International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Logistic Regression with Correction for HIV Self Reporting

Dorothy Nyakerario Riang'a[1], Samuel Musili Mwalili[2], Humphreys Murray[3]

[1]Department of Statistics and Actuarial Science, Jomo Kenyatta University of Agriculture and Technology (JKUAT), Nairobi, Kenya

Abstract: ***Self report is one of the widely used methods of collecting information regarding individuals' health status. However it is surprising that this issue has not received, in relative terms more sustained attention. In these surveys patients might simply be mistaken, misremember or exaggerate the material covered. Thus reliability of self-reported is tenuous and the exposure assessments on which the associations between exposures and disease occurrence rely on are subject to either measurement error in a quantitative variable or misclassification in a categorical variable. Relatively few methods are available to handle misclassified categorical exposure variable(s) in the context of logistic regression models. The statistical model for characterizing misclassification is given by the transition matrix \lambda from the true to the observed variable. In our research we aim at correcting the self reported data using the actual biomedical status of a sample of individuals from the given population. We exploit the relationship between the size of misclassification and bias in estimating the parameters of interest using logistic regression of self-report, the corrected logistic regression and the misclassified SIMEX (Simulation Extrapolation). We show that these methods are quite general and applicable to models with misclassified response and/or misclassified discrete regressors. We apply our methods to a study on the Kenya AIDS Indicator Survey data with a misclassified (Self-reported) longitudinal response.***

Keywords: SIMEX,MC-SIMEX, Misclassification Error, Logistic Regression

1. Introduction

Self report is one of the most widely used methods of collecting information regarding individuals' health status. It remains the field's most commonly used mode of assessment-by far, Paulhus and vazire (2007). Despite its popularity and demonstrated utility, the self-report method has been a frequent target of criticism from the early days of psychological assessment, Allport (1927) right up to the present Dunning et al (2005).

Examination of these processes requires burrowing deep into the affective and cognitive substrata of personality like motives in self perception Robins and John (1997). Thus reliability of self-reported data is tenuous resulting in measurement error (ME) in the data obtained through this method. Measurement error (ME) is common in biomedical and epidemiologic research.

When an exposure variable (or covariate) is analyzed as a categorical variable, the ME is generally referred to as 'misclassification'. Currently, a number of methods have been developed to handle different types of MEs, study designs and statistical or data settings. Some of these methods developed from fundamentally different formulations or paradigms, while others are major or minor extensions of existent methods.

A direct correction for misclassification bias is available for simple models by the matrix method; Kuha et al (2001). Misclassification in the response has been treated by Neuhaus (1999, 2002).Ordinal regression with misclassification in the response and with validation data has been handled by Mwalili et al (2005)

In this research we use a sample of the actual biomedical status to correct misclassification error in the self reported data. We review and compare available methods by analysing data from the Kenya AIDS Indicator Survey (KAIS), (2012), that could handle misclassified binary exposure variable(s) in the logistic regression model. We select three fundamentally different but practical methods:

1) Logistic regression of self-report,
2) Corrected logistic regression and
3) Misclassification SIMEX.

We apply the SIMEX idea to the case of misclassification. We show that it is a very general method and can be applied to misclassification of the response, of the discrete regressors, or to both.

2. Review of Previous Studies

Measurement error is pervasive in medical research. Carroll et al (2006), discussed examples arising from nutrition research, in which nutrition intake instruments (24-hour recall or food frequency questionnaires) are well known to be error prone; coronary kidney disease, in which an estimated glomerular filtration rate is often substituted for a genuine laboratory measurement; and pollution exposure studies, in which particulate concentrations at specified locations are used as surrogates for personal exposure. There is a large body of work addressing CD4 counts as noisy predictors for AIDS onset or progression, and, in an analogous context, Lin et al (2000), handle prostate-specific antigen levels as noisy predictors for prostate cancer onset.

Applications such as these have motivated statistical methods for handling measurement error in a wide variety of models. In their research they addressed an application arising from oral health research that leads to a clustered survival outcome (subject to either left or right censoring) and a discrete covariate measured with error. The approach combines the clustered survival measurement error models of Li and Lin (2000) with the misclassification covariate error of a general method

for dealing with misclassification in regression: the misclassification SIMEX.

Most currently available methods are suited for handling continuous covariate(s) in generalized linear models (GLMs) (e.g., linear or logistic regression) Freedman et al (2008); Messer and Natarajan (2008) while there have been fewer developments for applications with categorical covariate and/or censored outcome data. Heejung et al (1998) in their study reviewed and compared available methods by simulation and data analysis that could handle misclassified binary exposure variable(s) in the Cox proportional hazards regression model Cox (1972).

Multiple imputation (MI) was originally developed to solve missing data problems in statistics Little and Rubin (2002); Rubin (1976). Yet, considerable similarities in missing and mismeasured data have been noted and some methods can handle these two types of incomplete data together. The use of MI has been suggested as a bias correction method for a binary covariate subject to misclassification in the Cox model, Cole et al (2006).

3. Methodology

3.1. Logistic Regression of Self-report

Let S denote the self reported status given by

$$S = \begin{cases} \mathbf{1}\ if\ condition\ present \\ \\ \mathbf{0}\ if\ condition\ absent \end{cases} \quad (1)$$

The purpose of this is to regress $(S_i)\ for\ i = 1, 2, \dots, n$, where S_i is the self reported status outcome for ith subject with a dimensional vector covariate $\underline{x}_{ij} = (x_{i1}, x_{i2}, \dots\ x_{ip})$
$pr(S_i = 1/\underline{x}_{ij})$ Where $j = 1, \dots, p$ hence $\pi_i = g^{-1}(x_i^{'}, \beta)$, under the assumption that the outcomes S_i are independent given these probabilities.

We refer to $x^{'}\beta$ as the linear predictor where g^{-1} is the logit link function which yields the logistic random effects regression model $g(\theta) = \log(\frac{\theta}{1-\theta})$ and $g^{-1}(\theta) = \frac{e^{\theta}}{1+e^{\theta}}$, which transforms continuous values to range (0,1).We prefer to work with $logit^{-1}$ because it is natural to focus on mapping from the linear predictor of probabilities, rather than the reverse.

The log-likelihood

$$\iota(\beta) = \prod_{i=1}^{n} \pi_i^{S_i} (1 - \pi_i)^{1-S_i} \quad (2)$$

Therefore the likelihood is given by:

$$\iota(\beta) = \sum_{i=1}^{n} S_i \log \pi_i + \sum_{i=1}^{n} (1 - S_i) \log(1 - \pi_i) \quad (3)$$

3.2 Corrected Logistic Regression

Let S denoted self reported status subject to misclassification and let Y be the true underlying response variable. The misclassification process is expressed by misclassification probabilities

$$pr(S = j\ /\ Y = k) = \lambda_{jk} \qquad j, k = 0, 1 \quad (4)$$

Therefore
$\lambda_{00} = pr(S = 0\ /\ Y = 0)$,
$\lambda_{10} = pr(S = 1\ /\ Y = 0) = 1 - \lambda_{00}$,
$\lambda_{01} = pr(S = 0/\ Y = 1) = 1 - \lambda_{11}$,
$\lambda_{11} = pr(S = 1\ /\ Y = 1)$ .Resulting in misclassification matrix.

$$\lambda = \begin{pmatrix} \lambda_{00} & \lambda_{01} \\ \lambda_{10} & \lambda_{11} \end{pmatrix} = \begin{pmatrix} \lambda_{00} & 1 - \lambda_{11} \\ 1 - \lambda_{00} & \lambda_{11} \end{pmatrix}, \quad (5)$$

In an ideal situation when S=Y then it is referred to as perfect classification. Under non-differential misclassification, the effect of misclassification is described by sensitivity and specificity of S as a proxy measure of Y.

We will assume that $\lambda_{00} + \lambda_{11} < 1$ since values of $\lambda_{00}\ and\ \lambda_{11}$ larger than 0.5 indicate that misclassification process of the observed response S performs worse than chance. When response misclassification occurs, the true model for the observed dependent variable has the expression

$$E[S_i/x_{ij}] = pr(S_i = 1/x_{ij})\ where\ j = 1, \dots p \quad (6)$$

If there is no misclassification $\lambda_{00} = \lambda_{11} = 1$

$$= pr(S_i = 1/x_{ij}) = g^{-1}(x_i^{'}, \beta) \equiv pr(Y_i = 1/x_{ij}) \quad (7)$$

3.3 Misclassification SIMEX

Simulation and extrapolation (SIMEX) is another general method that can deal with additive measurement error in a continuous variable Cook and Stefanski (1994). This method consists of 'simulation' and 'extrapolation' steps, and is particularly useful for complex models with a simple measurement error structure. Later, SIMEX has been extended to handle misclassification of categorical variables and called the method, MC-SIMEX Kuchenhoff et al (2006). The key idea is that SIMEX estimates are obtained by adding additional measurement error to the data like re-sampling, establishing a trend of measurement error-induced bias over the variance of the added measurement error, and then extrapolating this trend back to the case of no measurement error.

For a binary covariate, the misclassification error can be described by the misclassification matrix λ instead of σ_u^2. Using a similar logic outlined above, the MC-SIMEX estimator can be defined by a parametric approximation of:

$$\alpha \longrightarrow \beta^*(\lambda^{\alpha}): f(1 + \alpha) \quad (8)$$

Where $\lambda = \begin{pmatrix} \lambda_{00} & 1 - \lambda_{11} \\ 1 - \lambda_{00} & \lambda_{11} \end{pmatrix}$,

And $\lambda^{\alpha} = \begin{pmatrix} \lambda_{00} & 1 - \lambda_{11} \\ 1 - \lambda_{00} & \lambda_{11} \end{pmatrix}^{\alpha}$ (9)

λ^{α} Can be expressed as $\lambda^{\alpha} = E\Lambda^{\alpha}E^{-1}$ via spectral decomposition, with Λ being the diagonal matrix of eigenvalues and E the corresponding matrix of eigenvectors. Note that for $\alpha = n$, an integer, $\alpha^{1+n} = \alpha^n * \alpha$ and that $\lambda^0 = I_{kXk}$. Expression $\alpha \longrightarrow \beta^*(\lambda^{\alpha})$ allows the SIMEX method to be applied to the misclassification problem. In this case we will denote the method as MC-SIMEX.

Then by performing a similar simulation step (i.e., generate pseudo data and compute the naïve estimators for each α) and extrapolation step (i.e., fit a curve for the relationship of $X = \alpha\ vs. Y = f(1 + \alpha)$ and find the Y value that corresponds to $X = \alpha = -1$ as in the SIMEX, the MC-SIMEX estimator is computed as $\hat{\beta}_{MC-SIMEX} = \hat{f} = (0)$.

3.3.1. Application of Misclassification Problems

We refer to a general regression problem with a response $\boldsymbol{Y}$ and with a discrete regressor $\boldsymbol{X}$ and further correctly specified regressors $\boldsymbol{Z}$. Since our procedure applies to misclassification of the response and the regressors we denote the possibly misspecified variables by $\boldsymbol{X}^*$or $\boldsymbol{Y}^*$, for the corresponding correctly measured (gold standard) variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. We describe our method for a misspecified regressor $\boldsymbol{X}$. Usually misclassification error is characterized by the misclassification matrix $\boldsymbol{\lambda}$, which is defined by its components $\boldsymbol{\lambda_{ij} = pr(X^* = i/X = j)}$. $\boldsymbol{\lambda}$ is a k × k matrix, where k is the number of possible outcomes for $\boldsymbol{X}$. The parameter of interest is $\boldsymbol{\beta}$, with the limit of the naïve estimator denoted by $\boldsymbol{\beta^*}$.

If $\boldsymbol{X}^*$ has misclassification $\boldsymbol{\lambda}$ in relation to matrix $\boldsymbol{X}$ and $\boldsymbol{X}^{**}$ is related to $\boldsymbol{X}^*$ by the misclassification matrix $\boldsymbol{\lambda^{\alpha}}$ then $\boldsymbol{X}^{**}$ is related to $\boldsymbol{X}$ by the misclassification matrix $\boldsymbol{\lambda^{1+\alpha}}$, when the two misclassification mechanisms are independent. For the function $\boldsymbol{\alpha \longrightarrow \beta^*(\lambda^{\alpha})}$ to be well defined, we need to ensure the existence of $\boldsymbol{\lambda^{\alpha}}$ and that it is a misclassification matrix for $\boldsymbol{\alpha \geq 0}$.

The MC-SIMEX algorithm consists in applying the misclassification matrix $\boldsymbol{\lambda^{\alpha}}$ to the misclassified variable in the simulation step. For the extrapolation step of the MC-SIMEX procedure we need a parametric approximation of (2) $\boldsymbol{\alpha \longrightarrow \beta^*(\lambda^{\alpha}) \simeq G(1 + \alpha, \Gamma)}$. In detail, the MC-SIMEX procedure consists of a simulation and an extrapolation step. Given data $\boldsymbol{(Y_i, X_i^*, Z_i)_{i=1}^n}$we denote the naive estimator by $\boldsymbol{\widehat{\beta}_{na}[(Y_i, X_i^*, Z_i)_{i=1}^n]}$.

3.3.2 Simulation Step

For a fixed grid of values $\boldsymbol{\alpha_1, \ldots, \alpha_m (\geq 0)}$ we simulate B new pseudo data sets by $\boldsymbol{X_{b,i}^*(\alpha_k) := MC[\lambda^{\alpha_k}](X_i^*), i = 1, \ldots n; b = 1, \ldots B; k = 1, \ldots m}$,where the misclassification operation MC [M] ($\boldsymbol{X_i^*}$) denotes the simulation of a variable given $\boldsymbol{X_i^*}$ with misclassification matrix M. Further, we define $\boldsymbol{\alpha_0 = 0, with\ \widehat{\beta}(\alpha_0) = \widehat{\beta}_{na}[(Y_i, X_i^*, Z_i)_{i=1}^n]}$ the estimate of $\boldsymbol{\beta}$ without further measurement error and

$\boldsymbol{\widehat{\beta}(\alpha_k) := B^{-1} \sum_{b=1}^{B} \widehat{\beta}_{na}[(Y_i, X_{b,i}^*(\alpha_k), Z_i)_{i=1}^n]}$, k=1,...,m. (10)

3.3.3. Extrapolation Step

Note that $\boldsymbol{\widehat{\beta}(\alpha_k)}$ is an average over naive estimators corresponding to data with misclassification matrix $\boldsymbol{\lambda^{1+\alpha_k}}$. The estimator $\boldsymbol{\widehat{\beta}}$ is obtained by fitting a parametric model $\boldsymbol{G(1 + \alpha, \Gamma)}$ by least squares on $\boldsymbol{[1 + \lambda_k, \widehat{\beta}(\alpha_k)]_{k=0}^m}$, yielding an estimator $\boldsymbol{\widehat{\Gamma}}$. The MC-SIMEX estimator is then given by $\boldsymbol{\widehat{\beta}_{SIMEX} := G(0, \widehat{\Gamma})}$ which corresponds to $\alpha = -1$. If β is a vector, the MC-SIMEX estimator can be applied on each component of β separately. The application of the MC-SIMEX procedure for a misclassified response Y or more complex misclassification settings is defined in the same way. The estimator $\boldsymbol{\widehat{\beta}_{SIMEX}}$ is consistent when the extrapolation function is correctly specified, that is, $\boldsymbol{\beta^*(\lambda^{\alpha}) = G(1 + \alpha, \Gamma)}$, for some parameter vector $\boldsymbol{\Gamma}$.

However, this is often not the case. When $\boldsymbol{G(1 + \alpha, \Gamma)}$ is a good approximation of $\boldsymbol{\beta^*(\lambda^{\alpha})}$ then approximate consistency will hold. To find a suitable candidate for the function $\boldsymbol{G(1 + \alpha, \Gamma)}$ we exploit the relationship between $\boldsymbol{\beta^*}$ and the misclassification parameter $\boldsymbol{\alpha}$ in the next section for some special cases.

4. Results and Discussion

4.1. Introduction

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. 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 estimated regression coefficients. There is a rich literature on how to correct for this misclassification bias, Gustafson (2003).

4.2. Data Extraction and Discussions

In our research we used data from KAIS (2012) which is Kenya AIDS Indicator Survey 2012. Kenya's second AIDS Indicator Survey (KAIS 2012) was conducted to monitor changes in the epidemic, evaluate HIV prevention, care, and treatment initiatives, and plan for an efficient and effective response to the HIV epidemic. KAIS 2012 was a cross-sectional 2-stage cluster sampling design, household-based HIV serologic survey that collected information on households as well as demographic and behavioral data from Kenyans aged 18 months to 64 years. Participants also provided blood samples for HIV serology and other related tests at the National HIV Reference Laboratory.

Among 9300 households sampled, 9189 (98.8)\% were eligible for the survey. Of the eligible households, 8035 (87.4)\% completed household-level questionnaires. Of 16,383 eligible individuals aged 15–64 years and emancipated minors aged less than 15 years in these households, 13,720 (83.7)\% completed interviews; 11,626 (84.7)\% of the interviewees provided a blood specimen. Of 6,302 eligible children aged 18 months to 14 years, 4340 (68.9)\% provided a blood specimen. Of the 2,094 eligible children aged 10–14 years, 1661 (79.3)\% completed interviews. KAIS 2012 provided representative data to inform a strategic response to the HIV epidemic in the country.

The table below has HIV status which is the true biomedical status used to correct hivstatus_selfreport data that is the misclassified variable. We have two categories: positive and negative individuals. Small n represents number of individuals in each category.

Table 1: Positivity Rates

Test		Positive	Negative
Hiv status	n	648	10978
Proportion %		5.57	94.43
Hivstatus_selfreport n		363	9248
Proportion %		3.78	96.22

It is known, that some participants lie when they are asked about their HIV status. Research indicates, that about 8% of hivstatus_ selfreport them as negative, so the misclassification matrix is defined by

Table 2: Misclassification Table

	Hiv Status		
		Negative	Positive
	Negative	7628	28
Hivstatus_selfreport	Positive	244	305
	Total	7872	333

Table 3: Misclassification Rates

	Hiv Status		
		Negative	Positive
	Negative	96.9	8.4
Hivstatus_selfreport	Positive	3.1	91.6

Here sensitivity which is defined as pr (S=1/Y=1) is 91.6 and specificity defined as pr(S=0/Y=0) is 96.9. For the correction of the misclassification we have to compare the simulation standard deviation (SE) of each estimator and the estimates: (a) under no misclassification (true. model), (b) when misclassification is ignored (naïve. model) and (c) when corrected for misclassification (SIMEX models).

We model the probability that $S = 1, pr(S_i = 1) = logit^{-1}(X_i\beta)$, under the assumption that the outcomes S_i are independent given these probabilities. We illustrate classical logistic regression with a simple analysis from the Kenya AIDS Indicator Surveys in 2012. For each respondent I in this data, we label $S = 1$ if the condition is present or 0 if condition is absent, excluding the self-reported indeterminate, never received results and never tested individuals.

The MC-SIMEX estimator is calculated for the log-linear extrapolation functions. The true estimator is calculated using the correctly measured data and the naive estimator using the possibly corrupted variables. Our approach is general since the only assumptions to be made are the availability of a consistent estimator for the model parameters in case of no misclassification and an estimator or exact knowledge of the misclassification matrix.

So MC-SIMEX is applicable to general regression models involving binary, ordinal, and count data subject to misclassification in either response or regressor. We compare the MC-SIMEX and the matrix method in the case of a misclassified binary regressor. The MC-SIMEX method gives a better coverage rate than the matrix method.

Moreover, it can handle more complex situations like the addition of confounders, differential misclassification, misclassification dependent on other variables, or simultaneous misclassification in more than one discrete variable.

The MC-SIMEX correction gave improved estimates even with high misclassification probabilities. The results presented in tables below show a better correction in the case of a binary regressor to the logistic regression with misclassified response.

Table 4: The naive parameter estimates ignoring misclassification.

	Naïve-model			
	Estimate	Std.Error	z vaue	pr(>|z|)
Sex:				
Male	0	.	.	.
Female	0.69894	0.12657	5.522	<0.001
Residence:				
Urban	0	.	.	.
Rural	0.46696	0.12499	3.736	0.0001
Age				
(15-24)	0	.	.	.
(25-34)	1.37825	0.24995	5.514	<0.001
(35-44)	2.01608	0.22708	8.878	<0.001
(45-54)	2.39337	0.23334	10.257	<0.001
(55-64)	2.29333	0.25285	9.070	<0.001
(65+)	1.83914	0.39067	4.708	<0.001
Provinces:				
Nairobi	0	.	.	.
Central	0.25551	0.25419	1.005	0.3148
Coast	0.16935	0.25119	0.674	0.5002
Eastern	0.02317	0.24935	0.093	0.9260
Nyanza	1.83338	0.21025	8.720	<0.001
RiftValley	0.07274	0.25034	0.291	0.7714
Western	0.61229	0.25215	2.428	0.0152

The socio demographic variables with zero values i.e (under sex we have male, for residence we have urban e.t.c) are the baseline variables we use to compare with the rest of the variable in each category. In the naive model which ignores misclassification we have HIV prevalence being higher in the female as compared with the male who were considered in our study, also it is higher among the residents in the rural areas as compared with the residents in urban areas.

For the age categories HIV prevalence was highest in the individuals aged between 45-54 years. Finally Nyanza province was found to have a higher HIV prevalence 1.833 as compared with the other provinces: 0.2555, 0.1694, 0.0232, 0.0727 and 0.6123 for Central, Coast, Eastern, Rift Valley and Western respectively.

Table 5: The corrected parameter estimate for MC-SIMEX model

	Estimate	Std.Error	t value	pr(>|t|)
Sex:				
Male	0	.	.	.
Female	1.7425	0.2133	8.170	<0.001
Residence:				
Urban	0	.	.	.
Rural	1.1951	0.2143	5.577	0.001
Age				
(15-24)	0	.	.	.
(25-34)	3.7212	0.4271	8.713	<0.001
(35-44)	5.3228	0.3942	13.504	<0.001
(45-54)	6.1910	0.3984	15.538	<0.001
(55-64)	5.9138	0.4310	13.723	<0.001
(65+)	4.8814	0.6619	7.375	<0.001
Provinces:				
Nairobi	0	.	.	.
Central	0.7191	0.4429	1.624	0.1045

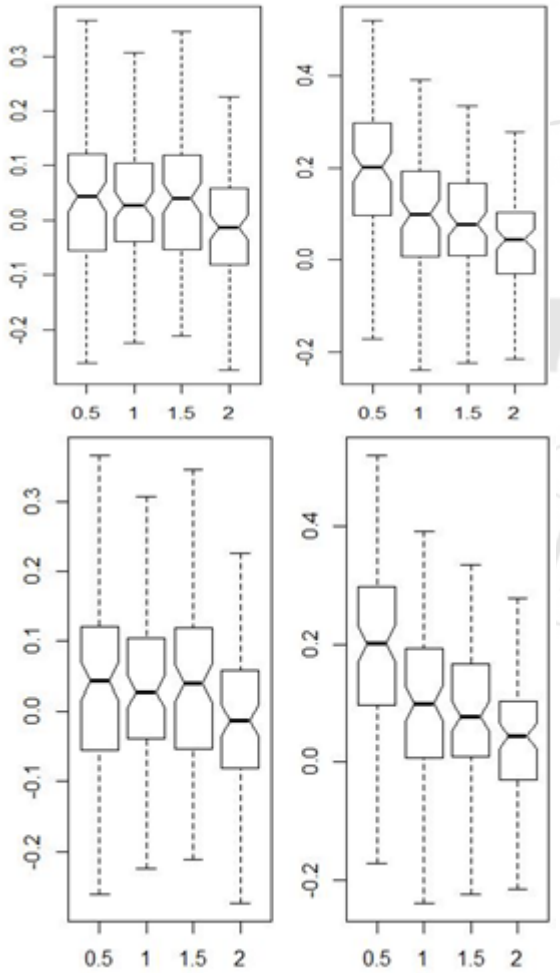
Coast	0.4639	0.4434	1.046	0.2955
Eastern	0.0591	0.4346	0.136	0.8919
Nyanza	4.3469	0.3671	11.840	<0.001
RiftValley	0.1644	0.4360	0.377	0.7062
Western	1.5722	0.4387	3.584	0.0003

The MC-SIMEX approach derives estimates of the model parameters for a general class of models (corrected for misclassication). It is also important to know the standard errors and the estimate value of the true.model and the MC-SIMEX model. In our findings the KAIS data was quite high; the correction of the parameter estimates is in all cases substantial and not to be ignored.

Most of the MC-SIMEX corrected estimates of the regression coefficient are larger in absolute size than the uncorrected version. Hence making the true.model a better model compared with the naive.model and also the MC-SIMEX model the best among them all.

4.3 Graphical Illustrations of the Analysis

4.3.1 Side by side box plots



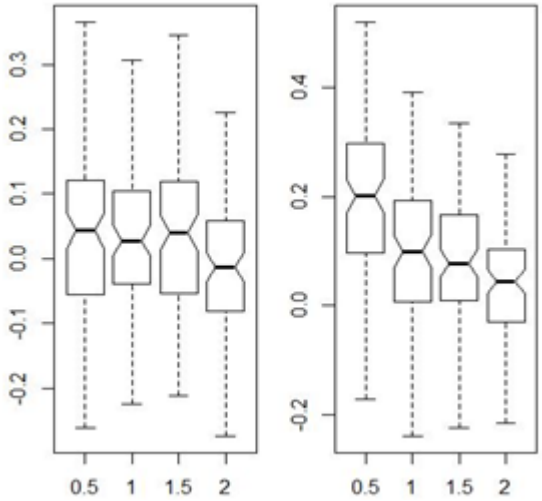


Figure 1: Box plots used to do comparison and show the relationship between the variables.

4.3.2 MC-SIMEX fit to the KAIS data

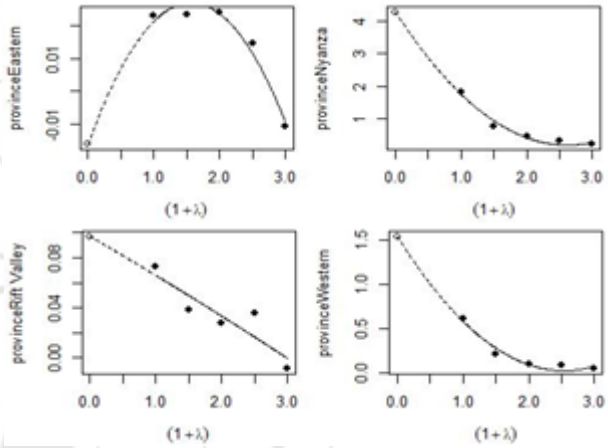


Figure 2: MC-SIMEX fit showing the effect of misclassification error in the province variable.

5. Conclusion and Recommendations

We have presented the misclassification SIMEX (MC-SIMEX) method for parameter estimation in regression models in the presence of misclassification. It is based on the SIMEX idea for additive normal covariate measurement error Cook and Stefanski (1994). The package SIMEX features easy to use functions for correcting estimation in regression models with measurement error or misclassification via the SIMEX– or MC-SIMEX–method. It provides fast and easy means to produce plots that illustrate the effect of misclassification on parameters. Several additional functions are available that help with various problems concerning misclassification.

The MC-SIMEX method has shown good results in our study. It reduces bias compared with the naive estimator and its performance is comparable to ML estimation, where it is feasible. So MC-SIMEX is applicable to general regression models involving binary, ordinal, and count data subject to misclassification in either response or regressor. Further, it can be used when the misclassification probabilities are estimated from a validation study, which will be the case in many practical situations.

One problem is the correct specification of the parametric form of the extrapolation function, which characterizes the relationship between the amount of misclassification and the limit of the naive estimator. Since the exact form is not available in most situations, for variance estimation and confidence Intervals we propose using a two-step bootstrap method in the case of uncertain knowledge of the misclassification matrix.

We also recommend the use of the package SIMEX which is easy to use for correcting estimation in regression models with misclassification via MC-SIMEX, in the epidemiological survey research.

6. Acknowledgement

I would first of all like to sincerely thank God, who made all the things possible by His grace which is bestowed upon me. Secondly my profound gratitude and deep regards to my supervisors **Dr. Samuel Mwalili** and **Mr. Humphreys Murray** who gave me a wonderful and golden opportunity to do this project and their exemplary guidance, monitoring and constant encouragement throughout.

A special feeling of gratitude to my loving parents, **Hudson** and **Elizabeth** whose words of encouragement and push for tenacity ring in my ears. My grandma **Nyankangi**, my dearies **Deborah** and **Brian** who have never left my side you are very special. Last, but not the least, I would like to thank my friends, chairperson and the Department of Statistics and Actuarial Science members for their moral support during the project. The blessing, help and guidance given by you all time to time shall carry me a long way in the journey of life.

References

[1] Allport, F. H. (1927). Self-evaluation: a problem in personal development. Mental Hygiene, 11, 570-583.
[2] Carroll, R. J., Küchenhoff, H., Lombard, F., and Stefanski, L. A. (1996). Asymptotics for the SIMEX Estimator in Nonlinear Measurement Error Models. Journal of the American Statistical Association, 91:242–250.
[3] Carroll, R., Ruppert, D., and Stefanski, L. A. (1995). Nonlinear Measurement Error Models. New York: Chapman and Hall
[4] Carroll, R., Maca, J., and Ruppert, D. (1999). Nonparametric regression in the presence of measurement error. Biometrika 86, 541–554.
[5] Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition.Chapman & Hall/CRC.
[6] Cook, J. R., and Stefanski, L. A. (1994). Simulation–Extrapolation Estimation in Parametric Measurement Error Models. Journal of the American Statistical Association, 89:1314–1328 .
[7] Dafni, U. G., and Tsiatis A. A. (1998). Evaluating surrogate markers of clinical outcome when measured with error. Biometrics. 54(4):1445-62.
[8] Dunning, D., Heath, C., and Suls, J. M. (2005). Flawed self-assessment: Implications for education, and the workplace. Psychological Science in the Public Interest, 5, 69-1.
[9] Elizabeth, H. S., and Dipankar, B. S. (2009). An investigation of the MC-SIMEX method with application to measurement error in periodontal outcomes. Stat Med. 28(28): 3523–3538.Marschak, J. (1939), "On combining market and budget data in demand studies", Econometrica, Vol. 7, pp 332-335
[10] Hu, P., Tsiatis, A. A., and Davidian, M. (1998).Estimating the parameters in the Cox model when covariate variables are measured with error. Biometrics.54(4):1407-19.
[11] Johnson, J. A. (2004). Impact of item characteristics on item and scale validity. Multivariate Behavioral Research, 39, 273-302.
[12] Kowalski, J., and Tu, X. M. (2002) A generalized estimating equation approach to modelling incompatible data formats with covariate measurement error: Application to human immunodeficiency virus immune markers. Journal of the Royal Statistical Society, Series C: Applied Statistics.51(1):91–114.
[13] Küchenhoff, H., Mwalili, S. M., and Lesaffre, E. (2005). A general method for dealing with misclassification in regression: the Misclassification SIMEX. Biometrics.
[14] Lin, X., Carroll, R. J. (2000). Nonparametric function Paul, S. (2004). Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments .Am J Epidemiol 159 (9): 911-912.
[15] Stefanski, L. A., and Cook, J. R. (1995). Simulation-Extrapolation: The Measuerement Error Jackknife. Journal of the American Statistical Association, 90: 1247–1256, 5.