For a strictly decreasing transformation, as it is the case for some transformations of the Box-Cox family, the integration with respect to z_2 must be performed by reversing the direction of integration, from $(T(0) - \beta_2^\top x_2)/\sigma$ to B_1. By applying the theorem on the inversion of the integration limits of a definite integral, we obtain a closed form of the integral similar to that of formula (13) with truncation value B_2 replaced by B_1.

$$
\begin{aligned}
\mathrm{P}(y>0) &= \int_{-\beta_1^\top x_1}^{\infty} \int_{\frac{T(0)-\beta_2^\top x_2}{\sigma}}^{B_2} \int_{-\beta_3^\top x_3}^{\infty} \frac{\phi(z_1,z_2,z_3;\rho)}{\Pi} dz_1 dz_2 dz_3 \\
&= \int_{-\infty}^{\beta_1^\top x_1} \int_{-B_2}^{\frac{\beta_2^\top x_2 - T(0)}{\sigma}} \int_{-\infty}^{\beta_3^\top x_3} \frac{\phi(z_1,z_2,z_3;\rho)}{\Pi} dz_1 dz_2 dz_3 \\
&= \frac{\Phi(\beta_1^\top x_1, (\beta_2^\top x_2 - T(0))/\sigma, \beta_3^\top x_3;\rho) - \Phi(\beta_1^\top x_1, -B_2, \beta_3^\top x_3;\rho)}{\Pi}.
\end{aligned} \tag{13}
$$

Therefore, for z_2 left-truncated at $B_1 = -\beta_2^\top x_2/\sigma$, which is the case for the truncated normal distribution (4), this closed form for $\mathrm{P}(y>0)$ simplifies to:

$$
\mathrm{P}(y>0) = \frac{\Phi(\beta_1^\top x_1, (\beta_2^\top x_2 - T(0))/\sigma, \beta_3^\top x_3;\rho)}{\Phi(\beta_2^\top x_2/\sigma)}, \tag{14}
$$

while for the un-truncated normal and the translated (inverse) log-normal distribution (6) it is written as:

$$
\mathrm{P}(y>0) = \Phi(\beta_1^\top x_1, (\beta_2^\top x_2 - T(0))/\sigma, \beta_3^\top x_3;\rho). \tag{15}
$$

To compute the density function $f_+(y)$ we first derive the joint density function of variables y_1^*, y and y_3^* by performing the change of variable $y_2^* = \Phi_3 y$ with $\Phi_3 = \Phi(\beta_3^\top x_3) = \mathrm{P}(I_3 = 1)$ on the joint density function (12), which leads to :

$$
f(y_1^*, y, y_3^*) = \frac{\Phi_3 T^{'}(\Phi_3 y)}{\sigma} \frac{\phi\left(y_1^* - \beta_1^\top x_1, (T(\Phi_3 y) - \beta_2^\top x_2)/\sigma, y_3^* - \beta_3^\top x_3;\rho\right)}{\Pi}. \tag{16}
$$

We must then integrate this transformed density function over the positive quadrant of the bi-dimensional support of variables y_1^* and y_3^*. To this purpose we rewrite this trivariate normal density function as the product of the marginal density function of y, written as:

$$
f(y) = \frac{\Phi_3 T^{'}(\Phi_3 y)}{\sigma} \frac{\phi\left((T(\Phi_3 y) - \beta_2^\top x_2)/\sigma\right)}{\Pi}, \tag{17}
$$

and of the conditional joint density function $f(y_1^*, y_3^* | y)$, which is bivariate normal with expectations, standard-deviations and correlation coefficient written as:

$$
\begin{aligned}
&\mu_{i|2}(y) = \beta_i^\top x_i + \rho_{i2}(T(\Phi_3 y) - \beta_2^\top x_2)/\sigma, \quad \sigma_{i|2} = \sqrt{1-\rho_{i2}^2}, \quad i = 1,3 \\
&\rho_{13|2} = \frac{\rho_{13} - \rho_{12}\rho_{23}}{\sqrt{1-\rho_{12}^2}\sqrt{1-\rho_{23}^2}}.
\end{aligned} \tag{18}
$$

Integrating this factorization of density function (16) with respect to positive values of y_1^* and y_3^* leads to the following closed form for $f_+(y)$:

$$
\begin{aligned}
f_+(y) &= f(y) \int_0^\infty \int_0^\infty f(y_1^*, y_3^* | y) dy_1^* dy_3^* \\
&= f(y) \int_{-\frac{\mu_{1|2}(y)}{\sigma_{1|2}}}^{\infty} \int_{-\frac{\mu_{3|2}(y)}{\sigma_{3|2}}}^{\infty} \phi\left(z_1, z_3; \rho_{13|2}\right) dz_1 dz_3 \\
&= f(y) \Phi\left(\frac{\mu_{1|2}(y)}{\sigma_{1|2}}, \frac{\mu_{3|2}(y)}{\sigma_{3|2}}; \rho_{13|2}\right).
\end{aligned} \tag{19}
$$

The probability distribution of the observed censored variable y for bivariate and univariate hurdle models can be derived from that of the trivariate model by eliminating hurdles 1, 3 and 1 and 3, respectively. By eliminating hurdle 1 we obtain:

$$\mathrm{P}(y>0) = \frac{\Phi\left((\beta_2^\top x_2 - T(0))/\sigma, \beta_3^\top x_3; \rho_{23}\right) - \Phi\left(-B_2, \beta_3^\top x_3; \rho_{23}\right)}{\Pi} \tag{20}$$

and

$$f_+(y) = \frac{\Phi_3 T^{'}(\Phi_3 y)}{\sigma} \frac{\phi\left((T(\Phi_3 y) - \beta_2^\top x_2)/\sigma\right)}{\Pi} \Phi\left(\frac{\mu_{3|2}(y)}{\sigma_{3|2}}\right), \tag{21}$$

while the elimination of hurdle 3 leads to:

$$\mathrm{P}(y>0) = \frac{\Phi\left(\beta_1^\top x_1, (\beta_2^\top x_2 - T(0))/\sigma; \rho_{12}\right) - \Phi\left(\beta_1^\top x_1, -B_2; \rho_{12}\right)}{\Pi} \tag{22}$$

and

$$f_+(y) = \frac{T^{'}(y)}{\sigma} \frac{\phi\left((T(y) - \beta_2^\top x_2)/\sigma\right)}{\Pi} \Phi\left(\frac{\mu_{1|2}(y)}{\sigma_{1|2}}\right). \tag{23}$$

Finally, removing both hurdles 1 and 3 leads to the univariate censured regression model with:

$$\mathrm{P}(y>0) = \frac{\Phi\left((\beta_2^\top x_2 - T(0))/\sigma\right) - \Phi\left(-B_2\right)}{\Pi} \tag{24}$$

and

$$f_+(y) = \frac{T^{'}(y)}{\sigma} \frac{\phi\left((T(y) - \beta_2^\top x_2)/\sigma\right)}{\Pi}. \tag{25}$$

The econometric framework described in the previous section provides a theoretical background for tackling the problems of model estimation, evaluation, selection and prediction within the statistical theory of classical inference. In `mhurdle` these statistical issues are tackled by assuming that data at hand are those observed on a sample of individuals selected in a large population using a sampling design generating a data base having the features of a random sample.[8]

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

[8]Data provided by surveys conducted by official statistical offices usually fulfill such requirement, at least as far as the survey is performed according to a random sampling design where no one of the modeled dependent variables was used to design the sampling scheme. On this important issue see Chapter 24 of Cameron and Trivedi (2005).

3 Likelihood function

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.[9]

For a random sample of n observations of the censored dependent variable y it is easy to derive its likelihood function from the results set out in previous section 2.2. As these observations are all independently drawn from the same conditional (on covariates x_1, x_2, x_3 and x_4) discrete-continuous distribution, which assigns a conditional probability mass $\mathrm{P}(y=0)$ to the observed value $y=0$ and a conditional density function $f_+(y)$ to the observed values $y>0$, the log-likelihood function for an observation y_i can be written as :

$$\ln L_i = \begin{cases} \ln \mathrm{P}(y_i=0) & \text{if} \quad y_i = 0 \\ \ln f_+(y_i) & \text{if} \quad y_i > 0 \end{cases} \tag{26}$$

and the log-likelihood for the entire sample:

$$\ln L = \sum_{i=1}^{n} \ln L_i = \sum_{i|y_i=0} \ln \mathrm{P}(y_i=0) + \sum_{i|y_i>0} \ln f_+(y_i). \tag{27}$$

A maximum likelihood estimate of model parameters is obtained by numerically maximizing this function with respect to all the unknown model parameters entering this function. Algorithms used in `mhurdle` to perform this numerical optimization are described later in section 4.2. To the extent that such an estimate is obtained, classical statistical inference about model parameters θ under the assumption of a correctly specified model is performed in `mhurdle` using the asymptotic approximation of the distribution of maximum likelihood estimators, which is normal, with expectation θ and covariance matrix $(nI_A(\theta))^{-1}$ where $I_A(\theta)$ stands for the asymptotic Fisher information matrix of a random sample of n observations.[10] For empirical applications this theoretical matrix can be consistently estimated using one of the following two matrices:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \ln L_i(\hat{\theta})}{\partial\theta\partial\theta^\top} \quad \text{or} \quad \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ln L_i(\hat{\theta})}{\partial\theta}\frac{\partial \ln L_i(\hat{\theta})}{\partial\theta^\top}$$

which limits (for $n\to\infty$) are called the Hessian and the gradient versions of $I_A(\theta)$. These matrices are directly provided by two standard iterative methods used to compute $\hat{\theta}$, namely the Newton-Raphson and BHHH methods mentioned in section 4.2.

4 Model evaluation and selection using goodness of fit measures

4.1 Model evaluation and selection using goodness of fit measures

To assess the goodness of fit of a `mhurdle` model estimate we can not rely on a single measure, as a sample of censored observations of dependent variable y consists of two sub-samples of incomparable observations, namely the sub-sample of zero observations and that of positive observations. The former is a sample of observations of a qualitative binary variable coding the outcome of the censoring mechanisms at work, while the latter is a sample of observations of a quantitative variable measuring the outcome of the demand function, when this demand is uncensured. As a consequence, a measure of goodness of fit for each of these two

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

[10] A more robust approach to classical statistical inference with respect to model specification errors will be presented later, in subsection 3.3.

samples of observations must be used as the goodness of fit is not measured the same way in qualitative binary response models as in quantitative response models.

To this purpose, we reformulate a `mhurdle` model as a two-part model. The first-part specifies a model for the censoring mechanism explaining the outcome of the qualitative binary variable:

$$d = \left\{ \begin{array}{lll} 0 & \text{if} & y = 0 \\ 1 & \text{if} & y > 0 \end{array} \right. \tag{28}$$

according to a probit model assigning the probability $\mathrm{P}(y = 0)$ to the censored observation ($d = 0$) and the probability $\mathrm{P}(y > 0) = 1 - \mathrm{P}(y = 0)$ to the uncensored observation ($d = 1$). Therefore, the log-likelihood for this first-part model is written as:

$$\ln L_0 = \sum_{i=1}^{n} [(1 - d_i) \ln \mathrm{P}(y_i = 0) + d_i \ln \mathrm{P}(y_i > 0)]. \tag{29}$$

The second-part specifies, conditionally to an uncensored response being observed, the demand model explaining the outcome of quantitative variable y according to the following conditional density function:

$$f(y|y > 0) = f_+(y)/\mathrm{P}(y > 0). \tag{30}$$

For this second-part model, the log-likelihood is written as:

$$\ln L_+ = \sum_{i=1}^{n} d_i [\ln f_+(y) - \ln \mathrm{P}(y_i > 0)]. \tag{31}$$

Adding-up the log-likelihoods of these two-part model leads to:

$$\begin{array}{rl} \ln L_0 + \ln L_+ = & \sum_{i=1}^{n} [(1 - d_i) \ln \mathrm{P}(y_i = 0) + d_i \ln f_+(y)] \\ = & \sum_{i|y_i=0} \ln \mathrm{P}(y_i = 0) + \sum_{i|y_i>0} \ln f_+(y_i) = \ln L, \end{array} \tag{32}$$

meaning that the two-part model is entirely consistent with an integrated specification of `mhurdle` model.

To develop a measure of fit for binary response models which may be intuitively interpreted in a similar way to the coefficient of determination R^2 in the linear regression model, Estrella (1998) suggests to rely on the relationship between R^2 and the classical likelihood-based test statistics, namely the likelihood ratio (LR), Wald (W) and Lagrange multiplier/score (LM) statistics. In the normal linear regression model, where such relationships holds exactly for the likelihood ratio statistics[11], it is written as:

$$R^2 = 1 - \exp\{-A\} \tag{33}$$

with $A = -(2/n) \ln(L_c/L_u)$ the average value of the likelihood ratio statistic, L_u the maximum value of the unconstrained likelihood, and L_c the maximum value of the likelihood constrained by the hypothesis of a model without covariates (intercept-only model).

From this relationships, it follows that $R^2 = 0$ for $A = 0$, which corresponds to the case where the fit of the unconstrained model is identical to that of the constrained model (worst fit), and $R^2 = 1$ when $A = \infty$, the upper bound of the likelihood ratio statistic achieved when the fit of the unconstrained model is perfect (best fit). Between these two limits R^2 increases with A according to a rescaling of A determined by a differential equation written as:

[11] For the Wald and Lagrange multiplier statistics such relationships holds only locally, in a neighborhood of the maintained hypothesis of an intercept-only model, and takes another functional form.

$$dR^2/(1-R^2) = dA. \tag{34}$$

These properties of relationships (33) were adopted by Estrella (1998) as a set of requirements for developing a R^2 analog of the measure of fit for a binary response model. Taking into account that, contrary to the case of a normal linear regression model, for a binary response model the average likelihood ratio statistic A varies between 0 (when $L_u = L_c$) and an upper finite bound $B = -(2/n)\ln L_c$ (when $L_u = 1$), a R^2 analog for a binary response model is obtained by solving the differential equation:

$$dR^2/(1-R^2) = dA/(1-A/B) \tag{35}$$

with the initial condition $R^2(0) = 0$. This solution:

$$R^2 = 1 - (1-A/B)^B \tag{36}$$

provides a rescaling of the likelihood ratio statistic for the binary response model that satisfies the condition $R^2(B) = 1$ and converge towards the rescaling (33) when $B \to \infty$.

To conclude, for assessing the goodness of fit of the two parts of a `mhurdle` model we shall use the following R^2-rescaling of likelihoods L_0 and L_+:

$$R_0^2 = 1 - (\ln L_{0c}/\ln L_{0u})^{(2/n)\ln L_{0c}} \quad \text{and} \quad R_+^2 = 1 - (L_{+c}/L_{+u})^{2/n_+} \tag{37}$$

with $n_+ = \sum_{i=1}^n d_i$ the sample size of uncensored observations.

Note that the estimation of an intercept-only `mhurdle` model is not generally possible in its original parametrization because in the absence of covariates many parameters are under-identified, and thus cannot be consistently estimated. To compute the maximum values of the constrained likelihoods L_{0c} and L_{+c}, we must respecify the probability distribution of dependent variable y by means of identifiable parameters. To this end we observe from formula (13) that $\mathrm{P}(y > 0)$ is a constant P_+, and from formula (19) that $f_+(y)$ is proportional to the density function of $T(y)$ which is normal with a constant expectation μ and a constant variance s^2. Hence, the log-likelihood of the constrained first-part model is written as:

$$\ln L_{0c} = n_+ \ln(1-P_+) + n_+ \ln P_+, \tag{38}$$

and provides the following maximum likelihood estimator for P_+: $\hat{P}_+ = n_+/n$. Similarly, the log-likelihood of the constrained second-part model is written as:

$$\ln L_{+c} = \sum_{i=1}^n d_i \ln \phi\left(\frac{T(y_i)-\mu}{s}\right) - n_+ \ln(sP_+). \tag{39}$$

leading to the following estimators for parameters μ and s^2: $\hat{\mu} = \sum_{i=1}^n d_i T(y_i)/n_+$ and $\hat{s}^2 = \sum_{i=1}^n d_i (T(y_i) - \hat{\mu})^2/n_+$.

The maximized values of constrained likelihoods L_{0c} and L_{+c} are obtained by inserting these estimators for parameters P_+, μ and s^2 into formulas (38) and (39). Note that for the logarithmic transformation (5) with a non-zero location parameter α, the constrained log-likelihood (39) must also be maximized with respect to this parameter. This maximization may be performed effectively by considering $\hat{\mu}$ and $\hat{s}^2$ as functions of α, and maximizing the concentrated log-likelihood obtained by inserting these conditional estimators into constrained log-likelihood (39), which becomes a function of the single parameter α written as:

$$\ln \max_{\mu,\sigma^2} L_{+c} = -\frac{n_+}{2}\left[1 + \ln\left(2\pi\frac{n_+}{n}\right) + \ln \hat{s}^2(\alpha)\right]. \tag{40}$$

Goodness of fit measures (37) can also be used for model selection, a decision problem dealing with the identification, among several model specifications used to explain the same dependent variable, of the one that is best suited to explain the sample of available observations. From this point of view, goodness of fit measures (37) allow to identify the model specification achieving the best in-sample fit. To be fair, the goodness of fit measure used for model selection must account for differences in the degree of model parametrization. Indeed, the value of the above goodness of fit measures can be improved by increasing model parametrization, because model estimation is performed by maximizing the model likelihood function from which functionally depend the goodness of fit measures we consider. Consequently, a penalty that increases with the number of model parameters should be added to formulas (37). To this purpose, we shall rely on THEIL(1971)'s correction of the coefficient of determination in the linear regression model. This leads to the following adjusted R^2-rescaling of likelihoods L_0 and L_+:

$$\bar{R}_0^2 = 1 - \frac{n-1}{n-K}(\ln L_{0c}/\ln L_{0u})^{(2/n)\ln L_{0c}} \quad \text{and} \quad \bar{R}_+^2 = 1 - \frac{n-K_0}{n-K}(L_{+c}/L_{+u})^{2/n_+} \tag{41}$$

with K and K_0 the number of parameters in the unconstrained model and the number of parameters in the constrained second-part model, respectively. Such adjustments may not be relevant in large samples but may be relevant in small samples.

4.2 Model selection using Vuong tests

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

This second model selection criterion relies on the use of a test proposed by Vuong (1989). According to the rationale of this test, the "best" parametric model specification among a collection of competing specifications is the one that minimizes the Kullback-Leibler Information Criterion (KLIC), namely a measure of the distance between the conditional probability mass/density function[12] $f(y|x;\theta)$ of a possibly misspecified parametric model and that of the true unknown model denoted by $h(y|x)$, defined by the following formula:

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

where $H(y,x)$ denotes the distribution function of the true joint distribution of (y,x) and θ_* the probability limit, with respect to $H(y,x)$, of the so called quasi-maximum likelihood estimator $\hat{\theta}$ obtained by applying the maximum likelihood method when $f(y|x;\theta)$ is misspecified. Note that KLIC criterion takes a minimum value of 0 when there is a value θ_0 of parameter vector θ such that $f(y|x;\theta_0) = h(y|x)$ almost everywhere[13], meaning that conditional probability function $f(y|x;\theta)$ is correctly specified. At the opposite, KLIC will take large values when $f(y|x;\theta)$ very poorly specifies the true conditional probability function $h(y|x)$.

As the application of this model selection criterion equals selecting the model specification for which the quantity:

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

is the largest, given two competing models with conditional probability 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:

[12] For mhurdle models, this conditional probability function is written as :

$$f(y|x) = \mathrm{P}(y=0|x)^{1-d} f_+(y|x)^d,$$

with d defined by formula (28).

[13] By "almost everywhere" we mean: for all points of the support of the true distribution of (y,x) except those for which the probability $\mathrm{P}_H[(y,x) : f(y|x;\theta_0) \neq h(y|x)] = 0$.

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

meaning that the two models are equivalent, against either:

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

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

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

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 for θ. Hence, $1/n$ times the log-likelihood ratio statistic:

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

is a consistent estimator for $E\left[\ln(f(y|x;\theta_*)/g(y|x;\pi_*))\right]$.

Therefore, an obvious test of H_0 consists in verifying whether a LR-based statistic differs from zero. The distribution of such a 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 centered on these pseudo-true values.

In his seminal paper Vuong (1989) suggests to use different LR-based statistic according to the relation linking the two competing models, which may be nested, strictly non-nested or overlapping models.[14]

In a recent paper Shi (2015) has criticized Vuong testing procedures for discriminating between competing models by showing that they can generate severe size distortions in finite samples. She develops a new test statistic that corrects these size distortions and which universally applies to nested, strictly non-nested and overlapping models. The rationale of Shi test statistic consists in correcting two finite sample size distortions present in the test statistic used by Vuong to discriminate two strictly non-nested models, namely:

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

with $\hat{\omega}^2$ a sample analogue consistent estimator of the variance of the asymptotic distribution of the numerator of statistic T_{LR}, which is $N(0,\omega^2)$ under the null hypothesis H_0.

The first size distortion is due to a bias of order $O(1/n)$ present in the expectation of the asymptotic distribution of the numerator of T_{LR} under H_0. This expectation is written as $-tr(V)/(2n)$ with $tr(V) = tr(A_F^{-1}B_F) - tr(A_G^{-1}B_G)$, where A_F and B_F are the limits of the Hessian and of the gradient version of the Fisher information matrix of model F_θ, respectively, and A_G and B_G those of model G_π. Therefore, for a model for which the information identity $A + B = O$ holds $\delta = -tr(A^{-1}B) = tr(I)$ is equal to the number of model parameters. This suggests to consider δ as a measure of the degree of model parametrization which increases with the number of model parameters and with the discrepancy of the information identity resulting from the nonlinearity of the model with respect to its parameters. As a consequence, the finite sample bias of statistic $LR(\hat{\theta},\hat{\pi})/n$ is given by $(-\delta_F + \delta_G)/(2n)$. It will be negligible for models of similar

[14] These classical Vuong tests have been programmed in our former version of `mhurdle` presented in Carlevaro et al. (2012).

degrees of parametrization but negative and of possibly significant magnitude when model F_θ is more heavily and non-linearly parameterized than model G_π. Hence, in order to avoid an over-rejection of H_0 in favor of the more parameterized model, Shi suggests to correct the Vuong statistic T_{LR} by subtracting to $LR(\hat{\theta}, \hat{\pi})$ a consistent estimator of its asymptotic bias under H_0, namely $-tr(\hat{V})/2$ with $\hat{V}$ a consistent estimator of matrix V.

The second adjustment of Vuong statistic T_{LR} suggested by Shi is intended to prevent the denominator $\hat{\omega}$ of this statistic from taking values close to zero with non negligible probability, which contributes to create a fat tail for the distribution of T_{LR}. This adjustment consists in adding a positive constant term to $\hat{\omega}^2$, written as $c\text{tr}(\hat{V}^2/n)$ with c a positive constant chosen according to a data-dependent method,[15] in order to avoid large values of T_{LR} when its denominator is close to zero while its numerator is not.

Finally the modified Vuong test by Shi, called by her a *non-degenerate Vuong test*, uses the following statistic:

$$T_{LR}^{mod} = \frac{(LR(\hat{\theta}, \hat{\pi}) + \text{tr}(\hat{V})/2)/\sqrt{n}}{\sqrt{\hat{\omega}^2 + c\text{tr}(\hat{V}^2)/n}}. \tag{46}$$

To completely remove the asymptotic size distortion, Shi proposes to compute a simulated critical value $\text{cv}(a)$, which has a well controlled asymptotic size, in the sense that, asymptotically and for any $c \geq 0$, the probability of rejecting H_0 when it is true is upper bounded by a.[16] By systematically changing the value of a in order to equalize the simulated critical value $\text{cv}(a)$ to the observed value of statistic T_{LR}^{mod}, it is possible to derive a simulated p-value allowing to perform a non-degenerate Vuong test at any controlled asymptotic size level.

For non-nested models the test is two sided. It is carried out by rejecting H_0 if the simulated p-value is less than the selected size level a, and in favor of either H_f if $T_{LR}^{mod} > 0$, or of H_g if $T_{LR}^{mod} < 0$.

For nested models the test is one sided because the nested model cannot be closer to the true unknown model than the nesting model according to the KLIC criterion. Supposing without loss of generality that model G_π is nested into model F_θ, the test is carried out by rejecting H_0 in favor of H_f if the simulated p-value is less than the selected size level a. Note that this test is a robust alternative to the usual LR-test for testing the statistical significance of a priori restrictions on the parameters of model F_θ, in the sense that it allows the unrestricted model to be misspecified.

A last comment deserves to be devoted to the choice of the size α of the test, which measures the probability of rejecting H_0 when it is true, and thus the risk of favoring, on grounds of statistical evidence, the use of one of two competing models while they are equivalent. This choice is routinely made by assuming a low value of α (5% or 1%) without referring to the operational meaning of the test result. But in a decision-making framework of selection between two models, the size of the test can be chosen as large as desired because if the two models are equivalent it is irrelevant to use one or the other of these two models. On the other end, the risk of the second kind, namely the probability of rejecting either H_f or H_g when one of these two models is true, must be minimized, because concluding that the two models are equivalent does not allow deciding which one of them should be used. Not rejecting H_0 simply point out that further sample information is needed to determine, with the desired degree of certainty, which of the two model alternatives is the more suitable for the intended uses.

But how to proceed if further data-gathering is impossible, which is the typical situation faced in econometric studies. In this situation the discrimination test between two models must be reformulated as a Simon (1943) symmetric test, namely a binary choice between models F_θ and G_π where none of these models is privileged as a null hypothesis of a classical asymmetric Neyman-Pearson test. This reformulation removes the irrelevant hypothesis of doubt H_0 from the decision-making framework by forcing the test to always conclude in favor of one of the two confronted models, which seems appropriate when there is no a priori

[15] Except for nested models, where c is set equal to zero

[16] For a two sided test this size is $P[|T_{LR}^{mod}| > \text{cv}(a)|H_0]$, and for a one sided test $P[T_{LR}^{mod} > \text{cv}(a)|H_0]$.

evidence to consider one of the two models more plausible than the other[17]. The test is carried out by first computing the value of statistic T_{LR} or T_{LR}^{mod}, then by rejecting model G_π in favor of model F_θ when this value is positive, and conversely by rejecting F_θ in favor of G_π when the value is negative.[18] The probability of committing an error with this symmetric test (accepting model F_θ when it is worse than G_π)[19] is a function of $D = E\left[\ln(f(y|x;\theta_*)/g(y|x;\pi_*))\right]$, since the tested hypothesis is a composite one leaving unspecified the absolute value of the distance D between the distributions of models F_θ and G_π. This risk is maximum for $D = 0$ and decreases with $|D|$. The exact value of this risk can be found by integrating the density function of statistics T_{LR} or T_{LR}^{mod} over its negative values[20].

4.3 Prediction and marginal effects

Prediction with a causal model like `mhurdle` of an observable dependent random variable z refers to defining a function of model covariates x providing, for any admissible value of x, a point estimate $\hat{z}$ of z whose prediction error $e = z - \hat{z}$ is minimized according to some criterion. In the frame of classical or frequentist statistical inference paradigm, an optimal conditional predictor of z, for a given value of covariates x, is defined as the one that minimizes the expectation of the squared conditional prediction error, namely $E[e^2|x]$ called the mean-square error of prediction. Using this criterion leads to select as optimal predictor for z its conditional expectation $E[z|x]$, called the best mean-square error predictor for z.

As defined by identity (2), the dependent observable variable y of `mhurdle` is a mixture of a fqualitative and a quantitative variables whose expectation has no actual meaning, the zero-value of y denoting the numerical value taken by the binary dummy variable d when the observation of latent variable y_2^* is censored. Therefore, for a predictive purpose, one must disentangle the qualitative component of y from its purely quantitative component. This can be done by solving the prediction problem of y in the framework of the two-part respecification of `mhurdle` models presented in section 3.2.

The first-part of this model explains the outcome of the qualitative component of y using the binary variable d, whose best mean-square error predictor is given by:

$$E[d|x] = 0 \times \mathrm{P}(y = 0) + 1 \times \mathrm{P}(y > 0) = \mathrm{P}(y > 0), \tag{47}$$

with $\mathrm{P}(y > 0)$ specified by formulas (13), (14), (15), (20), (22) and (24).

The second-part of the model explains the outcome of the quantitative component of y taking positive real values according to the density function (30). Accordingly, the best mean-square error predictor of $y > 0$ is given by:

$$E[y|y > 0, x] = \int_0^\infty y \frac{f_+(y)}{\mathrm{P}(y > 0)} dy = \frac{E[y|x]}{\mathrm{P}(y > 0)}, \tag{48}$$

with $f_+(y)$ specified by formulas (19), (21), (23) and (25).

For the special cases of transformations $T(\cdot)$ identity and logarithmic focused in this paper, closed analytical forms for $E[y|x]$ can be established[21].

[17] The assumption of a Bayesian probability of one half for the plausibility a priori of both confronted models may not be appropriate in the case of two nested models, where one is a special case of the other, and thus the former less plausible than the latter. For these models, an asymmetric Neyman-Person test where the restricted model represents the null hypothesis and the unrestricted model the alternative hypothesis seems more appropriate, in particular when the model parameters on which restrictions are tested have an economic meaning.

[18] For nested models, Simon's approach always leads to favor the nesting (unrestricted) model to the nested (restricted) one.

[19] There is no need to distinguish between error of the first and second kind, since an error of the first kind for testing H_f against H_g (rejecting model F_θ when it is better than model G_π) is an error of the second kind for testing H_g against H_f (accepting model G_π when it is worse than F_θ), and vice versa.

[20] For the Vuong statistic T_{LR} this distribution for testing H_f against H_g is $N(D, 1)$ with $D \geq 0$. Hence, the risk of rejecting model F_θ when it is better than model G_π is equal to $\mathrm{P}(N(0, 1) < -D) = \Phi(-D)$. This probability decreases from 1/2 to 0 when D increases from 0 to ∞.

[21] Proofs for the analytical forms (49) and (53) are available from the authors.

For the identity transformation $T(y_2^*) = y_2^*$, this analytical closed form for trivariate hurdle models with density function $f_+(y)$ specified by formula (19) is written as:

$$E[y|x] = \frac{\Phi_{123}}{\Phi_3}\beta_2^\top x_2 + \frac{\sigma}{\Phi_3}\left(\phi_2\Phi_{13|2} + \rho_{12}\phi_1\Phi_{23|1} + \rho_{23}\phi_3\Phi_{12|3}\right), \tag{49}$$

where

$$\begin{aligned}
&\Phi_{123} = \Phi\left(\beta_1^\top x_1, \frac{\beta_2^\top x_2}{\sigma}, \beta_3^\top x_3; \rho\right), \quad \phi_i = \phi(\beta_i^\top x_i), \quad i = 1,3, \quad \phi_2 = \phi\left(\frac{\beta_2^\top x_2}{\sigma}\right),\\
&\Phi_{ij|k} = \Phi\left(\frac{\mu_{i|k}}{\sigma_{i|k}}, \frac{\mu_{j|k}}{\sigma_{j|k}}; \rho_{ij|k}\right), \quad i \neq j \neq k, \quad \mu_{i|k} = \beta_i^\top x_i - \rho_{ik}\beta_k^\top x_k, \quad k = 1,3,\\
&\mu_{i|2} = \beta_i^\top x_i - \rho_{i2}\frac{\beta_2^\top x_2}{\sigma}, \quad \sigma_{i|k} = \sqrt{1-\rho_{ik}^2}, \quad \rho_{ij|k} = \frac{\rho_{ij} - \rho_{ik}\rho_{jk}}{\sigma_{i|k}\sigma_{j|k}}.
\end{aligned}$$

The closed forms for bivariate and univariate hurdle models with density function $f_+(y)$ specified by formulas (21), (23) and (25), are written as:

$$E[y|x] = \frac{\Phi_{23}}{\Phi_3}\beta_2^\top x_2 + \frac{\sigma}{\Phi_3}\left(\phi_2\Phi_{3|2} + \rho_{23}\phi_3\Phi_{2|3}\right), \tag{50}$$

$$E[y|x] = \Phi_{12}\beta_2^\top x_2 + \sigma\left(\phi_2\Phi_{1|2} + \rho_{12}\phi_1\Phi_{2|1}\right), \tag{51}$$

$$E[y|x] = \Phi_2\beta_2^\top x_2 + \sigma\phi_2, \tag{52}$$

respectively.

Similarly, closed analytical forms for $E[y|x]$ can be established in the case of the logarithmic transformation $T(y_2^*) = \ln(y_2^* + \alpha)$. For trivariate, bivariate and univariate hurdle models with density function $f_+(y)$ specified by formulas (19), (21), (23) and (25) these closed forms are written, respectively:

$$\mathrm{E}[y|x] = \frac{e^{\beta_2^\top x_2 + \frac{1}{2}\sigma^2}}{\Phi_3}\Phi\left(\beta_1^\top x_1 + \rho_{12}\sigma, \frac{\beta_2^\top x_2 - \ln\alpha}{\sigma} + \sigma, \beta_3^\top x_3 + \rho_{23}\sigma; \rho\right) - \frac{\alpha}{\Phi_3}\Phi_{123}, \tag{53}$$

with $\Phi_{123} = \Phi(\beta_1^\top x_1, (\beta_2^\top x_2 - \ln\alpha)/\sigma, \beta_3^\top x_3; \rho)$,

$$\mathrm{E}[y|x] = \frac{e^{\beta_2^\top x_2 + \frac{1}{2}\sigma^2}}{\Phi_3}\Phi\left(\frac{\beta_2^\top x_2 - \ln\alpha}{\sigma} + \sigma, \beta_3^\top x_3 + \rho_{23}\sigma; \rho_{23}\right) - \frac{\alpha}{\Phi_3}\Phi_{23}, \tag{54}$$

$$\mathrm{E}[y|x] = \frac{e^{\beta_2^\top x_2 + \frac{1}{2}\sigma^2}}{\Phi_3}\Phi\left(\beta_1^\top x_1 + \rho_{12}\sigma, \frac{\beta_2^\top x_2 - \ln\alpha}{\sigma} + \sigma; \rho_{12}\right) - \frac{\alpha}{\Phi_3}\Phi_{12}, \tag{55}$$

$$\mathrm{E}[y|x] = e^{\beta_2^\top x_2 + \frac{1}{2}\sigma^2}\Phi\left(\frac{\beta_2^\top x_2 - \ln\alpha}{\sigma} + \sigma\right) - \frac{\alpha}{\Phi_3}\Phi_2. \tag{56}$$

The marginal effects we consider refer to the change in the value of predictors (47) and (48) generated by a ceteris paribus increase of a model covariate, the value of the other model covariates being kept unchanged, as in a laboratory experiment.

The definition of the marginal effect of a `mhurdle` covariate x_0 on a dependent variable of the model, depends on the scale of measure of x_0. When x_0 is a continuous quantitative variable measured either on an interval

or on a ratio scale,[22] the marginal effect of x_0 on predictor (47) or (48) is the first derivative of the predictor with respect to x_0, namely the rate of change in the value of the predictor generated by a ceteris paribus infinitesimal increase of x_0. Similarly, if x_0 is a count variable measuring the size of a population of individuals or objects which can only vary by integers, the marginal effect will be defined as the finite change in the predictor induced by a ceteris paribus unit increase of x_0.

When x_0 is a qualitative explanatory variable that can be in one and only one of $K \geq 2$ different mutually exclusive and exhaustive situations called attributes, it is usual to code it as a column vector of K dummy variables $x_0 = [d_k]_{k=1,...,K}$, taking values $d_h = 1$ and $d_k = 0, \forall k \neq h$ when the observed attribute of x_0 is the one indexed by h. In this case, considering x_0 as the first (vector) element of one of `mhurdle` covariate vectors $x_i, i = 1, ..., 4$, the first term of the corresponding linear combination $\beta_i^\top x_i$ will take the value β_{ih}.

Yet, this dummy coding of a nominal explanatory variable[23] has an annoying drawback: because the K dummy variables are linearly related by the constraint $\sum_{k=1}^{K} d_k = 1$, the parameters $\beta_{ik}, k = 1, ..., K$ are not identified and therefore not estimable when vector β_i does contain an intercept term. To avoid this inconvenience, it is suggested to suppress from a K-attribute dummy coding one of the K dummies, the discarded attribute playing the role of a "reference attribute" with respect to which the impact on $\beta_i^\top x_i$ of one of the $K-1$ remaining attributes is measured. Indeed, substituting in $\sum_{k=1}^{K} \beta_{ik} d_k$ dummy variable d_K of reference attribute K by $d_K = 1 - \sum_{k=1}^{K} d_k$, we obtain $\sum_{k=1}^{K-1} (\beta_{ik} - \beta_{iK}) d_k + \beta_{iK}$, showing that the impact on $\beta_i^\top x_i$ of attribute k is measured by parameter $b_k = \beta_{ik} - \beta_{iK}$ called contrast with respect to attribute K, while the impact coefficient β_{iK} is included in vector β_i as an intercept, or incorporated to an already existing intercept of this vector. Consequently, the marginal effect of a change of x_0, from attribute h to attribute k, will be defined as the finite change in the predictor induced by a ceteris paribus change of the first term of $\beta_i^\top x_i$ from contrast b_{ih} to contrast b_{ik}, with obviously $b_K = 0$.

When x_0 is a nominal qualitative variable whose K attributes may be partially or totally ordered according to some meaningful criterion (for example the educational level measured and ranked according to the highest level of diploma awarded), it is possible to use a dummy coding allowing to assess the degree of monotonicity between a ceteris paribus marginal increment of x_0 from one attribute level to the next higher attribute level, and its impact on the dependent variable predictor. To this purpose, this impact generated by the marginal increment of x_0 from its attribute of rank $k-1$ to its attribute of immediately higher rank k must be measured by the following marginal contrasts: $\gamma_k = \beta_{ik} - \beta_{i,k-1}, k = 2, ..., K$. Hence, substituting in $\sum_{k=1}^{K} \beta_{ik} d_k$ parameters β_{ik} by the following solution of the former recurrence equations with respect to $\gamma_k, k = 2, ..., K$, namely $\beta_{i1} + \sum_{h=2}^{k} \gamma_h$ leads to the following recoding of the K-attribute dummy coding in terms of marginal contrasts: $\beta_{i1} + \sum_{k=2}^{K} \gamma_k \delta_k$ with $\delta_k = \sum_{h=k}^{K} d_h$. Consequently, the marginal effect of a change of x_0, from pre-ordered attributes $k-1$ to k, will be defined as the finite change in the predictor induced by a ceteris paribus change of the first term of $\beta_i^\top x_i$ from marginal contrast γ_{k-1} to marginal contrast γ_k, with $\gamma_1 = 0$.

So far, punctual prediction and marginal effects for a `mhurdle` independent variable have been tackled by assuming known the value of the model parameters. However, to be operational these formulas must be quantified by means of an estimate of these theoretical parameters. Using the maximum likelihood estimate $\hat{\theta}$ of model parameters θ presented in section 3.1, provide an implied maximum likelihood estimator for a column-vector $h(\theta)$ of best mean-square error predictors (47), (48), and/or marginal effects of these predictors, namely $h(\hat{\theta})$. This estimator is, as $\hat{\theta}$, a consistent and asymptotically normal estimator for $h(\theta)$ whose asymptotic distribution can be derived from a linearization of $h(\hat{\theta})$ around θ called the "delta method", which leads to the following asymptotic variance-covariance matrix of $h(\hat{\theta})$: $(1/n)(\partial h/\partial \theta^\top) I_A(\theta)^{-1} (\partial h/\partial \theta^\top)^\top$.

[22] A variable is said to be quantitative when the distance between two states of the variable can be measured, as for a physical magnitude, using the usual notion of distance between real numbers. This condition is fulfilled when a unit of measure, namely the distance between two particular states of the variable, can be used as a standard to express the distance between any pair of states as multiples of the unit of measure. This scale of measurement is said to be a cardinal or interval scale if the origin of the scale, numerically coded by zero, is arbitrary; it is said to be a measure or ratio scale if the origin of the scale is unique and not arbitrary; finally, it is said to be a count scale if its origin and unit of measure are both unique and not arbitrary.

[23] A qualitative variable is said to be nominal if its "scale of measure" allows to determine whether two observed states are equal or different; it is said to be partially ordered if some of its attributes may also be ordered; finally, it is said to be ordered if all its attributes may be ordered.

A last comment deserves to be devoted to how to present the estimates of marginal effects which are representative for the available sample. To this purpose it is customary to provide either the marginal effects computed at the sample average value of model covariates, or the average value of the sample of marginal effects computed at the model covariates values of the individuals of the sample. The first solution must be avoided when same model covariates are qualitative, because the sample average value of a dummy coded qualitative variable is no more a qualitative variable but a vector of proportions without a meaningful interpretation for a representative individual. Regarding the second solution, it provides an informative measure of location of the marginal distribution of a marginal effect, but that is not enough to characterize, by itself, the a priori unknown profileq of this distribution. Accordingly, it should be supplemented by other indicators allowing to characterize the typical shape of the distribution, in particular the bounds of the distribution range, the standard deviation and the three quartiles.

5 Software rationale

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. They are reported on an annual basis, in thousands of USD, and divided by the number of consumption units[24]. 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 households.

```
library("mhurdle")
data("Interview", package = "mhurdle")
```

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),

Among the numerous goods available in this data set, we choose the "fees and admissions" good (denoted `shows` in the data set), which is of particulary interest because the three competing hurdles are a priori relevant to explain censored observations.

```
round(c(mean = mean(Interview$shows),
        "% of 0" = mean(Interview$shows == 0),
        "mean pos" = mean(Interview$shows) / mean(Interview$shows > 0)),
      2)
```

```
##     mean   % of 0 mean pos
##     0.12     0.69     0.38
```

The consumption of this good is highly censored, as only less than a third of the households in the sample actually purchased it during the survey. The average consumption is \$120 a year, or \$380 for those who consumme.

[24] 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

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, that can be provided by the user, which may be necessary for complicated and highly parametric models, -`dist`: this argument indicates the functional form of the desired consumption equation, which may be either log-normal `"ln"` (the default), normal `"n"`, Box-Cox `"bc"`, or inverse Hyperbolic Sine `"ihs"`,

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

-`h2`: this boolean argument indicates whether the second hurdle is effective or not,

-`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`.

For sake of clarity, we'll denote the estimated models using à 5 digits name. The first one is a capital letter indicating the distribution used (either `N` for a normal distribution and `L` for a log-normal distribution), the second, third and fourth ones are 1 if the first, second and third hurdles are present, 0 if they are not and the fifth digit is either `D` or `I` if a dependent or an independent model is estimated.

To illustrate the use of `mhurdle`, we start with single equation tobit models, using a normal and a log-normal specification.

```
N010I <- mhurdle(shows ~ 0 | linc + smsa + age +
                     educ + size, data = Interview,
                 h2 = TRUE, dist = "n", method = "bhhh")
L010I <- update(N010I, dist = "ln")
```

We then consider the two other potential hurdles, namely the selection process and the infrequency of purchase. We only keep `linc` as a covariate for the desired consumption and the other covariates (`smsa`, `age`, `educ` and `size`) are used either in the first (selection model) or the third part (infrequency of purchase model) of the formula. Note that the `dist` argument is not provided (so that the default log-normal distribution is chosen) and that the lack of ressource hurdle is either maintained or removed by setting the `h2` argument respectively to `TRUE` or `FALSE`

```
L100D <- mhurdle(shows ~ smsa + age + educ + size |
                     linc, data = Interview,
                 h2 = FALSE, dist = "ln", corr = TRUE, method = "bhhh",
                 finalHessian = TRUE)
L100D2 <- update(L100D, start = coef(L100D), robust = FALSE)
L110D <- update(L100D, h2 = TRUE)
L110D2 <- update(L110D, start = coef(L110D), robust = FALSE)
```

```
L001D <- mhurdle(shows ~ 0 | linc | smsa + age +
                        educ + size, data = Interview,
                    h2 = FALSE, corr = TRUE, method = "bhhh",
                  finalHessian = TRUE)
L001D2 <- update(L001D, start = coef(L001D), robust = FALSE)
L011D <- update(L001D, h2 = TRUE)
L011D2 <- update(L011D, start = coef(L011D), robust = FALSE)
```

Finally, we estimate three equations models. The 4 socio-economic covariates that were previously either included in the first or the third parts of the formula are now split on these two equations. More precisely, we assume that `educ` and `size` are the determinants of the selection process and that `smsa` and `size` explain the frequency of purchase. The first model is a double-hurdle model ; selection and purchasing mechanism are effective and note that in this case, the `h2` argument is set to `FALSE`). The last two ones triple-hurdle models, with correlated and uncorrelated errors:

```
L101D <- mhurdle(shows ~ educ + size | linc |
                         smsa + age, data = Interview,
                    h2 = FALSE, method = "bhhh", corr = TRUE,
                  finalHessian = TRUE)
L101D2 <- update(L101D, start = coef(L101D), robust = FALSE)
L111D <- update(L101D, h2 = TRUE)
L111D2 <- update(L111D, start = coef(L111D), robust = FALSE)
L111I <- update(L111D, corr = FALSE)
L111I2 <- update(L111I, start = coef(L111I), robust = FALSE)
```

The results are presented in table **??**.

5.2 Tests

The position parameter (μ in table **??** enables to test directly the hypothesis that the second hurdle is operative, using a Wald-like test. For all the models, the hypothesis that the coefficient is equal to 0 is highly rejected. Note also that the only model which impose that the second hurdle is not operative (denoted `L101D` in the table) has the worst value of log-likelihood.

The coefficient of correlation is highly significant for the double-hurdle selection model, but not for the double-hurdle selection model. For the most general three equations model, only the coefficient of correlation for the first two equations is significant. Formal tests for this two nested models can be performed using a traditional log-likelihood ratio test:

```
library("lmtest")
lrtest(L111D, L111I)
```

or a Vuong test, which don't make the hypothesis that one of the competing model is the "true" model:

```
vuongtest(L111D, L111I, type = "nested")
```

Both tests don't reject the independence hypothesis at the 5% level, especially the Vuong test, the p-values being respectively equal to 0.076 and 0.2154219.

For non-nested models, one can use Vuong test. For example, for the double-hurdles selection and ptobit models, the test is performed using:

```
vuongtest(L110D, L011D)
```

The statistic is -2.173. Its negative sign indicate that the p-tobit model better fits the data compared to the selection model. The difference is significant at the 5% level, as the (two-ways) p-values equal 0.015, which leads to the conclusion that the ptobit model is significantly better than the selection model.

The parcimony principal leads to the selection of the ptobit double-hurdle model, as it is not significantly worst than the more general three-hurdle model.

Bibliography

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

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

Box, G. E. P., and D. R. Cox. 1964. "An Analysis of Transformations." *Journal of the Royal Statistical Society. Series B (Methodological)* 26 (2): 211–52.

Cameron, A. Colin, and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge University Press.

Chaze, Jean-Paul. 2005. "Assessing Household Health Expenditure with Box-Cox Censoring Models." *Health Economics* 14: 893–907.

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

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

Estrella, Arturo. 1998. "A New Measure of Fit for Equations with Dichotomous Dependent Variables." *Journal of Business & Economic Statistics* 16 (2): 198–205.

Henningsen, Arne. 2013. *CensReg: Censored Regression (Tobit) Models*. http://CRAN.R-project.org/package=censReg.

Hoareau, Stéphane. 2009. "Modélisation économétrique Des Dépenses de Consommation Censurées." PhD thesis, Faculty of Law; Economics, University of La Réunion.

Johnson, N. L. 1949. "Systems of Frequency Curves Generated by Methods of Translation." *Biometrika* 36 (1-2): 149–76.

Kleiber, Christian, and Achim Zeileis. 2008. *Applied Econometrics with R*. New York: Springer-Verlag. http://CRAN.R-project.org/package=AER.

Lankford, R. H., and J. H. Wyckoff. 1991. "Modeling Charitable Giving Using a Box-Cox Standard Tobit Model." *Review of Economics and Statistics* 73(3): 460–70.

Pudney, Stephen. 1989. *Modelling Individual Choice. The Econometrics of Corners, Kinks and Holes*. Oxford; New York: Basil Blackwell.

Shi, Xiaoxia. 2015. "A Nondegenerate Vuong Test." *Quantitative Economics*, 85–121.

Simon, Herbert. 1943. "Symmetric Tests of the Hypothesis That the Mean of One Normal Population Exceeds That of Another." *Annals of Mathematical Statistics* 14 (2): 149–54.

Smith, Murray. 2002. "Handbook of Applied Econometrics and Statistical Inference." In, chapter 25. New-York: Marcel Dekker.

Therneau, Terry. 2013. *Survival: Survival Analysis, Including Penalised Likelihood.* http://CRAN.R-project.org/package=survival.

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

Toomet, Ott, and Arne Henningsen. 2008. "Sample Selection Models in R: Package sampleSelection." *Journal of Statistical Software* 27 (7). http://www.jstatsoft.org/v27/i07/,%20http://CRAN.R-project.org/package=sampleSelection.

Vuong, Quang H. 1989. "Likelihood Ratio Tests for Selection and Non-Nested Hypotheses." *Econometrica* 57 (2): 397–33.

Zeileis, Achim, Christian Kleiber, and Simon Jackman. 2008. "Regression Models for Count Data in R." *Journal of Statistical Software* 27 (8): 1–25. http://www.jstatsoft.org/v27/i08.