AStA Adv Stat Anal (2008) 92: 391–404
DOI 10.1007/s10182-008-0086-x

ORIGINAL PAPER

SIMEX estimation in case of correlated measurement errors

Gerd Ronning · Martin Rosemann

Received: 24 January 2008 / Accepted: 18 September 2008 / Published online: 21 October 2008

Abstract The simulation-extrapolation (SIMEX) approach of Cook and Stefanski (J. Am. Stat. Assoc. 89:1314–1328, 1994) has proved to be successful in obtaining reliable estimates if variables are measured with (additive) errors. In particular for nonlinear models, this approach has advantages compared to other procedures such as the instrumental variable approach if only variables measured with error are available. However, it has always been assumed that measurement errors for the dependent variable are not correlated with those related to the explanatory variables although such scenario is quite likely. In such a case the (standard) SIMEX suffers from misspecification even for the simple linear regression model. Our paper reports first results from a generalized SIMEX (GSIMEX) approach which takes account of this correlation. We also demonstrate in our simulation study that neglect of the correlation will lead to estimates which may be worse than those from the naive estimator which completely disregards measurement errors.

Keywords Anonymization · Differential measurement error · Noise addition · Statistical disclosure

Results in this paper are related to the project "Wirtschaftsstatistische Paneldaten und faktische Anonymisierung," which is financed by Bundesministerium für Bildung und Forschung. We thank Olaf Hübler and Sandra Nolte for hints on relevant publications and Helmut Küchenhoff and Hans Schneeweiß for helpful comments on an earlier version. We also thank the two anonymous referees for their critical remarks.

G. Ronning (✉)
Wirtschaftswissenschaftliche Fakultät, Universität Tübingen, Mohlstrasse 36, 72074 Tübingen, Germany
e-mail: gerd.ronning@uni-tuebingen.de

M. Rosemann
Institute for Applied Research (IAW), Ob dem Himmelreich 1, 72074 Tübingen, Germany
e-mail: martin.rosemann@iaw.edu

1 Introduction

Measurement errors may exist in almost any empirical data set. At least this seems to be true if data are collected from individuals and firms. However, very seldom the exact characteristics of these errors are known. Therefore, the biassing effects on estimation and testing are mostly considered under the assumption of white noise added to the variable under discussion. This is in particular true for econometric models with "errors in variables" (EVM) where some or all regressors are measured with error (see, for example, Greene 2000, Chap. 9.5.2.) The discussion of the EVM included the possibility of these measurement errors being correlated.[1] However, the case that the measurement error for the dependent variable is correlated with those related to regressors is mostly disregarded although it seems not to be unlikely. For example, the dependent variable may be R&D, and the regressor variable is "research intensity" measured as R&D divided by sales. If research is measured with error, also the regressor is affected by measurement error. Another example is a share equation in demand analysis where income acts both as regressor and as defining the budget shares (see Ronning 1991).

In our research on the anonymization of micro data, we encountered such a situation under "experimental conditions" when studying the protection of these data by addition of stochastic noise. In particular, we considered multivariate mixture distributions which were applied to a set of (continuous) variables for which anonymization had to be achieved. As can be easily shown, such error distributions imply correlation of errors for different variables (see Ronning 2008). Note that in this case the correlation will be known (or communicated to the data users), whereas in usual empirical applications this parameter would have to be estimated.

Fuller (1980) provides a survey of different specifications regarding measurement errors including the case that the measurement error for the dependent variable is correlated with those related to regressors. Schaalje and Butts (1993) derive consistent estimators for this case. Results in the two papers are given under the assumption of joint normality of all measurement errors. However, the results can also be obtained from asymptotic considerations leading to so-called corrected estimators.[2]

Two alternative approaches are available: the instrumental variables (IV) approach and the SIMEX procedure. IV estimation has been advocated as a general recipe in case of measurement errors (see, for example, Greene 2000, Chap. 9.5.2). However, if measurement errors are correlated and only variables measured with error are available as instrumental variables, the postulate that instrument and measurement error should be uncorrelated (see, for example, Carroll et al. 2006, Chap. 6). cannot be satisfied. In such a case the simulation-extrapolation (SIMEX) procedure of Cook and Stefanski (1994) provides a more promising approach to estimation in case of measurement errors, in particular for *nonlinear* models. However, usually possible correlation of errors regarding the dependent variable and the regressors are not taken into account. Therefore this paper presents a generalization of the standard SIMEX

[1] See Klepper and Leamer (1984) as an early example of the discussion of EVM with emphasis on the identification of the model.

[2] See, for example, Carroll et al. (2006, Sects. 3.3.2 and 3.4.1), who also consider the case of linear models.

for this situation; for illustrative purpose, we restrict ourselves to estimation of the linear model.

The paper is organized as follows: In Sect. 2 the linear model with additive measurement errors is shortly described. Section 3 first sketches the standard SIMEX approach and then introduces the generalized SIMEX for correlated errors related to the dependent variable and the regressor(s). Section 4 contains the simulation results. We also discuss the effect of neglecting the error correlation in the SIMEX approach. In Sect. 5 some concluding remarks are added.

2 The model

We consider the following simple linear regression model:

$$y_i^* = \alpha + \beta x_i^* + \eta_i, \quad i = 1, \ldots, n, \tag{2.1}$$

with $E[\eta_i] = 0$ and $\text{var}[\eta_i] = \sigma_\eta^2$. Both x_i^* and y_i^* are measured with additive error:

$$x_i = x_i^* + u_i \tag{2.2}$$

and

$$y_i = y_i^* + v_i, \tag{2.3}$$

where the errors u_i and v_j satisfy, for all i and j,

$$\begin{aligned}
&E[u_i] = 0, \qquad \text{var}[u_i] = \sigma_u^2,\\
&E[v_i] = 0, \qquad \text{var}[v_i] = \sigma_v^2,\\
&\text{cov}[\eta_i, u_j] = \text{cov}[\eta_i, v_j] = 0,\\
&\text{cov}\big[x_i^*, u_j\big] = \text{cov}\big[x_i^*, v_j\big] = 0,\\
&\text{cov}\big[y_i^*, u_j\big] = \text{cov}\big[y_i^*, v_j\big] = 0,\\
&\text{cov}[u_i, u_j] = \text{cov}[v_i, v_j] = 0 \quad \text{for } i \neq j,
\end{aligned} \tag{2.4}$$

and in particular

$$\text{cov}[u_i, v_j] = \begin{cases} \sigma_{uv} & \text{if } i = j,\\ 0 & \text{if } i \neq j, \end{cases} \tag{2.5}$$

where σ_{uv} may be nonzero. Then, using the mismeasured variables x and y, the "naive" least squares estimator of β tends towards

$$\text{p-lim}\,\hat{\beta} = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}\beta + \frac{\sigma_{uv}}{\sigma_x^2 + \sigma_u^2}, \tag{2.6}$$

where σ_x^2 denotes the variance of x^*. Usually assuming that there is no correlation between u_i and v_i, the result is given without the second right-hand term, which may be seen as the bias from omitted variables represented by v_i. Note that this second term may offset the attenuation bias towards zero implied by the first term if β and σ_{uv}

have the same sign. This should be remembered later on when the simulation results in Sect. 4 are discussed. Note also that correct measurement of x implies consistent estimation of β even when y^* is measured with error since then $\sigma_u^2 = 0$ and $\sigma_{uv} = 0$ hold.

We also note in passing that the naive estimator of the parameter α satisfies

$$\text{p-lim}\,\hat{\alpha} = \text{p-lim}\big(\overline{y} - \hat{\beta}\overline{x}\big) = \alpha + \mu_x\big(\beta - \text{p-lim}\big(\hat{\beta}\big)\big), \tag{2.7}$$

where μ_x denotes the expected value of x^*, and p-lim$(\hat{\beta})$ is given by (2.6). Therefore, α will be estimated consistently if $\hat{\beta}$ is a consistent estimator. In the following, we therefore confine ourselves to the estimation of the parameter β.

Carroll et al. (2006, Sects. 2.5 and 3.2.4) consider the situation that only the regressor is measured with error and that the error u_i is correlated with the residual η_i. They call this an example of "differentiated measurement errors" since the error u_i contains information about the dependent variable y_i through the correlation. Note that this is observationally equivalent to the model given above in (2.1) to (2.4) since the error v_i affecting y_i may be seen as part of the equation error, that is, $y_i = y_i^*$ and $\eta_i = -v_i$. The assumption cov$[u_i, \eta_i] \neq 0$ then replaces the above specification that σ_{uv} from (2.5) may be different from zero. Therefore, our result (2.6) is also presented in Carroll et al. (2006, p. 47). However, joint normality is assumed there for x_i, u_i, and η_i, so that the result holds even for finite samples.

3 Generalized SIMEX approach

3.1 The standard approach

If only x_i^* is not correctly measured, the SIMEX approach will increase artificially the error by using $\lambda_j > 1$, $j = 1, \ldots, m$, implying that the least squares estimator $\hat{\beta}(\lambda_j)$ based on this generated regressor variable will tend towards

$$\text{p-lim}\,\hat{\beta}(\lambda_j) = \frac{\sigma_x^2}{\sigma_x^2 + \lambda_j \sigma_u^2}\beta. \tag{3.1}$$

The idea then is to estimate the model for each λ_j, fit the resulting estimates to some smooth function, and extrapolate this function to the point $\lambda = 0$.[3] Note that we use a notation which follows a suggestion by Helmut Küchenhoff and which is different from the one used in Cook and Stefanski (1994): In our notation $\lambda = 1$ denotes the case that x_i^* is mismeasured with error u_i as specified in (2.2). Augmentation of the error variance then is done by values of $\lambda > 1$ indicating relative changes of the error variance. For example, $\lambda = 1.50$ means that the variance of the measurement error is increased by 50 per cent.

[3] For this approach, see Carroll et al. (2006, Chap. 5).

However, in applying this estimation procedure the artificial increase of the variance of the measurement error has to be defined in terms of the mismeasured (observable) variable x_i. We therefore *define* the pseudo-regressor $x_i(\lambda_j)$ as

$$x_i(\lambda_j) = x_i + \sqrt{\lambda_j - 1}\, u_{i1}, \quad \lambda_j \geq 1,$$

or

$$x_i(\lambda_j) = x_i^* + u_{i0} + \sqrt{\lambda_j - 1}\, u_{i1},$$

i.e., $x_i = x_i^* + u_{i0}$, where u_{i0} is the measurement error, and u_{i1} is the SIMEX simulation error both being stochastically independent with $E[u_{i0}] = E[u_{i1}] = 0$ and $\text{var}[u_{i0}] = \text{var}[u_{i1}] = \sigma_u^2$. Therefore in the above example with $\lambda = 1.50$ we will use the factor $\sqrt{1.50 - 1} = 0.7071$.

For the variance of the artificial regressor, this implies

$$\text{var}\big[x_i(\lambda_j)\big] = \sigma_x^2 + \sigma_u^2 + (\lambda_j - 1)\sigma_u^2 = \sigma_x^2 + \lambda_j \sigma_u^2.$$

Therefore the variance tends to σ_x^2 (and the error variance $\text{var}[u_0 + u_1]$ tends to 0) as $\lambda \to 0$, whereas in the specification of Cook and Stefanski (1994) the parameter λ tends towards -1.

3.2 An operational GSIMEX procedure

We now assume that the two measurement errors u_i and v_i are correlated so that $\sigma_{uv} \neq 0$. We first thought of a *two-dimensional* generalization of the SIMEX procedure: Increase the error u_i by using the factors $\sqrt{\lambda_j - 1}, \lambda_j > 1$, and increase the error v_i by using the factors $\sqrt{\mu_k - 1}$, $\mu_k > 1$, $k = 1, \ldots, p$, holding the correlation between the measurement errors u and v fixed. More exactly, the correlation

$$\text{corr}[u_{i0}, v_{i0}] = \frac{\sigma_{uv}}{\sigma_u \sigma_v} \equiv \varrho_{uv}$$

should also hold in the simulation stage.

Formally, this procedure can be described as follows: Let v_{i0} be the measurement error and v_{i1} the SIMEX simulation error regarding y_i^* with $E[v_{i0}] = E[v_{i1}] = 0$ and $\text{var}[v_{i0}] = \text{var}[v_{i1}] = \sigma_v^2$. Then the covariance of the errors regarding x_i^* and y_i^* is given by

$$\text{cov}\big[u_{i0} + \sqrt{\lambda_j - 1}u_{i1}, v_{i0} + \sqrt{\mu_k - 1}v_{i1}\big] = \Big(1 + \sqrt{(\lambda_j - 1)(\mu_k - 1)}\Big)\sigma_{uv},$$

leading to

$$\text{corr}\big[u_{i0} + \sqrt{\lambda_j - 1}u_{i1}, v_{i0} + \sqrt{\mu_k - 1}v_{i1}\big] = \frac{(1 + \sqrt{(\lambda_j - 1)(\mu_k - 1)})\sigma_{uv}}{\sqrt{\lambda_j \mu_k}\sigma_u \sigma_v} < \varrho_{uv},$$

so that correlation is lower than it should be and varies with λ and μ. Here we have used the fact that for $\lambda > 0$, $\mu > 0$, $\lambda \neq \mu$, always

$$\frac{(1 + \sqrt{(\lambda - 1)(\mu - 1)})}{\sqrt{\lambda\mu}} < 1$$

holds.[4] For example, for $\lambda = 1.00$ and $\mu = 3.00$, we obtain

$$\frac{1}{\sqrt{3.00}}\varrho_{uv} = 0.5774\varrho_{uv},$$

so that correlation is reduced to almost half of the true correlation. Of course, we would obtain a GSIMEX estimate for β by fitting the three-dimensional points $(\lambda, \mu, \hat{\beta}(\lambda_j, \mu_k))$ and then extrapolate the fitted plane for $\lambda \to 0, \mu \to 0$. It is still unclear whether this approach with varying correlation is superior to the following, computationally much simpler, approach which we actually used: If we restrict the analysis to the case $\lambda = \mu$, we obtain from the formula above

$$\operatorname{cov}\left[u_{i0} + \sqrt{\lambda_j - 1}u_{i1}, v_{i0} + \sqrt{\lambda_j - 1}v_{i1}\right] = \left(1 + \sqrt{(\lambda_j - 1)(\lambda_j - 1)}\right)\sigma_{uv} = \lambda_j\sigma_{uv},$$

and

$$\operatorname{corr}\left[u_{i0} + \sqrt{\lambda_j - 1}u_{i1}, v_{i0} + \sqrt{\lambda_j - 1}v_{i1}\right] = \frac{\lambda_j\sigma_{uv}}{\lambda_j\sigma_u\sigma_v} = \varrho_{uv}.$$

Our "operational" GSIMEX approach therefore in the simulation phase considers only estimates for $\lambda = \mu$ and then extrapolates this (one-dimensional) function to the point $\lambda = 0$.[5] In other words, the estimate is based on

$$\text{p-lim}\,\hat{\beta}(\lambda_j) = \frac{\sigma_x^2}{\sigma_x^2 + \lambda_j\sigma_u^2}\beta + \frac{\lambda_j\sigma_{uv}}{\sigma_x^2 + \lambda_j\sigma_u^2}. \tag{3.2}$$

Note that this is *not* equivalent to the standard SIMEX procedure since we consider correlated errors in the simulation step. This can also be seen from comparing (3.1) for the "standard approach" with (3.2) above.

We would like to add some comments on modifications if the linear model has more than one regressor and all L regressors (and the dependent variable) are observed with error. For the naive estimator of the coefficient vector $\boldsymbol{\beta}$, we obtain (see, for example, Ronning 2008, Chap. 9)

$$\text{p-lim}\,\hat{\boldsymbol{\beta}} = \left(\operatorname{cov}[\mathbf{x}] + \operatorname{cov}[\mathbf{u}]\right)^{-1}\left(\operatorname{cov}[\mathbf{x}]\boldsymbol{\beta} + \operatorname{cov}[\mathbf{u}_x, v_y]\right), \tag{3.3}$$

where $\operatorname{cov}[\mathbf{x}]$ is the $(L \times L)$ covariance matrix of the L regressor variables, $\operatorname{cov}[\mathbf{u}]$ the corresponding covariance matrix of the errors related to the regressors, and

[4] The inequality above is equivalent to the following set of inequalities:

$$\begin{aligned}
\sqrt{(\lambda - 1)(\mu - 1)} &< \sqrt{\lambda\mu} - 1,\\
(\lambda - 1)(\mu - 1) &< \lambda\mu - 2\sqrt{\lambda\mu} + 1,\\
0 &< \lambda + \mu - 2\sqrt{\lambda\mu},\\
0 &< (\sqrt{\lambda} - \sqrt{\mu})^2.
\end{aligned}$$

[5] This procedure was suggested to us also by Helmut Küchenhoff.

$\mathrm{cov}[\mathbf{u}_x, v_y]$ is an L-dimensional vector containing the covariances $\mathrm{cov}[u_{i\ell}, v_i]$. Applying the GSIMEX procedure to all regressors with *identical* λ_j will result in the following bias function:

$$\text{p-lim}\,\hat{\boldsymbol{\beta}}(\lambda_j) = \big(\mathrm{cov}[\mathbf{x}] + \lambda\,\mathrm{cov}[\mathbf{u}]\big)^{-1}\big(\mathrm{cov}[\mathbf{x}]\boldsymbol{\beta} + \lambda_j\,\mathrm{cov}[\mathbf{u}_x, u_y]\big). \tag{3.4}$$

As in the case of simple regression, we would obtain estimates for different λ_j, $\lambda_j \geq 1$, then fit the points $(\lambda_j, \hat{\beta}_\ell(\lambda_j))$ for each ℓ separately to a (linear or quadratic) function and finally obtain a GSIMEX estimate of β_ℓ from the extrapolated function at $\lambda = 0$. It is evident that for the case of many regressors, the argument raised above with regard to setting $\lambda = \mu$ has even more appeal.

4 Simulation results

4.1 Results for the GSIMEX

In this section we will present simulation results concerning the estimation of the linear regression model (2.1) with specifications of the measurement errors u and v given in (2.2), (2.3), (2.4), and (2.5). As mentioned above in Sect. 2, we confine ourselves to the estimation of the parameter β.[6] Since the "naive" estimator using the observable variables will be biased in case of measurement errors, we will report this expression as well which will offer an indication whether the GSIMEX estimate performs better (i.e., is less biased). Following a suggestion of a referee, we also report estimation results in case of no measurement errors so that we get an impression of the magnitude of the bias from measurement errors.

The details of the simulation design are given in Table 1. Note that we vary the parameter setting both with regard to the correlation between u_i and v_i and to the regressor variance σ_x^2, where the errors u_i and v_i are assumed to be jointly normal. However, we restrict ourselves to two alternative values for both σ_x^2 and ϱ_{uv}. In particular, the correlation of $+0.9$ and -0.9 is chosen in order to show most clearly the effect of correlation on parameter estimation. Since the error variances σ_u^2 and σ_v^2 are fixed, a larger regressor variance implies a larger signal-to-noise ratio σ_x^2/σ_u^2 or, equivalently, smaller measurement errors for the regressor variable. The same remark applies to the dependent variable since $\mathrm{var}[y^*] = \beta^2\sigma_x^2 + \sigma_\eta^2$ and therefore the ratio $\mathrm{var}[y^*]/\sigma_v^2$ increases or, equivalently, measurement errors for the dependent variable become smaller if σ_x^2 rises. Figure 1 displays the bias functions from (3.2) for the six scenarios of our simulation study. The dotted vertical line separates the "simulation region" $\{\lambda \mid \lambda \geq 1\}$ from the "extrapolation region" $\{\lambda \mid \lambda < 1\}$. Note the offsetting effect of positive ϱ_{uv} which reduces the attenuation effect to a large extent. It is evident from this figure that a *linear* extrapolation will not work satisfactorily for "larger" measurement errors and in particular for $\varrho_{uv} < 0$.

[6] Since in our simulation study the regressor variable is normally distributed with $\mu_x = 0.00$ (see Table 1), the estimation of α in this special case is unaffected by the inconsistency of $\hat{\beta}$. See (2.7).

Table 1 Simulation design for GSIMEX

Number of simulation runs (wh_{sim})	50
Number of observations (n)	500
Number of simulations in SIMEX (wh_{ex})	30
Regression parameter α	-0.50
Regression parameter β	1.00
Equation error η	$\eta \sim N(0, \sigma_\eta^2)$
Regression parameter σ_η^2	0.25
Regressor variable x	$x \sim N(0, \sigma_x^2)$
Variance of regressor: σ_x^2	$\{1.00, 4.00\}$
Variance of u: σ_u^2	1.00
Variance of v: σ_v^2	1.00
Correlation of u and v: ϱ_{uv}	$\{-0.90, 0.00, +0.90\}$
Extrapolation vector $\boldsymbol{\lambda}$	1.00, 1.10, 1.25, 1.50, 2.00, 2.50, 3.00

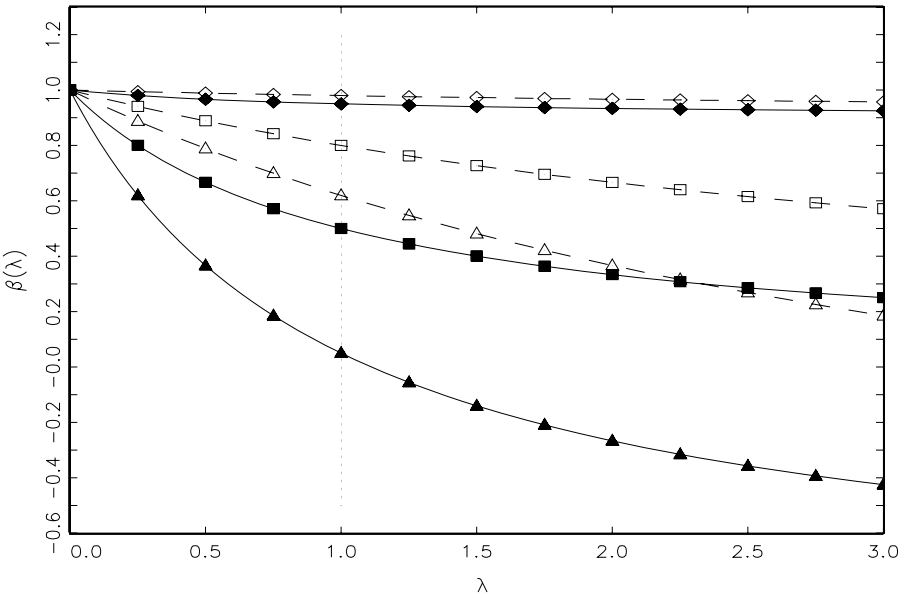

Fig. 1 Bias functions for β with $\sigma_u^2 = \sigma_v^2 = 1.00$. *Solid curves* and *solid symbols* show the bias functions for "large" measurement errors ($\sigma_x^2 = 1.00$); *dashed curves* and *non-solid* show those for "small" measurement error ($\sigma_x^2 = 4.00$). Each curve refers to a certain correlation: $\diamond/\blacklozenge \rightarrow \varrho_{uv} = 0.9$, $\square/\blacksquare \rightarrow \varrho_{uv} = 0.0$, $\triangle/\blacktriangle \rightarrow \varrho_{uv} = -0.9$

The simulation results are given in Table 2.[7] The parts A to D of the table present results for GSIMEX estimation, and part E at the bottom contains estimation results based on error-free data which show the expected results for the classical regression

[7] The lines headed by "corr. negl." will be discussed in Sect. 4.2.

Table 2 GSIMEX estimation of the linear model

Linear extrapolation						
ϱ_{uv}		Estimate	std. dev.	min.	Median	max.
A $\sigma_u^2 = \sigma_v^2 = 1.00$, $\sigma_x^2/\sigma_u^2 = 1.00$ (Large measurement errors)						
0.0	GSIMEX	0.5928	0.0502	0.4657	0.5890	0.7043
	naive	0.4901	0.0420	0.3867	0.4874	0.5759
0.9	GSIMEX	0.9559	0.0247	0.8970	0.9579	1.0080
	naive	0.9462	0.0204	0.8972	0.9478	0.9894
	corr. negl.	1.1504	0.0292	1.0430	1.1529	1.2002
−0.9	GSIMEX	0.2423	0.0494	0.1291	0.2420	0.3607
	naive	0.0449	0.0407	−0.0438	0.0435	0.1431
	corr. negl.	0.0561	0.0563	−0.0855	0.0549	0.1746
B $\sigma_u^2 = \sigma_v^2 = 1.00$, $\sigma_x^2/\sigma_u^2 = 4.00$ (Small measurement errors)						
0.0	GSIMEX	0.9045	0.0289	0.8341	0.9015	0.9740
	naive	0.7976	0.0263	0.7309	0.7968	0.8582
0.9	GSIMEX	0.9906	0.0136	0.9601	0.9893	1.0222
	naive	0.9803	0.0117	0.9541	0.9792	1.0080
	corr. negl.	1.1147	0.0200	1.0695	1.1118	1.1686
−0.9	GSIMEX	0.8184	0.0421	0.6970	0.8300	0.8916
	naive	0.6164	0.0375	0.5114	0.6227	0.6733
	corr. negl.	0.6999	0.0393	0.6137	0.7032	0.7906

model. For each scenario, we compute the "estimate" as the average of estimates from all simulation runs. We also use these values for computing standard deviation, minimum, median, and maximum.

We start by considering the case of no correlation, i.e., $\varrho_{uv} = 0$. This case could be handled by the "standard" SIMEX. However, we will use the GSIMEX with vector $\boldsymbol{\lambda}$ which will give identical results in this case. From the table we learn that the linear extrapolation function gives satisfactory results at least for smaller measurement errors, i.e., larger variance σ_x^2. The use of quadratic approximation improves the results. For larger measurement errors both approximations (linear and quadratic) do not work satisfactorily, although the quadratic extrapolation does a slightly better job than the linear extrapolation. However, for all four cases (linear versus quadratic extrapolation, $\sigma_x^2 = 1.00$ versus $\sigma_x^2 = 4.00$), the GSIMEX is less biased than the naive estimate.[8]

We now consider the results for $\varrho_{uv} \neq 0$. The most remarkable result is the much better performance in case of positive correlation. This is due to the fact that β and ϱ_{uv} have the same sign. However, this should not be overrated since the naive estimate also shows less bias in this case! From Fig. 1 it becomes clear that for positive correlation, even for larger measurement errors, the bias is negligible, and the bias

[8] Since the naive estimate is not based on an extrapolation function, results with regard to this estimation method should be identical for parts A and C or B and D. However, they differ due to the simulation error. For example for $\varrho_{uv} = 0.0$ we obtain 0.4901 in part A and 0.4954 in part C.

Table 2 (*Continued*)

Quadratic extrapolation						
ϱ_{uv}		Estimate	std. dev.	min.	Median	max.
C $\sigma_u^2 = \sigma_v^2 = 1.00$, $\sigma_x^2/\sigma_u^2 = 1.00$ (Large measurement errors)						
0.0	GSIMEX	0.7405	0.0583	0.6370	0.7398	0.8780
	naive	0.4954	0.0375	0.4196	0.4974	0.5815
0.9	GSIMEX	0.9753	0.0326	0.8989	0.9746	1.0670
	naive	0.9504	0.0202	0.9109	0.9478	1.0130
	corr. negl.	1.4171	0.0359	1.3163	1.4216	1.5150
−0.9	GSIMEX	0.5049	0.0864	0.3068	0.5072	0.7139
	naive	0.0500	0.0518	−0.0792	0.0534	0.1670
	corr. negl.	0.0708	0.0693	−0.0687	0.0604	0.2263
D $\sigma_u^2 = \sigma_v^2 = 1.00$, $\sigma_x^2/\sigma_u^2 = 4.00$ (Small measurement errors)						
0.0	GSIMEX	0.9739	0.0298	0.9166	0.9694	1.0323
	naive	0.8042	0.0237	0.7531	0.8029	0.8568
0.9	GSIMEX	0.9963	0.0166	0.9527	0.9965	1.0432
	naive	0.9807	0.0131	0.9521	0.9815	1.0148
	corr. negl.	1.1864	0.0212	1.1514	1.1838	1.2305
−0.9	GSIMEX	0.9460	0.0548	0.7965	0.9413	1.0638
	naive	0.6247	0.0389	0.5398	0.6249	0.7151
	corr. negl.	0.7429	0.0450	0.6206	0.7539	0.8523
	σ_x^2	Estimate	std. dev.	min.	Median	max.
E Estimation results from error-free data						
	1.00	1.0002	0.0203	0.9591	0.9976	1.0489
	4.00	1.0004	0.0093	0.9799	1.0000	1.0271

Remark: All estimation results refer to true parameter $\beta = 1.00$. For the meaning of "corr. negl." see Sect. 4.2

function is almost linear so that both approximating functions (linear and quadratic) should give good results!

Some final remarks concerning the case of negative correlation are in order: Fig. 1 shows that in this case the bias increases, and the bias function has a curvature in the extrapolation region $\{\lambda \mid \lambda < 1\}$, which is quite different from the curvature in the simulation region and which makes extrapolation much harder. This is true in particular for larger measurement errors. Therefore not unexpectedly the linear extrapolation function does not much improve the GSIMEX compared to the naive estimate, whereas the *quadratic* extrapolation function removes the bias to a considerable extent leading to an almost perfect estimate in part D of the table, which considers the case of 'smaller' measurement error. For larger measurement errors, a higher-order approximation might be useful but has not been tried in this study.

Please note that the GSIMEX estimation (or rather approximation) error is smallest for $\varrho_{uv} = 0.90$ in all four parts (A to D) of the table. This is caused by the almost

linear bias function which allows a reliable approximation and extrapolation. In three of four cases the estimation error is highest for the case of negative correlation where extrapolation is more difficult as explained above.

One final comment should be added which should be seen as a caveat regarding the generality of our simulation results which, in fact, depend very much on the "true" value of the coefficient β:[9] If we would choose a much greater value as, for example, $\beta = 10$ instead of $\beta = 1$, (2.6) would give

$$\text{p-lim}\,\hat{\beta} = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} 10 + \frac{\sigma_{uv}}{\sigma_x^2 + \sigma_u^2},$$

so that the offsetting effect of the second term will be much weaker and therefore results even for high positive correlation would not result in an (almost) straight line for the SIMEX procedure. However, the statement that in case of equal sign of β and σ_{uv} the attenuation effect of the first term in (2.6) will be partly offset by the second term, will remain valid.[10]

4.2 Neglect of error correlation in the SIMEX approach

We also considered the likely situation that the correlation of errors is not recognized by the statistician. In our simulation study we therefore considered the case that measurement error is only assumed for the regressor, whereas the error for the dependent variable and its correlation with the error for the regressor are neglected, that is, $\sigma_v^2 = 0$ instead of $\sigma_v^2 > 0$ and $\sigma_{uv} = 0$ instead of $\sigma_{uv} \neq 0$ is assumed. Therefore misspecified SIMEX estimation will be based on the following bias function:

$$\text{p-lim}\,\widehat{\beta^{\text{mis}}}(\lambda) = \frac{\sigma_x^2}{\sigma_x^2 + \lambda\sigma_u^2}\beta + \frac{\sigma_{uv}}{\sigma_x^2 + \lambda\sigma_u^2}. \tag{4.1}$$

Note that this function does not imply consistency since, for $\lambda \to 0$, we obtain

$$\lim_{\lambda \to 0} \frac{\sigma_x^2}{\sigma_x^2 + \lambda\sigma_u^2}\beta + \frac{\sigma_{uv}}{\sigma_x^2 + \lambda\sigma_u^2} = \beta + \varrho_{uv}\frac{\sigma_u}{\sigma_x}\frac{\sigma_v}{\sigma_x}. \tag{4.2}$$

Therefore the sign of the bias will be determined by the sign of the correlation, and the magnitude of the bias also depends on the two error ratios σ_u/σ_x and σ_v/σ_x. Our simulation results in Table 2 in the rows headed by "corr. negl." correspond to these findings: In all four cases the SIMEX procedure overestimates the true β if $\varrho_{uv} > 0$ and underestimates the coefficient if $\varrho_{uv} < 0$. We also note that in most cases the estimate is even worse than the "naive" estimate.

However, there is no general rule for this as becomes apparent from Figs. 2 and 3 which show the "correct" bias functions already presented in Fig. 1 together with the "misspecified" bias functions from (4.1). Figure 2 describes the situation for "large"

[9] We are most grateful to one of the referees for drawing our attention to this aspect.

[10] See also Carroll et al. (2006, p. 54), who show a graphical example.

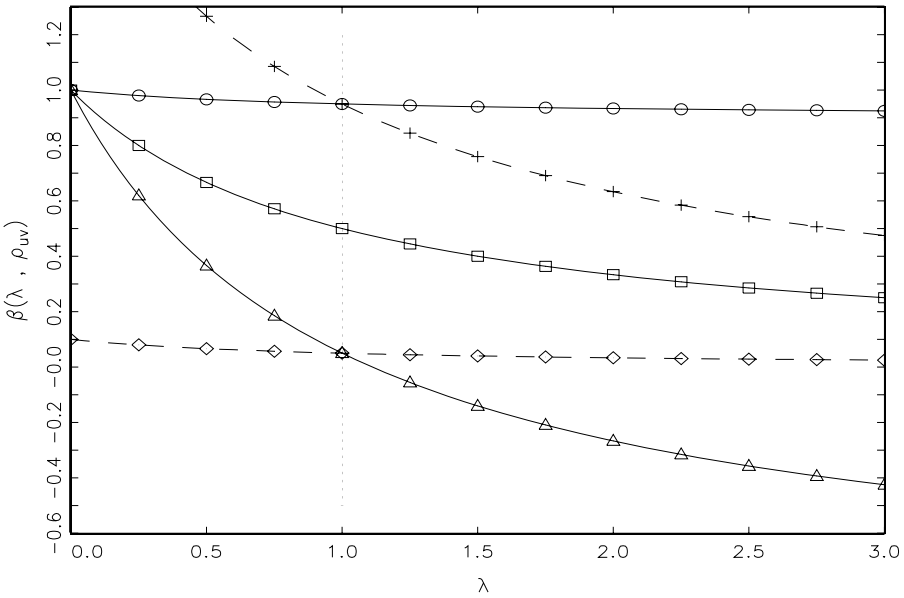

Fig. 2 Bias functions for β in case of large measurement errors ($\sigma_x^2 = 1.00$, $\sigma_u^2 = \sigma_v^2 = 1.00$). *Solid curves* show the "correct" bias functions (see (3.2)): $\bigcirc \rightarrow \varrho_{uv} = 0.90$, $\square \rightarrow \varrho_{uv} = 0.00$, $\triangle \rightarrow \varrho_{uv} = -0.90$. *Dashed curves* show the "misspecified" bias functions (see (4.1)): $+ \rightarrow \varrho_{uv} = 0.90$, $\diamondsuit \rightarrow \varrho_{uv} = -0.90$

measurement errors, and Fig. 3 has the corresponding results for "small" measurement errors. Note that we use the same scaling in both figures. Both functions for a certain ϱ_{uv} intersect at $\lambda = 1$ which also gives the "naive" estimate for this case. The "true" (misspecified) GSIMEX estimate (given by the dashed functions at $\lambda = 0$) may be more or less biased than the naive estimate: For example, in Fig. 3 with $\sigma_x^2 = 4.00$ the dashed curve for $\varrho_{uv} = 0.9$ gives $4.9/5.0 = 0.98$ for the naive estimate and $4.9/4.0 = 1.225$ for the misspecified GSIMEX, the latter being much more (and positively) biased. However, in the same figure the dashed curve for $\varrho_{uv} = -0.9$ gives $3.1/5.0 = 0.62$ for the naive estimate and $3.1/4.0 = 0.775$ for the misspecified GSIMEX, the latter being less biased in this case.

Of course, the computed (misspecified) GSIMEX will differ from the "true" (misspecified) GSIMEX due to the approximating functions used in this procedure. However, we find very similar simulation results in parts B and D of Table 2: For the case $\varrho_{uv} = 0.9$, $\sigma_x^2 = 4.00$, the naive estimate is given by 0.9803 (part B) and 0.9807 (part D) and the misspecified GSIMEX (correlation neglected) by 1.1147 (linear approximation) and 1.1864 (quadratic approximation), so that the quadratic approximation is even worse. For the case $\varrho_{uv} = -0.9$, $\sigma_x^2 = 4.00$, we find for the naive estimate 0.6164 (part B) and 0.6247 (part D) and for the misspecified GSIMEX (correlation neglected) 0.6999 in case of linear approximation and 0.7429 in case of quadratic approximation.

In case of larger measurement errors, the results given by Fig. 2 and parts A and C of Table 2 are qualitatively identical although the magnitude of bias increases. In

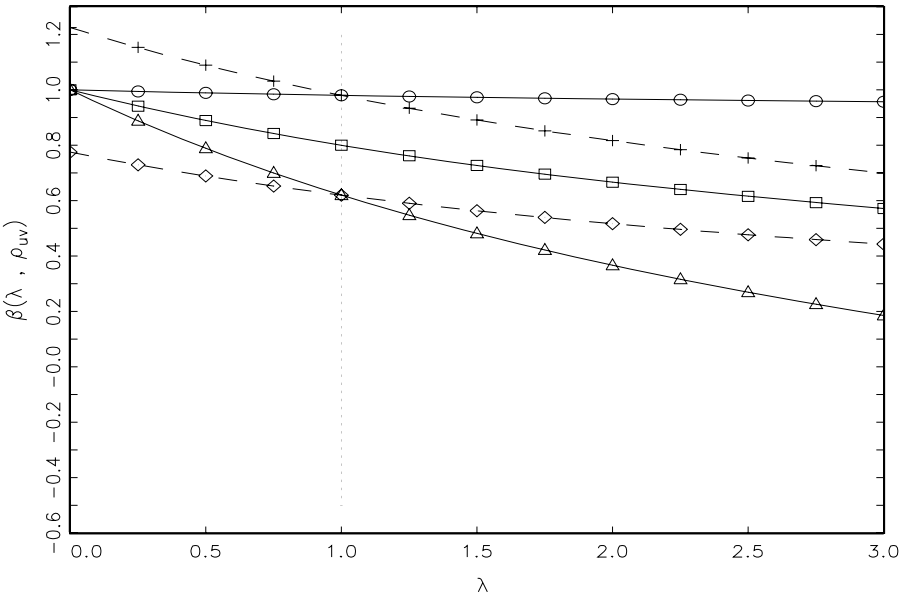

Fig. 3 Bias functions for β in case of small measurement errors ($\sigma_x^2 = 4.00$, $\sigma_u^2 = \sigma_v^2 = 1.00$). Solid curves show the "correct" bias functions (see (3.2)): $\bigcirc \rightarrow \varrho_{uv} = 0.90$, $\square \rightarrow \varrho_{uv} = 0.00$, $\triangle \rightarrow \varrho_{uv} = -0.90$. *Dashed curves* show the "misspecified" bias functions (see (4.1)): $+ \rightarrow \varrho_{uv} = 0.90$, $\diamondsuit \rightarrow \varrho_{uv} = -0.90$

case of negative correlation the misspecified bias function is almost linear and rather flat so that both the naive estimate and the misspecified SIMEX will lead to severely biased results with no clear order regarding their magnitudes for both approximating functions. For positive correlation, the overestimation of β is much more pronounced than for the case of small measurement errors discussed above. Of course, the effect of the magnitude of the "true" value of β on the simulation results discussed in Sect. 4.1 applies also here as becomes evident from (4.1).

5 Concluding remarks

We have considered parameter estimation where it is assumed that measurement errors for the dependent variable and for a single regressor are correlated. This scenario so far has not received much attention in the literature although it seems not unlikely. We propose a generalized SIMEX approach which takes account of the correlation in the simulation phase. We choose a procedure which holds correlation constant in the simulation phase which is computational simpler than an alternative procedure also discussed in the paper. It has the advantage that for the case of more than one explanatory variable, the measurement errors for all explanatory variables can be treated symmetrically. For illustrative purpose, we apply the approach to the linear model although its real potential lies in the estimation of nonlinear models.

Our simulation results in Sect. 4 show that this approach works more satisfactorily only if the coefficient β and the correlation coefficient have the same sign. The effect

of the magnitude of β should be noted. In case of opposite sign and large measurement errors the proposed GSIMEX procedure will fail. The important finding of the paper is that neglect of the error correlation when applying the standard SIMEX approach will give misleading results that in many cases may be worse than those from the naive estimator, which disregards the existence of measurement errors at all.

The paper has not discussed how the error structure could be estimated or tested. Of course, estimation will be possible only for the case of repeated measurement. We have also not provided any empirical data on the possible magnitude and sign of error correlation.[11] So far this has not been discussed in the literature. We also would like to extend our analysis to the case of *multiplicative* errors. All this will be a topic of further research.

References

Carroll, R.J., Ruppert, D., Stefanski, L.A., Crainiceanu, C.M.: Measurement Error in Nonlinear Models. A Modern Perspective, 2nd edn. Chapman and Hall, London (2006)

Cook, J.R., Stefanski, L.A.: Simulation-Extrapolation Estimation in Parametric Measurement Error Models. J. Am. Stat. Assoc. **89**, 1314–1328 (1994)

Fuller, W.A.: Properties of some estimators for the errors-in-variables model. Ann. Stat. **8**, 407–422 (1980)

Greene, W.H.: Econometric Analysis, 4th edn. Prentice Hall, Upper Saddle River (2000)

Klepper, S., Leamer, E.E.: Consistent sets of estimates for regressions with errors in variables. Econometrica **52**, 163–184 (1984)

Ronning, G.: Probleme bei der Schätzung fehlerbehafteter Anteilsgleichungen. Jahrb. Nationalökonomie Stat. **204**, 69–82 (1991)

Ronning, G.: Stochastische Überlagerung mit Hilfe der Mischungsverteilung. Manuscript, University of Tübingen (April 2007, revised June 2008)

Schaalje, G.B., Butts, R.A.: Some effects of ignoring measurement errors in straight line regression and prediction. Biometrics **49**, 1262–1267 (1993)

[11] In our own work on the anonymization of micro data we have the comfortable situation that we know the error structure.