# A Two Recursive Equation Model to Correct for Endogeneity in Latent Class Binary Probit Models

Mauricio Sarrias*
Universidad de Talca
mauricio.sarrias@utalca.cl

June 11, 2021

**Abstract**

This article proposes a two recursive equation model to correct for endogeneity in latent class Probit models. Concretely, it is assumed that there exists an endogenous and continuous variable defined as a predictor, while unobserved heterogeneity is conceptualized as a vector of parameters that varies across individuals following a discrete distribution. A Maximum Likelihood Estimator is provided to estimate the model parameters based on normally distributed random terms and a free code in R software is provided to carry out the estimation procedure. A small Monte Carlo experiment is carried out to analyze the properties of the estimator. Finally, the estimator is applied to analyzed the heterogeneous effects of weight on mental well-being.

*Key words:* instrumental variable; probit model; latent class; unobserved heterogeneity; MLE.

*Facultad de Economía y Negocios, Universidad de Talca, Talca, Chile. I am obliged to three anonymous referees. I also express my thanks to FONDECYT, Grant 1200522 for full financial support.

# 1 Introduction

The Latent Class (LC) model or Mixture Model has been an important econometric tool for disentangling unobserved heterogeneity in the sample. Under this approach, the coefficients for each explanatory variable are allowed to vary across the sample assuming that there exists segments or classes of individuals that have different parameters, but within each class the coefficients are homogeneous.[1] Disentangling the coefficients for different segments of individuals in the sample not only adds more realism to the modeling approach by detecting important features and insights about heterogeneous relationships, but also increases the explanatory power and reduces the bias of the estimated coefficients (Hess, 2014) if the true data generating process incorporates unobserved heterogeneity. Given this flexibility, LC models have been applied for modeling unobserved heterogeneity on the determinants of healthcare demand (Deb and Trivedi, 2002; Atella et al., 2004; d'Uva and Jones, 2009), subjective well-being measures (Clark et al., 2005; Palomino and Sarrias, 2019) travel demand (Gopimatj, 1997; Greene and Hensher, 2003), R&D and patent (Wang et al., 1998), the effect of direct marketing on purchases (Wedel and DeSarbo, 1995), and criminology (Nagin and Land, 1993) to mention a few.

Although modeling unobserved heterogeneity using a LC approach is an expanding field both in theoretical and applied grounds, another important empirical concern is the potential endogeneity due to omitted variables, measurement error, simultaneity, and/or self-selection: if the error term is correlated with the explanatory variables (or the treatment), then the coefficients representing the degree of heterogeneity across classes will be inconsistent even using large samples. In such a case, the instrumental variables method can be a solution to eliminate the potential endogeneity of the treatment variable, either continuous or discrete. For example, some researchers have already proposed models that incorporate both econometric issues for linear models. Heckman et al. (2006) examine the properties of instrumental variables applied to models with essential heterogeneity, that is, when the treatment effect varies with the treatment due to unobserved confounders. In this scenario, individuals might respond differently to the same treatment yielding substantial differences in the average treatment effect. Similarly, Florens et al. (2008) proposes a control function approach to identify the average treatment effect and the effect of the treatment on the treated in models with a continuous endogenous regressor whose impact is heterogeneous, whereas Moffitt (2008) proposes a nonparametric method of estimating marginal treatment effects in heterogeneous populations when the treatment is a dummy variable by assuming that the outcomes are a nonlinear function of participation probabilities.

More in line with this article, Bhat et al. (2014) provide a methodological innovation that allows to control for potential endogenous effects using instrumental variables techniques (for different type of endogenous variables, such as continuous and count variables) and parametric distribution for multi-dimensional choice systems using a

[1] For a deeper review on LC models or mixture models see for example Titterington (1990); Wedel and DeSarbo (1994); McLachlan and Peel (2004); Greene (2004); Greene and Hensher (2010).

Maximum Approximate Composite Marginal Likelihood approach for the estimation of the parameters (Bhat, 2011). In particular, they applied this joint model to analyze household-level decision on residential location, motorized vehicle ownership, and activity-travel patterns and found that these choice dimensions are inter-related, both through direct observed structural relationships and trough correlation across the error term, and that the endogeneity coming from self-selection might overestimate the treatment effects.

This article proposes a fully parametric method to deal with both unobserved heterogeneity at the individual level (represented by different coefficients for each individuals) and the potential endogeneity of a continuous variable in models where the dependent variable is dichotomous. Specifically, an instrumental variable latent class approach for the binary Probit (IVLC-Probit) model is derived by incorporating the characteristics of the traditional recursive two-equation models proposed by Heckman (1978), Amemiya (1978, 1979) and Rivers and Vuong (1988) to the LC-Probit model (see Greene, 2004, for a review of the LC-Probit model).[2] It is assumed that there exists an endogenous and continuous regressor defined as a predictor, which is partly determined by predetermined factors and instrumental variables, while unobserved heterogeneity is conceptualized as a vector of parameters that varies across individuals following a discrete distribution. Thus, individual heterogeneity is accommodated by making use of a discrete and fixed number, say $Q$, of separate classes (support points) with different coefficients in each class, while the degree of endogeneity might vary across classes. The proposed IVLC-Probit model extends the traditional IV-Probit by including unobserved heterogeneity using latent classes, and the LC-Probit by including the possibility that a continuous variable is endogenous in some or all classes. Therefore, the treatment, the strength of instrument and the degree of endogeneity are allowed to vary across groups of individuals. A conditional Maximum Likelihood Estimator (MLE) is provided to estimate the model parameters based on normally distributed random terms and a free code in R software is provided in Annex A to carry out the estimation procedure.[3] To assess the properties of the proposed estimator under small and large sample a small Monte Carlo experiment is performed. Finally, the advantages and disadvantages of the proposed method are analyzed through an empirical example where the potential heterogeneous effect of weight on mental well-being is analyzed.

The remainder of the paper is organized as follows. Section 2 presents the IVLC-Probit model along with the assumptions and Maximum Likelihood Estimator. Some issues regarding estimation, testing and post-estimation measures are provided in Section 3. Section 4 shows the results for the Monte Carlo experiment, while Section 5 provides

[2] One difference between the model proposed in this article and the model proposed by Bhat et al. (2014) is that their methodology allows dealing with multiple independent variables. The approach in this article only allows the modeling of a single dichotomous dependent variable. However, a difference with Bhat et al. (2014)'s formulation is that the proposed IVLC-Probit model also allows the inclusion of individual heterogeneity in the parameters by including latent classes, which in turn, may depend on covariates. In any case, both approaches deal with endogeneity.

[3] Several extensions have been made for the LC approach. For example, Atella et al. (2004) propose a latent class seemingly unrelated Probit model; Jedidi et al. (1993) propose a linear latent class model for censored dependent variables; and Vermunt and Van Dijk (2001) extend the LC for multilevel models.

the empirical application. Finally, Section 6 concludes.

# 2 Instrumental variable latent class probit model

## 2.1 Model and assumptions

To include the instrumental variables approach in a model with unobserved heterogeneity in the coefficients, some characteristics of the traditional IV-Probit model (Heckman, 1978; Amemiya, 1978; Smith and Blundell, 1986; Rivers and Vuong, 1988) are incorporated into the traditional LC-Probit model.[4] Specifically, this paper considers the following two recursive equation model defined by:[5]

$$\begin{aligned} y_{1i}^* &= \mathbf{x}_{1i}'\boldsymbol{\beta}_{1q} + \gamma_q y_{2i} + \epsilon_{iq} = \mathbf{x}_i'\boldsymbol{\beta}_q + \epsilon_{iq}, \quad (1)\\ y_{2i} &= \mathbf{z}_i'\boldsymbol{\delta}_q + \upsilon_{iq}, \quad \text{for } i = 1, ..., N, \quad q = 1, ..., Q, \quad (2) \end{aligned}$$

and

$$\begin{aligned} y_{1i} &= \mathbb{1}\left[y_{1i}^* > 0\right], \quad (3)\\ (\boldsymbol{\beta}_q, \boldsymbol{\delta}_q) &\sim g(\boldsymbol{\lambda}_q), \quad (4) \end{aligned}$$

where $y_{1i}^*$ is a latent (unobserved) dependent variable for individual $i$ and we only observe the binary variable $y_{1i}$ if and only if $\mathbb{1}\left[y_{1i}^* > 0\right]$;[6] $\mathbf{x}_i = (\mathbf{x}_{1i}', y_{2i})'$ is a $K \times 1$ column vector of explanatory variables where $y_{2i}$ is a **continuous and endogenous** variable and $\mathbf{x}_{1i}$ is a set of predetermined (exogenous) variables; $\mathbf{z}_i = (\mathbf{x}_{1i}', \mathbf{x}_{2i}')'$ is a $P \times 1$ vector of predetermined variables where $\mathbf{x}_{2i}$ is the vector of instruments (additional predetermined variables) for $y_2$; $\epsilon_{iq}$ and $\upsilon_{iq}$ are the unobservable error terms.

The vector $\boldsymbol{\beta}_q = (\boldsymbol{\beta}_{1q}', \gamma_q')'$ is a $K \times 1$ vector of true but unknown regression parameters—defined up to some scalar normalization—which varies across $q = 1, ..., Q$ classes or segments of individuals. This implies that the 'treatment effect', $\gamma$, is allowed to vary across classes. For example, the effect of the treatment might be zero for some group of individuals, whereas for some other group it might be positive or negative. It is further assumed that the coefficients in Equation (2) also vary across classes. That is, the instrument(s) can modify the endogenous variable in a heterogeneous way (either in magnitude or sign) in the population.

Let $\boldsymbol{\psi}_q = (\boldsymbol{\beta}_q', \boldsymbol{\delta}_q')'$ be a $(K + P) \times 1$ vector that collects the parameters for Equations (1) and (2). The parameters $\boldsymbol{\psi}_q$ are assumed to vary across individuals following a discrete but unknown distribution given by:

[4] Unlike more traditional articles, this article uses $i$ instead of $n$ as subscript for individuals.

[5] This model is considered as a recursive model since $y_{2i}$—the endogenous variable—appears in the equation for $y_{1i}^*$, but $y_{1i}^*$ does not appear in the equation for $y_{2i}$ (see Maddala, 2002).

[6] $\mathbb{1}$ is an indicator: $\mathbb{1}(A) = 1$ if A is true; 0 otherwise.

$$g(\boldsymbol{\psi}_q|\boldsymbol{\lambda}_q) = \begin{cases} \boldsymbol{\psi}_1 & \text{with probability } \pi_{i1}(\boldsymbol{\lambda}_1) \\ \boldsymbol{\psi}_2 & \text{with probability } \pi_{i2}(\boldsymbol{\lambda}_2) \\ \vdots & \vdots \\ \boldsymbol{\psi}_Q & \text{with probability } \pi_{iQ}(\boldsymbol{\lambda}_Q) \end{cases}, \tag{5}$$

where individual $i$ belongs to class $q$ with probability $\pi_{iq}$, such that $\sum_{q=1}^{Q}\pi_{iq} = 1$ and $\pi_{iq} > 0$; $\boldsymbol{\lambda}_q$ $(q = 1, ..., Q)$ is the $L \times 1$ vector of parameters that describe the class-probability assignment. Thus, individual heterogeneity is accommodated by making use of a discrete and fixed number $Q$ of separate classes of individuals—or support points—with different coefficients in each class. These classes can be though as a classification or segmentation of individuals, which are homogeneous in terms of their coefficients (Hess, 2014).[7]

Similar to Rivers and Vuong (1988), the following assumptions are made:

**Assumption 1 (Distribution)** *Assume that* $(\epsilon_{iq}, \upsilon_{iq})$ *are i.i.d and independent of* $\mathbf{z}_i$ *distributed as a bivariate normal distribution as follows:*

$$(\epsilon_{iq}, \upsilon_{iq})|\mathbf{z}_i \sim \mathrm{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{q,\epsilon} & \rho_q\sigma_{q,\epsilon}\sigma_{q,\upsilon} \\ \rho_q\sigma_{q,\epsilon}\sigma_{q,\upsilon} & \sigma^2_{q,\upsilon} \end{pmatrix}\right].$$

*Since* $y^*_{1i}$ *is latent, it is further assumed that* $\sigma_{q,\epsilon} = 1, \forall q = 1, ..., Q$.

**Assumption 2 (Data)** *The sequence of observed data* $\{y_{1i}, y_{2i}, \mathbf{x}_i, \mathbf{z}_i\}$ *is i.i.d.*

**Assumption 3 (Compact parameter space)** *The vector of parameters of the model* $\boldsymbol{\theta} = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_Q, \boldsymbol{\delta}_1, ..., \boldsymbol{\delta}_Q, \boldsymbol{\lambda}_2, ..., \boldsymbol{\lambda}_Q, \rho_1, ..., \rho_Q, \sigma_{1,\upsilon} ... \sigma_{Q,\upsilon})$ *is known to lie in the interior of a compact convex subset of* $\boldsymbol{\Theta}$.

**Assumption 4 (Rank condition for identification)** *The* $P \times K$ *matrix* $\mathbb{E}(\mathbf{z}_i\mathbf{x}'_i)$ *if full column rank, i.e., its rank equals* $K$.

**Assumption 5 (Identification II)** $\mathbb{E}(\mathbf{x}_i\mathbf{x}'_i)$ *exists and is nonsingular.*

**Assumption 6 (Class-assignment)** *Let* $\mathbf{h}_i$ *be a* $L \times 1$ *i.i.d vector having finite and nonsingular matrix* $\mathbb{E}(\mathbf{h}_i\mathbf{h}'_i)$. *Assume that there exists a latent continuous variable* $F^*_{iq}$ *that determines the class-assignment of individual* $i$ *in class* $q = 1, ..., Q$ *and is given by a linear function of* $\mathbf{h}_i$:

$$F^*_{iq} = \mathbf{h}'_i\boldsymbol{\lambda}_q + \xi_{iq}.$$

*Assuming that* $\xi_{iq}$ *are i.i.d Extreme Value Type I, the probability for individual* $i$ *to belong to a particular class* $q$ *is given by:*

$$\pi_{iq}(\boldsymbol{\lambda}_q) = \frac{\exp\left(\mathbf{h}'_i\boldsymbol{\lambda}_q\right)}{\sum_{c=1}^{Q}\exp\left(\mathbf{h}'_i\boldsymbol{\lambda}_c\right)}, \quad q = 1, ..., Q. \tag{6}$$

[7] As Greene (2004) states, a LC model can also be viewed as arising from a discrete, unobserved sorting of individuals into groups, each of which has its own set of characteristics.

*The parameters for some class are normalized to zero for identification of the probabilities.*

Although these assumptions are standard in recursive models, a brief discussion of the assumptions is needed. Assumption 1 (Distribution) states that, conditional on the class individuals belong to, the error terms follow a bivariate normal distribution, thus the model given by Equations (1)-(4) is fully parametric and can be estimated by Maximum Likelihood. Since $y_{1i}^*$ is latent, assuming that $\sigma_{q,\epsilon} = 1, \forall q = 1, ..., Q$ is needed for the identification of the parameters in Equation (1) (Amemiya, 1978).

The expectation and variance of the error terms in each class under Assumption 1 are:

$$\begin{aligned}\mathbb{E}(\epsilon_{iq}|\mathbf{z}_i) &= 0,\\ \mathbb{E}(\upsilon_{iq}|\mathbf{z}_i) &= 0,\\ \mathrm{Var}(\epsilon_{iq}|\mathbf{z}_i) &= \sigma^2_{q,\epsilon} = 1,\\ \mathrm{Var}(\upsilon_{iq}|\mathbf{z}_i) &= \sigma^2_{q,\upsilon},\end{aligned}$$

$\forall q = 1, ..., Q$. Whereas, the correlation is given by:

$$\mathrm{corr}(\epsilon_{iq}, \upsilon_{iq}|\mathbf{z}_i) = \rho_q, \quad \forall q = 1, ..., Q.$$

It is important to highlight that $\rho_q$ measures the degree of endogeneity of $y_{2i}$ in Equation (1) for each class. Thus, the model is flexible enough to allow different degrees of endogeneity across classes. Note also that if $\boldsymbol{\beta}_q = \boldsymbol{\beta}$, $\boldsymbol{\delta}_q = \boldsymbol{\delta}$, $\sigma^2_{q,\epsilon} = \sigma^2_{\epsilon}$, $\sigma^2_{q,\upsilon} = \sigma^2_{\upsilon}$ and $\rho_q = \rho$ then we have the traditional IV-Probit model. Similarly, if $\epsilon_i$ and $\upsilon_i$ are not correlated, $\rho_q = 0, \forall q = 1, ..., Q$, the model results in the traditional LC-Probit Model.

Assumption 2 (Data) simplifies the maximization of the log likelihood of the joint distribution, which is obtained as the product of the probability for each individual in the sample.

Since the sample log-likelihood is not globally concave, compactness (Assumption 3) is essential for existence and consistency of MLE (see Newey and McFadden, 1994; Chen et al., 2017, 2009). For example, the sample log-likelihood becomes unbounded as $\sigma_{q,\upsilon}$ goes to zero or $|\rho_q|$ tends toward 1.[8] Section 3.2 explains some transformations that can be applied to bound some of the parameters away from their boundary.

The rank condition for identification (Assumption 4) requires that the number of excluded exogenous variables be at least as great as the number of included endogenous variables. Assumption 5 (Identification II) is required for the existence of the log-likelihood function (Newey and McFadden, 1994).

Since the discrete mixing distribution, $\pi_{iq}(\boldsymbol{\lambda}_q)$ in Equation (5), is unknown in advance to the analyst, Assumption 6 (Class-assignment) states the functional form to model such probability to assign each individual to one an only one class. Using the semiparametric multinomial Logit formula for the class-assignment probability has become standard in the

[8] This can be also observed in the gradient derived in Annex B. If $\sigma_{q,\upsilon} = 0$ or $\rho_q = \pm 1$, then the score is undefined for the parameters in class $q$.

applied literature (see for example Greene and Hensher, 2003; Shen, 2009; Hess, 2014) and ensures that the probabilities sum to one. Note also that $\mathbf{h}_i$ in Equation (6) is a set of sociodemographic variables that determines the assignment of each individual in a given class. If $\mathbf{h}_i$ is omitted, the class probabilities simply become:

$$\pi_q = \frac{\exp\left(\lambda_q\right)}{\sum_{c=1}^{Q}\exp\left(\lambda_c\right)}, q = 1, ..., Q, \tag{7}$$

such that the class allocation probabilities are constant across individuals, $\pi_{iq} = \pi_q, \forall i = 1, ..., N$. For identification of the probabilities, in this paper I set $\boldsymbol{\lambda}_1 = \mathbf{0}$.

## 2.2 Maximum Likelihood Estimator

Given the class that individual $i$ belongs to, the joint distribution of $(y_{1i}, y_{2i})$, conditional on $\mathbf{z}_i$, is given by $f_q(y_{1i}, y_{2i}|\mathbf{z}_i)$, which can be written as $f_q(y_{1i}|y_{2i}, \mathbf{z}_i)f_q(y_{2i}|\mathbf{z}_i)$. To derive the joint distribution, note that under the normality of $(\epsilon_{iq}, \upsilon_{iq})$ and the normalization of the error in the latent equation (Assumption 1), we can write $\epsilon_{iq}|\upsilon_{iq} = \left[(\rho_q \cdot \sigma_{q,\upsilon})/\sigma^2_{q,\upsilon}\right]\upsilon_{iq} + \eta_{iq}$, where $\eta_{iq} \sim \mathrm{N}(0, 1-\rho_q^2)$ (Smith and Blundell, 1986; Wooldridge, 2010). Inserting this into Equation (1) generates the conditional model:

$$y^*_{1i}|y_{2i} = \mathbf{x}'_i\boldsymbol{\beta}_q + (\rho_q/\sigma_{q,\upsilon})\,\upsilon_{iq} + \eta_{iq}.$$

Since $\upsilon_{iq} = y_{2i} - \mathbf{z}'_i\boldsymbol{\delta}_q$ (from Equation 2) and $y_{1i} = \mathbb{1}\left[y^*_{1i} > 0\right]$, then the probability of individual $i$ conditional on $y_{2i}$ and $\mathbf{z}_i$ is:

$$\begin{aligned}\Pr(y_{1i} = 1|y_{2i}, \mathbf{z}_i) &= \Pr(y^*_{1i} > 0|y_{2i}, \mathbf{z}_i), \\ &= \Phi\left[\frac{\mathbf{x}'_i\boldsymbol{\beta}_q + \frac{\rho_q}{\sigma_{q,\upsilon}}(y_{2i} - \mathbf{z}'_i\boldsymbol{\delta}_q)}{\sqrt{1-\rho_q^2}}\right],\end{aligned} \tag{8}$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution. Using the fact that the normal distribution is symmetric, then the conditional density of $y_{1i}$ given $(y_{2i}, \mathbf{z}_i)$ can be written as:

$$f_q(y_{1i}|y_{2i}, \mathbf{z}_i) = \Phi\left[q_i \cdot \left(\frac{\mathbf{x}'_i\boldsymbol{\beta}_q + \frac{\rho_q}{\sigma_{q,\upsilon}}(y_{2i} - \mathbf{z}'_i\boldsymbol{\delta}_q)}{\sqrt{1-\rho_q^2}}\right)\right], \tag{9}$$

where $q_i = 2y_{1i} - 1$ (Greene, 2003). Since $y_{2i}|\mathbf{z}_i \sim \mathrm{N}(\mathbf{z}'_i\boldsymbol{\delta}_q, \sigma^2_{q,\upsilon})$ under Assumption 1, the conditional (on $q$) marginal distribution is:

$$f_q(y_{2i}|\mathbf{z}_i) = \frac{1}{\sigma_{q,\upsilon}}\phi\left(\frac{y_{2i} - \mathbf{z}'_i\boldsymbol{\delta}_q}{\sigma_{q,\upsilon}}\right), \tag{10}$$

where $\phi(\cdot)$ is the standard normal density function. Using (9) and (10), the unconditional (weighted) probability is:

$$\begin{aligned} f(y_{1i}, y_{2i}|\mathbf{z}_i; \boldsymbol{\theta}) &= \sum_{q=1}^{Q} \pi_{iq}(\mathbf{h}_i) \cdot f_q(y_{1i}, y_{2i}|\mathbf{z}_i), \\ &= \sum_{q=1}^{Q} \pi_{iq}(\mathbf{h}_i) \cdot f_q(y_{1i}|y_{2i}, \mathbf{z}_i) \cdot f_q(y_{2i}|\mathbf{z}_i), \\ &= \sum_{q=1}^{Q} \frac{\exp(\mathbf{h}_i'\boldsymbol{\lambda}_q)}{\sum_{c=1}^{Q} \exp(\mathbf{h}_i'\boldsymbol{\lambda}_c)} \Phi\left[q_i \cdot \left(\frac{\mathbf{x}_i'\boldsymbol{\beta}_q + \frac{\rho_q}{\sigma_{q,\upsilon}}(y_{2i} - \mathbf{z}_i'\boldsymbol{\delta}_q)}{\sqrt{1-\rho_q^2}}\right)\right] \frac{1}{\sigma_{q,\upsilon}} \phi\left(\frac{y_{2i} - \mathbf{z}_i'\boldsymbol{\delta}_q}{\sigma_{q,\upsilon}}\right). \end{aligned} \tag{11}$$

In general $f(y_{1i}, y_{2i}|\mathbf{z}_i; \boldsymbol{\theta})$ is called the finite *mixture* density and $\pi_{iq}(\mathbf{h}_i)$ is called the *mixing* distribution or mixing weights (Titterington, 1990; McLachlan and Peel, 2004).

The MLE is a value of the parameter vector that maximizes the log-likelihood function:

$$\widehat{\boldsymbol{\theta}} \equiv \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max} \sum_{i=1}^{N} \ln f(y_{1i}, y_{2i}|\mathbf{z}_i; \boldsymbol{\theta}),$$

where $\boldsymbol{\Theta}$ denotes the parameter space in which the parameter vector $\boldsymbol{\theta}$ lies.[9]

Under the assumptions made, the conditions of Theorem 2.5 in Newey and McFadden (1994) are met, so that $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$ as $n \to \infty$. Furthermore, $\widehat{\boldsymbol{\theta}}$ is asymptotically efficient and the variance-covariance matrix can be estimated using the estimated Hessian, the estimated expected Hessian, or the outer product of the score (Wooldridge, 2010).

# 3 Empirical and numerical issues

## 3.1 Marginal effects and posterior membership probability

The estimated coefficients of the IVLC-Probit model, as in any nonlinear model, do not have a direct interpretation. Thus, the marginal effects are required to give a quantitative meaning to the estimates.

Using Equation (8), the partial effect of $x_{ik}$ on the probability of a positive outcome, $y_i = 1$, can be computed for each class $q = 1, ..., Q$ as follows:

$$ME_{i,q}^{x_k} = \frac{\partial \Pr(y_i = 1|y_2, \mathbf{z}_i)}{\partial x_{ik}} = \phi\left[\frac{\mathbf{x}_i'\boldsymbol{\beta}_q + \gamma_q y_{i2} + \frac{\rho_q}{\sigma_{q,\upsilon}}(y_{2i} - \mathbf{z}_i'\boldsymbol{\delta}_q)}{\sqrt{1-\rho_q^2}}\right]\left[\frac{\beta_{qk} - \frac{\rho_q}{\sigma_{q,\upsilon}}\delta_{kq}}{\sqrt{1-\rho_q^2}}\right], \tag{12}$$

where $\phi(\cdot)$ is the pdf function of the standard normal distribution. The marginal effect for the endogenous variable $y_{i2}$ can be computed as:

[9] This estimation procedure is also known as the Latent-Variable approach (Walker et al., 2007), and according to Guevara (2015), it can be classified among the Full Information Maximum Likelihood (FIML) methods to address endogeneity.

$$ME_{i,q}^{y_2} = \frac{\partial \Pr(y_i = 1|y_2, \mathbf{z}_i)}{\partial y_{i2}} = \phi\left[\frac{\mathbf{x}_i'\boldsymbol{\beta}_q + \gamma_q y_{i2} + \frac{\rho_q}{\sigma_{q,\upsilon}}(y_{2i} - \mathbf{z}_i'\boldsymbol{\delta}_q)}{\sqrt{1-\rho_q^2}}\right]\left[\frac{\gamma_q + \frac{\rho_q}{\sigma_{q,\upsilon}}}{\sqrt{1-\rho_q^2}}\right]. \tag{13}$$

Both Equations (12) and (13) are similar to those for the IV-Probit model (see Skeels and Taylor, 2015), but they are now allowed to vary across classes. Note also that both $ME_{i,q}^{x_k}$ and $ME_{i,q}^{y_2}$ depend on the values of the variables. However, they can be computed using the estimated coefficients at the sample mean of the data or evaluate the marginal effects at every observation and use the sample average of the individual marginal effects. As noted by Greene (2003), both procedures produces similar values in large samples. But, in small samples this is not necessary true, favoring in general the average of the ME.

The marginal effects at the means of the variables can be computed as follows:

$$\begin{aligned} \widehat{ME}_q^{x_k} &= \phi\left[\frac{\bar{\mathbf{x}}'\widehat{\boldsymbol{\beta}}_q + \widehat{\gamma}_q\bar{y}_2 + \frac{\widehat{\rho}_q}{\widehat{\sigma}_{q,\upsilon}}(\bar{y}_2 - \bar{\mathbf{z}}'\widehat{\boldsymbol{\delta}}_q)}{\sqrt{1-\widehat{\rho}_q^2}}\right]\left[\frac{\widehat{\beta}_{qk} - \frac{\widehat{\rho}_q}{\widehat{\sigma}_{q,\upsilon}}\widehat{\delta}_{kq}}{\sqrt{1-\widehat{\rho}_q^2}}\right], \\ \widehat{ME}_q^{y_2} &= \phi\left[\frac{\bar{\mathbf{x}}'\widehat{\boldsymbol{\beta}}_q + \widehat{\gamma}_q\bar{y}_2 + \frac{\widehat{\rho}_q}{\widehat{\sigma}_{q,\upsilon}}(\bar{y}_2 - \bar{\mathbf{z}}'\widehat{\boldsymbol{\delta}}_q)}{\sqrt{1-\widehat{\rho}_q^2}}\right]\left[\frac{\widehat{\gamma}_q + \frac{\widehat{\rho}_q}{\widehat{\sigma}_{q,\upsilon}}}{\sqrt{1-\widehat{\rho}_q^2}}\right]. \end{aligned} \tag{14}$$

The average ME (AME) can be computed using the following expressions:

$$\begin{aligned} \widehat{AME}_q^{x_k} &= \frac{1}{n}\sum_{i=1}^n \phi\left[\frac{\mathbf{x}_i'\widehat{\boldsymbol{\beta}}_q + \widehat{\gamma}_q y_{2i} + \frac{\widehat{\rho}_q}{\widehat{\sigma}_{q,\upsilon}}(y_{2i} - \mathbf{z}_i'\widehat{\boldsymbol{\delta}}_q)}{\sqrt{1-\widehat{\rho}_q^2}}\right]\left[\frac{\widehat{\beta}_{qk} - \frac{\widehat{\rho}_q}{\widehat{\sigma}_{q,\upsilon}}\widehat{\delta}_{kq}}{\sqrt{1-\widehat{\rho}_q^2}}\right], \\ \widehat{AME}_q^{y_2} &= \frac{1}{n}\sum_{i=1}^n \phi\left[\frac{\mathbf{x}_i'\widehat{\boldsymbol{\beta}}_q + \widehat{\gamma}_q y_{i2} + \frac{\widehat{\rho}_q}{\widehat{\sigma}_{q,\upsilon}}(y_{2i} - \mathbf{z}_i'\widehat{\boldsymbol{\delta}}_q)}{\sqrt{1-\widehat{\rho}_q^2}}\right]\left[\frac{\widehat{\gamma}_q + \frac{\widehat{\rho}_q}{\widehat{\sigma}_{q,\upsilon}}}{\sqrt{1-\widehat{\rho}_q^2}}\right]. \end{aligned} \tag{15}$$

Using the individual probability of membership to a class $q$ defined in Equation (6), one can compute the weighted average of the marginal effects (WAME) by taking the expectation over the $q$ classes. Let $ME(\mathbf{x}^0, \boldsymbol{\theta})_q$ be the marginal effect for either $x_k$ or $y_2$ evaluated at data point $\mathbf{x}^0$ (either at the sample mean or any other statistic), then the (unconditional) weighted average marginal effect is given by:

$$\mathbb{E}\left[ME(\mathbf{x}^0, \boldsymbol{\theta})_q\right] = \int ME(\mathbf{x}^0, \boldsymbol{\theta})_q g(\boldsymbol{\psi}_q|\boldsymbol{\lambda}_q)d\boldsymbol{\psi}_q,$$

where $g(\boldsymbol{\psi}_q|\boldsymbol{\lambda}_q)$ is the unconditional distribution of parameters in the population (Equation (5)). The WAME can be estimated as follows:

$$\widehat{\mathbb{E}}\left[ME(\mathbf{x}^0, \boldsymbol{\theta})_q\right] = \sum_{q=1}^Q \pi_{iq}(\widehat{\boldsymbol{\lambda}}_q)ME(\mathbf{x}^0, \widehat{\boldsymbol{\theta}})_q. \tag{16}$$

The standard errors for Equation (14), (15) and (16) can be computed using the Delta Method or Bootstrap.

Researchers interested on obtaining the marginal effects for each individual can use the individual-specific posterior distribution. Using the Bayes' theorem it is possible to obtain an estimator of the posterior membership probability:

$$\widehat{w}_{iq}(\boldsymbol{\psi}_q|y_{1i}, y_{2i}, \mathbf{z}_i, \boldsymbol{\theta}) = \frac{\widehat{\pi_{iq}}(\widehat{\boldsymbol{\lambda}}_q)\widehat{f}_q(y_{1i}, y_{2i}|\mathbf{z}_i)}{\sum_{q=1}^{Q}\widehat{\pi_{iq}}(\widehat{\boldsymbol{\lambda}}_q)\widehat{f}_q(y_{1i}, y_{2i}|\mathbf{z}_i)}, \tag{17}$$

which gives the probability for individual $i$ belonging to class $q$ given observed choices. Thus, individuals can be assigned to specific class using the highest posterior $\widehat{w}_{iq}(\boldsymbol{\psi}_q|y_{1i}, y_{2i}, \mathbf{z}_i, \boldsymbol{\theta})$. Finally, the individual-specific marginal effects can be estimated using the posterior membership probabilities:

$$\widehat{\widehat{ME}}_i = \sum_{q=1}^{Q} ME(\mathbf{x}^0, \widehat{\boldsymbol{\theta}})_q \widehat{w}_{iq}(\boldsymbol{\psi}_q|y_{1i}, y_{2i}, \mathbf{z}_i, \boldsymbol{\theta}). \tag{18}$$

It is important to emphasize that individual-specific estimators should be used with caution when applied to cross-sectional databases. As shown by Revelt and Train (2000) for continuous unobserved heterogeneity and Sarrias and Daziano (2018) for latent class models individual-specific estimates are consistent when the number of observed choices is large enough.

## 3.2 Re-parametrization

Since the log-likelihood function is not globally concave, the parameters in the model are assumed to lie in the interior of a compact convex set for identification. However, during the optimization procedure some of the parameters might tend to the boundary points of the parameter space generating identifiability problems of the MLE.

To prevent such cases and make the estimation easier, the following transformations are used. First, to ensure $\sigma_{q,\upsilon} > 0$, we rather estimate $\ln\sigma_{q,\upsilon}$, such that:

$$\sigma_{q,\upsilon} = \exp(\ln\sigma_{q,\upsilon}). \tag{19}$$

Second, to force the correlation to remain in the $(-1, +1)$ interval, we estimate the monotonic inverse hyperbolic tangent (Greene, 2003):

$$\tau_q = \frac{1}{2}\log\left(\frac{1+\rho_q}{1-\rho_q}\right) = athanh(\rho_q),$$

where $\tau_q$ is unrestricted, and $\rho_q$ is obtained using the inverse of $\tau_q$:

$$\rho_q = \tau_q^{-1} = \frac{\exp(2\tau_q) - 1}{\exp(2\tau_q) + 1}. \tag{20}$$

After convergence, the original parameters $\rho_q$ and $\sigma_{q,\upsilon}$ can be recovered using Equations (19) and (20), respectively. The standard error of such transformations can be computed using, again, the Delta Method or bootstrap.

### 3.3 Initial values

A potential problem with the application of the ML approach is its convergence to local rather to global maxima in models involving latent classes (Titterington, 1990; McLachlan and Peel, 2004). Thus, to analyze the sensitivity of the optimization process to different starting values is important (Formann, 1992).

In this article, the following procedure to obtain the initial values is used:

- Run an IV-Probit model to obtain $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\delta}}$, $\widehat{\sigma}_v$ and $\widehat{\rho}$. These estimates are perturbed slightly to obtain the the initial values for each class.
- Run a LC-Probit model on Equation (1) to obtain $\widehat{\boldsymbol{\lambda}}_q$.

### 3.4 Testing endogeneity

The model is flexible enough to test whether $y_{2i}$ is endogenous for each class. Here we are interested in testing the null and alternative hypotheses:

$$\begin{aligned} H_0 &: \rho_q = 0 \\ H_1 &: \rho_q \neq 0 \end{aligned}$$

where $H_0$ corresponds to exogeneity of $y_{2i}$ in Equation (1) for class $q$. Once the parameters are estimated, the classical Wald test can be used.

A more simpler approach is to test:

$$\begin{aligned} H_0 &: athanh(\rho_q) = 0 \\ H_1 &: athanh(\rho_q) \neq 0 \end{aligned}$$

Both test are distributed as $\chi^2$ with one degree of freedom.

### 3.5 Number of classes

In applied research, deciding the number of classes (which is unknown) is critical.[10] Using the incorrect number of classes might lead to misspecification issues biasing the estimates (Dias, 2006; Sarrias and Daziano, 2018). Unfortunately, $Q$ cannot be determined using the traditional tests since the parameter space is not bounded. For example, suppose that we which to test the null-hypothesis $H_0 : Q$ against the alternative hypothesis $H_1 : Q + 1$. Since $H_0$ corresponds to a boundary of the parameter space for $H_1$, the classical likelihood ratio test is not asymptotically distributed as $\chi^2$ (Titterington, 1990; McLachlan and Peel, 2004).

Given this problem, researchers rely on goodness-of-fit measures such as the Akaike's Information criterion (AIC) or Bayes' Information criterion (BIC). The procedure consists in estimating the model using a differing number of classes, and then choose the model with the lowest information criterion. However, other authors claim that this procedure, although is objective, might result in a proliferation of parameters as the number of classes increases, jeopardizing interpretability and identifiability of the estimates (Hess, 2014).

[10] For a more comprehensive review see McLachlan and Peel (2004, chapter 6).

Thus, some researchers proceed using a combination of goodness-of-fit and parsimony. In any case, the correct number of classes is an important but very difficult issue which has not yet been empirically solved.

Given that the model proposed in this article presents the same drawbacks in terms of determining the correct number of classes, it is recommended that researchers spend considerable time evaluating the number of classes based on goodness-of-fit measures and the distribution of the individuals' marginal effects.

# 4 Monte Carlo experiment

In this Section, a small Monte Carlo experiment is undertaken to analyze the finite and large properties of the IVLC-Probit Model under a well-specified model. In particular, it examines whether the ML estimator is able to recover the true parameters under small and large samples under a just-identified model.

## 4.1 Design

The Monte Carlo setup in this article has its roots in Rivers and Vuong (1988) and Adkins et al. (2008). The simulations consider the following latent process with a single continuous endogenous variable $y_{2i}$:

$$y_{1i}^* = \beta_{1q} + \beta_{2q}x_{2i} + \gamma_q y_{2i} + \epsilon_{iq}.$$

It is also considered a just-identified case where the equation for $y_{2i}$ is given by:

$$y_{2i} = \delta_{1q} + \delta_{2q}x_{2i} + \delta_{3q}x_{3i} + \upsilon_{iq},$$

where $x_{3i}$ is the additional exogenous variable (instrument) for the identification of $y_{2i}$. The set of exogenous variables $(x_{2i}, x_{3i})$ are drawn from a multivariate normal distribution with zero means, variances equal to 1 and covariance of 0.5, being $x_{3i}$ the instrument for $y_{2i}$.

For this experiment, it is assumed that exist two groups of individual, $Q = 2$, such that the proportion of individuals in each class are 70% and 30%, respectively. Thus, the unobserved heterogeneity is modeled assuming that the parameters are distributed following a discrete distribution with probabilities $\pi_1 = 0.7$ and $\pi_2 = 0.3$ such that $\lambda_1 = 0$ and $\lambda_2 = \log(\pi_2/(1-\pi_2)) \approx -0.8473$ in Equation (7).

The error terms are generated as $\epsilon_{iq} = (\rho_q/\sigma_{q,\upsilon})\upsilon_{iq} + \eta_{iq}$, where $\upsilon_{iq} \sim \mathrm{N}(0, \sigma^2_{q,\upsilon})$ and $\eta_{iq} \sim \mathrm{N}(0, 1-\rho_q^2)$; whereas the variance of the reduced form equation for both classes is set at 1: $\sigma^2_{q,\upsilon} = 1, \forall q = 1, 2$. To analyze the finite properties under different degree of endogeneity in each class, we run 3 different experiments for different values of $\rho_q$ given in Table 1. The rest of the parameters are held fixed in each experiment. Their values are presented in Table 2.

(Insert Table 1 , about here)

(Insert Table 2 , about here)

Three sample sizes are considered: $N = 100, 1000, 5000$. Similar to Adkins et al. (2008), $S = 1000$ Monte Carlo samples are generated for each experiment and sample size. After running the experiments, we compute the following statistics or each parameter $\theta$:

- The bias:

$$Bias = \frac{1}{S}\sum_{s=1}^{S}\left(\widehat{\theta}_s - \theta\right).$$

- The average coverage across individuals:

$$Cov = \frac{1}{S}\sum_{s=1}^{S}\mathbb{1}\left[\widehat{\theta}_s - 1.96\cdot SE_{\widehat{\theta},s} < \theta < \widehat{\theta}_s + 1.96\cdot SE_{\widehat{\theta},s}\right].$$

  where $\theta$ is true parameter. Coverage probability is basically a way of assessing standard error performance by evaluating confidence intervals and their coverage probability. It is just the proportion of simulated samples for which the estimated confidence interval includes the true parameter. If the standard error of the estimates is computed correctly, then we should observe that the CI includes the true parameter in 95% of the simulated samples.

The simulations were done using R software and the Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimization algorithm. To increase the convergence speed and avoid local optimal, the gradient was also coded. The formulas for the gradient are presented in Annex B. The free function to estimate the model is presented in Annex A.

## 4.2 Results

Table 3 includes bias and coverage for each experiment design based on 1000 samples. It is further divided into sub tables based on the number of individuals in each sample: $N = 100, 1000$ and 5000. Panel A shows the results under a well-specified model, whereas panel B presents the results under misspecification, that is, when a LC-Probit Model is estimated, but endogeneity is not taken into account. E1, E2 and E3 refers to the experiments under different degree of endogeneity (see Table 1).

(Insert Table 3, about here)

Under a well-specified model (Panel A), the average bias is high when the sample size is small, $n = 100$, but it is substantially reduced when the sample size increases. Furthermore, and as expected in latent class models, the identification of the parameters is difficult in small samples (Garrett and Zeger, 2000). In fact, the ML did not converge for about 20% of the designs when $n = 100$. This result is also found by Sarrias and Daziano (2018) for the LC-Multinomial Logit model. In general, the parameters are usually underestimated when $n = 100$ and the degree of correlation is high, that is the estimator is too small on average compared to the true value in finite sample. For $n = 1000$, the bias is reduced, but the parameters ($\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$) for class 2 tend to be overestimated and have higher bias than parameters in class 1. This might be explained by the lower proportion of individuals in this class which might compromise the identifiability of the parameters. This also explains the higher bias of the correlation parameter for class 2 in small samples. This pattern holds when the sample size increase up to 5000. In general the degree of bias increases as the correlation parameter increases in small sample. Finally, coverage probability for the point-estimates also improves as $n$ increases: when the sample size is 5000, almost all the parameters has coverage close to the nominal 95%.

Looking at panel B of Table 3, it is clear that not considering the endogeneity of $y_{2i}$ results in an important amount of bias for the standard LC-Probit model. Increasing the sample even worsens the situation. Overall, the magnitude of the bias found here are similar to those found by Akin for the IV-Probit model under a well-specified and just identified model.

# 5 Empirical application: The effect of weight on mental well-being

## 5.1 Model and data

This section shows a brief empirical application of the IVLC-Probit model by analyzing the heterogeneous effect of body weight on happiness using a Chilean dataset.

Obesity is a very highly stigmatized condition and overweight individuals face social exclusion and discrimination in many areas of their life (Wardle and Cooke, 2005). As pointed out by Fabricatore and Wadden (2004), prejudice and discrimination towards obese individuals persist despite worldwide increases in the prevalence of obesity and the recognition of genetic contributions to body weight. As a result, it has been assumed that overweight people are more likely to report lower levels of mental well-being such as self-esteem, satisfaction with life and health and greater depressive symptoms (Wardle and Cooke, 2005; Granberg, 2011).

However, to disentangle the causal effect of weight on mental well-being is not an easy task since the relationship might be endogenous. For example, a lack of mental health may lead to eating disorders (eating too much or eating too little) and less physical activity which in turn affects weight, resulting in a reverse-causality problem. Another source of endogeneity is omitted variables. If individuals' unobserved characteristics influence mental well-being, as well as their weight, then the relationship would be biased. For example, one might expect that people who have experienced greater stress (or major shocks in their life) are more likely to gain more weight and in turn decrease their mental well-being.

To deal with these problems, researchers have relied primarily on the instrumental variables approach. For example, Katsaiti (2012) analyzes the relationship between obesity and happiness using height as instrument for weight. Using the Germany German Socio-Economic Panel, UK British Household Panel Survey and Australia Household, Income and Labour Dynamics she finds that obesity has a negative effect on the subjective well-being of individuals for the three countries. Other studies exploit genetic variation to identify the causal effect of weight on mental well-being. For example, Kivimäki et al. (2011) use a 2SLS approach to analyze the relationship between Body Mass Index (BMI) and depression using fat mass and obesity-associated (FTO) genotype as instruments for weight in UK. Using a male sample, they find that a significant marginal effect of BMI equal 1% using OLS and 1.7% using IV procedure. An opposite effect is found by Willage (2018) for US data. He find no effect of BMI on depression and counseling using also an index of genetic risk as an instrument for weight. Other studies using the BMI of a biological relative to instrument for the respondent's weight have also found mixed results. Sabia and Rees (2015) find that increased weight negatively affects women' mental health, while Bjørngaard et al. (2015) find higher weight increases depression but not anxiety.

The mixing results using the IV approach might be explained, among other things, by the fact that IV estimates capture the Local Average Treatment Effect (LATE) instead of the Average Treatment Effect (ATE). That is, the estimated IV coefficients are the effect for the sub-population of individuals who would increase their weight because they are genetically disposed to being overweight but would not increase their weight if they are not so disposed—also known as the compliers (Angrist et al., 1996; Heckman et al., 2006; Greve, 2016). However, individuals' weight might response differently to the instrument, and hence, its power might be heterogeneous across the population. On the other hand, although the instrument could have a homogeneous effect on the population, it is also possible that the effect of body weight on the level of happiness could be heterogeneous due to observed or unobserved factors. In both cases, the IVLC-Probit model might be a more suitable approach.

To allow for such possibilities, I estimate the following structural model:

$$
\begin{aligned}
h_i^* &= \mathbf{x}_i'\boldsymbol{\beta}_q + \gamma_q \mathrm{bmi}_i + \epsilon_{iq}, \\
\mathrm{bmi}_i &= \mathbf{x}_i'\boldsymbol{\delta}_{x,q} + \delta_{z,q} z_i + \upsilon_{iq}, \\
h_i &= \mathbb{1}\left[h_i^* > 0\right], \\
g(\boldsymbol{\psi}_i|\boldsymbol{\lambda}_q) &= \begin{cases} \boldsymbol{\psi}_1 & \text{with probability } \pi_{i1}(\boldsymbol{\lambda}_q) \\ \boldsymbol{\psi}_2 & \text{with probability } \pi_{i2}(\boldsymbol{\lambda}_q) \\ \vdots & \vdots \\ \boldsymbol{\psi}_Q & \text{with probability } \pi_{iQ}(\boldsymbol{\lambda}_q) \end{cases}, \qquad (21) \\
\pi_{iq}(\boldsymbol{\lambda}_q) &= \frac{\exp\left(\mathbf{h}_i'\boldsymbol{\lambda}_q\right)}{\sum_{c=1}^{Q}\exp\left(\mathbf{h}_i'\boldsymbol{\lambda}_c\right)}, \quad q = 1, ..., Q, \\
(\epsilon_{iq}, \upsilon_{iq})|\mathbf{z}_i &\sim \mathrm{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_q\sigma_{q,\upsilon} \\ \rho_q\sigma_{q,\upsilon} & \sigma_{q,\upsilon}^2 \end{pmatrix}\right].
\end{aligned}
$$

where $h_i^*$ is the continuous but unobserved mental well-being of individual $i$; $h_i$ is the observed satisfaction with life (our proxy for mental well-being) which equals 1 if and only if $h_i^* > 0$; bmi is the body mass index for individual $i$ computed as $\mathrm{weight}_{kg}/\mathrm{height}^2_{meters}$; and $\pi_{iq}$ is discrete probability of individual $i$ belonging to class $q$ assumed to be a function of a constant and gender, $\mathbf{h}_i = (1, \text{male})'$.[11] As controls, $\mathbf{x}_i$, I use age, age-squared, a dummy variable indicating whether the respondent is married, education measured as years of schooling, a dummy indicating high household income, and a dummy indicating whether the individuals experienced an stressful event in the last year.

I use as instrument a dummy variable indicating whether or not the respondent's closer relatives (children, parents, siblings) have diabetes. The identification assumption is that respondent's family history in terms of diabetes is related to weight gain through genetics (inclusion restriction), and that after all observable factors have been taken into account the respondent's family history on diabetes has no direct impact on the probability of satisfaction with life (exclusion restriction). Inclusion restriction is supported by Groop et al. (1996) and Sargeant et al. (2000) who found that individuals with a family history of diabetes have increased abdominal fat content and are at increased risk for obesity.

To estimate the model in Equation (21), I use the Chilean National Health (ENS) Survey 2009-2010. The ENS is a national and regional representative survey of a randomly chosen sample of 5,293 respondents and is carried out by the Ministry of Health.[12] The overall aim of the ENS is to determine the prevalence of priority health problems in the Chilean adult population (people over 15 years old) using questionnaires, bio-physiological measurements and laboratory tests. This allow us to obtain objective measures of weight and height and avoid measurement errors. The data were collected between spring 2009 and summer 2010 at two home visits; an interview followed by a nurse visit to collect bio-physiological measurements, urine and blood test.[13] The dependent variable in this study is obtained from the response to the question "*How satisfied are you with life?*". The answer ranges from 1 "Completely unsatisfied" to 7 "Completely satisfied", whose response is then recoded into a binary variable. Namely, the respondents are classified as being "Satisfied" if they replied 6, or 7, and being "Unsatisfied or neither satisfied nor unsatisfied" if they replied with an integer between 1 and 5.[14] Table 4 reports the summary statistics.

(Insert Table 4, about here)

### 5.2 Results

Table 5 shows the estimated coefficients for different Probit-model specifications.[15] The results for the standard Probit model indicate that higher BMI decreases the probability of being satisfied with life. However the point estimate is weekly significant at the 10%. Then, I estimate the LC Probit model assuming two classes ($Q = 2$) to test whether there exists unobserved and/or observed heterogeneity in the relationship between weight

[11] A dummy for gender is included in the class-assignment since the literature has shown that the effect of weight on mental health might be different for men and women (Palinkas et al., 1996; Sabia and Rees, 2015)

[12] Pregnant women and people who reported violent behavior were excluded from the random selection within the home.

[13] The raw data can be downloaded from `http://epi.minsal.cl/estudios-y-encuestas-poblacionales/encuestas-poblacionales/descarga-ens/`.

[14] The estimations were also carried out classifying 5, 6 and 7 as being satisfied with life; however, no significant differences were observed. The results are available from the author upon request.

[15] To avoid numerical issues in the optimization procedure, the variables were re-scaled.

and well-being.[16] The first class represents around 60% of the sample, whereas class two represents the remained 40%. Although the average coefficient between the two classes is close to the Probit estimate,[17] in neither of the two we can reject the null that weight does not matter in the probability of being satisfied with life. The remaining coefficients are very stable across classes and with the expected sign. Given these results, it seems that the lack of statistical significance of the BMI is not due to unobserved heterogeneity but rather to a problem of bi-directionality of the relationship, at least in this sample.

(Insert Table 5, about here)

Then, I analyze whether the BMI is endogenous by estimating an IV-Probit model using the indicator of a relative with diabetes as instrument for respondent's BMI. The instrument shows sufficient power, as revealed by the first-equation estimates and the Wald statistics for the null that $\delta_z = 0$: $\chi^2_1 = 31.3$.[18] The exogeneity test ($H_0 : \rho = 0$) also rejects the null at the 10%. The IV estimate is significantly more negative than the point estimate of the Probit model (-7.9 vs -0.7), corroborating previous works that show that after accounting for sources of endogeneity the impact of weight on mental health is higher (see for example Kivimäki et al., 2011; Katsaiti, 2012).[19]

To capture unobserved heterogeneous effects of the instrument and the treatment (BMI), I also estimate the IVLC-Probit model assuming two classes. The first class represents about 88% of the sample, and men are approximately 3 times more likely to belong this class, while the second class is the smallest with the 12% of the sample. Although the instrument has greater power in class 1—as revealed by the Wald statistic—the magnitude is higher for class 2: a respondent with a close relative with diabetes has, on average, 1.7 more BMI than those who do not. However, endogeneity seems to be more serious issue for the first class.

The estimated coefficient for BMI has an opposite between classes. While for the first class an increase in BMI is negatively related to the probability of being satisfied with life, in the second class higher body weight is positively related to mental well-being. While the result for the second class may be surprising, it may be explained by the happy "jolly-fat" hypothesis (Crisp and McGuiness, 1976). That is, individuals who are obese, under certain circumstances, tend to be happier than their slimmer peers. This may be explained by the high consumption of some nutrients, which can reduce or prevent certain depressive symptoms (Roberts et al., 2002). In the same vein, other studies have shown that, in general, being obese is more accepted in certain social groups (Palinkas et al.,

[16] Models with higher number of classes did not show lower values of Bayesian Information Criterion (BIC).

[17] The average coefficient is -0.6 and it is obtained by multiplying each coefficient in each class by the proportion of individuals in each of them.

[18] Since the first-stage equation for the endogenous explanatory variable is linear, it is also possible to test the power of the instrument using the traditional $F$-test. In this case, the first-stage $F$-statistic is 31.28, which is greater than the 10 (Stock et al., 2002).

[19] In the particular case of having more instruments than endogenous variables, overidentification tests further explored in Guevara (2018) for discrete choice models can be used to check the validity of the instruments.

1996), especially in those with less income, which may also explain the lack of significance of income in class 2.[20]

Figure 1 shows the estimated average marginal effects of BMI on the probability of being satisfied with life under some of the specifications in Table 1. The histogram shows the conditional estimates of the ME using the LCIV-Probit estimates using Equation (18). It can be seen that for most individuals the effect of an increase of one unit of BMI on the probability of being satisfied with life is approximately between -0.6 and -0.5. In fact, the weighted marginal effect (WAVE), estimated using Equation (16), is -0.45, that is, an increase of one unit of BMI decreases, on average, the probability of being satisfied with life by 45% when considering the whole sample. However, it is clear that there is considerable heterogeneity in the effect of the BMI. Although there are positive marginal effects, they represent a fairly low percentage of the population (12% according to the results in Table 5). On the other hand, it can be observed that the average marginal effect of the Probit model and the IV-Probit model may underestimate and overestimate the effect for a large part of the population, respectively.

(Insert Figure 1, about here)

As an additional robustness check, Figure 2 plots the kernel estimates of the estimated individual-specific marginal effects using the LCIV-Probit estimates with two (for comparison purposes) and three classes. It can be observed that two modes are generated in the domain of negatives values when assuming three classes (dashed line). It is also possible to deduce that the first class, whose values are in the negative domain and has higher mass when $Q = 2$, is the one that is decomposed into two new classes when $Q = 3$. That is, there are individuals with a weight penalty with values close to -0.8 and others with values closer to -0.3, while the number of individuals with a positive marginal effect remains relatively constant compared to the model with two classes. Despite this bi-modality in the negative values, the WAMEs are very close to each other: -0.45 and -0.40 with two and three class, respectively.

(Insert Figure 2, about here)

# 6 Conclusion

This article proposes an Instrumental Variable approach for Latent Class models where the dependent variable is dichotomous. The model is derived under the assumption that there exists a fixed number of $Q$ of separate classes with different coefficient in each class. The unobserved sorting of individuals into groups can also be modeled using socio-economic characteristics. Furthermore, it is assumed that there exists a continuous variable which is potentially endogenous for some classes. The parameters are estimated using the ML procedure considering that the error terms are distributed as bivariate normal.

[20] As a reviewer correctly stated, several control variables are potentially endogenous. As an additional robusteness check, I have also estimated the IVLC-Probit model using only age and age-squared as controls. The results (available upon request) show that the estimates are not sentitive to the exclusion of the potentially endogenous controls.

The main advantages of the model are that: (1) it allows to explicitly model unobserved heterogeneity when individuals are sorted into groups (classes), and each group has its own set of characteristics; (2) it allows for consistent estimation of the treatment effect for each group of individuals; (2) it also allows heterogeneous effects of the instruments on the endogenous continuous variable across classes; (3) different degree of endogeneity across classes; and (4) it also allows to recover the average treatment effect. The model also shares the main disadvantages of latent class models, which are intensified due to the large number of parameters that must be estimated. First, the model has $((K+P+2)\times Q+L\times(Q-1))$ parameters that must be estimated. Therefore, an exactly identified model with 5 explanatory variables, two variables in the assignment of classes, and three classes requires the estimation of 40 parameters. Therefore, the sample size necessary to achieve consistent estimators might be demanding (Holm and Pedersen, 2007). This result is corroborated by the Monte Carlo experiment: the estimator is highly biased in small samples and the identification of the parameters is difficult to achieve. However, with samples greater than 1000, the biased is substantially reduced, and coverage improves. The second key message is that, even with moderate degree of endogeneity in each class, the LC-Probit model is highly biased.

To show the main advantages of the IVLC-Probit model, I also provide an empirical example by analyzing the impact of body weight on the probability of being satisfied with life. The results shows that both the IV-Probit and LC-Probit model miss important information about the heterogeneity of the impact. First, the latent class model does not allow controlling for the potential endogeneity of the BMI, and gives non-significant coefficients. If we also consider the results of the Monte Carlo experiment, this bias might be considerable. Second, the IV-Probit model, although it allows detecting a certain degree of endogeneity and a greater effect than the standard Probit model, is not capable of detecting the heterogeneous effect that the BMI has on the sample analyzed. The results show that the marginal impact is highly heterogeneous in the sample, ranging from -0.6 to 0.4, with a 88% of the sample with a negative impact.

This work can be extended in several ways in future research. First, a Monte Carlo study would be interesting to analyze the properties of the ML estimator against other potential approach such as the Control Function as in Guevara and Hess (2019) and Guevara (2015). Second, the IV approach to latent class models presented in this article can be easily generalized to other limited dependent variables such as the Ordered Probit model or include a panel dimension of the dataset.

# References

Adkins, L. C. et al. (2008). Small sample performance of instrumental variables probit estimators: A monte carlo investigation.

Amemiya, T. (1978). The estimation of a simultaneous equation generalized probit model. *Econometrica: Journal of the Econometric Society*, pages 1193–1205.

Amemiya, T. (1979). The estimation of a simultaneous-equation tobit model. *International Economic Review*, pages 169–181.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

Atella, V., Brindisi, F., Deb, P., and Rosati, F. C. (2004). Determinants of access to physician services in italy: a latent class seemingly unrelated probit approach. *Health economics*, 13(7):657–668.

Bhat, C. R. (2011). The maximum approximate composite marginal likelihood (macml) estimation of multinomial probit-based unordered response choice models. *Transportation Research Part B: Methodological*, 45(7):923–939.

Bhat, C. R., Astroza, S., Sidharthan, R., Alam, M. J. B., and Khushefati, W. H. (2014). A joint count-continuous model of travel behavior with selection based on a multinomial probit residential density choice model. *Transportation Research Part B: Methodological*, 68:31–51.

Bjørngaard, J. H., Carslake, D., Nilsen, T. I. L., Linthorst, A. C., Smith, G. D., Gunnell, D., and Romundstad, P. R. (2015). Association of body mass index with depression, anxiety and suicide—an instrumental variable analysis of the hunt study. *PloS one*, 10(7):e0131708.

Chen, J. et al. (2017). Consistency of the mle under mixture models. *Statistical Science*, 32(1):47–63.

Chen, J., Li, P., et al. (2009). Hypothesis test for normal mixture models: The em approach. *The Annals of Statistics*, 37(5A):2523–2542.

Clark, A., Etilé, F., Postel-Vinay, F., Senik, C., and Van der Straeten, K. (2005). Heterogeneity in reported well-being: evidence from twelve european countries. *The Economic Journal*, 115(502):C118–C132.

Crisp, A. H. and McGuiness, B. (1976). Jolly fat: relation between obesity and psychoneurosis in general population. *Br Med J*, 1(6000):7–9.

Deb, P. and Trivedi, P. K. (2002). The structure of demand for health care: latent class versus two-part models. *Journal of health economics*, 21(4):601–625.

Dias, J. G. (2006). Model selection for the binary latent class model: A monte carlo simulation. In *Data science and classification*, pages 91–99. Springer.

d'Uva, T. B. and Jones, A. M. (2009). Health care utilisation in europe: new evidence from the echp. *Journal of health economics*, 28(2):265–279.

Fabricatore, A. N. and Wadden, T. A. (2004). Psychological aspects of obesity. *Clinics in dermatology*, 22(4):332–337.

Florens, J.-P., Heckman, J. J., Meghir, C., and Vytlacil, E. (2008). Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects. *Econometrica*, 76(5):1191–1206.

Formann, A. K. (1992). Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association*, 87(418):476–486.

Fox, J., Friendly, M., and Weisberg, S. (2013). Hypothesis tests for multivariate linear models using the car package. *The R Journal*, 5(1):39–52.

Garrett, E. S. and Zeger, S. L. (2000). Latent class model diagnosis. *Biometrics*, 56(4):1055–1067.

Gopimatj, D. A. (1997). Modeling heterogeneity in discrete choice processes: Application to travel demand. *Transportation Research Part A*, 1(31):86.

Granberg, E. (2011). Depression and obesity. *Handbook of the Social Science of Obesity.*

Greene, W. (2004). Convenient estimators for the panel probit model: Further results. *Empirical Economics*, 29(1):21–47.

Greene, W. H. (2003). *Econometric analysis.* Pearson Education India.

Greene, W. H. and Hensher, D. A. (2003). A latent class model for discrete choice analysis: contrasts with mixed logit. *Transportation Research Part B: Methodological*, 37(8):681–698.

Greene, W. H. and Hensher, D. A. (2010). *Modeling ordered choices: A primer.* Cambridge University Press.

Greve, J. (2016). Why do people with higher body weight earn lower wages? In *The Oxford Handbook of Economics and Human Biology.*

Groop, L., Forsblom, C., Lehtovirta, M., Tuomi, T., Karanko, S., Nissén, M., Ehrnström, B.-O., Forsén, B., Isomaa, B., Snickars, B., et al. (1996). Metabolic consequences of a family history of niddm (the botnia study): evidence for sex-specific parental effects. *Diabetes*, 45(11):1585–1593.

Guevara, C. A. (2015). Critical assessment of five methods to correct for endogeneity in discrete-choice models. *Transportation Research Part A: Policy and Practice*, 82:240–254.

Guevara, C. A. (2018). Overidentification tests for the exogeneity of instruments in discrete choice models. *Transportation Research Part B: Methodological*, 114:241–253.

Guevara, C. A. and Hess, S. (2019). A control-function approach to correct for endogeneity in discrete choice models estimated on sp-off-rp data and contrasts with an earlier fiml approach by train & wilson. *Transportation Research Part B: Methodological*, 123:224–239.

Heckman, J. J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica*, 46(4):931–959.

Heckman, J. J., Urzua, S., and Vytlacil, E. (2006). Understanding instrumental variables in models with essential heterogeneity. *The Review of Economics and Statistics*, 88(3):389–432.

Henningsen, A. and Toomet, O. (2011). maxlik: A package for maximum likelihood estimation in r. *Computational Statistics*, 26(3):443–458.

Hess, S. (2014). Latent class structures: taste heterogeneity and beyond. In *Handbook of choice modelling*. Edward Elgar Publishing.

Holm, A. and Pedersen, M. (2007). Latent class binary regression models: identification and estimation.

Jackson, C. H. et al. (2011). Multi-state models for panel data: the msm package for r. *Journal of statistical software*, 38(8):1–29.

Jedidi, K., Ramaswamy, V., and DeSarbo, W. S. (1993). A maximum likelihood method for latent class regression involving a censored dependent variable. *Psychometrika*, 58(3):375–394.

Katsaiti, M. S. (2012). Obesity and happiness. *Applied Economics*, 44(31):4101–4114.

Kivimäki, M., Jokela, M., Hamer, M., Geddes, J., Ebmeier, K., Kumari, M., Singh-Manoux, A., Hingorani, A., and Batty, G. D. (2011). Examining overweight and obesity as risk factors for common mental disorders using fat mass and obesity-associated (fto) genotype-instrumented analysis: The whitehall ii study, 1985–2004. *American journal of epidemiology*, 173(4):421–429.

Maddala, G. (2002). Limited-dependent and qualitative variables in econometrics.

McLachlan, G. J. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

Moffitt, R. (2008). Estimating marginal treatment effects in heterogeneous populations. *Annales d'Economie et de Statistique*, pages 239–261.

Nagin, D. S. and Land, K. C. (1993). Age, criminal careers, and population heterogeneity: Specification and estimation of a nonparametric, mixed poisson model. *Criminology*, 31(3):327–362.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245.

Palinkas, L. A., Wingard, D. L., and Barrett-Connor, E. (1996). Depressive symptoms in overweight and obese older adults: a test of the "jolly fat" hypothesis. *Journal of psychosomatic research*, 40(1):59–66.

Palomino, J. and Sarrias, M. (2019). The monetary subjective health evaluation for commuting long distances in chile: A latent class analysis. *Papers in Regional Science.*

Revelt, D. and Train, K. (2000). Customer-specific taste parameters and mixed logit: Households' choice of electricity supplier.

Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., Firth, D., and Ripley, M. B. (2013). Package 'mass'. *Cran R*, 538.

Rivers, D. and Vuong, Q. H. (1988). Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of econometrics*, 39(3):347–366.

Roberts, R. E., Strawbridge, W. J., Stephane, D., and Kaplan, G. A. (2002). Are the fat more jolly? *Annals of behavioral medicine*, 24(3):169–180.

Sabia, J. J. and Rees, D. I. (2015). Body weight, mental health capital, and academic achievement. *Review of Economics of the Household*, 13(3):653–684.

Sargeant, L., Wareham, N., and Khaw, K. (2000). Family history of diabetes identifies a group at increased risk for the metabolic consequences of obesity and physical inactivity in epic-norfolk: a population-based study. *International journal of obesity*, 24(10):1333–1339.

Sarrias, M. and Daziano, R. A. (2018). Individual-specific point and interval conditional estimates of latent class logit parameters. *Journal of choice modelling*, 27:50–61.

Shen, J. (2009). Latent class model or mixed logit model? a comparison by transport mode choice data. *Applied Economics*, 41(22):2915–2924.

Skeels, C. L. and Taylor, L. W. (2015). Prediction in linear index models with endogenous regressors. *The Stata Journal*, 15(3):627–644.

Smith, R. J. and Blundell, R. W. (1986). An exogeneity test for a simultaneous equation tobit model with an application to labor supply. *Econometrica: Journal of the Econometric Society*, pages 679–685.

Stock, J. H., Wright, J. H., and Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 20(4):518–529.

Titterington, D. (1990). Some recent research in the analysis of mixture distributions. *Statistics*, 21(4):619–641.

Vermunt, J. K. and Van Dijk, L. (2001). A nonparametric random-coefficients approach: The latent class regression model. *Multilevel Modelling Newsletter*, 13(2):6–13.

Walker, J. L., Ben-Akiva, M., and Bolduc, D. (2007). Identification of parameters in normal error component logit-mixture (neclm) models. *Journal of Applied Econometrics*, 22(6):1095–1125.

Wang, P., Cockburn, l. M., and Puterman, M. L. (1998). Analysis of patent data—a mixed-poisson-regression-model approach. *Journal of Business & Economic Statistics*, 16(1):27–41.

Wardle, J. and Cooke, L. (2005). The impact of obesity on psychological well-being. *Best Practice & Research Clinical Endocrinology & Metabolism*, 19(3):421–440.

Wedel, M. and DeSarbo, W. S. (1994). A review of recent developments in latent class regression models.

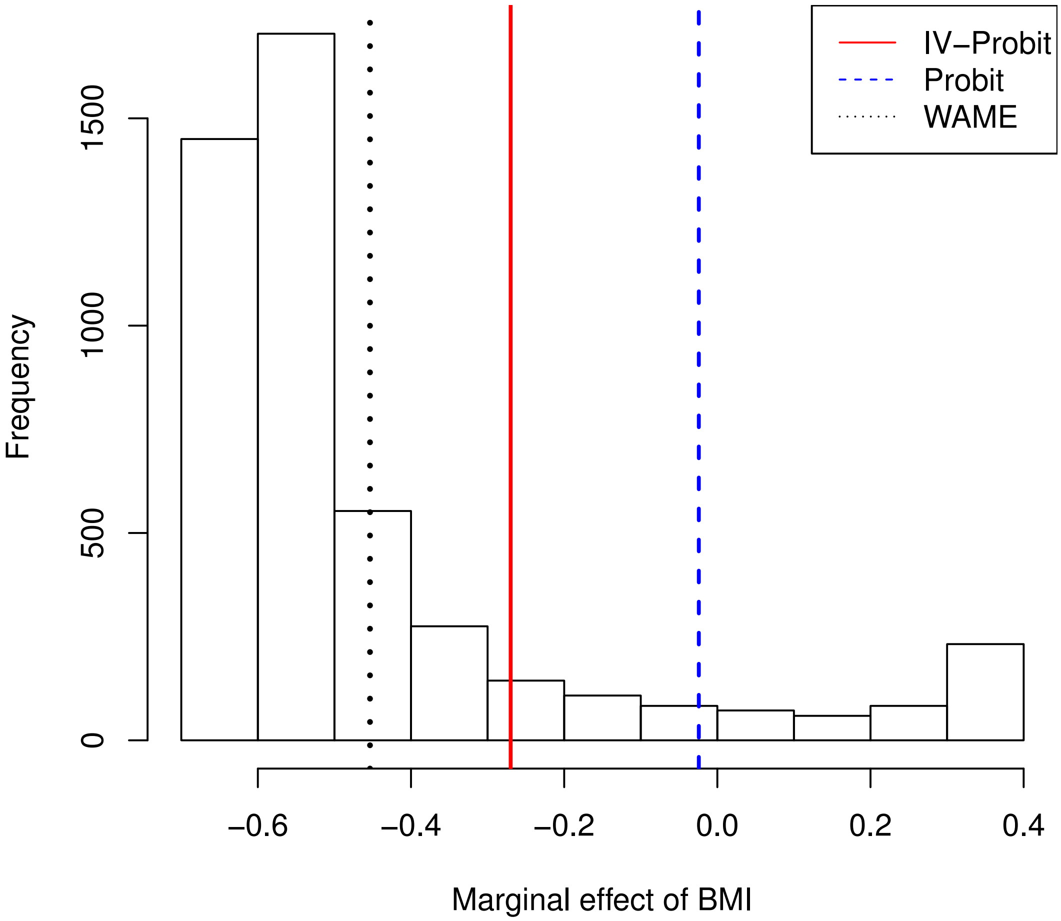
Wedel, M. and DeSarbo, W. S. (1995). A mixture likelihood approach for generalized linear models. *Journal of Classification*, 12(1):21–55.

Willage, B. (2018). The effect of weight on mental health: New evidence using genetic ivs. *Journal of Health Economics*, 57:113–130.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data.* MIT press.

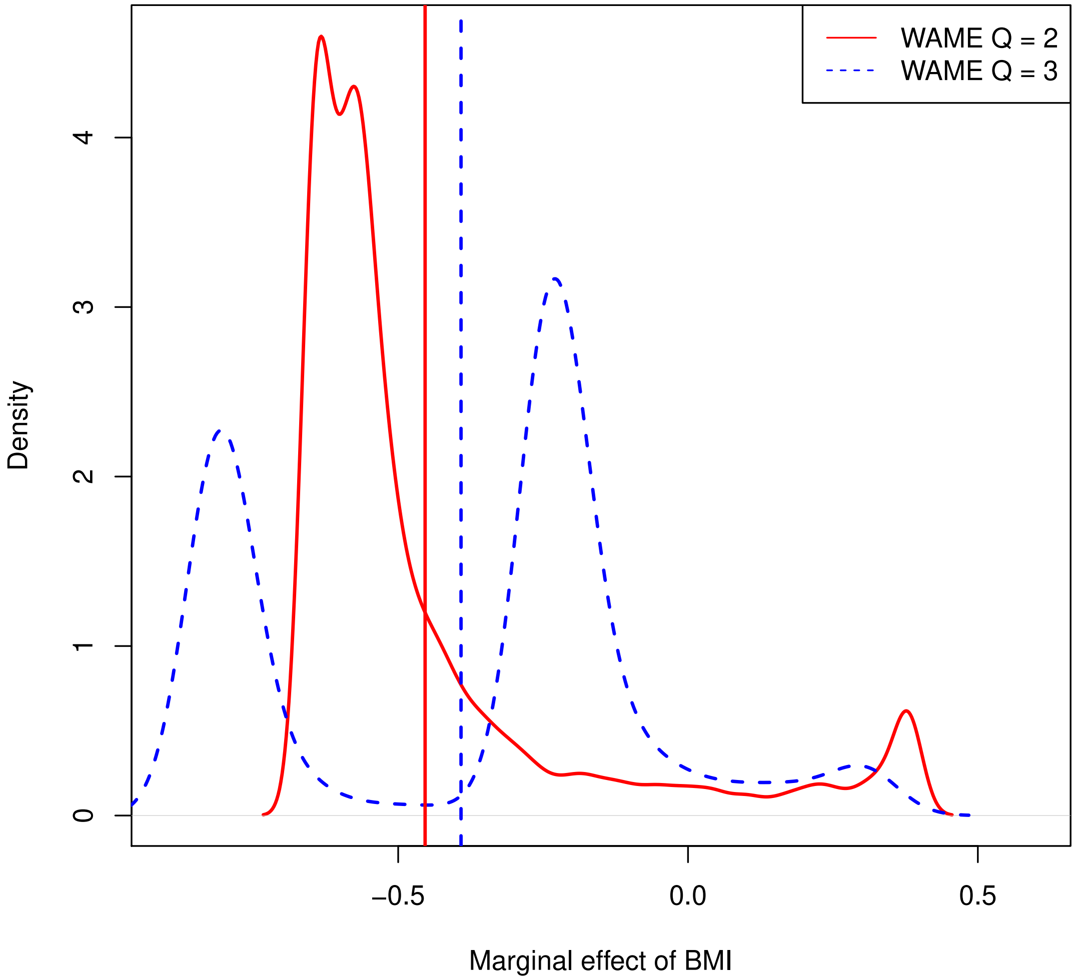
# Figures

Figure 1: Average marginal effects of an increase in one unit of BMI on the probability of being satisfied with life.



*Notes:* The histogram shows the estimated individual-specific marginal effects for the IVLC-Probit model. WAME is the (unconditional) weighted average marginal effect of the IVLC- Probit model.

Figure 2: Kernel estimate of the distribution for the estimated individual-specific marginal effects for the IVLC-Probit model: Comparison using 2 and 3 classes.



# Tables

Table 1: Degree of endogeneity in each experiment for each class.

| | Experiment 1 | | Experiment 2 | | Experiment 3 | |
|---|---|---|---|---|---|---|
| | $\rho$ | $\tau = 0.5 \times \ln\left(\frac{1+\rho}{1-\rho}\right)$ | $\rho$ | $\tau = 0.5 \times \ln\left(\frac{1+\rho}{1-\rho}\right)$ | $\rho$ | $\tau = 0.5 \times \ln\left(\frac{1+\rho}{1-\rho}\right)$ |
| Class 1 | -0.8 | -1.0986123 | -0.2 | -0.2027326 | -0.6 | -0.6931472 |
| Class 2 | 0.8 | 1.0986123 | 0.2 | -0.202732 | 0 | 0 |

Table 2: Parameters for the Monte Carlo experiments for each class.

| | $\beta_1$ | $\beta_2$ | $\gamma$ | $\delta_1$ | $\delta_2$ | $\delta_3$ | $\lambda$ | $\sigma$ | $\ln(\sigma)$ |
|---|---|---|---|---|---|---|---|---|---|
| Class 1 | -1 | -1 | -1 | -1 | -1 | -1 | - | 1 | 0 |
| Class 2 | 1 | 1 | 1 | 1 | 1 | 1 | -0.8472979 | 1 | 0 |

Table 3: Monte Carlo results.

| | | A: Results on a well-specified model | | | | | | B: Results on a misspecified model (LC) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | E1 | | E2 | | E3 | | E1 | | E2 | | E3 | |
| $N$ | Parameter | Bias | Cov | Bias | Cov | Bias | Cov | Bias | Cov | Bias | Cov | Bias | Cov |
| 100 | $\beta_{1,q=1}$ | -0.1859 | 91% | -0.1496 | 91% | -0.1100 | 91% | -1.6439 | 90% | -1.4055 | 89% | -1.2746 | 92% |
| | $\beta_{2,q=1}$ | -0.2478 | 91% | -0.1021 | 85% | -0.1026 | 90% | -1.2360 | 83% | -0.5974 | 81% | -1.9394 | 91% |
| | $\gamma_{q=1}$ | -0.2054 | 91% | -0.1217 | 93% | -0.1661 | 89% | 0.0394 | 60% | 0.4175 | 57% | -0.2835 | 66% |
| | $\beta_{1,q=2}$ | -0.6344 | 81% | -0.3746 | 78% | -0.3891 | 75% | 0.8874 | 92% | 1.4315 | 92% | 0.9889 | 92% |
| | $\beta_{2,q=2}$ | -0.6477 | 74% | -0.2235 | 82% | -0.2349 | 80% | 1.7198 | 99% | 3.0922 | 99% | 3.1678 | 100% |
| | $\gamma_{q=2}$ | -0.1684 | 81% | -0.5923 | 61% | -0.6105 | 63% | -1.2824 | 18% | -1.3181 | 21% | -1.2923 | 26% |
| | $\lambda_{q=2}$ | 0.1686 | 95% | 0.2070 | 93% | 0.1974 | 90% | 1.0987 | 21% | 1.1449 | 23% | 1.1111 | 17% |
| | $\ln(\sigma)_{q=1}$ | -0.1110 | 81% | -0.1149 | 78% | -0.0904 | 84% | | | | | | |
| | $\ln(\sigma)_{q=2}$ | -0.0775 | 80% | 0.0368 | 70% | 0.0556 | 70% | | | | | | |
| | $\tau_{q=1}$ | -0.1520 | 91% | 0.1982 | 85% | 0.1111 | 88% | | | | | | |
| | $\tau_{q=2}$ | -0.6950 | 69% | 2.6475 | 92% | 2.8328 | 89% | | | | | | |
| 1000 | $\beta_{1,q=1}$ | -0.0256 | 93% | -0.0277 | 93% | -0.0316 | 93% | -3.8116 | 26% | -4.0252 | 57% | -4.1657 | 36% |
| | $\beta_{2,q=1}$ | -0.0258 | 94% | -0.0244 | 94% | -0.0205 | 92% | -2.2476 | 18% | 1.7952 | 37% | -1.1016 | 26% |
| | $\gamma_{q=1}$ | -0.0215 | 93% | -0.0200 | 93% | -0.0215 | 92% | -0.7505 | 13% | 2.8386 | 24% | 0.3125 | 19% |
| | $\beta_{1,q=2}$ | 0.0408 | 92% | 0.3645 | 84% | 0.2693 | 86% | 0.3016 | 92% | 0.3775 | 89% | 0.2547 | 93% |
| | $\beta_{2,q=2}$ | 0.0075 | 89% | 0.3808 | 83% | 0.2141 | 86% | 0.5742 | 48% | 0.5509 | 75% | 0.5250 | 49% |
| | $\gamma_{q=2}$ | 0.1363 | 91% | 0.3064 | 88% | 0.2364 | 89% | -1.0568 | 1% | -1.2196 | 1% | -1.0939 | 1% |
| | $\lambda_{q=2}$ | 0.0059 | 100% | 0.0008 | 100% | 0.0040 | 99% | 1.1437 | 4% | 1.2868 | 11% | 1.2492 | 4% |
| | $\ln(\sigma)_1$ | -0.0064 | 93% | -0.0063 | 92% | -0.0070 | 93% | | | | | | |
| | $\ln(\sigma)_2$ | -0.0058 | 93% | -0.0002 | 93% | 0.0038 | 92% | | | | | | |
| | $\tau_{q=1}$ | -0.0030 | 92% | 0.0169 | 90% | 0.0149 | 92% | | | | | | |
| | $\tau_{q=2}$ | 0.2445 | 89% | 0.3416 | 87% | 0.0223 | 85% | | | | | | |
| 5000 | $\beta_{1,q=1}$ | -0.0046 | 94% | -0.0043 | 93% | -0.0053 | 93% | -3.4656 | 1% | -3.9469 | 9% | -2.8889 | 3% |
| | $\beta_{2,q=1}$ | -0.0020 | 93% | -0.0007 | 93% | -0.0011 | 95% | -2.3225 | 0% | 3.5442 | 3% | -1.4385 | 1% |
| | $\gamma_{q=1}$ | -0.0022 | 93% | -0.0009 | 92% | -0.0020 | 94% | -1.2180 | 0% | 3.1257 | 2% | -0.4534 | 0% |
| | $\beta_{1,q=2}$ | 0.0107 | 94% | 0.0591 | 93% | 0.0661 | 93% | 0.1081 | 80% | 0.2793 | 49% | 0.1607 | 78% |
| | $\beta_{2,q=2}$ | 0.0021 | 94% | 0.0466 | 93% | 0.0587 | 92% | 0.4476 | 9% | 0.2271 | 43% | 0.5090 | 18% |
| | $\gamma_{q=2}$ | 0.0360 | 94% | 0.0317 | 95% | 0.0487 | 94% | -0.9933 | 0% | -1.1852 | 0% | -1.0398 | 0% |
| | $\lambda_{q=2}$ | -0.0009 | 99% | -0.0009 | 99% | -0.0019 | 98% | 1.1100 | 4% | 1.1734 | 18% | 1.1489 | 10% |
| | $\ln(\sigma)_{q=1}$ | -0.0004 | 95% | -0.0003 | 95% | -0.0003 | 95% | | | | | | |
| | $\ln(\sigma)_{q=2}$ | -0.0010 | 94% | 0.0024 | 94% | 0.0018 | 94% | | | | | | |
| | $\tau_{q=1}$ | -0.0047 | 94% | -0.0033 | 93% | -0.0027 | 93% | | | | | | |
| | $\tau_{q=2}$ | -0.0299 | 95% | 0.0924 | 94% | -0.0231 | 94% | | | | | | |

*Notes: The results are averaged over the number of samples S for which the ML converged.*

Table 4: Summary statistics of variables used in the applied example. ENS survey.

| Variable | Mean | Std. Dev | Min | Max | N |
|---|---|---|---|---|---|
| Satisfied with life ( = 1) | 0.658 | 0.474 | 0 | 1 | 4763 |
| BMI (kg/meters$^2$) | 27.845 | 5.389 | 14.005 | 75.243 | 4763 |
| Age | 46.110 | 18.357 | 15 | 100 | 4763 |
| Married ( = 1) | 0.409 | 0.492 | 0 | 1 | 4763 |
| Education (years) | 9.775 | 4.209 | 0 | 22 | 4763 |
| High Income (=1) | 0.187 | 0.390 | 0 | 1 | 4763 |
| Stressful event ( = 1) | 0.106 | 0.308 | 0 | 1 | 4763 |
| Relative with diabetes ( = 1) | 0.303 | 0.460 | 0 | 1 | 4763 |
| Male ( = 1) | 0.403 | 0.491 | 0 | 1 | 4763 |

*Notes: The full ENS sample has 5,293 respondents. However, the sample used for estimation has 4,763 individuals after cleaning for missing values on the main covariates. Hihg income is defined as those individuals whose household income is greater than $650,000 CLP.*

Table 5: Estimates of the relationship between weight and satisfaction with life

| | Probit | | LC-Probit | | | | IV-Probit | | IVLC- Probit | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Est.* | *SE* | *Est.* | *SE* | *Est.* | *SE* | *Est.* | *SE* | *Est.* | *SE* | *Est.* | *SE* |
| | | | Class 1 | | Class 2 | | | | Class 1 | | Class 2 | |
| | | | | | *A: Satisfaction with life equation* | | | | | | | |
| Constant | $0.957^{***}$ | 0.158 | $0.762^{***}$ | 0.202 | $1.199^{***}$ | 0.266 | $2.399^{***}$ | 0.739 | $4.631^{***}$ | 0.674 | $-2.568^{**}$ | 1.152 |
| BMI/100 | $-0.706^{*}$ | 0.369 | $-0.628$ | 0.444 | $-0.615$ | 0.689 | $-7.883^{**}$ | 3.782 | $-20.830^{***}$ | 4.240 | $10.180^{***}$ | 3.636 |
| Age/10 | $-0.390^{***}$ | 0.058 | $-0.345^{***}$ | 0.073 | $-0.456^{***}$ | 0.096 | $-0.117$ | 0.170 | 0.357 | 0.226 | $-0.525^{***}$ | 0.189 |
| $\text{Age}^2/1000$ | $0.348^{***}$ | 0.058 | $0.320^{***}$ | 0.073 | $0.382^{***}$ | 0.097 | 0.098 | 0.157 | $-0.325$ | 0.206 | $0.495^{**}$ | 0.194 |
| Married | $0.162^{***}$ | 0.042 | 0.084 | 0.054 | $0.275^{***}$ | 0.071 | $0.189^{***}$ | 0.041 | $0.331^{***}$ | 0.053 | $-0.144$ | 0.122 |
| Educ /10 | $0.529^{***}$ | 0.057 | $0.541^{***}$ | 0.072 | $0.522^{***}$ | 0.096 | $0.400^{***}$ | 0.106 | $0.243^{*}$ | 0.141 | $0.638^{***}$ | 0.163 |
| High Income | $0.409^{***}$ | 0.058 | $0.293^{***}$ | 0.077 | $0.517^{***}$ | 0.094 | $0.397^{***}$ | 0.060 | $0.385^{***}$ | 0.093 | 0.149 | 0.183 |
| Stressful event | $-0.389^{***}$ | 0.062 | $-0.370^{***}$ | 0.074 | $-0.347^{***}$ | 0.118 | $-0.375$ | 0.064 | $-0.337^{***}$ | 0.097 | $-0.155$ | 0.195 |
| | | | | | *B: Endogenous variable equation* | | | | | | | |
| Constant | | | | | | | $0.207^{***}$ | 0.005 | $0.191^{***}$ | 0.005 | $0.272^{***}$ | 0.022 |
| Age/10 | | | | | | | $0.033^{***}$ | 0.002 | $0.032^{***}$ | 0.002 | $0.035^{***}$ | 0.009 |
| $\text{Age}^2/1000$ | | | | | | | $-0.030^{***}$ | 0.002 | $-0.029^{***}$ | 0.002 | $-0.034^{***}$ | 0.009 |
| Married | | | | | | | $0.005^{***}$ | 0.002 | $0.008^{***}$ | 0.002 | 0.001 | 0.006 |
| Educ/10 | | | | | | | $-0.012^{***}$ | 0.002 | $-0.006^{***}$ | 0.002 | $-0.031^{***}$ | 0.009 |
| High Income | | | | | | | 0.002 | 0.002 | $0.004^{**}$ | 0.002 | 0.003 | 0.009 |
| Stressful event | | | | | | | $-0.003$ | 0.002 | $-0.003$ | 0.002 | $-0.004$ | 0.009 |
| Relative's diabetes | | | | | | | $0.009^{***}$ | 0.002 | $0.005^{***}$ | 0.002 | $0.017^{***}$ | 0.006 |
| | | | | | *C: Class-Assignment Equation* | | | | | | | |
| Constant | | | | | $-6.605$ | 30.662 | | | | | $-0.928^{***}$ | 0.145 |
| Male | | | | | 12.675 | 43.265 | | | | | $-1.214^{***}$ | 0.153 |
| $\ln(\sigma_\upsilon)$ | | | | | | | $-2.970^{***}$ | 0.010 | $-3.330^{***}$ | 0.021 | $-2.667^{***}$ | 0.031 |
| $\text{athanh}(\rho)$ | | | | | | | $0.392^{*}$ | 0.229 | $0.980^{***}$ | 0.340 | $-0.787^{*}$ | 0.470 |
| Shares ($\pi$) | | | 60% | | 40% | | | | 88% | | 12% | |
| LL | -2866 | | -2849.969 | | | | -2864.82 | | -2859.361 | | | |
| Parameters | 8 | | 18 | | | | 18 | | 38 | | | |
| N | 4763 | | 4763 | | | | 4763 | | 4763 | | | |
| $H_0 : \rho = 0$ | | | | | | | 2.927 | | 8.313 | | 2.802 | |
| p-value | | | | | | | 0.087 | | 0.004 | | 0.094 | |
| $H_0 : \delta_z = 0$ | | | | | | | 31.332 | | 12.078 | | 7.671 | |
| p-value | | | | | | | 0.000 | | 0.000 | | 0.007 | |

*Notes: * $p < 0.1$*, ** $p < 0.05$, *** $p < 0.01$. For comparison purposes, the log-likelihood (LL) for IV-Probit and IVLC-Probit models are computed for the choice equation.*

# A Annex 1: Code in R to estimate the IVLC-Probit model

In this Section, I show how the IVLC-Probit model can be estimated using the function `LcIv` in `R` software. This function uses the analytical gradient to increase accuracy and convergence speed. The formulas for the gradient are presented in Annex B. I also show how to compute the marginal effects and perform hypothesis testing.

The function can be downloaded into `R` as follows:

```r
# Downloading the function into R
library("devtools")
di <- "https://bitbucket.org/mauricio1986/lciv/raw/88fdcc19a85d739b850e595f269710eac0878d71/lciv.R"
source_url(di)
```

## A.1 Creating an artificial data set

To show how to estimate the model, an artificial data set for $N = 10000$ individuals is created assuming the following latent process with a single continuous endogenous variable $y_{2i}$:

$$y^*_{1i} = \beta_{0q} + \beta_{1q}x_{1i} + \gamma_q y_{2i} + \epsilon_{iq}.$$

It is also considered a just-identified case where the reduced equation for $y_{2i}$ is given by:

$$y_{2i} = \delta_{0q} + \delta_{1q}x_{1i} + \delta_{2q}x_{2i} + \upsilon_{iq},$$

where $x_{2i}$ is the additional exogenous variable (instrument) for the identification of $y_{2i}$. It is further assumed that exist two groups of individual, $Q = 2$, such that the proportion of individuals in each class are 70% and 30%, respectively. Thus, the unobserved heterogeneity is modeled assuming that the parameters are distributed following a discrete distribution with probabilities $\pi_1 = 0.7$ and $\pi_2 = 0.3$ such that $\lambda_1 = 0$ and $\lambda_2 = \log(\pi_2/(1 - \pi_2)) \approx -0.8473$ in Equation (7).

The number of individuals, classes and the proportion of individuals in each class are defined as follows in R:

```r
## Globals
set.seed(1986)       # Set seed for random numbers
N <- 10000           # Number of individuals
Q <- 2               # Number of classes
prop <- c(0.7, 0.3) # Proportion of individuals in each class
```

The (included and excluded) exogenous variables are drawn from the following multivariate normal distribution:

$$(x_1, x_2) \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.5\\0.5 & 1\end{pmatrix}\right],$$

where $x_1$ is the included exogenous variable, whereas $x_2$ is the additional instrument. To create $N$ draws, we use the MASS package (Ripley et al., 2013):

```r
# Generate exogenous variables
library("MASS")  # Package to create draws from MVN
Sigma_x <- matrix(c(1, 0.5, 0.5, 1), nrow = 2)
mu_x    <- c(0, 0)
mvn_x <- mvrnorm(N, mu = mu_x, Sigma = Sigma_x)
x1 <- mvn_x[, 1] # Included exogenous variable
x2 <- mvn_x[, 2] # Instrument
```

To generate the error terms $\epsilon_{iq}$ and $\upsilon_{iq}$ for $q = 1, 2$, we assume that $\rho_1 = -0.6$ and $\rho_2 = 0$, whereas the variance of $\upsilon_{iq}$ are assumed to be $\sigma^2_{1,\upsilon} = \sigma^2_{2,\upsilon} = 1$. Thus, $y_2$ is exogenous for class two, but endogenous for class one.

```r
# Generate rho and sigma for each class
rho   <- c(-0.6, 0) # Degree of the endogeneity
sigma <- c(1, 1)    # Variance of upsilon

# Generate upsilon for each class
v1 <- rnorm(N * prop[1], 0, sigma[1])
v2 <- rnorm(N * prop[2], 0, sigma[2])

# Generate epsilon for each class
e1 <- (rho[1] / sigma[1]) * v1 + rnorm(N * prop[1], 0, sqrt(1 - rho[1]^2))
e2 <- (rho[2] / sigma[2]) * v2 + rnorm(N * prop[2], 0, sqrt(1 - rho[2]^2))
e  <- c(e1, e2)
v  <- c(v1, v2)
```

Table A.1: Parameters for the artificial data set

| | $\beta_0$ | $\beta_1$ | $\gamma$ | $\delta_0$ | $\delta_1$ | $\delta_2$ | $\lambda$ | $\sigma$ | $\ln(\sigma)$ | $\rho$ | $\tau$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Class 1 | -1 | -1 | -1 | -1 | -1 | -1 | - | 1 | 0 | -0.6 | -0.6931472 |
| Class 2 | 1 | 1 | 1 | 1 | 1 | 1 | -0.8472979 | 1 | 0 | 0 | 0 |

Assuming the true parameters given in Table A.1, the final data set is generated as follows:

```r
# True coefficients in the probit model
b0    <- c(rep(-1, N * prop[1]), rep(1, N * prop[2]))
b1    <- c(rep(-1, N * prop[1]), rep(1, N * prop[2]))
gamma <- c(rep(-1, N * prop[1]), rep(1, N * prop[2]))

# True coefficients in the reduced equation
```

```
d0    <- c(rep(-1,  N * prop[1]), rep(1, N * prop[2]))
d1    <- c(rep(-1,  N * prop[1]), rep(1, N * prop[2]))
d2    <- c(rep(-1,  N * prop[1]), rep(1, N * prop[2])) # Strength of the instrument

# Generate DGP
y2 <- d0 + d1 * x1 + d2 * x2     + v # Reduced equation
l1 <- b0 + b1 * x1 + gamma * y2 + e # Latent model for probit equation
y1 <- as.numeric(l1 > 0)            # Generate dummy dependent variable
data <- as.data.frame(cbind(y1, y2, x1, x2))
```

### A.2 Estimating the IVLC-Probit model

The function to estimate the model is `LcIv`, which can be used as follows:

```
# Estimate the IVLC- probit model
lciv <- LcIv(y1 + y2 ~ x1 + y2 | x1 + x2 | 1,
                   data = data,
                   Q = Q,
                   model = "lciv",
                   init.value = "ivp",
                   method = "bfgs",
                   iterlim = 500,
                   print.init = FALSE)

## Note: model just identified
## Estimating an IV probit for initial values
## Estimating an IVLC-Probit model
```

The argument `y1 + y2` $\sim$ `x1 + y2 | x1 + x2| 1` is the `formula` argument. The first term, `y1 + y2`, includes the dependent (dummy) variable and the endogenous continuous variable. The second term, `x1 + y2` includes the set of variables for the Probit equation, whereas the third term, `x1 + x2` includes the set of exogenous variables including also the instrument `x2`. Finally, the fourth part is reserved for the specification of the class-assignment. If the class assignment $\pi_{iq}$ is also determined by socio-economic characteristics, those variables can also be included here. In this example, `| 1` implies that only the $Q-1$ constants will be estimated (see Equation 7). The argument `data` is reserved for the dataset of class `data.frame`, which was created in the previous subsection. `Q` indicates the number of classes. The argument `model` is a string indicating which model is estimated. The current option allows to estimate two models: if `model = "lc"`, then a LC-Probit model will be estimated; if `lciv` then an IVLC-Probit model is estimated. `init.value` indicates the procedure to obtain the initial values. The current option of the function is to obtain the initial values from an IV-Probit model as explained in Section 3.3. The optimization algorithm can be managed using the argument `method`, which is passed on to the `maxLik` function (Henningsen and Toomet, 2011). Currently, either `"bfgs"` for Broyden-Fletcher-Goldfarb-Shanno, `"bhhh"` for Brendt-Hall-Hall-Hausman or `"nr"` for Newton-Raphson can be implemented. For more information about the optimization procedure see `help(maxLik)`. The maximum number of iterations can be modified by the argument `iterlim`. Finally, if `print.init = TRUE` then the initial values are printed.

The estimated parameters are printed using the `summary` command:

```
# Results
summary(lciv)

## --------------------------------------------
## Maximum Likelihood estimation
## BFGS maximization, 236 iterations
## Return code 0: successful convergence
## Log-Likelihood: -21546.82
## 17  free parameters
## Estimates:
##                             Estimate Std. error t value  Pr(> t)
## class.1.eq.1.(Intercept) -0.9374847  0.0547294 -17.129  < 2e-16 ***
## class.1.eq.1.x1          -0.9633648  0.0633170 -15.215  < 2e-16 ***
## class.1.eq.1.y2          -0.9783915  0.0465325 -21.026  < 2e-16 ***
## class.2.eq.1.(Intercept)  0.7868159  0.1563669   5.032 4.86e-07 ***
## class.2.eq.1.x1           0.7570621  0.1833175   4.130 3.63e-05 ***
## class.2.eq.1.y2           0.9082414  0.0814721  11.148  < 2e-16 ***
## class.1.eq.2.(Intercept) -1.0129516  0.0127908 -79.194  < 2e-16 ***
## class.1.eq.2.x1          -0.9992916  0.0144073 -69.360  < 2e-16 ***
## class.1.eq.2.x2          -0.9960593  0.0143875 -69.231  < 2e-16 ***
## class.2.eq.2.(Intercept)  0.9883354  0.0198913  49.687  < 2e-16 ***
## class.2.eq.2.x1           1.0279624  0.0222005  46.303  < 2e-16 ***
## class.2.eq.2.x2           0.9813914  0.0226187  43.388  < 2e-16 ***
## (class)2                 -0.8230652  0.0260241 -31.627  < 2e-16 ***
## class.1.lnsigma          -0.0069943  0.0091362  -0.766    0.444
## class.2.lnsigma           0.0002712  0.0141402   0.019    0.985
## class.1.athro            -0.7028942  0.0378096 -18.590  < 2e-16 ***
## class.2.athro            -0.0646351  0.1198353  -0.539    0.590
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## --------------------------------------------
```

The results report the point estimates for each equation in each class. The coefficients with the string `.eq.1` are the point estimates for the Probit Model (Equation (1)), whereas those with the string `.eq.2` correspond to the parameters for the reduced equation (Equation (2)). The estimate `(class)2` is the constant $\lambda$ for class 2 (the parameter for class 1 is assumed to be 0 for identification). Note that all the parameters are close to the true parameters given in Table 2.

### A.3 Obtaining the shares for each class

Using the previous estimates, we can compute the proportion of individuals in each class using Equation (7). The following commands show how to compute the point estimate for each share, their standard errors using the Delta Method with the `msm` package (Jackson et al., 2011), and their 95% confidence interval.

```r
# Compute the share of individual in each class and their CI
library("msm") # package for Delta Method

# Computing share
lambda_2  <- coef(lciv)["(class)2"]
share1    <- exp(0) / (exp(0) + exp(lambda_2))
share2    <- exp(lambda_2) / (exp(0) + exp(lambda_2))
share1
```

```
##  (class)2
## 0.6948866
```

```r
share2
```

```
##  (class)2
## 0.3051134
```

```r
# Computing standard errors
share1_se <- deltamethod(~ exp(0) / (exp(0) + exp(x13)),
                         coef(lciv),
                         vcov(lciv),
                         ses =  TRUE)
share2_se <- deltamethod(~ exp(x13) / (exp(0) + exp(x13)),
                         coef(lciv),
                         vcov(lciv),
                         ses =  TRUE)

# Computing 95% CI for share 1
round(cbind(share1 - qnorm(0.975) * share1_se,
            share1 + qnorm(0.975) * share1_se), digits = 4)
```

```
##            [,1]   [,2]
## (class)2 0.6841 0.7057
```

```r
# Computing 95% CI for share 2
round(cbind(share2 - qnorm(0.975) * share2_se,
            share2 + qnorm(0.975) * share2_se), digits = 4)
```

```
##             [,1]   [,2]
## (class)2 0.2943 0.3159
```

The estimates of the share of individuals in each class are $\widehat{\pi}_1 = 0.7$ with 95% confidence interval $(0.68, 0.71)$, and $\widehat{\pi}_2 = 0.3$ with 95% confidence interval $(0.29, 0.32)$. When computing the standard errors using `deltamethod` function, the coefficients are labeled `x1`, `x2`... Since $\lambda_2$ is the 13th coefficient, we use `x13`. For more information on how to use the `deltamethod` function use `help(deltamethod)`.

The posterior membership probabilities for each individual in the sample (Equation 17) are obtained as follows:

```
# Obtain posterior membership probabilities
wiq <- lciv$Piq / lciv$Pi # This is a N * Q matrix
```

where `lciv$Piq` delivers $\widehat{\pi_{iq}}(\widehat{\boldsymbol{\lambda}}_q)\widehat{f}_q(y_{1i}, y_{2i}|\mathbf{z}_i)$ whereas `lciv$Pi` gives $\sum_{q=1}^{Q}\widehat{\pi_{iq}}(\widehat{\boldsymbol{\lambda}}_q)\widehat{f}_q(y_{1i}, y_{2i}|\mathbf{z}_i)$. Using these probabilities, the user can also compute the percentage of individual in each class by assigning each individual to the class with the highest posterior membership probability and then computing the average individuals in each class:

```
# Compute the share of individual in each class using conditional probabilities
shares  <- table(apply(wiq, 1, function(x) which(x == max(x)))) / N
names(shares) <- paste("share q", 1:Q, sep = "=")
print(shares)

## share q=1 share q=2
##    0.7274    0.2726
```

Note that the share are also closed to the true share in both classes. Finally, some researchers might be also interested in the unconditional class probability $\pi_{iq}$. After the estimation, $\widehat{\pi}_{iq}$ can be computed as follows:

```
# Obtaining the unconditional probabilities for each class
piq <- lciv$Wiq # This is a N * Q matrix
colnames(piq) <- paste("Class", 1:Q, sep = "=")
head(piq)

##        Class=1   Class=2
## [1,] 0.6948866 0.3051134
## [2,] 0.6948866 0.3051134
## [3,] 0.6948866 0.3051134
## [4,] 0.6948866 0.3051134
## [5,] 0.6948866 0.3051134
## [6,] 0.6948866 0.3051134
```

Since class assignment was not modeled using variables at the individual level, the estimated probabilities for each class are the same for all the artificial individuals in the sample.

### A.4 Estimates for $\rho_q$, $\sigma_{q,v}$ and testing

The estimates for $\rho_q, q = 1, 2$, are obtained using Equation (20), as follows:

```
# Compute rho in each class and their CI
athro1  <- coef(lciv)["class.1.athro"]
athro2  <- coef(lciv)["class.2.athro"]
rho1    <- (exp(2 * athro1) - 1) / (exp(2 * athro1) + 1)
rho2    <- (exp(2 * athro2) - 1) / (exp(2 * athro2) + 1)
rho1

## class.1.athro
##    -0.6062016

rho2

## class.2.athro
##    -0.0645452

# Obtain SEs using deltamethod
rho1_se <- deltamethod(~ (exp(2 * x16) - 1) /  (exp(2 * x16) + 1),
                       coef(lciv),
                       vcov(lciv),
                       ses =  TRUE)
rho2_se <- deltamethod(~ (exp(2 * x17) - 1) /  (exp(2 * x17) + 1),
                       coef(lciv),
                       vcov(lciv),
                       ses =  TRUE)

# Computing 95% CI for rho_1
round(cbind(rho1 - qnorm(0.975) * rho1_se,
            rho1 + qnorm(0.975) * rho1_se), digits = 4)

##                  [,1]    [,2]
## class.1.athro -0.6531 -0.5593

# Computing 95% CI for rho_2
round(cbind(rho2 - qnorm(0.975) * rho2_se,
            rho2 + qnorm(0.975) * rho2_se), digits = 4)

##                  [,1]   [,2]
## class.2.athro -0.2984 0.1693
```

As expected, the estimated $\rho$ for class one is close to true parameter with with 95% confidence interval $(-0.65, -0.56)$. Since the confidence interval for $\rho_2$ includes zero, we can claim that there is not evidence that $y_2$ is endogenous for almost 30% of the sample.

The Wald test for each correlation parameter can be performed as follows using `car` package (Fox et al., 2013):

```r
# Wald test for rho_q
library("car")
linearHypothesis(lciv, c("class.1.athro"), test = "Chisq")

## Linear hypothesis test
##
## Hypothesis:
## class.1.athro = 0
##
## Model 1: restricted model
## Model 2: lciv
##
##   Df Chisq Pr(>Chisq)
## 1
## 2  1 345.6  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

linearHypothesis(lciv, c("class.2.athro"), test = "Chisq")

## Linear hypothesis test
##
## Hypothesis:
## class.2.athro = 0
##
## Model 1: restricted model
## Model 2: lciv
##
##   Df  Chisq Pr(>Chisq)
## 1
## 2  1 0.2909     0.5896
```

Given the results, we can reject the null for class one. However, for class two the test statistic is not significant (p-value $= 0.59$), revealing that there is not sufficient information in the sample to reject the null.

To compute $\sigma_{1,\upsilon}$ and $\sigma_{2,\upsilon}$, we follow the same procedure using Equation (19):

```
# Estimates of sigma for each class
sigma1   <- exp(coef(lciv)["class.1.lnsigma"])
sigma2   <- exp(coef(lciv)["class.2.lnsigma"])
sigma1

## class.1.lnsigma
##       0.9930301

sigma2

## class.2.lnsigma
##        1.000271

# Obtain standard errors
sigma1_se   <- deltamethod(~ exp(x14),
                           coef(lciv),
                           vcov(lciv),
                           ses =  TRUE)
sigma2_se   <- deltamethod(~ exp(x15),
                           coef(lciv),
                           vcov(lciv),
                           ses =  TRUE)

# Compute 95% CI
# Computing 95% CI for sigma_1
round(cbind(sigma1 - qnorm(0.975) * sigma1_se,
            sigma1 + qnorm(0.975) * sigma1_se), digits = 4)

##                   [,1]   [,2]
## class.1.lnsigma 0.9752 1.0108

# Computing 95% CI for sigma_2
round(cbind(sigma2 - qnorm(0.975) * sigma2_se,
            sigma2 + qnorm(0.975) * sigma2_se), digits = 4)

##                   [,1]  [,2]
## class.2.lnsigma 0.9725 1.028
```

### A.5 Estimating marginal effects

To show how to compute the marginal effects, we calculate the average marginal effect for the endogenous variable `y2` for both classes. To do so, we use Equation (15):

```
# Compute the AME for y2
ai <- lciv$ai    # N * Q matrix
gamma_hat <- coef(lciv)[c("class.1.eq.1.y2", "class.2.eq.1.y2")]
rho_hat   <- c(rho1, rho2)
sigma_hat <- c(sigma1, sigma2)
AME_hat  <- apply(dnorm(ai), 2, mean) *
       ((gamma_hat + (rho_hat / sigma_hat)) / sqrt(1 - rho_hat^2))
AME_hat

##    class:1    class:2
## -0.2447669  0.1841673
```

Here is important to mention that `ai` delivers:

$$\frac{\mathbf{x}_i'\widehat{\boldsymbol{\beta}}_q + \widehat{\gamma}_q y_{i2} + \frac{\widehat{\rho}_q}{\widehat{\sigma}_{q,v}}(y_{2i} - \mathbf{z}_i'\widehat{\boldsymbol{\delta}}_q)}{\sqrt{1-\widehat{\rho}_q^2}}$$

as an $N \times Q$ matrix and the function `dnorm` is the normal standard PDF, $\phi(\cdot)$. According to the results and holding $x_1$ fixed, an increase of one unit of $y_2$ decreases the probability of $y_1 = 1$ for class one in about 25 per cent, whereas it increases this probability in about 18 per cent for class two. Here is important to note that this heterogeneity in the marginal effect of $y_2$ is ignored in the traditional IV-Probit model, which might lead to misleading conclusions.

The weighted average marginal effect (Equation 16) is computed as follows:

```
# Compute weighted ME for y2
sum(apply(piq, 2, mean) * AME_hat)

## [1] -0.1138933
```

Since the highest percentage of individuals is in the first (70%) class, the WAVE is negative: on average, an increase of one unit of $y_2$ decreases the probability of $y_1 = 1$ for class one in about 11 per cent.

The standard error of the marginal effect can be computed using bootstrap (the Delta Method is more difficult in this case). In the following code, we use package `boot` to perform the simulation. First, we create the function `ame` which returns the AMEs. The first argument of this function should be the dataset. The second argument can be an index vector of the observations in the dataset. The example below assumes we wish to use all of our observations.

```
# Compute standard errors of AME using bootstrapping
library("boot") # package to perform the bootstrapping

# function to obtain the marginal effects
```

```r
ame <- function(data, indices) {
     d <- data[indices, ] # bootstrap sample
     fit <- LcIv(y1 + y2 ~ x1 + y2 | x1 + x2 | 1,
                   data = d,
                   Q = Q,
                   model = "lciv",
                   init.value = "ivp",
                   method = "bfgs",
                   iterlim = 500,
                   print.init = FALSE,
                   messages = FALSE)
     ai <- fit$ai
     gamma <- coef(fit)[c("class.1.eq.1.y2",
                          "class.2.eq.1.y2")]
     athro1  <- coef(fit)["class.1.athro"]
     athro2  <- coef(fit)["class.2.athro"]
     rho1    <- (exp(2 * athro1) - 1) / (exp(2 * athro1) + 1)
     rho2    <- (exp(2 * athro2) - 1) / (exp(2 * athro2) + 1)
     rho <- c(rho1, rho2)
     sigma1   <- exp(coef(fit)["class.1.lnsigma"])
     sigma2   <- exp(coef(fit)["class.2.lnsigma"])
     sigma <- c(sigma1, sigma2)
     AMEi <- apply(dnorm(ai), 2, mean) * ((gamma + (rho / sigma)) / sqrt(1 - rho^2))
     return(AMEi)
}
```

Once the function `ame` is defined, we can use the `boot` command, which executes the resampling of the database and calculation of the AMEs on these samples.

```r
# perform bootstrapping
set.seed(666)
results <- boot(data = data, statistic = ame, R = 200)
```

The results for `R = 200` bootstrap replicates are the following:

```r
# Results from bootstrapping
results

##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
```

```
## boot(data = data, statistic = ame, R = 200)
##
##
## Bootstrap Statistics :
##       original        bias    std. error
## t1* -0.2447669  4.748447e-06 0.003801056
## t2*  0.1841673 -8.755836e-03 0.033272194
```

`original` contains the values for the average marginal effect using the full dataset. `bias` is the difference between the mean of bootstrap realizations and the value in the original dataset. `std.error` is a standard error of bootstrap estimate, which equals deviation of bootstrap realizations.

We can also compute the 95%-CI intervals from the bootstrap samples using the `boot.ci` command:

```
# get 95% confidence intervals
boot.ci(results, type = "norm", index = 1)

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 200 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = results, type = "norm", index = 1)
##
## Intervals :
## Level      Normal
## 95%   (-0.2522, -0.2373 )
## Calculations and Intervals on Original Scale

boot.ci(results, type = "norm", index = 2)

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 200 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = results, type = "norm", index = 2)
##
## Intervals :
## Level      Normal
## 95%   ( 0.1277,  0.2581 )
## Calculations and Intervals on Original Scale
```

Finally, we can compute the individual-specific marginal effect using Equation (18):

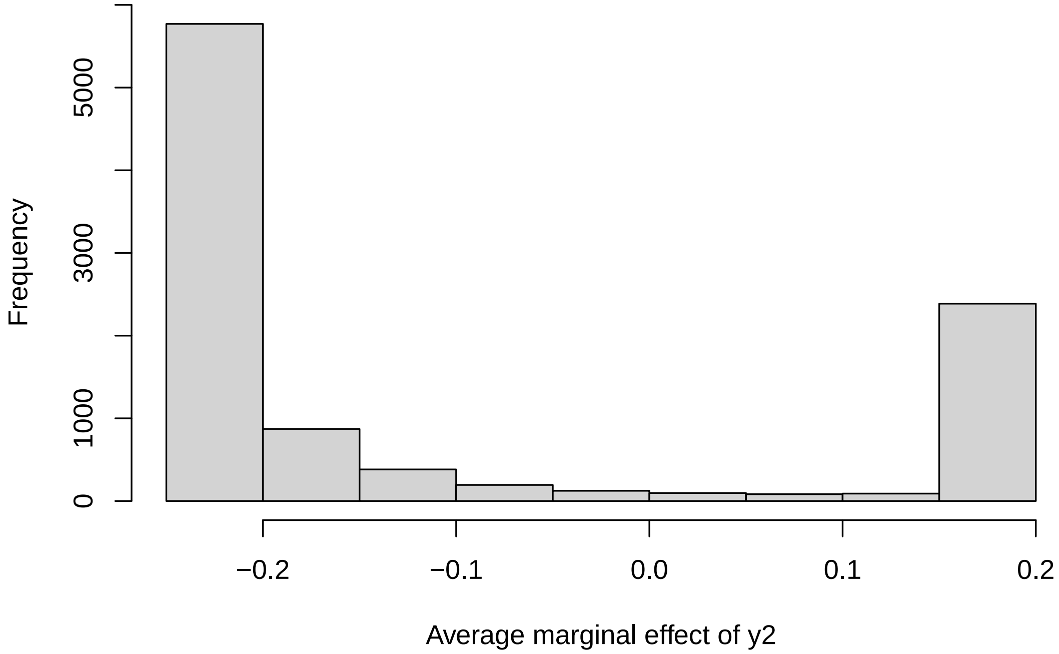


Figure A.1: Histogram for individual-specific marginal effects

```
# Estimate conditional marginal effect
cond_me <- wiq %*% AME_hat
summary(drop(cond_me))

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.24477 -0.24473 -0.22283 -0.11389  0.08412  0.18417
```

Note that the average of the conditional marginal effect is closed to the weighted ME. Figure 1 plots the distribution of the marginal effect for each individual.

```
# Plot distribution of individual-specific marginal effects
hist(cond_me,
     xlab = "Average marginal effect of y2",
     main = "")
```

## B Annex 2: Gradient Derivation

The function `LcIv` uses the numerical gradient to improve the convergence speed and avoid local maxima. This annex shows how the gradient is derived.

The log-likelihood for each individual is rewritten as follows to simplify the notation:

$$L_i(\boldsymbol{\theta}) = \ln P_i = \ln\left(\sum_{q=1}^{Q} w_{iq}\cdot P_{i|q}\right), \tag{B.1}$$

where:

$$\begin{aligned}
P_{i|q} &= \Phi(a_{iq})\cdot\frac{1}{\sigma_{q,\upsilon}}\phi\left(b_{iq}\right),\\
a_{iq} &= q_i\cdot\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}_q + \frac{\rho_q}{\sigma_{q,\upsilon}}(y_{2i}-\mathbf{z}_i'\boldsymbol{\delta}_q)}{\sqrt{1-\rho_q^2}}\right),\\
q_i &= 2y_{1i}-1,\\
b_{iq} &= \frac{y_{2i}-\mathbf{z}_i'\boldsymbol{\delta}_q}{\sigma_{q,\upsilon}},\\
\sigma_{q,\upsilon} &= \exp(\ln\sigma_{q,\upsilon}),\\
\rho_q &= \tau_q^{-1} = \frac{\exp(2\tau_q)-1}{\exp(2\tau_q)+1}.
\end{aligned}$$

Then, the contribution for each individual to the gradient can be computed as follows:

$$\begin{aligned}
\frac{\partial L_i}{\partial\boldsymbol{\beta}_q} &= \left(\frac{w_{iq}P_{i|q}}{P_i}\right)\left(\frac{\partial\ln P_{i|q}}{\partial\boldsymbol{\beta}_q}\right),\\
\frac{\partial L_i}{\partial\boldsymbol{\delta}_q} &= \left(\frac{w_{iq}P_{i|q}}{P_i}\right)\left(\frac{\partial\ln P_{i|q}}{\partial\boldsymbol{\delta}_q}\right),\\
\frac{\partial L_i}{\partial\boldsymbol{\lambda}_q} &= \frac{1}{P_i}\sum_{c=1}^{Q}P_{i|c}w_{ic}\left[\mathbb{1}\left(q=c\right)-w_{ic}\right]\mathbf{h}_i,\\
\frac{\partial L_i}{\partial\ln\sigma_{q,\upsilon}} &= \left(\frac{w_{iq}P_{i|q}}{P_i}\right)\left(\frac{\partial\ln P_{i|q}}{\partial\ln\sigma_{q,\upsilon}}\right)\\
\frac{\partial L_i}{\partial\tau_q} &= \left(\frac{w_{iq}P_{i|q}}{P_i}\right)\left(\frac{\partial\ln P_{i|q}}{\partial\tau_q}\right)
\end{aligned} \tag{B.2}$$

where:

$$\begin{aligned}
\frac{\partial\log P_{i|q}}{\partial\boldsymbol{\beta}_q} &= \frac{\phi(a_{iq})}{\Phi(a_{iq})}\cdot\frac{q_i}{\sqrt{1-\rho_q^2}}\mathbf{x}_i,\\
\frac{\partial\log P_{i|q}}{\partial\boldsymbol{\delta}_q} &= -\left[\frac{\phi(a_{iq})}{\Phi(a_{iq})}\frac{q_i(\rho_q/\sigma_{q,\upsilon})}{\sqrt{1-\rho_q^2}}+\frac{\phi'(b_{iq})}{\phi(b_{iq})}\frac{1}{\sigma_{q,\upsilon}}\right]\mathbf{z}_i,\\
\frac{\partial\log P_{i|q}}{\partial\ln\sigma_{q,\upsilon}} &= -\left[\frac{\phi(a_{iq})}{\Phi(a_{iq})}\frac{q_i\rho_q}{\sqrt{1-\rho_q^2}}+\frac{\phi'(b_{iq})}{\phi(b_{iq})}\right](y_{i2}-\mathbf{z}_i'\boldsymbol{\delta}_q)\exp(-\ln\sigma_{q,\upsilon})-1,
\end{aligned} \tag{B.3}$$

where $\phi'(z) = -z\phi(z)$. Finally, the derivative respect to $\tau_q$ is:

$$\frac{\partial \log P_{i|q}}{\partial \tau_q} = \frac{\phi(a_{iq})}{\Phi(a_{iq})} \cdot q_i \cdot d_{iq},$$

where:

$$d_{iq} = \left[ \frac{\exp(-\tau_q)\left[([y_{2i} - \mathbf{z}_i'\boldsymbol{\delta}_q] + \mathbf{x}_i'\boldsymbol{\beta}_q \sigma_{q,\upsilon}) \exp(2\tau_q) + [y_{2i} - \mathbf{z}_i'\boldsymbol{\delta}_q] - \mathbf{x}_i'\boldsymbol{\beta}_q \sigma_{q,\upsilon}\right]}{2\sigma_{q,\upsilon}} \right].$$