ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Küchenhoff, Helmut; Lederer, Wolfgang; Lesaffre, Emmanuel

Working Paper

Asymptotic Variance Estimation for the Misclassification SIMEX

Discussion Paper, No. 473

Provided in Cooperation with:
Collaborative Research Center (SFB) 386: Statistical Analysis of discrete structures - Applications in Biometrics and Econometrics, University of Munich (LMU)

Suggested Citation: Küchenhoff, Helmut; Lederer, Wolfgang; Lesaffre, Emmanuel (2006) : Asymptotic Variance Estimation for the Misclassification SIMEX, Discussion Paper, No. 473, Ludwig-Maximilians-Universität München, Sonderforschungsbereich 386 - Statistische Analyse diskreter Strukturen, München,
https://doi.org/10.5282/ubm/epub.1841

This Version is available at:
https://hdl.handle.net/10419/31104

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

Asymptotic Variance Estimation for the Misclassification SIMEX

Helmut Küchenhoff*, Wolfgang Lederer*, Emmanuel Lesaffre†

February 21, 2006

Abstract

Most epidemiological studies suffer from misclassification in the response and/or the covariates. Since ignoring misclassification induces bias on the parameter estimates, correction for such errors is important. For measurement error, the continuous analog to misclassification, a general approach for bias correction is the SIMEX (simulation extrapolation) originally suggested by Cook and Stefanski (1994). This approach has been recently extended to regression models with a possibly misclassified categorical response and/or the covariates by Küchenhoff et al. (2005), and is called the MC-SIMEX approach. To assess the importance of a regressor not only its (corrected) estimate is needed, but also its standard error. For the original SIMEX approach. Carroll et al. (1996) developed a method for estimating the asymptotic variance. Here we derive the asymptotic variance estimators for the MC-SIMEX approach, extending the methodology of Carroll et al. (1996). We also include the case where the misclassification probabilities are estimated by a validation study. An extensive simulation study shows the good performance of our approach. The approach is illustrated using an example in caries research including a logistic regression model, where the response and a binary covariate are possibly misclassified.

Keywords: misclassification, SIMEX approach, variance estimation

1 Introduction

It is well known that in linear regression analysis the regression coefficients can be severely biased when there is measurement error in continuous regressors or categorical regressors are subject to misclassification. Further, in nonlinear regression models, such as logistic regression, possibly misclassified categorical regressors as well as a possibly misclassified response can lead to severely biased

*Department of Statistics, Ludwig–Maximilians–Universität, München, Germany
†Biostatistical Centre, Catholic University Leuven, Belgium

estimated regression coefficients. There is a rich literature on how to correct for this misclassification bias, see e.g. Gustafson (2004).

In this paper we consider the method recently introduced by Küchenhoff et al. (2005), hereafter denoted by KML. The method is based on the simulation and extrapolation (SIMEX) approach of Cook and Stefanski (1994) for regression models with measurement error in a continuous regressor. The SIMEX idea is to exploit the relationship between the bias in parameter estimation and the amount of measurement error. The adaptation to the misclassification situation is called the MisClassification SIMEX (MC-SIMEX) approach. The MC-SIMEX approach is a computer intensive method that can take into account misclassification of a categorical response or of a categorical regressor or of both. The amount of misclassification is characterized by an exponent of the misclassification matrix. In the first step of the (MC)-SIMEX algorithm, data with a higher amount of measurement error (misclassification) are produced by simulation. In a second step an approximately unbiased estimator is achieved by extrapolating back to the case of no measurement error (no misclassification).

The MC-SIMEX approach delivers estimates of the model parameters for a general class of models (corrected for misclassification). But it is also important to know the standard errors of the corrected estimates. In KML two approaches were applied: (a) the approach of Stefanski and Cook (1995) and (b) a general bootstrap approach. The first approach was applied without any theoretical justification, but the simulation results revealed that the method has good sampling properties. The bootstrap approach, on the other hand, finds its justification from general principles but is time consuming. The bootstrap approach enjoys the advantage that it can easily take into account the uncertainty with which the misclassification probabilities have been estimated, in contrast to the approach of Stefanski and Cook (1995). Thus there is a need for a method which is theoretically justified and can calculate the standard errors in a less computer intensive manner allowing for the uncertainty of the estimated misclassification probabilities. The asymptotic variance estimation has been developed for the original SIMEX by Carroll et al. (1996). Here we transfer this strategy to the case of the MC-SIMEX approach. It turns out that most of the SIMEX theory can be used for the MC-SIMEX approach in a straightforward manner. However, there are some difficulties due to the differences between the characterization of additive measurement error by a simple variance and misclassification by a matrix with at least two parameters. Especially when the misclassification matrix must be estimated from a validation study, requires a careful inspection.

The paper is organized as follows. In Section 2 we give a short review of the MC-SIMEX method. In Section 3 we develop the asymptotic variance estimation method with a focus on the case of an estimated misclassification matrix. In Section 4 we analyze the behavior of our method by simulation. In Section 5 we give an application to data from an oral health study. Concluding remarks are found in Section 6.

2 Review of the MC-SIMEX method

2.1 The MC-SIMEX method

Suppose a general regression model with a response Y and a discrete regressor X. The MC-SIMEX approach can be applied to misclassification of the response as well as to misclassification of the regressors or misclassification in both. Let us denote the possibly corrupted variable by X^* or Y^*, for the corresponding correctly measured (gold standard) variable X and Y, respectively. We also assume that the model contains correctly specified regressors Z.

Suppose now that only X is subject to misclassification. The case where (only or also) the response is prone to misclassification is treated similarly. The misclassification process is described by the misclassification matrix Π, which is defined by its components

$$\pi_{ij} = P(X^* = i|X = j).$$

Π is a $k \times k$ matrix, where k is the number of possible outcomes for X. We denote the parameter of interest by β. If misclassification is ignored then the estimator is classically called *the naive estimator*. Let us denote the limit (when the sample size goes to infinity) by $\beta^*(\Pi)$, since it depends on the misclassification matrix. Further it is assumed that $\beta^*(I_{k\times k}) = \beta$, i. e. that the estimator is consistent when there is no misclassification (represented by identity matrix $I_{k\times k}$). When X is a binary covariate $\beta^*(\Pi)$ depends on the sensitivity $\pi_{11} = P(X^* = 1|X = 1)$ and the specificity $\pi_{00} = P(X^* = 0|X = 0)$, i. e. on

$$\Pi = \begin{pmatrix} \pi_{00} & 1 - \pi_{11} \\ 1 - \pi_{00} & \pi_{11} \end{pmatrix}.$$

The MC-SIMEX employs the function ($\lambda \geq 0$)

$$\lambda \longrightarrow \beta^*(\Pi^\lambda), \tag{1}$$

whereby $\Pi^\lambda := E\Lambda^\lambda E^{-1}$, Λ is the diagonal matrix of eigenvalues and E the corresponding matrix of eigenvectors. For $\lambda = n$, an integer, $\Pi^{1+n} = \Pi^n * \Pi$ and when $n = 0$, $\Pi^0 = I_{k\times k}$. The central idea of the MC-SIMEX method is to add extra misclassification to the possibly corrupted X^*. Namely, if X^* has misclassification Π in relation to matrix X and X^{**} is related to X^* by the misclassification matrix Π^λ then X^{**} is related to X by the misclassification matrix $\Pi^{1+\lambda}$, when the two misclassification mechanisms are independent.

The MC-SIMEX procedure consists of a simulation and an extrapolation step. Given data $(Y_i, X_i^*, Z_i)_{i=1}^n$ the naive estimator is denoted by $\hat{\beta}_{na}[(Y_i, X_i^*, Z_i)_{i=1}^n]$.

Simulation step

For a fixed grid of values $\lambda_1, \ldots, \lambda_m$, ($\geq 0$) B new pseudo data sets are simulated by

$$X^*_{b,i}(\lambda_k) := MC[\Pi^{\lambda_k}](X_i^*), \;\; i = 1, \ldots, n; \; b = 1, \ldots B; k = 1, \ldots, m.$$

where the misclassification operation $MC[M](X_i^*)$ denotes the simulation of a variable given X_i^* with misclassification matrix M. Further, for $\lambda_0 = 0$, with $\hat{\beta}_{\lambda_0} = \hat{\beta}_{na}\left[(Y_i, X_i^*, Z_i)_{i=1}^n\right]$ the estimate of β without further measurement error is obtained and

$$\hat{\beta}_{\lambda_k} := B^{-1} \sum_{b=1}^{B} \hat{\beta}_{na}\left[(Y_i, X_{b,i}^*(\lambda_k), Z_i)_{i=1}^n\right], k = 1, \ldots m. \tag{2}$$

Note that $\hat{\beta}_{\lambda_k}$ is an average over naive estimators corresponding to data with misclassification matrix $\Pi^{1+\lambda_k}$. Thus, $\hat{\beta}_{\lambda_k} = \beta^*(\widehat{\Pi^{1+\lambda_k}})$.

Extrapolation step

We use a parametric approximation $\beta^*(\Pi^\lambda) \approx \mathcal{G}(\lambda, \Gamma)$. Therefore we estimate the parameter Γ by least squares on $[1+\lambda_k, \hat{\beta}_{\lambda_k}]_{k=0}^m$, yielding an estimator $\hat{\Gamma}$. The MC-SIMEX estimator is then given by

$$\hat{\beta}_{SIMEX} := \mathcal{G}(0, \hat{\Gamma}). \tag{3}$$

If β is a vector, the MC-SIMEX estimator can be applied on each component of β separately.

2.2 The extrapolation function

The estimator $\hat{\beta}_{SIMEX}$ is consistent when the extrapolation function is correctly specified, i. e. $\beta^*(\Pi^\lambda) = \mathcal{G}(\lambda, \Gamma)$, for some parameter vector Γ. When $\mathcal{G}(\lambda, \Gamma)$ is a good approximation of $\beta^*(\Pi^\lambda)$ then approximate consistency will hold. KML looked at the relationship between β^* and the misclassification parameter λ for some special cases, also including the case when Y is prone to misclassification. This exercise indicated possibly suitable candidates for the function $\mathcal{G}(\lambda, \Gamma)$. The conclusion was that the extrapolation function was monotonic in all parameters and that it has a curvature which was well approximated by the quadratic extrapolation function. In some simple cases the function was exponential in λ, which was also approximately true in more complicated cases. Therefore, KML recommend to use a quadratic and a log linear extrapolation function given by $\mathcal{G}_Q(\lambda, \Gamma) := \gamma_0 + \gamma_1\lambda + \gamma_2\lambda^2$ and $\mathcal{G}_{LOG}(\lambda) := \exp(\gamma_0 + \gamma_1\lambda)$, respectively.

2.3 Simulation properties

The simulation results indicated that the MC-SIMEX method leads to substantial reduction of bias compared to the naive estimator. The addition of a confounder implied more attenuation than the case without a confounder, and consequently a poorer correction than in the case with no confounder. Finally, even in the complicated situation where both the response and the binary covariate were subject to misclassification and with an additional continuous confounder, the MC-SIMEX correction gave improved estimates even with high misclassification probabilities.

3 Asymptotic variance estimation

The asymptotic distribution theory for the original SIMEX approach has been developed by Carroll et al. (1996), hereafter denoted by CKLS. The idea is to view regression estimation as solving unbiased estimating equations, see e.g. Godambe (1991) for a basic reference on estimating equations. Using asymptotic expansion results, variance estimates are obtained in each step of the SIMEX procedure. These results can be combined by noting that extrapolation can be seen as a differentiable operator on the results from the Simulation step. At first we show how the results of CKLS can be transferred to the MC-SIMEX with known misclassification matrix Π. We show the derivation only for a possibly misclassified X, but the results also apply to the case of a possibly misclassified Y. Let the true value of the unknown parameter vector β be β_0. Estimating β_0 by maximum likelihood can be viewed as solving the estimating equation $\sum_{i=1}^{n} \psi(Y_i, X_i, Z_i, \beta) = 0$, if based on the true data. The calculation of the naive estimator with misclassification matrix Π implies solving the estimating equation

$$\sum_{i=1}^{n} \psi(Y_i, X_i^*, Z_i, \beta) = 0.$$

The limit of the naive estimator in the presence of misclassification Π for $n \to \infty$ is the solution of $E[\psi(Y, X^*, Z, \beta)] = 0$ and is denoted by $\beta^*(\Pi)$, as before. In the simulation step we calculate $\hat{\beta}_{b,\lambda_k}$ by solving $\sum_{i=1}^{n} \psi(Y_i, X_{b,i}^*(\lambda_k), Z_i, \beta) = 0$. Following CKLS we get an asymptotic expansion

$$\begin{aligned} \sqrt{n}\left[\hat{\beta}_{b,\lambda_k} - \beta^*\left(\Pi^{1+\lambda_k}\right)\right] &= -\mathcal{A}^{-1}\left[\Pi, \lambda_k, \beta^*\left(\Pi^{1+\lambda_k}\right)\right] && (4) \\ &\quad \times \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi\left[Y_i, X_{b,i}^*(\lambda_k), Z_i, \beta^*\left(\Pi^{1+\lambda_k}\right)\right] && (5) \end{aligned}$$

with

$$\mathcal{A}\left[\Pi, \lambda_k, \beta\right] := E\left[\frac{\partial}{\partial \beta}\psi\left(Y, MC\left[\Pi^{1+\lambda_k}\right](X), Z, \beta\right)\right].$$

Expansion (5) can be transferred to the mean $\hat{\beta}_{\lambda_k}$ in each simulation step and combining the simulation steps gives an asymptotic multivariate normal distribution for vec $\left[\hat{\beta}_{\lambda_k};\ k = 1, \ldots m\right]$ with mean vec $\left[\beta^*(\Pi^{\lambda_k+1});\ k = 1, \ldots m\right]$ and covariance matrix Σ, which is given by

$$\begin{aligned} \Sigma &= \mathcal{A}_{11}^{-1}\mathcal{C}_{11}\left(\mathcal{A}_{11}^{-1}\right)^T, && (6) \\ \mathcal{A}_{11} &= \text{diag}\left[\mathcal{A}\left[\Pi, \lambda_k, \beta^*\left(\Pi^{1+\lambda_k}\right)\right], k = 1, \ldots, m\right], && (7) \\ \mathcal{C}_{11} &= \text{cov}\left[\Psi_1\left(\Pi, \boldsymbol{\lambda}, \beta^*\left(\Pi^{1+\lambda_k}\right)\right)\right], && (8) \\ \Psi_i(\Pi, \boldsymbol{\lambda}, \beta) &:= \text{vec}\left[B^{-1}\sum_{b=1}^{B} \psi\left[Y_i, X_{b,i}^*(\lambda_k), Z_i, \beta\right], k = 1, \ldots, m\right], && (9) \end{aligned}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^T$.

Each of the above matrices can be estimated by their empirical counterparts.

For the extrapolation step the parameter vector Γ of the extrapolation function is estimated by least squares using the values $\hat{\beta}_{\lambda_k}$. The resulting estimate $\hat{\Gamma}$ is again asymptotically normal with covariance matrix

$$\Sigma(\hat{\Gamma}) \quad = \quad D(\Gamma)^{-1}\mathbf{s}(\Gamma)\Sigma\mathbf{s}(\Gamma)^T D(\Gamma)^{-1}, \tag{10}$$

with

$$\mathbf{s}(\Gamma) \quad = \quad \frac{\partial}{\partial\Gamma}\mathrm{vec}\left[\mathcal{G}(\lambda_1,\Gamma),\ldots,\mathcal{G}(\lambda_m,\Gamma)\right], \tag{11}$$

$$D(\Gamma) \quad = \quad \mathbf{s}(\Gamma)\mathbf{s}(\Gamma)^T, \tag{12}$$

where Γ is replaced by $\hat{\Gamma}$ in the above expression to calculate the covariance matrix in practice. The asymptotic variance of the SIMEX estimator $\hat{\beta}_{SIMEX} = \mathcal{G}(0,\hat{\Gamma})$ is derived by the delta method:

$$V_{asy}(\hat{\beta}_{SIMEX}) = \frac{\partial}{\partial\Gamma}\mathcal{G}(0,\Gamma)\Sigma(\Gamma)\frac{\partial}{\partial\Gamma}\mathcal{G}(0,\Gamma)^T.$$

For more details when $B = \infty$ or when the mean is replaced by the median, we refer to CKLS.

3.1 Estimated misclassification matrix

In most applications the misclassification matrix Π has to be estimated from validation data. We assume here that the validation data are entirely independent from the main study. For a possibly misclassified X a data set is available where for each subject the true value X_k and the possibly corrupted value X_k^* $(k = 1,\ldots,n_V)$ are recorded. The misclassification probabilities are estimated by

$$\hat{\pi}_{ij} = \frac{\sum_k I(X_k = j \text{ and } X_k^* = i)}{\sum_k I(X_k = j)}, \tag{13}$$

where $I(a) = 1$ if a is true and 0 otherwise. Let the vector of free parameters of the misclassification matrix be denoted by Π_V. For two categories, $\Pi_V = (\pi_{00},\pi_{11})$, i. e. the sensitivity and the specificity. The vector of estimates $\hat{\Pi}_V$ is also the solution of a system of estimating equations related to binomial experiments for each value of X. Combining these estimating equations with the estimating equations for $\mathrm{vec}\left[\hat{\beta}_{\lambda_k};\ k = 1,\ldots m\right]$ yields an asymptotic variance for the joint distribution of the vector $\left[\mathrm{vec}\left[\hat{\beta}_{\lambda_k};\ k = 1,\ldots m\right],\hat{\Pi}_V\right]$:

$$\Sigma = \frac{1}{n}\begin{pmatrix}\mathcal{A}_{11} & \mathcal{A}_{12}\\ 0 & \mathcal{A}_{22}\end{pmatrix}^{-1}\begin{pmatrix}\mathcal{C}_{11} & \mathcal{C}_{12}\\ \mathcal{C}_{12}^T & \mathcal{C}_{22}\end{pmatrix}\left(\begin{pmatrix}\mathcal{A}_{11} & \mathcal{A}_{12}\\ 0 & \mathcal{A}_{22}\end{pmatrix}^{-1}\right)^T. \tag{14}$$

Here $\mathcal{A}_{11}$ and $\mathcal{C}_{11}$ are given by (7) and (8), respectively. $\mathcal{C}_{12} = 0$, since the validation and the main data are independent. Further,

$$\begin{aligned}
\mathcal{A}_{12} &= n^{-1} \sum_{i=1}^{n} \left[E\left(\frac{\partial}{\partial \Pi_V} \Psi_i \left(\Pi, \boldsymbol{\lambda}, \beta^* \left(\Pi^{1+\lambda_k} \right) \right) \right) \right], \qquad (15) \\
\mathcal{A}_{22} &= -n^{-1} \mathrm{diag} \left(n_{V,j};\ \ j = 1, \ldots, \dim(\Pi_V) \right), \\
\mathcal{C}_{22} &= n^{-1} \mathrm{cov} \left[\mathrm{vec}(\hat{\Pi}_{V,j}\ n_{V,j}, j = 1, \ldots, \dim(\Pi_V) \right], \qquad (16)
\end{aligned}$$

where $n_{V,j}$ is the sample size for estimating the j-th component of Π_V in the validation study. The matrix $\mathcal{C}_{22}$ depends on how the validation study was set up. Examples are given in the appendix. The main practical difficulty is the handling of expression (15). Since the simulation process is a discrete process we use the case $B \to \infty$, i. e. we approximate $B^{-1} \sum_{b=1}^{B} \psi \left[Y_i, X^*_{b,i}(\lambda_k), Z_i, \beta^* \left(\Pi^{1+\lambda_k} \right) \right]$ by its expected value with respect to the simulation operation induced by Π^{λ_k}. Since this is a discrete process we get a finite sum for fixed i, which is differentiable. Then we are able to estimate the matrix $\mathcal{A}_{22}$ from our data. Note that the value of B can be chosen large enough, so the approximation performs well. The technical aspects including differentiation of Π^{λ} is given in the appendix, where we give detailed formulae for the case of logistic regression.

After estimating the matrix Σ we proceed with equation (10) using the corresponding submatrix for estimating the variance of the parameters.

4 Simulation study

4.1 Simulation study setup

The variance estimation for the MC-SIMEX procedure is validated by a simulation study. We focused on three different simulation setups:

A Logistic regression with a possibly misclassified binary response Y^* and a correctly measured binary or continuous covariate X.

B Logistic regression with a correctly measured binary response Y and a possibly misclassified binary covariate X^*, with and without a correctly measured continuous confounder Z.

C Logistic regression with a possibly misclassified binary response Y^* and a possibly misclassified binary covariate X^*, with and without a correctly measured confounder Z

We created the true binary covariate from a Bernoulli distribution with $P(X = 0) = P(X = 1) = 0.5$. The confounder was generated from a normal distribution with variance $\sigma^2 = 1$ and mean 0.5 for $X = 0$ and -0.5 for $X = 1$. For case A the continuous covariate is drawn from a standard normal distribution. The true response is generated as a random variable drawn from a Bernoulli

distribution with $P(Y=1) = 1/(1+\exp(-\beta_0 - \beta_X X))$ for case A, B and C. If a confounder was present we used $P(Y=1) = 1/(1+\exp(-\beta_0 - \beta_X X - \beta_Z Z))$ instead. The true coefficients are $\beta_0 = 0$ and $\beta_X = \beta_Z = 1$. The sample size for all simulations is 1000, which is justified by the fact that measurement error correction is usually done for large epidemiological studies.

We applied a misclassification operator on Y and X to obtain the misclassified variables Y^* and X^* respectively. We assumed rather high misclassification rates. Either the misclassification is symmetric with $\pi_{00} = \pi_{11} = 0.8$ or asymmetric with $\pi_{00} = 0.9, \pi_{11} = 0.7$. The misclassification matrix is estimated from a validation study with the sample sizes $n_{V,1} = n_{V,2} = 50$ and $n_{V,1} = n_{V,2} = 100$ respectively. The case of a known misclassification matrix is denoted by $n_{V,1} = n_{V,2} = \infty$.

The MC-SIMEX estimator is calculated for the quadratic and the log-linear extrapolation functions. The true estimator is calculated using the correctly measured data and the naive estimator using the possibly corrupted variables. The MC-SIMEX procedure was performed with $B = 100$, i. e. for each of the 1000 simulated data, $B = 100$ repetitions result in an extrapolated estimate. The variance was calculated with three different methods. The Cook and Stefanski jackknife method (SE_J) and our asymptotic methods without correction for the estimation of the misclassification matrix, hence the naive approach (SE_{AN}) and with correction (SE_A).

4.2 Results

For the performance of the variance estimation we have to compare the simulation standard deviation (SE) of each estimator: (a) under no misclassification (true model), (b) when misclassification is ignored (naive model) and (c) when corrected for misclassification (using two SIMEX models). We have denoted this standard error by SE_S for all three cases. We compared this standard error to the square root of the mean of the corresponding estimated variances (SE_{AN}, SE_A and SE_J). For case A (see Tables 1 and 2) the variance estimation (SE_{AN}, SE_A) works extremely well and is far superior to the jackknife estimator (SE_J), which is independent of the extrapolation function. Note that the variance estimation is biased downwards if the variability in the estimation in the validation study is not taken into account (SE_{AN} versus SE_A). For case B (see Tables 3 and 4) the corrected asymptotic variance estimation works very well but has a small tendency to produce outliers for the log–linear extrapolation function. The asymptotic variance estimate is always better than the jackknife estimate and therefore to be preferred. The variance estimates for a correctly measured covariate are nearly identical for all three methods of variance estimation. In the case of misclassification in both, the regressor and the response (case C), the asymptotic variance estimation works rather well, but suffers from the small tendency to produce outliers for the log-linear extrapolation function, but is in general much better than the jackknife variance estimation. For an additional confounder Z (Table 5) the variance estimation is not as good, but still far more accurate than the jackknife estimator. Although the misclassification in

Table 1: Simulation results (Case A): Logistic regression of a misclassified response Y^* on a binary covariate X. SE_S is the simulation standard error of the parameter estimates. SE_{AN} is the square root of the simulation mean of the asymptotic variance estimator without taking the validation study variance into account, SE_A the square root of the simulation mean of the asymptotic variance estimator taking the validation study variance into account. SE_J is the square root of the mean of the Cook and Stefanski Jackknife variance estimator.

	$(\pi_{00},\pi_{11})=(0.9,0.7)$, $(n_{V,1},n_{V,2})=(50,50)$					$(\pi_{00},\pi_{11})=(0.8,0.8)$, $(n_{V,1},n_{V,2})=(50,50)$				
	Mean	SE_S	SE_{AN}	SE_A	SE_J	Mean	SE_S	SE_{AN}	SE_A	SE_J
True Model	0.996	0.136				0.996	0.136			
Naive Model	0.563	0.131				0.570	0.132			
MC-SIMEX(Q)	0.848	0.212	0.197	0.211	0.179	0.899	0.227	0.208	0.225	0.186
MC-SIMEX(LOG)	0.888	0.241	0.206	0.243	0.179	0.971	0.271	0.226	0.272	0.186
	$(\pi_{00},\pi_{11})=(0.9,0.7)$, $(n_{V,1},n_{V,2})=(100,100)$					$(\pi_{00},\pi_{11})=(0.8,0.8)$, $(n_{V,1},n_{V,2})=(100,100)$				
MC-SIMEX(Q)	0.851	0.208	0.197	0.204	0.179	0.900	0.223	0.208	0.217	0.187
MC-SIMEX(LOG)	0.887	0.222	0.205	0.222	0.179	0.968	0.249	0.223	0.244	0.187
	$(\pi_{00},\pi_{11})=(0.9,0.7)$, $(n_{V,1},n_{V,2})=(\infty,\infty)$					$(\pi_{00},\pi_{11})=(0.8,0.8)$, $(n_{V,1},n_{V,2})=(\infty,\infty)$				
MC-SIMEX(Q)	0.849	0.200	0.197		0.180	0.901	0.210	0.208		0.186
MC-SIMEX(LOG)	0.886	0.207	0.203		0.180	0.962	0.226	0.221		0.186

the data was quite high, the correction of the parameter estimates is in all cases substantial and not to be ignored.

5 Application

5.1 The Signal-Tandmobiel® study

The Signal-Tandmobiel® study is a 6 year longitudinal oral health study conducted in Flanders (Belgium) and involving 4468 children. The children were examined annually for a period of six years (1996-2001). In this paper we look at the data of the first year of the study, hence the data of seven-year old children. Due to the practical organization of the sampling, the age of the children actually varied from 6.12 years to 8.09 years. Dental data was collected on e.g. tooth decay, presence of restorations, etc. together with data obtained from questionnaires filled in by the parents of the children on oral hygiene and dietary behavior. For a more detailed description of the Signal-Tandmobiel® study we

Table 2: Simulation results (Case A): Logistic regression of a misclassified response Y^* on a continuous covariate X. SE_S is the simulation standard error of the parameter estimates. SE_{AN} is the square root of the simulation mean of the asymptotic variance estimator without taking the validation study variance into account, SE_A the square root of the simulation mean of the asymptotic variance estimator taking the validation study variance into account. SE_J is the square root of the mean of the Cook and Stefanski Jackknife variance estimator.

	$(\pi_{00}, \pi_{11}) = (0.9, 0.7)$, $\left(n_{V,1}, n_{V,2}\right) = (50, 50)$					$(\pi_{00}, \pi_{11}) = (0.8, 0.8)$, $\left(n_{V,1}, n_{V,2}\right) = (50, 50)$				
	Mean	SE_S	SE_{AN}	SE_A	SE_J	Mean	SE_S	SE_{AN}	SE_A	SE_J
True Model	1.005	0.085				1.007	0.081			
Naive Model	0.555	0.074				0.531	0.071			
MC-SIMEX(Q)	0.852	0.146	0.119	0.143	0.104	0.854	0.145	0.122	0.144	0.104
MC-SIMEX(LOG)	0.899	0.184	0.125	0.184	0.104	0.925	0.194	0.132	0.196	0.104
	$(\pi_{00}, \pi_{11}) = (0.9, 0.7)$, $\left(n_{V,1}, n_{V,2}\right) = (100, 100)$					$(\pi_{00}, \pi_{11}) = (0.8, 0.8)$, $\left(n_{V,1}, n_{V,2}\right) = (100, 100)$				
MC-SIMEX(Q)	0.853	0.132	0.119	0.132	0.104	0.852	0.132	0.121	0.133	0.104
MC-SIMEX(LOG)	0.898	0.152	0.124	0.154	0.104	0.915	0.158	0.130	0.160	0.104
	$(\pi_{00}, \pi_{11}) = (0.9, 0.7)$, $\left(n_{V,1}, n_{V,2}\right) = (\infty, \infty)$					$(\pi_{00}, \pi_{11}) = (0.8, 0.8)$, $\left(n_{V,1}, n_{V,2}\right) = (\infty, \infty)$				
MC-SIMEX(Q)	0.852	0.123	0.119		0.104	0.854	0.121	0.121		0.104
MC-SIMEX(LOG)	0.893	0.127	0.123		0.104	0.911	0.128	0.128		0.104

Table 3: Simulation results (Case B): Logistic regression of a correctly measured response Y on a misclassified binary covariate X^*. SE_S is the simulation standard error of the parameter estimates. SE_{AN} is the square root of the simulation mean of the asymptotic variance estimator without taking the validation study variance into account, SE_A the square root of the simulation mean of the asymptotic variance estimator taking the validation study variance into account. SE_J is the square root of the mean of the Cook and Stefanski Jackknife variance estimator.

	$(\pi_{00}, \pi_{11}) = (0.9, 0.7)$, $\left(n_{V,1}, n_{V,2}\right) = (50, 50)$					$(\pi_{00}, \pi_{11}) = (0.8, 0.8)$, $\left(n_{V,1}, n_{V,2}\right) = (50, 50)$				
	Mean	SE_S	SE_{AN}	SE_A	SE_J	Mean	SE_S	SE_{AN}	SE_A	SE_J
True Model	0.996	0.136				1.000	0.140			
Naive Model	0.620	0.137				0.569	0.130			
MC-SIMEX(Q)	0.918	0.219	0.209	0.227	0.191	0.898	0.227	0.209	0.225	0.186
MC-SIMEX(LOG)	0.975	0.254	0.222	0.292*	0.191	0.973	0.270	0.225	0.272	0.186
	$(\pi_{00}, \pi_{11}) = (0.9, 0.7)$, $\left(n_{V,1}, n_{V,2}\right) = (100, 100)$					$(\pi_{00}, \pi_{11}) = (0.8, 0.8)$, $\left(n_{V,1}, n_{V,2}\right) = (100, 100)$				
MC-SIMEX(Q)	0.917	0.215	0.208	0.218	0.191	0.898	0.219	0.208	0.217	0.186
MC-SIMEX(LOG)	0.975	0.236	0.220	0.264*	0.191	0.964	0.247	0.223	0.245	0.186
	$(\pi_{00}, \pi_{11}) = (0.9, 0.7)$, $\left(n_{V,1}, n_{V,2}\right) = (\infty, \infty)$					$(\pi_{00}, \pi_{11}) = (0.8, 0.8)$, $\left(n_{V,1}, n_{V,2}\right) = (\infty, \infty)$				
MC-SIMEX(Q)	0.920	0.206	0.208		0.191	0.896	0.209	0.208		0.187
MC-SIMEX(LOG)	0.972	0.218	0.219		0.191	0.961	0.222	0.221		0.187

* due to outliers, the median is used instead of the mean.

Table 4: Simulation results (Case B): Logistic regression of a correctly measured response Y on a misclassified binary covariate X^* and a continuous confounder Z. SE_S is the simulation standard error of the parameter estimates. SE_{AN} is the square root of the simulation mean of the asymptotic variance estimator without taking the validation study variance into account, SE_A the square root of the simulation mean of the asymptotic variance estimator taking the validation study variance into account. SE_J is the square root of the mean of the Cook and Stefanski Jackknife variance estimator.

	$\beta_X,\ (\pi_{00},\pi_{11})=(0.9,0.7),$ $(n_{V,1},n_{V,2})=(50,50)$					$\beta_X,\ (\pi_{00},\pi_{11})=(0.8,0.8),$ $(n_{V,1},n_{V,2})=(50,50)$				
	Mean	SE_S	SE_{AN}	SE_A	SE_J	Mean	SE_S	SE_{AN}	SE_A	SE_J
True Model	1.007	0.162				1.007	0.162			
Naive Model	0.535	0.146				0.520	0.148			
MC-SIMEX(Q)	0.833	0.245	0.243	0.258	0.215	0.849	0.264	0.250	0.267	0.217
MC-SIMEX(LOG)	0.879	0.281	0.257	0.308*	0.215	0.920	0.313	0.273	0.299*	0.217
	$\beta_X,\ (\pi_{00},\pi_{11})=(0.9,0.7),$ $(n_{V,1},n_{V,2})=(100,100)$					$\beta_X,\ (\pi_{00},\pi_{11})=(0.8,0.8),$ $(n_{V,1},n_{V,2})=(100,100)$				
MC-SIMEX(Q)	0.826	0.240	0.241	0.248	0.215	0.844	0.254	0.249	0.257	0.217
MC-SIMEX(LOG)	0.866	0.262	0.251	0.280*	0.215	0.904	0.281	0.266	0.283*	0.217
	$\beta_X,\ (\pi_{00},\pi_{11})=(0.9,0.7),$ $(n_{V,1},n_{V,2})=(\infty,\infty)$					$\beta_X,\ (\pi_{00},\pi_{11})=(0.8,0.8),$ $(n_{V,1},n_{V,2})=(\infty,\infty)$				
MC-SIMEX(Q)	0.827	0.232	0.241		0.215	0.846	0.248	0.249		0.218
MC-SIMEX(LOG)	0.867	0.242	0.250		0.215	0.902	0.263	0.264		0.218
	$\beta_Z,\ (\pi_{00},\pi_{11})=(0.9,0.7),$ $(n_{V,1},n_{V,2})=(50,50)$					$\beta_Z,\ (\pi_{00},\pi_{11})=(0.8,0.8),$ $(n_{V,1},n_{V,2})=(50,50)$				
True Model	1.002	0.084				1.002	0.084			
Naive Model	0.842	0.075				0.841	0.076			
MC-SIMEX(Q)	0.919	0.089	0.089	0.092	0.082	0.921	0.092	0.090	0.094	0.082
MC-SIMEX(LOG)	0.867	0.079	0.079	0.079	0.082	0.866	0.080	0.080	0.080	0.082
	$\beta_Z,\ (\pi_{00},\pi_{11})=(0.9,0.7),$ $(n_{V,1},n_{V,2})=(100,100)$					$\beta_Z,\ (\pi_{00},\pi_{11})=(0.8,0.8),$ $(n_{V,1},n_{V,2})=(100,100)$				
MC-SIMEX(Q)	0.919	0.089	0.089	0.091	0.082	0.920	0.092	0.090	0.092	0.083
MC-SIMEX(LOG)	0.868	0.079	0.079	0.080	0.082	0.866	0.080	0.080	0.080	0.083
	$\beta_Z,\ (\pi_{00},\pi_{11})=(0.9,0.7),$ $(n_{V,1},n_{V,2})=(\infty,\infty)$					$\beta_Z,\ (\pi_{00},\pi_{11})=(0.8,0.8),$ $(n_{V,1},n_{V,2})=(\infty,\infty)$				
MC-SIMEX(Q)	0.919	0.088	0.089		0.082	0.921	0.091	0.090		0.083
MC-SIMEX(LOG)	0.868	0.079	0.080		0.082	0.866	0.080	0.080		0.083

* due to outliers, the median is used instead of the mean.

Table 5: Simulation results (Case C): Logistic regression of a misclassified response Y^* on a misclassified binary covariate X^*. SE_S is the simulation standard error of the parameter estimates. SE_{AN} is the square root of the simulation mean of the asymptotic variance estimator without taking the validation study variance into account, SE_A the square root of the simulation mean of the asymptotic variance estimator taking the validation study variance into account. SE_J is the square root of the mean of the Cook and Stefanski Jackknife variance estimator.

	$(\pi_{00}, \pi_{11}) = (0.9, 0.7)$, $\left(n_{V,1}, n_{V,2}\right) = (50, 50)$					$(\pi_{00}, \pi_{11}) = (0.8, 0.8)$, $\left(n_{V,1}, n_{V,2}\right) = (50, 50)$				
	Mean	SE_S	SE_{AN}	SE_A	SE_J	Mean	SE_S	SE_{AN}	SE_A	SE_J
True Model	1.006	0.134				1.003	0.140			
Naive Model	0.348	0.126				0.342	0.130			
MC-SIMEX(Q)	0.674	0.257	0.256	0.259	0.205	0.697	0.270	0.265	0.267	0.206
MC-SIMEX(LOG)	0.886	0.395	0.336	0.387*	0.205	0.971	0.437	0.375	0.454*	0.206
	$(\pi_{00}, \pi_{11}) = (0.9, 0.7)$, $\left(n_{V,1}, n_{V,2}\right) = (100, 100)$					$(\pi_{00}, \pi_{11}) = (0.8, 0.8)$, $\left(n_{V,1}, n_{V,2}\right) = (100, 100)$				
MC-SIMEX(Q)	0.675	0.249	0.255	0.256	0.206	0.702	0.272	0.266	0.267	0.207
MC-SIMEX(LOG)	0.873	0.348	0.327	0.349*	0.206	0.983	0.413	0.374	0.405*	0.207
	$(\pi_{00}, \pi_{11}) = (0.9, 0.7)$, $\left(n_{V,1}, n_{V,2}\right) = (\infty, \infty)$					$(\pi_{00}, \pi_{11}) = (0.8, 0.8)$, $\left(n_{V,1}, n_{V,2}\right) = (\infty, \infty)$				
MC-SIMEX(Q)	0.676	0.246	0.255		0.205	0.699	0.269	0.266		0.207
MC-SIMEX(LOG)	0.865	0.322	0.322		0.205	0.962	0.373	0.363		0.207

* due to outliers, the median is used instead of the mean.

Table 6: Simulation results (Case C): Logistic regression of a misclassified response Y^* on a misclassified binary covariate X^* and a continuous confounder Z. SE_S is the simulation standard error of the parameter estimates. SE_{AN} is the square root of the simulation mean of the asymptotic variance estimator without taking the validation study variance into account, SE_A the square root of the simulation mean of the asymptotic variance estimator taking the validation study variance into account. SE_J is the square root of the mean of the Cook and Stefanski Jackknife variance estimator.

	$\beta_X,\ (\pi_{00},\pi_{11})=(0.9,0.7),$ $(n_{V,1},n_{V,2})=(50,50)$					$\beta_X,\ (\pi_{00},\pi_{11})=(0.8,0.8),$ $(n_{V,1},n_{V,2})=(50,50)$				
	Mean	SE_S	SE_{AN}	SE_A	SE_J	Mean	SE_S	SE_{AN}	SE_A	SE_J
True Model	0.996	0.170				0.997	0.170			
Naive Model	0.276	0.138				0.271	0.132			
MC-SIMEX(Q)	0.553	0.280	0.284	0.285	0.223	0.570	0.284	0.292	0.293	0.222
MC-SIMEX(LOG)	0.738	0.421	0.371	0.420*	0.223	0.816	0.478	0.416	0.467*	0.222
	$\beta_X,\ (\pi_{00},\pi_{11})=(0.9,0.7),$ $(n_{V,1},n_{V,2})=(100,100)$					$\beta_X,\ (\pi_{00},\pi_{11})=(0.8,0.8),$ $(n_{V,1},n_{V,2})=(100,100)$				
MC-SIMEX(Q)	0.553	0.284	0.284	0.284	0.223	0.570	0.284	0.293	0.293	0.222
MC-SIMEX(LOG)	0.743	0.406	0.368	0.403	0.223	0.821	0.445	0.414	0.450	0.222
	$\beta_X,\ (\pi_{00},\pi_{11})=(0.9,0.7),$ $(n_{V,1},n_{V,2})=(\infty,\infty)$					$\beta_X,\ (\pi_{00},\pi_{11})=(0.8,0.8),$ $(n_{V,1},n_{V,2})=(\infty,\infty)$				
MC-SIMEX(Q)	0.555	0.281	0.284		0.225	0.573	0.284	0.293		0.224
MC-SIMEX(LOG)	0.733	0.375	0.361		0.225	0.817	0.416	0.408		0.224
	$\beta_Z,\ (\pi_{00},\pi_{11})=(0.9,0.7),$ $(n_{V,1},n_{V,2})=(50,50)$					$\beta_Z,\ (\pi_{00},\pi_{11})=(0.8,0.8),$ $(n_{V,1},n_{V,2})=(50,50)$				
True Model	1.005	0.081				1.005	0.081			
Naive Model	0.441	0.065				0.443	0.064			
MC-SIMEX(Q)	0.707	0.122	0.110	0.111	0.093	0.740	0.129	0.115	0.116	0.096
MC-SIMEX(LOG)	0.739	0.156	0.115	0.123	0.093	0.799	0.177	0.126	0.137	0.096
	$\beta_Z,\ (\pi_{00},\pi_{11})=(0.9,0.7),$ $(n_{V,1},n_{V,2})=(100,100)$					$\beta_Z,\ (\pi_{00},\pi_{11})=(0.8,0.8),$ $(n_{V,1},n_{V,2})=(100,100)$				
MC-SIMEX(Q)	0.707	0.120	0.110	0.111	0.093	0.740	0.122	0.115	0.115	0.096
MC-SIMEX(LOG)	0.736	0.137	0.114	0.117	0.093	0.794	0.149	0.124	0.129	0.096
	$\beta_Z,\ (\pi_{00},\pi_{11})=(0.9,0.7),$ $(n_{V,1},n_{V,2})=(\infty,\infty)$					$\beta_Z,\ (\pi_{00},\pi_{11})=(0.8,0.8),$ $(n_{V,1},n_{V,2})=(\infty,\infty)$				
MC-SIMEX(Q)	0.707	0.110	0.110		0.094	0.740	0.113	0.115		0.097
MC-SIMEX(LOG)	0.732	0.113	0.113		0.094	0.790	0.120	0.123		0.097

* due to outliers, the median is used instead of the mean.

refer to Vanobbergen et al. (2000).

In this application the response of interest is caries experience on tooth level, a binary variable equal to 1 if the tooth is decayed, missing due to caries or filled, and 0 otherwise and on subject level, i. e. whether there was caries in the seven-year old child or not. More specifically, we will look at two research questions. In the first research question we examine the relation between caries experience on subject level with the various dietary and brushing behavior variables. In the second research question we examine the relationship between caries experience on a permanent molar and caries experience on an adjacent deciduous molar. In the first research question we assume that only the response is subject to misclassification, i. e. we assume that the brushing and dietary variables were correctly reported. For the second research question, both the response as well as the caries regressor were subject to misclassification.

During the study period three calibration exercises involving 92, 32 and 24 children, respectively were devoted to the scoring of caries experience. At the end of each of the three calibration exercises the sensitivity and specificity of sixteen dental examiners vis-a-vis a gold standard (Dominique Declerck) was determined. For the first research question we based the misclassification probabilities on the first calibration exercise. Further, we lumped together the misclassification matrices of the sixteen dental examiners. Finally, we ignored the fact that some children were examined by two or more dental examiners. For the second research question we based the misclassification table for the deciduous molar on the first calibration exercise, while the data from the second calibration exercise was used for the permanent molar.

5.2 Research Question 1

The first research question evaluates the impact of dietary and brushing behavior on caries experience in the mouth. More specifically, as regressors we considered: age (years), gender (girl = 1), age at start of brushing (years), use of systematic fluoride supplements (regular use = 1), daily consumption of sugar containing drinks between meals (yes = 1), intake of in-between-meals (= 1 if greater than 2, 0 otherwise), and frequency of brushing (= 1 if less than twice a day, 0 otherwise). Because a reasonable portion of the parents did not fill in the questionnaires the analysis below is based on only 3303 children.

Table 8 shows the naive regression estimates of the logistic regression model, i. e. without taking into account the misclassification of the response. Further, the overall misclassification matrix of the examiners vis-a-vis the gold standard at the first calibration exercise is given in Table 7. As can be seen it is based on 142 evaluations.

Using the MC-SIMEX approach we corrected for possible misclassification of the response. The result of the correction mechanism is also shown in Table 8. Therein, we give two standard errors, the first based on the approach suggested above. The second is based on the adaptation of the approach of Stefanski and Cook (1995) to misclassification as explained in Küchenhoff et al. (2005).

Except for gender and brushing frequency all regressors are significant (at

Table 7: Research question 1: Table of misclassified teeth (and misclassification probabilities in percentage) for caries experience at subject level, column= gold standard, row= (pool of) dental examiner(s)

	Y	
Y^*	0	1
0	44 (89.8%)	5 (5.4%)
1	5 (11.2%)	88 (94.6%)

Table 8: Research question 1: Maximum likelihood estimates of the regressors of the naive and MC-SIMEX corrected (for misclassification of the response) logistic regression model. The log-linear extrapolation function is used. For the corrected model two standard errors are given: SE_J is based on the approach of Stefanski and Cook (1995), SE_A is based on the approach suggested above.

	Naive Estimates	MC-SIMEX Corrected Estimates		
Parameter	Estimate(SE)	Estimate	SE_A	SE_J
Intercept	−0.355 (0.118)	−0.474	0.135	0.140
Gender (girl)	0.026 (0.072)	0.037	0.083	0.085
Age (years)	0.302 (0.087)	0.361	0.108	0.103
Brushing frequency (< 2)	0.196 (0.107)	0.243	0.128	0.135
Age start brushing (years)	0.180 (0.034)	0.214	0.043	0.041
Fluoride supplement (yes)	−0.448 (0.072)	−0.496	0.081	0.087
Sugary drinks (yes)	0.318 (0.074)	0.385	0.091	0.086
Between meals (> 2)	0.238 (0.078)	0.293	0.095	0.088

$\alpha = 0.05$), without or with correction. All MC-SIMEX corrected estimates of the regression coefficients are larger in absolute size than their uncorrected version. As can be seen in Table 8, the two variance estimation methods give roughly the same results. Of course, in practice, it could be of importance sometimes which method is used, certainly if significance is interpreted strictly.

5.3 Research Question 2

In the second research question we look at the impact of the caries status of the deciduous molar 55 in the first year of the Signal-Tandmobiel® study on the caries status of the adjacent permanent molar 16 in the last year of the study, taking into account that for both teeth the caries status is possibly misclassified. Note that we used the European notation for the position of the teeth in the mouth. The deciduous teeth in the upper right (for the child) position are denoted by '5x'. The incisors in this quadrant are denoted by '51' and '52', the single canine by '53'. The two deciduous molars are denoted by '54' and '55' and are also called the deciduous first and second molar, respectively. The permanent teeth in the same quadrant and occupying the same positions as the deciduous teeth described above are denoted by '11', '12', '13', '14' and '15'. The molar '16' is called the six-year (permanent) molar, because it starts to emerge at the age of six. This molar is located next to deciduous molar '55'. Hence, caries problems on the deciduous molar '55' can and probably will have an effect on caries problems on the neighboring permanent molar. To establish this effect, we used a logistic regression model with regressors the caries status of the deciduous molar 55 and additionally age (years), gender (girl = 1). We did not include any brushing or dietary behavior variables because they could partially mask the relationship between the two caries experiences. Related analyses on this relationship, but without taking into account misclassification error, can be found in Leroy et al. (2005) and Komárek and Lesaffre (2006).

Now there are two misclassification matrices: one for the response (from the last calibration exercise based on 148 evaluations) and one for the caries regressor (from the first calibration exercise based on 134 evaluations), these are shown in Table 9.

Table 10 shows the naive regression estimates of the logistic regression model. Further, we applied three MC-SIMEX corrections to the model. The first correcting for misclassification of the caries regressor, the second for misclassification of the response and the third for misclassification of both the regressor and the response.

Again the MC-SIMEX corrected estimate of the regression coefficients are larger in absolute value than the uncorrected estimates. Gender is again not significant. The regression coefficient of the binary variable 'deciduous molar 55' (caries=1, no caries=0) is positive and significant. This implies that caries on the deciduous tooth increases the likelihood of caries at the neighboring permanent tooth. The interaction term between the binary tooth regressor and age was not significant in any of the analyses (uncorrected or corrected) and hence was not considered any further. Note that the estimated standard

Table 9: Research question 2: Table of misclassified teeth (and misclassification probabilities in percentage) on caries regressor (left) and on response (right), column= gold standard, row= (pool of) dental examiner(s)

	X regressor			Y response	
X^*	0	1		0	1
0	92 (98.9%)	5 (12.2%)		118 (96.7%)	4 (15.4%)
1	1 (1.1%)	36 (87.8%)		4 (3.3%)	22 (84.6%)

Table 10: Research question 2: Maximum likelihood estimates of the regressors of the naive and MC-SIMEX corrected (for misclassification of the response, regressor and both) logistic regression model. The log-linear extrapolation function is used. For the corrected model two standard errors are given: SE_J is based on the approach of Stefanski and Cook (1995), SE_A is based on the approach suggested above.

	Naive Estimates	MC-SIMEX Corrected Estimates		
		Corrected for response		
Parameter	Estimate(SE)	Estimate	SE_A	SE_J
Intercept	1.178 (0.723)	2.111	0.890	0.863
Gender (girl)	0.123 (0.078)	0.152	0.093	0.094
Age (years)	−0.366 (0.101)	−0.424	0.115	0.122
Deciduous molar 55	1.203 (0.080)	1.419	0.130	0.095
		Corrected for regressor		
Parameter	Estimate(SE)	Estimate	SE_A	SE_J
Intercept	1.178 (0.723)	1.085	0.732	0.741
Gender (girl)	0.123 (0.078)	0.134	0.078	0.080
Age (years)	−0.366 (0.101)	−0.367	0.104	0.104
Deciduous molar 55	1.203 (0.080)	1.378	0.130	0.093
		Corrected for both		
Parameter	Estimate(SE)	Estimate	SE_A	SE_J
Intercept	1.178 (0.723)	1.610	0.877	0.913
Gender (girl)	0.123 (0.078)	0.165	0.092	0.096
Age (years)	−0.366 (0.101)	−0.407	0.115	0.128
Deciduous molar 55	1.203 (0.080)	1.643	0.113	0.108

errors SE_A are higher than SE_J for the effect of the 'deciduous molar 55' in all models. This makes sense, since SE_J does not take the variability induced by the validation study into account. The negative and significant coefficient of age seems at first sight surprising, but age is recorded at the first visit. Thus, when the deciduous molar is examined at older age the time at risk for the permanent tooth is on average lower than when examined at younger age. Remember that the age at the first visit varies from 6 to 8 years.

For the three different kinds of correction it is easily seen, that for sole correction of the misclassification of the response (part 1 of Table 10) the correction is visible for all variables, whereas the correction for the misclassification of the regressor (part 2 of Table 10) changes the parameter estimate for the regressor quite strongly and lets the other parameter estimates stay nearly unchanged. The correction for both, response and regressor (part 3 of Table 10), is more or less the addition of the sole corrections.

6 Discussion

The asymptotic approach suggested here has a better theoretical foundation than the method of Stefanski and Cook (1995) and has shown good results in the simulation study. Further, it can be used when the misclassification probabilities are estimated from a validation study, which will be the case in many practical situations. Furthermore, the method can be applied to all regression models where the estimation is done by estimating equations, at least in principle, but the approach necessitates the calculation of derivatives which can become complicated or not feasible. In our example and in the simulation study we have used the rather complicated case of misclassification in the regressor and in the response simultaneously. For these kinds of studies other methods as maximum likelihood or the matrix method are not feasible.

We have developed an package for the statistical computing environment R developed by the R Development Core Team (2005), which features an implementation of the SIMEX method by Cook and Stefanski (1994) and MC-SIMEX method by Küchenhoff et al. (2005) including naive asymptotic and jackknife variance estimation, downloadable from
http://cran.r-project.org/src/contrib/Descriptions/simex.html

Acknowledgements

The third author acknowledges the partial support from the Interuniversity Attraction Poles Program P5/24 − Belgian State − Federal Office for Scientific, Technical and Cultural Affairs, and the partial support from the Research Grant OT/05/60, Catholic University Leuven.

Data collection of the Signal-Tandmobiel® study was supported by Unilever, Belgium. The Signal-Tandmobiel® project comprises the following partners: D. Declerck (Dental School, Catholic University Leuven), L. Martens (Dental

School, University Ghent), J. Vanobbergen (Oral Health Promotion and Prevention, Flemish Dental Association), P. Bottenberg (Dental School, University Brussels), E. Lesaffre (Biostatistical Centre, Catholic University Leuven) and K. Hoppenbrouwers (Youth Health Department, Catholic University Leuven; Flemish Association for Youth Health Care).

A Detailed formulae for logistic regression

For logistic regression, the estimation equation ψ has the form

$$\psi(Y_i, X_i, \beta) = \left\{Y_i - H\left(\beta_0 + \beta_X^T X_i\right)\right\} \begin{pmatrix} 1 \\ X_i \end{pmatrix},$$

with $H()$ being the logistic function. Using $H'(\cdot) = (1-H)H$, the derivation has the form:

$$\left(\frac{\partial}{\partial \beta^T}\right) \psi_i\left(Y_i, X_i, \beta\right) = -(1 - H(\beta_0 + \beta_X^T X_i))H(\beta_0 + \beta_X^T X_i) \begin{pmatrix} 1 \\ X_i \end{pmatrix} (1, X_i)$$

which is needed to estimate $\mathcal{A}_{11}$ from equation (7).

The matrix $\mathbf{s}(\Gamma)$ from equation (12) has for quadratic extrapolation the form

$$\mathbf{s}(\Gamma) = \begin{pmatrix} \begin{pmatrix} -1 & 0\cdots 0 \\ -\lambda_1 & 0\cdots 0 \\ -\lambda_1^2 & 0\cdots 0 \end{pmatrix} & \cdots & \begin{pmatrix} -1 & 0\cdots 0 \\ -\lambda_m & 0\cdots 0 \\ -\lambda_m^2 & 0\cdots 0 \end{pmatrix} \\ \vdots & \ddots & \vdots \\ \begin{pmatrix} 0\cdots 0 & -1 \\ 0\cdots 0 & -\lambda_1 \\ 0\cdots 0 & -\lambda_1^2 \end{pmatrix} & \cdots & \begin{pmatrix} 0\cdots 0 & -1 \\ 0\cdots 0 & -\lambda_m \\ 0\cdots 0 & -\lambda_m^2 \end{pmatrix} \end{pmatrix}$$

and for log-linear extrapolation

$$\mathbf{s}(\Gamma) = \begin{pmatrix} \begin{pmatrix} -\exp(\gamma_{01} + \gamma_{11}\lambda_1) & 0\cdots 0 \\ -\exp(\gamma_{01} + \gamma_{11}\lambda_1)\lambda_1 & 0\cdots 0 \end{pmatrix} & \cdots & \begin{pmatrix} -\exp(\gamma_{01} + \gamma_{11}\lambda_m) & 0\cdots 0 \\ -\exp(\gamma_{01} + \gamma_{11}\lambda_m)\lambda_m & 0\cdots 0 \end{pmatrix} \\ \vdots & \ddots & \vdots \\ \begin{pmatrix} 0\cdots 0 & -\exp(\gamma_{0p} + \gamma_{1p}\lambda_1) \\ 0\cdots 0 & -\exp(\gamma_{0p} + \gamma_{1p}\lambda_1)\lambda_1 \end{pmatrix} & \cdots & \begin{pmatrix} 0\cdots 0 & -\exp(\gamma_{0p} + \gamma_{1p}\lambda_m) \\ 0\cdots 0 & -\exp(\gamma_{0p} + \gamma_{1p}\lambda_m)\lambda_m \end{pmatrix} \end{pmatrix}.$$

where p is the number of parameters.

The derivative of the function $\mathcal{G}(0, \Gamma)$, which is needed for the application of the Δ–method, has for the quadratic extrapolation function the following form.

$$\frac{\partial}{\partial \Gamma}\mathcal{G}_Q(0, \Gamma) = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 \\ 0 & \cdots & 0 & -1 \\ 0 & \cdots & 0 & 1 \end{pmatrix}$$

For log-linear extrapolation it looks like

$$\frac{\partial}{\partial \Gamma}\mathcal{G}_{LOG}(0, \Gamma) = \begin{pmatrix} \exp\left(\gamma_{01} - \gamma_{11}\right) & 0 & \cdots & 0 \\ -\exp\left(\gamma_{01} - \gamma_{11}\right) & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \exp\left(\gamma_{0p} - \gamma_{1p}\right) \\ 0 & \cdots & 0 & -\exp\left(\gamma_{0p} - \gamma_{1p}\right) \end{pmatrix}.$$

A.1 Estimated misclassification Matrix

For a binary misclassified variable the matrix $\mathcal{C}_{22}$ has the form

$$\mathcal{C}_{22} = \text{diag}\left(\hat{\pi}_{00}(1-\hat{\pi}_{00})\frac{n_{V,0}}{n}, \hat{\pi}_{11}(1-\hat{\pi}_{11})\frac{n_{V,1}}{n}\right) \tag{17}$$

where $n_{V,j}$ is the sample size of the validation study for the case $Y = j$ or $X = j$, and n the sample size of the data set.

In the case of misclassified response the estimation equation Ψ is given by

$$\Psi(\cdot) = (\text{MC}[\Pi^{\lambda}](y_i) - H(x_i\beta))x_i^T$$

with x_i being the i-th row of the design matrix and

$$\Pi^{\lambda} = \frac{1}{1-\delta}\begin{pmatrix} 1-\pi_{11}+(1-\pi_{00})\delta^{\lambda} & (1-\pi_{11})(1-\delta^{\lambda}) \\ (1-\pi_{00})(1-\delta^{\lambda}) & 1-\pi_{00}+(1-\pi_{11})\delta^{\lambda} \end{pmatrix} \tag{18}$$

where $\delta = \det(\Pi) = \pi_{00}+\pi_{11}-1$. To estimate $\mathcal{A}_{12}$ the derivatives $(\partial\Pi^{\lambda}/\partial\Pi_V)\Psi(\cdot)$ are needed. For $B \to \infty$, fixed λ and a binary misclassified response, $\Psi(\cdot)$ can be written as

$$\begin{aligned}\Psi(\cdot) = &(\pi_{00}^{\lambda}(0-H(x_i\beta))x_i^T + (1-\pi_{00}^{\lambda})(1-H(x_i\beta))x_i^T)I_{y_i=0} \\ &+ (\pi_{11}^{\lambda}(1-H(x_i\beta))x_i^T + (1-\pi_{11}^{\lambda})(0-H(x_i\beta))x_i^T)I_{y_i=1}.\end{aligned} \tag{19}$$

with I being the indicator function

$$I_{y=a}\begin{cases} 1 & : \quad y = a \\ 0 & : \quad y \neq a \end{cases}$$

and x_i row i of the design matrix, e.g. for a simple linear model with intercept $(1, x_i, z_i^T)$.

The derivatives of Π^{λ} are needed:

$$\frac{\Pi_{00}^{\lambda}}{\partial\pi_{00}} = \frac{(\lambda(\pi_{00}-1)(\pi_{00}+\pi_{11}-2)\delta^{\lambda-1} + (\delta^{\lambda}-1)(\pi_{11}-1))}{(\pi_{00}+\pi_{11}-2)^2} \tag{20}$$

$$\frac{\Pi_{00}^{\lambda}}{\partial\pi_{11}} = \frac{\left(\lambda\left(\pi_{00}+\pi_{11}-2\right)\delta^{\lambda-1}-\delta^{\lambda}+1\right)(\pi_{00}-1)}{\left(\pi_{00}+\pi_{11}-2\right)^2} \tag{21}$$

$$\frac{\Pi_{11}^{\lambda}}{\partial\pi_{11}} = \frac{(\lambda(\pi_{11}-1)(\pi_{00}+\pi_{11}-2)\delta^{\lambda-1} + (\pi_{00}-1)(\delta^{\lambda}-1))}{(\pi_{00}+\pi_{11}-2)^2} \tag{22}$$

$$\frac{\Pi_{11}^{\lambda}}{\partial\pi_{00}} = \frac{(\pi_{11}-1)\left(\lambda\left(\pi_{00}+\pi_{11}-2\right)\delta^{\lambda-1}-\delta^{\lambda}+1\right)}{\left(\pi_{00}+\pi_{11}-2\right)^2} \tag{23}$$

This leads to an estimate for $\mathcal{A}_{12}$ for fixed λ

$$\mathcal{A}_{12,\lambda} = \left(n^{-1}\sum_{i=1}^{n}\frac{\partial\Pi_{11}^{\lambda}}{\partial\pi_{00}}x_iI_{y_i=1} - \frac{\partial\Pi_{00}^{\lambda}}{\partial\pi_{00}}x_iI_{y_i=0},\ n^{-1}\sum_{i=1}^{n}\frac{\partial\Pi_{11}^{\lambda}}{\partial\pi_{11}}x_iI_{y_i=1} - \frac{\partial\Pi_{00}^{\lambda}}{\partial\pi_{11}}x_iI_{y_i=0}\right) \tag{24}$$

References

Carroll, R., H. Küchenhoff, 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.

Cook, J. and L. Stefanski (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association 89*, 1314–1328.

Godambe, V. P. (1991). *Estimating Functions*. Clarendon Press, Oxford.

Gustafson, P. (2004). *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. Chapman & Hall, New York.

Komárek, A. and E. Lesaffre (2006). Bayesian semiparametric accelerated failure time model for paired doubly-interval-censored data. *accepted in Statistical Modelling 6*, 000–000.

Küchenhoff, H., S. Mwalili, and E. Lesaffre (2005). A general method for dealing with misclassification in regression: the misclassification simex. *Biometrics 61*, to appear.

Leroy, R., K. Bogaerts, E. Lesaffre, and D. Declerck (2005). Effect of caries experience in primary molars on cavity formation in the adjacent permanent first molar. *Caries Research 39*, 342–349.

R Development Core Team (2005). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Stefanski, L. and J. Cook (1995). Simulation-extrapolation: the measurement error jackknife. *Journal of the American Statistical Association 90*, 1247–1256.

Vanobbergen, J., L. Martens, E. Lesaffre, and D. Declerck (2000). The Signal-Tandmobiel® project – a longitudinal intervention health promotion study in Flanders (Belgium): baseline and first year results. *European Journal of Paediatric Dentistry 2*, 87–96.