Carpita, M., Ciavolino, E. A generalized maximum entropy estimator to simple linear measurement error model with a composite indicator. Adv Data Anal Classif 11, 139–158 (2017). https://doi.org/10.1007/s11634-016-0237-y

A generalized maximum entropy estimator to simple linear measurement error model with a composite indicator

- Maurizio Carpita & • Enrico Ciavolino

Abstract

We extend the simple linear measurement error model through the inclusion of a composite indicator by using the generalized maximum entropy estimator. A Monte Carlo simulation study is proposed for comparing the performances of the proposed estimator to his counterpart the ordinary least squares "Adjusted for attenuation". The two estimators are compared in term of correlation with the true latent variable, standard error and root mean of squared error. Two illustrative case studies are reported in order to discuss the results obtained on the real data set, and relate them to the conclusions drawn via simulation study.

Introduction

The measurement error that is present in observable variables is one of the most frequent (sometimes ignored) problems in statistical analyses. There is an extensive literature about this subject, and many models have been proposed to try to overcome it (Fuller 1987; Bollen 1989; Wansbeek and Maijer 2000; Carroll et al. 1995; Cheng and Van Ness 2010; Buonaccorsi 2010).

For example, it's well known that the existence of a measurement error in the independent variable of the simple linear regression model introduces a downward bias in the estimate of the slope parameter of interest: in this situation, a simple "correction for attenuation" or a more sophisticated structural equations model with multiple indicators are typically used as solutions for the problem. Another issue to be addressed in the case of the simple linear measurement error model is to obtain "good" estimates of dependent and independent latent variables that can be used in further analysis: some procedure, as the partial least squares with a structural equations model can be useful for this purpose. The Structural Equation Model (SEM) with multiple-indicator approach simultaneously allows to correct for measurement error and to estimate the latent variables.

Another possibility is the "two-step" procedure (Bollen 1989; Schumacker and Lomax 2004; Oberski and Satorra 2013): in the first step, each latent variable is defined as having the (unweighted or weighted) sum of some indicators (repeated observed variables) named the composite indicator, with its measurement error variance estimated separately based on the variances of the single observed variables or obtained from preliminary reliability studies; in the second step, the SEM is estimated while correcting for these estimated and fixed measurement error variances. The two-step approach to correction for measurement error has some advantages: it can reduce the complexity of the model and the number of parameters that must be estimated, and allows for the separation between reliability studies and more substantive research (Oberski and Satorra 2013; Carpita and Ciavolino 2014; Ciavolino et al. 2015b). In fact, composite indicators are very useful for statistical analyses and have many important applications in economic and social sciences (Saltelli 2007; OECD 2008; Foster et al. 2013; Paruolo et al. 2013; Brentari and Zuccolotto 2011).

For the simple linear measurement error model (MEM), Al-Nasser (2005) proposes a Generalized Maximum Entropy (GME) estimator, which allows one to abstract away from the additional assumptions that are made in the classical method: for example, the GME estimation approach does not require linearity of the model and normal distribution of the error terms.

In his study, Al-Nasser (2005) considers the case with only one observed variable for each dependent and independent latent variable. In this study we extend the Al-Nasser (2005) proposed GME estimation method to the case with multiple indicators for the independent latent variable. We show that, using the two-step approach, the reliability measures obtained in step one from the observed variables for the composite indicator can be profitably used in step two to define the two error variances and support points required by the GME approach (see Sect. 4.3); as well as to obtain for the structural parameter of interest a more efficient estimator (in terms of the root mean square error or RMSE criterion) with respect to the usual ordinary least squares “Adjusted for attenuation” (OLSA) estimator for the classical simple linear measurement error model with normal errors (Fuller 1987, chapter 1). The GME approach allows estimating the measurement errors that can be used to adjust the usual composite indicator of the latent variable.

The results of our extensive simulation show the better performance of the GME estimator with respect to the OLSA estimator and the usefulness of the correction with the GME estimated measurement errors for the composite indicator.

The paper is organised as follows: in Sect. 2, the simple linear regression model with a composite indicator as regressor is described; Sect. 3 presents ordinary least squares “Adjusted for attenuation” estimator (OLSA); in Sect. 4, the generalized maximum entropy (GME) estimator and its extension to the simple linear regression model with a composite indicator is presented, with the last section devoted to the estimation of the latent variables; Sect. 5 shows the simulation study and draws conclusions on the performance of the two approaches presented; Sect. 6 reports the results of the two illustrative case studies; conclusions and remarks are given in Sect. 7.

The simple linear MEM with a composite indicator as regressor

In this section we extend the simple linear MEM of Al-Nasser (2005) to the case of multiple indicators for the regressor.

Consider two latent random variables ξ and η, satisfying the following linear deterministic structural relationship:

$$\eta = \alpha + \beta \cdot \xi \qquad (1)$$

where α and β are unknown structural parameters. Usually, the slope β is of primary interest for the analysis.

From these two latent variables one can only obtain multiple indicators, i.e. observe the realization of 1+J random variables, with additive measurement errors ε and δj respectively, that are uncorrelated between them and with the latent variable ξ:

$$Y = \eta + \varepsilon \quad \text{and} \quad X_j = \xi + \delta_j \quad j = 1, 2, \ldots, J \qquad (2)$$

These 1+J reflective relations are represented (Fig. 1).

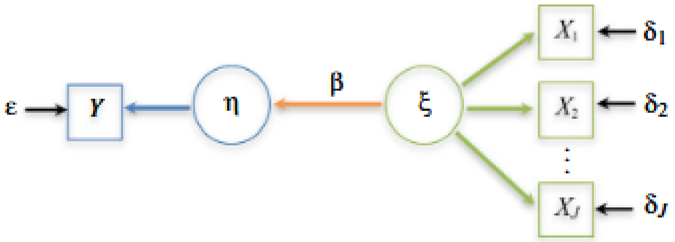


Fig. 1 Path diagram of the simple linear MEM with multiple indicators for the regressor (1)–(2)

As stated in the introduction, our goal is twofold: obtain "good" estimates (i) of the unknown structural parameter β and (ii) of the latent variables ξ and η.

The ordinary least squares "Adjusted for attenuation" estimator (OLSA)

Following Fuller (1987), which considers the case of a single indicator for the regressor, we extend the solution of the estimation problem to the simple linear MEM with multiple indicators described in the previous section.

Given the simple linear MEM in Eq. (2), the unobservable variable ξξ can be "estimated" with the average of X's, i.e. with the simple (equal weights) composite indicator (e.g. Decancq and Lugo 2013, section 4.2):

$$\hat{\xi} = \frac{1}{J}\sum_{j=1}^{J} X_j \qquad (3)$$

Considering the MEM in (1)–(2), for the average of the Xs in (3) we have:

$$\hat{\xi} = \xi + \delta \text{ with } \delta = \frac{1}{J}\sum_{j=1}^{J} \delta_j$$

and, therefore, we obtain the following model specification:

$$Y = \eta + \varepsilon = \alpha + \beta \cdot \xi + \varepsilon = \alpha + \beta \cdot (\hat{\xi} - \delta) + \varepsilon = \alpha + \beta \cdot \hat{\xi} + \nu \qquad (4)$$

Where ν=ε−β·δ. Using a simple random sample of n observations:

$$(x_{1i}, x_{2i}, \ldots, x_{Ji}, y_i) \quad i = 1, 2, \ldots, n \qquad (5)$$

with the standard ordinary least squares (OLS) method applied to the "empirical version" of the simple linear regression (1):

$$y_i = \alpha + \beta \cdot \hat{\xi}_i + v_i \quad i = 1, 2, \ldots, n$$

we can estimate the structural parameter β with the ratio:

$$\hat{\beta}_{OLS} = Cov(\hat{\xi}, y)/Var(\hat{\xi}) \qquad (6)$$

where $$\hat{C}ov(\hat{\xi}, y) = \frac{1}{n}\sum_{i=1}^{n} \hat{\xi}_i y_i - \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\xi}_i\right) \cdot \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)$$

and $$\hat{V}ar(\hat{\xi}) = \frac{1}{n}\sum_{i=1}^{n} \hat{\xi}_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\xi}_i\right)^2.$$

It’s well known that the estimator β^OLS has a downward bias that depends on the “size” of the measurement errors of the X’s, i.e. it depends on the reliability of ξ^.
Assuming equal true correlation between the Xj and ξ, i.e. Cor(Xj,ξ)=ρξ as for the classical parallel measurement error model (Nunnally and Bernstein 1994, chapter 6), the true reliability index of ξ^ is:

$$\kappa_{\xi} = \frac{Var(\xi)}{Var(\hat{\xi})} = \frac{J \cdot \rho_{\xi}^2}{1 + (J - 1) \cdot \rho_{\xi}^2} \qquad (7)$$

The true reliability index κξ take values in the interval between 0 (no reliability) and 1 (maximum reliability).
Using the sample of observations in (5), we can compute the average inter-correlation r¯X (as estimate of ρ2ξ) of the X’s, and the estimated reliability index:

$$\hat{\kappa}_{\xi} = \frac{J \cdot \bar{r}_X}{1 + (J - 1) \cdot \bar{r}_X} \qquad (8)$$

that is the general form of the Spearman-Brown prophecy formula, related to the Cronbach’s Alpha used in the classical item analysis (Nunnally and Bernstein 1994, formula 6.18 and 6.26).
Adopting the measurement error in variable approach, we obtain an unbiased estimate of β using the OLS Adjusted for attenuation (OLSA) estimator (Fuller 1987, formula 1.1.7):

$$\hat{\beta}_{OLSA} = \hat{\beta}_{OLS}/\hat{\kappa}_{\xi} \qquad (9)$$

The standard error (SE) of the OLSA estimator and the SE of the OLS estimator are related by the same relation in (9) (Fuller 1987, formula 1.1.12):

$$SE(\hat{\beta}_{OLSA}) = SE(\hat{\beta}_{OLS})/\hat{\kappa}_{\xi}.$$

In the next section we introduce a new estimator, that uses the estimated reliability index κ^ξ. for the estimate of the structural parameter β, and allows to “adjust” the estimates ξ^ and η^ of the latent variables ξ and η.

The generalized maximum entropy (GME) estimator

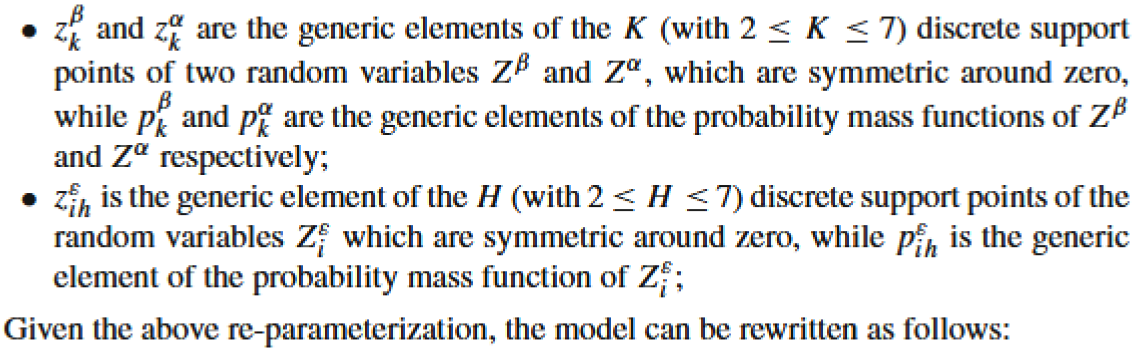
In this study we have developed the Generalized Maximum Entropy (GME) estimator proposed by Golan et al. (1996) for the simple linear MEM with multiple indicators described in Sect. 2. First we introduce the GME estimator and its properties in the case of the simple linear regression model (Sect. 4.1), then the simple linear MEM with a composite indicator is defined in the GME framework (Sect. 4.2), then the problem of the choice of the support points used in the estimation phase is faced (Sect. 4.3), finally the use of the GME residuals for the estimation of the latent variables is proposed.

The GME estimator and its properties

The GME approach applied to the regression model allows estimating at the same time all the parameters and the error terms of the model. More formally, let us consider the following simple linear regression model:
yi=α+xiβ+εi   i=1,2,…,n
the idea of GME is the re-parameterization of the model parameters (α;β) and the error terms (εi) as the expected values of discrete random variables:

$$\alpha = \sum_{k=1}^{K} z_k^{\alpha} p_k^{\alpha} \quad \beta = \sum_{k=1}^{K} z_k^{\beta} p_k^{\beta} \quad \varepsilon_i = \sum_{h=1}^{H} z_{ih}^{\varepsilon} p_{ih}^{\varepsilon} \quad i = 1, 2, \ldots, n$$

where (Golan et al. 1996):



- z_k^{β} and z_k^{α} are the generic elements of the K (with $2 \leq K \leq 7$) discrete support points of two random variables Z^{β} and Z^{α}, which are symmetric around zero, while p_k^{β} and p_k^{α} are the generic elements of the probability mass functions of Z^{β} and Z^{α} respectively;
- z_{ih}^{ε} is the generic element of the H (with $2 \leq H \leq 7$) discrete support points of the random variables Z_i^{ε} which are symmetric around zero, while p_{ih}^{ε} is the generic element of the probability mass function of Z_i^{ε};

Given the above re-parameterization, the model can be rewritten as follows:

$$y_i = \sum_{k=1}^{K} z_k^{\alpha} p_k^{\alpha} + x_i \sum_{k=1}^{K} z_k^{\beta} p_k^{\beta} + \sum_{h=1}^{H} z_{ih}^{\varepsilon} p_{ih}^{\varepsilon}$$

The support points z_k^{α}, z_k^{β} and z_{ih}^{ε} play an important role in the GME estimation procedure, and a relevant issue concerns the choice of their values. In particular, they may be set up by using some objective prior information, fixed ad hoc and/or by using the three sigma rule (Pukelsheim 1994).
The regression parameters α and β and the error terms εiεi are estimated by recovering the corresponding probability mass functions pβk, pαk and pεihp by the maximization of the Shannon’s entropy function (Shannon 1948) given a consistency constraint, that refers to the rewritten model, and adding up normalization constraints as in the following system:

$$H(P) = -\sum_{k=1}^{K} p_k^{\alpha} \cdot ln(p_k^{\alpha}) - \sum_{k=1}^{K} p_k^{\beta} \cdot ln(p_k^{\beta}) - \sum_{i=1}^{n} \sum_{h=1}^{H} p_{ih}^{\varepsilon} \cdot ln(p_{ih}^{\varepsilon})$$

$$y_i = \sum_{k=1}^{K} z_k^{\alpha} p_k^{\alpha} + x_i \sum_{k=1}^{K} z_k^{\beta} p_k^{\beta} + \sum_{h=1}^{H} z_{ih}^{\varepsilon} p_{ih}^{\varepsilon} \quad i = 1, 2, \ldots, n$$

$$\sum_{k=1}^{K} p_k^{\alpha} = 1 \; \sum_{k=1}^{K} p_k^{\beta} = 1 \; \sum_{h=1}^{H} p_{ih}^{\varepsilon} = 1 \quad i = 1, 2, \ldots, n$$

The solution of the above non-linear programming system cannot be given in closed form, and to get the final values a numerical optimization technique (successive quadratic programming method) may be used to compute probabilities. Asymptotic normality property and finite sample approximation can be obtained for the GME estimator (Golan et al. 1996, sections 6.6 and 7.2.1). The GME estimator is asymptotically equivalent to the least squares estimator, and an approximation of its standard error can be computed for finite samples (Golan et al. 1996, formula 7.2.1).

Looking more in depth, the discrete support points zβk and zαk and zεihz can affect the estimate of the parameters and the error terms of the regression model. In particular:

- For the two parameters α and β, the support points zβk and zαk are usually fixed in an interval as follow: [−−100; −−50; 0; 50; 100], with upper and lower limits quite large and symmetric around zero, when we don't have prior information on the real values of the regression parameters. When priors are available, the interval can be changed according with new limits (for instance if we know that the coefficient is usually positive) and symmetric around a new point. Since the interval of the zβk affects in a substantial way the estimates, a sensitivity analysis can be performed to evaluate the stability of the results.
- For error terms εi, the support points zεihε should reflect the variability of the dependent variable, so usually the vector is as follows: [−−3sysy; −−1.5sysy; 0; 1.5sysy; 3sysy], where sy is the estimate of the standard deviation σy (three sigma rule, Pukelsheim 1994).

With respect to other estimators, in general the GME shows some advantages (Golan et al. 1996; Ciavolino and Dahlgaard 2009):

1. does not require distributional error assumptions;
2. is robust for a general class of error distributions;
3. does excellent work with small samples and, in the case of multiple regression, when the number of observations is less than the number of variables, when the design matrix is affected by multi-collinearity;
4. allows to use inequality constraints in the estimation process;
5. allows to employ the set of empirical knowledge about the phenomenon studied and to evaluate its impact on the parameters estimation procedure.

Some limits can involve:

1. high computational cost;
2. estimation linked to the prior information on the data;
3. sensitivity analysis needed;
4. software not well developed.

In the next section we apply the GME approach to the case of the estimate of the parameters of the MEM with a composite indicator as regressor.

The simple linear MEM with a composite indicator in the GME framework

As reported above in the Sect. 3, for the average of the Xs in (3) we have:

$$\hat{\xi} = \xi + \delta \quad \text{with} \quad \delta = \frac{1}{J}\sum_{j=1}^{J} \delta_j, \tag{10}$$

and from (4) with a sample with n data, we consider the following model specification:

$$y_i = \eta_i + \varepsilon_i = \alpha + \beta \cdot \xi_i + \varepsilon_i = \alpha + \beta \cdot (\hat{\xi}_i - \delta_i) + \varepsilon_i \qquad i = 1, 2, \ldots, n. \tag{11}$$

For the model (11) the GME estimator is outlined by the reformulation of the intercept and slope coefficients and the two error terms in the form of expected values of four discrete random variables Zα, Zβ, Zδ and Zε:

$$y_i = \sum_{k=1}^{K} z_k^{\alpha} p_k^{\alpha} + \sum_{k=1}^{K} z_k^{\beta} p_k^{\beta} \cdot (\hat{\xi}_i - \sum_{h=1}^{H} z_h^{\delta} p_{ih}^{\delta}) + \sum_{h=1}^{H} z_h^{\varepsilon} p_{ih}^{\varepsilon} \qquad i = 1, 2, \ldots, n \tag{12}$$

As stated in the previous section, the support points zαkzkα, $z_k^{\alpha}, z_k^{\beta}, z_{ih}^{\varepsilon}$ and z_{ih}^{δ} of these discrete random variables are symmetric around zero, with the associated probability mass functions $p_k^{\alpha}, p_k^{\beta}, p_{ih}^{\delta}, p_{ih}^{\varepsilon}$, pδihpihδ, pεihpihε, which assume values in the interval (0, 1) and respect the following normalization constraints:

$$\sum_{k=1}^{K} p_k^{\alpha} = 1 \qquad \sum_{k=1}^{K} p_k^{\beta} = 1 \qquad \sum_{h=1}^{H} p_{ih}^{\delta} = 1 \qquad \sum_{h=1}^{H} p_{ih}^{\varepsilon} = 1 \qquad i = 1, 2, \ldots, n \tag{13}$$

The idea underlying the GME method is to estimate the unknown parameters and the error terms, by maximizing the following Shannon's entropy function:

$$\begin{aligned} H(p^{\alpha}, p^{\beta}, p^{\delta}, p^{\varepsilon}) = &- \sum_{k=1}^{K} p_k^{\alpha} ln(p_k^{\alpha}) - \sum_{k=1}^{K} p_k^{\beta} ln(p_k^{\beta}) \\ &- \sum_{i=1}^{n} \sum_{h=1}^{H} p_{ih}^{\delta} ln(p_{ih}^{\delta}) - \sum_{i=1}^{n} \sum_{h=1}^{H} p_{ih}^{\varepsilon} ln(p_{ih}^{\varepsilon}) \end{aligned} \tag{14}$$

under the consistency constraints (12) the normalization constraints equation (13).

As explained in the previous section, the solution of the above non-linear problem is not possible in closed form, and to get the final values a numerical optimization technique may be used to compute probabilities; the software used in our study to solve the optimization problem is General Algebraic Modeling System (GAMS); for more details see Golan et al. (1996) and Golan (2006).

Given the estimated probability distributions, it is possible to derive the estimates of the parameters as expected values:

$$\hat{\alpha} = \sum_{k=1}^{K} z_k^{\alpha} \hat{p}_k^{\alpha} \quad \text{and} \quad \hat{\beta} = \sum_{k=1}^{K} z_k^{\beta} \hat{p}_k^{\beta} \tag{15}$$

and for the two error terms:

$$\hat{\delta}_i = \sum_{h=1}^{H} z_h^{\delta} \hat{p}_{ih}^{\delta} \quad \text{and} \quad \hat{\varepsilon}_i = \sum_{h=1}^{H} z_h^{\varepsilon} \hat{p}_{ih}^{\varepsilon} \qquad i = 1, \ldots, n. \tag{16}$$

The choice of the parameters and the error support points is discussed in the next section.

The choice of support points for structural parameters and errors of the model

With the GME approach parameters and error terms are modeled as the expected values of probability mass functions defined over some support points supplied by the researcher. For the structural parameters, the standard choice is a 5 support points uniformly symmetric around 0, with high lower and upper bounds as −−100 and +100 (Golan et al. 1996). In our study we use this standard choice for the simulation (Sects. 5.3) and the two case studies (Sects. 6.1 and 6.2); however, a sensitivity analysis is performed to see if the estimates obtained depend on this choice (Sect. 6.3).

For the two error terms δδ and εε, the choice of the proper support points is more problematic, as it depends on the observed sample as well as any conceptual or empirical information about them. To specify the error bounds for δδ and εε we use the three sigma rule (Sect. 4.1).

Rewriting (10) in terms of variance of the error δδ, that is uncorrelated with ξξ as stated section (2), we have:

$$Var(\hat{\xi}) = Var(\xi) + Var(\delta) \tag{17}$$

and using the true reliability index defined in (7) we obtain:

$$Var(\hat{\xi}) = Var(\hat{\xi}) \cdot \kappa_{\xi} + Var(\delta) \tag{18}$$

Finally, using in (18) the sample variance $\hat{V}ar(\hat{\xi})$ and the sample reliability index $\hat{\kappa}_{\xi}$ defined in (8) we can write:

$$\hat{V}ar(\delta) = \hat{V}ar(\hat{\xi}) \cdot (1 - \hat{\kappa}_{\xi}) \tag{19}$$

In other terms, the variance of the error δ for the model (11) is estimated using sample variance and reliability index of the composite indicator ξ^.

An estimate of the variance of the error εε can easily be obtained from the coefficient of determination R2R2 of the simple linear regression in (11):

$$R^2 = \rho_{\xi y}^2 = 1 - Var(\varepsilon)/Var(y) \quad \text{so that} \quad Var(\varepsilon) = Var(y) \cdot (1 - \rho_{\xi y}^2)$$

where ρξy is the correlation between ξ and y.

Therefore, by computing the sample variance V^ar(y), the sample correlation C^or(ξ^,y) and the adjusted for attenuation estimate of ρξy (Fuller 1987, page 7):

$$\hat{\rho}_{\xi y} = \hat{C}or(\hat{\xi}, y)/\sqrt{\hat{\kappa}_{\xi}} \tag{20}$$

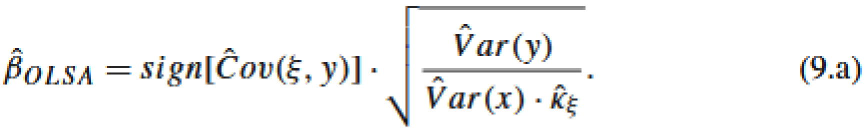
we obtain an estimate of the variance for the error ε:

$$\hat{V}ar(\varepsilon) = \hat{V}ar(y) \cdot (1 - \hat{\rho}_{\xi y}^2) \tag{21}$$

that is not necessarily positive, since the adjusted for attenuation estimate of ρξyρξy in (20) can be equal or greater than one. In this case, as suggested by Fuller (1987) page 7, we use the maximum likelihood estimator ρ^ξy=1ρ^ξy=1 so that (21) is modified in the following manner:

$$\hat{V}ar(\varepsilon) = 0 \qquad (21.a)$$

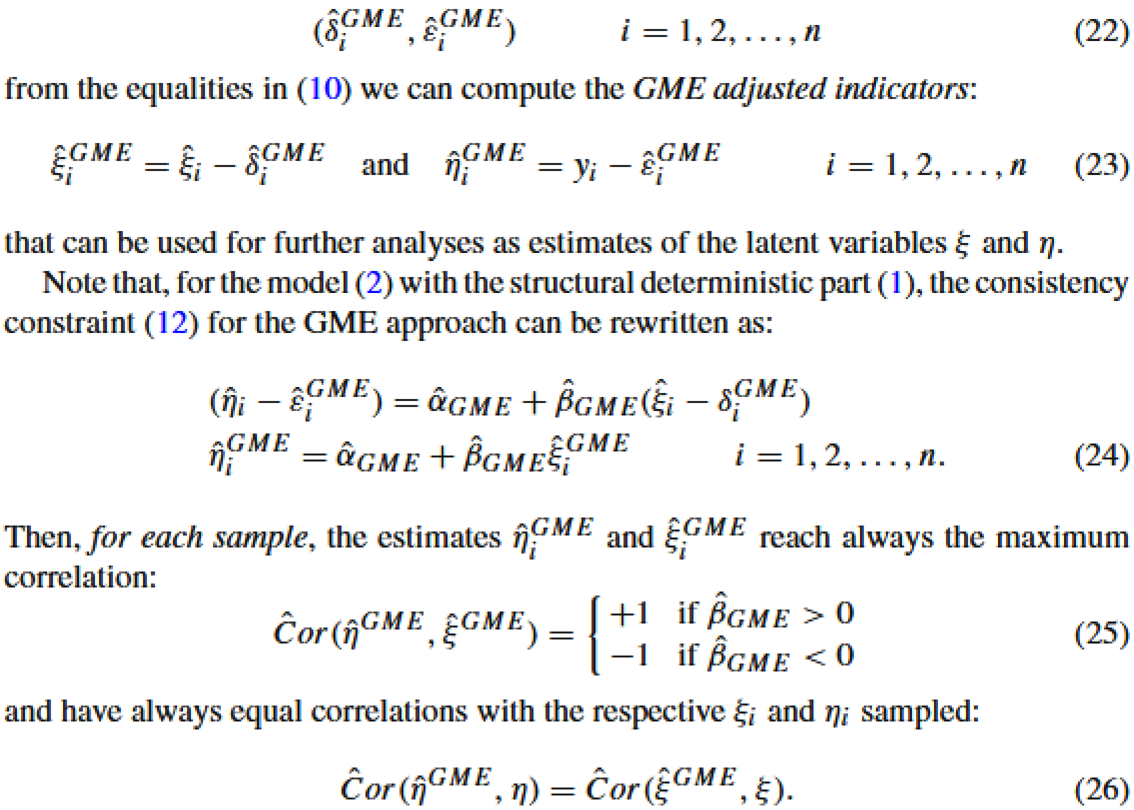
i.e. we assume a degenerate probability distribution in 0 for the error ε, and the maximum likelihood estimator of the slope parameter in (9) is modified in the following manner:



$$\hat{\beta}_{OLSA} = sign[\hat{C}ov(\xi, y)] \cdot \sqrt{\frac{\hat{V}ar(y)}{\hat{V}ar(x) \cdot \hat{\kappa}_{\xi}}}. \qquad (9.a)$$

Using GME residuals for the estimation of the latent variables

As the GME approach allows to estimate the measurement errors δδ and εε for each observation with:



$$(\hat{\delta}_i^{GME}, \hat{\varepsilon}_i^{GME}) \qquad i = 1, 2, \ldots, n \qquad (22)$$

from the equalities in (10) we can compute the *GME adjusted indicators*:

$$\hat{\xi}_i^{GME} = \hat{\xi}_i - \hat{\delta}_i^{GME} \quad \text{and} \quad \hat{\eta}_i^{GME} = y_i - \hat{\varepsilon}_i^{GME} \qquad i = 1, 2, \ldots, n \qquad (23)$$

that can be used for further analyses as estimates of the latent variables ξ and η.

Note that, for the model (2) with the structural deterministic part (1), the consistency constraint (12) for the GME approach can be rewritten as:

$$\begin{aligned} (\hat{\eta}_i - \hat{\varepsilon}_i^{GME}) &= \hat{\alpha}_{GME} + \hat{\beta}_{GME}(\hat{\xi}_i - \delta_i^{GME}) \\ \hat{\eta}_i^{GME} &= \hat{\alpha}_{GME} + \hat{\beta}_{GME}\hat{\xi}_i^{GME} \qquad i = 1, 2, \ldots, n. \end{aligned} \qquad (24)$$

Then, *for each sample*, the estimates $\hat{\eta}_i^{GME}$ and $\hat{\xi}_i^{GME}$ reach always the maximum correlation:

$$\hat{C}or(\hat{\eta}^{GME}, \hat{\xi}^{GME}) = \begin{cases} +1 & \text{if } \hat{\beta}_{GME} > 0 \\ -1 & \text{if } \hat{\beta}_{GME} < 0 \end{cases} \qquad (25)$$

and have always equal correlations with the respective ξ_i and η_i sampled:

$$\hat{C}or(\hat{\eta}^{GME}, \eta) = \hat{C}or(\hat{\xi}^{GME}, \xi). \qquad (26)$$

In the next section we present the results of a simulation and an application with real data, carried out with the objective to evaluate the performance of the GME estimator comparatively with the standard OLSA estimator in (9).

Simulation study

The object of this simulation is to compare the performance of the OLSA and the GME estimators for the structural parameter of the simple linear regression model under the classical assumption of normality and homoscedasticity of the error terms, using two small sample sizes (n= 30 and n= 60), J= 3 observed variables and different levels of reliability (κξ) for the latent exogenous variable, starting from 0.65 and arriving at 0.95, with a 0.05 step.

The performances are evaluated by means of the standard error, the root mean square error and the correlation with the true latent variable (Sect. 4.4). The following sections present the data generating process (Sect. 5.1), the simulation scenario (Sect. 5.2) and a synthesis of the results (Sect. 5.3).

Data generating process

To guarantee the desired reliabilities for the composite indicator ξ^ (Eq. 3), the random samples must be generated fixing the structural parameter β and the equal correlation of the three observed variables XjXj with the latent variable ξ. This goal can be reached solving the formula of the true reliability index κξκξ in (7) with respect to the true correlation ρξρξ between Xj and ξ:

$$\rho_\xi^2 = \frac{\kappa_\xi}{J-(J-1)\cdot\kappa_\xi} = \frac{\kappa_\xi}{3-2\cdot\kappa_\xi}. \qquad (27)$$

Then, the “latent” data are generated under the following classical conditions:
ξ ~N(0,σ2ξ),δh∼N(0,σ2δ),ε∼N(0,σ2ε)ξ∼N(0,σξ2),δh∼N(0,σδ2),ε∼N(0,σε2)
with the latent variable ξξ and the measurement errors δhδh and εε that are uncorrelated. Then we can assign suitable values for the four variances that depend only by the structural parameterββ and by the reliability index κξκξ.
For the measurement error model (2) we define the two restrictions:

$$\sigma_\xi^2 = \rho_\xi^2 \quad \text{and} \quad \sigma_\delta^2 = 1-\rho_\xi^2 \qquad (28)$$

so that each XhXh is a standard normal random variable and the composite indicator ξ^ξ^ in (3) has the requested reliability κξκξ.
From the reliability index

$$\kappa_\eta = \rho_\eta^2 = \frac{\sigma_\eta^2}{\sigma_Y^2} = \frac{\sigma_\eta^2}{\sigma_\eta^2+\sigma_\varepsilon^2} \qquad (29)$$

we obtain

$$\sigma_\varepsilon^2 = \sigma_\eta^2 \cdot \left(\frac{1}{\rho_\eta^2}-1\right). \qquad (30)$$

Therefore, once the correlation ρηρη between Y and ηη is fixed, the variance of the measurement error of Y depends on the variance of the dependent latent variable ηη defined in (1).
From (1) we have:

$$\sigma_\eta^2 = \beta^2\sigma_\xi^2 = \beta^2\rho_\xi^2$$

and therefore we obtain from (30) the third restriction:

$$\sigma_\varepsilon^2 = \beta^2\rho_\xi^2 \cdot \left(\frac{1}{\rho_\eta^2}-1\right) \qquad (31)$$

that depends only by β, ρξ and ρη, so that Y is a (non-standard) normal random variable with reliability index κηκη.
For the model (1)–(2) the restrictions (28) and (31) imply for the covariances that:

$$Cov(X_j, X_l) = \rho_\xi^2 \text{ and } \; Cov(X_j, Y) = \beta\rho_\xi^2$$

and for the correlations that:

$$Cor(X_j, X_l) = \rho_\xi^2 \text{ and } \; Cor(X_j, Y) = \rho_\xi\rho_\eta.$$

For the relations (27) and (29) between the squared correlations (ρ2ξ,ρ2η) and reliability indices (κξ,κη), the restrictions (28) and (31) ensure that repeated random samples of n observations (5) from the normal random variables X’s and Y satisfy on average the desired levels of reliability for the composite indicator ξ^ in (3) and Y.

Simulation scenario

To compare the performance of GME and OLSA estimators for the model (1)–(2), we have fixed the structural parameters to α = 0 and β = 0.5, the sample size n = 30, the correlation κη=ρ2η=0.8 and the true reliability index κξ = (0.65, 0.70, 0.75, 0.80, 0.85, 0.9, 0.95). We chose β = 0.5 as intermediate value between the estimates obtained for the two case studies presented in the next section.

To avoid outliers in the cases of low reliability and/or small sample sizes for the OLSA estimator (that's the ratio $\hat{\beta}_{OLSA} = \hat{\beta}_{OLS}/\hat{\kappa}_{\xi}$), only replications with all positive sample correlations are considered (so that the estimates of κξ are not near to zero).

Simulation results

The first part of Table 1 reports the simulation results about the sample distribution of the estimators of the regression coefficient β: the expectation (as average), the standard error (as standard deviation), and the root mean of squared error based on 2000 replications. Results are divided according to the GME (with support points for structural parameters and errors defined in Sect. 4.3), OLSA and OLS approaches, reported in the rows, and according to different levels of reliability, in the columns.

In the second part of the table, we reported the correlation between the true LV and the estimated LV with both the methods, so as to evaluate the ability of the GME to recover the measurement error.

The GME standard errors, for each level of reliability, are lower then the OLSA, but tend to be the same when the reliability is 0.95.

We must emphasize that the OLS has the lowest variability, but as reported in Fig. 2 on the left, the intervals don't include the real value (0.5) showing the bias of the estimator.

Table 1 Simulation results for the estimator of $\beta = 0.5$ (sample size $n = 30$ and 2000 replications)

Reliability (κ_ξ)	0.65	0.70	0.75	0.80	0.85	0.90	0.95
Averages							
$\hat{\beta}_{GME}$	0.497	0.502	0.501	0.502	0.502	0.494	0.502
$\hat{\beta}_{OLSA}$	0.510	0.511	0.507	0.503	0.498	0.495	0.491
$\hat{\beta}_{OLS}$	0.314	0.339	0.361	0.387	0.411	0.436	0.459
Standard errors							
$\hat{\beta}_{GME}$	0.112	0.102	0.095	0.084	0.071	0.058	0.054
$\hat{\beta}_{OLSA}$	0.132	0.118	0.108	0.093	0.076	0.064	0.055
$\hat{\beta}_{OLS}$	0.057	0.058	0.057	0.056	0.055	0.053	0.050
Root mean square errors							
$\hat{\beta}_{GME}$	0.112	0.102	0.095	0.084	0.071	0.058	0.054
$\hat{\beta}_{OLSA}$	0.132	0.119	0.108	0.093	0.076	0.064	0.055
$\hat{\beta}_{OLS}$	0.194	0.171	0.150	0.126	0.104	0.083	0.065
Correlation with the latent variable							
Simple mean with GME correction							
Average	0.881	0.887	0.894	0.905	0.916	0.929	0.945
Standard deviation	0.034	0.031	0.029	0.025	0.020	0.014	0.008
Simple mean							
Average	0.777	0.804	0.833	0.862	0.889	0.915	0.941
Standard deviation	0.064	0.055	0.049	0.038	0.029	0.019	0.010

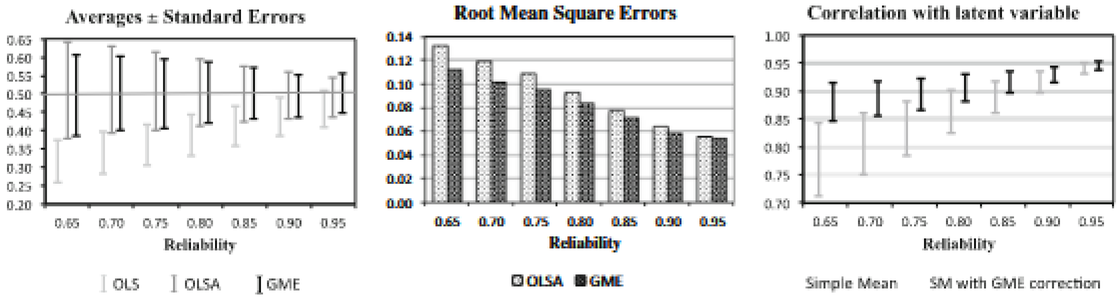


Fig. 2 Simulation statistics (averages, standard errors, RMSE and correlation with LV) for the estimator of $\beta = 0.5$ (sample size n= 30 and 2000 replications)

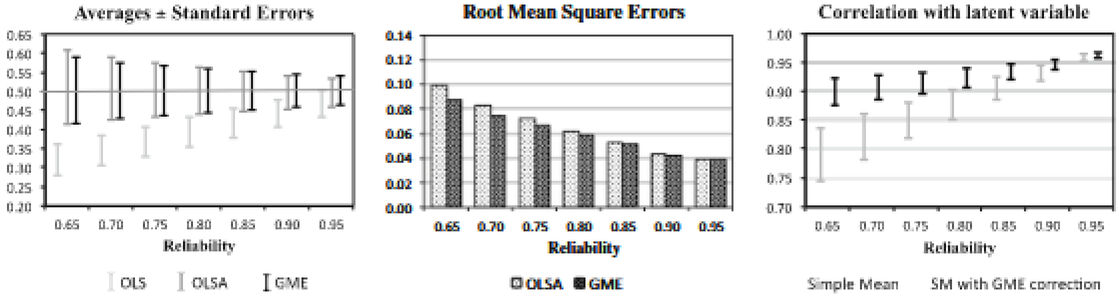


Fig. 3 Simulation statistics (averages, standard errors, RMSE and correlation with LV) for the estimator of $\beta = 0.5$ (sample size $n = 60$ and 2000 replications)

In order to evaluate the estimation accuracy, the deviation of the estimated parameters from the true values has been assessed through the root mean square errors (RMSE). The RMSEs computed show the GME outperforms OLSA in all the levels of reliability, indicating more efficiency for the estimator, and tends to become equal when the reliability is equal to 0.95.

In the end, we evaluated the ability to reproduce the true value of the defined LV, by considering the correlation between the estimated LV, which we called GME adjusted indicator (Eq. 23), and the true LV. The correlations are compared with the indicator obtained with the simple average of the X's variables.

We can notice there are differences between the adjusted indicator and the indicator obtained as simple average, except for the high levels of reliability. The practical relevance of this result is that the composite indicator with the GME correction is more useful for further applications (for example we can use it for a cluster analysis), as it's "more similar" in statistical terms to the unobservable variable (Fig. 3).

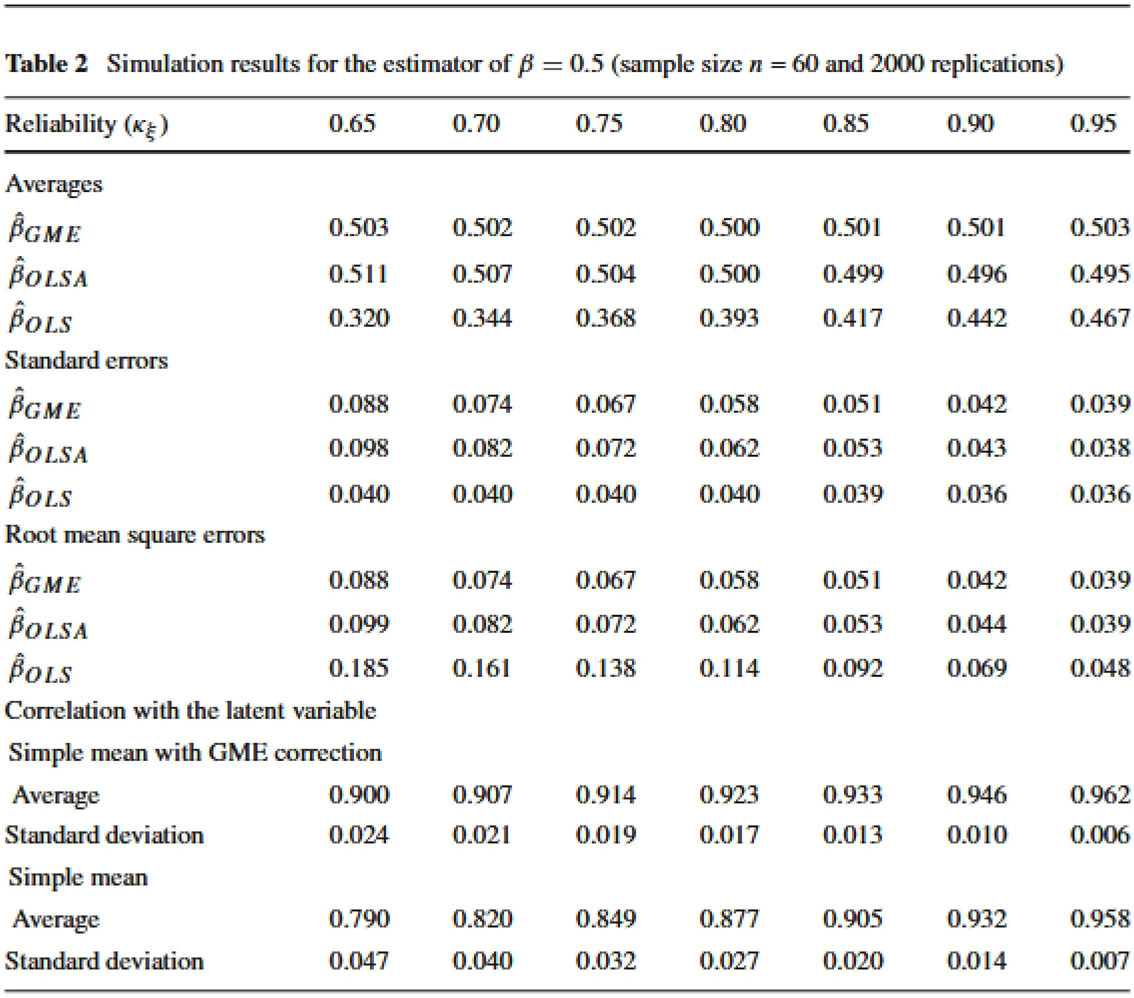


Table 2 Simulation results for the estimator of $\beta = 0.5$ (sample size $n = 60$ and 2000 replications)

Reliability (κ_ξ)	0.65	0.70	0.75	0.80	0.85	0.90	0.95
Averages							
$\hat{\beta}_{GME}$	0.503	0.502	0.502	0.500	0.501	0.501	0.503
$\hat{\beta}_{OLSA}$	0.511	0.507	0.504	0.500	0.499	0.496	0.495
$\hat{\beta}_{OLS}$	0.320	0.344	0.368	0.393	0.417	0.442	0.467
Standard errors							
$\hat{\beta}_{GME}$	0.088	0.074	0.067	0.058	0.051	0.042	0.039
$\hat{\beta}_{OLSA}$	0.098	0.082	0.072	0.062	0.053	0.043	0.038
$\hat{\beta}_{OLS}$	0.040	0.040	0.040	0.040	0.039	0.036	0.036
Root mean square errors							
$\hat{\beta}_{GME}$	0.088	0.074	0.067	0.058	0.051	0.042	0.039
$\hat{\beta}_{OLSA}$	0.099	0.082	0.072	0.062	0.053	0.044	0.039
$\hat{\beta}_{OLS}$	0.185	0.161	0.138	0.114	0.092	0.069	0.048
Correlation with the latent variable							
Simple mean with GME correction							
Average	0.900	0.907	0.914	0.923	0.933	0.946	0.962
Standard deviation	0.024	0.021	0.019	0.017	0.013	0.010	0.006
Simple mean							
Average	0.790	0.820	0.849	0.877	0.905	0.932	0.958
Standard deviation	0.047	0.040	0.032	0.027	0.020	0.014	0.007

Table 2 reports the same analysis we performed in Table 1, with a sample size equal to 60. We can see there are no significant differences in the simulations, that are more reliable, for the sample size effect, keeping an outperform of the GME over the OLSA in the presence of low levels of reliability.

Empirical evidences

In this section we apply the proposed approach to two real case studies, with two different samples sizes, regression coefficients and levels of reliability. The first case study refers to an example proposed by Fuller (1987), for the analysis of manager performance; the second one proposes the study of the relationship between the innovation output and business sophistication of the

European Union Countries. We discuss the results obtained on the real data set, and relate them to the conclusions drawn via the simulation study.

Fuller example

We use the data of Example 2.2.1 in Fuller (1987), where the responses of 55 managers of Iowa farmer co-operatives were analysed with a multiple regression model. The dependent variable Y, is the “Role performance of the manager”, X1 is “Knowledge of the economic phases of management”, X2 is the “Value orientation”, and X3 is “Past training”. We did not use a fourth regressor, the “Role satisfaction”, as it has a negative correlation with the others (see Fuller 1987 for details). We construct the composite indicators using the standardized value of the three Xs.

As in the simulation of the previous section, for the GME estimator we use the support points for structural parameters and errors defined in Sect. 4.3. The estimated reliability, regression coefficients and standard errors are obtained via bootstrap procedure with 2000 replications and bootstrap sample equal to 55.

Table 3 Fuller example results

Correlation matrix	X_1	X_2	X_3
Knowledge - X_1	1		
Value Orientation - X_2	0.375	1	
Past Training - X_3	0.284	0.467	1
Role Performance - Y	0.587	0.494	0.392
Mean corr. of Xs ($\overline{r}_X$)	0.375	Reliability ($\hat{\kappa}_\xi$)	0.643
Regression results			
$R^2 = 0.413$	*Estimate*	*Std.Err.*	*t Stat.*
$\hat{\beta}_{OLS}$	0.122	0.020	6.112
$\hat{\beta}_{OLSA}$	0.204	0.068	3.000
$\hat{\beta}_{GME}$	0.204	0.058	3.517

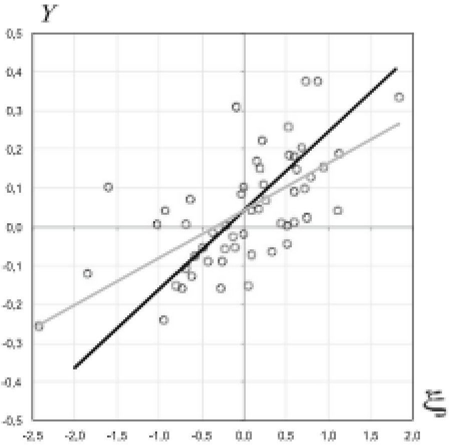


Table 3 sums up the estimation results. In the first part is reported the correlation matrix of the Xs variables and also the correlation with the Role Performance. The mean of the correlation is 0.375 with a corresponding Cronbach’s Alfa equal to 0.643. This number represents the estimate of the reliability, with a value a little bit under the standard threshold of the internal coherence, which is 0.70: considering the results of the simulation in Sect. 5 with the same sample size (Table 2, column 0.70), the expected improvement of the correlation of the LV with its estimate using the composite indicator with the GME correction in formula (23) with respect to the standard approach is about 10 % (0.82 for the simple mean versus 0.91 with the GME correction). The regression coefficients estimated with the two methods show there is no difference between the GME and OLSA, but the biased OLS estimate is about the 60 % of these (0.122 with respect to 0.204). The OLS and GME regression lines are reported in the figure respectively in gray and black. The GME estimator has an estimated standard error lower by about 15 % with respect to that of the OLSA estimator (0.058 with respect to 0.068), so that the value of the t-statistics is greater for the GME estimator.

Innovation example

The example reported concerns the 27 Countries of the European Union (EU) from the Global Innovation Index (GII) 2012 report, for the study innovation level in all Countries of the world. The GII project started in 2007 to define variables to catch innovation including social and business aspects and in way to define a rank and a decision support system identifying targeted policies and good practices for the countries that want to increase their level of innovation.

The GII in based on two sub-indices: the Innovation Input Sub-Index and the Innovation Output Sub-Index. The sub-indices are measured by using different variables that are called pillars. For the Input Sub-Index the pillars are: Institutions, Human capital and research, Infrastructure, Market sophistication, and Business sophistication. For the Output sub-Index the pillars are: Knowledge and technology outputs and Creative outputs. The overall GII score is defined as the simple average of the input and output innovation sub-indices (for details see Dutta 2012). Moreover, the Innovation Efficiency Index is calculated as the ratio of the Output and input Sub-Indices.

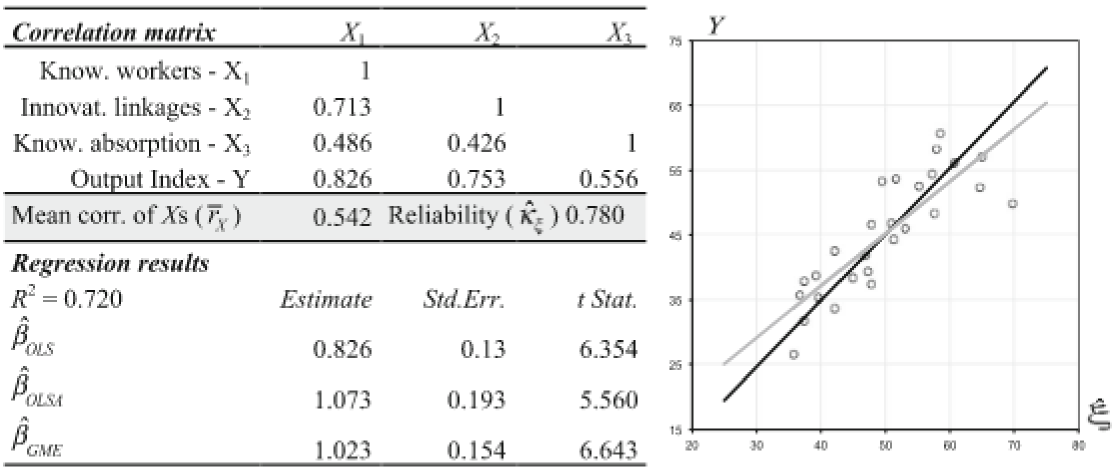
In our example, the relationship between the input and output innovation is defined through the specification of a regression model, where the explicative variable is the pillar of Business Sophistication and as dependent variable the Output sub-index. We selected the pillar Business Sophistication based on the correlation matrix, giving a good level of reliability, as an interesting knowledge input dimension.

As in the previous section, for the GME estimator we use the support points for structural parameters and errors defined in Sect. 4.3. The estimated reliability, regression coefficients and standard errors are obtained via bootstrap procedure with 2000 replications and bootstrap sample equal to 27.

Table 4 reports the correlation matrix of the manifest variables of the pillar Business Sophistication, with an estimated reliability equal to 0.78, above the conventionally accepted threshold: considering the results of the simulation in Sect. 5 with the same sample size (Table 1, column 0.80), the expected improvement of the correlation of the LV with its estimate using the composite indicator with the GME correction in formula (23) with respect to the standard approach is about 6 % (0.86 for the simple mean versus 0.91 with the GME correction).

The regression coefficients estimatedFootnote1 with the two methods show there is a little difference between the GME and OLSA approaches, but the biased OLS estimate is about 80 % of these (0.826 respect to about 1). The OLS and GME regression lines are reported in the figure respectively in gray and black.

Table 4 Innovation example results



Correlation matrix	X_1	X_2	X_3
Know. workers - X_1	1		
Innovat. linkages - X_2	0.713	1	
Know. absorption - X_3	0.486	0.426	1
Output Index - Y	0.826	0.753	0.556
Mean corr. of *X*s ($\overline{r}_X$)	0.542	Reliability ($\hat{\kappa}_\xi$) 0.780	
Regression results			
$R^2 = 0.720$	*Estimate*	*Std.Err.*	*t Stat.*
$\hat{\beta}_{OLS}$	0.826	0.13	6.354
$\hat{\beta}_{OLSA}$	1.073	0.193	5.560
$\hat{\beta}_{GME}$	1.023	0.154	6.643

Y
75
65
55
45
35
25
15
20
30
40
50
60
70
80
$\hat{\xi}$

The GME estimator has an estimated standard error lower by about 20 % with respect to that of the OLSA estimator (0.154 with respect to 0.193), so that the value of the t-statistics is greater for the GME estimator.

Sensitivity analysis

Since the estimations can be affected by the choice of support points, it is useful and recommended (Golan et al. 1996; Ciavolino and Dahlgaard 2009) to verify the sensitivity of the estimated results across different support values.

Results of the previous two sections are obtained with the standard choice of the support points [−−100, −−50, 0, 50, 100]. Table 5 shows the sensitivity analysis results with different numbers and values of support points. In the first five rows, the number of support points is equal to 5; in the second five rows, it is equal to 3. Moreover, the values of the support points vary from 1 to 100. Also in this case, we have used 2000 replications.

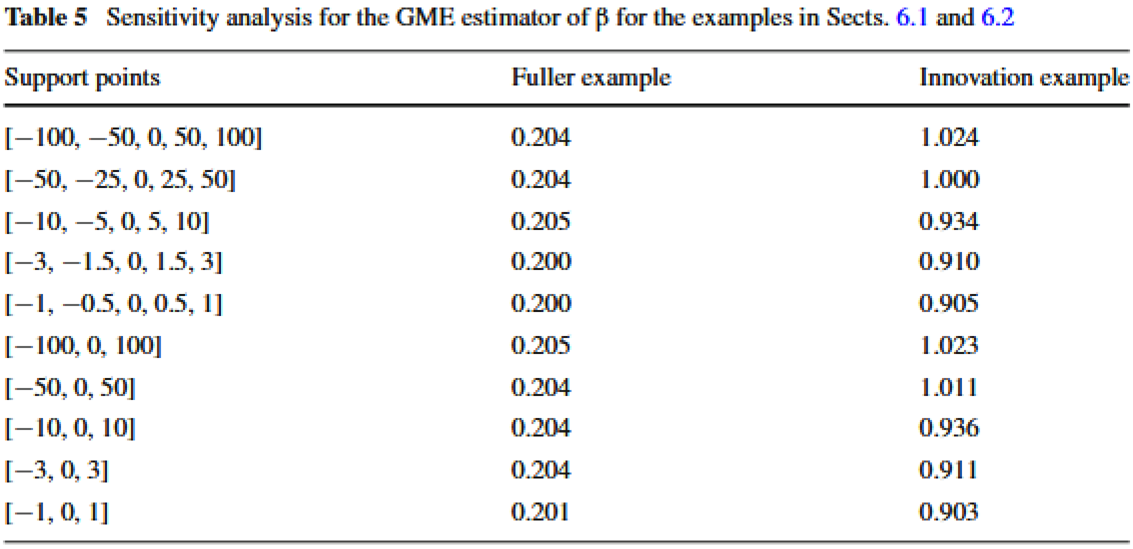


Table 5 Sensitivity analysis for the GME estimator of β for the examples in Sects. 6.1 and 6.2

Support points	Fuller example	Innovation example
[−100, −50, 0, 50, 100]	0.204	1.024
[−50, −25, 0, 25, 50]	0.204	1.000
[−10, −5, 0, 5, 10]	0.205	0.934
[−3, −1.5, 0, 1.5, 3]	0.200	0.910
[−1, −0.5, 0, 0.5, 1]	0.200	0.905
[−100, 0, 100]	0.205	1.023
[−50, 0, 50]	0.204	1.011
[−10, 0, 10]	0.204	0.936
[−3, 0, 3]	0.204	0.911
[−1, 0, 1]	0.201	0.903

The results of the sensitivity analysis show that the GME estimates of the structural parameter β, for both the empirical examples, present quite stable results according to the variation of the support points: for the Fuller Example, the GME estimate of β varies in the interval [0.200; 0.205]; for the Innovation Example, this estimate varies in the interval [0.905; 1.024].

In synthesis, for these two examples the results obtained with the GME estimator of β are not sensitive to the choice of support points.

Concluding remarks

The purpose of this paper was to extend the simple linear measurement error model to include a composite indicator adopting the GME estimator. The idea was to incorporate external information about the reliability of the composite indicator by the definition of the GME errors structure: this approach allows obtaining an estimate of the structural parameter (as the OLSA approach) as well as an estimate of the latent variables. In particular, we compared the performance of our GME estimator with the OLSA estimator, the standard estimator for the MEM under the normality assumption of the errors: following Al-Nasser (2005), to see if our approach worked well, we have experimented it in a standard context.

By means of a Monte Carlo simulation study these two different approaches have been compared in terms of standard error, root mean square errors and estimation accuracy. The model tested in the study was varied for sample sizes and reliability of the composite indicator. Simulation results suggest that the main differences between the two approaches are due to the reliability of the LV. Results indicate a lower variability in the estimates for the GME approach which tend to reach the OLSA approach when we increase the reliability. In term of RMSE, the GME outperforms OLSA for all the simulation conditions defined, but this difference is set to zero when $\kappa\xi= 0.95$.

Finally, the GME adjusted indicator is used to evaluate the aptitude of the proposed method to reproduce the true LV, showing GME is preferable in terms of parameter accuracy.

The standard error, the RMSE and the prediction accuracy are also compared with two sample sizes (n = 30 and n = 60), the results obtained are substantially equal.

The evaluation of the simple linear MEM with the composite indicator as regressor has been integrated with two case studies: The performance evaluation of 55 managers (Fuller example, 1997); the analysis of the innovation of the 27 EU Countries. We chose these two case studies because they represent two real scenarios with different sample sizes, levels of reliabilities and goodness of fit.

Both the examples lead to the same results obtained by the simulation, confirming the inclusion of external information in the model, relative to the composite indicator, and the use of the GME estimator, improving the parameters estimations of the simple linear MEM. We think that these results can be useful for an applied researcher, though they should be extended to more complex cases: the composite indicator obtained from discrete multiple indicators (Ciavolino and Carpita 2015), the composite indicator for the dependent variable and two or more independent variables, nonlinear models (Carpita and Manisera 2012), ordinal data (Pagani and Zanarotti 2015; Vezzoli and Manisera 2012), non-normal and eteroskedastic errors.

References

- Al-Nasser AD (2005) Entropy type estimator to simple linear measurement error models. Aust J Stat 34(3):283–294
- Bollen K (1989) Structural equations with latent variables. Wiley, New York
- Brentari E, Zuccolotto P (2011) The impact of chemical and sensorial characteristics on the market price of Italian red wines. Electron J Appl Stat Anal 4(2):265–276
- Buonaccorsi JP (2010) Measurement error models, methods and applications. Boca Raton: Chapman & Hall, CRC Press
- Carpita M, Manisera M (2012) Constructing indicators of unobservable variables from parallel measurements. Electron J Appl Stat Anal 5(3):320–326
- Carpita M, Ciavolino E (2014) MEM and SEM in the GME framework: statistical modelling of perception and satisfaction. Procedia Econ Financ 17:20–29
- Carroll RJ, Ruppert D, Stefanski LA (1995) Measurement error in nonlinear models. Chapman & Hall, London
- Cheng C-L, Van Ness JW (2010) Statistical regression with measurement error. Wiley, New York
- Ciavolino E, Al-Nasser AD (2009) Comparing generalized maximum entropy and partial least squares methods for structural equation models. J Nonparametr Stat 21(8):1017–1036
- Ciavolino E, Carpita M (2015) The GME estimator for the regression model with a composite indicator as explanatory variable. Qual Quant 49(3):955–965
- Ciavolino E, Carpita M, Al-Nasser AD (2015) Modeling the quality of work in the Italian social co-operatives combining NPCA-RSM and SEM-GME approaches. J Appl Stat 42(1):161–179
- Ciavolino E, Dahlgaard JJ (2009) Simultaneous equation model based on generalized maximum entropy for studying the effect of the management's factors on the enterprise performances. J Appl Stat 36(7):801–815
- Decancq K, Lugo MA (2013) Weights in multidimensional indices of wellbeing: an overview. Econ Rev 32(1):7–34

- Dutta S (2012) The global innovation index 2012: stronger innovation linkages for global growth. INSEAD, France
- Foster JE, McGillivray M, Suman S (2013) Composite indices: rank robustness, statistical association, and redundancy. Econ Rev 32(1):35–56
- Fuller WA (1987) Measurement errors models. Wiley, New York
- Golan A (2006) Information and entropy econometrics. A review and synthesis, foundation and trends®® in Econometrics. 2(1–2), 1–145
- Golan A, Judge G, Miller D (1996) A maximum entropy econometrics: robust estimation with limited data. Wiley, New York
- Madansky A (1959) The fighting of straight lines when both variables are subject to error. J Am Stat Assoc 55:173–205
- Nunnally JC, Bernstein IH (1994) Psychometric theory, 3rd edn. McGraw-Hill, New York
- Oberski DL, Satorra A (2013) Measurement error models with uncertainty about the error variance. Struct Equ Model 20:409–428
- Organisation for Economic Co-operation and Development (2008) Handbook on constructing composite indicators: methodology and user guide. Organisation for Economic Co-operation and Development, Paris
- Pagani L, Zanarotti M (2015) Some considerations to carry out a composite indicator for ordinal data. Electron J Appl Stat Anal 8(3):384–397
- Paruolo P, Saisana M, Saltelli A (2013) Ratings and rankings: voodoo or science? J R Stat Soc Ser A 176(3):609–634
- Pukelsheim F (1994) The three sigma rule. Am Stat 48(2):88–91
- Roeder K, Carroll RJ, Lindsay BG (1996) A semiparametric mixture approach to case–control studies with errors in covariables. J Am Stat Assoc 91:722–732
- Saltelli A (2007) Composite indicators between analysis and advocacy. Soc Indic Res 81:65–77
- Schumacker R, Lomax R (2004) A beginner’s guide to structural equation modeling. Lawrence Erlbaum, Mahwah
- Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423
- Vezzoli M, Manisera M (2012) Assessing item contribution on unobservable variables’ measures with hierarchical data. Electron J Appl Stat Anal 5(3):314–319
- Wansbeek T, Maijer E (2000) Measurement error and latent variables in econometrics. Elsevier, Amsterdam