Electronic Journal of Statistics
Vol. 12 (2018) 4032–4056
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1489

Categorizing a continuous predictor subject to measurement error[*]

Betsabé G. Blas Achic

Departamento de Estatística, Universidade Federal de Pernambuco, Av. Prof. Moraes Rego, 1235 – Cidade Universitária, Recife-PE-Brasil, CEP: 50670-901
e-mail: betsabe@de.ufpe.br

Tianying Wang[†]

Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143
e-mail: tianying@stat.tamu.edu

Ya Su

Department of Statistics, University of Kentucky, Lexington, KY, 40536-0082
e-mail: syusapp@gmail.com

Victor Kipnis and Kevin Dodd

Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, MD
e-mail: kipnisv@mail.nih.gov; doddk@mail.nih.gov

Raymond J. Carroll

Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143, and School of Mathematical Sciences, University of Technology Sydney, Broadway NSW 2007
e-mail: carroll@stat.tamu.edu

Abstract: Epidemiologists often categorize a continuous risk predictor, even when the true risk model is not a categorical one. Nonetheless, such categorization is thought to be more robust and interpretable, and thus their goal is to fit the categorical model and interpret the categorical parameters. We address the question: with measurement error and categorization, how can we do what epidemiologists want, namely to estimate the parameters of the categorical model that would have been estimated if the true predictor was observed? We develop a general methodology for such an analysis, and illustrate it in linear and logistic regression. Simulation studies are presented and the methodology is applied to a nutrition data set. Discussion of alternative approaches is also included.

Keywords and phrases: Categorization, differential misclassification, epidemiology practice, inverse problems, measurement error.

Received January 2018.

[*] Blas and Wang should be considered joint first authors.

[†] *Currently at*: Biostatistics Department, Columbia University

1. Introduction

Fitting models by categorizing a continuous risk predictor is a common practice in epidemiology. Among many recent examples, see [20, 19, 1, 5, 10] and [25]. A look at current issues of epidemiology journals will uncover many more examples. An important issue is that, generally in these problems, there are many covariates other than the main risk predictor.

The appeal of categorization in interpreting results is clear. If we have a risk predictor X, and we categorize it into J levels $(C_1, ..., C_J)$, one can compare the highest level of the predictor, C_J, to the lowest level, C_1, and if they are statistically significantly different, one can then conclude that it is better to be in the class that has the lowest risk, and quantify how much better.

One important technical point is that categorization implicitly posits an induced model based on the categorized variable X. In some cases, the induced model actually fits the data, e.g., when the response Y actually depends on X only through its categorized version, or if there are no other covariates, see the next paragraph. In other cases, and generally, the induced model does not fit the data, and we call this model *misspecified*. In particular, suppose that there are other covariates than X, say Z. Consider a binary response, Y, let $H(\cdot)$ be the logistic distribution function, and suppose that the true risk model in the continuous scale is $\mathrm{pr}(Y = 1|X, Z) = H\{m(X, Z, \boldsymbol{\beta})\}$ for some continuous function $m(\cdot)$. Then, even if there is no measurement error, if any of the covariates Z are related to Y in this continuous model, or if there is an interaction of X and Z on Y, categorizing X into J levels and plugging that into $m(X, Z, \boldsymbol{\beta})$ in place of X leads to a misspecified model as we have defined it. Measurement error in this context makes things even more difficult. When there is no measurement error, [26] gives a characterization of what is actually being estimated in misspecified models: while we do not emphasize it, our paper extends this characterization to the measurement error case. A relevant paper that first solved this particular problem is [14], which was also cited in [26].

This slightly different terminology is motivated by the following example. Suppose that Y is binary, there are no additional covariates Z, and simply define $\pi_j = \mathrm{pr}(Y = 1|X^* = j)$, where X^* is the categorized predictor. Then we can write, correctly, that $\mathrm{pr}(Y = 1|X^* = j) = H\{I(X^* = j)\theta_j\}$ by making the obvious identifications. Thus, categorization does result in an induced correctly specified logistic model, just not the one in the continuous scale. A logistic regression analysis of Y on the categories of X^* then will estimate θ_j consistently.

Our point is not to try to get epidemiologists to change their common practice. Instead, we study the effect of measurement error when a continuous predictor variable subject to measurement error is categorized. Our goal is to answer the question: with measurement error in this context, how can we (a) obtain consistent estimates of what epidemiologists would have obtained if X were actually observed; and (b) develop consistent standard errors.

We answer the question above in a general way. Section 2 gives basic technical background. Section 3 provides a general methodology for answering questions

(a) and (b) above. Section 4 presents simulation studies for linear and logistic regression that show the good behavior of our methodology, both in terms of bias and confidence interval coverage. Section 5 shows applications of our approach by using data from the Eating at America's Table Study [23]. Section 6 presents a discussion about other potential approaches to categorization and how those approaches compare to ours. Sketches of technical arguments are in the appendix.

Remark 1. As discussed above, categorization leads to a misspecified model. It is also well-known that such categorization generally leads to differential measurement error [11, 13, 3], and thus additional complications over simply fitting a measurement error model. Chapters 6.1–6.2 of [13] has a detailed discussion when the continuous variable is dichotomized, calling the result *differential by dichotomization*. We are thus assuming that the true risk model in a continuous variable X is not categorical in X. If it were, consult [13] and [3], who also discuss the issue of doing a measurement error analysis in this case, especially the difficult complex issues of computation and identifiability both theoretical and practical.

2. Data generating mechanism and basic ideas

2.1. *Illustration: A special case of linear regression*

It is instructive to consider a special case, namely linear regression. Doing so will set the stage for our general method. The response is Y, the scalar predictor subject to error is X, the observed scalar predictor is W, there are predictors Z measured without error, and we define $\widetilde{Z} = (1, Z^{\mathrm{T}})^{\mathrm{T}}$ to allow for an intercept. The regression model in the continuous predictor X is $Y = X\beta_1 + \widetilde{Z}^{\mathrm{T}}\beta_2 + \epsilon$, where ϵ is mean zero independent of (W, X, Z). There are $j = 1, ..., J$ categories $(C_1, ..., C_J)$: the number of categories J is set by the investigator, and is generally 3 (tertiles), 4 (quartiles) or 5 (quintiles), depending on the scientific field and the investigator's interests. Here $M(X, Z) = \{I(X \in C_1), ..., I(X \in C_J), Z^{\mathrm{T}}\}^{\mathrm{T}}$. If X could be observed, then we would also immediately obtain an estimate of $\boldsymbol{\beta} = (\beta_1, \beta_2^{\mathrm{T}})^{\mathrm{T}}$.

By [26], when X is observed, what epidemiologists estimate by using the categorized $M(X, Z)$ is $\boldsymbol{\Theta}$, where, based on the normal equations for the categorized predictor, $\boldsymbol{\Theta} = (\theta_1, ..., \theta_J, \boldsymbol{\Theta}_{J+1}^{\mathrm{T}})^{\mathrm{T}}$ is the solution to

$$\begin{aligned} 0 &= E[M(X, Z)\{Y - M^{\mathrm{T}}(X, Z)\boldsymbol{\Theta}\}] \\ &= E[M(X, Z)\{X\beta_1 + \widetilde{Z}^{\mathrm{T}}\beta_2 - M^{\mathrm{T}}(X, Z)\boldsymbol{\Theta}\}]. \end{aligned} \tag{1}$$

The estimate $\widehat{\boldsymbol{\Theta}}$ is the solution to $0 = n^{-1}\sum_{i=1}^{n} M(X_i, Z_i)\{Y_i - M^{\mathrm{T}}(X_i, Z_i)\boldsymbol{\Theta}\}$, and this is a consistent estimate of $\boldsymbol{\Theta}$. Comparisons between categories j and k for $j, k \leq J$, say, are $\widehat{\theta}_j - \widehat{\theta}_k$.

However, when X is not observable, estimating the solution to (1) has to be based solely on (Y, W, Z). In (1), it makes sense that if one believes the true regression model is linear in (X, Z), then, at some point, an estimate of $\boldsymbol{\beta}$ can be obtained via a measurement error analysis if there are sufficient data to do so.

Solving (1) based only on the observed W though is not so easy, and it is clear that some part of the relationship between W and X given Z is going to need to be specified, as it needs to be to do a general measurement error analysis. One way to do this is to define

$$\mathcal{G}(X, Z, \boldsymbol{\Theta}, \boldsymbol{\beta}) = M(X, Z)\{X\beta_1 + \widetilde{Z}^{\mathrm{T}}\beta_2 - M^{\mathrm{T}}(X, Z)\boldsymbol{\Theta}\}, \tag{2}$$

and then define $Q(W, Z, \boldsymbol{\Theta}, \boldsymbol{\beta}) = E\{\mathcal{G}(X, Z, \boldsymbol{\Theta}, \boldsymbol{\beta})|W, Z\}$. Since $0 = E\{Q(W, Z, \boldsymbol{\Theta}, \boldsymbol{\beta})\}$, $\boldsymbol{\Theta}$ can be estimated by solving

$$\begin{aligned} 0 = n^{-1}\textstyle\sum_{i=1}^{n}\big[&E\{M(X, Z)(X\beta_1 + \widetilde{Z}^{\mathrm{T}}\beta_2)|W_i, Z_i\} \\ &- E[\{M(X, Z)M^{\mathrm{T}}(X, Z)\}|W_i, Z_i]\boldsymbol{\Theta}\big]. \end{aligned}$$

Hence, in this simple case, for $j = 1, ..., J$ we will need to be able to calculate expectations of $XI(X \in C_j)$ given (W, Z) and the probability that $X \in C_j$ given (W, Z). As we will see, in general problems, we will need to estimate the expectations of other functions of X given (W, Z).

So, to summarize, to get a general solution, it appears that we will need to estimate (β_1, β_2) by a measurement error analysis and estimate expectations of specified functions of X given (W, Z).

Remark 2. Following on Remark 1, it is obvious that in the unlikely event that the true risk model is actually categorical in X, so that $E(Y|X, Z) = M^{\mathrm{T}}(X, Z)\boldsymbol{\beta}$, then model misspecification and differential measurement error both disappear, and one really needs just the probabilities that X is in the categories given (W, Z). As [13] and [3] discuss in detail, estimating such models is difficult because of model identifiability concerns. Often, papers dealing with this issue assume the existence of a validation data set, where X is actually observed on a subset of the data. [13] is a particularly good source for the difficulties we have mentioned and remedies using replication data. [3], page 314, who states that estimating the misclassification rates is "*most likely coming from internal validation data*" and also has a nice discussion.

2.2. Assumptions

Our work is very general, but even so, the algorithm is basically the same as in Section 2.1. Our methodology requires three basic assumptions, described below. We let X be the continuous predictor subject to measurement error, Z covariates measured exactly, W the mismeasured version of X, and Y the response.

Assumption 1. *When X is observed, the true response model in the continuous scale has parameters $\boldsymbol{\beta}$, such that there is an estimating function,* $\Phi_{\mathrm{true}}(Y, X,$

$Z, \boldsymbol{\beta})$ that identifies $\boldsymbol{\beta}$ and satisfies

$$0 = E\{\Phi_{\text{true}}(Y, X, Z, \boldsymbol{\beta})|X, Z\}. \tag{3}$$

Assumption 1 occurs in at least two circumstances.

Example 1. *(A) There are functions $m_1(X, Z, \boldsymbol{\beta})$ and $m_2(X, Z, \boldsymbol{\beta})$ such that $E(Y|X, Z) = m_1(X, Z, \boldsymbol{\beta})$ and the unbiased estimating function that would be used if X were observable is*

$$\Phi_{\text{true}}(Y, X, Z, \boldsymbol{\beta}) = m_2(X, Z, \boldsymbol{\beta})\{Y - m_1(X, Z, \boldsymbol{\beta})\}. \tag{4}$$

(B) There is a parametric model for Y given (X, Z).

Example 1(A) is very general, in that it includes traditional quasilikelihood models, nonlinear regression, generalized linear models, probit regression, etc. Crucially, it does not require a fully parametric model for the distribution of Y given (X, Z).

In our approach, as in linear regression in Section 2.1, we may need to obtain information about moments of specified functions of X given (W, Z). To do this, we will consider the setting in which there may be an external data set of N observations giving information on one set of parameters of the joint distribution, $\boldsymbol{\Lambda}_{\text{ext}}$: if there is no external study, $N = 0$ and $\boldsymbol{\Lambda}_{\text{ext}}$ does not exist. In addition, there is another set of the parameters, $\boldsymbol{\Lambda}_{\text{int}}$, that is estimated from the n observations in the internal data set.

Assumption 2. *When X is not observed, either (a) the distribution of X given (W, Z) is known up to parameters $\boldsymbol{\Lambda}_{\text{ext}}$ and $\boldsymbol{\Lambda}_{\text{int}}$ as described above, or (b) there is a function, $\mathcal{G}(X, Z, \boldsymbol{\Theta}, \boldsymbol{\beta})$ defined at (11) below, whose conditional expectation given (W, Z) depends on parameters $\boldsymbol{\Lambda}_{\text{ext}}$ and $\boldsymbol{\Lambda}_{\text{int}}$ and can be estimated. The parameter $\boldsymbol{\Lambda}_{\text{ext}}$ cannot be estimated by internal data, while the parameter $\boldsymbol{\Lambda}_{\text{int}}$ can be estimated by internal data. For both, there are unbiased estimating functions $V_{\text{ext},m}(\boldsymbol{\Lambda}_{\text{ext}})$ for the external data and $V_{\text{int},i}(\boldsymbol{\Lambda}_{\text{int}}, \boldsymbol{\Lambda}_{\text{ext}})$ for the internal data such that $E\{V_{\text{ext},m}(\boldsymbol{\Lambda}_{\text{ext}})\} = 0$ and $E\{V_{\text{int},i}(\boldsymbol{\Lambda}_{\text{int}}, \boldsymbol{\Lambda}_{\text{ext}})\} = 0$.*

For linear regression, $\mathcal{G}(X, Z, \boldsymbol{\Theta}, \boldsymbol{\beta})$ is given in (2).

If there are external data and $N > 0$, we estimate $\boldsymbol{\Lambda}_{\text{ext}}$ by solving the estimating equation

$$0 = N^{-1} \sum_{m=1}^{N} V_{\text{ext},m}(\boldsymbol{\Lambda}_{\text{ext}}). \tag{5}$$

In the internal data set, we estimate $\boldsymbol{\Lambda}_{\text{int}}$ by solving an estimating equation

$$0 = n^{-1} \textstyle\sum_{i=1}^{n} V_{\text{int},i}(\boldsymbol{\Lambda}_{\text{int}}, \widehat{\boldsymbol{\Lambda}}_{\text{ext}}). \tag{6}$$

There is also a very subtle issue that needs to be made explicit.

Assumption 3. *If external data are necessary for model identification, the parameter $\boldsymbol{\Lambda}_{\text{ext}}$ is transportable in the sense that this parameter is the same in the external and internal data sets.*

The issue of when parameters are transportable from an external data set to the internal data set is discussed in Chapter 2.2.4–2.2.5 of [4]. As they state, it is much better if there are sufficient internal data that external data need not be used, but this is not always the case.

2.3. General observations when X is observed

As argued in Section 1, the goal is to fit a model when X is categorized into J levels $(C_1, ..., C_J)$, and so we defined the dummy variables and Z together as $M(X, Z) = \{I(X \in C_1), ..., I(X \in C_J), Z^{\mathrm{T}}\}^{\mathrm{T}}$: our formulation allows more complex forms, including interactions. Suppose there are $i = 1, ..., n$ subjects in the primary/main/internal study, and suppose further that we observe (Y_i, X_i, Z_i). If X is observed, the analysis done on these categories will be based on replacing (X, Z) in (3)–(4) by $M(X, Z)$, and to make clear the categorization, we define a parameter $\mathbf{\Theta}$, set $\Phi_{\mathrm{cat}}\{Y_i, M(X_i, Z_i), \mathbf{\Theta}\} = \Phi_{\mathrm{true}}\{Y_i, M(X_i, Z_i), \mathbf{\Theta}\}$, and obtain $\widehat{\mathbf{\Theta}}$ by solving

$$0 = n^{-1}\textstyle\sum_{i=1}^{n}\Phi_{\mathrm{cat}}\{Y_i, M(X_i, Z_i), \mathbf{\Theta}\}. \tag{7}$$

More complex forms of (7) are easily accommodated.

Unlike in Assumption 1 and (3)–(4), except in the rare case that the categorized model is actually true, $0 \neq E[\Phi_{\mathrm{cat}}\{Y, M(X, Z), \mathbf{\Theta}\}|X, Z]$, a *conditional* expectation. This is a key part of the work in [26].

Despite the fact that the categorized model does not fit the data conditional on (X, Z), by standard estimating equation theory [26], the estimate formed by solving (7) has a limit as $n \to \infty$, $\mathbf{\Theta}$, which is the solution to

$$0 = E[\Phi_{\mathrm{cat}}\{Y, M(X, Z), \mathbf{\Theta}\}]. \tag{8}$$

It is important to observe that (8) is an *unconditional* expectation, not a conditional one.

If, instead of observing X, we observe its mismeasured version W, and if we replace X by W, we will of course generally inconsistently estimate both $\boldsymbol{\beta}$ and $\mathbf{\Theta}$.

2.4. Estimating the true parameter $\boldsymbol{\beta}$

In our approach, as in Section 2.1 for linear regression, we must estimate $\boldsymbol{\beta}$ in (3). There is of course a large literature on how to do this [13, 4, 3, 27]. Borrowing on that literature, from Assumptions 1–2, for an estimating function $\Phi(Y, W, Z, \boldsymbol{\beta}, \mathbf{\Lambda}_{\mathrm{int}}, \mathbf{\Lambda}_{\mathrm{ext}})$, the estimate, $\widehat{\boldsymbol{\beta}}$, is the solution to

$$0 \quad = \quad n^{-1}\textstyle\sum_{i=1}^{n}\Phi(Y_i, W_i, Z_i, \boldsymbol{\beta}, \widehat{\mathbf{\Lambda}}_{\mathrm{int}}, \widehat{\mathbf{\Lambda}}_{\mathrm{ext}}), \tag{9}$$

where $(\widehat{\mathbf{\Lambda}}_{\mathrm{int}}, \widehat{\mathbf{\Lambda}}_{\mathrm{ext}})$ are obtained from equations (5) and (6), respectively. Of course, the details and the form of $\Phi(\cdot)$ differ from case-to-case.

3. Methodology and asymptotic theory

3.1. Methodology: General case

The methodology is simple to explain at the general level. The target $\mathbf{\Theta}$ is defined as the solution to (8). However, we can rewrite (8) as

$$0 = E\left(E[\Phi_{\rm cat}\{Y, M(X,Z), \mathbf{\Theta}\}|W,Z]\right). \tag{10}$$

Since the distribution of Y given (X,Z) depends on $\boldsymbol{\beta}$, for notational completeness we define

$$\begin{aligned}\mathcal{G}(X,Z,\mathbf{\Theta},\boldsymbol{\beta}) &= E\left[\Phi_{\rm cat}\{Y, M(X,Z), \mathbf{\Theta}\}|X,Z\right] \qquad (11)\\ &= E\left[\Phi_{\rm cat}\{Y, M(X,Z), \mathbf{\Theta}\}|X,Z,\boldsymbol{\beta}\right];\\ Q(W,Z,\mathbf{\Theta},\boldsymbol{\beta},\mathbf{\Lambda}_{\rm int},\mathbf{\Lambda}_{\rm ext}) &= E\{\mathcal{G}(X,Z,\mathbf{\Theta},\boldsymbol{\beta})|W,Z\}. \qquad (12)\end{aligned}$$

Making the usual nondifferential measurement error assumption, i.e., that Y and W are independent given (X,Z),

$$0 = E\left\{Q(W,Z,\mathbf{\Theta},\boldsymbol{\beta},\mathbf{\Lambda}_{\rm int},\mathbf{\Lambda}_{\rm ext})\right\}. \tag{13}$$

Critically, (12) depends only on the observed covariates. Thus, if we have consistent estimates $(\widehat{\boldsymbol{\beta}},\widehat{\mathbf{\Lambda}}_{\rm int},\widehat{\mathbf{\Lambda}}_{\rm ext})$ of $(\boldsymbol{\beta},\mathbf{\Lambda}_{\rm int},\mathbf{\Lambda}_{\rm ext})$, then a consistent estimate, $\widehat{\mathbf{\Theta}}$, of $\mathbf{\Theta}$ solves

$$0 = n^{-1}\textstyle\sum_{i=1}^{n} Q(Z_i, W_i, \mathbf{\Theta}, \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{\Lambda}}_{\rm int}, \widehat{\mathbf{\Lambda}}_{\rm ext}). \tag{14}$$

In some cases, we do not have external data. Thus, we do not have $V_{\rm ext}$ and $\mathbf{\Lambda}_{\rm ext}$, and $V_{\rm int}$ and $\mathbf{\Theta}$ only depend on $\mathbf{\Lambda}_{\rm int}$.

Remark 3. The key question is how to compute $\mathcal{G}(X,Z,\mathbf{\Theta},\boldsymbol{\beta})$ in (11). In the fully general case (3), we require a parametric model for the distribution of Y given (X,Z), as in Example 1(B). However, in standard regression models of the form in (4) in Example 1(A), great simplification occurs, because in that case,

$$\Phi_{\rm cat}\{Y, M(X,Z), \mathbf{\Theta}\} = m_2\{M(X,Z),\mathbf{\Theta}\}\left[Y - m_1\{M(X,Z),\mathbf{\Theta}\}\right],$$

and thus

$$\mathcal{G}(X,Z,\mathbf{\Theta},\boldsymbol{\beta}) = m_2\{(X,Z),\mathbf{\Theta}\}\left[m_1(X,Z,\boldsymbol{\beta}) - m_1\{M(X,Z),\mathbf{\Theta}\}\right].$$

C.3 gives detailed formulae for linear and logistic regression.

Remark 4. Our method is closely related to the *expectation-correction method* of [27], Chapter 2.5.2, and less closely to the general corrected score methods first introduced by [17]. [27] has an excellent and comprehensive discussion of the correction methods in the literature. We do not have a score function per se, but we have a function, $\Phi_{\rm cat}\{Y, M(X,Z), \mathbf{\Theta}\}$, with the property that

$E[\Phi_{\rm cat}\{Y, M(X,Z), \boldsymbol{\Theta}\}] = 0$: importantly, it is not true that the conditional expectation $E[\Phi_{\rm cat}\{Y, M(X,Z), \boldsymbol{\Theta}\}|X, Z] \equiv 0$. Instead of our (11)–(12), the expectation-correction method uses as its estimating equation $E[\Phi_{\rm cat}\{Y, M(X, Z), \boldsymbol{\Theta}\}|Y, W, Z] = Q^{*}(Y, W, Z, \boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{\Lambda}_{\rm int}, \boldsymbol{\Lambda}_{\rm ext})$. The obvious distinction is that our function $Q(\cdot)$ does not involve Y explicitly, while the expectation-correction function $Q^{*}(\cdot)$ does involve Y. We used $Q(\cdot)$ and (11) because our assumptions allow $\mathcal{G}(\cdot)$ to be calculated explicitly, especially in Example 1(A), so that implementation is somewhat easier. In addition, in Example 1(A), there does not need to be a full likelihood, as would be required in the expectation-correction method, so there are actual differences in the methods.

3.2. Asymptotic Theory

Asymptotic theory for the parameter estimates is easily derived. Let $\boldsymbol{\Omega} = (\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{\Lambda}_{\rm int}, \boldsymbol{\Lambda}_{\rm ext})$ and let the true values of the parameters be denoted by $\boldsymbol{\Omega}$.

It is neater notation in this section to let $i = 1, ..., n$ denote the internal data, and $i = n + 1, ..., n + N$ denote the external data. For $i > n$, define $\Psi_i(\boldsymbol{\Omega}) = \{0, 0, 0, V^{\rm T}_{{\rm ext},i}(\boldsymbol{\Lambda}_{\rm ext})\}^{\rm T}$, while for $i \leq n$ define

$$\Psi_i(\boldsymbol{\Omega}) = \{Q^{\rm T}(W_i, Z_i, \boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{\Lambda}_{\rm int}, \boldsymbol{\Lambda}_{\rm ext}), \Phi^{\rm T}(Y_i, W_i, Z_i, \boldsymbol{\beta}, \boldsymbol{\Lambda}_{\rm int}, \boldsymbol{\Lambda}_{\rm ext}),\\ V^{\rm T}_{{\rm int},i}(\boldsymbol{\Lambda}_{\rm int}, \boldsymbol{\Lambda}_{\rm ext}), 0\}^{\rm T}.$$

If there are external data, the estimate $\widehat{\boldsymbol{\Omega}}$ solves $0 = \sum_{i=1}^{n+N} \Psi_i(\widehat{\boldsymbol{\Omega}})$. If there are no external data, then $N = 0$, $\boldsymbol{\Omega} = (\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{\Lambda}_{\rm int})$ and the zero element and $\boldsymbol{\Lambda}_{\rm ext}$ in the definition of $\Psi_i(\boldsymbol{\Omega})$ are removed.

By standard estimating equation results, we have the following results, which are shown in Appendices A.1 and A.2.

Lemma 1. If there are external data, i.e., $N > 0$, make Assumptions 1–3. Suppose that $N \to \infty$ and $n \to \infty$ such that $n/(n + N) \to b_{\rm lim}$, where $0 < b_{\rm lim} < 1$. Then

$$(n + N)^{1/2}(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}) \to {\rm Normal}\{0, A^{-1}B(A^{-1})^{\rm T}\},$$

where $A = b_{\rm lim}E\{\partial\Psi_1(\boldsymbol{\Omega})/\partial\boldsymbol{\Omega}^{\rm T}\} + (1 - b_{\rm lim})E\{\partial\Psi_{n+N}(\boldsymbol{\Omega})/\partial\boldsymbol{\Omega}^{\rm T}\}$ and $B = b_{\rm lim}{\rm cov}\{\Psi_1(\boldsymbol{\Omega})\} + (1 - b_{\rm lim}){\rm cov}\{\Psi_{n+N}(\boldsymbol{\Omega})\}$. In the definitions of A and B, the expectation and covariance matrix for $\Psi_1(\boldsymbol{\Omega})$ are computed in the internal data, while the expectation and covariance matrix for $\Psi_{N+n}(\boldsymbol{\Omega})$ are computed in the external data. Let $\widehat{C}_{\rm ext}$ be the sample covariance matrix of $\Psi_i(\widehat{\boldsymbol{\Omega}})$ for $i = n + 1, ..., n + N$ and let $\widehat{C}_{\rm int}$ be the sample covariance matrix of $\Psi_i(\widehat{\boldsymbol{\Omega}})$ for $i = 1, ..., n$. Consistent estimates of A and B are easily seen to be $\widehat{A} = (n + N)^{-1}\sum_{i=1}^{N+n}\partial\Psi_i(\widehat{\boldsymbol{\Omega}})/\partial\boldsymbol{\Omega}^{\rm T}$ and $\widehat{B} = \{n/(n + N)\}\widehat{C}_{\rm int} + \{N/(n + N)\}\widehat{C}_{\rm ext}$.

Lemma 2. If there are no external data, i.e., $N = 0$, make Assumptions 1–2. As $n \to \infty$,

$$n^{1/2}(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}) \to {\rm Normal}\{0, A^{-1}B(A^{-1})^{\rm T}\},$$

where $A = E\{\partial\Psi_1(\mathbf{\Omega})/\partial\mathbf{\Omega}^{\mathrm{T}}\}$ and $B = \mathrm{cov}\{\Psi_1(\mathbf{\Omega})\}$. In the definitions of A and B, the expectation and covariance matrix for $\Psi_1(\mathbf{\Omega})$ are computed in the internal data. Let $\widehat{C}_{\mathrm{int}}$ be the sample covariance matrix of $\Psi_i(\widehat{\mathbf{\Omega}})$ for $i = 1, ..., n$. Consistent estimates of A and B are easily seen to be $\widehat{A} = n^{-1}\sum_{i=1}^{n}\partial\Psi_i(\widehat{\mathbf{\Omega}})/\partial\mathbf{\Omega}^{\mathrm{T}}$ and $\widehat{B} = \widehat{C}_{\mathrm{int}}$.

Remark 5. While the calculations used in Lemmas 1–2 are standard, as a referee has pointed out, we are making the following kinds of assumptions to carry them through: weaker conditions can be constructed. All these conditions hold in our examples of linear and logistic regression with additive measurement error. There is a parameter which we have called in this subsection $\mathbf{\Omega} = (\mathbf{\Theta}, \boldsymbol{\beta}, \mathbf{\Lambda}_{\mathrm{int}}, \mathbf{\Lambda}_{\mathrm{ext}})$. For $i = 1, ..., n + N$, we have defined estimating functions $\Psi_i(\mathbf{\Omega})$, which we have defined in such a way that $E\{\Psi_i(\mathbf{\Omega})\} = 0$ for $i = 1, ..., n + N$: the expectations are unconditional, although in implementing the estimators we have exploited our Assumptions 1–3 to simplify the numerical calculations. Having done all this, we are now in the realm of estimating equation theory. Sufficient but not necessary conditions for our asymptotic theory to hold are the following.

- The parameter space is compact. This is not necessary but it is convenient for proving consistency.
- There is a unique $\mathbf{\Omega}$ in the parameter space such that $E\{\Psi_i(\mathbf{\Omega})\} = 0$ for all $i = 1, ..., n + N$.
- The estimating equations $\Psi_i(\mathbf{\Omega})$ are 3-times continuously and boundedly differentiable in the parameter space.
- The estimating equation $0 = \sum_{i=1}^{n+N} \Psi_i(\mathbf{\Omega})$ has a unique solution.
- The matrix $E\{\partial\Psi_i(\mathbf{\Omega})/\partial\mathbf{\Omega}^{\mathrm{T}}\}$ is of full rank within a neighborhood of the true parameter value.
- For sufficiently large (n, N), within a neighborhood of the true parameter value, $(n + N)^{-1}\sum_{i=1}^{n+N} \partial\Psi_i(\mathbf{\Omega})/\partial\mathbf{\Omega}^{\mathrm{T}}$ is of full rank with eigenvalues bounded away from 0 and $\pm\infty$.

Remark 6. The major new item here in verifying the assumptions mentioned in Remark 5 are the differentiability assumptions having to do with $Q(W, Z, \mathbf{\Theta}, \boldsymbol{\beta}, \mathbf{\Lambda}_{\mathrm{int}}, \mathbf{\Lambda}_{\mathrm{ext}})$ in (12). Let the conditional density/mass function of Y given (X, Z) be $f_{Y|X,Z}(\cdot, \boldsymbol{\beta}, \mathbf{\Lambda}_{\mathrm{int}}, \mathbf{\Lambda}_{\mathrm{ext}})$ and the conditional density/mass function of X given (W, Z) be $f_{X|W,Z}(\cdot, \mathbf{\Lambda}_{\mathrm{int}}, \mathbf{\Lambda}_{\mathrm{ext}})$. Let $d\nu(y)$ and $d\nu(x)$ be integrals/counts as the case requires. Then (12) can be written out as

$$\begin{aligned} &Q(W, Z, \mathbf{\Theta}, \boldsymbol{\beta}, \mathbf{\Lambda}_{\mathrm{int}}, \mathbf{\Lambda}_{\mathrm{ext}}) \\ &= \int \left\{ \int \Phi_{\mathrm{cat}}\{y, M(x, Z), \mathbf{\Theta}\} f_{Y|X,Z}(y \mid x, Z, \boldsymbol{\beta}, \mathbf{\Lambda}_{\mathrm{int}}, \mathbf{\Lambda}_{\mathrm{ext}}) d\nu(y) \right\} \\ &\qquad\qquad \times f_{X|W,Z}(x \mid W, Z, \mathbf{\Lambda}_{\mathrm{int}}, \mathbf{\Lambda}_{\mathrm{ext}}) d\nu(x). \end{aligned}$$

Then the non-standard differentiability assumptions in Remark 5 are really about the differentiability assumptions of $\Phi_{\mathrm{cat}}\{y, M(x, Z), \mathbf{\Theta}\}$, $f_{Y|X,Z}(\cdot, \boldsymbol{\beta}, \mathbf{\Lambda}_{\mathrm{int}}, \mathbf{\Lambda}_{\mathrm{ext}})$ and $f_{X|W,Z}(\cdot, \mathbf{\Lambda}_{\mathrm{int}}, \mathbf{\Lambda}_{\mathrm{ext}})$ with respect to the parameters.

4. Simulations: Logistic and linear regression

4.1. Logistic regression

4.1.1. Scenarios

For simplicity, we do our simulations in the case that there is no Z. For logistic regression, we assume that the true model is

$$\mathrm{pr}(Y=1|X) \quad = \quad H(\beta_0 + X\beta_1) = H\{(1,X)\boldsymbol{\beta}\}, \tag{15}$$

where $H(\cdot)$ is the logistic distribution function. Then we generate data as

$$W = X + U; \qquad X = \mathrm{Normal}(\mu_x, \sigma_x^2); \qquad U = \mathrm{Normal}(0, \sigma_u^2), \tag{16}$$

where X and U are independent. We set $\beta_0 = -0.42$ and set $\beta_1 = \log(1.5)$ in Table 1. We set $(\mu_x = 0, \sigma_x^2 = 1, \sigma_u^2 = 1)$, so that the measurement error variance is the same as the variance of X, and the classical attenuation coefficient is $\lambda = \sigma_x^2/(\sigma_x^2 + \sigma_u^2) = 0.50$. Solving (8) numerically, we find that $\boldsymbol{\Theta} = (-0.98, -0.64, -0.42, -0.21, 0.14)^{\mathrm{T}}$. In both cases, the main study sample size is $n = 500$.

We used the quintiles of the distribution of X to define the categories. This is because, as stated in the introduction, we have our goal is to obtain consistent estimates of what epidemiologists would have obtained if X were actually observed, in this case, the quintiles of X.

We did simulations in two cases:

1. External-Internal Data: The internal data has no replicates and the external data set has size $N = 300$ and $K = 2$ replicates for each observation. The nuisance parameters are $\boldsymbol{\Lambda}_{\mathrm{ext}} = \sigma_u^2$ and $\boldsymbol{\Lambda}_{\mathrm{int}} = (\mu_x, \sigma_x^2)$. We estimated σ_u^2 from the external data with replicates, and estimated μ_x, σ_x^2 using the internal data without any replicates. Standard errors were computed as in Lemma 1.
2. Internal Data Only: The internal data has $R = 2$ replicates and there are no external data $(K = 0)$. The nuisance parameters $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}_{\mathrm{int}} = (\mu_x, \sigma_x^2, \sigma_u^2)$. We estimated $(\mu_x, \sigma_x^2, \sigma_u^2)$ from the internal data with replicates. Standard errors were computed as in Lemma 2.

C.3 provides details of implementation.

4.1.2. Results

The results given below are similar, and indeed even more impressive, when the main study sample size n increases to $n = 1{,}000$, $2{,}000$ and $3{,}000$, and thus these are not displayed here. The results are also similar when β_1 is either smaller or larger. The same qualitative results are also found for $\boldsymbol{\Theta} = (\theta_1, ..., \theta_5)^{\mathrm{T}}$ individually (results not shown).

We fit the new approach and compare it with the naive method for the both cases described above. Our main interest is to estimate the log relative risk $\theta_5 - \theta_1$, which compares the effect of the category 5 with the effect of the category 1. In the two simulations, we computed (a) the log relative risk pretending that X is observed; (b) our method; and (c) the naive method that ignores measurement error. In the scenario of internal data with $R = 2$, the predictor used was the sample mean of the replicates.

Based on 1000 simulated data sets, in Table 1, we report the empirical average mean bias, asymptotic standard error, standard deviation, root mean squared error, and coverage rate of the nominal 95% confidence interval across the simulations.

From Table 1, we observe the following.

- The estimator using true X and our method both have little bias and provide near-nominal coverage.
- The naive estimator that ignores the measurement error is badly biased and attenuated towards zero. Consequently the coverage probabilities are near-zero and the root mean squared errors are quite inflated.
- With no internal replicates, i.e., $R = 1$, the root mean squared error of our method is naturally higher than if X had been observed, but not quite as high as would be expected in a continuous analysis. Indeed, in a continuous analysis with attenuation $\lambda = 0.50$, as in our simulation, one would expect a doubling of root mean squared error.

4.2. Linear regression

4.2.1. Scenarios

In this section, we do simulations based on simple linear regression with no Z, including homoscedastic and heteroscedastic cases.

We assume that the true model is

$$Y \quad = \quad \beta_0 + X\beta_1 + \epsilon = (1, X)\boldsymbol{\beta} + \epsilon, \tag{17}$$

Similarly, we generate data as

$$W = X + U; \qquad X = \text{Normal}(\mu_x, \sigma_x^2); \qquad U = \text{Normal}(0, \sigma_u^2).$$

We set $\beta_0 = 0$ and set $\beta_1 = 0.75$ and studied two cases: (a) homoscedastic with $\epsilon \sim N(0, 1)$; and (b) heteroscedastic with $\epsilon \sim N(0, 0.2 + 0.5x^2)$. The classical attenuation coefficient and sample size are the same as in Section 4.1. Solving (8) numerically, we find that $\boldsymbol{\Theta} = (-1.04, -0.40, 0.00, 0.40, 1.05)^{\mathrm{T}}$. C.2 provides implementation details.

4.2.2. Results

Similarly as before, our main interest is to estimate $\theta_5 - \theta_1$, which compares the effect of the category 5 with the effect of the category 1. In the two simulations, we computed $\theta_5 - \theta_1$ (a) pretending that X is observed; (b) our methods; and (c) the naive method that ignores measurement error. For the naive method, in internal data with $R = 2$, the predictor used is the sample mean of the replicates.

Based on 1000 simulated data sets, in Table 2, we report the empirical average mean bias, asymptotic standard error, standard deviation, root mean squared error, and coverage rate of the nominal 95% confidence intervals across the simulations.

From Table 2, we see that similar conclusions can be drawn as in Section 4.1. However, an interesting thing is in the heteroscedastic case, when noise ϵ has its variance related to X. Assuming that X is observed, the coverage rate of nominal 95% confidence intervals is low, because the heteroscedasticity is ignored. Using our method, we can get close to nominal coverage without knowing any information about the noise ϵ. Thus, this example shows that our method is very general as we stated in Example 1(A).

5. Empirical example

5.1. Data description

We illustrate our methods using data from the Eating at America's Table (EATS) Study [23], in which 964 participants completed multiple 24-hour recalls of diet. We consider the variable Fat Density, which is the percentage of calories coming from Fat. The response Y is either (i) the indicator of obesity, which means that a subject's body mass index (BMI, weight in kilograms divided by the square of height in meters) is 30 or greater. or (ii) the actual body mass index. We assume that W, is unbiased for usual intake X, and that $W = X + U$. It is reasonable in these data to take (a) X to be normally distributed, (b) that U is normally distributed; and (c) that X and U are independent, as we now describe. We used the methods described in [9] and Chapter 1.7 of [4], which also give the rationale for these methods. Specifically, for (a), as they suggest a qq-plot of the individual means for Fat Density looked acceptably normal, with skewness and kurtosis = -0.06 and 3.02, respectively, see the top panel of Figure 1. For (b), as they suggest, we took differences of the first and second Fat Density measurements, which had skewness (theoretically = 0) and kurtosis = -0.14 and 3.40, respectively: the somewhat higher kurtosis here is seen to be minor on the qq-plot, see the middle panel of Figure 1. Finally, for (c), they suggest analyzing the correlation between the individual-level mean and standard deviation = 0.06, and there was no obvious strong pattern when we plotted the data the latter against the former, see the bottom panel of Figure 1.

For numerical stability, our analysis in the continuous scale is uses centered and standardized W using $(15W - 5)/\sqrt{0.5}$. To illustrate an example of an internal and an external study, we randomly selected $N = 200$ subjects as the external study to have the first two 24-hour recalls, while using the remaining data as the main internal study. As in the simulation, we either set the number of recalls $R = 1,\ K = 2$, meaning the external study data were used to estimate the measurement error variance, for $R = 2,\ K = 0$, in which case the external data were not used.

5.2. *Results*

5.2.1. Logistic regression

As described in Section 4.1, we assume the true model defined by (15)–(16), and the respective two cases. In this application we again estimate the log relative risk $\theta_5 - \theta_1$. We fit both our new approach and the naive model that ignores measurement error when external data is and is not used.

In Table 3, we observe that when using the external data and only 1 observation in the internal data the estimate of the log relative risk $\theta_5 - \theta_1$ from our approach is 108% greater than the naive estimate, while when using internal data with two replicates our estimate of our approach is 32% greater than the naive estimate. This makes sense because the second case uses the mean of two replicates, hence has smaller measurement error variance, and thus the naive estimate will be closer to our method.

In both cases, the asymptotic standard error from our new method is greater than the naive method, which led to wider confidence intervals. This makes sense, because with a scalar covariate measured with error, correcting for measurement error bias usually increases estimated standard errors, while of course reducing bias.

5.2.2. Linear regression

Next we consider the linear model with body mass index as the response. All assumptions for W, X and U are the same as in Section 5.1. Moreover, we maintain the standardization and sampling scheme in Section 5.1: the results are presented in Table 4.

From Table 4, we observe similar conclusions as in logistic regression case. One point of particular interest is that in both scenarios (external-internal or internal data only), our estimator converges theoretically to the same value, and this is seen in the results. The naive method that ignores measurement error estimates different parameters because the measurement error variance is twice as large in the external-internal case as it is in the internal-only case.

6. Other approaches and the assumptions

6.1. Other approaches

We emphasize once more that it is common practice in epidemiology to categorize a continuous predictor, and we have given numerous citations of this practice. Generally, this practice results in a misspecified model.

Our goal is to correct the analysis so as to reproduce, asymptotically, the estimators that would have been obtained if there were no measurement error.

The problem has not been considered previously in the context that a continuous predictor has been categorized. Such categorization generally leads to differential measurement error [11, 13, 3], and thus additional complications over simply fitting a measurement error model.

While our paper is the first to consider the issue of how to correct an analysis to account for a continuous predictor that is categorized, there are of course other possible approaches, but none of them really avoids the basic issues we have discussed of what is needed to obtain consistent estimators with asymptotically correct inference in the case of measurement error.

- For example, one could assume that the true risk model is based upon the categorized truth, even if this is implausible in most contexts. One could further assume that the misclassification is nondifferential, which is incorrect if the true risk model is in the continuous scale [11, 13, 3]. There is a small literature on this problem. [13], especially Chapter 6.1, has remarks on the bias induced when a binary predictor is misclassified. [3], Chapter 6.7.7 and Chapter 6.14, has a detailed discussion of the issue, and provides a number of references to the problem. Both [13] and [3] show that a measurement error correction will require a distribution for the categorical X given (W, Z), sometimes called the reclassification rate, and both indicate that there are substantive issues, including identifiability, involved with estimating these models. For replication studies wherein W is measured repeatedly on a subset of the data, there is some evidence that 3 replicates will result in identifiability. However, both books emphasize the use of internal validation substudies, wherein one actually observes X in a substudy.

  If X_{cat} is the categorized truth, then one might attempt an analysis based on assuming a joint distribution of (Y, W, X_{cat}) given Z, but as in any measurement error model [4], the joint distribution requires (a) a distribution for Y given (X_{cat}, W, Z), and (b) the distribution of (W, X_{cat}) given Z. However, (a) actually depends on W, and thus that the modeling presents additional complications. In addition, (b) is no easier than ours, can be implausible and does not make fewer assumptions than we have done.
- Simulation-extrapolation, or SIMEX, [6, 22, 4] is a well-known approach to the creation of *approximately*, but not fully, consistent estimators for additive measurement error models of the form $W = X + Z^{\mathrm{T}}\alpha + U$, where U is independent of Z and can be homoscedastic or heteroscedastic but has replicates [8], and is generally taken to be normally distributed. This literature attempts to dispense with distributional assumptions for X for the continuous case, but is at best approximately correct. The fact that a categorized risk model is implausible, leading to differential measurement error, may also cause complications, but the use of SIMEX in this context is a worthwhile topic for further study. We also mention the MC-SIMEX procedure [16], which is appropriate for misclassified data where the misclassification probabilities can be estimated.

- It is also possible to change the paradigm entirely and avoid categorization, and all the issues related to categorization, by instead using Bsplines. Indeed, part of the reason sometimes given for categorizing a continuous predictor and not modeling a response linearly in the continuous X is that it could lead to unduly extreme comparisons for risk between the lowest and the highest values of X. The general thought is that this can be overcome by replacing the linear X by a Bspline in X. There are papers involving Bsplines and measurement error [2, 12, 18], and it appears that regression calibration can possible be used by calibrating each spline basis function. After the fitting, one could compare the Bspline fits at the 10^{th}, 30^{th}, 50^{th}, 70^{th} and 90^{th} percentiles of X to form versions of the tables found in epidemiology papers, but the interpretations are not fully comparable.

We showed how to solve this problem and given asymptotically consistent estimators with asymptotically correct standard errors. Assumption 2 is reasonable in other contexts than ours, for example, that X has a mixture-of-normals distribution and U is normally distributed [7].

6.2. Assumptions in the simulations and example

Readers of an initial version of this paper have noted that our simulations and data example use the assumption that the distribution of X given (W, Z) is normally distributed, but misinterpreted this fact into concluding that the approach is only applicable in that case. For the data example in Section 5, we justified the assumptions using known methods for model checking of measurement error models. Assumption 2 is widely used and reasonable in many other contexts than ours numerical work, for example, that X has a mixture-of-normals distribution and U is normally distributed [7]. Modeling via mixture distributions is a reasonable way to extend what we have done in the classical error case. See also [21] for the homoscedastic and heteroscedastic cases when the variance function and the distributions of X and U are modeled as mixture distributions.

Many papers in the literature also rely on the existence of validation data, where X is actually observed in a subset of the main data set. In that case, Assumption 2 is easily checked by model fitting and validation on the observed validation data subset.

6.3. Categorization

In Section 2.1, we stated that the number J of categories was set by the investigators, Usually, $J = 3$, 4 or 5, as seen by the examples cited in the introduction. In addition, setting the category limits is also an art, and may be based on (a) limits in the literature; (b) limits based on the error-prone instrument, such as the quintiles of a food frequency questionnaire or 24-hour recall; and (c) limits based on a measurement error analysis. Since our goal is to construct the analysis that would have been done if X could be observed, we use the latter.

Appendix A: Sketch of technical arguments

A.1. Argument for Lemma 1

We consider the case that there are external data used to estimate $\boldsymbol{\Lambda}_{\text{ext}}$ and that there are parameters $\boldsymbol{\Lambda}_{\text{int}}$. As in Section 3.2, the data for $i = 1, ..., n$ are for the internal data, while, for $i = n+1, ..., n+N$, are for the external data if such external data exist and are used. The functions $\Psi_i(\boldsymbol{\Omega})$ are also defined in Section 3.2. By Taylor series,

$$\begin{aligned}
&(n+N)^{-1/2}\textstyle\sum_{i=1}^{n+N}\left\{\Psi_i(\widehat{\boldsymbol{\Omega}}) - \Psi_i(\boldsymbol{\Omega})\right\} \\
&\quad = (n+N)^{-1/2}\textstyle\sum_{i=1}^{n+N}\left\{\dfrac{\partial\Psi_i(\boldsymbol{\Omega})}{\partial\boldsymbol{\Omega}}(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}) + o_p(\|\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\|)\right\} \\
&\quad = \left\{(n+N)^{-1}\textstyle\sum_{i=1}^{n+N}\dfrac{\partial\Psi_i(\boldsymbol{\Omega})}{\partial\boldsymbol{\Omega}}\right\}(n+N)^{1/2}(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}) + o_p(1).
\end{aligned}$$

For logistic regression and linear regression, the forms of $\Psi_i(\boldsymbol{\Omega})$ can be found in Appendix C. Thus,

$$\begin{aligned}
&(n+N)^{1/2}(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}) \\
= \;& -\left\{(n+N)^{-1}\sum_{i=1}^{N+n}\partial\Psi_i(\boldsymbol{\Omega})/\partial\boldsymbol{\Omega}\right\}^{-1}(n+N)^{-1/2}\sum_{i=1}^{N+n}\Psi_i(\boldsymbol{\Omega}) + o_p(1).
\end{aligned}$$

It is obvious that $(n+N)^{-1}\sum_{i=1}^{N+n}\partial\Psi_i(\boldsymbol{\Omega})/\partial\boldsymbol{\Omega} = A + o_p(1)$, and immediate that $(n+N)^{-1/2}\sum_{i=1}^{N+n}\Psi_i(\boldsymbol{\Omega}) \to \text{Normal}(0, B)$, where A and B are defined in Lemma 1.

A.2. Argument for Lemma 2

We consider the case that there are only parameters $\boldsymbol{\Lambda}_{\text{int}}$. As in Section 3.2, the data for $i = 1, ..., n$ are for the internal data. The functions $\Psi_i(\boldsymbol{\Omega})$ are also defined in Section 3.2. Then

$$\begin{aligned}
0 \;&=\; n^{-1/2}\sum_{i=1}^{n}\Psi_i(\widehat{\boldsymbol{\Omega}}) \\
&=\; n^{-1/2}\sum_{i=1}^{n}\Psi_i(\boldsymbol{\Omega}) + \left\{n^{-1}\sum_{i=1}^{n}\partial\Psi_i(\boldsymbol{\Omega})/\partial\boldsymbol{\Omega}\right\}n^{1/2}(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}) + o_p(1),
\end{aligned}$$

so that

$$n^{1/2}(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}) \;=\; -\left\{n^{-1}\sum_{i=1}^{n}\partial\Psi_i(\boldsymbol{\Omega})/\partial\boldsymbol{\Omega}\right\}^{-1} n^{-1/2}\sum_{i=1}^{n}\Psi_i(\boldsymbol{\Omega}) + o_p(1).$$

As in A.1, $n^{-1}\sum_{i=1}^{n}\partial\Psi_i(\boldsymbol{\Omega})/\partial\boldsymbol{\Omega} = A + o_p(1)$, and $n^{-1/2}\sum_{i=1}^{n}\Psi_i(\boldsymbol{\Omega}) \to \text{Normal}(0, B)$, where A and B are defined in Lemma 2.

Appendix B: Tables for simulations and EATS data analysis

Table 1
Simulation study for logistic regression in Section 4.1 with sample size $n = 500$ and, where applicable, the external study has sample size $N = 300$ and 2 replicates, while $\beta_0 = -0.42$, $\beta_1 = log(1.5)$. The target parameter, $\boldsymbol{\Theta} = (\theta_1, ..., \theta_5)^{\mathrm{T}}$, where θ_j is the parameter for the j^{th} category. Displayed are results for the estimation of the log relative risk, $\theta_5 - \theta_1$. Ext-Int Data is the case that external data are used to estimate the measurement error variance. Int Data is the case that the internal data have 2 replicates, and the Ignore ME estimator ignores the measurement error and is based on the mean of these replicates. Coverage is the coverage rate of nominal 95% confidence intervals. RMSE is the square root of the mean squared error.

		Log Relative Risk Analysis				
			Mean	Actual		
		mean	Estimated	Standard		
Data	Method	bias	Std. Err.	Deviation	RMSE	Coverage
X observed		0.016	0.304	0.301	0.301	95.2%
Ext-Int Data						
	Our Method	-0.005	0.41	0.402	0.402	94.5%
	Ignore ME	-0.453	0.251	0.256	0.520	0%
Int Data						
	Our method	0.005	0.361	0.323	0.323	95.9%
	Ignore ME	-0.287	0.268	0.266	0.391	80.2%

Table 2
Simulation study for linear regression in Section 4.2 with $n = 500$ and, where applicable, the external study has sample size $N = 300$ and 2 replicates, while $\beta_0 = 0$, $\beta_1 = 0.75$. The target parameter, $\boldsymbol{\Theta} = (\theta_1, ..., \theta_5)^{\mathrm{T}}$, where θ_j is the parameter for the j^{th} category. Displayed are results for the estimation of $\theta_5 - \theta_1$. Ext-Int Data is the case that external data are used to estimate the measurement error variance. Int Data is the case that the internal data have 2 replicates, and the Ignore ME estimator ignores the measurement error and is based on the mean of these replicates. Coverage is the coverage rate of nominal 95% confidence intervals. RMSE is the square root of the mean squared error.

		Results Analysis ($\theta_5 - \theta_1$)				
			Mean	Actual		
		mean	Estimated	Standard		
Data	Method	bias	Std. Err.	Deviation	RMSE	Coverage
			Homoscedastic $\epsilon \sim N(0, 1)$			
X observed		0.004	0.145	0.150	0.150	95.1%
Ext-Int Data						
	Our Method	0.013	0.249	0.233	0.233	95.8%
	Ignore ME	-0.814	0.139	0.142	0.826	0.1%
Int Data						
	Our method	-0.007	0.176	0.170	0.170	95.3%
	Ignore ME	-0.536	0.142	0.145	0.555	3.7%
			Heteroscedastic $\epsilon \sim N(0, 0.2 + 0.5x^2)$			
X observed		0.004	0.123	0.169	0.169	85.3%
Ext-Int Data						
	Our Method	0.011	0.261	0.245	0.245	95.9%
	Ignore ME	-0.814	0.122	0.135	0.825	0.1%
Int Data						
	Our Method	-0.010	0.197	0.189	0.189	95.9%
	Ignore ME	-0.537	0.123	0.141	0.555	1.8%

Table 3

Data analysis for logistic regression in Section 5. The target parameter, $\mathbf{\Theta} = (\theta_1, ..., \theta_5)^{\mathrm{T}}$, where θ_j is the parameter for the j^{th} category. Displayed are results for the estimation of the log relative risk, $\theta_5 - \theta_1$. Ext-Int Data is the case that external data are used only to estimate the measurement error variance, and the external data have 2 replicates. Int Data is the case that the internal data have 2 replicates, and the Ignore ME estimator ignores the measurement error and is based on the mean of these replicates. Asymptotic Std. Err. is the standard error estimate from the theory. CI is the nominal 95% confidence interval for the log relative risk. p-value is the p-value for the test that the log relative risk = 0.

		Log Relative Risk Analysis			
			Asymptotic		
Data	Method	Estimate	Std. Err.	95% CI	p-value
Ext-Int Data					
	Our Method	0.98	0.47	$(0.06, 1.90)$	0.036
	Ignore ME	0.47	0.24	$(0.00, 0.95)$	0.049
Int Data					
	Our Method	1.10	0.34	$(0.43, 1.77)$	0.001
	Ignore ME	0.83	0.22	$(0.39, 1.26)$	0.000

Table 4

Data analysis in for linear regression Section 5. The target parameter, $\mathbf{\Theta} = (\theta_1, ..., \theta_5)^{\mathrm{T}}$, where θ_j is the parameter for the j^{th} category. Displayed are results for the estimation of $\theta_5 - \theta_1$. Ext-Int Data is the case that external data are used only to estimate the measurement error variance, and the external data have 2 replicates. Int Data is the case that the internal data have 2 replicates, and the Ignore ME estimator ignores the measurement error and is based on the mean of these replicates. Asymptotic Std. Err. is the standard error estimate from the theory. CI is the nominal 95% confidence interval for $\theta_5 - \theta_1$. p-value is the p-value for the test that $\theta_5 - \theta_1 = 0$.

		Results Analysis $(\theta_5 - \theta_1)$			
			Asymptotic		
Data	Method	Estimate	Std. Err.	95% CI	p-value
Ext-Int Data					
	Our Method	0.59	0.18	$(0.24, 0.95)$	0.001
	Ignore ME	0.28	0.10	$(0.09, 0.47)$	0.004
Int Data					
	Our Method	0.56	0.13	$(0.30, 0.81)$	0.000
	Ignore ME	0.35	0.09	$(0.18, 0.52)$	0.000

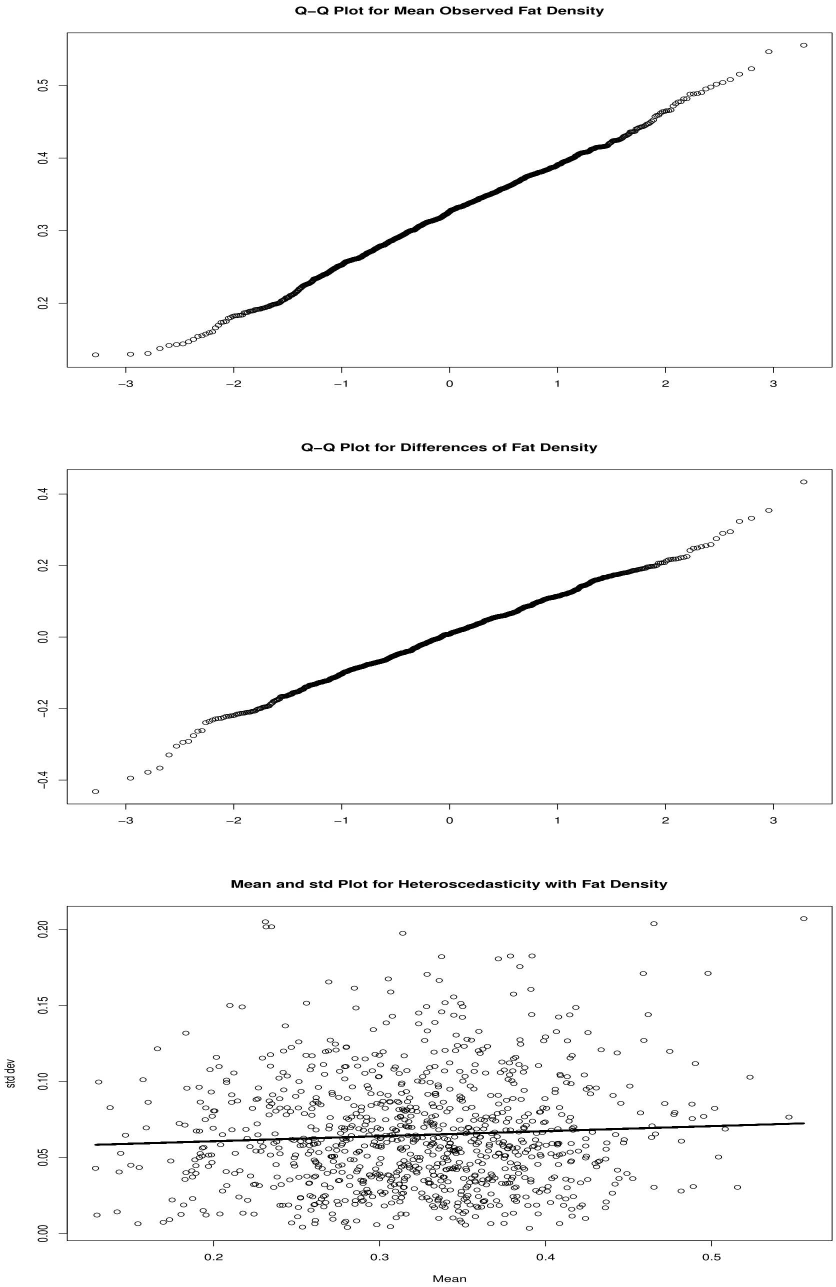

Fig 1. *EATS data of Section 5. Top panel: Normal qq-plot of the mean Fat Density over 4 recalls. This indicates that the mean Fat Density is approximately normally distributed and qualifies for the assumptions in our numerical example. Middle panel: Normal qq-plot of differences of observed Fat density, as a diagnosis that U is approximately normally distributed. Bottom panel: Mean and standard deviation plot to diagnose heteroscedasticity, showing that there is little heteroscedasticity in the measurement errors.*

Appendix C: Estimating equations for linear and logistic regression

C.1. Estimating the nuisance parameter $\boldsymbol{\Lambda}$

Here we only consider two cases among numerous possibilities. One is that the internal data consists of (Y_i, W_i, Z_i) for $i = 1, ...n$ and σ_u^2 is estimated from the external data using replicates W_{ik} for $k = 1, ..., K$ and $i = n+1, ..., n+N$. The second case is that the replicates are in the internal data.

C.1.1. External-internal data

For specificity, we consider the first case that the external data have no responses Y, are independent of the internal data. Suppose that we use external data only to estimate σ_u^2, and we observe $W_{ik} = X_i + U_{ik}$ for $k = 1, ..., K$ and $i = n+1, ..., n+N$. We use internal data to estimate μ_x, σ_x^2 without replicates. In the external data, let $\overline{W}_{i\cdot} = K^{-1}\sum_{k=1}^K W_{ik}$. Define $\widehat{\sigma}_{u,i}^2 = (K-1)^{-1}\sum_{k=1}^K (W_{ik} - \overline{W}_{i\cdot})^2$ to be the sample variance of the W_{ik} for a given i. Because $E\{(W_i - \mu_x)^2) = \sigma_x^2 + \sigma_u^2$, unbiased estimating equations for $(\boldsymbol{\Lambda}_{\text{ext}}, \boldsymbol{\Lambda}_{\text{int}}) = (\mu_x, \sigma_x^2, \sigma_u^2)$ are

$$\begin{aligned}
&\text{For } \mu_x: && n^{-1}\textstyle\sum_{i=1}^n (W_i - \mu_x) = 0;\\
&\text{For } \sigma_u^2: && N^{-1}\textstyle\sum_{i=n+1}^{n+N} (\widehat{\sigma}_{u,i}^2 - \sigma_u^2) = 0;\\
&\text{For } \sigma_x^2: && n^{-1}\textstyle\sum_{i=1}^n \{(W_i - \mu_x)^2 - \sigma_x^2 - \sigma_u^2\} = 0.
\end{aligned}$$

C.1.2. Internal data only

Suppose there is no external data, and we have replicates W_{ir} for $r = 1, ..., R$ in the internal data. Now we use internal data to estimate $\Lambda = (\mu_x, \sigma_x^2, \sigma_{uR}^2)$, and we observe $W_{ir} = X_i + U_{ir}$ for $r = 1, ..., R$ and $i = 1, ..., n$.

Define $\overline{W}_{i\cdot} = R^{-1}\sum_{r=1}^R W_{ir}$. Define $\widehat{\sigma}_{u,i}^2$ to be the sample variance of the W_{ir} within subject i, and define $\sigma_u^2/R = \sigma_{uR}^2$. The estimating equations are

$$\begin{aligned}
&\text{For } \mu_x: && n^{-1}\textstyle\sum_{i=1}^n (\overline{W}_{i\cdot} - \mu_x) = 0;\\
&\text{For } \sigma_{uR}^2: && n^{-1}\textstyle\sum_{i=1}^n (\widehat{\sigma}_{u,i}^2/R - \sigma_{uR}^2) = 0;\\
&\text{For } \sigma_x^2: && n^{-1}\textstyle\sum_{i=1}^n \{(\overline{W}_{i\cdot} - \mu_x)^2 - \sigma_x^2 - \sigma_{uR}^2\} = 0.
\end{aligned}$$

Since the two cases we considered are the same as in linear regression and logistic regression, the way we estimate Λ_{int} and Λ_{ext} are exactly the same. Then we will only give details for the estimating equations about $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}$ below.

C.2. Details for linear regression

C.2.1. Background

Here we give full details of our methodology for linear regression. As in Lemma 1, $\boldsymbol{\Omega} = (\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{\Lambda}_{\text{int}}, \boldsymbol{\Lambda}_{\text{ext}})$.

Let $\widetilde{Z} = (1, Z^{\mathrm{T}})^{\mathrm{T}}$. Here we consider the simple case of linear regression with the classical measurement error model in both the external and internal data sets to be

$$\begin{aligned} Y &= X\beta_1 + \widetilde{Z}^{\mathrm{T}}\boldsymbol{\beta}_2 = (X, \widetilde{Z}^{\mathrm{T}})\boldsymbol{\beta}; \\ W &= X + U; \qquad X = \mathrm{Normal}(\widetilde{Z}^{\mathrm{T}}\boldsymbol{\alpha}, \sigma_x^2); \qquad U = \mathrm{Normal}(0, \sigma_u^2). \end{aligned}$$

C.2.2. The forms of $\Phi(\cdot)$

In this linear model, denote the estimating equations for $\boldsymbol{\beta}$ as $\Phi(\cdot)$, we consider

$$\Phi(Y, W, Z, \boldsymbol{\beta}, \boldsymbol{\Lambda}_{\mathrm{int}}, \boldsymbol{\Lambda}_{\mathrm{ext}}) \quad = (1, W)^{\mathrm{T}}(Y - W\beta_1 - \widetilde{Z}^{\mathrm{T}}\beta_2) + (0, \beta_1\sigma_u^2)^{\mathrm{T}}.$$

C.2.3. The forms of $\Phi_{\mathrm{cat}}(\cdot)$ *and* $Q(\cdot)$

Since we assume the true model is $Y = (X, \widetilde{Z}^{\mathrm{T}})\boldsymbol{\beta}$, it is easy to see that categorical estimating function

$$\Phi_{\mathrm{cat}}\{Y, M^{\mathrm{T}}(X, Z)\boldsymbol{\Theta}\} = M(X, Z)[Y - M^{\mathrm{T}}(X, Z)\boldsymbol{\Theta}].$$

Hence, by simple calculations and following Remark 3, with $\boldsymbol{\Omega} = (\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{\Lambda}_{\mathrm{int}}, \boldsymbol{\Lambda}_{\mathrm{ext}})$,

$$Q(W, Z, \boldsymbol{\Omega}) \quad = \quad E\left[M(X, Z)\left\{(X, \widetilde{Z}^{\mathrm{T}})\boldsymbol{\beta} - M^{\mathrm{T}}(X, Z)\boldsymbol{\Theta}\right\}\middle| W, Z\right].$$

We used the `integrate` function in the R package `stats` to compute the integrals.

The estimating function for $\boldsymbol{\beta} = (\beta_0, \beta_1)$ is

$$\Phi(\boldsymbol{\beta}, \widehat{\boldsymbol{\Lambda}}) \quad = \quad n^{-1}\textstyle\sum_{i=1}^{n} E\left([Y_i - H\{m(X_i, \boldsymbol{\beta})\}]\partial m(X_i, \boldsymbol{\beta})/\partial\boldsymbol{\beta}^{\mathrm{T}}\middle| W_i\right).$$

The estimating function for $\boldsymbol{\Theta}$ is

$$Q(W_i, \boldsymbol{\Theta}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Lambda}}) \quad = \quad E\left[\begin{array}{c} m(X_i, \widehat{\boldsymbol{\beta}})I(X_i \in C_1) - \boldsymbol{\Theta}_1 I(X_i \in C_1) \\ \vdots \\ m(X_i, \widehat{\boldsymbol{\beta}})I(X_i \in C_J) - \boldsymbol{\Theta}_J I(X_i \in C_J) \end{array}\middle| W_i\right].$$

Asymptotic standard errors were estimated as in Lemma 1 and Lemma 2.

C.3. Details for logistic regression

C.3.1. Background

Here we give full details of our methodology for logistic regression. As in Lemma 1, $\boldsymbol{\Omega} = (\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{\Lambda}_{\mathrm{int}}, \boldsymbol{\Lambda}_{\mathrm{ext}})$.

As before, let $H(\cdot)$ denote the logistic distribution function and let $\widetilde{Z} = (1, Z^{\mathrm{T}})^{\mathrm{T}}$. Here we consider the special case of linear logistic regression with the classical measurement error model in both the external and internal data sets to be

$$\begin{aligned} \mathrm{pr}(Y=1|X,Z) &= H(X\beta_1 + \widetilde{Z}^{\mathrm{T}}\boldsymbol{\beta}_2) = H\{(X, \widetilde{Z}^{\mathrm{T}})\boldsymbol{\beta}\}; \\ W &= X + U; \quad X = \mathrm{Normal}(\widetilde{Z}^{\mathrm{T}}\boldsymbol{\alpha}, \sigma_x^2); \quad U = \mathrm{Normal}(0, \sigma_u^2). \end{aligned}$$

Different from the linear case in Section C.2, we consider the case where X depends on another covariate Z. There are numerous data structures possible, but we here present the external-internal and internal data only cases.

C.3.2. Settings

There are two settings of interest.

- There is no information about σ_u^2 in the internal data, so that the external parameter is the measurement error variance, $\boldsymbol{\Lambda}_{\mathrm{ext}} = \sigma_u^2$, while the internal parameters are $\boldsymbol{\Lambda}_{\mathrm{int}} = (\boldsymbol{\alpha}^{\mathrm{T}}, \sigma_x^2)^{\mathrm{T}}$.
- There are no external data, so that $\boldsymbol{\Lambda}_{\mathrm{ext}}$ is null, and the internal data with replicates allow estimation of $\boldsymbol{\Lambda}_{\mathrm{int}} = (\boldsymbol{\alpha}^{\mathrm{T}}, \sigma_u^2, \sigma_x^2)^{\mathrm{T}}$.

In both case, σ_u^2 (or σ_{uR}^2 in the internal data only case) are estimated the same as in C.1.1 and C.1.2, while the estimating function for $(\boldsymbol{\alpha}, \sigma_x^2)$ is

$$V_{\mathrm{int},i}(\boldsymbol{\Lambda}_{\mathrm{int}}, \boldsymbol{\Lambda}_{\mathrm{ext}}) = \left\{ \widetilde{Z}_i^{\mathrm{T}}(W_i - \widetilde{Z}_i^{\mathrm{T}}\boldsymbol{\alpha}), (W_i - \widetilde{Z}_i^{\mathrm{T}}\boldsymbol{\alpha})^2 - \sigma_x^2 - \sigma_u^2 \right\},$$

where $i = 1, ..., n$.

C.3.3. Estimating $\boldsymbol{\beta}$

In this section, we implement our method and give all estimating equations in the case where we have both external and internal data. In another case, where we only use internal data with replicates, all results below are still valid by removing $\boldsymbol{\Lambda}_{\mathrm{ext}}$.

Define $\lambda = \sigma_x^2/(\sigma_x^2 + \sigma_u^2)$. Then, given (W, Z), X is normally distributed with mean $\mu(W, Z, \boldsymbol{\Lambda}_{\mathrm{ext}}, \boldsymbol{\Lambda}_{\mathrm{int}}) = \widetilde{Z}^{\mathrm{T}}\boldsymbol{\alpha} + \lambda(W - \widetilde{Z}^{\mathrm{T}}\boldsymbol{\alpha})$ and variance $\lambda\sigma_u^2$. We write this conditional density as $f_{x|w,z}(x, w, z, \boldsymbol{\beta}, \boldsymbol{\Lambda}_{\mathrm{int}}, \boldsymbol{\Lambda}_{\mathrm{ext}})$.

There are multiple ways to estimate $\boldsymbol{\beta}$ from the observed data. Here we describe two of them.

- The first is regression calibration, in which X is replaced by its mean given (W, Z) and the linear logistic model is fit. Thus the regression calibration method has

$$\begin{aligned} \Phi(Y, W, Z, \boldsymbol{\beta}, \boldsymbol{\Lambda}_{\mathrm{int}}, \boldsymbol{\Lambda}_{\mathrm{ext}}) = \{&\mu(W, Z, \boldsymbol{\Lambda}_{\mathrm{ext}}, \boldsymbol{\Lambda}_{\mathrm{int}}), \widetilde{Z}\}^{\mathrm{T}} \\ &\times[Y - H\{\mu(W, Z, \boldsymbol{\Lambda}_{\mathrm{ext}}, \boldsymbol{\Lambda}_{\mathrm{int}})\beta_1 + \widetilde{Z}^{\mathrm{T}}\boldsymbol{\beta}_2\}]. \end{aligned}$$

- A second possibility, one that we used, is the following. By simple calculations, $\mathrm{pr}(Y=1|W,Z)=p(W,Z,\boldsymbol{\beta},\boldsymbol{\Lambda}_{\mathrm{int}},\boldsymbol{\Lambda}_{\mathrm{ext}})$, where

  $$p(W,Z,\boldsymbol{\beta},\boldsymbol{\Lambda}_{\mathrm{int}},\boldsymbol{\Lambda}_{\mathrm{ext}})=\int H\{(x,\widetilde{Z}^{\mathrm{T}})\boldsymbol{\beta}\}f_{x|w,z}(x,W,Z,\boldsymbol{\Lambda}_{\mathrm{int}},\boldsymbol{\Lambda}_{\mathrm{ext}})dx,$$

  a quantity that is easily computed in R using the `integrate` function in the R package `stats`. Denote $p_i=\mathrm{pr}(Y_i=1|W_i,Z_i)$. Thus, the loglikelihood $\propto n^{-1}\sum_{i=1}^{n}Y_i\log(p_i)+(1-Y_i)\log(1-p_i)$. We then use `optim` function in the R package `stats` to minimize the negative loglikelihood to estimate $\boldsymbol{\beta}$.

C.3.4. The forms of $\Phi_{\mathrm{cat}}(\cdot)$ and $Q(\cdot)$

Since we assume the true model is $\mathrm{pr}(Y=1|X,Z)=H\{(X,\widetilde{Z}^{\mathrm{T}})\boldsymbol{\beta}\}$, it is easy to see that categorical estimating function

$$\Phi_{\mathrm{cat}}\{Y,M^{\mathrm{T}}(X,Z)\boldsymbol{\Theta}\}=M(X,Z)[Y-H\{M^{\mathrm{T}}(X,Z)\boldsymbol{\Theta}\}].$$

Hence, with $\boldsymbol{\Omega}=(\boldsymbol{\Theta},\boldsymbol{\beta},\boldsymbol{\Lambda}_{\mathrm{int}},\boldsymbol{\Lambda}_{\mathrm{ext}})$, by simple calculations and following Remark 3,

$$Q(W,Z,\boldsymbol{\Omega}) \quad = \quad E\left(M(X,Z)\left[H\{(X,\widetilde{Z}^{\mathrm{T}})\boldsymbol{\beta}\}-H\{M^{\mathrm{T}}(X,Z)\boldsymbol{\Theta}\}\right]\Big|W,Z\right).$$

We used the `integrate` function in the R package `stats` to compute the integrals.

Acknowledgments

We thank the Associate Editor and two anonymous referees for their valuable comments. Blas and Wang should be considered as joint first authors. Blas's research was supported by a post-doctoral fellowship from the Brazilian Agency CNPq (201192/2015-2). The research of Wang and Carroll was supported by a grant from the National Cancer Institute (U01-CA057030). The R package `CCP` for implementing the methods has been placed on GitHub at https://github.com/tianyingw/CCP. The Eating at America's Table Study data in Section 5 can be obtained from a data transfer agreement with the National Cancer Institute: our R package can generate simulated data as in Section 4 as a check for reproducibility.

References

[1] Arem, H., Reedy, J., Sampson, J., Jiao, L., Hollenbeck, A. R., Risch, H., Mayne, S. T., and Stolzenberg-Solomon, R. Z. (2013). The Healthy Eating Index 2005 and risk for pancreatic cancer in the NIH–AARP Study. *Journal of the National Cancer Institute* **105**, 1298–1305.

[2] Berry, S. M., Carroll, R. J., and Ruppert, D. (2002). Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association* **97**, 457, 160–169. MR1947277

[3] Buonaccorsi, J. P. (2010). *Measurement Error: Models, Methods and Applications*. Chapman & Hall. MR2682774

[4] Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*. Chapman and Hall. MR2243417

[5] Chaix, B., Kestens, Y., Duncan, D. T., Brondeel, R., Méline, J., Aarbaoui, T. E., Pannier, B., and Merlo, J. (2016). A gps-based methodology to analyze environment-health associations at the trip level: case-crossover analyses of built environments and walking. *American Journal of Epidemiology* **184**, 8, 579–589.

[6] Cook, J. R. and Stefanski, L. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association* **89**, 1314–1328. MR1379467

[7] Cordy, C. B. and Thomas, D. R. (1997). Deconvolution of a distribution function. *Journal of the American Statistical Association* **92**, 1459–1465. MR1615256

[8] Devanarayan, V. and Stefanski, L. A. (2002). Empirical simulation extrapolation for measurement error models with replicate measurements. *Statistics & Probability Letters* **59**, 219–225. MR1932865

[9] Eckert, R. S., Carroll, R. J., and Wang, N. (1997). Transformations to additivity in measurement error models. *Biometrics* **53**, 262–272. MR1450184

[10] Evenson, K. R., Wen, F., and Herring, A. H. (2016). Associations of accelerometry-assessed and self-reported physical activity and sedentary behavior with all-cause and cardiovascular mortality among us adults. *American Journal of Epidemiology* **184**, 10, 621–632.

[11] Flegal, K. M., Keyl, P. M., and Nieto, F. J. (1991). Differential misclassification arising from nondifferential errors in exposure measurement. *American Journal of Epidemiology* **134**, 10, 1233–1246.

[12] Ganguli, B., Staudenmayer, J., and Wand, M. P. (2005). Additive models with predictors subject to measurement error. *Australian & New Zealand Journal of Statistics* **47**, 2, 193–202. MR2155119

[13] Gustafson, P. (2004). *Measurement Error and Misclassication in Statistics and Epidemiology*. Chapman and Hall/CRC. MR2005104

[14] Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 1, 221–233. MR0216620

[15] Kauermann, G. and Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association* **96**, 456, 1387–1396. MR1946584

[16] Lederer, W. and Küchenhoff, H. (2006). A short introduction to the simex and mcsimex. *The Newsletter of the R Project Volume 6/4*, October 2006, 26.

[17] Nakamura, T. (1990). Corrected score function for errors-in-variables models: methodology and application to generalized linear models. *Biometrika* **77**, 127–137. MR1049414

[18] Pham, T. H., Ormerod, J. T., and Wand, M. P. (2013). Mean field variational Bayesian inference for nonparametric regression with measurement error. *Computational Statistics & Data Analysis* **68**, 375–387. MR3103783

[19] Reedy, J., Wirfält, E., Flood, A., Mitrou, P. N., Krebs-Smith, S. M., Kipnis, V., Midthune, D., Leitzmann, M., Hollenbeck, A., Schatzkin, A., and others. (2010). Comparing 3 dietary pattern methods – cluster analysis, factor analysis, and index analysis – with colorectal cancer risk: the NIH–AARP Diet and Health Study. *American Journal of Epidemiology* **171**, 479–487. MR1131644

[20] Reedy, J. R., Mitrou, P. N., Krebs-Smith, S. M., Wirfält, E., Flood, A. V., Kipnis, V., Leitzmann, M., Mouwand, T., Hollenbeck, A., Schatzkin, A., and Subar, A. F. (2008). Index-based dietary patterns and risk of colorectal cancer: the NIH-AARP Diet and Health Study. *American Journal of Epidemiology* **168**, 38–48.

[21] Sarkar, A., Mallick, B. K., Staudenmayer, J., Pati, D., and Carroll, R. J. (2014). Bayesian semiparametric density deconvolution in the presence of conditionally heteroscedastic measurement errors. *Journal of Computational and Graphical Statistics* **23**, 1101–1125. MR3270713

[22] Stefanski, L. A. and Cook, J. R. (1995). Simulation-extrapolation: the measurement error jackknife. *Journal of the American Statistical Association* **90**, 1247–1256. MR1379467

[23] Subar, A. F., Thompson, F. E., Kipnis, V., Mithune, D., Hurwitz, P., McNutt, S., McIntosh, A., and Rosenfeld, S. (2001). Comparative validation of the Block, Willett, and National Cancer Institute food frequency questionnaires: The Eating at America's Table Study. *American Journal of Epidemiology* **154**, 1089–1099.

[24] Trentham-Dietz, A., Newcomb, P. A., B, E. S., Longnecker, M. P., Baron, J., Greenberg, E. R., and Willett, W. C. (1997). Body size and risk of breast cancer. *American Journal of Epidemiology* **145**, 11, 1011–1019.

[25] Wang, Y., Wellenius, G. A., Hickson, D. A., Gjelsvik, A., Eaton, C. B., and Wyatt, S. B. (2016). Residential proximity to traffic-related pollution and atherosclerosis in 4 vascular beds among African-American adults: Results from the Jackson Heart Study. *American Journal of Epidemiology* **184**, 10, 732–743.

[26] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25. MR0640163

[27] Yi, G. Y. (2017). *Statistical Analysis with Measurement Error or Misclassification: Strategy, Method and Application.* Springer. MR3676914

[28] Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44**, 1049–1060. MR0980999