Measurement error in regression; model-based bootstrap and penalized regressions.

Giovanni Romeo

Dissertation for the degree of PhD

Oslo Centre for Biostatistics and Epidemiology
Department of Biostatistics
Institute of Basic Medical Sciences
Faculty of Medicine
University of Oslo

Oslo, December 2018

Series of dissertations submitted to the
Faculty of Medicine, University of Oslo

ISBN 978-82-8377-495-5

Cover: Hanne Baadsgaard Utigard.
Print production: Reprosentralen, University of Oslo.

Acknowledgments

First of all, I would like to express my sincere gratitude to my supervisor Magne Thoresen for his continuous support, for his patience, motivation, and inspiration. His guidance helped me in all the time of research and writing of this thesis and articles. I would also like to deeply thank my co-supervisor John P. Buonaccorsi for his insightful comments and encouragement, and for all his stimulating questions and suggestions.

The Department of Biostatistics has been a great academic and social environment and contributed immensely to my personal and professional time in Norway. I thanks all the colleagues for creating such a stimulating and pleasant atmosphere.

My sincerest thanks to Morten Wang Fagerland and the Section of Biostatistics and Epidemiology of the Oslo University hospital for the opportunity they offered me, that allowed me to finish my thesis in serenity.

Most of all I want to thank my wife and soul mate, Veronica, for her endless love, constant patience and invaluable support. She always kept me going, and never stopped to believe in me.

List of papers

Paper I

Buonaccorsi, J.P., Romeo, G., and Thoresen, M. (2018). Model-based bootstrapping when correcting for measurement error with application to logistic regression. *Biometrics*, 74(1), 135-144.

Paper II

Romeo, G., Buonaccorsi, J.P., and Thoresen, M. (2018). Detection and correction of heteroscedasticity under measurement error with non-constant variance. Submitted to *Statistics in Medicine*.

Paper III

Romeo, G. and Thoresen, M. (2018). Model selection in high-dimensional noisy data: a simulation study. *Journal of Statistical Computation and Simulation*, conditionally accepted.

Contents

Abbreviations

Paper I	RC	Regression Calibration
Paper II	MR	Moment Reconstruction
	MM	Moment Matched
Paper III	MUS	Matrix Uncertainty Selector
	NCL	Non-Convex Lasso
	CoCoLasso	Conditionally Convex Lasso

Chapter 1

Introduction

This thesis is about measurement error. In almost all disciplines, it may not be possible to observe a variable accurately, for some reason, and therefore it is necessary to work with an error-prone version of that variable. Any measurement process can be affected by errors, due for example to the measuring instrument or to the sampling process.

Examples of measurement error can be found in many different scientific fields. For example, in epidemiology, biometric screening variables are useful for identifying the causes and risk factors of many diseases, and variables such as blood pressure, cholesterol, or other blood chemistry values typically involve some random variations or errors in their measurements. Another example might be the measurement of pollutants as exposure for a certain disease; here the source of measurement error in the sampling is due to both time and location of the measurements. Another source of measurement error is the self-reporting nature of a variable, a common problem in surveys. Furthermore, in recent decades, genetic and epigenetic studies have become increasingly more important in medical research, but the process of sequencing DNA typically involves some errors.

Additionally, measurement error can happen in categorical variables, in which case it is generally called misclassification. A typical example can be found in ecological studies when the presence of a species or the number of individuals is modelled as a function of habitat variables. Frequently, such variables are estimated for an average of the samples of the subareas within the region of interest, and, due to sampling, there will be some errors to the category finally assigned to the region.

When measurement error is present among the covariates of a regression model, there are three main reasons why measurement error cannot be ignored; it can

cause bias in the parameter estimation, interfere with variable selection and lead to a loss of power, leading to trouble in detecting relationships among variables.

A vast body of literature exists on measurement error. There are a number of excellent books, starting with one by Fuller [1], who wrote the first influential book focusing on linear regression models, and one by Carroll et al. [2] who treated measurement error in a much broader application context. Another book that gives wide treatment to the topic of measurement error and misclassification is by Buonaccorsi [3], who focuses on different topics from those in the aforementioned two books and places emphasis on a more applied approach. Additionally, it is worth mentioning a book by Gustafson [4], who presents a Bayesian perspective on measurement error treatment. However, in this thesis, I do not consider a Bayesian approach in measurement error treatment. Furthermore, Yi [5] has written one of the most recent books about measurement error and misclassification in medical research and epidemiological studies.

A plethora of correction methods for measurement error in both linear and non-linear models is available, among which the most popular method is regression calibration. However, despite the availability of techniques that can take into account measurement error, inference can be challenging for a variety of reasons. When available, analytical standard errors are asymptotic and thus, approximate and rely on some underlying assumptions. Additionally, most of the corrected estimators, despite being consistent or approximately consistent under certain conditions, are biased.

The bootstrap method is powerful and allows for inference when analytical alternatives are not available, but it has received limited attention in the measurement error context. The majority of the bootstrap applications in measurement error problems have only considered the simple bootstrap – the resampling of observations. Furthermore, the method is the only one implemented in the few statistical software packages that can handle measurement error.

In **Paper I**, we explore the use of model-based bootstrap when correcting for measurement error. We suggest new methodologies that not only offer some definitive advantages over the simple bootstrap and other standard methods of inference when it comes to estimation of standard errors, but also are able to estimate the bias of the corrected estimators.

A central assumption of regression theory is the homoscedasticity of the model error. Distribution theory, confidence intervals, and hypothesis testing all rely on this assumption. Standard asymptotic theory suggests that ignoring the presence of the heteroscedasticity, despite having no effect on the first-order properties of

the estimates of the regression parameters, will lead to inefficiency and faulty inference. There is a very large body of literature on measurement error theory, but most of it focuses on the correction of bias in the estimations of the coefficients. Far less consideration has been given to the analysis of residuals and assessment of homoscedasticity. Accordingly, in **Paper II**, we explore the available methods for residuals estimation, present a developed model-based bootstrap test for heteroscedasticity, and, through a practical application with binomial measurement error, we show how modelling heteroscedasticity can affect prediction intervals.

In recent decades, technological progress has led to a great abundance of data in many scientific fields. For example, in genetics, a new framework has been developed, in which the number of variables is larger than the number of observations. High-dimensional data analysis has had a tremendous growth in popularity, and a plethora of methods has been proposed for statistical modelling of, and inference in, high-dimensional data. Penalized regression methods are particularly good in this context and probably the most popular of these is the Lasso method [6]. The penalized regression methods have been extended, adapted, and improved for many different cases and application contexts.

Applying high-dimensional regression methods that do not correct for measurement error results in faulty inference, as demonstrated for the Lasso and the Dantzig selector [7, 8]. Consequently, correction for measurement error in penalized linear regression has recently been studied by various authors. Examples include the Non-Convex Lasso (NCL) by Loh and Wainwright [9] the Conditionally Convex Lasso (CoCoLasso) of Datta and Zou [10] the Matrix Uncertainty Selector (MUS) proposed by Rosenbaum and Tsybakov [7]. Despite their interesting theoretical properties, not all of the aforementioned methods have been compared from a practical viewpoint and therefore it is unclear which of them can offer advantages over the others in a real data application. In **Paper III**, we evaluate these methods and focus on situations that are relevant to a practical application context, in particular focusing on situations in which the measurement error distribution and dependence structure are not known and need to be estimated from the data.

The organization of this thesis is as follows. Chapter 2, introduces the measurement error in regression theory, provides an overview of the consequences of measurement error in linear regression and generalized linear model, and introduces some correction methods. Chapter 3 describes the simple bootstrap method and model-based bootstrap approaches. Chapter 4 introduces high-dimensional regressions, focusing on Lasso. Chapter 5 states the aims of the thesis and Chapter

6 presents a summary of the three papers that form the basis of the thesis. Finally, Chapter 7 discusses the contributions, strengths, and weaknesses of Papers I–III.

Chapter 2

Measurement error in regression

2.1 Background

The consequences of ignoring measurement error in statistical analysis can range from negligible to rather considerable. For example, when a covariate is affected by measurement error, it is not necessary to model measurement error when the purpose is prediction under certain conditions (section 2.6 of [2]). However, if the measurement error is large, the estimations of the coefficients and the variable selection of a statistical model will be greatly affected. Statistical analysis that is carried out by ignoring the presence of the measurement error is called a naive approach.

Measurement error can affect covariates in different ways. One of the most common ways of modelling is additive measurement error. The variable $x \in \mathbb{R}_n$, where n is the number of observed units, cannot be directly observed due to measurement error.

In the additive measurement error model, the unobservable variable x is altered by adding a random measurement error u, and the following is observed

$$w_i = x_i + u_i \quad \text{for} \quad i = 1, ..., n.$$

A key assumption is that the measurement error is centered on zero $E(u_{ij}) = 0$ such that w is an unbiased measurement of x_i, $E(w_i|x_i) = x_i$. The variance of the measurement error is σ_u^2 and, in many applications, is assumed to be constant for all the units and known. However, there are methods that allow to relax this assumption and estimate the measurement error variance. Another important assumption is that the measurement error u is assumed to be independent from the

error ϵ of the assumed statistical model.

It is possible to distinguish between the functional case, in which x_i is treated as fixed, and the structural case where x_i is the realization of the random variable X_i. In the latter case, the X_i are assumed to be independent and identically distributed (i.i.d.) with mean μ_X and covariance $\mathbf{\Sigma}_X$.

There are other alternatives to additive measurement error modelling, such as multiplicative measurement error. In multiplicative measurement error, the measurement error u will interfere multiplicatively and $w_i = x_i \cdot u_i$. This approach can also be used to model missing data, considering a dichotomous measurement error that will take value zero when data are missing and value one otherwise. However, in this thesis, I focus on additive measurement error.

2.2 Measurement error in simple linear regression

Regression analysis is the most common and widely used statistical model. To outline the problems caused by measurement error, I focus on simple linear regression but any conclusion can be easily extended to multiple linear regression.

The main model can be formulated as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, ..., n$$

where x_i is the true unobservable value of the predictor for the i-th individual. The model error ϵ_i is assumed to be i.i.d. as $N(0, \sigma_\epsilon^2)$ for $i = 1, ..., n$. If it is possible to observe x, the estimated coefficients will be $\hat{\beta}_1 = \sum_i (x_i - \bar{x})(y_i - \bar{y}) / \sum_i (x_i - \bar{x})^2$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, where $\bar{y} = \sum_i y_i / n$ and $\bar{x} = \sum_i x_i / n$. The estimator of the model error variance σ_ϵ^2 will be $\hat{\sigma}_\epsilon^2 = \sum_i (y_i - \hat{y}_i)^2 / (n - p)$, where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is the i-th fitted value. It is well known that these estimators are unbiased and, under normality, maximum likelihood estimators.

If measurement error is present, instead of x_i, w_i will be observed. The first and most natural approach is to use w in place of x - the naive approach.

The naive estimators of the coefficients and of the error variance are

$$\hat{\beta}_{1,naive} = S_{wy} / S_{ww}, \qquad \hat{\beta}_{0,naive} = \bar{y} - \hat{\beta}_{1,naive} \bar{w}$$

and

$$\hat{\sigma}^2_{\epsilon,naive} = \sum_i (y_i - \hat{\beta}_{0,naive} - \hat{\beta}_{1,naive} w_i)^2 / (n - 2)$$

where $\bar{w} = \sum_i w_i/n$, $\bar{y} = \sum_i y_i/n$,

$$S_{wy} = \frac{\sum_i (w_i - \bar{w})(y_i - \bar{y})}{n-1}, \qquad S_{ww} = \frac{\sum_i (w_i - \bar{w})^2}{n-1}.$$

It is well known that for simple linear regression the naive estimators will be biased. The observed variability in w overestimates the variability of the true x and, consequentially, the naive estimation of β_1 will approximately estimate $\kappa\beta_1$ instead of β_1, where $\kappa = \sigma_x^2/(\sigma_x^2 + \sigma_u^2)$ is the reliability ratio. The bias in the estimation of β_1 will also lead to bias in the estimation of β_0, and the estimator $\hat{\sigma}^2_{\epsilon,naive}$ has been demonstrated to be biased, too [1, 3].

The reliability ratio κ can only take values in the interval $[0, 1]$ with one corresponding to absence of measurement error, while zero corresponds to infinite measurement error variance. This will cause the naive estimate of the slope $\hat{\beta}_{1,naive}$ to underestimate always the true slope coefficient β_1, causing attenuation. Hence, even if biased towards zero, the naive ordinary least squares method preserves the sign of the regression coefficient asymptotically.

2.3 Measurement error in multiple linear regression

When multiple covariates are present and without measurement error, the multiple linear regression model assumes

$$y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij} + \epsilon_i = \boldsymbol{\beta}' \mathbf{x}_i + \epsilon_i$$

where $\boldsymbol{\beta}$ is the p dimensional vector of the coefficients (intercept and $p-1$ covariates), $\mathbf{x}_i = (1, x_1, ..., x_{p-1})$ is the collection of predictors and ϵ_i are the model errors, assumed as uncorrelated and with mean zero and variance σ_ϵ^2. As before, $\mathbf{x}_i$ denotes the fixed predictors in the functional case and the realization of the random vector $\mathbf{X}_i$ in the structural case. This model can be expressed in matrix form as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{X}$ is an $n \times p$ matrix with rows $\mathbf{x}_i$, $\mathbf{Y} = (y_1, ..., y_n)$ and $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)$ with $E(\boldsymbol{\epsilon}) = 0$ and $Cov(\boldsymbol{\epsilon}) = \sigma_\epsilon^2 I$. In the absence of measurement error, the estimated

coefficients are $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ and an unbiased estimator of the model error variance is $\hat{\sigma}^2_\epsilon = \sum_i (y_i - \hat{y}_i)^2/(n-p)$, where $\hat{y}_i = \hat{\boldsymbol{\beta}}'\mathbf{x}_i$.

The matrix of the covariates $\mathbf{X}$ cannot be directly observed due to measurement error. In the additive measurement error model, $\mathbf{X}$ is corrupted by adding measurement error $\mathbf{U}$, so what we observe is

$$\mathbf{W} = \mathbf{X} + \mathbf{U}$$

where $\mathbf{U}$ is an $n \times p$ random noise matrix with covariance matrix $\boldsymbol{\Sigma}_U$ (if an intercept is present in the model, the first column of $\mathbf{U}$ will be equal to zero). Also in this case, if $E(u_{ij}) = 0$ then $\mathbf{W}$ is an unbiased estimate of $\mathbf{X}$ so that $E(w_{ij}|x_{ij}) = x_{ij}\ \forall i, j$.

It is worth noting that it is possible to include in the model variables that are not affected by measurement error. If the k-th variable has been measured correctly, the corresponding column of $\mathbf{U}$ will be set equal to zero, as will the variance of the measurement error of the k-th variable, $\boldsymbol{\Sigma}_{U(k,k)} = 0$.

The naive estimations of the coefficients will be

$$\hat{\boldsymbol{\beta}}_{naive} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Y}.$$

By focusing on the non-intercept coefficients $\boldsymbol{\beta}_1$, it is possible to show the bias in $\hat{\boldsymbol{\beta}}_{1,naive}$

$$E(\hat{\boldsymbol{\beta}}_{1,naive}) = (\boldsymbol{\Sigma}_{XX} + \boldsymbol{\Sigma}_U)^{-1}\boldsymbol{\Sigma}_{XX}\boldsymbol{\beta}_1$$

where $\boldsymbol{\kappa} = (\boldsymbol{\Sigma}_{XX} + \boldsymbol{\Sigma}_U)^{-1}\boldsymbol{\Sigma}_{XX}$ is the reliability matrix. This result leads to an important conclusion: the measurement error in one of the variables may induce bias in the estimation of all coefficients, including those measured without error. If more covariates are affected by measurement error, the resulting bias may become rather complex and the effect of measurement errors may become difficult to describe.

2.4 Measurement error in generalized linear models

In many cases, the assumption of linearity between the covariates and the response variable or the assumption of normality of the response variable do not hold. Generalized linear models allow to relax some of the assumptions of the linear regression and can fit in a wider range of applications. In this chapter, I focus only on logistic regression, which is probably the most widely used non-linear regression

model, but what said can be extended to other generalized linear models as well. The logistic model assumes

$$P(Y_i = 1|\mathbf{x}_i) = E(Y_i|\mathbf{x}_i) = m(\mathbf{x}_i, \beta) = 1/(1 + e^{-\mathbf{x}_i'\beta})$$

where Y_i is the binary outcome and $m(.)$ is the link function.

As I have already mentioned, bias caused by measurement error is always in the form of attenuation for simple linear regression. However, it can become much more complex for multiple linear regression. When additive measurement error is present in logistic regression, attenuation frequently occurs [11]. In logistic regression with a single scalar x affected by additive and non-differential measurement error (i.e. the measurement error does not depend on y), the naive estimator may not preserve the sign of the true coefficient [12]]. Nevertheless, in most cases, the sign should be preserved in the naive estimation. However, this refers only to the correlation of y and w, and does not say anything about the structure of the relationship (section 3.6 of [2]). Making inference about the detailed relationship between y and x based on the observed relationship between y and w is in general a difficult problem. Moreover, the presence of multiple covariates will further complicate the structure of the measurement error effect.

2.5 Correction methods

Estimating the direction of the relationship between y and the covariates is important, but may be misleading if their magnitude is severely biased. Furthermore, the presence of measurement error will interfere with variable selection and inference for the coefficients. Correcting for measurement error can correct the bias in the estimated coefficients and subsequently lead to improvements in the power of inferences.

To correct for measurement error, additional information is required. The measurement error variance is an essential quantity for correction. Here, I assume it as known, but as pointed out in Section 2.6, in most of the cases it will be necessary to estimate the measurement error variance. There is a vast body of literature on techniques for correcting measurement error in linear regression, but here I focus mainly on regression calibration. For an extensive discussion on additional methods, see Fuller [1] and Cheng et al. [13].

The most commonly used correction method for linear regression is the unweighted moment corrected estimators method. The main idea behind the method

is to correct for the bias present in S^2_{ww} as an estimator of σ^2_x. The estimator for the simple linear regression case is defined as follows

$$\hat{\beta}_1 = S_{wy}/(S^2_{ww} - \sigma^2_u) \text{ and } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{w}$$

These corrected estimators of $\boldsymbol{\beta}$ also allows for correction of the estimation of σ^2_ϵ (section 4.5.1 [3]), leading to the method of moments estimator

$$\hat{\sigma}^2_\epsilon = \frac{\sum_i r_i^2}{n-2} - \hat{\beta}_1^2 \hat{\sigma}^2_u$$

where $r_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 w_i$ are the residuals using the naive measurement and the corrected coefficients.

2.5.1 Regression calibration

Under certain conditions, for linear regression the moment corrected estimates are equivalent to those obtained from regression calibration (RC), a widely used correction method that was originally developed to handle measurement error in generalized linear models, particularly for logistic regression. The idea behind this method is that, if the measurement error is non-differential and X_i is random, then $E(Y_i|W_i) = E(m(\boldsymbol{\beta}, X_i)|W_i)$. This implies, for the linear case, that $E(Y_i|W_i) = \beta_0 + \beta_1 E(X_i|W_i)$. In the non-linear case, an approximation is used, assuming $E(Y_i|W_i) \approx m(\boldsymbol{\beta}, E(X_i|W_i))$. So, the measurement error will be corrected by fitting the usual regression model using an estimate of $E(X_i|W_i)$ instead of W_i.

Typically, one will be using

$$\hat{x}_i = \hat{E}(X_i|W_i) = \bar{w} + \hat{\kappa}(w_i - \bar{w})$$

where $\hat{\kappa}$ is the estimated reliability ratio $\hat{\kappa} = \hat{\sigma}^2_x(\hat{\sigma}^2_x + \hat{\sigma}^2_u)^{-1}$.

Regression calibration is a very powerful method, despite its simplicity. It can considerably reduce the bias in the estimates and its simplicity makes it potentially applicable to any regression model. A simple approximation cannot always be accurate, and generally RC tends to be more useful for generalized linear models but can perform quite poorly for highly non-linear models.

Additionally, the RC method allows for corrected inference for the coefficients. An analytical approximation of the variance of the estimated coefficients of a generalized linear model is available (appendix B.3.1 of [2]), although bootstrap

resampling is more commonly used.

Moreover, the residuals of the regression of y on the RC $\hat{x}$ cannot be used to estimate the model error variance σ_ϵ^2. It can be demonstrated that the RC residuals are equal to the naive residuals and consequently, any residual analysis carried out on the RC residuals will correspond to a residual analysis where measurement error is ignored.

2.6 Estimation of the measurement error variance

Thus far, I have assumed that the measurement error variance σ_u^2 is known in the simple regression case and correspondingly for the covariance matrix $\mathbf{\Sigma}_U$ in the multiple regression case. However, in any practical applications, it is very unlikely that the measurement error variance will be known, and therefore in order to correct for measurement error, the variance will need to be estimated. Estimation of the measurement error matrix is often challenging and may require in-depth knowledge of the measurement error process. Generally, supplementary data are needed for such estimations and may be obtained from similar or complementary studies, but the more common approach is to estimate the measurement error matrix through replicated measurements of the covariates.

When replicated measurements are collected, we observe m_i replicated measurement of $\mathbf{x}_i$ for the i-th individual

$$\mathbf{W}_{ij} = \mathbf{x}_i + \mathbf{u}_{ij}, \qquad \text{for } i = 1, ..., n \text{ and } j = 1, ..., m_i$$

where the errors u_{ij} are independent with $E(u_{ij}) = 0$ and $Cov(u_{ij}) = \mathbf{\Sigma}_{u(1)}$ (the measurement error covariance matrix for one replicate for the i-th subject). Given that $E(\mathbf{u}_i) = 0$, the mean of the replicated measurements $\mathbf{W}_i = \sum_{j=1}^{m_i} \mathbf{W}_{ij}/m_i$ has expected values $\mathbf{x}_i$. The covariance matrix of $\mathbf{u}_i$ is $\mathbf{\Sigma}_{ui} = \mathbf{\Sigma}_{u(1)}/m_i$, which implies that even if $\mathbf{\Sigma}_{u(1)}$ is equal among all units, the measurement error variance in $\mathbf{W}_i$ will not be constant, with the exception of cases when the number of repeated measurements m_i are equal. It should be noted that the more replicated measurements m_i there are for the i-th subject, the less measurement error variance there will be in $\mathbf{W}_i$.

The covariance matrix $\mathbf{\Sigma}_{ui}$ can be estimated with $\hat{\mathbf{\Sigma}}_{ui} = \mathbf{S}_{W_i}/m_i$ where $\mathbf{S}_{W_i} = \sum_{j=1}^{m_i}(W_{ij} - W_i)(W_{ij} - \bar{W}_i)'/(m_i - 1)$ is the sample covariance matrix of the replicates of the i-th observation.

Finally, the average estimated measurement error covariance matrix $\hat{\boldsymbol{\Sigma}} = \sum_{i=1}^{n} \hat{\boldsymbol{\Sigma}}_{ui}/n$ can be defined. This quantity is an estimate for $\boldsymbol{\Sigma}_u$. It should be noted that to obtain this estimator, neither a constant per-replica measurement error variance, nor an equal number of replicates across the subjects is assumed.

2.7 Binomial measurement error

Binomial measurement error structure, and the inherent heteroscedasticity of the measurement error that it implies, constitute an interesting problem in the analysis of methylation data. This, in turn, partly motivated the study reported in **Paper II**.

DNA methylation analyses have recently received much attention. Methylation may occur in the DNA chain, and its presence has been shown to be strictly related to the regulation of gene expression. The study of such epigenetic data has been found relevant in the investigation of many biological disorders, such as cancer, autoimmune diseases, and neurodegenerative disorders [14].

If based on sequencing technology, methylation is measured by a number of reads (samples) for each given DNA position. The methylation rate is the proportion of such reads that are methylated. The intensity by which each DNA position has been sampled varies across the genome and also among individuals. This sampling process of methylation results in an error structure where the precision of the estimated methylation rates varies both with the true underlying proportion and the number of reads.

The topic of binomial measurement error has been addressed by Buonaccorsi [15], who, given the sampling method used in methylation sequencing, assumed a binomial structure for the measurement error.

In this context, x_i is the methylation rate for a specific DNA position in individual i. Being a proportion, x can only take values in the interval $[0, 1]$. For a specific methylation site, only m_i reads are available for the i-th individual and T_i will be the number of methylated reads among these m_i reads. T_i is a random variable and $T_i|x_i \sim Binomial(x_i, m_i)$ for $i = 1, ..., n$. The observed methylation rate is $W_i = T_i/m_i$. This will lead to $E(W_i|x_i) = x_i$ and consequently, the measurement error will be additive with measurement error variance

$$V(W_i|x_i) = x_i(1 - x_i)/m_i = \sigma_{ui}^2.$$

This measurement error structure perfectly represents the sampling process,

but implies a measurement error variance that depends on the values of x, being maximum for $x = 0.5$.

2.8 Heteroscedasticity of the model error and measurement error

As anticipated in the Introduction (Chapter 1), a central assumption of regression theory is the homoscedasticity of the model error. Ignoring the presence of the heteroscedasticity will lead to inefficient and faulty inference, such as when the model error depends on the covariates.

The analysis of the residuals and the problem of assessing and testing for heteroscedasticity in the presence of measurement error has received only limited attention to date. Examples of such publications are those by Carroll and Spiegelman [16], Fuller [1], Miller [17], and Buonaccorsi [3], some of which are discussed further in **Paper II**.

The presence of measurement error can severely influence standard residual analysis tools. The two quantities necessary for visual inspection, x and r, the residual of the regression of y on x, cannot be directly observed and have to be estimated. It is well known that the square sum of the naive residuals is a biased estimator of model error variance. Moreover, most correction methods focus only on correcting the bias in the estimated coefficients and consequently, the residuals of a corrected regression are not necessarily suitable for residual analysis, such as the residuals of the regression calibration method.

Furthermore, heteroscedasticity tests can be affected by the presence of measurement error. For example, well-known heteroscedasticity tests such as the White test [18], the test proposed by Glejser [19], and the Breush-Pagan test [20] rely on asymptotical properties guaranteed by the normality of the model error. The presence of the measurement error can heavily affect the distribution of the residuals, and consequently the distribution of any of these tests.

In the presence of measurement error, the detection of and testing for heteroscedasticity can face an additional problem: in real applications, it is quite common for the measurement error variance to be non-constant. In the presence of heteroscedastic measurement error, it may be even more challenging to evaluate whether the model error variance is constant or in some way dependent on the covariates.

Chapter 3

The bootstrap approach

The bootstrap approach generally encompasses all procedures that typically through some form of resampling of the data allow inference of the parameters of a statistical model as an alternative to traditional procedures. More details on what is described in this section can be found in, for example, the books by Efron and Tibshirani [21] and of Davison and Hinkley [22].

The bootstrap was first proposed by Efron in 1979 [23] as an automatic, computer-based procedure for estimating the standard error of an estimator. It does not require theoretical calculation and is available even if the problem cannot be solved analytically. This original procedure has been successively renamed as simple bootstrap.

3.1 Simple bootstrap

Suppose, in a common data analytical framework, we observe a random sample $\mathbf{x} = (x_1, x_2, ..., x_n)$ from an unknown probability distribution F and our parameter of interest is θ, linked to the distribution by a function t; $\theta = t(F)$. An estimator of θ is available as a function of the sample $\hat{\theta} = s(\mathbf{x})$; this function $s(.)$ may be the plug in estimate $t(\hat{F})$ but it is not necessary. Suppose, too, to be interested in $se_F(\hat{\theta})$ the standard deviation of $\hat{\theta}$ and that, for some reason, it is not available analytically.

The first step of a bootstrap procedure is to identify the bootstrap sample, a random sample of size n drawn from $\hat{F}$, where $\hat{F}$ is the empirical distribution of the data, with probability $1/n$ on each observed values x_i with $i = 1, ..., n$. We will indicate the bootstrap sample with $\mathbf{x}^* = (x_1^*, x_2^*, ..., x_n^*)$, where $\mathbf{x}^*$ is the

resampled version of $\mathbf{x}$. In other words, the bootstrap sample is a random sample with replacement from the population of n individuals that constitutes the original sample.

For each bootstrap dataset $\mathbf{x}^*$, it is possible to have a bootstrap replication of $\hat{\theta}$

$$\hat{\theta}^* = s(\mathbf{x}^*).$$

To obtain this quantity, $s(.)$ has been applied to $\mathbf{x}^*$, in the same way $s(.)$ has been applied to $\mathbf{x}$ to obtain $\hat{\theta}$.

The ideal bootstrap estimate of $se_F(\hat{\theta})$ is the standard error of θ for a random dataset of size n randomly sampled from $\hat{F}$, that we will call $se_{\hat{F}}(\hat{\theta}^*)$. The simple bootstrap approach can be seen as a non-parametric approach because it is based on $\hat{F}$, a non-parametric estimation of the population F.

Unfortunately, this ideal bootstrap estimate is generally impossible to obtain analytically. Rather, it is available only for linear regression and other simple settings in which the standard errors usually have a closed form. However, the following bootstrap algorithm allows a good numerical approximation of $se_{\hat{F}}(\hat{\theta}^*)$:

1. Extract B independent bootstrap samples of length n with replacement from the original sample $\mathbf{x}$
$$\mathbf{x}^{*1}, \mathbf{x}^{*2}, ..., \mathbf{x}^{*B}.$$

2. For each bootstrap sample, compute the estimate of the parameter of interest
$$\hat{\theta}^*(b) = s(\mathbf{x}^{*b}) \quad \text{for } b = 1, ..., B.$$

3. Estimate $se_F(\hat{\theta})$ with the standard deviation of the B bootstrap replications
$$\hat{se}_B = \left(\sum_{b=1}^{B} \frac{(\hat{\theta}^*(b) - \bar{\theta}_B)^2}{B-1} \right)^{\frac{1}{2}}$$
where $\bar{\theta}_B = \sum_{b=1}^{B} \hat{\theta}^*(b)/B$ is the bootstrap average of $\hat{\theta}$.

The final result $\hat{se}_B$ will be the bootstrap estimate of the standard error, where B is the number of bootstrap samples used. If an infinite amount of bootstrap sample are drawn, that quantity will converge to the ideal bootstrap estimate

$$\lim_{B \to \infty} \hat{se}_B = se_{\hat{F}}(\hat{\theta}^*).$$

Thus, the choice of B should be as large as possible. The literature contains several rules of thumb and suggestions about the ideal size of B, but given that bootstrap techniques are often associated with statistical models that are computationally expensive, B is often chosen, such that the total computational time of the procedure remains feasible.

It is worth emphasizing that even the ideal bootstrap estimate $se_{\hat{F}}(\hat{\theta}^*)$ (and its equivalent $\hat{se}_{\infty}$) can still have a considerable variability when estimating $se_F(\hat{\theta}^*)$ and this is due to the variability that $\hat{F}$ has in estimating F.

Additionally, the simple bootstrap procedure can be used to obtain a confidence interval estimation for θ in different ways. For example, the percentiles estimated from the distribution of $\hat{\theta}^*(b)$ are a simple estimator of the confidence intervals that offers the advantage of not assuming any distribution for $\hat{\theta}$. However, this procedure is not always the best and other possibilities are available (see, for example, Section 2.4 of [22] or Chapter 12 of [21]). Furthermore, the simple bootstrap can also be easily extended to the case when $\boldsymbol{\theta}$ is a vector of parameters and there is a need to estimate their covariance matrix.

3.2 Model-based bootstrap

The simple bootstrap has proven a revolutionary and powerful tool, but not without some limitations. Formally, the simple bootstrap is justified only if the units of observation are i.i.d. This means that the data have to be a random sample of units from a given population and have to be sampled with the same method and 'with equal sampling effort'. This concept is discussed further in the next section (3.3).

The model-based bootstrap does not need the above-described assumption. Rather, the procedure will explicitly mimic the various steps that lead to the data, assuming some parametric structure and making distributional assumptions about the random quantities. The model-based bootstrap of the standard error can be defined as $se_{\hat{F}_{mb}}(\hat{\theta}^*)$, where $\hat{F}_{mb}$ is an estimate of F. Generally, this estimate is obtained from a parametric model of the data.

The substantial difference between the simple bootstrap and the model-based bootstrap is in the sampling step. Instead of resampling with replacement from the data, the bootstrap sample $(x_1^*, x_2^*, ..., x_n^*)$ will be generated directly from $\hat{F}_{mb}$. Once B bootstrap samples are generated, one can proceed exactly as in steps 2 and 3 of the bootstrap algorithm presented in the preceding section: the statistics

of interest $\hat{\theta}^*(b)$ are calculated for each sample and then $\hat{se}_B$ is calculated as the standard deviation of the B bootstrap replications.

As illustrative example, we can consider the multiple linear regression $y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i$, with $i = 1, ..., n$, $\boldsymbol{\beta}$ the p-dimensional vector of parameters, $\mathbf{x}_i$ the p-dimensional vector of the covariates (assumed as fixed) and the model error $\epsilon_i \sim N(0, \sigma^2_\epsilon)$. Given the estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2_\epsilon$, the first step of the model-based bootstrap procedure will produce the b-th replication $\hat{\boldsymbol{\beta}}^*(b)$ as the regression of $\mathbf{y}^{*b}$ on $\mathbf{x}$. The bootstrapped response variable y_i^{*b} has been generated as $y_i^{*b} = \mathbf{x}_i\hat{\boldsymbol{\beta}} + \epsilon_i^{*b}$, where $\epsilon_i^{*b} \sim N(0, \hat{\sigma}^2_\epsilon)$.

The key difference between the model-based bootstrap and the simple bootstrap is that the source of 'simulated' variability comes from $\boldsymbol{\epsilon}^{*b}$. This quantity can be generated with different approaches; in our example, it is generated from a normal distribution, but alternatively it can be generated without distributional assumption or non-parametrically, for example by resampling the residuals.

Furthermore, the model-based bootstrap has the potential of providing a bootstrap estimate of bias. Since the resampling is explicitly done from a model that uses $\hat{\theta}$ as if it represents the true coefficients, the bootstrap average of $\hat{\theta}^*$ will be biased in a similar way of how $\hat{\theta}$ is biased with respect to θ. The bootstrap estimate of bias is defined as follows

$$Bias(boot) = \frac{\sum_{b_1}^{B} \hat{\theta}^*(b)}{B} - \hat{\theta}$$

where $\hat{\theta}$ is the original estimate of the parameter of interest. Bias estimates for other functions of the parameters can be computed in a similar fashion.

3.3 The importance of bootstrap for inference under measurement error

When measurement error is present, most of the correction methods, despite allowing for corrected estimates, do not allow for analytical estimation of the standard errors of the estimates. Consequently, when correcting for measurement error, it is extremely rare to perform inference on the parameters analytically and alternatives are necessary. One exception is the regression calibration, which has asymptotic solutions for the standard errors in linear regression (Section 2.5.3 of [1]) and in generalized linear regression case (appendix B.3.1 of [2]). However, this solution is difficult to derive and implement and, depending on the sample

size and other factors, it may not be more convenient than the bootstrap because of its computational issues.

The bootstrap is a widely used technique for performing inference when measurement error is present. As outlined in the introduction to Chapter 3.1, two bootstrap approaches are possible. Both approaches have advantages and disadvantages, but in general the model-based bootstrap is preferable. On the one hand, the simple bootstrap allows for easy resampling, without the need for assumptions about a measurement error structure (and correlation) or model error distribution. On the other hand, the model-based bootstrap, in addition to estimating variance, can estimate bias but it requires a precise model specification. A general strategy is suggested by Buonaccorsi (section 6.16 [3]): first, assess the presence of bias with model-based bootstrap and, if the bias is negligible *and* if the sample units can be assumed i.i.d., then use the simple bootstrap for further inference.

When replicated measurements are present, the simple bootstrap assumption has a wider implication. In such cases, 'equal sampling effort' implies an equal number of replicates for each observation. An unequal number of replicates can be considered only if the number of replicates is random in such a way that it does not depend on any other variable or parameter of the model.

In contrast to the simple bootstrap, which can be easily implemented, the model-based bootstrap requires more attention. Under the presence of the measurement error, the model-based bootstrap will explicitly mimic the various steps that lead to the data. In the bootstrap process, data will be generated by the response from the regression model as well as the replicated values of the mismeasured predictors. This raises two questions: (1) What x to use to generate the response with the regression model, and (2) How to generate the measurement error? These are two of the main topics examined in **Paper I**.

Chapter 4

High-dimensional regression

In recent decades, there has been a revolution in data analysis. In almost all areas of science, business, industry, and entertainment, a huge amount of data is sourced and collected. An example that fits well in this thesis is genetics and the analysis of epigenetics data: in such analyses, it is necessary to find the relationship between *thousands* of variables and a certain disease or treatment. Generally, data are available just on a few hundreds of individuals. The term high-dimensional statistics refers to statistical analysis in this specific context, where the number of variables (and of unknown parameters) p largely exceeds the sample size n, that is $p >> n$.

4.1 Background

In this section, I focus on the linear model. As described earlier, a linear regression model assumes that

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i, \qquad i = 1, ..., n$$

where β_0 and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ are the unknown parameters and ϵ_i is an error term. An estimate of the parameters is provided from the least squares method by the minimization of the following function

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2.$$

As long as $p < n$ the solution to this problem will be unique. Typically, all estimates provided from the least squares method are different from zero, making the interpretation of the model difficult if p is large. Furthermore, when $p > n$, there is an infinite set of solutions that makes the objective function equal to zero. Thus, these solutions will overfit the data.

A possible solution to the problem would be to carry out variable selection and estimation in two steps. However, even if an appropriate selection method is used, this will induce a high variability that is difficult to account for in the final estimation step.

Ridge regression [24] is a penalized regression method that allows estimation in high-dimensional data by shrinking the regression coefficients through imposing a quadratic penalization on their size. A complexity parameter λ is used to control the amount of shrinkage. The Ridge regression will provide a unique solution (for a given λ) with p non-zero estimates. Unfortunately, given that all the p variables will have an estimated effect, the estimated model will still be difficult to interpret.

However, other penalized regressions allow for variable selection and model estimation in a single step, such as the Lasso [6] and the Dantzig selector [25]. To make the process possible, it is essential to assume that *the world is not as complex as it might be* [26]; in other words, it is assumed that the model one wants to estimate is much less complicated than what we observe in the real world. For example, it is hoped that just a few of the ca.20,000 human genes are actually involved as a cause or risk factor for a given disease. This simplifying assumption is called sparsity: the vector of parameters $\boldsymbol{\beta}$, given n observations, can be estimated reasonably well only if some sort of sparsity is present in $\boldsymbol{\beta}$. Assume the number of true non-zero coefficients is s. As a general guideline, it can be said that high-dimensional statistical inference can lead to an acceptable level of accuracy if

$$log(p) \times s \ll n.$$

4.2 The Lasso

The Least Absolute Shrinkage and Selection Operator (Lasso), was originally proposed by Tibshirani [6]. This revolutionary method owes its popularity to its ability to perform variable selection and model estimation simultaneously. The provided *regularized* estimation vector will be sparse, according to a penalization

parameter.

Among the reasons why the Lasso became very popular for statistical applications on high-dimensional data are its statistical accuracy for prediction and its high computational feasibility. Furthermore, the Lasso is a penalized *likelihood* approach – a quite general method that can be used in a broad variety of models such as generalized linear models or proportional hazard models [27].

For simplicity, we will assume that the intercept of the model β_0 is equal to zero and the focus is on linear regression. Then, the Lasso can be defined as the solution to the optimization problem

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \text{ subject to } \|\boldsymbol{\beta}\|_1 \leq t$$

where $\|\cdot\|_1$ denotes the ℓ_1 norm and $\|\cdot\|_2$ denotes the ℓ_2 norm. Alternatively, the most common representation of Lasso minimization is the Lagrangian form

$$\hat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1\} \tag{4.1}$$

where λ is a given regularization parameter. With this representation, it is easier to see how λ acts as a trade-off between the ordinary least squares estimator and the penalization $\|\boldsymbol{\beta}\|_1$ on the parameters. For $\lambda = 0$ the Lasso will reduce to the ordinary least squares, while for $\lambda > \lambda_{max}$ the Lasso will apply maximum shrinkage, where λ_{max} is the value of λ after which the solution does not change. The choice of λ is further discussed in Section 4.2.2. The two formulations are equivalent; for each value of t there is a value of λ that leads to the same solution, as long as $\|\boldsymbol{\beta}\|_1 < t$.

Other penalizations on the parameters are possible, but the ℓ_1 norm has some interesting properties. If a generic ℓ_q norm is used, the value of q will lead to very different results. If $q > 1$, the solutions will lose the sparsity property, for example $q = 2$ leads to the Ridge regression. If $q < 0$ the solutions will be sparse, but the objective function will no longer be convex, the minimization will be very challenging computationally. The value $q = 1$ is the only case that leads to a sparse solution for a convex problem. The convexity will greatly simplify the computation and the sparse solution will make the results more readily interpretable.

However, some alternative penalizations still allow for sparsity. One example is the Dantzig selector, originally proposed by Candes et al. [25]. This penalized regression resembles the Lasso but instead of controlling the squared error loss, it controls the correlation of residuals with $\mathbf{X}$. However, the Lasso and Dantzig

selector are strictly connected, and under certain conditions they lead to the same solution.

4.2.1 Beyond the Lasso

The success and simplicity of the Lasso have inspired a plethora of variants, extensions, and alternatives. Many of them have focused on compensating for certain limitations of the original technique or adapting the method to particular application contexts. For example, Elastic Net regularization [28] combines Ridge and Lasso penalization, allowing the method to select at the same time variables that are strongly correlated, thereby improving the prediction accuracy. The group Lasso [29] allows for the treatment of pre-defined groups of variables as a single unit through the selection process, thus for example easing the inclusion of categorical variables as Lasso predictors.

Some of the Lasso extensions, such as the Smoothly Clipped Absolute Deviation penalty (SCAD) [30] have focused on improving a particular Lasso property. SCAD is a penalized regression model that, under certain conditions, can asymptotically correct for the bias present in Lasso estimates. Another example is the adaptive Lasso [31], a reweighted version of the Lasso that has been shown a valid alternative in scenarios when the Lasso variable selection is inconsistent.

4.2.2 Cross-validation

In this section, I discuss the choice of the tuning parameter λ for the Lasso, but the discussion is valid for any penalized regression method. The tuning parameter λ controls the complexity of the model; for smaller values of λ more parameters can take values different from zero and the model can adapt more closely to the data. The extreme case $\lambda = 0$ will free all the parameters from any constriction and will make the Lasso correspond to the least squares method and, consequently, unfeasible for $p > n$. Larger values of λ will lead to a sparser solution that is easier to interpret but it will not have such a close fit to the data.

The choice of λ is challenging. If the choice of if λ is too big, it can lead to failure to capture the most important variables in the data. Alternatively, if the choice of λ is too small, it can lead to overfitting. When overfitting occurs, the model will be adapted to the noise present in the data along with the main signal. In both cases, the prediction error of the selected model will be inflated.

Ideally, λ should be chosen as in intermediate value that balances the two problems. It is common practice to choose the λ that gives the most accurate model for the prediction of independent test data from the same population. Cross-validation is a procedure that, through artificial training and tests sets, can select the λ with the best prediction performances. More specifically, the full data are divided into K groups (typically 5 or 10); one group is designated as test data and the remaining $K-1$ are used as training data. The Lasso is then applied to the training data for a range of different λ values; for each of these fitted models, the response is predicted in the test set and for each value of λ a mean squared prediction error is obtained. This procedure is repeated K times, using each of the K groups as test data. A cross-validation curve can be obtained by averaging out the K estimates of the prediction error for each value of λ. The $\hat{\lambda}$ selected by the procedure will be the λ for which the cross-validation curve will take its minimum value.

Generally, the different values of λ are chosen from a range between zero and a value corresponding to the maximum shrinkage. For the Lasso, this value is available analytically (section 2.2.2 of [26]). Furthermore, the LARS algorithm [32] allows for simultaneous estimation of the Lasso for each possible value of λ, providing all possible Lasso solutions at the same time, at the cost of a negligible increase in computation time. This is possible because even if λ can take continuous values, the shrinkage process of the Lasso will still be discrete (i.e. all of the λ in a certain interval will lead to the same solution).

It is worth mentioning that the above-defined cross-validation procedure is a consolidated and widely used procedure for tuning parameter selection. However, the procedure provides a λ that is optimal for predictions but not for variable selection (Section 2.5.1 of [33]). To optimize for variable selection, typically a more strict λ would be needed. Alternatively, such a criterion should involve the likelihood and the number of parameters of the model, for example the Akaike information criterion, which is a difficult task for high dimensional regressions.

4.3 Measurement error and high-dimensional data

Thus far, the high-dimensional regression models have been formulated under the assumption that $\mathbf{X}$ is observable. However, when measurement error is present, only its error-prone version $\mathbf{W}$ is observed. Applying high-dimensional regression methods that do not account for measurement error on error-prone data will

result in misleading inference, as has been demonstrated for the Lasso and the Dantzig Selector [7, 8].

To show the bias in the estimation caused by the measurement error, consider the naive Lasso approach, plugging in $\mathbf{W}$ for $\mathbf{X}$ in the Lasso estimator

$$\hat{\boldsymbol{\beta}}(\lambda_n) = \arg\min_{\boldsymbol{\beta}} \{\|\mathbf{y} - \mathbf{W}\beta\|_2^2 + \lambda_n \|\boldsymbol{\beta}\|_1\}.$$

It is possible to demonstrate that this will yield to a biased loss function

$$E(\|\mathbf{y} - \mathbf{W}\boldsymbol{\beta}\|_2^2 \,|\mathbf{X}, \mathbf{y}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + n\boldsymbol{\beta}'\boldsymbol{\Sigma}_U\boldsymbol{\beta}$$

where $n\boldsymbol{\beta}'\boldsymbol{\Sigma}_U\boldsymbol{\beta}$ is the bias term.

In addition, Sørensen et al. [8] show that standard results for consistency of estimation and covariate selection no longer hold when the covariates are affected by measurement error. They also show, through simulations, that ignoring measurement error can lead to a large number of false positive selections. Furthermore, Rosenbaum and Tsybakov [7] observed the same behaviour for censored and missing data.

4.3.1 The penalized regressions correction methods

Correction for measurement error in penalized linear regression has been studied recently by various authors. The appeal of penalized regressions as correction methods is due to the great advantage offered by penalized regressions. If the measurement error is adequately modelled, it will be possible to correct the variable selection and the model estimation at the same time. Examples include Loh and Wainwright's Non-Convex Lasso [9] and Datta and Zou's CoCoLasso [10] both of which use the covariance matrix of the measurement error in the model. These models correct for measurement error, by including in the model the variance of the measurement error $\boldsymbol{\Sigma}_U$, and yielding estimators with good theoretical properties. However, this quantity is assumed to be known and in practice it is usually not known. The estimation of the covariance matrix of the measurement error requires additional data, as replicated measurement of the covariates, and can be computationally expensive or even unfeasible when the number of variables p increases.

An interesting alternative is the Matrix Uncertainty Selector proposed by Rosenbaum and Tsybakov [7]. The MUS can be seen as an evolution of the Dantzig

selector that can also take into account the measurement error in the model without needing any information about the measurement error variance, but rather by using a supplementary tuning parameter. The technique can account for measurement error without requiring any additional information besides $\mathbf{Y}$ and $\mathbf{W}$, at the cost of sacrificing some statistical properties.

Of the above-cited methods, the MUS has received considerable attention. The intention behind the Improved Matrix Uncertainty Selector [34] and Belloni et al.'s Conic estimator. [35] has been to improve some aspects of the MUS at the cost of a more complex structure or requirement of additional information, such as the measurement error covariance matrix. Furthermore, it is worth mentioning the work of Sørensen et al. [36], who have proposed an extension of the MUS to generalized linear models.

All of the above-discussed methods, despite yielding interesting theoretical properties, have not yet been compared from a practical viewpoint or in a real-data application, and therefore is unclear which of them can offer advantages over the others.

Chapter 5

Aims

The overall topic of this thesis is statistical methods for the correction of measurement error, focusing on the use of bootstrap and high-dimensional regressions. Specifically, the aims have been:

- to develop new model-based bootstrap methods for inference and bias estimation when correcting for measurement error
- to explore methods for model variance and residuals estimations when correcting for measurement error, with the aim of improving residuals analysis and developing a new model-based bootstrap test for heteroscedasticity detection
- to evaluate some penalized regression methods that can account for measurement error in high-dimensional linear regression, in a real application context in which the measurement error variance is not known and has to be estimated.

Chapter 6

Summary of the papers

The current chapter gives a short summary of the papers included in this thesis.

6.1 Paper I

Buonaccorsi, J.P., Romeo, G., and Thoresen, M. (2018). Model-based bootstrapping when correcting for measurement error with application to logistic regression. *Biometrics*, 74(1), 135-144.

The aim of the paper was to explore new methods for bias estimation of corrected estimators through the use of model-based bootstrap. In a regression model, when data are affected by measurement error, a wide range of correction methods is available to obtain corrected estimates, but obtaining standard errors and confidence intervals can still be challenging. Additionally, there is concern about the remaining bias in the corrected estimators.

The bootstrap approach is an option that can address the problems but it has received limited attention in this context to date. Usually, only the simple bootstrap has been employed, which does not allow for estimating bias and is not always formally justified. By contrast, model-based bootstrapping can potentially estimate bias, in addition to being robust to the original sampling and the measurement error structure.

Our contribution has been to investigate the model-based bootstrap estimation of bias in correction methods, focusing on logistic regression with replicate measurements, and on regression calibration as correction method. We developed new methods for model-based bootstrapping when correcting for measurement

error in logistic regression with replicated measurements. We have proposed a fully parametric bootstrap approach, that mimics every step of the data generating process, that makes distributional assumptions on model and measurement errors distributions, along with an alternative model-based bootstrap approach in which the measurement error is simulated with a non-parametric approach.

We have shown that the model-based bootstrap approach, despite not always being perfect, can offer some advantages over the simple bootstrap and other standard methods.

6.2 Paper II

Romeo, G., Buonaccorsi, J.P., and Thoresen, M. (2018). Detection and correction of heteroscedasticity under measurement error with non-constant variance. Submitted to *Statistics in Medicine*.

We studied the residual analysis when correcting for measurement error in linear regression. In particular, we focused on model error variance estimation and model assumptions assessment through residual analysis and suggested and evaluate some new methods. One of our main goals was to find a possible representation of the residuals by using a method that can correct for the presence of measurement error, and that can allow for heteroscedasticity detection, graphically. From our simulations, we concluded that the use of estimated x values based on the Moment Reconstruction (MR) [37] method is the best choice for residual analysis. The MR method was seen to provide an approximately unbiased estimate of the model error variance in all the simulated frameworks. Furthermore, the MR method outperformed all the other methods considered here in the estimation of the residuals and of the residual plot. We also developed a model-based bootstrap test for heteroscedasticity and through a practical application, we showed how estimating the variance function can affect the prediction interval for new units. These new methods were validated through a simulation study and applied to an example with methylation data, motivated by the peculiar heteroscedastic binomial measurement error of those data.

6.3 Paper III

Romeo, G. and Thoresen, M. (2018). Model selection in high-dimensional noisy data: a simulation study. *Journal of Statistical Computation and Simulation*, conditionally accepted.

The main purpose of the paper was to examine penalized regression with measurement error and to compare some of these methods in a practical context, in which the distribution and the variance of the measurement error are unknown and need to be estimated from data. We were particularly interested in to what extent this will influence the behaviour of methods that use such measurement error information compared with those that do not use it.

We focused on (1) the Non-Convex Lasso, because in many respects it is the most natural way to include the measurement error effect in the Lasso regression, (2) the CoCoLasso because it is strictly connected to the NCL, and (3) the Matrix Uncertainty Selector, because the method does not require any information about the measurement error distribution.

Our main contribution has been to compare the three methods through extensive simulation studies. We concluded that their performance is dependent on the different structures of measurement error and on the size of the active set.

Chapter 7

Discussion

In this thesis, I have covered different aspects of the analysis of data affected by measurement error. I have explored the use of bootstrap to perform inference on a measurement error corrected model, discussed the analysis of the residuals of a measurement error corrected regression, and evaluated high-dimensional variable selection methods in presence of measurement error.

In **Paper I** we developed new methodology for implementing model-based bootstrap methods when correcting for measurement error in logistic regression with repeated measurements. To implement the model-based bootstrap methods, we proposed the use of an estimated set of values that match the estimated mean and variance of the true, unobserved covariate. We also suggested two possible ways of generating replicates. Through a simulation study, we found that the model-based bootstrap can improve on the simple bootstrap when estimating bias. However, the bias was often only partially reduced, at the cost of a small increase in variability. Regarding the standard error estimation, the model-based bootstrap was often found adequate, but if the standard error was large, it did not perform as well as the simple bootstrap. However, the model-based bootstrap percentile intervals generally had the best overall coverage. We concluded that the model-based bootstrap offers some definite advantages over the simple bootstrap and other standard methods of inference. However, there are some drawbacks. First, in the model-based bootstrap procedure, the replicates are generated from an estimated model that does not take into account the uncertainty in the coefficient estimation. Furthermore, we did not consider the following alternatives because they were beyond the scope of our study. To improve the estimation of the bias, one possibility is to use the double bootstrap (section 3.9 of [22]). Another possibility is the approach used by Pfeffermann and Correa [38], who modelled the

bias of the target estimator as a function of the corresponding estimator obtained from bootstrap resamples, the original estimator, and the bootstrap estimation of the parameters governing the underlying model.

It is worth discussing the method we suggested for estimating x for the model-based bootstrap procedure. The method, called Moment Matched estimator, has been demonstrated to recover correctly the distribution of x up to the second moment [39], and therefore it is suitable for estimating x for the bootstrap procedure. However, other alternatives might have been taken into account. For example, in Paper II we show that the Moment Reconstruction is by far superior to the Moment Matched estimator in correctly recovering x and estimating the residuals of the corrected regression. However, this has been empirically demonstrated only for linear regression and the MR estimator cannot be immediately extended to generalized linear models; therefore, it is unknown whether its properties will hold. The main issue is that the MR estimator is built around the linear correlation of x and y. In a generalized linear model, considering only the linear correlation of x and y may not lead to a sufficient improvement on the MR above the Moment Matched estimator. One possibility would be to include in the definition of the MR the chosen link function and reconstruct x on the base of its correlation with the linear prediction. However, the Moment Matched has been demonstrated to recover correctly the second moment of x and is therefore adequate for the bootstrap procedure in generalized linear models, despite not being suitable for residual analysis.

In **Paper II**, we study residual analysis when correcting for measurement error, we develop a new model-based bootstrap test for heteroscedasticity and, through a practical application with binomial measurement error, we show how modelling heteroscedasticity can affect prediction intervals. The MR method has been demonstrated as the best choice among the alternatives we considered for estimating the model error variance, if constant, and the residuals of the regression. In the paper, we contribute both to residual analysis and hypothesis testing when measurement error is present and to the treatment of binomial types of measurement error. Although this type of measurement error has been already explored by Buonaccorsi [15], there was a need for a study of how the intrinsic heteroscedasticity of the error can affect the residuals of the model. Compared with previous works [15, 16], we had a broader picture of methods available for the estimation of x and of the residuals, and we investigated how the heteroscedasticity of the binomial measurement error can affect the residual analysis. However, in the context of prediction, further work is needed when correcting for measurement

error. We briefly explored the effect of modelling the heteroscedasticity on the prediction intervals. In the paper, through an example, we show that modelling the heteroscedasticity of the model error may locally improve predictions, at the cost of a slightly higher variability for higher values of the estimated variance function. However, it is still necessary to evaluate systematically the coverage and width of the prediction intervals. It should be evaluated how accounting for the heteroscedasticity improves the prediction intervals from the coverage perspective but also with respect to the width of the prediction intervals. Although it can be quite straightforward to test the coverage, the width is challenging in several respects, due to the varying width of the prediction intervals caused by the model error and the measurement error. Furthermore, non-parametric alternatives for the estimation of the variance function should be explored, as in the practical application we assumed the variance function was quadratic. However, given the unusual pattern in the corrected residuals, a non-parametric estimation of the variance function might have been more appropriate.

In general, a non-parametric estimation of the variance function will be more robust in estimating non-monotonic variance functions and would avoid functional assumptions. However, non-parametric models are generally sensitive to extreme values, which is a common problem if the square of the residuals is considered.

Another important problem related to the nature of x and its measurement error in the binomial case is the presence of observed zeroes and ones. Any of the prediction approaches that we considered would be negatively affected if the value of w in the new unit is equal to zero or one. This problem may be solved if a methodology is adequately extended to predict x from the observed w in the new units. However, this might be challenging, because the methods that were found superior in estimating x, such as the Moment Reconstruction, generally require the corresponding y. We suggest estimating x on the basis of distributional assumptions. A further possibility is to use model-based bootstrapping to obtain the prediction intervals for y. For both cases, further investigation is needed.

Model-based bootstrapping particularly requires further attention, beyond the applications described in Papers I and II. Potentially, the procedure can handle measurement error structures that are much more complicated than the one considered in this thesis. For example, it would be beneficial to understand to what extent the model-based bootstrap is able to model successfully the measurement error correlation with other elements of the model, such as x, the model error variance, or other parameters, particularly when this correlation needs to be estimated.

Paper III is the first systematic simulation study to investigate the performance of methods for measurement error correction in a truly high-dimensional situation. We focus on penalized linear regressions, and Non-Convex Lasso, CoCoLasso, and Matrix Uncertainty Selector are all systematically compared considering different measurement error distributions and correlation structure. Based on our results, none of these methods are globally preferable to the naive estimator. However, they have some advantages in certain situations. When no information about the measurement error variance is available, the MUS can drastically reduce the dimension of the active set if compared with the naive estimator, yet this is true only for a limited set of cases and the consistency of its variable selection is subject to rather strict criteria about the magnitude of the covariate correlation. When the measurement errors are correlated, the NCL was found able to correct naive variable selection and estimation, but only when $\mathbf{\Sigma}_U$ was correctly estimated. If the measurement error covariance matrix is estimated under wrong assumptions (for example, independence of the measurement errors) the NCL performance can become even worse than the performance of the naive method. The simulations reported by Chen and Caramanis [40] showed that the NCL can fall short in variable selection when the estimation of $\mathbf{\Sigma}_U$ deteriorates, supporting our findings. Lastly, we found that the CoCoLasso tended to overselect the number of influential variables, both compared with the naive Lasso and to the true Lasso, particularly for large active sets.

We found that the choice of the tuning parameters might be crucial for the NCL and the MUS. The poor performance of the MUS when the active set is large may be mainly due to the rule of thumb of the 'elbow point' suggested by Rosenbaum and Tsybakov [7] for the choice of its tuning parameter δ, thus an alternative criterion for this selection may be beneficial. We considered some alternative approaches to the elbow point. First, we tried to redefine the elbow point as the point closest to the origin on the curve of the number of selected variables versus δ. Unfortunately, using this approach did not change the performance of the MUS. Second, we considered some criteria related to the likelihood. However, the presence of measurement error made the task difficult and we considered it to be outside our scope.

Regarding the NCL, also the choice of the interval in which its non-convex loss function is minimized may be questionable. The results of our tests suggest that, due to the non-convex nature of the minimization, the selected penalization parameter will often be on the upper boundary of the cross-validation interval, even if this has been corrected for measurement error or chosen deliberately too

big. However, it seems that this problem does not affect the final performance of the NCL.

Regarding CoCoLasso, given the simplicity of its implementation, we attempted to improve it by exploiting the idea behind the adaptive Lasso. The adaptive Lasso has been developed to reduce the typical overselection of the ordinary Lasso, through a re-weighting of the regular Lasso estimator. The main idea was to reduce the overselection that the CoCoLasso does in most cases, by giving smaller weights to variables with smaller 'naive' effects. We reweighted estimates of the CoCoLasso in the same way as estimates of the regular Lasso are reweighted in the adaptive Lasso. Accordingly, we tried different criteria for the first step reweighting based on, for example, the naive ridge regression or the naive Lasso regression. However, none of the criteria demonstrated improvements on the original CoCoLasso in some preliminary simulations and are therefore not reported in the paper.

Another method that requires further investigation is the improved MUS [34]. This method is an extension of the MUS that, contrary to the MUS, include the measurement error covariance matrix in its formulation. The improved MUS has been demonstrated to improve on the regular MUS. Despite losing the interesting feature of the MUS, requiring $\mathbf{\Sigma}_u$ (or some estimate), it may offer some advantages over the NCL and need to be investigated further.

In this paper, we focused on penalized regressions, yet penalized regressions are not the only methods available to correct for measurement error in a high-dimensional context. It is worth mentioning that the idea behind NCL and CoCoLasso correction was to recover the behaviour of the Lasso we would have seen without measurement error. However, the Lasso is known to suffer from overselection, and therefore it may be worth considering other alternatives. For example, the two-stage non-penalized corrected least squares [41] separates model selection and estimation in two different steps, using the corrected least squares for the final estimation. Measurement Error Boosting [42] is an iterative functional gradient descent type algorithm that generates measurement error corrected variable selection paths and, in certain frameworks, it has been found better than the CoCoLasso. However, the comparison made by Brown et al. [42] was done in a context in which the measurement error variance was known. Chen and Caramanis [40] developed a modified version of the orthogonal matching pursuit algorithm, also to deal with measurement error in the covariates in high-dimensional regression. Although the alternatives were not considered because they were beyond the scope of our research, it would be interesting to compare these methods with the

penalized regressions considered in Paper III.

It is worth emphasizing that the study of methylation data – which motivated the research for Paper II – generally includes high-dimensional data, given the nature of methylation. An important step for the field would be to study the behaviour of high-dimensional penalized regression when binomial measurement error is present. Furthermore, the use of a correction method such as the MR, which exploits the estimated correlation of x and y to estimate x, should be considered in such contexts. The MR method can be studied in two directions: it can be extended to a high-dimensional context by considering the whole correlation matrix of y with $\mathbf{X}$, or by only considering the marginal correlations $corr(y, x_i)$ for the estimation of x_i. Another possibility is to exploit the possibility of reweighting the naive Lasso or the CoCoLasso using as set of weights the number of reads used for each methylation estimate.

Exploring the variability of the estimated coefficients of the high-dimensional correction methods was beyond the goals of Paper III. Even when measurement error is not present, incorporating the selection uncertainty into the uncertainty estimation of the parameters of a high-dimensional penalized regression method is a major problem. Analytical solutions for the variance of the coefficients are not available for the studied methods in focus in this thesis, as in many other penalized regression methods. The bootstrap can be extremely useful in this context, as both the simple bootstrap and the model-based bootstrap can offer solutions. The challenge is created by the computational burden of some penalized regressions and in particular their cross-validation procedures to choose the penalization parameters, which need to be systematically repeated in order to obtain bootstrap estimates. In many applications, this may not be feasible.

More generally, all of the methods as well as the studies presented in this thesis, despite being tested in a variety of frameworks, assumed rather simple measurement error structures. Caution and further testing are required for more complicated measurement error schemes, such as when w is a non-linear transformation of x. However, model-based bootstrap flexibility offers some definitive advantages in this context. As long as the measurement error structure can be assumed with a certain degree of precision, model-based bootstrap will still allow for inference and bias estimation for simpler correction methods.

References

[1] W. A. Fuller, *Measurement error models*. Wiley, 1987.

[2] R. J. Carroll, D. Ruppert, C. M. Crainiceanu, and L. A. Stefanski, *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC, 2006.

[3] J. P. Buonaccorsi, *Measurement error: models, methods, and applications*. Chapman and Hall/CRC, 2010.

[4] P. Gustafson, *Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments*. Chapman and Hall/CRC, 2004.

[5] G. Y. Yi, *Statistical Analysis with Measurement Error Or Misclassification*. Springer, 2016.

[6] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 58, no. 1, pp. 267–288, 1996.

[7] M. Rosenbaum and A. B. Tsybakov, "Sparse recovery under matrix uncertainty," *The Annals of Statistics*, vol. 38, no. 5, pp. 2620–2651, 2010.

[8] Ø. Sørensen, A. Frigessi, and M. Thoresen, "Measurement error in lasso: Impact and likelihood bias correction," *Statistica Sinica*, vol. 25, no. 2, pp. 809–829, 2015.

[9] B. P.-L. Loh and M. J. Wainwright, "High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity," *The Annals of Statistics*, vol. 40, no. 3, pp. 1637–1664, 2012.

[10] A. Datta, H. Zou, *et al.*, "Cocolasso for high-dimensional error-in-variables regression," *The Annals of Statistics*, vol. 45, no. 6, pp. 2400–2426, 2017.

[11] L. A. Stefanski and R. J. Carroll, "Covariate measurement error in logistic regression," *The Annals of Statistics*, pp. 1335–1351, 1985.

[12] G. T. Hwang and L. A. Stefanski, "Monotonicity of regression functions in structural measurement error models," *Statistics & Probability Letters*, vol. 20, no. 2, pp. 113–116, 1994.

[13] C.-L. Cheng and J. W. Van Ness, *Statistical regression with measurement error*. Oxford University Press, 1999.

[14] D. Schübeler, "Function and information content of DNA methylation," *Nature*, vol. 517, no. 7534, pp. 321–326, 2015.

[15] J. Buonaccorsi, A. Prochenka, M. Thoresen, and R. Ploski, "Correcting for binomial measurement error in predictors in regression with application to analysis of DNA methylation rates by bisulfite sequencing," *Statistics in medicine*, vol. 35, no. 22, pp. 3987–4007, 2016.

[16] R. Carroll and C. Spiegelman, "Diagnostics for nonlinearity and heteroscedasticity in errors-in-variables regression," *Technometrics*, vol. 34, no. 2, pp. 186–196, 1992.

[17] S. M. Miller, *The limiting behavior of residuals from measurement error regressions*. PhD thesis, Iowa State University, 1986.

[18] H. White, "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica: Journal of the Econometric Society*, vol. 48, no. 4, pp. 817–838, 1980.

[19] H. Glejser, "A new test for heteroskedasticity," *Journal of the American Statistical Association*, vol. 64, no. 325, pp. 316–323, 1969.

[20] T. S. Breusch and A. R. Pagan, "A simple test for heteroscedasticity and random coefficient variation," *Econometrica: Journal of the Econometric Society*, vol. 47, no. 5, pp. 1287–1294, 1979.

[21] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*. Chapman and Hall, 1993.

[22] A. C. Davison and D. V. Hinkley, *Bootstrap methods and their application*, vol. 1. Cambridge university press, 1997.

[23] B. Efron, "Bootstrap methods: Another look at the jackknife," *The Annals of Statistics*, vol. 7, pp. 1–26, 1 1979.

[24] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[25] E. Candes, T. Tao, *et al.*, "The dantzig selector: Statistical estimation when p is much larger than n," *The Annals of Statistics*, vol. 35, no. 6, pp. 2313–2351, 2007.

[26] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

[27] R. Tibshirani, "The lasso method for variable selection in the cox model," *Statistics in medicine*, vol. 16, no. 4, pp. 385–395, 1997.

[28] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[29] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[30] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[31] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American statistical association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[32] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, *et al.*, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[33] P. Bühlmann and S. Van De Geer, *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[34] M. Rosenbaum, A. B. Tsybakov, *et al.*, "Improved matrix uncertainty selector," in *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*, pp. 276–290, Institute of Mathematical Statistics, 2013.

[35] A. Belloni, M. Rosenbaum, and A. B. Tsybakov, "Linear and conic programming estimators in high dimensional errors-in-variables models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 79, no. 3, pp. 939–956, 2017.

[36] Ø. Sørensen, K. H. Hellton, A. Frigessi, and M. Thoresen, "Covariate selection in high-dimensional generalized linear models with measurement error," *Journal of Computational and Graphical Statistics*, no. to appear, 2018.

[37] L. S. Freedman, D. Midthune, R. J. Carroll, and V. Kipnis, "A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression," *Statistics in medicine*, vol. 27, no. 25, pp. 5195–5216, 2008.

[38] D. Pfeffermann and S. Correa, "Empirical bootstrap bias correction and estimation of prediction mean square error in small area estimation," *Biometrika*, vol. 99, no. 2, pp. 457–472, 2012.

[39] L. Thomas, L. Stefanski, and M. Davidian, "A moment-adjusted imputation method for measurement error models," *Biometrics*, vol. 67, no. 4, pp. 1461–1470, 2011.

[40] Y. Chen and C. Caramanis, "Noisy and missing data regression: Distribution-oblivious support recovery," *Proceedings of the 30th International Conference on Machine Learning*, vol. 28, pp. 383–391, 2013.

[41] A. Kaul, H. L. Koul, A. Chawla, and S. N. Lahiri, "Two stage non-penalized corrected least squares for high dimensional linear models with measurement error or missing covariates," *arXiv preprint:1605.03154*, 2016.

[42] B. Brown, T. Weaver, and J. Wolfson, "Meboost: Variable selection in the presence of measurement error," *arXiv preprint:1701.02349*, 2017.

I

Paper I

 DOI: 10.1111/biom.12730

Model-Based Bootstrapping When Correcting for Measurement Error with Application to Logistic Regression

John P. Buonaccorsi,[1,*] **Giovanni Romeo**,[2] **and Magne Thoresen**[2]

[1]Department of Mathematics and Statistics, University of Massachusetts, Amherst, Massachusetts, U.S.A.
[2]Department of Biostatistics, Oslo Centre for Biostatistics and Epidemiology, University of Oslo, Norway
email: johnpb@math.umass.edu

Summary. When fitting regression models, measurement error in any of the predictors typically leads to biased coefficients and incorrect inferences. A plethora of methods have been proposed to correct for this. Obtaining standard errors and confidence intervals using the corrected estimators can be challenging and, in addition, there is concern about remaining bias in the corrected estimators. The bootstrap, which is one option to address these problems, has received limited attention in this context. It has usually been employed by simply resampling observations, which, while suitable in some situations, is not always formally justified. In addition, the simple bootstrap does not allow for estimating bias in non-linear models, including logistic regression. Model-based bootstrapping, which can potentially estimate bias in addition to being robust to the original sampling or whether the measurement error variance is constant or not, has received limited attention. However, it faces challenges that are not present in handling regression models with no measurement error. This article develops new methods for model-based bootstrapping when correcting for measurement error in logistic regression with replicate measures. The methodology is illustrated using two examples, and a series of simulations are carried out to assess and compare the simple and model-based bootstrap methods, as well as other standard methods. While not always perfect, the model-based approaches offer some distinct improvements over the other methods.

Key words: Bootstrap; Logistic regression; Measurement error; Replication.

1. Introduction

In fitting regression models, measurement error in predictors may result in biased estimates of coefficients and other incorrect inferences. This problem has received a tremendous amount of attention (see, e.g., Buonaccorsi, 2010; Carroll et al., 2006; Gustafson, 2004; Fuller, 1987 for general treatments) and a plethora of correction techniques have been proposed for both linear and non-linear models. Inferences can be challenging for a variety of reasons. While analytical standard errors are available for some methods these are approximate in nature and usually involve some underlying assumptions. Relatedly, Wald type confidence intervals based on these standard errors rely on approximate normality and unbiasedness of the estimator involved. An additional concern is that the corrected estimators are not unbiased; rather, most are either consistent, or approximately consistent, under suitable conditions. A data driven way of assessing potential bias in either the corrected estimators or naive estimators, which ignore the measurement error, is desirable.

One obvious tool for attacking these problems is the bootstrap, which has received limited attention in the measurement error context. The majority of the applications of the bootstrap in measurement error problems have used simple with replacement resampling of observations. This is used in STATA, one of the few statistical software packages that easily handles measurement error in regression. The simple bootstrap is only formally justified with a random sample and "equal sampling effort" in the measurement error process. Equally important is the fact that with non-linear models the simple bootstrap does not necessarily allow for a bootstrap estimate of bias. This is well known without measurement error where an alternative approach is model-based bootstrapping.

The goal of this article is to examine the use of model-based bootstrap methods when correcting for additive measurement error in regression with replicate measures of the unobserved true values. Measurement error introduces some substantial challenges that are not present when using the model-based bootstrap in regression without measurement error. Our focus is on logistic regression, an important model which shares with other generalized linear models the feature that implementing the model-based bootstrap in the presence of measurement error reduces to two concerns; what to use for true values and how to generate replicates of the error-prone values.

In Section 2, we first describe the regression and measurement error models involved along with laying out the regression calibration method for correcting for measurement error. We concentrate on the most popular correction method, regression calibration, in order to focus the discussion on the bootstrap rather than a comparison of estimators. Section 3 explores the bootstrap methods, first addressing when the simple bootstrap is appropriate and then developing new model-based methods. Two examples are presented in Section 4 followed by simulations in Section 5 and a discussion in Section 6.

2. Models and Methods

2.1. *Models*

The logistic model assumes $P(Y_i = 1|\mathbf{x}_i) = E(Y_i|\mathbf{x}_i) = m(\mathbf{x}_i, \boldsymbol{\beta}) = 1/(1 + e^{-\mathbf{x}_i'\boldsymbol{\beta}})$, where Y_i is a binary outcome and $\mathbf{x}_i$ denotes the fixed predictors on the ith observation. When viewed as random, $\mathbf{X}_i$ will denote the random vector and $\mathbf{x}_i$ its realization. We distinguish between the so-called functional case, where the $\mathbf{x}_i$ are treated as fixed, and the "structural" case, taken here to mean the $\mathbf{X}_i$ are independent and identically distributed (i.i.d.) with mean $\boldsymbol{\mu}_X$ and covariance matrix $\boldsymbol{\Sigma}_X$.

Instead of observing $\mathbf{x}_i$, we observe $\mathbf{W}_i = \mathbf{x}_i + \mathbf{u}_i$, with $E(\mathbf{u}_i|\mathbf{x}_i) = \mathbf{0}$ and $V(\mathbf{u}_i|\mathbf{x}_i) = \boldsymbol{\Sigma}_{ui}$. The fact that $E(\mathbf{u}_i|\mathbf{x}_i) = \mathbf{0}$ means the measurement error is additive or, equivalently, $\mathbf{W}_i$ is unbiased for $\mathbf{x}_i$. In applications, there are many ways that $\mathbf{W}_i$ may arise and the manner in which it arises will determine both the nature of the measurement error covariance matrix, $\boldsymbol{\Sigma}_{ui}$, and how it is estimated. We investigate the frequently studied situation where there are k_i replicate measures, $\{\mathbf{W}_{ij}, j = 1 \text{ to } k_i\}$ on the ith observation, where $\mathbf{W}_{ij} = \mathbf{x}_i + \mathbf{u}_{ij}$ and the $\mathbf{u}_{ij}$ are independent with $E(\mathbf{u}_{ij}) = 0$ and $Cov(\mathbf{u}_{ij}) = \boldsymbol{\Sigma}_{ui(1)}$ (the measurement error covariance matrix for a single replicate on the ith subject). The mean, $\mathbf{W}_i = \sum_{j=1}^{k_i} \mathbf{W}_{ij}/k_i$ has expected value $\mathbf{x}_i$, so $E(\mathbf{u}_i) = \mathbf{0}$ with measurement error covariance matrix $\boldsymbol{\Sigma}_{ui} = \boldsymbol{\Sigma}_{ui(1)}/k_i$, estimated by $\hat{\boldsymbol{\Sigma}}_{ui} = \mathbf{S}_{Wi}/k_i$, where $\mathbf{S}_{Wi} = \sum_{j=1}^{k_i}(\mathbf{W}_{ij} - \mathbf{W}_i)(\mathbf{W}_{ij} - \mathbf{W}_i)'/(k_i - 1)$ is the sample covariance matrix of the replicates on the ith observation.

If some predictors are measured perfectly, the elements in the rows and columns of the covariance matrix associated with perfectly measured predictors are all set equal to 0. Although not specified explicitly in the notation, $\boldsymbol{\Sigma}_{ui(1)}$ is the covariance matrix of a replicate *given* $\mathbf{x}_i$ and, as such, it could depend on $\mathbf{x}_i$ in some (usually unspecified) manner.

2.2. *Correction Techniques*

When correcting for measurement error in non-linear models, a variety of methods have been proposed. As noted in the introduction, we will focus on regression calibration.

The mean $\bar{\mathbf{W}} = \sum_i \mathbf{W}_i/n$ is an unbiased estimate of $\bar{\mathbf{x}} = \sum_i \mathbf{x}_i/n$ in the functional case and of $\boldsymbol{\mu}_X$ in the structural case. We can also construct an unbiased estimator of the covariance matrix of the true predictors,

$$\hat{\boldsymbol{\Sigma}}_X = \mathbf{S}_W - \hat{\boldsymbol{\Sigma}}_u, \tag{1}$$

where $\hat{\boldsymbol{\Sigma}}_u = \sum_{i=1}^n \hat{\boldsymbol{\Sigma}}_{ui}/n$ is the average estimated measurement error covariance matrix, and $\mathbf{S}_W$ is the sample covariance matrix of the $\mathbf{W}_i$'s. Like the mean vector this can be interpreted two ways. In the functional case, $E(\hat{\boldsymbol{\Sigma}}_X) = \mathbf{S}_x = \sum_i(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'/(n-1)$ while in the structural case $E(\hat{\boldsymbol{\Sigma}}_X) = \boldsymbol{\Sigma}_X$. Note that these results about the expectation of $\hat{\boldsymbol{\Sigma}}_X$ *do not depend on either a constant per-replicate measurement error covariance matrix nor on an equal number of replicates across subjects.*

Regression calibration was originally developed in handling measurement error in generalized linear models, and logistic regression in particular; see Carroll et al. (2006) or Buonaccorsi (2010) for overviews. It is based on the idea that if $\mathbf{X}_i$ is random and the measurement error is non-differential (the model for the error in $\mathbf{x}$ does not depend on Y) then $E(Y_i|\mathbf{W}_i) = E(m(\boldsymbol{\beta}, \mathbf{X}_i)|\mathbf{W}_i)$. In the linear case this means $E(Y_i|\mathbf{W}_i) = \beta_0 + \boldsymbol{\beta}_1' E(\mathbf{X}_i|\mathbf{W}_i)$. In non-linear models the approximation $E(Y|\mathbf{W}_i) \approx m(\boldsymbol{\beta}, E(\mathbf{X}_i|\mathbf{W}_i))$ is used. This argues for correcting for measurement error by fitting the regression model but instead of $\mathbf{W}_i$ using an imputed value, which is an estimate of $E(\mathbf{X}_i|\mathbf{W}_i)$. Formally, regression calibration is based on assuming random $\mathbf{X}$'s.

If $\mathbf{X}_i$ is distributed normal with mean $\boldsymbol{\mu}_X$ and covariance matrix $\boldsymbol{\Sigma}_X$ and the measurement error is normal with mean 0 and covariance matrix $\boldsymbol{\Sigma}_{ui}$, then $E(\mathbf{X}_i|\mathbf{W}_i) = \boldsymbol{\mu}_X + \boldsymbol{\kappa}_i(\mathbf{W}_i - \boldsymbol{\mu}_X)$, where $\boldsymbol{\kappa}_i = \boldsymbol{\Sigma}_X(\boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_{ui})^{-1}$. Without the normality assumption $\boldsymbol{\mu}_X + \boldsymbol{\kappa}_i(\mathbf{W}_i - \boldsymbol{\mu}_X)$ is the best linear predictor of $\mathbf{X}_i$ given $\mathbf{W}_k$. The simplest form of regression calibration regresses Y_i on

$$\hat{\mathbf{x}}_i = \bar{\mathbf{W}} + \hat{\boldsymbol{\kappa}}(\mathbf{w}_i - \bar{\mathbf{W}}) \tag{2}$$

where $\hat{\boldsymbol{\kappa}} = \hat{\boldsymbol{\Sigma}}_X(\hat{\boldsymbol{\Sigma}}_X + \hat{\boldsymbol{\Sigma}}_u)^{-1}$ is the estimated reliability matrix for the ith subject using the estimated average measurement error covariance matrix $\hat{\boldsymbol{\Sigma}}_u$ in place of $\boldsymbol{\Sigma}_{ui}$. The estimators using this $\hat{bx}_i$ will be referred to simply as RC estimators.

Note that (2) could be applied with unequal per-replicate covariance matrices and/or unequal numbers of replicates. With unequal numbers of replicates and random $\mathbf{X}$'s, an alternative is to use weighted estimates of $\boldsymbol{\mu}_X$ and $\boldsymbol{\Sigma}_X$ (e.g., Carroll et al., 2006, p. 71) which leads to additional ways to impute. We limit the discussion here to equal numbers of replicates for space considerations.

An alternative to (2) is to replace $\hat{\kappa}$ with $\hat{\kappa}_i = \hat{\boldsymbol{\Sigma}}_X(\hat{\boldsymbol{\Sigma}}_X + \mathbf{S}_{Wi}/k)^{-1}$, assuming k replicates, which utilizes a subject specific per-replicate measurement error covariance matrix. However, we have found through examples and simulations in various settings (see, e.g., Buonaccorsi, 2010, 2016) that the resulting RC estimators often behave rather poorly.

Analytical approaches are available for obtaining an estimate of the covariance of the RC estimates, and associated standard errors, as outlined in equations (B.9) and (B.10) in Appendix B.3 of Carroll et al (2006). These are implemented in the rcal command in STATA with the default being the "information-type" asymptotic covariance with an option (one we prefer) that uses the robust sandwich type estimated covariance.

3. Bootstrapping

The reasons to consider bootstrapping for inferences are well known (see, e.g., Efron and Tibshirani, 1993 or Davison and Hinkley, 1997 among others) including the potential for estimating the bias of a proposed estimator.

There are two bootstrapping strategies that can be employed in regression contexts. The simple bootstrap refers to simply resampling observations, while the model-based bootstrap explicitly incorporates the regression model. Both generate B (large) bootstrap samples with the estimates obtained from the bth sample denoted by $\hat{\boldsymbol{\beta}}_b$. The bootstrap estimate of the standard error of the jth coefficient is the standard deviation of the B values $\hat{\beta}_{1j}, \ldots, \hat{\beta}_{Bj}$, sometimes

obtained with some amount of trimming to reduce the effect of outliers. For example, STATA's rcal procedure trims 1% from each end by default.

The simplest bootstrap confidence interval is the percentile method which uses the $(\alpha/2)$th and $(1-\alpha/2)$th percentiles of the B bootstrap values. We will also consider bias corrected (BC) intervals (e.g., Efron and Tibshirani, 1993, p. 185), which are targeted to correcting for median bias. There are also bias corrected accelerated intervals (e.g., Davison and Hinkley, 1997, p. 204 or Efron and Tibshirani, 1993, p. 184) but we do not pursue these further here both because of the computational demands and since determination of the acceleration constant in a complex situation like this is worthy of separate investigation.

3.1. *The Simple Bootstrap*

In each bootstrap sample, one resamples n times with replacement from the original data. An observation carries with it $\mathbf{D}_i = (Y_i, \mathbf{W}_i, \mathbf{S}_{Wi}, k_i)$ and the bth bootstrap consists of $(Y_{bi}, \mathbf{W}_{bi}, \mathbf{S}_{bWi}, k_{bi})$, for $i = 1$ to n, chosen with replacement from the original $\mathbf{D}_1, \ldots, \mathbf{D}_n$.

Formally, the simple bootstrap is only justified if the $\mathbf{D}_1, \ldots, \mathbf{D}_n$ are i.i.d. This means that we have a random sample of units from some population, with the same method and "equal sampling effort" for each observations to obtain the mismeasured value $\mathbf{W}_i$ and its estimated measurement error covariance matrix. With replication, equal sampling effort means either an equal number of replicates for all observations or, if the number of replicates are unequal, they are random in a way that does not depend on the position in the sample, i.

3.2. *Model-Based Bootstrapping*

Here, the bootstrap is implemented by explicitly mimicking the various steps leading to the data; generating responses from the regression model as well as generating replicate values of the error-prone predictors. To implement this we need a set of estimated true values, $\hat{\mathbf{x}}_{T1}, \ldots \hat{\mathbf{x}}_{Tn}$ which we discuss further after the overall strategy is outlined.

For the bth bootstrap sample and for $i = 1$ to n first generate

$$Y_{bi} \sim g(\hat{\mathbf{x}}_{Ti}, \hat{\boldsymbol{\beta}}), \tag{3}$$

where g is a specified distribution corresponding to a generalized linear model with mean $m(\hat{\mathbf{x}}_{Ti}, \hat{\boldsymbol{\beta}})$. The distribution is just the Bernoulli distribution for logistic regression with $P(Y_{bi} = 1) = m(\hat{\mathbf{x}}_{Ti}, \hat{\boldsymbol{\beta}})$. If there were no measurement error then $\hat{\mathbf{x}}_{Ti} = \mathbf{x}_i$ and the model-based bootstrap would stop here. Other generalized linear models can be handled in similar fashion. However, models of the form $Y_i = m(\mathbf{x}_i, \boldsymbol{\beta}) + \epsilon_i$ where the distribution of the ϵ_i is not fully specified by knowing $\mathbf{x}_i$ and $\boldsymbol{\beta}$ (as happens for GLIMs) require explicitly generating an error term; see Buonaccorsi (2010, p. 217) for further discussion.

With replication, the next step generates

$$\mathbf{W}_{bij} = \hat{\mathbf{x}}_{Ti} + \mathbf{U}_{bij} \tag{4}$$

for $j = 1$ to k_i, where the $\mathbf{U}_{bij}$ are i.i.d. with mean 0 and covariance matrix $\hat{\boldsymbol{\Sigma}}_{ui(1)}$, or the true $\boldsymbol{\Sigma}_{ui(1)}$ itself (see the empirical option below). Recall that k_i is the number of replicates for the ith observation and $\hat{\boldsymbol{\Sigma}}_{ui(1)}$ is the estimated per-replicate covariance matrix for the ith observation. From this obtain $\mathbf{W}_{bi} = \sum_j \mathbf{W}_{bij}/k_i$, $\mathbf{S}_{Wbi} = \sum_{j=1}^{k_i}(\mathbf{W}_{bij} - \mathbf{W}_{bi})(\mathbf{W}_{bij} - \mathbf{W}_{bi})'/(k_i - 1)$, and $\hat{\boldsymbol{\Sigma}}_{bui} = \mathbf{S}_{Wbi}/k_i$. There are two issues that need to be addressed.

The first issue is what to use for the $\hat{\mathbf{x}}_{Ti}$'s, the estimated true values. Treating the $\mathbf{x}$'s as fixed, Buonaccorsi (2010) discussed two options, one to simply use $\mathbf{W}_i$ (as done in a number of his examples), the other to use the $\hat{\mathbf{x}}_i$'s employed in regression calibration. However, neither of these turn out to be a very good choice. Ideally we would like each $\hat{\mathbf{x}}_{Ti}$ to equal the corresponding true $\mathbf{x}_i$, which is of course impossible with measurement error. In lieu of that, one strategy is to choose the $\hat{\mathbf{x}}_{Ti}$'s so they at least have a mean and covariance equal to the estimated mean and covariance of the true $\mathbf{x}_i$'s (or $\mathbf{X}_i$'s in the structural case). If we use $\hat{\mathbf{x}}_{Ti} = \mathbf{W}_i$, these values have a covariance matrix $\mathbf{S}_W$ that overestimates $\mathbf{S}_x$ or $\boldsymbol{\Sigma}_X$, while use of the $\hat{\mathbf{x}}_i$'s from (2) leads to underestimation. Our proposal is to use

$$\hat{\mathbf{x}}_{Ti} = \bar{\mathbf{W}} + (\hat{\boldsymbol{\kappa}})^{1/2}(\mathbf{W}_i - \bar{\mathbf{W}}) \tag{5}$$

where $\hat{\boldsymbol{\kappa}}^{1/2} = (\mathbf{S}_W - \hat{\boldsymbol{\Sigma}}_u)^{1/2}\mathbf{S}_W^{-1/2} = \hat{\boldsymbol{\Sigma}}_X^{1/2}\mathbf{S}_W^{-1/2}$. These values have mean $\bar{\mathbf{W}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}_X = \mathbf{S}_W - \hat{\boldsymbol{\Sigma}}_u$. The above is easily modified to accommodate unequal numbers of replicates.

For the univariate setting, the use of the estimated true values in (5) for model-based bootstrapping was proposed by Buonaccorsi et al. (2016). It was also used by Zheng and Frey (2005) for making inferences in a one-way random effects model and by Thomas et al. (2011) when using moment matching to impute in measurement error problems. Hutchison et al. (2003) use a related model-based approach to try and estimate the bias in the naive estimator and then correct for it in a linear mixed model. Linder and Babu (1994) also considered a version of a model-based bootstrap using modified residuals but in the limited situation where there is no error in the equation and the ratio of the measurement error variances is known.

The second major issue is how to generate the error $\mathbf{U}_{bij}$ in (4). We explore two approaches.

- Parametric model based (PMB). Assume the measurement errors are normal and generate replicates $\mathbf{U}_{bij}$ distributed normal with mean $\mathbf{0}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}_{ui(1)} = \mathbf{S}_{Wi}$ (an estimate of the per-replicate measurement covariance for the ith observation).
- Empirical nonparametric (EMP). This approach borrows the idea behind the empirical SIMEX approach proposed by Devanarayan and Stefanski (2002). Consider $\mathbf{U}_{bij} = \sum_{m=1}^{k_i} c_{bijm}\mathbf{W}_{im}$, where $\mathbf{C}_{bij} = (c_{bij1}, \ldots, c_{bijk_i})'$ is a random normalized contrast vector with $\mathbf{C}_{bij}'\mathbf{1} = 0$ and $\mathbf{C}_{bij}'\mathbf{C}_{bij} = 1$. This results in $\mathbf{U}_{bij}$ having mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_{ui(1)}$ (exactly, without knowing what $\boldsymbol{\Sigma}_{ui(1)}$ is!). There is no parametric assumption on the measurement errors here but

if the original measurement errors are normally distributed then $\mathbf{U}_{bij}$ will also be normally distributed.

3.3. *Estimating Bias*

The easiest (and standard) way to estimate bias in the jth estimated coefficient is with $\widehat{Bias}_j = (\sum_{b=1}^{B} \hat{\beta}_{bj}/B) - \hat{\beta}_j$, where $\hat{\beta}_j$ is the original estimate of interest. The mean bias can be sensitive to outliers so an alternative is to estimate median bias via $\widehat{Bias}_{med,j} = MED_j - \hat{\beta}_j$, where MED_j is the median of the B estimates of the jth coefficient, $\hat{\beta}_{1j}, \ldots, \hat{\beta}_{Bj}$. The corresponding bias corrected estimator (e.g., Efron and Tibshirani, 1993, p. 138) is $\hat{\beta}_j - \widehat{Bias}_j = 2\hat{\beta}_j - (\sum_{b=1}^{B} \hat{\beta}_{bj}/B)$ (or similarly $\hat{\beta}_j - \widehat{Bias}_{med,j}$).

Even without measurement error the simple bootstrap does not provide a valid estimate of bias in non-linear models, including logistic regression. The reason is that the estimators are not "plug-in" estimators; see, for example, Efron and Tibshirani (1993). The model-based bootstrap has the potential of providing a bootstrap estimate of bias since we explicitly resample from a model using the $\hat{\boldsymbol{\beta}}$ as if it were the true coefficients.

One might also want to assess the bias of the naive estimator, one reason being to try and form a bias corrected estimator without an explicit correction technique. However, the bias may be a function of the true coefficient, so using the naive estimates in (3) leads to a problem in that a misestimate of bias will be used. A better way to estimate bias is using the RC estimate of $\boldsymbol{\beta}$ in (3) and then comparing the mean of the naive estimates over the bootstrap samples to the RC estimate. However, this defeats the purpose of trying to correct the naive estimate without an explicit correction method.

4. Examples

4.1. *Cholesterol/Heart Disease*

The first example, presented in Section 7.2.2 of Buonaccorsi (2010), uses a subset of 1475 individuals from the Framingham Heart Study in which there were two measures of cholesterol on each individuals, treated as replicates. A logistic model is used for the occurrence of heart disease as a function of true mean cholesterol. The estimated mean cholesterol ranged between 175 and 325 mg/dl and the per-replicate variability varied greatly among individuals with the estimated measurement error standard deviation (per-replicate sd/$\sqrt{2}$) associated with the mean ranging from 4.5 to 86.

We present results only for β_1. Table 1 shows results first using the naive method with standard analytical standard errors (AN) and associated Wald inferences followed by simple and model-based bootstrap analyses. This is followed by similar results for the RC estimate. The correction for measurement error leads to a substantial 30% change in β_1, from the naive estimate of .0102 to the corrected estimate of .0133. (This corresponds, e.g., to a change in the odds ratio of 2.705–3.785 for a 100 mg/dl change in cholesterol.)

The CIs associated with the analytical robust (ROB) and information based (INF) are Wald intervals. The model-based methods, B-PMB and B-EMP, both use the $\hat{x}_{Ti}$'s (5) as true values, with replicate measurement errors generated as described at the end of Section 3.2. Recall that PMB uses normal replicates while the EMP method is nonparametric. The B-MW method is like B-PMB but with the observed W_i as true values, as used by Buonaccorsi (2010, p. 238). As noted in Section 3.2 this underestimates the standard error.

Looking at the results for the RC estimate, there is a remarkable amount of agreement among the results for the simple and two model-based bootstraps using the $\hat{\mathbf{x}}_{Ti}$'s; see the Supplementary Material for a plot of fitted bootstrap distributions. In addition, the analytical standard errors are very close to those from the different bootstrap methods. The amount of agreement is somewhat surprising given the differences in the motivation for the different methods. Notice that all of the methods estimate no bias in the RC estimator using either the mean, trimmed mean, or median of the bootstrap samples.

Using either the mean or median, the simple bootstrap estimates the bias of the naive estimator to be $.0102 - .0102 = .0000$. However, this is not a valid way to estimate bias, as noted earlier. Using the B-PMB and B-EMP methods to estimate the bias of the naive estimate (see the end of Section 3.3) leads to an estimated bias of $.0102 - .0133 = -.0031$, corresponding exactly to the change observed in moving from the naive estimate to the RC estimate.

4.2. *Gypsy Moths and Defoliation*

Ecological examples abound where logistic regression is used to model a binary outcome (absence of a species, success or failure of a nest, etc.) as a function of habitat variables associated with the unit of interest. Frequently, however, those habitat variables are estimated with an average over sampled subareas within the spatial region of interest. To illustrate the behavior of the bootstrap with a small sample size (typical in many ecological studies) with substantial among replicate variability, we examine the relationship between gypsy moth egg mass density (x) and defoliation over $n = 18$ plots, each 60 ha in size where 10, 15, or 20 replicates (subsamples of size .01 ha) are sampled in a plot. These data are courtesy of Sandy Liebhold and discussed in more detail in Buonaccorsi (2010, p. 75). For our purposes, we use only the first 10 replicates ($k_i = 10$) since we are concentrating on the case with an equal number of replicates. W_i is the mean egg density over the 10 subplots/replicates on the ith plot. For illustration, we dichotomize the outcome defoliation by setting it equal to 1 if the average defoliation is over 50 on the plot (we will call this "severe" defoliation) and 0 otherwise, and ignore any potential misclassification. For $Y = 0$, the W_i (mean egg mass density) ranged from 8.3 to 72.0 with a mean of 33.4 and a median of 29.4, with corresponding values of 22.1 to 88.3, 46.1, and 29.2 respectively, for $Y = 1$. The measurement error standard deviation, $S_{Wi}/\sqrt{10}$, ranged from 2.31 to 28.32.

Results for estimating β_1 appear in Table 2, showing the correction for measurement error leading to a dramatic change in the fitted model. For the RC estimates, the estimated standard errors differ considerably between analytical and bootstrap values. There are also some substantial differences among bootstrap methods and within each method trimming makes a big difference due to outliers. The trimmed standard errors are all still substantially larger than the analytical standard errors. The percentile intervals from the two model-based

Table 1

Inferences for β_1 for cholesterol example. SE= standard error; SE(T)= trimmed standard error for bootstrap methods; AN= analytical based SE; ROB= analytical robust SE; INF= analytical information based SE; B= bootstrap; SIM= simple; PMB= parametric normal model based; EMP= empirical; MW= normal model based with W's as true values; CI= confidence interval; CI(BC)= bias corrected CI for bootstrap methods.

				Bootstrap			
Method	Estimate	SE	SE(T)	Mean	Median	CI	CI(BC)
Naive							
AN	.0102	.0029	.	.	.	[.0045, .0159]	.
B-SIM	.0102	.0028	.0026	.0102	.0102	[.0047, .0155]	[.0047, .0156]
B-PMB	.0102	.0029	.0028	.0102	.0102	[.0045, .0159]	[.0109, .0202]
B-EMP	.0102	.0029	.0027	.0102	.0102	[.0046, .0158]	[.0109, .0202]
RC							
ROB	.0133	.0036	.	.	.	[.0063, .0204]	.
INF	.0133	.0038	.	.	.	[.0058, .0209]	.
B-SIM	.0133	.0036	.0034	.0133	.0133	[.0062, .0203]	[.0062, .0204]
B-PMB	.0133	.0039	.0036	.0133	.0133	[.0059, .0208]	[.0060, .0208]
B-EMP	.0133	.0038	.0035	.0133	.0133	[.0060, .0209]	[.0063, .0211]
B-MW	.0133	.0032	.0030	.0133	.0133	[.0070, .0195]	[.0071, .0196]

bootstrap methods are somewhat close, but quite different than the percentile intervals from the simple bootstrap or the Wald intervals using analytical standard errors.

The bootstrap estimates of mean bias in the RC estimates differ wildly among the simple and model-based methods due to outliers. The estimated median biases of the model-based methods (which are more valid than those from the simple bootstrap) are .008 (PMB) and .007 (EMP), pointing toward moderate positive bias.

As with the first example, estimating the bias of the naive estimator based on the simple bootstrap is misleading. The model-based estimates based on the median leads to bias estimates of $.027 - .0327 = -.006$ (PMB) and $.0263 - .0327 = -.007$ (EMP) which are more in line with the correction of .010 in going from the naive (.023) to RC (.033) estimate.

The story here is more complicated than our first example, showing the challenges in dealing with a small sample size with substantial measurement error. We recommend the use of the model-based bootstrap results for these data.

5. Simulations

Here, we provide some initial assessment of the performance of the various bootstrap methods and of the analytical standard errors and associated Wald confidence intervals. There are obviously an endless number of scenarios that could be considered. We will discuss the situation with random X and $k = 2$ replicates in full here and summarize some results from additional settings in Section 5.5.

The X_i are random with mean $\mu_X = 250$ and standard deviation $\sigma_X = 25$ with X either normally distributed or $X = 250 + 25 * Z$ where Z follows a normalized chi-square distribution with 3 degrees of freedom, leading to a skewed distribution. Based on the cholesterol example, we take $\beta_0 = -5$ while $\beta_1 = .0133$ (the corrected estimate from the example) or .03. The sample size is $n = 100$ or $n = 500$ with a per-replicate standard deviation of measurement error of $\sigma_{u(1)} = 17.677$ (overall reliability = .8) or $\sigma_{u(1)} = 34.355$ (reliability = .5). Lastly, the measurement error is either normal with mean 0 and standard error $\sigma_{u(1)}$ or comes from a standardized chi-square distribution with 3 degrees of freedom, rescaled to have standard deviation $\sigma_{u(1)}$. Overall there are 32

Table 2

Inferences for β_1 for egg mass/defoliation example. Labeling as in Table 1.

Method	Est	SE	SE(T)	Bmean	Bmed	CI	CI(BC)
Naive							
AN	.0229	.0210				[−.0183, .0641]	
B-SIM	.0229	.0547	.0228	.0274	.0249	[−.0291, .0783]	[−.0370, .0708]
B-PMB	.0229	.3964	.0313	.0461	.0269	[−.0219, .1276]	[−.0101, .2048]
B-EMP	.0229	.0365	.0298	.0323	.0263	[−.0179, .1169]	[−.0076, .1572]
RC							
ROB	.0327	.0238				[−.0139, .0794]	
INF	.0327	.0281				[−.0224, .0877]	
B-SIM	.0327	.0887	.0356	.0355	.037	[−.0604, .1146]	[−.1004, .0934]
B-PMB	.0327	.4849	.0558	.0703	.0407	[−.0350, .2416]	[−.048, .1632]
B-EMP	.0327	.2124	.0519	.0580	.0400	[−.029, .2053]	[−.0473, .1619]

Table 3

Random X, $k = 2$. "Med= 'True'" median and RBias= relative median bias of RC estimator based on 5000 simulations. Followed by estimated coverage rates of confidence intervals (1000 simulations). T= Wald interval/true values; N= Wald interval/ naive values; W= $\hat{\beta}_1 + 1.96SE_{rob}$, $\hat{\beta}_1$ = RC estimate, SE_{rob} = robust standard error; W-S (Wald STATA)= $\hat{\beta}_1 + 1.96SE_{boot,trim}$, $SE_{boot,trim}$ = trimmed bootstrap standard error from simple bootstrap; SIM, PMB, and EMP= percentile intervals from simple, normal model based, and empirical model based, respectively; BC, bias corrected. N indicates normal distribution, C indicates use of a standardized chi-square distribution.

				Dist											BC		
Sim	n	κ	β_1	X	U	Med	Rbias%	T	N	W	W-S	SIM	PMB	EMP	SIM	PNB	EMP
1	100	.8	.0133	N	N	0.0139	4.5	.956	.938	.930	.945	.929	.946	.947	.937	.941	.944
2	500	.8	.0133	N	N	0.0133	−0.05	.947	.909	.949	.929	.941	.946	.949	.950	.940	.943
3	100	.5	.0133	N	N	0.0133	0.12	.951	.878	.959	.979	.949	.969	.967	.958	.948	.943
4	500	.5	.0133	N	N	0.0133	−0.21	.957	.556	.962	.954	.955	.964	.965	.961	.958	.961
5	100	.8	.0300	N	N	0.0305	1.52	.939	.893	.933	.946	.915	.948	.947	.935	.929	.928
6	500	.8	.0300	N	N	0.0296	−1.40	.955	.729	.941	.926	.945	.949	.956	.943	.940	.939
7	100	.5	.0300	N	N	0.0305	1.59	.960	.544	.967	.979	.936	.985	.985	.951	.949	.948
8	500	.5	.0300	N	N	0.0291	−2.93	.968	.025	.952	.949	.950	.980	.981	.954	.962	.954
9	100	.8	.0133	N	C	0.0139	4.59	.968	.955	.935	.955	.923	.921	.921	.939	.947	.946
10	500	.8	.0133	N	C	0.0138	3.97	.948	.931	.944	.936	.945	.929	.926	.948	.958	.963
11	100	.5	.0133	N	C	0.0154	16.37	.966	.887	.974	.992	.958	.962	.955	.971	.947	.941
12	500	.5	.0133	N	C	0.0150	13.09	.962	.651	.956	.947	.944	.934	.931	.950	.959	.961
13	100	.8	.0300	N	C	0.0260	−13.37	.957	.872	.911	.925	.934	.977	.974	.946	.918	.913
14	500	.8	.0300	N	C	0.0255	−15.00	.966	.603	.882	.868	.902	.888	.883	.906	.898	.897
15	100	.5	.0300	N	C	0.0223	−25.57	.943	.408	.876	.912	.911	.962	.960	.917	.934	.928
16	500	.5	.0300	N	C	0.0215	−28.34	.964	.003	.778	.757	.858	.874	.873	.848	.914	.912
17	100	.8	.0133	C	N	0.0138	3.35	.946	.936	.915	.937	.911	.936	.936	.930	.928	.933
18	500	.8	.0133	C	N	0.0133	0.26	.960	.915	.953	.940	.948	.961	.962	.953	.937	.934
19	100	.5	.0133	C	N	0.0125	−5.70	.958	.882	.966	.987	.944	.984	.982	.951	.957	.954
20	500	.5	.0133	C	N	0.0128	−4.74	.951	.541	.957	.948	.951	.970	.973	.951	.950	.950
21	100	.8	.0300	C	N	0.0314	4.68	.958	.931	.953	.964	.943	.954	.955	.957	.967	.968
22	500	.8	.0300	C	N	0.0304	1.48	.939	.762	.940	.924	.934	.933	.932	.939	.942	.946
23	100	.5	.0300	C	N	0.0326	8.68	.953	.626	.966	.980	.958	.989	.988	.970	.972	.968
24	500	.5	.0300	C	N	0.0318	4.55	.953	.087	.957	.945	.937	.966	.962	.939	.969	.965
25	100	.8	.0133	C	C	0.0138	3.45	.974	.966	.944	.961	.938	.937	.943	.950	.950	.955
26	500	.8	.0133	C	C	0.0139	4.16	.956	.922	.944	.938	.940	.930	.928	.950	.950	.946
27	100	.5	.0133	C	C	0.0142	6.93	.969	.905	.970	.989	.933	.960	.958	.951	.948	.943
28	500	.5	.0133	C	C	0.0145	8.85	.953	.617	.946	.941	.932	.952	.949	.941	.942	.941
29	100	.8	.0300	C	C	0.0270	−10.10	.950	.870	.901	.917	.921	.965	.963	.934	.933	.934
30	500	.8	.0300	C	C	0.0260	−13.27	.951	.665	.893	.872	.927	.929	.931	.925	.917	.919
31	100	.5	.0300	C	C	0.0248	−17.46	.955	.537	.923	.955	.937	.984	.985	.945	.956	.956
32	500	.5	.0300	C	C	0.0234	−22.07	.937	.025	.845	.831	.884	.936	.941	.880	.929	.934

combinations, numbered as indicated in the results in Table 3. For each combination, 1000 data sets are generated. However, for obtaining the "true" standard errors and biases used in Sections 5.2 and 5.3 and the performance of the estimators discussed in Section 5.1, 5000 simulations were used (but the computational demands were excessive to use 5000 simulations for everything). All of the bootstrap methods use $B = 1000$ bootstrap samples.

5.1. *Performance of Estimators*

A comparison of the estimators is not our main objective. Boxplots in the Supplementary Material show the performance of the naive and RC estimators of β_1 along with that of the estimator based on true values, as a benchmark and the first portion of Table 3 shows the median and the relative median bias of the RC estimator. The mean and median bias were often similar, but outliers influenced the mean in a couple of scenarios (see the Supplementary Material) so the median is given here. These results show the well known fact that the naive estimator can be significantly biased downward and that the RC estimator has less bias but more variability. While in many cases the RC estimator has a bias similar to that based on using true x's, if they could be observed, there are scenarios (chosen intentionally in order to assess how well we can estimate bias) where there is a substantial amount of bias in the RC estimator; see in particular simulations 11–16 and 29–32. The four most severe cases (15, 16, 31, and 32) have low reliability (large measurement error), a big effect ($\beta_1 = .03$) and the distribution of the measurement error is skewed.

5.2. *Estimating the Standard Error of the RC Estimator*

With random X and constant measurement error variance, the simple bootstrap is applicable. As noted earlier, there are some problems with outliers, so we use the trimmed standard deviation from 5000 simulated data sets as the "true" standard error. The top of Figure 1 shows the median of the estimated standard error using the robust analytical and three bootstrap methods (also using 1% trimming) versus the "true" standard error over all 32 scenarios. Figure 2 shows further details for simulations 9–16 with additional figures in the Supplementary Material. The robust estimator of the standard error always does rather well and the estimated SE from the simple bootstrap also does fairly well. Here, model-based estimates tend to overestimate the standard error when it is large, with some reasons for this touched on in the discussion. We note that this problem largely disappears with more replicates; see the bottom of Figure 1.

5.3. *Estimating Bias*

The mean bias is sometimes quite erratic due to outliers, so Figure 3 shows the median of the estimated median bias of the RC estimator versus the "true" median bias for the RC estimator (from 5000 simulations). Further details on the bias estimation is shown in the boxplots in the Supplementary Material. The model-based bootstrap methods are preferred over the simple bootstrap for estimating bias since the latter may estimate that there is negligible bias, even when there is some. This problem is even more severe when using the simple bootstrap to estimate the bias of a naive estimator; see the Supplementary Material for details. In those places

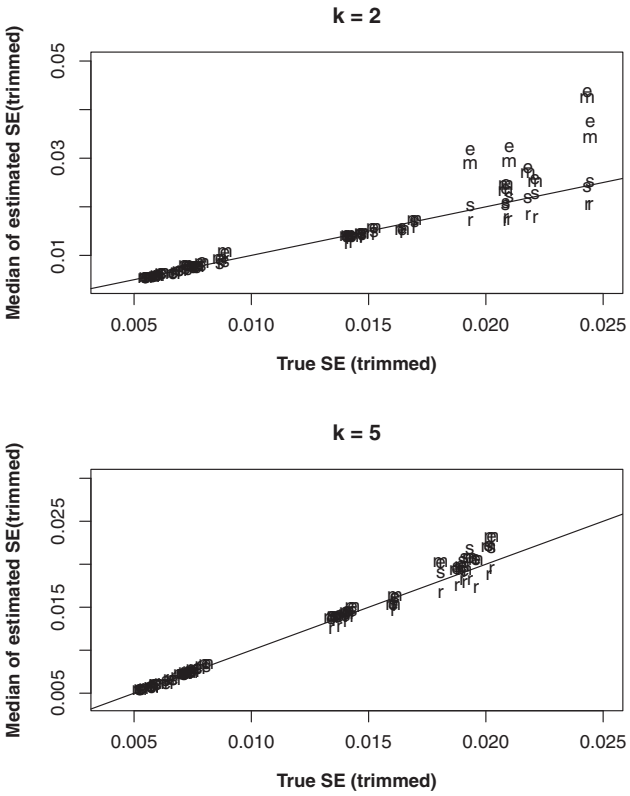

Figure 1. Random X. Median of estimated standard error (bootstrap trimmed) versus "true" SE (trimmed). r, analytical robust standard error; s, simple bootstrap; m, model based with normal measurement errors; e, empirical bootstrap.

where the simple bootstrap is at its worse, when there is substantial negative bias in the RC estimator, the model-based bootstrap methods also encounter some difficulty but they still do better than the simple bootstrap. The corresponding bias corrected estimators based on the model-based bootstrap do result in less biased, but somewhat more variable, estimators, with some extreme values, which are directly related to a corresponding extreme value of the RC estimator. Supporting figures appear in the Supplementary Material.

5.4. *Performance of Confidence Intervals for β_1*

Table 3 shows the estimated coverage rates for a variety of confidence intervals for β_1, with a target rate of .95. The true and naive intervals use standard Wald intervals based on the true and error prone values, respectively, while the others are based on the RC estimator.

The confidence intervals from using the true values always do quite well, as expected and, also as expected, the naive interval can do quite badly, sometimes disastrously so. The Wald interval using the robust standard error does fairly well in most cases, but does encounter some serious problems in simulations 13–16 and 29–32. The Wald-STATA intervals are sometimes better and sometimes worse than the Wald interval. In the eight aforementioned cases it is better in some cases, but in others, offers little, or no, improvement.

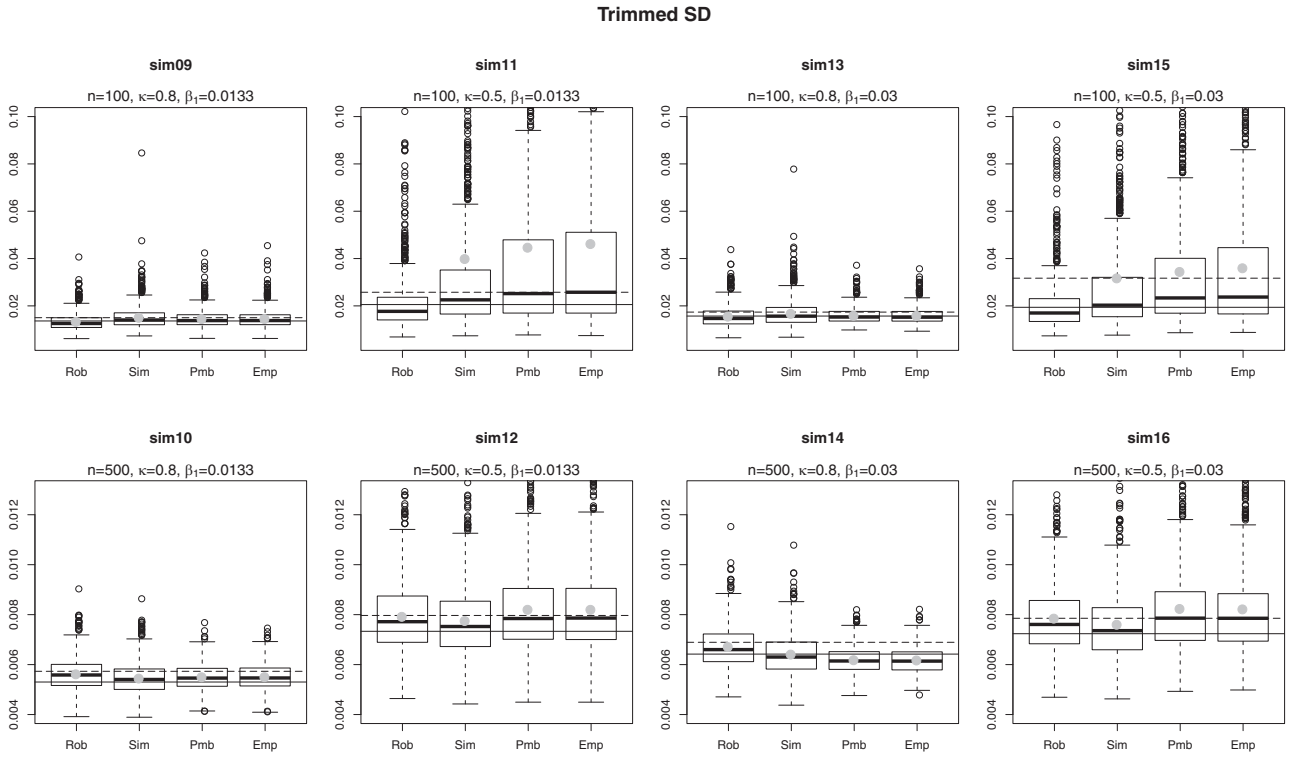

Figure 2. Random $X, k = 2$. Boxplots showing performance of estimated standard errors (trimmed) for simulations 9–16; gray dot, mean; black line, median. Dashed line, simulated SE from 5000 simulations, untrimmed; solid line, simulated SE, trimmed.

Focusing on the percentile intervals, the simple bootstrap does not do quite as well as the model-based methods. There are cases where the coverage rates from the simple bootstrap are too small (simulations 5, 15, 29, and 32) but both the model-based methods do fairly well. All three bootstrap methods fall substantially short of the desired .95 in simulations 14 and 16. The bias corrected intervals did offer some improved coverage rates when the standard percentile methods fell short but did not always improve matters; for example, simulations 14 and 30 among others.

The Supplementary Material provides summary statistics on the length of the various intervals. For $n = 500$ the lengths are fairly similar for the various methods, although the model-based bootstrap intervals are a bit larger when the measurement error distribution is skewed and $\kappa = .5$. When $n = 100$ the Wald and Wald-S intervals have shorter mean and median length than the other procedures, but the Wald intervals sometimes lead to wildly wide intervals in some scenarios. Still with $n = 100$, the model-based intervals are generally longer throughout than those based on the simple bootstrap, which is what leads to better coverage rates. The bias corrected percentile intervals, while similar in median length to their non-bias corrected counterparts in some cases, they are longer in some others and more susceptible to outliers.

5.5. *Additional Simulations*

A number of additional simulations were run with details on the set-up as well as extensive results appearing in the Supplementary Material. The first of these were identical to the preceding simulations except with $k = 5$ replicates and the per-replicate variance rescaled so the reliability stayed at .5 or .8. As noted earlier (see Figure 1), the model-based methods do a better job of estimating the SE than with $k = 2$. The model-based methods still have some trouble estimating bias, although they do somewhat better than when $k = 2$ and they still do better than the simple bootstrap. When the error associated with a replicate is skewed, the error for the mean becomes more normal with $k = 5$ than with $k = 2$ and so the performance of the confidence intervals there are similar to results for corresponding cases with U distributed normal in Table 3. The lengths of the confidence intervals behave similarly as for $k = 2$ with the important exception that the model-based bootstrap intervals are now either similar in length or generally shorter than those from the simple bootstrap. The Wald intervals are not as susceptible to outliers.

The second set of simulations return to $k = 2$, but use a fixed set of true values ($n = 100$ or 500), based on selected cases from the original Framingham data. The per-replicate measurement error variance is either constant or changing over observations. The conclusions for estimating bias or the standard error and the performance of confidence intervals are very similar to those from the random X setting. The performance of the simple bootstrap in estimating the standard error is somewhat surprising given the x values are fixed, but this is mostly likely due to the fact that there are many (100 or 500) different x values; see Chernick (2008, Ch. 4) also

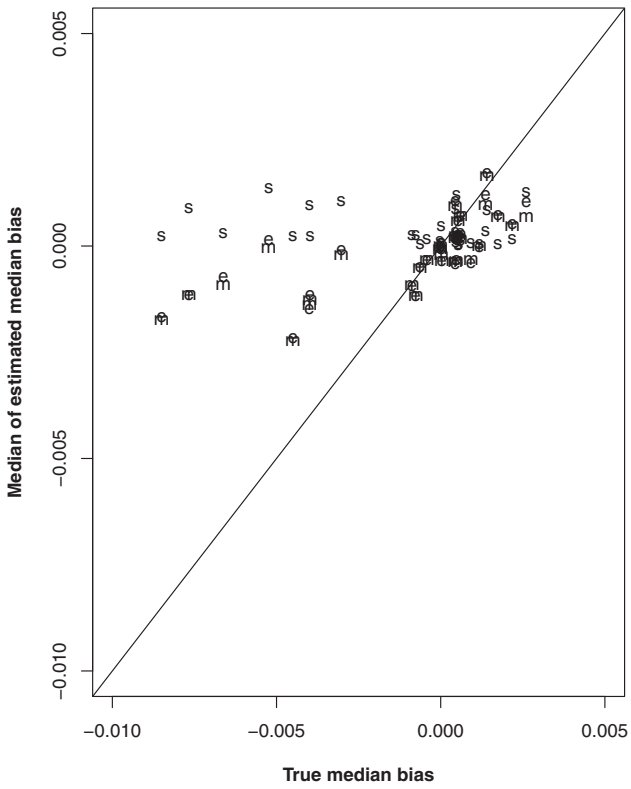

Figure 3. Random $X, k = 2$. Median of estimated median bias versus "true" median bias for RC estimator. s, simple bootstrap; m, model based with normal measurement errors; e, empirical bootstrap.

for some discussion on the robustness of simple bootstrapping in regression problems. However, despite the fact that the model-based method does not always do well estimating the standard errors the model-based percentile confidence intervals routinely outperform the intervals from the simple bootstrap and the standard Wald interval.

6. Discussion

We have developed some new methodology for implementing model-based bootstrap methods when correcting for measurement error in logistic regression (more generally, generalized linear models) with replicates. These are an alternative to the simple bootstrap, which is not always justified and is also not designed to estimate bias in non-linear models. To implement the model-based methods, we propose the use of an estimated set of true values which match the estimated mean and variance of the underlying true values (see equation 5) and suggested two ways to generate replicates, one using a normal model with estimated measurement error variances, the other an empirical method which exploits a clever technique originally introduced for use with SIMEX.

In a series of simulations, we found (i) the model-based bootstrap improved on the simple bootstrap in estimating bias in either the naive or RC estimator. As expected, the simple bootstrap tended to estimate there was little bias, even when bias was present. There were a few settings, however where the model-based methods, while better than the simple bootstrap, still had some difficulty in estimating bias. The corresponding model-based bias corrected estimators reduced some of the bias in the RC estimator, when present, although they were a bit more variable. (ii) While often adequate, the model-based methods did not do as well as the simple bootstrap and analytical methods in estimating a large standard error. However, this did not always translate into poor performance of the associated percentile or bias corrected confidence intervals. (iii) The model-based bootstrap percentile intervals generally had the best overall coverage rates. This improved performance was achieved without giving up much, if anything, in terms of having wider intervals. There were a number of settings where the Wald intervals based on either the analytical standard error or a trimmed standard error from the simple bootstrap, or the percentile interval from the simple bootstrap, fell considerably short in their coverage rates. Still there were some settings where the model-based percentile intervals, as well at their bias corrected counterpart, also fell somewhat short, essentially when there was substantial bias in the RC estimator.

Overall the model-based bootstrap methods offer some definite advantages over the other, more standard, methods of inference and we recommend their use. However, there are still some shortcomings and there are a few potential reasons for that. One is uncertainty in the estimated coefficients that enter into (3). This is a potential problem even in regression problems without measurement error where the bias may be a function of the true coefficients. A second reason is the uncertainty in the estimated true values, which enter into generating both the Y's in (3) and the replicates in (4). Based on some simulations (results not shown) where we used the true x's in place of the $\hat{x}_{Ti}$'s but the estimation of bias did not improve much, this does not appear to be a major factor. A third reason is uncertainty in the replicate measures that go into the EMP method and into estimating the measurement error variances used in the normal-based PMB method. There is an additional issue with the EMP method, which we expected to outperform the PMB method, especially when the errors were non-normal. While it does generate replicates that unconditionally have measurement error covariance matrix $\mathbf{\Sigma}_{ui(1)}$ (even without knowing what it is) and the replicates are conditionally independent, they are unconditionally dependent since they each use the common estimated true value $\hat{\mathbf{x}}_{Ti}$. (It is easy, using double expectations, to show that unconditionally $Cov(\mathbf{W}_{bi1}, \mathbf{W}_{bi2}) = Cov(\hat{\mathbf{x}}_{Ti})$). This has the potential to compromise the performance of the EMP method somewhat.

Our goal was to introduce some relatively basic model-based methods, but there are a number of additional strategies beyond the scope of this article that could be explored to try and improve the methods. For improving estimation of bias, one alternative is the use of the double bootstrap (Davison and Hinkley, 1997, p. 104). Another is to try and model the bias over plausible values of the underlying parameters and then use the resulting model for bias correction (see, e.g., Pfefferman and Correa, 2012). If one was comfortable with the assumption of constant per-replicate variance then a pooled estimate of the per-replicate standard deviation rather than the subject specific values could be used in the PMB method. There are also other nonparametric

model-based bootstraps that could be considered including the use of an estimate of the distribution of the true values with random X's to generate true values. This was used by Carroll et al. (2011) in estimating a shape-constrained nonparametric density and regression.

7. Supplementary Materials

Tables and Figures referenced in Sections 4 and 5 are available with this article at the *Biometrics* website on Wiley Online Library.

Acknowledgments

We are grateful to Sandy Liebhold and Ray Carroll for use of data going into the defoliation and cholesterol examples.

References

Buonaccorsi, J. P. (2010). *Measurement Error: Models, Methods and Applications*. Boca Raton, FL: Chapman and Hall/CRC.

Buonaccorsi, J. P., Prochenka, A., Thoresen, M., and Ploski, R. (2016). Correcting for binomial measurement error in predictors in regression with application to analysis of DNA methylation rates by bisulfite sequencing. *Statistics in Medicine* **35**, 3987–4007.

Carroll, R. J., Stefanski, L. A., Ruppert, D., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models*, Second Edition. Boca Raton, FL: Chapman and Hall/CRC.

Carroll, R. J., Delaigle, A., and Hall, P. (2011). Testing and estimating shape-constrained nonparametric density and regression in the presence of measurement error. *Journal of the American Statistical Association* **106**, 191–202.

Chernick, M. R. (2008). *Bootstrap Methods: A Guide for Practicioners and Researchers*, Second Edition. Hoboken, NJ: John Wiley and Sons.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

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

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York, USA: Chapman & Hall.

Fuller, W. A. (1987). *Measurement Error Models*. New York: John Wiley.

Gustafson, P. (2004). *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. Boca Raton, FL: Chapman and Hall/CRC.

Hutchison, D., Morrison, J., and Felgate, R. (2003). Bootstrapping the effects of measurement errors. *Multilevel Modelling Newsletter* **15**, 2–10.

Linder, E. and Babu, G. J. (1994). Bootstrapping the linear functional relationship with known error variance ratio. *Scandinavian Journal of Statistics* **21**, 21–39.

Pfefferman, D. and Correa, S. (2012). Empirical bootstrap bias correction and estimation of prediction mean square error in small area estimation. *Biometrika* **99**, 457–472.

Thomas, L., Stefanski, L., and Davidian, M. (2011). A moment adjusted imputation method for measurement error models. *Biometrics* **67**, 1461–1470.

Zheng, J. and Frey, C. (2005). Quantitative analysis of variability and uncertainty with known measurement error: Methodology and case study. *Risk Analysis* **25**, 663–675.

Received October 2016. *Revised February* 2017. *Accepted April* 2017.