

The Multiplicative Simulation-Extrapolation Approach

Sandra Lechner*

University of Konstanz

Center for Quantitative Methods and Survey Research

This version: March 1, 2007

Abstract

We develop a new general approach for handling multiplicative measurement error in continuous covariates in linear and nonlinear regression models. We apply the Simulation-Extrapolation (SIMEX) approach, which is a simulation based method of estimating and reducing the bias due to additive measurement error, to the case of multiplicative measurement error. We do not apply a logarithmic transformation, so that the multiplicative measurement error model becomes an additive one, but we show how to modify the SIMEX approach, in order to use the multiplicative measurement error model as such. Multiplying the measurement error by additional measurement error allows us to infer in which way the estimation bias is affected by the increase of variance of the measurement error. In the extrapolation step, the estimated parameters are modelled as a function of the magnitude of the variance of the measurement error and extrapolated to the case of no measurement error. We apply our method to the case of data masking, in order to obtain parameter estimates of the true data generating process, if the data are multiplied by an additional measurement error. We produce Monte-Carlo evidence on how the reduction of data quality can be minimized by masking.

JEL classification: C21, J24, J31

Keywords: multiplicative error-in-variables, SIMEX, disclosure limitation

*Corresponding author. Department of Economics, Box D124, University of Konstanz, 78457 Konstanz, Germany. Phone ++49-7531-88-3214, Fax -4450, email: sandra.lechner@uni-konstanz.de. The idea for this paper arose from our work for the scientific advisory board "Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten" (Actual Anonymization of Individual Economic Data) of the German Federal Statistical Office. For helpful comments we like to thank Winfried Pohlmeier, Ingmar Nolte and Gerd Ronning. Financial support by the DFG is gratefully acknowledged. The usual disclaimer applies.

1 Introduction

Statistical offices collect many types of microdata which contain highly sensitive information, whose confidentiality has to be protected against disclosure. The goal of the statistical agencies is therefore not only to provide a maximum amount of information to the empirical researcher, but also to guaranty a maximum amount of confidentiality and privacy to the individual respondents. In order to solve this trade-off, the data collecting institutions use so-called data disclosure limitation techniques, which can be regarded as a data filter that transforms the true data generating process.

Most of the masking procedures are concerned with the protection of continuous variables,¹ where addition of an independent noise term to the continuous covariates is the simplest way of data masking.² In the case of noise addition this leads to the well-known error-in-variables problem, for which the effects of measurement error on the properties of linear estimators are well understood and discussed in the literature (see e.g. Fuller (1987)). The monograph of Carroll, Ruppert, and Stefanski (1995) surveys various approaches to errors-in-variables for nonlinear models, and special aspects are treated by Amemiya (1985), Hausman, Newey, and Powell (1995), Lee and Sepanski (1995), Hong and Tamer (2003), and Schennach (2004) for example. However data collecting institutions complain that this disclosure limitation technique does not protect enough the data. Additive measurement error indeed only slightly modifies the original value, especially when the original value is high. This means that the probability of reidentifying/disclosing the individual information for those observations is not minimized.

This paper is concerned with multiplicative measurement error, which is considered by the data collecting institutions as more suited to protect data against disclosure. Multiplicative measurement error has the advantage that the original observation is proportionally modified so that a single observation with a higher value, which is typically subject to a higher disclosure risk, is better masked as through additive measurement error. Furthermore, multiplicative measurement error conserves the

¹See, for example, Domingo-Ferrer and Torra (2002) and Pohlmeier, Ronning, and Wagner (2005) for such procedures and the latter also for their effects on estimation properties.

²See for example Lechner and Pohlmeier (2004), and for the case of discrete variables see Ronning, Rosenmann, and Strotmann (2005). They analyze the effect on the estimates of a probit model, when post-randomization is applied to the binary dependent variable, in order to protect it against disclosure.

structural zeros contained in the dataset.

Less attention has been given in the literature to multiplicative measurement error. Hwang (1986) derived a consistent estimator for the slope parameter in a linear regression model in the presence of multiplicative measurement error, by correcting the asymptotic bias of the least squares estimator. The only assumption that he makes about the measurement error is that it is independent and identically distributed. Iturria, Carroll, and Firth (1999) consider a polynomial regression model in the presence of multiplicative measurement error. They compare two methods differing in their assumptions about the distribution of the measurement error, and the distribution of the unobserved variable. They derive a consistent estimator and asymptotic standard error using M-estimation.

The Simulation-Extrapolation (SIMEX) method developed by Cook and Stefanski (1994) is well suited for estimating and reducing the bias due to additive measurement error. We apply the SIMEX method to the case of multiplicative measurement error. As Hwang (1986) points out, the first thought of a statistician when he see a multiplicative measurement error model is usually to apply a logarithmic transformation, so that the multiplicative measurement error model becomes an additive one. The underlying problem of this approach is that the logarithmic transformation destroys the linearity of the model, so that the common methods developed to correct the effects of measurement error on the properties of linear estimators generally fail.³ In this paper, we show how the SIMEX method has to be modified in order to be applied to the case of multiplicative measurement error, without using a logarithmic transformation.

The outline of the paper is as follows. Section 2 gives a short review of the SIMEX method, and shows how it can be extended to the case of multiplicative measurement error. Section 3 presents evidence on the performance of our method. Based on the results of a Monte-Carlo experiment for the linear and probit regression model for different values of the measurement error variance, we show that the multiplicative SIMEX method nicely corrects for the estimation bias introduced by data masking through multiplicative measurement error. Section 4 concludes and gives an outlook about further research.

³See Hwang (1986).

2 The SIMEX Method

2.1 SIMEX for Additive Measurement Error

The SIMEX method is a simulation based method of estimating and reducing the bias due to additive measurement error. SIMEX is a two-step estimation procedure consisting of a simulation step and an extrapolation step. The key idea is to use the information from an incremental addition of measurement error to the mismeasured (masked) data W_i using computer simulated random errors. Adding additional measurement error to the data by the simulation exercise allows the statistician to infer in which way the estimation bias is affected by the increase of the variance of the measurement error. This is the so-called simulation step. In the extrapolation step, the estimated parameters are modelled as a function of the magnitude of the variance of the measurement error and extrapolated to the case of no measurement error.

Suppose that the explanatory variable X_i contains sensitive information, which should be protected against disclosure. Rather than observing X_i , we observe a masked explanatory variable W_i defined as:

$$W_i = X_i + u_i, \quad i = 1, \dots, N, \quad (2.1)$$

where u_i is an independent random variable with $E[u_i | X_i] = 0$ and $V[u_i | X_i] = \sigma_u^2$, that is added to the original variable in order to mask it.

In the simulation step of the SIMEX algorithm, additional measurement error is added to the covariate measured with error. We generate B new covariates $W_{i,b}(\lambda_t)$ by the rule:

$$W_{i,b}(\lambda_t) = W_i + \sqrt{\lambda_t} u_{i,b}, \quad b = 1, \dots, B, \quad t = 1, \dots, T, \quad i = 1, \dots, N, \quad (2.2)$$

where $0 = \lambda_0 < \lambda_1 < \lambda_2 < \dots < \lambda_T = 2$,⁴ are given parameters controlling for the variance of the measurement error, and $\{u_{i,b}\}_{b=1}^B$ are iid computer simulated normal random numbers with mean zero and variance σ_u^2 . Note that for each λ_t the simulation step creates B additional datasets (replication samples) with the same

⁴The value $\lambda_T = 2$ is recommended by Carroll, Ruppert, and Stefanski (1995)

dependent variable Y_i and the explanatory variable $W_{i,b}(\lambda_t)$ whose variance

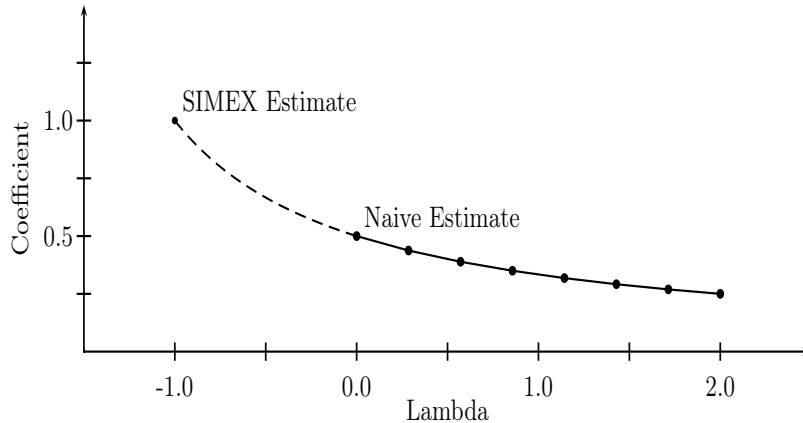
$$V[W_{i,b}(\lambda_t)] = \sigma_x^2 + (1 + \lambda_t)\sigma_u^2, \quad (2.3)$$

increases with the control parameter λ_t . Let $\hat{\beta}_b(\lambda_t)$ denote the vector of naive estimates obtained by regression of Y on $W_b(\lambda_t)$ for each λ_t . Given the B estimates for each λ_t , we can compute an average estimate $\hat{\beta}(\lambda_t) = \frac{1}{B} \sum_{b=1}^B \hat{\beta}_b(\lambda_t)$. The estimates $\hat{\beta}(\lambda_t)$ are depicted as the filled circles in Figure 1 below.

In the extrapolation step each component of the vector $\hat{\beta}(\lambda_t)$ is modelled as a function of λ_t for $\lambda_t \geq 0$. The SIMEX estimator is defined as the extrapolation of $\hat{\beta}(\lambda_t)$ to $\hat{\beta}(\lambda_t = -1)$, which represents the bias free estimate of β .

To estimate the variance, one can either use the delta method (Carroll, Kuechenhoff, Lombard, and Stefanski (1996)), the jackknife type (Stefanski and Cook (1995)) or the bootstrap. Carroll, Kuechenhoff, Lombard, and Stefanski (1996) derive the asymptotic distribution of the SIMEX estimator for parametric models.

Figure 1: Illustration of the Extrapolation Step



Note that in the case of disclosure limitation, this method can be implemented without additional expenditure for the data collecting institutions. They only have to provide the variance (univariate case) σ_u^2 or the variance-covariance matrix (mul-

tivariate case) Σ_{uu} to the data user. The information about the variance on the measurement error is sufficient to get a consistent estimate of the parameters of interest, and does not increase the probability of reidentifying a single unit.

2.2 SIMEX for Multiplicative Measurement Error

Let us now consider a multiplicative measurement error model defined as:

$$W_i = X_i * u_i, \quad (2.4)$$

where u_i is an independent random variable with $E[u_i | X_i] = 1$ and $V[u_i | X_i] = \sigma_u^2$, that is multiplied with the original variable. We suppose that the expectation of u_i given X_i is 1, because it is noticeable that the mean of the masked data is equal to the mean of the original one.

For multiplicative measurement error we transform only the simulation step of the SIMEX algorithm. Now, in the simulation step, the mismeasured variable is multiplied by an additional measurement error. We generate B new covariates $W_{i,b}(\lambda_t)$ by the rule:

$$W_{i,b}(\lambda_t) = W_i * u_{i,b}^{\lambda_t}, \quad b = 1, \dots, B, \quad t = 1, \dots, T, \quad i = 1, \dots, N, \quad (2.5)$$

where $0 = \lambda_0 < \lambda_1 < \lambda_2 < \dots < \lambda_T$, are given integers controlling for the variance of the measurement error, and $\{u_{i,b}\}_{b=1}^B$ are iid computer simulated log-normally distributed random numbers with mean one and variance σ_u^2 . Note that for each λ_t the simulation step creates B additional datasets (replication samples) with the same dependent variable Y_i and the explanatory variable $W_{i,b}(\lambda_t)$ with variance⁵

$$V[W_{i,b}(\lambda_t)] = E\left[\left(u_{i,b}^{\lambda_t+1}\right)^2\right] E[X_i^2] - E[X_i]^2. \quad (2.6)$$

Let $\hat{\beta}_b(\lambda_t)$ denote the vector of naive estimates obtained by regression of Y on $W_b(\lambda_t)$ for each λ_t . Given the B estimates for each λ_t , we can compute an average estimate $\hat{\beta}(\lambda_t) = \frac{1}{B} \sum_{b=1}^B \hat{\beta}_b(\lambda_t)$.

As in the additive case, each component of the vector $\hat{\beta}(\lambda_t)$ is modelled as a function

⁵See appendix A.1 for more explanations

of λ_t for $\lambda_t \geq 0$ in the extrapolation step. The multiplicative SIMEX (M.SIMEX) estimator is defined as the extrapolation of $\hat{\beta}(\lambda_t)$ to $\hat{\beta}(\lambda_t = -1)$, which represents the bias free estimate of β .

We use the bootstrap method in order to calculate the variance of the M.SIMEX estimator. This approach is very computer intensive, because the M.SIMEX estimator has to be calculated for every bootstrap sample.

Note that as for the case of additive and or multiplicative measurement error, the data collecting institutions only have to provide the variance-covariance matrix Σ_{uu} to the data user, which is sufficient to get a consistent estimate of the parameters of interest, in the presence of additive or multiplicative measurement error.

Carroll, Kuechenhoff, Lombard, and Stefanski (1996) show that the SIMEX estimator is consistent if the exact extrapolation function is used. This is also true for the M.SIMEX estimator. The M.SIMEX estimator is consistent when the extrapolation function is correctly specified, or approximately consistent, if the extrapolation function is a good approximation of the relationship between $\hat{\beta}(\lambda_t)$ and λ_t ⁶. For the linear regression model for example, this relationship can explicitly be determined. Suppose that we consider the following bivariate linear regression model, where the explanatory variable X_i is measured with error

$$Y_i = \alpha + \beta X_i + \varepsilon_i. \quad (2.7)$$

Suppose we observe $W_i = X_i u_i$ instead of X_i . If we use the M.SIMEX method, we regress Y_i of $W_{i,b} = W_i u_i^{\lambda_t}$ and we get the following naive OLS estimators of the slope parameter and the constant:

$$\hat{\beta}_{naive}(\lambda_t) = \frac{\sum_i (W_{i,b} - \bar{W}_b)(Y_i - \bar{Y})}{\sum_i (W_{i,b} - \bar{W}_b)^2}, \quad (2.8)$$

$$\hat{\alpha}_{naive}(\lambda_t) = \bar{Y} - \hat{\beta}_{naive}(\lambda_t) \bar{W}_b. \quad (2.9)$$

Simplifying this expression a little bit, we can write both estimators in dependence

⁶See Carroll, Kuechenhoff, Lombard, and Stefanski (1996)

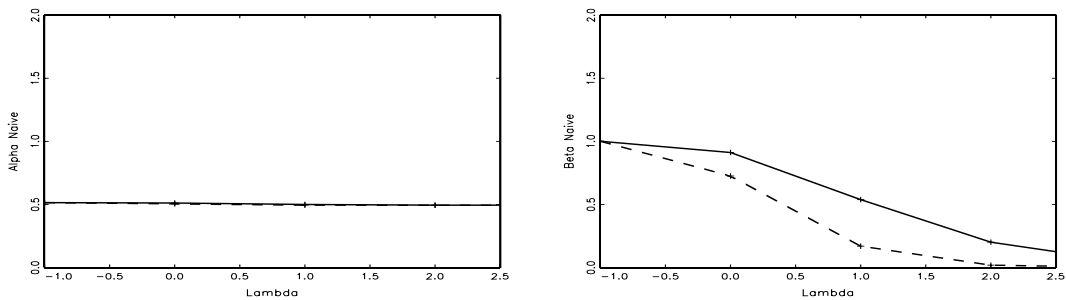
of λ_t :⁷

$$\hat{\beta}_{naive}(\lambda_t) = \frac{\sigma_x^2}{\mathbb{E} \left[(u_{i,b}^{\lambda_t+1})^2 \right] \mathbb{E} [X_i^2] - \mathbb{E} [X_i]^2} \beta, \quad (2.10)$$

$$\hat{\alpha}_{naive}(\lambda_t) = \bar{Y} - \hat{\beta}_{naive}(\lambda_t) \bar{X}. \quad (2.11)$$

In Figure 2 we plot the function $\hat{\beta}_{x^m}(\lambda_t)$ for different value of λ_t , and for $N = 1000$.

Figure 2: Relationship between the naive estimator $\hat{\alpha}(\lambda_t)$ and $\hat{\beta}(\lambda_t)$ and λ_t



We display the relationship between $\hat{\alpha}(\lambda_j)$ (left) and $\hat{\beta}(\lambda_j)$ (right) and λ_j for the linear regression model. We suppose that the true values of α and β are 0.5 and 1. In both plots, the solid line corresponds to a value of σ_u^2 of 0.1 and the dotted line to a value of $\sigma_u^2 = 0.3$.

For the linear regression model, we are able to calculate directly the exact extrapolation function, and are not forced to use an approximation of it.⁸ In this case we get a consistent estimate of the parameter of interest. For more complex models, like probit models, where the relationship between between $\hat{\beta}(\lambda_t)$ and λ_t can not be specified explicitly, we use an approximation of the extrapolation function, and get a so-called approximately consistent estimator, see Carroll, Kuechenhoff, Lombard, and Stefanski (1996).

⁷See Appendix A.4

⁸Nevertheless, for the Monte-Carlo Experiment we use three different specifications of the extrapolation function in order to find the one which is more appropriate to reduce the bias.

3 Simulation Study

The Monte-Carlo experiment illustrates the quantitative effect of multiplicative measurement error. We focus on the linear regression model and on the probit model.

We assume that the bivariate linear regression model is given by

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i, \quad (3.1)$$

where the error term ε_i is normally distributed with mean 0 and variance 1. To protect the data against disclosure, we disturb the explanatory variable X_{1i} , which is normally distributed with mean 2 and variance equal to 1, by a multiplicative error u_i following a log-normal distribution with mean 1 and variance $\sigma_u^2 = \{0.01, 0.04, 0.1, 0.3\}$.

For the standard probit model, we consider the same simulation design, and define:

$$Y_i = \begin{cases} 1, & \text{if } Y_i^* = \beta_0 + \beta_1 X_{1i} + \varepsilon_i > 0, \\ 0, & \text{if } Y_i^* = \beta_0 + \beta_1 X_{1i} + \varepsilon_i \leq 0. \end{cases} \quad (3.2)$$

For each simulation we suppose that the true values for β are $\beta_0 = -1$ and $\beta_1 = 0.25$. For the SIMEX approach, we suppose that the λ_t 's take the value $0 = \lambda_0 < \lambda_1 = 1 < \lambda_2 = 2 < \lambda_3 = 3 < \lambda_4 = 4$, and generate for each value of λ_t , $B=50$ new covariates. Our Monte-Carlo results are based on two different samples of size $N = 100$ and $N = 1000$, which are replicated $R = 1000$ times.

For the SIMEX approach we use three different specifications of the extrapolation function, $G(\lambda, \gamma)$, in order to see how strongly the estimates depend on this function. We use a linear extrapolation function, where $G(\lambda, \gamma) := \gamma_0 + \gamma_1 \lambda$, the quadratic one which is suggested by Cook and Stefanski (1994), where $G(\lambda, \gamma) := \gamma_0 + \gamma_1 \lambda + \gamma_2 \lambda^2$ and a nonlinear extrapolation function, where $G(\lambda, \gamma) := \gamma_0 + \gamma_1 (\gamma_2 + \lambda)^{-1}$. For the linear and the quadratic ones, this gives a system of equations $L\gamma = B$:

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 1 & \lambda_1 & \lambda_1^2 \\ \vdots & \vdots & \vdots \\ 1 & \lambda_T & \lambda_T^2 \end{bmatrix}, \quad B = \begin{bmatrix} \hat{\beta}(\lambda_0) \\ \hat{\beta}(\lambda_1) \\ \vdots \\ \hat{\beta}(\lambda_T) \end{bmatrix}, \quad \gamma = \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \end{bmatrix},$$

where the last column of the matrix L , and the last line of the vector γ are dropped for the linear extrapolation function. Solving the system of equations yields the following result:

$$\hat{\gamma} = (L'L)^{-1}L'B. \quad (3.3)$$

In the nonlinear case, we suppose that there exist three points $(\hat{\beta}(\lambda_0), \lambda_0)$, $(\hat{\beta}(\lambda_1), \lambda_1)$ and $(\hat{\beta}(\lambda_2), \lambda_2)$ where the function is fitted. Suppose that $\lambda_0 = 0$, $\lambda_1 = 1$, and $\lambda_2 = 2$, this means that this choice corresponds to a subset of $\Lambda = \{\lambda_0, \lambda_1, \dots, \lambda_T\}$. The solution is given by:

$$\begin{aligned} \hat{\gamma}_0 &= \frac{\hat{\beta}(\lambda_0)(\hat{\beta}(\lambda_2) - \hat{\beta}(\lambda_1)) - \hat{\beta}(\lambda_2)(\hat{\beta}(\lambda_1) - \hat{\beta}(\lambda_0))}{\hat{\beta}(\lambda_2) - 2\hat{\beta}(\lambda_1) + \hat{\beta}(\lambda_0)}, \\ \hat{\gamma}_1 &= 2 \frac{(\hat{\beta}(\lambda_1) - \hat{\beta}(\lambda_0))(\hat{\beta}(\lambda_0) - \hat{\beta}(\lambda_2))(\hat{\beta}(\lambda_2) - \hat{\beta}(\lambda_1))}{(\hat{\beta}(\lambda_2) - 2\hat{\beta}(\lambda_1) + \hat{\beta}(\lambda_0))^2}, \\ \hat{\gamma}_2 &= 2 \frac{\hat{\beta}(\lambda_1) - \hat{\beta}(\lambda_2)}{\hat{\beta}(\lambda_2) - 2\hat{\beta}(\lambda_1) + \hat{\beta}(\lambda_0)}. \end{aligned}$$

The SIMEX estimator is obtained by evaluating the extrapolation function at the point $\lambda = -1$, i.e.:

$$\begin{aligned} \hat{\beta}_{M_SIMEX_L} &= (1, -1)\hat{\gamma} \quad \text{for the linear case,} \\ \hat{\beta}_{M_SIMEX_Q} &= (1, -1, 1)\hat{\gamma} \quad \text{for the quadratic case,} \\ \hat{\beta}_{M_SIMEX_NL} &= \hat{\gamma}_0 + \hat{\gamma}_1(\hat{\gamma}_2 - 1)^{-1} \quad \text{for the nonlinear case.} \end{aligned}$$

Since the SIMEX approach is very computer intensive, we decide to use 50 bootstrap replications for each Monte-Carlo replication, in order to get an estimate of the standard deviation of the estimator.

Table 1 contains the results of the Monte Carlo simulations for the sample sizes $N = 100$ and 1000 for the linear regression model, and Table 2 contains the Monte Carlo results for the probit model for the sample sizes $N = 100$ and 1000 .⁹ M_SIMEX_L , respectively M_SIMEX_Q and M_SIMEX_NL denote the multiplicative SIMEX estimator obtained using a linear extrapolation function, respectively a quadratic and a nonlinear extrapolation function. In order to have a benchmark

⁹We use OLS estimation to obtain the estimates for the linear regression model, and Maximum-Likelihood to get the estimates of the probit model.

of how strongly the measurement error bias the estimates we also report the naive estimate (Naive) on the mismeasured data and the true estimate based on the original data set without multiplicative measurement error (True) which provide insights on how close our estimates come to the true one's.

For the linear and the probit models, we see that the bias of the M_SIMEX estimators is somewhat larger than the bias of OLS and ML estimates when the data set is not masked for $N = 100$ and $N = 1000$ when the variance of the multiplicative error u_i is small, i.e. $\sigma_u^2 = 0.01$ or 0.04 . As expected, the bias is somewhat larger for the small sample than for $N = 1000$. The same can be said about the root mean squared error (RMSE) which decreases considerably with a larger sample size. The RMSE is not far from the corresponding RMSE from the OLS estimator and the ML estimator of the original data for the M_SIMEX_Q and M_SIMEX_{NL} estimators for both sample sizes for small values of σ_u^2 . For the linear model, the bias increases with the value of σ_u^2 for both sample sizes. In this case we can note that for medium ($\sigma_u^2 = 0.1$) or large ($\sigma_u^2 = 0.3$) values of the variance of the multiplicative error the bias of the M_SIMEX estimator is much higher when the linear and the nonlinear extrapolation functions are used than the quadratic one. However this bias becomes unacceptably large for large values of σ_u^2 , for each specification of the extrapolation function.

The relative standard error, RELSE, is defined as the ratio of the average standard error of the estimator over the number of completed MC replications to the empirical standard deviation of the estimator. Indeed, when the number of replications tends to infinity, the standard error of the estimates converges to the true standard error, for a finite N . A deviation of RELSE from 1 provides information about the accuracy of the estimation of the standard error based on the asymptotic distribution. In small samples the standard errors of the estimates cannot be estimated with great precision. But with a larger sample size ($N = 1000$) the relative estimated standard error comes close to its desired value 1, when σ_u^2 is small. The estimated standard errors become a little less precise when the value of the multiplicative measurement error becomes larger for both sample sizes (see Table 1).

Overall we can conclude that the quality of the M_SIMEX estimates comes close to the best case estimates for the original data, for moderate multiplicative measurement errors.

		Estimation Results of the linear regression model															
σ_u^2		$N = 100$								$N = 1000$							
		β_0				β_1				β_0				β_1			
		Mean	Bias	RMSE	RELSE	Mean	Bias	RMSE	RELSE	Mean	Bias	RMSE	RELSE	Mean	Bias	RMSE	RELSE
0.01	True	-0.997	0.003	0.215	1.044	0.248	-0.002	0.096	1.049	-0.999	0.001	0.071	0.999	0.250	0.000	0.031	1.008
	Naive	-0.971	0.029	0.213	1.047	0.235	-0.015	0.095	1.050	-0.975	0.025	0.073	1.002	0.238	-0.012	0.033	1.006
	M_SIMEX _L	-1.043	-0.043	0.240	1.041	0.274	0.024	0.112	1.040	-1.050	-0.050	0.093	0.999	0.279	0.029	0.046	1.003
	M_SIMEX _Q	-0.993	0.007	0.219	1.041	0.245	-0.005	0.098	1.045	-0.997	0.003	0.072	1.001	0.249	-0.001	0.032	1.003
	M_SIMEX _{NL}	-0.983	0.017	0.216	1.069	0.242	-0.008	0.097	1.139	-0.987	0.013	0.072	0.998	0.244	-0.006	0.032	1.003
0.04	True	-0.997	0.003	0.215	1.044	0.248	-0.002	0.096	1.049	-0.999	0.001	0.071	0.999	0.250	0.000	0.031	1.008
	Naive	-0.910	0.090	0.221	1.040	0.205	-0.045	0.100	1.040	-0.915	0.085	0.108	1.002	0.208	-0.042	0.051	0.999
	M_SIMEX _L	-0.997	0.003	0.235	1.032	0.252	0.002	0.110	1.030	-1.007	-0.007	0.078	0.999	0.264	0.016	0.040	0.999
	M_SIMEX _Q	-1.012	-0.012	0.242	1.046	0.258	0.008	0.113	1.045	-1.022	-0.022	0.082	1.001	0.257	0.007	0.037	0.999
	M_SIMEX _{NL}	-0.949	0.050	0.224	1.201	0.223	-0.027	0.106	1.051	-0.954	0.046	0.085	0.999	0.227	-0.023	0.039	0.998
0.1	True	-1.004	-0.004	0.234	0.962	0.250	0.000	0.103	0.979	-0.999	0.001	0.071	0.999	0.250	0.000	0.031	1.008
	Naive	-0.836	0.164	0.260	0.964	0.167	-0.083	0.120	0.962	-0.827	0.173	0.183	0.997	0.164	-0.086	0.090	0.983
	M_SIMEX _L	-0.896	0.104	0.251	0.954	0.200	-0.050	0.115	0.953	-0.888	0.111	0.131	0.998	0.197	-0.053	0.062	0.992
	M_SIMEX _Q	-0.979	0.020	0.271	0.958	0.243	-0.007	0.128	0.957	-0.976	0.024	0.086	1.000	0.244	-0.006	0.040	0.992
	M_SIMEX _{NL}	-0.910	0.090	0.257	1.026	0.198	-0.052	0.153	0.769	-0.898	0.102	0.124	0.998	0.197	-0.053	0.062	0.991
0.3	True	-0.997	0.003	0.237	0.951	0.250	0.000	0.104	0.971	-0.999	0.001	0.071	0.999	0.250	0.000	0.031	1.008
	Naive	-0.683	0.317	0.360	0.947	0.095	-0.155	0.170	0.940	-0.681	0.319	0.323	0.985	0.091	-0.159	0.160	0.945
	M_SIMEX _L	-0.702	0.298	0.349	0.949	0.106	-0.144	0.163	0.961	-0.699	0.301	0.306	1.000	0.100	-0.150	0.151	0.978
	M_SIMEX _Q	-0.796	0.204	0.317	0.955	0.156	-0.094	0.148	0.961	-0.799	0.201	0.214	0.994	0.154	-0.096	0.102	0.977
	M_SIMEX _{NL}	-0.805	0.195	0.376	1.049	0.142	-0.108	0.202	0.882	-0.786	0.214	0.226	1.068	0.137	-0.113	0.117	0.980

Table 1: Estimation results for the linear regression model. This table contains the results of the Monte-Carlo simulations for sample size $N = 100$ on the left part and for $N = 1000$ on the right part, for the true and the naive OLS-estimator, the multiplicative SIMEX-estimator for three different specifications of the extrapolation function (linear, quadratic and nonlinear) for different values of the measurement error term $\sigma_u^2 \in \{0.01, 0.04, 0.1, 0.3\}$, .

σ_u^2	Method	Estimation Results of the Probit Model							
		β_0				β_1			
		Mean	Bias	RMSE	RELSE	Mean	Bias	RMSE	RELSE
$N = 100$									
0.01	True	-1.026	-0.026	0.328	0.972	0.256	0.006	0.140	0.988
	Naive	-0.994	0.006	0.315	0.982	0.241	-0.009	0.135	0.994
	M.SIMEX _L	-1.074	-0.074	0.366	1.051	0.283	0.033	0.162	1.061
	M.SIMEX _Q	-1.023	-0.023	0.336	1.055	0.255	0.005	0.144	1.069
	M.SIMEX _{NL}	-1.005	-0.005	0.352	1.097	0.248	-0.002	0.141	1.191
0.04	True	-1.026	-0.026	0.328	0.972	0.256	0.006	0.140	0.988
	Naive	-0.921	0.079	0.302	0.987	0.207	-0.043	0.131	0.995
	M.SIMEX _L	-1.011	-0.011	0.340	1.101	0.254	0.004	0.152	1.539
	M.SIMEX _Q	-1.037	-0.037	0.364	1.137	0.265	0.015	0.161	1.751
	M.SIMEX _{NL}	-0.970	0.030	0.325	1.204	0.228	-0.022	0.141	1.265
$N = 1000$									
0.01	True	-1.000	0.000	0.095	1.038	0.251	0.001	0.041	1.054
	Naive	-0.972	0.028	0.097	1.031	0.237	-0.013	0.042	1.047
	M.SIMEX _L	-1.052	-0.052	0.118	1.029	0.279	0.029	0.055	1.043
	M.SIMEX _Q	-1.000	0.000	0.099	1.032	0.250	0.000	0.042	1.045
	M.SIMEX _{NL}	-0.986	0.014	0.097	1.030	0.244	-0.006	0.041	1.046
0.04	True	-1.003	-0.003	0.095	1.039	0.252	0.002	0.041	1.053
	Naive	-0.906	0.093	0.127	1.039	0.205	-0.045	0.057	1.054
	M.SIMEX _L	-0.998	0.002	0.101	1.041	0.253	0.003	0.045	1.056
	M.SIMEX _Q	-1.023	-0.023	0.109	1.046	0.264	0.014	0.050	1.060
	M.SIMEX _{NL}	-0.950	0.050	0.106	1.042	0.225	-0.024	0.047	1.055

Table 2: Estimation results for the probit regression model. This table contains the results of the Monte-Carlo simulations for sample size $N = 100$ and $N = 1000$, for the true and the naive ML-estimator, the multiplicative SIMEX-estimator for three different specifications of the extrapolation function (linear, quadratic and nonlinear) for different values of the measurement error term $\sigma_u^2 \in \{0.01, 0.04\}$, .

4 Conclusion

In this paper we proposed using the multiplicative SIMEX (M_SIMEX) approach for parameter estimation in regression models in the presence of multiplicative measurement error. Our approach is quite general since the only assumption to be made is the existence of a consistent estimator for the model parameters in the case of no anonymization of the data set. This means that this method can be applied to various regression models, linear or nonlinear. In order to apply the proposed method, the data collecting institution only has to provide the variance covariance matrix of the multiplicative measurement error to the data user. As for the original SIMEX method proposed by Cook and Stefanski (1994), the major problem here is the specification of the extrapolation function, because in some situation only an approximation of this function is available, so that the multiplicative SIMEX approach is only approximately consistent. In some cases, it is possible to calculate the true extrapolation function, and we show that a quadratic extrapolation function is a good approximation most of the time. We used the bootstrap method to calculate the variance of the estimator. The Monte-Carlo experiments show that the proposed method reduces the bias due to the presence of multiplicative measurement error compared to the naive one, and seems to be a helpful tool to obtain consistent parameter estimates when data have been masked through multiplicative measurement error.

However, there remain further research topics. First of all, we will in near the future derive the asymptotic properties of our proposed estimator. Second, a trade-off analysis between bias and efficiency for different specifications of the multiplicative measurement error should be carried out. This would shed light on the relation between disclosure risk and estimation quality. Third, it seems appropriate to compare the efficiency of our proposed approach with the traditional one which consists of applying a logarithmic transformation, so that the multiplicative measurement error problem becomes an additive one. Finally, it would be useful to investigate if a combination of statistical estimation methods. Ronning and Rosemann (2006) show how to combine the additive SIMEX method and the PRAM technique to correct the estimates in the presence of additive measurement error in the explanatory variable and misclassification in the dependent variable. Another possibility would be for example to combine the multiplicative SIMEX approach and the PRAM technique to correct the estimates for the presence of multiplicative measurement error in the

explanatory variable and misclassification in the dependent variable.

References

- AMEMIYA, T. (1985): “Instrumental Variable Estimator for the Non-linear Errors in Variable Model,” *Journal of Econometrics*, 28, 273–289.
- CARROLL, R., H. KUECHENHOFF, F. LOMBARD, AND L. STEFANSKI (1996): “Asymptotics for the Simex Estimator in Structural Measurement Error Models,” *Journal of the American Statistical Association*, 91, 242–250.
- CARROLL, R., D. RUPPERT, AND L. STEFANSKI (1995): *Measurement Error in Nonlinear Models*. Chapman and Hall.
- COOK, AND STEFANSKI (1994): “A Simulation Extrapolation Method for Parametric Measurement Error Models,” *Journal of the American Statistical Association*, 89, 1314–1328.
- DOMINGO-FERRER, J., AND V. TORRA (2002): “Disclosure Control Methods and Information Loss for Microdata,” in *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, Amsterdam: North-Holland, 93-112.
- FULLER, W. (1987): *Measurement Error Models*. Wiley.
- HAUSMAN, J., W. NEWEY, AND J. POWELL (1995): “Nonlinear Errors in Variables Models,” *Journal of Econometrics*, 41, 159–185.
- HONG, H., AND E. TAMER (2003): “A Simple Estimator for Nonlinear Error in Variables Models,” .
- HWANG, J. T. (1986): “Multiplicative Errors-in-Variables Models with Applications to Recent Data Released by the U.S. Department of Energy,” *Journal of the American Statistical Association*, 81, 680–688.
- ITURRIA, S. J., R. J. CARROLL, AND D. FIRTH (1999): “Polynomial Regression and Estimating Functions in the Presence of Multiplicative Measurement Error,” *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61, 547 – 561.
- LECHNER, S., AND W. POHLMEIER (2004): “To Blank or Not to Blank? A Comparison of the Effects of Disclosure Limitation Methods on Nonlinear Regression Estimates,” in *Domingo-Ferrer J. and Torra V : Privacy in Statistical Databases*, CASC Project Final Conference, PSD 2004, LNCS 3050, Springer.
- LEE, L., AND J. SEPANSKI (1995): “Estimation of Linear and Nonlinear Error in Variables Models using Validation Data,” *Journal of the American Statistical Association*, 90, 130–140.

- POHLMEIER, W., G. RONNING, AND J. WAGNER (2005): “Econometrics of Anonymized Micro Data,” *Sonderband der Jahrbücher für Nationalökonomie und Statistik*, 225(5).
- RONNING, G., AND M. ROSEMAN (2006): “Estimation of the Probit Model From Anonymized Micro Data,” *Work Session on Statical Data Confidentiality, Geneva, 9-11 November 2005. Monograph of Official Statistics. Eurostat, Luxembourg 2006, S. 207 - 216. Geneva, 9 - 11 November 200.*
- RONNING, G., M. ROSEMAN, AND H. STROTMANN (2005): “Post-Randomization under Test: Estimation of the Probit Model,” *Jahrbuecher fuer Nationaloekonomie und Statistik*, 225(5), 544–566.
- SCHENNACH, S. M. (2004): “Estimation of Nonlinear Models with Measurement Error,” *Econometrica*, 72(1), 33–75.
- STEFANSKI, L., AND J. COOK (1995): “Simulation-Extrapolation: The Measurement Error Jackknife,” *Journal of the American Statistical Association*, 90(432), 1247–1256.

A Appendix

A.1 Variance of the mismeasured variable

The variance of the product of two independent random variables X and Y is defined as:

$$V[XY] = E[X]^2 V[Y] + E[Y]^2 V[X] + V[X] V[Y] \quad (\text{A.1})$$

Applying this formula to equation (2.5) we get:

$$\begin{aligned} V[W_{i,b}(\lambda_t)] &= V[X_i * u_{i,b}^{\lambda_t+1}] \\ &= E[X_i]^2 V[u_{i,b}^{\lambda_t+1}] + E[u_{i,b}^{\lambda_t+1}]^2 V[X_i] + V[X_i] V[u_{i,b}^{\lambda_t+1}] \\ &= E[X_i]^2 V[u_{i,b}^{\lambda_t+1}] + V[X_i] + V[X_i] V[u_{i,b}^{\lambda_t+1}] \\ &= V[u_{i,b}^{\lambda_t+1}] (E[X_i]^2 + V[X_i]) + V[X_i] \\ &= V[u_{i,b}^{\lambda_t+1}] E[X_i^2] + V[X_i] \\ &= E[(u_{i,b}^{\lambda_t+1})^2] E[X_i^2] - E[X_i]^2 \end{aligned} \quad (\text{A.2})$$

Using the fact that u_i are iid and $E[u_i] = 1$, so that $E[u_{i,b}^{\lambda_t+1}]^2 = 1$. In the other lines the simplifications are due to the replacement of $V[X_i]$, (resp. $V[u_{i,b}^{\lambda_t+1}]$), by $E[X_i^2] - E[X_i]^2$, (resp. $E[(u_{i,b}^{\lambda_t+1})^2] - E[u_{i,b}^{\lambda_t+1}]^2$).

A.2 Lognormal distribution

If $X \sim N(\mu, \sigma^2)$, then e^X is log normally distributed with density function

$$f_X(x) = \frac{1}{x} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right).$$

Note: We write $X \sim \text{LogN}(\mu, \sigma^2)$.

The expectation of X is given by:

$$\text{E}[X] = e^{\mu + \frac{\sigma^2}{2}}$$

and the variance of X is given by:

$$\text{V}[X] = e^{2\mu + \sigma^2} (e^{\sigma^2} - 1).$$

A.3 Value for the variance the multiplicative error term U

As mentioned before, we suppose that $u \sim \text{LogN}(\mu, \sigma^2)$, and we want u to have expectation 1. This means that during the simulation we suppose that $\mu = -\sigma^2/2$. In the following Table, you will find the value of μ and the value of the variance of u , for different value of σ^2

σ^2	μ	$\text{E}[u]$	$\text{V}[u]$
0.1	-0.05	1	0.1052
0.2	-0.10	1	0.2214
0.3	-0.15	1	0.3498
0.4	-0.20	1	0.4918
0.5	-0.25	1	0.6487
0.6	-0.30	1	0.8221
0.7	-0.35	1	1.0137
0.8	-0.40	1	1.2255
0.9	-0.45	1	1.4596
1.0	-0.50	1	1.7183

Table 3: Moments of the lognormal distribution for different value of σ^2

A.4 Extrapolation Function for the linear regression model

We start with equation (2.8), and suppose that $b = 1$. By the Weak Law of Large Number we have

$$\begin{aligned}
\frac{1}{N} \sum_i (W_{i,b} - \bar{W}_b)^2 &= \frac{1}{N} \sum_i (W_i u_i^{\lambda_t} - \overline{W u^{\lambda_t}})^2 \\
&= \frac{1}{N} \sum_i \left(X_i^2 u_i^{2(\lambda_t+1)} - \bar{X}^2 \right) \\
&= \mathbf{E} \left[(u_{i,b}^{\lambda_t+1})^2 \right] \mathbf{E} [X_i^2] - \mathbf{E} [X_i]^2 \tag{A.3}
\end{aligned}$$

$$\begin{aligned}
\frac{1}{N} \sum_i (W_{i,b} - \bar{W}_b)(Y_i - \bar{Y}) &= \frac{1}{N} \beta \sum_i (W_{i,b} - \bar{W}_b)(X_i - \bar{X}) + \frac{1}{N} \beta \sum_i (W_{i,b} - \bar{W}_b)(\varepsilon_i - \bar{\varepsilon}) \\
&= \frac{1}{N} \beta \sum_i (W_{i,b} - \bar{W}_b)(X_i - \bar{X}) \tag{1} \\
&= \beta \left(\frac{1}{N} \sum_i W_{i,b} X_i - \bar{X} \bar{W}_b \right) \\
&= \beta (\mathbf{E} [W_{i,b} X_i] - \mathbf{E} [X] \mathbf{E} [W_b]) \\
&= \beta (\mathbf{E} [W_i u_i^{\lambda_t} X_i] - \mathbf{E} [X] \mathbf{E} [W_i u_i^{\lambda_t}]) \\
&= \beta \mathbf{E} [u_i^{\lambda_t+1}] [\mathbf{E} [X_i^2] - \mathbf{E} [X_i]^2] \\
&= \beta \mathbf{V} [X_i]. \tag{2} \tag{A.4}
\end{aligned}$$

(1) \longrightarrow because $X_i \perp \varepsilon_i, u_i$ and $u_i \perp \varepsilon_i$

(2) $\longrightarrow \mathbf{E} [u_i^{\lambda_t+1}] = 1$

This means that

$$\text{plim} \hat{\beta}_{naive}(\lambda_t) = \frac{\sigma_x^2}{\mathbf{E} [(u_{i,b}^{\lambda_t+1})^2] \mathbf{E} [X_i^2] - \mathbf{E} [X_i]^2} \beta.$$