Regression calibration for Cox regression under heteroscedastic measurement error — Determining risk factors of cardiovascular diseases from error-prone nutritional replication data

T. Augustin*
Munich
Department of Statistics

A. Döring
Neuherberg
GSF–National Research Center
for Environment and Health

D. Rummel
Munich
Department of Statistics

Abstract

For instance nutritional data are often subject to severe measurement error, and an adequate adjustment of the estimators is indispensable to avoid deceptive conclusions. This paper discusses and extends the method of regression calibration to correct for measurement error in Cox regression. Special attention is paid to the modelling of quadratic predictors, the role of heteroscedastic measurement error, and the efficient use of replicated measurements of the surrogates. The method is used to analyze data from the German part of the MONICA cohort

*Corresponding author: T. Augustin, Department of Statistics, University of Munich, Ludwigstr. 33, D-80539 Munich, Germany, thomas@stat.uni-muenchen.de

study on cardiovascular diseases. The results corroborate the importance of taking into account measurement error carefully.

Keywords: Error-in-variables; replication data; heteroscedastic measurement error in quadratic variables; Cox model; regression calibration; MONICA/KORA study.

1 Introduction

A widespread problem in applying regression analysis is the presence of measurement error. Often the variables of interest, called *ideal variables* or *gold standard*, cannot be observed directly or measured correctly, and one has to be satisfied with so called *surrogates* (often also named *indicators* or *proxies*), i.e., with somehow related, but different variables. If one ignores the difference between the ideal variables in the model and the observable variables and just plugs in the surrogates instead of the variables ('naive estimation'), then all the estimators must be suspected to be severely biased. Error-in-variables modelling provides a methodology, which is serious about that fact. Based on an error model describing the relation between ideal variables and surrogates, it develops procedures to adjust for the measurement error. Recent surveys on measurement error modelling, also containing many examples from different fields of application, include Cheng and van Ness [1], who concentrate on linear models, and Carroll, Ruppert, and Stefanski [2], Stefanski [3], and Van Huffel and Lemmerling [4], who are concerned with non-linear models.

It should be stressed explicitly that the topic of measurement error is not simply a matter of sloppy research; quite often the 'true value' is unascertainable eo ipso. A typical example is the recording of the protein intakes in surveys on eating habits and their influence on certain diseases. Though much attention is paid to the high quality of the questionnaire and the subsequent procedures, a considerable random distortion in the data can not be avoided. Below we analyze data from the WHO MONICA Augsburg substudy on the surveillance of dietary intake, see [5, 6]. This study, which is embedded into the WHO MONICA project

(MONItoring of trends and determinants in CArdiovascular disease), is concerned with the question whether changes in dietary intake can explain trends in the incidence and mortality of cardiac infarctions. Indeed severe error is present in the measurements of the animal and plant protein intake from a seven day food diary, and so applying Cox regression without adjusting for the error could lead to wrong conclusions. The MONICA Augsburg study is currently continued as the KORA study (Cooperative health research in the area of Augsburg).

Recently the quality of Swedish nutrition data was investigated in [7], where the reproducibility of food frequency measurements of a sample of respondents to the Swedish MONICA study was considered. It may be mentioned that, if such local studies were combined and compared, additional measurement error would arise: It is quite important to take into account the variation in these aggregated observations (cf. [8]).

Covariate measurement error correction in Cox regression is currently an area of intensive and fruitful research (see, in particular, [9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19] — as well as the survey and comparison of basic approaches in [20]).

In this paper we rely on a variant of the regression calibration approach, which is one of the most universal methods to correct for measurement error (see [2, Chapter 3] for a general description). Its basic idea is to run a standard analysis where the unobservable variables are replaced by values predicted from the observable ones. For Cox regression, regression calibration type methods were introduced by Prentice ([21]) and were studied and developed further in [22, 23, 24, 9] and [17].

Here we adapt and extend this method taking into account three general methodological issues, which also deserve special attention in the data analyzed below:

- *Heteroscedastic* measurement error. Recent research in nutritional epidemiology strongly suggests that the measurement error must be expected to vary considerably among the different study participants (cf., e.g., [25, pp. 33-48]).

- The presence of *replication data*. The protein intake measurements are based on diaries, where all food intake had to be recorded in great detail for seven days. Taking for every individual the errors in these measurements as independently and identically distributed gives us the opportunity to estimate the error variances.

- The *non-linearity* of the influence. Pre-studies showed that the effect of protein intake on morbidity and mortality could be nonlinear: both types of extreme intakes, very high as well as very low intakes, could be detrimental, and so it is of great importance to work with quadratic predictors. While introducing non-linearity in the covariates does not encounter much difficulty in the error-free situation, under measurement error it is often hard or even impossible to handle non-linear terms. (For the problems already arising in the linear polynomial model see, e.g., [26]. For some models a general result [27, Theorem 1] can be used to prove even the non-existence of a so-called corrected score function.)

As shown below, the convenience of regression calibration is maintained in this extended setting; still the core parts of the estimation can be done by standard software packages. Applying this correction method shows a complex relationship between naive and corrected estimates. After having adjusted for measurement error, some of the estimates change substantially, others do not. Sometimes there is a high deattenuation, sometimes the absolute values even get smaller. Since, however, regression calibration is known to be only an approximative correction method, reducing the bias but not necessarily producing consistent estimators, we understand our analysis more as a motivation for further methodological development than as the last word on the topic.

The paper is organized as follows: The next section describes our modelling of the replication data. Section 3 adapts the idea of regression calibration to replication data and to quadratic predictors. The application to the MONICA data is reported in Section 4, while Section 5 concludes by sketching some topics for further research.

2 Survival Data with Replicated Covariate Measurements

2.1 The Main Setting

Let n be the sample size and $T_1, \ldots, T_n$ the lifetimes, which may be subject to noninformative independent censorship in the sense of, e.g., [28]. For every $i = 1, \ldots, n$ we split the vector of covariates into a vector X_i and a vector Z_i. All error-prone variables are collected in X_i, while Z_i consists of the correctly measured variables. Let all elements of X_i be measured on a metrical scale, Z_i may contain metrical and categorical covariates in 0/1-coding. Both types of covariates should not be time-varying. With the application below in mind, we additionally consider another vector, denoted by $X_i^{②}$, which contains the squared elements of X_i.

We assume that Cox's ([29]) proportional hazard model describes the relationship between the lifetimes and the covariates; the individual hazard rate $\lambda(t|X_i, Z_i)$ has the form

$$\lambda(t|X_i, Z_i) = \lambda_0(t) \cdot \exp\left(\beta_1' X_i + \beta_2' X_i^{②} + \beta_Z' Z_i\right), \qquad (1)$$

with the unspecified baseline hazard rate $\lambda_0(t)$ and the regression parameter vector $\beta = (\beta_1', \beta_2', \beta_Z')'$.

For X_i, i.e., plant and animal protein in the application discussed below, replicated measurements $W_{i1}, \ldots, W_{ik}$, $k > 1$ (later on, k=7) are available for every unit i. We assume them to follow the additive error model

$$W_{ij} = X_i + U_{ij}, \qquad j = 1, \ldots, k, \quad i = 1, \ldots, n, \qquad (2)$$

and make the usual assumptions: The errors (U_{ij}), $j = 1, \ldots, k$, $i = 1, \ldots, n$ should have zero mean (see however the last paragraphs of this section) and they should be independent among each other as well as of $X_1, \ldots, X_n$ and $T_1, \ldots, T_n$. It will prove important to allow for heteroscedasticity of the errors, where, for i fixed, $U_{i1}, \ldots, U_{ik}$ are i.i.d., but the covariance matrix Σ_i may vary among the units $i = 1, \ldots, n$. The common covariance matrix in the homoscedastic case will be denoted by Σ.

In a naive analysis, for every unit i, the individual average

$$\overline{W}_i := \frac{1}{k} \sum_{j=1}^{k} W_{ij}$$

would function as the surrogate for X_i. Additionally defining

$$\overline{U}_i := \frac{1}{k} \sum_{j=1}^{k} U_{ij}$$

leads us back to the classical error model

$$\overline{W}_i = X_i + \overline{U}_i, \quad i = 1, \ldots, n, \tag{3}$$

with $\mathbb{E}(\overline{U}_i) = 0$ and $\mathbb{V}(\overline{U}_i) = \frac{1}{k} \cdot \Sigma_i$. The particular attractiveness of replication data is based on the fact that the measurement error variances can be estimated from the data. Therefore, in contrast to most cases relying on the classical error model, it is possible here to avoid additional assumptions, which are quite often difficult to justify.

2.2 A Note on Systematic Measurement Error

Before addressing this topic, the assumption $\mathbb{E}(U_{ij}) = 0$ deserves some attention. If it is violated, i.e. if systematic measurement error with $\mathbb{E}(U_{ij}) = a \neq 0$ with a unknown is present, then it becomes important to distinguish whether the covariates act merely linearly or also in a nonlinear way. In order to bring out this point most clearly, concentrate on the following special case: X_i is one-dimensional, there are no error-free covariates Z_i, and there is only a deterministic error a so that (3) reads as

$$\overline{W}_i = X_i + a, \quad i = 1, \ldots, n.$$

In the case of no quadratic influence, where $\beta_2 \equiv 0$ a priori, Relation (1) can be written as

$$\begin{aligned} \lambda_0(t) \cdot \exp(\beta_1 \overline{W}_i) &= \lambda_0(t) \cdot \exp(\beta_1 a + \beta_1 X_i) \\ &=: \lambda_0^*(t) \cdot \exp(\beta_1 X_i)\,. \end{aligned} \tag{4}$$

Therefore, the naive partial likelihood estimator based on replacing X_i by $\overline{W}_i$ still estimates β_1 consistently, and a bias only occurs in the estimation of $\lambda_0(t)$, where the naive standard methods estimate $\lambda_0^*(t) = \lambda_0(t) \cdot \exp(\beta_1 a)$ instead of $\lambda_0(t)$ itself. If, however, quadratic terms are taken into account, then we have to consider

$$\begin{aligned} &\lambda_0(t) \cdot \exp(\beta_1 \overline{W}_i + \beta_2 \overline{W}_i^2) \\ =\ &\lambda_0(t) \cdot \exp(\beta_1 a + \beta_1 X_i + \beta_2 X_i^2 + 2\beta_2 a X_i + \beta_2 a^2) \\ =:\ &\lambda_0^{**}(t) \cdot \exp(\beta_1 X_i + \beta_2 X_i^2 + 2\beta_2 a X_i)\,, \end{aligned}$$

and also inconsistencies in the estimation of the regression parameters must be expected.

3 Regression Calibration under Replication Data

3.1 The Basic Concept

Regression calibration (cf., in particular, [2, Chapter 3]) is an universally applicable, easy-to-handle method to adjust for measurement error. The main idea is to utilize the surrogate $\overline{W}_i$, together with the error-free variable Z_i, to predict the corresponding value of the unobservable variable X_i, and then to proceed with a standard analysis where X_i is replaced by its prediction $\widehat{X_i}$.

Applying this concept, the vector $(X_i', Z_i')'$ of covariates is assumed to be i.i.d., with unknown mean vector $(\mu_X', \mu_Z')'$ and unknown covariance matrix

$$\begin{pmatrix} \Sigma_{X,X} & \Sigma_{X,Z} \\ \Sigma_{X,Z}' & \Sigma_{Z,Z} \end{pmatrix}.$$

Based on Relation (3), the best linear prediction of X_i given $\overline{W_i}$ and Z_i is

$$\begin{aligned} \widehat{X_i} &= \mu_X + (\Sigma_{X,X}\ \Sigma_{X,Z}) \\ &\times \begin{pmatrix} \Sigma_{X,X} + \frac{1}{k}\Sigma_i & \Sigma_{X,Z} \\ \Sigma_{X,Z}' & \Sigma_{Z,Z} \end{pmatrix}^{-1} \begin{pmatrix} \overline{W}_i - \mu_X \\ Z_i - \mu_Z \end{pmatrix}. \end{aligned} \tag{5}$$

If additionally X_i, U_i and Z_i are Gaussian then (5) is exactly the conditional expectation of X_i given W_i and Z_i.

Under replication data all nuisance parameters in (5), i.e. the parameters μ_X, μ_Z, Σ_X and Σ_Z of the distribution of $(X_i', Z_i')'$ as well as the measurement error variances Σ_i, can be estimated powerfully from the data. We firstly adopt the procedure for the homoscedastic case ($\Sigma_i \equiv \Sigma$), taken from [2, p. 47f.], and then present the generalization to the heteroscedastic case.

3.2 The Case of Homoscedastic Measurement Error

Equation (3) immediately suggests the overall mean

$$\overline{W} := \frac{1}{n}\sum_{i=1}^{n} \overline{W}_i \tag{6}$$

as an unbiased estimator for μ_X; analogously μ_Z is estimated by $\overline{Z} := \frac{1}{n}\sum_{i=1}^{n} Z_i$.

In order to derive estimators for the other parameters, it is illuminating to embed the situation under homoscedastic measurement error into the theory of design of experiments. Then (2) is reinterpreted as a one-factorial model with a random effect (e.g., Toutenburg ([30], pp. 147-150)), yielding the estimators

$$\widehat{\Sigma} = \frac{1}{n(k-1)} \cdot \sum_{i=1}^{n}\sum_{j=1}^{k} \left(W_{ij} - \overline{W}_i\right)\left(W_{ij} - \overline{W}_i\right)' \tag{7}$$

$$\widehat{\Sigma}_{X,X} = \left(\frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(\overline{W}_i - \overline{W}\right)\left(\overline{W}_i - \overline{W}\right)'\right) - \frac{1}{k} \cdot \widehat{\Sigma} \tag{8}$$

$$\widehat{\Sigma}_{X,Z} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(\overline{W}_i - \overline{W}\right)\left(Z_i - \overline{Z}\right)' \tag{9}$$

$$\widehat{\Sigma}_{Z,Z} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(Z_i - \overline{Z}\right)\left(Z_i - \overline{Z}\right)'. \tag{10}$$

3.3 The Case of Heteroscedastic Measurement Error

Under heteroscedastic measurement error, the relation

$$\begin{aligned} \mathbb{V}(U_{ij}) = \mathbb{V}(U_{ij}|X_i) &= \mathbb{V}(X_i|X_i) + \mathbb{V}(U_{ij}|X_i) \\ &= \mathbb{V}(X_i + U_{ij}|X_i) \\ &= \mathbb{V}(W_{ij}|X_i) \end{aligned}$$

plays a central role. It provides

$$\widehat{\Sigma}_i = \frac{1}{(k-1)} \cdot \sum_{j=1}^{k} \left(W_{ij} - \overline{W}_i\right)\left(W_{ij} - \overline{W}_i\right)', \quad i = 1, \ldots, n, \tag{11}$$

as an estimator for the error covariance matrices Σ_i at the individual level. $\Sigma_{X,Z}$ and $\Sigma_{Z,Z}$ are estimated in the same way as in (9) and in (10). To get an idea how to estimate $\Sigma_{X,X}$ it is helpful to apply the covariance decomposition formula

$$\begin{aligned} \mathrm{Cov}(\overline{W}_i[l_1], \overline{W}_i[l_2]) &= \mathrm{Cov}\left(\mathbb{E}\left(\overline{W}_i[l_1]\,|X_i\right), \mathbb{E}\left(\overline{W}_i[l_2]\,\middle|\, X_i\right)\right) \\ &+ \mathbb{E}\left(\mathrm{Cov}\left(\overline{W}_i[l_1], \overline{W}_i[l_2]\,\middle|\, X_i\right)\right) \end{aligned}$$

to every pair $(\overline{W}_i[l_1], \overline{W}_i[l_2])$ of components of $\overline{W}_i$. For the covariance matrices this finally yields, in somewhat informal notation, the relation

$$\mathbb{V}(\overline{W}_i) = \mathbb{V}\left(\mathbb{E}(\overline{W}_i|X_i)\right) + \mathbb{E}\left(\mathbb{V}(\overline{W}_i|X_i)\right) = \mathbb{V}(X_i) + \mathbb{E}\left(\mathbb{V}(\overline{U}_i|X_i)\right),$$

which suggests to generalize (8) by using the pooled version

$$\widehat{\Sigma}_{X,X} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(\left(\overline{W}_i - \overline{W}\right)\left(\overline{W}_i - \overline{W}\right)' - \frac{1}{k} \cdot \widehat{\Sigma}_i\right). \tag{12}$$

3.4 Calibrating the Quadratic Part

The most consequential way to deal with the quadratic part $X_i^{②}$ is to replace every component $(X_i[l])^2$ of $X_i^{②}$ by the square $\left(\widehat{X_i}[l]\right)^2$ of the corresponding component $\widehat{X_i}[l]$ of $\widehat{X_i}$. Alternatively to this procedure, which is also pursued in the analysis below, one could prefer to calibrate $(X_i[l])^2$ 'directly' by an appropriate approximation to $\mathbb{E}((X_i[l])^2|W_i, Z_i)$. By means of the relation

$$\begin{aligned}\mathbb{E}\Big((X_i[l])^2|W_i, Z_i\Big) &= \Big(\mathbb{E}\big(X_i[l]|W_i, Z_i\big)\Big)^2 + \mathbb{V}(X_i[l]|W_i, Z_i) \\ &\approx \Big(\widehat{X_i}[l]\Big)^2 + \mathbb{V}(X_i[l]|W_i, Z_i)\end{aligned}$$

and arguments very similar to (4), both approaches lead to the same estimator for β as long as $\mathbb{V}(X_i[l]|W_i, Z_i)$ does not depend on i. This is the case, for instance, if under homoscedastic Gaussian measurement error $(X_i', Z_i')'$ are Gaussian, too.

4 Application to the MONICA Data

4.1 The Data

Within the WHO MONICA project (MONItoring of trends and determinants in CArdiovascular disease) also the influence of nutrition was considered. We analyze data from a panel of the WHO MONICA substudy on the surveillance of dietary intake, conducted in 1984/1985 in Southern Germany, which is currently continued as the KORA study (Cooperative health research in the area of Augsburg), see [5, 6]. A subpopulation of 899 male respondents, aged from 45 to 65, filled in a comprehensive diary. For seven consecutive days all meals had been listed in detail. By using a nutritional data base also containing standard recipes, nutritional variables were derived from the raw data given in everyday units like ladle or gram of certain ingredients. Among other questions the role of plant protein intake (PLANT in the tables below) and animal protein intake (ANIMAL) was investigated. Though high attention has been paid to the exactness of the measurement procedure, substantial error in the calculation of protein intake is unavoidable, and so we applied the correction methods developed above to adjust for it.

By a mortality and morbidity follow-up for more than 10 years, the respondents' first cardiac infarctions (total number 71 of 858 observations) and deaths (114 cases of 892 observations[1]) had been registered.[2] The main interest focused on the influence protein intake had on the response variable which was defined as age at the event. In the analysis also confounders were incorporated, namely cholesterol (mg/dl) (CHOL), daily alcohol consumption (g/day) (ALC) as continuous variables, as well as hypertension (HYPER) and smoking[3] (SMOKER) as categorical variables (1=yes, 0=no). The measurement error in these variables may be expected to be quite low compared to that in the protein intakes, and so the confounders were treated as error free. The estimated regression coefficients are written in the form $\hat{\beta}[VARIABLE]$, i.e., $\hat{\beta}[PLANT]$, $\hat{\beta}[ANIMAL]$, etc.

4.2 The Results

Table 1 and Table 2 summarize the results of naive and corrected proportional hazards regression. The first two columns belong to the naive analysis, which used the seven-days averages of calculated animal protein intake and of calculated plant protein intake as surrogates for the true corresponding intake. They contain the naive estimates and the p-values based on them.[4] Column 3 and 5 report the corrected estimates after having adjusted for homoscedastic measurement error by the methods of Subsection 3.2, and for heteroscedastic measurement error along the lines of Subsection 3.3, respectively. In Column 4 and 6 also "approximative p-values" are given, which, however, have to be used with particular reservation here. They are based on the standard errors which usual software calculates after every X_i was replaced by the corre-

[1] The number of overall observations slightly differs for the two events, because for some units there was no information about morbidity, but it could be found out whether they died or survived the follow-up period.

[2] The median of the follow-up times with respect to the occurrence of infarction was 2302 days for the cases and 3996 days for the censored observations. The median of the follow-up times concerning the death event was 2598.5 days for the cases and 4006 days for the censored observations, respectively.

[3] In this analysis persons who are currently smoking or are ex-smokers were summarized into the smoker category.

[4] It may be noted explicitly that not only the naive estimators of the regression parameters are inconsistent, but also the estimators of the standard error.

sponding $\widehat{X_i}$; they are only meant to give a very rough impression and should not be taken literally. Correct estimators for the standard error of regression calibration estimators are not straightforwardly found (cf. [2, Section 3.5 and Subsection 3.12.2]), and so we used those easy available values as a rule of thumb to judge the significance. Though they are not correct, they still should give an impression of the correct magnitude.

In order to illustrate the overall influence of animal and plant protein intake on morbidity and mortality, it is helpful to look at the functions

$$f(x) = \hat{\beta}[ANIMAL] \cdot x + \hat{\beta}[(ANIMAL)^2] \cdot x^2 \quad (13)$$

$$g(y) = \hat{\beta}[PLANT] \cdot y + \hat{\beta}[(PLANT)^2] \cdot y^2 \,. \quad (14)$$

They describe the effect of the animal protein intake x, and of the plant protein intake y, respectively, on the predictor in the hazard function in (1). The domains of x and y are chosen such that they cover approximately the whole range of the observed values. These functions are plotted below in Figure 3, where the dotted and dashed line corresponds to the naive estimation. The results, after having adjusted for homoscedastic or heteroscedastic measurement error, are plotted by thin and thick solid lines, respectively.

4.2.1 The Naive Analysis

For the naive analysis the seven-days averages of calculated animal protein intake and of calculated plant protein intake were used as surrogates for the true corresponding intake in a proportional hazards regression. The naive analysis judges the linear and quadratic terms for animal protein to be significant at the five percent level, and cholesterol to have a highly significant influence on morbidity. For mortality the estimates $\hat{\beta}[PLANT]$ and $\hat{\beta}[(PLANT)^2]$ are significant at least at the ten percent level, and hypertension becomes highly significant.

	naive estimation		homoscedastic error		heteroscedastic error	
	estimate	p-value	estimate	p-value	estimate	p-value
$ANIMAL$	$-5.62 \cdot 10^{-5}$	0.0366	$-1.07 \cdot 10^{-4}$	0.0424	$-5.49 \cdot 10^{-5}$	0.2785
$PLANT$	$-2.77 \cdot 10^{-5}$	0.7887	$-1.32 \cdot 10^{-5}$	0.9298	$-6.60 \cdot 10^{-5}$	0.6415
$(ANIMAL)^2$	$4.68 \cdot 10^{-10}$	0.0100	$9.09 \cdot 10^{-10}$	0.0161	$5.27 \cdot 10^{-10}$	0.1743
$(PLANT)^2$	$1.47 \cdot 10^{-11}$	0.9938	$-4.10 \cdot 10^{-10}$	0.8822	$4.51 \cdot 10^{-10}$	0.8697
$CHOL$	$8.28 \cdot 10^{-3}$	0.0008	$8.14 \cdot 10^{-3}$	0.0010	$7.81 \cdot 10^{-3}$	0.0015
$HYPER$	$4.60 \cdot 10^{-1}$	0.0627	$4.70 \cdot 10^{-1}$	0.0578	$4.71 \cdot 10^{-1}$	0.0573
$SMOKER$	$8.79 \cdot 10^{-1}$	0.0328	$8.68 \cdot 10^{-1}$	0.0344	$8.35 \cdot 10^{-1}$	0.0406
ALC	$8.00 \cdot 10^{-5}$	0.9831	$5.01 \cdot 10^{-5}$	0.9895	$-2.03 \cdot 10^{-5}$	0.9957

Table 1: Estimates for the influence on morbidity

	naive estimation		homoscedastic error		heteroscedastic error	
	estimate	p-value	estimate	p-value	estimate	p-value
$ANIMAL$	$-1.01 \cdot 10^{-5}$	0.7296	$-1.54 \cdot 10^{-5}$	0.7862	$-6.43 \cdot 10^{-6}$	0.8932
$PLANT$	$-1.15 \cdot 10^{-4}$	0.0604	$-1.61 \cdot 10^{-4}$	0.0669	$-1.72 \cdot 10^{-4}$	0.0339
$(ANIMAL)^2$	$4.55 \cdot 10^{-11}$	0.8358	$8.23 \cdot 10^{-11}$	0.8501	$3.41 \cdot 10^{-11}$	0.9298
$(PLANT)^2$	$2.07 \cdot 10^{-9}$	0.0447	$2.94 \cdot 10^{-9}$	0.0500	$3.16 \cdot 10^{-9}$	0.0323
$CHOL$	$7.34 \cdot 10^{-4}$	0.7387	$6.77 \cdot 10^{-4}$	0.7597	$5.78 \cdot 10^{-4}$	0.7934
$HYPER$	$5.43 \cdot 10^{-1}$	0.0047	$5.41 \cdot 10^{-1}$	0.0049	$5.42 \cdot 10^{-1}$	0.0049
$SMOKER$	$6.79 \cdot 10^{-1}$	0.0231	$6.77 \cdot 10^{-1}$	0.0236	$6.97 \cdot 10^{-1}$	0.0204
ALC	$3.00 \cdot 10^{-3}$	0.2924	$3.06 \cdot 10^{-3}$	0.2832	$2.86 \cdot 10^{-3}$	0.3162

Table 2: Estimates for the influence on mortality

The decisive question following the naive analysis now is: are these result still valid if one takes into account the substantial measurement error which is naturally inherent in the protein intake?

4.2.2 Adjusting for Homoscedastic Measurement Error

First the homoscedastic error model is considered. In order to obtain corrected estimates the regression calibration method based on (5) and the estimators from (7) to (10) are applied. Column 3 and 4 of Table 1 report the corrected estimates for the influence on morbidity. In comparison to the naive estimates the effects of animal protein are estimated about twice as high; this results in the thin solid line in Figure 3a) below. The point of minimal risk (x=59038) is about the same as in the naive analysis (x=60079), and also the zeros are equal in essence, but the curve is much steeper. $\hat{\beta}[PLANT]$ is half as high as the naive estimate. Now $\hat{\beta}[(PLANT)^2]$ has a negative sign, too. The corresponding function $g(y)$, which is depicted as the thin solid line in Figure 3b), is concave and decreasing in y in a monotone way: the higher the plant intake the higher is the reduction of the risk by an additional unit of intake.

The role of the confounders is more or less the same. The estimated strong influence of hypertension and smoking is confirmed. The regression parameter for alcohol intake changes its sign, but it remains insignificant.

Turning to mortality (cf. Table 2), the absolute values of the regression parameters of the linear and the quadratic terms in the protein variables become higher by factors between 1.4 and 1.8, the effects of the confounders remain unchanged in essence. Figure 3c) and Figure 3d) show the corresponding curves, which are of the same shape as those from the naive analysis, but run steeper again.

4.2.3 Adjusting for Heteroscedastic Measurement Error

As discussed above, the presence of replication data also allows, for every unit i, $i = 1, \ldots, n$, to estimate the covariance matrix Σ_i of the error variable in animal protein intake and in plant protein intake at the individual level (cf. Equation (11)). Even if one takes into account that only seven observations are available to estimate Σ_i, the variation in the estimated variances (Figure 1 and Figure 2) is high enough that a detailed study of heteroscedastic measurement error is promising.

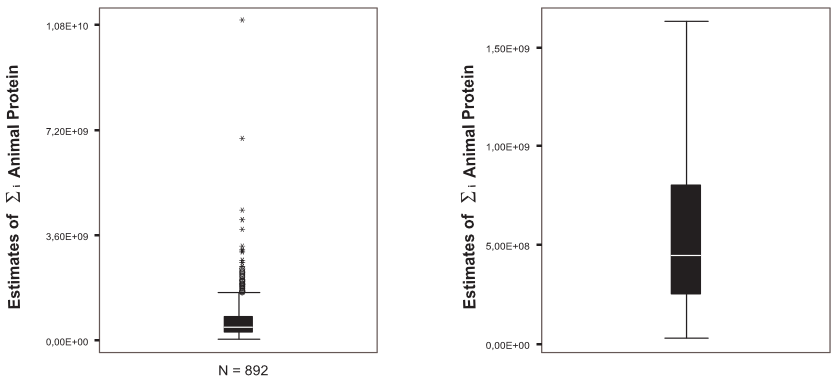

Figure 1: Estimated individual error variances for animal protein intake: overall and detail figure.

The last two columns in Table 1 refer to the corrected estimates for morbidity, the corresponding curves are shown by the thick solid lines in Figure 3a) and Figure 3b).

Compared to the analysis assuming homoscedastic measurement error, the absolute values of the estimates of the regression coefficients for the linear and the quadratic terms in animal protein intake are attenuated, indeed they are even closer to the results from the naive analysis. The curve grows flatter (cf. Figure 3a), the point of minimal risk and the second zero are shifted to the left: from about x=60000 to x=52408, and from about x=120000 to x=104098, respectively. In contrast to this, the quadratic nature of the influence of plant protein becomes much clearer. The

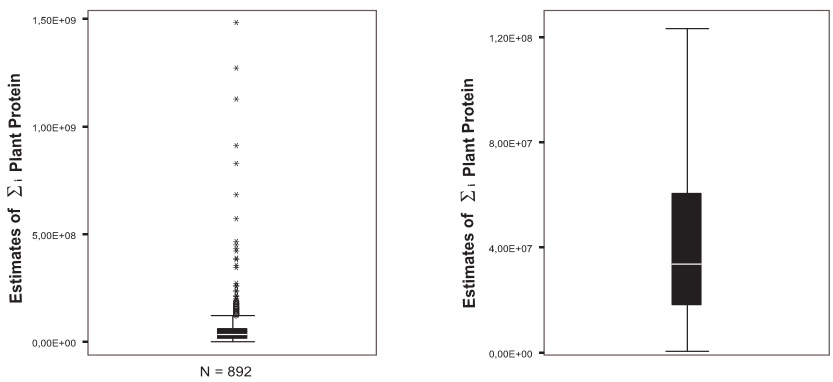

Figure 2: Estimated individual error variances for plant protein intake: overall and detail figure.

regression coefficient for the quadratic term now again has a positive sign, its value is about 30 times as high as in the naive analysis.

As can also be seen in Figure 3b), the risk is still decreasing with increasing plant protein intake, but now the curve is clearly convex: the relative gain in risk reduction becomes the smaller the higher the intake is, and there would be a border value (outside the domain of the data, at y=73241), where further intake would increase the risk again. Correcting for heteroscedastic measurement error in the estimation of mortality confirms the results obtained from the homoscedastic error model for plant protein intake (cf. also Figure 3d)). The absolute values of the estimated coefficients of animal protein intake are by the factor 2.4 lower, which results in a much flatter curve in Figure 3c).

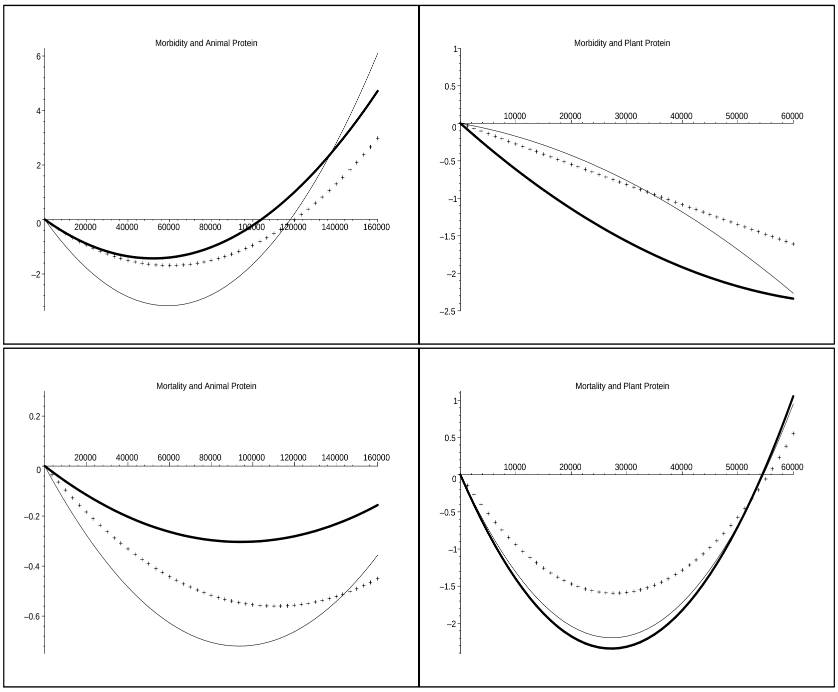

Figure 3: Estimated overall influence of animal/plant protein intake on morbidity and mortality (cf. (13) and (14)), calculated from the naive estimates (dotted line), from the estimates after having corrected for homoscedastic measurement error (thin solid line), and from the estimates after having corrected for heteroscedastic measurement error (thick solid line), respectively.

It is also worth mentioning that morbidity and mortality differ with respect to the consequences a certain amount of protein intake has. High plant protein intake considerably reduces the risk of cardiac infarction, but increases the risk of death. In the case of animal protein the intake which minimizes the risk of death (x=94263 for the heteroscedastic error model) has already a rather high risk for cardiac infarction.

5 Concluding Remarks

We discussed an extended version of regression calibration to correct for possibly heteroscedastic measurement error in Cox regression with a quadratic predictor when replication data are available. This method was applied to a part of the MONICA Augsburg survey to study the influence of eating habits on cardiovascular diseases.

It has become clear how important it is to take into account measurement error carefully. In particular under heteroscedastic measurement error there is a complex relationship between naive and corrected estimation, which may alter the estimates substantially. Nevertheless, the results reported here must be taken only as a first step towards a comprehensive analysis, suggesting and motivating further research in several directions. Four topics should be mentioned explicitly:

First of all, it must not be forgotten that the regression calibration method is only an approximate method, reducing the bias of naive analysis but not necessarily producing consistent estimators. Furthermore, the parameter estimates have to be interpreted in relative terms because correct estimators for their standard errors are missing. To derive such appropriate estimators is demanding (cf. [2, Section 3.5 and Subsection 3.12.2]), an interesting alternative would be bootstrapping.

Secondly, alternative correction methods should be applied, in order to justify, or to correct, the preliminary results obtained here. Of special interest here is a so-to-say dynamic regression calibration procedure, developed by [17], where at every failure time only those units are taken into account which still are under risk (cf. also [31]).

Another powerful method to correct for homoscedastic measurement error in the Cox model was developed by Nakamura ([32]) and extended to heteroscedastic error by [19]. However, prior to applying this method, further theoretic development is needed, in order to be able to model the quadratic influence of the covariates. The inherent restriction to linear predictors is also the main hurdle for an application of Huang and Wang's nonparametric functional correction method (cf. [15]), which would provide an appealing alternative to utilize replicated measurements.

There are good reasons to doubt the assumption made above that the measurement error should be independent of the true protein intake, and so more complex error models deserve special attention (cf., e.g., [33, 34, 35]).

The third issue to keep in mind is that valuable insights in the data may be gained by applying accelerated failure time models instead of Cox's proportional hazards model. Techniques for measurement error correction in such survival models have not yet received much attention. One of the very rare exceptions is [36] where Nakmura illustrates his general method of so-called corrected score functions with members of the exponential family. His approach is generalized to possibly censored Weibull distributed lifetimes in [37]. A procedure to correct for covariate measurement error in the nonparametric log-linear lifetime model is suggested by [38], while [39, Chapter 5f.] proposes two methods for corrected quasi-likelihood estimation in arbitrary parametric accelerated failure time models. As discussed there, the latter approaches need some non-standard treatment of censored observations, but have, on the other hand, the advantage of being able to take also error-prone lifetimes into account.

The last item to be mentioned is the most difficult one: Eating habits may change! Even if the X_i to be measured by the diary could be determined exactly, this measurement would only stem from a cursory glance at a process developing over time. Morbidity and mortality is also affected by the intake before as well as after the recording of the eating habits. This leads to the superposition of the heteroscedastic measurement error treated here with a complex kind of measurement error where a time-dependent covariate is only observed at a certain time point. (Compare for this also [40], considering a Cox model where a time dependent covariate is only observed irregularly.)

Acknowledgements

We are very grateful to Helmut Küchenhoff and Hans Schneeweiß for many helpful comments. Thomas Augustin and David Rummel thank the Deutsche Forschungsgemeinschaft (DFG) for financial support.

6 References

1. Cheng C-L, van Ness JW. *Statistical regression with measurement error.* Arnold: London, 1999.

2. Carroll RJ, Ruppert D, Stefanski LA. *Measurement error in nonlinear models.* Chapman and Hall: London, 1995.

3. Stefanski LA. Measurement error models. *Journal of the American Statistical Association* 2000; **95**(452): 1353-1358.

4. Van Huffel S, Lemmerling P (eds). *Total least squares and errors-in-variables modeling: analysis, algorithms and applications.* Kluwer: Dordrecht, 2002.

5. Döring A, Kußmaul B. *Ernährungsdeterminanten des Herzinfarktrisikos.* Report GSF-Fe-7629. GSF — National Research Center for Environment and Health: Neuherberg, 1997.

6. Winkler G, Döring A, Keil U. Selected nutrient intakes of middle-aged men in Southern Germany: Results from the WHO MONICA Augsburg Dietary Survey of 1984/ 1985. *Annals of Nutrition and Metabolism* 1991; **35**(5): 284-291.

7. Johansson I, Hallmans G, Wikman A, Biessy C, Riboli E, Kaaks R. Validation and calibration of food-frequency questionnaire measurement in the Northern Sweden health and disease cohort. *Public Health Nutrition* 2002; **5**(3): 487-496.

8. Kulathinal SB, Kuulasmaa K, Gasbarra D. Estimation of an errors-in-variables regression model when the variances of the measurement errors vary between the observations. *Statistics in Medicine* 2002; **21**(8): 1089-1101.

9. Wang CY, Hsu L, Feng ZD, Prentice RL. Regression calibration in failure time regression. *Biometrics* 1997; **53**(1): 131-145.

10. Hu P, Tsiatis A, Davidian M. Estimating the parameters in the Cox model when covariate variables are measured with error. *Biometrics* 1998; **54**(4): 1407-1419.

11. Buzas JS. Unbiased scores in proportional hazards regression with covariate measurement error. *Journal of Statistical Planning and Inference* 1998; **67**(2): 247-257.

12. Kong FH, Huang W, Li X. Estimating survival curves under proportional hazards model with covariate measurement errors. *Scandinavian Journal of Statistics* 1998; **25**(4): 573-587.

13. Kong FH. Adjusting regression attenuation in the Cox proportional hazards model, *Journal of Statistical Planning and Inference* 1999; **79**(1): 31-44.

14. Kong FH, Gu M. Consistent estimation in Cox proportional hazards model with covariate measurement errors. *Statistica Sinica* 1999; **9**(4): 953-969.

15. Huang Y, Wang CY. Cox regression with accurate covariates unascertainable: A nonparametric-correction approach. *Journal of the American Statistical Association* 2000; **95**(452): 1209-1219.

16. Zhou H, Wang CY. Failure time regression with continuous covariates measured with error. *Journal of the Royal Statistical Society Series B* 2000; **62**(4): 657-665.

17. Xie SX, Wang CY, Prentice RL. A risk set calibration method for failure time regression by using a covariate reliability sample. *Journal of the Royal Statistical Society Series B* 2001; **63**(4): 855-870.

18. Hu C, Lin DY. Cox regression with covariate measurement error. *Scandinavian Journal of Statistics* 2002; **29**(4): 637-655.

19. Augustin T. An exact corrected log-likelihood function for Cox's proportional hazards model under measurement error and some extensions. To appear in: *Scandinavian Journal of Statistics* 2003.

20. Augustin T, Schwarz R. Cox's proportional hazards model under covariate measurement error — A review and comparison of methods. In: S. Van Huffel and P. Lemmerling

(eds): *Total Least Squares and Errors-in-Variables Modeling: Analysis, Algorithms and Applications.* Kluwer: Dordrecht, 2002; pp 175-184.

21. Prentice RL. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* 1982; **69**(2): 331-342.

22. Pepe MS, Self SG, Prentice RL. Further results in covariate measurement errors in cohort studies with time to response data. *Statistics in Medicine* 1989; **8**(9): 1167-1178.

23. Clayton DG. Models for the analysis of cohort and case-control studies with inaccurately measured exposures. In: JH Dwyer, M Feinleib, P Lipsert et al. (eds): *Statistical Models for Longitudinal Studies of Health.* Oxford University Press: New York, 1991; pp 301-331.

24. Hughes MD. Regression dilution in the proportional hazards model. *Biometrics* 1993; **49**(4): 1056-1066.

25. Willett W. *Nutritional epidemiology.* Oxford University Press: New York, 1998[2].

26. Cheng C-L, Schneeweiß H. The polynomial regression with errors in the variables. *Journal of the Royal Statistical Society Series B* 1998; **60**(1): 189-199.

27. Stefanski LA. Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models. *Communications in Statistics — Theory and Methods* 1989; **18**(12): 4335-4358.

28. Kalbfleisch JD, Prentice RL. *The Statistical analysis of failure time data.* Wiley: New York, 2002[2].

29. Cox DR. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society Series B* 1972; **34**: 187-220.

30. Toutenburg H. *Statistical analysis of designed experiments.* Springer: New York, 2002[2].

31. Wang CY, Xie SX, Prentice RL. Recalibration based on an approximate relative risk estimator in cox regression with missing covariates. *Statistica Sinica* 2001; **11**(4): 1081-1104.

32. Nakamura T. Proportional hazards model with covariates subject to measurement error. *Biometrics* 1992; **48**(3): 829-838.

33. Heitmann BL, Lissner L. Dietary underreporting by obese individuals — is it specific or non-specific? *British Medical Journal* 1995; **311**(7011): 986-989.

34. Prentice RL. Measurement error and results from analytic epidemiology: dietary fat and breast cancer. *Journal of the National Cancer Institute* 1996; **88**(23): 1738-1747.

35. Carroll RJ, Freedman LS, Kipnis V, Li L. A new class of measurement error models, with applications to dietary data. *Canadian Journal of Statistics* 1998; **26**(3): 467-477.

36. Nakamura T. Corrected score functions for errors-in-variables models: Methodology and application to generalized linear models. *Biometrika* 1990; **77**(1): 127-137.

37. Gimenez P, Bolfarine H, Colosimo EA. Estimation in Weibull regression model with measurement error. *Communications in Statistics — Theory and Methods* 1999; **28**(2): 495-510.

38. Wang Q. Estimation of linear error-in-covariables models with validation data under random censorship. *Journal of Multivariate Analysis* 2000; **74**(2): 245-266.

39. Augustin T. *Survival analysis under measurement error.* Habilitation (post-doctorial) thesis. Department of Statistics: University of Munich, 2002.

40. de Bruijne MHJ, le Cessie S, Kluin-Neemans HC, van Houwelingen HC. On the use of Cox regression in the presence of an irregularly observed time-dependent covariate. *Statistics in Medicine* 2001; **20**(24): 3817-3829.