JAME
E-ISSN:2791-7401
Volume : 4, Issue 1, 2024
URL: https://journals.gen.tr/jame

Received: 2023-12-27
Accepted: 2024-04-25

RESEARCH ARTICLE

A Solution to Errors-in-variables Bias in Multivariate Linear Regression using Compact Genetic Algorithms

Mehmet Hakan Satman[1] **Erkin Diyarbakırlıoğlu**[2]

[1] Prof. Dr., Istanbul University, Faculty of Economics, Department of Econometrics, Türkiye, email: mhsatman@gmail.com
[2] IAE Paris-Est, University of Paris-Est - Cr´eteil, (IRG, EA 2354), Place de la Porte des Champs, France, e-mail: erkin.d@u-pec.fr

Abstract

We address the classical errors-in-variables (EIV) problem in multivariate linear regression with N dependent variables where each left-hand-side variable is a function of a common predictor X subject to measurement error. Our contribution consists in employing the remaining $N-1$ regressions as *extra information* to obtain a filtered version of the mismeasured series X. We test the performance of our approach using simulations whereby we control for different cases like low vs. high R^2 models, small vs. large sample or small vs. large measurement error variances. The results suggest that the multivariate-Compact Genetic Algorithm (mCGA) approach yields estimates with lower mean-square-errors (MSEs). The MSEs are decreasing as the number of dependent variables increases. When there is no measurement error, our method gives results similar to those that would have been obtained by ordinary least-squares.

Keywords: Multivariate regression, Errors-in-variables, Compact genetic algorithms

JEL codes: B23, C13, C63, C61

Citation: SATMAN, M.H. & DİYARBAKIRLIOĞLU, E. (2024). A Solution to Errors-in-variables Bias in Multivariate Linear Regression using Compact Genetic Algorithms. *Journal of Applied Microeconometrics (JAME)*. 4(1),31-64, DOI: 10.53753/jame.2293

Corresponding Author:
Mehmet Hakan Satman
E-mail: mhsatman@gmail.com

1 Introduction

Errors-in-variables (EIV) occur when the observations of one or more variables in a regression model do not match their true values and, consequently, contain a measurement error. Basically, EIV in the left and/or right-hand-side variables in a statistical model can be read as *the observation equals the true values plus measurement error* with side effects ranging from "mild" to "severe" to the researcher. The econometric treatment of EIV is at the origin of a rich yet inconclusive literature going back as far as to Frisch, Berkson or Durbin's pioneering works (Frisch, 1934; Berkson, 1950; Durbin, 1954). More recent and exhaustive treatments of the topic include, among others, Feng et al. (2020), Racicot (2015), Chen et al. (2011), Buonaccorsi (2010), Davidson and MacKinnon (2004, ch. 8), Hausman (2001), Bound et al. (2001), Hyslop and Imbens (2001), Cheng and Van Ness (1999), Dagenais and Dagenais (1997), Griliches (1987) or Fuller (1987), only to cite a few. This paper's primary objective is to contribute to this literature by addressing the classical EIV model in the context of a linear regression setting with multiple dependent variables where a single independent variable X, possibly measured with error, is linked to N dependent variables $Y_1, ..., Y_N$.

Specifically, we extend the work of Satman and Diyarbakirlioglu (2015) who develop a modern approach to deal with EIV that requires no extra information nor additional data to mitigate the bias generated by the measurement error in the independent variable. We consider this feature of our approach as the central block that marks off our work from previous studies in the field. A detailed exposition of existing methods would inevitably extend the paper's scope beyond acceptable limits, so we outline a concise discussion in due course.

A first, and rather naive, way of addressing the measurement error consists in simply ignoring the problem by admitting that it is a difficult one to solve

and extra data may not be available to the researcher. A second approach, known as *Berkson's approach* (Berkson, 1950) considers the observed X values as predetermined given, say, a controlled scientific study or under a laboratory setting. Then, one would consider the observed values of X swinging around their true equivalents due to measurement error. Under this setup, it is then reasonable to assume that the measurement error is no longer correlated with the observed values of the independent variable, which enables one to show that the traditional least-squares estimator of the slope remains still unbiased even when X is mismeasured (Durbin, 1954, p. 24-25). When it comes to social and economic phenomena, however, such controlled experiment settings do not truly exist outside the laboratory.

Another approach relies on correcting the bias in the estimates assuming that the variance of the measurement error in the predictor or that of the unobserved predictor is known. This would then make it possible to derive unbiased estimates of model parameters using the *signal-to-total* variance, typically known as the *reliability ratio*. (Fuller, 1987, p. 5-6) gives a list of some situations where the reliability ratio can be considered as known like IQ test scores. Such situations where one would plausibly assume that the reliability ratio is known are however mostly limited to survey studies in which the data about a particular feature of a set of respondents are obtained over repeated studies of the same nature over time and space.

Given the shortcomings associated with the approaches described briefly above and to the extent that EIV naturally induces a specification error in regression models, instrumental-variables estimation of EIV models constitutes the central prescription to address the issue. The main idea of IV-based processing of EIV consists in using instruments correlated with the true but unobserved values of the predictor *and* uncorrelated with the measurement

error, see, among others, Fuller (1987), Davidson and MacKinnon (2004), Carroll et al. (2006) or Wooldridge (2010) for further developments on the IV-estimation of EIV models. As noted by (Buonaccorsi, 2010, p. 130), the instruments are supposed to carry independent information about the mismeasured predictor which can be used to obtain estimates of the coefficients. That being said, the validity condition of good instruments just stated previously is ironically the unique but also the most critical potential drawback associated with the IV-estimation of EIV models to the extent that poor instruments lead to even more serious consequences (Wooldridge, 2013, p. 499).

The bottomline is that existing approaches commonly require additional information in the form of either better data or valid instruments (Shalabh et al., 2010, p. 718). While there are nice examples where the researcher comes up with an ingenious solution to overcome the impact of a badly measured variable like the studies on the estimates of the economic return to schooling (Ashenfelter and Krueger, 1994; Harmon and Walker, 1995), such additional information may not be available in other situations or there may be no consensus in the field as to what makes an instrument a good one (Klepper and Leamer (1984, p. 163), Dagenais and Dagenais (1997, p. 194)). The approach proposed here is free from such considerations. It does not require any out-of-the-system information about the data-generating process to mitigate the EIV problem. This is a key feature in that the unique *extra* information required lies in the additional dependent variables of the system. We conceive the mismeasured variable $X^{\star}$ as consisting of two blocks, one deterministic and the other stochastic where the deterministic part refers to the true but unobserved portion of it. We then devise an optimization problem that minimizes the squared deviations from the expectation of the response variable conditional on the estimated values of the mismeasured predictor. The latter

variable in turn is the result of an auxiliary dummy regression of the initial variable subject to measurement error. The key difference with the initial Satman and Diyarbakirlioglu (2015) study is that we consider the case where several dependent variables are connected to the same independent variable, potentially measured with error.[1]

We employ numerical experiments to gain insights into the performance of our method. Following a parsimonious strategy to devise the simulations, we consider 36 different configurations to control for features like the relative ratio of regression vs. measurement error variances, low vs. high R^2 models, the sample size and the number of dependent variables. For each configuration, we repeat the estimations 1,000 times. We report the bias, the variance and the mean-square-error of the coefficient estimates for the first dependent variable.[2] The results are promising. The algorithm does capture and correct the bias due to the measurement error in the independent variable, which, when ignored, distorts seriously the parameter estimates. The bias in the slope tends to vanish as the sample size or the number of left-hand-side variables increases. This comes with some cost in the increase of the estimator's variance but overall the increase in the variance is largely offset by the decrease in the bias. This, collectively, yields much smaller MSEs, giving further credit to our approach.

The paper is organized as follows. After a brief discussion of the classical EIV model, section 2 describes our algorithm and gives a short discussion some of its important features. Section 3 presents the simulation results. In section 4,

[1] One should also highlight that the solution presented in Satman and Diyarbakirlioglu (2015) can be considered as a specific case of the setup we develop therein with the number of dependent variables set to 1.

[2] This choice is motivated by the fact that it would be nearly impossible to report every single intercept and slope estimate for each Y_i in the model as this would make the size of the paper cross the acceptable limits. That being said, we saved the entire output from each set of estimations. We have also performed the procedure for 18 additional configurations where we controlled for the performance of our approach when there is no measurement error in the independent variable. For sake of brevity, we do not report the results of these additional simulations in the paper. These results are available upon request.

we give a simple example to illustrate the implementation of our algorithm. Section 5 concludes.

2 Methodology

2.1 The multivariate CGA Algorithm

We first give a sketch of the consequences of the classical EIV model and present our methodology afterwards.

Consider the population model $Y_t^\star = \beta_0 + \beta_1 X_t^\star + \epsilon_t$ for $t = 1, ..., T$ with $\epsilon \sim iid(0, \sigma_e^2)$. The classical errors-in-variables (EIV) model is introduced by assuming that the observations on $Y^\star$ and/or $X^\star$ are recorded with error as $Y_t = Y_t^\star + \nu_t$ and/or $X_t = X_t^\star + \delta_t$ where ν and δ are observation, or measurement errors on Y and X respectively.[3]. While the cost of ν is limited to an inflated variance of the regression error, matters are different when it comes to δ.[4] Assuming $Var(\nu) = 0$, $Var(\delta) > 0$, $E(X^\star\delta) = 0$ and $E(\delta\epsilon) = 0$[5], simple algebra shows that the model can now be expressed as $Y_t = \beta_0 + \beta_1 X_t + \omega_t$ where $\omega_t = \epsilon_t - \beta_1\delta_t$.

Thus, we obtain a *composite* regression error and the predictor X becomes correlated with the new disturbance term as $Cov(X_t, \omega_t) = Cov((X_t^\star + \delta_t), (\epsilon_t - \beta_1\delta_t)) = -\beta_1 Var(\delta)$. This implies that the least-squares estimate will be biased and inconsistent even in large samples.[6] In addition, given the probability limit plim $\widehat{\beta}_1^{LS} = \beta_1 + \frac{Cov(X_t,\omega_t)}{Var(X_t)}$, which is also commonly expressed as plim $\widehat{\beta}_1^{LS} = \beta_1 \left(\frac{Var(X_t^\star)}{Var(X_t^\star)+Var(\delta_t)}\right)$, it can be seen that the slope

[3] Typically, these equations read "the observation is the sum of the true value plus *measurement error*"

[4] If $Var(\nu) > 0$ and $Var(\delta) = 0$, the model can be rewritten as $Y_t = \beta_0 + \beta_1 X_t^\star + (\epsilon_t + \nu_t)$. The new disturbance term is $\epsilon + \nu$. The least-squares estimates of parameters will still be unbiased.

[5] One last assumption holds that the measurement error, by definition, has zero mean, $E(\delta) = 0$.

[6] To see the non-zero covariance between X and ω, note that $E(X_t) = E(X_t^\star + \delta_t) = X_t^\star$ because $E(\delta_t) = 0$. Next, substituting $\omega_t = \epsilon_t - \beta_1\delta_t$ back into $Cov(X_t, \omega_t) = E\left[\left(X_t^\star + \delta_t - \bar{X}^\star\right)(\epsilon_t - \beta_1\delta_t)\right]$ and developing the terms, we obtain $Cov(X_t, \omega_t) = -\beta_1 Var(\delta)$.

estimate is downwards biased as long as $Var(X_t^\star)+Var(\delta_t) > Var(X_t^\star)$.[7] The bias in $\widehat{\beta}$ is known as the *least-squares* attenuation, which Hausman (2001) refers to as the *iron law of econometrics* – "the magnitude of the estimate is usually smaller than expected". The attenuation in $\widehat{\beta}$ also suggests that the bias gets worse when $Var(\delta_t)$ increases relative to $Var(X_t^\star)$.[8] Finally, when there is more than one predictor subject to error, it is no longer possible to derive exact formulas to express neither the sign nor the magnitude of the bias in the slope coefficients because the measurement error on a particular X_t spills over to other model parameters, raising a further puzzling issue, which Cragg (1994) qualifies as the *contamination effect*.

We now turn to our approach and extend the previous setup to accommodate for N dependent variables specified as a function of the same predictor variable. We assume the population model is $\mathbf{Y} = \mathbf{1}_T\boldsymbol{\beta}_0^\top + \mathbf{X}^\star\boldsymbol{\beta}_1^\top + \boldsymbol{\epsilon}$.[9] The true values of $\mathbf{X}$ are not directly given but observed as $\mathbf{X} = \mathbf{X}^\star + \boldsymbol{\delta}$ where $\boldsymbol{\delta}$ is a $T \times 1$ vector of measurement errors. The multivariate EIV model can be rewritten as $\mathbf{Y} = \mathbf{1}_T\boldsymbol{\beta}_0^\top + \mathbf{X}\boldsymbol{\beta}_1^\top + \boldsymbol{\omega}$ where $\boldsymbol{\omega} = \boldsymbol{\epsilon} - \boldsymbol{\delta}\boldsymbol{\beta}_1^\top$ is the $T \times N$ matrix of composite error terms, which make the least-squares estimation of the N slope estimates inconsistent and biased in the same way it does when $N = 1$. To describe how our algorithm works, consider the first two equations of the system that relates the first and second dependent variables Y_{tj}, $j = 1, 2$ to X_t:

$$Y_{t1} = \beta_{01} + \beta_{11}X_t + \omega_{t1}$$

$$Y_{t2} = \beta_{02} + \beta_{12}X_t + \omega_{t2}$$

[7] Consistent estimation of the slope using generalized least-squares is actually possible if the value of the *reliability ratio* $\lambda = Var(X^*)/\left(Var(X^*) + Var(\delta)\right)$ is known. This is however a big "if" because the true value of the reliability ratio is also unknown outside controlled experiment settings (Buonaccorsi, 2010).

[8] The results of our simulations also highlight this fact whereby we pinpoint the case of a high ratio of measurement error variance to independent variable variance.

[9] $\mathbf{Y}$ is a $T \times N$ matrix that contains T observations for N dependent variables, $\mathbf{X}^\star$ is a T-vector of the observations on the true values of the independent variable, $\mathbf{1}$ is a conforming vector of ones, $\boldsymbol{\beta}_0$ is a N-vector of intercepts, $\boldsymbol{\beta}_1$ is a N-vector of slopes, and $\boldsymbol{\epsilon}$ is a $T \times N$ matrix of residuals.

The objective is to estimate the parameters β_{01} and β_{11} for the first equation (as well as σ^2_ω), but also the parameters β_{02} and β_{12} for the second equation too, and so on for any additional Y. The departure point of the extension we propose in this paper relative to the original approach developed in Satman and Diyarbakirlioglu (2015) consists in employing the additional variables Y_{ti} to obtain a new series $\widehat{X}_t^{mCGA}$ for the regressor that can be seen as a *filtered* version of the true yet unobserved values $X_t^\star$. This can be achieved by running the following auxiliary regression of the observed X_t on a set of m dummy variables as,

$$X_t = \alpha_0 + \alpha_1 D_{t1} + \cdots + \alpha_m D_{tm} + \eta_t \tag{1}$$

where α_j are unknown parameters that must be estimated, D_{tj} are $j = 1, ..., m$ dummy variables and η_t are regression residuals. Just like any other regression model one would conceive, this auxiliary regression breaks down the observed series X_t into two components, one deterministic and one random. By construction, the stochastic part η is an estimate of the measurement error δ while the deterministic part represents the series $\widehat{X}_t^{mCGA}$. With no closed-form solution available, the fitted coefficients $\widehat{\alpha}_j$ are devised as solution to the following problem,

$$\underset{\{D_1,\ldots,D_m\}}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}\left(Y_{ti} - \left(\widehat{\beta}_{0i} + \widehat{\beta}_{1i}\widehat{X}_t^{mCGA}\right)\right)^2 \tag{2}$$

where $\widehat{X}_t^{mCGA}$ are themselves the fitted values of the original variable obtained from the auxiliary regression as,

$$\widehat{X}_t^{mCGA} = \widehat{\alpha}_0 + \widehat{\alpha}_1 D_{t1} + \cdots + \widehat{\alpha}_m D_{tm} \tag{3}$$

Finally, the series $\widehat{X}_t^{mCGA}$ is plugged back into the system to estimate Y_{ti} as,

$$\begin{aligned}
\widehat{Y}_{t1} &= \widehat{\beta}_{01} + \widehat{\beta}_{11}\widehat{X}_t^{mCGA} \\
\widehat{Y}_{t2} &= \widehat{\beta}_{02} + \widehat{\beta}_{12}\widehat{X}_t^{mCGA} \\
\vdots \quad &\quad \vdots \\
\widehat{Y}_{tN} &= \widehat{\beta}_{0N} + \widehat{\beta}_{1N}\widehat{X}_t^{mCGA}
\end{aligned} \tag{4}$$

for each $i = 1, ..., N$. Equation (2) defines a quadratic objective function subject to the constraint defined in equation (3). With T observations and $T \times m$ unknown binary values[10], the problem admits theoretically an infinite number of solutions for there is no explicit rules about the appropriate number of dummies that must be used. In their original paper, Satman and Diyarbakirlioglu (2015) address this issue by observing the behaviour of the estimated intercept and slope coefficients. They note that the MSE's of the estimates tend to stabilize about $m = 10$. We follow the same empirical rule in this paper too and use 10 as the default value of this parameter. Besides, even if the number of dummies was known, the results of the algorithm should still be seen as *approximations* sharing, nonetheless, the important feature of systematically smaller MSE's for the estimated regression parameters.[11]

2.2 Discussion

Having set up the mechanics of our approach, we briefly discuss some of its main building blocks.

[10] Recall that m is the number of dummy variables in the auxiliary regression.

[11] One should bear this feature of our method in mind: The procedure does not yield a single *exact* output, the results are likely to vary, at least marginally, from one iteration to another. That does not however mean that the algorithm does not converge, so we conceive these *approximations* as *solutions* of the system.

First, unlike the mainstream literature on EIV models, we conjecture that the additional information to mitigate the EIV bias can be found within the relationship between the set of N dependent variables and the predictor X. One should also underline that the procedure does not require further assumptions about the stochastic behaviour of X, such as its distributional properties. That is one of the key features of the approach initially adopted by Satman and Diyarbakirlioglu (2015), which we aim to develop further in this work, to the extent that standard methods generally require outside information to address the EIV problem. However, as pointed out by (Buonaccorsi, 2010, p. 4-5), getting such extra information either in the form of better data or instrumental variables that satisfy many conditions can be difficult.

Second, we explain briefly the reason why we implement a (compact) genetic algorithm-based solution. Equations (2) and (3) form together a two-stage discrete optimization problem whose objective is to minimize the squared deviations from the conditional expectation of the independent variable on a set of dummy variables and model parameters. Regression of the error-prone variable onto these dummies aims to break down this variable into a clean, but unobserved component and another one that captures the measurement error in X. Since the decision variables of the optimization problem take exclusively binary values, e.g. $D_{tm} \in \{0, 1\}$, a genetic algorithm (GA) happens to be one natural solver to estimate the dummy coefficients α of the auxiliary regression. Developed by pioneering studies like Holland (1975), Holland (1987) or Goldberg (1989), among others, a GA mimics the process of natural selection with, consequently, a related vocabulary borrowing extensively from the Theory of Evolution. A typical GA starts by encoding an array of randomly selected candidate solutions in binary forms, assimilated to *chromosomes*, each member of a larger *population*. The chances a chromosome survives for mating with

another one to generate an *offspring* is determined by a *fitness value*, which is a score associated with an objective function. Iterations continue until no incremental improvement is obtained in terms of the fitness value.

A potential issue associated with GAs concerns the computational difficulties associated with the optimization of the objective function. This is where Compact Genetic Algorithms (CGAs) may be of practical help as they are designed to overcome the issue of computational memory one would face when working with a GA (Harik et al., 2006). Although one would not assert that CGAs are superior to GAs in reaching the global optimum, they represent several advantages. Specifically, in a CGA, candidate solutions are sampled from a given population using a probability vector rather than screening the entire population. The number of iterations is defined with respect to the the population size (Harik et al., 1999). The absence of genetic operators and the sampling strategy employed by a CGA make it a member of Estimation of Distribution Algorithms (EDA) as it always converges to a probability vector through iterations (Pelikan et al., 2002; Baluja, 1994; Larranaga, 2002). Therefore, our choice in implementing the CGA is simply motivated by the fact that the algorithm provides a suitable method to solve the discrete optimization problem defined in equation (3), yet it should be acknowledged that another optimizer handling a similar problem would also be used instead.

Finally, we present some practical, but equally important, aspects of our methodology.[12] Given a set of T observations on $i = 1, ..., N$ dependent variables Y_i and one independent variable X observed with some error δ, the procedure is initialized by setting two user-defined parameters; namely, (1) the number of dummy variables m used in the auxiliary regression specified in equation (3), and (2) the population size. Although there are no specific

[12] We provide in the appendix a pseudo-code of our entire algorithm and an R package (Satman and Diyarbakirlioglu, 2022) including all necessary functions to perform the calculations is readily available on CRAN repositories.

guidelines concerning an adequate value for m, Satman and Diyarbakirlioglu (2015) show using simulations that the mean-square-errors of the slope $\widehat{\beta}_1^{CGA}$ and intercept $\widehat{\beta}_0^{CGA}$ estimators stabilize around $m = 10$.[13] For a given m, the iterations begin with a *probability vector* that represents, to speak CGA, a *chromosome*. For example, a 4-m length chromosome like,

$$P = [0.8, 0.1, 0.7, 0.2]$$

tells that the probability of getting the first dummy equal to 1 is 0.8, the probability of $D_2 = 1$ is 0.1, and so on.[14] Accordingly, given the P in this example, sampling a chromosome like $C = [1, 0, 1, 0]$ is much more likely than sampling another chromosome like $C' = [0, 1, 0, 1]$. Once the number of dummies is chosen, which we set to 10, iterations begin with a probability vector whose elements are initially all equal to 0.5, guaranteeing that no specific dummy coefficient is favoured relative to others. In the next step, the procedure samples two *parents* using the initial P, say C_1 and C_2. The *winner* is the one with the lowest score of the *cost function*, which is specified as the sum of the squared residuals of the corresponding dummy regression. Once the winner C is determined, the vector P is updated using the formula,

$$P_{i+1} = \begin{cases} P_i + \frac{1}{\text{pop. size}} & \text{if } C_i^{winner} = 1 \\ P_i - \frac{1}{\text{pop. size}} & \text{if } C_i^{winner} = 0 \end{cases} \tag{5}$$

Given the new P_{i+1}, the process moves forward by sampling new parents, generating new offsprings and updating thereby P_i. Iterations continue until all

[13] See Satman and Diyarbakirlioglu (2015), figure 1, p. 3225. We also follow the same empirical rule suggested by the authors in the original paper and set $m = 10$ in our applications.

[14] The term *probability vector* should then not be understood in the sense the elements of the vector must sum up to 1. Instead, each element of P defines the probability that the corresponding dummy variable to be equal to 1.

elements of the vector P take either the value of 1 or 0. Note that the *population size* is there for updating iteratively to the probability vector until the stability condition for the auxiliary dummy variables regression is obtained.[15]

3 Simulations

3.1 Setup

We investigate the statistical properties of our approach by Monte Carlo simulations. We specify our data-generating process as follows:

$$\begin{aligned} Y_{ti} &= 5 + 5X_t + \epsilon_{ti} \\ X_t &= X_t^{\star} + \delta_t \\ \epsilon_t &\sim iid\ N\left(0, \sigma_{\epsilon}^2\right) \\ \delta_t &\sim iid\ N\left(0, \sigma_{\delta}^2\right) \end{aligned}$$

The index $t = 1, \ldots, T$ shows the sample size and $i = 1, ..., N$ the number of left-hand-side variables. Each configuration is described by four parameters: (1) The number of left-hand-side variables N, (2) the sample size T, (3) The regression error variance σ_{ϵ}^2 and, (4) the measurement error variance σ_{δ}^2.

We choose three different values for $N \in \{2, 5, 25\}$ to construct the multivariate regression setting and three different sample sizes as $T \in \{30, 50, 100\}$. The measurement error δ is introduced as $X = X^{\star} + \delta$. ϵ and δ are both generated as *iid* normal random variables with zero-mean and constant variance as $\epsilon \sim N(0, \sigma_{\epsilon}^2)$ and $\delta \sim N(0, \sigma_{\delta}^2)$. Regarding the "regression error ϵ & measurement error δ" pairs, we distinguish four configurations as we set $\sigma_{\epsilon} \in \{1, 3\}$

[15] The choice of the population size takes into consideration the trade-off between the *convergence speed* vs. the risk of a local optimum trap. Again, we follow the recommendations of Satman and Diyarbakirlioglu (2015) who suggest that the population size should be 20 or higher. That is said, the authors also note that beyond this limit, the population size has negligible effect on the results. For the record, this parameter is set to 40 in our applications.

together with $\sigma_\delta \in \{0.5, 0.9\}$.[16] Therefore, we distinguish between small vs. large samples as well as low vs. high R^2 models by considering different pairs of regression error vs. measurement error variances, σ_ϵ vs. σ_δ. This allows to study the performance of our method with small vs. large attenuation bias.

Collectively, there are 36 different configurations, which we run 1,000 times each. We thus report the results using a total of 36,000 simulated regressions. We also repeat the experiment for 18 additional configurations[17] whereby we control for the case with no measurement error in X and keep other parameters constant. Our objective is to verify the accuracy of the process in the idealized case and to compare the least-squares with the mCGA method. We find no significant difference between the statistical properties of the two methods. The mCGA therefore conveys no erroneous signal in the absence of measurement error.

3.2 Simulation results

We report our results in Tables 1 to 3. These tables show the results by the number of dependent variables, i.e. $N = 2$, 5, and 25, respectively. We calculate for each configuration the parameter bias as $E(\widehat{\theta}) - \theta$, the variance $Var(\widehat{\theta})$ and the $MSE = \left(E(\widehat{\theta}) - \theta\right)^2 + Var(\widehat{\theta})$ of the intercept β_0 and slope β_1 estimates. We also provide two additional tables in which we organize the results by σ_ϵ&σ_δ pairs and for increasing number of observations T to enable a complementary reading of our numerical experiments. These are given in Tables 4 and 5. As a supplement, we also provide a graphical summary of part of the output in Figures 1 and 2 where we show the bias and the MSE scores of the slope estimates, broken down by the number of dependent variables.

We can make several observations on the basis of our numerical experiment.

[16] We also consider the case with no measurement error by setting $\sigma_\delta = 0$. For sake of brevity, we do not report the results of these configurations, which are available upon request.
[17] The results are available upon request.

First, we observe, regardless of the configuration, that the least-squares estimate of the slope suffers from the attenuation bias when the predictor is subject to measurement error. For example, when we set $\sigma_\epsilon = 1$ and $\sigma_\delta = 0.5$, one would expect that the least-squares estimate of the slope to be biased downwards by 20% relative to the true value of the parameter, which implies that β_1 set initially to 5 will be cut down to 4. This observation holds indeed for the case $(\sigma_\epsilon = 1, \sigma_\delta = 0.5)$ in Tables 1 to 3 regardless of the sample size chosen. The bias in β_1 is even more pronounced when the variability of the measurement error increases relative to that of the regression error.

Second, the performance of our method in mitigating the attenuation bias in the slope is noticeable. In some cases, especially for the $\sigma_\epsilon = 1$ & $\sigma_\delta = 0.5$ pair, the algorithm comes up with an estimate of the slope fairly close to the true value of the parameter. In addition, we observe, as one would expect from our setup, even better-behaved results for the bias in β_1 as we increase the number of dependent variables across Tables 2 and 3. That is said, the decrease in the bias of the slope estimate is not homogeneous as for larger values of the regression error variance. To sum up, the simulations yield a systematically lower bias of the CGA-estimate of the slope $\widehat{\beta}_1^{mCGA}$ relative to that the least-squares estimate $\widehat{\beta}_1^{LS}$ for all configurations.

Third, we look at the variance and mean-square errors (MSE) of the estimates. Overall, we note that the variance of the estimates remains stable across different simulation configurations; while the number of left-hand-side variables has seemingly no effect on the variance, the sample size appears to significantly flatten the variance of the mCGA estimates within a given simulation configuration. Given the decrease in the bias, this results in lower MSE associated with $\widehat{\beta}_1^{mCGA}$, as suggested in Tables 2 to 3 for any configuration one would consider. For a given number of dependent variables N and the σ_ϵ

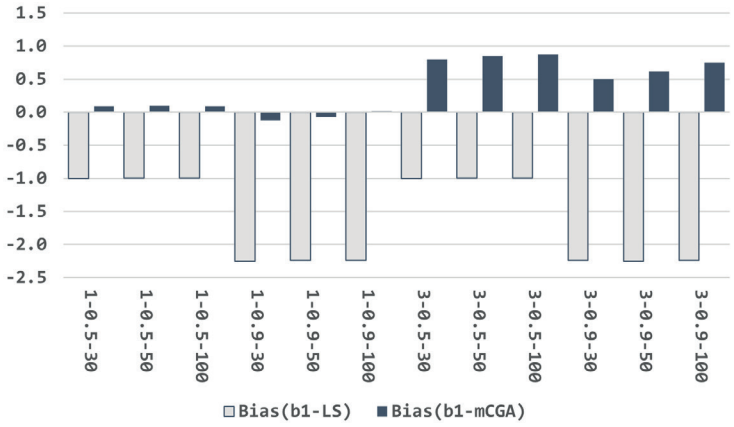
& σ_δ pair, we observe that the MSEs decrease systematically as the sample size increases. This can be easily observed in Table 4. The same observations also hold for the MSE values of the intercept estimates. On the other hand, when we consider the results for the intercept in Tables 1 to 3 and 5, we note that, as expected, the least-squares estimator tends to outperform the mCGA as $\widehat{\beta}_0^{LS}$ remains unbiased with minimum variance even with measurement error. $\widehat{\beta}_0^{mCGA}$, however, bears bias and variance values comparable to those of the least-squares. For example, in the first row of table 3 with $\sigma_\epsilon = 1$ and $\sigma_\delta = 0.5$, we read the bias in $\widehat{\beta}_0^{mCGA}$ as 0.0072 while it is -0.0032 for the LS. In addition, the MSEs of $\widehat{\beta}_0$'s are insensitive to the simulation configurations, seemingly independent of the number of dependent variables and decreasing as the sample size increases. We also note that an increase in the measurement error standard deviation and that of the disturbances tend to degrade the statistical properties of the intercept estimator while an increase in σ_δ has more destructive effects than an increase in σ_ϵ.

To conclude this section, we also give a graphical summary of the message carried out by our numerical experiments. Specifically, in figures 1 and 2, we show using bar charts how the bias and the MSE of estimated β_1's change when one applies the m-CGA estimator (dark bars) relative to least-squares (grey bars). We consider three panels to distinguish the three different values we chose for the number of dependent variables N. The x-axis labels consist of three consecutive numbers that define a given simulation configuration, namely (1) the regression error standard deviation σ_ϵ, (2) the measurement error standard deviation σ_δ, and (3) the sample size T.

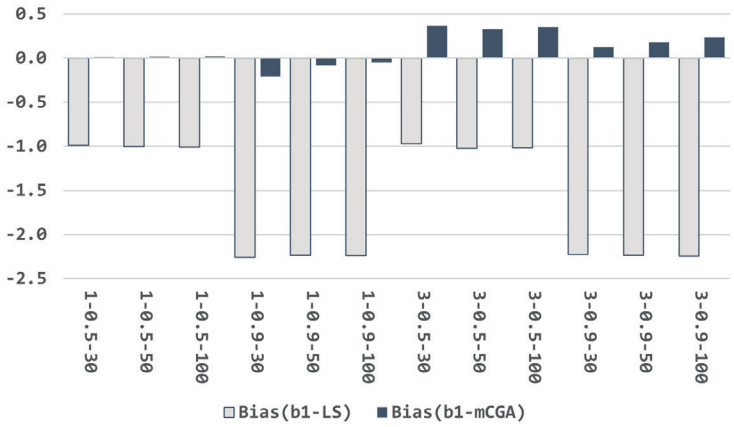
The bars are of the same length regardless of these values for the bias and MSE scores associated with the LS estimates. In a nutshell, the shorter the bars, the better the results, which is the case for every configuration we consider in terms of both bias and MSE of the estimates: The CGA estimates of

Figure 1: Comparing Bias($\widehat{\beta}_1$), OLS vs. mCGA

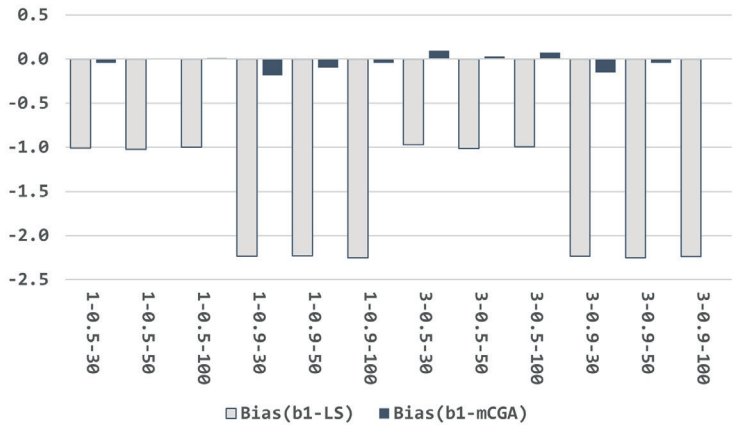
(a) Panel A: N = 2



(b) Panel B: N = 5



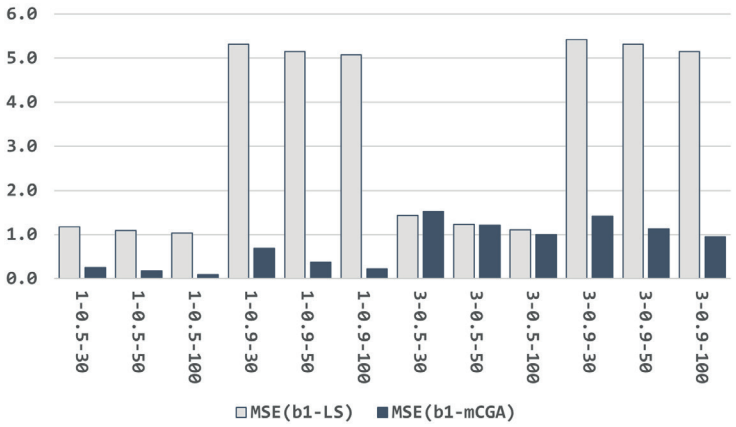
(c) Panel C: N = 25



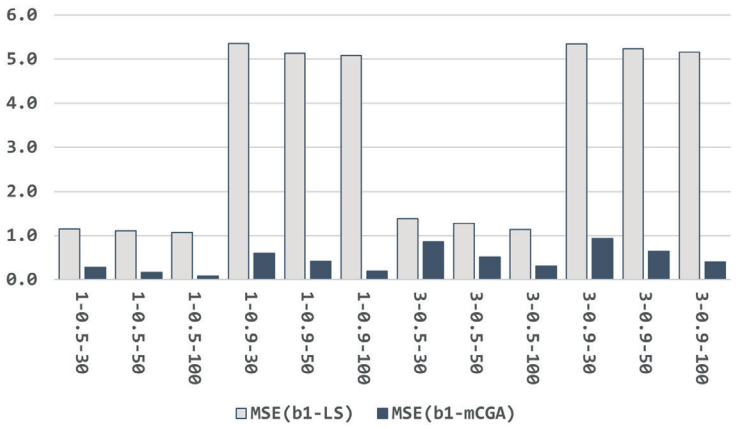
the regression slope have better statistical properties than those of their least-squares equivalents. The downward bias caused by the measurement error is

Figure 2: Comparing MSE($\widehat{\beta}_1$), OLS vs. mCGA

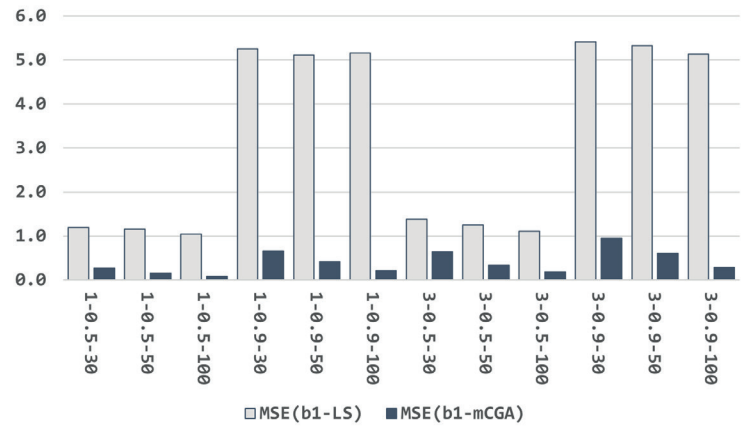
(a) Panel A: N = 2



(b) Panel B: N = 5



(c) Panel C: N = 25



noticeable based on the plots on the left-side of the figures. The CGA estimator is on the other hand successful in pulling the estimate back to its original

value in such a way that Bias $\left(\widehat{\beta}_1\right)$ nearly disappears as suggested in panel C of figure 1 where we consider the case with 25 left-hand-side variables in the multivariate regression.

4 Empirical illustration

We provide a simple empirical illustration of our approach. We prefer an idealized setup for ease of tractability of the results in a multivariate multiple regression model with two response variables and three predictor variables, one of which is measured with error. For $t = 1, ..., 25$ observations and the ith response variable, $i = 1, 2$, we consider the following data generating process:

$$\begin{aligned} Y_{ti} &= \alpha + \beta X_t^{\star} + \gamma_1 W_{t1} + \gamma_2 W_{t2} + \epsilon_{ti} \\ \epsilon_i &\sim iid\ N(0,1) \\ X^{\star}, W_1, W_2 &\sim iid\ N(0,1) \end{aligned}$$

Table 6 shows our artificial dataset. The model parameters α, β, γ_1 and γ_2 are all equal to 5. Therefore, the population regression function for the first response variable is $Y_{t1} = 5 + 5X_t^{\star} + 5W_{t1} + 5W_{t2} + \epsilon_{t1}$. Then, we introduce the measurement error in $X^{\star}$ using $X_t = X_t^{\star} + \delta_t$ where $\delta \sim iid\ N\left(0, 0.5^2\right)$.

We focus on the β coefficient associated with the variable X and fit the following regressions to analyze the behaviour of the coefficient β in,

$$Y_{t1} = \begin{cases} \alpha_i + \beta X_t^{\star} + \gamma_1 W_{t1} + \gamma_2 W_{t2} + \epsilon_{t1} & \text{Model 1: No EIV} \\ \alpha_i + \beta X_t + \gamma_1 W_{t1} + \gamma_2 W_{t2} + \omega_{t1} & \text{Model 2: EIV} \\ \alpha_i + \beta X_t^{mCGA} + \gamma_1 W_{t1} + \gamma_2 W_{t2} + u_{t1} & \text{Model 3: mCGA} \end{cases}$$

The first model is the initial case with no measurement error whereby the least-squares method is expected to yield a BLUE estimator of β. The second model involves the errors-in-variables case where we expect an attenuation bias by 80% in β. The true beta being equal to 5, the fitted beta with error in X should be close to $\beta \times \left(1/(1+0.5^2)\right) = 4$.[18] Model 3 shows the case where we implement our method to mitigate the errors-in-variables bias. Table 7 summarizes the estimation results.

The message of the example is conspicuous. With no error in $X^\star$, on the first column of table 7, the OLS yields virtually "perfect" results as long as we consider the idealized where several assumptions of the estimator hold within the simulation setting from the very beginning. When we add the measurement error δ and run the model using X, as reported by model 2, the point estimate of β is downsized, unsurprisingly, by more than 20%, going down from 5.048 to 3.931, with a standard error nearly twice as much as the one found by the OLS. Finally, the mCGA estimator is remarkably successful as it pulls the β associated with the mismeasured variable back to 4.662.

Additional practical and important observations concern the pairwise relationships between the variables of interest. We mentioned earlier in section **??** that the principal feature of the algorithm we devise consists in filtering out the variable X into two components as $X = \widehat{X}^{mCGA} + \hat{\eta}$: The random component stands for the estimate of the additive measurement error δ while the deterministic component $\widehat{X}^{mCGA}$ matches the *fitted* X, which we use in the second-stage regressions.[19] These two parts must then be uncorrelated. This is indeed the case: The sample correlation between the measurement error δ and the fitted errors is $Cor(\delta, \hat{\eta}) = 0.8786$, suggesting that the algorithm

[18] The calculation is possible thanks to the knowledge about the measurement error and regression error variances.

[19] There are other instances in the EIV literature following a similar *two-stage path* like the one we introduce here. See, among others, Dagenais and Dagenais (1997), Racicot (2015).

comes up with an accurate estimate of the measurement error. In addition, by assumptions of the classical EIV model, we expect the measurement error δ to be uncorrelated with the true values of the predictor $X^{\star}$. The weak sample correlation between the two series in our data validates this insight: $Cor(\delta, X^{\star}) = 0.2007$. Finally, and above all else, the filtered series $\widehat{X}^{mCGA}$ has a very strong correlation with the true values $X^{\star}$ (assumed unobserved). The correlation between $\widehat{X}^{mCGA}$ and $X^{\star}$ is 0.9726. In words, the method comes up with a *clean series* for the variable of interest, by providing very close to the true but unobserved series.

5 Conclusion

This paper addresses the classical errors-in-variables problem in multivariate linear regression by introducing a compact genetic algorithm-based estimator designed to mitigate the EIV bias. We build on the original work by Satman and Diyarbakirlioglu (2015). The authors developed a framework that considers the measurement error problem within a constrained convex optimization setting and generates a cleaner version of the error-prone regressor with no outside information. This paper extends their idea in a multivariate regression system involving N response variables, where each variable is a function of the same regressor and, doing so, aims to take advantage of the additional information provided by the $N-1$ variables to obtain better-behaved estimates of model coefficients.

In the same spirit as the original paper, our approach consists of a two-stage optimization process in which the first stage comes up with a filtered version of the independent variable through an auxiliary dummy-variables regression. The new series is then plugged back into the initial model in the second stage to mitigate the EIV problem. We perform extensive simulation analyses to

assess our approach. We consider several control parameters like the sample size, the number of dependent variables of the multivariate regression or the regression vs. measurement error variances. We also provide a simple empirical application of our method again using simulated data to further highlight the accuracy of our approach. To summarize, the results overall suggest that the inclusion of additional response variables as extra information reduces the bias at the expense of a relatively tolerable increase in the variance. That is said, the increase in the variance is largely offset as we observe systematically smaller MSE's in all simulation configurations, endorsing the performance of our approach.

There are several options for future studies. A direct extension would focus on an in-depth investigation of the statistical properties of our estimator. Although the slight increase in the variance is substantially offset by lower bias in the coefficient estimates, yielding, collectively, systematically lower mean-square-errors, studying other features of our method like the consistency, decision error probabilities or robustness is equally desirable. Therefore, we consider such a simulation-driven study as a starting point for future work to provide further credit to the framework we aim to develop. Another avenue for future work concerns the empirical ground so that one would check the CGA estimator in action with real data.[20] There are of course several instances in different disciplines in which the model involves a linear relationship between several response variables each function of the same set of regressors. For example, a particularly interesting case in financial economics is the so-called *factor pricing models* where many dependent variables, i.e. returns on a set of assets or portfolios, are modelled as a linear function of a given set of independent variables, i.e. risk factors. In a recent study conducted by

[20] As a matter of fact, the main issue related to empirical work is that it is rarely possible to know about the population model, making the comparison of the results with those obtained from simulations difficult.

Diyarbakirlioglu et al. (2022), the authors provide a real-world data application of the original method developed in Satman and Diyarbakirlioglu (2015). Specifically, they focus on the impact of the measurement error on the *market risk factor* for a large number of test assets across three popular asset pricing models, namely the Capital Asset Pricing Model, the Fama-French three-factor model and the Fama-French five-factor model. We thus leave the investigation of the behaviour of our method in these financial models, which are basically multivariate-multiple regressions, for future work.

Disclosure statement

We hereby certify that the material presented in the manuscript is the authors' original work, currently not being considered for publication elsewhere. The content therefore reflects the authors' own research in a truthful and complete manner. There are no conflicts of interest.

References

Ashenfelter, O. and Krueger, A. (1994). Estimates of the economic returns to schooling from a new sample of twins. *American Economic Review*, 84(5):1157–73.

Baluja, S. (1994). Population-based incremental learning: A method for integrating genetic search based function optimization and competitive learning.

Berkson, J. (1950). Are there two regressions? *Journal of the Americal Statistical Association*, 45(250):164–180.

Bound, J., Brown, C., and Mathiowetz, N. (2001). Measurement error in survey data. In Heckman, J. J. and Leamer, E. E., editors, *Handbook of Econometrics*, volume 5, pages 3705–3843. Elsevier. Section: 59.

Buonaccorsi, J. P. (2010). *Measurement Error: Models, Methods, and Applications*. Chapman & Hall.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman & Hall.

Chen, X., Hong, H., and Nekipelov, D. (2011). Nonlinear models of measurement errors. *Journal of Economic Literature*, 49(4):901–937.

Cheng, C.-L. and Van Ness, J. W. (1999). *Statistical Regression with Measurement Error*. Wiley.

Cragg, J. G. (1994). Making good inferences from bad data. *Canadian Journal of Economics*, 27(4):776–800.

Dagenais, M. G. and Dagenais, D. L. (1997). Higher moment estimators for linear regression models with errors in the variables. *Journal of Econometrics*, 76(1):193–221.

Davidson, R. and MacKinnon, J. G. (2004). *Econometric Theory and Methods*. Oxford University Press.

Diyarbakirlioglu, E., Desban, M., and Lajili-Jarjir, S. (2022). Asset pricing models with measurement error problems: A new framework with compact genetic algorithms. *Finance*, 43(2):1–78.

Durbin, J. (1954). Errors in variables. *International Statistical Review*, 22(1):23–32.

Feng, Z., Zhang, J., and Chen, Q. (2020). Statistical inference for linear regression models with additive distortion measurement errors. *Statistical Papers*, 61(6):2483–2509.

Frisch, R. (1934). *Statistical confluence analysis by means of complete regression systems*. Universitetets Okonomiske Instituut.

Fuller, W. A. (1987). *Measurement Error Models*. John Wiley & Sons, Inc.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc.

Griliches, Z. (1987). Economic data issues. In Griliches, Z. and Intriligator, M. D., editors, *Handbook of Econometrics*, volume 3, chapter 25, pages 1465–1514. Elsevier, 1 edition.

Harik, G. R., Lobo, F. G., and Goldberg, D. E. (1999). The compact genetic algorithm. *IEEE Transactions on Evolutionary Computation*, 3(4):287–297.

Harik, G. R., Lobo, F. G., and Sastry, K. (2006). *Linkage Learning via Probabilistic Modeling in the Extended Compact Genetic Algorithm (ECGA)*, pages 39–61. Springer Berlin Heidelberg, Berlin, Heidelberg.

Harmon, C. and Walker, I. (1995). Estimates of the economic return to schooling for the united kingdom. *American Economic Review*, 85(5):1278–86.

Hausman, J. A. (2001). Mismeasured variables in econometric analysis: Problems from the right and problems from the left. *Journal of Economic*

Perspectives, 15(4):57–67. Publisher: American Economic Association.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*. University of Michigan Press.

Holland, J. H. (1987). Genetic algorithms and classifier systems: Foundations and future directions. In *ICGA*, pages 82–89.

Hyslop, D. R. and Imbens, G. W. (2001). Bias from classical and other forms of measurement error. *Journal of Business & Economic Statistics*, 19(4):475–481.

Klepper, S. and Leamer, E. E. (1984). Consistent sets of estimates for regressions with errors in all variables. *Econometrica*, 52(1):163–183.

Larranaga, P. (2002). A review on estimation of distribution algorithms. In *Estimation of distribution algorithms*, pages 57–100. Springer.

Pelikan, M., Goldberg, D. E., and Lobo, F. G. (2002). A survey of optimization by building and using probabilistic models. *Computational Optimization and Applications*, 21(1):5–20.

Racicot, F.-E. (2015). Erreurs de mesure sur les variables économiques et financières. *La Revue des Sciences de Gestion*, 3-4(267-268):79–103.

Satman, M. H. and Diyarbakirlioglu, E. (2015). Reducing errors-in-variables bias in linear regression using compact genetic algorithms. *Journal of Statistical Computation and Simulation*, 85(16):3216–3235.

Satman, M. H. and Diyarbakirlioglu, E. (2022). *eive: An Algorithm for Reducing Errors-in-variables Bias in Linear Regression*. R package version 3.1.0.

Shalabh, Garg, G., and Misra, N. (2010). Consistent estimation of regression coefficients in ultrastructural measurement error model using stochastic prior information. *Statistical Papers*, 51(3):717–748.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross-Section and Panel Data*. MIT Press.

Wooldridge, J. M. (2013). *Introductory Econometrics: A Modern Approach*. South-Western Cengage Learning.

Table 1: Simulation results, $N = 2$

		$\widehat{\beta}_0^{LS}$			$\widehat{\beta}_0^{mCGA}$			$\widehat{\beta}_1^{LS}$			$\widehat{\beta}_1^{mCGA}$		
		bias	var	MSE	bias	var	MSE	bias	var	MSE	bias	var	MSE
T = 30													
σ_ϵ	σ_δ												
1.0	0.5	-0.0100	0.2129	0.2130	-0.0144	0.2717	0.2719	-1.0021	0.1736	1.1778	0.0876	0.2419	0.2495
1.0	0.9	-0.0001	0.4900	0.4900	-0.0222	0.8026	0.8031	-2.2521	0.2413	5.3134	-0.1240	0.6710	0.6863
3.0	0.5	-0.0097	0.4880	0.4881	-0.0178	0.6680	0.6683	-1.0029	0.4324	1.4381	0.8016	0.8770	1.5195
3.0	0.9	0.0324	0.7209	0.7220	0.0294	1.1841	1.1850	-2.2389	0.4123	5.4251	0.4999	1.1643	1.4142
T = 50													
σ_ϵ	σ_δ												
1.0	0.5	-0.0018	0.1215	0.1215	0.0019	0.1580	0.1580	-0.9963	0.1004	1.0930	0.0937	0.1704	0.1792
1.0	0.9	0.0290	0.2489	0.2498	0.0101	0.4336	0.4337	-2.2391	0.1383	5.1518	-0.0721	0.3637	0.3689
3.0	0.5	0.0186	0.3179	0.3182	0.0071	0.4276	0.4276	-0.9966	0.2343	1.2275	0.8482	0.4902	1.2096
3.0	0.9	-0.0023	0.4195	0.4195	-0.0444	0.7585	0.7604	-2.2527	0.2444	5.3189	0.6207	0.7463	1.1315
T = 100													
σ_ϵ	σ_δ												
1.0	0.5	-0.0053	0.0626	0.0626	-0.0068	0.0763	0.0764	-0.9941	0.0484	1.0366	0.0914	0.0804	0.0887
1.0	0.9	0.0028	0.1230	0.1230	-0.0146	0.2121	0.2123	-2.2369	0.0684	5.0720	0.0161	0.2227	0.2230
3.0	0.5	-0.0055	0.1580	0.1581	-0.0043	0.2005	0.2005	-0.9982	0.1096	1.1060	0.8713	0.2374	0.9965
3.0	0.9	-0.0489	0.2259	0.2283	-0.0465	0.3787	0.3809	-2.2424	0.1995	5.1480	0.7515	0.3820	0.9468

The table shows the simulation results for the model $Y_{ti} = 5 + 5X_t + \epsilon_{ti}$, where $i = 1, 2$ is the number of dependent variables and $t = 1, ..., T$ the sample size. We consider four σ_ϵ & σ_δ pairs for the regression error and the measurement error in X_t's as $X_t = X_t^\star + \delta_t$. Each configuration has been run 1000 times. Table entries show the bias, variance and MSE of the regression intercept and slope estimates from least squares and mCGA methods.

Table 2: Simulation results, $N = 5$

		$\widehat{\beta}_0^{LS}$			$\widehat{\beta}_0^{mCGA}$			$\widehat{\beta}_1^{LS}$			$\widehat{\beta}_1^{mCGA}$		
		bias	var	MSE	bias	var	MSE	bias	var	MSE	bias	var	MSE
T = 30													
σ_ϵ	σ_δ												
1.0	0.5	-0.0195	0.1971	0.1974	-0.0140	0.2486	0.2488	-0.9876	0.1713	1.1467	0.0131	0.2748	0.2750
1.0	0.9	-0.0036	0.4200	0.4200	-0.0020	0.6908	0.6908	-2.2589	0.2557	5.3582	-0.2095	0.5540	0.5979
3.0	0.5	-0.0296	0.4699	0.4708	-0.0187	0.5622	0.5626	-0.9715	0.4332	1.3770	0.3680	0.7208	0.8562
3.0	0.9	0.0044	0.6706	0.6706	-0.0058	1.0244	1.0244	-2.2260	0.3907	5.3458	0.1230	0.9173	0.9324
T = 50													
σ_ϵ	σ_δ												
1.0	0.5	-0.0131	0.1167	0.1169	-0.0066	0.1461	0.1461	-1.0052	0.1030	1.1135	0.0166	0.1650	0.1653
1.0	0.9	-0.0091	0.2602	0.2602	0.0023	0.4494	0.4494	-2.2357	0.1361	5.1346	-0.0812	0.4095	0.4161
3.0	0.5	0.0066	0.2818	0.2819	0.0048	0.3353	0.3353	-1.0225	0.2312	1.2766	0.3303	0.4015	0.5106
3.0	0.9	0.0244	0.4208	0.4214	0.0015	0.6434	0.6434	-2.2359	0.2381	5.2375	0.1811	0.6141	0.6469
T = 100													
σ_ϵ	σ_δ												
1.0	0.5	-0.0018	0.0603	0.0603	-0.0033	0.0744	0.0744	-1.0065	0.0543	1.0674	0.0229	0.0823	0.0828
1.0	0.9	-0.0051	0.1177	0.1177	-0.0147	0.2136	0.2138	-2.2379	0.0749	5.0832	-0.0501	0.1953	0.1978
3.0	0.5	-0.0072	0.1346	0.1346	-0.0079	0.1665	0.1666	-1.0177	0.1043	1.1400	0.3513	0.1835	0.3069
3.0	0.9	-0.0009	0.1982	0.1982	-0.0036	0.3197	0.3197	-2.2449	0.1185	5.1580	0.2349	0.3429	0.3981

The table shows the simulation results for the model $Y_{ti} = 5 + 5X_t + \epsilon_{ti}$, where $i = 1, ..., 5$ is the number of dependent variables and $t = 1, ..., T$ the sample size. We consider four σ_ϵ & σ_δ pairs for the regression error and the measurement error in X_t's as $X_t = X_t^\star + \delta_t$. Each configuration has been run 1000 times. Table entries show the bias, variance and MSE of the regression intercept and slope estimates from least squares and CGA methods.

Table 3: Simulation results, $N = 25$

		$\widehat{\beta}_0^{LS}$			$\widehat{\beta}_0^{mCGA}$			$\widehat{\beta}_1^{LS}$			$\widehat{\beta}_1^{mCGA}$		
		bias	var	MSE	bias	var	MSE	bias	var	MSE	bias	var	MSE
T = 30													
σ_ϵ	σ_δ												
1.0	0.5	-0.0032	0.2222	0.2222	0.0072	0.2547	0.2547	-1.0087	0.1717	1.1891	-0.0445	0.2671	0.2691
1.0	0.9	0.0071	0.4104	0.4104	-0.0080	0.6660	0.6661	-2.2353	0.2545	5.2509	-0.1845	0.6258	0.6598
3.0	0.5	-0.0154	0.5158	0.5161	-0.0117	0.5778	0.5779	-0.9725	0.4306	1.3763	0.0969	0.6313	0.6407
3.0	0.9	0.0110	0.6817	0.6818	0.0325	0.9573	0.9583	-2.2341	0.4196	5.4106	-0.1492	0.9235	0.9457
T = 50													
σ_ϵ	σ_δ												
1.0	0.5	0.0125	0.1198	0.1199	0.0119	0.1476	0.1477	-1.0225	0.1086	1.1540	0.0037	0.1592	0.1593
1.0	0.9	-0.0308	0.2465	0.2475	-0.0144	0.3983	0.3985	-2.2302	0.1392	5.1129	-0.0969	0.4075	0.4169
3.0	0.5	-0.0126	0.2707	0.2709	-0.0108	0.3081	0.3082	-1.0134	0.2248	1.2519	0.0283	0.3352	0.3360
3.0	0.9	0.0156	0.4292	0.4294	0.0074	0.6543	0.6544	-2.2545	0.2419	5.3246	-0.0456	0.6059	0.6079
T = 100													
σ_ϵ	σ_δ												
1.0	0.5	0.0083	0.0601	0.0602	0.0121	0.0730	0.0731	-0.9960	0.0500	1.0420	0.0127	0.0785	0.0787
1.0	0.9	0.0224	0.1279	0.1284	0.0263	0.2120	0.2127	-2.2544	0.0749	5.1572	-0.0433	0.2070	0.2089
3.0	0.5	0.0085	0.1469	0.1470	0.0117	0.1607	0.1609	-0.9946	0.1172	1.1065	0.0717	0.1768	0.1820
3.0	0.9	-0.0040	0.2054	0.2054	-0.0017	0.3125	0.3125	-2.2386	0.1234	5.1349	0.0053	0.2884	0.2885

Notes: The table shows the simulation results for the model $Y_{ti} = 5 + 5X_t + \epsilon_{ti}$, where $i = 1, ..., 25$ is the number of dependent variables and $t = 1, ..., T$ the sample size. We consider four σ_ϵ & σ_δ pairs for the regression error and the measurement error in X_t's as $X_t = X_t^\star + \delta_t$. Each configuration has been run 1000 times. Table entries show the bias, variance and MSE of the regression intercept and slope estimates from least squares and CGA methods.

Table 4: Simulation results for β_1

σ_ϵ	σ_δ	T	Bias $\widehat{\beta}_1^{LS}$	Bias $\widehat{\beta}_1^{mCGA}$	$Var(\widehat{\beta}_1^{LS})$	$Var(\widehat{\beta}_1^{mCGA})$	MSE $\widehat{\beta}_1^{LS}$	MSE $\widehat{\beta}_1^{mCGA}$
			Panel A: Number of dependent variables, N = 2					
1	0.5	30	-1.0021	0.0876	0.1736	0.2419	1.1778	0.2495
		50	-0.9963	0.0937	0.1004	0.1704	1.0930	0.1792
		100	-0.9941	0.0914	0.0484	0.0804	1.0366	0.0887
1	0.9	30	-2.2521	-0.1240	0.2413	0.6710	5.3134	0.6863
		50	-2.2391	-0.0721	0.1383	0.3637	5.1518	0.3689
		100	-2.2369	0.0161	0.0684	0.2227	5.0720	0.2230
3	0.5	30	-1.0029	0.8016	0.4324	0.8770	1.4381	1.5195
		50	-0.9966	0.8482	0.2343	0.4902	1.2275	1.2096
		100	-0.9982	0.8713	0.1096	0.2374	1.1060	0.9965
3	0.9	30	-2.2389	0.4999	0.4123	1.1643	5.4251	1.4142
		50	-2.2527	0.6207	0.2444	0.7463	5.3189	1.1315
		100	-2.2424	0.7515	0.1995	0.3820	5.1480	0.9468
			Panel B: N = 5					
1	0.5	30	-0.9876	0.0131	0.1713	0.2748	1.1467	0.2750
		50	-1.0052	0.0166	0.1030	0.1650	1.1135	0.1653
		100	-1.0065	0.0229	0.0543	0.0823	1.0674	0.0828
1	0.9	30	-2.2589	-0.2095	0.2557	0.5540	5.3582	0.5979
		50	-2.2357	-0.0812	0.1361	0.4095	5.1346	0.4161
		100	-2.2379	-0.0501	0.0749	0.1953	5.0832	0.1978
3	0.5	30	-0.9715	0.3680	0.4332	0.7208	1.3770	0.8562
		50	-1.0225	0.3303	0.2312	0.4015	1.2766	0.5106
		100	-1.0177	0.3513	0.1043	0.1835	1.1400	0.3069
3	0.9	30	-2.2260	0.1230	0.3907	0.9173	5.3458	0.9324
		50	-2.2359	0.1811	0.2381	0.6141	5.2375	0.6469
		100	-2.2449	0.2349	0.1185	0.3429	5.1580	0.3981
			Panel C: N = 25					
1	0.5	30	-1.0087	-0.0445	0.1717	0.2671	1.1891	0.2691
		50	-1.0225	0.0037	0.1086	0.1592	1.1540	0.1593
		100	-0.9960	0.0127	0.0500	0.0785	1.0420	0.0787
1	0.9	30	-2.2353	-0.1845	0.2545	0.6258	5.2509	0.6598
		50	-2.2302	-0.0969	0.1392	0.4075	5.1129	0.4169
		100	-2.2544	-0.0433	0.0749	0.2070	5.1572	0.2089
3	0.5	30	-0.9725	0.0969	0.4306	0.6313	1.3763	0.6407
		50	-1.0134	0.0283	0.2248	0.3352	1.2519	0.3360
		100	-0.9946	0.0717	0.1172	0.1768	1.1065	0.1820
3	0.9	30	-2.2341	-0.1492	0.4196	0.9235	5.4106	0.9457
		50	-2.2545	-0.0456	0.2419	0.6059	5.3246	0.6079
		100	-2.2386	0.0053	0.1234	0.2884	5.1349	0.2885

The table shows the bias, variance and the mean-square error of the regression slope estimates $\widehat{\beta}_1^{mCGA}$. The data generating process is specified as $Y_{ti} = 5 + 5X_t + \epsilon_{ti}$, where $i = 1, ..., N$ is the number of dependent variables and $t = 1, ..., T$ the sample size. A given configuration is described by four parameters, namely the standard deviation of regression error σ_ϵ, the standard deviation of measurement error σ_δ, the sample size T, and the number of dependent variables N.

Table 5: Simulation results for β_0

σ_ϵ	σ_δ	T	Bias $\widehat{\beta}_0^{LS}$	Bias $\widehat{\beta}_0^{mCGA}$	$Var(\widehat{\beta}_0^{LS})$	$Var(\widehat{\beta}_0^{mCGA})$	MSE $\widehat{\beta}_0^{LS}$	MSE $\widehat{\beta}_0^{mCGA}$
			Panel A: Number of dependent variables, N = 2					
1	0.5	30	-0.0100	-0.0144	0.2129	0.2717	0.2130	0.2719
		50	-0.0018	0.0019	0.1215	0.1580	0.1215	0.1580
		100	-0.0053	-0.0068	0.0626	0.0763	0.0626	0.0764
1	0.9	30	-0.0001	-0.0222	0.4900	0.8026	0.4900	0.8031
		50	0.0290	0.0101	0.2489	0.4336	0.2498	0.4337
		100	0.0028	-0.0146	0.1230	0.2121	0.1230	0.2123
3	0.5	30	-0.0097	-0.0178	0.4880	0.6680	0.4881	0.6683
		50	0.0186	0.0071	0.3179	0.4276	0.3182	0.4276
		100	-0.0055	-0.0043	0.1580	0.2005	0.1581	0.2005
3	0.9	30	0.0324	0.0294	0.7209	1.1841	0.7220	1.1850
		50	-0.0023	-0.0444	0.4195	0.7585	0.4195	0.7604
		100	-0.0489	-0.0465	0.2259	0.3787	0.2283	0.3809
			Panel B: N = 5					
1	0.5	30	-0.0195	-0.0140	0.1971	0.2486	0.1974	0.2488
		50	-0.0131	-0.0066	0.1167	0.1461	0.1169	0.1461
		100	-0.0018	-0.0033	0.0603	0.0744	0.0603	0.0744
1	0.9	30	-0.0036	-0.0020	0.4200	0.6908	0.4200	0.6908
		50	-0.0091	0.0023	0.2602	0.4494	0.2602	0.4494
		100	-0.0051	-0.0147	0.1177	0.2136	0.1177	0.2138
3	0.5	30	-0.0296	-0.0187	0.4699	0.5622	0.4708	0.5626
		50	0.0066	0.0048	0.2818	0.3353	0.2819	0.3353
		100	-0.0072	-0.0079	0.1346	0.1665	0.1346	0.1666
3	0.9	30	0.0044	-0.0058	0.6706	1.0244	0.6706	1.0244
		50	0.0244	0.0015	0.4208	0.6434	0.4214	0.6434
		100	-0.0009	-0.0036	0.1982	0.3197	0.1982	0.3197
			Panel C: N = 25					
1	0.5	30	-0.0032	0.0072	0.2222	0.2547	0.2222	0.2547
		50	0.0125	0.0119	0.1198	0.1476	0.1199	0.1477
		100	0.0083	0.0121	0.0601	0.0730	0.0602	0.0731
1	0.9	30	0.0071	-0.0080	0.4104	0.6660	0.4104	0.6661
		50	-0.0308	-0.0144	0.2465	0.3983	0.2475	0.3985
		100	0.0224	0.0263	0.1279	0.2120	0.1284	0.2127
3	0.5	30	-0.0154	-0.0117	0.5158	0.5778	0.5161	0.5779
		50	-0.0126	-0.0108	0.2707	0.3081	0.2709	0.3082
		100	0.0085	0.0117	0.1469	0.1607	0.1470	0.1609
3	0.9	30	0.0110	0.0325	0.6817	0.9573	0.6818	0.9583
		50	0.0156	0.0074	0.4292	0.6543	0.4294	0.6544
		100	-0.0040	-0.0017	0.2054	0.3125	0.2054	0.3125

The table shows the bias, variance and the mean-square error of the regression intercept estimates $\widehat{\beta}_0^{mCGA}$. The data generating process is specified as $Y_{ti} = 5 + 5X_t + \epsilon_{ti}$, where $i = 1, ..., N$ is the number of dependent variables and $t = 1, ..., T$ the sample size. A given configuration is described by four parameters, namely the standard deviation of regression error σ_ϵ, the standard deviation of measurement error σ_δ, the sample size T, and the number of dependent variables N.

Table 6: Artificial multivariate errors-in-variables model dataset

t	Y_1	Y_2	$X^\star$	δ	X	W_1	W_2
1	7.882	7.940	-0.560	-0.843	-1.404	0.253	1.026
2	2.539	2.518	-0.230	0.419	0.189	-0.029	-0.285
3	6.229	6.554	1.559	0.077	1.635	-0.043	-1.221
4	12.755	12.140	0.071	-0.569	-0.499	1.369	0.181
5	2.872	3.752	0.129	0.627	0.756	-0.226	-0.139
6	21.141	22.631	1.715	0.213	1.928	1.516	0.006
7	0.702	1.939	0.461	-0.148	0.313	-1.549	0.385
8	-1.923	-0.214	-1.265	0.448	-0.817	0.585	-0.371
9	5.027	4.984	-0.687	0.439	-0.248	0.124	0.644
10	3.668	0.696	-0.446	0.411	-0.035	0.216	-0.220
11	14.102	15.809	1.224	0.344	1.568	0.380	0.332
12	10.380	8.311	0.360	0.277	0.637	-0.502	1.097
13	5.896	8.254	0.401	-0.031	0.370	-0.333	0.435
14	-1.225	0.740	0.111	-0.153	-0.042	-1.019	-0.326
15	3.125	1.162	-0.556	-0.190	-0.746	-1.072	1.149
16	20.721	21.122	1.787	-0.347	1.440	0.304	0.994
17	12.578	12.210	0.498	-0.104	0.394	0.448	0.548
18	-4.015	-4.947	-1.967	-0.633	-2.599	0.053	0.239
19	9.129	8.464	0.701	1.084	1.786	0.922	-0.628
20	18.666	18.088	-0.473	0.604	0.131	2.050	1.361
21	-5.678	-6.326	-1.068	-0.562	-1.629	-0.491	-0.600
22	2.353	1.839	-0.218	-0.201	-0.419	-2.309	2.187
23	12.071	13.250	-1.026	-0.233	-1.259	1.006	1.533
24	-3.625	-1.269	-0.729	0.390	-0.339	-0.709	-0.236
25	-4.853	-7.984	-0.625	-0.042	-0.667	-0.688	-1.026

The table shows the dataset used in the empirical example. Y_1 and Y_2 are the response variables. $X^\star$ refers to the true regressor, with no measurement error. X is the error-prone regressor, defined as $X = X^\star + \delta$. W_1 and W_2 are additional regressors, generated with no errors-in-variables.

Table 7: Estimation results

	Model 1: OLS	Model 2: EIV	Model 3: mCGA
β	5.048	3.931	4.662
	(0.174)	(0.418)	(0.208)
γ_1	4.86	4.592	5.489
	(0.169)	(0.481)	(0.215)
γ_2	4.776	5.199	5.118
	(0.197)	(0.561)	(0.255)
α	4.789	4.434	4.436
	(0.169)	(0.476)	(0.217)
Observations	25	25	25
Adjusted R^2	0.989	0.918	0.982

The table shows the estimation results for the regression model $Y_{ti} = \alpha + \beta X_t^{\star} + \gamma_1 W_{t1} + \gamma_2 W_{t2} + \epsilon_{ti}$ using the artificial dataset for the first response variable Y_1. The population parameters are all equal to 5. Model 1 refers to the initial case where the true observations on $X^{\star}$ are assumed to be available. Model 2 considers the classical errors-in-variables case where the values $X^{\star}$ are observed with error as $X = X^{\star} + \delta$. Model 3 shows the output obtained using the mCGA method.