Variance-covariance matrix estimation in Biogeme

Michel Bierlaire

June xx, 2017

Report TRANSP-OR 1706xx
Transport and Mobility Laboratory
School of Architecture, Civil and Environmental Engineering
Ecole Polytechnique Fédérale de Lausanne
`transp-or.epfl.ch`

Series on Biogeme

Abstract

The package PythonBiogeme (`biogeme.epfl.ch`) is designed to estimate the parameters of various models using maximum likelihood estimation. In addition to report the values of the estimated coefficients, it also reports various statistics associated with these parameters. This document explains how they are calculated.

Let $\mathcal{L}(\beta)$ be the likelihood function to be maximized, and

$$H = \nabla^2 \mathcal{L}(\beta^*) \tag{1}$$

its second derivatives matrix calculated at the estimated parameters β^*. The Rao-Cramer bound is defined as

$$R_{\text{Rao}} = -H^{-1}. \tag{2}$$

The standard errors reported in the Biogeme output files are the square root of the diagonal entries of this matrix R_{Rao}. The rest of the matrix is reported in the section *Correlation of coefficients* in the column *Covariance*. A normalized version is reported in the column *Correlation*, knowing that

$$\text{Corr}(\beta_i, \beta_j) = \frac{\text{Cov}(\beta_i, \beta_j)}{\sqrt{\text{Var}(\beta_i)\,\text{Var}(\beta_j)}}. \tag{3}$$

If β is an estimated parameter and σ its reported standard error, the reported t-test is calculated as

$$t = \beta/\sigma. \tag{4}$$

The reported p value is calculated as

$$p = 2(1 - \Phi(t)), \tag{5}$$

where $\Phi(t)$ is the Cumulative Distribution Function (CDF) of a standardized normal random variable, estimated at the value t.

For the Exogenous Maximum Likelihood Estimator (ESML), the log likelihood function to be maximized is defined as

$$\sum_n \log P_n(\beta). \tag{6}$$

This is the estimator considered when the variable BIOGEME_OBJECT.WEIGHT is not defined or, equivalently, when it is equal to 1.0 for each observation. The score of an observation n is defined as

$$s_n = \nabla \log P_n(\beta), \tag{7}$$

that is the gradient (vector of first derivatives) of the contribution to the log likelihood function of observation n. The BHHH matrix is defined as

$$B = \sum_n s_n s_n^T, \tag{8}$$

where s_n^T is the transposed of s_n, so that $s_n s_n^T$ is a K by K matrix of rank 1. The robust variance-covariance matrix is calculated as

$$R_{\text{robust}} = R_{\text{Rao}} B R_{\text{Rao}}. \tag{9}$$

The "robust" statistics reported in the Biogeme output files are computed as described above, based on the matrix R_{robust} is used instead of R_{Rao}.

For the Weighted Exogenous Maximum Likelihood Estimator (WESML), the log likelihood function to be maximized is defined as

$$\sum_n w_n \log P_n(\beta). \tag{10}$$

The score of an observation n is defined as above, and the BHHH matrix is defined as

$$B = \sum_n w_n s_n s_n^T. \tag{11}$$

References