Estimating choice models and latent variables with Biogeme

Michel Bierlaire

October 11, 2024

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

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. But it can also be used to extract indicators from an estimated model. 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 3.xxx, but should remain valid for future versions.

1 Data

Incorporating latent variables into choice models requires more than just choice data; it necessitates additional information. By definition, latent variables are unobserved, but we assume access to indirect indicators. Respondents are typically asked to express their level of agreement or disagreement with certain statements. For example, if we are interested in assessing the perception of comfort in public transportation, we might include statements like, "Using public transportation for commuting is comfortable" or "It is difficult to work on my laptop while using public transportation." These responses serve as proxies for the underlying latent variables.

In the data set that we are using as an example in this document (**?**), various categories of questions have been asked:

Environment : "Fuel price should be increased to reduce congestion and air pollution."

Mobility : "I use the time of my trip in a productive way."

Residential choice : "I like living in a neighborhood where a lot of things happen."

Life style : "I would like to spend more time with my family and friends."

And the responses are coded as follows:

- strongly disagree: 1,
- disagree: 2,
- neutral: 3,
- agree: 4,
- strongly agree: 5,

- not aplicable: 6.

We refer the reader to the online documentation for the comprehensive list of questions. In the rest of the document, we assume that R indicators have been collected for N individuals, and we denote I_{in} the indicator number i collected for individual n.

> **<u>Internal note</u>:** Notations: R is already used for the number of draws in Monte-Carlo simulation. Also, in the notations of the book, indicators are denoted by m, not by I as used in the text. We need to decide. Moreover, I is used for the identity matrix.

2 Exploratory Factor Analysis

The factor analysis model is a measurement model that relates the L latent constructs that we are interested in with the observed data:

$$I_{in} = \lambda_{i0} + \sum_{\ell=1}^{L} \lambda_{i\ell} x^*_{\ell n} + \nu_{in},$$

where I_{in} is the indicator i measured for individual n, $\lambda_{i\ell}$, $\ell = 0, \ldots, L$ are parameters called "loading factors", $x^*_{\ell n}$ is the latent variable ℓ for individual n, and ν_{in} is a disturbance term. All disturbance terms are assumed to be independent from one another, normally distributed, and independent from the latent variables $x^*_{\ell n}$. In matrix form, we have

$$I_n = \lambda_0 + \tilde{\Lambda}\tilde{x}^*_n + \nu_n,$$

where $I_n, \lambda_0, \nu_n \in \mathbb{R}^R$, $\Lambda \in \mathbb{R}^{R \times L}$, $\tilde{x}^*_n \in \mathbb{R}^L$. If we define $x^*_0 = 1$, we can simplify it as

$$I_n = \Lambda x^*_n + \nu_n.$$

Therefore, the variance covariance-matrix Σ_n of the observations is

$$\Sigma_n = \Lambda \operatorname{Cov}(x^*_n) \Lambda^T + I \operatorname{Var}(\mu_n).$$

It is common to assume that the latent variables are uncorrelated and of unit variance, that is, $\operatorname{Cov}(x^*_n) = I$, so that

$$\Sigma_n = \Lambda\Lambda^T + I \operatorname{Var}(\mu_n).$$

Note that, for any orthogonal matrix Q, replacing Λ by ΛQ provides the exact samevariance covariance matrix. Indeed, an orthogonal matrix is such that $QQ^T = I$, so that $\Lambda Q Q^T \Lambda^T = \Lambda\Lambda^T$. Such a matrix Q is usually called a "rotation".