Estimating choice models with latent variables with Biogeme

Michel Bierlaire Moshe Ben-Akiva Joan Walker

August 19, 2025

Report TRANSP-OR xxxxxx
Transport and Mobility Laboratory
School of Architecture, Civil and Environmental Engineering
Ecole Polytechnique Fédérale de Lausanne
transp-or.epfl.ch

Series on Biogeme

Contents

The package Biogeme (`biogeme.epfl.ch`) is designed to estimate the parameters of various models using maximum likelihood estimation. It is particularly designed for discrete choice models. In this document, we present how to estimate choice models involving latent variables.

We assume that the reader is already familiar with discrete choice models, and has successfully installed Biogeme. This document has been written using Biogeme 3.3.0.

1 Models and notations

The literature on discrete choice models with latent variables is vast (Walker, 2001, Ashok et al., 2002, Greene and Hensher, 2003, Ben-Akiva et al., 2002, to cite just a few). We start this document by a short introduction to the models and the notations.

1.1 Structural equations

A *latent variable* is a variable that cannot be directly observed. It is typically modeled using a **structural equation**, which expresses the latent variable as a function of observed (explanatory) variables and an error term. A general form of such a structural equation is:

$$x^*_{nk} = x^*(x_n; \psi_k) + \omega_{nk}, \tag{1}$$

where n indexes individuals, x^*_{nk} is the kth latent variable of interest, x_n is a vector of observed explanatory variables, ψ_k is a vector of parameters to be estimated, and ω_{nk} is a stochastic error term.

A common specification assumes a linear functional form with i.i.d. normally distributed error terms:

$$x^*_{nk} = \sum_s \psi_{sk} x_{ns} + \sigma_{\omega k} \omega_{nk}, \tag{2}$$

where $\omega_{nk} \sim N(0, 1)$, and $\sigma_{\omega k}$ is a scaling parameter for the error term.

In discrete choice models, for example, the utility U_{in} that individual n associates with alternative i is a typical example of a latent variable.

Information about latent variables is obtained indirectly through *measurements*, which are observable manifestations of the underlying latent constructs. For example, in discrete choice models, utility is not directly observed but is inferred from the choices individuals make. The relationship between a latent variable and its associated measurements is described by **measurement equations**. The specific form of these equations depends on the nature of the observed measurements (e.g., continuous, or ordinal).

1.2 Measurement equations: the continuous case

Since latent variables cannot be directly observed, analysts rely on indirect measurements to infer their values. A common approach involves asking respondents to rate the perceived magnitude of the latent construct on an arbitrary scale. For example: *"How would you rate the level of pain that you are experiencing, from 0 (no pain) to 10 (worst pain imaginable)?"*

Each such rating is referred to as an *indicator*, indexed by $\ell = 1, \ldots, L_n$, and is modeled using a **measurement equation**. This equation relates the observed indicator to the latent variables and, sometimes, other explanatory variables:

$$I_{n\ell} = I_\ell(x_n, x^*_n; \lambda_\ell) + \upsilon_{n\ell}, \ \forall \ell = 1, \ldots, L_n, \forall n, \tag{3}$$

where $I_{n\ell}$ denotes the response provided by individual n for indicator ℓ, x^*_n is the latent variable of interest (e.g., pain perception), x_n is a vector of observed explanatory variables (such as

socio-demographic characteristics), λ_ℓ is a vector of parameters to be estimated, and $\upsilon_{n\ell}$ is the random error term.

A common specification of the measurement function assumes linearity and i.i.d. normally distributed errors:

$$I_{n\ell} = \lambda_{\ell 0} + \sum_k \lambda_{\ell k} x^*_{nk} + \sigma_{\upsilon\ell}\upsilon_{n\ell}, \quad \forall \ell, \tag{4}$$

where $\lambda_{\ell k}$ are unknown parameters to be estimated, $\sigma_{\upsilon\ell}$ is an indicator-specific scale parameter, and $\upsilon_{n\ell} \sim N(0,1)$.

If we observe a vector of continuous indicators $I_n = (I_{n1}, \ldots, I_{nL_n})$ for individual n, the contribution to the likelihood function, *conditional on the latent variables* x^*_n, is given by the product:

$$\prod_{\ell=1}^{L_n} \phi\left(\frac{I_{n\ell} - \lambda_{\ell 0} - \sum_k \lambda_{\ell k} x^*_{nk}}{\sigma_{\upsilon\ell}}\right), \tag{5}$$

where $\phi(\cdot)$ denotes the probability density function (pdf) of the standard normal distribution.

If other types of observations are available for the same individual (such as discrete choices), the corresponding components of the likelihood can be multiplied with the expression above. Once all relevant components are combined, the latent variables must be integrated out, as discussed later.

If the continuous indicators are the only data available for individual n, the contribution to the unconditional likelihood becomes:

$$\int_{x^*_n} \left[\prod_{\ell=1}^{L_n} \phi\left(\frac{I_{n\ell} - \lambda_{\ell 0} - \sum_k \lambda_{\ell k} x^*_{nk}}{\sigma_{\upsilon\ell}}\right)\right] f(x^*_n)\, dx^*_n, \tag{6}$$

where $f(x^*_n)$ is the pdf of the vector of latent variables x^*_n. As this integral does not have a closed-form expression, it is approximated using Monte Carlo integration (see Bierlaire, 2019 for a discussion about performing Monte-Carlo integration with Biogeme).

1.3 Measurement equation: the ordinal case

Another type of indicator arises when respondents are asked to evaluate a statement using an ordinal scale. A typical context for this type of measurement is the use of a Likert scale (Likert, 1932), where individuals express their degree of agreement or disagreement with a given statement. For example:

> *"I believe that my own actions have an impact on the planet."*
> Response options: strongly agree (2), agree (1), neutral (0), disagree (−1), strongly disagree (−2).

To model these types of indicators, we represent the observed measurement as an *ordered discrete variable* $I_{n\ell}$, which takes values in a finite, ordered set $\{j_1, j_2, \ldots, j_{M_\ell}\}$. The measurement equation involves two stages:

Step 1: Latent response formulation. We first define a continuous response variable, as explained in Section 1.2, except that it happens to be unobserved (latent) in this case:

$$I^*_{n\ell} = I^*_\ell(x_n, x^*_n; \lambda_\ell) + \upsilon_{n\ell}, \tag{7}$$

where $I^*_{n\ell}$ is a continuous latent variable underlying the reported response, x^*_n is a vector of relevant latent variables (e.g., environmental concern), x_n is a vector of observed explanatory variables (e.g., age, income), λ is a vector of parameters to be estimated, and $\upsilon_{n\ell}$ is a random error term.

Step 2: Discretization via thresholds. Since $I^*_{n\ell}$ is not observed, we relate it to the reported discrete measurement $I_{n\ell}$ through a set of threshold parameters:

$$I_{n\ell} = \begin{cases} j_1 & \text{if } I^*_{n\ell} < \tau_1, \\ j_2 & \text{if } \tau_1 \leqslant I^*_{n\ell} < \tau_2, \\ \vdots & \\ j_m & \text{if } \tau_{m-1} \leqslant I^*_{n\ell} < \tau_m, \\ \vdots & \\ j_M & \text{if } \tau_{M_\ell - 1} \leqslant I^*_{n\ell}, \end{cases} \tag{8}$$

where $\tau_1, \ldots, \tau_{M_\ell - 1}$ are threshold parameters to be estimated, satisfying the ordering constraint:

$$\tau_1 \leqslant \tau_2 \leqslant \cdots \leqslant \tau_{M_\ell - 1}. \tag{9}$$

Note that it is customary to use the same set of parameters for all individuals n and all indicators ℓ, which explains the absence of these indices on the parameter τ.

Defining $\tau_0 = -\infty$ and $\tau_{M_\ell} = +\infty$, it simplifies to

$$I_{n\ell} = j_m \text{ if } \tau_{m-1} \leqslant I^*_{n\ell} < \tau_m,\ m = 1, \ldots, M_\ell. \tag{10}$$

It is often advantageous to impose a symmetric structure on the definition of the thresholds. In addition, it is more convenient from an estimation standpoint to parameterize the thresholds in terms of differences and to constrain these differences to be positive. For example, when $M_\ell = 4$, the thresholds can be defined as follows:

$$\begin{aligned} \tau_1 &= -\delta_1 - \delta_2, \\ \tau_2 &= -\delta_1, \\ \tau_3 &= \ \ \delta_1, \\ \tau_4 &= \ \ \delta_1 + \delta_2, \end{aligned}$$

where $\delta_1 > 0$ and $\delta_2 > 0$ are the parameters to be estimated. This parameterization guarantees that the thresholds are strictly ordered and symmetrically centered around zero, which facilitates both identification and interpretation. To enforce positivity, we estimate the logarithm of the underlying parameters, that is $\delta'_i = \ln(\delta_i)$, so that

$$\begin{aligned} \tau_1 &= -\exp(\delta'_1) - \exp(\delta'_2), \\ \tau_2 &= -\exp(\delta'_1), \\ \tau_3 &= \ \ \exp(\delta'_1), \\ \tau_4 &= \ \ \exp(\delta'_1) + \exp(\delta'_2). \end{aligned}$$

If we consider a linear specification,

$$I^*_{n\ell} = \lambda_{\ell 0} + \sum_k \lambda_{\ell k} x^*_{nk} + \sigma_{v\ell} v_{n\ell}, \quad \forall \ell, \tag{11}$$

where the error term $v_{n\ell} \sim N(0, 1)$, the contribution of each indicator ℓ for each observation n

to the likelihood function, *conditional on the latent variables*, is defined as follows:

$$\begin{aligned}
\Pr(I_{n\ell} = j_m | x_n^*, x_n; \lambda_\ell, \Sigma_{v\ell}) &= \Pr(\tau_{m-1} \leqslant I_{n\ell}^* \leqslant \tau_m) \\
&= \Pr(I_{n\ell}^* \leqslant \tau_m) - \Pr(I_{n\ell}^* \leqslant \tau_{m-1}), \\
&= \Pr\left(v_{n\ell} \leqslant \frac{\tau_m - \lambda_{\ell 0} - \sum_k \lambda_{\ell k} x_{nk}^*}{\sigma_{v\ell}}\right) \\
&\quad - \Pr\left(v_{n\ell} \leqslant \frac{\tau_{m-1} - \lambda_{\ell 0} - \sum_k \lambda_{\ell k} x_{nk}^*}{\sigma_{v\ell}}\right), \qquad (12) \\
&= \Phi\left(\frac{\tau_m - \lambda_{\ell 0} - \sum_k \lambda_{\ell k} x_{nk}^*}{\sigma_{v\ell}}\right) \\
&\quad - \Phi\left(\frac{\tau_{m-1} - \lambda_{\ell 0} - \sum_k \lambda_{\ell k} x_{nk}^*}{\sigma_{v\ell}}\right)
\end{aligned}$$

where j_m is the observed category for respondent n and indicator ℓ.

This specification is known as the *ordered probit model* and is widely used for modeling ordinal responses that depend on latent constructs.

If we observe a vector of continuous indicators $I_n = (I_{n1}, \ldots, I_{nL_n})$ for individual n, the contribution to the likelihood function, *conditional on the latent variables* x_n^*, is given by:

$$\prod_{\ell=1}^{L_n} \Pr(I_{n\ell} = j_m | x_n^*, x_n; \lambda_\ell, \Sigma_{v\ell}). \qquad (13)$$

As in the continous case, if other types of observations are available for the same individual (such as choices), the corresponding components of the likelihood can be multiplied with the expression above. Once all relevant components are combined, the latent variables must be integrated out, as discussed later.

If the continuous indicators are the only data available for individual n, the contribution to the unconditional likelihood becomes:

$$\int_{x_n^*} \left[\prod_{\ell=1}^{L_n} \Pr(I_{n\ell} = j_m | x_n^*, x_n; \lambda_\ell, \Sigma_{v\ell})\right] f(x_n^*)\, dx_n^*, \qquad (14)$$

where $f(x_n^*)$ is the pdf of the vector of latent variables x_n^*. Again, this integral is approximated using Monte-Carlo integration.

2 The MIMIC model

The Multiple Indicators Multiple Causes (MIMIC) model is a structural equation modeling framework designed to analyze relationships involving latent variables. In a MIMIC model, the latent variable is simultaneously influenced by a set of observed explanatory variables (the "multiple causes") and reflected in several observed indicators (the "multiple indicators"). This dual structure enables the analyst to capture both the determinants and the manifestations of latent constructs, such as attitudes, preferences, or psychological traits. A seminal introduction to the MIMIC model is provided by Jöreskog and Goldberger (1975), who formalized its use within the broader class of structural equation models. —-

Under the specifications in (2) and (4), the latent vector and indicators are jointly normal, so the latent variables can be integrated out analytically. Let $\nu = (\lambda_{\ell 0})_\ell$, $\Lambda = [\lambda_{\ell k}]_{\ell,k}$, $\Theta = \text{diag}(\sigma_{v\ell}^2)_\ell$, and $\Psi = \text{diag}(\sigma_{\omega k}^2)_k$. Write the structural part as $x_n^* = \Gamma x_n + \zeta_n$ with $\zeta_n \sim \mathcal{N}(\mathbf{0}, \Psi)$

(so $\Gamma_{ks} = \psi_{sk}$). Then, conditionally on $\mathbf{x}_n^*$, we have $\mathbf{I}_n \mid \mathbf{x}_n^* \sim \mathcal{N}(\boldsymbol{\nu} + \Lambda\mathbf{x}_n^*, \Theta)$. Marginalizing $\mathbf{x}_n^*$ yields the multivariate normal

$$\mathbf{I}_n \mid \mathbf{x}_n \sim \mathcal{N}\big(\, \boldsymbol{\mu}_n,\ \Sigma \,\big), \qquad \boldsymbol{\mu}_n = \boldsymbol{\nu} + \Lambda\Gamma\mathbf{x}_n, \quad \Sigma = \Lambda\Psi\Lambda^\top + \Theta.$$

Therefore, the contribution of individual n to the likelihood is

$$L_n \;=\; (2\pi)^{-L_n/2}\,|\Sigma|^{-1/2}\,\exp\Big(-\tfrac{1}{2}\big(\mathbf{I}_n - \boldsymbol{\mu}_n\big)^\top \Sigma^{-1}\big(\mathbf{I}_n - \boldsymbol{\mu}_n\big)\Big)\,, \tag{15}$$

and the sample likelihood is $\prod_n L_n$. Hence, no Monte-Carlo integration is required for the continuous-indicator part: the conjugate normal–normal structure delivers a closed form. (If additional non-Gaussian components—e.g., discrete choices—are included, their likelihood terms multiply (15); only those non-Gaussian parts may require numerical integration.)

—- We first investigate the continuous case, that is, the model specification that involves the structural equations (2) and the measurement equations (3).

In this specific context, the formulation (6) of the likelihood happens to simplify. Indeed, both the structural and measurement equations are linear and the error terms are assumed to be normally distributed. Because linear transformations and sums of normal random variables remain normal, the pair (x_n^*, I_n) is jointly normal. Integrating out the latent variables x_n^* therefore yields a closed-form multivariate normal distribution for the indicators I_n, with a mean shifted according to the structural equation and a covariance matrix equal to the sum of the measurement-error variances and the variance propagated from the latent variables. In other words, the integral in (6) collapses to the standard multivariate normal density, without requiring any numerical approximation.

It can be seen by using (2) into (3), to obtain

$$\begin{aligned} I_{n\ell} &= \lambda_{\ell 0} + \sum_k \lambda_{\ell k}\Big(\sum_s \psi_{sk} x_{ns} + \sigma_{\omega k}\omega_{nk}\Big) + \sigma_{v\ell} v_{n\ell}, \quad \forall \ell, \\ &= \lambda_{\ell 0} + \sum_s \lambda'_{\ell s} x_{ns} + \sigma'_{v\ell} v'_{n\ell}, \quad \forall \ell, \end{aligned} \tag{16}$$

where $\lambda'_{\ell s} = \sum_k \lambda_{\ell k}\psi_{sk}$, and $\sigma'_{v\ell}v'_{n\ell} = \sigma_{v\ell}v_{n\ell} + \sum_k \lambda_{\ell k}\sigma_{\omega k}\omega_{nk}$.

In order to normalize the model, we associate each latent variable k with one specific indicator ℓ_k, different across latent variables. The following normalization can be done:

- As the error terms of the structural and measurement equations are confounded, we set $\sigma_{\omega k} = 0$, for each latent variable k.
- As the units of the latent variables are arbitrary, we set the coefficient of latent variable to one[1] in the corresponding measurement equation, and the scale parameter to zero: $\lambda_{\ell_k k} = 1$, $\sigma_{v\ell_k} = 0$ for each latent variable k.

Note that the structural equations do not include an intercept, so that there is no need to mormalize the intercept of the corresponding measurement equation.

In contrast, this simplification does not apply when the indicators are discrete. In that case, the measurement equations involve threshold-crossing representations that express the probability (12) of each observed category as differences of normal cumulative distribution functions. These nonlinear functions of the latent variables break the convenient normal–linear structure: the conditional distribution of the indicators given x_n^* is no longer Gaussian. As a consequence, the integral over x_n^* in the likelihood cannot be evaluated in closed form.

[1] Or -1 if the statement implies a decrease of the latent variable, as illustrated in the case study.

3 Choice Model with Latent Variables

Consider a discrete choice model. We illustrate the methodology using a logit specification for the choice component, although other models — such as the nested logit or cross-nested logit — could also be employed. Under the logit framework, the probability that individual n chooses alternative i is given by

$$P(i \mid x_n; \beta) = \frac{e^{\mu V_{in}(x_n;\beta)}}{\sum_{j \in \mathcal{C}_n} e^{\mu V_{jn}(x_n;\beta)}}, \tag{17}$$

where μ is the (unknown) scale parameter, and $V_{in}(x_n; \beta)$ denotes the systematic utility associated with alternative i for individual n, as a function of observed explanatory variables x_n and a parameter vector β.

To enhance the behavioral realism of the model, we extend the utility specification by incorporating latent variables. This leads to the following choice probability conditional on the latent variables x_n^*:

$$P(i \mid x_n^*, x_n; \beta) = \frac{e^{\mu V_{in}(x_n^*, x_n;\beta)}}{\sum_{j \in \mathcal{C}_n} e^{\mu V_{jn}(x_n^*, x_n;\beta)}}, \tag{18}$$

where x_n^* denotes the vector of latent psychological constructs influencing choice behavior.

From a modeling perspective, the latent variables are included in the utility function in the same way as observed covariates. For this reason, it is useful to first specify the choice model assuming that the latent variables are observed. Once this conditional model is established, the latent nature of these variables can be addressed.

The first approach relies on the structural equations defined in Equation (2), with parameter values estimated from the MIMIC model:

$$\widehat{x}_{nk}^* = \widehat{\psi}_{0k} + \sum_s \widehat{\psi}_{sk} x_{ns} + \sigma_{\omega k} \omega_{nk}, \tag{19}$$

where $\widehat{\psi}$ is the vector of estimated parameters of the structural equations. Note that the scale parameters $\sigma_{\omega k}$ cannot be identified in the MIMIC model and are therefore normalized to 1, as they are confounded with the scale of the choice model.

Given this specification, the choice model becomes, conditional on the structural error terms ω_n:

$$P(i \mid \omega_n, x_n; \beta) = \frac{e^{\mu V_{in}(\omega_n, x_n;\beta)}}{\sum_{j \in \mathcal{C}_n} e^{\mu V_{jn}(\omega_n, x_n;\beta)}}, \tag{20}$$

where ω_n is the vector of normally distributed structural errors.

To obtain the unconditional choice probability, we integrate over the distribution of the latent variables. This results in a mixture of logit models:

$$P(i \mid x_n; \beta) = \int_{\omega_n} \frac{e^{\mu V_{in}(\omega_n, x_n;\beta)}}{\sum_{j \in \mathcal{C}_n} e^{\mu V_{jn}(\omega_n, x_n;\beta)}} \phi(\omega_n)\, d\omega_n, \tag{21}$$

where $\phi(\cdot)$ is the probability density function of the standard normal distribution. This integral is approximatied using Monte-Carlo integration.

This sequential estimation approach — first estimating the MIMIC model, then the choice model — yields consistent parameter estimates. However, it is not statistically efficient. Intuitively, observed choices provide additional information about the latent variables, which is not exploited in a two-step procedure. We now turn to the simultaneous estimation of all components — structural equations, measurement equations, and the choice model — based on

the full likelihood function. This integrated approach leverages all available information and improves the statistical efficiency of the estimates.

In this context, the likelihood function incorporates the complete set of observations for each individual. This includes not only the observed choice but also responses to the psychometric indicators. Conditional on the latent variables, the contribution of individual n to the likelihood function is the product of the ordered probit probabilities for each indicator, as specified in Equation (12), and the choice probability given in Equation (18):

$$P(i_n \mid x_n^*, x_n; \beta) \prod_\ell \mathbb{P}(I_{n\ell} = j_m \mid x_n^*, x_n; \lambda_\ell, \Sigma_{v\ell}).$$

Substituting the structural equations (2), the expression becomes conditional on ω_n, and involves now the parameters of the structural equations as well:

$$P(i_n \mid \omega_n, x_n; \beta, \psi, \sigma) \prod_\ell \mathbb{P}(I_{n\ell} = j_m \mid \omega_n, x_n; \lambda_\ell, \Sigma_{v\ell}, \psi, \sigma).$$

To obtain the full contribution of individual n to the likelihood function, we integrate this expression over the distribution of the latent variables:

$$\int_{\omega_n} P(i_n \mid x_n^*, x_n; \beta) \prod_\ell \mathbb{P}(I_{n\ell} = j_m \mid x_n^*, x_n; \lambda_\ell, \Sigma_{v\ell}) \phi(\omega_n)\, d\omega_n,$$

where $\phi(\cdot)$ is the standard normal density. As this integral has no closed form, it is approximated via Monte Carlo integration.

4 A case study

This example focuses on the estimation of a mode choice model for residents of Switzerland, using revealed preference data. The data were collected as part of a research project aimed to assess the market potential of combined mobility solutions — particularly in urban agglomerations — by identifying the factors that influence individuals in their choice of transport mode (Bierlaire et al., 2011).

The survey was conducted between 2009 and 2010 on behalf of CarPostal, the public transport operator of the Swiss Postal Service. Its primary objective was to collect data on travel behavior in low-density areas, which represent the typical service environment of CarPostal. In addition to revealed preference data, the survey includes several psychometric indicators, enabling the incorporation of latent variables into the model specification.

The data file as well as its description is available on the Biogeme webpage.

We first estimate a MIMIC model involving two latent variables. The first one captures a "car-centric" attitude. The second one captures a "urban preference attitude". The car-centric attitude captures the extent to which individuals exhibit a strong preference for private car use as their primary mode of transportation. This latent construct reflects values such as independence, flexibility, comfort, and perceived status associated with driving. Individuals with a high car-centric attitude are more likely to perceive cars as the most practical and desirable means of travel, often resisting modal shift to public transport or active mobility. The urban preference attitude represents the degree to which individuals value characteristics associated with dense, mixed-use urban environments. This includes a positive perception of walkability, access to local services, efficient public transport, and vibrant public spaces. This attitude is typically associated with environmental awareness, social interaction, and a preference for compact urban living.

4.1 Psychometric indicators

The psychmometric indicators selected to capture the car-centric attitude are:

Envir01 Fuel price should be increased to reduce congestion and air pollution.

Envir02 More public transportation is needed, even if taxes are set to pay the additional costs.

Envir03 Ecology disadvantages minorities and small businesses.

Envir04 People and employment are more important than the environment.

Mobil09 Taking the bus helps making the city more comfortable and welcoming.

Mobil11 It is difficult to take the public transport when I carry bags or luggage.

Mobil14 When I take the car I know I will be on time.

Mobil16 I do not like changing the mean of transport when I am traveling.

Mobil17 If I use public transportation I have to cancel certain activities I would have done if I had taken the car.

LifSty08 For me the car is only a practical way to move.

The psychmometric indicators selected to capture the urban preference attitude are:

ResidCh01 I like living in a neighborhood where a lot of things happen.

ResidCh02 The accessibility and mobility conditions are important for the choice of housing.

ResidCh03 Most of my friends live in the same region I live in.

ResidCh05 I would like to live in the city center of a big city.

ResidCh06 I would like to live in a town situated in the outskirts of a city.

ResidCh07 I would like to live in the countryside.

Mobil07 In general, for my activities, I always have a usual mean of transport.

Mobil24 I have always used public transports all my life.

4.2 Structural equations

For the structural equations, we use the linear specification (2) with the following explanatory variables. For the car-centric attitude $\mathfrak{x}^*_{n,\text{car}}$, we have:

- age_65_more = age >= 65
- ScaledIncome = CalculatedIncome / 1000
- moreThanOneCar = NbCar > 1
- moreThanOneBike = NbBicy > 1
- individualHouse = HouseType == 1
- haveChildren = (FamilSitu == 3)+ (FamilSitu == 4)> 0

- haveGA = GenAbST == 1
- highEducation = Education >= 6

And for the urban preference attitude $\mathfrak{x}^*_{\mathfrak{n},\text{urban}}$, we have:

- childCenter = ((ResidChild == 1)+ (ResidChild == 2))> 0
- childSuburb = ((ResidChild == 3)+ (ResidChild == 4))> 0
- highEducation = Education >= 6
- artisans = SocioProfCat == 5
- employees = SocioProfCat == 6
- age_30_less = age <= 30
- haveChildren = (FamilSitu == 3)+ (FamilSitu == 4)> 0
- UrbRur
- individualHouse = HouseType == 1

4.3 Measurement equations

For each individual $\mathfrak{n}$ and each indicator ℓ described in Section 4.1, we introduce a latent continuous response variable, as outlined in Section 1.3. This latent response captures the unobserved propensity underlying the observed ordinal response on a Likert scale.

For the indicators associated with the car-centric attitude, the latent response is modeled as:

$$\mathrm{I}^*_{\mathfrak{n}\ell} = \lambda_{0\ell} + \lambda_{1\ell}\mathfrak{x}^*_{\mathfrak{n},\text{car}} + \lambda_{2\ell}\upsilon_{\mathfrak{n}\ell}, \tag{22}$$

where $\lambda_{0\ell}$ is an intercept term, $\lambda_{1\ell}$ is the loading on the latent variable $\mathfrak{x}^*_{\mathfrak{n},\text{car}}$, $\lambda_{2\ell}$ scales the stochastic component, and $\upsilon_{\mathfrak{n}\ell}$ is a random error term.

The indicator `Envir01` is selected for the normalization of the measurement model. Individuals with a stronger car-centric attitude are expected to be more likely to *disagree* with the corresponding statement. Accordingly, the loading $\lambda_{1\ell}$ is expected to be negative, and fixed to -1 to establish the direction of the latent construct. The scale parameter $\lambda_{2\ell}$ is normalized to 1 to ensure identifiability of the model.

Similarly, for the indicators capturing the urban-preference attitude, we specify:

$$\mathrm{I}^*_{\mathfrak{n}\ell} = \lambda_{0\ell} + \lambda_{1\ell}\mathfrak{x}^*_{\mathfrak{n},\text{urban}} + \lambda_{2\ell}\upsilon_{\mathfrak{n}\ell}, \tag{23}$$

with analogous interpretation of the parameters.

The indicator `ResidCh01` is selected for the normalization of this measurement model. Individuals with a stronger urban-preference attitude are expected to be more likely to *agree* with the corresponding statement. Accordingly, the loading $\lambda_{1\ell}$ is expected to be positive, and fixed to 1 to establish the direction of the latent construct. The scale parameter $\lambda_{2\ell}$ is normalized to 1 to ensure identifiability of the model.

The thresholds for the ordered probit model are defined as described in Section 1.3:

$$\begin{aligned}
\tau_1 &= -\delta_1 - \delta_2, \\
\tau_2 &= -\delta_1, \\
\tau_3 &= \delta_1, \\
\tau_4 &= \delta_1 + \delta_2,
\end{aligned}$$

where $\delta_1 > 0$ and $\delta_2 > 0$ are estimated.

4.4 Implementation notes

The Biogeme implementation is structured across three dedicated files: one defining the variables associated with each latent construct, one specifying the structural equations, and one containing the measurement equations. These files are not only used in the estimation of the MIMIC model, but are also employed in the subsequent estimation of choice models incorporating latent variables. This design ensures full consistency in the specification of the structural and measurement equations across all stages of the analysis.

The file defining the relevant variables is described in Section 6.1. It specifies two sets, car_indicators and urban_indicators, which list the indicators associated with each latent variable. It also identifies the indicators used for the normalization of the model. In addition, the file defines two dictionaries, each mapping a latent variable to the set of explanatory variables used in its corresponding structural equation. These dictionaries associate variable names with Biogeme expressions used to compute them.

The file defining the structural equations is described in Section 6.2. It includes two functions — one for each latent variable — that construct the corresponding structural equations. These functions operate in two modes. When called without arguments, they return Biogeme expressions involving parameters to be estimated. This mode is used, for example, in the MIMIC model. Alternatively, if a dictionary of parameter values is provided as input, the functions substitute the given values and treat them as fixed, enabling evaluation based on previously estimated parameters. This mode is used for the sequential estimation of the choice model, for example.

The file containing the measurement equations is described in Section 6.3. It begins by organizing the parameters to be estimated into dictionaries: one for the intercepts (intercepts), two for the coefficients associated with each latent variable (car_coefficients and urban_coefficients), and one for the scale parameters (sigma_star). The required normalizations are also applied at this stage.

The function generate_model_terms is responsible for constructing the expression representing the contribution of the latent variables to the measurement equation for a given indicator. It is designed to accommodate the possibility that an indicator may be influenced by multiple latent variables, even though this situation does not arise in the present example.

Finally, the function generate_measurement_equations produces the complete set of measurement equations in the form of a dictionary. This dictionary maps each potential response value of an indicator to the corresponding expression for its likelihood contribution.

4.5 The MIMIC model

With all generic components defined, we now turn to the script used to estimate the parameters of the MIMIC model, presented in Section 6.4. After retrieving the relevant elements discussed above, the script constructs the joint log likelihood function of all observed indicators. This is done by first creating a dictionary that maps each indicator to its individual log likelihood contribution, which are then aggregated into a single expression representing the total log likelihood.

This combined likelihood expression is subsequently linked to the database using Biogeme, enabling the maximum likelihood estimation of the model parameters. The general statistics of the estimation are reported in Table 1 on the following page. The estimated parameters of the structural equations are reported in Table 2 on the next page. The estimated parameters of the measurement equations are reported in Table 3 on page 12 for the car-centric attitude, and in Table 4 on page 13 for the urban-preference attitude. Finally, the estimated values of the differences between thresholds are reported in Table 5 on page 13.

Number of estimated parameters	69
Sample size	1899
Init log likelihood	-109217.9
Final log likelihood	-43773.72
Akaike Information Criterion	87685.45
Bayesian Information Criterion	88068.33

Table 1: Mimic model: general statistics

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	car_struct_age_65_more	0.158	0.0664	2.38	0.0174
2	car_struct_ScaledIncome	-0.0333	0.00707	-4.71	2.53e-06
3	car_struct_moreThanOneCar	0.614	0.0584	10.5	0.0
4	car_struct_moreThanOneBike	-0.341	0.0584	-5.84	5.33e-09
5	car_struct_individualHouse	-0.0939	0.0517	-1.82	0.0693
6	car_struct_haveChildren	-0.044	0.0502	-0.875	0.382
7	car_struct_haveGA	-0.599	0.0843	-7.1	1.23e-12
8	car_struct_highEducation	-0.328	0.0585	-5.61	1.99e-08
9	urban_struct_childCenter	0.0979	0.0294	3.33	0.00087
10	urban_struct_childSuburb	0.0892	0.0239	3.72	0.000196
11	urban_struct_highEducation	0.0341	0.0153	2.24	0.0253
12	urban_struct_artisans	-0.0986	0.035	-2.81	0.00491
13	urban_struct_employees	-0.0398	0.0174	-2.28	0.0224
14	urban_struct_age_30_less	0.162	0.0594	2.72	0.00645
15	urban_struct_haveChildren	-0.0246	0.0124	-1.99	0.0466
16	urban_struct_UrbRur	0.104	0.0333	3.13	0.00175
17	urban_struct_IndividualHouse	0.0277	0.014	1.98	0.0475

Table 2: MIMIC model: estimated parameters of the structural equations

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	meas_intercept_Envir01	-0.829	0.0739	-11.2	0.0
2	meas_intercept_Envir02	0.42	0.0309	13.6	0.0
3	car_meas_b_Envir02	-0.45	0.0557	-8.08	6.66e-16
4	meas_sigma_star_Envir02	0.945	0.0228	41.4	0.0
5	meas_intercept_Envir03	-0.393	0.0337	-11.6	0.0
6	car_meas_b_Envir03	0.647	0.0592	10.9	0.0
7	meas_sigma_star_Envir03	0.879	0.0212	41.4	0.0
8	meas_intercept_Envir04	-0.543	0.0313	-17.3	0.0
9	car_meas_b_Envir04	0.38	0.0601	6.32	2.67e-10
10	meas_sigma_star_Envir04	0.811	0.0204	39.7	0.0
11	meas_intercept_Mobil09	0.822	0.0311	26.4	0.0
12	car_meas_b_Mobil09	-0.34	0.0511	-6.65	2.85e-11
13	meas_sigma_star_Mobil09	0.873	0.0241	36.2	0.0
14	meas_intercept_Mobil11	0.408	0.035	11.6	0.0
15	car_meas_b_Mobil11	0.541	0.059	9.17	0.0
16	meas_sigma_star_Mobil11	0.963	0.0249	38.7	0.0
17	meas_intercept_Mobil14	-0.154	0.0308	-5.0	5.66e-07
18	car_meas_b_Mobil14	0.615	0.056	11.0	0.0
19	meas_sigma_star_Mobil14	0.85	0.021	40.6	0.0
20	meas_intercept_Mobil16	0.167	0.0331	5.04	4.6e-07
21	car_meas_b_Mobil16	0.478	0.0568	8.42	0.0
22	meas_sigma_star_Mobil16	0.93	0.0227	41.0	0.0
23	meas_intercept_Mobil17	0.179	0.032	5.59	2.24e-08
24	car_meas_b_Mobil17	0.404	0.0567	7.13	1.01e-12
25	meas_sigma_star_Mobil17	0.934	0.0239	39.1	0.0
26	meas_intercept_LifSty08	1.22	0.04	30.5	0.0
27	car_meas_b_LifSty08	-0.227	0.062	-3.66	0.000251
28	meas_sigma_star_LifSty08	0.957	0.0317	30.2	0.0

Table 3: MIMIC model: estimated parameters of the measurement equations for the car-centric attitude

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	meas_intercept_ResidCh01	-0.591	0.0682	-8.67	0.0
2	meas_intercept_ResidCh02	1.18	0.177	6.66	2.71e-11
3	urban_meas_b_ResidCh02	1.53	0.456	3.35	0.000815
4	meas_sigma_star_ResidCh02	0.95	0.0245	38.8	0.0
5	meas_intercept_ResidCh03	-0.284	0.224	-1.27	0.205
6	urban_meas_b_ResidCh03	-1.24	0.582	-2.13	0.0328
7	meas_sigma_star_ResidCh03	0.959	0.0223	43.1	0.0
8	meas_intercept_ResidCh05	-0.388	0.298	-1.3	0.194
9	urban_meas_b_ResidCh05	3.02	0.775	3.89	9.88e-05
10	meas_sigma_star_ResidCh05	1.15	0.0417	27.6	0.0
11	meas_intercept_ResidCh06	0.873	0.407	2.15	0.0318
12	urban_meas_b_ResidCh06	3.59	1.05	3.43	0.000603
13	meas_sigma_star_ResidCh06	1.06	0.0267	39.8	0.0
14	meas_intercept_ResidCh07	-0.156	0.338	-0.461	0.645
15	urban_meas_b_ResidCh07	-3.25	0.879	-3.69	0.000223
16	meas_sigma_star_ResidCh07	0.929	0.0283	32.8	0.0
17	meas_intercept_Mobil07	0.633	0.117	5.41	6.41e-08
18	urban_meas_b_Mobil07	-0.607	0.298	-2.04	0.0415
19	meas_sigma_star_Mobil07	0.833	0.0258	32.4	0.0
20	meas_intercept_Mobil24	0.841	0.203	4.14	3.46e-05
21	urban_meas_b_Mobil24	1.45	0.524	2.77	0.00566
22	meas_sigma_star_Mobil24	1.12	0.0288	38.8	0.0

Table 4: MIMIC model: estimated parameters of the measurement equations for the urban-preference attitude

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	δ_1	0.307	0.00614	49.9	0.0
2	δ_2	0.936	0.0167	55.9	0.0

Table 5: MIMIC model: thresholds of the measurement equations

4.6 The choice model

The transportation mode choice model developed for the case study presented in Section 4 considers three alternatives:

- public transportation,
- private car,
- slow modes (e.g., walking, biking).

The utility functions are specified as linear combinations of explanatory variables and include alternative-specific constants (intercepts). The full specification is provided in Table 6, where

- MarginalCostPT_scaled = MarginalCostPT / 10,
- CostCarCHF_scaled = CostCarCHF / 10,
- PurpHWH = TripPurpose == 1,
- PurpOther = TripPurpose == 1,
- TimePT_scaled = TimePT / 200,
- TimeCar_scaled = TimeCar / 200,
- distance_km_scaled = distance_km / 5.

The choice model adopts the logit formulation (17). The model is normalized using a money-metric specification: the coefficient of the cost variable, interacted with the "home–work–home" trip purpose, is fixed to -1. This ensures that the utility scale is expressed in monetary units.

Parameter	Pub. transp.	Car	Slow modes
asc_pt	1	0	0
asc_car	0	1	0
beta_cost_hwh	MarginalCostPT_scaled ∗ PurpHWH	CostCarCHF_scaled ∗ PurpHWH	0
beta_cost_other	MarginalCostPT_scaled ∗ PurpOther	CostCarCHF_scaled ∗ PurpOther	0 0
beta_time_pt	TimePT_scaled	0	0
beta_waiting_time	WaitingTimePT	0	0
beta_time_car	0	TimeCar_scaled	0
beta_dist	0	0	distance_km_scaled

Table 6: Specification of the utility functions of the choice model

The script is described in Section 6.5 and the results of the estimation are reported in Table 7 on the next page.

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	choice_asc_pt	-0.132	0.262	-0.505	0.614
2	choice_asc_car	0.408	0.283	1.44	0.149
3	choice_beta_cost_other	-0.465	0.142	-3.27	0.00106
4	choice_beta_time_pt	-1.58	0.658	-2.39	0.0167
5	choice_beta_time_car	-5.14	1.67	-3.08	0.00209
6	choice_beta_waiting_time	-0.018	0.00789	-2.28	0.0228
7	choice_beta_dist	-0.983	0.317	-3.1	0.00192
8	scale_choice_model	1.17	0.275	4.24	2.24e-05

Number of estimated parameters	8
Sample size	1899
Final log likelihood	-1230.014
Akaike Information Criterion	2476.028
Bayesian Information Criterion	2520.421

Table 7: Choice model: estimated parameters

4.7 Sequential estimation

The objective now is to incorporate the latent variables estimated in the MIMIC model into the choice model. We investigate a specification in which the latent variables interact with the alternative-specific constants. This specification, summarized in Table 8, involves random components — namely, the two latent variables — and thus transforms the choice model into a *mixture of logit models*.

Parameter	Pub. transp.	Car	Slow modes
asc_pt	1	0	0
asc_car	0	1	0
beta_cost_hwh	MarginalCostPT_scaled × PurpHWH	CostCarCHF_scaled × PurpHWH	0
beta_cost_other	MarginalCostPT_scaled × PurpOther	CostCarCHF_scaled × PurpOther	0 0
beta_time_pt	TimePT_scaled	0	0
beta_waiting_time	WaitingTimePT	0	0
beta_time_car	0	TimeCar_scaled	0
beta_dist	0	0	distance_km_scaled
car_centric_pt_cte	$\chi^*_{n,\text{car}}$	0	0
car_centric_car_cte	0	$\chi^*_{n,\text{car}}$	0
urban_life_pt_cte	$\chi^*_{n,\text{urban}}$	0	0
urban_life_car_cte	0	$\chi^*_{n,\text{urban}}$	0

Table 8: Specification of the utility functions of the choice model with latent variables

The script is presented in Section 6.6. The latent variables are constructed using two components: the deterministic part derived from the structural equations with parameter values estimated from the MIMIC model, and a stochastic term capturing unobserved heterogeneity:

```
car_centric_attitude = build_car_centric_attitude(
    estimated_parameters=struct_betas
) + Draws('car_error_term', 'NORMAL_MLHS_ANTI')
```

Since the latent variables are integrated out via Monte Carlo methods, the random term is defined using a draw generator. In this case, the generator produces draws from a standard normal distribution using Modified Latin Hypercube Sampling (MLHS), a variance-reduction technique that ensures a more uniform coverage of the distribution's support (Hess et al., 2006). The keyword ANTI specifies the use of antithetic draws, meaning that for each draw ω, its symmetric counterpart $-\omega$ is also included. This technique improves numerical stability and accelerates convergence by reducing the variance of the estimator.

Once constructed, the latent variables are incorporated into the utility functions exactly like any other explanatory variable. The estimated parameters for this sequential estimation are reported in Table 9 on the next page.

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	choice_asc_pt	0.436	0.417	1.05	0.295
2	choice_asc_car	1.55	0.588	2.64	0.00824
3	choice_beta_cost_other	-0.416	0.148	-2.81	0.00492
4	choice_beta_time_pt	-1.95	0.865	-2.26	0.024
5	choice_beta_waiting_time	-0.02	0.0101	-1.98	0.0476
6	choice_beta_time_car	-5.58	2.03	-2.75	0.00602
7	choice_beta_dist	-1.13	0.407	-2.77	0.00562
8	choice_car_centric_pt_cte	0.652	0.265	2.46	0.014
9	choice_urban_life_pt_cte	0.0223	0.0826	0.27	0.787
10	choice_car_centric_car_cte	1.62	0.52	3.11	0.00187
11	choice_urban_life_car_cte	-0.338	0.18	-1.87	0.061
12	scale_choice_model	1.11	0.298	3.72	0.0002

Number of estimated parameters	12
Sample size	1899
Final log likelihood	-1191.766
Akaike Information Criterion	2407.532
Bayesian Information Criterion	2474.121
Number of draws	10000

Table 9: Choice model with latent variables: sequential estimation

4.8 Simultaneous estimation

We now proceed with the simultaneous estimation of all components of the model: the choice model, the structural equations, and the measurement equations.

The specifications of each component remain consistent with those previously described. The key difference lies in the treatment of the scale parameters in the structural equations, which are now estimated rather than fixed. This modification is motivated by the inclusion of the latent variables in the utility specification of the choice model. As a result, the error terms in the structural equations can now be disentangled from those in the measurement equations.

The script implementing this specification is described in Section 6.7. As in the sequential case, the latent variables are constructed from two components: the deterministic part obtained from the structural equations, and a stochastic term capturing unobserved heterogeneity:

```
sigma_car_structural = Beta('sigma_car_structural', 0.1, None, None, 0)
car_centric_attitude = build_car_centric_attitude() +
    sigma_car_structural * Draws('car_error_term', 'NORMAL_MLHS_ANTI')
```

Compared to the sequential estimation approach, two main differences arise. First, the scale parameter sigma_car_structural is now explicitly estimated. Second, the structural parameters are no longer fixed to the values obtained from the MIMIC model but are jointly estimated within the full model. This ensures internal consistency across all components.

The conditional likelihood function now combines the contributions of both the choice model and the measurement model:

```
cond_prob = logit(V, None, Choice) * likelihood_indicator
```

Simultaneous estimation presents significant numerical challenges due to the complexity of the composite likelihood expression. In particular, the computation of derivatives becomes demanding. For this reason, additional options are specified when instantiating the BIOGEME object:

```
BIOGEME(
    ...,
    calculating_second_derivatives='never',
    numerically_safe=True,
    max_iterations=5000,
)
```

The option calculating_second_derivatives='never' instructs Biogeme to skip the calculation of second derivatives, which often fails due to numerical instability. As a result, statistical inference is performed using the BHHH approximation matrix (Berndt et al., 1974) instead of the Rao-Cramer bound. The numerically_safe=True flag activates additional safeguards to avoid numerical issues, especially when computing the logarithm of expressions approaching zero — typically the case when multiplying several small probability terms. Lastly, the number of maximum iterations is increased to allow the optimization algorithm sufficient time to converge to the desired level of precision.

Although all parameters have been estimated simultaneously, they are reported in separate groups to improve clarity and readability. General estimation statistics are presented in Table 10 on the following page. The parameters of the structural equations are provided in Table 11 on the next page, while those of the measurement equations are reported in Tables 12–14. Finally, the parameters of the choice model are shown in Table 15 on page 21. Finally, Table 16 on page 22 compares the parameters of the choice model (i) without latent variables, (ii) with latent variables and sequential estimation, and (iii) with latent variables and simultaneous estimation.

Number of estimated parameters	83
Sample size	1899
Final log likelihood	-43695.23
Akaike Information Criterion	87556.45
Bayesian Information Criterion	88017.03
Number of draws	10000

Table 10: Simultaneous model: general statistics

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	car_struct_age_65_more	0.175	0.104	1.69	0.0902
2	car_struct_ScaledIncome	-0.046	0.00962	-4.78	1.73e-06
3	car_struct_moreThanOneCar	0.945	0.0802	11.8	0.0
4	car_struct_moreThanOneBike	-0.505	0.0872	-5.79	6.96e-09
5	car_struct_individualHouse	-0.126	0.0775	-1.62	0.104
6	car_struct_haveChildren	-0.0314	0.0757	-0.415	0.678
7	car_struct_haveGA	-0.922	0.117	-7.89	3.11e-15
8	car_struct_highEducation	-0.451	0.0765	-5.89	3.82e-09
9	sigma_car_structural	1.17	0.0495	23.7	0.0
10	urban_struct_childCenter	0.199	0.0499	3.99	6.73e-05
11	urban_struct_childSuburb	0.139	0.035	3.97	7.31e-05
12	urban_struct_highEducation	0.0112	0.0342	0.327	0.743
13	urban_struct_artisans	-0.184	0.072	-2.55	0.0107
14	urban_struct_employees	-0.105	0.0309	-3.39	0.000689
15	urban_struct_age_30_less	0.386	0.0472	8.18	2.22e-16
16	urban_struct_haveChildren	-0.0622	0.0301	-2.07	0.0385
17	urban_struct_UrbRur	0.239	0.0324	7.36	1.83e-13
18	urban_struct_IndividualHouse	0.0371	0.0307	1.21	0.227
19	sigma_urban_structural	-0.456	0.0342	-13.3	0.0

Table 11: Simultaneous model: estimated parameters of the structural equations

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	meas_intercept_Envir01	-1.13	0.104	-10.9	0.0
2	meas_intercept_Envir02	0.0434	0.055	0.789	0.43
3	car_meas_b_Envir02	-0.457	0.0294	-15.5	0.0
4	meas_sigma_star_Envir02	1.11	0.0342	32.5	0.0
5	meas_intercept_Envir03	0.0879	0.058	1.52	0.13
6	car_meas_b_Envir03	0.476	0.0296	16.0	0.0
7	meas_sigma_star_Envir03	1.04	0.0317	32.7	0.0
8	meas_intercept_Envir04	-0.332	0.0449	-7.38	1.55e-13
9	car_meas_b_Envir04	0.33	0.0255	12.9	0.0
10	meas_sigma_star_Envir04	0.994	0.0293	33.9	0.0
11	meas_intercept_Mobil09	0.709	0.0488	14.5	0.0
12	car_meas_b_Mobil09	-0.346	0.0272	-12.7	0.0
13	meas_sigma_star_Mobil09	1.07	0.0319	33.5	0.0
14	meas_intercept_Mobil11	1.11	0.0635	17.5	0.0
15	car_meas_b_Mobil11	0.49	0.0317	15.5	0.0
16	meas_sigma_star_Mobil11	1.13	0.0363	31.1	0.0
17	meas_intercept_Mobil14	0.404	0.0577	6.99	2.67e-12
18	car_meas_b_Mobil14	0.522	0.0274	19.1	0.0
19	meas_sigma_star_Mobil14	0.941	0.0297	31.6	0.0
20	meas_intercept_Mobil16	0.736	0.0574	12.8	0.0
21	car_meas_b_Mobil16	0.455	0.027	16.8	0.0
22	meas_sigma_star_Mobil16	1.1	0.0345	31.8	0.0
23	meas_intercept_Mobil17	0.709	0.0586	12.1	0.0
24	car_meas_b_Mobil17	0.439	0.0289	15.2	0.0
25	meas_sigma_star_Mobil17	1.1	0.036	30.7	0.0
26	meas_intercept_LifSty08	1.45	0.0533	27.2	0.0
27	car_meas_b_LifSty08	-0.0685	0.0294	-2.33	0.0198
28	meas_sigma_star_LifSty08	1.25	0.0391	32.0	0.0

Table 12: Simultaneous model: estimated parameters of the measurement equations for the car-centric attitude

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	meas_intercept_ResidCh01	-0.894	0.0669	-13.4	0.0
2	meas_intercept_ResidCh02	0.425	0.063	6.75	1.52e-11
3	urban_meas_b_ResidCh02	0.87	0.107	8.16	4.44e-16
4	meas_sigma_star_ResidCh02	1.18	0.0407	29.1	0.0
5	meas_intercept_ResidCh03	0.251	0.044	5.71	1.15e-08
6	urban_meas_b_ResidCh03	0.0226	0.0747	0.302	0.762
7	meas_sigma_star_ResidCh03	1.27	0.0413	30.7	0.0
8	meas_intercept_ResidCh05	-3.06	0.17	-18.0	0.0
9	urban_meas_b_ResidCh05	2.44	0.207	11.8	0.0
10	meas_sigma_star_ResidCh05	0.954	0.0506	18.8	0.0
11	meas_intercept_ResidCh06	-1.31	0.0972	-13.5	0.0
12	urban_meas_b_ResidCh06	1.56	0.144	10.8	0.0
13	meas_sigma_star_ResidCh06	1.27	0.0416	30.4	0.0
14	meas_intercept_ResidCh07	2.25	0.125	18.0	0.0
15	urban_meas_b_ResidCh07	-1.93	0.16	-12.0	0.0
16	meas_sigma_star_ResidCh07	0.895	0.036	24.9	0.0
17	meas_intercept_Mobil07	1.3	0.0544	23.8	0.0
18	urban_meas_b_Mobil07	-0.378	0.0763	-4.95	7.46e-07
19	meas_sigma_star_Mobil07	1.08	0.0299	36.2	0.0
20	meas_intercept_Mobil24	0.209	0.0601	3.48	0.000495
21	urban_meas_b_Mobil24	0.395	0.101	3.9	9.56e-05
22	meas_sigma_star_Mobil24	1.46	0.0473	30.9	0.0

Table 13: Simultaneous model: estimated parameters of the measurement equations for the urban-preference attitude

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	delta_1	0.402	0.00842	47.8	0.0
2	delta_2	1.24	0.0243	51.0	0.0

Table 14: Simultaneous model: thresholds of the measurement equations

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	choice_asc_pt	-0.75	0.273	-2.74	0.00605
2	choice_asc_car	0.755	0.241	3.14	0.0017
3	choice_beta_cost_other	-0.413	0.0802	-5.15	2.6e-07
4	choice_beta_time_pt	-1.8	0.371	-4.86	1.18e-06
5	choice_beta_waiting_time	-0.0158	0.00632	-2.5	0.0124
6	choice_beta_time_car	-5.26	0.645	-8.16	4.44e-16
7	choice_beta_dist	-1.29	0.144	-8.92	0.0
8	choice_car_centric_pt_cte	0.113	0.114	0.983	0.325
9	choice_urban_life_pt_cte	0.764	0.367	2.08	0.0372
10	choice_car_centric_car_cte	0.678	0.13	5.21	1.86e-07
11	choice_urban_life_car_cte	-0.28	0.326	-0.86	0.39
12	scale_choice_model	1.06	0.114	9.32	0.0

Table 15: Choice model with latent variables: simultaneous estimation

Table 16: Comparison of the parameters of the choice model

	Parameter name	Choice only Coef./(SE)	Sequential Coef./(SE)	Simultaneous Coef./(SE)
1	choice_asc_pt	-0.132	0.436	-0.75^{***}
		(0.262)	(0.417)	(0.273)
2	choice_asc_car	0.408	1.55^{***}	0.755^{***}
		(0.283)	(0.588)	(0.241)
3	choice_beta_cost_other	-0.465^{***}	-0.416^{***}	-0.413^{***}
		(0.142)	(0.148)	(0.0802)
4	choice_beta_time_pt	-1.58^{**}	-1.95^{**}	-1.8^{***}
		(0.658)	(0.865)	(0.371)
5	choice_beta_waiting_time	-0.018^{**}	-0.02^{**}	-0.0158^{**}
		(0.00789)	(0.0101)	(0.00632)
6	choice_beta_time_car	-5.14^{***}	-5.58^{***}	-5.26^{***}
		(1.67)	(2.03)	(0.645)
7	choice_beta_dist	-0.983^{***}	-1.13^{***}	-1.29^{***}
		(0.317)	(0.407)	(0.144)
8	choice_car_centric_pt_cte		0.652^{**}	0.113
			(0.265)	(0.114)
9	choice_urban_life_pt_cte		0.0223	0.764^{**}
			(0.0826)	(0.367)
10	choice_car_centric_car_cte		1.62^{***}	0.678^{***}
			(0.52)	(0.13)
11	choice_urban_life_car_cte		-0.338^{*}	-0.28
			(0.18)	(0.326)
12	scale_choice_model	1.17^{***}	1.11^{***}	1.06^{***}
		(0.275)	(0.298)	(0.114)
Number of observations		1899	1899	1899
Number of parameters		8	12	83
Akaike Information Criterion		2476.0	2407.7	87556.5
Bayesian Information Criterion		2520.4	2474.3	88017.0

Standard errors: ***: $p < 0.01$, **: $p < 0.05$, *: $p < 0.1$

5 Conclusion

Choice models with latent variables offer a powerful and flexible framework for capturing complex behavioral mechanisms underlying decision-making. By incorporating unobserved psychological constructs such as attitudes, and perceptions, these models extend the explanatory power of traditional discrete choice models. They allow researchers to account for systematic heterogeneity in behavior that is not directly observed in the data, thereby enhancing both the behavioral realism and predictive performance of the models.

Despite their potential, these models are inherently more complex to specify, estimate, and interpret. It is therefore recommended to proceed incrementally. A practical and effective strategy is to begin by developing and estimating the choice model and the MIMIC model independently. This allows the analyst to ensure that both components are correctly specified and empirically supported.

Once the separate models have been validated, the next step is to explore their integration through sequential estimation. In this stage, the latent variables generated from the MIMIC model are incorporated into the utility specification of the choice model. This provides valuable insights into how these latent constructs influence behavior, while still maintaining manageable computational complexity.

Only after the specification has been refined and the results from the sequential estimation are deemed satisfactory should one proceed to the simultaneous estimation of all components. This final step — though computationally more demanding — offers the benefit of statistical efficiency by leveraging all available information jointly. It also provides a more coherent treatment of the latent variables, since their estimation is informed not only by the indicators, but also by the observed choices.

6 Complete specification files

The following specification files have been used for the estimation of the presented results. They have been developed for Biogeme 3.3.0. It is possible that the syntax should be slightly adapted for future versions of Biogeme.

6.1 relevant_data.py

```
"""

Relevant data for the hybrid choice model
=========================================

This file identifies the relevant data for the hybrid choice model, that are shared by several
    specifications.
Michel Bierlaire, EPFL
Thu May 15 2025, 15:47:42
"""

from biogeme.data.optima import (
    HouseType,
    ScaledIncome,
    SocioProfCat,
    UrbRur,
    age,
    age_65_more,
    childCenter,
    childSuburb,
    haveChildren,
    haveGA,
    highEducation,
    individualHouse,
    moreThanOneBike,
    moreThanOneCar,
)
from biogeme.expressions import Expression

# %%
# Indicators for the car centric attitude.

car_indicators = {
    'Envir01',
    'Envir02',
    'Envir03',
    'Envir04',
    'Mobil09',
    'Mobil11',
    'Mobil14',
    'Mobil16',
    'Mobil17',
    'LifSty08',
}
normalized_car = 'Envir01'

# %%
# indicators for the urban preference attitude
urban_indicators = {
    'ResidCh01',
    'ResidCh02',
    'ResidCh03',
    'ResidCh05',
    'ResidCh06',
    'ResidCh07',
    'Mobil07',
    'Mobil24',
}
normalized_urban = 'ResidCh01'


# %%
# Latent variable for the car centric attitude
car_explanatory_variables: dict[str, Expression] = {
    'age_65_more': age_65_more,
    'ScaledIncome': ScaledIncome,
    'moreThanOneCar': moreThanOneCar,
    'moreThanOneBike': moreThanOneBike,
    'individualHouse': individualHouse,
    'haveChildren': haveChildren,
    'haveGA': haveGA,
    'highEducation': highEducation,
}
# %%
# Latent variable for the urban preference attitude
urban_explanatory_variables: dict[str, Expression] = {
    'childCenter': childCenter,
    'childSuburb': childSuburb,
    'highEducation': highEducation,
    'artisans': SocioProfCat == 5,
    'employees': SocioProfCat == 6,
    'age_30_less': age <= 30,
    'haveChildren': haveChildren,
    'UrbRur': UrbRur,
    'IndividualHouse': HouseType == 1,
}

# %%
```

```python
# Dict of all explanatory variables
all_explanatory_variables = car_explanatory_variables | urban_explanatory_variables
```

6.2 structural_equations.py

```python
from relevant_data import car_explanatory_variables, urban_explanatory_variables

from biogeme.expressions import Beta, Expression, MultipleSum


# %%
# Structural equation: car centric attitude
def build_car_centric_attitude(
    estimated_parameters: dict[str, float] | None = None,
) -> Expression:
    """Builds the expression for the structural equation of the car centric attitude

    :param estimated_parameters: if not None, provides the value of the parameters for
        the direct calculation.
    :return: the expression for the structural equation
    """
    if estimated_parameters is None:
        car_struct_coefficients = {
            variable_name: Beta(f'car_struct_{variable_name}', 0.0, None, None, 0)
            for variable_name in car_explanatory_variables.keys()
        }
    else:
        car_struct_coefficients = {
            variable_name: estimated_parameters[f'car_struct_{variable_name}']
            for variable_name in car_explanatory_variables.keys()
        }

    car_centric_attitude = MultipleSum(
        [
            car_struct_coefficients[variable_name] * variable_expression
            for variable_name, variable_expression in car_explanatory_variables.items()
        ]
    )
    return car_centric_attitude


# %%
# Latent variable for the urban preference attitude
# Structural equation
def build_urban_preference_attitude(
    estimated_parameters: dict[str, float] | None = None,
) -> Expression:
    """Builds the expression for the structural equation of the urban preference
        attitude

    :param estimated_parameters: if not None, provides the value of the parameters for
        the direct calculation
    :return: the expression for the structural equation
    """
    if estimated_parameters is None:
        urban_struct_coefficients = {
            variable_name: Beta(f'urban_struct_{variable_name}', 0.0, None, None, 0)
            for variable_name in urban_explanatory_variables.keys()
        }
    else:
        urban_struct_coefficients = {
            variable_name: estimated_parameters[f'urban_struct_{variable_name}']
            for variable_name in urban_explanatory_variables.keys()
        }

    urban_life_attitude = MultipleSum(
        [
            urban_struct_coefficients[variable_name] * variable_expression
            for variable_name, variable_expression in urban_explanatory_variables.items()
        ]
    )
    return urban_life_attitude
```

6.3 measurement_equations.py

```python
from biogeme.exceptions import BiogemeError
from biogeme.expressions import Beta, Expression, NormalCdf

from relevant_data import (
    car_indicators,
    normalized_car,
    normalized_urban,
    urban_indicators,
)

# %%
# Measurement equations.

# %%
# All indicators
all_indicators = car_indicators | urban_indicators


# %%
# Intercepts
intercepts: dict[str, Beta | float] = {
    k: Beta(f'meas_intercept_{k}', 0, None, None, 0) for k in all_indicators
}

```

```python
# %%
# coefficients for the car centric attitude latent variable
car_coefficients: dict[str, Beta | float] = {
    k: Beta(f'car_meas_b_{k}', 0, None, None, 0) for k in car_indicators
}

# %%
# The indicator that is used to normalize the latent variable "car centric attitude" is
# "Fuel price should be increased to reduce congestion and air"
# Therefore, we normalize the coefficient to -1.
car_coefficients[normalized_car] = -1.0

# %%
# coefficients for the urban preference attitude latent variable
urban_coefficients: dict[str, Beta | float] = {
    k: Beta(f'urban_meas_b_{k}', 0, None, None, 0) for k in urban_indicators
}
urban_coefficients[normalized_urban] = 1.0

# %%
# Scale parameters of the error terms.
sigma_star: dict[str, Beta | float] = {
    k: Beta(f'meas_sigma_star_{k}', 1, 1.0e-4, None, 0) for k in all_indicators
}

# %%
# Normalization of the scale parameters
sigma_star[normalized_car] = 1.0
sigma_star[normalized_urban] = 1.0


# %%
# Contribution of the latent variables to the measurement equations
def generate_model_terms(
    indicator: str,
    car_centric_attitude: Expression,
    urban_preference_attitude: Expression,
) -> Expression:
    """Returns the contribution of the latent variables to the measurement equation
    for a given indicator."""
    is_car = indicator in car_indicators
    is_urban = indicator in urban_indicators

    if is_car and is_urban:
        # Indicator is influenced by both latent variables
        return (
            car_coefficients[indicator] * car_centric_attitude
            + urban_coefficients[indicator] * urban_preference_attitude
        )
    if is_car:
        # Indicator is influenced by the car-centric latent variable
        return car_coefficients[indicator] * car_centric_attitude
    if is_urban:
        # Indicator is influenced by the urban preference latent variable
        return urban_coefficients[indicator] * urban_preference_attitude
    raise BiogemeError(f'Unknown indicator: {indicator}')


def generate_measurement_equations(
    car_centric_attitude: Expression, urban_preference_attitude: Expression
) -> dict[int, Expression | float]:
    """

    :param car_centric_attitude: expression for the latent variable.
    :param urban_preference_attitude: expression for the latent variable.
    :return: a dict associating each level of the Liker scale with the Expression
        calculating the corresponding probability.

    The first one is normalized to -1. Indeed, we expect the respondents with a higher
    car centric attitude to disagree more with the statement
    "Fuel price should be increased to reduce congestion and air pollution."
    (corresponding to a low value of the indicator) compared to those who have a
    lower car centric attitude. In other words, when the car centric attitude
    increases, we expect the value of the indicator to decrease.
    """
    models = {
        k: intercepts[k]
        + generate_model_terms(
            k,
            car_centric_attitude=car_centric_attitude,
            urban_preference_attitude=urban_preference_attitude,
        )
        for k in all_indicators
    }

    # Symmetric threshold.
    delta_1 = Beta('delta_1', 0.3, 1.0e-3, None, 0)
    delta_2 = Beta('delta_2', 0.8, 1.0e-3, None, 0)
    tau_1 = -delta_1 - delta_2
    tau_2 = -delta_1
    tau_3 = delta_1
    tau_4 = delta_1 + delta_2

    # %%
    # Ordered probit models.
    tau_1_residual = {
        indicator: (tau_1 - models[indicator]) / sigma_star[indicator]
        for indicator in all_indicators
    }
    tau_2_residual = {
        indicator: (tau_2 - models[indicator]) / sigma_star[indicator]
        for indicator in all_indicators
    }
```

```
    tau_3_residual = {
        indicator: (tau_3 - models[indicator]) / sigma_star[indicator]
        for indicator in all_indicators
    }
    tau_4_residual = {
        indicator: (tau_4 - models[indicator]) / sigma_star[indicator]
        for indicator in all_indicators
    }

    dict_prob_indicators = {
        indicator: {
            1: NormalCdf(tau_1_residual[indicator]),
            2: NormalCdf(tau_2_residual[indicator])
            - NormalCdf(
                tau_1_residual[indicator],
            ),
            3: NormalCdf(tau_3_residual[indicator])
            - NormalCdf(
                tau_2_residual[indicator],
            ),
            4: NormalCdf(tau_4_residual[indicator])
            - NormalCdf(
                tau_3_residual[indicator],
            ),
            5: 1 - NormalCdf(tau_4_residual[indicator]),
            6: 1.0,
            -1: 1.0,
            -2: 1.0,
        }
        for indicator in all_indicators
    }

    return dict_prob_indicators
```

6.4 plot_b01_mimic.py

```
"""

MIMIC (Multiple Indicators Multiple Causes) model
=================================================

THe MIMIC model involves two latent variables: "car centric" attitude, and "urban preference" attitude.
Michel Bierlaire, EPFL
Fri May 16 2025, 10:32:30

"""

import biogeme.biogeme_logging as blog
from IPython.core.display_functions import display
from biogeme.biogeme import BIOGEME
from biogeme.data.optima import (
    read_data,
)
from biogeme.database import Database
from biogeme.expressions import (
    Elem,
    MultipleSum,
    Variable,
    log,
)
from biogeme.results_processing import (
    get_pandas_estimated_parameters,
)

from measurement_equations import all_indicators, generate_measurement_equations
from read_or_estimate import read_or_estimate
from structural_equations import (
    build_car_centric_attitude,
    build_urban_preference_attitude,
)

# %%
# Structural equation: car centric attitude
car_centric_attitude = build_car_centric_attitude()

# %%
# Structural equation: urban preference
urban_preference_attitude = build_urban_preference_attitude()

logger = blog.get_screen_logger(level=blog.INFO)
logger.info('Example b01one_latent_regression.py')

dict_prob_indicators = generate_measurement_equations(
    car_centric_attitude=car_centric_attitude,
    urban_preference_attitude=urban_preference_attitude,
)

# %%
# We calculate the joint probability of all indicators
log_proba = {
    indicator: log(Elem(dict_prob_indicators[indicator], Variable(indicator)))
    for indicator in all_indicators
}
log_likelihood = MultipleSum(log_proba)

# %%
# Read the data
database: Database = read_data()

# %%
# Create the Biogeme object.
```

```
the_biogeme = BIOGEME(database, log_likelihood)
the_biogeme.model_name = 'b01_mimic'

# %%
# If estimation results are saved on file, we read them to speed up the process.
# If not, we estimate the parameters.
results = read_or_estimate(the_biogeme=the_biogeme, directory='saved_results')

# %%
print(f'Estimated betas: {results.number_of_parameters}')
print(f'final log likelihood: {results.final_log_likelihood:.3f}')
print(f'Output file: {the_biogeme.html_filename}')

# %%
pandas_results = get_pandas_estimated_parameters(estimation_results=results)
display(pandas_results)
```

6.5 plot_b02_choice_only.py

```
"""

Estimation of choice model without latent variables
===================================================

Michel Bierlaire, EPFL
Thu May 15 2025, 15:23:42
"""

import biogeme.biogeme_logging as blog
from IPython.core.display_functions import display
from biogeme.biogeme import BIOGEME
from biogeme.data.optima import (
    Choice,
    CostCarCHF_scaled,
    MarginalCostPT_scaled,
    PurpHWH,
    PurpOther,
    TimeCar_scaled,
    TimePT_scaled,
    WaitingTimePT,
    distance_km_scaled,
    read_data,
)
from biogeme.expressions import Beta
from biogeme.models import loglogit
from biogeme.results_processing import (
    get_pandas_estimated_parameters,
)

from read_or_estimate import read_or_estimate

logger = blog.get_screen_logger(level=blog.INFO)

# %%
# Choice model: parameters
choice_asc_car = Beta('choice_asc_car', 0.0, None, None, 0)
choice_asc_pt = Beta('choice_asc_pt', 0, None, None, 0)
choice_asc_sm = Beta('choice_asc_sm', 0, None, None, 1)
choice_beta_cost_hwh = Beta('choice_beta_cost_hwh', -1, None, None, 1)
choice_beta_cost_other = Beta('choice_beta_cost_other', 0, None, None, 0)
choice_beta_dist = Beta('choice_beta_dist', 0, None, None, 0)
choice_beta_time_car = Beta('choice_beta_time_car', 0, None, 0, 0)
choice_beta_time_pt = Beta('choice_beta_time_pt', 0, None, 0, 0)
choice_beta_waiting_time = Beta('choice_beta_waiting_time', 0, None, None, 0)
scale_choice_model = Beta('scale_choice_model', 1, 1.0e-5, 10, 0)

# %%
# Definition of utility functions:
V0 = scale_choice_model * (
    choice_asc_pt
    + choice_beta_time_pt * TimePT_scaled
    + choice_beta_waiting_time * WaitingTimePT
    + choice_beta_cost_hwh * MarginalCostPT_scaled * PurpHWH
    + choice_beta_cost_other * MarginalCostPT_scaled * PurpOther
)

V1 = scale_choice_model * (
    choice_asc_car
    + choice_beta_time_car * TimeCar_scaled
    + choice_beta_cost_hwh * CostCarCHF_scaled * PurpHWH
    + choice_beta_cost_other * CostCarCHF_scaled * PurpOther
)

V2 = scale_choice_model * (choice_asc_sm + choice_beta_dist * distance_km_scaled)

# %%
# Associate utility functions with the numbering of alternatives
V = {0: V0, 1: V1, 2: V2}

# %%
# We integrate over omega using numerical integration
log_likelihood = loglogit(V, None, Choice)

# %%
# Read the data
database = read_data()

# %%
# Create the Biogeme object
the_biogeme = BIOGEME(database, log_likelihood)
the_biogeme.model_name = 'b02_choice_only'
```

```
# %%
# If estimation results are saved on file, we read them to speed up the process.
# If not, we estimate the parameters.
results = read_or_estimate(the_biogeme=the_biogeme, directory='saved_results')

# %%
print(results.short_summary())

# %%
# Get the results in a pandas table
pandas_results = get_pandas_estimated_parameters(
    estimation_results=results,
)
display(pandas_results)
```

6.6 plot_b03_sequential.py

```
"""

Choice model with latent variables: sequential estimation
=========================================================

Mixture of logit.
Measurement equation for the indicators.
Sequential estimation.

Michel Bierlaire, EPFL
Thu May 15 2025, 15:34:13
"""

import sys

import biogeme.biogeme_logging as blog
from IPython.core.display_functions import display
from biogeme.biogeme import BIOGEME
from biogeme.data.optima import (
    Choice,
    CostCarCHF_scaled,
    MarginalCostPT_scaled,
    PurpHWH,
    PurpOther,
    TimeCar_scaled,
    TimePT_scaled,
    WaitingTimePT,
    distance_km_scaled,
    read_data,
)
from biogeme.expressions import Beta, Draws, MonteCarlo, log
from biogeme.models import logit
from biogeme.results_processing import (
    EstimationResults,
    get_pandas_estimated_parameters,
)

from read_or_estimate import read_or_estimate
from structural_equations import (
    build_car_centric_attitude,
    build_urban_preference_attitude,
)

logger = blog.get_screen_logger(level=blog.INFO)

# %%
# Read the estimates from the structural equation estimation.
MODEL_NAME = 'b01_mimic'
try:
    mimic_results = EstimationResults.from_yaml_file(
        filename=f'saved_results/{MODEL_NAME}.yaml'
    )
except FileNotFoundError:
    print(
        f'Run first the script {MODEL_NAME}.py in order to generate the '
        f'file {MODEL_NAME}.yaml, and move it to the directory saved_results'
    )
    sys.exit()
struct_betas = mimic_results.get_beta_values()

# %%
# Read the estimates from the structural equation estimation.
CHOICE_MODEL_NAME = 'b02_choice_only'
try:
    choice_results = EstimationResults.from_yaml_file(
        filename=f'saved_results/{CHOICE_MODEL_NAME}.yaml'
    )
except FileNotFoundError:
    print(
        f'Run first the script {CHOICE_MODEL_NAME}.py in order to generate the '
        f'file {CHOICE_MODEL_NAME}.yaml, and move it to the directory saved_results'
    )
    sys.exit()
choice_betas = choice_results.get_beta_values()

# %%
# Structural equation: car centric attitude

car_centric_attitude = build_car_centric_attitude(
    estimated_parameters=struct_betas
) + Draws('car_error_term', 'NORMAL_MLHS_ANTI')


```

```
# %%
# Latent variable for the urban preference

urban_life_attitude = build_urban_preference_attitude(
    estimated_parameters=struct_betas
) + Draws('urban_error_term', 'NORMAL_MLHS_ANTI')

# %%
# Choice model

# %%
# Parameter from the original choice model
choice_asc_car = Beta('choice_asc_car', choice_betas['choice_asc_car'], None, None, 0)
choice_asc_pt = Beta('choice_asc_pt', choice_betas['choice_asc_pt'], None, None, 0)
choice_beta_cost_hwh = -1.0
choice_beta_cost_other = Beta(
    'choice_beta_cost_other', choice_betas['choice_beta_cost_other'], None, None, 0
)
choice_beta_dist = Beta(
    'choice_beta_dist', choice_betas['choice_beta_dist'], None, None, 0
)
choice_beta_waiting_time = Beta(
    'choice_beta_waiting_time', choice_betas['choice_beta_waiting_time'], None, None, 0
)
choice_beta_time_car = Beta(
    'choice_beta_time_car', choice_betas['choice_beta_time_car'], None, 0, 0
)
choice_beta_time_pt = Beta(
    'choice_beta_time_pt', choice_betas['choice_beta_time_pt'], None, 0, 0
)
scale_choice_model = Beta(
    'scale_choice_model', choice_betas['scale_choice_model'], 1.0e-5, None, 0
)


# %%
# Parameter affected by the latent variables.

# %%
# Alternative specific constants
choice_car_centric_car_cte = Beta('choice_car_centric_car_cte', 0, None, None, 0)
choice_car_centric_pt_cte = Beta('choice_car_centric_pt_cte', 0, None, None, 0)
choice_urban_life_car_cte = Beta('choice_urban_life_car_cte', 0, None, None, 0)
choice_urban_life_pt_cte = Beta('choice_urban_life_pt_cte', 0, None, None, 0)

# %%
# Definition of utility functions:
V0 = scale_choice_model * (
    choice_asc_pt
    + choice_beta_time_pt * TimePT_scaled
    + choice_beta_waiting_time * WaitingTimePT
    + choice_beta_cost_hwh * MarginalCostPT_scaled * PurpHWH
    + choice_beta_cost_other * MarginalCostPT_scaled * PurpOther
    + choice_car_centric_pt_cte * car_centric_attitude
    + choice_urban_life_pt_cte * urban_life_attitude
)

V1 = scale_choice_model * (
    choice_asc_car
    + choice_beta_time_car * TimeCar_scaled
    + choice_beta_cost_hwh * CostCarCHF_scaled * PurpHWH
    + choice_beta_cost_other * CostCarCHF_scaled * PurpOther
    + choice_car_centric_car_cte * car_centric_attitude
    + choice_urban_life_car_cte * urban_life_attitude
)

V2 = scale_choice_model * choice_beta_dist * distance_km_scaled

# %%
# Associate utility functions with the numbering of alternatives
V = {0: V0, 1: V1, 2: V2}

# %%
# Conditional on the latent variables, we have a logit model (called the kernel)
cond_prob = logit(V, None, Choice)

# %%
# We integrate over omega using numerical integration
log_likelihood = log(MonteCarlo(cond_prob))

# %%
# Read the data
database = read_data()

# %%
# Create the Biogeme object
the_biogeme = BIOGEME(database, log_likelihood, number_of_draws=10_000)
the_biogeme.model_name = 'b03_sequential'

# %%
# If estimation results are saved on file, we read them to speed up the process.
# If not, we estimate the parameters.
results = read_or_estimate(the_biogeme=the_biogeme, directory='saved_results')

# %%
print(results.short_summary())

# %%
# Get the results in a pandas table
pandas_results = get_pandas_estimated_parameters(
    estimation_results=results,
)
display(pandas_results)
```

6.7 plot_b03_simultaneous.py

```
"""

Choice model with latent variables: simultaneous estimation
===========================================================

Mixture of logit.
Measurement equation for the indicators.
Sequential estimation.

Michel Bierlaire, EPFL
Fri May 16 2025, 15:53:52
"""

from IPython.core.display_functions import display

import biogeme.biogeme_logging as blog
from biogeme.biogeme import BIOGEME
from biogeme.data.optima import (
    Choice,
    CostCarCHF_scaled,
    MarginalCostPT_scaled,
    PurpHWH,
    PurpOther,
    TimeCar_scaled,
    TimePT_scaled,
    WaitingTimePT,
    distance_km_scaled,
    read_data,
)
from biogeme.expressions import (
    Beta,
    Draws,
    Elem,
    MonteCarlo,
    MultipleProduct,
    Variable,
    log,
)
from biogeme.models import logit
from biogeme.results_processing import (
    get_pandas_estimated_parameters,
)
from measurement_equations import all_indicators, generate_measurement_equations
from read_or_estimate import read_or_estimate
from structural_equations import (
    build_car_centric_attitude,
    build_urban_preference_attitude,
)

logger = blog.get_screen_logger(level=blog.INFO)

# %%
# Structural equation: car centric attitude
sigma_car_structural = Beta('sigma_car_structural', 0.1, None, None, 0)
car_centric_attitude = build_car_centric_attitude() + sigma_car_structural * Draws(
    'car_error_term', 'NORMAL_MLHS_ANTI'
)


# %%
# Latent variable for the urban preference

sigma_urban_structural = Beta('sigma_urban_structural', 0.1, None, None, 0)
urban_preference_attitude = (
    build_urban_preference_attitude()
    + sigma_urban_structural * Draws('urban_error_term', 'NORMAL_MLHS_ANTI')
)

# %%
# Choice model

# %%
# Parameter from the  choice model
choice_asc_car = Beta('choice_asc_car', 0, None, None, 0)
choice_asc_pt = Beta('choice_asc_pt', 0, None, None, 0)
choice_beta_cost_hwh = -1.0
choice_beta_cost_other = Beta('choice_beta_cost_other', 0, None, None, 0)
choice_beta_dist = Beta('choice_beta_dist', 0, None, None, 0)
choice_beta_waiting_time = Beta('choice_beta_waiting_time', 0, None, None, 0)
choice_beta_time_car = Beta('choice_beta_time_car', 0, None, 0, 0)
choice_beta_time_pt = Beta('choice_beta_time_pt', 0, None, 0, 0)
scale_choice_model = Beta('scale_choice_model', 1, 1.0e-5, None, 0)


# %%
# Parameter affected by the latent variables.

# %%
# Alternative specific constants
choice_car_centric_car_cte = Beta('choice_car_centric_car_cte', 1, None, None, 0)
choice_car_centric_pt_cte = Beta('choice_car_centric_pt_cte', 1, None, None, 0)
choice_urban_life_car_cte = Beta('choice_urban_life_car_cte', 1, None, None, 0)
choice_urban_life_pt_cte = Beta('choice_urban_life_pt_cte', 1, None, None, 0)

# %%
# Definition of utility functions:
V0 = scale_choice_model * (
    choice_asc_pt
    + choice_beta_time_pt * TimePT_scaled
    + choice_beta_waiting_time * WaitingTimePT
    + choice_beta_cost_hwh * MarginalCostPT_scaled * PurpHWH
```

```
    + choice_beta_cost_other * MarginalCostPT_scaled * PurpOther
    + choice_car_centric_pt_cte * car_centric_attitude
    + choice_urban_life_pt_cte * urban_preference_attitude
)

V1 = scale_choice_model * (
    choice_asc_car
    + choice_beta_time_car * TimeCar_scaled
    + choice_beta_cost_hwh * CostCarCHF_scaled * PurpHWH
    + choice_beta_cost_other * CostCarCHF_scaled * PurpOther
    + choice_car_centric_car_cte * car_centric_attitude
    + choice_urban_life_car_cte * urban_preference_attitude
)

V2 = scale_choice_model * choice_beta_dist * distance_km_scaled

# %%
# Associate utility functions with the numbering of alternatives
V = {0: V0, 1: V1, 2: V2}

# %%
# Measurement equations
dict_prob_indicators = generate_measurement_equations(
    car_centric_attitude=car_centric_attitude,
    urban_preference_attitude=urban_preference_attitude,
)

# %%
# We calculate the joint probability of all indicators
proba = {
    indicator: Elem(dict_prob_indicators[indicator], Variable(indicator))
    for indicator in all_indicators
}
likelihood_indicator = MultipleProduct(proba)
# %%
# Conditional on the latent variables, we have a logit model (called the kernel)
cond_prob = logit(V, None, Choice) * likelihood_indicator

# %%
# We integrate over omega using numerical integration
log_likelihood = log(MonteCarlo(cond_prob))

# %%
# Read the data
database = read_data()

# %%
# Create the Biogeme object
the_biogeme = BIOGEME(
    database,
    log_likelihood,
    number_of_draws=10_000,
    calculating_second_derivatives='never',
    numerically_safe=True,
    max_iterations=5000,
)
the_biogeme.model_name = 'b03_simultaneous'

# %%
# If estimation results are saved on file, we read them to speed up the process.
# If not, we estimate the parameters.
results = read_or_estimate(the_biogeme=the_biogeme, directory='saved_results')

# %%
print(results.short_summary())

# %%
# Get the results in a pandas table
pandas_results = get_pandas_estimated_parameters(
    estimation_results=results,
)
display(pandas_results)
```

7 Description of the variables

THe following table describes the variables involved in the models described in this document.

Name	Description
TimePT	The duration of the loop performed in public transport (in minutes).
WaitingTimePT	The total waiting time in a loop performed in public transports (in minutes).
TimeCar	The total duration of a loop made using the car (in minutes).
MarginalCostPT	The total cost of a loop performed in public transports, taking into account the ownership of a seasonal ticket by the respondent. If the respondent has a "GA" (full Swiss season ticket), a seasonal ticket for the line or the area, this variable takes value zero. If the respondent has a half-fare travelcard, this variable corresponds to half the cost of the trip by public transport..
CostCarCHF	The total gas cost of a loop performed with the car in CHF.
TripPurpose	The main purpose of the loop: 1 =Work-related trips; 2 =Work- and leisure-related trips; 3 =Leisure related trips. -1 represents missing values
UrbRur	Binary variable, where: 1 =Rural; 2 =Urban.
distance_km	Total distance performed for the loop.
age	Age of the respondent (in years) -1 represents missing values.
ResidChild	Main place of residence as a kid (< 18), 1 is city center (large town), 2 is city center (small town), 3 is suburbs, 4 is suburban town, 5 is country side (village), 6 is countryside (isolated), -1 is for missing data and -2 if respondent didn't answer to any opinion questions.
NbCar	Number of cars in the household.-1 for missing value.
NbBicy	Number of bikes in the household. -1 for missing value.
HouseType	House type, 1 is individual house (or terraced house), 2 is apartment (and other types of multi-family residential), 3 is independent room (subletting). -1 for missing value.
Income	Net monthly income of the household in CHF. 1 is less than 2500, 2 is from 2501 to 4000, 3 is from 4001 to 6000, 4 is from 6001 to 8000, 5 is from 8001 to 10'000 and 6 is more than 10'001. -1 for missing value.
CalculatedIncome	Net monthly income of the household in CHF, calculated as a continuous variable. The value is the center of the interval of the corresponding incone category.
FamilSitu	Familiar situation: 1 is single, 2 is in a couple without children, 3 is in a couple with children, 4 is single with your own children, 5 is in a colocation, 6 is with your parents and 7 is for other situations. -1 for missing values.

SocioProfCat	To which of the following socioprofessional categories do you belong? 1 is for top managers, 2 for intellectual professions, 3 for freelancers, 4 for intermediate professions, 5 for artisans and salespersons, 6 for employees, 7 for workers and 8 for others. -1 for missing values.
GenAbST	Is equal to 1 if the respondent has a GA (full Swiss season ticket) and 2 if not.

Education	Highest education achieved. As mentioned by Wikipedia in English: "The education system in Switzerland is very diverse, because the constitution of Switzerland delegates the authority for the school system mainly to the cantons. The Swiss constitution sets the foundations, namely that primary school is obligatory for every child and is free in public schools and that the confederation can run or support universities." (source: Education in Switzerland (Wikipedia), accessed April 16, 2013). It is thus difficult to translate the survey that was originally in French and German. The possible answers in the survey are: 1. Unfinished compulsory education: education is compulsory in Switzerland but pupils may finish it at the legal age without succeeding the final exam. 2. Compulsory education with diploma. 3. Vocational education: a three or four-year period of training both in a company and following theoretical courses. Ends with a diploma called "Certificat fédéral de capacité" (i.e., "professional baccalaureate") (reference: Certificat fédéral de capacité (Wikipedia) - in French). 4. A 3-year generalist school giving access to teaching school, nursing schools, social work school, universities of applied sciences or vocational education (sometime in less than the normal number of years). It does not give access to universities in Switzerland. 5. High school: ends with the general baccalaureate exam. The general baccalaureate gives access automatically to universities. 6. Universities of applied sciences, teaching schools, nursing schools, social work schools: ends with a Bachelor and sometimes a Master, mostly focus on vocational training. 7. Universities and institutes of technology: ends with an academic Bachelor and in most cases an academic Master. 8. PhD thesis.

Table 17: Description of variables

References

Ashok, K., Dillon, W. R. and Yuan, S. (2002). Extending discrete choice models to incorporate attitudinal and other latent variables, *Journal of Marketing Research* **39**(1): 31–46.

Ben-Akiva, M., Walker, J., Bernardino, A. T., Gopinath, D. A., Morikawa, T. and Polydoropoulou, A. (2002). Integration of choice and latent variable models, *Perpetual motion: Travel behaviour research opportunities and application challenges* pp. 431–470.

Berndt, E. K., Hall, B. H., Hall, R. E. and Hausman, J. A. (1974). Estimation and inference in nonlinear structural models, *Annals of Economic and Social Measurement* **3/4**: 653–665.

Bierlaire, M. (2019). Monte-carlo integration with pandasbiogeme, *Technical Report TRANSP-OR 191231*, Lausanne, Switzerland.

Bierlaire, M., Curchod, A., Danalet, A., Doyen, E., Faure, P., Glerum, A., Kaufmann, V., Tabaka, K. and Schuler, M. (2011). Projet de recherche sur la mobilité combinée, rapport définitif de l'enquête de préférences révélées, *Technical Report TRANSP-OR 110704*, Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.

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.

Hess, S., Train, K. E. and Polak, J. W. (2006). On the use of a modified latin hypercube sampling (MLHS) method in the estimation of a mixed logit model for vehicle choice, *Transportation Research Part B: Methodological* **40**(2): 147–163. DOI: https://doi.org/10.1016/j.trb.2004.10.005.

Jöreskog, K. G. and Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable, *Journal of the American Statistical Association* **70**(351a): 631–639. DOI: 10.1080/01621459.1975.10482485.

Likert, R. (1932). A technique for the measurement of attitudes, *Archives of psychology* **140**: 1–55.

Walker, J. L. (2001). *Extended discrete choice models: integrated framework, flexible error structures, and latent variables*, PhD thesis, Massachusetts Institute of Technology.