Katholieke Universiteit Leuven
Faculteit Wetenschappen
Departement Wiskunde

Bayesian and Frequentist Approaches to Correct for Misclassification Error *with Applications to Caries Research*

Samuel Musili MWALILI

Promotoren: Prof. Dr. E. Lesaffre
Prof. Dr. J. Beirlant

Proefschrift ingediend tot
het behalen van de graad van
Doctor in de Wetenschappen

October 2006

ISBN 90-8649-050-6
D/2006/10.705/47

In Memory of Dr. Elijah Muthengi Mukuni

"And I–in righteousness I shall see your face; when I awake, I shall be satisfied with seeing your likeness"

Psalm 17:15

Dankwoord

Graag wil ik hier iedereen bedanken die heeft meegeholpen aan het voltooien van dit proefschrift.
In de eerste plaats bedank ik mijn beide promotoren. Prof. E. Lesaffre, hoofd van het Biostatistisch Centrum, heeft mij geïntroduceerd in het domein van de groep sequentile methoden. Door de vele discussies over het onderwerp en de aanmoedigingen tijdens de moeilijke momenten heeft hij mij weten te motiveren om toch telkens door te bijten. Ik dank ook Prof. J. Beirlant om mede promotor te zijn van deze thesis. Bijzandere dank aan Prof. H. Küchenhoff en Prof. D. Declerk.
Dank ook aan alle leden van de jury, Prof. An Carbonez, Prof. Anne-Marie De Meyer, Prof. Dominique Declerck, Prof. Geert Molenberghs, Prof. Geert Verbeke en Prof. Helmut Küchenhoff voor de kritische lezing van dit werk die tot een serieuze verbetering leidde.
Mijn (ex-)collega's van het Biostatistisch centrum, Wendimagegn, Roos, Steffen, Spyridoula, Silvia, María-José, Luwis, Kris, Kirsten, Jeannine, Francis, Dimitris, Arnošt, Ann, Anhar en Alejandro wil ik voor de aangename werkomgeving gedurende de laatste vijf jaren danken.
Deze thesis zou niet kunnen ontstaan zonder de financiële steun van de onderzoeksbeurzen van de Katholieke Universteit Leuven. De steun van de beurzen OT/00/35 wordt diep geapprecieerd.
Tot slot bedank ik mijn familie en vrienden. Extra dank gaat uit naar D. Karel en familie van Mian, Jaak en Fien. In het bijzonder wil ik ook mijn ouders en mijn vrouw Telesia bedanken, zonder wie ik dit proefschrift onmogelijk tot een goed einde had weten te brengen.

Bedankt,
Samuel.

Shukrani

Ninaanza kwa kumshukuru mungu kwa kuwawezesha wasimamizi na washauri wangu profesa E. Lesaffre, J. Beirlant, D. Declerk kwa kukubali kwao kutoa ushauri na mapendekezo ya kazi yangu ya utafiti. Ninawashuru pia taasisi nzima ya chuo Kikuu cha Kikatoliki cha Leuven kwa msaada wa mali na kiroho.

Kwa kuwa si rahisi kuwashukuru watu wote kwa pamoja, ninaomba kuwataja wawakilishi wote walionisaidia katika shughuli nzima ya utafiti.

Kwanza ninawashukuru uongozi na jumuiya ya shule ya Sekondari ya Kalimani, pamoja na usharika wa jumuiya ya kanisa la A.I.C. Kyambusya kwa msaada wa kiroho.

Pili nina washukuru marafiki zangu, pamoja na familia yangu, hasa mke wangu Telesia kwa kunipatia mawazo na mapendekezo.

Mwisho ni vigumu kuwataja watu wote kwa mchango wao wa maoni na mapendekezo, ninaomba kila mmoja apokee shukurani zangu kwa kuniwezesha kukamilisha kazi hii.

Ahsanteni,
Samueli.

Acknowledgements

Over the years it has been a honour to encounter many people who have given me more of their time, companionship, professional and personal help, and above all, providing a conducive environment that gave me the possibility to complete this thesis.

I would first of all like to thank my promotor, Prof. Dr. E. Lesaffre, who arduously assisted in times of need, and for guidance that improved the quality of this thesis considerably. It has been a privilege and pleasure to work with you. I am deeply grateful to co-promoter Prof. Dr. J. Beirlant. A debt of gratitude is owed to Prof. Dr. D. Declerk and Prof. Dr. H. Küchenhoff whose collaboration has been very important to my work. I am grateful for financial support of Research Grant OT/00/35, Catholic University Leuven.

My countless thanks to the Kenyan community in Leuven, even though I cannot mention you individually, you surely provided a *home away from home* atmosphere.

I would like to express my appreciation to Prof. Dr. R. O. Odhiambo and H. M. Humpreys for the indirect and continued support.

I am grateful to the least in Kalimani Secondary School. I am thankful for thoughts and prayers of members of the A.I.C. Kyambusya.

I am deeply indebted to my mother, Laeli, my father, Mwalili, my granny, Malia, and to all family members for their sleepless nights, constant support and encouragement. I am grateful to my twin brother Lewis alias Coco for his spiritual support. I am also indebted to Peace and Angela for being part of the family.

I would like to give my special thanks to my wife, Telesia. She proved that love is the key that unlocks the bars of impossibility.

Thank you,

Samuel.

The majority of this thesis is based largely on the following papers:

Chapter 4: Lesaffre, E., Mwalili, S. M. and Declerck, D. (2004). Analysis of caries experience taking inter-observer bias and variability into account. *Journal of Dental Research* **83**, 951–955.

Chapter 5: Mwalili, S. M., Lesaffre, E. and Declerck, D. (2006). A Bayesian ordinal logistic regression model to correct for inter-observer measurement error in a geographical oral health study. *Journal of the Royal Statistical Society, Series C* **1**, 77–93.

Chapter 6: Mwalili, S. M., Lesaffre, E. and Declerck, D. (2006). The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research. *Statistical Methods in Medical Research* **15**, in press.

Chapter 7: Küchenhoff, H., Mwalili, S. M. and Lesaffre, E. (2006). A general method for dealing with misclassification in regression: the misclassification SIMEX. *Biometrics* **62**, 85–96.

Chapter 8: Lesaffre, E., Küchenhoff, H., Mwalili, S. M. and Declerck, D. (2006). On the estimation of the misclassification table for finite count data with an application in caries research. In revision.

Chapter 9: Mwalili, S. M., Lesaffre, E., Küchenhoff, H. and Declerck, D. (2006). Modeling multivariate binary data subject to misclassification error. Submitted.

The author was also co-author of the following papers:

Vanobbergen, J., Declerck, D., Mwalili, S. M. and Martens, L. (2004). The effectiveness of a 6-year oral health education programme for primary schoolchildren. *Community Dentistry & Oral Epidemiology,* **32** 173–182.

Kallu, R., Vinckier F., Politis, C., Mwalili, S. M. and Willems, G. (2005). Tooth transplantations: a descriptive retrospective study. *International Journal of Oral & Maxillofacial Surgery* **34**, 745–55.

Contents

Preface

Since long, methods have been developed in epidemiologic studies to correct for the effects of measurement and scoring errors on the exposure-disease relation. Ignoring measurement and scoring errors has a possible biasing effect on this relation. For this reason we observe an increasing use of these correction methods in epidemiology, especially in nutritional studies which aim to relate the risk of various diseases on, say, measures for dietary behaviour such as saturated fat intake. However, in other medical areas, such as in caries research, we argue that the correction for scoring errors is largely underused.

Chapter 1 gives a general overview of measurement error and misclassification in epidemiological studies. When the variable under consideration is continuous, then one speaks of measurement error. In the case of a discrete variable, scoring with error leads to misclassification. More attention has been paid in the literature to adjustment for measurement error. Both correction for measurement error and/or misclassification involves validation data. These validation data are extra collected data where the true and possibly corrupted variables are measured for each subject.

In this thesis, the applications are taken from caries research. Important issues in caries research are described in Chapter 2. This chapter also introduces the Signal Tandmobiel® study, which provided the trigger to study methods dealing with misclassification errors. The

dmft-index, a popular measure for caries experience in the primary dentition, will mainly be used in this thesis as a response variable.

In Chapter 3 general principles of frequentist and Bayesian model fitting are reviewed.

In large oral health surveys, clinical measurements are often obtained from several examiners. This raises the issue of examiner bias and variability in measuring. Often kappa values are reported to indicate agreement between scorers. In Chapter 4, the limitations of this statistic are presented. Further, an alternative approach to correct for scoring errors, based on the sensitivity and specificity of the scoring behaviour of the dental examiners relative to a benchmark scorer, is proposed within either the frequentist or the Bayesian paradigm. We applied this approach to the caries experience data of the seven-year-old children from the Signal Tandmobiel® study. In this application, the measurement of interest is a binary outcome, that is, no-caries (0) vs. caries (1). We also accounted for the uncertainty with which the correction terms are estimated. In Chapter 5 we extend this approach to an ordinal score. In the application the ordinal score is obtained by splitting up the dmft-index. The correction process is now based on a matrix of misclassification probabilities. Various models for estimating the misclassification structure are considered.

In Chapter 6 models for count data subject to misclassification are expressed. Here we look at the zero-inflated negative binomial regression model, and illustrate how correction for misclassification can be achieved. We also consider the approach of Albert, Hunsberger, and Biro (1997) to model the correction terms. Our approach is applied to the dmft-index. Similarly, Chapter 7 describes models for bounded count data, focusing mainly on the zero-inflated beta-binomial regression model. The outcome of interest is then the dmft-score restricted

to the 8 deciduous molars. In addition to the approach of Albert et al. (1997), the double binomial approach is used to model the misclassification process. The double binomial approach is based on the fact that to determine a count the examiner needs to evaluate all items that make up that count. We also suggest various extensions of this approach which might mimic better the scoring behavior of the examiner relative to a benchmark scorer.

In Chapter 8, a general approach for handling misclassification in discrete covariates or responses in regression models is developed. The simulation and extrapolation (SIMEX) method, which was originally designed for handling additive covariate measurement error, is developed for misclassification and is called the MC-SIMEX method. We show that our method is quite general and applicable to models with misclassified response and/or misclassified discrete regressors.

In Chapter 9, a general approach to multivariate binary data subject to misclassification error is proposed. The misclassification problem is simplified by expressing the unobserved true response in terms of the observed and latent data. In this application, the response is a multivariate binary vector indicating absence or presence of caries in the primary molars.

Finally, a general conclusion and ideas for future research are given in Chapter 10.

Notations

In this section we give brief explanation of the notations used in this thesis. For precise definitions, see the text.

$\boldsymbol{\beta}$	coefficients in the regression model
$\boldsymbol{\beta}_b(\lambda)$	simulated estimator used in MC-SIMEX
$f_{Y\|X,Z}$ or $[Y\|X,Z]$	conditional density of Y given (X,Z)
$\mathcal{G}$	extrapolation function in MC-SIMEX
Π	a matrix of misclassification probabilities
(λ_1,λ_0)	sensitivity and specificity of Y^* as surrogate for Y
$\boldsymbol{M}$	a matrix of misclassification frequencies
π	probability of success in logistic regression
$\boldsymbol{\theta}$	unknown vector of parameters
W	a latent variable underlying the misclassification process
X	the error-prone covariate
X^*	the observed value of the mismeasured covariate
Y	the error-prone response
Y^*	the observed value of the mismeasured response
Z	covariate measured without error

CHAPTER 1

Measurement Error and Misclassification: An Overview

1.1 Introduction

Epidemiological studies aim to estimate the impact of exposure to a risk factor on the development of a disease. To quantify the effect of exposure it is important that the risk factor as well as the variable that measures the disease are recorded without error. Unfortunately, this is often not the case. If measurement error is present the association between the risk factor and the disease will be underestimated and one speaks of a dilution effect. Many epidemiological study areas suffer from this dilution effect. For instance, when the impact of dietary habits on the incidence of coronary heart disease is of interest one typically inquires subjects using a food questionnaire. Unfortunately, a food questionnaire can never measure the true food intake of a subject. Often there will be underreporting of the exact food intake but also there will be a larger variability in the reported food

intake from questionnaires than in the true food intake. This implies that the relationship with the risk of coronary heart disease might be underestimated if based on the reported food intake. The same phenomenon occurs when examining the impact of passive smoking on, for example, lung cancer. Not only the risk factors (covariates) are subject to measurement error, also the disease variable (response) can be corrupted, as we shall see here in this thesis.

When the data are categorical, measurement error is called *misclassification error*. Measurement error and hence also misclassification error results in biased estimates of the parameters of interest. Any regression analysis that treats the error-corrupted variable to be the same as the true variable of interest is referred to as *naive*, since it ignores the effect of measurement error in the parameter estimation. It is well known that the effect of measurement error is to bias parameter estimates. For example, in simple linear regression, measurement error on a covariate will bias the slope estimator in the direction of 0. This type of bias is referred to as *attenuation* or *attenuation to the null*.

Despite the detrimental effect that measurement and misclassification error have on the estimation process, unbiased estimates of the parameters can still be obtained if one has an idea of the magnitude and the direction of the error. This knowledge could possibly be elicited from experts and combined in a Bayesian way to correct for bias. This correction can also be achieved if validation data are available. Such data are obtained from a validation study where, on a subset of subjects, not only the possibly corrupted value is known but also the true value. Correction for measurement/misclassification error can be done in a frequentist and a Bayesian way. This will be the topic of this thesis, where the problem will be specifically applied

to caries research.

1.2 Existing literature

The problem of covariate measurement error in regression analysis has been extensively studied for the last two decades. Fuller (1987) presents early research in linear regression with covariates subject to measurement error and/or misclassification. Models that take measurement error in the covariates into account are called *errors-in-variables* models.

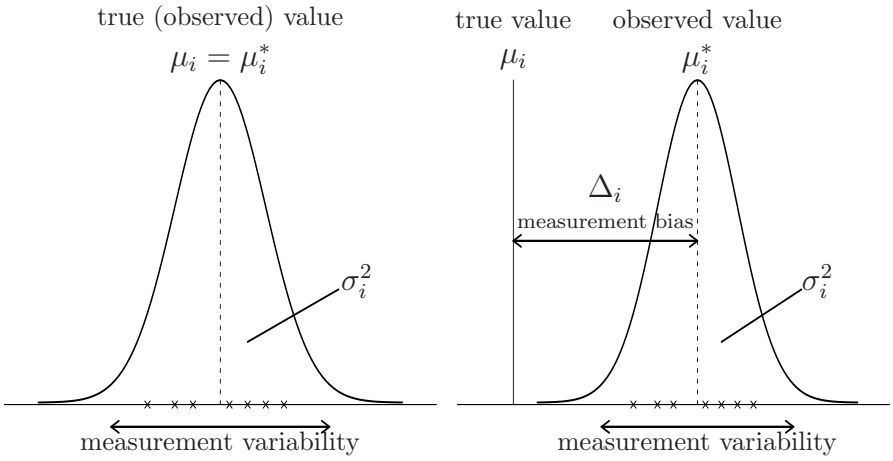
The early literature on the effects of misclassification include Bross (1954); Goldberg (1975); Greenland (1980); Greenland and Robins (1985); Greenland (1988); Birkett (1992) and a (bibliography) review by Chen (1989). A general overview of measurement error including misclassification error is given by Willett (1989). More recent reviews on both measurement error and misclassification are included in a bibliography by Carroll, Ruppert, and Stefanski (1995) and also in Gustafson (2004).

We will now summarize the most important issues in measurement/ misclassification error problems.

1.3 Types of errors in epidemiologic studies

Sources of measurement error can be grouped into two general types: *random* and *systematic*. The distinction is that for random measurement error, the mean of many repeated measures will approach the true value. On the other hand for systematic measurement error, the mean of the repeated measurements does not approach the true value. In Figure 1.1 measurement variability, σ_i^2, describes a typical spread in

the values of repeated measurements for subject i, and hence it represents the random measurements fluctuating around the average value. Therefore, under random error (see Figure 1.1(a)), measurements fluctuate at random around the individual's true value, μ_i. Examples of random error are the day-to-day variation in dietary intake and day-to-day measurements of systolic blood pressure. On the other hand, systematic measurement error is the measurement bias, i.e. the offset between the observed value, μ_i^*, and the true value, μ_i, equal to Δ_i (see Figure 1.1(b)).



(a) ***Random measurement error.*** **(b)** ***Systematic measurement error.***

Figure 1.1: *A graphical view of measurement error for subject i. The dispersion of the distribution determines the measurement variance – random error, while the offset between the observed value and the true value is bias of the measurement – systematic error.*

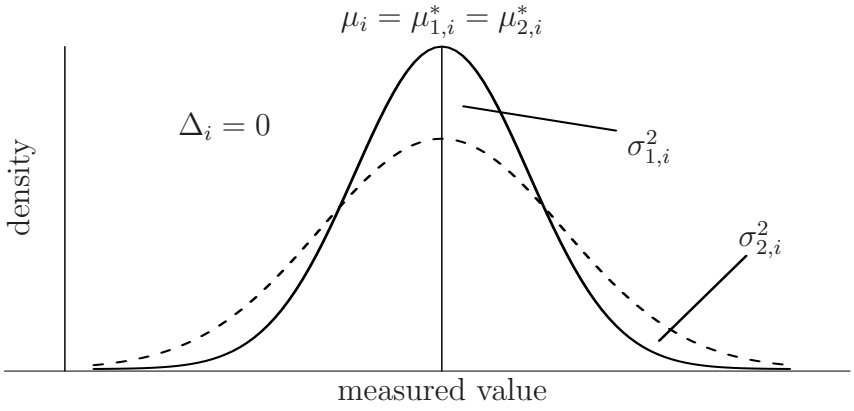
Systematic measurement error happens frequently in many (epidemiologic) studies and can have many causes, but typically it happens because of a miscalibrated instrument. Systematic measurement error is also likely to occur when a standardized food questionnaire,

in which a common food item is omitted, is used. In that case all individuals are affected in the same direction, but not necessarily to the same degree since the use of these foods will differ among subjects.

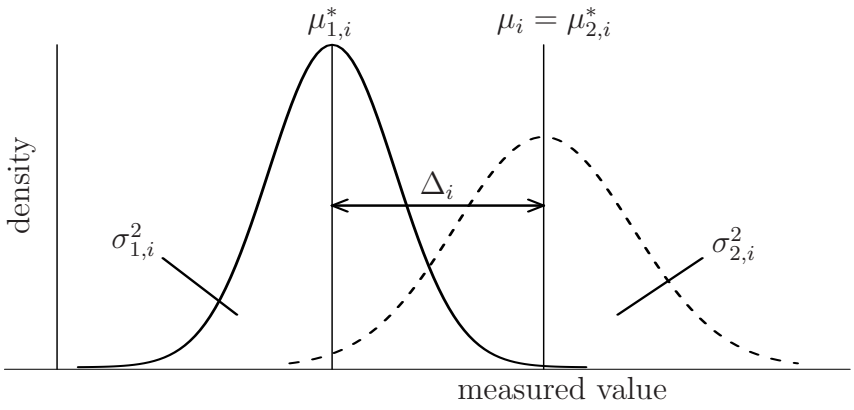
Figure 1.2(a) shows two measurements exhibiting random error but differing in precision. It is clear that random error is simply the variability around the true value, μ_i, measured by $\sigma^2_{1,i}$ (Measurement 1) or by $\sigma^2_{2,i}$ (Measurement 2). Now $\Delta_i = 0$. In Figure 1.2(b) a measuring instrument is shown that experiences both random and systematic measurement error. Thus now $\Delta_i \neq 0$. While in the previous example it was clear which measuring instrument is to be preferred, now the situation is less clear. It will not only depend on the magnitude of Δ_i and the ratio $\sigma^2_{1,i}/\sigma^2_{2,i}$ but also on how well Δ_i, $\sigma^2_{1,i}$ and $\sigma^2_{2,i}$ can be estimated.

1.4 The effect of measurement error and misclassification on estimation

In this section we present the basic concepts used in the measurement error literature. First we point out the distinction between differential and non-differential error, then explore the effect of measurement error and misclassification on the estimation of parameters, and finally review the existing methods which correct for error-prone data. The methods are classified into two main classes: functional and structural error. In this thesis we shall consider both the case of a covariate and a response measured with error.



(a) ***Measurement*** 1 ***(solid line) is more precise than Measurement*** 2 ***(broken line). The true value being measured is*** μ_i.



(b) ***Measurement*** 1 ***(solid line) is more precise than Measurement*** 2 ***(broken line) but presents bias. The true value being measured is*** μ_i.

Figure 1.2: *Comparison of bias and precision of two measurement processes on the basis of the probability distribution of the measurements for subject* i.

1.4.1 Measurement error

1.4.1.1 Non-differential and differential

Suppose the true response is denoted by Y and the possibly corrupted response by Y^*. Consider a regression of the response Y on covariates

X and Z. When, given the true response Y, the conditional distribution of the possibly corrupted response Y^* does not depend on the covariates (X, Z) then the measurement error process is said to be *non-differential*, i.e. $f_{Y^*|Y,X,Z} = f_{Y^*|Y}$, where $f_{w|v}$ denotes the conditional distribution of w given v. In other words, for a non-differential measurement error process the covariates do not provide any extra information about the response measured with error over and above what is given by the true response. On the other hand, if $f_{Y^*|Y,X} \neq f_{Y^*|Y}$ then measurement error in response is said to be *differential*. Under non-differential measurement error the possibly corrupted Y^* is called a *surrogate* for Y. If measurement error is due solely to instrument or laboratory-analysis error, then it can often be argued that the error is non-differential. Non-differential misclassification often arises in prospective studies, where the true and observed covariates are assessed before response is measured.

For a covariate, non-differential measurement error occurs when the observed covariate X^*, as a measure of X, has no information about the response Y, in addition to the information already contained in the true covariate X and other correctly measured covariates Z. In other words, the conditional distribution of Y given (Z, X, X^*) is the same as that of Y given (Z, X), that is $f_{Y|Z,X,X^*} = f_{Y|Z,X}$. When $f_{Y|Z,X,X^*} \neq f_{Y|Z,X}$ the error is called differential. Differential misclassification is more likely to occur in case-control studies, where the response is obtained first and subsequently ascertained covariates may be subject to recall bias.

Under non-differential measurement error the observed covariate X^* is said to be a surrogate for the true covariate X. The same terminology is used for misclassification error.

1.4.1.2 Effect of measurement error

In this section we explore the effect of response or covariate measurement error in simple and multiple regression.

Response measurement error: There is an extensive literature on the effect of measurement error on estimation of parameters, e.g.: Mote and Anderson (1965); Goldberg (1975); Greenland (1980); Stefanski (1985); Greenland (1988); Yoshimura (1991); Birkett (1992); Fuller (1987); Carroll et al. (1995). Most of the papers focus on errors in covariates. The problem with measurement in the response has received less attention, largely because response measurement error may be *ignorable* in some cases. By ignorable we mean that the model for the true response holds also for the error-prone response, with the same parameters values, except for some variance parameters which are changed due to measurement error. For example, in a linear regression model with equation

$$Y \;=\; \beta_0 \;+\; \boldsymbol{\beta}_x' \boldsymbol{X} + \epsilon,$$

where Y represents the true response variable and $\boldsymbol{X}$ the covariate vector measured without error, ϵ represents the inter-subject variability with variance σ_e^2. Let $\hat{\boldsymbol{\beta}}$ be the estimate of $\boldsymbol{\beta}_x$ from the regression of Y on $\boldsymbol{X}$. If instead $Y^* = Y + U$ is measured with error, where U is independent of $(Y, \boldsymbol{X})$ with mean zero and variance σ_u^2, then

$$Y^* \;=\; \beta_0 \;+\; \boldsymbol{\beta}_x' \boldsymbol{X} \;+\; (\epsilon \;+\; U)$$

implying that $\mathrm{var}(Y^*) \;=\; \sigma_u^2 + \sigma_e^2$ and $\hat{\boldsymbol{\beta}}^* \;=\; \hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}^*$ is the estimate of $\boldsymbol{\beta}_x$ from the regression of Y^* on $\boldsymbol{X}$. Clearly, in this example, the proxy response Y^* is unbiased for Y, i.e. $\mathrm{E}(Y^*) = \mathrm{E}(Y)$ but the

variance is increased. In this case measurement error is ignorable but still present since variance increases.

On the other hand, suppose that $Y = \beta_0 + \boldsymbol{\beta}_x' \boldsymbol{X} + \epsilon$ and that $Y^* = \alpha_0 + \alpha_1 Y + U$, $\alpha_0 \neq 0, \alpha_1 \neq 1$, with U independent of $(Y, \boldsymbol{X})$. Then the observed data follow a linear model of the form:

$$Y^* = \alpha_0 + \alpha_1\beta_0 + \alpha_1\boldsymbol{\beta}_x'\boldsymbol{X} + (\epsilon + U), \quad \alpha_0 \neq 0, \alpha_1 \neq 1$$

implying that $\text{var}(Y^*) = \sigma_u^2 + \sigma_e^2$ and $\hat{\boldsymbol{\beta}}^* = \widehat{\alpha_1\boldsymbol{\beta}}$. In this case $\text{E}(Y^*) \neq \text{E}(Y)$, i.e. Y^* is biased for Y, hence measurement error is not ignorable.

The attenuation effect for a simple nonlinear model is similar to that for simple linear regression. However, the effect of measurement error in nonlinear models is in general more complex than for linear models.

Measurement error in covariates: Fuller (1987) and Carroll et al. (1995) present a comprehensive overview on the effect of measurement error on covariates in linear and nonlinear regression models. In linear regression, measurement error on covariates will not only attenuate the parameter estimate but also induce increased variability around the regression line.

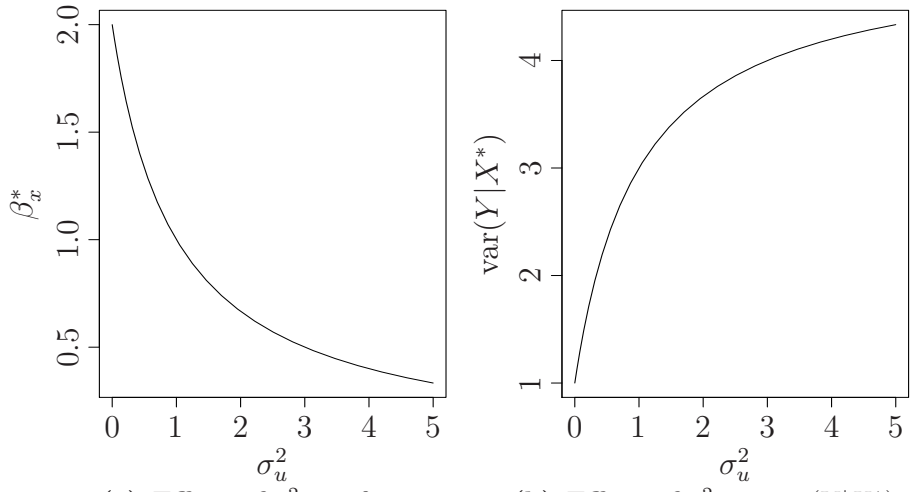
More specifically, suppose $Y = \beta_0 + \beta_x X + \epsilon$, where X has mean μ_x and variance σ_x^2, and the equation error ϵ is independent of X and has mean 0 and variance σ_ϵ^2. Assuming a classical additive measurement error model, i.e. $X^* = X + U$ where U is independent of X with mean 0 and variance σ_u^2, then a naive ordinary least squares regression estimates $\beta_{x^*} = \lambda\beta_x$, where

$$\lambda = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} < 1.$$

Figure 1.3(a) shows the effect of σ_u^2 on the estimation of β_x. Clearly the ordinary least squares regression of Y on X^* produces an estimator that is attenuated to 0. The attenuating factor λ is also called a *reliability ratio*. The variance of Y given X^* is

$$\text{var}(Y|X^*) = \sigma_\epsilon^2 + \lambda\beta_x^2\sigma_u^2.$$

As can be seen in Figure 1.3(b), the variability around the regression increases as σ_u^2 increases. Thus, in this example measurement error causes two problems: attenuation of the slope parameter, and increased variability around the regression line.



(a) ***Effect of*** σ_u^2 ***on*** β_x**.** **(b)** ***Effect of*** σ_u^2 ***on*** **var**$(Y|X^*)$**.**

Figure 1.3: *Illustrating the effect of measurement error on a simple linear regression:* $Y = \beta_0 + \beta_x X + \epsilon$ *for different values of* σ_u^2*. For this example* $\sigma_\epsilon^2 = \sigma_x^2 = 1$*, and* $(\beta_0, \beta_x) = (-1, 2)$*.*

In multiple regression the effect of measurement error is more complicated if at least one additional covariate is correlated with the surrogate covariate. For example, suppose that the multiple regression model for Y given $\boldsymbol{X}$ and $\boldsymbol{Z}$ is $Y = \beta_0 + \boldsymbol{\beta}_x'\boldsymbol{X} + \boldsymbol{\beta}_z'\boldsymbol{Z} + \epsilon$, where $\boldsymbol{X}$

is a matrix of multiple covariates measured with error and $\boldsymbol{Z}$ is a set of covariates measured without error. For the additive measurement error model $\boldsymbol{X}^* = \boldsymbol{X} + \boldsymbol{U}$, the naive ordinary least square estimator consistently estimates $(\boldsymbol{\beta}_x^*, \boldsymbol{\beta}_z^*)$ different from $(\boldsymbol{\beta}_x, \boldsymbol{\beta}_z)$ with

$$\begin{pmatrix} \boldsymbol{\beta}_x^* \\ \boldsymbol{\beta}_z^* \end{pmatrix} = \begin{pmatrix} \Sigma_{xx} + \Sigma_{uu} & \Sigma_{xz} \\ \Sigma_{zx} & \Sigma_{zz} \end{pmatrix}^{-1} \left\{ \begin{pmatrix} \Sigma_{xx} & \Sigma_{xz} \\ \Sigma_{zx} & \Sigma_{zz} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_x \\ \boldsymbol{\beta}_z \end{pmatrix} + \begin{pmatrix} \Sigma_{u\epsilon} \\ \mathbf{0} \end{pmatrix} \right\},$$

where Σ_{st} is in general the covariance matrix between S and T. In general when more than one covariate is measured with error in a linear regression model the effect of a naive analysis is unpredictable.

For generalized linear models, such as logistic regression and Poisson regression, the effect of measurement error is much the same as in linear regression. This implies that a relative risk or an odds ratio is affected in largely the same way as coefficients in a linear regression model.

1.4.1.3 Correction for measurement error

Below we give a short description of the models for adjusting for measurement error, which are grouped into functional and structural models.

Functional and structural models: Methods for correcting for measurement error are classified into two broad classes, namely *functional* and *structural* models. Consider a measurement error model relating Y to X with possibly error-prone X^* observed instead of X. Models with fixed X are called functional models, whereas models in which Xs are

regarded as random variables are called structural models. This leads to *functional modeling* with only minimal assumptions made about the distribution of Xs. Further, in *structural modeling*, parametric models are assumed for the distribution of the random Xs. At first sight functional modeling is to be preferred because it is based on only few assumptions. Structural modeling, on the other hand, allows for more complex models through the distributional assumption of X. Functional approach is quite efficient in some problems, e.g., linear regression with additive measurement error. On the other hand, in other problems for example regression analysis with replicated data, structural modeling is more efficient.

First, we describe examples of functional models, including the method-of-moments, regression calibration and simulation extrapolation (SIMEX). Secondly, we present structural methods which in general are likelihood-based using either a frequentist or a Bayesian approach.

Functional modeling: Suppose that the reliability ratio λ is known in simple linear regression then an estimate of β_x corrected for measurement error is given by dividing the ordinary least squares estimate by $\hat{\beta}_{x^*}/\lambda$, where $\hat{\beta}_{x^*}$ is the naive ordinary least square estimate of β_x. If the reliability ratio is unknown and we are given the estimates of the measurement error variance, $\hat{\sigma}_u^2$, and the sample variance of X^*, $\hat{\sigma}_{x^*}^2$, then the consistent estimate of the reliability ratio is $\hat{\lambda} = (\hat{\sigma}_{x^*}^2 - \hat{\sigma}_u^2)/\hat{\sigma}_{x^*}^2$. Thus the estimate of $\hat{\beta}_x$ is $\hat{\beta}_{x^*}/\hat{\lambda}$. This approach is known as *method-of-moments*, because the ordinary least squares estimate of $\hat{\beta}_{x^*}$ and the reliability ratio $\hat{\lambda}$ depend only on the moments of the observed data. This method can be extended to a generalized linear model (GLM), see, e.g., Fuller (1987) and Carroll et al. (1995).

The most popular and also more general functional approaches are *regression calibration* and *simulation extrapolation* (SIMEX). Rosner, Willett, and Spiegelmann (1989, 1990) developed regression calibration methods for logistic regression. Further, Carroll and Stefanski (1990) and Gleser (1990) describe the general idea about the regression calibration method. The rationale behind regression calibration is to replace X by the predicted value of X from a regression of X on (X^*, Z). The intuition about regression calibration is best seen in simple linear regression: $Y = \beta_0 + \beta_x X + \beta_z Z + \epsilon$ with $X^* = X + U$ where, given X, U is independent of (Y, Z). Then

$$\begin{aligned}
\mathrm{E}[Y|X^*, Z] &= \mathrm{E}\left\{\mathrm{E}[Y|X^*, X, Z]|X^*, Z\right\} \\
&= \mathrm{E}\left\{\mathrm{E}[Y|X, Z]|X^*, Z\right\} \\
&= \mathrm{E}\left[\beta_0 + \beta_x X + \beta_z Z|X^*, Z\right] \\
&= \beta_0 + \beta_x \mathrm{E}\left[X|X^*, Z\right] + \beta_z Z
\end{aligned}$$

Thus, X is replaced by an estimate $\widehat{X}$ from a regression model predicting X from X^* and Z.

Cook and Stefanski (1994) suggested the SIMEX (acronym for simulation extrapolation) method. SIMEX estimates are obtained by adding additional measurement error to the data in a resampling stage, creating a measurement error trend, and extrapolating back to the case with no measurement error. This approach is intuitively appealing and relatively easy to implement. However, it can be sensitive to the choice of the extrapolation function (see, e.g., Küchenhoff and Carroll, 1997).

Structural modeling: Carroll et al. (1984) consider a binary regression model when some of the covariates are measured with error. This is an

example of random effects modeling because the marginal likelihood is a product of the observed data likelihood and the distribution of the error. The unobserved covariates measured with error can be assumed independently and normally distributed. Given the sample estimates of the mean and the covariance matrix of the corresponding normal distribution then the binary regression is completed by relating the covariates to the response via either a logit or a probit link function. This structural approach, however, requires replication in estimating the measurement error variance. Note that the functional maximum likelihood estimate of the regression coefficient from this binary regression is not consistent, even when the measurement error variance is known.

Other examples of likelihood based structural models include Schafer (1993) for probit regression with measurement error in a covariate, Richardson and Gilks (1993b) and Wang, Carroll, and Liang (1996) for logistic regression with a mismeasured covariate in epidemiological studies, Küchenhoff and Carroll (1997) in change of point problems with a mismeasured covariate and Carroll, Roeder, and Wasserman (1999) on flexible parametric for measurement error in a covariate. Examples of structural models for logistic regression with a mismeasured covariate using a Bayesian approach include Stephens and Dellaportas (1991), Schmid and Rosner (1993), Richardson and Gilks (1993a) and Müller and Roeder (1997).

1.4.2 Misclassification

In this section we look on the effect of misclassification in a covariate and a response. Further, we provide a brief overview on the correction for misclassification. Note that the definition of non-differential and differential misclassification is similar to that of measurement er-

ror (Section 1.4.1.1), only that for misclassification the variables are categorical.

1.4.2.1 Effect of misclassification

Below we explore the effect of (a) response misclassification on the estimation of the difference of population proportions in two exposure groups, and (b) binary covariate misclassification on estimation of a regression coefficient in simple linear regression.

Misclassification in response: Let Y^* denote a binary response variable subject to misclassification, and Y the true binary response variable. The misclassification process is expressed by the misclassification probabilities

$$\Pr\left(Y^* = j|Y = k\right) \;=\; \pi_{jk}, \;\; j,k = 0,1.$$

The misclassification probabilities may be collected in the following 2×2 matrix:

$$\Pi = \begin{pmatrix} \pi_{00} & 1-\pi_{11} \\ 1-\pi_{00} & \pi_{11} \end{pmatrix}. \tag{1.1}$$

Under non-differential misclassification, the effect of misclassification is described by *sensitivity* and *specificity* of Y^* as a proxy measure of Y. The parameter $\pi_{11} = \Pr(Y^* = 1|Y = 1)$ is called the sensitivity of the measuring instrument and $\pi_{00} = \Pr(Y^* = 0|Y = 0)$ is the specificity.

Suppose the interest focuses on the difference between two population proportions, say $\delta = (\pi_1 - \pi_2)$ where π_j $(j = 1, 2)$ is the proportion of population j. For example this could be the difference of population proportions of disease cases between two exposure groups. The

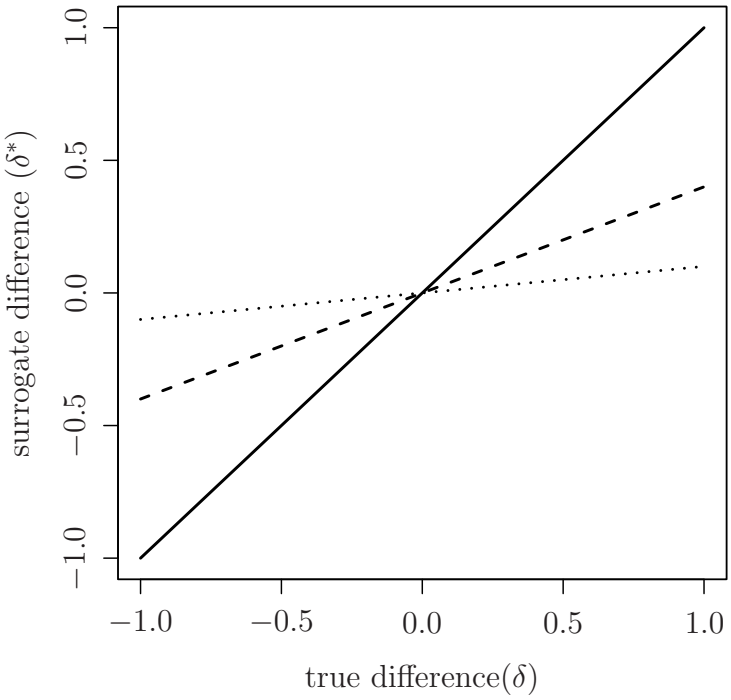


Figure 1.4: *This plot shows the effect of sensitivity* π_{11} *and specificity* π_{00} *on the difference of proportions between two exposure groups for a misclassified disease measurement. The larger the misclassification, corresponding to smaller values of* π_{00} *and* π_{11}*, the more the line (surrogate difference) is pulled towards zero. In this example* $\pi_{00} = \pi_{11} = 0.8$ *(solid line);* $\pi_{11} = 0.6$*,* $\pi_{00} = 0.7$ *(dashed line); and* $\pi_{00} = 0.5$*,* $\pi_{11} = 0.6$ *(dotted line).*

sample proportion p_j $(j = 1, 2)$ is an unbiased point estimator of π_j if Y is measured without error. However, if Y is recorded with misclassification, characterized by misclassification matrix (1.1), it can be shown that the difference between two proportions is

$$\delta^* = (\pi_{00} + \pi_{11} - 1)\delta.$$

In practice, the sensitivity (π_{11}) and specificity (π_{00}) are often larger than 0.5, so that the surrogate difference is always smaller than the

true difference.

Figure 1.4 demonstrates the effect of π_{00} and π_{11} on the estimation of the difference of proportions between two populations. Thus, the effect of non-differential misclassification is to attenuate δ towards the null.

A similar attenuation effect of non-differential misclassification mechanism is seen for other risk measures like a relative risk or an odds ratio. However, if we assume that sensitivities and specificities vary over the covariate groups, i.e. under differential misclassification, the bias can be away or towards the null. Summarized, in general, the effect of response misclassification is similar to the effect of measurement error in response.

Misclassification in covariates In this section we explore the effect of covariate misclassification in linear models. For example, consider a continuous response Y and binary predictor X. We can write

$$\mathrm{E(Y|X)} = \beta_0 + \beta_{\mathrm{x}}\mathrm{X}.$$

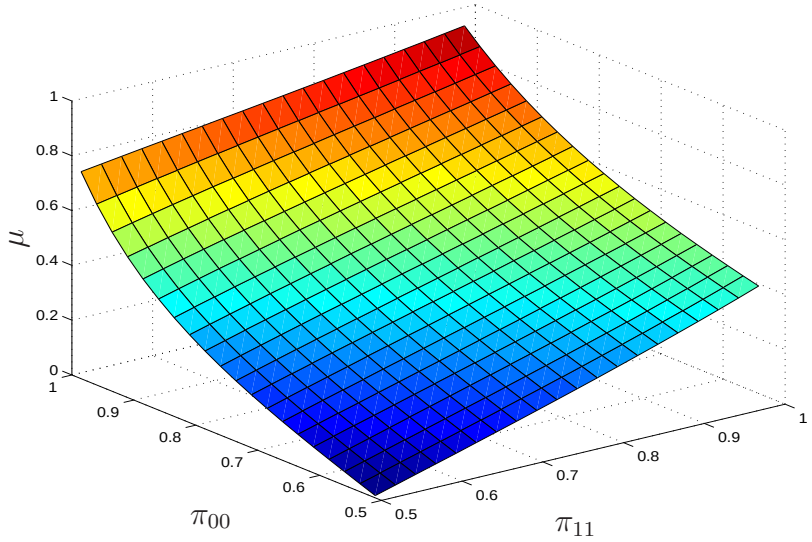
Now suppose that one observes X^*, a misclassified version of X with misclassification matrix (1.1).

It can be easily shown that the naive regression of Y on X^* estimates not β_x but $\beta_{x*} = \mu\beta_x$ with (e.g., Gustafson, 2004)

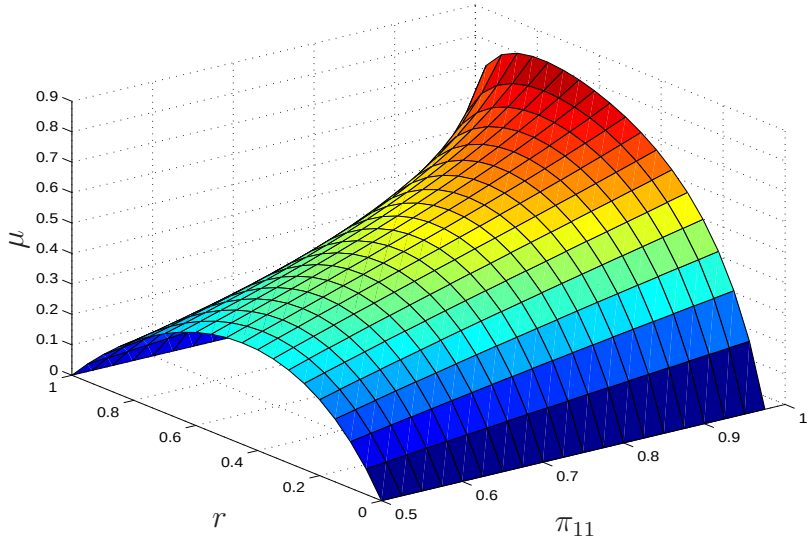
$$\mu = (\pi_{11} + \pi_{00} - 1)\frac{r(1-r)}{r^*(1-r^*)},$$

where $r = \Pr(X = 1)$ and

$$r^* = \Pr(X^* = 1) = (1 - \pi_{00}) + (\pi_{11} + \pi_{00} - 1)r.$$



(a) *Varying* π_{11} *and* π_{00} *for* $r = 0.3$.



(b) *Varying* r *and* π_{11} *for* $\pi_{00} = 0.8$.

Figure 1.5: *The effect of* π_{11}, π_{00} *and* r *on attenuation of the regression coefficient in simple linear regression with misclassified binary covariate.*

Note that the attenuating factor, μ, is also a function of the prevalence of the risk factor.

As can be seen in Figure 1.5(a), for a fixed r, the higher the sensitivity or specificity the higher the attenuation factor. Further, as shown in Figure 1.5(b), the attenuation factor varies quadratically as a function of r. In particular, when sensitivity is equal to specificity the attenuation factor is higher and symmetric around $r = 0.5$.

In general, for any set $\{r, \pi_{00}, \pi_{11}\}$ such that $\pi_{00} + \pi_{11} - 1 < 1.0$, the attenuation factor is less than 1 implying that the effect of non-differential misclassification in the binary covariate is again to attenuate the estimate of the regression coefficient to the null. Therefore, the effect of misclassification of a categorical covariate is similar to measurement error in a continuous covariate.

1.4.2.2 Correction for misclassification

The most straightforward approach to correct for misclassification is via the *matrix method*. This method follows a functional modeling approach since no distribution assumptions are made about the possibly error-prone variable. The matrix method exploits the relation $\boldsymbol{p}^* = \boldsymbol{\pi}\boldsymbol{p}$, where $\boldsymbol{p}^*$ and $\boldsymbol{p}$ are the observed proportions for the possibly corrupted and the true categorical variable in the main data, respectively and $\boldsymbol{\pi}$ is the misclassification matrix. The matrix method estimate of $\boldsymbol{p}$ is given by $\hat{\boldsymbol{p}} = \hat{\boldsymbol{\pi}}^{-1}\hat{\boldsymbol{p}}^*$. The main analysis is performed on the transformed probabilities $\hat{\boldsymbol{p}}$. A variance estimator for $\hat{\boldsymbol{p}}$ can be obtained using the *delta-method* (Greenland, 1980). The matrix method is straightforward to compute, but it has the disadvantage that the estimated probabilities are not constrained to lie between 0 and 1, a consequence of matrix inversion.

Gustafson (2004) describes an example of structural modeling in which a misclassified categorical covariate X is treated as a random variable in a Bayesian model. The misclassified covariate X is assumed to have two imperfect surrogate variables X_1^* and X_2^*, which are assumed conditionally independent given the actual variable X. The resulting likelihood is in fact a product of binomial probabilities for observing the pair (X_1^*, X_2^*). The Bayesian estimation of the model is simplified by including the unobserved covariate X into the posterior, i.e. by assigning prior distributions to the unknown quantities. This is an example of data augmentation. Other examples of structural models include: Liu and Liang (1991) for the misclassification of categorical covariates, Whittemore and Gong (1991) for the misclassified Poisson counts, Greenland and Brenner (1993) and Prescott and Garthwaite (2002) using a frequentist approach to misclassification of a binary risk factor, and Neuhaus (1999, 2002) using a frequentist approach to data subject to response misclassification.

1.5 Validation studies and the gold standard

1.5.1 Validation studies

An ideal epidemiological studies would measure all variables without error. Unfortunately, this is for many studies not possible for various reasons. For example, the error-free measurement may be very expensive to apply to all subjects in the main study. Further, due to practical or ethical considerations, it may not be possible to measure all study subjects. Consequently, measurements are taken by a standard instrument subject to error.

To correct for attenuation in the parameter estimates, one needs information about the error structure. If the measurement error struc-

ture is known the correction is straightforward. In general, however, the measurement error structure is unknown and is estimated from a suitable subset of auxiliary data. The auxiliary data sources can be grouped into two broad categories: an *internal* subset of the main study, and an *external* set of independent studies. Within each of these two broad categories there are three data types, namely, validation data, replication data and instrumental data.

1.5.1.1 Validation data

In a validation study the true variable is observed together with the possibly corrupted data. An internal validation is ideal because it allows for the direct examination of the error structure and typically leads to much greater precision in parameter estimation. On the other hand, with external validation one must assume that the same error structure also applies to the main data, i.e. a subset of the parameters of the measurement error model could be transportable. For example, if examiners with a similar level of training are working in different centres, then it is reasonable to expect that the distribution of the error is independent of the centre and the examiner making the measurement.

An internal validation study will be referred to below as a validation study. In a validation study, the response (or explanatory) variable of each subject is measured by two methods: first, by the primary but error-prone method used on all subjects in the main study; and second, by a 'gold standard' method used on only a subset of subjects in the main study. The validation data therefore contain the true and error-prone variable(s), possibly with other correctly measured variables.

1.5.1.2 Replication data

In replication data, replicates of the mismeasured response/covariate are available. Replicated or repeated data are useful when it is impossible to measure the true variable exactly. For example, when a covariate measured with error represents long-term systolic average blood pressure or long-term average nutrient intake. Usually, one would make replicate measurements if there were good reasons to believe that the replicated mean is a better estimate of the true variable than a single observation. In additive measurement error models, replication data can be used to estimate the variance of the measurement error.

1.5.1.3 Instrumental data

It is not always possible to obtain replicates or validation data and thus direct estimation of the measurement error variance is sometimes impossible. In the absence of information about measurement error variance, estimation of the regression model parameters is still possible provided that the data contain an instrumental variable, in addition to the mismeasured covariate. The instrumental variable is a second surrogate measurement of the mismeasured variable obtained by an independent method. This additional variable is correlated with the true explanatory variable but not measured with error, and can be used to predict the true explanatory variable. For example, in an epidemiologic study of skin cancer and arsenic exposure, the water arsenic measurements are possibly error-corrupted while the toenail arsenic measurements are interpreted as the instrumental variable.

1.5.2 Gold standard versus benchmark scorer

A key assumption in a validation study is that true data are measured by a *gold standard*, a measuring instrument that is assumed to be error-free. However, in practice the measurements are often made by a *benchmark scorer*, an experienced examiner or a tested measuring instrument which is assumed to be error-free or is nearly so, in relation to error-prone measurement. In this case, the corrected overall study results are valid since there is no distinction between the measurements of the benchmark scorer and the gold standard.

A benchmark scorer may, however, be an 'alloyed'-gold standard, i.e. a too imperfect approximation of the gold standard. In this situation, the corrected estimates are attenuated. For example, when the errors of the alloyed-gold standard and the usual method of measurement are negatively correlated, the corrected estimates will tend to overcorrect beyond the true value (Wacholder, Armstrong, and Hartge, 1993). This finding suggests that it is possible that the methods for correcting estimates for the effect of measurement error might introduce more bias than they are correcting.

Spiegelman, Schneeweiss, and McDermott (1997) give an example of using medical records to validate the self-reported prescription drug use, in the study of exogenous oestrogen as a risk factor for endometrial cancer. This may not be a true measurement, because the apparent low accuracy of the self-report may result from errors in the records rather than in the self-reports.

This is a clear warning to safeguard against using any measuring instrument, no matter how much better than the available measure, as a proxy to a gold standard. Nevertheless, an experienced benchmark scorer (or a duly tested measuring instrument) can replace the gold standard if the resulting measurements are believed or shown to match

that of the gold standard with negligible error.

CHAPTER 2

Caries Research

2.1 Dental caries

Dental caries (decay) is in fact the dissolution (demineralisation) of tooth enamel by acids from bacterial origin. Bacteria, e.g., *Streptococcus mutans* species present in dental plaque (a bio-film of bacteria, food remnants and other debris) metabolize dietary sugars to produce organic acids which attack the enamel surface of the tooth. On the other hand, exposure to alkali, such as sodium bicarbonate in saliva, reverses this process and aids in remineralisation. Therefore, enamel is in a dynamic state of de- and re-mineralisation, and when demineralisation predominates, cavitation occurs (see, e.g., Moyniham, 2000). Cavitation of the tooth enamel surface occurs, followed by the spread into the dentine and then eventually to the pulp.

In most westernized countries, the occurrence of dental caries has declined over the past decades (see, e.g., Spencer, 1997; Carvalho, van Nieuwenhuysen, and D'Hoore, 2001; Pieper and Schulte, 2004) leading to a skewed distribution of the degree of caries experience.

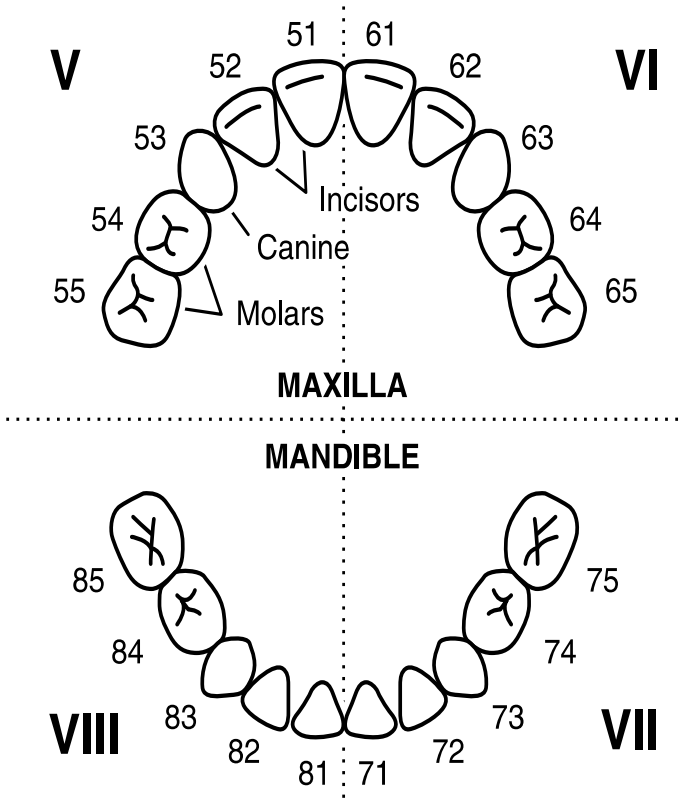


Figure 2.1: *Tooth indices for the primary dentition.*

For this reason, the majority of children have no or only a few cavitated lesions, while for a minority the degree of caries experience still remains relatively high (Hausen, 1997). The polarized distribution of the degree of caries experience heeds for high risk strategies to target preventive care or treatment to those individuals or groups identified as being at high risk for future caries development.

The first set of teeth appearing in the mouth are called deciduous teeth. A child can have maximally 20 deciduous teeth. Figure

2.1 shows the European way of indexing deciduous teeth. A deciduous tooth can have four or five surfaces as shown in Figure 2.2. The incisors and canines have four surfaces, namely mesial (M), buccal (B), distal (D) and lingual (L), whereas molars have an occlusal (O) surface in addition. Molars are more likely to become decayed than other teeth because the occlusal surface (due to the complex anatomy) is most vulnerable to a caries attack. Consequently, there could be a considerable different susceptibility to caries development of the different teeth in the mouth.

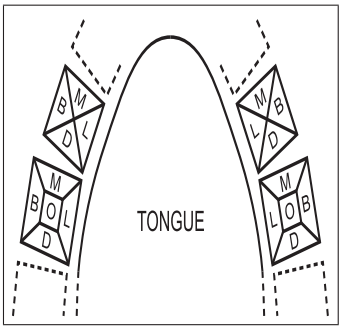


Figure 2.2: *Tooth surfaces label*

Saliva, bacteria and diet play an important role in caries activity. Consequently, there is a variation in the degree of dental caries experience between subjects caused by differences in biological factors like the bacterial strains in the mouth and differences in the level of oral hygiene and dietary habits. Demographic, behavioral, and environmental factors may also be associated with caries activity. For example, tap water may contain fluoride and since fluoride is assumed to have a protective effect against caries development, geographical differences in caries experience may occur.

2.2 Measuring the degree of caries experience

The dmft-index is probably the simplest but certainly the most commonly used index for reporting dental caries experience. The dmft index quantifies the dental health status based on the total sum of decayed (d), missing due to extractions because of caries (m) and filled (f) teeth in the primary dentition (see, e.g., Klein, Palmer, and

Knutson, 1938; Klein and Palmer, 1941). An analogous index, the dmfs-index, is the sum of the surfaces which are either decayed, missing or filled due to caries. The dmft (dmfs) score for any individual can range from 0 to 20 (0 to 88). The lower the index, the better the dental health of the subject. For permanent teeth, the corresponding index is DMFT for caries on teeth level and DMFS for caries on surface level. The DMFT (DMFS) score for any individual can range from 0 to 32 (0 to 148).

The dmft(s)/DMFT(S) indices have some limitations. First, the dmft (DMFT) index (and also the dmfs (DMFS)) do not indicate which teeth are affected such that much detail is lost by working with this aggregate measure. Secondly, the dmft(s)/DMFT(S)-index is an exaggerated score for caries experience because, if a tooth is missing, all surfaces of that tooth are considered as diseased (Benigeri, Payette, and Brodeur, 1998). Nevertheless, the indices provide a reasonably accurate history of changes in the degree of dental caries experience because of their widespread use over the past decades. Thus, at least for this reason, we are motivated to examine these indices in this thesis.

Measuring caries experience is a yes/no decision as to whether caries as a disease is present in, for example, a particular tooth surface. The result obtained depends on the *diagnostic threshold*, i.e. an arbitrary cut-off level, to decide on what to classify as diseased and what to classify as sound. Caries experience scoring is based on four levels of lesion severity: d_4 (dentine caries with pulpal involvement), d_3 (limited dentine caries), d_2 (enamel cavity) and d_1 (white or brown-spot initial lesions without cavitation) as displayed in Figure 2.3. See, for example, Pitts and Fyffe (1988), Fyffe et al. (2000) and Pitts (2004). These scores permit the calculation of the dmft/dmfs indices at vari-

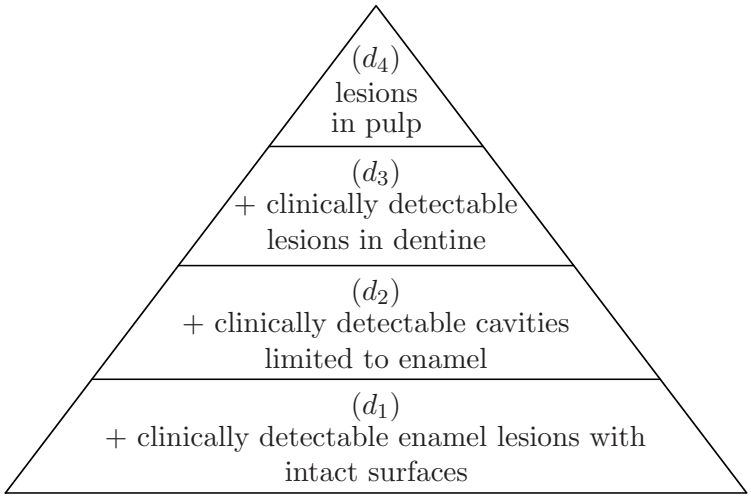


Figure 2.3: *Caries diagnostic thresholds.*

ous levels of diagnosis (d_1, d_2, d_3, d_4). Note that the severity of caries experience will not only depend on the level of diagnosis used, but also on the diagnostic instrument used: visual, radiological or advanced diagnostic tools. The accuracy of the measurement of interest will thus vary depending on which level of lesion severity is used.

2.3 Signal Tandmobiel® study

2.3.1 Sample and study design

The Signal Tandmobiel® study[1] is a population-based study. This study involved a sample representative of 4468 Flemish primary school-

[1]The partners in this collaborative project are: the Dental Departments (Paediatric and Preventive Dentistry) of the Catholic University of Leuven, the University of Ghent and the Free University of Brussels; the Youth Health Department of the Catholic University of Leuven; the Biostatistical Centre of the Catholic University of Leuven; the Working Group of Oral Health of the Flemish Dental Association and the Flemish Association for Youth Health Care.

children (2315 boys and 2153 girls). The sample represents 7.3% of the children born in 1989 in Flanders and first examined in 1996. At the first examination the average age of the children was 7.1 years (standard deviation = 0.4) and varied from 6.12 years to 8.09 years.

The children were randomly drawn through stratified cluster sampling without replacement. The selection units were the schools, stratified by province and educational system. Thus, the target population was divided into 15 different strata, comprising the three types of the Belgian educational system, namely private schools (mainly catholic schools), state schools and municipal schools, for the five provinces of Flanders. Schools were selected with a probability proportional to the number of children in the first year of primary school. By this sampling scheme, the children were selected with equal probability. They were examined annually during their primary school time (1996–2001). They underwent an oral health examination in a mobile dental clinic by one of the 16 examiners and the parent completed a questionnaire.

Three groups were sampled as shown in Figure 2.4. For Group A, the intervention group (initially 4468 children) – the children were followed over 6 consecutive years and they were subjected to an oral health educational programme. In Group B, a longitudinal control group (initially 520 children) of children age matched to group A – the children were examined in the first and sixth year but did not receive the oral health educational programme. Group C, comprises a different cross-sectional control group each year (approximately 500 children) – each year a different group was selected and they were included to be used for a cross-sectional comparison with the longitudinal A group. Note that the first year children of group B also serve as the first C group.

In this thesis we will use only the data from Group A. Hence, the

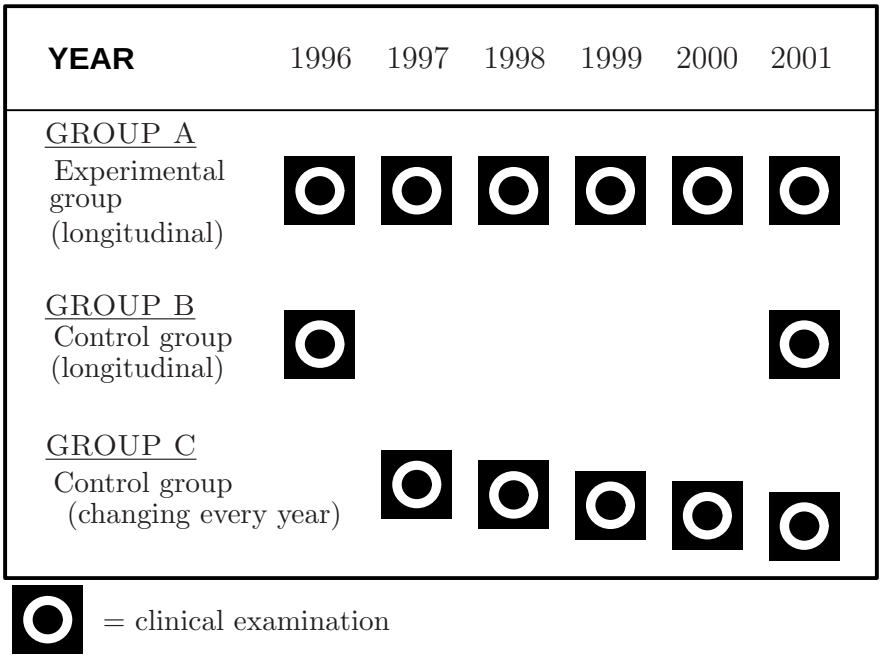


Figure 2.4: *Study groups and examination periods for children participating in the Signal Tandmobiel® study.*

data from Group A will be referred to below as the Signal Tandmobiel® data. In this study caries experience was measured at level d_3 of lesion severity (see Section 2.2).

Clinical data were collected by dental examiners on the level of oral hygiene, gingival condition, dental trauma, prevalence and extent of enamel developmental defects, fluorosis, tooth decay, the presence of restorations, missing teeth, the stage of tooth eruption and orthodontic treatment need, all using established criteria, as recommended by the WHO report (1987) and based on the diagnostic criteria for caries prevalence surveys published by the British Association for the Study of Community Dentistry (BASCD) (Pitts, Evans, and Pine, 1997). Besides the detailed data, information was also collected on oral hygiene

and dietary habits, the use of fluorides, dental attendance, medical history and social demographic background of the children. These extra data were obtained from questionnaires completed by parents and school medical centres. For a more detailed description of the Signal Tandmobiel® study we refer to Vanobbergen et al. (2000).

2.3.2 Variables included

For most of the analyses, the degree of caries experience in the primary dentition, i.e. the dmft-index will be used in this thesis as a response variable in a regression model. In particular, we will treat the response either as a categorical (binary/ordinal) or as a count variable. Throughout the thesis, we are interested in establishing the determinants of caries experience taking into account the potential misclassification by the dental examiners involved in the study with respect to the benchmark examiner. The scoring was based on the d_3 level of caries experience.

Various explanatory variables were considered. We considered: age (years), gender (girl = 1), geographical location (in terms of the $x-$ and $y-$coordinates) of the school that the child attends, age at start of brushing (years), systemic use of fluorides (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).

Observe that the geographical components expressed in terms of $x-$ and $y-$coordinates represent also the stratification variable 'province'. 'Age at start of brushing' reflects the age of the child when it first started brushing while the systemic use of fluoride means that tablets or drops were regularly used as fluoride supplements. Further, daily consumption of sugar containing drinks was compared to the use of

water, milk, etc.

2.4 Calibration exercises

Dentist-examiners were specifically trained at baseline and participated in calibration exercises according to the guidelines of training and calibration published by the British Association for the Study of Community Dentistry (see, Pine, Pitts, and Nugent, 1997). The examiners were calibrated by observing the same group of children (all tooth surfaces) and comparing their results to the scores obtained by the benchmark examiner. Calibration exercises were conducted in children with a variety of pathology present, including untreated caries, recurrent caries and fillings, nevertheless making sure that some caries free children were also included.

In the Signal Tandmobiel® study much attention was paid to the selection and training of the 16 dental examiners. In order to maintain a high level of intra- and inter-examiner reliability, calibration exercises were carried out twice a year for all examiners involved. During the study period (1996-2001), 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 each dental examiner vis-a-vis the benchmark examiner (Dominique Declerck)[2] was determined. In this study, since the scoring of caries experience was based on d_3 level then it implies that the benchmark scorer is approximate to gold standard. The detection of d_1 and d_2 lesions requires a more careful clinical diagnosis and the evaluation of radiographs, and hence benchmark scorer would be far from gold standard.

[2]Professor, School of Dentistry, Catholic University Leuven, Belgium.

2.5 The validation data of the Signal Tandmobiel® study

The validation data of the Signal Tandmobiel® study is obtained from the above calibration exercises. In the Signal Tandmobiel® study sixteen dental examiners were involved. The three calibration exercises for scoring caries experiences yield a misclassification table for each examiner.

In the literature it is often recommended that the validation data constitute a random sample of the main study, thereby establishing an internal validation data set. But in the Signal Tandmobiel® validation data the children were not sampled at random from the main study. Rather, a school was selected (and all seven-year old children examined) where a relatively high prevalence for caries experience could be expected. Although the validation study is not internal, since the children belong to the same population as those of the main study and the dental examiners are also the same, the misclassification probabilities can be unbiasedly estimated using the validation data. But, for these children no questionnaire data were available so their true scores could not be included in likelihood.

In Chapters 4, 5 and 6 the validation data is combined from the three caries calibration exercises. In Chapter 7 and the succeeding chapters, we considered the validation data of the first calibration exercise.

2.6 Some preliminary results

Vanobbergen et al. (2001) fitted a logistic regression model to the dichotomised (dmft $= 0$ versus dmft $\neq 0$) response on a set of risk

indicators (described in Section 2.3.2), but instead of the $x-$ and $y-$ coordinates they used 4 dichotomous variables to indicate the five Flemish provinces to which the school of the child belonged. The variables found to be playing an important role as determinants of caries experience in the primary dentition in Flanders for the seven-year old children are: geographic area (province), oral hygiene habits (age at start of brushing, frequency of brushing), the use of fluorides supplements and fluoridated toothpaste, and dietary habits (daily use of sugar containing drinks between meals, in-between-meal snacks). Further, the geographical variable (province) showed that living in the Western part of Flanders seems to be a protective factor. The

Table 2.1: *Maximum likelihood estimates of the multiple logistic regression model fitted to the binary dmft score (caries experience versus no caries experience)*

Parameter	Estimate†	95% CI‡	
Intercept	−0.398(0.119)	−0.632	−0.163
x-ordinate	0.195(0.037)	0.122	0.268
y-ordinate	−0.025(0.037)	−0.098	0.048
Gender (girl)	0.028(0.072)	−0.113	0.169
Age (years)	0.326(0.089)	0.151	0.500
Brushing frequency (< 2 days)	0.229(0.107)	0.018	0.439
Age at start brushing (years)	0.200(0.035)	0.132	0.268
Fluoride supplement (yes)	−0.446(0.072)	−0.588	−0.304
Sugary drinks (yes)	0.310(0.074)	0.166	0.455
Between meals (> 2)	0.210(0.078)	0.056	0.363

†Standard error in parenthesis.
‡CI = Confidence interval.

probability of remaining caries free was 14% higher for children living in the Western part compared with the Eastern part of Flanders.

Table 2.1 shows the results of fitting multiple logistic regression controlling for the geographical component using $x-$ and $y-$coordinates. The results clearly indicate a significant geographical trend, i.e. the East-West gradient in prevalence of caries experience, with higher scores in the Eastern part of Flanders (the province of Limburg). The positive regression coefficient for age implies that the older the child the higher the risk of having caries. Similarly, the positive regression coefficient for age at start of brushing implies that the later the child starts brushing the higher the risk of having caries. Daily use of sugar containing drinks between meals and the consumption of more than two between-meals snacks per day are also important risk factors. Moreover, regular use of fluoride supplements and brushing more than once a day, are important protective factors.

CHAPTER
3 An Overview of Frequentist and Bayesian Approaches Applied to Random Effects Logistic Regression

3.1 Introduction

In this chapter we present an overview of frequentist and Bayesian approaches with application to the random effects logistic regression. The rationale is to briefly explore the methods used for parameter estimation throughout the thesis, and also to highlight the software used in model fitting.

Further, we study the random effects logistic regression not only because of its simple form but also due to its wide application in epidemiological analysis. In addition, it serves as a typical introduction to the models described in Chapters 4 and 5.

3.2 The logistic random-effects model

Examples of clustered data include repeated measurements in longitudinal studies, multiple observations in the same mouth in oral health research and family data in genetic studies. It is well known that observations within a cluster are more alike than observations of different clusters. Random effects logistic regression model is an extension of classical logistic regression model, in which random effects are introduced into the linear predictor, to model the clustering effect.

Specifically, suppose we have binary responses Y_{ij}, $i = 1, \cdots, N$ and $j = 1, \cdots, n_i$, where Y_{ij} denotes the jth measurement on the ith subject (cluster). The random-intercept logistic regression model has the following form:

$$\log\left(\frac{\Pr(Y_{ij}=1|u_i)}{1-\Pr(Y_{ij}=1|u_i)}\right) = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + u_i, \tag{3.1}$$

where $\boldsymbol{x}_{ij}$ is a $d-$dimensional vector of covariates with $\boldsymbol{\beta}$ the corresponding vector of regression coefficients, and u_i is a random variable expressing the ith subject-specific level. The term u_i is called a random intercept (of the ith subject) and it is assumed to follow a standard normal distribution with mean 0 and variance σ^2, i.e. $u_i \sim \mathcal{N}(0, \sigma^2)$. The random effects can be expressed in a standardized form, namely $u_i = \sigma z_i$, $\sigma > 0$, where $z_i \sim \mathcal{N}(0, 1)$. Thus, model (3.1) may be rewritten as

$$\log\left(\frac{\Pr(Y_{ij}=1|z_i)}{1-\Pr(Y_{ij}=1|z_i)}\right) = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \sigma z_i, \quad \sigma > 0. \tag{3.2}$$

It is assumed that, conditionally on z_i, the likelihood terms involving the ith subject are independent. The evaluation of the marginal likelihood (expressing likelihood of the observed vector of responses)

for the ith subject involves integrating out the random intercept and is equal to

$$L_i \equiv L(\boldsymbol{Y}_i|\boldsymbol{\beta}, \sigma) = \int \prod_{j=1}^{n_i} \Pr(Y_{ij} = y_{ij}|\boldsymbol{\beta}, \sigma, z_i)\phi(z_i)dz_i, \tag{3.3}$$

where $\boldsymbol{Y}_i = (Y_{i1}, \cdots, Y_{in_i})'$ is the random vector of measurements for the ith subject, y_{ij} is the corresponding observed value of the response, and $\phi(\cdot)$ is the standard normal density. The total (marginal) likelihood function for $\boldsymbol{\beta}$ and σ is the product of the N terms in expression (3.3) and hence

$$L \equiv L(\boldsymbol{Y}|\boldsymbol{\beta}, \sigma) = \prod_{i=1}^{N} L(\boldsymbol{Y}_i|\boldsymbol{\beta}, \sigma) = \prod_{i=1}^{N} L_i, \tag{3.4}$$

where $\boldsymbol{Y} = (\boldsymbol{Y}'_i, \cdots, \boldsymbol{Y}'_N)'$ is the total vector of the responses.

Calculating the marginal likelihood requires the evaluation of integral (3.3). However, there is no analytic solution available and thus numerical approximations are needed. For parameter estimation we consider both a frequentist and a Bayesian approach, as described below.

3.3 Frequentist approach

The parameters of the random-effects logistic regression can be estimated by maximizing the marginal likelihood (3.4). Among the various methods of approximating the integrals that have been proposed in the literature the most widely used is the Gauss-Hermite quadrature, which we present in the next section.

3.3.1 Gaussian quadrature

The Gauss-Hermite quadrature procedure implies that the integral (3.3) is approximated by the weighted sum

$$L_i = \int_{-\infty}^{\infty} \Pr(z)\phi(z) \approx \sum_{q=1}^{Q} w_q \Pr(z_q), \tag{3.5}$$

where

$$\Pr(z) = \prod_{j=1}^{n_i} \Pr(Y_{ij} = y_{ij}|\boldsymbol{\beta}, \sigma, z),$$

with z_q denoting the nodes (or quadrature points), which are solutions to the Qth order Hermite-polynomials and w_q are appropriately chosen weights. In general, the higher the order Q the more accurate the approximation will be. The main disadvantage of the (non-adaptive) Gaussian quadrature is that the quadrature points z_q are based on $\phi(z)$, which is independent of $\Pr(z)$ in the integral. That is, $\Pr(z)$ is calculated at points $\hat{z}$ irrespective of the range of where $\Pr(z)$ is relatively high. In general, a larger number of quadrature points (Q) is necessary for adequate accuracy when the variance of the random effect is large.

It is useful to rescale and shift the quadrature points such that more quadrature points lie in the region of interest. This is done in the so-called adaptive Gaussian quadrature.

3.3.2 Adaptive Gaussian quadrature

Adaptive Gaussian quadrature centers the quadrature nodes with respect to the mode of the function being integrated, and scales them according to the estimated curvature at the mode. Liu and Pierce (1994) suggested an adaptive Gauss-Hermite technique using the adaptive

quadrature idea (see also Molenberghs and Verbeke, 2005).

Typically, adaptive Gaussian quadrature needs fewer evaluation points than non-adaptive Gaussian quadrature. However, adaptive Gaussian quadrature requires the calculation of the mode for each individual in the data set, hence a numerical maximization of functions of the form (3.5). This implies that adaptive Gaussian quadrature is in general much more time consuming than non-adaptive Gaussian quadrature.

Both the non-adaptive and adaptive Gaussian quadrature procedure have been implemented in the SAS procedure `NLMIXED` (SAS© Institute Inc., 1999–2001). Results from the frequentist analyses will not be shown in the cases where they are practically the same as those from the Bayesian analyses.

3.4 Bayesian approach

The Bayesian approach differs from the frequentist approach in treating parameters as random variables and using data to update prior knowledge about the parameters. The prior distribution of the parameters expresses the prior knowledge of these parameters. An *informative* prior expresses specific or definite information about a variable, whereas an *uninformative* prior expresses vague information about a variable. When more than one parameters are involved, often the prior distribution of each parameter is assumed to be a priori independent. In the random-effects logistic regression model a prior distribution needs to be specified for $\boldsymbol{\beta}$ and σ^2. Also here we assumed prior independence. The corresponding prior densities are represented by $p(\boldsymbol{\beta})$ and $p(\sigma^2)$, respectively. Using Bayes' Theorem, the prior information can be combined with the likelihood to yield the posterior distribution

$p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y})$. That is,

$$p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) = \frac{\prod_{i=1}^{N} \int \prod_{j=1}^{n_i} \Pr(Y_{ij} = 1|\boldsymbol{\beta}, u_i)\phi(u_i|\sigma^2)du_i \, p(\boldsymbol{\beta})p(\sigma^2)}{\int \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} \Pr(Y_{ij} = 1|\boldsymbol{\beta}, u_i)\phi(u_i|\sigma^2)du_i \, p(\boldsymbol{\beta})p(\sigma^2)d\boldsymbol{\beta}d\sigma^2}.$$

Note that the denominator is a normalizing constant independent of $\boldsymbol{\beta}$ and σ^2, so that the estimators, such as the posterior mode, can be derived from the numerator. That is,

$$p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) \propto \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} \Pr(Y_{ij} = 1|\boldsymbol{\beta}, u_i)\phi(u_i|\sigma^2)du_i \, p(\boldsymbol{\beta})p(\sigma^2). \tag{3.6}$$

3.4.1 Bayesian Data Augmentation

It is difficult to establish the posterior distribution (3.6) because it involves an intractable integral. Instead, we use a sampling-based approach using the idea of Data Augmentation to obtain the marginal posterior density $p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y})$.

The idea of a Bayesian Data Augmentation procedure is that we augment the vector of unknowns $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2)'$ by the (latent) random effects $\boldsymbol{u}$. Thus we augment the vector of parameters $\boldsymbol{\theta}$ with (latent but unknown) data $\boldsymbol{u}$ and work with the posterior distribution of $(\boldsymbol{\theta}, \boldsymbol{u})$ given $\boldsymbol{y}$. In other words the $\boldsymbol{u}$ are also parameters. The marginal posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$ is related to the posterior $p(\boldsymbol{\theta}, \boldsymbol{u}|\boldsymbol{y})$ as

follows:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \int p(\boldsymbol{\theta}, \boldsymbol{u}|\boldsymbol{y}) d\boldsymbol{u}.$$

All marginal posterior properties (e.g., summary statistics) of $\boldsymbol{\theta}$ are the same regardless of whether they are obtained directly from $p(\boldsymbol{\theta}|\boldsymbol{y})$ or from first $p(\boldsymbol{\theta}, \boldsymbol{u}|\boldsymbol{y})$ and then marginalized over $\boldsymbol{u}$. Thus, inference can be based on the joint posterior density $p(\boldsymbol{\theta}, \boldsymbol{u}|\boldsymbol{y})$. For our random effects logistic regression example the $p(\boldsymbol{\theta}, \boldsymbol{u}|\boldsymbol{y})$ is given by

$$p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{u}|\boldsymbol{y}) \propto \prod_{i=1}^{N} \prod_{j=1}^{n_i} \Pr(Y_{ij} = 1|\boldsymbol{\beta}, u_i) \phi(u_i|\sigma^2) p(\boldsymbol{\beta}) p(\sigma^2). \quad (3.7)$$

Expression (3.7) is easier to work with as seen in the next section, posterior information is obtained via Markov chain Monte Carlo (MCMC) sampling.

3.4.2 Markov chain Monte Carlo sampling

Markov chain Monte Carlo (MCMC) methods are a class of algorithms for sampling the posterior distribution based on constructing a Markov chain that has the desired distribution as its limiting (stationary) distribution. The idea of MCMC sampling is to simulate a random walk in the space of parameters of interest, $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_d)'$, which converges to the joint posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$. The samples are drawn sequentially, with the distribution of the sampled draws depending on the last value drawn; hence, the draws form a Markov chain. The states of the chain after a large number of iterations is then used as a sample from the desired posterior distribution.

MCMC techniques have several attractive features. From a prac-

tical point of view, they are in general relatively easy to construct and often not too difficult to implement. In addition, they are often the only available techniques for exploring high dimensional problems. We present below the basic Metropolis(-Hastings) algorithm and the Gibbs sampler.

The Metropolis algorithm

Given a target posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$, known up to a normalizing constant, the Metropolis algorithm creates a sequence of random vectors $(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \cdots)$ whose distribution converges to the target distribution. Each sequence can be considered a random walk whose stationary distribution is $p(\boldsymbol{\theta}|\boldsymbol{y})$. The algorithm proceeds as follows (see, e.g.,Tierney, 1994; Gelman, Carlin, Stern, and Rubin, 1995). Start with some initial value $\boldsymbol{\theta}^0$. For $t = 1, 2, \cdots$, obtain $\boldsymbol{\theta}^{(t)}$ from $\boldsymbol{\theta}^{(t-1)}$ using the following steps:

1. Sample a *candidate point* $\boldsymbol{\theta}^*$ from a *proposal distribution* at time t, $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)})$. The proposal distribution distribution must be *symmetric*; that is, $q(\boldsymbol{\theta}_a|\boldsymbol{\theta}_b) = q(\boldsymbol{\theta}_b|\boldsymbol{\theta}_a)$ for all $\boldsymbol{\theta}_a$ and $\boldsymbol{\theta}_b$.

2. Calculate the ratio of the densities,

$$r = \frac{p(\boldsymbol{\theta}^*|\boldsymbol{y})}{p(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{y})}.$$

3. Set

$$\boldsymbol{\theta}^{(t)} = \begin{cases} \boldsymbol{\theta}^* & \text{with probablity } \min(r, 1), \\ \boldsymbol{\theta}^{(t-1)} & \text{otherwise.} \end{cases}$$

The algorithm requires the ability to draw $\boldsymbol{\theta}^*$ from the proposal (jumping) distribution $q(\boldsymbol{\theta}^*|\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$.

The Metropolis-Hastings algorithm

The Metropolis-Hastings (M-H) algorithm generalizes the basic Metropolis algorithm, described above, in two ways. First, the proposal distribution q needs no longer to be symmetric. That is, there is no requirement that $q(\boldsymbol{\theta}_a|\boldsymbol{\theta}_b) = q(\boldsymbol{\theta}_b|\boldsymbol{\theta}_a)$. Secondly, to correct for the asymmetry in the proposal density the acceptance ratio is now (see, e.g., Tierney, 1994; Chib and Greenberg, 1995; Gelman et al., 1995)

$$r = \frac{p(\boldsymbol{\theta}^*|\boldsymbol{y})q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{y})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)})}.$$

Allowing an asymmetric proposal distribution can be useful in increasing the speed of the random walk.

Gibbs sampler

The Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990; Gilks, 1996) is a MCMC algorithm that has been found very useful in multidimensional problems. It is defined in terms of subvectors of $\boldsymbol{\theta}$. At each iteration t, the Gibbs sampler cycles through the subvectors of $\boldsymbol{\theta}$, drawing θ_j from the conditional distribution given all the remaining components of $\boldsymbol{\theta}$:

$$p_j(\theta_j|\boldsymbol{\theta}_{(-j)}^{(t-1)}, \boldsymbol{y}),$$

where $\boldsymbol{\theta}_{(-j)}$ represents all the components of $\boldsymbol{\theta}$, except for θ_j, i.e. $\boldsymbol{\theta}_{(-j)} = (\theta_1, \cdots, \theta_{j-1}, \theta_{j+1}, \cdots, \theta_d)'$. This suggests the following MCMC scheme.

1. Generate $\theta_1^{(t)}$ from $p_1(\theta_1|\theta_2^{(t-1)}, \theta_3^{(t-1)}, \cdots, \theta_d^{(t-1)}, \boldsymbol{y})$

2. Generate $\theta_2^{(t)}$ from $p_2(\theta_2|\theta_1^{(t)}, \theta_3^{(t-1)}, \cdots, \theta_d^{(t-1)}, \boldsymbol{y})$

$\vdots$

d. Generate $\theta_d^{(t)}$ from $p_d(\theta_d|\theta_1^{(t)}, \theta_2^{(t)}, \cdots, \theta_{d-1}^{(t)}, \boldsymbol{y})$

At the completion of these steps, the vector $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, \cdots, \theta_d^{(t)})'$ provides the simulated value of $\boldsymbol{\theta}$ at the tth iteration of sampling. The d steps of this Gibbs sampling scheme completes one iteration of the simulation method.

After a large number, T, of iterations, we obtain $\boldsymbol{\theta}^{(T)}$. Geman and Geman (1984) have shown that under mild conditions, the joint distribution $\boldsymbol{\theta}^{(T)}$ converges at an exponential rate to $p(\boldsymbol{\theta}|\boldsymbol{y})$ as $T \to \infty$. The desired joint posterior distribution, $p(\boldsymbol{\theta}|\boldsymbol{y})$, can be approximated by the empirical distribution of M values $\boldsymbol{\theta}^{(t)}$ for $t = T+1, T+2, \cdots, T+M$, where T is large enough so that the Gibbs sampler has converged and M is chosen to give sufficient precision to the empirical distribution of interest.

3.4.3 Posterior samples from our logistic example

In the random-effects logistic regression, we are interested in the joint posterior density $p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{u}|\boldsymbol{y})$ and its marginal posterior densities $p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y})$ and $p(\boldsymbol{u}|\boldsymbol{y})$. Let $[U|V, W]$ denotes the conditional distribution of U given V and W. Then $\boldsymbol{\beta}$ and σ^2 can be sampled by Gibbs sampling from the full conditionals $[\boldsymbol{\beta}|\sigma^{2(t-1)}, \boldsymbol{u}^{(t-1)}, \boldsymbol{y}]$, $[\sigma^2|\boldsymbol{\beta}^{(t)}, \boldsymbol{u}^{(t-1)}]$ and $[\boldsymbol{u}|\boldsymbol{\beta}^{(t)}, \sigma^{2(t)}]$. The elegance of this approach (using the Data Augmentation approach) is that it is easy to sample from each of the conditionals.

The conditional $[\boldsymbol{\beta}|\sigma^2, \boldsymbol{u}, \boldsymbol{y}] = [\boldsymbol{\beta}|\boldsymbol{u}, \boldsymbol{y}]$ as long as $p(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{\beta})p(\sigma^2)$, i.e. under prior independence of $\boldsymbol{\beta}$ and σ^2. Similarly, $[\sigma^2|\boldsymbol{\beta}, \boldsymbol{u}] = [\sigma^2|\boldsymbol{u}]$. The conditional $[\boldsymbol{u}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{y}]$ does not simplify.
The assumed prior distributions are:

(a) a multivariate normal prior on $\boldsymbol{\beta}$, i.e.

$$p(\boldsymbol{\beta}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{B}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right],$$

where $\boldsymbol{\beta}_0$ denotes the prior mean, and $\boldsymbol{B}_0^{-1}$ the prior precision matrix. For a reasonably vague prior, it is common to assume that $\boldsymbol{\beta}_0 = 0$ and $\boldsymbol{B}_0^{-1} = 10^{-3}\mathrm{I}_d$, where I_d is a $d \times d$ identity matrix.

(b) the inverse-gamma prior density for σ^2, i.e.

$$p(\sigma^2) = \frac{\xi^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-(\alpha+1)}\exp\left(\frac{-\xi}{\sigma^2}\right),$$

which has hyperparameters (α, ξ). A convenient parameterization is a scaled inverse-χ^2 with scale σ_0^2 and ν_0 degrees of freedom (e.g., Gelman et al., 1995). That is, the distribution of σ^2 is taken to be the distribution of $\sigma_0^2\nu_0/W$, where W is a $\chi^2_{\nu_0}$ random variable.

The sampling method for each of the conditionals is now specified.

• $[\boldsymbol{\beta}|\boldsymbol{u}^{(t-1)}, \boldsymbol{y}]$

Given $u_i^{(t-1)}$ $(i = 1, \cdots, N)$, the random-effects logistic model reduces to the simple logistic model with offset $u_i^{(t-1)}$ for each observation. The posterior conditional density is given by

$$p(\boldsymbol{\beta}|\boldsymbol{u}^{(t-1)}, \boldsymbol{y}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{B}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right] \times \prod_{i=1}^{N}\prod_{j=1}^{n_i}\frac{\exp\left[\boldsymbol{x}_{ij}'\boldsymbol{\beta} + u_i^{(t-1)}\right]}{1+\exp\left[\boldsymbol{x}_{ij}'\boldsymbol{\beta} + u_i^{(t-1)}\right]}.$$

This expression, $p(\boldsymbol{\beta}|\boldsymbol{u}^{(t-1)}, \boldsymbol{y})$, does not have an explicit form, i.e. it does not have a kernel of any known standard distribution. The exact posterior distribution of $\boldsymbol{\beta}$ can be obtained by the Metropolis algorithm with proposal density (e.g., Johnson and Albert, 1999)

$$q(\boldsymbol{\beta}^{(t-1)}|\boldsymbol{\beta}^*) = \mathcal{N}\left(\boldsymbol{\beta}^*|\boldsymbol{\beta}_u^{(t-1)}, \tau_\beta \boldsymbol{V}_u\right),$$

where τ_β is a scalar that is adjusted in order to obtain a reasonable acceptance rate. That is, to sample from $p(\boldsymbol{\beta}|\boldsymbol{u}^{(t-1)}, \boldsymbol{y})$, we find $\hat{\boldsymbol{\beta}}_u$ and $\hat{\boldsymbol{V}}_u$ by performing a GLM regression of $\boldsymbol{y}$ on $\boldsymbol{x}$ using the simulated values $\boldsymbol{u}^{(t-1)}$ as offsets and then generate a random variate $\boldsymbol{\beta}^*$ from a multivariate normal distribution, $N(\hat{\boldsymbol{\beta}}_u, \tau_\beta \hat{\boldsymbol{V}}_u)$, and set $\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^*$ with probability

$$\min\left(\frac{p(\boldsymbol{\beta}^*|\boldsymbol{u}^{(t-1)}, \boldsymbol{y})}{p(\boldsymbol{\beta}^{(t-1)}|\boldsymbol{u}^{(t-1)}, \boldsymbol{y})}, 1\right),$$

otherwise, set $\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^{(t-1)}$. Any acceptance rate between 10%-40% ought to perform close to optimal (e.g., Roberts and Rosenthal, 2001).

• $[\sigma^2|\boldsymbol{u}^{(t-1)}]$

With the inverse-gamma prior on σ^2, and hence a scaled inverse-χ^2 prior distribution, the resulting posterior density is

$$\begin{aligned} p(\sigma^2|\boldsymbol{u}^{(t-1)}) &\propto p(\sigma^2)p(\boldsymbol{u}|\sigma^2) \\ &= \left(\sigma^2\right)^{-(N+\nu_0)/2+1} \exp\left(-\frac{1}{2\sigma^2}\left(\nu_0\sigma_0^2 + \sum_{i=1}^{N} u_i^{2(t-1)}\right)\right) \end{aligned}$$

Thus,

$$\sigma^2|\boldsymbol{u}^{(t-1)} \sim \text{Inv}-\chi^2 \left(N+\nu_0, \frac{\left(\nu_0\sigma_0^2 + \sum_{i=1}^{N} u_i^{2(t-1)}\right)}{N+\nu_0} \right)$$

which is also a scaled inverse-χ^2 distribution. Hence, it is possible to sample σ^2 directly.

• $[\boldsymbol{u}|\boldsymbol{\beta}^{(t)}, \sigma^{2(t)}, \boldsymbol{y}]$

Unfortunately, the conditional distribution $[\boldsymbol{u}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{y}]$ does not have a closed form. Its posterior density is

$$p(u_i|\boldsymbol{\beta}^{(t)}, \sigma^{2(t)}, \boldsymbol{y}_i) \propto \exp\left(-\frac{u_i^2}{2\sigma^{2(t)}}\right) \times \prod_{j=1}^{n_i} \left[\frac{\exp\left(\boldsymbol{x}'_{ij}\boldsymbol{\beta}^{(t)} + u_i\right)}{1+\exp\left(\boldsymbol{x}'_{ij}\boldsymbol{\beta}^{(t)} + u_i\right)}. \right]$$

As described by Zeger and Karim (1991), we can use the Metropolis algorithm to generate the posterior samples of $u_i^{(t)}$ with

$$q(u_i^*|u_i^{(t-1)}) = N(\hat{u}_i^{(t-1)}, \tau_u^{-1}\hat{v}_i^{(t-1)}),$$

where τ_u is a tuning parameter for adjusting the Metropolis acceptance rate. The maximum value of $p(u_i|\boldsymbol{\beta}^{(t)}, \sigma^{2(t)}, \boldsymbol{y})$ occurs at

$$\hat{u}_i^{(t-1)} = \left(\sum_{j=1}^{n_i} v_{ij} + \sigma^{-2(t)} \right)^{-1} \sum_{j=1}^{n_i} v_{ij} \left(y_{ij}^* - x'_{ij}\boldsymbol{\beta}^{(t)} \right)$$

and its curvature is

$$\hat{v}_i^{(t-1)} = \left(\sum_{j=1}^{n_i} v_{ij} + \sigma^{-2(t)} \right)^{-1},$$

where $v_{ij} = \text{var}(Y_{ij}|u_i)$, $y_{ij}^* = \eta_{ij} + \frac{d\mu_{ij}}{d\eta_{ij}}(y_{ij} - \mu_{ij})$, $\eta_{ij} = x'_{ij}\boldsymbol{\beta} + u_i$ and $\mu_{ij} = \text{E}(Y_{ij}|u_i$. Generating u_i, $(i = 1 \cdots, N)$, from its full conditional distribution is the most time-consuming step because y_{ij}^* and v_{ij} depend on u_i; hence the actual mode and curvature must be obtained by iterating the equations for $\hat{u}_i$ and $\hat{v}_i$.

The WinBUGS software, described in the next section, uses Gibbs sampling and a Metropolis-within-Gibbs routine to draw MCMC samples from complex statistical models.

3.4.4 WinBUGS for Bayesian inference

`WinBUGS` (MS Windows operating system version of the `BUGS`: **B**ayesian analysis **U**sing **G**ibbs **S**ampling) is flexible software for the Bayesian analysis of complex statistical models using MCMC methods. The software is currently distributed electronically from the BUGS project web site.

`http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml`

More details can be obtained from `WinBUGS`' extensive user manual (Spiegelhalter, Thomas, Best, and Lunn, 2003).

The versatility of the `WinBUGS` package allows for a wide variety of posterior models. Firstly, it is possible to sample from a large number of statistical models including, for example, the Bernoulli, Poisson, normal, multinomial and gamma distributions. Secondly, with recent developments in the software, it is also possible to draw samples from non-standard distributions using the "ones" trick (see `WinBUGS`'s manual). Thirdly, various sampling algorithms are implemented including Gibbs sampler, Metropolis,

slice sampler etc.

A Bayesian p-value can be obtained from the WinBUGS program by using the `step( )` function. The function creates a Boolean variable that counts the number of simulations in which, for example, $\beta_x > 0$ is true. As a result, the 0/1 values from the `step( )` function can be used to compute left- or right-tail areas. For example, if we have M samples of $p_{\beta_x} = \texttt{step}(\beta_x)$ after T burn-in samples, an equal-tail two-sided Bayesian p-value of β_x is given by

$$2 \min \left(1 - \frac{1}{M} \sum_{t=T+1}^{T+M} p_{\beta_x}^{(t)}, \frac{1}{M} \sum_{t=T+1}^{T+M} p_{\beta_x}^{(t)} \right).$$

While the MCMC algorithms, e.g. using `WinBUGS`, have the potential to be quicker than the numerical approximations, the convergence rate of any algorithm cannot be guaranteed. In the next section we describe methods for assessing the rate of convergence of MCMC samples.

3.4.5 MCMC convergence

Convergence is diagnosed when the chains have 'forgotten' their initial values, and the output from all chains is indistinguishable. Geweke (1992) proposed a convergence diagnostic for MCMC samples based on a test for equality of the means of the first and last part of a *single* chain (by default the first 10% and the last 50%). Geweke's approach involves calculation of the sample mean and asymptotic variance in each window, the latter being determined by spectral density estimation. His convergence diagnostic Z is the difference between these two means divided by the asymptotic standard error of their difference. As the chain length $\rightarrow \infty$, the sampling distribution of the chain has converged. Hence values of $Z \rightarrow \mathcal{N}(0, 1)$ which fall in the extreme tails of a standard normal distribution, ± 2, suggest that the chain has not fully converged.

Gelman and Rubin (1992) and Gelman (1996) proposed a general approach to monitoring convergence of MCMC output in which two or more parallel chains are run with starting values that are overdispersed relative to the posterior distribution. Convergence for multiple chains may be assessed using Gelman-Rubin scale factor reduction factors that compare variation of the samples parameter values within and between chains. It is based on a comparison of within-chain and between-chain variances, and is similar to a classical analysis of variance. To measure the variability of sample $\theta_j^{(t)}$ within the chain $(j = 1, \cdots, J)$ define

$$V_j = \sum_{t=T+1}^{T+M} \left(\theta_j^{(t)} - \bar{\theta}_j\right)^2 / (M-1)$$

over M iterations after an initial burn-in of T iterations, where $\bar{\theta}_j$ is the average of $\theta_j^{(t)}$ $(t = T+1, \cdots, T+M)$. Ideally, the burn-in period is the initial set of samples where the effect of initial parameter values tails off. Convergence is therefore assessed from $T+1$ to $T+M$. Variability within chains V_W is the average of V_js. Between chain variance is measured by

$$V_B = \frac{M}{J-1} \sum_{j=1}^{J} \left(\bar{\theta}_j - \bar{\theta}\right)^2$$

where $\bar{\theta}$ is the average of $\bar{\theta}_j$s. The scale factor reduction (SRF) compares a pooled estimator of $\mathrm{Var}(\theta)$, given by $V_P = V_B/M + MV_W/(M-1)$, to V_W. More specifically, $\mathrm{SRF} = \sqrt{V_P/V_W}$ with values under 1.2 (Congdon, 2003, p. 19) indicating convergence.

More recently Brooks and Gelman (1997) proposed a convergence statistic known as Brooks-Gelman-Rubin (BGR). This is a ratio of parameter interval lengths, where for chain j the length of the $100(1-\alpha)\%$ interval for parameter θ is obtained, i.e. the gap between 0.5α and $(1-0.5\alpha)$ points from M simulated values. This provides J within-chain interval lengths, with mean I_w. For the pooled output of MJ samples, the same $100(1-\alpha)\%$

interval I_B is obtained. The ratio I_B/I_W should converge to one if there is convergent mixing over different chains.

The above MCMC convergence diagnostics are implemented in two R-packages, namely CODA (**C**onvergence **D**iagnosis and **O**utput **A**nalysis) and BOA (**B**ayesian **O**utput **A**nalysis). These packages are downloadable from `http://cran.r-project.org/`. The packages compute convergence diagnostics and statistical and graphical summaries for the MCMC samples. Even though BOA is designed to be faster and more efficient than CODA, it is not flexible in terms of data manipulation than CODA. That is, CODA offers more analysis options and better graphical tools than BOA.

CHAPTER

4 Analysis of Binary Data Subject to Response Misclassification

4.1 Introduction

Whenever multiple examiners are involved in an epidemiological study kappa values (Cohen, 1960), denoted as κ, are reported. High values of κ indicate that the examiners agree much in their scoring behaviour. However, since the introduction of Cohen's kappa, several 'paradoxes' in its interpretation have been pointed out (Cicchetti and Feinstein, 1990; Feinstein and Cicchetti, 1990). Further, the measures of agreement do not indicate the impact of the bias and variability of scoring of the examiners on the estimates of the regression coefficients of an epidemiological regression model. Instead, when a gold standard or benchmark scorer is available, the appropriate measures are the sensitivity and specificity of each examiner vis-à-vis the gold standard (or benchmark scorer).

From the literature on errors-in-variables (see, e.g., Carroll et al., 1995 and Gustafson, 2004), it is known that when covariates are measured with error (with misclassification as a special case), this will result in biasedly es-

timated regression coefficients. When the model is nonlinear, like a logistic regression model, attenuation also occurs when the response is measured with error. The idea is then to correct for this misclassification using the sensitivity and specificity values.

In this thesis, we wish to correct for examiners' misclassification in the prevalence of caries experience. More specifically, we show here how the scores of different dental examiners can be corrected in a logistic regression model using data from calibration exercises. In this chapter, we will explain the concepts of correcting for misclassification for a binary logistic regression model and apply it to the Signal Tandmobiel® study. Simplifying caries experience to a binary response has the advantage that correction for misclassification is done in a relatively simple manner, which is useful given the sparseness of the validation data. The following rule was used to construct a binary response y based on the dmft-index:

$$y = \begin{cases} 0 & \text{if the dmft-index is 0 (no caries experience),} \\ 1 & \text{if the dmft-index is } > 0. \end{cases}$$

From a dental point of view it is of interest to examine the geographical trend in the prevalence of caries experience in Flanders. For our geographical analysis we have taken the data of the first year survey of the Signal Tandmobiel® study, which will be referred to below as a cross-sectional oral health study. The geographical trend in the prevalence of caries experience was examined by including the (standardized) $(x, y)-$co-ordinate of the municipality of the school to which the child belongs. Additionally age (centred) and gender were included as covariates.

Since children of the same school share some common characteristics, 'school' was included in the model as a random effect (see Mwalili, Lesaffre, and Declerck, 2005) measuring the between school variation. A preliminary fit (Section 2.6) from a random effects logistic regression model showed a significant East-West gradient in the prevalence of caries experience. A possible cause for the apparent geographical trend in caries experience is a

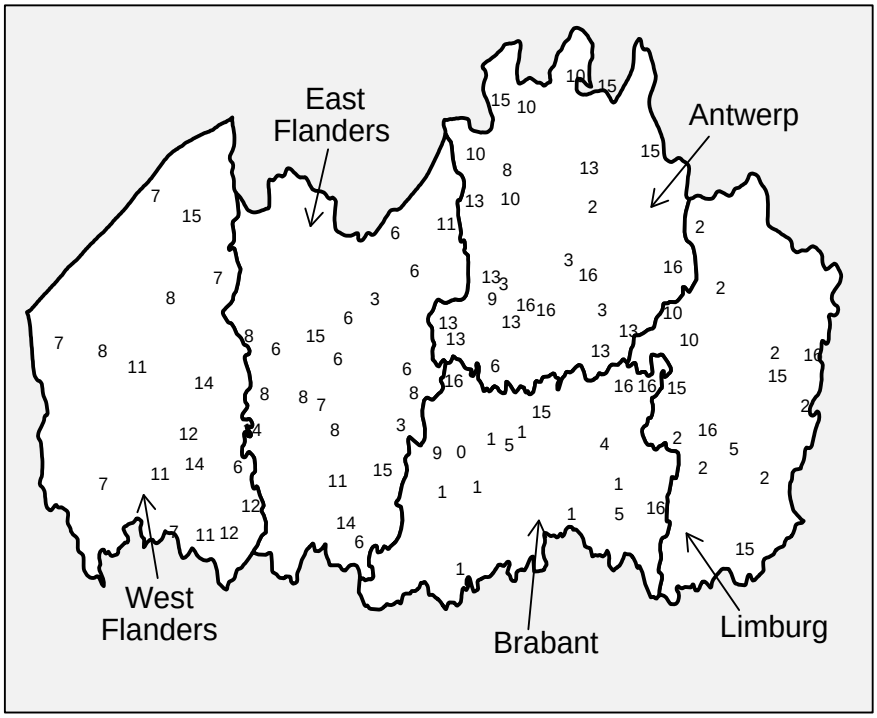


Figure 4.1: *Map of Flanders showing the distribution of the examiners scoring in the different schools (numbers indicate examiners that scored the children of a particular school located at the corresponding* $(x, y)-$*coordinate) in the five provinces of Flanders.*

different scoring behaviour of the 16 dental examiners. Indeed, in Figure 4.1 it is clearly seen that each examiner was active in a relatively restricted geographical area. Thus, a legitimate question was whether the geographical trend in the prevalence of caries experience was genuine or due to the different scoring behaviour of the examiners. In Figure 4.1 the benchmark examiner is an experienced dentist, indicated by the symbol '0' near the centre of the map, who is assumed to be error free for measuring the dmft-index. But, even if the benchmark examiner were not error free, it is still of interest to know whether the geographical trend would remain if only one examiner (say the benchmark examiner) had scored the caries experience

alone.

A classical way to take a confounder into account is to include it in the (logistic) regression model. Thus, we considered three binary regression models, namely,

Model 1: random effects logistic regression without examiners' term;

Model 2: random effects logistic regression with fixed examiners' term;

Model 3: random effects logistic regression corrected for examiners' misclassification.

Since the ith child is scored by the jth examiner and is located in the kth school, sometimes the notation y_i, y_{ij} or y_{ik} will be used, whichever is appropriate.

4.2 Measures of agreement, bias and variability

The kappa statistic is used to measure inter-observer agreement. It determines how strongly two observers agree by comparing the probability of the two agreeing by chance with the observed agreement. Cohen's kappa is a coefficient of agreement for binary outcomes (0, 1) and is equal to (Cohen, 1960):

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

where p_o is the observed proportion of agreement between the two scores and p_e is the agreement obtained purely by chance. Using the misclassification matrix

$$\begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{pmatrix} \end{array}$$

and $n_{11} + n_{12} + n_{21} + n_{22} = n$, we obtain $p_0 = (n_{11} + n_{22})/n$ and $p_e = [(n_{11} + n_{12})(n_{11} + n_{21}) + (n_{21} + n_{22})(n_{12} + n_{22})]/n^2$. In terms of the cell

frequencies the estimate of κ simplifies to

$$\hat{\kappa} = \frac{2(n_{11}n_{22}-n_{12}n_{21})}{(n_{11}+n_{12})(n_{12}+n_{22})+(n_{11}+n_{21})(n_{21}+n_{22})}.$$

The possible values of κ range from -1 (complete disagreement) through 0 (no agreement above that expected by chance) to $+1$ (perfect agreement). Kappa statistics, even when high (> 0.8), do not rule out that the results of an epidemiological analysis are biased when different examiners are involved, as will be shown in our analysis below. Further, kappa statistics do not distinguish between bias and variability. The misclassification matrices below illustrate this. They all correspond to $\kappa = 0.6$.

$$\begin{array}{c} \\ 0 \\ 1 \end{array}\!\!\begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 37 & 13 \\ 3 & 27 \end{pmatrix} \end{array} \qquad \begin{array}{c} \\ 0 \\ 1 \end{array}\!\!\begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 35 & 11 \\ 5 & 29 \end{pmatrix} \end{array} \qquad \begin{array}{c} \\ 0 \\ 1 \end{array}\!\!\begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 32 & 8 \\ 8 & 32 \end{pmatrix} \end{array}$$

(1) (2) (3)

$$\begin{array}{c} \\ 0 \\ 1 \end{array}\!\!\begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 29 & 5 \\ 11 & 35 \end{pmatrix} \end{array} \qquad \begin{array}{c} \\ 0 \\ 1 \end{array}\!\!\begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 27 & 3 \\ 13 & 37 \end{pmatrix} \end{array}$$

(4) (5)

In the first matrix, the examiner (row) clearly underscores caries experience compared with the gold standard (column), while in the fifth matrix the reverse is true. In the third matrix, $\kappa = 0.6$ because of scoring variability.

When a gold standard is available, it is preferable to estimate the sensitivity (π_{11}) and specificity (π_{00}) of each examiner vis-à-vis the gold standard. In the above notation, $\hat{\pi}_{11} = n_{22}/(n_{12}+n_{22})$ estimates the probability that the examiner rates caries experience when the gold standard also rated caries experience, while $\hat{\pi}_{00} = n_{11}/(n_{11}+n_{21})$ estimates the probability that the examiner did not rate caries experience when the gold standard did not.

The sensitivity and specificity can differentiate between bias and variability. Indeed, the sensitivities of the above tables are 27/40, 29/40, 32/40, 35/40, and 37/40, respectively, and the specificities are 37/40, 35/40, 32/40, 29/40, and 27/40. The sensitivity and specificity of each examiner vis-à-vis the benchmark scorer will be used as correction terms for the binary regression model.

Further, to illustrate that the errors in a binary response can produce biased measures of association and loss of precision in estimation, consider an hypothetical population with exposure-disease classification shown in Table 4.1. The true and the observed disease (binary response) status are denoted by Y and Y^*, respectively, with the exposure (binary covariate) denoted by X. The estimated log odds ratio measuring the association of X with Y is $\hat{\beta} = 0.539$, with asymptotic standard error (ASE) = 0.3318 and a Wald statistic, $z^2 = (\hat{\boldsymbol{\beta}}/ASE)^2 = 2.638$ ($p = 0.1044$), for the test of null hypothesis $H_0 : \boldsymbol{\beta} = 0$.

However, we do not observe Y directly but rather a possibly error-corrupted version Y^*. Suppose that the misclassification probabilities for observing Y^* given Y are

$$\boldsymbol{\pi} = \begin{pmatrix} 0.9 & 0.2 \\ 0.1 & 0.8 \end{pmatrix}.$$

That is, given $Y = 0$ we observe $Y^* = 1$ with probability 0.1 and given $Y = 1$ we observe $Y^* = 0$ with probability 0.2. These misclassification probabilities leads to observed counts in Table 4.1.

Table 4.1: *Classification of hypothetical data*

True classification				Observed classification			
	$X = 0$	$X = 1$			$X = 0$	$X = 1$	
$Y = 0$	30	20	70	$Y^* = 0$	41	34	75
$Y = 1$	70	80	130	$Y^* = 1$	59	66	125
	100	100	200		100	100	200

A naive analysis of the observed counts in Table 4.1 yields as estimate $\hat{\beta}_* = 0.299$, with asymptotic standard error $(\text{ASE}_*) = 0.2931$ and Wald statistic $z_*^2 = 1.043$ $(p = 0.3071)$. The correct likelihood of the Y involves the response misclassification probabilities (described in Section 4.4). The corrected maximum likelihood estimate of the regression coefficient then becomes $\hat{\beta}_{cor} = 0.539$, with asymptotic standard error $(\text{ASE}_{cor}) = 0.5393$ and Wald statistic $z_{cor}^2 = 1.00$ $(p = 0.3176)$. The regression coefficients are here the estimates of the log odds ratios. Thus, the corrected estimate of the log odds ratio is unbiased for the true log odds estimate but it is estimated with less precision, implying that errors in the response lead to attenuation and loss of estimation efficiency.

4.3 The logistic random-effects model

The purpose is to regress a binary response y_{ik} for the ith subject $(i = 1, \cdots, n_k)$ within cluster k $(k = 1, \cdots, N)$ on a d-dimensional vector of covariates $\boldsymbol{x}_i$ pertaining to the ith subject. Let $\boldsymbol{\beta}$ be the corresponding vector of regression coefficients (fixed effects). Further, let $\boldsymbol{y} = (\boldsymbol{y}_1', \cdots, \boldsymbol{y}_N')'$, where $\boldsymbol{y}_k = (y_{1k}, \cdots, y_{n_k,k})'$, and $\boldsymbol{x} = (\boldsymbol{x}_1', \cdots, \boldsymbol{x}_N')'$.

The common approach to explicitly model the effect of clusters (here 'school') is to use a generalized linear mixed model (GLMM). The GLMM is an extension of the class of generalized linear model (GLM) (McCullagh and Nelder, 1989) by adding random effects to the linear predictor (see, e.g. Zeger, Liang, and Albert, 1988; Neuhaus, Kalbfleisch, and Hauck, 1991). That is, given a vector $\boldsymbol{u}_k$ of the random effects specific to the kth cluster, for the ith subject, the conditional density of y_{ik} is in GLM form. In other words, one assumes that

$$\pi_{ik} \equiv \Pr\left(y_{ik} = 1 | \boldsymbol{x}_i, \boldsymbol{u}_k\right) = g^{-1}\left(\boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{u}_k\right), \tag{4.1}$$

where $\boldsymbol{z}_i$ is a specified q-dimensional vector of covariates. Further, the

model assumes that the random effects follow a distribution G with mean zero. The logit link function

$$g(\theta) = \log\left[\frac{\theta}{(1-\theta)}\right]$$

yields the logistic random-effects regression model. The marginal likelihood of the logistic random-effects regression model is given by

$$L(\boldsymbol{\beta}, G) = \prod_{k=1}^{N} \int_{\Re^q} \prod_{i=1}^{n_k} \pi_{ik}^{y_{ik}} (1-\pi_{ik})^{1-y_{ik}} dG(\boldsymbol{u}_k).$$

In particular, if we assume that $\boldsymbol{u}_k$ is an independent random vector from a multivariate normal distribution with mean $\mathbf{0}$ and variance $\boldsymbol{D}$, i.e. $\boldsymbol{u}_k \sim \mathcal{N}(\mathbf{0}, \boldsymbol{D})$ then

$$L(\boldsymbol{\beta}, D) = \prod_{k=1}^{N} \int_{\Re^q} \prod_{i=1}^{n_k} \pi_{ik}^{y_{ik}} (1-\pi_{ik})^{1-y_{ik}} \times \frac{1}{\sqrt{2\pi|\boldsymbol{D}|}} \exp\left(-\frac{1}{2}\boldsymbol{u}_k' \boldsymbol{D}^{-1} \boldsymbol{u}_k\right) d\boldsymbol{u}_k. \tag{4.2}$$

The model parameters are obtained by maximizing the marginal likelihood (4.2).

Note that for our application we considered a logistic regression with random school intercepts, so that $u_k \sim \mathcal{N}(0, \sigma^2)$. Observe that the factor "school" was included into the logistic regression model, even though, from a dental viewpoint it is only marginally necessary. This is so because caries experience is examined here for seven year old children and the prevalence of caries experience at this age is largely the result of the dietary and brushing behaviour in the past, i.e. in the pre-school period. Hence, a large "cluster" effect of school was not anticipated, but we kept "school" in the model as a random effect to take account for its clustering effect on caries experience in (mainly the last part of) the first year. Further, children from

the same school may be more similar, e.g. in the socio-economic aspects, than children from different schools.

4.3.1 Fitting a random effects logistic model

The random effects logistic regression can be fitted with a frequentist and a Bayesian approach as explained below.

4.3.1.1 Frequentist approach

In the frequentist approach, the log-likelihood (4.2) can be evaluated and maximized numerically. The natural choice is the Gauss-Hermite quadrature method (e.g., Liu and Pierce, 1994). We used the SAS procedure `NLMIXED`, as explained in Section 3.3.

4.3.1.2 Bayesian approach

In a Bayesian analysis of the random effects GLM (4.1), the parameters $\boldsymbol{\beta}$ and σ^2 are random variables. Let $p(\boldsymbol{\beta})$ and $p(\sigma^2)$ represent the prior distributions for $\boldsymbol{\beta}$ and σ^2. The posterior distribution $p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{u}|\boldsymbol{y})$ is given by

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{u}|\boldsymbol{y}) &\propto \prod_{k=1}^{N}\prod_{i=1}^{n_k} p(y_{ik}|u_k, \boldsymbol{\beta}, \sigma^2) \times p(u_k|\boldsymbol{\beta}, \sigma^2) \times p(\boldsymbol{\beta}) \times p(\sigma^2) \\ &= \prod_{k=1}^{N}\prod_{i=1}^{n_k} \pi_{ik}^{y_{ik}}(1-\pi_{ik})^{1-y_{ik}} \times \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{u_k^2}{2\sigma^2}) \times \\ &\qquad p(\boldsymbol{\beta}) \times p(\sigma^2) \end{aligned} \tag{4.3}$$

The conditional distributions of the parameters simplify because the GLMM is an example of a hierarchical Bayes model. It therefore possible to sample alternatingly from the conditionals $[\boldsymbol{\beta}|\boldsymbol{u}, \boldsymbol{y}]$, $[\sigma^2|\boldsymbol{u}]$ and $[\boldsymbol{u}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{y}]$, where $[s|t]$ denotes the conditional density of s given t.

We performed a Bayesian analysis using the `WinBUGS` (version 1.4) program. The specific prior distributions for this random-effects logistic model are:

(a) For the regression coefficient β_s ($s = 1, \cdots, d$), a vague normal prior was assumed i.e, $\beta_s \sim \mathcal{N}(0, 10^6)$.

(b) The prior distribution for σ^2 was taken as $IG(10^{-3}, 10^{-3})$ but a sensitivity analysis was also performed later (see Section 4.6.3) because of the known problem with this prior in hierarchical models (e.g., Gelman et al., 2004).

4.3.2 Application to the Signal Tandmobiel® study

Geographical differences in, for example, caries experience are often reported (Nadanovsky and Sheiham, 1994; Tickle et al., 2003). The analysis of determining factors for these differences is of utmost importance and facilitates the introduction of region-specific measures and/or interventions.

The results from (**Model 1** – without examiners' effect), shown in the left panel of Table 4.2, clearly indicate a significant East-West gradient in the prevalence of caries experience, being higher in the Eastern part of Flanders (the province of Limburg) (Figure 4.1). Other covariates, e.g. indicating the degree of deprivation for the area where the schools belong to, have also been included in the model (see Table 2.1 in Chapter 2). However, they did not have an appreciable effect on the East-West gradient. This means that the geographical trend is probably not due to sociological factors. The same conclusion could be drawn for fluoride intake from tap water and the intake of fluoride tablets. So, the question remained what the possible cause is for this phenomenon.

When adding 'examiner' into the random 'school' effects logistic model as a fixed effect (**Model 2**), the geographical East-West trend was clearly diminished; see the right panel of Table 4.2. In fact, the regression coefficient of the $(x, y)-$co-ordinate expresses the East-West gradient relative

Table 4.2: *Parameter estimates from the uncorrected random effects logistic model (4.1) predicting the prevalence of caries experience, controlling for the geographical effect in two ways (using* `WinBUGS Program 4.1`*)*

		without examiners' fixed effects			with examiners' fixed effects		
		Estimate	95% CI[a]		Estimate	95% CI[a]	
Parameter		(SE)	2.5	97.5	(SE)	2.5	97.5
Intercept		0.310(0.055)	0.20	0.42	0.239(0.456)	−0.63	1.14
x-ordinate		0.176(0.045)	0.09	0.27	0.109(0.078)	−0.04	0.26
y-ordinate		−0.025(0.046)	−0.12	0.07	−0.016(0.050)	−0.11	0.08
Gender		−0.052(0.066)	−0.18	0.08	−0.049(0.067)	−0.18	0.08
Age		0.342(0.090)	0.16	0.52	0.321(0.091)	0.14	0.50
	1				−0.088(0.487)	−1.05	0.85
	2				0.324(0.486)	−0.64	1.25
	3				0.260(0.485)	−0.71	1.19
	4				−0.149(0.489)	−1.13	0.80
E	5				0.254(0.508)	−0.75	1.24
X	6				0.049(0.476)	−0.89	0.96
A	7				−0.169(0.492)	−1.14	0.78
M	8				−0.093(0.492)	−1.06	0.85
I	9				0.540(0.504)	−0.46	1.51
N	10				0.198(0.509)	−0.81	1.19
E	11				−0.341(0.487)	−1.30	0.59
R	12				0.259(0.514)	−0.75	1.25
S	13				0.051(0.497)	−0.93	1.01
	14				0.346(0.496)	−0.63	1.30
	15				−0.191(0.471)	−1.14	0.72
	16				0.320(0.485)	−0.64	1.25
σ^2		0.160(.041)	0.09	0.25	0.124(0.039)	0.06	0.21

[a] CI = Credible interval

to the geographical area where the dental examiner was active. Thus there seems to be no local geographical East-West trend in caries experience. However, we argue that this is not the most appropriate way to correct for an examiner effect because it does not take into account the scoring bias and/or variability of the examiners. To take the examiners' effect into account properly, we opted for a corrected binary regression model as described in the next section.

4.4 Corrected binary random effects model

4.4.1 General framework

We consider a non-differential misclassification model where the misclassification probabilities are

$$\lambda_0 = \Pr(Y_{ik}^* = 1|Y_{ik} = 0) \text{ and } \lambda_1 = \Pr(Y_{ik}^* = 0|Y_{ik} = 1).$$

The probability $1-\lambda_0$ is the specificity of the measurement Y^*, while $1-\lambda_1$ is the sensitivity. We will assume that $\lambda_0+\lambda_1 < 1$ since values of λ_0 and λ_1 larger than 0.5 indicate that the misclassification process of the observed response Y^* performs worse than chance.

When response misclassification occurs, the true model for the observed dependent variable has the expression

$$\begin{aligned}
\mathrm{E}(Y_{ik}^*|\boldsymbol{x}_i) &\equiv \Pr(Y_{ik}^* = 1|\boldsymbol{x}_i) \\
&= \sum_y \Pr(Y_{ik}^* = 1|Y_{ik} = y) \Pr(Y_{ik} = y|\boldsymbol{x}_i) \\
&= \lambda_0 \Pr(Y_{ik} = 0|\boldsymbol{x}_i) + (1-\lambda_1) \Pr(Y_{ik} = 1|\boldsymbol{x}_i) \\
&= \lambda_0 \left(1 - \Pr(Y_{ik} = 1|\boldsymbol{x}_i)\right) + (1-\lambda_1) \Pr(Y_{ik} = 1|\boldsymbol{x}_i) \\
&= \lambda_0 + (1-\lambda_0-\lambda_1)\, g^{-1}\left(\boldsymbol{x}_i'\boldsymbol{\beta} + u_k\right) \qquad (4.4)
\end{aligned}$$

where g is the link function, namely: logit, probit or complementary log-log cumulative density function. Expression (4.4) collapses to the usual binary regression expression, $g^{-1}(\boldsymbol{x}_i'\boldsymbol{\beta} + u_k)$, when there is no misclassification ($\lambda_0 = \lambda_1 = 0$).

4.4.2 The approach proposed by Neuhaus

The misclassified binary regression has been analyzed by Neuhaus (1999) and Neuhaus (2002) for GLM and GLMM, respectively. If Y follows a

GLM, from (4.4) we have

$$\Pr(Y_i^* = 1|\boldsymbol{x}_i) \;=\; \lambda_0 + (1 - \lambda_0 - \lambda_1)\, g^{-1}(\eta_i)\,, \tag{4.5}$$

where $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$. Note that $\Pr(Y_{ik}^* = 1|\boldsymbol{x}_i)$ in expression (4.5) depends on $\boldsymbol{x}_i$ only through the the linear predictor η_i. Thus, performing simple algebra one gets

$$\eta_i \;=\; g\left[\frac{\Pr(Y_i^* = 1|\boldsymbol{x}_i) - \lambda_0}{1 - \lambda_0 - \lambda_1}\right] \;=\; g^*\left[\Pr(Y_i^* = 1|\boldsymbol{x}_i)\right].$$

When Y_i follows a GLM with link function g and the misclassification probabilities are known and independent of covariates, so does Y_i^* but with a modified link function, g^*.

For the case of simple binary regression $\Pr(Y = 1|X) \;=\; g^{-1}(\beta_0 + \beta_1 X)$. Neuhaus (1999) gives approximate bias factors, ψ, for the relationship between β_1 and β_1^*. That is, $\beta_1^* \;=\; \psi\beta_1$. For instance,

$$\begin{aligned}
\psi \;&=\; 1 - \lambda_0 - \lambda_1 && \text{if } g \text{ is linear;}\\
\psi \;&=\; \tfrac{(1-\lambda_0-\lambda_1)\exp(\beta_0)}{[\lambda_0+(1-\lambda_1)\exp(\beta_0)][1-\lambda_0+\lambda_1\exp(\beta_0)]} && \text{if } g \text{ is logistic;}\\
\psi \;&=\; \tfrac{(1-\lambda_0-\lambda_1)\phi(\beta_0)}{\phi[\Phi^{-1}\{\lambda_0+(1-\lambda_0-\lambda_1)\Phi(\beta_0)\}]} && \text{if } g \text{ is probit.}
\end{aligned}$$

Thus the bias due to misclassified binary responses are all towards the null and substantial when either λ_0 or λ_1 are greater than 0.1.

Neuhaus (2002) extended the misclassified GLM to misclassification in GLMM. When Y_{ik} follows a GLMM with link function g and the misclassification probabilities are known and independent of covariates and random effects, it can be shown that the possibly error corrupted response Y_i^* will also follow a GLMM but with a modified link function, g^*.

4.4.3 Our approach

In our approach we formulate the model in terms of the corrected probability in expression (4.4). Let $\pi_{i(j)k}$ be the expected proportion of successes for the ith child scored by the jth examiner ($j = 1, \cdots, 16$) in the kth school. Further, λ_{j0} and λ_{j1} are the misclassification probabilities corresponding to the jth examiner. Thus, the vector of misclassification parameters is

$$\boldsymbol{\lambda} = (\lambda_{10}, \lambda_{20}, \cdots, \lambda_{16,0}, \lambda_{11}, \lambda_{21} \cdots, \lambda_{16,1})'.$$

The corrected success probability is

$$\psi_{i(j)k} = \lambda_{0j} + (1 - \lambda_{0j} - \lambda_{1j})\, \pi_{i(j)k}, \tag{4.6}$$

so that the corrected likelihood of the random intercepts binary regression model is given by

$$L(\boldsymbol{\beta}, \sigma^2) = \prod_{k=1}^{N} \int \prod_{i=1}^{n_k} [\psi_{i(j)k}]^{y_{ik}} [1 - \psi_{i(j)k}]^{y_{ik}} \times -\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} u_k^2\right) du_k. \tag{4.7}$$

We use the notation $\pi_{i(j)k}$ and $\psi_{i(j)k}$ since the ith child is scored by the jth examiner and is located in the kth school.

4.5 Estimating the parameters under misclassification

4.5.1 Frequentist approach

The parameters of interest are $\boldsymbol{\beta}$ and σ^2. These parameters could be estimated by maximizing the likelihood function given by expression (4.7) with respect to $\boldsymbol{\beta}$ and σ^2 and with respect to the $\boldsymbol{\lambda}$ parameters using the

data set combining the cross-sectional and the validation data. For instance the SAS procedure `NLMIXED` could be used for this by writing a dedicated program employing the delta method for the propagation of errors. However, our experience with the `NLMIXED` procedure is that it can be quite cumbersome to achieve convergence even when based on adaptive Gaussian quadrature (see, e.g., Lesaffre and Spiessens, 2001).

4.5.2 Bayesian approach

4.5.2.1 Likelihood and prior for the cross-sectional data

The likelihood of the cross-sectional data for a binary response measured with error is obtained from (4.7). We denote the corresponding density for all observations as $f(\boldsymbol{y}_M^*|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\lambda})$, where $\boldsymbol{y}_M^*$ is the total vector of observed binary caries experience responses over all children in the main (cross-sectional) study. The vague prior distributions for the parameters were described in Section 4.3.1.2.

A prior distribution for $\boldsymbol{\lambda}$ needs to be specified also, but the $\boldsymbol{\lambda}$-parameters actually pertain to the validation data. Hence, we specify a prior distribution for $\boldsymbol{\lambda}$ for the validation data set which, when combined with the observed validation data alone, then results in a posterior distribution $p(\boldsymbol{\lambda}|\boldsymbol{y}_V, \boldsymbol{y}_V^*)$ for $\boldsymbol{\lambda}$, where $\boldsymbol{y}_V$ and $\boldsymbol{y}_V^*$ is the total vector of the true (benchmark scorer) and possibly corrupted (examiner) binary responses from the validation study, respectively. The posterior distribution $p(\boldsymbol{\lambda}|\boldsymbol{y}_V, \boldsymbol{y}_V^*)$ could be used as a prior for the likelihood of the cross-sectional data.

4.5.2.2 Likelihood and prior for the validation data

The validation data provide information for estimating $\boldsymbol{\lambda}$ because we are provided with the surrogate responses $\boldsymbol{y}_V^*$ and the corresponding true (benchmark scorer) responses $\boldsymbol{y}_V$, for each of the 16 examiners. Let $\boldsymbol{M}_j$ denote the 2×2 matrix corresponding to the jth examiner with entries m_{jab} $(a, b = 0, 1)$, whereby m_{jab} is the frequency of scoring in the validation data

an "a" by the jth examiner when the benchmark scorer assigns a score "b".

It is reasonable to assume prior independence among the misclassification rates $(\lambda_{j0}, \lambda_{j1})$, $j = 1, \cdots, 16$. It is known from a classical Bayesian theory that the resulting posterior density of a Bernoulli variable with a Beta prior distribution is a Beta density. We therefore assign non-informative independent beta distributions to the misclassification probabilities, i.e. the prior distributions are Beta$(1, 1)$.

4.5.2.3 Posterior distribution $p(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}_M^*, \boldsymbol{y}_V, \boldsymbol{y}_V^*)$

For a given $\tilde{\boldsymbol{\lambda}}$ the Bayesian analysis of the cross-sectional data yield $p(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}_M^*, \tilde{\boldsymbol{\lambda}})$, and the posterior estimates obtained by `WinBUGS` are conditional on the imputed value for $\tilde{\boldsymbol{\lambda}}$. On the other hand, the validation data result in the posterior distribution $p(\boldsymbol{\lambda} | \boldsymbol{y}_V, \boldsymbol{y}_V^*)$. That is,

$$\begin{aligned} \lambda_{j0} &\sim \text{Beta}(m_{j10} + 1, m_{j00} + 1), \\ \lambda_{j1} &\sim \text{Beta}(m_{j01} + 1, m_{j11} + 1). \end{aligned} \tag{4.8}$$

This posterior distribution could thus be used as a prior distribution for $\boldsymbol{\lambda}$ in the Bayesian analysis of the cross-sectional data. However, it was not immediately clear how to do this in an elegant way using `WinBUGS` and for general prior distributions for $\boldsymbol{\lambda}$. Instead, we opted to process the cross-sectional data and the validation data simultaneously. That is, at each iteration of the Markov chain of the validation data we obtained an estimate of $\boldsymbol{\lambda}$ and this estimate was imputed into the Markov chain pertaining to the cross-sectional data. This procedure enabled us in a simple way to take into account the variability with which $\boldsymbol{\lambda}$ is estimated from the validation data. In fact, our procedure samples from

$$p(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}_M^*, \boldsymbol{y}_V, \boldsymbol{y}_V^*) = \int p(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}_M^*, \boldsymbol{\lambda}) \cdot p(\boldsymbol{\lambda} | \boldsymbol{y}_V, \boldsymbol{y}_V^*) d\boldsymbol{\lambda}.$$

Hence our procedure estimates the marginal posterior distribution of $(\boldsymbol{\beta}, \sigma^2)$ taking into account the uncertainty with which the misclassification parameters are estimated. This procedure can be applied to any model, not just to binary logistic regression model and with any prior distribution for $\boldsymbol{\lambda}$ (at least those available in `WinBUGS`, but the approach can of course also be used outside `WinBUGS`), without major modifications.

4.6 Application to the Signal Tandmobiel® study

In spite of the considerable efforts that are undertaken to calibrate examiners involved in oral health surveys, variability in scoring cannot be avoided. Since examiners often operate in well-defined geographical areas, the presence of possible bias can influence results considerably when the research question has a geographical nature. The methodology presented here offers an opportunity to refine current analytical approaches, allowing more reliable conclusions to be drawn.

In this section we present the results from the corrected binary analysis with application to Signal Tandmobiel® study. First, we show the frequentist estimates of simple kappa statistics and the Bayesian posterior estimates of the sensitivity and specificity in validation data. Second, we present the results of the main analysis corrected for misclassification, more specifically the posterior summary statistics of the regression parameters. The correction was done considering the misclassification rates for each examiner because each examiner was active in a quite limited geographical area.

4.6.1 Analysis of the validation data

Due to the relatively small number of children used per calibration exercise we combined here the validation data from the three caries calibration exercises. Table 4.3 shows the combined caries (0/1) validation data. Observe

Table 4.3: *The caries validation data combined from the 1996, 1998 & 2000 caries calibration exercises.*

Examiner	Misclassification frequencies			
j	m_{j00}	m_{j01}	m_{j10}	m_{j11}
1	17	2	2	12
2	18	0	2	15
3	19	2	1	6
4	10	0	2	12
5	15	1	2	15
6	16	0	2	13
7	8	0	1	11
8	17	1	1	12
9	1	0	1	7
10	21	0	0	14
11	10	0	1	8
12	15	1	1	13
13	16	1	2	13
14	15	2	2	10
15	20	2	0	6
16	17	0	0	13

that by pooling the data of the calibration exercises we actually underestimated the possible systematic bias of the examiners in the first year since it would be expected that the examiners became better calibrated in due time.

The scores on prevalence of caries experience of each of the 16 dental examiners were compared with the scores obtained by the benchmark examiner. As can be seen in Table 4.4, based on the scheme of agreement levels proposed by Landis and Koch (1977), all examiners had an excellent agreement with the benchmark examiner (κ above 0.80) except for the examiners 1, 3, 9 and 14 who had only a "substantial agreement" (κ between 0.61 and 0.75). Note that the upper bounds of the estimated 95% confidence intervals from the SAS (version 8.2) procedure *FREQ* for some κs are greater than 1 because of the rough asymptotic approximation.

Table 4.4: *Simple κ measuring agreement between the benchmark scorer and each of the 16 dental examiners when scoring caries experience in the calibration exercises, obtained from SAS (version 8.2) procedure* `FREQ`

	Estimate	95% CI[a]			Estimate	95% CI[a]	
	(SE)	LCI	UCI		(SE)	LCI	UCI
κ_1	0.752(0.116)	0.524	0.980	κ_9	0.609(0.340)	-0.057	1.274
κ_2	0.885(0.078)	0.732	1.039	κ_{10}	1.000(0.000)	1.000	1.000
κ_3	0.727(0.147)	0.439	1.015	κ_{11}	0.894(0.103)	0.693	1.095
κ_4	0.833(0.111)	0.615	1.051	κ_{12}	0.866(0.091)	0.687	1.045
κ_5	0.818(0.100)	0.623	1.014	κ_{13}	0.811(0.104)	0.608	1.014
κ_6	0.870(0.088)	0.698	1.043	κ_{14}	0.716(0.132)	0.457	0.974
κ_7	0.898(0.099)	0.704	1.092	κ_{15}	0.811(0.127)	0.563	1.059
κ_8	0.868(0.091)	0.690	1.045	κ_{16}	1.000(0.000)	1.000	1.000

[a]CI = Confidence interval; LCI = lower CI; UCI = upper CI.

Further, to compare scores on prevalence of caries experience of each of the 16 dental examiners with the scores obtained by the benchmark examiner, we computed the specificities and sensitivities from the validation data of the prevalence of the caries experience. The estimates of sensitivity and specificity for the 16 examiners are shown in Table 4.5. We opted for Bayesian estimation to avoid the embarrassing asymptotic properties of frequentist estimation, as seen above.

The misclassifications suggest that the prevalence of caries experience, as scored by the dental examiners, is possibly biased. Hence there is a need to correct for this bias.

4.6.2 Regression analysis of the main data with correction

Bayesian analysis was performed using the `WinBUGS` (version 1.4) program. Observe that our `WinBUGS` program simultaneously estimates the parameters of model main model and of misclassification model and hence is based on two Markov chains. Namely, one chain pertains to the validation data, sampling the conditional classification probabilities. In the other chain the

Table 4.5: *Posterior summary statistics of the misclassification rates from model (4.8) (using* `WinBUGS Program 4.2`*)*

Examiner	Specificity($\hat{\pi}_{00}$)			Sensitivity($\hat{\pi}_{11}$)		
	Estimate	95% CI[a]		Estimate	95% CI[a]	
	(SE)	2.5	97.5	(SE)	2.5	97.5
1	0.856(0.075)	0.682	0.967	0.812(0.095)	0.593	0.957
2	0.863(0.071)	0.697	0.969	0.941(0.055)	0.794	0.999
3	0.909(0.060)	0.762	0.988	0.699(0.139)	0.398	0.924
4	0.787(0.105)	0.547	0.950	0.929(0.066)	0.756	0.998
5	0.842(0.082)	0.654	0.964	0.888(0.073)	0.713	0.986
6	0.851(0.077)	0.672	0.967	0.933(0.063)	0.766	0.998
7	0.818(0.111)	0.557	0.974	0.923(0.072)	0.736	0.998
8	0.900(0.065)	0.740	0.987	0.867(0.085)	0.662	0.982
9	0.500(0.224)	0.094	0.907	0.888(0.101)	0.625	0.997
10	0.957(0.041)	0.850	0.999	0.937(0.059)	0.780	0.998
11	0.846(0.096)	0.618	0.979	0.900(0.091)	0.663	0.997
12	0.889(0.072)	0.712	0.985	0.875(0.080)	0.681	0.983
13	0.850(0.078)	0.670	0.966	0.875(0.080)	0.682	0.984
14	0.842(0.082)	0.653	0.965	0.786(0.105)	0.546	0.951
15	0.955(0.043)	0.839	0.999	0.701(0.138)	0.398	0.923
16	0.947(0.050)	0.813	0.999	0.934(0.062)	0.770	0.998

[a]CI = Credible interval

parameters $\boldsymbol{\beta}$ and σ^2 are sampled, employing the sampled $\boldsymbol{\lambda}$ from the first chain. Hence our MCMC analysis consists of an 'imputation step', i.e. where the correction factors are imputed from the MCMC analysis pertaining to the validation data, and an estimation step whereby, given the imputed correction term, the parameters of the logistic regression model are sampled.

We have chosen to use `WinBUGS`'s `cut` function on the misclassification parameters. This function prevents the cross-sectional data from giving feed-back on the estimation of the parameters $\boldsymbol{\lambda}$. Although in principle the main data could also be used to provide information (though little) on $\boldsymbol{\lambda}$ (by not using the `cut` option) we preferred not to do so for the following reasons.

(a) Our approach resembles better the classical, frequentist, approach

where the correction terms are estimated from the validation data and are imputed into the logistic regression analysis in a second step.

(b) The use of the `cut` function reduced the necessary time for convergence in our case by a factor of more than 10, one of the reasons being that with the `cut` function much fewer iterations are needed to attain convergence.

(c) For the initially fitted models the posterior estimates of the model parameters were practically the same regardless of using the `cut` function or not.

Five initially overdispersed chains were initiated. After the first 5000 iterations, Gelman and Rubin's shrinkage factors as well as Geweke's Z-scores and QQ-plots (Best, Cowles, and Vines, 1996) were evaluated for every chain in batches of 5000 iterations. This was continued until 20,000 iterations were seen, where convergence was diagnosed for all regression parameters in three of the five chains. The posterior summary statistics were computed after obtaining convergence (which was judged by the three diagnostic procedures) from an extra 10,000 iterations.

The posterior estimates of the regression parameters from the corrected binary logistic model are shown in Table 4.6. The posterior means of the geographical regression parameters are usually larger in absolute value than the corresponding means from the uncorrected model. This shows an improvement in the parameter estimates as they are pulled away from the null. However, the standard errors of the estimates are increased as a result of the sampling variability in estimating $\boldsymbol{\lambda}$. But, more importantly, here the East-West gradient remains significant (in a Bayesian sense).

4.6.3 Sensitivity analysis

We performed a sensitivity analysis for the choice of the prior distributions for $\boldsymbol{\beta}$ and σ^2. First, a sensitivity analysis was performed by changing the

Table 4.6: *Parameter estimates from the corrected random-effects simple logistic regression model (4.7) in combination with misclassification model (4.8) predicting the prevalence of caries experience (using* `WinBUGS Program 4.3`*).*

Parameter	Estimate (SE)	95% CI[a]		Bayesian[b] $p-$value
		2.5	97.5	
Intercept	0.438(0.158)	0.135	0.759	0.0035
x-ordinate	0.290(0.117)	0.064	0.525	0.0046
y-ordinate	0.011(0.099)	−0.184	0.208	0.3651
Gender	−0.074(0.151)	−0.376	0.222	0.2211
Age	0.519(0.155)	0.219	0.830	0.0008
σ^2	0.606(0.156)	0.353	0.965	<0.001

[a]CI = Credible interval.

[b]The Bayesian $p-$value is calculated as explained in Section 3.4.4.

prior distribution for σ^2 from $IG(10^{-3}, 10^{-3})$ to a Uniform$(0, 100)$ distribution for σ giving practically the same results. Second, changing the prior distribution of the regression coefficients from a normal to a t-distribution with 4 degrees of freedom also gave very similar results.

4.7 Discussion

Large-scale epidemiologic studies necessarily involve multiple examiners, due to a large number of subjects to be examined and some unavoidable organizational aspects, like geographical locations. This implies that the collected data are inevitably subject to measurement error, and hence it is necessary to investigate the impact of such errors.

In the analysis, we have opted for a Bayesian approach for two reasons. First, the Bayesian approach allows for the incorporation of oral health knowledge into the statistical analysis. Although we have not done so here, we believe that this is an important feature of the approach. Indeed, the validation datasets are most often quite small, implying that the correction

terms are then (relatively) poorly estimated. In that case, any external useful oral health information can improve the stability of the estimated correction terms. Secondly, the Bayesian software provides a flexible way to fit quite complex statistical models and to switch from one model to another with a limited amount of extra work, usually implying much less analytical work, which can be quite cumbersome once one deviates from classical statistical approaches.

Finally, despite the fact that a gold standard was not available in our study, but only a benchmark examiner, our analysis is not invalidated. Indeed, our regression coefficients estimate a binary logistic regression model as if all children were scored by the same individual, in this case the benchmark examiner. Of course, if the benchmark examiner also scores with error, then some attenuation will still be present in the analysis.

CHAPTER

5 Analysis of Ordinal Data Subject to Response Misclassification

5.1 Introduction

Ordinal variables are common in epidemiological research. Examples are severity of a disease (none, mild, moderate, severe), agreement ratings (disagree, undecided, agree), smoking status (nonsmoker, light smoker, heavy smoker), and so on. However, an ordinal variable represents often a subjective qualification and is therefore more prone to misclassification than a numerical variable. As seen above, this misclassification needs to be taken into account in statistical analyses.

Response misclassification has been considered only for the binary data (e.g., Hausman, Abrevaya, and Scott-Morton, 1998; Neuhaus, 1999, 2002). Although the extension to ordinal misclassification does not pose any major methodological obstacles it has not been considered in the literature until recently. In this chapter we illustrate the ordinal response misclassification using a categorized score for caries experience. The dmft-index for the ith

child at school k scored by examiner j was split up according to:

$$y_{ijk} = \begin{cases} 1 & \text{if the dmft-index for the } i\text{th child is 0 (no caries experience),} \\ 2 & \text{if the dmft-index for the } i\text{th child is 1,} \\ 3 & \text{if the dmft-index for the } i\text{th child is in (1,4],} \\ 4 & \text{if the dmft-index for the } i\text{th child is in (4,20].} \end{cases}$$

5.2 Cumulative logit random effects model

Let the ordinal response Y_{ik} take possible values in $\{1, ..., R\}$. The threshold model for an ordinal response posits a latent variable S, such that one observes $Y_{ik} = r$ if S is between α_{r-1} and α_r. Suppose that S has a cumulative density function (CDF) $G(s - \eta)$, with η related to covariates by

$$\eta_{ik} = \boldsymbol{x}'_{ik}\boldsymbol{\beta} + \boldsymbol{z}'_{ik}\boldsymbol{u}_k,$$

where $\boldsymbol{x}_{ik}$ is a $d-$dimensional vector of known covariates with fixed regression coefficient $\boldsymbol{\beta}$ and $\boldsymbol{z}_{ik}$ is a $q-$dimensional vector of known covariates for a vector $\boldsymbol{u}_k \sim \mathcal{N}(0, \boldsymbol{D})$ of random effects accounting for within-cluster (school) correlation. Then CDF for y_{ik} conditional on $\boldsymbol{u}_k$ is modeled by

$$\Pr\left(Y_{ik} \leq r|\boldsymbol{u}_k\right) = \Pr\left(S \leq \alpha_r|\,\boldsymbol{u}_k\right) = G(\alpha_r - \boldsymbol{x}'_{ik}\boldsymbol{\beta} - \boldsymbol{z}'_{ik}\boldsymbol{u}_k).$$

The inverse of the CDF of $G(\cdot)$ serves as the link function. Further, α_r is the rth ordered category cut-off parameter, satisfying $\alpha_1 < \alpha_2 \cdots < \alpha_{R-1}$ and depending on the values of the regression variables and the random components.

The most popular ordinal regression model, with logit link

$$G(\cdot) = \frac{\exp(\cdot)}{1 + \exp(\cdot)},$$

is the *cumulative logit* model. A random-effect version has the expression

(see, e.g., Hartzel, Agresti, and Caffo, 2001)

$$\text{logit}\left[\Pr(Y_{ik} \leq r|\boldsymbol{u}_k)\right] = \eta_{ikr} = \alpha_r - \boldsymbol{x}'_{ik}\boldsymbol{\beta} - \boldsymbol{z}'_{ik}\boldsymbol{u}_k. \tag{5.1}$$

It is assumed here that the effect of covariates is the same for all logits. This is called the *proportional odds* assumption. Thus the probability of subject i in cluster k being classified in category r of the ordinal caries experience response is

$$\pi_{ikr} = \Pr\left(Y_{ik} = r|\boldsymbol{u}_k\right) = G\left(\eta_{ikr}\right) - G\left(\eta_{ik,r-1}\right)$$

with $\eta_{ik0} = -\infty$.

For subject i in cluster k define $w_{ikr} = 1$ if $Y_{ik} = r$ $(r = 1, \cdots, R)$ and $w_{ikr} = 0$ otherwise. Then $\boldsymbol{w}_{ik} = (w_{ik1}, \cdots, w_{ikR})'$ is a $R-$dimensional vector following the multinomial distribution:

$$\boldsymbol{w}_{ik} \sim \text{Multinomial}\big(1, \boldsymbol{\pi}_{ik}\big),$$

where $\boldsymbol{\pi}_{ik} = (\pi_{ik1}, \cdots, \pi_{ikR})'$ with $\pi_{ikR} = 1 - \sum_{r=1}^{R-1} \pi_{ikr}$. Let $f_y(y_{ik}, \boldsymbol{\alpha}, \boldsymbol{\beta}|\boldsymbol{u})$ be the multinomial mass function and ϕ be the multivariate normal density function with mean $\mathbf{0}$ and covariance $\boldsymbol{D}$. The marginal likelihood for the cumulative logit random effects model is thus

$$l(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{D}) = \sum_{k=1}^{N} \log \int\limits_{\Re^q} \left[\prod_{i=1}^{n_k} f_y(y_{ik}, \boldsymbol{\alpha}, \boldsymbol{\beta}|\boldsymbol{u}_k)\right] \phi(\boldsymbol{u}_k|\boldsymbol{D}) d\boldsymbol{u}_k. \tag{5.2}$$

5.2.1 Fitting a cumulative random effects logit model

For the Signal Tandmobiel® data, the ordinal response variable takes $R = 4$ possible values, and thus, the ordered category cut-off parameters are α_1, α_2, α_3. Let $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)'$. Similar to Chapter 4, we considered here an ordinal logit regression model with random school intercepts, u_k with $u_k \sim \mathcal{N}(0, \sigma^2)$.

5.2.1.1 Frequentist approach

The computation of the maximum likelihood estimate (MLE) of the parameter vector $(\boldsymbol{\alpha}', \boldsymbol{\beta}', \sigma_u^2)$ is a complex task: the likelihood (5.2) is, in general, not analytically tractable. However, maximum likelihood estimation can be done using the SAS procedure `NLMIXED`.

5.2.1.2 Bayesian approach

For a Bayesian approach, the unknown parameters are treated as random variables. Let $p(\boldsymbol{\alpha})$, $p(\boldsymbol{\beta})$ and $p(\sigma^2)$ be prior distributions for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and σ^2, respectively. Then the posterior distribution $p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{u}|\boldsymbol{y})$ is given by

$$\begin{aligned} p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{u}|\boldsymbol{y}) &\propto \prod_{k=1}^{N}\prod_{i=1}^{n_k} p(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{u}|\boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\alpha}) p(\boldsymbol{\beta}) p(\sigma^2) \\ &= \prod_{k=1}^{N}\prod_{i=1}^{n_k} f_y(y_{ik}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{x}_i, u_k) \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{u_k^2}{2\sigma^2}) \\ &\qquad \times p(\boldsymbol{\alpha}) p(\boldsymbol{\beta}) p(\sigma^2) \end{aligned} \tag{5.3}$$

Bayesian estimation of expression (5.3) is possible via Gibbs and Metropolis-Hastings (MH) sampling. We can sample from `WinBUGS` through the fully conditional posteriors $[\boldsymbol{\alpha}|\boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{y}]$, $[\boldsymbol{\beta}|\boldsymbol{u}, \boldsymbol{\alpha}, \boldsymbol{y}]$ and $[\sigma^2|\boldsymbol{u}]$ in a similar way to the Bayesian analysis in Chapter 4.

The assumed prior distributions chosen for this ordinal random-effects logit model are as follows:

(a) for the regression coefficient β_s $(s = 1, \cdots, d)$, a vague normal prior was assumed i.e, $\beta_s \sim \mathcal{N}(0, 10^6)$;

(b) the prior distribution for σ^2 was taken as $IG(10^{-3}, 10^{-3})$ but for the same reason as in Chapter 4 a sensitivity analysis was also performed (see Section 5.5.3);

(c) a normal vague prior is taken for the first category cutoff, i.e. $\alpha_1 \sim$

$\mathcal{N}(0, 10^6)$ and combined with a truncated normal vague prior for the other category cutoffs, i.e. $\alpha_2 \sim \mathcal{N}(0, 10^6)I(\alpha_1, \alpha_3)$ and $\alpha_3 \sim \mathcal{N}(0, 10^6)I(\alpha_2, +\infty)$, where $I(a, b)$ is the truncation function for the interval (a, b). Note that α_2 and α_3 have truncated prior distributions because of the constraint $\alpha_1 < \alpha_2 < \alpha_3$.

5.2.2 Application to the Signal Tandmobiel® study

The left hand side (LHS) of Table 5.1 shows the result of fitting model (5.1) to the cross-sectional caries experience data by using `WinBUGS` (version 1.4) without taking into account examiners' effect. The results clearly indicate a significant East-West gradient in the degree of caries experience, being higher in the Eastern part of Flanders (the province of Limburg) (Figure 5.1).

When adding 'examiner' to model (5.1) as a fixed effect, to account for its confounding effect, the geographical East-West trend was clearly attenuated but remained significant; see the right hand side (RHS) of Table 5.1. Observe that for the binary response (Chapter 4) the geographical East-West trend vanished after adding the fixed examiners' terms to the model. Thus, this demonstrates that the ordinal response provides extra information over and above the binary response. Thus there seems to be a genuine (local) geographical East-West trend in the degree of caries experience. However, following the same argument as in Chapter 4, this approach is not the best choice to correct for examiner misclassification.

Model (5.1) assumes that the probability of scoring r on y is the same for all examiners, and hence it ignores possible different scoring behaviour of the examiners. However, it became clear during the conduct of the Signal Tandmobiel® study that some examiners had the tendency to overscore or underscore the dmft-index compared with the benchmark examiner. In Figure 5.1 the over- and underscoring behaviour of the dental examiners shows an East-West gradient which is similar to that of the caries experience. Clearly, this over- and underscoring behaviour can have an effect

on the East-West geographical trend in caries experience. To take the examiners' effect into account properly, we opted for an ordinal logistic measurement error model that is described below.

Table 5.1: *Parameter estimates from the random-intercepts multinomial logit (5.1) model predicting the degree of caries experience, controlling for the geographical effect in two ways (using* `WinBUGS Program 5.1`*)*

		without examiners' fixed effects			with examiners' fixed effects		
		Estimate	95% CI[a]		Estimate	95% CI[a]	
Parameter		(SE)	2.5	97.5	(SE)	2.5	97.5
α_1		−0.317(0.053)	−0.42	−0.21	−0.217(0.520)	−1.30	0.68
α_2		0.164(0.053)	0.06	0.27	0.265(0.520)	−0.82	1.16
α_3		1.321(0.057)	1.21	1.43	1.426(0.521)	0.34	2.32
x-ordinate		0.198(0.042)	0.11	0.28	0.146(0.073)	0.00	0.29
y-ordinate		−0.017(0.044)	−0.10	0.07	−0.002(0.048)	−0.09	0.09
Gender		−0.062(0.061)	−0.18	0.06	−0.059(0.060)	−0.18	0.06
Age		0.309(0.082)	0.15	0.47	0.287(0.083)	0.12	0.45
	1				−0.023(0.546)	−1.15	0.94
	2				0.360(0.545)	−0.79	1.32
	3				0.230(0.543)	−0.90	1.18
	4				−0.161(0.546)	−1.29	0.80
E	5				0.231(0.564)	−0.92	1.23
X	6				0.139(0.535)	−0.97	1.07
A	7				−0.079(0.550)	−1.21	0.90
M	8				0.044(0.550)	−1.09	1.02
I	9				0.502(0.557)	−0.65	1.50
N	10				0.119(0.560)	−1.04	1.12
E	11				−0.262(0.546)	−1.38	0.71
R	12				0.185(0.569)	−0.97	1.20
S	13				−0.002(0.553)	−1.16	0.97
	14				0.345(0.551)	−0.79	1.33
	15				−0.133(0.533)	−1.24	0.80
	16				0.335(0.541)	−0.79	1.28
σ^2		0.154(0.035)	0.09	0.23	0.132(0.035)	0.07	0.21

[a]CI = Credible interval.

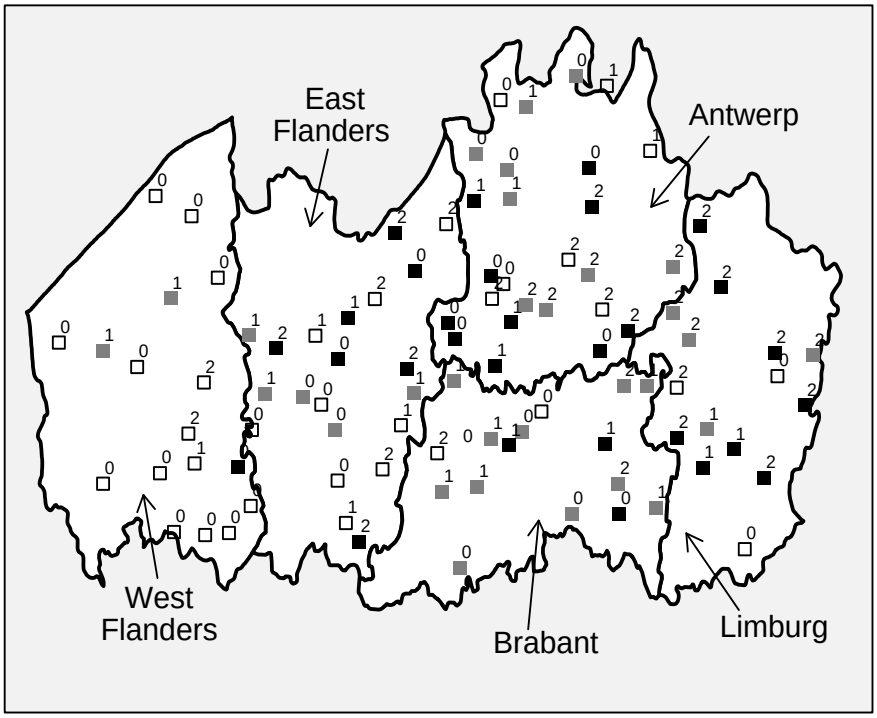


Figure 5.1: *Map of Flanders with level of caries experience and over- and underscoring of dental examiners. Caries experience was split into 3 categories according to quartiles of the mean dmft scores obtained per school and coded as* 0 *(minimum to* Q_1*),* 1 *(*Q_1 *to* Q_3*), or* 2 *(above* Q_3*). The over- and underscoring of the examiner is indicated with the symbols* □*,* ■*, and* ■*. The symbol* □ *signifies that the dental examiner scoring the respective school underscored 5% to 15% compared with the benchmark examiner in the calibration exercises. The symbol* ■ *signifies between 5% under- and 5% overscoring, and the symbol* ■ *signifies at least 5% overscoring (up to 18%).*

5.3 Corrected cumulative logit random effects

Under non-differential misclassification, i.e. assuming that the distribution of Y^* does not depend on $\boldsymbol{X}$ given Y, the corrected model for observed

ordinal response has the expression

$$\begin{aligned}\Pr(Y^* = y^* | \boldsymbol{X} = \boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\lambda}) \\ = \textstyle\sum_y \Pr(Y = y | \boldsymbol{X} = \boldsymbol{x}; \boldsymbol{\theta}) \cdot \boldsymbol{\pi}_{y^*|y}(\boldsymbol{\lambda}),\end{aligned} \tag{5.4}$$

where $\boldsymbol{\pi}_{y^*|y}(\boldsymbol{\lambda}) = \Pr(Y^* = y^* | Y = y; \boldsymbol{\lambda})$.

Using the child and examiner subscripts i and j, respectively, the probability of scoring caries experience as s (taking into account the random effect of school) is given by

$$\begin{aligned}\Pr(Y_i^* = s | \boldsymbol{x}_i, u_k; \boldsymbol{\theta}, \boldsymbol{\lambda}) = \\ \sum_{r=1}^{4} \left[\boldsymbol{\pi}_{s|r,j}(\boldsymbol{\lambda}) \times \Pr\left(Y_i = r | \boldsymbol{x}_i, u_k, \boldsymbol{\theta}\right) \right]\end{aligned} \tag{5.5}$$

for $s = 1, \ldots, 4$, where $u_k \sim \mathcal{N}(0, \sigma^2)$ is the random intercept pertaining to the school k of child i, and $\boldsymbol{\theta}' = (\boldsymbol{\alpha}', \boldsymbol{\beta}', \sigma)$. The subscript j on the classification probabilities $\boldsymbol{\pi}(\boldsymbol{\lambda})$ indicates that these probabilities possibly depend on the dental examiners.

For the model introduced in Section 5.2 the probability of observing $Y_i = r$ is given by

$$\begin{aligned}\Pr(Y_i = r | \boldsymbol{x}_i, u_k, \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma) = \\ \begin{cases} G(\alpha_1 - \boldsymbol{x}_i'\boldsymbol{\beta} - u_k) & \text{if } r = 1, \\ G(\alpha_2 - \boldsymbol{x}_i'\boldsymbol{\beta} - u_k)) - G(\alpha_1 - \boldsymbol{x}_i'\boldsymbol{\beta} - u_k) & \text{if } r = 2, \\ G(\alpha_3 - \boldsymbol{x}_i'\boldsymbol{\beta} - u_k) - G(\alpha_2 - \boldsymbol{x}_i'\boldsymbol{\beta} - u_k) & \text{if } r = 3, \\ 1 - G(\alpha_3 - \boldsymbol{x}_i'\boldsymbol{\beta} - u_k) & \text{if } r = 4. \end{cases}\end{aligned}$$

where, as before, $G(\cdot) = \exp(\cdot)/[1 + \exp(\cdot)]$ is the logistic CDF.

5.4 Estimating the parameters

5.4.1 Frequentist approach

The parameters of interest are $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, σ^2. These parameters could be estimated by maximizing the likelihood corresponding to model (5.5) with respect to $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, σ^2 and with respect to the $\boldsymbol{\lambda}$ parameters using the data set combining the cross-sectional and the validation data. For instance the SAS procedure `NLMIXED` could be used for this by writing a dedicated program employing the delta method for the propagation of errors.

5.4.2 Bayesian approach

5.4.2.1 Likelihood and prior for the cross-sectional data

The likelihood for the cross-sectional data for an ordinal response measured with error is obtained from (5.5). We denote the corresponding density for all observations as $f(\boldsymbol{y}_M^*|\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\lambda})$, where $\boldsymbol{y}_M^*$ is the total vector of observed ordinal caries experience responses over all children in the main (cross-sectional) study. The vague prior distributions for the parameters were described in Section 5.2.1.2. As explained in Section 4.5.2.1 validation data results in a posterior distribution $p(\boldsymbol{\lambda}|\boldsymbol{y}_V, \boldsymbol{y}_V^*)$ for $\boldsymbol{\lambda}$, which could be used as a prior for the likelihood of the cross-sectional data.

5.4.2.2 Likelihood and prior for the validation data

The validation data provide information for estimating $\boldsymbol{\lambda}$ because we are provided with the surrogate responses $\boldsymbol{y}_V^*$ and the corresponding true (benchmark examiner) responses $\boldsymbol{y}_V$, for each of the 16 examiners. Let $\boldsymbol{M}_j$ denote the matrix corresponding to the jth examiner with entries m_{jab} $(a, b = 1, \cdots, 4)$ whereby m_{jab} is the frequency of scoring in the validation data an "a" by the jth examiner when the benchmark examiner assigns a score "b." In first instance we assumed that the distribution $f(\boldsymbol{y}_V^*|\boldsymbol{y}_V, \boldsymbol{\lambda})$

is the product of 16 multinomial distributions. That is, we assumed for the jth examiner that the bth column of $\boldsymbol{M}_j$, i.e. $\boldsymbol{m}_{jb}$, has the distribution:

$$\boldsymbol{m}_{jb} \sim \text{Mult}\big(m_{j+b}, \boldsymbol{\pi}_{jb}\big), \tag{5.6}$$

where $\boldsymbol{\pi}'_{jb} = (\pi_{j1b}, \pi_{j2b}, \pi_{j3b}, \pi_{j4b})$ and $m_{j+b} = \sum_{a=1}^{4} m_{jab}$. Thus we assumed that the conditional probabilities $\boldsymbol{\pi}_{jb}$ of the different examiners are not related. A Dirichlet prior with parameters $(1, 1, 1, 1)$ is a natural choice for a vague prior for $\boldsymbol{\pi}_{jb}$. This model involves the estimation of 12×16 $\boldsymbol{\lambda} = (\boldsymbol{\pi}_{1,(1,1)}, \boldsymbol{\pi}_{1,(2,1)}, \boldsymbol{\pi}_{1,(3,1)}, \ldots, \boldsymbol{\pi}_{16,(4,4)})'$-parameters, where $\boldsymbol{\pi}_{j,(a,b)}$ pertains to the ath row of the probability vector $\boldsymbol{\pi}_{jb}$, and it implicitly assumes that the jth examiner could react quite differently when, say, the true score is 1 (1st column of $\boldsymbol{M}_j$) than when, say, the true score is 2 (2nd column of $\boldsymbol{M}_j$). However, as will be seen in the Section 5.5.2, no convergence was obtained with this (basic) model and further modeling was necessary to reduce the number of parameters. A number of misclassification models were tried out and we discuss four of them here.

The *first* misclassification model assumes that the jth examiner is scoring in-between "the worst score" and "the benchmark examiner score" in a similar manner for all benchmark examiner scores. In this sense, the worst scorer assigns scores furthest away from the truth (benchmark examiner score). In that case $\boldsymbol{m}_{jb}$ has the distribution:

$$\boldsymbol{m}_{jb} \sim \text{Mult}\big(m_{j+b}, w_j v_b + (1 - w_j)\boldsymbol{\pi}_b\big) \tag{5.7}$$

where w_j is an examiner specific coefficient taking values in $[0, 1]$, v_b is a vector of size 4 with the bth element equal to 1 and 0 otherwise and $\boldsymbol{\pi}_b$ represents the vector of conditional probabilities of the worst scorer when the true score is b. Now $\boldsymbol{\lambda}' = (w_1, \ldots, w_{16}, \boldsymbol{\pi}_{1,1}, \ldots, \boldsymbol{\pi}_{4,3})$, where $\boldsymbol{\pi}_{a,b}$ pertains to the ath row of the probability vector $\boldsymbol{\pi}_b$. Thus, there are $12 + 16 = 28$ parameters to estimate now. Observe that model (5.7) indeed locates each examiner in-between the benchmark examiner (v_b) and the

worst scorer ($\boldsymbol{\pi}_b$). Again vague priors for the parameters could be taken, i.e. for the w_j a natural choice for the vague prior is the uniform distribution on [0,1], while for $\boldsymbol{\pi}_b$ a Dirichlet prior with parameters $(1, 1, 1, 1)$ could be taken.

The previous misclassification model assumes that the scoring of the different examiners shows no relationship. Yet, all dental examiners received in the past a similar training, say, at college and at the calibration exercises prior to collecting the validation data. Hence for our *second* misclassification model we assumed that, although the dental examiners do not score in exactly the same manner, they might share similar experiences and therefore the w_j might have a common distribution. In other words we assumed a hierarchical misclassification model. One could for instance assume that $logit(w_j) \sim \mathcal{N}(\mu_w, \sigma_w^2)$ whereby the prior distribution for μ_w is $\mathcal{N}(0, 10^6)$ while for $\sigma_w^2 \sim IG(10^{-3}, 10^{-3})$ could be taken. Now $\boldsymbol{\lambda}' = (\mu_w, \sigma_w, \boldsymbol{\pi}_{1,1}, \ldots, \boldsymbol{\pi}_{4,3})$ which implies that now we need to estimate 18 parameters for the validation data.

Up to now, the ordinal nature of the scores has not been used in the misclassification model. One might argue that it is improbable that the dental examiners would score far away from the benchmark examiner. Therefore we assumed in the *third* misclassification model that it is impossible that the dental examiner assigns a score which differs from the benchmark examiner score with more than three classes. More specifically, the third model assumes that for each examiner there is a "latent" score which differs from column to column (score "b"), denoted as c_{jb}, and which yields a score "a" with a probability given by

$$\boldsymbol{m}_{jb} \sim \text{Mult}\big(m_{j+b}, \boldsymbol{\pi}_{jb}\big), \tag{5.8}$$

where now

$$\pi_{jab} =$$
$$\begin{cases} \Phi_{\tau_b}(\zeta_{ab} - c_{jb}) & \text{if } a = 1, b = 1, 2, 3; a = 2, b = 4; \\ \Phi_{\tau_b}(\zeta_{ab} - c_{jb}) - \Phi_{\tau_b}(\zeta_{a-1,b} - c_{jb}) & \text{if } a = 2, b = 1, 2, 3; a = 3, b = 2, 4; \\ 1 - \Phi_{\tau_b}(\zeta_{ab} - c_{jb}) & \text{if } a = 3, b = 1; a = 4, b = 2, 3, 4; \\ 0 & \text{if } a = 1, b = 4; a = 4, b = 1. \end{cases}$$

In other words, we assume a latent continuous scoring scale for each examiner with a cumulative normal density Φ_{τ_b} with mean zero and standard deviation τ_b. Hence, the terms c_{jb} are latent random examiner scores. In this way the distance between the scores is maximally 3. We assigned a uniform prior, ranging from -3 to $+3$, on the *first* ζ in each category and set succeeding ζ to be larger than the preceding ζ, i.e., $\zeta_{ab} = \zeta_{a,b-1} + \delta_{ab}$ where $\delta_{ab} \sim U[0, 3]$. Now $\boldsymbol{\lambda}' = (\tau_1, \ldots, \tau_4, \zeta_{11}, \cdots, \zeta_{43}, c_{11}, \cdots, c_{16,1}, \delta_{11}, \cdots, \delta_{42})$. Thus we have now $4 + 6 + 6 + 4 = 20$ parameters to estimate.

In the *fourth* misclassification model we assumed that the conditional misclassification probabilities are given by (Albert et al., 1997)

$$\pi_{a|b} = \begin{cases} \frac{1}{1+\sum_{c \neq b} g(c|b)} & \text{if } a = b, \\ \frac{g(a|b)}{1+\sum_{c \neq b} g(c|b)} & \text{if } a \neq b, \end{cases} \tag{5.9}$$

with $g(a|b)$ being a positive-valued function of a given b. As special cases for $g(a|b)$ they considered :

$$\log g(a|b) = \zeta_0, \tag{5.10}$$
$$\log g(a|b) = \zeta_0 + \zeta_1 |a - b|, \tag{5.11}$$
$$\log g(a|b) = \zeta_0 + \zeta_1 (a - b) \mathrm{I}(a > b) + \zeta_2 (b - a) \mathrm{I}(a < b), \tag{5.12}$$

where $\mathrm{I}(x)$ is an indicator function: $\mathrm{I}(x) = 1$ if x is true and 0 otherwise.

Misclassification model (5.10) assumes that there is a constant misclassification probability irrespective of how far the scored value of the examiner lies from the true value of the gold standard. Misclassification model (5.11) assumes, with a negative value for ζ_1, that the probability of misclassifying 'b' as 'a' decreases in a symmetrical way when 'a' moves away from 'b'. Finally, for a negative ζ_1 and a negative ζ_2, misclassification model (5.12) describes in a similar manner as model (5.11) the scoring process of the examiner relative to the benchmark examiner.

We consider here one simple extension of the model, namely:

$$\log g(a|b) = \zeta_0(b) + \zeta_1(a-b)\mathrm{I}(a > b) + \zeta_2(b-a)\mathrm{I}(a < b), \tag{5.13}$$

where $\zeta_0(b)$ is a simple function of the true value of the count, say $\zeta_0(b) = \zeta_{00} + \zeta_{01}I(b > 0)$. This extension allows to differentiate the performance of the examiner with respect to the specificity π_{00} and the other diagonal elements of the misclassification matrix. Models (5.10), (5.11), (5.12) and (5.13) will be referred to below as the symmetric $1p$ (one parameter), the symmetric $2p$ (two parameter), the asymmetric $3p$ (three parameter) and the asymmetric $4p$ (four parameter) model, respectively. The prior distribution for ζ_t $(t = 0, 1, 2)$ is $\mathcal{N}(0, 10^6)$. For this misclassification model, the misclassification parameter vector, $\boldsymbol{\lambda}$, is therefore a vector of the $\tilde{\boldsymbol{\zeta}}-$coefficients. As can be seen in Table 5.4 there are $4 \times 1 + 3 \times 6 + 2 \times 6 + 1 \times 3 = 37$ parameters to estimate.

In general, combined with the validation data, the posterior distribution $p(\boldsymbol{\lambda}|\boldsymbol{y}_V, \boldsymbol{y}_V^*)$ is obtained, which could be used as a prior for the model of the cross-sectional data. However, we have chosen to simultaneously estimate the cross-sectional and the misclassification parameters, for reasons explained below.

5.4.2.3 Posterior distribution $p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}_M^*, \boldsymbol{y}_V, \boldsymbol{y}_V^*)$

As explained in the previous chapter, we process the cross-sectional data and the validation data simultaneously. So that the sampling mechanism in `WinBUGS` alternates between the posterior density from the cross-sectional data posterior $p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}_M^*, \tilde{\boldsymbol{\lambda}})$ and the posterior density from the validation data $p(\boldsymbol{\lambda} | \boldsymbol{y}_V, \boldsymbol{y}_V^*)$. Thus we are sampling from

$$p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}_M, \boldsymbol{y}_V, \boldsymbol{y}_V^*) = \int p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}_M, \boldsymbol{\lambda}) \cdot p(\boldsymbol{\lambda} | \boldsymbol{y}_V, \boldsymbol{y}_V^*) d\boldsymbol{\lambda}.$$

Thefore our sampling procedure estimates the marginal posterior distribution of $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2)$ taking into account the uncertainty with which the misclassification parameters are estimated.

5.5 Application to the Signal Tandmobiel® study

In this section we present the results from the corrected ordinal analysis applied to the Signal Tandmobiel® study. Firstly, we show the frequentist estimates of weighted kappa statistics and the Bayesian posterior estimates of the ws in validation data. Secondly, we present the results of the main analysis corrected for misclassification.

5.5.1 Analysis of the validation data

A classical way to express the difference between the benchmark scorer and the 16 examiners is to show a measure of agreement like the weighted κ (κ_w) (Agresti, 1990, page 367):

$$\kappa_w = \frac{\sum_a \sum_b w_{ab} \pi_{ab} - \sum_a \sum_b w_{ab} \pi_{i+} \pi_{+b}}{1 - \sum_a \sum_b w_{ab} \pi_{a+} \pi_{+b}},$$

where π_{ab} denotes here the probability that the examiner scores an "a" for caries experience (row) while the benchmark examiner scores a "b"

(column), $\pi_{a+} = \sum_a \pi_{ab}$, $\pi_{+b} = \sum_b \pi_{ab}$ and the weights $w_{ab} = 1 - (a - b)^2/(I-1)^2$; $a, b = 1, \cdots, 4$; $I = 4$.

Table 5.2: *Weighted $\kappa(\kappa_w)$ measuring agreement between the benchmark scorer and each of the 16 dental examiners when scoring caries experience in the calibration exercises, obtained from SAS (version 8.2) procedure* FREQ

	Estimate	95% CI[a]			Estimate	95% CI[a]	
	(SE)	LCI	UCI		(SE)	LCI	UCI
κ_{w_1}	0.755(0.097)	0.566	0.944	κ_{w_9}	0.816(0.114)	0.593	1.040
κ_{w_2}	0.891(0.051)	0.791	0.992	$\kappa_{w_{10}}$	1.000(0.000)	1.000	1.000
κ_{w_3}	0.777(0.115)	0.552	1.002	$\kappa_{w_{11}}$	0.913(0.060)	0.795	1.031
κ_{w_4}	0.799(0.080)	0.642	0.956	$\kappa_{w_{12}}$	0.881(0.058)	0.768	0.995
κ_{w_5}	0.815(0.072)	0.673	0.957	$\kappa_{w_{13}}$	0.768(0.077)	0.616	0.919
κ_{w_6}	0.929(0.050)	0.830	1.028	$\kappa_{w_{14}}$	0.812(0.087)	0.641	0.983
κ_{w_7}	0.960(0.040)	0.882	1.038	$\kappa_{w_{15}}$	0.860(0.104)	0.656	1.063
κ_{w_8}	0.890(0.063)	0.767	1.013	$\kappa_{w_{16}}$	1.000(0.000)	1.000	1.000

[a]CI = Confidence interval. LCI = lower CI; UCI = upper CI

The weighted kappas for the 16 examiners involved in Signal Tandmobiel® study using the pooled validation data from the combined 1996, 1998 and 2000 caries calibration exercises are shown in Table 5.2. Based on the scheme of agreement levels proposed by Landis and Koch (1977) all examiners had an excellent agreement with the benchmark examiner (κ_w above 0.80) except for the examiners 1, 3 and 13 who had "only" a substantial agreement (κ_w between 0.60 and 0.79). Note that the upper bounds of the estimated 95% confidence intervals from the SAS (version 8.2) procedure FREQ for some κs are greater than 1 because of the rough asymptotic approximation.

The posterior mean of the examiner-specific coefficient 'w_j' expressing the position of the jth examiner with respect to the benchmark examiner and the worst scorer is shown in Table 5.3, for the first and second misclassification models, i.e. given by expression (5.7). Remember that an examiner with a value of w that is close to 1 has excellent agreement with the benchmark examiner. Our results to some extent confirm the conclusion obtained

Table 5.3: *Posterior summary statistics of the examiner-specific coefficients ws of the first and second misclassification model (5.7), estimated from the corresponding (using* `WinBUGS Program 5.2`*).*

Parameter	Results for $w_j \sim dbeta(1,1)$			$\text{logit}(w_j) \sim \mathcal{N}(\mu_w, \sigma_w^2)$		
	Estimate	95% CI[a]		Estimate	95% CI[a]	
	(SE)	2.5%	97.5%	(SE)	2.5%	97.5%
w_1	0.334(.167)	0.034	0.660	0.506(.146)	0.192	0.758
w_2	0.560(.171)	0.180	0.844	0.647(.134)	0.343	0.861
w_3	0.293(.193)	0.013	0.703	0.515(.172)	0.149	0.805
w_4	0.435(.178)	0.076	0.760	0.571(.142)	0.248	0.807
w_5	0.357(.178)	0.037	0.701	0.534(.149)	0.204	0.782
w_6	0.728(.145)	0.385	0.941	0.740(.120)	0.464	0.920
w_7	0.710(.183)	0.263	0.961	0.724(.142)	0.387	0.933
w_8	0.625(.166)	0.245	0.888	0.681(.129)	0.390	0.884
w_9	0.440(.223)	0.040	0.849	0.595(.168)	0.216	0.869
w_{10}	0.904(.092)	0.658	0.998	0.823(.109)	0.579	0.967
w_{11}	0.514(.219)	0.071	0.883	0.636(.158)	0.268	0.888
w_{12}	0.467(.192)	0.078	0.806	0.602(.149)	0.264	0.843
w_{13}	0.343(.168)	0.037	0.668	0.511(.145)	0.194	0.758
w_{14}	0.338(.187)	0.027	0.709	0.529(.159)	0.180	0.802
w_{15}	0.536(.217)	0.082	0.891	0.646(.156)	0.282	0.889
w_{16}	0.869(.125)	0.535	0.997	0.794(.125)	0.504	0.963
μ_w				0.606(.470)	−0.314	1.403
σ_w^2				0.732(.387)	0.093	1.446

[a] CI = Credible interval.

Table 5.4: *The selected misclassification models for the 16 dental examiners (for the fourth validation model)*[a]

Examiner	1	2	3	4	5	6	7	8
Misc model	$A3p$	$A3p$	$A3p$	$S2p$	$S2p$	$S2p$	$S1p$	$S2p$
Examiner	9	10	11	12	13	14	15	16
Misc model	$S2p$	$S1p$	$S2p$	$A3p$	$A4p$	$A3p$	$A3p$	$S1p$

[a] $S1p$ = symmetric $1p$; $S2p$ = symmetric $2p$; $A3p$ = asymmetric $3p$; $A4p$ = asymmetric $4p$

from the κ-statistics. As expected, estimates from the hierarchical model are shrunk.

For the fourth misclassification model (of Albert et al. (1997)) we observe that for many examiners we need to take the symmetric ($2p$) and asymmetric ($3p$) misclassification model and could not simplify the model (to receptively a $1p$ or $2p$ model). See Table 5.4 for the choice (based on AIC) of the misclassification models for each examiner separately. Observe that the asymmetric misclassification model $4p$ was chosen only once.

5.5.2 Regression analysis of the main data with correction

The parameter estimates of model (5.5) and (5.6) had not converged after 20,000 iterations, showing a very high autocorrelation for most of the parameters probably because of an excessive number of parameters to estimate for the misclassification model (5.6) in relation to the available validation data, rendering it an unidentifiable model. Therefore we turned to misclassification model (5.7). The `WinBUGS` program simultaneously estimates the parameters of model (5.5) and of model (5.7) and hence is based on two Markov chains. The convergence of the regression parameters was done using diagnostic procedures similar to the previous chapter.

The posterior estimates of the regression parameters from the corrected model corresponding to the first misclassification model are shown in Table 5.5. The posterior means of the geographical regression parameters are usually larger in absolute value than the corresponding means from the uncorrected model. This shows an improvement in the parameter estimates as they are pulled away from the null. However, the standard errors of the estimates are increased as a result of the sampling variability in estimating $\boldsymbol{\lambda}$. But, more importantly, the East-West gradient remains significant (in a Bayesian sense).

Table 5.6 shows the posterior estimates of the regression coefficients from the corrected model corresponding to hierarchical misclassification model. This gives even more improved parameter estimates. Again, the East-West gradient remains significant.

The posterior estimates of the regression parameters from the corrected

Table 5.5: *Parameter estimates from the corrected random-effects ordinal logistic regression model (5.5) in combination with the first misclassification model, expression (5.7) with $w_j \sim dbeta(1,1)$, predicting the degree of caries experience (using* `WinBUGS Program 5.3`*)*

Parameter	Estimate (SE)	95% CI[a]		Bayesian[b] $p-$value
		2.5%	97.5%	
α_1	−0.4391(0.1184)	−0.6762	−0.2097	0.0006
α_2	0.0683(0.1112)	−0.1496	0.2878	0.2670
α_3	1.1550(0.1270)	0.9003	1.4010	0.0000
x-coordinate	0.2252(0.0537)	0.1190	0.3315	0.0000
y-coordinate	−0.0215(0.0521)	−0.1231	0.0797	0.3450
Gender	−0.0659(0.0998)	−0.2567	0.1357	0.2421
Age	0.3334(0.0955)	0.1471	0.5241	0.0004
σ^2	0.2156(0.0497)	0.1309	0.3251	0.0000

[a]CI = Credible interval.

[b]The Bayesian $p-$value is calculated as the fraction of the number of times that a parameter is positive or negative, whichever is appropriate

model corresponding to the third misclassification model (5.8) are shown in Table 5.7. These estimates are somewhat smaller in absolute value than the posterior estimates from the other two corrected models, but the same conclusion with regard to the East-West gradient holds.

Table 5.8 shows the posterior estimates of the regression coefficients from the corrected model corresponding to the fourth misclassification model (5.9) with the examiner-specific misclassification models given in Table 5.4. The standard errors for the category cut-off are smaller compared to other corrected models but the standard errors for the regression coefficient are somewhat larger. Again, the same conclusion with regard to the East-West gradient holds.

In all misclassification models, the East-West gradient remains (significant). Hence, we can now claim that there is a genuine East-West gradient taking into account the differential scoring behaviour of the dental examiners.

Table 5.6: *Parameter estimates from the corrected random-effects ordinal logistic regression model (5.5) in combination with second misclassification model, expression (5.7) with* $\text{logit}(w_j) \sim \mathcal{N}(\mu_w, \sigma_w^2)$*, predicting the degree of caries experience (using* `WinBUGS Program 5.4`*).*

Parameter	Estimate (SE)	95% CI[a]		Bayesian[b] $p-$value
		2.5%	97.5%	
α_1	$-0.4302(0.1237)$	-0.6809	-0.1912	0.0003
α_2	$0.0538(0.1144)$	-0.1759	0.2805	0.3132
α_3	$1.1430(0.1311)$	0.8758	1.3960	0.0000
x-ordinate	$0.2278(0.0522)$	0.1279	0.3345	0.0000
y-ordinate	$-0.0239(0.0519)$	-0.1269	0.0804	0.3205
Gender	$-0.0693(0.1017)$	-0.2643	0.1322	0.2396
Age	$0.3380(0.0954)$	0.1524	0.5259	0.0000
σ^2	$0.2146(0.0499)$	0.1281	0.3232	0.0000

[a] CI = Credible interval.

[b] The Bayesian $p-$value is calculated as the fraction of the number of times that a parameter is positive or negative, whichever is appropriate

For the four misclassification models one can produce posterior estimates of the misclassification tables allowing (more stable) posterior estimates of the overscoring and underscoring behaviour of the dental examiners. More specifically for each examiner we calculated the difference between the expected probability of overscoring versus the benchmark examiner with the expected probability of underscoring. These estimates give us insight about the estimated correction terms. In Figure 5.2 we show for the third misclassification model the scatterplot of the expected differences for the 16 dental examiners versus the observed differences. We observe that the estimated differences are relatively close to the observed differences and show minor shrinkage compared with the observed differences.

The scatter plots for the other three misclassification models show much more shrinkage (not shown). Hence we can expect that for the first two misclassification models the correction for the differential scoring behaviour

Table 5.7: *Parameter estimates from the corrected random-effects ordinal logistic regression model (5.5) in combination with third misclassification model (5.8) predicting the degree of caries experience (using* `WinBUGS Program 5.5`*).*

Parameter	Estimate (SE)	95% CI[a]		Bayesian[b] $p-$value
		2.5%	97.5%	
α_1	$-0.4174(0.1222)$	-0.6684	-0.1834	0.0006
α_2	$0.1189(0.1075)$	-0.0938	0.3306	0.1274
α_3	$1.3230(0.1186)$	1.0800	1.5510	0.0000
x-ordinate	$0.2244(0.0539)$	0.1188	0.3309	0.0000
y-ordinate	$-0.0170(0.0498)$	-0.1134	0.0832	0.3490
Gender	$-0.0678(0.0944)$	-0.2549	0.1173	0.2284
Age	$0.3201(0.0912)$	0.1398	0.4979	0.0000
σ^2	$0.1966(0.0439)$	0.1198	0.2921	0.0000

[a] CI = Credible interval.

[b] The Bayesian $p-$value is calculated as the fraction of the number of times that a parameter is positive or negative, whichever is appropriate

of the examiners will be less than for the third model as a correction is more needed when the underscoring and/or overscoring is more pronounced and the correction terms are based on the estimated misclassification table. From Tables 5.5, 5.6, 5.7 and 5.8 we can see that in absolute value the regression coefficients corresponding to the third and fourth misclassification model are the smallest, confirming the more pronounced correction of the regression coefficients of the cross-sectional model with these misclassification models.

5.5.3 Bayesian sensitivity analysis

We performed for the ordinal logistic regression model in combination with the four misclassification models, sensitivity analyses with respect to the choice of the prior distributions. Remember though that the first mis-

Table 5.8: *Parameter estimates from the corrected random-effects ordinal logistic regression model (5.5) in combination with fourth misclassification model (5.9) predicting the degree of caries experience (using* `WinBUGS Program 5.6`*).*

Parameter	Estimate (SE)	95% CI[a]		Bayesian[b] $p-$value
		2.5%	97.5%	
α_1	$-0.335(0.073)$	-0.480	-0.195	0.0000
α_2	$0.008(0.071)$	-0.133	0.148	0.9128
α_3	$1.143(0.081)$	0.985	1.302	0.0000
x-ordinate	$0.216(0.059)$	0.104	0.334	0.0004
y-ordinate	$-0.008(0.054)$	-0.113	0.099	0.8638
Gender	$-0.067(0.079)$	-0.223	0.090	0.3914
Age	$0.319(0.098)$	0.129	0.513	0.0010
σ^2	$0.228(0.054)$	0.138	0.349	0.0000

[a] CI = Credible interval.

[b] The Bayesian $p-$value is calculated as the fraction of the number of times that a parameter is positive or negative, whichever is appropriate

classification model did not yield convergence. For instance, a sensitivity analysis was performed by changing the prior distribution for σ^2 from $IG(10^{-3}, 10^{-3})$ to a Pareto$(0.5, 0.01)$ distribution for $1/\sigma^2$ giving practically the same results. Further, a sensitivity analysis by changing the prior distribution of the regression coefficients from a normal to a t-distribution with 4 degrees of freedom also gave very similar results. However since the prior distributions for the parameters of the validation data in the first two models were the classical uniform priors for each parameter separately we felt that a sensitivity analysis based on other vague priors was less compelling. Of course, we could have taken informative priors expressing our believe that lumping together the four calibration exercises underestimates the under- and overscoring behaviour of the examiners. This can easily be done by taking another Beta-distribution for the "w" parameters and another Dirichlet prior for the conditional classification probabilities. For the

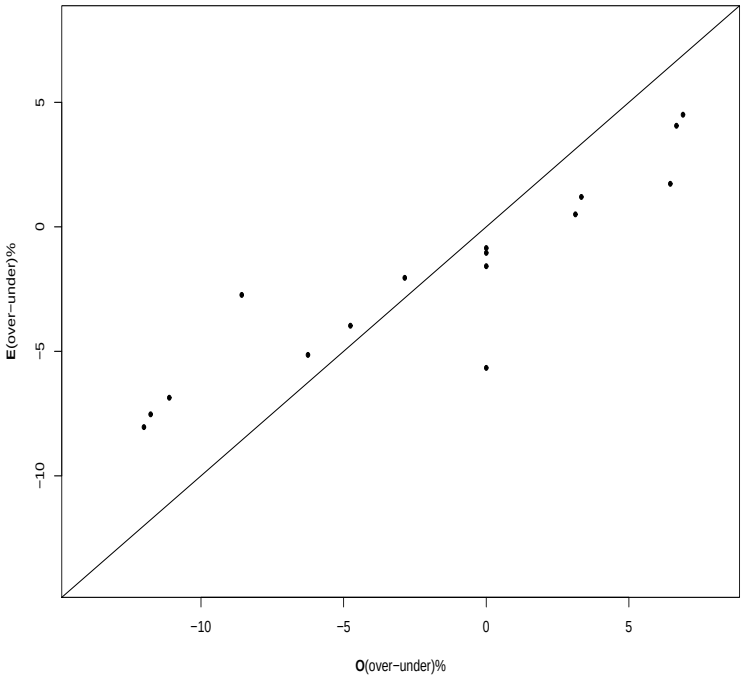


Figure 5.2: *Scatter plot of* $\boldsymbol{O}$(overscored − underscored)% *on the x-axis (i.e. the difference of the sum of the elements in the lower diagonal from the sum of the elements in the upper diagonal of the* 4×4 *observed misclassification matrix* $\boldsymbol{M}$ *divided by the sum of all elements of* $\boldsymbol{M}$ *times 100) versus* $\boldsymbol{E}$(overscored − underscored)% *on the y-axis (i.e. the corresponding value based on the expected value of the third misclassification model (5.8))*

second misclassification model the $logit - normal$ assumption of the w_j was relaxed to the $logit - t(4)$ assumption with the same prior distributions for μ_w and σ_w. Additionally, the uniform priors for μ_w and σ_w were changed to $\mathcal{N}(0, 0.05)$ and $IG(1, 1)$. For the third misclassification model we also changed the distribution of $first\ \lambda$ in each category and δ to $\mathcal{N}(0, 0.01)$ and $IG(1, 1)$, respectively. Changing the prior distributions and the likelihood models did not result in appreciable differences.

Finally, we monitored the *Deviance Information Criteria*, or DIC (Spiegel-

halter, Carlin, Best, and van der Linde, 2002) and related statistics to compare and evaluate the complexity of the misclassification models. The DIC-values of the corrected models, pertaining to the main analysis, corresponding to the last four models are 11229.6, 11220.6, 11363.2 and 11446, respectively. This suggests that the second (hierarchical) model may be the most simplest one.

5.6 Discussion

The estimates from the corrected ordinal regression are typically more precise than the estimates from the corrected binary regression (Chapter 4) because an ordinal response contains more information than a binary response. Further, the ordinal response allows for various misclassification models as compared to the binary response.

As a general conclusion we can now state that the East-West gradient remains important under the different statistical models we considered for the cross-sectional and validation data.

CHAPTER
6 Analysis of Count Data Subject to Response Misclassification

6.1 Introduction

The most popular model for count data is the Poisson model. This model can involve covariates leading to a Poisson regression model. When the counts show more variability than the Poisson model, then overdispersion can be modeled by assuming that the parameter of the Poisson model has itself a distribution which varies over the subjects. A popular model accounting for overdispersion is the negative binomial model when the Poisson mean has a gamma distribution. However, often the counts show an excess of zeroes compared to what is expected from the negative binomial model. To account for this inflated number of zeroes one traditionally assumes that the distribution is a mixture of the negative binomial distribution and a degenerate distribution at zero. This model is then coined as the zero-inflated negative binomial (regression) model. These models and random-effects versions thereof are described in Hall and Berenhaut (2002) and in the references of that paper.

Unfortunately, in practice counts are often recorded with error due to over- or underreporting of the count or, in other words, due to misclassification of the count. To obtain unbiased estimates of the regression coefficients a correction for misclassification is needed. Bratcher and Stamey (2002) and Stamey, Young, and Bratcher (2004) considered Bayesian approaches to correct for misclassification for a Poisson model. Using a binomial model for the number of false negatives and a Poisson model for the number of false positives, they obtained a closed form expression for the possibly corrupted counts. But, to our knowledge, no approach has been suggested to correct for misclassification for more complicated models for counts.

In this chapter we present various count data models, and mainly focus on the zero-inflated negative binomial (ZINB) model. To illustrate our approach we fitted the ZINB (regression) model to the dmft-index. Although, the dmft-index is bounded, the ZINB model gave an excellent fit to the distribution of the dmft scores obtained from the Signal Tandmobiel® study.

Further, we regressed the dmft-index on brushing and dietary behavior covariates, as well as on the geographical co-ordinates of the school to which the child belonged. As in previous chapters, the misclassification was fitted in a pooled and an examiner-specific manner.

6.2 Models for count data

6.2.1 Introduction

In this section, we consider and compare different models for count data – both allowing for overdispersion and taking into account also the possibility of extra zeros. In the Poisson model, the mean and variance are assumed equal. However, when the variance exceeds the mean one speaks of overdispersion. For overdispersed count data the generalized Poisson (GP), negative binomial (NB) or Poisson-inverse Gaussian (PIG) distribution can be used. These models allow for overdispersion by adding a scale

parameter.

Further, in order to account for extra zero observations in the count data we also considered the zero-inflated counterparts, namely, Poisson (ZIP), zero-inflated generalized Poisson (ZIGP), zero-inflated negative binomial (ZINB) and zero-inflated Poisson-inverse Gaussian (ZIPIG). The zero-inflated models assume a mixture distribution with one part following the reference distribution and the other part being degenerate at zero. We describe these models in the following section.

6.2.2 Poisson model

The Poisson distribution is the standard distribution for modeling count data. The Poisson distribution assumes for each observation Y_i $(i = 1, 2, \cdots, N)$ that

$$\Pr(Y_i = y_i|\psi) \;=\; \frac{\psi^{y_i}\exp(-\psi)}{y_i!} \qquad y_i \;=\; 0,\, 1,\, 2, \cdots, \tag{6.1}$$

where $\psi \;=\; \mathrm{E}(Y_i)$ is the mean value of Y_i. A key property of the Poisson distribution is that the mean and variance are equal, thus

$$\mathrm{E}(Y_i) \;=\; \mathrm{Var}(Y_i) = \psi. \tag{6.2}$$

The Poisson distribution arises when $Y_i \;=\; \sum_{s=1}^{S} Z_{si}$, whereby Z_{si} are independent binary random variables with $\Pr(Z_{si}) = \pi$ for $s = 1, \cdots, S$; $i = 1, \cdots, N$. A Poisson regression model is obtained by allowing ψ to depend on covariates through a link function. A common choice is the log link, i.e.

$$\log(\psi_i) \;=\; \boldsymbol{x}_i'\boldsymbol{\beta}, \tag{6.3}$$

where $\boldsymbol{x}_i$ is a d-dimensional vector of covariates and $\boldsymbol{\beta}$ is the corresponding vector of regression coefficients.

In caries research, the Poisson distribution is often inappropriate for modeling the dmft- and dmfs-index, because of the (high) correlation be-

tween the binary random variables. Further, the distribution of the dmft- and dmfs-index assumes finite values in contrast to the Poisson distribution. Thus, at best a distribution of dmft- and dmfs-index can be described by a truncated (at 20 and 88 for the dmft- and dmfs-index, respectively) Poisson distribution. When ψ is small, ignoring truncation has only a minor impact.

We observed overdispersion for the dmft-index in the Signal Tandmobiel® study. We therefore look here for extensions of the Poisson distributions which allow $\mathrm{Var}(Y_i) > \mathrm{E}(Y_i)$. Each time we critically examine the appropriateness of the distribution to model the distribution of the dmft-index.

6.2.3 Generalized Poisson (GP) model

An alternative to the standard Poisson distribution is the generalized Poisson (GP) distribution (Satterthwaite, 1942; Consul and Jain, 1973; Consul, 1989), also known as the Lagrangian Poisson distribution (Johnson, Kotz, and Kemp, 1992). The generalized Poisson (GP) distribution, for count response Y_i, with parameters ψ and ω is given by:

$$\begin{aligned}\Pr(Y_i = y_i|\psi,\omega) \;=\; (1-\omega)\psi\frac{\{(1-\omega)\psi+\omega y_i\}^{y_i-1}}{y_i!} \\ \times \exp\left(-(1-\omega)\psi-\omega y_i\right),\end{aligned} \tag{6.4}$$

where $\psi > 0$, and $0 \leq \omega \leq 1$. When $\omega = 0$ the GP distribution reduces to the Poisson distribution. The GP distribution (6.4) has mean ψ and variance $\psi(1-\omega)^{-1}$. Therefore this distribution may be suitable for count data with a sample variance (substantially) larger than the mean. However, increasing only the variance does not guarantee that the distribution of dmft-index is better approximated. Finally, a generalized Poisson regression model is obtained by relating the mean response ψ to a vector of covariates, $\boldsymbol{x}_i$, using e.g. the log-linear model (6.3).

6.2.4 Continuous mixture of Poisson distributions

Another approach to model overdispersion is by using a Poisson mixture model. We consider here a continuous mixture of Poisson distribution; in particular, the negative binomial (NB) and Poisson-inverse Gaussian (PIG) distributions. By allowing ψ to vary according to a *mixing density function* f the following probability distribution function is obtained

$$\Pr(Y_i = y_i|\theta) = \int_0^\infty \frac{\psi^{y_i}}{y_i!} \exp(-\psi) f(\psi|\theta) d\psi. \tag{6.5}$$

6.2.4.1 Negative binomial (NB) model

A negative binomial (NB) distribution is a continuous mixture of Poisson distributions, which allows the Poisson mean ψ to be gamma distributed (i.e., $f(\cdot)$ is a gamma distribution in expression (6.5)) and in this way overdispersion is modeled. Observe that this distribution is also useful when the count is made of correlated binary random variables which is the case for the dmft-index. More specifically, if $\psi \sim \text{Gamma}(\tau, \tau/\mu)$ then the NB distribution is obtained. The NB distribution is given by (see, e.g. Booth, Casella, Friedl, and Hobert, 2003)

$$\Pr(Y_i = y_i|\mu, \tau) = \frac{\Gamma(y_i + \tau)}{y_i!\Gamma(\tau)} \left(\frac{\tau}{\mu + \tau}\right)^\tau \left(\frac{\mu}{\mu + \tau}\right)^{y_i} \tag{6.6}$$
$$y_i = 0, 1, \cdots ; \ \mu, \tau > 0,$$

where $\mu = E(Y_i)$, τ is a shape parameter which quantifies the amount of overdispersion. The variance of Y_i for the NB distribution is $\mu + \mu^2/\tau$. Clearly, the NB distribution approaches a Poisson distribution when τ tends to ∞ (no overdispersion). Finally, the mean response μ can be related to a vector of covariates, $\boldsymbol{x}_i$, through a log-linear model (6.3), which then gives rise to a NB regression model.

6.2.4.2 Poisson-Inverse Gaussian (PIG) model

Dean, Lawless, and Willmot (1989) consider the Poisson-inverse Gaussian (PIG) regression model for insurance claims data. If $f(\psi|\theta)$ in (6.5) is the pdf of an inverse Gaussian (IG) distribution then

$$f(\psi|\theta) \equiv f(\psi|\mu,\alpha) = \sqrt{\frac{\alpha}{2\pi\psi^3}} \exp\left\{-\frac{\alpha(\psi-\mu)^2}{2\mu^2\psi}\right\},$$

where α is the shape parameter and $\mu, \alpha > 0$. The PIG distribution has a complicated expression, see Sichel (1974), Stein, Zucchini, and Juritz (1987) and Willmot (1987).

The mean and the variance of the PIG distribution are μ and $\mu(1+\mu/\zeta)$, respectively, where $\zeta = \sqrt{\mu^2+\alpha^2} - \mu^2$. The PIG distribution approaches the standard Poisson distribution as α tends to ∞. Finally, the mean response μ can be related to a vector of covariates, $\boldsymbol{x}_i$, through a log-linear model (6.3), which then gives rise to a PIG regression model.

6.2.5 Zero-inflated Models

The distribution of dmft-index in Signal Tandmobiel® study shows an excess of zeroes, see Figure 6.1. Two types of zeros can occur: one comes from the zero state and the other from the standard count distribution state. For example, some children may have a zero observation purely by chance, while others have a zero observation because they are protected by some genetic factors. Thus, the resulting distribution is a mixture of a standard count model, such as the Poisson or NB distribution, with one that is degenerate at zero (e.g., Lambert, 1992). For a general treatment of finite mixture distributions, we refer to McLachlan and Peel (2000).

Let $f(y_i|\boldsymbol{\theta})$ be a distribution function for count data, such as the Poisson and NB distribution, with unknown parameters $\boldsymbol{\theta}$. The zero-inflated

distribution with extra an proportion $p \in (0,1)$ of zeros is obtained from

$$Y_i = \begin{cases} 0, & \text{with probablity } p, \\ f(Y_i = y_i|\boldsymbol{\theta}) & \text{with probablity } 1-p. \end{cases}$$

More specifically, the zero-inflated $f(y_i|\boldsymbol{\theta})-$distribution, denoted as $\mathrm{ZI}f(y_i|\boldsymbol{\theta})$, is given by (Lambert, 1992; Johnson et al., 1992)

$$\Pr(Y_i = y_i|p,\boldsymbol{\theta}) = \begin{cases} p + (1-p)f(Y_i = 0|\boldsymbol{\theta}), \\ (1-p)f(Y_i = y_i|\boldsymbol{\theta}), \qquad y_i = 1, 2, \cdots . \end{cases} \tag{6.7}$$

The mean and variance of the $\mathrm{ZI}f(y_i|\boldsymbol{\theta})-$distribution are given by

$$\mathrm{E}_{\mathrm{zif}}(Y_i|p,\boldsymbol{\theta}) = (1-p)\,\mathrm{E}_{\mathrm{f}}(Y_i|\boldsymbol{\theta})$$

and

$$\begin{aligned} \mathrm{Var}_{\mathrm{zif}}(Y_i|p,\boldsymbol{\theta}) &= (1-p)\left[\mathrm{E}_{\mathrm{f}}(Y_i^2|\boldsymbol{\theta})\right] - \left[(1-p)\,\mathrm{E}_{\mathrm{f}}(Y_i|\boldsymbol{\theta})\right]^2 \\ &= (1-p)\left\{\mathrm{Var}_{\mathrm{f}}(Y_i|\boldsymbol{\theta}) + p[\mathrm{E}_{\mathrm{f}}(Y_i|\boldsymbol{\theta})]^2\right\}. \end{aligned}$$

When $p = 0$, $\mathrm{E}_{\mathrm{zif}}(Y_i|p,\boldsymbol{\theta}) = \mathrm{E}_{\mathrm{f}}(Y_i|\boldsymbol{\theta})$ the mean under $f(Y_i|\boldsymbol{\theta})$, and $\mathrm{Var}_{\mathrm{zif}}(Y_i|p,\boldsymbol{\theta}) = \mathrm{Var}_{\mathrm{f}}(Y_i|\boldsymbol{\theta})$ the variance under $f(Y_i|\boldsymbol{\theta})$.

The zero-inflated regression model relates μ and p to covariates, i.e.

$$\log(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta} \text{ and } \mathrm{logit}(p_i) = \boldsymbol{z}_i'\boldsymbol{\gamma},\ (i = 1, \cdots, n) \tag{6.8}$$

where $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ are d- and q-dimensional vectors of covariates pertaining to the ith subject, and with $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ the corresponding vector of regression coefficients, respectively. The vectors of covariates $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ could be the same or different. The zero-inflated log-likelihood given the observed data is obtained from plugging in equation (6.7) the dependence of the parameters on the covariates by using equation (6.8).

The zero-inflated model (6.7) can used in combination with any model

Table 6.1: *The zero-inflated distributions for the counts with* $E(Y_i) = (1-p)\mu$.

Distribution	Density function
ZIP	$\begin{cases} p + (1-p)\exp(-\mu), & y_i = 0, \\ (1-p)\exp(-\mu)\mu^{y_i}/y_i!, & y_i > 0. \end{cases}$ $\text{var}(Y_i) = (1-p)\mu(1+p\mu)$
ZIGP	$\begin{cases} p + (1-p)\frac{1}{\exp((1-\omega)\mu)}, & y_i = 0, \\ (1-p)\{(1-\omega)\mu\left[(1-\omega)\mu + \omega y_i\right]\}/\{y_i!\exp\left[(1-\omega)\mu + \omega y_i\right]\}, & y_i > 0. \end{cases}$ $\text{var}(Y_i) = (1-p)\mu[1/(1-\omega) + p\mu]$
ZINB	$\begin{cases} p + (1-p)(1+\mu/\tau)^{-\tau}, & y_i = 0, \\ (1-p)\{\Gamma(y_i+\tau)(1+\mu/\tau)^{-\tau}\}/\{\Gamma(y_i+1)\Gamma(\tau)(1+\tau/\mu)^{y_i}\}, & y_i > 0. \end{cases}$ $\text{var}(Y_i) = (1-p)\mu(1+p\mu+\mu/\tau)$
ZIPIG[†]	$\begin{cases} p + (1-p)\exp(\zeta-\alpha), & y_i = 0, \\ (1-p)(\mu\zeta/\alpha)\Pr(Y_i = 0), & y_i = 1, \\ (1-p)(2\mu\zeta/\alpha^2)\frac{2y_i-3}{2y_i}\Pr(Y_i = y_i - 1) + \frac{(\mu\zeta/\alpha)^2}{y_i(y_i-1)}\Pr(Y_i = y_i - 2), & y_i \geq 2. \end{cases}$ $\text{var}(Y_i) = (1-p)\mu(1+p\mu+\mu/\zeta)$

[†]The probabilities are computed recursively.

for count data. The zero-inflated distributions, together with their variance, arising from the Poisson, GP, NB and PIG count models are listed in Table 6.1.

6.2.6 Application to the Signal Tandmobiel® study

In this section we fit the distribution of the dmft-index. Actually the dmft-index is bounded but models for count data can be applied since the expected mean of the dmft-index is relatively small (mean = 2.243). As can be seen in Figure 6.1, the distribution of the dmft-index is markedly skewed, with the majority of the children having a low score for caries experience and a minority with a high score. About 44% (= (1913/4351) × 100%) of 7-year-old children presented without any sign of caries experience.

Further, from Figure 6.1 it is clear that the estimated Poisson distribution does not fit the observed distribution of the dmft-index well, especially for the low values of the dmft-index. Indeed, the distribution of the dmft-index is overdispersed with respect to a Poisson distribution. For a Poisson distribution one would expect var(y)/mean(y) to be approximately 1, but here we obtain for var(dmft)/mean(dmft) = 3.53. Böhning et al. (1999) suggested the zero-inflated Poisson (ZIP) distribution to model the DMFT-index and concluded that it gives a reasonable fit to the observed distribution. However, a ZIP model is not appropriate when the non-zero part of the distribution is overdispersed with respect to a Poisson distribution.

We fitted all the count data models described above to the dmft-index using `WinBUGS` (version 1.4). The assumed prior distributions for the parameters of interest are: (a) ψ, μ, $\tau \sim IG(10^{-3}, 10^{-3})$ and, (b) p, $\omega \sim$ Uniform$(0, 1)$.

The posterior summary statistics are given in Table 6.2. The expected mean for all models is practically the same as the observed mean. The performance of the model is thus determined by the estimated variance. In this application, the Poisson and the GP models underestimate the observed variance, whereas the NB and the P-IG models overestimate the

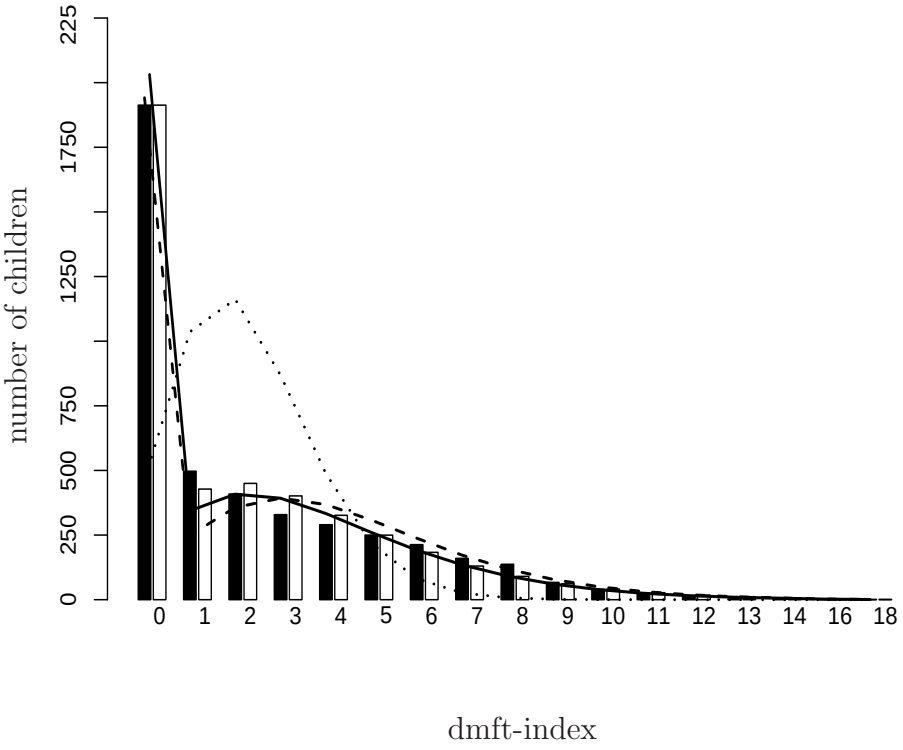


Figure 6.1: *Distribution of the dmft-index among 7-year old Flemish children,* ■ *observed,* □ *fitted from ZINB model; the dotted line shows the fit of the Poisson model, the broken line shows the fit of the pooled corrected ZINB model and the solid line shows the fit of the examiner-specific corrected ZINB model combined with the Albert et al.'s approach.*

observed variance.

Table 6.3 shows the fitted zero-inflated models to the distribution of the dmft-index. Again, the expected mean from all models are practically equal to the observed mean. The number of caries-free children is always estimated exactly equal to 1913 (observed frequency), implying that the zero-inflated models perfectly fit the caries-free group. The estimated variances from the ZINB and ZIPIG fits are closer to the observed variance compared to those of the ZIP and ZIGP fits. Observe that the ZINB model is better than ZIPIG model based on the Deviance Information Criterion

Table 6.2: *Observed distribution of the dmft-index together with expected frequencies obtained by fitting the Poisson, generalized Poisson (GP), negative binomial (NB) and Poisson-inverse Gaussian (PIG) distributions (using* `WinBUGS Program 6.1`*).*

		Expected frequencies			
dmft-index	Observed	Poisson	GP	NB	PIG
0	1913	462.0	1687.0	1773.0	1584.0
1	497	1036.0	896.9	788.8	1033.0
2	409	1162.0	529.3	491.7	573.1
3	329	868.4	338.3	335.7	335.6
4	290	487.0	228.5	239.2	212.3
5	250	218.5	160.5	174.6	143.2
6	213	81.7	116.1	129.6	101.3
7	160	26.2	86.0	97.3	74.3
8	137	7.3	64.8	73.6	56.0
9	67	1.8	49.5	56.1	43.2
10	40	0.4	38.4	43.0	33.9
11	23	0.1	30.0	33.1	26.9
12	15	0.0	23.7	25.5	21.7
13	3	0.0	18.8	19.8	17.6
14	2	0.0	15.1	15.3	14.5
16	2	0.0	9.8	9.3	12.1
18	1	0.0	6.5	5.7	10.1
Mean	2.243	2.243	2.245	2.244	2.244
Variance	7.924	2.243	5.321	11.320	14.450
π_0†	0.440	0.106	0.392	0.411	0.369
DIC		22933	17370	17151	17689

†The proportion of caries free children.

(DIC) (Spiegelhalter et al., 2002), i.e. the ZINB model has a lower DIC than the ZIPIG model. This motivates the use of the ZINB model here, as was done by Lewsey and Thomson (2004) in their study. The ZINB model fit to the dmft-index distribution of the Signal Tandmobiel® study, can be seen in Figure 6.1 and is close to being perfect. In addition, the compu-

Table 6.3: *Observed distribution of the dmft-index together with expected frequencies obtained by fitting the zero-inflated Poisson (ZIP), zero-inflated generalized Poisson (ZIGP), zero-inflated negative binomial (ZINB) and zero-inflated Poisson-inverse (ZIPIG) Gaussian distributions (using* `WinBUGS Program 6.2`*).*

		Expected frequencies			
dmft-index	Observed	ZIP	ZIGP	ZINB	ZIPIG
0	1913	1913.0	1913.0	1913.0	1913.0
1	497	193.0	416.5	428.5	398.1
2	409	378.5	455.2	449.8	463.3
3	329	494.9	409.6	401.2	423.4
4	290	485.4	331.1	326.1	339.4
5	250	380.9	250.6	249.5	252.2
6	213	249.1	181.7	182.9	179.3
7	160	139.6	127.8	130.0	124.1
8	137	68.5	88.0	90.2	84.5
9	67	29.9	59.7	61.4	57.1
10	40	11.7	40.0	41.1	38.4
11	23	4.2	26.6	27.2	25.7
12	15	1.4	17.6	17.8	17.2
13	3	0.4	11.6	11.6	11.6
14	2	0.1	7.6	7.4	7.8
16	2	0.0	3.2	3.0	5.2
18	1	0.0	1.4	1.2	3.5
Mean	2.243	2.242	2.243	2.243	2.242
Variance	7.924	6.011	8.779	8.302	8.323
π_0†	0.440	0.440	0.440	0.440	0.440
DIC		17505	16810	16794	16836

†The proportion of caries free children.

tation of the ZINB model is more efficient than that for the ZIPIG model since for the latter model the probabilities require recursive computations. Thus we shall focus on the ZINB distribution for modeling the dmft-index.

6.3 The ZINB regression model

6.3.1 The ZINB regression model formulation

The ZINB log likelihood, relating the parameters to the covariates by using equation (6.8), given the observed data can be derived from the ZINB distribution (see Table 6.1). Hence for the total sample the (minus) log-likelihood of the ZINB regression model is

$$\begin{aligned}
&\mathcal{L}_z(\boldsymbol{\beta}, \boldsymbol{\gamma}, \tau; \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}) = \\
&\quad \sum_{i=1}^{N} \log\left(1 + e^{\boldsymbol{z}_i'\boldsymbol{\gamma}}\right) - \\
&\quad \sum_{i:y_i=0} \log\left(e^{\boldsymbol{z}_i'\boldsymbol{\gamma}} + \left(\frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}} + \tau}{\tau}\right)^{-\tau}\right) + \\
&\quad \sum_{i:y_i>0} \left(\tau\, \log(\frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}} + \tau}{\tau}) + y_i\, \log(1 + e^{-\boldsymbol{x}_i'\boldsymbol{\beta}}\, \tau)\right) + \\
&\quad \sum_{i:y_i>0} \Big(\log\Gamma(\tau) + \log\Gamma(1 + y_i) - \log\Gamma(\tau + y_i)\Big),
\end{aligned} \tag{6.9}$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$ and $\boldsymbol{Z} = (\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n)$.

6.3.2 Parameter estimation

6.3.2.1 Frequentist approach

Parameter estimation can be carried out by the BFGS algorithm (Appendix A.2.3) as described in Nocedal and Wright (1999, pp. 193–201). This technique is a quasi-Newton optimization method implemented in the *optim* *R*-software package.

Parameter estimation in the likelihood approach was done using R-software calling C++ routines for fast computation of the likelihood and the first derivative. This has been implemented in our R-library `zicounts`, which is downloadable from `http://cran.r-project.org/`.

6.3.2.2 Bayesian approach

We performed a Bayesian analysis using a `WinBUGS` (version 1.4) program similar to the previous chapters. The assumed prior distributions for the ZINB and ZIBB regression models are as follows:

(a) For the regression coefficients β_s $(s = 1, \cdots, d)$ and γ_s $(s = 1, \cdots, q)$, a vague normal prior was assumed i.e., $\beta_s \sim \mathcal{N}(0, 10^6)$

(b) For the shape parameter τ, a vague inverse gamma prior was used i.e., $\tau \sim IG(10^{-3}, 10^{-3})$.

6.3.3 Application to the Signal Tandmobiel® study

We wish to examine again the geographical trend in caries experience in Flanders but now for the dmft-index as response, taking into account various risk factors. In the previous chapters an East-West gradient for caries experience at the age of seven in Flanders was demonstrated using the binarized and ordinal indicator for caries experience. Here we applied the ZINB regression model with dmft-index as response and $\boldsymbol{x}_i \equiv \boldsymbol{z}_i$, i.e. we assumed the same covariates effect on both the NB and zero part of the ZINB regression model.

The results of fitting the ZINB regression model using `WinBUGS` (version 1.4) are shown in Table 6.4. Similar to the previous chapters, we report results from the analysis (upper panel of Table 6.4) controlling only for the geographical components, age and gender. Overall we obtain similar results as before, but now the information is a bit more refined. Namely, a significant positive regression coefficient of the $x-$ordinate in the negative binomial part of the model reveals a significant East-West gradient in the degree of caries experience. On the other hand, a significant negative regression coefficient of the $x-$ordinate in the zero-inflated part of the model signifies a significant East-West gradient in the prevalence of caries. Thus, in fact the ZINB regression model allows to assess both the degree and prevalence of caries experience.

Table 6.4: *Parameter estimates from the multiple ZINB regression model predicting the dmft-index (using* `WinBUGS Program 6.3`*).*

Parameter	Negative binomial part			Zero-inflated part		
	Estimate (SE)	95% CI[a] 2.5%	97.5%	Estimate (SE)	95% CI[a] 2.5%	97.5%
without brushing & dietary habits terms						
Intercept	1.237(0.030)	1.177	1.295	−0.501(0.066)	−0.632	−0.375
x-ordinate	0.064(0.020)	0.024	0.104	−0.165(0.043)	−0.250	−0.081
y-ordinate	−0.038(0.022)	−0.083	0.005	0.059(0.045)	−0.030	0.146
Gender (girl)	0.018(0.040)	−0.059	0.094	0.014(0.085)	−0.153	0.181
Age (years)	0.069(0.050)	−0.028	0.170	−0.376(0.107)	−0.586	−0.167
τ	2.623(0.236)	2.190	3.115			
with brushing & dietary habits terms						
Intercept	1.038(0.068)	0.913	1.183	0.143(0.149)	−0.152	0.434
x-ordinate	0.068(0.021)	0.026	0.110	−0.194(0.046)	−0.283	−0.105
y-ordinate	−0.033(0.021)	−0.073	0.009	0.013(0.046)	−0.078	0.104
Gender (girl)	0.037(0.039)	−0.039	0.113	−0.008(0.088)	−0.182	0.165
Age (years)	0.060(0.051)	−0.038	0.160	−0.350(0.110)	−0.567	−0.139
Brushing frequency (< 2)	−0.008(0.055)	−0.116	0.098	−0.291(0.140)	−0.572	−0.020
Age start brushing (years)	0.029(0.018)	−0.008	0.064	−0.233(0.045)	−0.321	−0.147
Systemic fluoride (yes)	−0.081(0.041)	−0.162	−0.003	0.483(0.088)	0.312	0.658
Sugary drinks (yes)	0.199(0.044)	0.113	0.285	−0.269(0.090)	−0.442	−0.089
Between meals (> 2)	0.033(0.042)	−0.054	0.114	−0.228(0.097)	−0.422	−0.039
τ	2.721(0.241)	2.270	3.220			

[a]CI = Credible interval.

We posed the question whether the East-West gradient could be explained by an East-West gradient of the brushing and dietary habits. As can be seen in Table 6.4 (lower panel), the East-West gradient ($x-$ordinate) remains significant in both parts of the ZINB regression model. This implies that the East-West gradient in caries experience cannot be explained by a possible difference in the brushing and dietary habits. The consumption of sugar containing drinks is significant in both parts of the model, implying that it has an impact on the prevalence and the degree of caries experience. Except for gender and the $y-$ordinate, the other covariates are significant only in the zero-inflated part. Thus they only have an impact on the prevalence. The negative regression coefficient for age at start of brushing implies that the later the child starts brushing the lower the probability of being caries-free. In addition, the children who brushed their teeth inadequately and those who took more than two in-between-meals have a lower probability of being caries free. Finally, the use of systemic fluoride supplements increased the chance of being caries free.

6.4 Correction for misclassification in the main study

6.4.1 Correcting for misclassification in a regression model

Interest lies in relating Y to covariates, but if Y^* is observed instead, then the relationship will be distorted. The possibly error-corrupted count response Y^* is related to the true unobservable responses Y through

$$\Pr(Y^* = r|\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\pi}(\boldsymbol{x})) = \sum_{s=0}^{K} \Pr(Y^* = r|Y = s, \boldsymbol{x}) \Pr(Y = s|\boldsymbol{x}, \boldsymbol{\theta}), \tag{6.10}$$

where $\boldsymbol{\theta}$ contains the vector of regression coefficients and model parameters relating the true counts to the regressors, K is the maximal value of Y, and $\boldsymbol{\pi}(\boldsymbol{x})$ represents the misclassification matrix under possibly differential

misclassification. Expression (6.10) consists of:

(a) the misclassification model for Y^* given the true response and covariates, i.e. $\Pr(Y^* = r|Y = s, \boldsymbol{x})$;

(b) the underlying main model of interest, i.e. $\Pr(Y = y|\boldsymbol{x}, \boldsymbol{\beta})$.

When misclassification is non-differential, the covariates provide no information about Y^* over and above what is provided by Y, so that (6.10) becomes

$$\Pr(Y^* = r|\boldsymbol{X} = \boldsymbol{x}; \boldsymbol{\theta}; \boldsymbol{\pi}) = \sum_{s=0}^{K} \Pr(Y = s|\boldsymbol{X} = \boldsymbol{x}; \boldsymbol{\theta})\pi_{rs}, \tag{6.11}$$

where $\pi_{rs} = \Pr(Y^* = r|Y = s)$. Expression (6.11) can be applied to any misclassified count data distribution by replacing $\Pr(Y = y|\boldsymbol{x}, \boldsymbol{\beta})$ with the appropriate distribution.

Suppose that there are n observations in the main data, i.e. $\{Y_1^*, \ldots, Y_n^*\}$, and that an extra m pairs of observations $\{(Y_i^*, Y_i), i = (n+1), \ldots, (n+m)\}$ constitute the validation data set, either being a random subsample from the main data or sampled to increase the efficiency in unbiasedly estimating the misclassification probabilities π_{rs} $(r, s = 0, \ldots, K)$. The estimated probabilities $\hat{\pi}_{rs}$ from the misclassification model (see Section 6.5) are imputed in equation (6.11), to estimate the parameter vector $\boldsymbol{\theta}$ using, e.g. a maximum likelihood procedure, yielding $\hat{\boldsymbol{\theta}}$.

6.4.2 Variability of the corrected estimates

For a likelihood-based method, the asymptotic covariance matrix of $\boldsymbol{\theta}$ can be derived from the second order derivatives of the log-likelihood at the final iteration, where the likelihood is derived from expression (6.11) replacing the unknown misclassification probabilities by their estimates obtained from the validation study. However, this approach does not take the sampling variability of $\hat{\pi}_{rs}$ $(r, s = 0, \ldots, K)$ into account.

The total likelihood, combining the main data and the validation data, is given by (the dependence on covariates is omitted for convenience):

$$\prod_{i=1}^{n} \Pr(Y_i^*|\boldsymbol{\theta}, \boldsymbol{\pi}) \prod_{i=n+1}^{n+m} \Pr(Y_i^*, Y_i|\boldsymbol{\phi}, \boldsymbol{\pi}), \tag{6.12}$$

where the first term is obtained from (6.11). Further, the second term splits up in the products $\prod_{i=n+1}^{n+m} \Pr(Y_i^*|Y_i; \boldsymbol{\pi})$ and $\prod_{i=n+1}^{n+m} \Pr(Y_i|\boldsymbol{\phi})$. The first product represents the misclassification probabilities and the second product pertains to the true counts. When the validation study is a random sample of the main study, $\boldsymbol{\phi} \equiv \boldsymbol{\theta}$. In this case, the main data set consists of $N = n + m$ observations. The m observations then contribute $\prod_{i=n+1}^{n+m} \Pr(Y_i|\theta)$ to the main likelihood. When the validation data is not a subsample of the main study, the relation of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ is not always clear. Often, one needs to assume that the validation data does not provide any information about the main parameter of interest ($\boldsymbol{\theta}$). Given the nature of the validation data of the Signal Tandmobiel® study (Section 2.5), this will be assumed also here. The total likelihood in (6.11), as a function $\boldsymbol{\theta}$ and Π, now becomes

$$\prod_{i=1}^{n} \Pr(Y_i^*|\boldsymbol{\theta}, \boldsymbol{\pi}) \prod_{i=n+1}^{n+m} \Pr(Y_i^*|Y_i; \boldsymbol{\pi}). \tag{6.13}$$

The second derivative matrix at the maximum likelihood estimate of $\boldsymbol{\theta}$ and Π, obtained from likelihood (6.13), yields the asymptotic covariance matrix of the estimate of $\boldsymbol{\theta}$ taking the variability into account with which π_{rs} is estimated. This approach has been implemented in Chapters 4 and 5. In Chapter 5 we have seen how a Bayesian approach can handle the variability of estimating Π. However, here the `WinBUGS` Development Interface (WBDev) was used, which enables the implementation of user defined functions into the `WinBUGS` system via compiled Pascal code. There are two main reasons for doing this: first, function evaluations asso-

ciated with user defined components can be computed much more quickly than with their BUGS-language counterparts; and second, the full flexibility of a general-purpose computer language is available for specifying each new component, for instance, function specification via the BUGS language (mainly using the `step( )` function) can be somewhat cumbersome and sometimes infeasible.

When multiple examiners are involved we could either assume: (a) that the misclassification matrix $\boldsymbol{\pi}$ is the same for all examiners and hence that it is pooled over the examiners; or (b) that the misclassification matrix varies with the examiner.

6.5 Misclassification model

A general assumption we can make is that sth column, $\boldsymbol{m}_s$, of the misclassification table with entries m_{rs} follows a multinomial distribution:

$$\boldsymbol{m}_s \sim \text{Multinomial}\big(m_{+s}, \boldsymbol{\pi}_s\big). \tag{6.14}$$

where $m_{+s} = \sum_{r=0}^{K} m_{rs}$. The multinomial estimate of π_{rs}, i.e. $\hat{\pi}_{rs} = m_{rs} / \sum_{r=0}^{K} m_{rs}$ is one possibility to estimate the misclassification probabilities. However, for a sparse table $\boldsymbol{M}$, which is often the case with counts, the multinomial estimates $\hat{\pi}_{rs}$ are either determined with high variability or do not exist, say when the benchmark examiner did not score 's' in the validation data. Clearly some modeling of the misclassification probabilities is needed to overcome this problem.

Others, (see, e.g. Espeland and Odoroff, 1985; Espeland and Hui, 1987), have suggested a log-linear misclassification model for categorical data. However, when the log-linear misclassification model is applied to count data it too experiences computational difficulties, typically if not all true counts have been observed in the validation study.

To overcome the computational difficulties with the multinomial and

log-linear misclassification models, we used Albert et al.'s approach for estimating the conditional misclassification probabilities as described in Chapter 5. Although the method has been suggested for ordinal data, this model might also be appropriate for counts. Even though this approach is reasonable for the dmft-index it is difficult to use for counts in general, especially if the number of observed levels is large.

The approach of Albert et al. (1997) implies a drastic reduction of the number of parameters to estimate. This is particularly important here since we need to estimate these parameters for each examiner separately. Another advantage of this approach is that, even when the sth column of $\boldsymbol{M}$ contains only zeros, the model allows for estimation of π_{rs}. The unknown misclassification parameter ($\boldsymbol{\lambda}$) is a vector of the $\tilde{\boldsymbol{\zeta}}-$coefficients (Section 5.4.2.2). Thus, there are 1 to 4 parameters to estimate depending on whether one chooses the symmetric $1p$ (one parameter), the symmetric $2p$ (two parameter), the asymmetric $3p$ (three parameter) or the asymmetric $4p$ (four parameter) model. Observe that for an examiner-specific correction the number of parameters in the misclassification model will range from 16 to 48. We refer also to Chapter 7 where we evaluated Albert et al.'s approach for bounded count data.

6.6 Application to the Signal Tandmobiel® study

6.6.1 Analysis of the validation data

The pooled (over the examiners) misclassification matrix and the examiner-specific misclassification matrices were estimated. For the frequentist approach and the pooled misclassification matrix, the asymmetric misclassification structure $3p$ is chosen if based on Akaike's Information Criterion (AIC) (Sakamoto, Ishiguro, and Kitagawa, 1986) and the same is true for the Bayesian approach using DIC, see Table 6.5. Indeed, the difference of AIC is equal to the difference of DIC comparing any two models in Table

Table 6.5: *The ML (*`using R Program 6.4-a`*) and Bayesian (using* `WinBUGS Program 6.4-b`*) parameter estimates of both symmetric and asymmetric misclassification model from pooled Albert et al.'s approach.*

		Likelihood approach			
Model		Estimate(SE)	95% CI[a]		AIC[b]
Symmetric $1p$	α_0	$-4.416(0.119)$	-4.650	-4.182	964
Symmetric $2p$	α_0	$-0.806(0.240)$	-1.277	-0.335	635
	α_1	$-1.253(0.153)$	-1.553	-0.953	
Asymmetric $3p$	α_0	$-0.483(0.259)$	-0.990	0.024	556
	α_1	$-2.153(0.253)$	-2.650	-1.657	
	α_2	$-0.676(0.153)$	-0.976	-0.376	
Asymmetric $4p$	α_{00}	$-0.376(0.411)$	-1.181	0.429	557
	α_{01}	$-0.129(0.382)$	-0.877	0.620	
	α_1	$-2.205(0.301)$	-2.794	-1.615	
	α_2	$-0.668(0.155)$	-0.971	-0.365	
		Bayesian approach			
Model		Estimate(SE)	95% CI[c]		DIC[d]
Symmetric $1p$	α_0	$-4.418(0.119)$	-4.659	-4.190	524
Symmetric $2p$	α_0	$-0.792(0.246)$	-1.280	-0.290	195
	α_1	$-1.270(0.157)$	-1.608	-0.978	
Asymmetric $3p$	α_0	$-0.466(0.256)$	-0.986	0.027	115
	α_1	$-2.190(0.255)$	-2.692	-1.708	
	α_2	$-0.696(0.154)$	-1.012	-0.407	
Asymmetric $4p$	α_{00}	$-0.334(0.406)$	-1.083	0.458	117
	α_{01}	$-0.150(0.386)$	-0.928	0.581	
	α_1	$-2.266(0.302)$	-2.902	-1.718	
	α_2	$-0.690(0.161)$	-1.034	-0.397	

[a]CI = Confidence interval; [b]AIC = Akaike Information Criterion.
[c]CI = Credible interval; [d]DIC = Deviance Information Criterion

6.5. For the examiner-specific correction we observe that for many examiners we need to take the asymmetric misclassification model $3p$ and could not simplify the model (to a $1p$ or $2p$ model), again based on AIC. See Table 6.6 for the choice of the misclassification models for each examiner separately. Observe that the asymmetric misclassification model $4p$ was never chosen.

Table 6.6: *The selected misclassification model for the 16 dental examiners from examiner-specific Albert et al.'s approach.*†

Examiner	1	2	3	4	5	6	7	8
Misc model	$A3p$	$S1p$	$S2p$	$A3p$	$A3p$	$S1p$	$A3p$	$S2p$
Examiner	9	10	11	12	13	14	15	16
Misc model	$S1p$	$S2p$	$A3p$	$S1p$	$A3p$	$A3p$	$S2p$	$S1p$

†$S1p$ = symmetric $1p$; $S2p$ = symmetric $2p$; $A3p$ = asymmetric $3p$.

6.6.2 Analysis of the main data

6.6.2.1 Fitting the corrected distribution of the dmft-index

The ZINB model was corrected in a pooled and in an examiner-specific way. As can be seen in Figure 6.1 the correction mechanism using the model without covariates does not give very different results. The proportion of caries-free children from the pooled and from the examiner-specific correction are about 45% and 47% respectively, which is a slight increase over the observed 44%.

6.6.2.2 Regression analysis of the main data with correction

The results of fitting the corrected ZINB regression model to the dmft-index of the Signal Tandmobiel® data are shown in Table 6.7 for the pooled correction and in Table 6.8 for the examiner specific correction. For the pooled correction, the East-West gradient ($x-$ordinate) remains significant in both parts of the corrected ZINB model. In contrast, for the examiner-specific correction the East-West gradient vanishes in the negative binomial part.

A possible and perhaps speculative explanation of the discrepancy between the results of the two corrections is that the pooled correction ignores the fact that the misclassification model has become differential. Indeed, our regression model contains the geographical co-ordinates of the school

Table 6.7: *Parameter estimates from the pooled corrected multiple ZINB regression model predicting the dmft-index combined with 3p-asymmetric Albert et al.'s misclassification model (using* WinBUGS Program 6.5*).*

Parameter	Negative binomial part			Zero-inflated part		
	Estimate (SE)	95% CI[a] 2.5%	97.5%	Estimate (SE)	95% CI[a] 2.5%	97.5%
without brushing & dietary habits terms						
Intercept	1.426(0.036)	1.357	1.496	−0.296(0.078)	−0.450	−0.145
x-ordinate	0.052(0.019)	0.015	0.089	−0.179(0.045)	−0.269	−0.091
y-ordinate	−0.032(0.021)	−0.072	0.009	0.068(0.046)	−0.023	0.158
Gender (girl)	0.018(0.043)	−0.068	0.102	0.016(0.101)	−0.185	0.213
Age (years)	0.027(0.048)	−0.067	0.121	−0.453(0.112)	−0.673	−0.232
τ	5.319(0.834)	3.957	7.200			
with brushing & dietary habits terms						
Intercept	1.283(0.070)	1.143	1.425	0.422(0.163)	0.113	0.742
x-ordinate	0.054(0.019)	0.017	0.092	−0.215(0.049)	−0.312	−0.120
y-ordinate	−0.029(0.020)	−0.070	0.011	0.019(0.050)	−0.079	0.117
Gender (girl)	0.034(0.042)	−0.047	0.117	−0.006(0.101)	−0.201	0.193
Age (years)	0.014(0.048)	−0.082	0.108	−0.460(0.118)	−0.692	−0.228
Brushing frequency (< 2)	−0.016(0.052)	−0.117	0.086	−0.300(0.150)	−0.600	−0.008
Age start brushing (years)	0.012(0.018)	−0.024	0.047	−0.261(0.050)	−0.360	−0.163
Systemic fluoride (yes)	−0.027(0.042)	−0.109	0.054	0.591(0.099)	0.398	0.783
Sugary drinks (yes)	0.159(0.044)	0.075	0.245	−0.329(0.101)	−0.524	−0.127
Between meals (> 2)	0.024(0.041)	−0.054	0.105	−0.251(0.103)	−0.458	−0.053
τ	5.506(0.885)	4.071	7.501			

[a]CI = Credible interval.

Table 6.8: *Parameter estimates from the examiner-specific corrected multiple ZINB regression model predicting the dmft-index combined with Albert et al.'s misclassification model using* `WinBUGS Program 6.6`*).*

Parameter	Negative binomial part			Zero-inflated part		
	Estimate (SE)	95% CI[a] 2.5%	97.5%	Estimate (SE)	95% CI[a] 2.5%	97.5%
without brushing & dietary habits terms						
Intercept	1.317(0.035)	1.246	1.386	−0.331(0.077)	−0.484	−0.182
x-ordinate	0.039(0.024)	−0.008	0.086	−0.171(0.052)	−0.275	−0.070
y-ordinate	−0.022(0.025)	−0.070	0.026	0.087(0.050)	−0.012	0.185
Gender (girl)	0.025(0.047)	−0.067	0.117	0.007(0.100)	−0.187	0.202
Age (years)	0.007(0.058)	−0.106	0.119	−0.395(0.117)	−0.627	−0.167
τ	3.555(0.426)	2.810	4.479			
with brushing & dietary habits terms						
Intercept	1.153(0.082)	0.997	1.312	0.324(0.165)	−0.005	0.637
x-ordinate	0.042(0.023)	−0.005	0.087	−0.203(0.054)	−0.310	−0.096
y-ordinate	−0.019(0.025)	−0.067	0.030	0.044(0.055)	−0.065	0.151
Gender (girl)	0.044(0.045)	−0.043	0.133	−0.006(0.099)	−0.203	−0.195
Age (years)	−0.009(0.057)	−0.120	0.099	−0.390(0.122)	−0.635	−0.139
Brushing frequency (< 2)	−0.002(0.058)	−0.117	0.115	−0.303(0.151)	−0.607	−0.018
Age start brushing (years)	0.017(0.021)	−0.025	0.056	−0.243(0.049)	−0.337	−0.150
Systemic fluoride (yes)	−0.047(0.047)	−0.139	0.044	0.554(0.097)	0.368	0.746
Sugary drinks (yes)	0.185(0.049)	0.090	0.281	−0.304(0.101)	−0.497	−0.099
Between meals (> 2)	0.012(0.045)	−0.075	0.100	−0.227(0.105)	−0.434	−0.027
τ	3.658(0.442)	2.877	4.613			

[a]CI = Credible interval.

to which the child pertains and this is confounded with the dental examiner (actually with the dummy variable that indexes the dental examiner) who has examined the child. Hence, as we have indicated in Section 6.4.1, we obtained a differential misclassification process even when each dental examiner misclassified the score in a non-differential manner.

The ZINB regression model together with the examiner-specific correction shows more overdispersion ($\hat{\tau} = 3.9$) than that of the pooled correction ($\hat{\tau} = 5.7$), since the variance of the ZINB distribution is equal to $(1 - \hat{p})\hat{\mu}(1 + \hat{p}\hat{\mu} + \hat{\mu}/\hat{\tau})$. Thus, the examiner-specific correction preserves the negative binomial structure of the dmft-index more than the pooled correction.

6.7 Discussion

As indicated in the introduction, fallible count data is a problem occurring in probably all applications areas, not only in the dental or the medical area. For example, Bratcher and Stamey (2002) applied their method in a large scale epidemiological study on death rates due to congenital anomalies. Further, Stamey et al. (2004) corrected for misclassification of counts in an animal abundance study.

Because Stamey et al. (2004) assumed a particular misclassification process and a Poisson model for the infallible counts, a Poisson distribution for the fallible counts was obtained. In our approach we have chosen for the ZINB model by inspecting the distribution of the fallible counts. In a second step we then assumed that the ZINB models are also suitable for the infallible counts. It must be said, though, that it is unclear what model must be assumed for the infallible counts and what statistical assumptions must be made about the misclassification process to obtain a ZINB distribution for the fallible counts. Indeed, attempts to combine a ZINB distribution with a misclassification process of the same type, yields a much more complex model than the ZINB distribution.

CHAPTER
7

Analysis of Bounded Count Data Subject to Response Misclassification

7.1 Introduction

In this chapter we present bounded count data models with correction for misclassification error. An obvious candidate for the distribution of bounded counts is the binomial distribution. When the bounded counts show more variability than the binomial, then overdispersion can be modeled by assuming that the success probability in the binomial model has a beta distribution. This gives rise to a beta-binomial model. However, when the bounded counts show an excess of zeroes compared to what is expected from the beta-binomial model, a zero-inflated beta-binomial (ZIBB) model is used. Due to extra zeros in scoring caries experience we shall focus on the ZIBB (regression) model.

In this chapter, we look at the dmft-score restricted to the 8 deciduous molars (teeth x4 and x5, with x=5,6,7,8 – see Figure 2.1), denoted by $\text{dmft}_{4,5}$-score. Among other things, we are interested in the effect of dietary and brushing behavior on the $\text{dmft}_{4,5}$-score. The ZIBB (regression) model

was thus fitted to the $\text{dmft}_{4,5}$-index, which is a dmft-index based on the first and second primary molars. The $\text{dmft}_{4,5}$-index was regressed on brushing and dietary behavior covariates, as well as on the geographical co-ordinates of the school to which the child belonged.

When the misclassification probabilities can be estimated unbiasedly, then a suitable correction mechanism can result in (nearly) unbiased estimated parameters. However, from our experience in oral health research we must conclude that validation studies are often too small implying that the corrective terms are estimated with high variability. Especially with (finite) count data the misclassification matrix is most often sparse. Hence when estimated from a multinomial model, some of the misclassification probabilities cannot be determined or the estimated misclassification probabilities will be quite unreliable yielding corrected estimated parameters with high variability and thus unclear scientific conclusions.

Here, we suggest to base the estimation of the misclassification probabilities on the *double binomial* (DB) approach or extensions thereof. The DB approach exploits the nature of a count. Namely, in order to obtain a count the sum needs to be made of binary scores, each of which is prone to misclassification. Thus, one could estimate the misclassification probabilities for count data from the misclassification table of the binary scores making up the count.

The DB can only be applied when the misclassification process of the binary scores is done independently and does not depend on the label of the binary score. These assumptions have been implicitly taken for granted in a number of applications (see e.g. Paulino, Soares, and Neuhaus, 2003; Paulino, Silva, and Achcar, 2005). However, when the counts are determined within a subject, it is not immediately clear that these simplifying assumptions hold in practice. Further, when the misclassification process is differential, this needs to be taken into account also. This is exemplified by Luan et al. (2005). Finally, Paulino et al. (2003, 2005) assumed a (random-effects) binomial regression model, which easily combines with the

misclassification model. However, in the dental example examined here, a more complex model had to be assumed. Thus, we argue that there is a need for an approach to analyze finite count data in the presence of misclassification that: (a) expresses the misclassification process in an appropriate manner; (b) maximizes the efficiency in estimating the misclassification probabilities; and (c) combines the misclassification process with an appropriate and possibly complex model for the measurement process. The misclassification was fitted in a pooled and an examiner-specific manner.

7.2 Models for bounded count data

7.2.1 Beta-binomial model

Since the response $\text{dmft}_{4,5} \equiv Y_i = Z_{i1} + \ldots + Z_{iK}$ where $K = 8$ is a finite count, a natural candidate for the distribution of Y_i is the binomial distribution. However, the Z_{ik} $(k = 1, \ldots, K)$ are correlated. In that case, the beta-binomial distribution is a possible choice, given by (Skellam, 1948; Altham, 1978; Prentice, 1986):

$$\Pr(Y_i = y_i) = \frac{\binom{K}{y_i} \prod_{h=0}^{y_i-1} (\pi + \tau h) \prod_{h=0}^{K-y_i-1} (1 - \pi + \tau h)}{\prod_{h=0}^{K-1} (1 + \tau h)} \qquad (7.1)$$

with mean $K\pi$ and variance $K\pi(1-\pi)[1+(K-1)\delta]$, where $\delta = \tau/(1+\tau)$.

7.2.2 Zero-inflated beta-binomial model

In the Signal Tandmobiel® study, large frequencies of zeros are observed relative to what is predicted by the beta-binomial distribution. Therefore, we have chosen for the zero-inflated beta-binomial (ZIBB) model assuming for the distribution of the count a mixture of a beta-binomial distribution

and a point mass at zero. The ZIBB distribution is given by

$$\Pr(Y_i = y_i) = \begin{cases} p + (1-p)g(0) & \text{if } y_i = 0; \\ (1-p)g(y_i) & \text{if } y_i > 0, \end{cases} \tag{7.2}$$

where $g(y_i)$ is the beta-binomial distribution defined by expression (7.1). The ZIBB distribution has mean $(1-p)K\pi$ and variance $(1-p)K\pi[(1-\pi)(1+(K-1)\delta)+K\pi p]$.

7.2.3 Application to the Signal Tandmobiel® study

Table 7.1: *Observed and expected distribution of the dmft$_{4,5}$ with expected frequencies obtained by fitting a beta-binomial (BB) and zero-inflated beta-binomial (ZIBB) distribution (using* `WinBUGS Program 7.1`*).*

dmft$_{4,5}$	Observed	Expected	
		BB	ZIBB
0	1972	1935.0	1972.0
1	496	616.0	501.4
2	397	407.9	400.4
3	353	317.0	340.8
4	295	264.9	297.5
5	273	230.9	261.8
6	225	207.2	229.2
7	179	190.7	195.6
8	161	181.6	152.7
Mean	2.019	1.983	2.020
Variance	6.032	6.073	6.050
π_0†	0.453	0.440	0.453
DIC		108	68

†The proportion of caries free children.

In this section we fitted the BB and ZIBB distributions to the dmft$_{4,5}$ using `WinBUGS` (version 1.4). The assumed prior distributions for the

parameters of interest are: (a) $\tau \sim IG(10^{-3}, 10^{-3})$ and, (b) $\text{logit}(p)$, $\text{logit}(\pi) \sim \mathcal{N}(0, 10^6)$.

Table 7.1 shows the fitted BB and ZIBB models to the distribution of the $\text{dmft}_{4,5}$. The fitted mean and variance are very close to the observed values for both models. Further, the number of caries-free children is estimated exactly equal to 1972 (observed frequency). The ZIBB model is to be preferred over the BB model based on the DIC.

7.3 The ZIBB regression model

7.3.1 The ZIBB regression model formulation

The ZIBB regression model relates the parameters π and p of the ZIBB distribution to covariates as follows:

$$\text{logit}(\pi_i) = \boldsymbol{x}_i'\boldsymbol{\beta} \text{ and } \text{logit}(p_i) = \boldsymbol{z}_i'\boldsymbol{\gamma}, \ (i = 1, \cdots, N). \tag{7.3}$$

The ZIBB log likelihood, relating the parameters to the covariates by using equation (7.3), can be derived from equation (7.2). Parameter estimation using a frequentist or a Bayesian approach is done as described for the ZINB regression model in Chapter 6.

7.3.2 Application to the Signal Tandmobiel® study

The results of fitting the ZIBB regression model to the $\text{dmft}_{4,5}$ with covariates using `WinBUGS` (version 1.4) are shown in Table 7.2. Overall we obtain similar results in the beta-binomial part of the ZIBB regression model as in the negative binomial part of the ZINB regression model of Chapter 6 (apart from the fact that now the regression coefficients have a different meaning). For the degenerate part (explaining 0) only age at start of brushing was important. However, we omitted this variable from that part of the model to enhance the comparability between this model and the fitted

models corrected for misclassification (see Section 7.7.2.2).

Table 7.2: *Posterior estimates of the uncorrected ZIBB regression model fitted to the* $dmft_{4,5}$ *(using* `WinBUGS Program 7.2`*).*

Parameter	Estimate (SE)	95% CI[a]	
		2.5 %	97.5%
without $(x, y)-$*co-ordinate terms*			
Intercept	−1.414(0.116)	−1.647	−1.184
Gender (girl)	0.059(0.063)	−0.062	0.183
Age (years)	0.306(0.075)	0.162	0.455
Brushing frequency (< 2)	0.087(0.087)	−0.080	0.253
Age start brushing (years)	0.150(0.028)	0.095	0.205
Systemic fluoride (yes)	−0.429(0.064)	−0.551	−0.303
Sugary drinks (yes)	0.382(0.067)	0.256	0.517
Between meals (> 2)	0.160(0.065)	0.032	0.284
p	0.177(0.035)	0.099	0.242
τ	0.582(0.040)	0.506	0.667
with $(x, y)-$*co-ordinate terms*			
Intercept	−1.432(0.111)	−1.652	−1.220
x-ordinate	0.194(0.031)	0.136	0.255
y-ordinate	−0.042(0.032)	−0.105	0.021
Gender (girl)	0.052(0.060)	−0.066	0.169
Age (years)	0.335(0.076)	0.186	0.488
Brushing frequency (< 2)	0.106(0.085)	−0.064	0.273
Age start brushing (years)	0.160(0.029)	0.105	0.217
Systemic fluoride (yes)	−0.433(0.063)	−0.556	−0.308
Sugary drinks (yes)	0.376(0.064)	0.247	0.502
Between meals (> 2)	0.116(0.065)	−0.013	0.240
p	0.173(0.037)	0.089	0.230
τ	0.574(0.038)	0.504	0.652

[a]CI = Credible interval.

7.4 Correcting for misclassification in a ZIBB regression model

Correction for misclassification in a ZIBB regression model follows the same arguments as in Section 6.4.1. Indeed, we only need to replace $\Pr(Y = s|\cdot)$ in expression (6.11) by the ZIBB distribution. When multiple examiners are involved, the misclassification matrix changes from examiner to examiner.

7.5 Misclassification models of a finite count

Let $Y = \sum_{k=1}^{K} Z_k$ be the "true" count as determined by the benchmark scorer and $Y^* = \sum_{k=1}^{K} Z_k^*$ be the possibly corrupted observed count established by an examiner. Z_k and Z_k^* are the true and possibly corrupted binary scores, respectively which make up the respective counts. Further, let $\pi_{rs}(\boldsymbol{x}) = \Pr(Y^* = r|Y = s, \boldsymbol{x})$ $(r, s = 0, \ldots, K)$ with $\sum_{r=0}^{K} \pi_{rs}(\boldsymbol{x}) = 1$ represent the misclassification probabilities constituting the vector $\boldsymbol{\pi}_s(\boldsymbol{x}) = (\pi_{0s}(\boldsymbol{x}), \pi_{1s}(\boldsymbol{x}), \cdots, \pi_{Ks}(\boldsymbol{x}))'$ and the misclassification matrix $(\boldsymbol{\pi}_0(\boldsymbol{x}), \ldots, \boldsymbol{\pi}_K(\boldsymbol{x}))$. Suppose a $(K + 1) \times (K + 1)$ misclassification table is obtained from validation data with entries m_{rs} with $\sum_{r,s=0}^{K} m_{rs} = m$ whereby m_{rs} represents the number of subjects classified as $Y = s$ by the benchmark scorer and $Y^* = r$ by the examiner and n is the total number of subjects involved in the validation study.

For (finite) count data the misclassification matrix is often sparse. Hence if estimated from a multinomial model, the estimated misclassification probabilities are quite unreliable yielding corrected estimated parameters with high variability. Albert et al. (1997) suggested a fairly general approach for modeling misclassification probabilities of an ordinal variable. This approach could also be used for finite count data. Further, we suggest here a third approach, called the *double binomial approach* which exploits the nature of a count. Namely, in order to obtain a count the sum needs to be made of binary scores, each of which is prone to misclassification.

7.5.1 The multinomial approach

Let us assume independence of the subjects. Then the sth column, $\boldsymbol{m}_s$, of the misclassification table with entries m_{rs} follows a multinomial distribution:

$$\boldsymbol{m}_s \sim \text{Multinomial}\big(m_{+s}, \boldsymbol{\pi}_s(\boldsymbol{x})\big). \tag{7.4}$$

For a non-differential misclassification process, the multinomial estimate of π_{rs} is $\hat{\pi}_{rs} = m_{rs}/\sum_{r=0}^{K} m_{rs}$ and has variance $\pi_{rs}(1-\pi_{rs})/\sum_{r=0}^{K} m_{rs}$. However, the variance can be high and the estimate does not exist when the benchmark examiner does not score 's' in the validation data. When the misclassification process is differential the dependence on the covariates needs to be modeled.

7.5.2 The double binomial approach

In order to obtain a count, one needs to score the binary indicators Z_k $(k = 1, \ldots, K)$. Hence, it is likely that the validation data provide a misclassification table for each Z_k. Suppose for a non-differential misclassification process that $\alpha_k = \Pr(Z_k^* = 1 | Z_k = 1)$, $\beta_k = \Pr(Z_k^* = 0 | Z_k = 0)$ $(k = 1, \ldots, K)$ represent the sensitivity and specificity for Z_k, respectively of the examiner relative to a benchmark examiner. α_k and β_k can be estimated from the corresponding 2×2 misclassification table established in the validation study with entries $m_{k,rs}$ with $\sum_{r,s=0}^{1} m_{k,rs} = m$ as follows: $\hat{\alpha}_k = \frac{m_{k,11}}{m_{k,01}+m_{k,11}}$ and $\hat{\beta}_k = \frac{m_{k,00}}{m_{k,00}+m_{k,10}}$ $(k = 1, \ldots, K)$.

The above assumptions imply a binomial model for the sensitivity and for the specificity on the binary score. Further, it is assumed in first instance that the misclassification process is non-differential. Therefore, the basic double binomial approach (see below) will be based on the following three simplifying assumptions:

Assumption A1: scoring Z_k is done independently from scoring Z_l with $k \neq l$

Assumption A2: the scoring behavior of the examiner does not depend on k

Assumption A3: the scoring behavior of the examiner does not depend on the subject (non-differential misclassification process)

For the dental example, the assumptions $A1$ and $A2$ imply that the scoring of teeth is done equally well or bad for all teeth and that in an independent manner. Hence, when $A1$ to $A3$ are satisfied, $\alpha_k = \alpha_Z$ and $\beta_k = \beta_Z$ and are estimated by $\hat{\alpha}_Z = \frac{\sum_{k=1}^{K} m_{k,11}}{\sum_{k=1}^{K}[m_{k,01}+m_{k,11}]}$ and $\hat{\beta}_Z = \frac{\sum_{k=1}^{K} m_{k,00}}{\sum_{k=1}^{K}[m_{k,00}+m_{k,10}]}$, respectively.

Under the above simplifying assumptions, one can determine the misclassification table for Y based on the misclassification table for Z_k ($k = 1, \ldots, K$), which is assumed to be equal for all k. Namely

$$\pi_{rs} = \sum_{n=N_0}^{N_1} \binom{s}{n} \binom{K-s}{r-n} \alpha_Z^n (1-\alpha_Z)^{(s-n)} (1-\beta_Z)^{(r-n)} \beta_Z^{(K-s-r+n)}, \quad (7.5)$$

where the bounds $N_0 = \max(r-(K-s), 0)$ and $N_1 = \min(r, s)$ arise from the fact that expression (7.5) is derived from the distribution of two independent binomial distributions, $\text{Bin}(s, \alpha_Z)$ and $\text{Bin}(K-s, 1-\beta_Z)$. In the dental example, the first binomial distribution expresses the probability that the examiner scores n teeth as decayed from the s teeth that the benchmark examiner has scored decayed. The second binomial distribution expresses the probability that the examiner scores $(r-n)$ teeth as decayed from the $(K-s)$ teeth that the benchmark examiner has scored not decayed. Plugging the estimates $\hat{\alpha}_Z$, $\hat{\beta}_Z$ in expression (7.5) yields estimates $\tilde{\pi}_{rs}$ $(r, s = 0, \ldots, K)$ and hence the vectors $\tilde{\boldsymbol{\pi}}_s = (\tilde{\pi}_{0s}, \ldots, \tilde{\pi}_{Ks})'$ $(s = 0, \ldots, K)$.

7.5.3 Extensions of the DB approach

Assumptions $A1$ to $A3$ might not hold in practice. But, since the DB approach is based on two binomial distributions in principle all types of

extensions of binomial models could be used. We will discuss below some natural extensions of the DB model. The DB model can be described as follows:

$$\Pr(Z_1^*, \ldots, Z_K^* | Z_1, \ldots, Z_K) = \prod_{k=1}^{K} \Pr(Z_k^* | Z_k), \tag{7.6}$$

with $\Pr(Z_k^* = 1 | Z_k = 1) = \alpha_Z$ and $\Pr(Z_k^* = 0 | Z_k = 0) = \beta_Z$. Thus, extensions of the DB approach can be formulated as extensions of (7.6).

Extension *E*1: the sensitivity and specificity of the binary scores depend on covariates (Begg, 1987). For instance, when diagnosing oral cancer the sensitivity of detecting the disease might be higher for smokers than for non-smokers because the physician is more alerted for a smoker. Thus, we assume that (7.6) holds, but that $\alpha_Z \equiv \alpha_Z(\boldsymbol{x})$, $\beta_Z \equiv \beta_Z(\boldsymbol{x})$. In other words, we assume that $A1$ and $A2$ are satisfied, but not $A3$. In this case, the misclassification process is called differential.

Extension *E*2: model (7.6) is extended to

$$\Pr(Z_1^*, \ldots, Z_K^* | Z_1, \ldots, Z_K) = \prod_{k=1}^{K} \Pr(Z_k^* | Z_k, f(Z_1, \ldots, Z_K)).$$

Thus, the misclassification process depends on a global summary statistic of the binary scores, i.e. $\alpha_Z \equiv \alpha_Z(f(Z_1, \ldots, Z_K))$, $\beta_Z \equiv \beta_Z(f(Z_1, \ldots, Z_K))$. The motivation behind this extension is best seen in the caries example. When $f(Z_1, \ldots, Z_K) = \sum_{k=1}^{K} Z_k$ is large there is much caries in the mouth. It is conceivable that in such a mouth there might be some confusion of when a tooth is decayed or not.

Extension *E*3: the scoring is dependent, i.e. $\Pr(Z_1^*, \ldots, Z_K^* | Z_1, \ldots, Z_K)$ does not split up in a product. Thus, assumption $A1$ is relaxed, for a motivation in oral health studies see e. g. Hujoel, Moulton,

and Loesche (1990). There exists a variety of models for correlated binary random variables, see e.g. Rudolfer (1990). A convenient way to introduce correlation is to assume that, given a subject, the scoring is independent but that the sensitivity and specificity depend on the subject (unknown) characteristics. More formally, assume that α and β have a distribution depending on the subject's unknown characteristics, given by the random vector $\boldsymbol{b}$ and that

$$\Pr(Z_1^*, \ldots, Z_K^* | Z_1, \ldots, Z_K, \boldsymbol{b}) = \prod_{k=1}^{K} \Pr(Z_k^* | Z_k, \boldsymbol{b}),$$

and $\Pr(Z_1^*, \ldots, Z_K^* | Z_1, \ldots, Z_K) = \int \prod_{k=1}^{K} \Pr(Z_k^* | Z_k, \boldsymbol{b}) f(\boldsymbol{b}) d\boldsymbol{b}$. There are two natural candidates for the distribution of α and β. Firstly, assume that α and β each have a Beta density and that they are independent of each other. In that case $\Pr(Z_1^*, \ldots, Z_K^* | Z_1, \ldots, Z_K) = \int \prod_{Z_k=1} \Pr(Z_k^* | Z_k, \alpha) B(\alpha) d\alpha \int \prod_{Z_k=0} \Pr(Z_k^* | Z_k, \beta) B(\beta) d\beta$, where $B(\cdot)$ represents a Beta-density. Thus, $\alpha \equiv b_1$ and $\beta \equiv b_2$, where $\boldsymbol{b}' = (b_1, b_2)$. This generalizes the binomial distribution $\mathrm{Bin}(K, \pi)$ to a beta-binomial distribution $\mathrm{BB}(K, \pi, \tau)$, with mean π and variance τ (see Section 7.3 for an expression). Further, this implies that in expression (7.5), $\mathrm{Bin}(s, \alpha_Z)$ and $\mathrm{Bin}(K - s, 1 - \beta_Z)$ are replaced by $\mathrm{BB}(s, \alpha_Z, \tau_\alpha)$ and $\mathrm{BB}(K - s, \beta_Z, \tau_\beta)$, respectively. Secondly, since α and β have a distribution depending on the subject, it is natural to assume that they are also correlated. Correlation can be introduced by first taking the logit transform of α and β, i.e. $b_1 = \mathrm{logit}(\alpha)$, $b_2 = \mathrm{logit}(\beta)$ and then assuming that $\boldsymbol{b}' = (b_1, b_2) \sim \mathcal{N}(\boldsymbol{\mu_b}, \Sigma_{\boldsymbol{b}})$.

Extension *E*4: sensitivity and specificity depend on k, i.e. α_k, β_k ($k = 1, \ldots, K$). For instance, in caries research it is known that detecting caries experience in molars is more difficult than in other teeth.

In the sequel the basic DB model will also be denoted by $E0$. To test whether the E0 model needs to be extended to Ex ($x = 1, \ldots, 4$) a

likelihood ratio test can be employed. To choose between the extensions, Akaike's Information Criterion might be used. Further, the above extensions could be combined making the misclassification process even more general. The availability of a battery of models for the misclassification process helps in obtaining unbiased estimates of the misclassification probabilities while maintaining high efficiency. For all of the extensions, the estimated misclassification probabilities will still be denoted by $\tilde{\boldsymbol{\pi}}_s$.

7.6 Simulation study

A simulation study was set up to evaluate the performance of the DB approach and its extensions. More specifically, we have set up two types of simulations: (1) Evaluating the efficiency of the $E0$ model (basic DB approach) when the true model is in fact Ex ($x = 0, \ldots, 4$) in comparison to the true extension and in comparison to the multinomial and Albert et al.'s model. This evaluates the basic DB approach in estimating the correction terms; (2) Evaluating the efficiency of the DB approach as above, but when estimating the main model parameters. This efficiency comparison was set up to evaluate the practical gain in estimating the main model parameters when using the most efficient procedure for estimating the correction terms.

7.6.1 Setup of the simulation study

7.6.1.1 First simulation study

In the first simulation study, the sample size of the validation study was fixed to $N = 100$. Each scenario was sampled 1000 times. We formulated below the four extensions of the basic DB approach as logistic models.

Namely:

$$\begin{aligned}
\text{logit}[\Pr(Z^*_{k,i}=1)] &= \gamma_0+\gamma_1 Z_{k,i} && (7.7)\\
\text{logit}[\Pr(Z^*_{k,i}=1)] &= \gamma_0+\gamma_1 Z_{k,i}+\xi_1 X_{1,i}+\ldots+\xi_q X_{q,i} && (7.8)\\
\text{logit}[\Pr(Z^*_{k,i}=1)] &= \gamma_0+\gamma_1 Z_{k,i}+\delta f(Z_{1i},\ldots,Z_{Ki}) && (7.9)\\
\text{logit}[\Pr(Z^*_{k,i}=1)] &= \gamma_0+\gamma_1 Z_{k,i}+\gamma_2 T_{2,ki}+\ldots+\gamma_K T_{K,ki} && (7.10)
\end{aligned}$$

where $i=1,\ldots,N$ and $k=1,\ldots,K$ and the dependence on the covariates is omitted from the expression for convenience. The true and possibly misclassified counts are obtained by making the sums $\sum_{k=1}^{K} Z_{k,i}$ and $\sum_{k=1}^{K} Z^*_{k,i}$, where $K=8$ has been taken.

Two values for the prevalence p_M at mouth level were considered, i.e. $p_M=0.10$ and $p_M=0.30$. This implies a prevalence at tooth level equal to $p_T=1-(1-p_M)^{(1/K)}$ yielding $p_T=0.013$ and $p_T=0.043$, respectively.

Model $E0$, i.e. the basic DB approach, corresponds to model (7.7) with $\gamma_0=-2.94$ and $\gamma_1=5.14$, so we obtain $\alpha_Z=0.90$ and $\beta_Z=0.95$.

In extension $E1$, see equation (7.8), we have taken $q=2$. The regression vector ξ is taken equal to $(0.7,-0.3,0.4)'$ for a minimal variation of the sensitivities and specificities and equal to $(1.5,-0.6,0.9)'$ for a moderate variation. Further, for the two covariates we assume that $X_1\sim N(3,0.2)$ and $X_2\sim Bernoulli(0.6)$. These two distributions were inspired by the covariates considered in the Signal Tandmobiel® study (namely, "age at start brushing" and "gender"). α_Z and β_Z vary over the subjects and hence assumption $A3$ is violated here.

Extension $E2$, see equation (7.9), expresses that α_Z and β_Z depend on the values of $Z_1,\ldots,Z_K$ and thus assumption $A1$ is violated now. More specifically, we have taken $f(Z_1,\ldots,Z_K)=\sum_{k=1}^{K} Z_k$. For a minimal variation of the sensitivities and specificities we have taken $\delta=0.1$ and for a moderate variation $\delta=0.2$.

In extension $E3$, assumption $A1$ is relaxed by allowing the α_Z and β_Z to have a distribution varying by subject. For a minimal variation of

the sensitivities and specificities we take $\boldsymbol{\mu}'_{\boldsymbol{b}} = (\text{logit}(\alpha_Z), \text{logit}(\beta_Z))$, e.g. $\boldsymbol{\mu}'_{\boldsymbol{b}} = (2.197, 2.944)$ for $\alpha_Z = 0.90$ and $\beta_Z = 0.95$, and

$$\Sigma_{\boldsymbol{b}} = \begin{pmatrix} 0.050 & 0.035 \\ 0.035 & 0.030 \end{pmatrix}.$$

For a moderate variation the values of $\Sigma_{\boldsymbol{b}}$ are doubled.

Finally, in extension $E4$, see equation (7.10), the variables $T_{j,k} = 1$ if $j = k$ and 0 otherwise for $j, k = 1, \ldots, K$. They express the fact that the α_k and β_k differ over k and hence that assumption $A2$ is violated. For a small variation, we have taken for $K = 6$, $\gamma = (-0.3, -0.3, 0.3, 0.3, 0.3)'$ and for $K = 8$ we have taken $\gamma = (-0.3, -0.3, -0.3, 0.3, 0.3, 0.3, 0.3)'$. For a moderate variation, these values are doubled.

We have assumed that the true binary scores Z_k are independent ($\rho = 0$) and as well as that they are related ($\rho = 0.7$). However, since the simulation results for the two values of ρ are similar, we report only the results for $\rho = 0.7$. To generate the correlated binary scores we used the method of Dunn and Davies (1998).

For the models Ex ($x = 0, \ldots, 4$), the multinomial model and Albert et al.'s model, we calculated α_W and β_W: the average, the median, the SD, the Mean Squared Error (MSE) and the 95% confidence range. We calculated also the estimates for all elements of the misclassification matrix, i.e. π_{rs}, but reporting all of these results would be overwhelming. Alternatively, we calculated the discrepancy measure $D = \sum_{s=0}^{K} \sum_{r=0}^{K} (\pi_{rs} - p_{rs})^2/(K+1)$, where π_{rs} is the true misclassification probability and $p_{rs} = \hat{\pi}_{rs}, \tilde{\pi}_{rs}$, respectively. A chi-square type of statistic is also possible but it would give too much weight to the small (and unimportant) true misclassification probabilities and is therefore not reported here.

To determine the true misclassification probabilities and consequently the value of α_W and β_W an approximative method was used. Namely, we approximate the true misclassification probabilities using the multinomial method determined on a validation study of size $200,000$.

We report below only the case of moderate variability together with $K = 8$ since these results are sufficient to deliver the message.

7.6.1.2 Second simulation study

In the second simulation study, we examined the effect of the basic DB approach and its extensions on the estimation of the main model parameters and compared their performance to the multinomial approach. For all cases, the size of the main study is 1000 and 1000 simulations were performed for each scenario. We considered as main model a binomial regression model given by $Y \sim \text{Binom}(K, p_Y)$ with $\text{logit}(p_Y) = \beta_0 + \beta_X X + \beta_Z Z$, where X is a binary covariate with success probability p_X and Z is an independent continuous normal variate with mean 0 and standard deviation SD_Z. We have varied the values of p_X and SD_Z. This was done to examine the effect of the precision with which the regression coefficient is estimated in the model without scoring errors on the relative gain of the DB approach. For instance, when SD_Z is large it is known that the regression coefficient of Z is estimated with more precision than when SD_Z is small. Therefore, it is expected that the gain of the DB approach will be better seen for a relatively large value of SD_Z.

The size of the validation study was fixed at 100. Further, we sampled the validation data such that :

(a) there is *equal* probability for scoring $Y = s$, i.e. $\Pr(Y = s) = 1/(K + 1)$;

(b) there is *unequal* probability for $Y = s$, i.e. $\Pr(Y = s) = [2(K + 1 - s)]/[(K + 1)(K + 2)]$ which is decreasing in s;

(c) the validation study is a random sample of the main study.

Sampling and estimation was done under the different DB approaches. More specifically, we sampled from extension Ex and estimated the parameters with $E0$ and Ex. Thus, when sampling was done under extension Ex, estimation was done under the same model.

7.6.2 Simulation results

7.6.2.1 First simulation study

In Table 7.3 the simulation results are shown for the sensitivity (α_W) and in Table 7.4 the simulation results for the specificity (β_W) are given. More specifically, we show the estimated sensitivity and specificity for the two values of the prevalence when estimated with the basic DB , the multinomial and Albert et al.'s approach. when sampling is done under the models Ex ($x = 0, \ldots, 4$). We observe that in all cases α_W and β_W are estimated unbiasedly for the first two approaches. For a low prevalence the specificity is estimated with less variability than the sensitivity, while the reverse is true for the higher prevalence. In all cases the variability in estimation is lower for the DB approach than with the multinomial and Albert et al.'s approach.

Table 7.3: *Simulation results for sensitivity: K= 8, N = 100 with moderately varying sensitivity α_k and specificity β_k around 90%, 95% respectively for each of the four extensions (E0 corresponds to the basic DB approach). Prev represents the prevalence at mouth level. All values are expressed in percentages.*

			Double binomial (E0)		Multinomial			Albert et al.		
Ext	Prev	SN†	Mean(Median)/ 95% Range	SD/ MSE	Mean(Median)/ 95% Range	SD/ MSE	$RE^{\ddagger}_M$	Mean(Median)/ 95% Range	SD/ MSE	$RE^{\sharp}_A$
E0	10	93.3	93.3(94.1) [78.0, 100.0]	6.59 43.4	93.5(100.0) [72.7, 100.0]	8.36 69.8	161	93.2(93.9) [75.0, 100.0]	7.99 63.9	147
E1	10	94.1	94.5(95.1) [80.5, 100.0]	5.95 35.5	94.2(100.0) [75.0, 100.0]	7.75 60.1	169	94.1(100.0) [75.0, 100.0]	7.70 59.2	167
E2	10	92.5	91.8(92.8) [74.3, 100.0]	7.54 57.4	91.9(92.3) [70.0, 100.0]	9.11 83.2	145	92.0(92.4) [70.0, 100.0]	8.83 78.1	136
E3	10	93.3	93.3(94.0) [77.9, 100.0]	6.45 41.6	93.4(95.0) [75.0, 100.0]	8.03 64.5	155	92.9(94.7) [70.0, 100.0]	9.18 84.1	202
E4	10	92.8	92.5(93.3) [75.0, 100.0]	7.45 55.6	92.8(93.8) [70.0, 100.0]	8.83 77.9	140	92.1(92.3) [73.3, 100.0]	8.32 69.8	126
E0	30	94.0	94.0(94.4) [86.7, 100.0]	3.30 10.9	94.0(94.3) [83.3, 100.0]	4.42 19.6	180	92.8(93.3) [82.8, 100.0]	4.45 21.8	200
E1	30	94.8	95.0(95.3) [88.2, 100.0]	2.86 8.2	94.8(95.8) [85.2, 100.0]	4.12 16.9	206	93.8(94.2) [84.5, 100.0]	4.11 17.7	216
E2	30	93.1	93.0(93.4) [85.1, 98.5]	3.70 13.7	93.1(93.3) [82.9, 100.0]	4.73 22.3	163	92.9(94.7) [70.0, 100.0]	4.78 22.9	167
E3	30	94.1	93.8(94.0) [86.9, 98.9]	3.12 9.8	94.0(94.2) [84.0, 100.0]	4.32 18.7	191	92.8(93.3) [83.9, 100.0]	4.31 20.2	206
E4	30	93.4	92.6(92.8) [85.2, 98.6]	3.58 13.4	93.5(93.9) [83.3, 100.0]	4.53 20.5	153	91.3(91.7) [80.0, 100.0]	4.74 27.0	201

†SN = true sensitivity at mouth level.
‡$RE_M = MSE_{Mult}/MSE_{DB} \times 100$
♯$RE_A = MSE_{Albert}/MSE_{DB} \times 100$

Table 7.4: *Simulation results for specificity: K= 8, N = 100 with moderately varying sensitivity α_k and specificity β_k around 90%, 95% respectively for each of the four extensions (E0 corresponds to the basic DB approach). Prev represents the prevalence at mouth level. All values are expressed in percentages.*

			Double binomial (E0)		Multinomial			Albert et al.		
Ext	Prev	SP†	Mean(Median)/ 95% Range	SD/ MSE	Mean(Median)/ 95% Range	SD/ MSE	$RE^{\ddagger}_{M}$	Mean(Median)/ 95% Range	SD/ MSE	$RE^{\sharp}_{A}$
E0	10	66.1	66.6(66.5)	4.33	66.4(66.3)	4.95		66.6(66.7)	4.58	
			[58.6, 75.7]	18.9	[57.0, 76.3]	24.6	130	[57.6, 75.8]	21.0	111
E1	10	61.2	61.1(61.2)	4.33	61.3(61.4)	5.12		61.3(61.6)	4.79	
			[52.7, 69.6]	18.8	[51.6, 71.4]	26.2	139	[52.0, 70.7]	23.0	122
E2	10	71.6	71.9(71.8)	4.40	72.0(72.2)	4.88		72.0(72.0)	4.39	
			[63.2, 80.6]	19.4	[62.5, 81.0]	24.0	123	[63.3, 80.8]	19.4	100
E3	10	65.7	65.3(65.3)	4.47	65.5(65.6)	4.92		66.1(66.0)	4.82	
			[57.0, 73.4]	20.1	[56.2, 74.7]	24.2	120	[56.6, 75.0]	23.6	117
E4	10	71.6	71.8(71.8)	4.27	71.7(71.8)	4.81		71.7(72.0)	4.78	
			[63.7, 80.6]	18.3	[61.9, 81.1]	23.1	126	[61.8, 80.6]	23.0	126
E0	30	66.5	66.2(66.4)	4.34	66(66.2)	5.82		67.8(68.0)	4.73	
			[57.6, 75.0]	18.9	[54.3, 77.0]	34.1	180	[58.2, 77.0]	24.1	128
E1	30	61.4	60.9(60.9)	4.49	61.4(61.4)	5.82		62.9(63.0)	4.94	
			[52.2, 70.1]	20.3	[50.0, 72.5]	33.9	167	[52.6, 72.1]	28.3	139
E2	30	71.9	71.1(71.1)	4.48	71.7(72.0)	5.54		66.1(66.0)	4.82	
			[62.3, 80.0]	20.6	[60.0, 82.6]	30.7	149	[56.6, 75.0]	23.6	115
E3	30	65.6	65.3(65.0)	4.43	65.5(65.3)	5.58		67.0(67.0)	4.78	
			[57.5, 74.5]	19.7	[55.2, 77.1]	31.2	158	[57.3, 76.0]	24.6	125
E4	30	71.5	71.9(71.8)	4.42	71.6(71.7)	5.57		72.9(72.7)	4.52	
			[63.9, 80.9]	19.7	[60.9, 82.4]	31.0	157	[63.4, 82.0]	22.4	114

†SP = true specificity at mouth level.
‡$RE_M = MSE_{Mult}/MSE_{DB} \times 100$
♯$RE_A = MSE_{Albert}/MSE_{DB} \times 100$

The discrepancy measure D of the misclassification probabilities is given in Table 7.5. More specifically, the measure D is calculated for each simulated scenario and descriptive statistics over the 1000 simulations are reported. We observe that again the basic DB approach is the winner under most scenarios, with the most important gain for the higher value of the prevalence.

7.6.2.2 Second simulation study

Tables B.1 and B.2 (Appendix B) show the simulation results for a binomial regression with equal and unequal probability of scoring $Y = 0, 1, \cdots, K$ in the validation data.

The simulation results indicate that when SD_Z is relatively low, i.e. when the precision of estimating the true regression coefficient in the data

Table 7.5: *Simulation results for misclassification probabilities: estimate (over 1000 simulation samples) of the discrepancy measure* $D = \sum\sum(\pi_{rs} - p_{rs})^2/(K+1)$*, where* π_{rs} *is the true and* p_{rs} *the estimated misclassification probability over of the four extensions (K = 8, N = 100; E0 corresponds to the basic DB approach).*

		Double binomial (E0)				Multinomial				Albert et al.			
			Quantile				Quantile				Quantile		
Ext	Prev	Mean	25%	50%	75%	Mean	25%	50%	75%	Mean	25%	50%	75%
E0	10	9.3	7.8	8.4	9.8	17.2	15.0	16.3	17.9	10.4	7.1	12.2	12.3
E1	10	12.9	12.2	12.5	13.2	23.9	21.2	22.6	25.3	9.1	5.9	10.7	10.8
E2	10	17.5	14.9	16.4	18.9	25.0	22.5	23.9	26.0	9.1	6.3	10.7	10.9
E3	10	9.4	8.3	8.7	9.8	18.6	16.0	17.1	19.8	8.4	5.3	10.1	10.2
E4	10	9.3	7.6	8.5	10.1	15.1	12.2	14.0	16.4	8.6	6.4	9.9	10.0
E0	30	1.1	0.6	0.8	1.2	17.1	14.4	16.2	18.8	8.8	6.1	10.1	10.2
E1	30	1.4	1.1	1.2	1.5	18.3	15.3	17.2	20.5	8.1	5.6	9.2	9.4
E2	30	1.7	0.9	1.4	2.2	20.3	16.6	19.0	23.2	9.7	7.0	10.8	11.0
E3	30	1.3	0.8	1.1	1.5	18.9	16.0	17.8	20.8	10.1	7.3	11.3	11.5
E4	30	0.7	0.3	0.5	0.8	16.4	13.4	15.6	19.1	8.7	6.8	9.5	9.7

set without scoring errors is relatively low, then the DB approaches are roughly equivalent (although practically always better than) to the multinomial approach. In contrast, when the precision of estimating the true regression is high, there is much gain in using the DB approaches as compared to the multinomial method. Further, the multinomial method shows a serious bias in estimating the regression coefficients when SD_Z is high.

We also observe that the variability with which the parameters are estimated with the DB approach does not depend on the marginal probability distribution of Y. The same seems to be true for the multinomial approach. However, the latter approach clearly suffers from computational difficulties when the marginal probability of Y is not uniform. The results are sometimes dramatic when the validation study is a random sample of the main study, see Table B.3.

Finally, we observe that when the correct extension is used the performance of the DB approach is best. That is, when sampling is done under Ex and estimation is done under Ex, then the MSE is the lowest. However, the performance of the basic DB approach is relatively close to the extension, certainly in view of its difference with the multinomial approach.

7.7 Application to the Signal Tandmobiel® study

7.7.1 Analysis of the validation data

In previous chapters the validation data were combined from the three caries calibration exercises. Here we look at the validation data of the first calibration based on ninety-two children. First we present results from the analysis of the pooled validation data. Then we give results from the analysis of the examiner-specific validation data.

7.7.1.1 Analysis of the pooled validation data

Table 7.6: *Overall misclassification table for $dmft_{4,5}$, column= benchmark examiner, row= (pool of) dental examiner(s)*

	Y									
Y^*	0	1	2	3	4	5	6	7	8	
0	32	1	3	0	0	0	0	0	0	36
1	2	13	2	1	0	0	0	0	0	18
2	0	1	5	2	3	0	1	0	0	12
3	0	0	2	4	1	1	1	0	0	9
4	0	0	0	0	2	1	2	0	0	5
5	0	0	0	0	1	3	1	0	0	5
6	0	0	0	0	0	1	2	2	0	5
7	0	0	0	0	0	0	0	1	1	2
8	0	0	0	0	0	0	0	0	0	0
	34	15	12	7	7	6	7	3	1	92

In Table 7.6 the observed misclassification table for $dmft_{4,5}$ with respect to the benchmark examiner is given. Clearly, this is a very sparse table. In Table 7.7 scoring caries experience by the dental examiners on tooth level is compared with the scores of the benchmark examiner. From this table we can obtain $\hat{\alpha}_Z = 133/154 = 0.86$ and $\hat{\beta}_Z = 418/433 = 0.97$. These values are plugged in (7.5) together with $K = 8$ and yields the estimated misclassification probabilities for the basic DB method. Combined with the marginal totals from Table 7.6 the estimated frequencies of misclassification

are obtained. The fitted table, shown in Table 7.8, indicates that the expected frequency of cell $(0, 0)$ under the basic DB method is 26 which is about 25% lower than the observed frequency.

Table 7.7: *Misclassification table at tooth level, column= benchmark examiner, row= (pool of) dental examiner(s)*

	Z		
Z^*	0	1	
0	418	21	439
1	15	133	148
	433	154	587

Table 7.8: *Expected misclassifications (probability of misclassification $\times 100$) of $dmft_{4,5}$ based on the observed sensitivity ($\hat{\alpha}_Z = 133/154 = 0.86$) & specificity ($\hat{\beta}_Z = 418/433 = 0.97$). Column = benchmark examiner, row = dental examiner(s)*

	Y									
Y^*	0	1	2	3	4	5	6	7	8	
0	26(75)	2(11)	0(2)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	28
1	7(22)	10(70)	2(19)	0(4)	0(1)	0(0)	0(0)	0(0)	0(0)	19
2	1(3)	3(17)	8(64)	0(26)	1(7)	0(2)	0(0)	0(0)	0(0)	15
3	0(0)	0(2)	2(13)	1(59)	2(32)	1(11)	0(3)	0(1)	0(0)	8
4	0(0)	0(0)	0(1)	2(10)	4(53)	2(35)	1(15)	0(5)	0(1)	8
5	0(0)	0(0)	0(0)	3(1)	0(7)	3(47)	3(38)	1(18)	0(7)	7
6	0(0)	0(0)	0(0)	1(0)	0(0)	1(5)	3(41)	1(39)	0(22)	5
7	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(3)	1(36)	1(39)	2
8	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(1)	0(31)	0
	34	15	12	6	7	7	7	3	1	92

In Figure 7.1 the observed tooth-specific specificities and sensitivities are plotted as a function of the $dmft_{4,5}$-index. From this figure there is some evidence that the specificity and the sensitivity depend on the actual value of the $dmft_{4,5}$-index, namely they are higher for $dmft_{4,5} = 0$ and 1. A possible explanation for the dependence of sensitivity and specificity on $dmft_{4,5}$, is that when there is (almost) no caries experience in the mouth caries might be easier to distinguish from no-caries, while in a

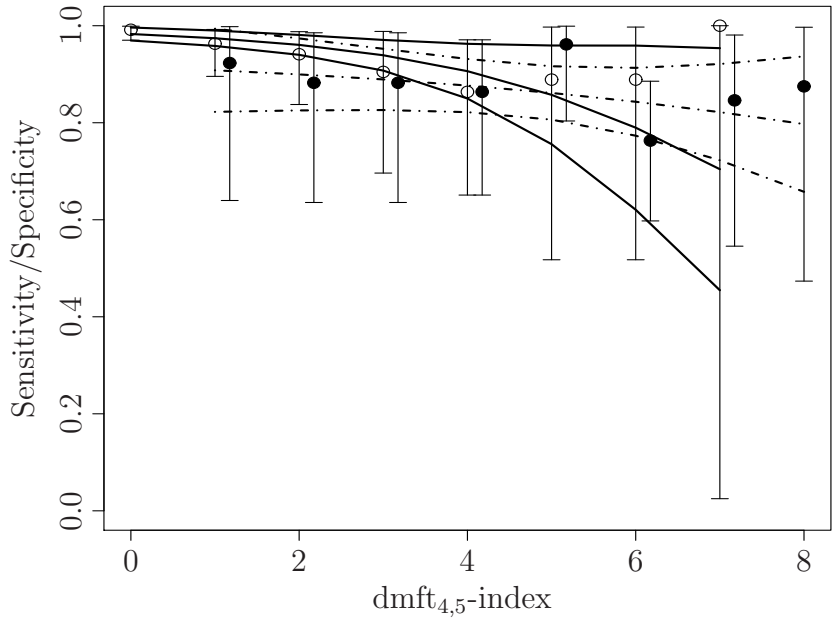


Figure 7.1: *Tooth-specific specificity and sensitivity (+ 95% CI) as a function of* $dmft_{4,5}$. *Open (filled) circles represent the observed specificities (sensitivities). The lines correspond to fitted values from model (7.12), the solid lines represent specificity and the dashed lines sensitivity. The inner lines represent the average curve, the outer lines show the 95% pointwise boundary values*

mouth with considerable caries experience the dental examiner might be distracted somewhat easier. We fitted a logistic regression model predicting the scoring behavior of the examiners as a function of the true score and $\text{dmft}_{4,5}$-index:

$$\begin{aligned}\text{logit}(\Pr(Z_k^* = 1|Z_k, \text{dmft}_{4,5})) &= \\ -4.2 + 6.3Z_k + 0.43\text{dmft}_{4,5} - 0.45Z_k \times \text{dmft}_{4,5}, &\qquad (7.11)\end{aligned}$$

where we have omitted the subscript k in $\text{dmft}_{4,5}$ for convenience. This model actually represents two models:

$$\begin{aligned} \text{logit}(\hat{\alpha}_k) &= 2.6 - 0.14 \times \text{dmft}_{4,5}, \\ \text{logit}(\hat{\beta}_k) &= 4.1 - 0.45 \times \text{dmft}_{4,5}. \end{aligned} \tag{7.12}$$

Model (7.12) corresponds to extension E2. The regression coefficient (SE) indicates that the specificity decreases with $\text{dmft}_{4,5}$, but the negative dependence of the sensitivity on $\text{dmft}_{4,5}$ is not so pronounced. Hence, we used model (7.12) to provide the pooled correction terms for the main model.

Table 7.9: *Expected misclassifications (probability of misclassification* $\times 100$*) of* $dmft_{4,5}$ *with specificities & sensitivities estimated from model (7.12). Column = benchmark examiner, row = dental examiner(s)*

	Y									
Y^*	0	1	2	3	4	5	6	7	8	
0	30(88)	1(7)	0(1)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	31
1	4(11)	12(78)	2(14)	0(2)	0(1)	0(0)	0(0)	0(0)	0(0)	18
2	0(1)	2(14)	8(68)	1(20)	0(5)	0(1)	0(0)	0(0)	0(0)	11
3	0(0)	0(1)	2(16)	4(59)	2(24)	1(9)	0(3)	0(1)	0(1)	9
4	0(0)	0(0)	0(1)	1(17)	4(50)	2(28)	1(13)	0(7)	0(4)	8
5	0(0)	0(0)	0(0)	0(2)	1(17)	3(43)	2(32)	1(20)	0(14)	7
6	0(0)	0(0)	0(0)	0(0)	0(3)	1(16)	3(37)	1(35)	0(29)	5
7	0(0)	0(0)	0(0)	0(0)	0(0)	0(3)	1(13)	1(30)	1(34)	3
8	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(2)	0(7)	0(18)	0
	34	15	12	6	7	7	7	3	1	92

Table 7.9 shows the estimated misclassification probabilities $\Pr(Y_2^* = r|Y_2 = s) \times 100$ and the estimated frequencies using model (7.12). Our estimated frequencies are now closer to the observed values than for the basic DB method, especially for the small (true) values of $\text{dmft}_{4,5}$. The value of $D \times 100$ for the extension $E2$ and the multinomial approach are 15.4 and 20.3, respectively.

7.7.1.2 Analysis of the examiner-specific validation data

We fitted a logistic regression model predicting the scoring behavior of each examiner as a function of the true score and $\text{dmft}_{4,5}$:

$$\begin{aligned} \text{logit}(\hat{\alpha}_{jk}) &= \hat{a}_{0j} + \hat{a}_{1j} \times \text{dmft}_{4,5}, \\ \text{logit}(\hat{\beta}_{jk}) &= \hat{b}_{0j} + \hat{b}_{1j} \times \text{dmft}_{4,5}. \end{aligned} \tag{7.13}$$

where $j = 1, \cdots, 16$ indexes the examiner. The regression coefficients in the logistic model (7.13) are treated as random examiner effects and are assumed to follow independent normal distributions, i.e. $a_{tj} \sim N(\mu_{a_t}, \sigma^2_{a_t})$, $b_{tj} \sim N(\mu_{b_t}, \sigma^2_{b_t})$ for $t = 0, 1$ and $j = 1, \cdots, 16$. The hyperparameters are assigned vague prior distributions, i.e. $\mu_{a_t}, \mu_{b_t} \sim N(0, 10^6)$ and $\sigma^2_{a_t}, \sigma^2_{b_t} \sim IG(10^3, 10^3)$ for $t = 0, 1$ and $j = 1, \cdots, 16$.

Table 7.10: *Posterior summary statistics of the hyperparameters of the random coefficients from the examiner-specific logistic regression (7.13).*

	Estimate	95% CI[a]			Estimate	95% CI[a]	
	(SE)	2.5%	97.5%		(SE)	2.5%	97.5%
μ_{a_0}	4.210(0.337)	3.583	4.904	$\sigma^2_{a_0}$	0.058(0.128)	0.001	0.373
μ_{a_1}	−0.425(0.122)	−0.658	−0.178	$\sigma^2_{a_1}$	0.033(0.051)	0.001	0.163
μ_{b_0}	2.096(0.527)	1.100	3.085	$\sigma^2_{b_0}$	0.080(0.171)	0.001	0.527
μ_{b_1}	−0.012(0.110)	−0.215	0.205	$\sigma^2_{b_1}$	0.008(0.012)	0.001	0.038

[a] CI = Credible interval.

The posterior mean estimate of the random examiner-slopes $({a_{1j}}'s, {b_{1j}}'s)$ together with their 95% credible interval are displayed in Figures 7.2. It is evident that the specificity for majority of the examiners depend on the actual value of the $\text{dmft}_{4,5}$-index, but the dependence of the sensitivity on $\text{dmft}_{4,5}$ is not significant for all examiners. Table 7.10 below shows the posterior summary statistics of the hyperparameters of the random coefficients. The posterior mean estimate of $\boldsymbol{\mu} = (\mu_{a_0}, \mu_{a_1}, \mu_{b_0}, \mu_{b_1})'$ is similar to the estimated regression coefficients of the logistic model (7.12).

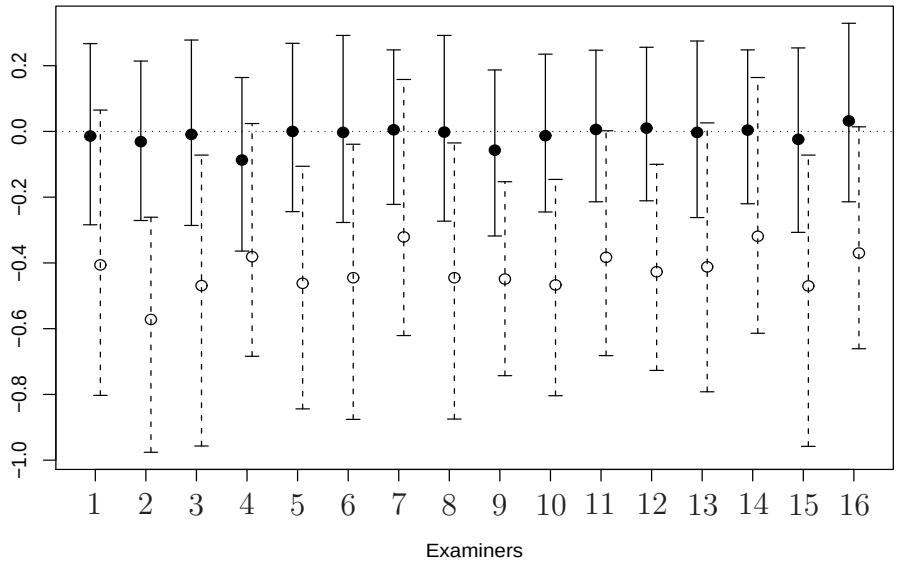


Figure 7.2: *Examiner-specific coefficients (+ 95% credible interval) from model (7.13). The solid and the dashed lines represent 95% credible interval of $\hat{a}_{1j}$'s and $\hat{b}_{1j}$'s, respectively. Open (filled) circles represent the posterior mean of $\hat{a}_{1j}$'s ($\hat{b}_{1j}$'s).*

7.7.2 Analysis of the main data

7.7.2.1 Fitting the corrected distribution of the $\text{dmft}_{4,5}$-index

The observed prevalence is 55%. To estimate the true prevalence of caries experience we need to assume a model for $\text{dmft}_{4,5}$ and correct for misclassification using either of the three methods discussed above. The ZIBB distribution gave a good fit to the observed $\text{dmft}_{4,5}$ so that our calculations are based on this distribution.

As can be seen in Figure 7.3 the three pooled correction mechanisms do not give very different results for the model without covariates. The corrected estimates of prevalence (and 95% credible interval) for the three correction methods are: (a) using the multinomial method: 60.1% (54.2 – 66.7), (b) using Albert *et al.*'s method: 57.8% (55.4 – 60.7) and (c) using extension $E2$: 59.4% (55.5 – 64.4). The estimated prevalence from both

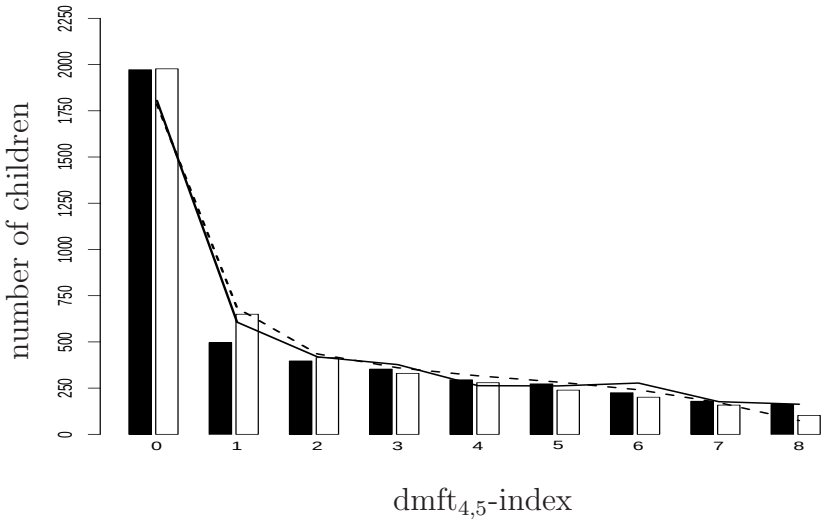


Figure 7.3: *Distribution of the $dmft_{4,5}$-index among 7-year old Flemish children: observed values (■); and fitted values from the pooled corrected ZIBB model combined with Albert* et al.*'s (□), the DB under extension E2 (continuous line) and multinomial (broken line) misclassification model.*

methods are quite close, but the estimate of $E2$ presents less variability than the multinomial estimate.

7.7.2.2 Regression analysis of the main data with correction

Table 7.11 shows the results from the pooled corrected ZIBB regression models based on the three approaches. First, observe that the fitted value for p is much lower for the multinomial and the DB (extension E2) method than for the method proposed by Albert *et al.*. This implied that no covariates turned out to be important to predict the extra-zeros in the first two approaches.

From Table 7.11 we observe that the effects of the covariates are roughly similar to that of the model without correction, except that the effect of taking in-between-meals vanishes for the multinomial correction, and, prac-

Table 7.11: *Posterior estimates of the pooled corrected ZIBB regression model fitted to the* $dmft_{4,5}$ *(using* `WinBUGS Program 7.3`*).*

Parameter	Double binomial Estimate SE	Multinomial Estimate SE	Albert *et al.* Estimate SE
Intercept	−1.592(0.143)	−1.779(0.333)	−1.274(0.194)
Gender (girl)	0.053(0.097)	0.056(0.223)	0.061(0.107)
Age (years)	0.337(0.081)	0.477(0.164)	0.348(0.088)
Brushing frequency (< 2)	0.111(0.118)	0.174(0.256)	0.102(0.124)
Age start brushing (years)	0.157(0.043)	0.237(0.099)	0.166(0.046)
Systemic fluoride (yes)	−0.465(0.086)	−0.683(0.230)	−0.477(0.091)
Sugary drinks (yes)	0.388(0.088)	0.519(0.229)	0.430(0.108)
Between meals (> 2)	0.171(0.080)	0.212(0.169)	0.180(0.087)
p	0.004(0.017)	0.128(0.161)	0.143(0.103)
τ	1.302(1.124)	1.143(3.810)	0.769(0.153)
$^{a}\beta_0$	−4.128(0.417)		
$^{a}\beta_Z$	6.655(0.791)		
$^{a}\beta_{dmft_{4,5}}$	0.453(0.130)		
$^{a}\beta_{Z*dmft_{4,5}}$	−0.589(0.180)		
$^{b}\alpha_0$		−0.088(0.438)	
$^{b}\alpha_1$		−2.382(0.496)	
$^{b}\alpha_2$		−0.656(0.226)	

[a] Coefficients of the logistic regression model (7.11) from the corrected model.
[b] Coefficients of the $3p-$asymmetric Albert *et al.* method from the corrected model.

tically all regression coefficients increased in absolute value. The increase (in absolute value) is the highest for the multinomial correction method. With respect to the variability of the estimated regression coefficients we observe that for all corrected regression coefficients, the 95% credible interval increased in size with respect to the model without correction. On the other hand, the 95% credible interval increased in size with respect to the uncorrected model. But, the median increase in size is about 1.5 for extension $E2$, while for the Albert *et al.* and multinomial correction this increase is about 1.8 and 2.5, respectively.

The regression coefficients from the beta-binomial part of the ZIBB

regression model have the same interpretation as for the negative binomial part in the ZINB regression model. For the degenerate part (explaining 0) only age at start of brushing was important. However, we omitted this variable from the model to enhance simple interpretation of the model.

Table 7.12: *Posterior estimates of the examiner-specific corrected ZIBB regression model in combination with double binomial misclassification model predicting the* $dmft_{4,5}$ *(using* `WinBUGS Program 7.4`*).*

Parameter	Estimate (SE)	CI[a] %	
		2.5%	97.5%
without $(x, y)-$*co-ordinate terms*			
Intercept	−1.581(0.134)	−1.824	−1.295
Gender (girl)	0.050(0.075)	−0.099	0.197
Age (years)	0.330(0.080)	0.174	0.488
Brushing frequency (< 2)	0.109(0.097)	−0.082	0.298
Age start brushing (years)	0.159(0.033)	0.098	0.228
Systemic fluoride (yes)	−0.457(0.072)	−0.599	−0.318
Sugary drinks (yes)	0.387(0.079)	0.238	0.550
Between meals (> 2)	0.173(0.072)	0.031	0.314
p	0.049(0.074)	0.000	0.207
τ	0.963(0.188)	0.630	1.372
with $(x, y)-$*co-ordinate terms*			
Intercept	−1.454(0.178)	−1.809	−1.110
x-ordinate	0.217(0.048)	0.126	0.315
y-ordinate	−0.044(0.039)	−0.121	0.031
Gender (girl)	0.053(0.084)	−0.112	0.222
Age (years)	0.385(0.090)	0.214	0.562
Brushing frequency (< 2)	0.129(0.110)	−0.088	0.345
Age start brushing (years)	0.186(0.039)	0.112	0.265
Systemic fluoride (yes)	−0.491(0.080)	−0.651	−0.339
Sugary drinks (yes)	0.434(0.089)	0.260	0.609
Between meals (> 2)	0.130(0.079)	−0.025	0.289
p	0.177(0.085)	0.000	0.305
τ	0.822(0.201)	0.526	1.317

[a]CI = Credible interval.

The results from examiner-specific correction in combination with double binomial misclassification model are shown in Table 7.12. The examiner-specific correction have standard errors that are larger than those corresponding to pooled correction. The regression coefficients have the same interpretation as those of the pooled correction, except that the East-West gradient remains significant for the model including $(x, y)-$co-ordinate. However, there was no significant gain in terms of overdispersion after the correction, presented by $\hat{\tau}$. The excess of caries-free children ($\hat{p}$) is now much larger for examiner-specific correction (5% and 18% for the model without and with $(x, y)-$co-ordinate, respectively) compared to the pooled estimate (<1%). The validation data sets need to be large to avoid such changes in the parameter estimates.

7.8 Discussion

Correction for misclassification can only work efficiently if the correction terms are estimated with high precision. This necessitates that the validation study is large enough. To increase the efficiency with which the misclassification probabilities are estimated some modeling of these probabilities seems necessary. In this chapter, we have suggested to describe the misclassification process in a simple statistical way by the double binomial method. The gain in efficiency but also the decrease in bias, compared to multinomial modeling, can be large if assumptions $A1$-$A3$ are roughly satisfied. We admit that it is not clear how often these assumptions will hold in practice. But from the simulations we can conclude that moderate violation of these assumptions seems of less importance. Further, we have shown that our approach can easily be extended when the assumptions do not hold.

CHAPTER 8

General Misclassification Model using Simulation-Extrapolation

8.1 Introduction

Models for adjusting for measurement error can be grouped into two classes, namely functional and structural models, as described in Chapter 1. In this chapter we adopt a functional modeling approach whereby no distributional assumptions are made about the misclassified data. In particular, 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.

In the previous chapters we analyzed the prevalence and degree of caries experience at child level, i.e. with a single summarized outcome for each child. In this chapter, we model the prevalence of caries experience at tooth level, implying multiple observations from different teeth for each child. But, we are also interested in assessing the geographical trend in the prevalence of caries experience in Flanders for the tooth level dmft-index, taking into account the various risk factors.

8.2 SIMEX for Measurement Error in Continuous Data

The SIMEX (SIMulation EXtrapolation) method is a general method to correct for measurement error via simulation. This approach is applicable when the measurement error variance is known or can be estimated (accurately), e.g., from validation data or replicated measurements. The SIMEX method was first suggested by Cook and Stefanski (1994) and further developed by Stefanski and Cook (1995) and Carroll and Küchenhoff (1995). This method is designed for estimating a parameter β in a general regression problem with additive measurement error in one regressor. An estimator, which is consistent when all variables are measured without error, is assumed to be available. This estimator is usually called the naive estimator, when it is used in spite of measurement error.

The SIMEX method uses the relationship of the size of measurement error, i.e. the measurement error variance σ_u^2, to the bias of the effect estimators when ignoring measurement error. In other words, SIMEX estimator is obtained by adding additional measurement error to the observed data in a resampling stage, establishing a trend of the induced bias versus the variance of the added measurement error, and then extrapolating back to the case of no measurement error. So we define the function

$$\sigma_u^2 \longrightarrow \beta^*(\sigma_u^2) =: \mathcal{G}(\sigma_u^2),$$

whereby β^* is the limit to which the naive estimator converges as the sample size $n \to \infty$. Consistency then implies that $\mathcal{G}(0) = \beta$ and in many cases $\mathcal{G}(\sigma_u^2)$ declines in absolute value when σ_u^2 increases. $\mathcal{G}(\sigma_u^2)$ corresponds to the attenuation of the estimated effect induced by measurement error. The SIMEX method is based on a parametric approximation of the function $\mathcal{G}(\sigma_u^2) \approx \mathcal{G}(\sigma_u^2, \Gamma)$, e.g. for a quadratic approximation $\mathcal{G}(\sigma_u^2, \Gamma) = \gamma_0 + \gamma_1 \sigma_u^2 + \gamma_2 (\sigma_u^2)^2$.

The idea behind the SIMEX method is best illustrated in a simple linear measurement error regression model. Suppose that the regression

model of interest is $EY|X = \beta_0 + \beta_x X$ and that $X^* = X + \sigma_u U$ rather than X, is observed where U has mean zero and variance 1, and the measurement error variance σ_u^2 is known. As seen in Chapter 1, the ordinary least squares regression does not estimate β_x but instead $\beta_{x^*} = \{\sigma_x^2/(\sigma_x^2+\sigma_u^2)\}\beta_x$, where σ_x denotes the variance of X. Suppose that one repeatedly adds, by simulation, additional error with mean zero and variance $\sigma_u^2\lambda$ to X^* resulting in X^{**}, for fixed $\lambda \geq 0$, so that the variance of the X^{**} is $(\sigma_u^2 + \sigma_u^2\lambda) = (1 + \lambda)\sigma_u^2$. Then, an ordinary least squares regression of Y on X^{**} consistently estimates

$$\beta_{x^*}^*(\lambda) = \frac{\sigma_x^2}{\sigma_x^2 + (1+\lambda)\sigma_u^2}\beta_x.$$

Observe that $\beta_{x^*}^*(-1) = \beta_x$, a case of no measurement error. The idea is therefore to fit a regression model of $\beta_{x^*}^*(\lambda)$ against λ, and then extrapolate back to $\lambda = -1$.

In general, for a given data set, the method adds, by simulation, extra measurement error with variance $\lambda\sigma_u^2$ to the error prone variable. The resulting measurement error is then $(1+\lambda)\sigma_u^2$ leading to an estimator which converges to $\mathcal{G}[(1 + \lambda)\sigma_u^2)]$ for naive estimation. Repeating this simulation for a fixed grid of λs yields an estimator $\hat{\Gamma}$ of the parameters of $\mathcal{G}(\sigma_u^2, \Gamma)$ e.g. by least squares. In the extrapolation step the function $\mathcal{G}(\sigma_u^2, \Gamma)$ is extrapolated back to 0. The SIMEX estimator is defined by $\mathcal{G}(0, \hat{\Gamma})$, i.e. setting $\lambda = -1$ in $\mathcal{G}[(1+\lambda)\sigma_u^2)]$. If $\mathcal{G}(\sigma_u^2, \Gamma)$ is a good approximation to the true function $\mathcal{G}(\sigma_u^2)$ then the SIMEX procedure is approximately consistent. This has been verified in many cases. For the variance estimation three methods are available: Delta method (Carroll et al., 1996), jackknife type (Stefanski and Cook, 1995) and the bootstrap.

8.3 Misclassification SIMEX for Measurement Error in Discrete Data

We consider a general regression problem with a response Y and with a discrete regressor X and further correctly specified regressors Z. Usually misclassification error is characterized by a $k \times k$ misclassification matrix Π, which is defined in Section 1.4.2.1. The parameter of interest is β, with the limit of the naive estimator denoted by β^*. The existence of β^* and its determination can be assumed by the theory of misspecified models (see e.g. White, 1982), depending on the model and on the misclassification matrix. Therefore we denote it as $\beta^*(\Pi)$. Further we assume that $\beta^*(I_{k\times k}) = \beta$, i. e. that the estimator is consistent when there is no misclassification (represented by identity matrix $I_{k\times k}$).

For SIMEX we define the function ($\lambda \geq 0$)

$$\lambda \longrightarrow \beta^*(\Pi^\lambda), \tag{8.1}$$

whereby $\Pi^\lambda := \Xi\Lambda^\lambda\Xi^{-1}$, Λ is a diagonal matrix of eigenvalues and Ξ the corresponding matrix of eigenvectors. Note that for $\lambda = n$, an integer, $\Pi^{1+n} = \Pi^n * \Pi$ and that $\Pi^0 = I_{k\times k}$. Expression (8.1) allows the SIMEX method to be applied to the misclassification problem, in this case we will denote the method as MC-SIMEX. 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. For the function (8.1) to be well–defined we need to ensure the existence of Π^λ and that it is a misclassification matrix for $\lambda \geq 0$.

In the $0-1$ case this is equivalent to $\det(\Pi) = \pi_{00} + \pi_{11} - 1 > 0$. This is fulfilled, if $\pi_{00} > 0.5$ and $\pi_{11} > 0.5$, which should hold for any useful measurement of X, see Gastwirth (1987). In the case of three or more categories a possible problem is that some of the entries of Π^λ are negative

or that Π has negative eigenvalues. We discuss these problems in Appendix A.3 and give concrete examples in Section 8.4.

The MC-SIMEX algorithm consists in applying the misclassification matrix Π^λ to the misclassified variable in the simulation step. For the extrapolation step of the MC-SIMEX procedure we need a parametric approximation of (8.1):

$$\lambda \quad \longrightarrow \quad \beta^*(\Pi^\lambda) \approx \mathcal{G}(1+\lambda, \Gamma).$$

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

8.3.1 Simulation step

For a fixed grid of values $\lambda_1, \ldots, \lambda_m$, (≥ 0) we simulate B new pseudo data sets 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, we define $\lambda_0 = 0$, 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 and

$$\hat{\beta}(\lambda_k) := B^{-1} \sum_{b=1}^{B} \hat{\beta}_{na}\left[(Y_i, X_{b,i}^*(\lambda_k), Z_i)_{i=1}^n\right], k = 1, \ldots m. \tag{8.2}$$

8.3.2 Extrapolation step

Note that $\hat{\beta}(\lambda_k)$ is an average over naive estimators corresponding to data with misclassification matrix $\Pi^{1+\lambda_k}$. The estimator $\hat{\beta}$ is obtained by fitting a parametric model $\mathcal{G}(1+\lambda, \Gamma)$ 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

$$\hat{\beta}_{SIMEX} := \mathcal{G}(0, \hat{\Gamma}), \tag{8.3}$$

which corresponds to $\lambda = -1$. 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 $\hat{\beta}_{SIMEX}$ is consistent when the extrapolation function is correctly specified, i. e. $\beta^*(\Pi^\lambda) = \mathcal{G}(1 + \lambda, \Gamma)$, for some parameter vector Γ. However, this is often not the case. When $\mathcal{G}(1 + \lambda, \Gamma)$ is a good approximation of $\beta^*(\Pi^\lambda)$ then approximate consistency will hold. To find a suitable candidate for the function $\mathcal{G}(1 + \lambda, \Gamma)$ we exploit the relationship between β^* and the misclassification parameter λ in the next section for some special cases.

8.4 Calculation of the extrapolation function

We start with the regressor or the response measured without error and evaluate the effect of misclassification based on the matrix Π^λ on the estimated regression coefficients using the naive method. Hence in contrast to the previous section where we started with X^* and needed to take $\lambda = -1$ to find the MC-SIMEX estimate; here we start with X so that $\lambda = 0$ results in the MC-SIMEX estimate because we now work with $\mathcal{G}(\lambda, \Gamma)$.

8.4.1 Linear model

The simplest case is a linear model with one misclassified binary covariate $X \in \{0, 1\}$

$$\mathrm{E}(Y|X) = \beta_0 + \beta_1 X. \tag{8.4}$$

Here, β_1 is just the difference $E(Y|X=1) - E(Y|X=0)$, i. e. this is the situation of a two sample t-test.

Assuming that we observe X^*, which is related to X by the misclassification matrix Π, and denoting the marginal probability $\Pr(X=1)$ by π_x we get

$$\begin{aligned}\mathrm{E}(Y|X^*) &= \mathrm{E}[\mathrm{E}(Y|X)|X^*] = \mathrm{E}(\beta_0 + \beta_1 X|X^*) = \beta_0 + \beta_1 \mathrm{E}(X|X^*) \\ &= \beta_0 + \beta_1 X^* \Pr(X=1|X^*=1) + \beta_1(1-X^*)\Pr(X=1|X^*=0) \\ &= \beta_0^* + \beta_1^* X^*,\end{aligned}$$

where it can be shown that

$$\begin{aligned}\beta_0^* &= \beta_0 + \beta_1 \frac{(1-\pi_{11})\,\pi_x}{\pi_{00} - \delta\pi_x} \text{ and} \\ \beta_1^* &= \beta_1 \frac{\delta\ (1-\pi_x)\,\pi_x}{(1-\pi_{00}+\delta\ \pi_x)\,(\pi_{00}-\delta\ \pi_x)}, \\ \text{with } \delta &= \det(\Pi) = \pi_{00} + \pi_{11} - 1.\end{aligned} \tag{8.5}$$

Since we are interested in the effect of misclassification depending on the exponent of the misclassification matrix we have to evaluate Π^λ. To ensure the existence of Π^λ, we assume that $\det(\Pi) > 0$. This is a reasonable assumption, which is e.g. true, if $\pi_{00} > 0.5$ and $\pi_{11} > 0.5$. Then the matrix Π^λ is:

$$\Pi^\lambda = \Xi\Lambda\Xi^{-1} \tag{8.6}$$

where

$$\Xi = \begin{pmatrix} \frac{1-\pi_{11}}{1-\pi_{00}} & 1 \\ 1 & -1 \end{pmatrix}$$

is a matrix of the eigenvectors spanning the misclassification matrix Π, and

$$\Lambda = \begin{pmatrix} 1 & 0 \\ 0 & (\pi_{00}+\pi_{11}-1)^\lambda \end{pmatrix}$$

is the diagonal matrix of the eigenvalues. We can show that expression (8.6) simplifies to

$$\Pi^{\lambda} = \frac{1}{1-\delta} \begin{pmatrix} 1-\pi_{11}+(1-\pi_{00})\delta^{\lambda} & (1-\pi_{11})(1-\delta^{\lambda}) \\ (1-\pi_{00})(1-\delta^{\lambda}) & 1-\pi_{00}+(1-\pi_{11})\delta^{\lambda} \end{pmatrix}. \quad (8.7)$$

To evaluate the function $\lambda \longrightarrow \beta_1^*(\Pi^{\lambda})$ we have to insert the components of Π^{λ} in (8.5). For example, when $\pi_{00} = \pi_{11}$ and $\pi_x = 0.5$, then $\beta_1^*(\lambda) = \beta_1(2\pi_{11}-1)^{\lambda}$ and when $\pi_{00} = 1$, then $\beta_1^*(\lambda) = \beta_1 \frac{1-\pi_x}{1-\pi_x \pi_{11}^{\lambda}}$. In Figure 8.1(a) we plot the function $\beta_1^*(\lambda)$ for different values of the misclassification matrix, clearly demonstrating the attenuating effect of measurement error, here indicated by λ.

8.4.2 Estimation of a probability

The problem of estimating probability $\Pr(Y=1)$ can be treated as a binary regression model without covariates:

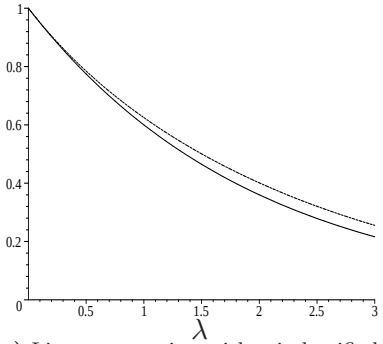
$$\Pr(Y=1) \quad = \quad g^{-1}(\beta_0), \quad (8.8)$$

where g is the link function, e.g. logit, probit etc. Instead of Y we observe Y^* with misclassification matrix Π. Then

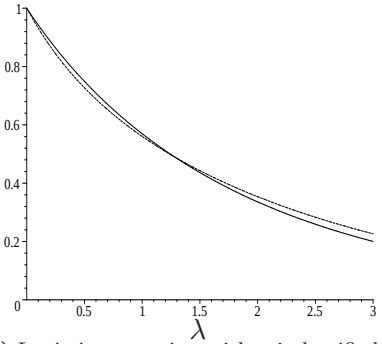
$$\Pr(Y^*=1) \quad = \quad \pi_{11}\Pr(Y=1) + (1-\pi_{00})(1-\Pr(Y=1)). \quad (8.9)$$

Solving (8.9) for $\Pr(Y=1)$ yields the ML estimator when $\Pr(Y^*=1)$ is estimated from the data by the relative frequency and π_{00} and π_{11} are known or estimated from validation data. The properties of the ML-estimator are well known (see e.g. Gastwirth (1987) and Stefanski (1992)). To use the MC-SIMEX approach we define $\beta_0^* := g\left[\Pr(Y^*=1)\right]$ and we get

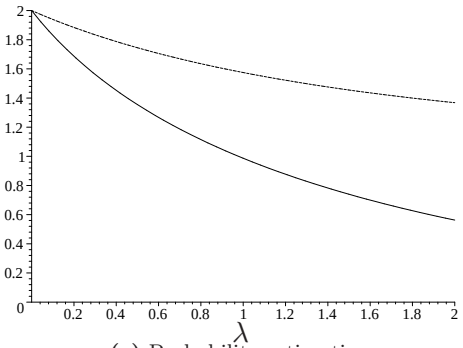
$$\beta_0^*(\lambda) \quad = \quad g\left[\pi_{11}^{(\lambda)} g^{-1}(\beta_0) + (1-\pi_{00}^{(\lambda)})(1-g^{-1}(\beta_0))\right]. \quad (8.10)$$



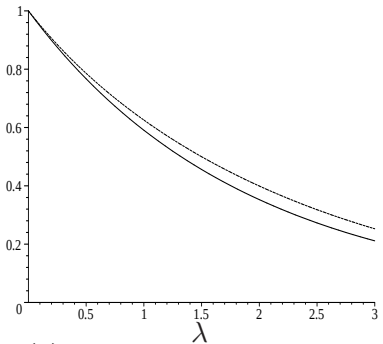
(a) Linear regression with misclassified X.



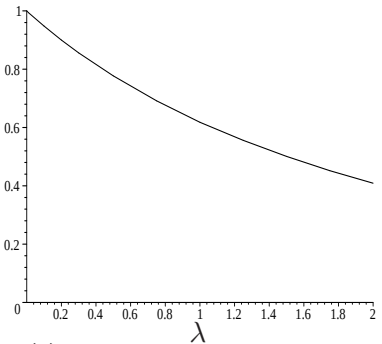
(b) Logistic regression with misclassified Y.



(c) Probability estimation.



(d) Logistic regression with misclassified X.



(e) Ordinal logistic regression with misclassified Y.

Figure 8.1: *Limit of the naive estimator (y−axis) depending on the exponent* λ *of the misclassification matrix* Π. *In (c) the true parameter is* $\beta_1 = 2$, *while in all other plots the true regression parameter is* $\beta_1 = 1$. *In the first four plots the misclassification matrix is given by* $\pi_{00} = \pi_{11} = 0.8$ *(solid line) and* $\pi_{00} = 0.7, \pi_{11} = 0.9$ *(dotted line). In the last plot, the misclassification matrix is given by* $\pi_{ii} = 0.8$ *and* π_{ij}*=0.1* $(i \neq j)$.

Here, $\pi_{ij}^{(\lambda)}$ are the components of the matrix Π^{λ}, see (8.7). In Figure 8.1(c) we display the function $\beta_0(\lambda)$ for the logistic link and for the case of $\beta_0 = 2$, which relates to $\Pr(Y = 1) = 0.88$.

8.4.3 Binary regression with a misclassified response

We now explore the case of a binary regression with a misclassified response:

$$\Pr(Y = 1|X) \quad = \quad g^{-1}(\beta_0 + \beta_1 X).$$

When we observe Y^* with misclassification matrix Π then we know from Section 4.4 that the observed regression is

$$\Pr(Y^* = 1|X) = (1 - \pi_{00}) + (\pi_{00} + \pi_{11} - 1)g^{-1}(\beta_0 + \beta_1 X)$$

and that the observed model is not logistic, but can be written as a GLM with a modified link function. This yields the following approximations of the function $\beta^*(\lambda)$ (Section 4.4.2):

For the identity link:

$$\beta_1^* = \delta^{(\lambda)}\beta_1.$$

For the logistic link

$$\beta_1^* = \frac{\delta^{\lambda}\exp(\beta_0)}{\left((1 - \pi_{01}^{(\lambda)})\exp(\beta_0) + \pi_{10}^{(\lambda)}\right)\left(\pi_{01}^{(\lambda)}\exp(\beta_0) + 1 - \pi_{10}^{(\lambda)}\right)}\beta_1.$$

In the case of a binary X, β is simply estimated by the log odds ratio. Therefore β_1^* can be calculated exactly. In Figure 8.1(b) we display the function $\beta_1^*(\lambda)$ for different values of the misclassification matrix and $\beta_0 = 0$ and $\beta_1 = 1$.

8.4.4 Logistic regression with a misclassified binary covariate

In simple logistic regression with a misclassified binary covariate, which corresponds to the case of a 2×2 table, we are able to calculate the observed odds ratios by the matrix method (Section 1.4.2.2) when the marginal probabilities are given. In Figure 8.1(d) we give the result for one particular example.

8.4.5 Misclassification of a response with three categories

We explore the case of an ordinal response variable with three categories ($k = 1, 2, 3$) with a single covariate. The cumulative logit model, see Agresti (1984), is given by

$$\Pr(Y \leq k \mid X) = G(\alpha_k + \beta X), \;\; k = 1, 2, 3, \tag{8.11}$$

where α_k are the threshold parameters and $G(t) = [1 + \exp(-t)]^{-1}$. Assuming that Y is misclassified with the 3×3-matrix Π we can calculate the observed model by

$$\begin{aligned}
\Pr(Y^* \leq 1 \mid X) &= \pi_{11} G(\alpha_1 + \beta X) + \pi_{12}[G(\alpha_2 + \beta X) - G(\alpha_1 + \beta X)] + \\
&\quad \pi_{13}[1 - G(\alpha_2 + \beta X)] \\
\Pr(Y^* \leq 2 \mid X) &= \Pr(Y^* \leq 1 \mid X) + \pi_{21} G(\alpha_1 + \beta X) + \\
&\quad \pi_{22}[G(\alpha_2 + \beta X) - G(\alpha_1 + \beta_x)] + \pi_{23}[1 - G(\alpha_2 + \beta X)]
\end{aligned}$$

To find the vector $\gamma^* := (\alpha_1^*, \alpha_2^*, \beta^*)$, the Kullback Leibler distance from the observed model to model (8.11) has to be minimized. This yields the following system of equations to be solved:

$$\int Z' D\Sigma^{-1} \begin{pmatrix} \Pr(Y^* = 1|x) - G_1 \\ \Pr(Y^* = 2|x) - (-G_1 + G_2) \end{pmatrix} df_x = 0,$$

with

$$\begin{aligned} Z &= \begin{pmatrix} 1 & 0 & x \\ 0 & 1 & x \end{pmatrix}, \\ D &= \begin{pmatrix} G_1(1-G_1) & -G_1(1-G_1) \\ 0 & G_2(1-G_2) \end{pmatrix}, \\ \Sigma &= \begin{pmatrix} G_1(1-G_1) & -G_1G_2 \\ -G_1G_2 & G_2(1-G_2) \end{pmatrix}, \\ G_i &:= G(\alpha_i^* + \beta^* x). \end{aligned}$$

In Figure 8.1(e) we give the function $\beta^*(\lambda)$ for a special setting with $\alpha_1 = 1, \alpha_2 = 2, \beta = 1$. Here X is assumed to be binary, which reduces the above integral to a sum. The results are similar in other settings. In principle the result can be well approximated by using a discrete distribution for X.

8.4.6 Consequences for the extrapolation function

In all cases we explored, the extrapolation function was monotonic in all parameters. The function has a curvature which should be well approximated by the quadratic extrapolation function, see Figure 8.1. In some simple cases the function is exponential in λ, which is also approximately true in more complicated cases. Therefore, we also use a log linear extrapolation function $\mathcal{G}_{LOG}(\lambda) := \exp(\gamma_0 + \gamma_1\lambda)$. The quadratic and linear extrapolation functions are given by $\mathcal{G}_Q(\lambda, \Gamma) := \gamma_0 + \gamma_1\lambda + \gamma_2\lambda^2$ and $\mathcal{G}_L(\lambda) := \gamma_0 + \gamma_1\lambda$, respectively.

8.5 Variance estimation of β_{SIMEX}

Since the methods of Carroll et al. (1996) and Stefanski and Cook (1995) are based on expansions where the additivity of the measurement error is essential, they do not simply apply to the variance estimation in the MC-SIMEX approach. However, the variance estimation of the approach of Stefanski and Cook (1995) showed good simulation results, so we use it for our analysis in our example. It is described in Appendix B and will be referred to $\mathrm{V_{ST}}$ with corresponding standard error $\mathrm{SE_{ST}} = \sqrt{\mathrm{V_{ST}}}$.

Another approach is the application of the bootstrap. It allows to take the uncertainty about the misclassification matrix into account. This can be done by a two stage bootstrap procedure: In the first step the misclassification matrix is estimated by a bootstrap sample from a validation study or by a parametric bootstrap step. Then the MC-SIMEX procedure is done on a bootstrap sample from the main data using the misclassification matrix from the first step. This is repeated in the usual way to find bootstrap variances and bootstrap confidence intervals. A 95% bootstrap confidence interval for β is computed as $\hat{\beta} \pm 1.96 \cdot \widehat{\mathrm{SE}}_b(\hat{\beta})$, where $\widehat{\mathrm{SE}}_b(\hat{\beta})$ is the estimated standard deviation of $\hat{\beta}$ from bootstrap samples. Note that this approach is rather time consuming, since MC-SIMEX has to be calculated for every bootstrap sample.

8.6 Simulation Study

8.6.1 Simulation study set up

The MC-SIMEX procedure is evaluated for different settings having a rather high probability of misclassification ($\pi_{00} = 0.8, \pi_{11} = 0.8$ and $\pi_{00} = 0.9, \pi_{11} = 0.7$). In the case of a binary response with misclassification, we compared MC-SIMEX estimates to the estimates from the maximum likelihood (ML) approach of Neuhaus (1999). Whereas in the case of a misclassified binary regressor we compared the MC-SIMEX estimates to the

estimates from the matrix method (Matrix) presented in Section 1.4.2.2. We focussed on logistic regression and considered the following cases that may be anticipated to occur in practice:

A. Misclassified binary response Y^* and a binary or a continuous covariate X. The MC-SIMEX estimates are compared to the ML estimates.

B. Correctly measured binary response Y and a binary misclassified covariate X^* with and without an additional correctly measured continuous confounder variable Z. The MC-SIMEX estimates are compared to the matrix method estimates for the case without the confounder.

C. Correctly measured binary response Y with differential misclassification in a binary covariate X^* dependent on Y. The MC-SIMEX estimates are compared to the matrix method estimates.

D. Misclassification both in a binary response Y^* and in a binary covariate X^* with and without an additional correctly measured continuous confounder Z.

We performed 200 simulations each time with a sample size of 1000. One motivation for this large sample size is that correction for measurement error is usually done for large epidemiological studies. We generated the true binary covariate X from a Bernoulli distribution with $\Pr(X = 0) = \Pr(X = 1) = 0.5$, whereas the confounder variable Z is generated from a normal distribution with mean equal to 0.5 for $X = 0$ or -0.5 for $X = 1$, and variance 1. The continuous covariate in case A is randomly drawn from a standard normal distribution. The true response is generated as a Bernoulli random variable with $\Pr(Y = 1) = 1/\left(1 + \exp(-\beta_0 - \beta_X X)\right)$ for cases A & C, or $\Pr(Y = 1) = 1/\left(1 + \exp(-\beta_0 - \beta_X X - \beta_Z Z\right)$ for cases B & D. We apply a misclassification operation (see Section 8.3.1) on Y and X to obtain the misclassified variables Y^* and X^*, respectively. Note

that for Case C (differential misclassification) the misclassification can be represented by a 4×4 matrix, i.e.

$$\Pi = \begin{pmatrix} \pi_{000} & \pi_{001} & 0 & 0 \\ \pi_{010} & \pi_{011} & 0 & 0 \\ 0 & 0 & \pi_{100} & \pi_{101} \\ 0 & 0 & \pi_{110} & \pi_{111} \end{pmatrix},$$

where $\pi_{ijk} = \Pr(X^* = j | X = k, Y = i)$; $i, j, k = 0, 1$. The top-left and the bottom-right blocks correspond to misclassification of X for $Y = 0$ and for $Y = 1$, respectively.

We calculated the MC-SIMEX estimator for the three extrapolation functions (linear, quadratic & log-linear), the naive estimator and the 'true' estimator using the (unobserved) benchmark variable. The true estimator is obtained by regressing the correctly measured Y on the correctly measured X (and/or Z). The MC-SIMEX procedure was performed with $B = 100$ (Section 8.3.1). Hence, for each of the 200 simulations the $B = 100$ repetitions result in an extrapolated point and a (Stefanski and Cook, 1995) variance estimate for each parameter. Hereby, one can calculate the mean of the 200 point estimates, and the mean of the 200 variance estimates $\mathrm{V_{ST}}$ and hence the standard error $\mathrm{SE_{ST}}$.

8.6.2 Simulation results of MC-SIMEX method

The simulation results indicate that our method leads to substantial reduction of bias compared to the naive estimator. In the tables below, the MC-SIMEX standard error is given by $\mathrm{SE} = \mathrm{SE}_{ST} = \sqrt{\frac{1}{200} \sum_{i=1}^{200} \mathrm{V}_{ST,i}}$, where $\mathrm{V}_{ST,i}$ is the Stefanski and Cook (1995) variance estimate of the ith simulation; whereas for *True model*, *Naive model* and *ML* (maximum likelihood) $\mathrm{SE} = \sqrt{\frac{1}{200} \sum_{i=1}^{200} \mathrm{V}_{M,i}}$, where $\mathrm{V}_{M,i}$ is the maximum likelihood variance estimate of the ith simulation. The nature of the sensitivity (π_{11}) and the

Table 8.1: *Simulation results for Case A: Logistic regression of a misclassified response Y on a binary or a continuous covariate X. The true regression coefficients are $\beta_0 = 0$ & $\beta_X = 1$, only the results for β_X are shown (based 200 simulations each with sample size = 1000).*

		$(\pi_{00}, \pi_{11}) = (0.9, 0.7)$				$(\pi_{00}, \pi_{11}) = (0.8, 0.8)$			
Estimator		Estimate	SE	RMSE[a]	CR[b]	Estimate	SE	RMSE[a]	CR[b]
X is a binary covariate									
True model	β_X	0.989	0.135	0.135	0.955	0.989	0.135	0.135	0.955
Naive model	β_X	0.559	0.128	0.459	0.075	0.573	0.129	0.446	0.125
MC-SIMEX(L)	β_X	0.702	0.179	0.348	0.550	0.731	0.186	0.327	0.645
MC-SIMEX(Q)	β_X	0.842	0.179	0.239	0.865	0.903	0.186	0.210	0.940
MC-SIMEX(LOG)	β_X	0.875	0.179	0.218	0.895	0.966	0.186	0.189	0.970
ML	β_X	1.004	0.240	0.240	0.950	1.003	0.237	0.237	0.970
X is a continuous covariate									
True model	β_X	1.002	0.083	0.083	0.945	1.002	0.083	0.083	0.945
Naive model	β_X	0.553	0.071	0.453	0.000	0.530	0.069	0.475	0.000
MC-SIMEX(L)	β_X	0.698	0.104	0.320	0.055	0.676	0.103	0.340	0.040
MC-SIMEX(Q)	β_X	0.848	0.104	0.184	0.720	0.852	0.103	0.181	0.750
MC-SIMEX(LOG)	β_X	0.884	0.104	0.155	0.820	0.903	0.103	0.142	0.880
ML	β_X	1.013	0.159	0.159	0.980	1.017	0.165	0.166	0.975

[a]RMSE = Root mean square error.
[b]CR = Coverage rate, based on SE.

Table 8.2: *Simulation results for Case B: Logistic regression of a correctly measured response Y on a binary misclassified covariate X with and without a correctly measured continuous confounder Z. The true regression coefficients were $\beta_0 = 0, \beta_X = 1$ & $\beta_Z = 1$. The results are not shown for β_0 (based 200 simulations each with sample size = 1000).*

		$(\pi_{00}, \pi_{11}) = (0.9, 0.7)$				$(\pi_{00}, \pi_{11}) = (0.8, 0.8)$			
Estimator		Estimate	SE	RMSE[a]	CR[b]	Estimate	SE	RMSE[a]	CR[b]
Without confounder									
True model	β_X	0.993	0.135	0.135	0.955	0.993	0.135	0.135	0.955
Naive model	β_X	0.616	0.137	0.408	0.145	0.593	0.132	0.428	0.080
MC-SIMEX(L)	β_X	0.778	0.189	0.292	0.735	0.757	0.190	0.308	0.630
MC-SIMEX(Q)	β_X	0.916	0.189	0.207	0.930	0.928	0.190	0.203	0.935
MC-SIMEX(LOG)	β_X	0.971	0.189	0.192	0.955	0.995	0.190	0.190	0.965
Matrix	β_X	0.987	0.181	0.181	0.930	1.009	0.229	0.229	0.980
With an additional continuous confounder Z									
True model	β_X	1.025	0.166	0.168	0.965	1.025	0.166	0.168	0.965
Naive model	β_X	0.535	0.151	0.488	0.120	0.513	0.148	0.509	0.070
MC-SIMEX(L)	β_X	0.676	0.216	0.389	0.545	0.654	0.216	0.408	0.525
MC-SIMEX(Q)	β_X	0.829	0.216	0.275	0.880	0.830	0.216	0.275	0.910
MC-SIMEX(LOG)	β_X	0.864	0.216	0.255	0.915	0.884	0.216	0.245	0.935
True model	β_Z	1.011	0.085	0.086	0.935	1.011	0.085	0.086	0.935
Naive model	β_Z	0.846	0.076	0.172	0.440	0.843	0.076	0.174	0.445
MC-SIMEX(L)	β_Z	0.869	0.082	0.155	0.650	0.865	0.082	0.158	0.610
MC-SIMEX(Q)	β_Z	0.923	0.082	0.113	0.875	0.921	0.082	0.114	0.860
MC-SIMEX(LOG)	β_Z	0.871	0.082	0.153	0.650	0.867	0.082	0.157	0.625

[a]RMSE = Root mean square error.
[b]CR = Coverage rate, based on SE.

Table 8.3: *Simulation results for Case C: Logistic regression of a correctly measured response Y on a binary covariate X with differential misclassification dependent on Y. The true regression coefficients were $\beta_0 = 0$ & $\beta_X = 1$, only the results for β_X are shown (based 200 simulations each with sample size = 1000).*

Estimator		Estimate	SE	RMSE[a]	CR[b]
$(\pi_{000}, \pi_{011}, \pi_{100}, \pi_{111}) = (.9, .7, .7, .8)$					
True model	β_X	1.008	0.135	0.135	0.945
Naive model	β_X	1.203	0.138	0.246	0.705
MC-SIMEX(L)	β_X	1.120	0.200	0.233	0.885
MC-SIMEX(Q)	β_X	1.042	0.200	0.205	0.945
MC-SIMEX(LOG)	β_X	1.133	0.200	0.241	0.875
Matrix	β_X	1.020	0.332	0.332	0.990
$(\pi_{000}, \pi_{011}, \pi_{100}, \pi_{111}) = (.8, .8, .75, .75)$					
True model	β_X	1.008	0.135	0.135	0.945
Naive model	β_X	0.561	0.132	0.458	0.055
MC-SIMEX(L)	β_X	0.721	0.193	0.339	0.570
MC-SIMEX(Q)	β_X	0.913	0.193	0.212	0.920
MC-SIMEX(LOG)	β_X	0.995	0.193	0.193	0.950
Matrix	β_X	1.022	0.355	0.356	0.995

[a]RMSE = Root mean square error.
[b]CR = Coverage rate,based on SE.

specificity (π_{00}) does not have a great impact on the misclassification as can be seen in Tables 8.1 and 8.2. The results presented in Table 8.1 show a better correction in the case of binary regressor than in the case of a continuous regressor to the logistic regression with misclassified response. Furthermore, it is clear from Table 8.2 that the addition of a confounder leads to more attenuation than the case without a confounder, and consequently a poorer correction than in the case with no confounder.

As can be seen in Table 8.3, we observe that differential misclassification can lead to attenuation but in different directions, e.g. away from (*left panel*) or towards (*right panel*) the true estimate. Table 8.4 shows the results from the logistic regression with misclassification in both the response and

Table 8.4: *Simulation results for Case D: Logistic regression of a misclassified response Y on a misclassified binary covariate X with additional continuous confounder Z. The true regression coefficients were $\beta_0 = 0$, $\beta_X = 1$ & $\beta_Z = 1$, the results are not shown for β_0 (based 200 simulations each with sample size = 1000).*

		$(\pi_{00}, \pi_{11})_x = (\pi_{00}, \pi_{11})_y = (0.9, 0.7)^\dagger$				$(\pi_{00}, \pi_{11})_x = (\pi_{00}, \pi_{11})_y = (0.8, 0.8)^\dagger$			
Estimator		Estimate	SE	RMSE[a]	CR[b]	Estimate	SE	RMSE[a]	CR[b]
Without a confounder									
True model	β_X	0.984	0.135	0.136	0.955	0.984	0.135	0.136	0.955
Naive model	β_X	0.344	0.130	0.669	0.005	0.337	0.128	0.675	0.000
MC-SIMEX(L)	β_X	0.449	0.206	0.588	0.145	0.441	0.208	0.596	0.135
MC-SIMEX(Q)	β_X	0.669	0.206	0.389	0.775	0.695	0.208	0.369	0.835
MC-SIMEX(LOG)	β_X	0.866	0.206	0.245	0.925	1.008	0.208	0.208	0.965
With a continuous confounder Z									
True model	β_X	0.971	0.165	0.167	0.955	0.971	0.165	0.167	0.955
Naive model	β_X	0.256	0.139	0.757	0.000	0.253	0.137	0.759	0.000
MC-SIMEX(L)	β_X	0.332	0.221	0.703	0.035	0.328	0.222	0.708	0.070
MC-SIMEX(Q)	β_X	0.517	0.221	0.531	0.580	0.537	0.222	0.513	0.675
MC-SIMEX(LOG)	β_X	0.704	0.221	0.369	0.960	0.774	0.222	0.316	0.990
True model	β_Z	1.002	0.085	0.085	0.940	1.002	0.085	0.085	0.940
Naive model	β_Z	0.441	0.064	0.563	0.000	0.442	0.064	0.562	0.000
MC-SIMEX(L)	β_Z	0.554	0.094	0.455	0.000	0.563	0.097	0.448	0.005
MC-SIMEX(Q)	β_Z	0.702	0.094	0.313	0.260	0.736	0.097	0.281	0.435
MC-SIMEX(LOG)	β_Z	0.713	0.094	0.302	0.285	0.768	0.097	0.252	0.510

[a]RMSE = Root mean square error.

[b]CR = Coverage rate, based on SE.

†Subscripts x and y indicate the misclassification matrix Π for the regressor and response, respectively.

the binary covariate with or without an additional continuous confounder. The addition of the confounder further attenuates the naive estimate of β_X. Nonetheless, the MC-SIMEX correction gives improved estimates even for this complicated situation and even with this rather high misclassification probability, although some bias remains.

8.6.3 Comparison to alternative methods

The performance of the MC-SIMEX method to the ML approach (Chapter 4) and matrix method is compared with regard to bias, root mean square error (RMSE), and coverage rate. The coverage rate is computed as the proportion of situations that the 95% confidence interval (95% CI) includes the true value of β, with 95% CI $= \hat{\beta} \pm 1.96 \cdot \widehat{\mathrm{SE}}(\hat{\beta})$, where $\widehat{\mathrm{SE}}(\hat{\beta})$ depends on the actual model used.

In the case of a logistic regression of a misclassified binary response regressed on either a correctly measured binary or a continuous covariate, the ML method performs better in terms of the Bias and coverage rate than the MC-SIMEX method as shown in Table 8.1. However, the performance of the ML estimates is similar to the MC-SIMEX in terms of the RMSE.

In Table 8.2 we compare the MC-SIMEX and the Matrix method in the case of a misclassified binary regressor. Compared to Table 8.1, the two approaches perform better in terms of bias reduction. Moreover, the MC-SIMEX method (logarithmic and quadratic extrapolation) gives a better coverage rate than the Matrix method, though, the RMSE's are comparable. Finally, we compared the MC-SIMEX method to the Matrix method in the case of differential misclassification in the binary regressor (see Table 8.3). Again, the two methods are comparable in terms of bias. However, the Matrix method has a higher RMSE.

As a general conclusion we suggest that the MC-SIMEX method substantially reduces bias compared to the naive estimator and its performance is comparable to ML- or Matrix estimation when it is feasible. It is also applicable to complicated situations as non differential misclassification error

and yields good results.

8.7 Application to the Signal Tandmobiel® Study

The response of interest is a binary variable, coded 1 if the tooth is decayed, missing due to caries or filled, and 0 otherwise. Thus, the analysis here is at a tooth level, not a child level analysis as in the previous chapters. Indeed, this is the first time were a modeling tooth level outcome. Furthermore, we will investigate whether the East-West gradient is also present in the four first permanent molars in the first year of the study, and whether the trend remains or changes in time. In addition to the tooth level binary response, a child-level binary response, combined binary response (over the 4 first molars), was constructed – which is 1 if there is caries on at least one of the 4 teeth.

The prevalence of caries experience on the four first molars was modeled with a GEE analysis (SAS version 8.2 PROC GENMOD, SAS© Institute Inc., 1999–2001) as a function of several covariates and tooth. The covariates of interest were x-, and y-ordinate (both standardized), year $= 0, 1, \cdots, 5$ (0 = baseline, 5 = end of study), gender (*girl = 1*), upper jaw tooth dummy (U), and right side tooth dummy (R).

8.7.1 Logistic regression on a combined teeth response

We fitted a logistic regression model on the combined binary response using as covariates: $x-$ordinate, $y-$ordinate, year, gender and the interaction of year with $x-$ordinate, using GEE with a MC-SIMEX correction for misclassification. The correction was based on a misclassification matrix for each examiner – each child's probability for caries experience was corrected using the corresponding misclassification matrix. In addition, we applied a two stage bootstrap procedure to MC-SIMEX as explained in Section 8.5. The results of MC-SIMEX with and without bootstrap from PROC GENMOD are shown in Table 8.5.

The results suggest a baseline East-West gradient in prevalence of caries experience but now on the first permanent molar. However, the difference established at baseline remains the same over the years (non-significant x-ordinate×year). In addition, year and gender are important predictors of caries experience – the older the child the more caries, with girls being more affected than boys. The results are comparable to the results of Chapters 4 and 5. But gender was not significant in the ordinal logistic model. As expected, the bootstrap standard deviations are in general larger than the Stefanski and Cook (1995) standard errors. This extra variation in bootstrap estimates is a result of taking into the account the uncertainty in estimating the misclassification structure.

Table 8.5: *The mean and the confidence limits from the MC-SIMEX and the bootstrapped MC-SIMEX parameter estimates from a GEE model on the combined binary response from first molars. Only the estimates from the log-linear extrapolation are shown.*

Parameter	Estimate	Estimate	95% CI	
Parameter	SE	SE	Lower	Upper
	Naive[a]	MC-SIMEX[b]		
intercept	−1.884(0.047)	−1.858(0.053)	−1.962	−1.754
x-ordinate	0.096(0.041)	0.097(0.043)	0.013	0.182
y-ordinate	−0.014(0.029)	−0.013(0.038)	−0.088	0.063
gender	0.164(0.057)	0.238(0.067)	0.108	0.369
year	0.313(0.009)	0.446(0.009)	0.429	0.463
x-ordinate×year	0.019(0.009)	0.012(0.008)	−0.004	0.029
		Bootstrapped MC-SIMEX[c]		
intercept		−2.398(0.188)	−2.766	−2.031
x-ordinate		0.078(0.033)	0.014	0.142
y-ordinate		−0.018(0.055)	−0.126	0.089
gender		0.245(0.089)	0.070	0.420
year		0.530(0.062)	0.409	0.651
x-ordinate×year		0.015(0.021)	−0.025	0.056

[a]GEE standard errors from PROC GENMOD are given in parenthesis.
[b]Stefanski and Cook (1995) standard are given in parenthesis
[c]Bootstrap standard deviation are given in parenthesis.

8.7.2 GEE analysis on tooth level response

A GEE analysis with a logistic link and with a MC-SIMEX correction for misclassification was performed on tooth level. The included covariates are found in Table 8.6 and were obtained from the initial variable selection procedure. Again, the correction was done using a misclassification matrix for each examiner. We did not consider the bootstrap in this model because the MC-SIMEX results with and without bootstrap in Section 8.7.1 were very close. Further, bootstrapping now would be too time consuming because of the large size of the data at tooth level. The results from PROC GENMOD with a MC-SIMEX correction for misclassification are shown in Table 8.6.

Table 8.6: *MC-SIMEX parameter estimates from a GEE model to caries prevalence at tooth (first molar) level using different misclassification probabilities matrix for each examiner.*

Parameter	Naive	MC-SIMEX		
		Linear	Quadratic	Loglinear
	Estimate SE	Estimate SE	Estimate SE	Estimate SE
intercept	−2.756(0.062)	−2.964(0.070)	−3.029(0.093)	−2.998(0.111)
x-ordinate	0.146(0.041)	0.129(0.047)	0.122(0.060)	0.132(0.060)
y-ordinate	−0.016(0.030)	−0.004(0.033)	0.016(0.040)	−0.011(0.040)
gender	0.183(0.060)	0.237(0.068)	0.235(0.081)	0.288(0.083)
year	0.325(0.012)	0.407(0.014)	0.494(0.017)	0.507(0.023)
U	−0.049(0.067)	−0.043(0.071)	−0.114(0.105)	−0.038(0.083)
R	−0.008(0.061)	−0.005(0.071)	−0.056(0.092)	−0.011(0.185)
U×R	−0.057(0.086)	−0.080(0.094)	−0.131(0.133)	−0.079(0.132)
x-ordinate×year	0.018(0.008)	0.018(0.008)	0.022(0.011)	0.018(0.009)
U×year	0.001(0.017)	−0.001(0.018)	−0.004(0.026)	0.001(0.022)
R×year	−0.025(0.015)	−0.024(0.018)	−0.070(0.020)	−0.015(0.026)
U×R×year	0.045(0.022)	0.049(0.025)	0.120(0.032)	0.145(0.030)

The corrected parameter estimates are in absolute value all larger than the corresponding naive estimates as shown in Table 8.6, thereby adjusting for the attenuation effect due to misclassification of the caries experience. But, it is not immediately clear from Figure 8.2, to decide which of the

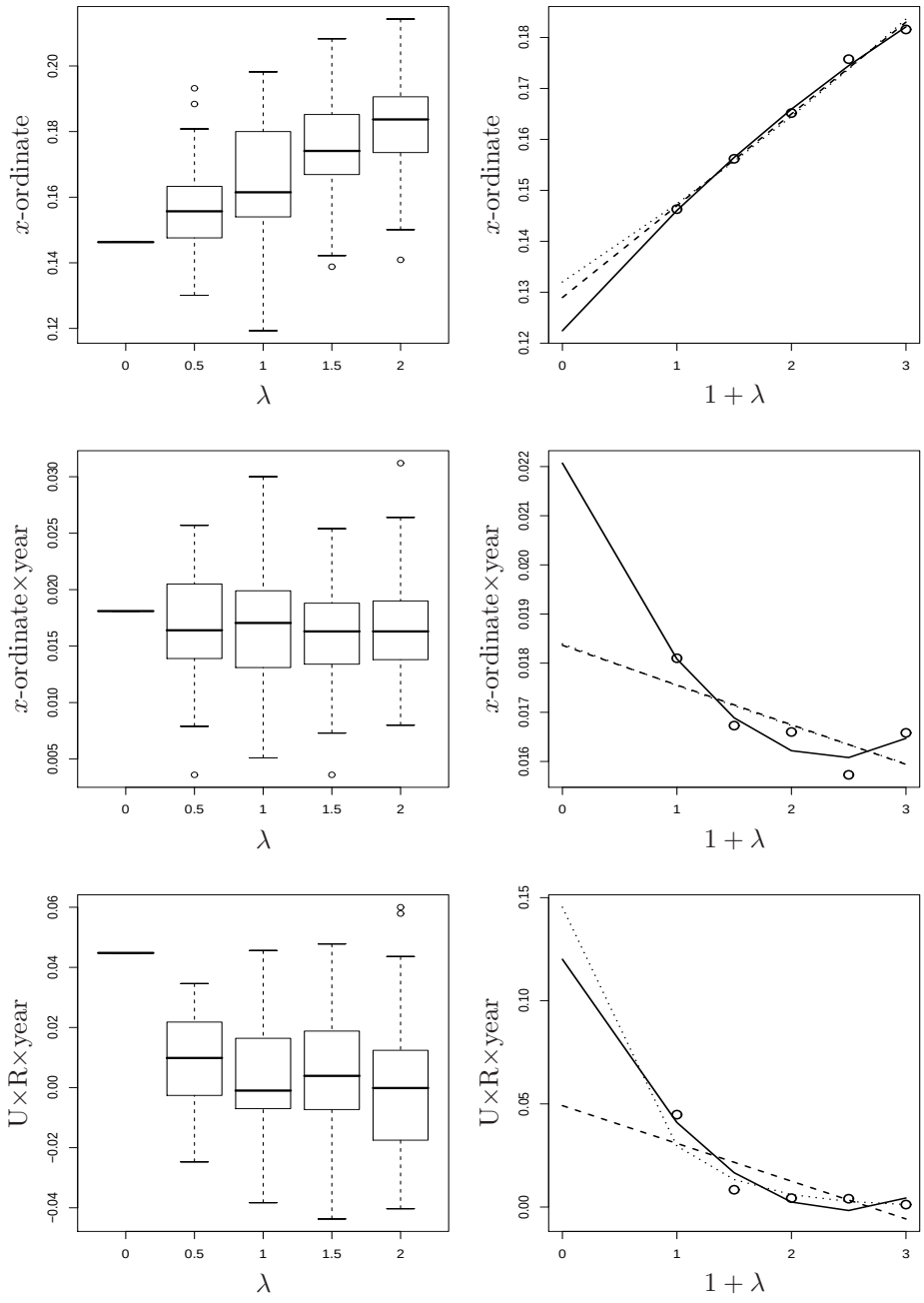


Figure 8.2: *The left panel shows the the box plots of the simulated data, whereas right panel shows the fitted extrapolants – linear(dashed line), quadratic (solid line) and log-linear (dotted line), for the MC-SIMEX fit to the Signal Tandmobiel® data.*

three extrapolation functions should be chosen.

In this analysis the importance of year, gender and the East-West gradient are confirmed. Further, now the East-West gradient seems to be slightly more pronounced in the later years (significant x-ordinate×year). Finally, the regression coefficient of the interaction U×R×year has increased considerably with MC-SIMEX with only a moderate increase of its standard error. The reason for this is not clear.

8.8 Discussion

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 simulation and extrapolation idea for additive normal covariate measurement error (Cook and Stefanski, 1994). Our approach is very 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. 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.

Note that the MC-SIMEX can also be applied in the simple case of prevalence estimation, see Section 8.4.2. While one would usually prefer the ML estimation, in some cases this estimate can be out of $[0; 1]$. Lew and Levy (1989) discuss this problem and propose a Bayesian solution. Like in other papers on prevalence estimation (see also Stefanski (1992)) they exploit the functional relationship between the observed and the true prevalence. In contrast, the idea of MC-SIMEX is a parametric approximation of the observed prevalence as a function of the parameter λ. So the

MC-SIMEX could serve as an alternative in prevalence estimation, if the ML estimator is not in $[0; 1]$.

Like for the original SIMEX 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, SIMEX-methods in general are only approximately consistent. We have shown by calculations of the true extrapolation function in some concrete models, that a log linear or a quadratic function is a good approximation. Furthermore, in a concrete data set the results of the simulation step give an indication of the form of the extrapolation function.

For variance estimation and confidence intervals we have proposed using a two step bootstrap method in the case of uncertain knowledge of the misclassification matrix. When the misclassification matrix is known, the method by Stefanski and Cook (1995) can be applied, since it has shown good results in the simulation. However, this method still lacks theoretical foundation in the misclassification case.

In our simulation study the MC-SIMEX method has shown good results. It reduces bias compared to the naive estimator and its performance is comparable to ML-estimation, where it is feasible.

CHAPTER
9 Analysis of Multivariate Binary Data Subject to Response Misclassification

9.1 Introduction

We propose in this chapter a general approach to correct for misclassification error when interest lies in regressing a multivariate binary vector on covariates. The approach is applied to the caries experience on tooth level taking into account that the misclassification process now depends on the type of tooth.

In this respect, Neuhaus (2002) suggested both a GEE and a GLMM to adjust for misclassification error in correlated binary data. Paulino et al. (2005) considered a random intercept logistic regression model for multivariate binary data and combined it with a binomial model for misclassification. However, the assumptions of the above approaches are too rigid for the misclassification process in scoring caries experience. For instance, their model assumes equal misclassification rates for all teeth. Using our approach, these assumptions can be relaxed.

9.2 Neuhaus' approach for possibly misclassified multivariate binary data

9.2.1 Classical approach to analyse multivariate binary data

Consider a sample of subjects (clusters) indexed by $i = 1, \cdots, N$, and let $T_i = T$ be the number of responses on subject i. Denote the tth binary response of ith subject by Y_{it}. Further let $\boldsymbol{x}_{it}$ denote a $d \times 1$ vector of fixed covariates for the tth response, $t = 1, 2, \cdots, T$.

The most common methods to model clustered data are: (a) a multivariate marginal parametric approach incorporating the correlation structure of the responses; (b) a subject specific parametric model like the generalized linear mixed model (GLMM); and (c) the GEE approach, which specifies the marginal or the population averaged distribution of the response without explicitly specifying the correlation structure.

A commonly used model for multivariate binary data is the multivariate probit (MVP) model, see for example Chib and Greenberg (1998) and Chen and Dipak (1998). The MVP model is formulated in terms of Gaussian latent variables. Let $\boldsymbol{Z}_i = (z_{i1}, z_{i2}, \cdots, z_{iT})'$ denote a T-variate normal vector with distribution $\boldsymbol{Z}_i \sim \mathcal{N}_T(\boldsymbol{x}_i'\boldsymbol{\beta}, \boldsymbol{R})$ and let Y_{it} be 0 or 1 according to the sign of z_{it}, i.e.

$$y_{it} = \mathrm{I}(z_{it} > 0) \quad (t = 1, \cdots, T),$$

where $\mathrm{I}(A)$ is the indicator function for the event A. According to the MVP model,

$$\begin{aligned} \Pr(\boldsymbol{Y}_i = \boldsymbol{y}_i) &\equiv \Pr(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \cdots, Y_{iT} = y_{iT} | \boldsymbol{\beta}, \boldsymbol{R}, \boldsymbol{x}_i) \\ &= \int_{A_{i1}} \cdots \int_{A_{iT}} \phi_T(\boldsymbol{Z}_i | \boldsymbol{x}_i'\boldsymbol{\beta}, \boldsymbol{R}) d\boldsymbol{Z}_i, \end{aligned} \tag{9.1}$$

where $\phi_T(\boldsymbol{Z}|\boldsymbol{\mu}, \boldsymbol{R})$ is the T-variate normal density with mean $\boldsymbol{\mu}$ and cor-

relation matrix $\boldsymbol{R}$, and A_{it} is the interval $(-\infty, 0)$ when $y_{it} = 0$ and the interval $(0, +\infty)$ when $y_{it} = 1$. Other similar multivariate models include, the multivariate t-link and logistic regression model proposed by Chen and Dipak (1998) and O'Brien and Dunson (2004), respectively.

The GLMM for binary data was introduced in Chapter 4. The model assumes that the random effects $\boldsymbol{u}_i$ follow a distribution G with mean zero. It is usually assumed that $\boldsymbol{u}_i$ follow a multivariate normal distribution with mean $\mathbf{0}$ and variance $\boldsymbol{D}$, i.e. $\boldsymbol{u}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{D})$. Further, it is assumed that, conditionally on $\boldsymbol{u}_i$, the likelihood terms involving the ith cluster are independent. The total (marginal) likelihood is a product of the marginal likelihood for the N clusters, integrating out the random effects, and is equal to (e.g. Stiratelli, Laird, and Ware, 1984)

$$L(\boldsymbol{\beta}, \boldsymbol{D}) = \prod_{i=1}^{N} \int \prod_{t=1}^{T} \boldsymbol{p}_{it}^{Y_{it}} (1 - \boldsymbol{p}_{it})^{Y_{it}} |\boldsymbol{D}|^{-1/2} \exp\left(-\frac{1}{2} \boldsymbol{u}_i' \boldsymbol{D}^{-1} \boldsymbol{u}_i\right) d\boldsymbol{u}_i. \quad (9.2)$$

The model parameters are obtained by maximizing the marginal likelihood (9.2). Paulino et al. (2005) used a random intercept logistic model which is a simplified version of equation (9.2).

Liang and Zeger (1986) introduced the GEE approach as a method of dealing with correlated data when, except for the correlation among responses for each dimension, the data can be modeled as a GLM. The GEE approach models the marginal expectation, specifically it assumes that

$$\Pr(Y_{it} = 1 | \boldsymbol{x}_{it}') = g^{-1}(\boldsymbol{x}_{it}' \boldsymbol{\beta}),$$

where g is the link function. This approach uses a working correlation matrix substituting the true correlation structure among the multiple observations. For a more detailed description of parameter estimation using GEE we refer to Liang and Zeger (1986) and Zeger et al. (1988).

It is known that $\boldsymbol{\beta}$ in a GLMM and a GEE model have a different interpretation. Consider a random intercept logistic regression. In GLMM,

$\boldsymbol{\beta}$ measures the change in conditional logit of the probability of a positive response for an individual conditional on $\boldsymbol{u}_i$. On the other hand, in GEE, $\boldsymbol{\beta}$ measures the change in logit in fraction of the positive response. In other words, regression coefficients in a GLMM measure the change in logit for a subject conditional on the underlying heterogeneity, whereas in GEE the regression coefficients are interpreted marginally, i.e. in reference to the population (see Neuhaus et al., 1991).

Above methods, however, assume that the clustered binary data are measured without error, which may lead to biased results. In the next section we explore these methods in the presence of misclassification errors for the response.

9.2.2 Neuhaus' approach to model misclassified multivariate binary data

In practice we do not observe Y_{it} but rather an error prone version Y_{it}^*. We assume tentatively that the probability of the misclassification of Y_{it}^* only depends on the true response (non-differential assumption). That is,

$$\Pr(Y_{it}^*|Y_{it}, Y_{is}, \boldsymbol{x}_{it}, \boldsymbol{x}_{is}) = \Pr(Y_{it}^*|Y_{it}), \quad (s = 1, 2, \cdots, T, s \neq t).$$

It can be easily shown that

$$\Pr(Y_{it}^* = 1|\boldsymbol{x}_{it}) = \lambda_0 + (1 - \lambda_0 - \lambda_1)g^{-1}(\boldsymbol{x}_{it}'\boldsymbol{\beta}),$$

where $\lambda_0 = \Pr(Y_{it}^* = 1|Y_{it} = 0)$ (1 − specificity) and $\lambda_1 = \Pr(Y_{it}^* = 0|Y_{it} = 1)$ (1 − sensitivity), so that if Y_{it} follows a GLM with link function g, Y_{it}^* will follow a GLM with a modified link function g^* (see Neuhaus, 2002). In particular,

$$g^*\left\{\Pr(Y_{it} = 1|\boldsymbol{x}_{it})\right\} = g\left\{\frac{\Pr(Y_{it} = 1|\boldsymbol{x}_{it}) - \lambda_0}{1 - \lambda_0 - \lambda_1}\right\}. \tag{9.3}$$

Thus, Neuhaus (2002) concluded that the observed response Y_{it}^* can be analyzed by a GEE method with link function g^*. He also concluded for a GLMM that the observed response will also follow a GLMM with the link g^* assuming that the misclassification probabilities do not depend on the random effects.

The corrected GEE and GLMM above, however, are based on some simplifying assumptions. Applied to caries research, Neuhaus' assumptions imply that

Assumption A1: scoring caries experience (response) is the same for all teeth, i.e. the misclassification rates (sensitivities and specificities) are the same for all teeth.

Assumption A2: scoring teeth caries experience occurs independently. That is, there is no association in the sensitivities and specificities for the different teeth in the mouth.

Assumption A1 is probably not valid for scoring caries experience since scoring of mandibular teeth is more difficult than of maxillary teeth as the former involves indirect sight (using a dental mirror). Further, Hujoel et al. (1990) pointed out that site-specific diagnostic tests within a subject are often dependent, implying that assumption A2 probably does not hold for scoring caries experience in the mouth.

Figure 9.1 shows that the pooled (over-teeth) estimator of the sensitivity and specificity may be inappropriate to describe the misclassification process in the Signal Tandmobiel® study. Indeed, the point-wise 95% credible intervals do not always cover the pooled point estimates of the tooth-specific sensitivity and specificity. Hence, there is a need for a tooth-specific correction for misclassification in a tooth-level analysis.

In this chapter we describe a general approach to misclassified multivariate binary data, which allows for response dependent misclassification parameters, i.e. relaxing assumption A1. In the discussion (Section 9.6), we will also provide suggestions for relaxing assumption A2 and we will

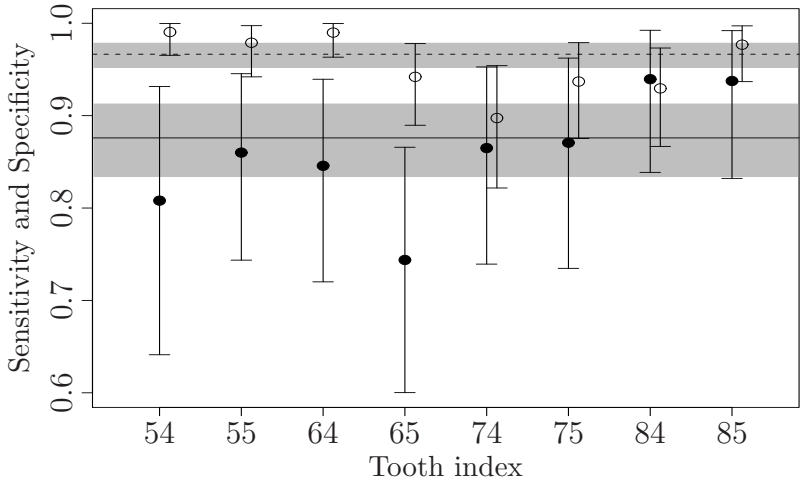


Figure 9.1: *Tooth-specific sensitivity (•) and tooth-specific specificity (∘) (± 95% credible interval lines) of the $dmft_{4,5}$-index. The dashed and the solid line show the mean of the pooled specificity and the pooled sensitivity over all teeth, respectively. The gray band is the 95% credible interval for the pooled estimate.*

address the problem when A2 is violated.

9.3 A general approach to analyse misclassified multivariate binary data

Below we develop a model which relaxes assumption A1. Define the tooth-specific misclassification rates as $\lambda_{0t} = \Pr(Y_{it}^* = 1|Y_{it} = 0)$ (1 − specificity of the tth response) and $\lambda_{1t} = \Pr(Y_{it}^* = 0|Y_{it} = 1)$ (1 − sensitivity of the tth response) for $t = 1, \cdots, T$. Further, let p_{it} be the success probability of the Bernoulli variable Y_{it}. We know that $\Pr(Y_{it}^* = 1|p_{it}, \lambda_{0t}, \lambda_{1t}) = \lambda_{0t} + (1 - \lambda_{0t} - \lambda_{1t})p_{it}$. The success probability p_{it} can be allowed to vary as a function of covariates, e.g.,

$$p_{it} \equiv p_{it}(\boldsymbol{\beta}) = g^{-1}(\boldsymbol{x}_{it}'\boldsymbol{\beta}).$$

The multivariate binary response $\boldsymbol{Y}_i = (Y_{i1}, Y_{i1}, \cdots, Y_{iT})'$ can take 2^T possible response values. In the case of a univariate binary response, the observed data Y_i^* are expressed in terms of the true data Y_i by

$$\Pr(Y_i^* = y_i^* | \boldsymbol{x}_i) = \sum_{y_i} \Pr(Y_i^* = y_i^* | Y_i = y_i) \Pr(Y_i = y_i | \boldsymbol{x}_i).$$

We can use here a similar representation for the multivariate binary responses. That is,

$$\begin{aligned}
&\Pr(Y_{i1}^* = y_{i1}^*, Y_{i2}^* = y_{i2}^*, \cdots, Y_{iT}^* = y_{iT}^* | \boldsymbol{x}_i) \\
&\quad = \sum_{y_{i1}} \sum_{y_{i2}} \cdots \sum_{y_{iT}} \Pr(Y_{i1}^* = y_{i1}^*, Y_{i2}^* = y_{i2}^*, \cdots, Y_{iT}^* = y_{iT}^* | Y_{i1} = y_{i1}, \\
&\qquad Y_{i2} = y_{i2}, \cdots, Y_{iT} = y_{iT}) \Pr(Y_{i1} = y_{i1}, Y_{i2} = y_{i2} \cdots, Y_{iT} = y_{iT} | \boldsymbol{x}_i) \\
&\quad \text{(Assuming A2)} \\
&\quad = \sum_{y_{i1}} \sum_{y_{i2}} \cdots \sum_{y_{iT}} \left(\prod_{t=1}^{T} \Pr(Y_{it}^* = y_{it}^* | Y_{it} = y_{it}) \right) \\
&\qquad \times \Pr(Y_{i1} = y_{i1}, Y_{i2} = y_{i2} \cdots, Y_{iT} = y_{iT} | \boldsymbol{x}_i) \qquad (9.4)
\end{aligned}$$

Expression (9.4) involves 2^T evaluations for each subject. With a moderate to large value of T, calculating the resulting likelihood will thus be computationally intensive. For this reason we suggest below a different approach assuming a latent random variable thereby involving the data augmentation algorithm. In this approach one can use the EM algorithm or a Bayesian method for parameter estimation. Here we used a Monte Carlo version of the EM algorithm.

The idea of data augmentation to correct for measurement error in covariates was introduced by Kuha (1997). He suggested the use of imputation to estimate the parameters of a regression model where some of the covariates are subject to measurement error. Further, Rekaya, Weigel, and Gianola (2001) suggested a Bayesian data augmentation approach to misclassified binary response, but with a simplified misclassification struc-

ture, i.e. with sensitivity equal to specificity. We extend the ideas of data augmentation to a more complex problem of modeling misclassified multivariate binary responses in which sensitivity and specificity are allowed to vary with the response.

Assume that the latent variable W_{it} for observation "it" is defined as $W_{it} = 1$ if Y_{it} is misclassified and 0 otherwise. W_{it} is the missing information linking Y_{it}^* and Y_{it}. The use of W_{it} simplifies the model formulation by expressing Y_{it} as a function of Y_{it}^* and W_{it} through the relation

$$Y_{it} \;=\; (1 - W_{it})Y_{it}^* \;+\; W_{it}(1 - Y_{it}^*), \tag{9.5}$$

so that $Y_{it} \;=\; Y_{it}^*$ when there is no misclassification. W_{it} is a Bernoulli variable with success probability

$$\Pr(W_{it} = 1|Y_{it}) = \begin{cases} \lambda_{0t} & \text{if } Y_{it} = 0, \\ \lambda_{1t} & \text{if } Y_{it} = 1. \end{cases} \tag{9.6}$$

The joint probability distribution of $\boldsymbol{W}$ and $\boldsymbol{Y}$, given $\boldsymbol{\beta}$, is

$$\Pr(\boldsymbol{W}, \boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{x}) \;= \\ \prod_{i=1}^{N}\prod_{t=1}^{T} \Pr(W_{it} = 1|Y_{it})^{W_{it}} \left[1 - \Pr(W_{it} = 1|Y_{it})\right]^{1-W_{it}} p_{it}^{Y_{it}} \left[1 - p_{it}\right]^{1-Y_{it}},$$

where $\boldsymbol{W} = (\boldsymbol{W}'_1, \cdots, \boldsymbol{W}'_N)'$ with $\boldsymbol{W}_i = (W_{i1}, \cdots, W_{iT})'$. The joint probability distribution of W_{it} and Y_{it}^*, given $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ is

$$\begin{aligned} \Pr(W_{it}, Y_{it}^*|\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{x}) \;&\equiv\; \Pr(W_{it}, Y_{it} = [1 - W_{it}]Y_{it}^* + W_{it}[1 - Y_{it}^*]|\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{x}) \\ &=\; \Pr(W_{it} = 1|Y_{it})^{W_{it}} \left[1 - \Pr(W_{it} = 1|Y_{it})\right]^{1-W_{it}} \times \\ &\quad p_{it}^{(1-W_{it})Y_{it}^* + W_{it}(1-Y_{it}^*)} \left[1 - p_{it}\right]^{1-(1-W_{it})Y_{it}^* - W_{it}(1-Y_{it}^*)}. \end{aligned} \tag{9.7}$$

With the latent (unobserved) variable $\boldsymbol{W}$, we can use the data augmentation algorithm (Tanner and Wong, 1987) for parameter estimation. The

algorithm essentially consists of two (iterative) steps, namely (a) the $\mathcal{S}$-step (simulation): the latent variables and hence the true data are sampled; and (b) the $\mathcal{P}$-step (posterior): based on the true data generated from the $\mathcal{S}$-step new estimates of the parameters are obtained. The $\mathcal{S}$-step obtains the conditional expected value of $\boldsymbol{W}$ given $(\boldsymbol{Y}^*, \boldsymbol{x})$, given the current value of the other parameters $(\hat{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\lambda}})$, where $\hat{\boldsymbol{\beta}}$ is the current estimate of $\boldsymbol{\beta}$ and $\widetilde{\boldsymbol{\lambda}}$ is the value of the misclassification rates obtained, say, from validation data. That is, in the $\mathcal{S}$-step the following is calculated:

$$\omega \equiv \mathrm{E}[W|Y^*, \boldsymbol{x}, \widehat{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\lambda}}] = \Pr(W = 1|Y^*, \boldsymbol{x}, \widehat{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\lambda}}). \tag{9.8}$$

From expressions (9.6) and (9.7), we have:

$$\Pr(W_{it} = 1|Y_{it}^*, \cdot) = \frac{\Pr(W_{it} = 1, Y_{it}^*|\cdot)}{\Pr(Y_{it}^*|\cdot)} \tag{9.9}$$

$$= \frac{\Pr(W_{it} = 1|Y_{it}) p_{it}^{1-Y_{it}^*}(1-p_{it})^{Y_{it}^*}}{\Pr(W_{it} = 1|Y_{it}) p_{it}^{1-Y_{it}^*}(1-p_{it})^{Y_{it}^*} + \Pr(W_{it} = 0|Y_{it}) p_{it}^{Y_{it}^*}(1-p_{it})^{1-Y_{it}^*}}$$

$$= \frac{\lambda_{0t}(1-p_{it})}{\lambda_{0t}(1-p_{it}) + (1-\lambda_{1t})p_{it}} Y_{it}^* + \frac{\lambda_{1t} p_{it}}{(1-\lambda_{0t})(1-p_{it}) + \lambda_{1t} p_{it}}(1 - Y_{it}^*).$$

To obtain the estimate of $\boldsymbol{\beta}$, start with an initial estimate $\boldsymbol{\beta}^{(0)}$ say, from a naive analysis. Given $\widehat{\boldsymbol{\beta}}^{(k)}$ of $\boldsymbol{\beta}$ at iteration k we then proceed as follows. For iteration $(k+1)$:

$\mathcal{S}$-step: Draw $W_{it}^{(k+1)}$ from a Bernoulli distribution with success probability $\Pr(W_{it} = 1|Y_{it}^*, \boldsymbol{x}_{it}, \widehat{\boldsymbol{\beta}}^{(k)}, \widetilde{\boldsymbol{\lambda}})$ based on (9.9) for $(t = 1, \cdots, T)$ and $(i = 1, \cdots, N)$ and determine $\boldsymbol{Y}^{(k+1)}$ from (9.5).

$\mathcal{P}$-step: Apply a multivariate binary regression model on $\boldsymbol{Y}^{(k+1)}$ to obtain a new estimate $\widehat{\boldsymbol{\beta}}^{(k+1)}$.

The $\mathcal{S}$- and $\mathcal{P}$-steps are alternated until convergence. In the $\mathcal{P}$-step, regression models for correlated binary data such as the multivariate-logistic, -t link and -probit regression models can be applied. We observe that the

$\mathcal{P}$-step is the most time consuming step of this approach since it involves fitting a regression model for the correlated binary data in every iteration.

The variances and covariances of the estimated regression coefficients are obtained from the inverse of the information matrix, $\boldsymbol{I}_{\boldsymbol{Y}}(\widehat{\boldsymbol{\beta}})$. It can be shown that for the misclassified binary regression (e.g., Louis, 1982; Magder and Hughes, 1997)

$$\begin{aligned}\boldsymbol{I}_{\boldsymbol{Y}}(\widehat{\boldsymbol{\beta}}) &= \boldsymbol{I}_{\boldsymbol{Y}^*}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{I}_{\boldsymbol{Y}^*|\boldsymbol{Y}}(\widehat{\boldsymbol{\beta}}) \\ &= \sum_i \boldsymbol{X}_i' \boldsymbol{V}_i \boldsymbol{X}_i, \end{aligned} \tag{9.10}$$

where $\boldsymbol{X}_i$ is the $T \times d$ regression matrix for the ith individual. For the logit link function $\boldsymbol{V}_i$ is a $T \times T$ diagonal matrix with general element $[p_{it}(1-p_{it}) - \omega_{it}(1-\omega_{it})]$. For the probit link function, the variances and covariances could be evaluated as above but using the logit–probit relation: $\boldsymbol{x}_i'\boldsymbol{\beta}$ (from logit) $\simeq 0.627\boldsymbol{x}_i'\boldsymbol{\beta}$ (from probit), for a detailed description, see Genton (2004, p. 135). Alternatively, these (co)variances could be calculated by taking the second order derivatives of the log-likelihood now assuming the probit model.

Further, the covariance matrix of the responses can be directly obtained from the naive estimate of the responses' covariance. We know that $\Pr(Y_{it}^* = r) = \lambda_{rt} + (1 - \lambda_{0t} - \lambda_{1t}) \Pr(Y_{it} = r)$ for $r = 0, 1$. Hence, it immediately follows that the covariance between Y_{it} and Y_{is}, $t \neq s$, is (e.g. Neuhaus, 2002)

$$\begin{aligned}\operatorname{Cov}(Y_{it}^*, Y_{is}^*) &= \operatorname{Cov}\left[(1 - \lambda_{0t} - \lambda_{1t})Y_{it},\, (1 - \lambda_{0s} - \lambda_{1s})Y_{is}\right] \\ &= (1 - \lambda_{0t} - \lambda_{1t})\,(1 - \lambda_{0s} - \lambda_{1s})\operatorname{Cov}(Y_{it}, Y_{is}), \quad t \neq s.\end{aligned}$$

$$\text{Hence, } \operatorname{Cov}(Y_{it}, Y_{is}) = \frac{1}{(1 - \lambda_{0t} - \lambda_{1t})\,(1 - \lambda_{0s} - \lambda_{1s})}\operatorname{Cov}(Y_{it}^*, Y_{is}^*), \quad t \neq s.$$

Correspondingly, the estimator of the corrected covariance is given by

$$\widehat{\text{Cov}}(Y_{it},\, Y_{is}) = \frac{1}{\left(1-\tilde{\lambda}_{0t}-\tilde{\lambda}_{1t}\right)\left(1-\tilde{\lambda}_{0s}-\tilde{\lambda}_{1s}\right)}\widehat{\text{Cov}}(Y^*_{it}, Y^*_{is}),$$

where $\widehat{\text{Cov}}(Y^*_{it}, Y^*_{is})$ is the naive estimator of the covariance between Y^*_{it} and Y^*_{is}, $\tilde{\lambda}_{0t}$ and $\tilde{\lambda}_{1t}$ are respectively the estimates of the specificity and sensitivity for tooth t. Indeed, $\widehat{\text{Cov}}(Y_{it}, Y_{is}) = \widehat{\text{Cov}}(Y^*_{it}, Y^*_{is})$ when there is no misclassification (i.e. $\tilde{\lambda}_{0t} = \tilde{\lambda}_{1t} = 0$ for all ts). To fully take account of all uncertainty one needs to incorporate also the variability with which λ_{0t} and λ_{1t} are estimated, as indicated in the previous chapters.

Our approach will be referred below to as MC-DA (correction for misclassification using data augmentation). In principle, a MC-DA method for relaxing the second assumption A2, can be developed. However, this extension to relax assumption A2 is a topic for further research, see also Section 9.6.

In the next section we perform a simulation study to assess the performance of our approach as compared to existing methods.

9.4 Simulation Study

9.4.1 Simulation study set up

The MC-DA procedure is evaluated for different scenarios with $T = 4$ binary responses. We considered the following cases:

A. equal misclassification with $\boldsymbol{\lambda}'_0 = (0.9, 0.9, 0.9, 0.9)$ and $\boldsymbol{\lambda}'_1 = (0.8, 0.8, 0.8, 0.8)$;

B. moderately unequal misclassification with $\boldsymbol{\lambda}'_0 = (0.9, 0.7, 0.7, 0.8)$ and $\boldsymbol{\lambda}'_1 = (0.8, 0.8, 0.75, 0.75)$. This scenario is motivated by the observed sensitivities and specificities in the Signal Tandmobiel® validation data;

C. large(r) unequal misclassification with $\boldsymbol{\lambda}_0' = (0.7, 0.8, 0.6, 0.9)$ and $\boldsymbol{\lambda}_1' = (0.6, 0.8, 0.9, 0.7)$.

It is important to note, however, that in the simulations study the misclassification probabilities were not estimated from a validation study, but were directly imputed into the MC-DA procedure. Thus, we assumed that λ_{0t} and λ_{1t} $(t = 1, \cdots, 4)$ are known.

We performed 200 simulations each time with a sample size of 1000. This sample size is motivated by the fact that correction for misclassification is usually done for large epidemiological studies. We generated the true binary responses Y_is through latent variables $\boldsymbol{Z}_i$s such that $Y_{it} = 1\,(Y_{it} = 0)$ if $Z_{it} > 1\,(Z_{it} \leq 0)$ with $\boldsymbol{Z}_i = (Z_{i1}, \cdots, Z_{i4})' \sim N(\boldsymbol{x}_i'\boldsymbol{\beta}, \boldsymbol{R})$, where $\boldsymbol{\beta}$ is a vector of the regression coefficient of $\boldsymbol{x}_i$ and $\boldsymbol{R}$ is a 4×4 correlation matrix. In particular, we have set

$$\boldsymbol{\beta} = (-2.00, 0.35, 0.45, -1.50, 0.32, 0.48, -1.50, 0.28, 0.52, -2.00, 0.25, 0.55)'$$

comprising the intercept and the coefficients of the simulated covariates X_{i1} and X_{i2}, respectively, for each of the four response variables, where $X_{i1} \sim \mathcal{N}(7, 0.4)$ and $X_{i2} \sim$ Bernoulli(0.6). The simulated covariates mimic the *age* and *gender* in our Signal Tandmobiel® application. The values of the regression coefficients have been chosen arbitrarily. We chose the correlation matrix with correlation equal to 0.6 among all responses.

9.4.2 Simulation results of MC-DA method

The simulation results indicate that our method leads to considerable reduction of bias compared to the naive estimator, as can be seen in Tables 9.1, 9.2 and 9.3. The point estimates remain fairly stable irrespective of the choice of misclassification probabilities. Thus the MC-DA method gives improved estimates even under high misclassification probability. However, the standard errors increase as misclassification gets worse, i.e. from scenario A to C.

Table 9.1: *Simulation results for Case A: Multivariate probit regression of a misclassified response Y on a continuous (X_1) and binary (X_2) covariates with equal misclassification rates, namely $\boldsymbol{\lambda}_0' = (0.9, 0.9, 0.9, 0.9)$ & $\boldsymbol{\lambda}_1' = (0.8, 0.8, 0.8, 0.8)$. The results are based 200 simulations each with sample size = 1000.*

					Corrected					
			Naive		Neuhaus			MC-DA		
Par[a]	Res[b]	True	Estimate (SE)	CR[‡]	Estimate (SE)	RMSE[†]	CR[‡]	Estimate (SE)	RMSE[†]	CR[‡]
β_0	1	−2.00	−1.204(0.718)	81	−1.947(1.272)	1.273	96	−1.925(1.436)	1.438	97
	2	−1.50	−0.730(0.741)	82	−1.536(1.575)	1.575	94	−1.657(1.617)	1.625	96
	3	−1.50	−0.928(0.720)	87	−1.661(1.317)	1.327	96	−1.379(1.473)	1.478	98
	4	−2.00	−1.446(0.707)	88	−1.924(1.043)	1.046	94	−2.082(1.194)	1.197	96
β_{X_1}	1	0.35	0.198(0.102)	70	0.344(0.182)	0.182	96	0.340(0.219)	0.219	97
	2	0.32	0.156(0.106)	63	0.325(0.226)	0.226	94	0.345(0.286)	0.287	96
	3	0.28	0.159(0.103)	76	0.302(0.189)	0.190	96	0.263(0.232)	0.233	97
	4	0.25	0.163(0.101)	87	0.238(0.148)	0.148	94	0.261(0.175)	0.175	96
β_{X_2}	1	0.45	0.254(0.083)	33	0.445(0.145)	0.145	96	0.446(0.170)	0.170	97
	2	0.48	0.222(0.085)	13	0.513(0.187)	0.190	92	0.515(0.237)	0.240	97
	3	0.52	0.295(0.083)	22	0.540(0.150)	0.151	96	0.538(0.178)	0.179	95
	4	0.55	0.386(0.082)	47	0.571(0.119)	0.121	92	0.555(0.130)	0.130	96

[a]Par = Parameter.
[b]Res = Response variable.
[†]RMSE = Root mean square error.
[‡]CR = Coverage rate (%), based on SE.

Table 9.2: *Simulation results for Case B: Multivariate probit regression of a misclassified response Y on a continuous (X_1) and binary (X_2) covariates with moderately unequal misclassification rates, namely $\boldsymbol{\lambda}_0' = (0.9, 0.7, 0.7, 0.8)$ and $\boldsymbol{\lambda}_1' = (0.8, 0.8, 0.75, 0.75)$. The results are based 200 simulations each with sample size = 1000.*

					Corrected					
			Naive		Neuhaus			MC-DA		
Par[a]	Res[b]	True	Estimate (SE)	CR[‡]	Estimate (SE)	RMSE[†]	CR[‡]	Estimate (SE)	RMSE[†]	CR[‡]
β_0	1	−2.00	−1.097(0.719)	77	−1.947(1.257)	1.258	96	−2.005(1.494)	1.494	96
	2	−1.50	−0.086(0.761)	53	−1.457(2.250)	2.250	93	−1.354(2.369)	2.373	96
	3	−1.50	−0.231(0.717)	59	−1.821(2.262)	2.285	93	−1.452(2.222)	2.223	96
	4	−2.00	−1.117(0.703)	76	−1.905(1.325)	1.328	98	−2.127(1.412)	1.418	98
β_{X_1}	1	0.35	0.185(0.102)	65	0.344(0.179)	0.179	96	0.352(0.184)	0.184	95
	2	0.32	0.081(0.108)	38	0.315(0.323)	0.323	93	0.333(0.330)	0.330	95
	3	0.28	0.070(0.102)	49	0.327(0.325)	0.328	93	0.290(0.326)	0.326	94
	4	0.25	0.130(0.100)	77	0.236(0.186)	0.187	98	0.269(0.209)	0.210	97
β_{X_2}	1	0.45	0.240(0.084)	26	0.445(0.148)	0.148	96	0.451(0.158)	0.158	97
	2	0.48	0.162(0.085)	2	0.529(0.315)	0.319	92	0.508(0.313)	0.314	98
	3	0.52	0.184(0.083)	2	0.545(0.238)	0.239	96	0.531(0.241)	0.241	96
	4	0.55	0.308(0.082)	17	0.568(0.160)	0.161	94	0.553(0.161)	0.161	95

[a]Par = Parameter.
[b]Res = Response variable.
[†]RMSE = Root mean square error.
[‡]CR = Coverage rate (%), based on SE.

Table 9.3: *Simulation results for Case C: Multivariate probit regression of a misclassified response Y on a continuous (X_1) and binary (X_2) covariates with large(r) unequal misclassification rates, namely $\boldsymbol{\lambda}_0' = (0.7, 0.8, 0.6, 0.9)$ and $\boldsymbol{\lambda}_1' = (0.6, 0.8, 0.9, 0.7)$. The results are based 200 simulations each with sample size = 1000.*

					Corrected					
			Naive		Neuhaus			MC-DA		
Par[a]	Res[b]	True	Estimate (SE)	CR[‡]	Estimate (SE)	RMSE[†]	CR[‡]	Estimate (SE)	RMSE[†]	CR[‡]
β_0	1	−2.00	−0.712(0.697)	60	−1.182(1.355)	1.583	90	−2.040(1.539)	1.540	97
	2	−1.50	−0.489(0.703)	70	−0.826(2.882)	2.960	99	−1.165(2.752)	2.772	98
	3	−1.50	−0.372(0.776)	71	−0.747(1.183)	1.402	88	−1.513(1.569)	1.569	96
	4	−2.00	−1.538(0.686)	92	−1.530(1.035)	1.137	32	−2.124(1.378)	1.384	96
β_{X_1}	1	0.35	0.151(0.099)	53	0.289(0.195)	0.204	63	0.354(0.203)	0.203	97
	2	0.32	0.073(0.100)	30	0.267(0.410)	0.413	71	0.340(0.397)	0.397	95
	3	0.28	0.145(0.111)	77	0.236(0.169)	0.175	52	0.287(0.171)	0.172	97
	4	0.25	0.161(0.098)	88	0.213(0.146)	0.151	46	0.256(0.180)	0.181	96
β_{X_2}	1	0.45	0.202(0.085)	17	0.359(0.180)	0.202	53	0.454(0.198)	0.198	96
	2	0.48	0.118(0.080)	1	0.561(0.543)	0.549	69	0.511(0.454)	0.455	94
	3	0.52	0.268(0.084)	20	0.436(0.155)	0.176	44	0.533(0.190)	0.191	97
	4	0.55	0.338(0.086)	27	0.489(0.136)	0.149	38	0.554(0.138)	0.138	97

[a]Par = Parameter.
[b]Res = Response variable.
[†]RMSE = Root mean square error.
[‡]CR = Coverage rate (%), based on SE.

9.4.3 Comparison to Neuhaus' method

The performance of MC-DA method is compared to the approach of Neuhaus (2002), both for the GEE approach, with regard to bias, root mean square error (RMSE), and coverage rate. Under equal misclassification probabilities (Scenario A), both methods perform equally good in terms of bias and coverage rate, as shown in Table 9.1. However, the RMSEs of the MC-DA method are larger.

For moderately unequal misclassification probabilities (Scenario B) the MC-DA method performs better in terms of bias reduction as can be seen in Table 9.2. The coverage rates and RMSEs are comparable between the two methods, but the MC-DA estimates are somewhat larger.

Results from the extreme situation (Scenario C) in terms of misclassification probabilities are shown in Table 9.3. The Neuhaus' method gives biased results compared to MC-DA method, and has also poor coverage rate, though, the RMSEs are comparable between the two methods.

As a general conclusion, we suggest that the MC-DA method can notably reduce the bias compared to the naive estimator and its performance is comparable to Neuhaus' method estimation under unequal misclassification process. Further, the MC-DA method has shown to perform better in terms of bias than Neuhaus' approach when the misclassification probabilities are very unequal, though the two methods are equally good in terms of RMSE.

9.5 Application to the Signal Tandmobiel® study

We are now interested in modeling on tooth level. In particular, we are interested in the comparison of the effect of the covariates on left versus right, upper versus lower, and first versus second primary molars. The corrected models that were fitted were pooled over the examiners due to computational difficulties arising from the sparseness of the validation data. We shall thus not focus here on the interpretation of the (x, y) geographical

coordinates, though, we keep these components in the analysis.

We considered two corrections: (a) Neuhaus' correction with GEE approach, where the sensitivity and specificity are pooled over the teeth; and (b) MC-DA correction with GEE regression model, where sensitivity and specificity are tooth-specific. The GEE estimation was done using the SAS (Version 9.1) procedure `GENMOD`.

Results from tooth-level regression analyses both with and without correction for misclassification are displayed in Table 9.4. Overall, the corrected estimates are larger in absolute value than the corresponding estimates from the naive regression model. This shows a gain in the parameter estimates as they are pulled away from the null. However, the standard errors of the estimates are increased as a result of misclassification errors. While the point estimates are similar in the Neuhaus' approach and MC-DA method, the latter shows larger standard errors.

After correction for misclassification the following covariates were significantly (overall) associated with the risk of caries (Table 9.4): age of the child, use of systemic fluoride supplements, age at the start of brushing, and consumption of sugar containing drinks. On the other hand, frequency of brushing, and daily intake of at least two in-between-meals were not significant. Below we give an interpretation of the results on tooth-level.

The coefficients of age at the start of brushing are, in general, larger for the right molars than the left molars. The positive coefficients for age at start of brushing implies that the later the child starts brushing the higher the probability of presenting caries, but this probability is somewhat higher for the right molars. However, this effect is not statistically significantly different (Wald Chi-square = 6.56, df=4; $P = 0.161$). For a description of the Wald Chi-square statistic, see Appendix A.4. The other risk factors did not show an appreciable difference between the left and the right molars.

In relation to mandibular and maxilla teeth, the results indicate that the coefficients of age at the start of brushing are smaller for the mandibular than for the maxilla teeth. This implies that the later the child start

Table 9.4: *Parameter estimates from Neuhaus & MC-DA corrected GEE regression analysis of caries experience controlling for the tooth-dependent covariates effect.*

		Naive	Corrected	
		GEE	Neuhaus (GEE)	MC-DA (GEE)
Parameter	Tooth	Estimate (SE)	Estimate (SE)	Estimate (SE)
Intercept	54	−1.205(0.083)	−1.281(0.105)	−1.074(0.132)
	55	−0.976(0.079)	−1.007(0.097)	−0.974(0.132)
	64	−1.152(0.084)	−1.225(0.103)	−1.082(0.130)
	65	−0.973(0.081)	−1.006(0.097)	−1.285(0.154)
	74	−0.947(0.076)	−0.968(0.093)	−2.119(0.155)
	75	−1.110(0.079)	−1.167(0.097)	−1.496(0.141)
	84	−0.960(0.076)	−0.986(0.093)	−1.272(0.135)
	85	−0.922(0.078)	−0.938(0.094)	−0.938(0.126)
x-ordinate	54	0.089(0.027)	0.105(0.032)	0.105(0.041)
	55	0.081(0.025)	0.096(0.029)	0.092(0.041)
	64	0.117(0.026)	0.140(0.031)	0.126(0.040)
	65	0.095(0.025)	0.112(0.029)	0.166(0.048)
	74	0.092(0.024)	0.107(0.028)	0.226(0.048)
	75	0.135(0.025)	0.160(0.029)	0.164(0.044)
	84	0.119(0.024)	0.140(0.028)	0.155(0.042)
	85	0.144(0.025)	0.169(0.029)	0.145(0.039)
y-ordinate	54	−0.019(0.026)	−0.023(0.033)	−0.025(0.041)
	55	−0.008(0.025)	−0.008(0.030)	−0.019(0.041)
	64	−0.022(0.026)	−0.027(0.032)	−0.010(0.041)
	65	−0.022(0.025)	−0.027(0.030)	−0.067(0.048)
	74	−0.005(0.024)	−0.006(0.029)	−0.024(0.048)
	75	−0.042(0.025)	−0.047(0.030)	−0.058(0.044)
	84	−0.002(0.024)	−0.001(0.029)	−0.018(0.042)
	85	−0.040(0.024)	−0.046(0.029)	−0.035(0.039)
Gender (girl)	54	0.025(0.051)	0.031(0.062)	0.019(0.080)
	55	0.012(0.049)	0.014(0.057)	0.035(0.079)
	64	0.045(0.051)	0.060(0.061)	0.092(0.079)
	65	−0.002(0.049)	−0.001(0.057)	0.008(0.093)
	74	−0.003(0.047)	−0.004(0.055)	−0.236(0.094)
	75	0.127(0.048)	0.152(0.057)	0.290(0.085)
	84	−0.018(0.047)	−0.021(0.055)	0.008(0.082)
	85	0.068(0.048)	0.078(0.056)	0.086(0.076)
Age (years)	54	0.256(0.062)	0.306(0.078)	0.182(0.099)
	55	0.186(0.059)	0.217(0.071)	0.222(0.099)
	64	0.242(0.063)	0.294(0.077)	0.214(0.097)
	65	0.217(0.060)	0.258(0.072)	0.242(0.115)
	74	0.167(0.058)	0.196(0.068)	0.180(0.117)
	75	0.110(0.059)	0.131(0.071)	0.110(0.105)
	84	0.162(0.058)	0.191(0.068)	0.169(0.102)
	85	0.221(0.058)	0.257(0.070)	0.186(0.094)

Continued on next page

Table 9.4: ***continued from previous page***

Parameter	Tooth	Naive GEE Estimate (SE)	Corrected Neuhaus (GEE) Estimate (SE)	Corrected MC-DA (GEE) Estimate (SE)
Brushing frequency (< 2)	54	0.039(0.076)	0.045(0.088)	0.022(0.117)
	55	0.094(0.071)	0.109(0.081)	0.047(0.117)
	64	0.003(0.076)	0.000(0.088)	0.015(0.116)
	65	0.107(0.071)	0.125(0.081)	0.287(0.139)
	74	−0.059(0.070)	−0.069(0.080)	−0.066(0.138)
	75	0.087(0.070)	0.101(0.081)	0.097(0.126)
	84	0.105(0.068)	0.123(0.078)	0.140(0.119)
	85	0.125(0.068)	0.147(0.080)	0.134(0.111)
Age start brushing (years)	54	0.067(0.024)	0.079(0.029)	0.067(0.038)
	55	0.054(0.022)	0.063(0.027)	0.060(0.038)
	64	0.090(0.024)	0.107(0.028)	0.094(0.037)
	65	0.084(0.023)	0.098(0.027)	0.105(0.045)
	74	0.110(0.022)	0.127(0.026)	0.100(0.045)
	75	0.146(0.022)	0.171(0.027)	0.191(0.040)
	84	0.113(0.021)	0.132(0.026)	0.146(0.038)
	85	0.073(0.022)	0.085(0.026)	0.069(0.036)
Systemic fluoride (yes)	54	−0.172(0.052)	−0.207(0.064)	−0.183(0.081)
	55	−0.179(0.049)	−0.212(0.059)	−0.220(0.081)
	64	−0.244(0.052)	−0.295(0.063)	−0.240(0.080)
	65	−0.242(0.050)	−0.286(0.059)	−0.457(0.094)
	74	−0.180(0.048)	−0.209(0.056)	−0.368(0.095)
	75	−0.303(0.049)	−0.359(0.058)	−0.492(0.087)
	84	−0.195(0.048)	−0.228(0.056)	−0.277(0.083)
	85	−0.277(0.049)	−0.324(0.058)	−0.307(0.077)
Sugary drinks (yes)	54	0.242(0.054)	0.292(0.067)	0.257(0.082)
	55	0.249(0.051)	0.293(0.061)	0.278(0.082)
	64	0.155(0.053)	0.187(0.065)	0.177(0.081)
	65	0.218(0.051)	0.259(0.061)	0.320(0.095)
	74	0.248(0.049)	0.289(0.058)	0.335(0.099)
	75	0.198(0.050)	0.235(0.060)	0.356(0.088)
	84	0.254(0.049)	0.298(0.058)	0.328(0.085)
	85	0.187(0.050)	0.217(0.059)	0.198(0.078)
Between meals (> 2)	54	0.106(0.055)	0.121(0.066)	0.068(0.086)
	55	0.055(0.052)	0.064(0.061)	0.062(0.086)
	64	0.134(0.054)	0.158(0.065)	0.127(0.085)
	65	0.001(0.052)	0.000(0.062)	0.083(0.101)
	74	0.082(0.051)	0.092(0.059)	0.044(0.102)
	75	0.025(0.052)	0.029(0.061)	0.003(0.092)
	84	0.151(0.050)	0.174(0.059)	0.194(0.088)
	85	0.092(0.051)	0.107(0.060)	0.106(0.082)

brushing the higher the risk of caries attack on the maxilla molars than in the mandibular molars. Again, the effect of age at start brushing is not statistically significantly different (Wald Chi-square = 7.291 df=4; $P = 0.121$). In addition, use of systemic fluoride supplements appears to be more protective for the maxilla molars than for the mandibular molars. However, again this different effect is shown to be not statistically significant (Wald Chi-square = 9.10, df=4; $P = 0.058$).

For the first and second molars, use of systemic fluoride supplements appears to be more protective for the second than for the first molars against caries attack. This observation is however not statistically significant with Wald Chi-square = 4.22 (df=4, $P = 0.376$).

9.6 Discussion

We have proposed a general and flexible approach for parameter estimation in correlated binary regression models in the presence of misclassification. This approach is simple in that it can be fitted using the standard software such as SAS and R.

The interesting feature of this approach is that we can easily relax the assumption of equal misclassification probabilities in scoring caries experience (A1). In our simulation study the MC-DA method has shown good results in terms of bias reduction but comparable to the approach of Neuhaus (2002) in terms of RMSE for unequal misclassification probabilities. However, our method, being an iterative procedure, is very computer intensive; for example in application to Signal Tandmobiel® the Neuhaus' approach took about 1 hour, while our procedure took over 36 hours. Thus the Neuhaus' approach is appropriate for GEE (and GLLM) regression models under equal or moderately unequal misclassification probabilities.

Our method, though computer intensive, can be used for extreme misclassification process since it has been shown to outperform the Neuhaus' approach as misclassification gets worse. The major advantage of our

method to the Neuhaus' approach is that it is robust to the choice of the correlated binary regression model. In particular, the approach of Neuhaus is only limited to the GEE and GLMM methods, whereas the MC-DA procedure can be applied, in addition to the GEE and GLMM methods, with any multivariate binary regression model, for example, the MVP, multivariate-t and logistic regression. Thus our method is more general since it allows for a complex and variety of multivariate binary regression model at the posterior step.

Wrongly assuming independent scoring (A2) will not bias the estimates in contrast to inadvertently assuming equal scoring (A1), but the standard errors will be clearly different. However, when A2 is violated, the misclassification parameters will be estimated with artificially too high precision. Consequently, the corrected estimates of the main model will also be estimated with artificially too high precision. Relaxing A2 requires a multivariate model describing the misclassification process, thereby assuming dependence between the scores. Some possible models might include the MVP, multivariate-t or multivariate logistic model.

Chapter 10

General Conclusions and Further Research

In this thesis, we have developed Bayesian and frequentist methodologies to correct for misclassification errors in discrete data. This chapter presents some general conclusions and topics for further research.

10.1 General conclusions

Statistical analysis of oral health data present quite a challenging task. Caries experience data have a complex hierarchical structure, i.e. within each individual there are several teeth each with 4 or 5 surfaces. This necessitates the application of statistical methods that adequately take into account the correlated structure of such data. An additional challenge in the analysis of oral health data is the problem of misclassification error, especially in the scoring of caries experience, since (dental) examiners are prone to misclassification.

To reduce the variability in scoring caries experience, the dental examiners involved in the Signal Tandmobiel® study were trained at baseline

and participated in calibration exercises. At the end of these calibration exercises the examiner consistency in measuring caries experience was evaluated. Despite these calibration exercises, they nonetheless seem to have considerable residual misclassification. This inspired the development of techniques to correct for potential misclassification errors due to the examiners' scoring variability. The validation data generated from the calibration exercises offered us a possibility to estimate correction terms for this purpose.

In this work, methodologies to correct for misclassification errors when dealing with possibly corrupted binary, ordinal and count data have been developed. Chapters 4 and 5 were devoted to the modeling of binary and ordinal data, taking into account the possible misclassification of the response by the dental examiners. The second part of this thesis (Chapters 6 and 7) is focused on models for misclassified count data, whereas the final part (Chapters 8 and 9), we developed more general methods for misclassified multivariate binary data. However, the MC-SIMEX (Chapters 8) and MC-DA (Chapters 9) approaches are computationally intensive compared to the methodologies developed in the previous chapters.

The main focus of this research has been to identify and describe the impact of the response misclassification on the covariates' effect. For example, in the Signal Tandmobiel® study, it was important to investigate whether the East-West gradient in the degree of caries experience was genuine or induced by the potential misclassification of the dental examiners. The results from the various regression models considered here revealed a significant geographical East-West trend in caries experience despite the potential misclassification of the response by dental examiners. The corrected estimates of the other risk factors, in general, revealed a stronger relationship to caries experience compared to the naive estimates.

10.2 Further Research

Further research could focus on some of the topics considered in this thesis. We also suggest further refinements in modeling caries data.

In Chapter 5, a misclassification model that takes into account the ordinal nature of the scores has been considered among other misclassification models. This model assumed a univariate latent continuous scoring scale that varies from column to column of the misclassification matrix. Further, it is based on the conditional distributions of the observed response given the true response for each of the columns. However, this model does not provide a direct link between these conditional distributions. A possible extension and topic for further research is to introduce a bivariate ordinal latent model and derive these conditional distributions. This may provide a parsimonious way of linking all the columns.

In Chapters 6 and 7, we proposed the ZINB and ZIBB regression model for modeling unbounded and bounded count data, respectively. However, the resulting models from the corrected ZINB and ZIBB models are not in a standard form, as in the case of the Stamey et al. (2004), where a Poisson model for the infallible counts results into a Poisson distribution for the fallible counts. Derivation of the analytical expression of the corrected ZINB (and ZIBB) distribution remains a topic of further research.

A key assumption made when using a validation study to correct for bias due to misclassification or measurement error is that a measurement by a gold standard is available in addition to the error-prone measurement. In this work we assumed that a benchmark scorer (or a gold standard) is available. However, in the absence of a gold standard, the standard estimation techniques for misclassified data cannot be appropriately determined. In this case a *latent-class approach* (Formann, 1994), which is a statistical technique that allows for estimation of sensitivity and specificity when there is no gold standard, could be used.

Measurements on caries experience are usually taken at the tooth or tooth surface level. However, the statistical analysis is most often carried

out on an aggregated level of the child, i.e. using the dmft(s)-index . The dmft(s)-index has been criticized by some authors. For example, Birch (1986) points out that the dmft-index is a too rough measure of caries experience ignoring the individual characteristic of the teeth. For the dmfs-index there is a problem that when a tooth is extracted, automatically all surfaces are assumed to have experienced the disease; clearly an unrealistic assumption (Benigeri et al., 1998). Determining the actual distribution of caries experience at the tooth surface, taking into account (a) the complex hierarchical data structure and (b) the possible misclassification error, is a topic of further research.

Finally, as described in Chapter 2, the degree of caries experience depends on the diagnostic threshold, i.e. on the level of lesion severity. The misclassification process will thus vary depending on which level of lesion severity is used. Any further research in modeling the effect of misclassification should assess the effect of the diagnostic thresholds on the relationship between caries experience and the risk factors.

APPENDIX

A Some theoretical details

This appendix describes some theoretical details supplementing the explanations given in the text. Section A.1 presents the derivatives of the misclassification model suggested by Albert et al. (1997). In Section A.2 the derivatives of the zero-inflated negative binomial regression with and without correction for misclassification are given. Section A.3 gives some technical details regarding the misclassification SIMEX. Finally, in Section A.4 we describe the Wald Chi-square statistic for testing the difference of regression coefficients.

A.1 Derivatives of the Albert *et al.* misclassification model

The first order derivative of the likelihood resulting from expression (5.9) with respect to the $\tilde{\boldsymbol{\zeta}} = (\zeta_0, \zeta_0, \zeta_0)'$-parameters for symmetric and asymmetric misclassification models.

A.1.1 Symmetric misclassification models

Symmetric $1p$

$$\frac{\partial \mathcal{L}_m}{\partial \zeta_0} = -\sum_a \sum_b m_{ab} \left(\frac{\Big(\mathrm{I}(a \neq b) - 1 \Big) V + 1}{V} \right),$$

where $V = 1 + \sum_{c \neq b} g(c|b)$.

Symmetric $2p$

$$\begin{aligned}
\frac{\partial \mathcal{L}_m}{\partial \zeta_0} &= -\sum_a \sum_b m_{ab} \left(\frac{\Big(\mathrm{I}(a \neq b) - 1 \Big) V + 1}{V} \right), \\
\frac{\partial \mathcal{L}_m}{\partial \zeta_1} &= -\sum_a \sum_b m_{ab} \left(\frac{\mathrm{I}(a \neq b)|a-b|V - \sum_{c \neq b} |c-b| g(c|b)}{V} \right).
\end{aligned}$$

A.1.2 Asymmetric misclassification models

Asymmetric $3p$

$$\begin{aligned}
\frac{\partial \mathcal{L}_m}{\partial \zeta_0} &= -\sum_a \sum_b m_{ab} \left(\frac{\Big(\mathrm{I}(a \neq b) - 1 \Big) V + 1}{V} \right), \\
\frac{\partial \mathcal{L}_m}{\partial \zeta_1} &= -\sum_a \sum_b m_{ab} \left(\frac{\mathrm{I}(a \neq b)(a-b)\,\mathrm{I}(a > b)V - \sum_{c \neq b} (c-b)\,\mathrm{I}(c > b) g(c|b)}{V} \right), \\
\frac{\partial \mathcal{L}_m}{\partial \zeta_2} &= -\sum_a \sum_b m_{ab} \left(\frac{\mathrm{I}(a \neq b)(b-a)\,\mathrm{I}(a < b)V - \sum_{c \neq b} (b-c)\,\mathrm{I}(c < b) g(c|b)}{V} \right).
\end{aligned}$$

Asymmetric $4p$

$$
\begin{aligned}
\frac{\partial \mathcal{L}_m}{\partial \zeta_{00}} &= -\sum_a \sum_b m_{ab} \left(\frac{\Big(\mathrm{I}(a \neq b) - 1\Big)V + 1}{V} \right), \\
\frac{\partial \mathcal{L}_m}{\partial \zeta_{01}} &= -\sum_a \sum_b m_{ab}\, \mathrm{I}(b > 0) \left(\frac{\Big(\mathrm{I}(a \neq b) - 1\Big)V + 1}{V} \right), \\
\frac{\partial \mathcal{L}_m}{\partial \zeta_1} &= -\sum_a \sum_b m_{ab} \left(\frac{\mathrm{I}(a \neq b)(a-b)\,\mathrm{I}(a > b)V - \sum\limits_{c \neq b} (c-b)\,\mathrm{I}(c > b) g(c|b)}{V} \right), \\
\frac{\partial \mathcal{L}_m}{\partial \zeta_2} &= -\sum_a \sum_b m_{ab} \left(\frac{\mathrm{I}(a \neq b)(b-a)\,\mathrm{I}(a < b)V - \sum\limits_{c \neq b} (b-c)\,\mathrm{I}(c < b) g(c|b)}{V} \right).
\end{aligned}
$$

A.2 Derivatives of the ZINB regression model

In this section we present the first derivative with respect to $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \tau)$ for ZINB regression model as well as the corrected ZINB regression model. Further, we state the BFGS algorithm.

A.2.1 The ZINB regression model

The first order derivative of the minus log-likelihood of ZINB regression model, $\mathcal{L}_z$, expression (6.9) with respect to the $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \tau)'$-parameters:

$$
\frac{\partial \mathcal{L}_z}{\partial \beta_j} = \begin{cases} \displaystyle\sum_{i=1}^{n} \left(\frac{\mu_i\, \tau}{(\mu_i + \tau)\,(1 + q_i\, r_i)} \boldsymbol{x}_{ij} \right), & y = 0, \\ \displaystyle\sum_{i=1}^{n} \left(1 - \frac{\tau + y}{\mu_i + \tau} \right) \tau\, \boldsymbol{x}_{ij} & y > 0. \end{cases}
$$

$$
\frac{\partial \mathcal{L}_z}{\partial \gamma_j} = \begin{cases} \displaystyle\sum_{i=1}^{n} \left(\frac{1}{1 + q_i\, r_i} \boldsymbol{z}_{ij} - \frac{1}{1 + q_i} \right), & y = 0, \\ \displaystyle\sum_{i=1}^{n} \left(\frac{q_i}{1 + q_i} z_{ij} \right), & y > 0. \end{cases}
$$

$$\frac{\partial \mathcal{L}_z}{\partial \log \tau} = \begin{cases} \sum_{i=1}^{n} \left(\frac{-\mu_i + (\mu_i+\tau)/\tau \log r_i}{(\mu_i+\tau)\,(1+q_i\, r_i)} \tau \right) & y = 0, \\ \sum_{i=1}^{n} \left(-1 + \frac{\tau+y}{\mu_i+\tau} + \log(\frac{\mu_i+\tau}{\tau}) + \psi(\tau) - \psi(\tau+y) \right) \tau, & y > 0, \end{cases}$$

where $\mu_i = e^{\boldsymbol{x}_i'\boldsymbol{\beta}}$, $r_i = \left(\frac{\mu_i+\tau}{\tau}\right)^{\tau}$, $q_i = e^{\boldsymbol{z}_i'\boldsymbol{\gamma}}$ and $\psi(x) = \frac{\delta \log \Gamma(x)}{\delta x}$.

A.2.2 The corrected ZINB regression model

The first order derivative of $\mathcal{L}_c$ (the total likelihood of the corrected zero-inflated negative binomial distribution adjusted for misclassification in a pooled manner or when only one examiner is involved)

$$\mathcal{L}_c(\boldsymbol{\beta}, \boldsymbol{\gamma}, \tau; \boldsymbol{y}*, \boldsymbol{X}, \boldsymbol{Z}, \Pi) = -\sum_{i=1}^{n} \log \left\{ \frac{e^{\boldsymbol{z}_i'\boldsymbol{\gamma}} + \left(\frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\tau} + \tau \right)^{-\tau}}{1+e^{\boldsymbol{z}_i'\boldsymbol{\gamma}}} \pi_{(y_i^*|0)} + \sum_{y=1}^{K} \left(\frac{\Gamma(\tau+y)\left(1+e^{\boldsymbol{z}_i'\boldsymbol{\gamma}}\right)\left(\frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{\tau} + \tau \right)^{-\tau}}{\Gamma(\tau)\,\Gamma(1+y)\left(1+e^{-\boldsymbol{x}_i'\boldsymbol{\beta}}\,\tau\right)^{y}} \pi_{(y_i^*|y)} \right) \right\}.$$

with respect to the $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \tau)'$-parameters:

$$\frac{\partial \mathcal{L}_c}{\partial \beta_j} = -\sum_{i=1}^{n} \frac{\left(\frac{\mu_i \tau}{(\mu_i+\tau) r_i}\right) \pi_{(y_i^*|0)}}{(1+q_i)\, D_i} \boldsymbol{x}_{ij} - \sum_{i=1}^{n} \left(\sum_{y=1}^{K} \frac{\frac{\tau}{(\mu_i+\tau) r_i} \boldsymbol{x}_{ij}\, \Gamma(\tau+y)\, (\mu_i - y)\, \pi_{(y_i^*|y)}}{(1+q_i)\, (1+\tau/\mu_i)^{y}\, \Gamma(\tau)\, \Gamma(1+y) D_i} \boldsymbol{x}_{ij} \right),$$

$$\frac{\partial \mathcal{L}_c}{\partial \gamma_j} = -\sum_{i=1}^{n} \frac{q_i\, (r_i - 1)\, \boldsymbol{z}_{ij}\, \pi_{(y_i^*|0)}}{(1+q_i)^2\, r_i D_i} + \sum_{i=1}^{n} \left(\sum_{y=1}^{K} \frac{q_i\, \boldsymbol{z}_{ij}\, \Gamma(\tau+y)\, \pi_{(y_i^*|y)}}{(1+q_i)^2\, r_i\, (1+\mu_i\,\tau)^{y}\, \Gamma(\tau)\, \Gamma(1+y) D_i} \right),$$

$$\frac{\partial \mathcal{L}_c}{\partial \log \tau_j} = \sum_{i=1}^{n} \frac{\frac{\tau}{(\mu_i+\tau)r_i}\left(-\mu_i + (\mu_i+\tau)\log(\frac{\mu_i+\tau}{\tau})\right)\pi_{(y_i^*|0)}}{(1+q_i)\,D_i} - \sum_{i=1}^{n}\left(\sum_{y=1}^{K} \frac{\frac{\tau}{(\mu_i+\tau)r_i}\Gamma(\tau+y)\,\pi_{(y_i^*|y)}\triangle_i}{(1+q_i)\,\tau\,(1+\mu_i)^y\,\Gamma(\tau)\,\Gamma(1+y)D_i}\right),$$

where $\triangle_i = \left\{\mu_i - (\mu_i+\tau)\log(\frac{\mu_i+\tau}{\tau}) + (\mu_i+\tau)\left[-\psi(\tau)+\psi(\tau+y)\right] - y\right\}$, and D_i is the ZINB likelihood of the ith individual.

A.2.3 The BFGS algorithm

The BFGS (Broyden Fletcher Goldfarb Shanno) algorithm is a quasi-Newton method which was suggested independently by Broyden (1970), Fletcher (1970), Goldfarb (1970) and Shanno (1970). It is one of the most efficient quasi-Newton technique for unconstrained optimization:

$$\min f(x),\ x \in \Re^q.$$

Given the staring point $x_0 \in \Re^q$, convergence tolerance $\epsilon > 0$, and inverse Hessian approximation $H_0 \in \Re^{q\times q}$ then this algorithm can be summarized as follows:

$k \leftarrow 0$;

while $\|\nabla f_k\| > \epsilon$;

Carry out a line search in the direction

$$p_k = H_k \nabla f_k;$$

Set $x_{k+1} = x_k + \alpha_k p_k$, where α_k is obtained from the line search;

Define $s_k = x_{k+1} - x_k$, $y_k = \nabla f_{k+1} - \nabla f_k$, and $\rho_k = \frac{1}{y_k^T s_k}$;

Compute

$$H_{k+1} = (I - \rho_k s_k y_k^T) H_k (I - \rho_k y_k s_k^T) + \rho_k s_k s_k^T$$

$k \leftarrow k+1$;

end while.

The BFGS algorithm is implemented using the R function *optim*, and by choosing *method*="BFGS".

A.3 Misclassification SIMEX

A.3.1 Existence of Π^λ

Given a misclassification matrix Π the existence of the function $\lambda \longrightarrow \Pi^\lambda$ for $\lambda \geq 0$ can be solved with the theory of Markov processes: If Π is regarded as the transition matrix of time continuous Markov discrete process at time $t = 1$, then the transition matrix at time λ is given Π^λ. The existence of the Π^λ is equivalent to the existence of a Markov process with transition matrix Π at time t=1. This problem is known as the embedding problem (see e.g. Israel, Rosenthal, and Wei, 2001; Carette, 1995). For two categories the existence is equivalent to $\det(\Pi) = \pi_{00} + \pi_{11} - 1 > 0$, which should be fulfilled in most practical problems. For higher dimensions a sufficient condition is that the matrix

$$\log(\Pi) := \Xi \log(\Lambda) \Xi^{-1}$$

is a matrix with positive off diagonal elements, where Ξ is the matrix of eigenvectors and $\log(\Lambda)$ is the diagonal matrix of logarithms of the eigenvalues. This condition can be easily checked and will hold in most practical cases. A more general condition for existence of Π^λ which also applies for multiple or negative eigenvalues is the convergence of the series

$$\sum_{i=1}^{\infty} (\Pi - I)^i / i!,$$

which holds, e.g. if all diagonal elements are bigger than 0.5. In that case the nonnegativity of Π^λ has still to be checked. One example for such a matrix is

$$\Pi = \begin{pmatrix} 0.80 & 0.15 & 0.00 \\ 0.20 & 0.70 & 0.20 \\ 0.00 & 0.15 & 0.80 \end{pmatrix}.$$

Then $\Xi \cdot \log(\Lambda) \cdot \Xi^{-1}$ has negative components and e.g.

$$\Pi^{1/2} = \begin{pmatrix} 0.89 & 0.09 & -0.01 \\ 0.12 & 0.82 & 0.12 \\ -0.01 & 0.09 & 0.89 \end{pmatrix}.$$

The main reason for the problem is that $\pi_{13} = 0$, which leads to the nonexistence of Π^λ for $\lambda < 1$, Israel et al. (2001) discuss methods for finding a matrix with existing roots, which is close to the given matrix. In that example the matrix

$$\Pi = \begin{pmatrix} 0.77 & 0.15 & 0.03 \\ 0.20 & 0.70 & 0.20 \\ 0.03 & 0.15 & 0.77 \end{pmatrix}$$

has existing roots. Note that the existence Π^λ for $\lambda < 1$ is not a condition, which should hold in general for misclassification matrices. It can also happen that the matrix $\hat{\Pi}^\lambda$ does not exist for $\lambda < 1$, when Π is estimated by a validation study, while there is no problem with the true unknown matrix Π. Then we propose to use the method from Israel et al. (2001) as a useful approximation.

A.3.2 Variance estimation by the method of Stefanski

For a given misclassification matrix and a fixed grid of values $\lambda_1, \ldots, \lambda_m$ we calculate the variance of the estimators in one simulation sample:

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

and $V_{sim}(0) := 0$ We also use the naive (information matrix) estimation of the variance for every calculation of the naive estimator, denoted by $\hat{V}_{naive}\left(\hat{\beta}_{na}\left[(Y_i, X^*_{b,i}(\lambda_k), Z_i)_{i=1}^n\right]\right)$:

$$V_{na}(\lambda_k) := B^{-1} \sum_{b=1}^{B} \hat{V}_{naive}\left(\hat{\beta}_{na}\left[(Y_i, X^*_{b,i}(\lambda_k), Z_i)_{i=1}^n\right]\right).$$

Note that $V(\hat{\beta}_{sim}) \approx V(\hat{\beta}_{true}) + V(\hat{\beta}_{sim} - \hat{\beta}_{true})$, and $V(\hat{\beta}_{sim} - \hat{\beta}_{true}) = -\lim_{\lambda \to -1} V_{na}(\lambda)$ (*see* e.g. Carroll et al., 1995, Chapter 4). Hence, the Stefanski and Cook (1995) variance estimator is given by

$$\mathrm{V_{ST}} = \lim_{\lambda \to -1} \left(V_{na}(\lambda) - V_{sim}(\lambda)\right).$$

Hence, the variance estimation is given by extrapolating the function

$$\lambda \longrightarrow V_{na}(\lambda) - V_{sim}(\lambda)$$

to $\lambda = -1$ by a quadratic extrapolation function. For a detailed description, see Stefanski and Cook (1995).

A.4 Wald chi-square test

Let any nonsingular set of contrasts of a vector of regression coefficients $\boldsymbol{\beta}$ be given by the r columns $C_1, \cdots, C_r$ of a contrast matrix C. Then $\text{var}(C'\boldsymbol{\beta}) = C'\Sigma C$, where Σ is the covariance matrix of $\boldsymbol{\beta}$. The Wald Chi-square statistic for testing the hypothesis $C'\boldsymbol{\beta} = 0$ is given by

$$W_C = \left(C'\hat{\boldsymbol{\beta}}\right)' \left[C'\hat{\Sigma}C\right]^{-1} \left(C'\hat{\boldsymbol{\beta}}\right),$$

where $\hat{\boldsymbol{\beta}}$ is the estimate $\boldsymbol{\beta}$ and $\hat{\Sigma}$ is its estimated covariance matrix. The asymptotic distribution of W_C is χ^2_r, where r is the rank of C.

Appendix B

Simulation results for the double binomial extensions

Here we present selected simulation results of the double binomial simulation study described in Section 7.6.1. Table B.1 shows the simulation results for misclassified binomial regression when the validation is sampled with *equal* probability for scoring $Y = s$, i.e. $\Pr(Y = s) = 1/(K + 1)$, whereas Table B.2 shows the simulation results when the validation is sampled with *unequal* probability for $Y = s$, i.e. $\Pr(Y = s) = [2(K + 1 - s)]/[(K + 1)(K + 2)]$. Finally, Table B.3 shows the results from binomial regression subject to misclassification error when the validation data is a random sub-sample of the main data.

Table B.1: *Simulation results for binomial regression: K= 8, N = 100 with moderately varying sensitivity α_k and specificity β_k around 90%, 95% respectively for each of the four extensions (E0 corresponds to the basic DB approach). Case of $P(Y = s) = 1/(K + 1)$ in the validation data. p_X is the success probability of the binary regressor, SD_Z is the standard deviation of the normal continuous regressor.*

		β_0		β_X		β_Z	
Ext (p_X,SD_Z)		Mean(SD)	MSE	Mean(SD)	MSE	Mean(SD)	MSE
	True	0.000(0.031)	—	−0.998(0.046)	—	0.996(0.238)	—
E0 (0.5, 0.1)	Naive	−0.100(0.031)	0.011	−0.849(0.045)	0.024	0.835(0.242)	0.084
	Mult	0.001(0.072)	0.005	−0.992(0.082)	0.007	0.982(0.289)	0.084
	E0	0.002(0.059)	0.003	−0.999(0.061)	0.004	0.992(0.289)	0.084
	True	0.000(0.039)	—	−0.999(0.051)	—	1.001(0.030)	—
E0 (0.6, 1)	Naive	−0.114(0.039)	0.015	−0.806(0.050)	0.040	0.803(0.028)	0.040
	Mult	−0.005(0.078)	0.006	−0.981(0.082)	0.007	0.982(0.066)	0.005
	E0	0.001(0.072)	0.005	−1.002(0.070)	0.005	1.004(0.050)	0.002
	True	−0.001(0.075)	—	−1.001(0.095)	—	1.000(0.024)	—
E0 (0.7, 5)	Naive	−0.181(0.061)	0.036	−0.475(0.076)	0.282	0.470(0.016)	0.281
	Mult	−0.032(0.117)	0.015	−0.816(0.145)	0.055	0.812(0.107)	0.047
	E0	−0.000(0.121)	0.015	−1.001(0.132)	0.018	0.996(0.072)	0.005
	True	−0.001(0.130)	—	−1.001(0.152)	—	1.002(0.033)	—
E0 (0.8, 10)	Naive	−0.200(0.089)	0.047	−0.273(0.104)	0.541	0.271(0.010)	0.534
	Mult	−0.044(0.153)	0.025	−0.624(0.202)	0.183	0.622(0.135)	0.162
	E0	−0.001(0.192)	0.035	−0.994(0.208)	0.043	0.997(0.086)	0.007
	True	0.001(0.031)	—	−1.002(0.046)	—	1.009(0.236)	—
E1 (0.5, 0.1)	Naive	−0.058(0.031)	0.004	−0.842(0.047)	0.028	0.842(0.238)	0.085
	Mult	−0.002(0.073)	0.005	−0.991(0.085)	0.007	0.994(0.291)	0.085
	E0	0.000(0.059)	0.003	−1.003(0.063)	0.004	1.013(0.289)	0.083
	E1	0.003(0.044)	0.002	−1.004(0.060)	0.004	0.997(0.281)	0.080
	True	−0.000(0.040)	—	−1.000(0.054)	—	1.002(0.031)	—
E1 (0.6, 1)	Naive	−0.066(0.037)	0.006	−0.798(0.051)	0.043	0.794(0.028)	0.044
	Mult	−0.004(0.076)	0.006	−0.978(0.087)	0.008	0.978(0.069)	0.005
	E0	−0.001(0.071)	0.005	−1.003(0.073)	0.005	1.005(0.052)	0.003
	E1	0.003(0.049)	0.002	−1.001(0.070)	0.005	0.998(0.040)	0.002
	True	0.001(0.072)	—	−1.001(0.091)	—	1.002(0.025)	—
E1 (0.7, 5)	Naive	−0.105(0.057)	0.015	−0.465(0.070)	0.292	0.464(0.016)	0.289
	Mult	−0.015(0.119)	0.014	−0.811(0.150)	0.058	0.810(0.106)	0.048
	E0	0.004(0.125)	0.016	−1.004(0.140)	0.020	1.002(0.071)	0.005
	E1	0.014(0.124)	0.026	−1.016(0.131)	0.017	0.996(0.066)	0.004
	True	0.000(0.127)	—	−1.004(0.151)	—	1.002(0.033)	—
E1 (0.8, 10)	Naive	−0.118(0.087)	0.021	−0.269(0.099)	0.55	0.268(0.010)	0.538
	Mult	−0.037(0.155)	0.026	−0.594(0.195)	0.205	0.594(0.134)	0.184
	E0	−0.009(0.188)	0.036	−0.992(0.213)	0.045	0.994(0.086)	0.007
	E1	0.006(0.172)	0.029	−1.005(0.198)	0.039	1.006(0.051)	0.003
	True	−0.001(0.032)	—	−0.999(0.047)	—	0.998(0.235)	—
E2 (0.5, 0.1)	Naive	−0.050(0.033)	0.003	−0.897(0.049)	0.013	0.887(0.242)	0.071
	Mult	−0.003(0.065)	0.004	−0.993(0.077)	0.006	0.988(0.280)	0.078
	E0	−0.005(0.055)	0.003	−1.034(0.062)	0.005	1.032(0.283)	0.082
	E2	−0.003(0.037)	0.001	−0.998(0.055)	0.003	0.996(0.277)	0.078
	True	−0.001(0.041)	—	−1.002(0.055)	—	1.002(0.032)	—
E2 (0.6, 1)	Naive	−0.055(0.041)	0.005	−0.865(0.055)	0.022	0.861(0.031)	0.021
	Mult	−0.004(0.070)	0.005	−0.985(0.079)	0.006	0.985(0.061)	0.004
	E0	−0.005(0.069)	0.005	−1.043(0.074)	0.007	1.045(0.051)	0.005
	E2	−0.006(0.046)	0.002	−0.993(0.064)	0.004	0.997(0.038)	0.001

Continued on next page

Table B.1: ***continued from previous page***

Ext (p_X,SD_Z)		β_0		β_X		β_Z	
		Mean(SD)	MSE	Mean(SD)	MSE	Mean(SD)	MSE
	True	0.002(0.074)	—	−1.004(0.088)	—	1.001(0.025)	—
E2 (0.7, 5)	Naive	−0.082(0.063)	0.011	−0.566(0.076)	0.198	0.562(0.019)	0.194
	Mult	−0.008(0.108)	0.012	−0.866(0.140)	0.039	0.859(0.100)	0.03
	E0	0.006(0.118)	0.014	−1.078(0.129)	0.022	1.072(0.066)	0.009
	E2	0.003(0.090)	0.008	−0.998(0.106)	0.011	1.000(0.039)	0.001
	True	0.001(0.128)	—	−1.004(0.144)	—	1.003(0.032)	—
E2 (0.8, 10)	Naive	−0.094(0.094)	0.018	−0.338(0.105)	0.454	0.339(0.014)	0.442
	Mult	−0.026(0.148)	0.023	−0.680(0.194)	0.142	0.683(0.138)	0.122
	E0	−0.003(0.189)	0.036	−1.075(0.213)	0.050	1.077(0.078)	0.011
	E2	0.014(0.172)	0.030	−1.013(0.192)	0.037	1.00(0.047)	0.002
	True	−0.001(0.032)	—	−0.999(0.048)	—	1.006(0.227)	—
E3 (0.5, 0.1)	Naive	−0.104(0.033)	0.012	−0.843(0.049)	0.026	0.841(0.227)	0.079
	Mult	−0.005(0.071)	0.005	−0.990(0.084)	0.007	0.995(0.282)	0.080
	E0	−0.001(0.059)	0.003	−1.001(0.064)	0.004	1.008(0.274)	0.075
	E3	−0.003(0.037)	0.001	−0.995(0.057)	0.003	1.016(0.271)	0.074
	True	0.001(0.039)	—	−1.002(0.052)	—	1.002(0.031)	—
E3 (0.6, 1)	Naive	−0.117(0.038)	0.015	−0.801(0.050)	0.043	0.796(0.029)	0.043
	Mult	−0.004(0.078)	0.006	−0.980(0.084)	0.008	0.979(0.069)	0.005
	E0	0.001(0.074)	0.005	−1.005(0.074)	0.005	1.006(0.053)	0.003
	E3	0.003(0.049)	0.002	−1.001(0.066)	0.004	1.002(0.041)	0.002
	True	0.005(0.074)	—	−1.005(0.092)	—	1.000(0.024)	—
E3 (0.7, 5)	Naive	−0.183(0.059)	0.039	−0.466(0.073)	0.296	0.462(0.016)	0.290
	Mult	−0.031(0.117)	0.015	−0.806(0.145)	0.061	0.800(0.100)	0.050
	E0	0.002(0.121)	0.015	−1.000(0.135)	0.018	0.994(0.070)	0.005
	E3	0.006(0.095)	0.009	−1.004(0.122)	0.015	1.000(0.041)	0.002
	True	−0.008(0.128)	—	−0.993(0.145)	—	1.005(0.033)	—
E3 (0.8, 10)	Naive	−0.210(0.092)	0.049	−0.257(0.104)	0.553	0.265(0.010)	0.548
	Mult	−0.058(0.156)	0.027	−0.594(0.188)	0.195	0.605(0.129)	0.176
	E0	−0.004(0.195)	0.035	−0.990(0.222)	0.042	1.001(0.087)	0.008
	E3	−0.014(0.249)	0.029	−1.009(0.196)	0.039	1.008(0.056)	0.003
	True	0.001(0.031)	—	−1.001(0.047)	—	0.986(0.239)	—
E4 (0.5, 0.1)	Naive	−0.214(0.032)	0.047	−0.837(0.048)	0.029	0.810(0.240)	0.089
	Mult	−0.006(0.082)	0.007	−0.990(0.090)	0.008	0.968(0.297)	0.088
	E0	−0.017(0.065)	0.005	−1.008(0.064)	0.004	0.988(0.295)	0.087
	E4	0.002(0.041)	0.002	−1.000(0.059)	0.003	0.977(0.283)	0.080
	True	0.003(0.038)	—	−1.003(0.050)	—	1.000(0.030)	—
E4 (0.6, 1)	Naive	−0.231(0.038)	0.056	−0.794(0.051)	0.046	0.786(0.029)	0.046
	Mult	−0.002(0.085)	0.007	−0.986(0.092)	0.009	0.981(0.072)	0.006
	E0	−0.006(0.078)	0.006	−1.021(0.077)	0.006	1.017(0.056)	0.003
	E4	−0.003(0.053)	0.003	−0.994(0.071)	0.005	1.001(0.042)	0.002
	True	0.003(0.075)	—	−1.005(0.091)	—	1.001(0.024)	—
E4 (0.7, 5)	Naive	−0.332(0.060)	0.116	−0.449(0.073)	0.314	0.447(0.015)	0.308
	Mult	−0.046(0.124)	0.018	−0.815(0.143)	0.057	0.810(0.102)	0.047
	E0	0.007(0.131)	0.017	−1.042(0.138)	0.020	1.038(0.073)	0.007
	E4	0.007(0.095)	0.009	−1.008(0.115)	0.013	1.004(0.046)	0.002
	True	0.005(0.125)	—	−1.005(0.142)	—	1.002(0.032)	—
E4 (0.8, 10)	Naive	−0.360(0.089)	0.141	−0.250(0.102)	0.580	0.254(0.010)	0.560
	Mult	−0.087(0.160)	0.034	−0.604(0.197)	0.199	0.610(0.134)	0.172
	E0	0.009(0.221)	0.036	−1.021(0.264)	0.043	1.032(0.083)	0.008
	E4	−0.008(0.179)	0.032	−1.004(0.239)	0.041	1.006(0.059)	0.003

Table B.2: *Simulation results for binomial regression: K= 8, N = 100 with moderately varying sensitivity α_k and specificity β_k around 90%, 95% respectively for each of the four extensions ((E0 corresponds to the basic DB approach). Case of $P(Y = s)$ descending in s in the validation data. p_X is the success probability of the binary regressor, SD_Z is the standard deviation of the normal continuous regressor.*

Ext (p_X,SD_Z)		β_0		β_X		β_Z	
		Mean(SD)	MSE	Mean(SD)	MSE	Mean(SD)	MSE
	True	0.002(0.031)	—	−1.002(0.048)	—	0.989(0.242)	—
E0 (0.5, 0.1)	Naive	−0.098(0.031)	0.011	−0.854(0.047)	0.024	0.836(0.232)	0.077
	Mult (889)†	−0.001(0.075)	0.006	−0.994(0.081)	0.007	0.977(0.283)	0.080
	E0	0.003(0.062)	0.004	−1.005(0.062)	0.004	0.994(0.278)	0.077
	True	−0.001(0.038)	—	−1.000(0.053)	—	1.002(0.031)	—
E0 (0.6, 1)	Naive	−0.116(0.037)	0.015	−0.806(0.051)	0.040	0.803(0.029)	0.040
	Mult (882)†	−0.019(0.080)	0.007	−0.964(0.085)	0.009	0.965(0.069)	0.006
	E0	0.002(0.076)	0.006	−1.000(0.072)	0.005	1.003(0.051)	0.003
	True	0.004(0.070)	—	−1.008(0.091)	—	1.002(0.024)	—
E0 (0.7, 5)	Naive	−0.177(0.060)	0.037	−0.477(0.074)	0.287	0.470(0.015)	0.282
	Mult (869)†	−0.134(0.129)	0.036	−0.697(0.153)	0.120	0.691(0.123)	0.112
	E0	0.001(0.129)	0.017	−1.008(0.139)	0.019	0.999(0.073)	0.005
	True	0.001(0.129)	—	−1.004(0.147)	—	1.002(0.033)	—
E0 (0.8, 10)	Naive	−0.206(0.091)	0.051	−0.268(0.100)	0.552	0.272(0.011)	0.534
	Mult (887)†	−0.185(0.155)	0.058	−0.453(0.181)	0.336	0.455(0.126)	0.316
	E0	0.000(0.195)	0.038	−1.004(0.214)	0.046	0.999(0.092)	0.009
	True	−0.000(0.031)	—	−1.000(0.049)	—	1.003(0.233)	—
E1 (0.5, 0.1)	Naive	−0.059(0.031)	0.004	−0.842(0.049)	0.027	0.833(0.228)	0.081
	Mult (891)†	−0.006(0.071)	0.005	−0.990(0.083)	0.007	0.994(0.276)	0.076
	E0	0.001(0.062)	0.004	−1.004(0.064)	0.004	1.003(0.275)	0.076
	E1	0.001(0.036)	0.001	−1.005(0.058)	0.003	0.974(0.270)	0.073
	True	−0.000(0.038)	—	−1.002(0.051)	—	1.002(0.031)	—
E1 (0.6, 1)	Naive	−0.068(0.037)	0.006	−0.798(0.051)	0.044	0.792(0.029)	0.045
	Mult (888)†	−0.022(0.080)	0.007	−0.964(0.081)	0.008	0.963(0.065)	0.006
	E0	−0.002(0.074)	0.006	−1.002(0.072)	0.005	1.002(0.051)	0.003
	E1	−0.003(0.047)	0.002	−0.998(0.063)	0.004	0.999(0.042)	0.002
	True	0.001(0.077)	—	−1.003(0.093)	—	1.001(0.025)	—
E1 (0.7, 5)	Naive	−0.105(0.062)	0.015	−0.467(0.076)	0.293	0.465(0.015)	0.289
	Mult (885)†	−0.122(0.138)	0.034	−0.678(0.157)	0.130	0.679(0.119)	0.118
	E0	0.006(0.133)	0.018	−1.003(0.140)	0.020	0.999(0.073)	0.005
	E1	−0.009(0.130)	0.017	−0.988(0.133)	0.018	0.987(0.070)	0.005
	True	−0.000(0.127)	—	−1.002(0.146)	—	1.002(0.033)	—
E1 (0.8, 10)	Naive	−0.118(0.091)	0.022	−0.268(0.102)	0.55	0.268(0.010)	0.538
	Mult (884)†	−0.159(0.164)	0.052	−0.452(0.184)	0.337	0.451(0.126)	0.319
	E0	0.010(0.225)	0.047	−1.003(0.242)	0.060	1.003(0.090)	0.008
	E1	−0.003(0.145)	0.021	−0.998(0.166)	0.028	1.004(0.051)	0.003
	True	0.001(0.032)	—	−1.002(0.047)	—	0.997(0.231)	—
E2 (0.5, 0.1)	Naive	−0.047(0.033)	0.003	−0.900(0.048)	0.013	0.889(0.238)	0.068
	Mult (878)†	0.001(0.068)	0.005	−0.994(0.077)	0.006	0.976(0.275)	0.076
	E0	0.023(0.058)	0.004	−1.037(0.061)	0.005	1.033(0.277)	0.078
	E2	−0.001(0.038)	0.001	−0.998(0.054)	0.003	0.989(0.271)	0.073
	True	−0.001(0.037)	—	−0.999(0.052)	—	1.000(0.031)	—
E2 (0.6, 1)	Naive	−0.056(0.038)	0.004	−0.862(0.053)	0.022	0.858(0.031)	0.021
	Mult (876)†	−0.015(0.071)	0.005	−0.973(0.079)	0.007	0.972(0.063)	0.005
	E0	0.023(0.069)	0.005	−1.041(0.073)	0.007	1.043(0.051)	0.004
	E2	−0.007(0.051)	0.003	−0.991(0.065)	0.004	0.994(0.043)	0.003

Continued on next page

Table B.2: ***continued from previous page***

Ext (p_X,SD_Z)		β_0		β_X		β_Z	
		Mean(SD)	MSE	Mean(SD)	MSE	Mean(SD)	MSE
	True	−0.001(0.075)	—	−1.000(0.093)	—	1.000(0.025)	—
E2 (0.7, 5)	Naive	−0.086(0.062)	0.011	−0.561(0.078)	0.199	0.561(0.018)	0.194
	Mult (882)†	−0.106(0.124)	0.026	−0.736(0.149)	0.092	0.741(0.122)	0.082
	E0	0.035(0.125)	0.017	−1.079(0.136)	0.025	1.078(0.071)	0.011
	E2	0.003(0.093)	0.009	−1.007(0.116)	0.014	1.002(0.041)	0.002
	True	0.003(0.128)	—	−1.002(0.150)	—	1.002(0.031)	—
E2 (0.8, 10)	Naive	−0.096(0.096)	0.019	−0.335(0.106)	0.455	0.339(0.014)	0.440
	Mult (895)†	−0.133(0.159)	0.044	−0.520(0.194)	0.269	0.526(0.152)	0.250
	E0	0.042(0.192)	0.039	−1.080(0.213)	0.052	1.088(0.082)	0.014
	E2	0.001(0.166)	0.027	−1.006(0.185)	0.034	1.003(0.055)	0.003
	True	0.000(0.032)	—	−1.002(0.047)	—	0.998(0.239)	—
E3 (0.5, 0.1)	Naive	−0.102(0.032)	0.012	−0.847(0.048)	0.026	0.838(0.241)	0.084
	Mult (901)†	−0.006(0.075)	0.006	−0.988(0.081)	0.007	0.983(0.286)	0.082
	E0	0.002(0.063)	0.004	−1.005(0.062)	0.004	1.004(0.290)	0.084
	E3	0.001(0.034)	0.001	−1.000(0.056)	0.003	0.992(0.269)	0.072
	True	0.000(0.039)	—	−1.000(0.051)	—	1.001(0.031)	—
E3 (0.6, 1)	Naive	−0.118(0.038)	0.015	−0.799(0.051)	0.043	0.796(0.029)	0.043
	Mult(888)†	−0.023(0.081)	0.007	−0.962(0.086)	0.009	0.964(0.069)	0.006
	E0	0.002(0.077)	0.006	−1.003(0.071)	0.005	1.006(0.054)	0.003
	E3	0.003(0.052)	0.003	−0.996(0.063)	0.004	0.997(0.042)	0.002
	True	−0.001(0.075)	—	−1.003(0.093)	—	1.003(0.025)	—
E3 (0.7, 5)	Naive	−0.189(0.060)	0.039	−0.458(0.072)	0.301	0.463(0.016)	0.292
	Mult(888)†	−0.148(0.131)	0.039	−0.660(0.149)	0.14	0.662(0.120)	0.131
	E0	−0.000(0.131)	0.017	−0.999(0.143)	0.02	1.002(0.077)	0.006
	E3	0.001(0.097)	0.009	−0.998(0.123)	0.015	1.000(0.041)	0.002
	True	−0.002(0.134)	—	−1.000(0.152)	—	1.004(0.032)	—
E3 (0.8, 10)	Naive	−0.201(0.094)	0.049	−0.267(0.105)	0.548	0.265(0.010)	0.546
	Mult (877)†	−0.183(0.157)	0.057	−0.436(0.181)	0.351	0.437(0.116)	0.334
	E0	−0.001(0.201)	0.036	−1.002(0.208)	0.043	0.999(0.090)	0.008
	E3	0.005(0.162)	0.026	−1.011(0.188)	0.036	1.006(0.055)	0.003
	True	−0.001(0.033)	—	−0.999(0.047)	—	0.991(0.237)	—
E4 (0.5, 0.1)	Naive	−0.216(0.034)	0.047	−0.835(0.050)	0.029	0.826(0.245)	0.087
	Mult (880)†	−0.011(0.086)	0.007	−0.986(0.088)	0.008	0.987(0.300)	0.090
	E0	−0.007(0.072)	0.005	−1.005(0.066)	0.004	1.007(0.299)	0.090
	E4	0.004(0.048)	0.002	−1.005(0.063)	0.004	0.982(0.287)	0.083
	True	0.001(0.038)	—	−1.003(0.053)	—	0.999(0.031)	—
E4 (0.6, 1)	Naive	−0.233(0.038)	0.057	−0.793(0.053)	0.047	0.786(0.029)	0.046
	Mult (891)†	−0.024(0.084)	0.008	−0.969(0.087)	0.009	0.964(0.069)	0.006
	E0	0.005(0.082)	0.007	−1.021(0.080)	0.007	1.018(0.054)	0.003
	E4	0.002(0.054)	0.003	−1.004(0.071)	0.005	1.004(0.042)	0.002
	True	−0.005(0.072)	—	−0.995(0.090)	—	1.001(0.025)	—
E4 (0.7, 5)	Naive	−0.337(0.062)	0.114	−0.446(0.076)	0.308	0.447(0.015)	0.307
	Mult (900)†	−0.174(0.142)	0.048	−0.660(0.154)	0.136	0.662(0.121)	0.130
	E0	0.022(0.148)	0.023	−1.034(0.150)	0.024	1.038(0.079)	0.008
	E4	0.001(0.103)	0.011	−1.009(0.123)	0.015	1.001(0.041)	0.002
	True	−0.007(0.130)	—	−0.999(0.151)	—	1.003(0.032)	—
E4 (0.8, 10)	Naive	−0.358(0.088)	0.131	−0.252(0.098)	0.567	0.254(0.010)	0.560
	Mult (875)†	−0.235(0.153)	0.075	−0.420(0.176)	0.366	0.417(0.114)	0.356
	E0	0.019(0.226)	0.042	−1.033(0.226)	0.050	1.031(0.081)	0.007
	E4	−0.002(0.212)	0.028	−0.995(0.225)	0.050	1.007(0.058)	0.003

†Number of simulation samples (over 1000) of which the multinomial approach was estimable

Table B.3: *Simulation results for binomial regression: K= 8, N = 100 with moderately varying sensitivity α_k and specificity β_k around 90%, 95% respectively for E0 (basic DB approach), when the validation data is a random sub-sample of the main data.*

Ext (p_X,SD_Z)		β_0		β_X		β_Z	
		Mean(SD)	MSE	Mean(SD)	MSE	Mean(SD)	MSE
	True	0.000(0.030)	—	−1.001(0.046)	—	0.992(0.241)	—
(0.5, 0.1)	Naive	−0.101(0.031)	0.011	−0.851(0.046)	0.025	0.832(0.240)	0.083
	Mult (227)†	−0.016(0.064)	0.004	−0.977(0.074)	0.006	0.934(0.274)	0.079
	E0	−0.001(0.059)	0.003	−1.002(0.061)	0.004	0.989(0.285)	0.081
	True	−0.002(0.038)	—	−0.999(0.051)	—	1.000(0.031)	—
(0.6, 1)	Naive	−0.117(0.037)	0.015	−0.806(0.050)	0.04	0.802(0.030)	0.040
	Mult (887)†	−0.018(0.072)	0.005	−0.964(0.081)	0.008	0.964(0.067)	0.006
	E0	−0.002(0.070)	0.005	−0.999(0.069)	0.005	1.001(0.052)	0.003
	True	0.001(0.076)	—	−1.005(0.091)	—	1.002(0.025)	—
(0.7, 5)	Naive	−0.181(0.062)	0.037	−0.473(0.074)	0.289	0.471(0.015)	0.282
	Mult (931)†	−0.014(0.150)	0.023	−0.908(0.150)	0.032	0.906(0.102)	0.020
	E0	0.003(0.124)	0.015	−1.006(0.132)	0.017	1.003(0.071)	0.005
	True	0.002(0.127)	—	−1.007(0.147)	—	1.002(0.034)	—
(0.8, 10)	Naive	−0.198(0.090)	0.048	−0.279(0.103)	0.541	0.271(0.011)	0.534
	Mult (563)†	−0.005(0.237)	0.056	−0.834(0.247)	0.091	0.822(0.173)	0.063
	E0	0.001(0.182)	0.033	−1.011(0.205)	0.042	1.004(0.082)	0.007

†Number of simulation samples (over 1000) of which the multinomial approach was estimable

APPENDIX C Software

In this appendix, we only present some of the selected but important WinBUGS' programs used for model fitting in this thesis. Provided also are the programs of the WinBUGS Development Interface (WBDev), which enables the implementation of user defined functions into the WinBUGS system. The WBDev *folder* should be copied to the WinBUGS' root directory. All the programs can be obtained from

`http://med.kuleuven.be/biostat/software/software.htm`

C.1 Programs for Chapter 4

Program 4.1 fits the uncorrected random effects logistic model (4.1) predicting the prevalence of caries experience, controlling for the geographical effect (see Table 4.2).

Program 4.2 fits the misclassification rates from model (4.8) (see Table 4.5).

Program 4.3 fits the corrected random-effects simple logistic regression model (4.7) in combination with misclassification model (4.8) predicting the prevalence of caries experience (see Table 4.6).

C.2 Programs for Chapter 5

Program 5.1 fits the random-intercepts multinomial logit model (5.1) predicting the degree of caries experience, controlling for the geographical effect (see Table 5.1).

Program 5.2 fits the examiner-specific coefficients ws of the first and second misclassification model (5.7), estimated from the corresponding (see Table 5.3).

Program 5.3 fits the corrected random-effects ordinal logistic regression model (5.5) in combination with the first misclassification model, expression (5.7) with $w_j \sim dbeta(1,1)$, predicting the degree of caries experience (see Table 5.5)

Program 5.4 fits the corrected random-effects ordinal logistic regression model (5.5) in combination with second misclassification model, expression (5.7) with $\text{logit}(w_j) \sim \mathcal{N}(\mu_w, \sigma_w^2)$, predicting the degree of caries experience (see Table 5.6).

Program 5.5 fits the corrected random-effects ordinal logistic regression model (5.5) in combination with third misclassification model (5.8) predicting the degree of caries experience (see Table 5.7).

Program 5.6 fits the corrected random-effects ordinal logistic regression model (5.5) in combination with fourth misclassification model (5.9) predicting the degree of caries experience (see Table 5.8).

C.3 Programs for Chapter 6

Program 6.1 fits the Poisson, generalized Poisson (GP), negative binomial (NB) and Poisson-inverse Gaussian (PIG) distributions (see Table 6.2).

Program 6.2 fits the zero-inflated Poisson (ZIP), zero-inflated generalized Poisson(ZIGP), zero-inflated negative binomial (ZINB) and zero-inflated Poisson-inverse (ZIPIG) Gaussian distributions (see Table 6.3).

Program 6.3 fits the multiple ZINB regression model predicting the dmft-index (see Table 6.4).

Program 6.4-a is a R program that fits the symmetric and asymmetric misclassification model from pooled Albert *et al.*'s approach (see 6.5).

Program 6.4-b is a `WinBUGS` program that fits the symmetric and asymmetric misclassification model from pooled Albert *et al.*'s approach (see 6.5).

Program 6.5 fits the pooled corrected multiple ZINB regression model predicting the dmft-index combined with 3p-asymmetric Albert *et al.*'s misclassification model (see Table 6.7).

Program 6.6 fits the examiner-specific corrected multiple ZINB regression model predicting the dmft-index combined with Albert *et al.*'s misclassification model (see Table 6.8).

C.4 Programs for Chapter 7

Program 7.1 fits the distribution of the $\text{dmft}_{4,5}$ with expected frequencies obtained by fitting a beta-binomial (BB) and zero-inflated beta-binomial (ZIBB) distribution (see Table 7.1).

Program 7.2 fits uncorrected ZIBB regression model fitted to the $\text{dmft}_{4,5}$ (see Table 7.2).

Program 7.3 fits the pooled corrected ZIBB regression model fitted to the $\text{dmft}_{4,5}$ (see Table 7.11).

Program 7.4 fits the examiner-specific corrected ZIBB regression model fitted to the $\text{dmft}_{4,5}$ (see Table 7.12).

Bibliography

Agresti, A. (1984). *Analysis of Ordinal Categorical Data*. Wiley, New York.

Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.

Albert, P. S., Hunsberger, S. A. and Biro, F. M. (1997). Modeling repeated measures with monotonic ordinal responses and misclassification, with applications to studying maturation. *Journal of the American Statistical Association*, **92**, 1304–1311.

Altham, P. M. E. (1978). Two generalizations of the binomial distribution. *Applied Statistics*, **27**, 162–167.

Begg, C. B. (1987). Biases in the assessment of diagnostic tests. *Statistics in Medicine*, **6**, 411–423.

Benigeri, M., Payette, M. and Brodeur, J. M. (1998). Comparison between the DMF indices and two alternative composite indicators of dental health. *Community Dentistry Oral Epidemiology*, **26**, 303–309.

Best, N., Cowles, K. M. and Vines, K. (1996). *Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output*. Cambridge, UK.

Birch, S. (1986). Measuring dental health improvements on the DMF-index. *Community Dental Health*, **3**, 303–311.

Birkett, N. J. (1992). Effect of nondifferential misclassification on the estimates of odds ratios with multiple levels of exposure. *American Journal of Epidemiology*, **136**, 356–362.

Booth, J. G., Casella, G., Friedl, H. and Hobert, J. P. (2003). Negative binomial loglinear mixed models. *Statistical Modelling*, **3**, 179–191.

Böhning, D., Dietz, E., Schlattman, P., Mendonça, L. and Kirchner, U. (1999). The zero-inflated Poisson model and the decayed, missing and filled

teeth index in dental epidemiology. *Journal of the Royal Statistical Society, Series A*, **162**, 195–209.

Bratcher, T. and Stamey, J. (2002). Estimation of Poisson rates with misclassified counts. *Biometrical Journal*, **44**, 946–956.

Brooks, S. P. and Gelman, A. (1997). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, **7**, 434–455.

Bross, I. (1954). Misclassification in 2 × 2 tables. *Biometrics*, **10**, 488–495.

Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms, Parts I and II. *Journal of the Institute of Mathematics and Its Applications*, **6**, 222–236.

Carette, P. (1995). Characterizations of embeddable 3 × 3 matrices with a negative eigenvalue. *New York Journal of Mathematics*, **1**, 120–129.

Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models*. Chapman & Hall, London.

Carroll, R. J., Spiegelman, C. H., Lan, K. K., Bailey, K. T. and Abbott, R. D. (1984). On errors-in-the variables for binary regression models. *Biometrika*, **71**, 643–648.

Carroll, R. and Küchenhoff, H. (1995). Approximate methods for regression models with errors in the covariates. In Härdle, W., Klinke, S. and Turlach, B. A., editors, *XploRe: An Interactive Statistical Computing Environment*, pages 239–248. Springer.

Carroll, R., Küchenhoff, H., Lombard, F. and Stefanski, L. (1996). Asymptotics for the SIMEX estimator in structural measurement error models. *Journal of the American Statistical Association*, **91**, 242–250.

Carroll, R., Roeder, K. and Wasserman, L. (1999). Flexible parametric measurement error models. *Biometrics*, **55**, 44–54.

Carroll, R. and Stefanski, L. (1990). Approximate quasi-likelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association*, **85**, 652–663.

Carvalho, J., van Nieuwenhuysen, J. and D'Hoore, W. (2001). The decline in dental caries among Belgian children between 1983 and 1998. *Community Dentistry and Oral Epidemiology*, **29**, 55–61.

Chen, M.-H. and Dipak, K. D. (1998). Bayesian modelling of correlated binary responses via scale mixture of multivariate normal link functions. *Sankhyā*, **60**, 322–343.

Chen, T. T. (1989). A review of method for misclassified categorical data in epidemiology. *Statistics in Medicine*, **8**, 1095–1165.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *American Statistician*, **49**, 327–335.

Chib, S. and Greenberg, E. (1998). Bayesian analysis of multivariate probit models. *Biometrika*, **85**, 347–361.

Cicchetti, D. V. and Feinstein, A. R. (1990). High agreement but low kappa: II. resolving the two paradoxes. *Journal of Clinical Epidemiology*, **43**, 551–558.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **XX (1)**, 37–46.

Congdon, P. (2003). *Applied Bayesian Modelling*. Wiley, New Jersey.

Consul, P. C. (1989). *Generalized Poisson Distribution: Properties and Applications*. Decker, New York.

Consul, P. C. and Jain, G. C. (1973). A generalization of the Poisson distribution. *Technometrics*, **15**, 791–799.

Cook, J. R. and Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, **90**, 1314–1328.

Dean, C. B., Lawless, J. and Willmot, G. E. (1989). A mixed Poisson-inverse gaussian regression model. *Canadian Journal of Statistics*, **17**, 171–182.

Dunn, A. D. and Davies, S. J. (1998). A note on generating correlated binary variables. *Biometrika*, **85**, 487 – 490.

Espeland, M. A. and Hui, S. L. (1987). A general approach to analyzing epidemiological data that contain misclassification errors. *Biometrics*, **43**, 1001–1012.

Espeland, M. A. and Odoroff, C. L. (1985). Log-linear models for doubly sampled categorical data fitted by the EM algorithm. *Journal of the American Statistical Association*, **80**, 663–670.

Feinstein, A. R. and Cicchetti, D. V. (1990). High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, **43**, 543–549.

Fletcher, R. (1970). A new approach to variable metric algorithms. *Computer Journal*, **13**, 317–322.

Formann, A. K. (1994). Measurement errors in caries diagnosis: some further latent class models. *Biometrics*, **50**, 865–871.

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

Fyffe, H. E., Deery, C. H., Nugent, Z. J., Nuttall, N. M. and Pitts, N. B. (2000). Effect of diagnostic threshold on the validity and reliability of epidemiological caries diagnosis using the Dundee Selectable Threshold Method for caries diagnosis (DSTM). *Community Dentistry and Oral Epidemiology*, **28**, 42–51.

Gastwirth, J. (1987). The practical precision of medical screening procedures: application to polygraph and AIDS antibodies test data. *Statistical Science*, **2**, 110–114.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

Gelman, A. (1996). Inference and monitoring convergence. In Gilks, W. R., Richardson, S. and Spiegelhalter, D. J., editors, *Markov chain Monte Carlo in practice*, pages 132–143. Chapman and Hall: London.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. (1995). *Bayesian Data Analysis*. Chapman and Hall, London.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. (2004). *Bayesian Data Analysis*. Chapman and Hall, London.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–511.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Genton, M. G. (2004). *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*. Chapman & Hall/CRC.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In Bernardo, J.-M., Berger, J. O., Dawid, A. P. and Smith, A. F. M., editors, *Bayesian Statistics*, volume 4, pages 169–193. Proceedings of the Fourth Valencia International Meeting, Oxford: Clarendon Press.

Gilks, W. R. (1996). Full conditional distributions. In Gilks, W. R., Richardson, S. and Spiegelhalter, D. J., editors, *Markov chain Monte Carlo in practice*, pages 75–88. Chapman and Hall: London.

Gleser, L. J. (1990). Improvement of the naive approach to estimation in nonlinear errors-in-variables regression models. In Brown, P. J. and Fuller, W. A., editors, *Statistical Analysis of Measurement Error Models and Application*. American Mathematics Society, Providence.

Goldberg, J. D. (1975). The effects of misclassification on the bias in the difference between two proportions and the relative odds in the fourfold table. *Journal of the American Statistical Association*, **70**, 561–567.

Goldfarb, D. (1970). A family of variable metric methods derived by variational means. *Mathematics of Computation*, **24**, 23–26.

Greenland, S. (1980). The effects of misclassification in the presence of covariates. *American Journal of Epidemiology*, **112**, 564–569.

Greenland, S. (1988). Variance estimation for epidemiologic effect estimates under misclassification. *Statistics in Medicine*, **7**, 745–757.

Greenland, S. and Brenner, H. (1993). Correcting for non-differential misclassification in ecologic analyses. *Applied Statistics*, **42**, 117–126.

Greenland, S. and Robins, J. M. (1985). Confounding and misclassification in the presence of covariates. *American Journal of Epidemiology*, **122**, 495–506.

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

Hall, D. and Berenhaut, K. (2002). Score tests for heterogeneity and overdispersion in zero-inflated Poisson and binomial regression models. *Canadian Journal of Statistics*, **30**, 415–430.

Hartzel, J., Agresti, A. and Caffo, B. (2001). Multinomial logit random effects models. *Statistical Modelling*, **1**, 81–102.

Hausen, H. (1997). Caries prediction – state of the art. *Community Dentistry and Oral Epidemiology*, **25**, 87–96.

Hausman, J. A., Abrevaya, J. and Scott-Morton, F. M. (1998). Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics*, **87**, 239–269.

Hujoel, P. P., Moulton, L. H. and Loesche, W. J. (1990). Estimation of sensitivity and specificity of site-specific diagnostic tests. *Journal of Peridontal Research*, **25**, 193–196.

Israel, R., Rosenthal, J. and Wei, J. (2001). Finding generators for Markov chains via empirical transition matrices, with applications to credit ratings. *Mathematical Finance*, **11**, 245–265.

Johnson, N., Kotz, S. and Kemp, A. W. (1992). *Univariate Discrete Distributions*. Wiley, New York.

Johnson, V. E. and Albert, J. H. (1999). *Ordinal Data Modeling*. Springer, New York.

Klein, H. and Palmer, C. E. (1941). Studies on dental caries. *Journal of Dental Research*, **20**, 203–216.

Klein, H., Palmer, C. E. and Knutson, J. W. (1938). Studies on dental caries. I. Dental status and dental needs of elementary school children. *Public Health Report*, **53**, 751–765.

Küchenhoff, H. and Carroll, R. J. (1997). Segmented regression with errors in predictors: semiparametric and parametric methods. *Statistics in Medicine*, **16**, 169–188.

Kuha, J. (1997). Estimation by data augmentation in regression models with continuous and discrete covariates measured with error. *Statistics in Medicine*, **16**, 189–201.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–14.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174.

Lesaffre, E. and Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random-effects model: an example. *Applied Statistics*, **50**, 325–335.

Lew, R. and Levy, P. (1989). Estimation of prevalence on the basis of screening tests. *Statistics in Medicine*, **8**, 1225–1230.

Lewsey, J. D. and Thomson, W. M. (2004). The utility of the zero-inflated Poisson and zero-inflated negative binomial models: a case study of cross-sectional and longitudinal DMF data examining the effect of socio-economic status. *Community Dentistry and Oral Epidemiology*, **32**, 183–189.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized estimating equation models. *Biometrika*, **73**, 13–22.

Liu, Q. and Pierce, D. A. (1994). A note on Gauss-Hermite quadrature. *Biometrics*, **81**, 624–629.

Liu, X. and Liang, K.-Y. (1991). Adjustment for non-differential misclassification error in generalized linear models. *Statistics in Medicine*, **10**, 1197–1211.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **44**, 226–233.

Luan, X., Pan, W., Gerberich, S. and Carlin, B. (2005). Does it always help to adjust for misclassification of a binary outcome in logistic regression? *Statistics in Medicine*, **24**, 2221–2234.

Magder, L. S. and Hughes, J. P. (1997). Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology*, **146**, 195–203.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, London.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons, Inc., New York.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer.

Mote, V. L. and Anderson, R. L. (1965). An investigation on the effect of misclassification on the properties of χ^2-tests in the analysis of categorical data. *Biometrika*, **52**, 95–109.

Moyniham, P. (2000). Food and factors that protect against dental caries. *Nutrition Bulletin*, **25**, 281–286.

Müller, P. and Roeder, K. (1997). A Bayesian semiparametric model for case-control studies with errors in variables. *Biometrics*, **84**, 523–537.

Mwalili, S., Lesaffre, E. and Declerck, D. (2005). A Bayesian ordinal logistic regression model to correct for inter-observer measurement error in a geographical oral health study. *Journal of the Royal Statistical Society, Series C*, **54**, 77–93.

Nadanovsky, P. and Sheiham, A. (1994). The relative contribution of dental services to the changes and geographical variations in caries status of 5- and 12-year-old children in England and Wales in the 1980s. *Community Dent Health*, **11**, 215–223.

Neuhaus, J. M. (1999). Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*, **86**, 843–855.

Neuhaus, J. M. (2002). Analysis of clustered and longitudinal binary data subject to response misclassification. *Biometrics*, **58**, 675–683.

Neuhaus, J. M., Kalbfleisch, J. D. and Hauck, W. W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review*, **59**, 25–35.

Nocedal, J. and Wright, S. (1999). *Numerical Optimization*. Springer-Verlag, New York.

O'Brien, S. M. and Dunson, D. B. (2004). Bayesian multivariate logistic regression. *Biometrics*, **60**, 739–746.

Paulino, C. D., Silva, G. and Achcar, J. A. (2005). Bayesian analysis of correlated misclassified binary data. *Computational Statistics & Data Analysis*, **49**, 1120–1131.

Paulino, C. D., Soares, P. and Neuhaus, J. (2003). Binomial regression with misclassification. *Biometrics*, **59**, 670–675.

Pieper, K. and Schulte, A. G. (2004). The decline in dental caries among 12-year-old children in Germany between 1994 and 2000. *Community Dental Health*, **21**, 199–206.

Pine, C. M., Pitts, N. B. and Nugent, Z. J. (1997). British Association for the Study of Community Dentistry (BASCD) guidance on sampling for surveys of child dental health. A BASCD coordinated dental epidemiology programme quality standard. *Community Dent Health*, **14**, (Suppl 1):10–17.

Pitts, N. B. (2004). Modern concepts of caries measurement. *Journal of Dental Research*, **83 (Spec Iss) C**, 43–47.

Pitts, N. B., Evans, D. J. and Pine, C. M. (1997). British Association for the Study of Community Dentistry (BASCD) diagnostic criteria for caries prevalence surveys - 1996/97. *Community Dental Health*, **14**, (Suppl 1):6–9.

Pitts, N. B. and Fyffe, H. E. (1988). The effect of varying diagnostic thresholds upon clinical caries data for a low prevalence group. *Journal of Dental Research*, **67**, 592–596.

Prentice, R. L. (1986). Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *Journal of the American Statistical Association*, **81**, 321–327.

Prescott, G. J. and Garthwaite, P. H. (2002). A simple Bayesian analysis of misclassified binary data with validation substudy. *Biometrics*, **58**, 454–458.

Rekaya, R., Weigel, K. A. and Gianola, D. (2001). Threshold model for misclassified binary responses with applications to animal breeding. *Biometrics*, **57**, 1123–1129.

Richardson, S. and Gilks, W. R. (1993a). A Bayesian approach to measurement error problems in epidemiology using conditional independence models. *American Journal of Epidemiology*, **138**, 430–432.

Richardson, S. and Gilks, W. R. (1993b). Conditional independence models for epidemiological studies with covariates measured with error. *Statistics in Medicine*, **12**, 1703–1722.

Roberts, G. O. and Rosenthal, J. S. (2001). Optimal scaling for various Metropolis–Hastings algorithms. *Statistical Science*, **16**, 351–367.

Rosner, B., Willett, W. and Spiegelmann, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine*, **8**, 1051–1070.

Rudolfer, S. M. (1990). A Markov Chain model of extrabinomial variation. *Biometrika*, **77**, 255–264.

Sakamoto, Y., Ishiguro, M. and Kitagawa, G. (1986). *Akaike Information Criterion Statistics*. D. Reidel Publishing Company.

SAS© Institute Inc. (1999–2001). *The SAS System for Windows.* Cary, NC, USA.

Satterthwaite, F. E. (1942). Generalized Poisson distribution. *Annals of Mathematical Statistics*, **13**, 410–417.

Schafer, D. W. (1993). Likelihood analysis for probit regression with measurement errors. *Biometrics*, **80**, 899–904.

Schmid, C. and Rosner, B. (1993). A Bayesian approach to logistic regression model having measurement error following a mixture distribution. *Statistics in Medicine*, **12**, 1141–1153.

Shanno, D. F. (1970). Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, **24**, 647–657.

Sichel, H. S. (1974). On a distribution representing sentence-length in written prose. *Journal of the Royal Statistical Society, Series A*, **137**, 25–34.

Skellam, J. G. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society, Series B*, **10**, 257–261.

Spencer, A. J. (1997). Skewed distribution - new outcome measures. *Community Dentistry and Oral Epidemiology*, **25**, 52–59.

Spiegelhalter, D., Carlin, B., Best, N. and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, **64**, 583–640.

Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D. (2003). *WinBUGS Version 1.4 User Manual.* MRC Biostatistics Unit, UK. URL `http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml`.

Spiegelman, D., Schneeweiss, S. and McDermott, A. (1997). Measurement error correction for logistic regression models with an 'alloyed gold standard'. *American Journal of Epidemiology*, **145**, 184–196.

Stamey, J., Young, D. and Bratcher, T. (2004). Bayesian predictive probability functions for count data that are subject to misclassification. *Biometrical Journal*, **46**, 572–578.

Stefanski (1992). Monotone likelihood ratio of a "faulty-inspection" distribution. *The American Statistician*, **46**, 110–114.

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

Stefanski, L. (1985). The effects of measurement error on parameter estimation. *Biometrica*, **72**, 583–592.

Stein, G. Z., Zucchini, W. and Juritz, J. M. (1987). Parameter estimation for the Sichel distribution and its multivariate extension. *Journal of the American Statistical Association*, **82**, 938–944.

Stephens, D. A. and Dellaportas, P. (1991). Bayesian analysis of generalised linear models with covariate measurement error. In Bernado, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M., editors, *Bayesian Statistics*, volume 4, pages 813 – 830. Oxford University Press, Oxford.

Stiratelli, R., Laird, N. and Ware, H. (1984). Random-effects models for serial observations with binary response. *Biometrics*, **40**, 961–971.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior density by data augmentation. *Journal of the American Statistical Association*, **82**, 528–540.

Tickle, M., Milsom, K. M., Jenner, T. M. and Blinkhorn, A., S. (2003). The geodemographic distribution of caries experience in neighboring and non-fluoridated populations. *Journal of Public Health Dentistry*, **63**, 92–98.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics*, **22**, 1701–1728.

Vanobbergen, J., Martens, L., Lesaffre, E., Bogaerts, K. and Declerck, D. (2001). Assessing the indicators for dental caries in the primary dentition. *Community Dentistry and Oral Epidemiology*, **29**, 424–434.

Vanobbergen, J., Martens, L., Lesaffre, E. and Declerck, D. (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.

Wacholder, S., Armstrong, B. and Hartge, P. (1993). Validation studies using alloyed gold standard. *American Journal of Epidemiology*, **137**, 1251–1258.

Wang, N., Carroll, R. J. and Liang, K. Y. (1996). Quasilikelihood estimation in measurement error models with correlated replica. *Biometrics*, **52**, 401–411.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1–25.

Whittemore, A. S. and Gong, G. (1991). Poisson regression with misclassified counts: application to cervical cancer mortality rates. *Applied Statistics*, **40**, 81–93.

Willett, W. (1989). An overview of issues related to the correction of non-differential exposure measurement error in epidemiologic studies. *Statistics in Medicine*, **8**, 1031–1041.

Willmot, G. (1987). The Poisson-inverse Gaussian distribution as an alternative to the negative binomial. *Scandinavian Actuarial Journal*, **87**, 113–127.

Yoshimura, I. (1991). The effect of measurement error on the dose-response curve. In Hoel, D. and Yanagawa, T., editors, *Biostatistics in Cancer Risk Assessment*, pages 141–145.

Zeger, S. L. and Karim, R. M. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association*, **86**, 79–86.

Zeger, S. L., Liang, K.-Y. and Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, **44**, 1049–1060.