AN ABSTRACT OF THE DISSERTATION OF

Vicente J. Monleon for the degree of Doctor of Philosophy in Statistics presented on November 22, 2005.
Title: Regression Calibration and Maximum Likelihood Inference for Measurement Error Models

Abstract approved:

Daniel W. Schafer

Regression calibration inference seeks to estimate regression models with measurement error in explanatory variables by replacing the mismeasured variable by its conditional expectation, given a surrogate variable, in an estimation procedure that would have been used if the true variable were available. This study examines the effect of the uncertainty in the estimation of the required conditional expectation on inference about regression parameters, when the true explanatory variable and its surrogate are observed in a calibration dataset and related through a normal linear model. The exact sampling distribution of the regression calibration estimator is derived for normal linear regression when independent calibration data are available. The sampling distribution is skewed and its moments are not defined, but its median is the parameter of interest. It is shown that, when all random variables are normally distributed, the regression calibration estimator is equivalent to maximum likelihood provided a natural estimate of variance is non-negative. A check for this equivalence is useful in practice for judging the suitability of regression calibration. Results about relative efficiency are provided for both external and internal calibration data. In some cases maximum likelihood is substantially more efficient than regression calibration. In general, though, a more important concern when the necessary conditional expectation is uncertain, is that inferences based on approximate normality and estimated standard errors may be misleading. Bootstrap and likelihood-ratio inferences are preferable.

Regression Calibration and Maximum Likelihood Inference for Measurement Error Models

by
Vicente J. Monleon

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented November 22, 2005
Commencement June 2006

Doctor of Philosophy dissertation of Vicente J. Monleon presented on November 22, 2005

APPROVED

Major Professor, representing Statistics

Chair of the Department of Statistics

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Vicente J. Monleon, Author

ACKNOWLEDGEMENTS

Many people have made it possible for me to reach this point. I can not name all of them, but they all have my sincere gratitude and appreciation. I owe much gratitude to my graduate advisor, Dr. Dan Schafer, for his guidance, advice and patience. He was always available to discuss my work, help me see the ‘big picture’, and suggest ways out of the many dead-ends where I found myself. I thank Drs. D. Birkes, D. Edge, A. Gitelman, L. Madsen, F. Ramsey and T. Wood for their important contributions at critical points during my graduate program, and for taking their valuable time to serve as committee members. The staff, faculty and students of the Department of Statistics contributed significantly to my education.

I could not have completed my program without the financial support of the Department of Statistics and, later, the Student Career Experience Program of the USDA Forest Service. D. Azuma and J. Fried, my supervisors at the PNW Research Station, showed patience and understanding, and allowed great flexibility in my work hours and duties.

I wish to thank family and friends for their support during all this time. As always, my parents, Vicente Monleon and Maria Jose Moscardo, have been unconditionally supportive, even though they may never understand why I had to go back to school. My children, Santiago and Carmen, granted me permission to go and ‘work with the computer’ too many weekends. Lastly, I could have never completed this program without the sacrifices and help of my wife, Carolina Hooper. It is difficult to express my appreciation to her but, in my mind, she deserves this as much as I do. May we enjoy the fruits of this effort for many years.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

Figure | Page

LIST OF FIGURES (Continued)

LIST OF APPENDICES

LIST OF APPENDIX TABLES

LIST OF APPENDIX TABLES (Continued)

Regression Calibration and Maximum Likelihood Inference for Measurement Error Models

1. INTRODUCTION

The overall topic of this dissertation is inference for regression models, particularly linear regression models, when one or more explanatory variables are measured with error. The dissertation work largely concerns the properties of a method called "regression calibration," which has emerged as a major tool for these models, and its performance relative to likelihood-based approaches. Regression calibration (RC) has received increasing usage for scientific problems, largely because of its simplicity, transparency and intuitive appeal. The most simple and transparent form, however, ignores both the uncertainty due to estimation of the regression calibration model, and the problems that arise when the regression of interest is not linear. While the latter issue has received considerable attention, the former remains largely ignored and is a focal point of this work.

1.1. The problem of regression with measurement error

Suppose interest is in the regression of a response variable, Y, on an explanatory variable, X, which is observed only through an imprecise measurement or surrogate variable, W. Regression with measurement error refers to the problems that arise from using W instead of X in the model of interest. In general, ignoring the measurement error results in biased estimates of the parameters of the regression of Y on X, so a variety of methods have been proposed to correct or reduce the bias. This problem was noted and studied as early as 1877 by Adcock. Recent reviews include Fuller (1987) for the classical linear model, Carroll, Ruppert and Stefanski (1995) for non-linear models, and Gustafson (2004) from a Bayesian perspective.

Assessing the impact and possible corrections of the measurement error requires an understanding of the measurement process. This involves the formulation of a conceptual model for the relationship between the true variable X and its surrogate W. A primary distinction is between classical and Berkson error models. Classical error structure arises when $W = X + \varepsilon$, and ε is independent of X. This structure is appropriate, for example, for an imprecise measurement device that adds noninformative noise to the true value of X. Berkson error structure arises when $X = W + \varepsilon$ and ε is independent of W (Berkson 1950). This structure was initially proposed for controlled experiments, in which a nominal level of a treatment was prescribed to an experimental unit, but the actual level applied was the nominal level plus some noninformative noise. In the simple linear model, classical error results in an attenuation of the regression slope, while Berkson error allows for unbiased estimation of the regression parameters (see, for example, Madansky 1959, Cochran 1965). In more complex cases, there is not a simple pattern and both structures can result in attenuation, inflation, or can even induce a curvature in the regression of Y on W (Fuller 1987, Reeves et al. 1998, Schafer and Gilbert, *in press*).

For the classical error model, if $Y \mid X$, $W \mid X$, and X are all normal, the parameters of the distribution of $Y \mid X$ are not identifiable without additional assumptions or data (Fuller 1987). Only Y and W are observed and, while the joint distribution of Y and W contains six parameters, the bivariate normal distribution is completely determined by only five parameters. Although for other distributional assumptions the parameters of the simple linear regression model with measurement error are identifiable (Reiersol 1950), additional information is necessary to practically estimate the parameters of interest. This information can take several forms, including replicate measurements on some observations or a calibration study in which the true X is observed. This study focuses on the latter, including external calibration studies, in which observations (x_i , w_i), independent from the primary study are available, and internal calibration studies, in which observations (y_i, x_i, w_i) are available for a subset of cases.

1.2. The regression calibration estimator

One of the most popular methods to approximate measurement error models, regression calibration, uses whatever method of estimation would have been appropriate if X were observed exactly, but with the missing X replaced by $E(X|W)$ (Carroll et al. 1995, Chapter 3). There are different forms of this estimator for different types of information available for arriving at $E(X|W)$, and there are refinements available for particular models. Versions of the RC estimator were proposed by Prentice (1982) for Cox proportional hazards regression models, Armstrong (1985) for generalized linear models, and Rosner, Willet and Spiegelman (1989) for logistic regression.

Due to its transparency and ease of use, the RC estimator has emerged as one of the more important tools for dealing with measurement errors. It has seen considerable use in epidemiology, where the exposure variables associated with a disease are difficult to measure precisely. For example, in a prospective study of the effect of fat on the risk of breast cancer, X was the long-term average intake of fat (Willett et al. 1992). This variable was imprecisely assessed with a semi-quantitative food questionnaire administered to over 90,000 women. Then, a validation study was conducted on 173 participants, who completed four, one-week diet records. The model for $E(X|W)$ was estimated from this subset with values of both X and W, and then used to estimate $E(X|W)$ for the 90,000 individuals in the primary data, for whom only W was available. As typical with this type of study, the correlation between nutrient intakes calculated from the questionnaire and the 'gold standard' was relatively low, ranging between 0.4 and 0.6 (Willett et al. 1988). In radiation epidemiology, to determine the effect of radiation exposure for the atomic bomb survivors or uranium miners, X was the dose of radiation, and $E(X|W)$ was estimated with a combination of physical and biological models, and empirical data (Pierce et al. 1990, 1992; Stram et al. 1999).

Often, RC is used without explicit acknowledgement. For example, in cosmology, estimation of the Hubble constant involves the regression of a galaxy's recession velocity on its distance from Earth (Freedman et al. 2001). Distance is estimated as a function of a galaxy's apparent luminosity, which is measured directly, and its intrinsic luminosity, which is estimated from other variables such as rotational velocity. The models to predict intrinsic luminosity are linear regressions, calibrated with a small sample of nearby galaxies. In this example, X is the true distance; W includes the ancillary variables used to predict intrinsic luminosity; and Z, the apparent luminosity, is an explanatory variable virtually free of measurement error. The missing X is replaced by $E(X|W, Z)$.

Regression calibration emerges naturally when the regression of Y on X is linear, because if $E\left(Y \mid X\right) = \beta_0 + \beta_1 X$, then $E\left(Y \mid W\right) = E\left[E\left(Y \mid X\right) \mid W\right] = \beta_0 + \beta_1 E\left(X \mid W\right)$. If $E(X|W)$ is known, usual regression tools for the regression of Y on $E(X|W)$ may be used to estimate β_0 and β_1, with appropriate attention to weights dictated by $Var(X|W)$. However, if the regression of Y on X is not linear, this form of RC is, in general, an approximation to the model of interest. The conditions under which the approximation is almost exact depend on the particular model. For example, for logistic regression with a linear model relating X and W, RC is approximately unbiased if either $\beta_1^2 Var\left(X \mid W\right)$ is small or $P(Y=1 \mid X)$ is small and $f(X|W)$ is normal (Kuha 1994). When the degree of non-linearity in the regression of Y on X is large, several improvements to the simple RC estimator have been proposed, based on Taylor expansions and the assumption of small measurement error variance (Carroll and Stefanski 1990).

In addition to its simplicity and transparency, RC is attractive because it relies on minimal assumptions on the distribution of the explanatory variables. Because of its emerging popularity, though, we believe it is appropriate to critically examine its shortcomings, and to better understand the situations in which extra care is needed. Of particular interest here is the role of the uncertainty in the estimation of $E(X|W)$ because, typically, it is not $E(X|W)$, but an estimate of it that is used in place of X. Using an estimate of $E(X|W)$ instead of the true value results in an extra component of variability to the estimator of the regression coefficients, so standard errors of estimated regression coefficients may be adjusted to account for this additional uncertainty. However, the effect of the additional variability on the estimator's properties is often ignored. There are additional problems with the simple version of RC unless either the degree of non-linearity in the regression of Y on X or the measurement error variance are small. These problems may be magnified after accounting for the additional uncertainty in the estimation of $E(X|W)$.

1.3. The effect of the uncertainty on the estimation of $E(X|W)$

The typical application of RC estimates $E(X|W)$, replaces X by its estimated expectation given W, and runs a standard analysis. Since $E(X|W)$ is not known exactly, it becomes an imprecise measurement, subject itself to the problem of regression with measurement error. The RC estimator should show an improvement compared to the naïve regression of Y on W, because the estimate of $E(X|W)$ should be closer to the true $E(X|W)$ than W is to X. However, if the estimated expectation is not sufficiently close to the actual value, the resulting estimators may be seriously biased, even in the linear model. Most discussions of RC ignore the effect of the uncertainty on the estimation of $E(X|W)$, either by assuming that $E(X|W)$ is known, or assuming that it is consistent and basing inference on the asymptotic distribution of the RC estimator. The latter may be problematic, because the sample size used to estimate $E(X|W)$ can be very small, or at least much smaller than that of the main study.

A conceptual model for incorporating the uncertainty in the estimation of $E(X|W)$, describes the RC and similar methods as a combination of classical and Berkson error structures (Tosteson and Tsiatis 1988, Reeves et al. 1998, Stram and Kopecky 2003, Schafer and Gilbert, *in press*). If $E(X|W)$ is known, RC is a mapping of W into a Berkson error structure, because $X = E(X|W) + \varepsilon$, and ε is independent of $E(X|W)$. However, the uncertainty on $E(X|W)$ is best described as classical error, because it arises from the variability in the sampling distribution of the estimator.

Reeves et al. (1998) propose a representation based on a latent, unobservable random variable to encapsulate both the Berkson and classical error structures. In RC, the latent variable is *E(X|W)* and the model can be written as:

$$X = E\left(X \mid W\right) + \varepsilon_B; \qquad \hat{E}\left(X \mid W\right) = E\left(X \mid W\right) + \varepsilon_C,$$

where $\hat{E}\left(X \mid W\right)$ is the estimated value of *E(X|W)*, and ε_B and ε_C are the Berkson and classical measurement error components. Then, they specify the relationship between the random variables $\left(E\left(X \mid W\right), \varepsilon_B, \varepsilon_C\right)$ as being either independently and normally distributed or mutually uncorrelated, and assume that observations from different subjects are independent. They show that, for simple linear regression, the slope parameter is attenuated, as is to be expected from the classical error component.

This model fails to recognize that all observations in the study may share the same model to estimate *E(X|W)*. Thus, while each observation may have a unique and independent deviation from its expected value ε_B, the deviation of the estimated expectation from the true value, ε_C, would be correlated among different observations, if not fully functionally related. This 'shared error' component was noted by Stram and Kopecky (2003) and Schafer and Gilbert (*in press*).

To further clarify this issue, assume that that $E\left(X \mid W\right) = \alpha_0 + \alpha_1 W$, and that $\left(\alpha_0, \alpha_1\right)$ are estimated from a calibration study where both *X* and *W* are observed. The estimate of *E(X|W)* for each observation in the main study is based on the same parameter estimates $\left(\hat{\alpha}_0, \hat{\alpha}_1\right)$, obtained from the same calibration data. Therefore, ε_C, while different depending of the values of *W*, is functionally dependent among all observations. The estimation of $\left(\alpha_0, \alpha_1\right)$ involves a classical error component but, for a particular study, only one realization from the sampling distribution of the parameter estimates is observed. From a measurement error perspective, the effect may be closer to that of a biased measurement device, than to random measurement noise.

The approach followed in this study is to derive the sampling distribution of the RC estimator and discuss its properties. Unfortunately, analytical results are only possible in the simplest cases, so most of the work focuses in a linear, normal model with an independent calibration study. The results, however, shed light on other models as well. As it will be shown, the RC estimator tends to be inflated away from zero, rather than attenuated as would be expected if the error associated with the estimate of $E(X|W)$ followed the classical model.

1.4. Regression calibration and maximum likelihood

Regression calibration is basically a method-of-moments-like estimator, and it is unusual in statistical data analysis that a method of moments is preferred over maximum likelihood. However, likelihood methods have not been used extensively in the analysis of measurement error models, in part because of difficult computations and concerns about robustness, but also because of the belief that, in many statistical models, simpler methods such as regression calibration perform just as well as likelihood methods (Carroll et al. 1995). However, Carroll et al. (1995) note that there is little documentation to support this belief.

In some instances, if $E(X \mid W)$ is known, the RC and ML estimators are the same or very close. For example, if the regressions of Y on X and of W on X are both linear, and $Y \mid X$, $W \mid X$ and X are all normally distributed, then the RC and ML estimators give the same estimates. ML and RC also give identical results if $Y \mid X$ is Bernoulli and the regression of Y on X is linear in X, and approximately the same results for linear hazard regression (Prentice 1982, Pepe et al. 1989, Schafer et al. 2001).

There is a natural concern over the robustness of ML inferences against possible distributional misspecifications. While RC typically only requires assumptions about the first two moments of the distributions, ML requires the specification of the distribution of variables that may not even be observed. In a typical application, under the assumption of non-differential error, the joint distribution of Y, X and W is decomposed as follows:

$$f(Y, X, W) = f(Y \mid X) f(W \mid X) f(X),$$

because non-differential error implies that $f(Y \mid X, W) = f(Y \mid X)$.

This model requires the specification of three distributions, which in epidemiology have been described as the disease model, $f(Y \mid X)$, the measurement model, $f(W \mid X)$, and the exposure model, $f(X)$ (Clayton 1992). The specification of $f(X)$, is the most problematic, because it describes the distribution of the risk factor X in the population, which typically is not even observed. Addressing those concerns, several approaches for flexible structural and semiparametric modeling have been proposed (Roeder et al. 1996; Carroll et al. 1999; Schafer 2001, 2002).

When the information about the measurement process comes from a validation study, where X and W are observed, ML becomes more attractive. Then, the likelihood can be expressed conditional on the observed values of W, as $f(Y, X \mid W) = f(Y \mid X) f(X \mid W)$. It is not necessary to specify $f(X)$, only $f(X \mid W)$, and the proposed distribution can be checked against data from the validation study.

Maximum likelihood can be relatively difficult to implement, because obtaining $f(Y \mid W)$ requires solving a complex integral. If the regressions of Y on X and of X on W are both linear, and $Y \mid X$ and $X \mid W$ are both normally distributed, the integral has a closed form solution. For other models, it has to be solved through numerical analysis or simulation. The variety of types of information about the measurement error process makes the implementation case specific, requiring retooling of the algorithms and programs for each application.

Not only is RC easy and transparent, but the alternative of ML requires stronger distributional assumptions, which may be difficult to check, and possibly difficult computation. However, even though likelihood analysis may not be an easy 'off-the shelf' solution for a wide range of problems, it may be worth the additional difficulty. If data collection and study involve significant time and cost, then the additional effort involved in a likelihood analysis would be small for realizing increased flexibility, greater efficiency and more powerful tests and confidence intervals.

1.5. Organization of the dissertation

Chapter 2 examines the sampling distribution of the RC estimator in the simple case when the regressions of Y on X and of X on W are both linear, $Y \mid X$ and $X \mid W$ are both normally distributed, and information about the measurement process comes from an independent calibration study in which X and W are observed. We consider this model partly because of motivating data problems with this structure, but partly because this simple setting permits some theoretical investigations into the effect of uncertainty in estimating $E(X \mid W)$ on the RC estimator. In this case, it is also possible to obtain MLEs in closed forms or using readily available software, so we compare the efficiency of the estimators for small sample sizes using simulation.

Chapter 3 considers the same settings, but when information about the measurement process comes from an internal calibration study in which Y, X and W are observed. In this case, there is not an agreement about the implementation of the RC method, and several estimators have been proposed. In addition, it is not possible to obtain a closed form solution for the MLE, although it can easily be obtained using standard software. We will rely in extensive simulations to compare the performance of the different estimators and associated confidence intervals, both under the correct and misspecified distributional assumptions.

2. REGRESSION CALIBRATION INFERENCE FOR MEASUREMENT ERROR MODELS WITH AN INDEPENDENT CALIBRATION STUDY

Vicente J. Monleon
Daniel W. Schafer

Department of Statistics
Oregon State University

2.1 Abstract

Regression calibration seeks to estimate regression models with measurement error in explanatory variables. The mismeasured explanatory variable is replaced by its conditional expectation, given a surrogate variable, in an estimation procedure that would have been used if the true value were available. This study examines the effect of the uncertainty in the estimation of this conditional expectation on inference about regression parameters, when the true explanatory variable and its surrogate are observed in an independent calibration study and related through a normal linear model. The sampling distribution of the regression calibration estimator is skewed and its moments are not defined, but its median is approximately the parameter of interest. As the sample size of the calibration study increases, it converges to a normal distribution centered on the target parameter. The maximum likelihood estimator, assuming that the distributions are properly specified, is bounded and more efficient than the regression calibration estimator. Likelihood ratio inferences are more accurate and efficient than those based on approximate normality and estimated standard errors. The performance of regression calibration inference approaches that of likelihood inference as the calibration sample size increases, and it approaches at a faster rate for small measurement error variance.

2.2. Introduction

Because of its transparency, ease of use, and apparently good operational characteristics, the technique known as regression calibration (RC) has emerged as an important tool for estimation of regression parameters in the presence of explanatory variable measurement errors. Let Y represent the response variable, X an explanatory variable of interest that is measured with error, W a measurement or surrogate for X, and Z additional explanatory variables free of measurement error. The regression calibration estimator, in its most transparent form, uses the regression estimator that would have been used if X were known exactly (for linear, generalized linear, or failure time regression models), but with the missing X replaced by $E(X|W, Z)$ (Carroll et al. 1995, Ch. 3). Versions of the RC estimator were proposed by Prentice (1982) for Cox proportional hazards regression models, Armstrong (1985) for generalized linear models, and Rosner, Willet and Spiegelman (1989, 1990) for logistic regression.

Although regression calibration can be used in many disciplines, two notable areas of application are nutritional epidemiology (e.g. Willett et al. 1992, Binham et al. 2003, van Gils et al. 2005) and radiation health epidemiology (e.g. Pierce et al. 1990, Stram et al. 1999, Schafer et al. 2001). The former primarily involves logistic and failure time regression models for binary health responses on diet and nutrition explanatory variables, which are measured imprecisely. The latter involves failure time regression models that are linear or quadratic functions of dose of radiation, which is observed through an imprecise estimate.

It is easy to see the rationale for regression calibration in simple linear regression. If the regression of Y on X is linear, $E\left(Y \mid X\right) = \beta_0 + \beta_1 X$, then $E\left(Y \mid W\right) = E\left[E\left(Y \mid X\right) \mid W\right] = \beta_0 + \beta_1 E\left(X \mid W\right)$. Since the coefficients in the regression of Y on $E(X|W)$ are the same as those in the regression of interest, $E(Y|X)$, practical attention can be switched to the regression of Y on $E(X|W)$. This also shows that the naïve regression of Y on W will lead to biased estimation of the regression coefficients if W is not the same as $E(X|W)$, as is the case under the classical measurement error model. In the classical model, W is the sum of X and a random measurement error that is independent of X, and the estimated slope of the regression line is biased towards zero (see, for example, Madansky 1959, Cochran 1965). On the other hand, under the Berkson error model, in which $E(X|W) = W$ follows from the model definition, the usual estimators of regression coefficients are unbiased (Berkson 1950).

If $E(Y|X, Z)$ is not linear in X, the simple substitution of $E(X|W, Z)$ in place of X in the regression model $E(Y|X, Z)$ leads to an approximate model for $E(Y|W, Z)$. Estimation based on this substitution is, in general, biased and inconsistent (Carroll et al. 1995). However, under additional assumptions that depend on the particular model, it is approximately consistent, and the approximation may be improved with a second-order approximation to $E(Y|X, Z)$ about $X = E(X|W, Z)$ (Carroll and Stefanski 1990, Kuha 1994). For generalized linear models, $Var(Y|W)$ will not have the same form as $Var(Y|X)$, and some attention to proper "weighting" is necessary for efficient estimation. If the distribution of Y given X and Z is Poisson and W is a "classical" measurement of X, for example, the distribution of Y given W and Z is not Poisson and, in particular, the variance is greater than the mean.

For the following models, regression calibration and maximum likelihood give the same estimates:

1. The "everything normal" linear structural model when either $E(X|W)$ or the reliability coefficient, defined as $\lambda = \sigma_X^2/(\sigma_X^2 + \sigma_{W|X}^2)$, is known. For this model, $Y \mid X \sim N\left(\beta_0 + \beta_1 X, \sigma_{Y|X}^2\right)$, $W \mid X \sim N\left(\alpha_0 + \alpha_1 X, \sigma_{W|X}^2\right)$, and $X \sim N\left(\mu, \sigma_X^2\right)$. Let $W^* = E\left(X \mid W\right) = \mu\left(1-\lambda\right) + \lambda W$. If either λ or W* are known, the maximum likelihood estimator of β_1 is $SS_{W^*Y} / SS_{W^*W^*}$, where the SS's are sums of squares or cross products indicated by their subscripts. This can be seen by equating the five sufficient statistics from the bivariate normal distribution of Y and W to their

expectations based on five unknown parameters. This estimator is equivalent to the least squares estimator for the linear regression of Y on X, but with X replaced by $E(X|W)$.

2. Bernoulli linear model when $E(X|W)$ is known. $Y \mid X \sim Bin(1, \pi)$; $\pi = \beta_0 + \beta_1 X$. It must be true that $Y \mid W \sim Bin(1, \pi^*)$; $\pi^* = \beta_0 + \beta_1 E(X \mid W)$, so maximum likelihood based on $Y \mid W$ is equivalent to using maximum likelihood based on $Y \mid X$ but with $E(X|W)$ used in place of X.

In addition, there is a model for time to response data in which regression calibration and maximum *partial* likelihood are approximately the same: the linear proportional hazards model when $E(X|W)$ is known. As shown in Prentice (1982), Pepe et al. (1989), and Schafer et al. (2001, appendix), the hazard function for a waiting time as a function of the measurement W is approximately the expected value of the hazard as a function of X, conditional on W. The approximation is good for the "rare disease" case. If the hazard function is linear in X, then the induced hazard is linear in $E(X|W)$. Therefore, any method, such as maximum partial likelihood based on W will be the same as it would be for X but with X replaced by $E(X|W)$.

These last two models are not broadly useful, but they are important for radiation research where there is theoretical and empirical justification for probability and hazard rate models that are linear in radiation dose (Pierce et al. 1992).

These equivalences of regression calibration and likelihood estimators, and the unbiasedness of the RC estimator when $E(Y|X, Z)$ is linear in X, are only true if the conditional mean $E(X|W, Z)$ is known, which is seldom the case. The properties of the regression calibration method are obviously more complicated when the uncertainty in $E(X|W, Z)$ is acknowledged. This paper focuses on the role of uncertainty in $E(X|W, Z)$ in regression calibration inference.

Uncertainty in $E(X|W, Z)$ is not negligible in most epidemiological studies. For example, for diet and health research from the Nurse's Health Study, a primary data set consists of almost 90,000 nurses. Investigations consider the regression of health outcomes on explanatory variables, X, associated with diet, such as total fat intake. Surrogates, W, for this type of variable are measured on all nurses in the primary study, and $E(X|W, Z)$ is taken to be an estimated mean from a regression model fit to a calibration dataset of 173 nurses (Willett et al. 1992). The estimated correlation between X and W is low – typically, between 0.4 and 0.6 (Willett et al. 1988). The values used as $E(X|W, Z)$ for each of the nurses in the primary data set therefore contain a component of uncertainty due to the sampling error from the regression estimation on the calibration set.

An interesting aspect of the problem is that this uncertainty component is shared by all the nurses in the study, since the estimate of $E(X|W, Z)$ for each individual in the primary data set is based on the same estimated regression equation. Fraser and Stram (2001) examined the effect of this kind of "shared" uncertainty in estimating $E(X|W, Z)$ on the power of tests based on regression calibration. Through simulations, they noted the need for sample sizes that were considerably larger than those usually available.

Because regression calibration is used for important statistical problems in epidemiology, we feel it is appropriate to explore in more detail further practical effects associated with the uncertainty in $E(X|W, Z)$. We are interested in both linear and nonlinear regression models for $E(Y|X, Z)$, and also in various types of data structures that permit the estimation of $E(X|W, Z)$. To start, though, we consider linear regression of Y on X and Z, with a model for $E(X|W, Z)$ estimated from an external calibration data set. This means that a data set is available with observations on X and W and Z (but not Y), separate from the primary data set. W may be a measurement of X (in which case the calibration data set is typically referred to as an external validation study) or a surrogate variable that is associated with X and which can be used to predict X. While the linear model with an independent calibration set situation is of interest in itself, it is studied here as one of the more transparent structures for isolating the effect of the uncertainty in the estimate of $E(X|W, Z)$, for a first step in exploring the use of regression calibration with inexact calibration more generally. It is, of course, essential that $E(X|W, Z)$ is the same in the primary data set as in the calibration data set or, in other words, the calibration equation must be portable. This is an important practical issue, but will not be considered further in this paper.

The RC estimator in some models is also a method-of-moments estimator (Appendix A) and is like a method-of-moments estimator more generally. It is often easy to implement and it is relatively transparent while alternatives, such as maximum likelihood, usually are not. However, maximum likelihood may efficiently combine the information from several sources, such as the primary and calibration studies. Therefore, in this study, we wish to further examine its relative efficiency and the accuracy of inferences based on its approximate normality in light of the uncertainty in $\mathrm{E}(X|W, Z)$.

This paper is organized as follows. Section 2.2 describes the model of interest, the regression calibration estimator, and the methods typically used to estimate its standard error for approximate tests and confidence intervals. Section 2.3 discusses the exact sampling distribution of the RC calibration estimator for a simple linear-linear model when $Y \mid X$ and $X \mid W$ are normally distributed. Relaxing the normality assumption, it provides an approximation to the bias of the RC estimator when its expectation is defined. Section 2.4 compares the RC and maximum likelihood (ML) estimators, and discusses the properties of the MLE. Section 2.5 presents the results of a set of simulations devised to explore the effect of the uncertainty in $E(X|W)$ on the performance of the RC and ML estimators and associated confidence intervals. Section 2.6 summarizes the main conclusions of this report.

2.3. The regression calibration estimator for linear regression when $E(X \mid W)$ is estimated from an independent calibration dataset.

Consider the model:

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{2.1}$$

$$X = \alpha_0 + \alpha_1 W + \delta \tag{2.2}$$

where Y is the response variable; X is the explanatory variable; W is a surrogate for X; $(\beta_0, \beta_1, \alpha_0, \alpha_1)$ are unknown regression coefficients; and (ε, δ) are random variables.

It follows that:

$$Y = (\beta_0 + \alpha_0\beta_1) + \alpha_1\beta_1 W + (\varepsilon + \beta_1\delta) \tag{2.3}$$

Let $\gamma_0 = \beta_0 + \alpha_0\beta_1$ and $\gamma_1 = \alpha_1\beta_1$ be the coefficients of the regression of Y on W.

Suppose that (1) there is a primary sample consisting of observations (y_i, w_i), $i = 1,...,n$, and an independent calibration sample consisting of observations (x_i, w_i), $i = n+1,...,n+m$; (2) (ε, δ) are independent random errors with means equal to 0; (3) random variables associated with different values of i are independent of one another; and (4) the error structure is non-differential, meaning $f(Y \mid X, W) = f(Y \mid X)$.

Notice that because of the existence of the calibration data set it is not necessary to make distributional assumptions for X, as long as E(X|W) has the same form in the primary and calibration data sets. If there are additional explanatory variables Z, free of measurement error, then all expectations should also be conditional on Z; but that notation will be suppressed.

For this model, the regression calibration estimator of β_1 can be defined following two different but equivalent approaches. One approach consists of estimating (α_0, α_1) by $(\hat{\alpha}_0, \hat{\alpha}_1)$ from the external calibration sample, calculating $\hat{x}_i = \hat{\alpha}_0 + \hat{\alpha}_1 w_i$ for each observation in the primary sample, and estimating the slope of the regression of y_i on $\hat{x}_i$ using least squares (e.g., Carroll et al. 1995, Chapter 3). Then, the regression calibration estimator takes the form (see also Appendix B.1):

$$\hat{\beta}_{1,RC} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{x}_i - \bar{\hat{x}})}{\sum_{i=1}^{n}(\hat{x}_i - \bar{\hat{x}})^2} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{\alpha}_1 w_i - \hat{\alpha}_1 \bar{w})}{\sum_{i=1}^{n}(\hat{\alpha}_1 w_i - \hat{\alpha}_1 \bar{w})^2} = \frac{\hat{\gamma}_1}{\hat{\alpha}_1} \tag{2.4}$$

where $\hat{\gamma}_1$ and $\hat{\alpha}_1$ are the least squares estimators of the slope of *Y* on *W*, based on the primary data, and of *X* on *W*, based on the calibration data, respectively.

The other approach arrives at the same estimator directly from equation (2.3), by solving $\gamma_1 = \alpha_1\beta_1$ for β_1 and substituting γ_1 and α_1 by their respective estimators (Rosner et al. 1989). More generally, as long as the regression of *X* on *W* is linear, both approaches yield $\hat{\beta}_{1,RC}$. This is true, for example, when the regression of *Y* on *X* is a generalized linear model. This is shown in Appendix B, as is the form of the estimator when there are additional explanatory variables measured with or without error (see also Thurston et al. 2003).

Without additional distributional assumptions, it is not possible to derive the sampling distribution of the RC estimator. Asymptotic properties are based on the theory of stacked estimating equations (Carroll et al. 1995, Appendix). In an asymptotic setting in which *n* and *m* both increase to infinity, the sampling distribution of the RC estimator converges to a normal distribution (Carroll and Stefanski 1990). The mean of this distribution is β_1 and, under the additional assumption that the variances of (ε, δ) are constant, the variance is

$$Var\left(\hat{\beta}_{1,RC}\right) = \frac{\gamma_1^2}{\alpha_1^4}\sigma_{\hat{\alpha}}^2 + \frac{1}{\alpha_1^2}\sigma_{\hat{\gamma}}^2 \tag{2.5}$$

where $\sigma_{\hat{\gamma}}^2 = Var\left(\hat{\gamma}_1\right)$ and $\sigma_{\hat{\alpha}}^2 = Var\left(\hat{\alpha}_1\right)$. Tests and confidence intervals are based on this asymptotic distribution, with unknown parameters replaced by their estimates, but the bootstrap method can also be used (Carroll et al. 1995, Rosner et al. 1989).

Confidence intervals based on asymptotic normality have been widely used and have been implemented in readily available software (Spiegelman et al. 1997). However, although asymptotically correct as $m \to \infty$, they may not have very desirable finite sample properties. While the interval is symmetric about $\hat{\beta}_{1,RC}$, the actual sampling distribution of $\hat{\beta}_{1,RC}$ can be very skewed, even for relatively large sample sizes. A bootstrap confidence interval, based on the percentiles of the bootstrap replications, should perform better in this case. In addition, the expectation of the estimated variance given by (2.5) typically does not exist. Therefore, for any finite sample size, a confidence interval with non-zero coverage probability (e.g. a 95% CI) has an expected infinite length, a common feature of this type of confidence interval in the measurement error problem (Gleser and Hwang 1987). The bootstrap confidence interval is not immune to this problem because, as it will be shown, the moments of the distribution of the regression calibration estimator are not defined (Athreya 1987). The practical consequences of this are that the width of both types of confidence intervals can be very large and erratic if the sampling distribution of $\hat{\alpha}_1$ has positive mass at 0. This will become apparent in the simulations of Section 6.

2.4. Exact sampling distribution of the regression calibration estimator

2.4.1. Sampling distribution when Y|X and X|W are normally distributed

Suppose that the variables ε and δ in (2.1) and (2.2) follow normal distributions with means 0 and constant, positive variances $\sigma^2_{Y|X}$ and $\sigma^2_{X|W}$, respectively. Then, the distribution of $Y|W$ is

$$y_i \mid w_i \sim N\left(\beta_0 + \alpha_0\beta_1 + \alpha_1\beta_1 w_i, \sigma^2_{Y|X} + \beta_1^2\sigma^2_{X|W}\right)$$

As before, let $\gamma_0 = \beta_0 + \alpha_0\beta_1$, $\gamma_1 = \alpha_1\beta_1$ and $\sigma^2_{Y|W} = \sigma^2_{Y|X} + \beta_1^2\sigma^2_{X|W}$ be the parameters of the distribution of Y given W. First notice that the least squares estimators of $\hat{\gamma}_1$ (from the regression of Y on W in the primary data) and $\hat{\alpha}_1$ (from the regression of X on W in the calibration data) are normally and independently distributed. The regression calibration estimator is the ratio of these two, so its sampling distribution is that of a ratio of two independent normal random variables. The probability density function, cumulative distribution function and asymptotic properties are discussed in Appendix C, and other parametrizations are given by Hinkley (1969) and Marsaglia (1965). Some features of this distribution are:

1. It depends on three parameters: a scale parameter $\eta = \sqrt{\sigma_{\hat{\alpha}}^2 / \sigma_{\hat{\gamma}}^2}$, where $\sigma_{\hat{\gamma}}^2 = Var(\hat{\gamma}_1)$ and $\sigma_{\hat{\alpha}}^2 = Var(\hat{\alpha}_1)$; and $\tau_\gamma = \gamma_1 / \sqrt{\sigma_{\hat{\gamma}}^2}$ and $\tau_\alpha = \alpha_1 / \sqrt{\sigma_{\hat{\alpha}}^2}$, the reciprocals of the coefficient of variation of the sampling distributions of $\hat{\gamma}_1$ and $\hat{\alpha}_1$, respectively. The parameter τ_α , which depends only on the calibration study, plays an important role in the behavior of the RC estimator.
2. Its moments are not defined. This is a common feature of estimators in the measurement error problem that are derived using the method of moments (Fuller 1987). The sampling distribution of $\hat{\alpha}_1$ has positive mass at 0, so the distribution of $\hat{\beta}_{1,RC}$ is heavy tailed. Therefore, in theory, the RC estimator can attain very large values and behave erratically. In addition, it can be difficult to compare alternative estimators based on their moment properties, such as bias and mean square error.
3. Although the mean is not defined, the median is approximately equal to the target parameter, β_1 (Appendix C).
4. It is symmetric when either τ_γ or τ_α are 0. In general, though, it is skewed away from 0. Therefore, symmetric confidence intervals based on the asymptotic normality of the RC estimator may not be appropriate for finite sample sizes.
5. It can be unimodal or bimodal. In the latter case, it has a positive and a negative mode, but one of the modes may be insignificant (Marsaglia 1965). Absurd modal values of very large magnitude and the opposite sign to that expected are possible.

Figure 2.1 shows the exact sampling distribution of the RC estimator for a selection of situations, and table G2 (Appendix G) shows the parameters of those distributions. The figures illustrate that the sampling distribution of the RC estimator is very skewed for small *m* (and negative values of the estimator are possible). As *m*

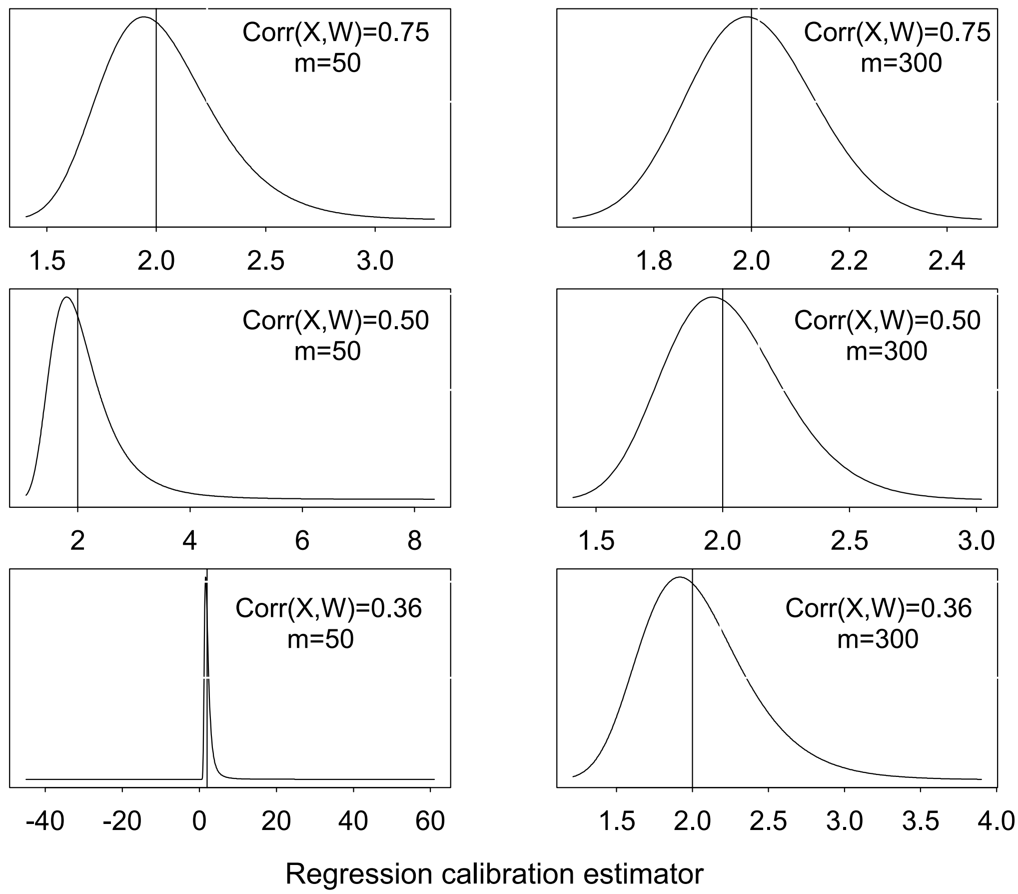

Figure 2.1. Probability density function of the sampling distribution of the RC estimator for several choices of $\sigma^2_{X|W}$ and calibration sample size *m*. The true value of the slope of the regression of *Y* on *X*, β_1, is 2.The X axis has been scaled to cover from the 0.001 to the 0.999 quantiles of the distribution.

increases, it converges to a normal distribution centered about β_1. It converges to normality at a faster rate for small $\sigma^2_{X|W}$.

Asymptotically, as $m \to \infty$ for fixed *n*, the distribution of the RC estimator converges in distribution to a normal distribution with mean β_1 and variance

$\sigma_{\hat{\gamma}}^2 / \alpha_1^2$ (Appendix C.2). More interestingly, as $\tau_\alpha \to \infty$ or, equivalently, if values of $\hat{\alpha}_1$ close to 0 are unlikely,

$$F\left(\hat{\beta}_{1,RC} \mid W\right) \to \Phi\left\{ \frac{\alpha_1 \hat{\beta}_{1,RC} - \gamma_1}{\left(\sigma_{\hat{\alpha}}^2 \hat{\beta}_{1,RC}^2 + \sigma_{\hat{\gamma}}^2\right)^{1/2}} \right\},$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal random variable. This distribution is a useful approximation to the true sampling distribution for large τ_α but, for finite τ_α, it is an improper distribution (Hinkley 1969). Although skewed away from 0, its median is β_1 (Appendix C.2).

2.4.2. Properties of the sampling distribution more generally

The results presented in the previous section only require that the sampling distributions of $\hat{\gamma}_1$ and $\hat{\alpha}_1$ be normal, and that the regression calibration estimator be defined as their ratio. Therefore, they apply more generally when the regression of *Y* on *X* (2.1) and *X* on *W* (2.2) include additional variables, but the errors still follow a normal distribution. They will also be approximately true for any kind of regression for which the estimated regression coefficients have approximately normal sampling distributions. In most epidemiological studies, the size of the primary sample is very large so, if the regression of *Y* on *W* follows a generalized linear model, the distribution of $\hat{\gamma}_1$ would be approximately normal. Typically, the regression of *X* on *W* is linear, so that the distribution of $\hat{\alpha}_1$ from the calibration study is approximately normal, and the results discussed in the previous section will apply approximately.

The RC estimator is the ratio of two estimators of slope. If the expectation of the ratio exists and if $E(\hat{\gamma}_1) = \gamma_1$ and $E(\hat{\alpha}_1) = \alpha_1$, and both have the same sign, it follows from Jensen's inequality that:

$$E\left(\hat{\beta}_{1,RC}\right) = E\left(E\left(\hat{\gamma}_1 / \hat{\alpha}_1 \middle| \hat{\alpha}_1\right)\right) = \gamma_1 E\left(1/\hat{\alpha}_1\right) \geq \gamma_1 / \alpha_1 = \beta_1 \geq 0$$

Likewise, if γ_1 and α_1 have different signs, $E\left(\hat{\beta}_{1,RC}\right) \le \beta_1 \le 0$. Therefore, under these conditions, any bias in the regression calibration estimator due to sampling error in *E(X | W)* would be described as inflation. This inflation effect on the regression coefficient is surprising, since it the opposite of the expected attenuation in the classical measurement error setting (Madansky 1959, Cochran 1965).

If is assumed that the expectation of $\hat{\beta}_{1,RC}$ exists, the bias of the RC estimator can be explored with a second-order Taylor expansion of $\hat{\beta}_1$ around $\beta_1 = \gamma_1/\alpha_1$:

$$E\left(\hat{\beta}_{1,RC}\right) \approx \frac{\gamma_1}{\alpha_1} + \frac{\gamma_1}{\alpha_1^3} Var\left(\hat{\alpha}_1\right) = \beta_1\left[1 + \frac{1}{\tau_\alpha^2}\right]$$

Thus, the relative bias depends only on the calibration study, and only through τ_α. If α_1 is estimated using least squares and $Var\left(X \mid W\right) = \sigma_{X|W}^2$ is constant, then $\tau_\alpha^2 = \left(m-1\right) S_{Wc}^2 \dfrac{\alpha_1^2}{\sigma_{X|W}^2}$, where $S_{Wc}^2 = \dfrac{1}{m-1}\sum_{i=n+1}^{n+m}\left(w_i - \overline{w}_c\right)^2$. The bias depends on the sample size of the calibration study, the relative magnitude of α_1^2 and $\sigma_{X|W}^2$, and the sample variance of *W* in the calibration study. In many epidemiological studies of diet-disease association, the relationship between *X* and *W* is weak, with R^2 in the range of 0.1 to 0.5, so that α_1 tends to be small compared with $\sigma_{X|W}^2$. Then, a large calibration sample size may be needed to reduce the bias. Even a greater sample size may be needed to reduce the variance and increase power (Fraser and Stram 2001).

The parameter τ_α^2 can also be written as $\tau_\alpha^2 = \left(m-1\right)\dfrac{S_{Wc}^2}{Var\left(W\right)}\dfrac{\rho^2}{1-\rho^2}$, where ρ is the correlation between *X* and *W* (Appendix C.3). Therefore, if S_{Wc}^2 is close to its expectation, *Var(W)*, then $\tau_\alpha^2 \approx \left(m-1\right)\dfrac{\rho^2}{1-\rho^2}$. In the diet-disease studies, ρ is typically between 0.3 and 0.7, so that $\frac{\rho^2}{1-\rho^2}$ is between 0.1 and 0.96, and increases rapidly as ρ becomes closer to 1.

Simulation results, detailed in Section 2.6, also indicate that one of the effects of uncertainty in E(X | W) is inflation in the estimated coefficient of X.

2.5. The relationship between regression calibration and maximum likelihood in the normal-normal model

The joint density function of Y, observed in the primary study, and X, observed in an independent calibration study, conditional on the observed values of W, is:

$$f\left(\mathbf{y},\mathbf{x}\,|\,\mathbf{w}\right)=\prod_{i=1}^{n} f\left(y_i\,|\,w_i\right)\prod_{i=n+1}^{n+m} f\left(x_i\,|\,w_i\right)$$

Typically, $f\left(Y\,|\,W\right)$ cannot be obtained analytically, so numerical integration may be necessary. However, under the normal distributional assumptions of section 2.4.1, there is a closed form solution for the log likelihood:

$$\begin{aligned} l\left(\boldsymbol{\beta},\boldsymbol{\alpha},\sigma^2_{Y|X},\sigma^2_{X|W}\right) &= -\frac{n}{2}\log\left(\sigma^2_{Y|X}+\beta_1^2\sigma^2_{X|W}\right)-\frac{m}{2}\log\left(\sigma^2_{X|W}\right) \\ &\quad -\frac{1}{2\left(\sigma^2_{Y|X}+\beta_1^2\sigma^2_{X|W}\right)}\sum_{i=1}^{n}\left[y_i-\left(\beta_0+\beta_1\alpha_0+\beta_1\alpha_1 w_i\right)\right]^2 \\ &\quad -\frac{1}{2\sigma^2_{X|W}}\sum_{i=n+1}^{n+m}\left[x_i-\left(\alpha_0+\alpha_1 w_i\right)\right]^2 \end{aligned} \tag{2.6}$$

The regression calibration estimators are shown in Appendix D.1 to be the unconstrained solution to the likelihood equations. The estimators of $\left(\alpha_0,\alpha_1,\sigma^2_{X|W}\right)$ are the MLEs of those parameters based on the calibration data alone, and the estimator of $\left(\beta_0,\beta_1,\sigma^2_{Y|X}\right)$ is

$$\left(\hat{\beta}_0,\hat{\beta}_1,\hat{\sigma}^2_{Y|X}\right)=\left[\hat{\gamma}_0-\frac{\hat{\gamma}_1}{\hat{\alpha}_1}\hat{\alpha}_0,\frac{\hat{\gamma}_1}{\hat{\alpha}_1},\hat{\sigma}^2_{Y|W}-\left(\frac{\hat{\gamma}_1}{\hat{\alpha}_1}\right)^2\hat{\sigma}^2_{X|W}\right]$$

where $\left(\hat{\gamma}_0, \hat{\gamma}_1, \hat{\sigma}^2_{Y|W}\right)$ are the MLEs of the parameters of the distribution of *Y* given *W*, based on the primary data only.

The estimator of $\sigma^2_{Y|X}$ can result in a negative estimate. To find the MLEs, the likelihood has to be maximized under the constraint that $\hat{\sigma}^2_{Y|W} - \left(\hat{\gamma}_1/\hat{\alpha}_1\right)^2 \hat{\sigma}^2_{X|W} > 0$. Equivalently, if $\hat{\sigma}^2_{Y|X} < 0$, $\sigma^2_{Y|X}$ can be set to 0 in (2.6), yielding a new set of likelihood equations and estimators (Appendix D.2). The solution for β_1 are the roots of the following quadratic equation:

$$\begin{aligned} n\left[SS^2_{XWc} - SS_{XX}\left(SS_{WpWp} + SS_{WcWc}\right)\right]\hat{\beta}^2_1 + (n-m)\,SS_{XWc}SS_{YWp}\hat{\beta}_1 \\ - m\left[SS^2_{YWp} - SS_{YY}\left(SS_{WpWp} + SS_{WcWc}\right)\right] = 0 \end{aligned} \tag{2.7}$$

where the *SS*'s denote are sums of squares or cross products indicated by their subscripts. The two solutions to this quadratic equation are real numbers, one positive and one negative, and are different from the regression calibration estimator.

The MLE, $\hat{\beta}_{1,MLE}$, equals $\hat{\gamma}_1/\hat{\alpha}_1$, the RC estimator, if $\hat{\sigma}^2_{Y|W} - \left(\hat{\gamma}_1/\hat{\alpha}_1\right)^2 \hat{\sigma}^2_{X|W} > 0$, or one of the roots of eq. (2.7) if $\hat{\sigma}^2_{Y|W} - \left(\hat{\gamma}_1/\hat{\alpha}_1\right)^2 \hat{\sigma}^2_{X|W} \le 0$. In fact, $\hat{\beta}_{1,MLE}$ is the minimum, in absolute value, of those two estimates and it is bounded (Appendix D.3). This is practically relevant, because the RC estimator can be very unstable and reach very large values when $\hat{\alpha}_1$ is close to 0, as indicated by the lack of moments of its sampling distribution. There is a close relationship between a large absolute value of the RC estimated regression coefficient and a negative estimate of $\sigma^2_{Y|X}$, because as $\hat{\alpha}_1$ becomes small, $\hat{\beta}_{1,RC} = \hat{\gamma}_1/\hat{\alpha}_1$ is large and $\hat{\sigma}^2_{Y|W} - \left(\hat{\gamma}_1/\hat{\alpha}_1\right)^2 \hat{\sigma}^2_{X|W} < 0$. Several solutions have been proposed to improve the behavior of other method-of-moments type of estimators in the context of the measurement error problem (Fuller 1987). These methods impose a bound on the estimator, a result obtained with maximum likelihood in a less ad-hoc manner.

Combining the constrained and unconstrained forms, the log likelihood equation, can be written from eq. (2.6) as:

$$\begin{aligned} l_c\left(\boldsymbol{\beta},\boldsymbol{\alpha},\sigma^2_{Y|X},\sigma^2_{X|W}\right) &= l\left(\boldsymbol{\beta},\boldsymbol{\alpha},\sigma^2_{Y|X},\sigma^2_{X|W}\right) I_{\left(\hat{\sigma}^2_{Y|X}>0\right)} \\ &+ l\left(\boldsymbol{\beta},\boldsymbol{\alpha},\sigma^2_{Y|X}=0,\sigma^2_{X|W}\right)\left(1-I_{\left(\hat{\sigma}^2_{Y|X}\le 0\right)}\right) \end{aligned} \quad . \tag{2.8}$$

where $I_{\left(\hat{\sigma}^2_{Y|X}>0\right)}$ is an indicator variable that depends only on the data and takes the value 1 if $\hat{\sigma}^2_{Y|X} = \hat{\sigma}^2_{Y|W} - \left(\hat{\gamma}_1/\hat{\alpha}_1\right)^2 \hat{\sigma}^2_{X|W} > 0$ (in which the parameters with hats are the regression calibration estimators). In practice, equation (2.6) can be parametrized as a function of $\log\left(\sigma^2_{Y|X}\right)$ and $\log\left(\sigma^2_{X|W}\right)$, and maximized using standard software (such as the S-Plus 'nlmin' or R 'nlm' functions).

Under standard regularity conditions, the estimators $\left(\hat{\alpha}_1,\hat{\gamma}_1,\hat{\sigma}^2_{X|W},\hat{\sigma}^2_{Y|W}\right)$ are consistent. Therefore, as $n\to\infty$ and $m\to\infty$, $\hat{\gamma}_1/\hat{\alpha}_1 \to \beta_1$, $\hat{\sigma}^2_{X|W}\to\sigma^2_{X|W}$, $\hat{\sigma}^2_{Y|W}\to\sigma^2_{Y|W}=\sigma^2_{Y|X}+\beta_1^2\sigma^2_{X|W}$, and:

$$P\left[\hat{\sigma}^2_{Y|W} - \left(\hat{\gamma}_1/\hat{\alpha}_1\right)^2 \hat{\sigma}^2_{X|W} > 0\right] \to P\left(\sigma^2_{Y|X}>0\right)=1$$

Thus, asymptotically, only the unconstrained likelihood equation is relevant, and the ML and RC estimators are asymptotically equivalent. However, for *m* or *n* finite, $P\left(\hat{\sigma}^2_{Y|X}<0\right)$ may not be negligible, even for large sample sizes (Appendix E). The sampling distribution of the MLE based on $l\left(\boldsymbol{\beta},\boldsymbol{\alpha},\sigma^2_{Y|X}=0,\sigma^2_{X|W}\right)$ is not asymptotically normal, and the asymptotic distribution of the likelihood ratio test statistic does not converge to a χ^2, because one of the nuisance parameters is in the boundary of the parameter space (Lehmann and Casella 1998).

The profile log likelihood function of β_1 can be obtained analytically, by evaluating the log likelihood at the conditional MLE of $\left(\alpha_0,\alpha_1,\beta_0,\sigma^2_{Y|X},\sigma^2_{X|W}\right)$ for fixed β_1 (Appendix F). As with the log likelihood, the profile log likelihood has two components, depending on whether $\hat{\sigma}^2_{Y|X}(\beta_1) = \hat{\sigma}^2_{Y|W}(\beta_1) - \beta_1^2 \hat{\sigma}^2_{X|W}(\beta_1)$ is positive or negative. In the latter case, the profile likelihood function is calculated under $\sigma^2_{Y|X} = 0$, resulting is a function that has a vertical asymptote at 0, and two local maximums. As a result, the profile likelihood for β_1 can be bimodal or unimodal, depending on the values of the parameters and the sample sizes of the primary and calibration studies. Figure 2.2 shows some results for simulated data.

The parameter space of β_1 is limited by the non-negativity of the variance of $Y|W$, so that $|\beta_1| \leq \sqrt{\sigma^2_{Y|W} / \sigma^2_{X|W}}$. Thus, $\hat{\sigma}^2_{Y|X}(\beta_1)$ eventually becomes negative as β_1 increases in magnitude, and so it is set to 0 when calculating the profile likelihood function. This restriction is reflected in the profile log-likelihood function, which drops rapidly as β_1 increases over its allowable range (Fig. 2.2). From a practical perspective, this property imposes a desirable bound in the allowable values of the MLE of β_1 but, on the other hand, confidence intervals calculated by inverting a likelihood ratio test may not perform properly. A confidence interval can be obtained by calculating the set of values of β_1 'not rejected' by a likelihood ratio test, but some values in this set may result in a negative estimate of $\sigma^2_{Y|X}(\beta_1)$. The sampling distribution of a likelihood ratio test statistic when β_1 is outside its parameter space may not be well approximated by a χ^2. This topic will be explored further with simulations.

2.6. Simulations

A primary objective of the simulations is to examine the effect of the uncertainty in the estimation of $E(X \mid W)$ on the operating characteristics of the regression calibration estimator and on the performance of tests and confidence intervals based on its approximate normality. Part of this is based on comparisons with likelihood analysis. The conditions and details of the simulations are described in Appendix G.

Figure 2.2. Log profile likelihood of β_1 for simulated data with different calibration sample sizes (m). The log profile likelihood of β_1 (solid line), restricted so that $\hat{\sigma}^2_{Y|X} > 0$, is the result of combining two functions: a log profile likelihood that does not place any restrictions on $\hat{\sigma}^2_{Y|X}$ (dashed line) and a log profile likelihood that sets $\sigma^2_{Y|X} = 0$ (dotted line). The true value of β_1 is 2 and the parameter space of β_1 is $|\beta_1| < \sqrt{\sigma^2_{Y|W} / \sigma^2_{X|W}} = 2.154$ (dashed vertical lines). The solid vertical line denotes the MLE. The maximum of the unrestricted log profile likelihood is the RC estimator (not shown). As the sample size of the calibration study increases, the profile log likelihood becomes unimodal. The parameters used in this example are $\left(\alpha_0, \alpha_1, \beta_0, \beta_1, \sigma^2_{Y|X}, \sigma^2_{X|W}\right) = \left(0, 1, 1, 2, 0.6^2, 0.75^2\right)$ and n = 1000.

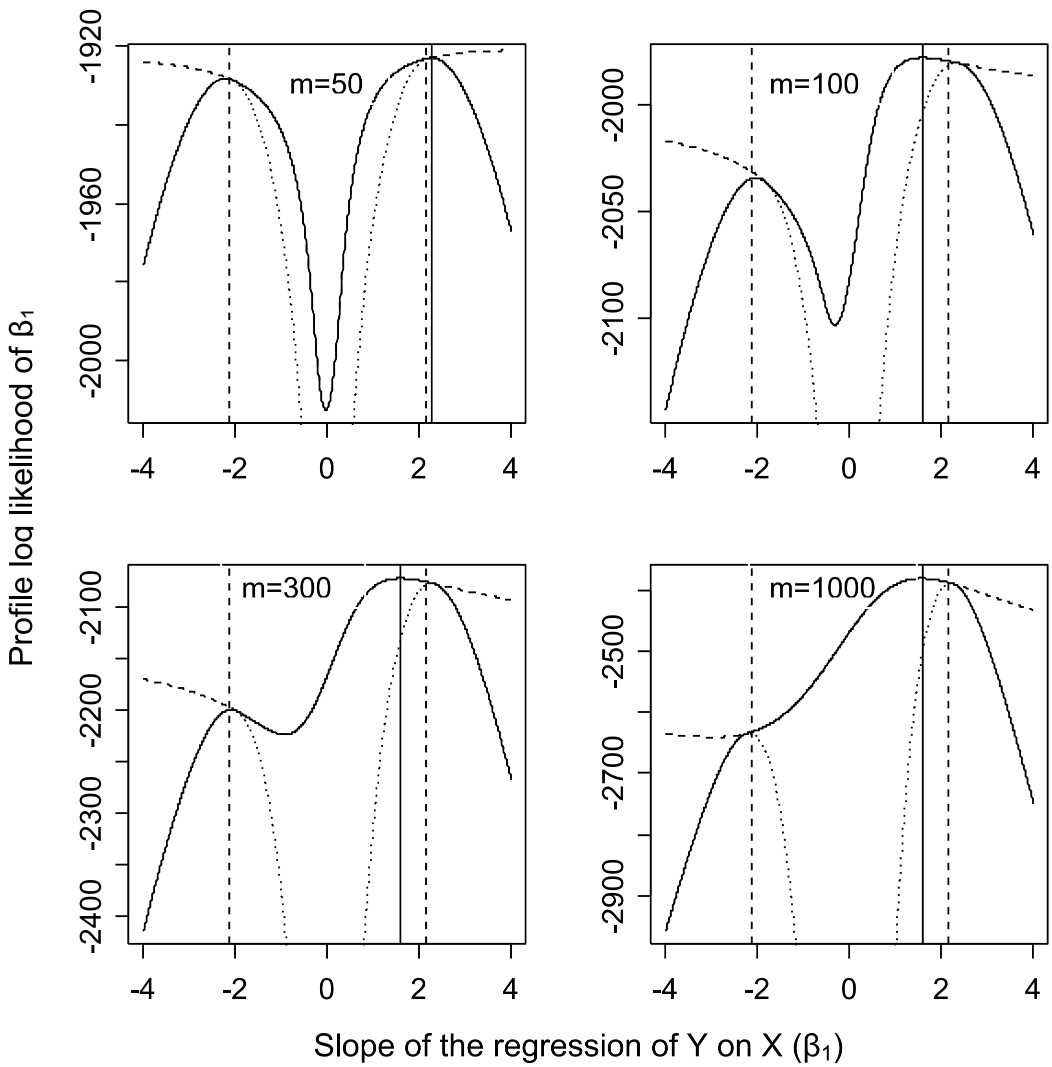
m=50
m=100
m=300
m=1000
-1920
-1960
-2000
-2000
-2050
-2100
-2100
-2200
-2300
-2400
-2500
-2700
-2900
-4
-2
0
2
4
Profile log likelihood of β_1
Slope of the regression of Y on X (β_1)

Figure 2.2.

2.6.1. Performance of the estimators

Figures 2.3-2.5 (and tables G.3 and G.4, in Appendix G) show the average, RMSE and selected quantiles of the Monte Carlo approximation to the sampling distribution of the regression calibration and maximum likelihood estimators after 2500 simulations. The results support the theoretical results from previous sections in the following ways. First, the median of the RC estimator is about equal to the target value (Fig. 2.5) but the mean tends to be greater than the target value (Fig. 2.3). The bias, however, decreases rapidly as the sample size of the calibration study is increased, even when the measurement error variance is large. Second, the behavior of the RC estimator depended mostly on the value of $\tau_{\alpha} = \alpha_1 \Big/ \sqrt{\sigma^2_{\hat{\alpha}}}$. Settings with similar values of this parameter (Table G.2, appendix G) have similar distributional characteristics (Table G.3, appendix G). Third, when the probability of a negative estimate of the $\sigma^2_{Y|X}$ is small, the RC estimator and MLE are about the same, as the results for $\beta_1 = 2$ and *Corr(X | W)* = 0.76, and for $\beta_1 = 1/2$, *Corr(X | W)* = 0.50, and $m \geq 300$, indicate.

The MLE is substantially more efficient and less erratic than the RC estimator when the correlation between *X* and *W* is low or moderate and the calibration sample size small, as evident in figure 2.4. The bias of the MLE, and its RMSE to a much greater degree, are smaller than those of the RC estimator (Figs. 2.3 and 2.4), mostly because the distribution of the RC estimator has heavy tails and some very large values. This is reflected in the quantiles of the respective distributions (Fig. 2.5). The 2.5% percentile and the median of the distributions are virtually identical, but the 97.5% percentile of the distribution of the RC estimator is always much larger than that of the MLE. Large values of the RC estimator correspond to small values of $\hat{\alpha}_1$, which are more likely to occur when *m* is small and $\sigma^2_{X|W}$ is large. Those values also result in negative estimates of $\sigma^2_{Y|X}$, so the MLE is one of the roots of eq. (2.7) instead of being identical to the RC estimator.

Even though the MLE is more efficient than the RC estimator when $\hat{\sigma}^2_{Y|X} \le 0$, estimators based on the likelihood function with $\sigma^2_{Y|X}$ set to 0 are not consistent and their distribution is not asymptotically normal, as apparent in figure 2.6. In the situation shown there, with *n* fixed at 1000, $P\left(\hat{\sigma}^2_{Y|X} \le 0\right) \xrightarrow{m\to\infty} 0.20$, and the MLE is frequently based on $l\left(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2_{Y|X} = 0, \sigma^2_{X|W}\right)$ even for very large *m*. The distribution of the RC estimator is very skewed for *m*=100, but is nearly normal for *m*=1000, while that of the MLE does not approach normality. Figure 2.6 also illustrates the superior performance of the MLE: its spread is less than that of the RC estimator, and the proportion of samples greater than the upper limit of the parameter space was much smaller.

2.6.2. Performance of the confidence intervals

Since the sampling distributions of estimators in this model are often skewed, there is some concern about tests and confidence intervals based on approximate normality and estimated standard errors. Simulations were used here to compare confidence intervals based on the approximate normality of the regression calibration estimator (labeled as “RC-Wald” in the figures), bootstrap confidence intervals for the RC estimator (labeled as “RC-bootstrap” in the figures), and confidence intervals derived by inverting the likelihood ratio test (see Appendix G for details). Figure 2.7 shows that likelihood ratio-based 95% confidence intervals tend to be substantially narrower than the other two, except when the correlation between *X* and *W* was large and, therefore, the measurement error small.

In general, RC-Wald intervals are shorter than RC-bootstrap intervals, especially for small *m*. However, the lengths of both intervals are very variable for moderate and large measurement error and small sample sizes. The asymptotic variance of the sampling distribution of the RC estimator (eq 2.5) depends on $1/\hat{\alpha}_1^4$, and so can become very large for $\hat{\alpha}_1$ close to 0. The bootstrap intervals also perform

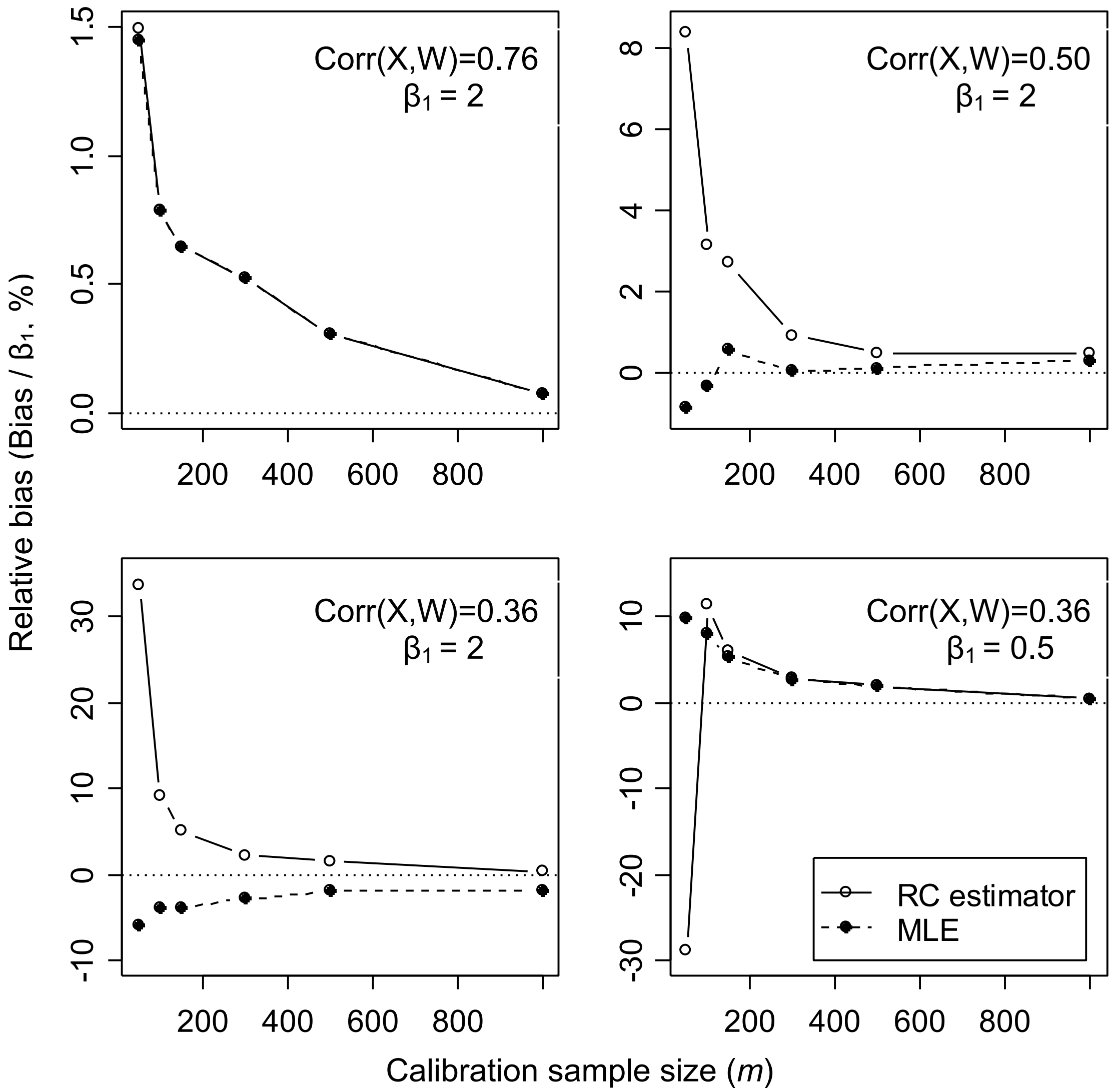

Figure 2.3. Relative bias of the RC and ML estimators of the slope in simple linear regression with measurement error. The distributions of $Y|X$ and $X|W$ are normal, the primary sample size is n = 1000 the calibration sample size is m. The results are based on 2500 simulated primary and calibration samples.

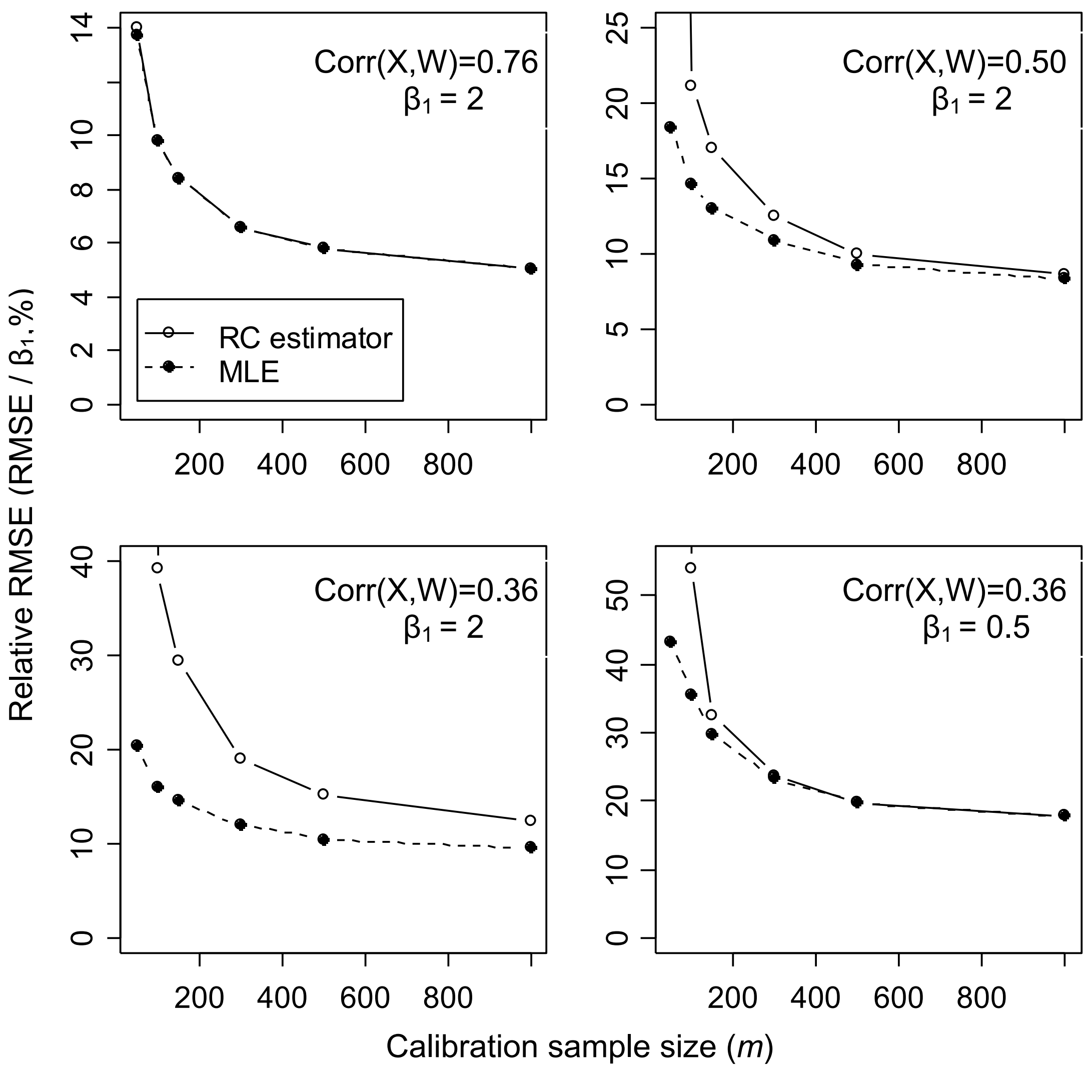

Figure 2.4. Relative root mean squared error of the RC and ML estimators of the slope in simple linear regression with measurement error. The distributions of $Y|X$ and $X|W$ are normal, the primary sample size is n = 1000 the calibration sample size is m. The results are based on 2500 simulated primary and calibration samples.

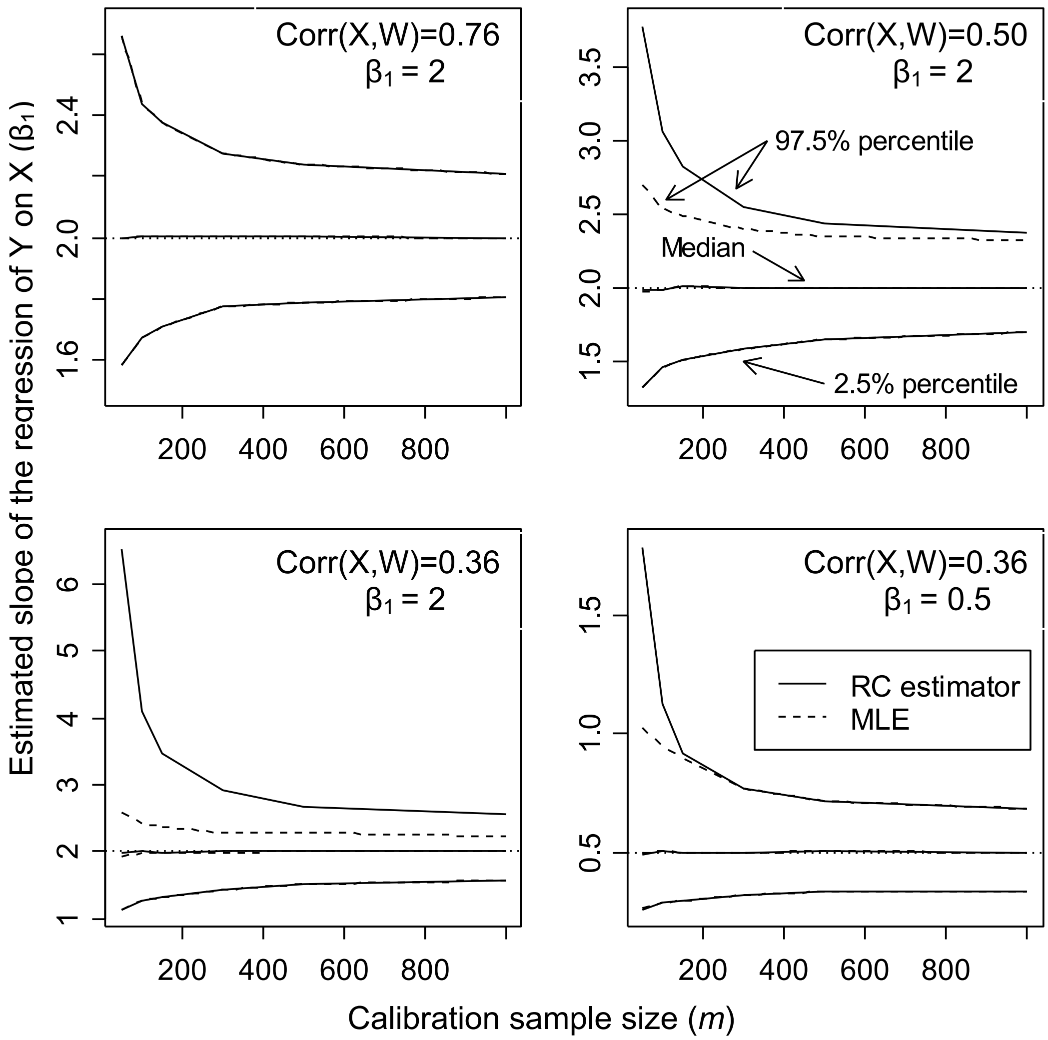

Figure 2.5. Median and 2.5% and 97.5% percentiles of the Monte Carlo sampling distribution of the RC and ML estimators of the slope in simple linear regression with measurement error. The distributions of $Y|X$ and $X|W$ are normal, the primary sample size is n = 1000 the calibration sample size is m. The results are based on 2500 simulated primary and calibration samples. The 2.5% percentile and median were almost identical for both estimators, only the 97.5 % percentile differed.

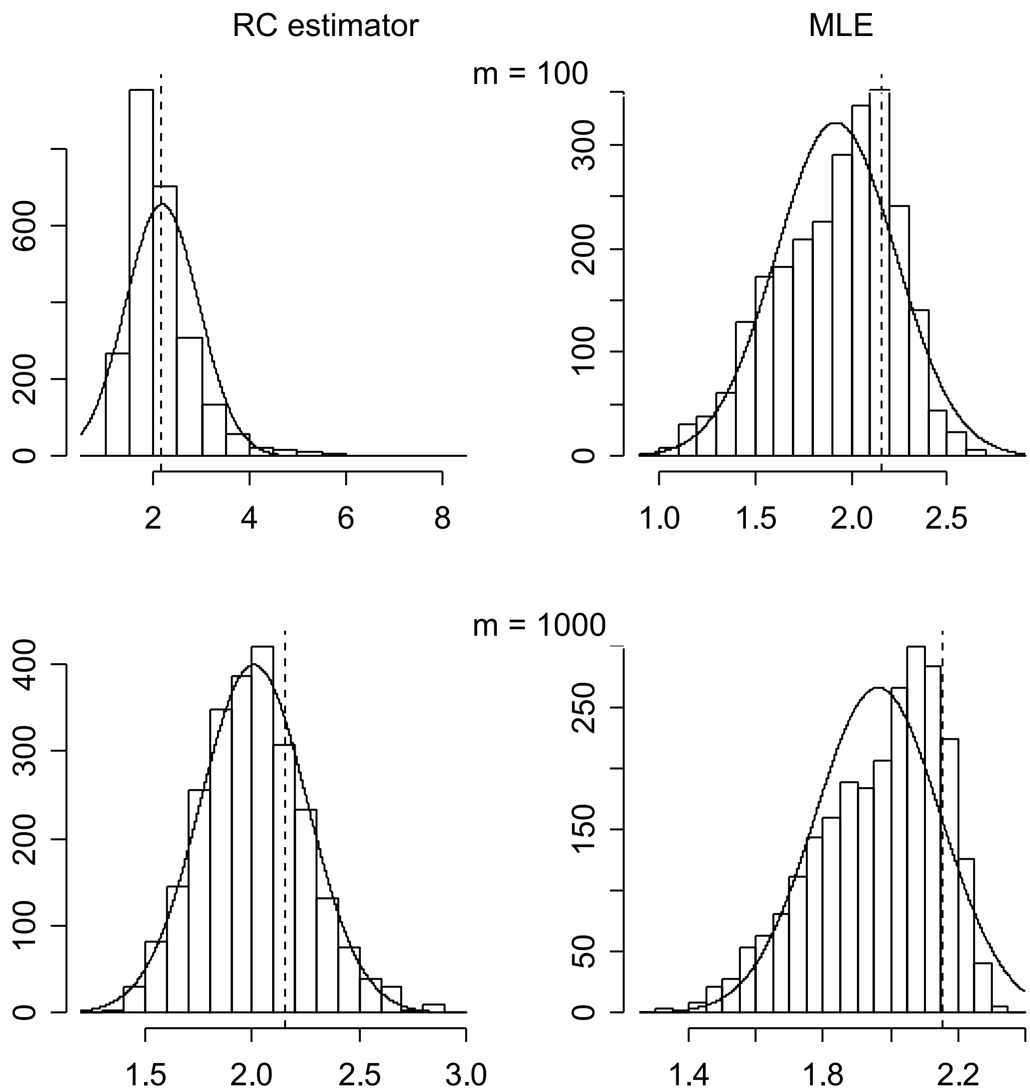

Figure 2.6. Histogram of the estimated sampling distribution of the RC and ML estimators of slope in simple linear regression with measurement error. The distributions of $Y|X$ and $X|W$ are normal, the primary sample size is n = 1000, *Cov(X,W)*=0.36, $\beta_1 = 2$, and the calibration sample size is m. The results are based on 2500 simulated primary and calibration samples. The solid line is the density of a normal distribution with the same moments as the estimated distribution. The vertical dashes line defines the upper limit of the parameter space of β_1.

poorly in that situation, because of the lack of moments of sampling distribution of the RC estimator. Surprisingly, the BCa bootstrap confidence intervals do not perform better than the much simpler intervals based on the percentiles on the bootstrapped distribution, even though the former is based on a higher order approximation to the true distribution function of the RC estimator (Tables G.7 and G.7) (Efron and Tibshirani 1993).

Of somewhat more concern, though, is the actual coverage rate and, in light of the skewness involved, the one-sided error rates on either side individually. Figures 2.8 and 2.9 show the non-coverage rates of the 95% confidence intervals on the upper and lower sides (corresponding to the separate upper and lower type I error rates).

The coverage of the LRT and RC-bootstrap intervals tend to be closer to the nominal coverage rates than that of the RC-Wald intervals. The error rate of the RC-Wald confidence interval is greater than the nominal 5% for sample sizes up to 300 when the correlation between *X* and *W* is low or moderate, while the RC-bootstrap and LRT intervals tend to be conservative for small sample sizes. As the calibration sample size increased, the three methods converge to the nominal error rate. The only exception is the error rate for the LRT interval for $\beta_1 = 2$ and *Corr(X, W)* = 0.36, which remained at approximately 3% even as *m* increased to 1000 and further (data not shown). In this scenario, the asymptotic distribution of the MLE is not approximately normal and, therefore, it is not surprising that the null distribution of the LRT statistic is not well approximated by a χ^2 distribution. However, even though the error rate of the LRT-based intervals is conservative, they were substantially shorter than intervals based on the RC estimator. At *m*=1000, the mean and median length of the intervals based on the RC estimator is approximately 40% greater than that of the LRT based interval. The difference is even greater for smaller calibration sample sizes (Tables G4-G8, appendix G).

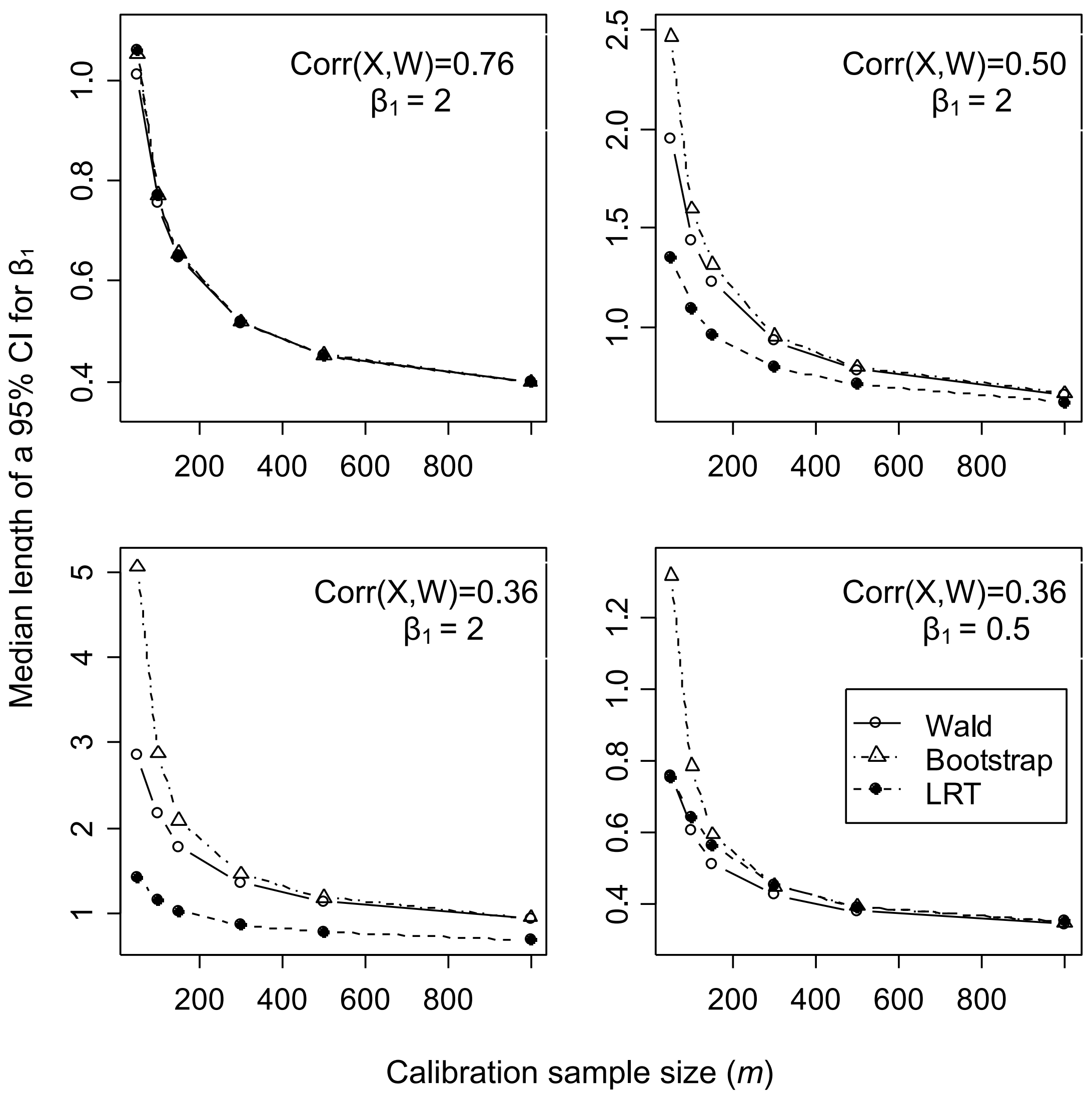

Figure 2.7. Median of the Monte Carlo distribution of length of the estimated 95% confidence interval of slope in simple linear regression with measurement error. The distributions of $Y|X$ and $X|W$ are normal, the primary sample size is n = 1000 the calibration sample size is m. The results are based on 2500 simulated primary and calibration samples. The intervals are based on a Wald approximation, bootstrap percentile, and inversion of the likelihood ratio test.

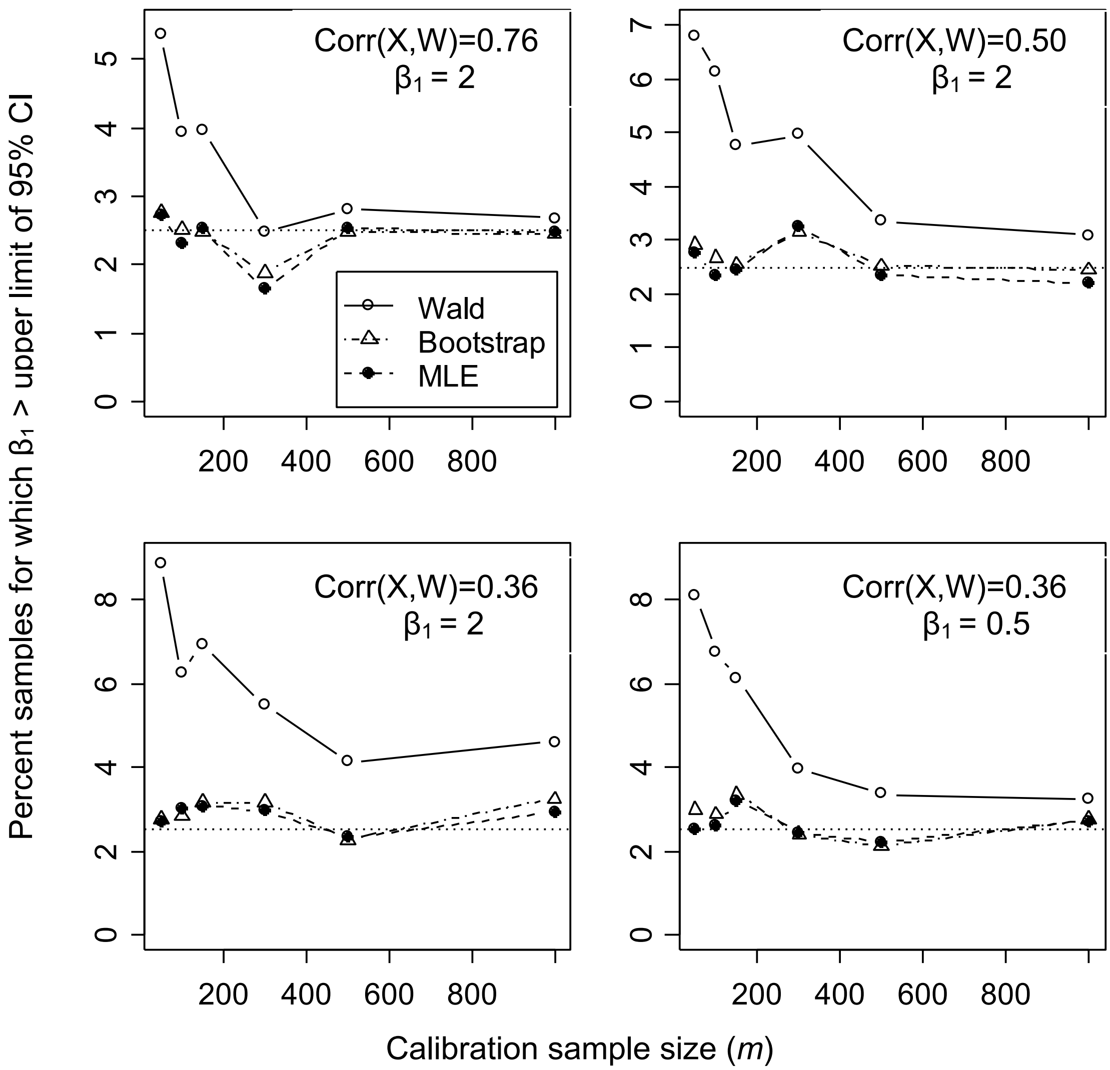

Figure 2.8. Percent of samples for which the true slope in simple linear regression with measurement error is greater than the upper limit of a 95% confidence interval. The distributions of $Y \mid X$ and $X \mid W$ are normal, with a primary sample size of n = 1000 and a calibration sample size of m. The results are based on 2500 simulated primary and calibration samples. The intervals are based on a Wald approximation, bootstrap percentile, and inversion of the likelihood ratio test.

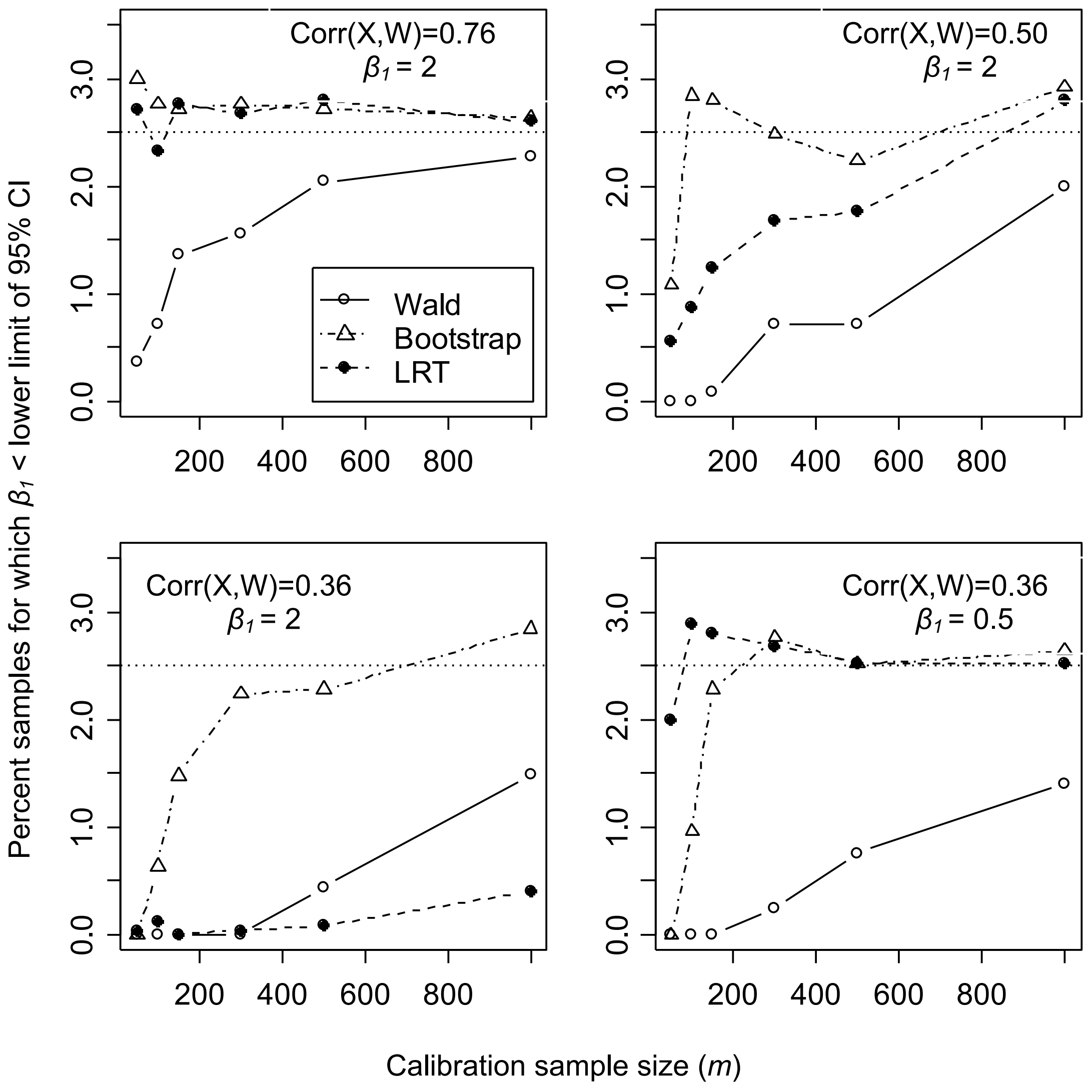

Figure 2.9. Percent of samples for which the true slope in simple linear regression with measurement error is less than the lower limit of a 95% confidence interval. The distributions of $Y \mid X$ and $X \mid W$ are normal, with a primary sample size of n = 1000 and a calibration sample size of m. The results are based on 2500 simulated primary and calibration samples. The intervals are based on a Wald approximation, bootstrap percentile, and inversion of the likelihood ratio test.

The most striking difference, however, is that most of the ‘misses’ of the RC-Wald intervals are in one direction, even in the best case scenario of small measurement error variance and large calibration sample size (Figs. 2.8 and 2.9). The true value of the parameter tends not to be included in the RC-Wald confidence interval when greater than the RC estimator. This may have important consequences in the interpretation of the results of epidemiological studies. For example, the calibration sample size for the Nurse’s Health Study was 173 nurses (Willett et al 1992). For m = 150, $\beta_1 = 2$ and a correlation between true and surrogate measurements of 0.75, these simulations show that the actual overall error rate of 95% CI is 5.4%, but the true parameter is greater than the upper bound of the interval 4.0% of the time, and smaller only 1.4% on the time. When the correlation is decreased to 0.50, those figures are 4.8% and 0.1%, respectively, and when the correlation is decreased further to 0.36, they are 6.9% and 0.0%. Therefore, for the settings considered here, there will be a tendency to report erroneous results when the estimated effects are less than the true effect. Intervals that allowed for asymmetric sampling distribution of the estimator, such as RC-bootstrap and LRT intervals, tend to have a more even rejection rate. In the case of the RC-bootstrap, the better performance in terms of symmetry and coverage came with decreased efficiency in terms of length, compared with the RC-Wald and, specially, the LRT interval.

2.7. Conclusions

Regression calibration is a transparent and straightforward method to estimate the parameters of regression models with measurement error in the explanatory variables when information about the measurement error process comes from an independent calibration study. In addition, if the distribution of $Y \mid X$ and $X \mid W$ are both normal, and the RC estimate of *var(Y|X)* is positive, RC and ML estimators are equivalent. However, if the RC estimate of *var(Y|X)* is negative, the RC estimator of the slope of the regression of Y on X can be very unstable and attain unreasonably large values, while the MLE is bounded and closer to the target parameter. Therefore, in practice, one may compute the RC estimator and verify that the estimate of *var(Y|X)*>0. If not, then one may switch to the MLE, which can be calculated with the formulae provided in this study or a maximization routine from standard software. Negative estimates of *var(Y|X)* are likely for small calibration sample size and large measurement error, but are also likely if β_1 is close, in magnitude, to $\sqrt{\sigma^2_{Y|W} / \sigma^2_{X|W}}$.

Even though the RC and ML estimator may be equivalent in some instances, likelihood ratio inferences can be substantially more accurate and powerful than those based on approximate normality and estimated standard errors. The sampling distribution of the RC estimator converges to a normal distribution centered on the target parameter but, for finite sample sizes, it can be very skewed. As a result, one-sided error rates of confidence intervals based on the asymptotic normality are inaccurate. Confidence intervals calculated from the percentiles of the bootstrapped sampling distribution of the RC estimator were more accurate than those based on approximate normality, in terms of symmetry and overall error rate, with only a relatively small increase in length.

Inverting a likelihood ratio test resulted in the shortest confidence intervals, while keeping an error rate close to the nominal rate. The efficiency gains where a consequence of the skewness of the sampling distribution of the RC estimator, and of the bound imposed by the constraint that the estimated *Var(Y | X)*>0. However, under this constraint and for some values of the parameters, the sampling distribution of the LRT statistic cannot be approximated by a χ^2 distribution, even for large sample sizes. This problem is apparent for small sample sizes in general, but becomes increasingly important when the true slope of the regression of *Y* on *X* is close, in magnitude, to the ratio of the standard deviations of *Y* and *X*, given the surrogate variable. Simulations indicated that, in those cases, LRT confidence intervals were conservative.

Although this study focuses on RC inference when both the distribution of *Y* given *X* and of *X* given its surrogate are normal, the findings apply more generally. The RC estimator can often be written as the ratio of the slope of the regression of *Y* on W to the slope of the regression of X on W. Whenever the estimators of those slopes are approximately normal, such as when the regression of interest is a generalized linear model, the results in this paper regarding the sampling distribution of the RC estimator and implications for inference remain approximately valid. In those cases, however, the RC and ML estimators may not be equivalent. Furthermore, since the superior efficiency of the MLE depends to a large degree on the positiveness of the estimated $Var(Y \mid X)$, it is not clear how its relative performance would be when the regression of Y on X is not normal.

There is a natural concern over the robustness of maximum likelihood inferences against possible distributional misspecifications. However, the existence of a calibration study in which both the true explanatory variable and its surrogate are observed should alleviate some of those concerns. While full likelihood analysis in measurement error problems typically requires the specification of three probability distributions (the response distribution, $Y \mid X$, the measurement error distribution, $W \mid X$, and the distribution of the true explanatory variable, X), when a calibration study is available, only the response distribution and the distribution of the explanatory variable, conditional on the surrogate, have to be specified. Furthermore, the adequacy of the assumed distributions can be assessed with the calibration data. Likelihood analysis can be challenging computationally for realistic distributional assumptions but, given the potential for substantial gains in efficiency and accuracy, we believe that likelihood inference should be considered for the scenarios discussed in this study.

2.8. References

Armstrong, B. 1985. Measurement error in generalized linear models. *Communications in Statistics Series B* 14: 529-544.

Athreya, K.B. 1987. Bootstrap of the mean in the infinite variance case. *Annals of Statistics* 15: 724-731.

Berkson, J. 1950. Are there two regressions? *Journal of the American Statistical Association* 45: 164-180.

Binham, S.A., Day, N.E., Luben, R., and others. 2003. Dietary fiber in food and protection against colorectal cancer in the European Prospective Investigation into Cancer and Nutrition. *Lancet* 361: 1496-1501.

Carroll, R.J., Ruppert, D., and Stefanski, L.A. 1995. *Measurement Error in Nonlinear Models*. Chapman & Hall, London.

Carroll, R.J., and Stefanski, L.A. 1990. Approximate quasi-likelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association* 85: 652-663.

Cochran, W.G. 1965. Errors of measurement in statistics. *Technometrics* 10: 637-665.

Efron, B. and Tibshirani, R.J. 1993. *An introduction to the bootstrap*. Chapman & Hall. New York. 436 p.

Fraser, G.E. and Stram, D.O. 2001. Regression calibration in studies with correlated variables measured with error. *American Journal of Epidemiology* 2001: 836-844.

Fuller, W.A. 1987. *Measurement Error Models*. John Wiley & Sons, New York.

Glesser, L.J. and Hwang, J.T. 1987. The nonexistence of $100(1-\alpha)\%$ confidence sets of finite expected diameter in errors-in-variables and related models. *Annals of Statistics* 15: 1351-1362,

Hinkley, D.V. 1969. On the ratio of two correlated normal random variables. *Biometrika* 56: 635-639.

Kuha, J. 1994. Corrections for exposure measurement error in logistic regression models with an application to nutritional data. *Statistics in Medicine* 13: 1135-1148.

Lehmann, E.L. and Casella, G. 1998. *Theory of Point Estimation*, 2nd Edition. Springer Verlag, New York.

Madansky, A. 1959. The fitting of straight lines when both variables are subject to error. *Journal of the American Statistical Association* 54: 173-205.

Marsaglia, G. 1965. Ratios of normal variables and ratios of sums of uniform variables. *Journal of the American Statistical Association* 60: 193-204.

Pepe, M. S., Self, S. G. and Prentice, R. L. 1989. Further results in covariate measurement errors in cohort studies with time to response data. *Statistics in Medicine* 8: 1167-1178.

Pierce, D.A., Stram, D.O., and Vaeth, M. 1990. Allowing for random errors in radiation dose estimates for the atomic bomb survivors. *Radiation Research* 123: 275-284.

Pierce, D.A., Stram, D.O., Vaeth, M., and Schafer, D.A. 1992. Some insights into the errors in variables problem provided by consideration of radiation dose-response analyses for the A-bomb survivors. *Journal of the American Statistical Association* 87: 351-359.

Prentice, R.L. 1982. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* 69: 331-342.

Rosner, B., Willet, W.C., and Spiegelman, D. 1989. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine* 8: 1051-1069.

Rosner, B., Spiegelman, D., and Willet, W.C. 1990. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *American Journal of Epidemiology* 132: 734-745.

Schafer, D.W. 2002. Likelihood analysis and flexible structural modeling for measurement error model regression. *Journal of Statistical Computing and Simulation* 72: 33-45.

Schafer, D. W., Lubin, J.H., Ron, E., Stovall, M., and Carroll, R.J. 2001. Thyroid cancer following scalp irradiation: a reanalysis accounting for uncertainty in dosimetry. *Biometrics* 57: 689-697.

Spiegelman, D., McDermott, A., and Rosner, B. 1997. Regression calibration method for correcting measurement error bias in nutritional epidemiology. *American Journal of Clinical Nutrition* 65 (suppl): 1179S-1186S.

Stram, D.O., and Kopecky, K.J. 2003. Power and uncertainty analysis of epidemiological studies of radiation-related disease risk in which dose estimates are based on a complex dosimetry system: some observations. *Radiation Research* 160: 408-417,

Stram, D.O., Langholz, B., Huberman, M., and Thomas, D.C. 1999. Correction for exposure measurement error in a reanalysis of lung cancer mortality for the Colorado Plateau uranium miners cohort. *Health Physics* 77: 265-275.

Thurston, S.W., Spiegelman, D, and Ruppert, D. 2003. Equivalence of regression calibration methods for external validation data. *Journal of Statistical Planning and Inference* 113: 527-539.

van Gils, C.H., Peeters, P.H.M., Bueno de Mesquita, H.B., and others. 2005. Consumption of vegetables and fruits and risk of breast cancer. *Journal of the American Medical Association* 293: 183-193.

Willett, W.C., Hunter, D.J., Stampfer, M.J., Colditz, G., Manson, J.E., Spiegelman, D., Rosner, B., Hennekens, C.H., and Speizer, F.E. 1992. Dietary fat and fiber in relation to breast cancer: an eight year follow up. *Journal of the American Medical Association* 268: 2037-2044.

Willett, W.C., Sampson, L., Browne, M.L., Stampfer, M.J., Rosner, B., Hennekens, C.H., and Speizer, F.E. 1988. The use of a self-administered questionnaire to assess diet four years in the past. *American Journal of Epidemiology* 127: 188-199.

3. REGRESSION CALIBRATION INFERENCE FOR MEASUREMENT ERROR MODELS WITH AN INTERNAL CALIBRATION STUDY

Vicente J. Monleon
Daniel W. Schafer

Department of Statistics
Oregon State University

3.1 Abstract

Regression calibration has emerged as a major tool to estimate regression models with measurement error in explanatory variables. The mismeasured explanatory variable is replaced by its conditional expectation, given a surrogate variable, in an estimation procedure that would have been used if the explanatory variable were known exactly. This study examines the effect of the uncertainty in the estimation of this conditional expectation on inference about regression parameters in the linear model, and the relative performance of regression calibration and likelihood inference, when the true explanatory variable is observed in a subset of the data. The estimator proposed by Spiegelman, Carroll and Kipnis (2001), defined as an inverse-variance weighted average of the estimator of the regression parameters obtained from the calibration data alone, and the regression calibration estimator obtained by treating the internal calibration data as an external calibration study, is almost as efficient as the maximum likelihood estimator (MLE). Other regression calibration estimators are less efficient, often substantially so, and can be less efficient than the estimator of the regression parameters obtained from the calibration data alone. The estimators are not affected by moderate departures from the assumption of normality of the regression of the true explanatory variable on its surrogate. For small sample sizes, inference based on the asymptotic normality of the sampling distribution of the regression calibration estimators can be improved by using the bootstrap or likelihood ratio tests and confidence intervals.

3.2. Introduction

Regression calibration (RC) has received increasing usage for problems that involve estimation of regression parameters in the presence of explanatory variable measurement errors, largely because of its simplicity, transparency and intuitive appeal. Suppose interest is in the regression of a response variable, *Y*, on an explanatory variable, *X*, which is observed only through an imprecise measurement or surrogate variable, *W*. The RC estimator uses whatever method of estimation would have been appropriate if *X* were observed exactly, but with the missing *X* replaced by $E(X|W)$ (Carroll et al. 1995, Chapter 3). Versions of the RC estimator were proposed by Prentice (1982) for Cox proportional hazards regression models, Armstrong (1985) for generalized linear models, and Rosner, Willet and Spiegelman (1989) for logistic regression.

Regression calibration has seen considerable use in nutritional and occupational epidemiology, where the exposure variables associated with a disease are difficult to measure precisely. For example, in a prospective study of the effect of fat intake on the risk of breast cancer, *X* was the long-term average intake of fat, assessed with a semi-quantitative food questionnaire administered to over 90,000 women (Willett et al. 1992). A validation study was conducted on 173 participants, who completed four, one-week diet records. The model for $E(X|W)$ was estimated from this subset with values of both *X* and *W*, and then used to estimate $E(X|W)$ for the 90,000 individuals in the primary data, for whom only *W* was available. As typical with this type of study, the correlation between nutrient intakes calculated from the questionnaire and the 'gold standard' was relatively low, ranging between 0.4 and 0.6 (Willett et al. 1988). Other examples include the effect of dietary fiber in the incidence of colon cancer (Binham et al. 2003), and of radiation exposure for the atomic bomb survivors (Pierce et al. 1990) or uranium miners (Stram et al. 1999).

Regression calibration emerges naturally when the regression of *Y* on *X* is linear, because if $E\left(Y \mid X\right) = \beta_0 + \beta_1 X$, then $E\left(Y \mid W\right) = E\left[E\left(Y \mid X\right) \mid W\right] =$

$\beta_0 + \beta_1 E\left(X \mid W\right)$. If $E(X|W)$ is known, usual regression tools for the regression of Y on $E(X|W)$ may be used to estimate β_0 and β_1, with appropriate attention to weights dictated by $Var(X|W)$. However, if the regression of Y on X is not linear, this form of RC is, in general, an approximation to the model of interest and the RC estimator is usually inconsistent. The conditions under which the approximation is almost exact depend on the particular model. When the degree of non-linearity in the regression of Y on X is large, several improvements to the simple RC estimator have been proposed, based on Taylor expansions and the assumption of small measurement error variance (Carroll and Stefanski 1990, Kuha 1994).

In some instances, if $E(X|W)$ is known, the RC and maximum likelihood (ML) estimators are the same or very close. For example, if the regressions of Y on X and of W on X are both linear, and $Y \mid X$, $W \mid X$ and X are all normally distributed, or if $Y \mid X$ is Bernoulli and the regression of Y on X is linear in X, then the RC and ML estimators are identical. They give approximately the same results for linear hazard regression models (Prentice 1982, Pepe et al. 1989, and Schafer et al. 2001).

These equivalences of RC and ML estimators, and the unbiasedness of the RC estimator when $E(Y|X)$ is linear in X, are only true if the conditional mean $E(X|W)$ is known, which is seldom the case. Problems with the simple version of RC when the degree of non-linearity in the regression of Y on X or the measurement error variance are large, may also be magnified if $E(X|W)$ is not known. While standard errors of estimated regression coefficients can be adjusted to account for the additional uncertainty in the estimate of $E(X|W)$, the effect on the estimator's properties has not received much attention. This paper focuses on the role of uncertainty in $E(X|W)$ in regression calibration inference.

Estimating measurement error models, whether using RC or other techniques, requires additional information about the measurement error process. It is well known that for the classical error model, if $Y \mid X$, $W \mid X$, and X are all normal, the parameters of the distribution of $Y \mid X$ are not identifiable without additional assumptions or data

(Fuller 1987). Although for other distributional assumptions the parameters of the simple linear regression model with measurement error are identifiable (Reiersol 1950), additional information is necessary to practically estimate the parameters of interest. This information can take several forms, including replicate measurements on some observations or a calibration study in which the true X is observed. In either case, the study can be an internal study, in which Y and either replicate measurements or the true X and W are observed in a subset of the data, or an external study, in which Y is not observed. An internal calibration dataset allows for increased precision, the assessment of both the model relating Y to X and the error structure and eliminates concerns about transportability of models for $E(X|W)$.

Regression calibration is often easy to implement and transparent, and relies on minimal assumptions on the distribution of the explanatory variables. Alternatives such as maximum likelihood, on the other hand, usually require difficult computations and stronger assumptions, although several approaches for flexible structural and semiparametric modeling have been developed (Roeder et al. 1996; Carroll et al. 1999; Schafer 2001, 2002). However, ML may efficiently and automatically combine the information from several sources, such as primary and calibration studies. Because RC is becoming increasingly popular, we believe it is appropriate to critically examine its properties and performance, relative to likelihood-based approaches, in light of the uncertainty in estimating $E(X|W)$. We are interested in complex models for $E(Y|X)$, but in this study we consider linear regression of Y on X, with a model for $E(X|W)$ estimated from an internal calibration data set. The term "internal calibration dataset" establishes an analogy with the closely-related data structure in which there is a primary dataset (with observations on Y and W) and an independent "external calibration dataset" (with observations on X and W). In fact, a more sensible terminology for the problem of interest here specifies a single dataset of observations Y and W, with true X available in a subset. Nevertheless, we will use the terminology "internal calibration" to be consistent with recent studies on this topic.

This paper is organized as follows. Section 3.3 describes the model and reviews alternative regression calibration estimators and the methods typically used to estimate their standard error for approximate tests and confidence intervals. Section 3.4 discusses likelihood-based inference and compares the asymptotic efficiency of the RC and ML estimators. Section 3.5 presents the results of a set of simulations devised to explore the effect of the uncertainty in *E(X | W)* on the performance of the RC and ML estimators and associated confidence intervals, and the robustness of RC and ML inference to departures from the specified assumptions. Section 3.6 summarizes the main conclusions of this report.

3.3. The regression calibration estimator for linear regression

Consider the model:

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{3.1}$$

$$X = \alpha_0 + \alpha_1 W + \delta \tag{3.2}$$

where *Y* is the response variable; *X* is the explanatory variable; *W* is a surrogate for *X*; $(\beta_0, \beta_1, \alpha_0, \alpha_1)$ are unknown regression coefficients; and (ε, δ) are random variables.

Suppose that (1) there is a primary sample consisting of observations (y_i, w_i), $i = 1, \ldots, n$, and an internal calibration sample consisting of observations (y_i, x_i, w_i), $i = n+1, \ldots, n+m$; (2) (ε, δ) are independent random errors with means equal to 0 and variances $\sigma^2_{Y|X}$ and $\sigma^2_{X|W}$, respectively; (3) random variables with different values of *i* are independent; and (4) the error structure is non-differential, meaning *f(Y | X, W) = f(Y | X)*. It follows that:

$$Y = (\beta_0 + \alpha_0 \beta_1) + \alpha_1 \beta_1 W + (\varepsilon + \beta_1 \delta) \tag{3.3}$$

Let $\gamma_0 = \beta_0 + \alpha_0\beta_1$ and $\gamma_1 = \alpha_1\beta_1$ be the coefficients of the regression of *Y* on *W*, and $\sigma^2_{Y|W} = \sigma^2_{Y|X} + \beta_1^2 \sigma^2_{X|W}$.

There may be additional explanatory variables Z, free of measurement error. Then all expectations and probability densities should also be conditional on Z, but that notation will be suppressed.

There is not a universally accepted approach for defining the RC estimator for internal calibration data. Several estimators have been proposed (Rosner et al. 1989, 1990; Carroll et al. 1995; Spiegelman et al. 2001), but all of them are weighted averages of the estimators of $\boldsymbol{\beta}$ obtained from the calibration data alone ($\hat{\boldsymbol{\beta}}_{\mathbf{INT}}$), and the RC estimator obtained by treating the internal calibration data as an external calibration study ($\hat{\boldsymbol{\beta}}_{\mathbf{EXT}}$) (Thurston et al. 2005). Neither $\hat{\boldsymbol{\beta}}_{\mathbf{INT}}$ nor $\hat{\boldsymbol{\beta}}_{\mathbf{EXT}}$ makes full use of the available information, but both are consistent, the latter as $n \to \infty$ and $n/m \to k$, k a positive constant (Carroll and Stefanski 1990). Therefore, as long as all the elements are consistent, the resulting weighted average will also be consistent and could be more efficient than either $\hat{\boldsymbol{\beta}}_{\mathbf{INT}}$ or $\hat{\boldsymbol{\beta}}_{\mathbf{EXT}}$.

The estimators that will be considered here are the 'as external' estimator, $\hat{\boldsymbol{\beta}}_{\mathbf{EXT}}$, proposed by Rosner, Willet and Spiegelman (1989); the estimator proposed by Carroll, Ruppert and Stefanski (1995), denoted by $\hat{\boldsymbol{\beta}}_{\mathbf{CRS}}$; and the estimator proposed by Spiegelman, Carroll and Kipnis (2001), denoted by $\hat{\boldsymbol{\beta}}_{\mathbf{SCK}}$. In what follows, let $\mathbf{X_c}$, $\mathbf{W_c}$ and $\mathbf{W_p}$ be the design matrices for the calibration and primary data, respectively, $\mathbf{A} = \begin{bmatrix} 1 & \alpha_0 \\ 0 & \alpha_1 \end{bmatrix}$, and $\hat{\mathbf{A}}$ the matrix $\mathbf{A}$ with the unknown parameters substituted by their estimators. The estimator of $\boldsymbol{\beta}$ based on the calibration data alone is $\hat{\boldsymbol{\beta}}_{\mathbf{INT}} = \left(\mathbf{X}_\mathbf{c}'\mathbf{X}_\mathbf{c}\right)^{\mathbf{-1}} \mathbf{X}_\mathbf{c}'\mathbf{y}_\mathbf{c}$, and $Var\left(\hat{\boldsymbol{\beta}}_{\mathbf{INT}}\right) = \sigma^2_{Y|X} \left(\mathbf{X}_\mathbf{c}'\mathbf{X}_\mathbf{c}\right)^{\mathbf{-1}}$.

<u>'As external' estimator, $\hat{\boldsymbol{\beta}}_{\mathbf{EXT}}$</u>

The 'as external' estimator is obtained by first estimating (α_0, α_1) by $(\hat{\alpha}_0, \hat{\alpha}_1)$ from the calibration sample, calculating $\hat{x}_i = \hat{\alpha}_0 + \hat{\alpha}_1 w_i$ for each observation in the primary sample, and estimating the slope of the regression of y_i on $\hat{x}_i$ using least squares. The design matrix for this model can by written as $\hat{\mathbf{X}} = \mathbf{W_p A}$, so that:

$$\hat{\boldsymbol{\beta}}_{\mathbf{EXT}} = \hat{\mathbf{A}}^{-1}\left(\mathbf{W}_{\mathbf{p}}'\mathbf{W}_{\mathbf{p}}\right)^{-1}\mathbf{W}_{\mathbf{p}}'\mathbf{y}_{\mathbf{p}} = \hat{\mathbf{A}}^{-1}\hat{\boldsymbol{\gamma}} \tag{3.4}$$

where $\hat{\boldsymbol{\gamma}}$ is the least squares estimators of the slope of *Y* on *W*, based on the primary data (Rosner et al. 1989, Carroll et al. 1995, Thurston et al. 2003, Chapter 2 of this Dissertation). Using the delta method, the variance of $\hat{\boldsymbol{\beta}}_{\mathbf{EXT}}$ can be approximated by

$$Var\left(\hat{\boldsymbol{\beta}}_{\mathbf{EXT}}\right) \approx \mathbf{A}^{-1}\left[\beta_{1,EXT}^2\sigma_{X|W}^2\left(\mathbf{W}_{\mathbf{c}}'\mathbf{W}_{\mathbf{c}}\right)^{-1} + \sigma_{Y|W}^2\left(\mathbf{W}_{\mathbf{p}}'\mathbf{W}_{\mathbf{p}}\right)^{-1}\right]\mathbf{A}^{-T} \tag{3.5}$$

In epidemiology, this estimator has been used mostly for logistic models for binary health responses, when the calibration sample is relative small and the disease rare (for example, Willett et al. 1992). Since there are only a few health events in the calibration sample, the efficiency gains from incorporating information provided by $\hat{\boldsymbol{\beta}}_{\mathbf{INT}}$ may be negligible. However, Spiegelman et al. (1997) also advocated the use of this estimator for linear and failure time regression.

The moments of the sampling distribution of $\hat{\beta}_{1,EXT}$ and of its estimated variance are not defined (Chapter 2, this Dissertation), because the sampling distribution of $\hat{\alpha}_1$ has positive mass at 0. For small calibration sample sizes and large measurement error variance, this results in an erratic behavior of $\hat{\beta}_{1,EXT}$ and associated confidence intervals. However, as the calibration sample size increases, the sampling distribution of $\hat{\beta}_{1,EXT}$ converges to a normal centered on the target parameter.

<u>Carroll, Ruppert and Stefanski (CRS) estimator</u>, $\hat{\boldsymbol{\beta}}_{\mathbf{CRS}}$

This estimator is obtained by fitting a single regression of *Y* to the true *X* when available, and to the estimated *E(X|W)* otherwise. The parameters of the regression of *X* on *W* are estimated from the calibration data. Carroll et al. (1995) also suggest including a dummy variable indicating whether *X* is observed or not. The model under consideration is:

$$E\begin{bmatrix}\mathbf{y_p}\\ \mathbf{y_c}\end{bmatrix}=\begin{bmatrix}\mathbf{W_p}\hat{\mathbf{A}}\\ \mathbf{X_c}\end{bmatrix}\boldsymbol{\beta}_{\mathbf{CRS}},$$

Therefore,

$$\begin{gathered}\hat{\boldsymbol{\beta}}_{\mathbf{CRS}}=\left(\hat{\mathbf{A}}'\mathbf{W}_\mathbf{p}'\mathbf{W}_\mathbf{p}\hat{\mathbf{A}}+\mathbf{X}_\mathbf{c}'\mathbf{X}_\mathbf{c}\right)^{-1}\left(\hat{\mathbf{A}}'\mathbf{W}_\mathbf{p}'\mathbf{y}_\mathbf{p}+\mathbf{X}_\mathbf{c}'\mathbf{y}_\mathbf{c}\right)=\\ \left(\hat{\mathbf{A}}'\mathbf{W}_\mathbf{p}'\mathbf{W}_\mathbf{p}\hat{\mathbf{A}}+\mathbf{X}_\mathbf{c}'\mathbf{X}_\mathbf{c}\right)^{-1}\left(\hat{\mathbf{A}}'\mathbf{W}_\mathbf{p}'\mathbf{W}_\mathbf{p}\hat{\mathbf{A}}\hat{\boldsymbol{\beta}}_{\mathbf{EXT}}+\mathbf{X}_\mathbf{c}'\mathbf{X}_\mathbf{c}\hat{\boldsymbol{\beta}}_{\mathbf{INT}}\right)\end{gathered}\tag{3.6}$$

The asymptotic variance of this estimator is (Thurston et al. 2005):

$$\begin{gathered}Var\left(\hat{\boldsymbol{\beta}}\right)=\left(\mathbf{A}'\mathbf{W}_\mathbf{p}'\mathbf{W}_\mathbf{p}\mathbf{A}+\mathbf{X}_\mathbf{c}'\mathbf{X}_\mathbf{c}\right)^{-1}\times\\ \left[\left(\mathbf{A}'\mathbf{W}_\mathbf{p}'\mathbf{W}_\mathbf{p}\mathbf{A}\right)Var\left(\hat{\boldsymbol{\beta}}_{\mathbf{EXT}}\right)\left(\mathbf{A}'\mathbf{W}_\mathbf{p}'\mathbf{W}_\mathbf{p}\mathbf{A}\right)+\left(\mathbf{X}_\mathbf{c}'\mathbf{X}_\mathbf{c}\right)Var\left(\hat{\boldsymbol{\beta}}_{\mathbf{INT}}\right)\left(\mathbf{X}_\mathbf{c}'\mathbf{X}_\mathbf{c}\right)\right]\times\\ \left(\mathbf{A}'\mathbf{W}_\mathbf{p}'\mathbf{W}_\mathbf{p}\mathbf{A}+\mathbf{X}_\mathbf{c}'\mathbf{X}_\mathbf{c}\right)^{-1}\end{gathered}\tag{3.7}$$

<u>Spiegelman, Carroll and Kipnis (SCK) estimator</u>, $\hat{\boldsymbol{\beta}}_{\mathbf{SCK}}$.

This estimator is an inverse-variance weighted average of $\hat{\boldsymbol{\beta}}_{\mathbf{INT}}$ and $\hat{\boldsymbol{\beta}}_{\mathbf{EXT}}$:

$$\hat{\boldsymbol{\beta}}_{SCK}=\left[\hat{V}ar\left(\hat{\boldsymbol{\beta}}_{\mathbf{EXT}}\right)^{-1}+\hat{V}ar\left(\hat{\boldsymbol{\beta}}_{\mathbf{INT}}\right)^{-1}\right]^{-1}\left[\hat{V}ar\left(\hat{\boldsymbol{\beta}}_{\mathbf{EXT}}\right)^{-1}\hat{\boldsymbol{\beta}}_{\mathbf{EXT}}+\hat{V}ar\left(\hat{\boldsymbol{\beta}}_{\mathbf{INT}}\right)^{-1}\hat{\boldsymbol{\beta}}_{\mathbf{INT}}\right]\tag{3.8}$$

Spiegelman et al. (2001) note that, as long as the weights are estimated accurately, this choice of weights gives the estimator with the minimum asymptotic variance among all unbiased linear combinations of $\hat{\boldsymbol{\beta}}_{\mathbf{INT}}$ and $\hat{\boldsymbol{\beta}}_{\mathbf{EXT}}$. The approximate asymptotic variance is (Spiegelman et al. 2001):

$$\begin{gathered}Var\left(\hat{\boldsymbol{\beta}}_{SCK}\right)=\left[Var\left(\hat{\boldsymbol{\beta}}_{\mathbf{EXT}}\right)^{-1}+Var\left(\hat{\boldsymbol{\beta}}_{\mathbf{INT}}\right)^{-1}\right]^{-1}=\\ \left\{\mathbf{A}'\left[\beta_{1,EXT}^2\sigma_{X|W}^2\left(\mathbf{W}_\mathbf{c}'\mathbf{W}_\mathbf{c}\right)^{\mathbf{-1}}+\sigma_{Y|W}^2\left(\mathbf{W}_\mathbf{p}'\mathbf{W}_\mathbf{p}\right)^{\mathbf{-1}}\right]^{-1}\mathbf{A}+\sigma_{Y|X}^{-2}\left(\mathbf{X}_\mathbf{c}'\mathbf{X}_\mathbf{c}\right)\right\}^{-1}\end{gathered}\tag{3.9}$$

Focusing only on the slope β_1, the SCK estimator and its estimated variance are:

$$\hat{\beta}_{1,SCK} = \frac{\hat{\alpha}_1^3 \hat{\sigma}^2_{\hat{\beta}_{1,INT}}}{\hat{\gamma}_1^2 \hat{\sigma}^2_{\hat{\alpha}_1} + \hat{\alpha}_1^2 \hat{\sigma}^2_{\hat{\gamma}_1} + \hat{\alpha}_1^4 \hat{\sigma}^2_{\hat{\beta}_{1,INT}}} \hat{\gamma}_1 + \frac{\hat{\gamma}_1^2 \hat{\sigma}^2_{\hat{\alpha}_1} + \hat{\alpha}_1^2 \hat{\sigma}^2_{\hat{\gamma}_1}}{\hat{\gamma}_1^2 \hat{\sigma}^2_{\hat{\alpha}_1} + \hat{\alpha}_1^2 \hat{\sigma}^2_{\hat{\gamma}_1} + \hat{\alpha}_1^4 \hat{\sigma}^2_{\hat{\beta}_{1,INT}}} \hat{\beta}_{1,INT} \tag{3.10}$$

$$\hat{V}ar\left(\hat{\beta}_{1,SCK}\right) = \frac{\hat{\sigma}^2_{\hat{\beta}_{1,INT}} \left(\hat{\gamma}_1^2 \hat{\sigma}^2_{\hat{\alpha}_1} + \hat{\alpha}_1^2 \hat{\sigma}^2_{\hat{\gamma}_1}\right)}{\hat{\gamma}_1^2 \hat{\sigma}^2_{\hat{\alpha}_1} + \hat{\alpha}_1^2 \hat{\sigma}^2_{\hat{\gamma}_1} + \hat{\alpha}_1^4 \hat{\sigma}^2_{\hat{\beta}_{1,INT}}} \tag{3.11}$$

where $\hat{\sigma}^2_{\hat{\beta}_{1,INT}}$, $\hat{\sigma}^2_{\hat{\alpha}_1}$ and $\hat{\sigma}^2_{\hat{\gamma}_1}$ are the estimated variances of $\hat{\beta}_{1,INT}$, $\hat{\alpha}_1$ and $\hat{\gamma}_1$, respectively. Note that as $\hat{\alpha}_1 \to 0$, $\hat{\beta}_{1,SCK}$ converges to $\hat{\beta}_{1,INT}$. The moments of the sampling distribution of $\hat{\beta}_{1,SCK}$ may not be defined, however, because the joint sampling distribution $\hat{\alpha}_1$ and $\hat{\gamma}_1$ has positive mass at (0, 0). In practice, it is unlikely that both $\hat{\alpha}_1$ and $\hat{\gamma}_1$ become close to 0, unless γ_1 and, therefore, β_1 is close to 0.

Regression calibration test and confidence intervals are usually based on the asymptotic normality of the sampling distribution of the estimators, with the unknown parameters of their asymptotic variances substituted by their estimators. Bootstrap inferences can also be used (Carroll et al. 1995).

Thurston et al. (2005) compared the asymptotic efficiency of the RC estimators when the regressions of *Y* on *X* and *X* on *W* are linear. Asymptotically, they found that the SCK estimator is uniformly more efficient than the 'as external' and CRS estimators. The asymptotic standard errors of the three estimators are approximately the same when the correlation between *X* and *W* is close to 1, otherwise $\hat{\beta}_{1,EXT}$ is less efficient than $\hat{\beta}_{1,SCK}$ and $\hat{\beta}_{1,CRS}$. When the correlation between *Y* and *X* is low and the proportion of data in the validation study is large, the standard error of $\hat{\beta}_{1,CRS}$ was approximately the same as that of $\hat{\beta}_{1,SCK}$. Thurston et al. (2005) do not study the finite sample performance of the alternative RC estimators, or compare their performance to the ML or internal estimators.

Spiegelman et al. (2000, 2001) compared the bias, mean squared error, size, power and coverage probabilities of $\hat{\beta}_{1,SCK}$, $\hat{\beta}_{1,EXT}$ and the MLE for a logistic regression with covariate misclassification and measurement error, using simulation. The MLE was generally superior to the RC estimators for all the examined criteria, but $\hat{\beta}_{1,SCK}$ was a clear improvement over $\hat{\beta}_{1,EXT}$. The SCK estimator was nearly as efficient as the MLE for the largest validation sample size that they examined (346 subjects). In their simulations, likelihood inference relied upon the asymptotic normality of the sampling distribution of the MLE, with the asymptotic standard error estimated from the empirical information matrix. If the profile likelihood is asymmetric, inference based on the likelihood ratio may perform better, in terms of efficiency and coverage, than Wald-based inference.

3.4. Regression calibration and maximum likelihood in the normal-normal model

3.4.1. Maximum likelihood in the normal-normal model

The joint density function of *Y*, observed in the primary study, and *Y* and *X*, observed in an internal calibration dataset, conditional on the observed values of *W,* can be decomposed as:

$$f\left(\mathbf{y},\mathbf{x}\,|\,\mathbf{w}\right)=\prod_{i=1}^{n} f\left(y_i\,|\,w_i\right)\prod_{i=n+1}^{n+m} f\left(y_i\,|\,x_i\right) f\left(x_i\,|\,w_i\right)$$

Typically, $f\left(Y\,|\,W\right)$ cannot be obtained analytically, so numerical integration may be necessary. However, if the conditions stated in section 3.3 hold and, in addition, $\left(\varepsilon,\delta\right)$ from equations 3.1 and 3.2 are normally distributed, the log likelihood function is:

$$l\left(\boldsymbol{\beta},\boldsymbol{\alpha},\sigma_{Y|X}^2,\sigma_{X|W}^2\right)=-\frac{n}{2}\log\left(\sigma_{Y|X}^2+\beta_1^2\sigma_{X|W}^2\right)-\frac{m}{2}\log\left(\sigma_{Y|X}^2\right)-\frac{m}{2}\log\left(\sigma_{X|W}^2\right)$$
$$-\frac{1}{2\left(\sigma_{Y|X}^2+\beta_1^2\sigma_{X|W}^2\right)}\sum_{i=1}^{n}\left(y_{p,i}-\beta_0-\beta_1\alpha_0-\beta_1\alpha_1 w_i\right)^2 \tag{3.12}$$
$$-\frac{1}{2\sigma_{Y|X}^2}\sum_{i=n+1}^{n+m}\left(y_i-\beta_0-\beta_1 x_i\right)^2-\frac{1}{2\sigma_{X|W}^2}\sum_{i=n+1}^{n+m}\left(x_i-\alpha_0-\alpha_1 w_i\right)^2$$

There is not a closed form solution for the MLE, but equation (3.12) can be easily maximized using standard software (such as the S-Plus 'nlmin' or R 'nlm' functions).

When inference is based on an external, rather than internal, calibration study, the estimator of $\sigma_{Y|X}^2$ obtained by maximizing the likelihood function can be negative. If $\hat{\sigma}_{Y|X}^2 > 0$, the RC and ML estimators are equivalent. Otherwise, imposing the constraint that $\hat{\sigma}_{Y|X}^2 > 0$ induces a bound on the MLE, while the RC estimator can behave erratically (Chapter 2, this Dissertation). As a result, the MLE is more efficient than the RC estimator but, for some values of the parameters, the null sampling distribution of the likelihood ratio test statistic cannot be approximated by a χ^2 distribution, even for large sample sizes. When an internal calibration dataset is available, however, the estimator of $\sigma_{Y|X}^2$ obtained by maximizing equation (3.12) is always positive. Therefore, neither the bound, nor the problems with inference associated with setting $\sigma_{Y|X}^2$ to 0 would be present in this case.

The profile log likelihood function of β_1 can be obtained by evaluating the log likelihood at the conditional MLE of $\left(\alpha_0,\alpha_1,\beta_0,\sigma_{Y|X}^2,\sigma_{X|W}^2\right)$ for fixed β_1. Figure 3.1 shows some examples of the profile likelihood for simulated data. Although the profile likelihood can be bimodal for small sample sizes, it becomes unimodal as the calibration sample size is increased. Even for relatively large sample sizes, however, the profile likelihood can be relatively asymmetric around the MLE, suggesting that inference based on the asymptotic normality of the MLE may not be adequate

(Fig.3.1). For very small sample sizes, a confidence interval obtained by the set of values of β_1 'not rejected' by a likelihood ratio test may include two disjoint intervals.

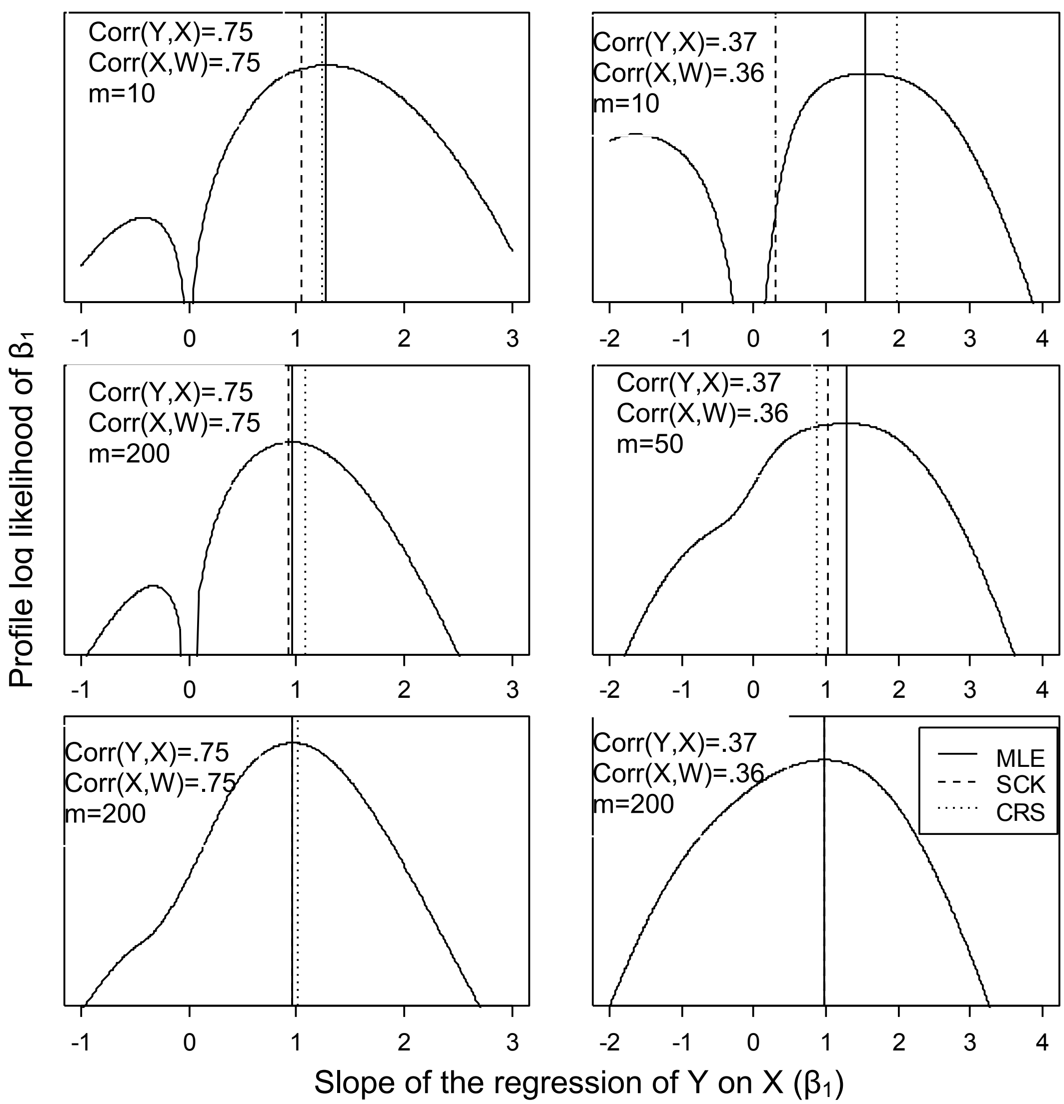

Figure 3.1. Log profile likelihood of β_1 for simulated data with different calibration sample sizes (m) and correlations between Y and X, and X and W. In this example, the true value of β_1 is 1. The vertical lines indicate the location of the ML, SCK and CRS estimators.

In simulations, this occurred about 13% of the time when the correlation between *Y* and *X* and *X* and *W* were both low, and the calibration sample consisted of 10 observations. As the sample size increased to 25 observations, or the correlation between *Y* and *X* to 0.75, the proportion of simulated samples with disjoint intervals decreased to approximately 1%, and it was 0 for all other scenarios.

3.4.1. Asymptotic relative efficiency of the regression calibration and maximum likelihood estimators

Thurston et al. (2005) provide formulae for the asymptotic variance of the 'as external', CRS and SCK estimators of β_1 (Appendix H). The asymptotic variances are functions of the correlation between *Y* and *X*, ρ_{YX}, between *X* and *W*, ρ_{XW}, and the ratio of the sample sizes of the calibration and primary datasets, scaled $Var(Y)/\left[nVar(X)\right]$. They showed that the 'as external' estimator is uniformly less efficient that the CRS and SCK, and therefore will not be considered here.

When internal calibration data is available, it is always possible to estimate the coefficients of the regression of *Y* on *X* with the calibration data alone, but the primary data can contribute additional information and result in increased efficiency. Therefore, $\hat{\beta}_{1,INT}$ provides a useful baseline to compare the relative efficiency of the other estimators, and to judge the gains incurred by adding the primary study data. The variance of $\hat{\beta}_{1,INT}$ can be written as:

$$Var\left(\hat{\beta}_{1,INT}\right) = \frac{\sigma_Y^2}{n\sigma_X^2}\left[\frac{n}{m}\left(1-\rho_{YX}^2\right)\right]$$

The asymptotic standard error of the MLE can be approximated through a Monte Carlo simulation, setting a very large *m* and *n*. Figure 3.2 shows the asymptotic standard deviation of the CRS, SCK, and the estimated standard error of the MLE, relative to the standard deviation of $\hat{\beta}_{1,INT}$, when m / n = 1/5. Asymptotically, the MLE is more efficient than any of the RC estimators. As the measurement error decreases, the efficiency of the CRS and SCK estimators approaches that of the MLE. The most surprising result, though, is that the CRS estimator can be less efficient than the estimator calculated from the internal validation alone, ignoring the information from the primary study altogether, if the correlation between Y and X is medium to high.

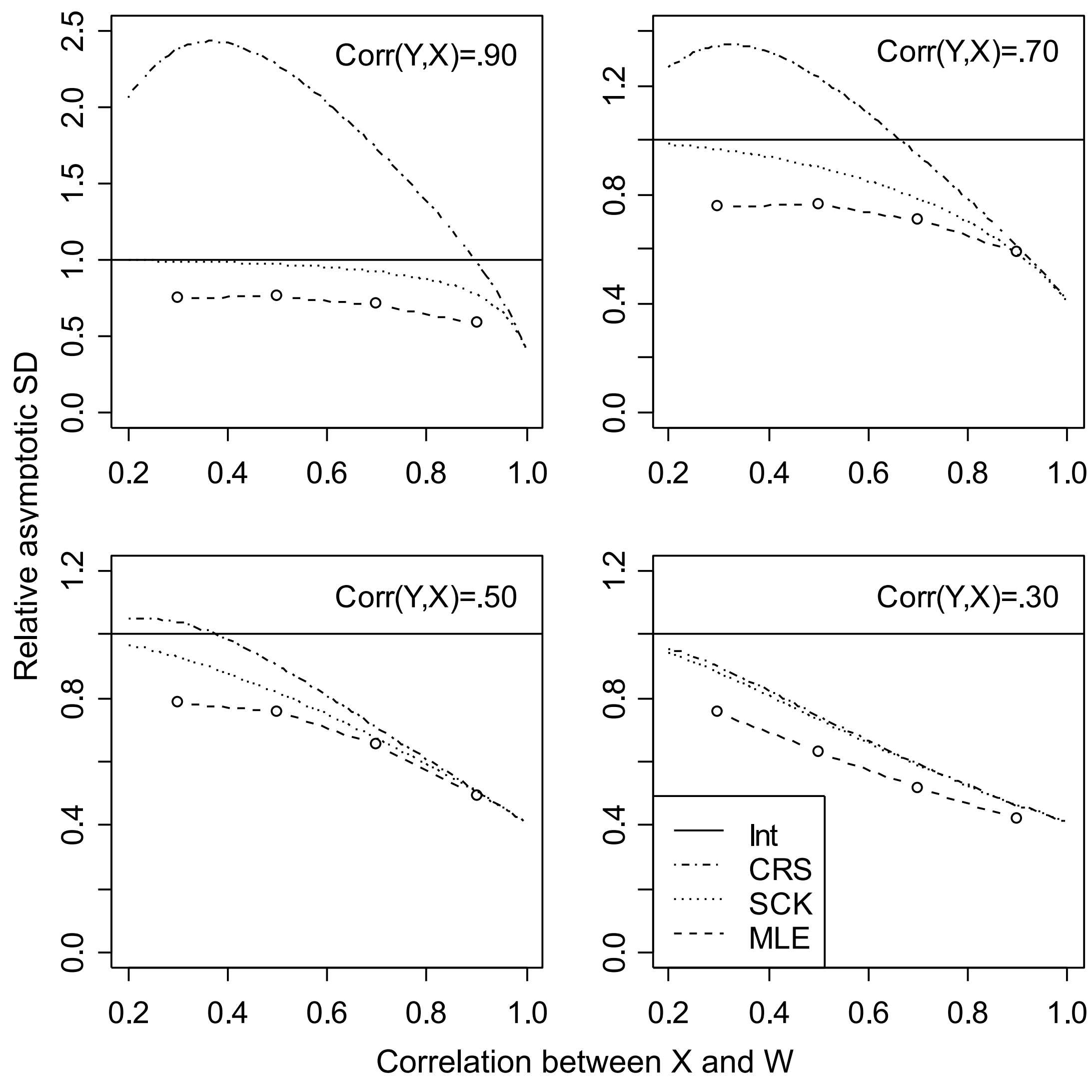

Figure 3.2. Asymptotic standard deviations of the CRS, SCK estimators, and standard errors of the MLE, relative to the standard deviation of the estimator based on the calibration data alone. The ratio of the calibration to primary study sample size is 1/5. The standard error of the MLE is estimated from a Monte Carlo simulation with a calibration sample size of 5000.

3.5. Simulations

3.5.1. Study design

The goal of this simulation is to investigate how the properties of the estimators and confidence intervals varied with the sample size of the calibration study and strength of the association between *Y* and *X*, and *X* and *W*, measured by their correlations. Since all estimators incorporate information from the calibration and primary data, we designed a complete factorial study in which the correlation between *Y* and *X*, and between *X* and *W* was approximately 0.36 or 0.75. These values correspond to the low and high ends of the correlations between the gold standard and surrogate observed in studies of association between diet and disease (Willet et al. 1988). The calibration sample size ranged between 10 and 500 observations. The values of the parameters used in the simulations are shown in Table 3.1.

For each scenario, we calculated the internal, 'as external', CRS, SCK and ML estimators. We also estimated confidence intervals based on the asymptotic normality and approximate asymptotic standard errors of the sampling distribution of the various RC estimators (denoted by 'Wald'), on the percentiles of the bootstrap distribution of the RC estimators (Efron and Tibshirani 1993), and on the inversion of a likelihood ratio test. The results are based on 2000 simulated primary and calibration samples. Detailed results from the simulations are included in tables I.1 – I.6, appendix I.

Table 3.1. Model parameters of the simulation scenarios. The parameters not shown in the table are $(\alpha_0, \alpha_1, \beta_0, \beta_1) = (0,1,1,1)$, n = 1000 and m = (10, 25, 50, 100, 200, 500).

Corr(X , W)	$\sigma^2_{X\|W}$	*Corr(Y , X)*	$\sigma^2_{Y\|X}$
High – 0.75	0.25^2	High – 0.75	0.34^2
		Low – 0.36	1.00^2
Low – 0.36	0.75^2	High – 0.75	0.70^2
		Low – 0.37	2.00^2

.

3.5.2. Performance of the estimators.

The SCK and ML estimators are at least as efficient as the alternatives in all the scenarios examined, but are not very different from each other (Fig. 3.3-3.5). In terms of RMSE (Fig. 3.4) and the spread of the sampling distribution (Fig. 3.5), the MLE is slightly more efficient than the SCK estimator when the correlation between *X* and *W* is 0.36, but indistinguishable when the correlation is 0.75. However, when the calibration sample size is small, the MLE tends to be positively biased, while the SCK tends to be negatively biased. If *Corr(X, W)* is low, these biases can be relatively large. The SCK estimator is an inverse-variance weighted average of $\hat{\beta}_{1,INT}$ and $\hat{\beta}_{1,EXT}$. The latter is proportional to $\hat{\alpha}_1^{-1}$, so large values of $\hat{\beta}_{1,EXT}$ are associated with small values of $\hat{\alpha}_1$. However, its estimated variance is proportional to $\hat{\alpha}_1^{-4}$. Therefore, smaller values of $\hat{\beta}_{1,EXT}$ would tend to receive a greater weight than large values, which would explain the negative bias.

The estimator of β_1 from the calibration data alone performs relatively well, and often better than some of the other estimators. Of course, $\hat{\beta}_{1,INT}$ is unbiased for all sample sizes, while the other estimators are not. The RC estimators were clearly more efficient than $\hat{\beta}_{1,INT}$ only when the correlation between *Y* and *X* is low and that between *X* and *W* high. In all other cases, the RMSE of $\hat{\beta}_{1,INT}$ is less than that of $\hat{\beta}_{1,EXT}$ for all sample sizes, often substantially so, and close to or less than that of the CRS estimator. While the RMSE of the SCK and ML estimators is less than that of $\hat{\beta}_{1,INT}$, the improvement is sometimes negligible.

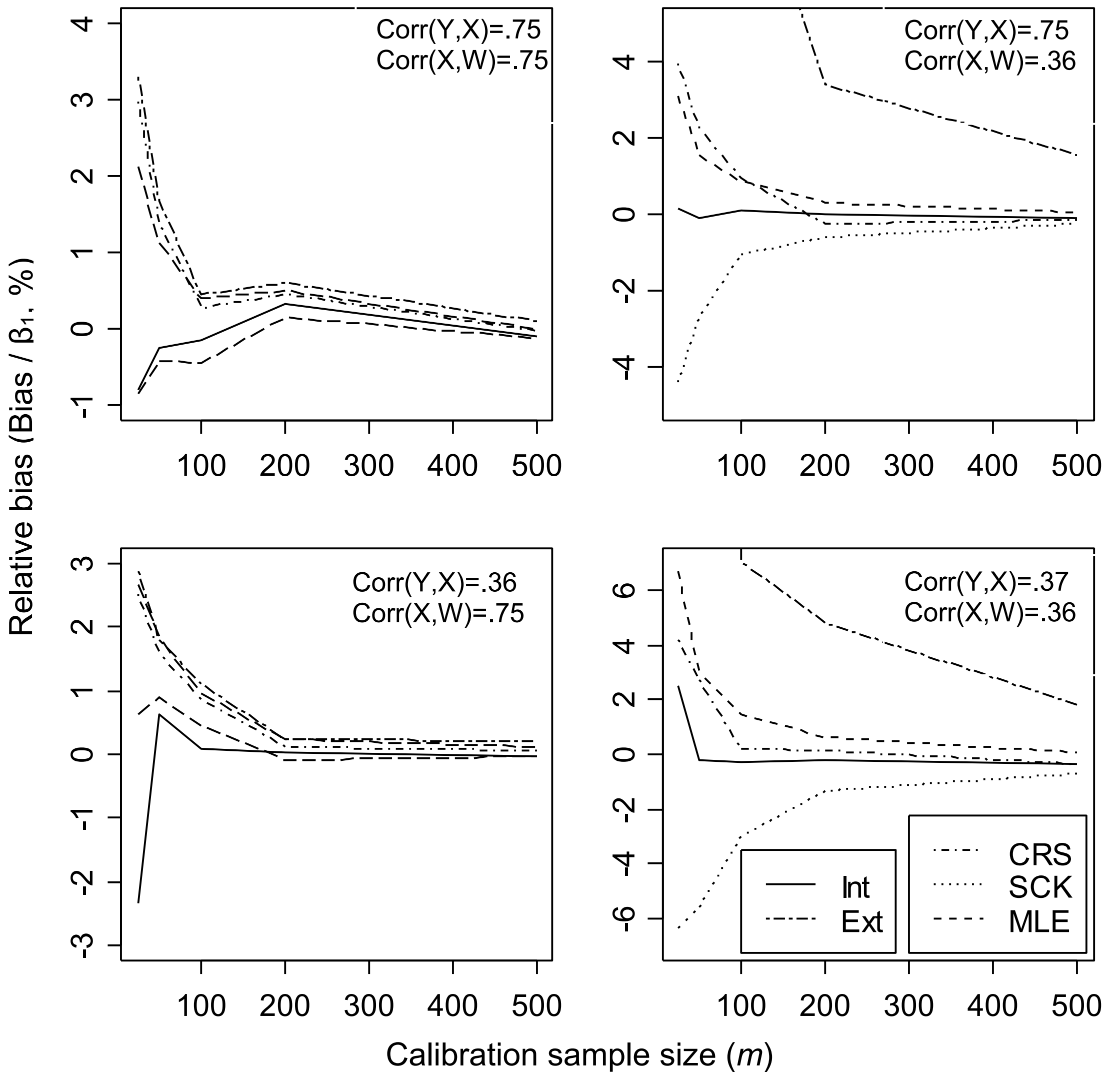

Figure 3.3. Relative bias of the internal, 'as external', CRS, SCK and ML estimators of the slope in simple linear regression with measurement error. The distributions of $Y|X$ and $X|W$ are normal, with a primary sample size of $n =$ 1000 and a calibration sample size of $m \geq 25$. The results are based on 2000 simulated primary and calibration samples.

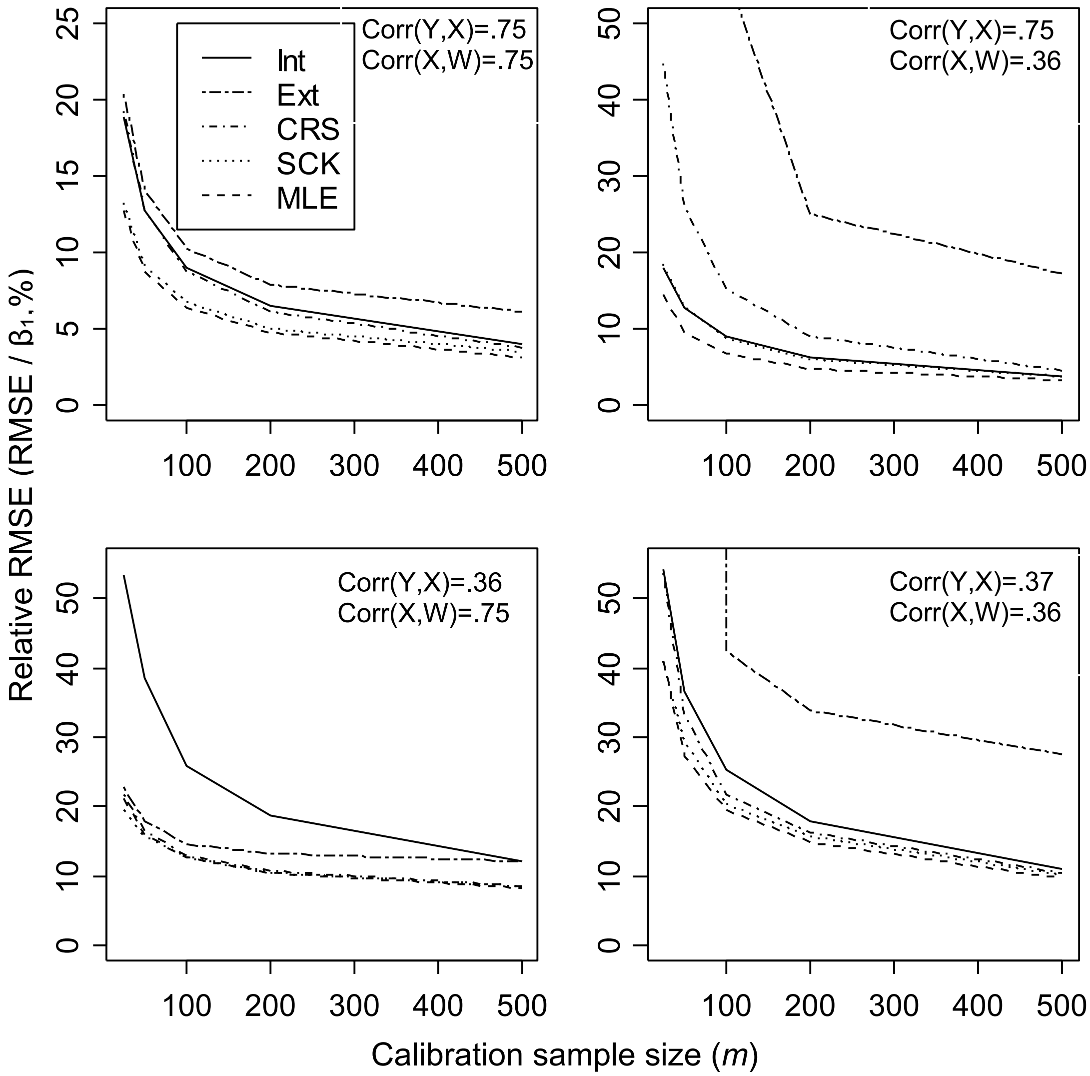

Figure 3.4. Relative root mean squared error of the internal, 'as-external', CRS, SCK and ML estimators of the slope in simple linear regression with measurement error. The distributions of $Y|X$ and $X|W$ are normal, with a primary sample size of $n = 1000$ and a calibration sample size of $m \geq 25$. The results are based on 2000 simulated primary and calibration samples.

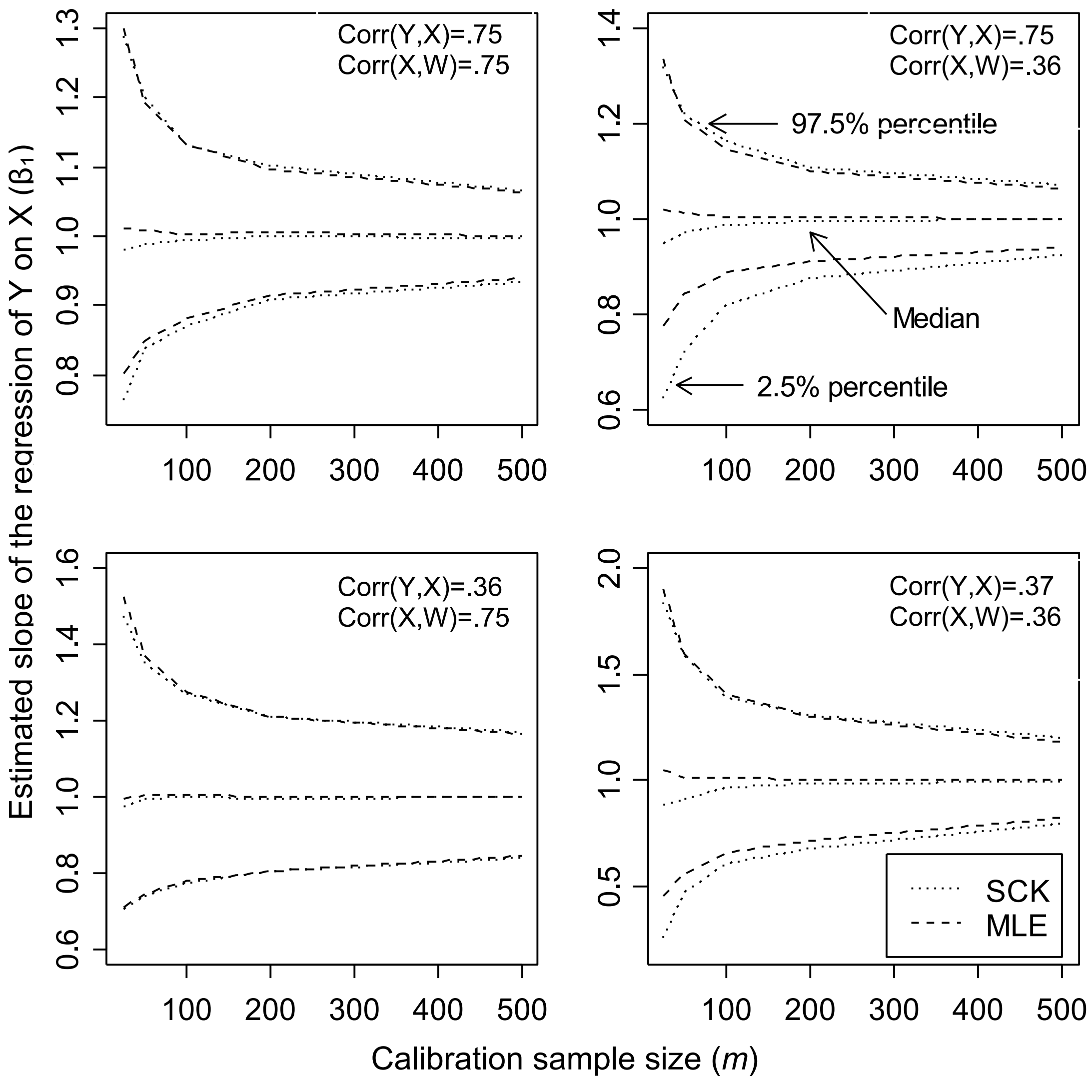

Figure 3.5. Selected percentiles of the Monte Carlo sampling distribution of the SCK and ML estimators of the slope in simple linear regression (with true slope equal one) with measurement error. The distributions of $Y|X$ and $X|W$ are normal, with a primary sample size of n = 1000 and a calibration sample size of $m \geq 25$. The results are based on 2000 simulated primary and calibration samples.

3.5.3. Performance of the confidence intervals.

Intervals based on inverting a LRT test are, on average, shorter than intervals based on the asymptotic normality of the sampling distribution of the SCK estimator or the bootstrap, even though sometimes they can be very close (Fig 3.6). The average length of the exact interval of $\hat{\beta}_{1,INT}$ is always greater that that of the SCK or ML estimators. However, the shorter CIs of the SCK and ML estimators are associated with a substantial increase in their error rate, especially for small sample sizes (Figs. 3.7-3.8). For example, when the correlation between X and W is low and m=50, the total error rate of the Wald-based 95% confidence interval is almost 10%, even though the average length of the exact confidence interval of $\hat{\beta}_{1,INT}$ is only 35% greater than that of the Wald-based confidence interval. Likelihood ratio and bootstrap confidence intervals have coverage rates that are closer to their nominal rates.

When the calibration sample is small, the 'misses' of the SCK and LRT confidence intervals tend to be greater in one direction. However, the true value of the parameter tends not to be included in the SCK intervals when greater than the SCK estimator, while the opposite is true for the LRT intervals. This difference may be explained by the negative and positive bias of the SCK and ML estimators, respectively.

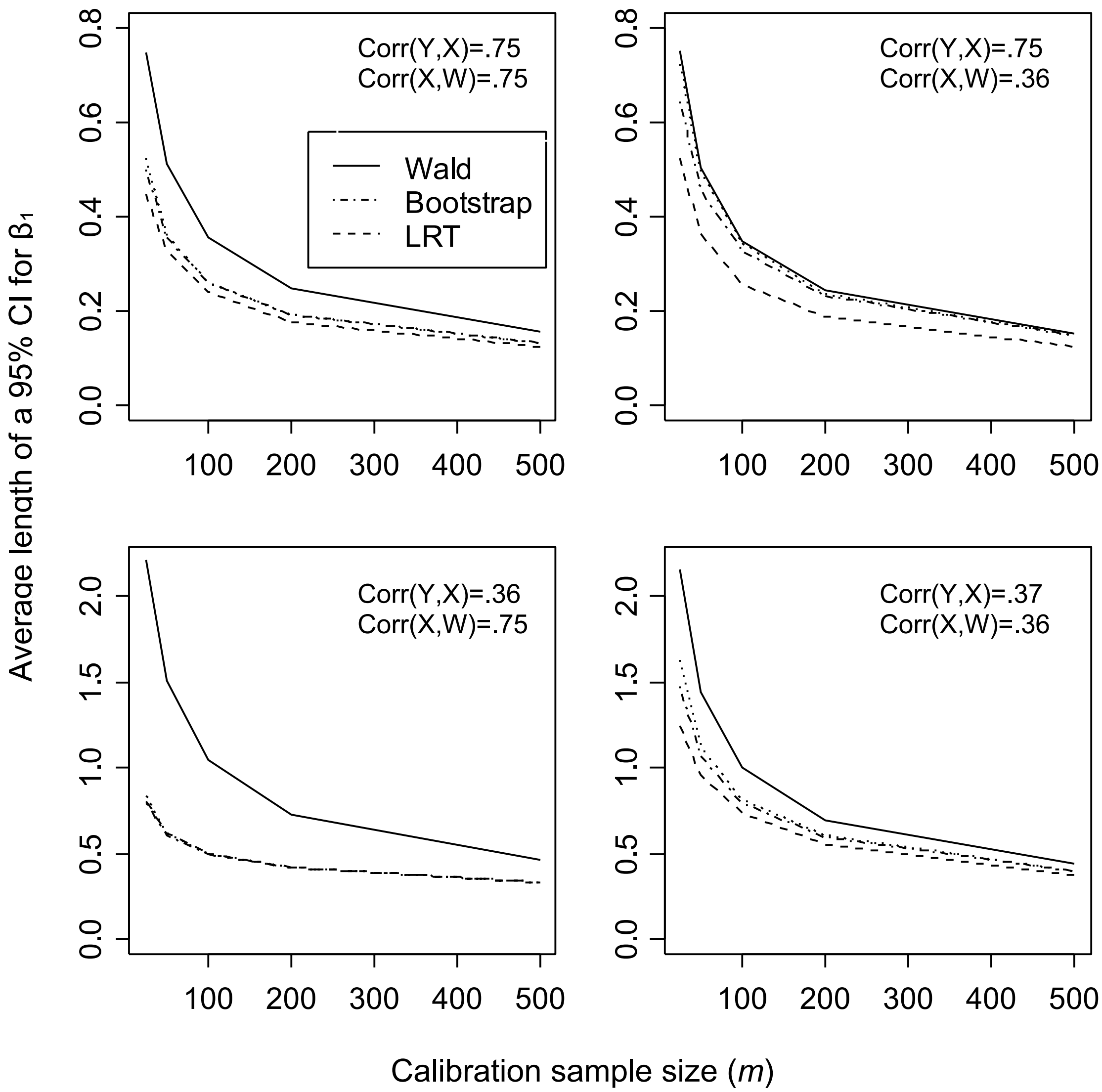

Figure 3.6. Average length of the estimated 95% confidence interval of the slope in simple linear regression with measurement error. The distributions of $Y|X$ and $X|W$ are normal, the primary sample size is n = 1000 the calibration sample size is m. The intervals are based on the t-distribution for the internal estimator; on a Wald approximation and bootstrap percentile for the SCK estimator; and on inverting the likelihood ratio test. The results are from 2000 simulated primary and calibration samples.

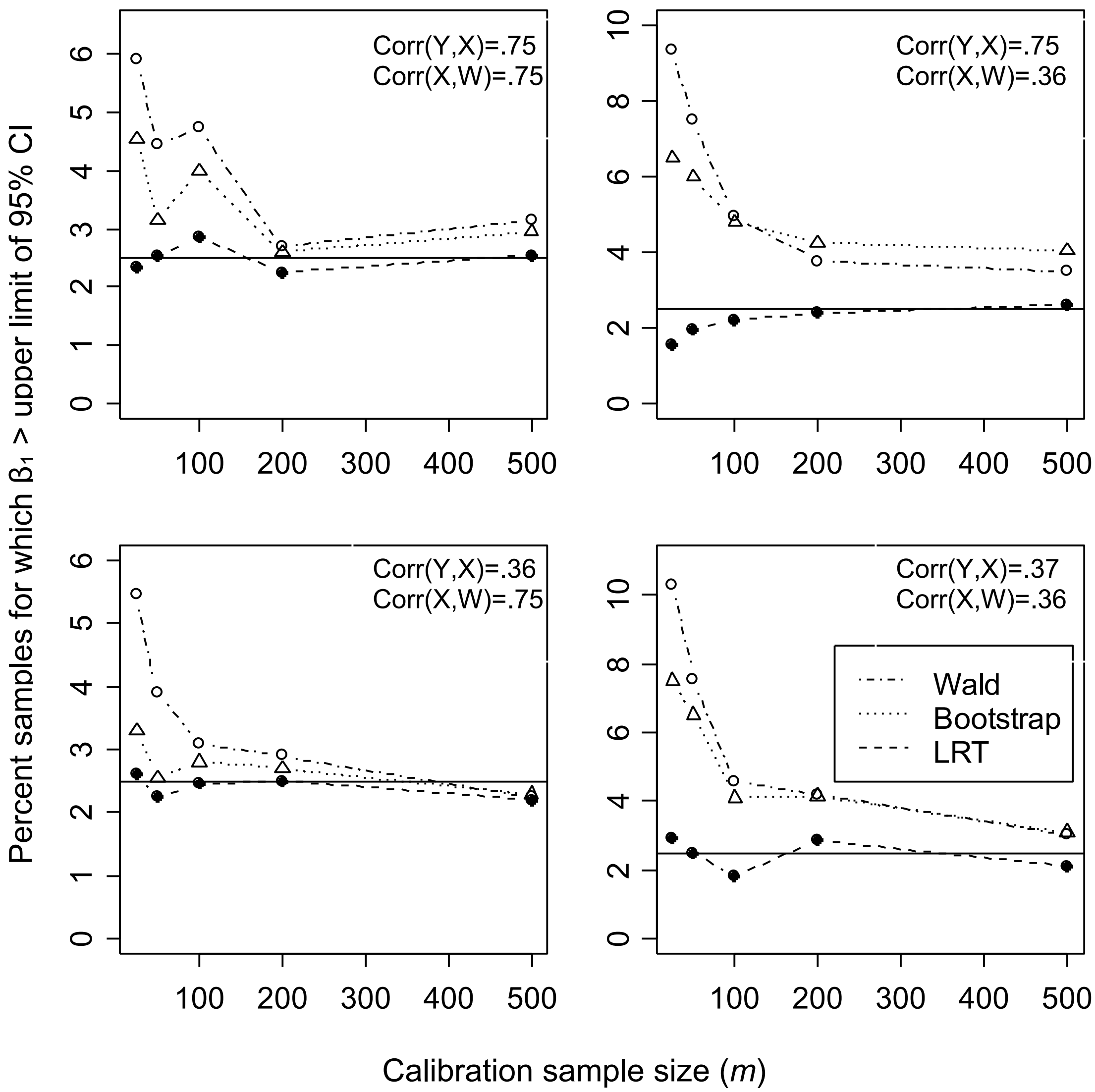

Figure 3.7. Percent of samples for which the true slope in simple linear regression with measurement error is greater than the upper limit of a 95% confidence interval. The distributions of $Y \mid X$ and $X \mid W$ are normal, with a primary sample size of n = 1000 and a calibration sample size of m. The intervals are based on a Wald approximation and bootstrap percentile for the SCK estimator, and inversion of the likelihood ratio test. The results are from 2000 simulated primary and calibration samples.

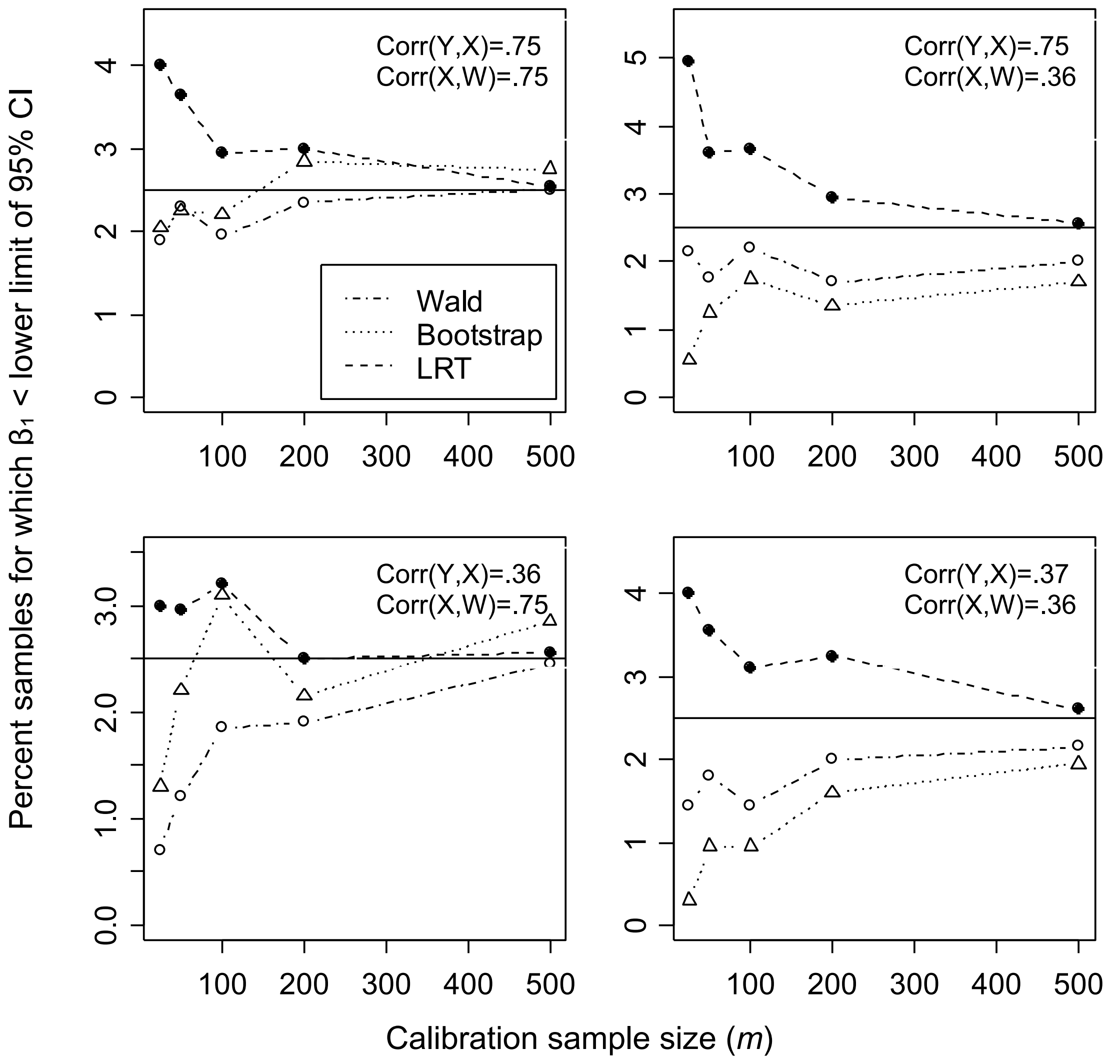

Figure 3.8. Percent of samples for which the true slope in simple linear regression with measurement error is less than the lower limit of a 95% confidence interval. The distributions of $Y \mid X$ and $X \mid W$ are normal, with a primary sample size of n = 1000 and a calibration sample size of m. The intervals are based on a Wald approximation and bootstrap percentile for the SCK estimator, and inversion of the likelihood ratio test. The results are from 2000 simulated primary and calibration samples.

3.5.4. Robustness.

There is a natural concern over the robustness of maximum likelihood inferences against distributional misspecifications. The existence of a calibration study in which the response, explanatory and surrogate variables are observed alleviates those concerns somewhat, because the assumed distributions can be checked against observed data. Nevertheless, the assessment may be difficult if the sample size and departures from the assumed distribution are small. In this section, we check the performance of the RC and ML estimators when the distribution of the error term of the regression of X on W follows a t-distribution with 5 degrees of freedom, or a lognormal distribution. The settings are the same as in the previous sections, and the t and lognormal distributions have been scaled to keep the same correlation between X and W and Y and X as before. Figure 3.9 shows the normal probability plot of those distributions when $\sigma^2_{X|W} = 0.75^2$.

The misspecifications of the distribution of the error term of the regression of X on W considered in this study have very little effect in any of the measurement of performance examined (Fig. 3.10-3.14 and tables I.5-I.7). Although there is a slight increase in bias and decrease in efficiency, which tends to be greater for the lognormal than for the t-distribution, the magnitude of the changes are minimal. Compared with the scenarios when the distribution of X on W was correctly specified, the relative performances of the estimators remain unchanged. Therefore, likelihood inference does not seem to be affected by small departures from the normal assumption of the distribution of X given W examined in this study.

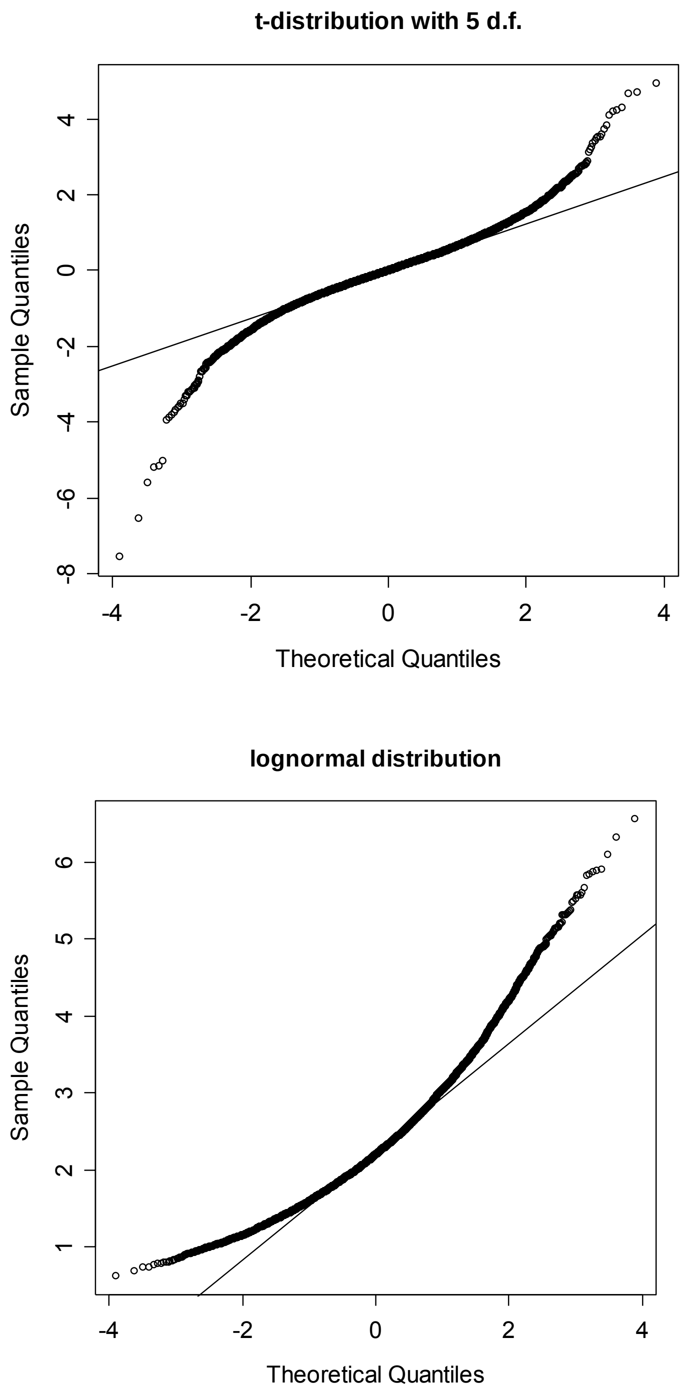

Figure 3.9. Normal probability plots of the distribution of the error term in the regression of X on W.

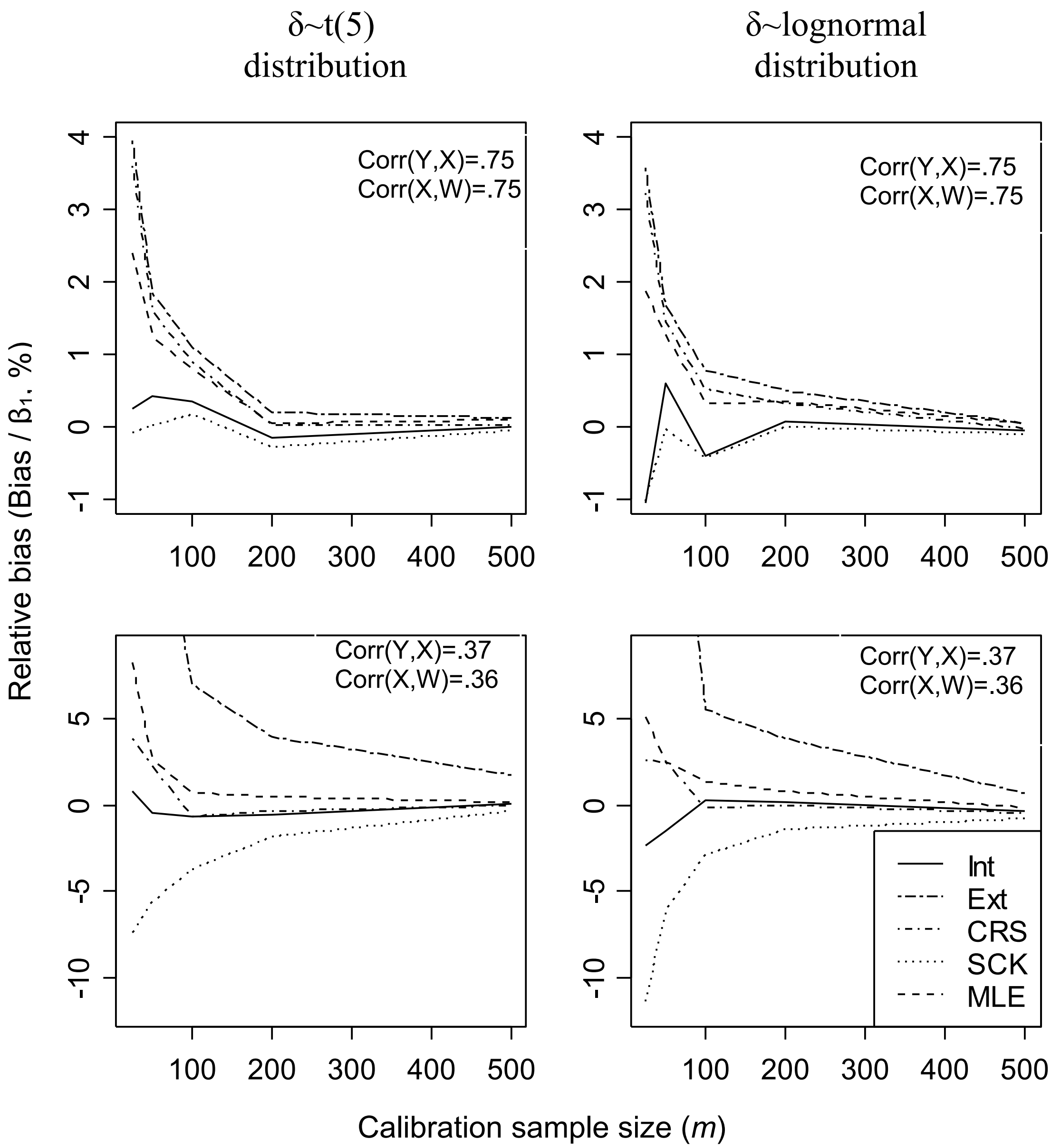

Figure 3.10. Robustness results: Relative bias of the internal, RC and ML estimators of the slope in simple linear regression with measurement error. The distributions $Y|X$ is normal, and the distribution of the regression error of $X|W$ is either a t-distribution with 5 d.f. or a lognormal distribution. The primary sample size is $n = 1000$ and the internal calibration sample size is m. The results are based on 2000 simulated primary and calibration samples.

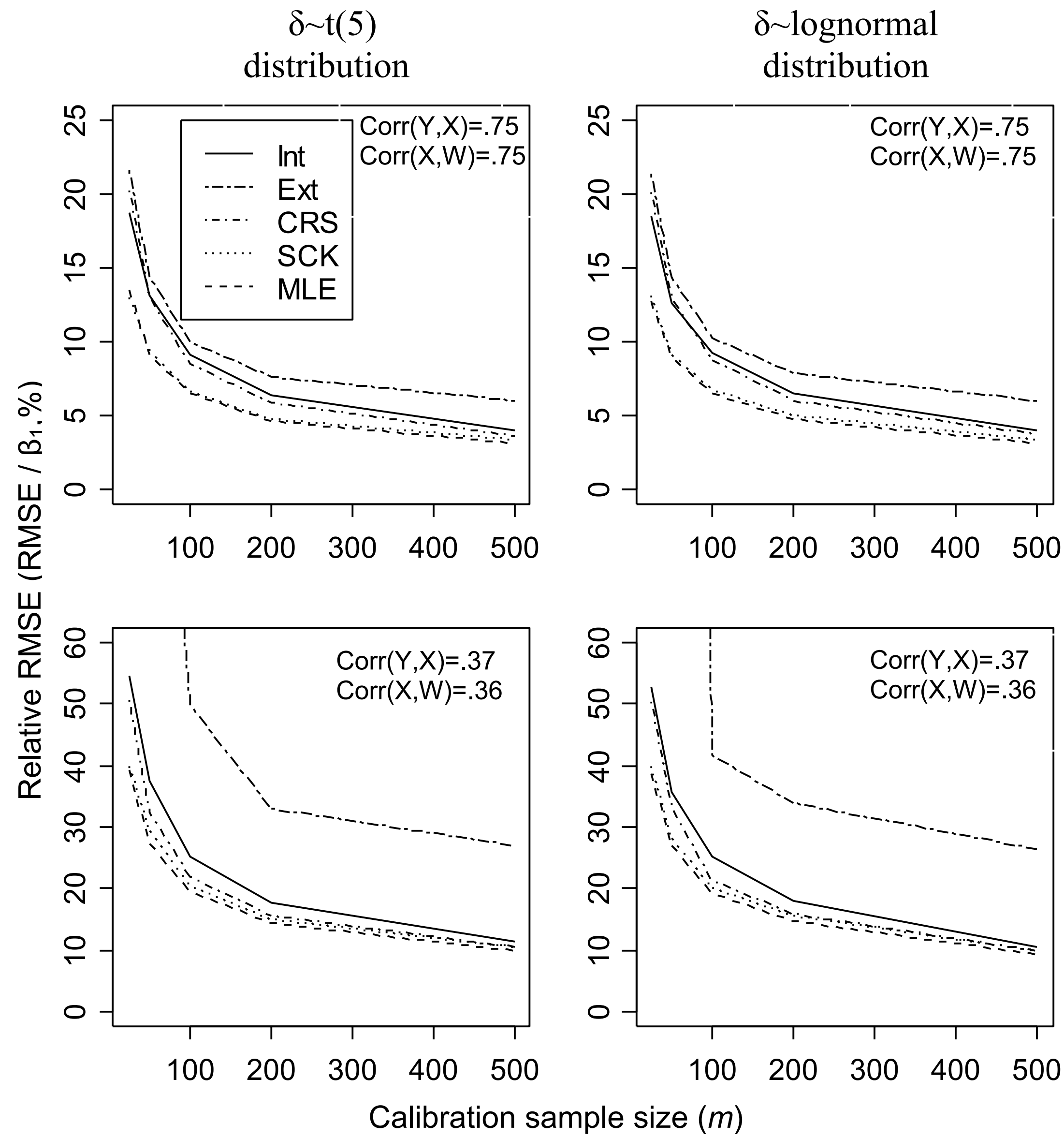

Figure 3.11. Robustness results: Relative root mean squared error of the internal, RC and ML estimators of the slope in simple linear regression with measurement error. The distributions $Y|X$ is normal, and the distribution of the regression error of $X|W$ is either a t-distribution with 5 d.f. or a lognormal distribution. The primary sample size is n = 1000 and the internal calibration sample size is m. The results are based on 2000 simulated primary and calibration samples.

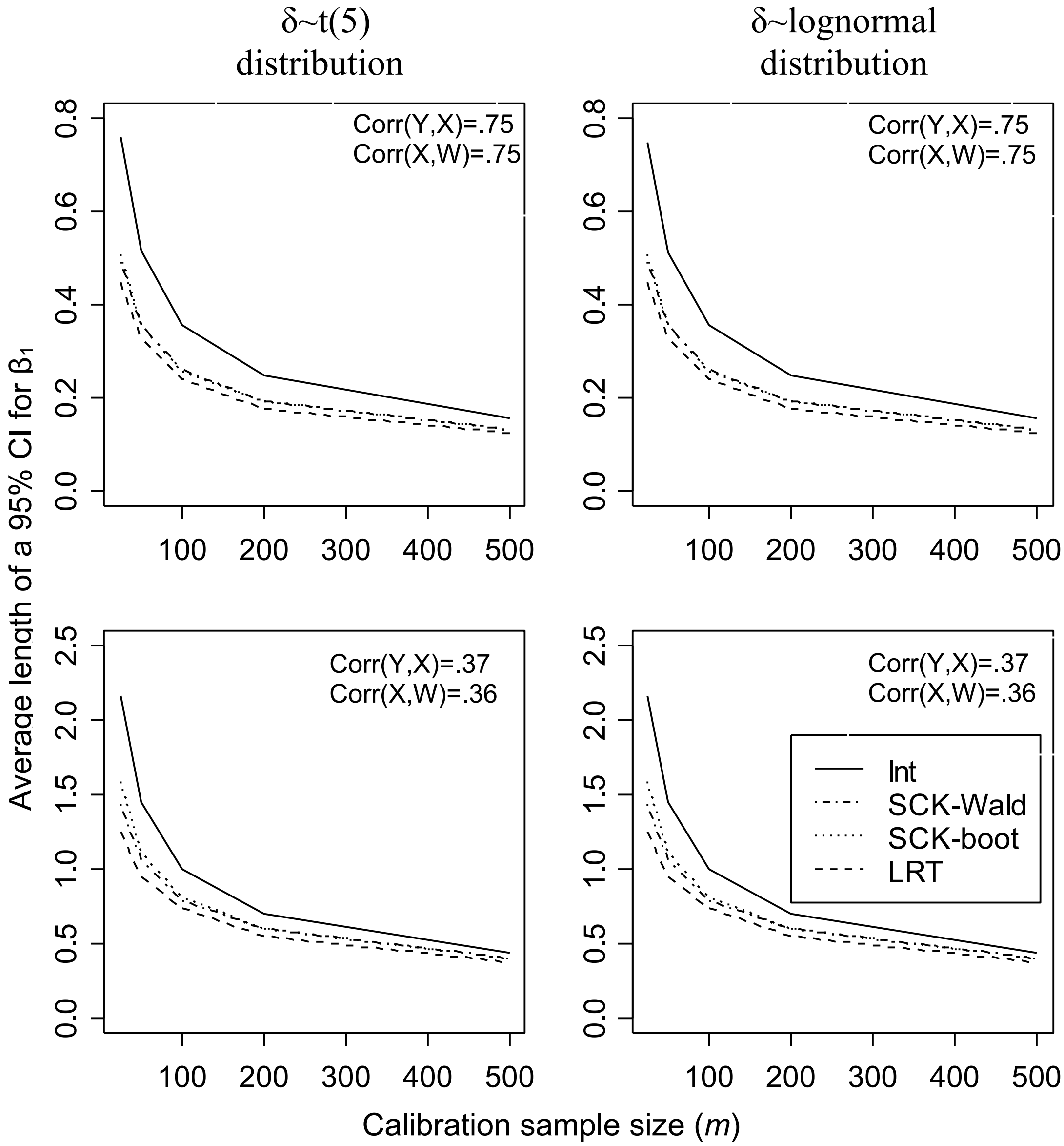

Figure 3.12. Robustness results: Average length of the estimated 95% confidence interval of the slope in simple linear regression with measurement error. The distributions $Y|X$ is normal, and the distribution of the regression error of $X|W$ is either a t-distribution with 5 d.f. or a lognormal distribution. The intervals are based on the t-distribution for the internal estimator; on a Wald approximation and bootstrap percentile for the SCK estimator; and on inverting the likelihood ratio test. The results are from 2000 simulated primary and calibration samples.

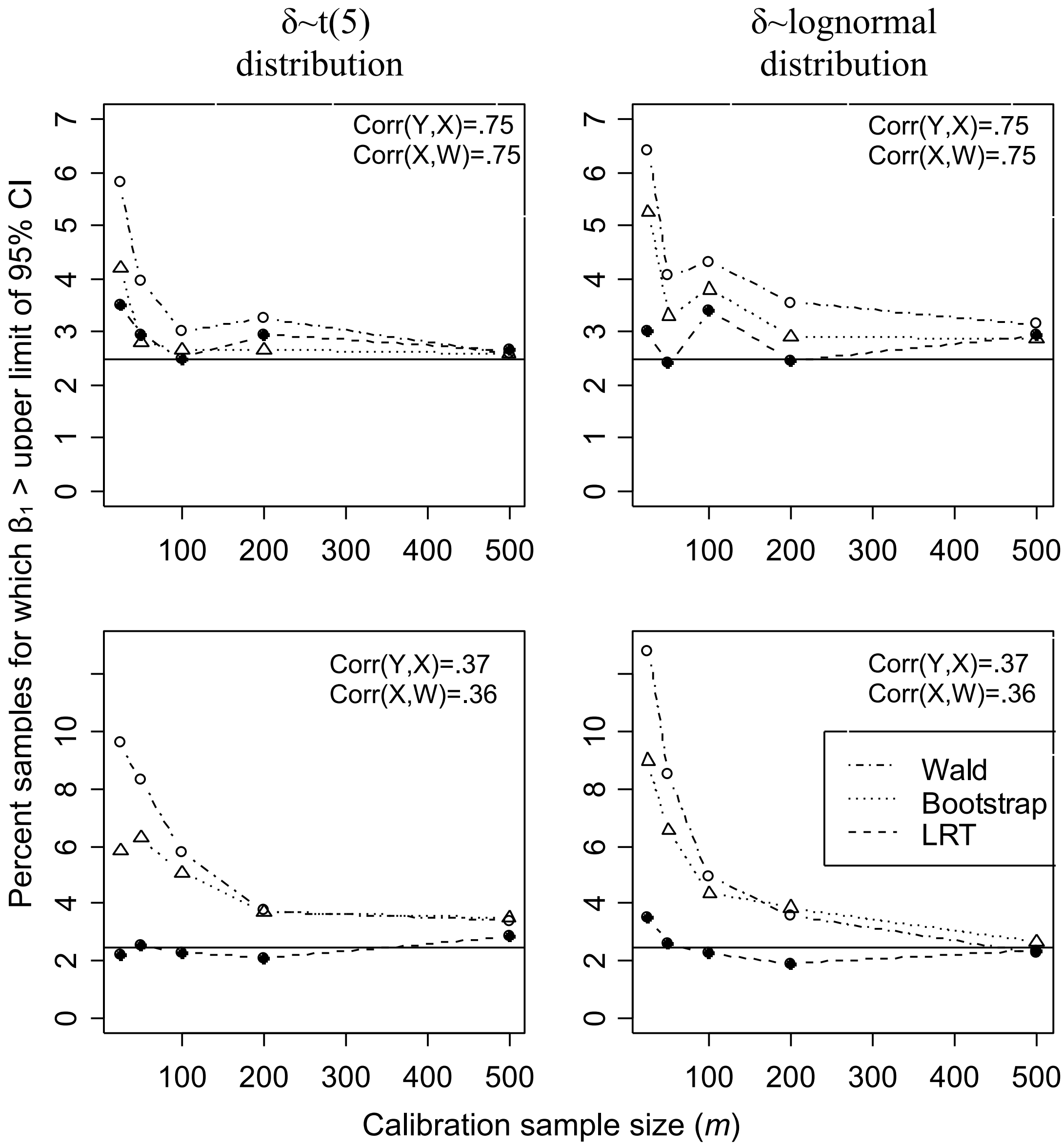

Figure 3.13. Robustness results: Percent of samples for which the true slope in simple linear regression with measurement error is greater than the upper limit of a 95% confidence interval. The distributions $Y|X$ is normal, and the distribution of the regression error of $X|W$ is either a t-distribution with 5 d.f. or a lognormal distribution. The primary sample size is n = 1000 and the internal calibration sample size is m. The intervals are based on a Wald approximation and bootstrap percentile for the SCK estimator, and inversion of the likelihood ratio test. The results are from 2000 simulated primary and calibration samples.

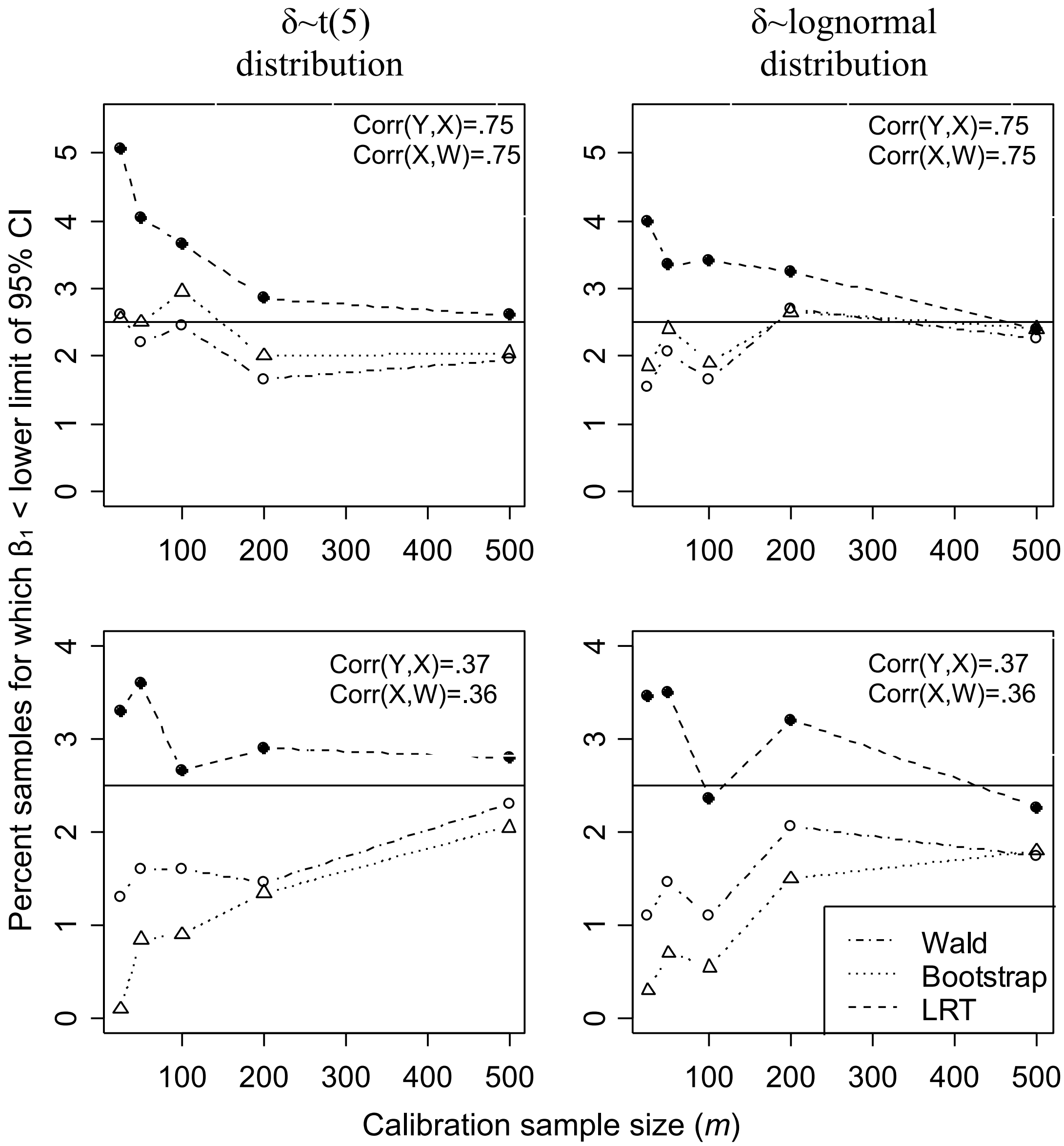

Figure 3.14. Robustness results: Percent of samples for which the true slope in simple linear regression with measurement error is less than the lower limit of a 95% confidence interval. The distributions $Y|X$ is normal, and the distribution of the regression error of $X|W$ is either a t-distribution with 5 d.f. or a lognormal distribution. The primary sample size is $n = 1000$ and the internal calibration sample size is m. The intervals are based on a Wald approximation and bootstrap percentile for the SCK estimator, and inversion of the likelihood ratio test. The results are from 2000 simulated primary and calibration samples.

3.6. Conclusions

If the distribution of the response variable given the true explanatory variable, and the explanatory variable given the surrogate are both normal, the maximum likelihood estimator is asymptotically more efficient than the proposed regression calibration estimators. However, simulation results indicate that the estimator proposed by Spiegelman, Carroll and Kipnis (2001) is nearly as efficient as the maximum likelihood estimator, particularly when the measurement error is small. This estimator is defined as an inverse-variance weighted average of the estimator of the regression parameters obtained from the calibration data alone, and the regression calibration estimator obtained by treating the internal calibration data as an external calibration study. It is easy to compute with existing software, and does not require special iterative calculations, even when the distributions involved are not normal. Therefore, from a practical point of view, it is a sensible alternative to the maximum likelihood estimator.

Confidence intervals based on the asymptotic normality and estimated standard errors of Spiegelman's et al. (2001) estimator, however, tend to have a high error rate for small calibration sample sizes. Inferential properties can be improved by using the bootstrap, but this reduces the simplicity and appeal of the estimator. Confidence intervals calculated by inverting a likelihood ratio test are somewhat more efficient and accurate than the bootstrap.

Other seemingly sensible estimators can be substantially less efficient than the MLE, and may even result in an actual loss of efficiency compared with estimators obtained from the calibration data alone. This is the case, for example, with the estimator obtained from the regression of Y on X when available, and on the estimated $E(X \mid W)$ otherwise. If the correlation between the response and true explanatory variable is high, this estimator is less efficient than the estimator that ignores the primary study altogether.

Maximum likelihood inference requires the specification of the distributions of $Y|X$ and $X|W$, while regression calibration only requires specification of the moments of the distributions. Simulations show that the maximum likelihood estimator, however, does not seem to be affected by small misspecification of the error of the regression of X on W. Its efficiency, both in overall terms and relative to the other estimator considered here, remained virtually unchanged. Although Spiegelman's et al. (2001) estimator is simple and transparent, because of the better inferential properties and relative robustness, we believe that likelihood inference may be worth the extra difficulty.

3.7. References

Armstrong, B. 1985. Measurement error in generalized linear models. *Communications in Statistics Series B* 14: 529-544.

Binham, S.A., Day, N.E., Luben, R., and others. 2003. Dietary fiber in food and protection against colorectal cancer in the European Prospective Investigation into Cancer and Nutrition. *Lancet* 361: 1496-1501.

Carroll, R.J., Roeder, K, and Wasserman, L. 1999. Flexible parametric measurement error models. *Biometrics* 55: 44-54.

Carroll, R.J., Ruppert, D., and Stefanski, L.A. 1995. *Measurement Error in Nonlinear Models*. Chapman & Hall, London.

Carroll, R.J., and Stefanski, L.A. 1990. Approximate quasi-likelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association* 85: 652-663.

Efron, B. and Tibshirani, R.J. 1993. *An introduction to the bootstrap*. Chapman & Hall. New York. 436 p.

Fuller, W.A. 1987. *Measurement Error Models*. John Wiley & Sons, New York.

Kuha, J. 1994. Corrections for exposure measurement error in logistic regression models with an application to nutritional data. *Statistics in Medicine* 13: 1135-1148.

Pepe, M. S., Self, S. G. and Prentice, R. L. 1989. Further results in covariate measurement errors in cohort studies with time to response data. *Statistics in Medicine* 8: 1167-1178.

Pierce, D.A., Stram, D.O., and Vaeth, M. 1990. Allowing for random errors in radiation dose estimates for the atomic bomb survivors. *Radiation Research* 123: 275-284.

Pierce, D.A., Stram, D.O., Vaeth, M., and Schafer, D.A. 1992. Some insights into the errors in variables problem provided by consideration of radiation dose-response analyses for the A-bomb survivors. *Journal of the American Statistical Association* 87: 351-359.

Prentice, R.L. 1982. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* 69: 331-342.

Riersol, O. 1950. Identifiablity of a linear relation between variables that are subject to error. *Econometrica* 18: 375-389.

Roeder, K., Carroll, R.J., and Lindsay, B.G. 1996. A semiparametric mixture approach to case-control studies with errors in covariates. *Journal of the American Statistical Association* 91: 722-732.

Rosner, B., Willet, W.C., and Spiegelman, D. 1989. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine* 8: 1051-1069.

Rosner, B., Spiegelman, D., and Willet, W.C. 1990. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *American Journal of Epidemiology* 132: 734-745.

Schafer, D.W. 2001. Semiparametric maximum likelihood for measurement error model regression. *Biometrics* 57: 53-61.

Schafer, D.W. 2002. Likelihood analysis and flexible structural modeling for measurement error model regression. *Journal of Statistical Computing and Simulation* 72: 33-45.

Schafer, D. W., Lubin, J.H., Ron, E., Stovall, M., and Carroll, R.J. 2001. Thyroid cancer following scalp irradiation: a reanalysis accounting for uncertainty in dosimetry. *Biometrics* 57: 689-697.

Spiegelman, D., Carroll, R.J., and Kipnis, V. 2001. Efficient regression calibration for logistic regression in main study / internal validation study designs with an imperfect reference instrument. *Statistics in Medicine* 20: 139-160.

Spiegelman, D., McDermott, A., and Rosner, B. 1997. Regression calibration method for correcting measurement error bias in nutritional epidemiology. *American Journal of Clinical Nutrition* 65 (suppl): 1179S-1186S.

Spiegelman, D., Rosner, B., and Logan, R. 2000. Estimation and inference for logistic regression with covariate misclassification and measurement error in main study / validation study designs. . *Journal of the American Statistical Association* 95: 51-61.

Stram, D.O., Langholz, B., Huberman, M., and Thomas, D.C. 1999. Correction for exposure measurement error in a reanalysis of lung cancer mortality for the Colorado Plateau uranium miners cohort. *Health Physics* 77: 265-275.

Thurston, S.W., Spiegelman, D, and Ruppert, D. 2003. Equivalence of regression calibration methods for external validation data. *Journal of Statistical Planning and Inference* 113: 527-539.

Thurston, S.W., Williams, P.L., Hauser, R., Hu, H., Hernandez-Avila, M., and Spiegelman, D. 2005. A comparison of regression calibration approaches for designs with internal validation data. *Journal of Statistical Planning and Inference* 131: 175-190.

Willett, W.C., Hunter, D.J., Stampfer, M.J., Colditz, G., Manson, J.E., Spiegelman, D., Rosner, B., Hennekens, C.H., and Speizer, F.E. 1992. Dietary fat and fiber in relation to breast cancer: an eight year follow up. *Journal of the American Medical Association* 268: 2037-2044.

Willett, W.C., Sampson, L., Browne, M.L., Stampfer, M.J., Rosner, B., Hennekens, C.H., and Speizer, F.E. 1988. The use of a self-administered questionnaire to assess diet four years in the past. *American Journal of Epidemiology* 127: 188-199.

4. CONCLUSIONS

Regression calibration is a simple, transparent and intuitive approach to estimate regression models with measurement error. In their monograph, Carroll, Ruppert and Stefanski (1995, p.142) note that "traditional folklore" suggests that, in many statistical models, simpler methods such as regression calibration perform just as well as likelihood methods, but they remark that there is little documentation to support this belief. Since Carroll et al. (1995) made this observation, a few studies have compared versions of regression calibration and likelihood methods, some supporting the folklore (for example, Spiegelman et al. 2001, Thoresen and Laake 2000), while others indicating that the maximum estimator can be substantially more efficient (for example, Spiegelman et al. 2000, Suh and Schafer 2002). Those studies rely on simulation of a limited number of scenarios, because analytical results or extensive simulation are not feasible in the complex settings that they consider.

The results from chapter 2 partially support this folklore, because for normal linear regression with an independent calibration dataset, the regression calibration estimator is equivalent to the maximum likelihood estimator, provided a natural estimate of variance is non-negative. In practice, this result offers a check for judging the suitability of regression calibration. However, if the estimate of this variance is negative, the regression calibration estimator can be very unstable and attain unreasonably large values, while the MLE is bounded and closer to the target parameter. Negative variance estimates are likely if the uncertainty on the estimation of $E(X \mid W)$, the conditional expectation of the true explanatory variable given its surrogate, is large. However, they are also likely if β_1 is close, in magnitude, to $\sqrt{\sigma^2_{Y|W} / \sigma^2_{X|W}}$, where β_1 is the slope of the regression of the response variable Y on the true explanatory variable X, and $\sigma^2_{Y|W}$ and $\sigma^2_{X|W}$ the variance of the regression of Y on the surrogate variable W and the measurement error variance, respectively. This situation reflects a strong relationship between *Y* and *X*, and a relatively large measurement error variance.

Asymptotically, the regression calibration and maximum likelihood estimators and their sampling distributions are equivalent, and converge to a normal distribution. However, for finite sample sizes, the sampling distribution of the regression calibration estimator can be very skewed, and even bimodal. One-sided error rates of confidence intervals based on the asymptotic normality and estimated standard errors are inaccurate. Two-sided type I error rates may be more accurate, but this is misleading because it is due to the one-sided error rates tend to be too high in one tail and too low in the other. In addition, the moments of the sampling distribution of the estimated asymptotic variance are not defined, resulting in large and erratic values for small calibration sample sizes. Confidence intervals calculated from the percentiles of the bootstrapped sampling distribution of the regression calibration estimator is more accurate than those based on approximate normality, in terms of symmetry and overall error rate, with only a relatively small increase in length.

Inference based on the likelihood ratio results in the shortest confidence intervals, while keeping a coverage rate close to the nominal rate. The efficiency gains arise in part from the bound imposed by constraining that the estimated variance of the conditional distribution of *Y* on *X* to be positive. However, under this constraint and for some values of the parameters, the sampling distribution of the LRT statistic cannot be approximated by a χ^2 distribution. This problem is apparent for small sample sizes in general, but becomes increasingly important when the true slope of the regression of *Y* on *X* is close, in magnitude, to the ratio of the standard deviations of *Y* and *X*, given the surrogate variable. Simulations indicated that, in those cases, LRT confidence intervals were conservative.

Chapter 3 discusses the case when the true explanatory variable is observed in a subset of the data. The maximum likelihood estimator is asymptotically more efficient than the proposed regression calibration estimators, but the difference decreases as the measurement error variance decreases. Through an extensive simulation study, we show that if the distribution of the response variable given the true explanatory variable, and the explanatory variable given the surrogate are both normal, the estimator proposed by Spiegelman, Carroll and Kipnis (2001) is nearly as efficient as the maximum likelihood estimator for a range of calibration sample sizes. This estimator is defined as an inverse-variance weighted average of the estimator of the regression parameters obtained from the calibration data alone, and the regression calibration estimator obtained by treating the internal calibration data as an external calibration study. Spiegelman's et al. (2001) estimator can be computed with little more than standard software and, therefore, from a practical point of view, it is a sensible choice. However, confidence intervals based on the asymptotic normality and estimated standard errors of Spiegelman's et al. (2001) estimator, tend to have a high error rate for small calibration sample sizes. Inferential properties can be improved by using the bootstrap, but this reduces the simplicity and appeal of the estimator. Confidence intervals calculated by inverting a likelihood ratio test are somewhat more efficient and accurate than the bootstrap.

Other seemingly sensible versions of the regression calibration estimator can be substantially less efficient than the MLE, and may even result in an actual loss of efficiency compared with simple estimators obtained from the calibration data alone. If the correlation between the response and true explanatory variable is high, the estimator derived from the regression of Y on X when available, and on the estimated $E(X \mid W)$ otherwise, is less efficient than the estimator that ignores the primary study altogether.

Although this study focuses on regression calibration inference when both the distribution of Y given X and of X given its surrogate are normal, the findings apply more generally. For generalized linear models with an external calibration dataset, the regression calibration estimator can be written as the ratio of the slope of the regression of Y on W to the slope of the regression of X on W. Since the estimators of those slopes are approximately normal, the results in this paper regarding the sampling distribution of the regression calibration estimator and implications for inference remain approximately valid. However, in those cases, as when the calibration dataset is internal, the regression calibration and maximum likelihood estimators may not be equivalent.

Even though regression calibration estimators are simple and, in some instances, equivalent to the maximum likelihood estimator, likelihood ratio inference is shown to be more accurate and efficient than inference based on the approximate normality of the regression calibration estimator. The existence of a calibration study in which both the true explanatory variable and its surrogate are observed alleviate some of the concerns over the robustness of maximum likelihood inferences against possible distributional misspecifications. When a calibration study is available, the distribution of the true explanatory variable does not have to be specified, and the adequacy of all the assumed distributions can be assessed with the calibration data. For complex cases, however, likelihood analysis can be challenging computationally for realistic distributional assumptions. However, if data collection and study involved significant time and cost, the additional effort in doing a likelihood analysis would be small for realizing greater efficiency and more powerful tests and confidence intervals.

BIBLIOGRAPHY

Adcock, R.J. 1877. Note on the method of least squares. *Analyst* 4: 183-184.

Armstrong, B. 1985. Measurement error in generalized linear models. *Communications in Statistics Series B* 14: 529-544.

Athreya, K.B. 1987. Bootstrap of the mean in the infinite variance case. *Annals of Statistics* 15: 724-731.

Berkson, J. 1950. Are there two regressions? *Journal of the American Statistical Association* 45: 164-180.

Binham, S.A., Day, N.E., Luben, R., and others. 2003. Dietary fiber in food and protection against colorectal cancer in the European Prospective Investigation into Cancer and Nutrition. *Lancet* 361: 1496-1501.

Carroll, R.J., Roeder, K, and Wasserman, L. 1999. Flexible parametric measurement error models. *Biometrics* 55: 44-54.

Carroll, R.J., Ruppert, D., and Stefanski, L.A. 1995. *Measurement Error in Nonlinear Models*. Chapman & Hall, London.

Carroll, R.J., and Stefanski, L.A. 1990. Approximate quasi-likelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association* 85: 652-663.

Clayton, D.G. 1992. Models for the analysis of cohort and case-control studies with inaccurately measured exposures. *In*: Dwyer, J.H., Feinleib, N., Lippert, P., and Hoffmeister, H. (eds.). *Statistical Models for Longitudinal Studies of Health.* Oxford University Press, New York. pp 301-325.

Cochran, W.G. 1965. Errors of measurement in statistics. *Technometrics* 10: 637-665.

Efron, B. and Tibshirani, R.J. 1993. *An introduction to the bootstrap*. Chapman & Hall. New York. 436 p.

Fraser, G.E. and Stram, D.O. 2001. Regression calibration in studies with correlated variables measured with error. *American Journal of Epidemiology* 2001: 836-844.

Freedman, W.L., B.F. Madore, B.K. Gibson, L. Ferrarese, D.D. Kelson, S. Sakai, J.R. Mould, R.C. Kennicutt Jr., H.C. Ford, J.A. Graham, J.P. Huchra, S.M.G. Hughes, G.D. Illingworth, L. Macri, P.B. Stetson. 2001. Final results from the Hubble space telescope key project to measure the Hubble constant. *Astrophysical Journal* 553: 47-72

Fuller, W.A. 1987. *Measurement Error Models*. John Wiley & Sons, New York.

Glesser, L.J. and Hwang, J.T. 1987. The nonexistence of $100(1-\alpha)\%$ confidence sets of finite expected diameter in errors-in-variables and related models. *Annals of Statistics* 15: 1351-1362,

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

Hinkley, D.V. 1969. On the ratio of two correlated normal random variables. *Biometrika* 56: 635-639.

Kuha, J. 1994. Corrections for exposure measurement error in logistic regression models with an application to nutritional data. *Statistics in Medicine* 13: 1135-1148.

Lehmann, E.L. and Casella, G. 1998. *Theory of Point Estimation*, 2nd Edition. Springer Verlag, New York.

Madansky, A. 1959. The fitting of straight lines when both variables are subject to error. *Journal of the American Statistical Association* 54: 173-205.

Marsaglia, G. 1965. Ratios of normal variables and ratios of sums of uniform variables. *Journal of the American Statistical Association* 60: 193-204.

Pepe, M. S., Self, S. G. and Prentice, R. L. 1989. Further results in covariate measurement errors in cohort studies with time to response data. *Statistics in Medicine* 8: 1167-1178.

Pierce, D.A., Stram, D.O., and Vaeth, M. 1990. Allowing for random errors in radiation dose estimates for the atomic bomb survivors. *Radiation Research* 123: 275-284.

Pierce, D.A., Stram, D.O., Vaeth, M., and Schafer, D.A. 1992. Some insights into the errors in variables problem provided by consideration of radiation dose-response analyses for the A-bomb survivors. *Journal of the American Statistical Association* 87: 351-359.

Prentice, R.L. 1982. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* 69: 331-342.

Reeves, G.K., Cox, D.R., Darby, S.C., and Whitley, E. 1998. Some aspects of measurement error in explanatory variables for continuous and binary regression models. *Statistics in Medicine* 17: 2157-2177.

Riersol, O. 1950. Identifiablity of a linear relation between variables that are subject to error. *Econometrica* 18: 375-389.

Roeder, K., Carroll, R.J., and Lindsay, B.G. 1996. A semiparametric mixture approach to case-control studies with errors in covariates. *Journal of the American Statistical Association* 91: 722-732.

Rosner, B., Willet, W.C., and Spiegelman, D. 1989. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine* 8: 1051-1069.

Rosner, B., Spiegelman, D., and Willet, W.C. 1990. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *American Journal of Epidemiology* 132: 734-745.

Schafer, D.W. 2001. Semiparametric maximum likelihood for measurement error model regression. *Biometrics* 57: 53-61.

Schafer, D.W. 2002. Likelihood analysis and flexible structural modeling for measurement error model regression. *Journal of Statistical Computing and Simulation* 72: 33-45.

Schafer, D.W. and Gilbert, E.S. Statistical implications of dose uncertainty in radiation dose-response analysis. *Radiation Research* (in press)

Schafer, D. W., Lubin, J.H., Ron, E., Stovall, M., and Carroll, R.J. 2001. Thyroid cancer following scalp irradiation: a reanalysis accounting for uncertainty in dosimetry. *Biometrics* 57: 689-697.

Spiegelman, D., Carroll, R.J., and Kipnis, V. 2001. Efficient regression calibration for logistic regression in main study / internal validation study designs with an imperfect reference instrument. *Statistics in Medicine* 20: 139-160.

Spiegelman, D., McDermott, A., and Rosner, B. 1997. Regression calibration method for correcting measurement error bias in nutritional epidemiology. *American Journal of Clinical Nutrition* 65 (suppl): 1179S-1186S.

Spiegelman, D., Rosner, B., and Logan, R. 2000. Estimation and inference for logistic regression with covariate misclassification and measurement error in main study / validation study designs. . *Journal of the American Statistical Association* 95: 51-61.

Stram, D.O., and Kopecky, K.J. 2003. Power and uncertainty analysis of epidemiological studies of radiation-related disease risk in which dose estimates are based on a complex dosimetry system: some observations. *Radiation Research* 160: 408-417,

Stram, D.O., Langholz, B., Huberman, M., and Thomas, D.C. 1999. Correction for exposure measurement error in a reanalysis of lung cancer mortality for the Colorado Plateau uranium miners cohort. *Health Physics* 77: 265-275.

Suh, E.Y. and Schafer, D.W. 2002. Semiparametric maximum likelihood for nonlinear regression with measurement errors. *Biometrics* 58: 448-453.

Thoresen, M. and Laake, P. 2000. A simulation study of measurement error correction in logistic regression. *Biometrics* 56: 868-872.

Thurston, S.W., Spiegelman, D, and Ruppert, D. 2003. Equivalence of regression calibration methods for external validation data. *Journal of Statistical Planning and Inference* 113: 527-539.

Thurston, S.W., Williams, P.L., Hauser, R., Hu, H., Hernandez-Avila, M., and Spiegelman, D. 2005. A comparison of regression calibration approaches for designs with internal validation data. *Journal of Statistical Planning and Inference* 131: 175-190.

Toteson, T.D. and Tsiatis, A.A. 1988. The asymptotic relative efficiency of score tests in a generalised linear model with surrogate covariates. *Biometrika* 75: 507-514.

van Gils, C.H., Peeters, P.H.M., Bueno de Mesquita, H.B., and others. 2005. Consumption of vegetables and fruits and risk of breast cancer. *Journal of the American Medical Association* 293: 183-193.

Willett, W.C., Hunter, D.J., Stampfer, M.J., Colditz, G., Manson, J.E., Spiegelman, D., Rosner, B., Hennekens, C.H., and Speizer, F.E. 1992. Dietary fat and fiber in relation to breast cancer: an eight year follow up. *Journal of the American Medical Association* 268: 2037-2044.

Willett, W.C., Sampson, L., Browne, M.L., Stampfer, M.J., Rosner, B., Hennekens, C.H., and Speizer, F.E. 1988. The use of a self-administered questionnaire to assess diet four years in the past. *American Journal of Epidemiology* 127: 188-199.

APPENDICES

Appendix A. Regression calibration and method of moments

Suppose that (y_i, w_i), $i = 1,...,n$, are observed, and that

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

with $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2_{Y|X}$, and $Cov(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$.

Suppose a classical error structure, with $w_i = x_i + \delta_i$, $E(\delta_i) = 0$, $Var(\delta_i) = \sigma^2_M$, $Cov(\delta_i, \delta_j) = 0$ for $i \neq j$, and ε_i and δ_i independent of each other.

Suppose that $E(x_i \mid w_i) = \alpha_0 + \alpha_1 w_i$, known. Let w^*_1 represent $E(x_i \mid w_i)$.

Then, the RC estimator of β_1 is $\hat{\beta}_{1,RC} = \dfrac{S_{W^*Y}}{S_{W^*W^*}}$, where $S_{W^*Y} = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(w^*_i - \overline{w}^*\right)\left(y_i - \overline{y}\right)$ and $S_{W^*W^*} = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(w^*_i - \overline{w}^*\right)^2$. This is equivalent to $\hat{\beta}_{1,RC} = \dfrac{1}{\alpha_1}\dfrac{S_{WY}}{S_{WW}}$. If and ε_i and δ_i normally distributed, then $\alpha_1 = \lambda = \dfrac{\sigma^2_X}{\sigma^2_X + \sigma^2_M}$, the reliability of the measurements, and $E\left(\dfrac{S_{WY}}{S_{WW}}\right) = \dfrac{\beta_1\sigma^2_X}{\sigma^2_X + \sigma^2_M}$. Therefore, if $E(x_i \mid w_i)$ or α_1 are known, $E\left(\hat{\beta}_{1,RC}\right) = \beta_1$. In addition, the RC and ML estimators are equivalent.

Without any distributional assumptions, $\hat{\beta}_{1,RC}$ is a method of moments estimator, because $E(S_{WY}) = \beta_1\sigma^2_X$ and $E(S_{WW}) = \sigma^2_X + \sigma^2_M$, so $E(\lambda S_{WW}) = \sigma^2_X$.

Appendix B. Derivation of the regression calibration estimator.

B.1. Linear model

Let Y denote the response variable, X the explanatory variable measured with error, W a surrogate variable for X, and $\mathbf{Z}$ a set of q other explanatory variables. Suppose that:

$$Y = \beta_0 + \beta_1 X + \mathbf{Z}'\boldsymbol{\beta}_{\mathbf{z}} + \varepsilon \tag{B.1}$$

$$X = \alpha_0 + \alpha_1 W + \mathbf{Z}'\boldsymbol{\alpha}_{\mathbf{z}} + \delta \tag{B.2}$$

$$Y = \gamma_0 + \gamma_1 W + \mathbf{Z}'\boldsymbol{\gamma}_{\mathbf{z}} + \zeta \tag{B.3}$$

where $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \boldsymbol{\beta}'_{\mathbf{z}}\right)'$, $\boldsymbol{\alpha} = \left(\alpha_0, \alpha_1, \boldsymbol{\alpha}'_{\mathbf{z}}\right)'$ and $\boldsymbol{\gamma} = \left(\gamma_0, \gamma_1, \boldsymbol{\gamma}'_{\mathbf{z}}\right)'$ are unknown regression coefficients; and $(\varepsilon, \delta, \zeta)$ are random variables.

Interest is in inference about β_1. The study data consists on a primary sample $\left(y_i, w_i, \mathbf{z}'_i\right)$, $i = 1,...,n$, and an independent calibration sample consisting of observations $\left(x_i, w_i, \mathbf{z}'_i\right)$, $i = n+1,...,n+m$. Assume that observations with different values of i are independent. The subindex p indicates the primary study, and c the calibration study. Thus, the available data are $\left(\mathbf{y_p}, \mathbf{w_p}, \mathbf{Z_p}\right)$ and $\left(\mathbf{x_c}, \mathbf{w_c}, \mathbf{Z_c}\right)$.

The RC estimator is obtained after substituting $x_{p,i}$ in (B.1) by an estimator of its expectation, given $w_{p,i}$ and $\mathbf{z}_{p,i}$. Let this estimator be $\hat{x}_{p,i} = \hat{\alpha}_0 + \hat{\alpha}_1 w_{p,i} + \mathbf{z}'_{\mathbf{p,i}}\hat{\boldsymbol{\alpha}}_{\mathbf{z}}$, where $\hat{\boldsymbol{\alpha}} = \left(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\boldsymbol{\alpha}}_{\mathbf{z}}\right)'$ is an estimator of $\boldsymbol{\alpha}$. Then, the model actually fitted is:

$$E\left(Y \mid W = w_i, \mathbf{Z} = \mathbf{z_i}\right) = \beta_{0,RC} + \beta_{1,RC}\hat{x}_{p,i} + \mathbf{z}'_{p,i}\boldsymbol{\beta}_{\mathbf{z,RC}} = \begin{pmatrix}1 & w_{p,i} & \mathbf{z}'_{\mathbf{p,i}}\end{pmatrix}\mathbf{A}\boldsymbol{\beta}_{\mathbf{RC}}$$

where $\hat{\mathbf{A}} = \begin{pmatrix} 1 & \hat{\alpha}_0 & \mathbf{0_{1xq}} \\ 0 & \hat{\alpha}_1 & \mathbf{0_{1xq}} \\ \mathbf{0_{qx1}} & \hat{\boldsymbol{\alpha}}_{\mathbf{z}} & I_q \end{pmatrix}$, which will be assumed to be positive definite w.p.1, and $\boldsymbol{\beta}_{\mathbf{RC}} = \left(\beta_{0,RC}, \beta_{1,RC}, \boldsymbol{\beta}'_{\mathbf{z,RC}}\right)'$. Let $\mathbf{W_p} = \left(\mathbf{1}, \mathbf{w_p}, \mathbf{Z_p}\right)$. The RC estimator is:

$$\hat{\boldsymbol{\beta}}_{\mathbf{RC}} = \left(\hat{\mathbf{A}}'\mathbf{W}_{\mathbf{p}}'\mathbf{W}_{\mathbf{p}}\hat{\mathbf{A}}\right)^{-1}\hat{\mathbf{A}}'\mathbf{W}_{\mathbf{p}}'\mathbf{y}_{\mathbf{p}} = \hat{\mathbf{A}}^{-1}\left(\mathbf{W}_{\mathbf{p}}'\mathbf{W}_{\mathbf{p}}\right)^{-1}\mathbf{W}_{\mathbf{p}}'\mathbf{y}_{\mathbf{p}} = \hat{\mathbf{A}}^{\mathbf{-1}}\hat{\boldsymbol{\gamma}}$$

$$\begin{pmatrix} \hat{\beta}_{0,RC} \\ \hat{\beta}_{1,RC} \\ \hat{\boldsymbol{\beta}}_{\mathbf{z,RC}} \end{pmatrix} = \begin{pmatrix} \hat{\gamma}_0 - \hat{\alpha}_0\hat{\alpha}_1^{-1}\hat{\gamma}_1 \\ \hat{\alpha}_1^{-1}\hat{\gamma}_1 \\ \hat{\boldsymbol{\gamma}}_{\mathbf{z}} - \hat{\boldsymbol{\alpha}}_{\mathbf{z}}\hat{\alpha}_1^{-1}\hat{\gamma}_1 \end{pmatrix}$$

B.2. Generalized linear model with canonical link

Suppose that $Y|X$, $\mathbf{Z}$ follows a distribution from a regular exponential family with mean $E\left(Y \mid X, \mathbf{Z}\right)$ and a scale parameter φ. Let

$$g\left[E\left(Y \mid X = x_i, \mathbf{Z} = \mathbf{z_i}\right)\right] = \beta_0 + \beta_1 x_i + \mathbf{z}_{\mathbf{i}}'\boldsymbol{\beta}_{\mathbf{z}}$$

$$X = \alpha_0 + \alpha_1 W + \mathbf{Z}'\boldsymbol{\alpha}_{\mathbf{z}} + \delta$$

where $g\left(\cdot\right)$ is the canonical link function for the distribution of $Y \mid X$, $\mathbf{Z}$, and the symbol definitions and data structure are as in the previous section.

Let $\mathbf{w_i} = \left(1 \quad w_i \quad \mathbf{z}_{\mathbf{i}}'\right)'$. Write the log likelihood for the regression of Y on W and Z, based on the primary data, is:

$$l\left(\boldsymbol{\gamma}\right) \propto \sum_{i=1}^{n}\left\{\left[y_i\mathbf{w}_{\mathbf{i}}'\boldsymbol{\gamma} - b\left(\mathbf{w}_{\mathbf{i}}'\boldsymbol{\gamma}\right)\right]/\left(\varphi/v_i\right)\right\}$$

where $b\left(\cdot\right)$ is the cumulant generating function and v_i is a known weight.

Note that

$$\mathbf{w}_{\mathbf{i}}'\hat{\mathbf{A}} = \left(1 \quad \hat{\alpha}_0 + \hat{\alpha}_1 w_{p,i} + \mathbf{z}_{\mathbf{p,i}}'\hat{\boldsymbol{\alpha}}_{\mathbf{z}} \quad \mathbf{z}_{\mathbf{p,i}}'\right) = \left[1 \quad \hat{E}\left(X \mid W = w_i, \mathbf{Z} = \mathbf{z_i}\right) \quad \mathbf{z}_{\mathbf{p,i}}'\right] = \hat{\mathbf{x}}_{\mathbf{i}}'$$

Multiply terms in $l\left(\boldsymbol{\gamma}\right)$ by $\hat{\mathbf{A}}\hat{\mathbf{A}}^{\mathbf{-1}}$, as shown below:

$$l\left(\boldsymbol{\gamma}\right) \propto \sum_{i=1}^{n}\left\{\left[y_i\mathbf{w}_{\mathbf{i}}'\hat{\mathbf{A}}\hat{\mathbf{A}}^{\mathbf{-1}}\boldsymbol{\gamma} - b\left(\mathbf{w}_{\mathbf{i}}'\hat{\mathbf{A}}\hat{\mathbf{A}}^{\mathbf{-1}}\boldsymbol{\gamma}\right)\right]/\left(\varphi/v_i\right)\right\} = \sum_{i=1}^{n}\left\{\left[y_i\hat{\mathbf{x}}_{\mathbf{i}}'\hat{\mathbf{A}}^{\mathbf{-1}}\boldsymbol{\gamma} - b\left(\hat{\mathbf{x}}_{\mathbf{i}}'\hat{\mathbf{A}}^{\mathbf{-1}}\boldsymbol{\gamma}\right)\right]/\left(\varphi/v_i\right)\right\}$$

Therefore, the coefficients of the generalized lineal model $g\left[E\left(Y \mid X = \hat{x}_i, \mathbf{Z} = \mathbf{z_i}\right)\right]$, i.e., the RC calibration estimator, are $\hat{\mathbf{A}}^{\mathbf{-1}}\boldsymbol{\gamma}$. An alternative proof was given by Thurston et al (2003).

Appendix C. Distribution of the ratio of two independent, normally distributed random variables

C.1. Density and cumulative distribution functions

Let $\hat{\gamma}_1 \sim N\left(\gamma_1, \sigma_{\hat{\gamma}}^2\right)$, $\hat{\alpha}_1 \sim N\left(\alpha_1, \sigma_{\hat{\alpha}}^2\right)$, independent, where

$\sigma_{\hat{\gamma}}^2 = Var\left(\hat{\gamma}_1\right) = \dfrac{\sigma_{Y|W}^2}{(n-1)S_{W_p}^2}$, $\sigma_{\hat{\alpha}}^2 = Var\left(\hat{\alpha}_1\right) = \dfrac{\sigma_{X|W}^2}{(m-1)S_{W_c}^2}$, $S_{W_p}^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}\left(w_i - \overline{w}_p\right)^2$ and

$S_{W_c}^2 = \dfrac{1}{m-1}\sum_{i=n+1}^{n+m}\left(w_i - \overline{w}_c\right)^2$

Then, $\hat{\beta}_{1,RC} = \hat{\gamma}_1 / \hat{\alpha}_1$ is the ratio of two independent normal variables. The distribution of this ratio has been discussed by Hinkley (1969) and Marsaglia (1965). Here, I will discuss a different parametrization.

The joint distribution of $\hat{\gamma}_1$ and $\hat{\alpha}_1$ is:

$$f\left(\hat{\gamma}_1, \hat{\alpha}_1 \mid W\right) = \frac{1}{2\pi\sqrt{\sigma_{\hat{\gamma}}^2 \sigma_{\hat{\alpha}}^2}} \exp\left\{-\frac{1}{2}\left[\frac{\left(\hat{\gamma}_1 - \gamma_1\right)^2}{\sigma_{\hat{\gamma}}^2} + \frac{\left(\hat{\alpha}_1 - \alpha_1\right)^2}{\sigma_{\hat{\alpha}}^2}\right]\right\}$$

The probability density function of $\hat{\beta}_{1,RC}$ is obtained by solving:

$$f\left(\hat{\beta}_{1,RC} \mid W\right) = \frac{1}{2\pi\sqrt{\sigma_{\hat{\gamma}}^2 \sigma_{\hat{\alpha}}^2}} \int_{-\infty}^{\infty} \left|\hat{\alpha}_1\right| \exp\left\{-\frac{1}{2}\left[\frac{\left(\hat{\beta}_{1,RC}\hat{\alpha}_1 - \gamma_1\right)^2}{\sigma_{\hat{\gamma}}^2} + \frac{\left(\hat{\alpha}_1 - \alpha_1\right)^2}{\sigma_{\hat{\alpha}}^2}\right]\right\} d\hat{\alpha}_1$$

which yields

$$f\left(\hat{\beta}_{1,RC} \mid W\right)=\frac{\sigma_{\hat{\alpha}}}{\sigma_{\hat{\gamma}}} e^{-\frac{1}{2}\left(\frac{\gamma_1^2}{\sigma_{\hat{\gamma}}^2}+\frac{\alpha_1^2}{\sigma_{\hat{\alpha}}^2}\right)} \frac{1}{\pi\left(\frac{\sigma_{\hat{\alpha}}^2}{\sigma_{\hat{\gamma}}^2}\hat{\beta}_{1,RC}^2+1\right)}\times$$

$$\left\{1+\sqrt{\frac{\pi}{2}}\left(\frac{\frac{\sigma_{\hat{\alpha}}}{\sigma_{\hat{\gamma}}}\hat{\beta}_{1,RC}\frac{\gamma_1}{\sigma_{\hat{\gamma}}}+\frac{\alpha_1}{\sigma_{\hat{\alpha}}}}{\sqrt{\frac{\sigma_{\hat{\alpha}}^2}{\sigma_{\hat{\gamma}}^2}\hat{\beta}_{1,RC}^2+1}}\right) e^{\frac{1}{2}\left(\frac{\frac{\sigma_{\hat{\alpha}}}{\sigma_{\hat{\gamma}}}\hat{\beta}_{1,RC}\frac{\gamma_1}{\sigma_{\hat{\gamma}}}+\frac{\alpha_1}{\sigma_{\hat{\alpha}}}}{\sqrt{\frac{\sigma_{\hat{\alpha}}^2}{\sigma_{\hat{\gamma}}^2}\hat{\beta}_{1,RC}^2+1}}\right)^2}\left[2\Phi\left(\frac{\frac{\sigma_{\hat{\alpha}}}{\sigma_{\hat{\gamma}}}\hat{\beta}_{1,RC}\frac{\gamma_1}{\sigma_{\hat{\gamma}}}+\frac{\alpha_1}{\sigma_{\hat{\alpha}}}}{\sqrt{\frac{\sigma_{\hat{\alpha}}^2}{\sigma_{\hat{\gamma}}^2}\hat{\beta}_{1,RC}^2+1}}\right)-1\right]\right\}$$

where $\Phi\left(t\right)$ is the cumulative distribution function of a standard normal random variable.

This function depends only on 3 parameters, $\tau_{\gamma}=\gamma_1\Big/\sqrt{\sigma_{\hat{\gamma}}^2}$, $\tau_{\alpha}=\alpha_1\Big/\sqrt{\sigma_{\hat{\alpha}}^2}$ and the scale parameter $\eta=\sqrt{\sigma_{\hat{\alpha}}^2/\sigma_{\hat{\gamma}}^2}$. Therefore, it can be re-expressed as:

$$f\left(\hat{\beta}_{1,RC} \mid W\right)=\eta e^{-\frac{1}{2}\left(\tau_{\gamma}^2+\tau_{\alpha}^2\right)}\frac{1}{\pi\left(\eta^2\hat{\beta}_{1,RC}^2+1\right)}\times$$

$$\left\{1+\sqrt{\frac{\pi}{2}}\left(\frac{\tau_{\gamma}\eta\hat{\beta}_{1,RC}+\tau_{\alpha}}{\sqrt{\eta^2\hat{\beta}_{1,RC}^2+1}}\right)e^{\frac{1}{2}\left(\frac{\tau_{\gamma}\eta\hat{\beta}_{1,RC}+\tau_{\alpha}}{\sqrt{\eta^2\hat{\beta}_{1,RC}^2+1}}\right)^2}\left[2\Phi\left(\frac{\tau_{\gamma}\eta\hat{\beta}_{1,RC}+\tau_{\alpha}}{\sqrt{\eta^2\hat{\beta}_{1,RC}^2+1}}\right)-1\right]\right\}$$

The cumulative distribution function can be obtained by integrating the density:

$$F\left(\hat{\beta}_{1,RC} \mid W\right)=L\left\{\frac{\tau_{\gamma}-\tau_{\alpha}\eta\hat{\beta}_{1,RC}}{\left(\eta^2\hat{\beta}_{1,RC}^2+1\right)^{1/2}},-\tau_{\alpha};\frac{\eta\hat{\beta}_{1,RC}}{\left(\eta^2\hat{\beta}_{1,RC}^2+1\right)^{1/2}}\right\}$$

$$+L\left\{-\frac{\tau_{\gamma}-\tau_{\alpha}\eta\hat{\beta}_{1,RC}}{\left(\eta^2\hat{\beta}_{1,RC}^2+1\right)^{1/2}},\tau_{\alpha};\frac{\eta\hat{\beta}_{1,RC}}{\left(\eta^2\hat{\beta}_{1,RC}^2+1\right)^{1/2}}\right\}$$

where L is the cumulative distribution function of a standard bivariate normal distribution:

$$L\{h,k;\gamma\} = \frac{1}{2\pi\sqrt{1-\gamma^2}} \int_{-\infty}^{h}\int_{-\infty}^{k} \exp\left\{-\frac{x^2 - 2\gamma xy + y^2}{2\left(1-\gamma^2\right)}\right\} dxdy .$$

C.2. Asymptotics

There are two asymptotic distributions of interest: when $m \to \infty$ and when $\tau_\alpha \to \infty$. Note that when both $n \to \infty$ and $m \to \infty$, the RC estimator converges to β_1, since $\hat{\gamma}_1 \to \gamma_1$ and $\hat{\alpha}_1 \to \alpha_1$. When $n \to \infty$, $\hat{\gamma}_1 \to \gamma_1$, but the distribution of the RC estimator is proportional to the inverse of a normal distribution.

If $m \to \infty$, assuming that the variance of W is finite,

$$\eta = \frac{\sigma_{\hat{\alpha}}}{\sigma_{\hat{\gamma}}} = \sqrt{\frac{\sigma^2_{X|W} \big/ (m-1) S^2_{W_c}}{\sigma^2_{Y|W} \big/ (n-1) S^2_{W_p}}} = \sqrt{\frac{(n-1) S^2_{W_p} \sigma^2_{X|W}}{(m-1) S^2_{W_c} \sigma^2_{Y|W}}} \to 0$$

$$\tau_\alpha = \frac{\alpha_1}{\sigma_{\hat{\alpha}}} = \frac{\alpha_1}{\sqrt{\sigma^2_{X|W} \big/ (m-1) S^2_{W_c}}} \to \infty$$

$$L\left\{-\frac{\tau_\gamma - \tau_\alpha \eta \hat{\beta}_{1,RC}}{\left(\eta^2 \hat{\beta}^2_{1,RC} + 1\right)^{1/2}}, \tau_\alpha ; \frac{\eta \hat{\beta}_{1,RC}}{\left(\eta^2 \hat{\beta}^2_{1,RC} + 1\right)^{1/2}}\right\} \to \Phi\left(-\tau_\gamma + \frac{\alpha_1}{\sigma_{\hat{\gamma}}} \hat{\beta}_{1,RC}\right)$$

$$L\left\{\frac{\tau_\gamma - \tau_\alpha \eta \hat{\beta}_{1,RC}}{\left(\eta^2 \hat{\beta}^2_{1,RC} + 1\right)^{1/2}}, -\tau_\alpha ; \frac{\eta \hat{\beta}_{1,RC}}{\left(\eta^2 \hat{\beta}^2_{1,RC} + 1\right)^{1/2}}\right\} \to 0$$

Then, $F\left(\hat{\beta}_{1,RC} \mid W\right) \to \Phi\left(-\tau_\gamma + \frac{\alpha_1}{\sigma_{\hat{\gamma}}} \hat{\beta}_{1,RC}\right)$ and $\hat{\beta}_{1,RC} \to N\left(\beta_1, \frac{\sigma^2_{\hat{\gamma}}}{\alpha_1^2}\right)$

As $\tau_\alpha \to \infty$:

$$L\left\{\frac{\tau_\gamma - \tau_\alpha \eta \hat{\beta}_{1,RC}}{\left(\eta^2 \hat{\beta}^2_{1,RC} + 1\right)^{1/2}}, -\tau_\alpha ; \frac{\eta \hat{\beta}_{1,RC}}{\left(\eta^2 \hat{\beta}^2_{1,RC} + 1\right)^{1/2}}\right\} \to 0$$

So that:

$$F\left(\hat{\beta}_{1,RC} \mid W\right) \to \Phi\left\{\frac{\alpha_1 \hat{\beta}_{1,RC} - \gamma_1}{\left(\sigma^2_{\hat{\alpha}} \hat{\beta}^2_{1,RC} + \sigma^2_{\hat{\gamma}}\right)^{1/2}}\right\}$$

The median of this distribution is β_1, because for $\hat{\beta}_{1,RC} = \gamma_1/\alpha_1$,

$\dfrac{\alpha_1(\gamma_1/\alpha_1)-\gamma_1}{\left(\sigma_{\hat{\alpha}}^2(\gamma_1/\alpha_1)^2+\sigma_{\hat{\gamma}}^2\right)^{1/2}} = 0$ and $\Phi\left\{\dfrac{\alpha_1\hat{\beta}_{1,RC}-\gamma_1}{\left(\sigma_{\hat{\alpha}}^2\hat{\beta}_{1,RC}^2+\sigma_{\hat{\gamma}}^2\right)^{1/2}}\right\} = \Phi\{0\}$. This distribution is a useful approximation to the true sampling distribution for large τ_α but, for finite τ_α, it is an improper distribution (Hinkley 1969).

C.3. Alternative form of τ_α

The parameter τ_α was defined as $\tau_\alpha = \alpha_1\big/\sqrt{\sigma_{\hat{\alpha}}^2}$, where α_1 is the slope of the regression of X on W and $\sigma_{\hat{\alpha}}^2$ is the variance of the estimator of α_1. Then,

$$\alpha_1 = \frac{Cov(X,W)}{Var(W)},\ \sigma_{\hat{\alpha}}^2 = \frac{\sigma_{X|W}^2}{(n-1)S_{Wc}^2},\ \sigma_{X|W}^2 = Var(X) - \frac{Cov(X,W)^2}{Var(W)}$$

$$\tau_\alpha^2 = \frac{\alpha_1^2}{\sigma_{\hat{\alpha}}^2} = (m-1)S_{Wc}^2\frac{Cov(X,W)^2\big/Var(W)}{Var(X)Var(W)-Cov(X,W)^2} = (m-1)\frac{S_{Wc}^2}{Var(W)}\frac{\rho^2}{1-\rho^2},$$

where ρ is the correlation between X and W.

Appendix D. Maximum likelihood estimation

D.1. Unrestricted likelihood

Suppose that:

$$y_i \mid x_i, w_i, \mathbf{z_i} = \beta_0 + \beta_1 x_i + \mathbf{z_i'}\boldsymbol{\beta}_\mathbf{z} + \varepsilon_i, \qquad i = 1,...,n$$

$$x_i \mid w_i, \mathbf{z_i} = \alpha_0 + \alpha_1 w_i + \mathbf{z_i'}\boldsymbol{\alpha}_\mathbf{z} + \delta_i, \qquad i = n+1,...,n+m$$

where $\mathbf{z_i}$ is a vector of explanatory variables measured without error, and $(\varepsilon_i, \delta_i)$ are distributed normally and independently, with mean (0, 0) and variances $\left(\sigma^2_{Y|X}, \sigma^2_{X|W}\right)$.

Then, $y_i \mid w_i, \mathbf{z_i}$ is normally distributed with mean and variance:

$$E\left(y_i \mid w_i, \mathbf{z_i}\right) = \left(\beta_0 + \beta_1\alpha_0\right) + \beta_1\alpha_1 w_i + \mathbf{z_i'}\left(\beta_1\boldsymbol{\alpha}_\mathbf{z} + \boldsymbol{\beta}_\mathbf{z}\right)$$

$$Var\left(y_i \mid w_i, \mathbf{z_i}\right) = \sigma^2_{Y|X} + \beta_1^2\sigma^2_{X|W}$$

Denote the coefficients the regression of y_i on w_i and $\mathbf{z_i}$ as $\gamma_0 = \beta_0 + \beta_1\alpha_0$, $\gamma_1 = \beta_1\alpha_1$, $\boldsymbol{\gamma}_\mathbf{z} = \beta_1\boldsymbol{\alpha}_\mathbf{z} + \boldsymbol{\beta}_\mathbf{z}$.

The log-likelihood is:

$$l\left(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2_{Y|X}, \sigma^2_{X|W}\right) = -\frac{m+n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^2_{Y|X} + \beta_1^2\sigma^2_{X|W}\right) - \frac{m}{2}\log\left(\sigma^2_{X|W}\right)$$
$$-\frac{1}{2\left(\sigma^2_{Y|X} + \beta_1^2\sigma^2_{X|W}\right)}\sum_{i=1}^{n}\left\{y_i - \left[\beta_0 + \beta_1\alpha_0 + \beta_1\alpha_1 w_i + \mathbf{z_i'}\left(\beta_1\boldsymbol{\alpha}_\mathbf{z} + \boldsymbol{\beta}_\mathbf{z}\right)\right]\right\}^2 \qquad \text{(D.1)}$$
$$-\frac{1}{2\sigma^2_{X|W}}\sum_{i=n+1}^{n+m}\left\{x_i - \left(\alpha_0 + \alpha_1 w_i + \mathbf{z_i'}\boldsymbol{\alpha}_\mathbf{z}\right)\right\}^2$$

Taking derivatives with respect to $\left(\beta_0, \beta_1, \boldsymbol{\beta}_\mathbf{z}, \alpha_0, \alpha_1, \boldsymbol{\alpha}_\mathbf{z}, \sigma^2_{Y|X}, \sigma^2_{X|W}\right)$ and equating to 0, the likelihood equations are:

$$\overline{y}_p - \hat{\beta}_0 - \hat{\beta}_1\hat{\alpha}_0 - \hat{\beta}_1\hat{\alpha}_1\overline{w}_p - \overline{\mathbf{z}}_\mathbf{p}'\left(\hat{\beta}_1\hat{\boldsymbol{\alpha}}_\mathbf{z} + \hat{\boldsymbol{\beta}}_\mathbf{z}\right) = 0 \qquad \text{(D.2)}$$

$$
\begin{aligned}
&-n\hat{\beta}_1\hat{\sigma}^2_{X|W} + \frac{\hat{\beta}_1\hat{\sigma}^2_{X|W}}{\hat{\sigma}^2_{Y|X} + \hat{\beta}^2_1\hat{\sigma}^2_{X|W}} \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1\hat{\alpha}_0 - \hat{\beta}_1\hat{\alpha}_1 w_i - \mathbf{z}'_{\mathbf{p,i}}\left(\hat{\beta}_1\hat{\boldsymbol{\alpha}}_{\mathbf{z}} + \hat{\boldsymbol{\beta}}_{\mathbf{z}}\right)\right)^2 \\
&+n\hat{\alpha}_0\left(\overline{y}_p - \hat{\beta}_0 - \hat{\beta}_1\hat{\alpha}_0 - \hat{\beta}_1\hat{\alpha}_1\overline{w}_p - \overline{\mathbf{z}}'_{\mathbf{p}}\left(\hat{\beta}_1\hat{\boldsymbol{\alpha}}_{\mathbf{z}} + \hat{\boldsymbol{\beta}}_{\mathbf{z}}\right)\right) \\
&+\hat{\alpha}_1\sum_{i=1}^{n} w_i\left(y_i - \hat{\beta}_0 - \hat{\beta}_1\hat{\alpha}_0 - \hat{\beta}_1\hat{\alpha}_1 w_i - \mathbf{z}'_{\mathbf{p,i}}\left(\hat{\beta}_1\hat{\boldsymbol{\alpha}}_{\mathbf{z}} + \hat{\boldsymbol{\beta}}_{\mathbf{z}}\right)\right) \\
&+\hat{\boldsymbol{\alpha}}'_{\mathbf{z}}\sum_{i=1}^{n} \mathbf{z}_{\mathbf{i}}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1\hat{\alpha}_0 - \hat{\beta}_1\hat{\alpha}_1 w_i - \mathbf{z}'_{\mathbf{i}}\left(\hat{\beta}_1\hat{\boldsymbol{\alpha}}_{\mathbf{z}} + \hat{\boldsymbol{\beta}}_{\mathbf{z}}\right)\right) = 0
\end{aligned}
\tag{D.3}
$$

$$
\sum_{i=1}^{n} \mathbf{z}_{\mathbf{i}}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1\hat{\alpha}_0 - \hat{\beta}_1\hat{\alpha}_1 w_i - \mathbf{z}'_{\mathbf{i}}\left(\hat{\beta}_1\hat{\boldsymbol{\alpha}}_{\mathbf{z}} + \hat{\boldsymbol{\beta}}_{\mathbf{z}}\right)\right) = 0 \tag{D.4}
$$

$$
\begin{aligned}
&\frac{n\hat{\beta}_1}{\hat{\sigma}^2_{Y|X} + \hat{\beta}^2_1\hat{\sigma}^2_{X|W}}\left(\overline{y}_p - \hat{\beta}_0 - \hat{\beta}_1\hat{\alpha}_0 - \hat{\beta}_1\hat{\alpha}_1\overline{w}_p - \overline{\mathbf{z}}'_{\mathbf{p}}\left(\hat{\beta}_1\hat{\boldsymbol{\alpha}}_{\mathbf{z}} + \hat{\boldsymbol{\beta}}_{\mathbf{z}}\right)\right) \\
&+\frac{m}{\hat{\sigma}^2_{X|W}}\left(\overline{x}_c - \hat{\alpha}_0 - \hat{\alpha}_1\overline{w}_c - \overline{\mathbf{z}}'_{\mathbf{c}}\hat{\boldsymbol{\alpha}}_{\mathbf{z}}\right) = 0
\end{aligned}
\tag{D.5}
$$

$$
\begin{aligned}
&\frac{\hat{\beta}_1}{\hat{\sigma}^2_{Y|X} + \hat{\beta}^2_1\hat{\sigma}^2_{X|W}}\sum_{i=1}^{n} w_i\left(y_i - \hat{\beta}_0 - \hat{\beta}_1\hat{\alpha}_0 - \hat{\beta}_1\hat{\alpha}_1 w_i - \mathbf{z}'_{\mathbf{i}}\left(\hat{\beta}_1\hat{\boldsymbol{\alpha}}_{\mathbf{z}} + \hat{\boldsymbol{\beta}}_{\mathbf{z}}\right)\right) \\
&+\frac{1}{\hat{\sigma}^2_{X|W}}\sum_{i=n+1}^{n+m} w_i\left(x_i - \hat{\alpha}_0 - \hat{\alpha}_1 w_i - \mathbf{z}'_{\mathbf{i}}\hat{\boldsymbol{\alpha}}_{\mathbf{z}}\right) = 0
\end{aligned}
\tag{D.6}
$$

$$
\begin{aligned}
&\frac{\hat{\beta}_1}{\hat{\sigma}^2_{Y|X} + \hat{\beta}^2_1\hat{\sigma}^2_{X|W}}\sum_{i=1}^{n} \mathbf{z}_{\mathbf{i}}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1\hat{\alpha}_0 - \hat{\beta}_1\hat{\alpha}_1 w_i - \mathbf{z}'_{\mathbf{i}}\left(\hat{\beta}_1\hat{\boldsymbol{\alpha}}_{\mathbf{z}} + \hat{\boldsymbol{\beta}}_{\mathbf{z}}\right)\right) \\
&+\frac{1}{\hat{\sigma}^2_{X|W}}\sum_{i=n+1}^{n+m} \mathbf{z}_{\mathbf{i}}\left(x_i - \hat{\alpha}_0 - \hat{\alpha}_1 w_i - \mathbf{z}'_{\mathbf{i}}\hat{\boldsymbol{\alpha}}_{\mathbf{z}}\right) = 0
\end{aligned}
\tag{D.7}
$$

$$
-n + \frac{1}{\hat{\sigma}^2_{Y|X} + \hat{\beta}^2_1\hat{\sigma}^2_{X|W}}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1\hat{\alpha}_0 - \hat{\beta}_1\hat{\alpha}_1 w_i - \mathbf{z}'_{\mathbf{i}}\left(\hat{\beta}_1\hat{\boldsymbol{\alpha}}_{\mathbf{z}} + \hat{\boldsymbol{\beta}}_{\mathbf{z}}\right)\right)^2 = 0 \tag{D.8}
$$

$$
\begin{aligned}
&-\frac{n\hat{\beta}^2_1}{\hat{\sigma}^2_{Y|X} + \hat{\beta}^2_1\hat{\sigma}^2_{X|W}} - \frac{m}{\hat{\sigma}^2_{X|W}} + \frac{1}{\left(\hat{\sigma}^2_{X|W}\right)^2}\sum_{i=n+1}^{n+m}\left(x_i - \hat{\alpha}_0 - \hat{\alpha}_1 w_i - \mathbf{z}'_{\mathbf{i}}\hat{\boldsymbol{\alpha}}_{\mathbf{z}}\right)^2 \\
&+\frac{\hat{\beta}^2_1}{\left(\hat{\sigma}^2_{Y|X} + \hat{\beta}^2_1\hat{\sigma}^2_{X|W}\right)^2}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1\hat{\alpha}_0 - \hat{\beta}_1\hat{\alpha}_1 w_i - \mathbf{z}'_{\mathbf{i}}\left(\hat{\beta}_1\hat{\boldsymbol{\alpha}}_{\mathbf{z}} + \hat{\boldsymbol{\beta}}_{\mathbf{z}}\right)\right)^2 = 0
\end{aligned}
\tag{D.9}
$$

D.1.1. Solution to likelihood equations for $\left(\alpha_0, \alpha_1, \boldsymbol{\alpha}_{\mathbf{z}}, \sigma^2_{X|W}\right)$

Substituting eq. (D.2) on eq. (D.5), eq. (D.4) on eq. (D.7), eq. (D.8) on eq. (D.9), and assuming that $\hat{\sigma}^2_{X|W} \neq 0$, the following equations are obtained:

$$\overline{x}_c - \hat{\alpha}_0 - \hat{\alpha}_1 \overline{w}_c - \overline{\mathbf{z}}'_{\mathbf{c}} \hat{\boldsymbol{\alpha}}_{\mathbf{z}} = 0 \tag{D.10}$$

$$\sum_{i=n+1}^{n+m} \mathbf{z}_{\mathbf{i}} \left(x_i - \hat{\alpha}_0 - \hat{\alpha}_1 w_i - \mathbf{z}'_{\mathbf{i}} \hat{\boldsymbol{\alpha}}_{\mathbf{z}} \right) = 0 \tag{D.11}$$

$$-m + \frac{1}{\hat{\sigma}^2_{X|W}} \sum_{i=n+1}^{n+m} \left(x_i - \hat{\alpha}_0 - \hat{\alpha}_1 w_i - \mathbf{z}'_{\mathbf{i}} \hat{\boldsymbol{\alpha}}_{\mathbf{z}} \right)^2 = 0 \tag{D.12}$$

Substituting eqs. (D.2), (D.4) and (D.7) on eq. (D.3) yields:

$$\hat{\alpha}_1 \sum_{i=1}^{n} w_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 \hat{\alpha}_0 - \hat{\beta}_1 \hat{\alpha}_1 w_i - \mathbf{z}'_{\mathbf{i}} \left(\hat{\beta}_1 \hat{\boldsymbol{\alpha}}_{\mathbf{z}} + \hat{\boldsymbol{\beta}}_{\mathbf{z}} \right) \right) = 0 \tag{D.13}$$

Then, under the additional assumption that $\hat{\alpha}_1 \neq 0$, substituting eq. (D.13) on eq. (D.6) yields:

$$\sum_{i=n+1}^{n+m} w_i \left(x_i - \hat{\alpha}_0 - \hat{\alpha}_1 w_i - \mathbf{z}'_{\mathbf{i}} \hat{\boldsymbol{\alpha}}_{\mathbf{z}} \right) = 0 \tag{D.14}$$

Equations (D.10), (D.11), (D.12) and (D.14) are the likelihood equations for $\left(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\boldsymbol{\alpha}}_{\mathbf{z}}, \hat{\sigma}^2_{X|W}\right)$ based on the calibration data alone. Therefore, the estimator is obtained by ignoring the primary data and relying on the calibration study alone.

D.1.2 Solution to likelihood equations for $\left(\beta_0, \beta_1, \boldsymbol{\beta}_{\mathbf{z}}, \sigma^2_{Y|X}\right)$

Equations (D.2), (D.13), (D.4) and (D.8) are the likelihood equations for the regression of Y on W and Z. If the estimators of the parameters of that regression are $\left(\hat{\gamma}_0, \hat{\gamma}_1, \hat{\boldsymbol{\gamma}}_{\mathbf{z}}, \hat{\sigma}^2_{Y|W}\right)$, then from the definition of $\left(\gamma_0, \gamma_1, \boldsymbol{\gamma}_{\mathbf{z}}\right)$ and the invariance property of the MLE, it follows that:

$$\hat{\beta}_1 = \hat{\gamma}_1 / \hat{\alpha}_1 \tag{D.15}$$

$$\hat{\beta}_0 = \hat{\gamma}_0 - \hat{\beta}_1 \hat{\alpha}_0 \tag{D.16}$$

$$\hat{\boldsymbol{\beta}}_{\mathbf{z}} = \hat{\boldsymbol{\gamma}}_{\mathbf{z}} - \hat{\beta}_1 \hat{\boldsymbol{\alpha}}_{\mathbf{z}} \tag{D.17}$$

$$\hat{\sigma}^2_{Y|X} = \hat{\sigma}^2_{Y|W} - \hat{\beta}_1^2 \hat{\sigma}^2_{X|W} \tag{D.18}$$

Therefore, if $\hat{\sigma}^2_{Y|X} + \hat{\beta}_1^2 \hat{\sigma}^2_{X|W} \neq 0$, $\hat{\sigma}^2_{X|W} \neq 0$, and $\hat{\alpha}_1 \neq 0$, the solutions to the likelihood equations for $\left(\boldsymbol{\alpha}, \sigma^2_{X|W}\right)$ are the MLEs obtained from the calibration data only. The solutions to the likelihood equations for $\left(\boldsymbol{\beta}, \sigma^2_{Y|X}\right)$ are the RC estimators. However, these estimators are the MLEs only if $\hat{\sigma}^2_{Y|X} = \hat{\sigma}^2_{Y|W} - \left(\hat{\gamma}_1 / \hat{\alpha}_1\right)^2 \hat{\sigma}^2_{X|W} > 0$.

D.2. Likelihood equations when $\hat{\sigma}^2_{Y|X} < 0$

If $\hat{\sigma}^2_{Y|X} < 0$, typically $\sigma^2_{Y|X}$ is set to 0 and the remaining parameters are estimated under this assumption. This is equivalent to maximizing the log-likelihood under the restriction that $\hat{\sigma}^2_{Y|X} > 0$. For simplicity, I will derive the estimator when there are no additional covariates **z**. Then, the log likelihood becomes:

$$\begin{aligned}
l\left(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2_{Y|X} = 0, \sigma^2_{X|W}\right) &= -\frac{n}{2}\log\left(\beta_1^2 \sigma^2_{X|W}\right) - \frac{m}{2}\log\left(\sigma^2_{X|W}\right) \\
&\quad - \frac{1}{2\beta_1^2 \sigma^2_{X|W}} \sum_{i=1}^{n} \left[y_i - \left(\beta_0 + \beta_1\alpha_0 + \beta_1\alpha_1 w_i\right)\right]^2 - \frac{1}{2\sigma^2_{X|W}} \sum_{i=n+1}^{n+m} \left[x_i - \left(\alpha_0 + \alpha_1 w_i\right)\right]^2 \\
&= -\frac{n}{2}\log\left(\beta_1^2\right) - \frac{n+m}{2}\log\left(\sigma^2_{X|W}\right) \\
&\quad - \frac{1}{2\beta_1^2 \sigma^2_{X|W}} \sum_{i=1}^{n} \left(y_{p,i} - \beta_0 - \beta_1\alpha_0 - \beta_1\alpha_1 w_{p,i}\right)^2 - \frac{1}{2\sigma^2_{X|W}} \sum_{i=n+1}^{n+m} \left(x_i - \alpha_0 - \alpha_1 w_i\right)^2
\end{aligned} \tag{D.19}$$

Taking derivatives and simplifying under $\hat{\beta}_1 \hat{\sigma}^2_{X|W} \neq 0$, the likelihood equations are:

$$\bar{y}_p - \hat{\beta}_0 - \hat{\beta}_1\hat{\alpha}_0 - \hat{\beta}_1\hat{\alpha}_1 \bar{w}_p = 0 \tag{D.20}$$

$$-n+\frac{1}{\hat{\beta}_1\hat{\sigma}^2_{X|W}}\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1\hat{\alpha}_0-\hat{\beta}_1\hat{\alpha}_1 w_i\right)^2 + \frac{\hat{\alpha}_1}{\hat{\beta}_1\hat{\sigma}^2_{X|W}}\sum_{i=1}^{n} w_i\left(y_i-\hat{\beta}_0-\hat{\beta}_1\hat{\alpha}_0-\hat{\beta}_1\hat{\alpha}_1 w_i\right)=0 \quad \text{(D.21)}$$

$$\bar{x}_c-\hat{\alpha}_0-\hat{\alpha}_1\bar{w}_c=0 \quad \text{(D.22)}$$

$$\frac{1}{\hat{\beta}_1}\sum_{i=1}^{n} w_i\left(y_i-\hat{\beta}_0-\hat{\beta}_1\hat{\alpha}_0-\hat{\beta}_1\hat{\alpha}_1 w_i\right)+\sum_{i=n+1}^{n+m} w_i\left(x_i-\hat{\alpha}_0-\hat{\alpha}_1 w_i\right)=0 \quad \text{(D.23)}$$

$$-\left(n+m\right)+\frac{1}{\hat{\beta}_1\hat{\sigma}^2_{X|W}}\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1\hat{\alpha}_0-\hat{\beta}_1\hat{\alpha}_1 w_i\right)^2 + \frac{1}{\hat{\sigma}^2_{X|W}}\sum_{i=n+1}^{n+m}\left(x_i-\hat{\alpha}_0-\hat{\alpha}_1 w_i\right)^2=0 \quad \text{(D.24)}$$

The solution to these equations is tedious, but after some algebra they yield:

$$n\left[SS^2_{XWc}-SS_{XX}\left(SS_{WpWp}+SS_{WcWc}\right)\right]\hat{\beta}_1^2+\left(n-m\right)SS_{XWc}SS_{YWp}\hat{\beta}_1 - m\left[SS^2_{YWp}-SS_{YY}\left(SS_{WpWp}+SS_{WcWc}\right)\right]=0 \quad \text{(D.25)}$$

$$\hat{\alpha}_1=\frac{\frac{1}{\hat{\beta}_1}SS_{YWp}+SS_{XWc}}{SS_{WpWp}+SS_{WcWc}} \quad \text{(D.26)}$$

$$\hat{\sigma}^2_{X|W}=\frac{SS_{XX}-\hat{\alpha}_1 SS_{XWc}}{m} \quad \text{(D.27)}$$

$$\hat{\alpha}_0=\bar{x}_c-\hat{\alpha}_1\bar{w}_c \quad \text{(D.28)}$$

$$\hat{\beta}_0=\bar{y}_p-\hat{\beta}_1\hat{\alpha}_0-\hat{\beta}_1\hat{\alpha}_1\bar{w}_p \quad \text{(D.29)}$$

where SS denote sum of cross products indicated by their subscripts.

D.3. The MLE of β_1 is bounded

Define the MLE of β_1 as $\hat{\beta}_1=\hat{\gamma}_1/\hat{\alpha}_1$ if $\hat{\sigma}^2_{Y|W}-\left(\hat{\gamma}_1/\hat{\alpha}_1\right)^2\hat{\sigma}^2_{X|W}>0$, or the solution to (D.25):

$$n\left[SS_{XWc}^2 - SS_{XX}\left(SS_{WpWp} + SS_{WcWc}\right)\right]\hat{\beta}_1^2 + (n-m)SS_{XWc}SS_{YWp}\hat{\beta}_1$$
$$-m\left[SS_{YWp}^2 - SS_{YY}\left(SS_{WpWp} + SS_{WcWc}\right)\right] = 0$$

if $\hat{\sigma}_{Y|W}^2 - \left(\hat{\gamma}_1/\hat{\alpha}_1\right)^2 \hat{\sigma}_{X|W}^2 \le 0$. $\left(\hat{\gamma}_1, \hat{\sigma}_{Y|W}^2\right)$ and $\left(\hat{\alpha}_1, \hat{\sigma}_{X|W}^2\right)$ are the MLEs of the parameters of the distribution of *Y* given *W*, based on the primary data only, and of *X* given *W*, based on the calibration data only, respectively.

Then, assuming that the cross products $SS_{XX} > 0$, $SS_{WpWp} > 0$ and $SS_{WcWc} > 0$, the MLE of β_1 is bounded.

Proof:

The proof has three parts. First, it will be shown that the roots of equation (D.25), are both real, one positive and one negative. Second, it will be shown that the MLE is the smallest, in magnitude, of $\hat{\gamma}_1/\hat{\alpha}_1$ or one of the roots of (D.25). Third, it will be show that the roots of (D.25) are finite.

Let $A = n\left[SS_{XWc}^2 - SS_{XX}\left(SS_{WpWp} + SS_{WcWc}\right)\right]$, $B = (n-m)SS_{XWc}SS_{YWp}$, and $C = m\left[SS_{YWp}^2 - SS_{YY}\left(SS_{WpWp} + SS_{WcWc}\right)\right]$.

(1) *Existence of real roots of eq. (D.25).* By Holder's inequality, $SS_{XWc}^2 \le SS_{XX}SS_{WcWc}$ and $SS_{YWp}^2 \le SS_{YY}SS_{WpWp}$. Therefore, A<0 and C<0. The argument of the square root of the formula for the roots to equation (D.25) is $B^2 + 4A*C \ge 0$, so both roots are real numbers. In addition, $|B|$ is smaller than or equal to the positive solution of $\sqrt{B^2 + 4A*C}$. Therefore, one root is positive and the other negative.

(2) *MLE is the smallest, in magnitude, of $\hat{\gamma}_1/\hat{\alpha}_1$ or the solution to (D.25).* Let

$$f(\beta_1) = n\beta_1^2\left[SS_{XWc}^2 - SS_{XX}\left(SS_{WpWp} + SS_{WcWc}\right)\right] + (n-m)\beta_1 SS_{XWc}SS_{YWp} - m\left[SS_{YWp}^2 - SS_{YY}\left(SS_{WpWp} + SS_{WcWc}\right)\right] \tag{D.30}$$

Re-express the condition $\hat{\sigma}_{Y|W}^2 - \left(\hat{\gamma}_1/\hat{\alpha}_1\right)^2 \hat{\sigma}_{X|W}^2$ in terms of sums of squares as:

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1\hat{\alpha}_0-\hat{\beta}_1\hat{\alpha}_1 w_i\right)^2-\frac{\hat{\gamma}_1^2}{\hat{\alpha}_1^2}\frac{1}{m}\sum_{i=n+1}^{n+m}\left(x_i-\hat{\alpha}_0-\hat{\alpha}_1 w_i\right)^2=$$

$$\frac{1}{n}\left(SS_{YY}-\frac{SS^2{}_{YWp}}{SS_{WpWp}}\right)-\left(\frac{SS_{YWp}/SS_{WpWp}}{SS_{XWc}/SS_{WcWc}}\right)^2\frac{1}{m}\left(SS_{XX}-\frac{SS^2{}_{XWc}}{SS_{WcWc}}\right)=$$

$$m\left(SS_{YY}SS_{WpWp}-SS^2{}_{YWp}\right)-n\frac{SS^2{}_{YWp}SS_{WcWc}}{SS^2{}_{XWc}SS_{WpWp}}\left(SS_{XX}SS_{WcWc}-SS^2{}_{XWc}\right) \qquad \text{(D.31)}$$

because $\dfrac{\hat{\gamma}_1}{\hat{\alpha}_1}=\dfrac{SS_{YWp}/SS_{WpWp}}{SS_{XWc}/SS_{WcWc}}$.

Then,

$$f\left(\hat{\gamma}_1/\hat{\alpha}_1\right)=n\left(\frac{\hat{\gamma}_1}{\hat{\alpha}_1}\right)^2\left[SS^2_{XWc}-SS_{XX}\left(SS_{WpWp}+SS_{WcWc}\right)\right]$$

$$+\left(n-m\right)\frac{\hat{\gamma}_1}{\hat{\alpha}_1}SS_{XWc}SS_{YWp}-m\left[SS^2_{YWp}-SS_{YY}\left(SS_{WpWp}+SS_{WcWc}\right)\right]=$$

$$\left[m\left(SS_{YY}SS_{WpWp}-SS^2{}_{YWp}\right)-n\frac{SS^2{}_{YWp}SS_{WcWc}}{SS^2{}_{XWc}SS_{WpWp}}\left(SS_{XX}SS_{WcWc}-SS^2_{XWc}\right)\right]\left(1+\frac{SS_{WcWc}}{SS_{WpWp}}\right)$$

$$f\left(\hat{\gamma}_1/\hat{\alpha}_1\right)=\left[\hat{\sigma}^2_{Y|W}-\left(\hat{\gamma}_1/\hat{\alpha}_1\right)^2\hat{\sigma}^2_{X|W}\right]\left(1+\frac{SS_{WcWc}}{SS_{WpWp}}\right) \qquad \text{(D.32)}$$

The second derivative of $f\left(\beta_1\right)$ is negative, therefore, the quadratic function has a maximum. Since both roots are real, the value of $f\left(\beta_1\right)$ at the maximum is positive. Therefore, if $f\left(\hat{\gamma}_1/\hat{\alpha}_1\right)>0$, $\hat{\gamma}_1/\hat{\alpha}_1$ is bounded by the roots of (D.25). If $f\left(\hat{\gamma}_1/\hat{\alpha}_1\right)<0$, $\hat{\gamma}_1/\hat{\alpha}_1$ is smaller than the smallest root of (D.25), or greater than the greatest root.

If $\hat{\sigma}^2_{Y|W}-\left(\hat{\gamma}_1/\hat{\alpha}_1\right)^2\hat{\sigma}^2_{X|W}>0$, then $\hat{\gamma}_1/\hat{\alpha}_1$ is the MLE and, by (D.32), is bounded by the roots of (D.25). The opposite holds if $\hat{\sigma}^2_{Y|W}-\left(\hat{\gamma}_1/\hat{\alpha}_1\right)^2\hat{\sigma}^2_{X|W}<0$.

(3) *The roots of (D.25) are bounded*. The denominator of the formula for the roots of the quadratic equation (D.25) is:

$$\left|2n\left(SS_{XWc}^2 - SS_{XX}SS_{WcWc} - SS_{XX}SS_{WpWp}\right)\right| \geq 2nSS_{XX}SS_{WpWp}$$

because $SS_{XWc}^2 \leq SS_{XX}SS_{WcWc}$. Since the numerator is finite, the roots are of D.25 are bounded.

Appendix E. $\mathbf{P\left(\hat{\sigma}^2_{Y|W}-\left(\hat{\gamma}_1/\hat{\alpha}_1\right)^2\hat{\sigma}^2_{X|W}>0\right)}$

The expression above can be restated as

$$P\left(\hat{\sigma}^2_{Y|W}-\left(\hat{\gamma}_1/\hat{\alpha}_1\right)^2\hat{\sigma}^2_{X|W}>0\right)=P\left(\frac{\hat{\gamma}_1^2}{\hat{\sigma}^2_{Y|W}}<\frac{\hat{\alpha}_1^2}{\hat{\sigma}^2_{X|W}}\right)$$

Under the assumption of section 2.3.1., the distributions of $\left(\hat{\alpha}_1^2,\hat{\gamma}_1^2,\hat{\sigma}^2_{X|W},\hat{\sigma}^2_{Y|W}\right)$ are:

$$\frac{(m-1)S^2_{Wc}}{\sigma^2_{X|W}}\hat{\alpha}_1^2\sim\chi^2\left(1,\frac{(m-1)S^2_{Wc}\alpha_1^2}{\sigma^2_{X|W}}\right),\ \frac{(n-1)S^2_{Wp}}{\sigma^2_{Y|W}}\hat{\gamma}_1^2\sim\chi^2\left(1,\frac{(n-1)S^2_{Wp}\gamma_1^2}{\sigma^2_{Y|W}}\right)$$

$$\frac{m\hat{\sigma}^2_{X|W}}{\sigma^2_{X|W}}\sim\chi^2_{(m-2)}.\ \frac{n\hat{\sigma}^2_{Y|W}}{\sigma^2_{Y|W}}\sim\chi^2_{(n-2)}$$

Then,

$$\frac{(n-1)S^2_{Wp}\hat{\gamma}_1^2}{\left[n/(n-2)\right]\hat{\sigma}^2_{Y|W}}\sim F\left(1,n-2,\frac{(n-1)S^2_{Wp}\gamma_1^2}{\sigma^2_{Y|W}}\right)$$

$$\frac{(m-1)S^2_{Wc}\hat{\alpha}_1^2}{\left[m/(m-2)\right]\hat{\sigma}^2_{X|W}}\sim F\left(1,m-2,\frac{(m-1)S^2_{Wc}\alpha_1^2}{\sigma^2_{X|W}}\right)$$

As $m\to\infty$,

$$P\left(\frac{\hat{\gamma}_1^2}{\hat{\sigma}^2_{Y|W}}<\frac{\hat{\alpha}_1^2}{\hat{\sigma}^2_{X|W}}\right)\xrightarrow{m\to\infty}P\left(\frac{\hat{\gamma}_1^2}{\hat{\sigma}^2_{Y|W}}<\frac{\alpha_1^2}{\sigma^2_{X|W}}\right)$$

$$=P\left(F\left(1,n-2,\frac{(n-1)S^2_{Wp}\gamma_1^2}{\sigma^2_{Y|W}}\right)<\frac{(n-1)S^2_{Wp}\alpha_1^2}{\left[n/(n-2)\right]\sigma^2_{X|W}}\right)$$

This asymptotic probability of $\hat{\sigma}^2_{Y|W}-\left(\hat{\gamma}_1/\hat{\alpha}_1\right)^2\hat{\sigma}^2_{X|W}<0$ can be large. For example, for the simulations of section 2.6, if $\beta_1=2$, the asymptotic probability as $m\to\infty$ is negligible for $\sigma^2_{X|W}=0.25^2$, 0.01 for $\sigma^2_{X|W}=0.5^2$, and 0.20 for $\sigma^2_{X|W}=0.75^2$. If $\beta_1=0.5$ and $\sigma^2_{X|W}=0.75^2$, the asymptotic probability is negligible.

Appendix F. Profile likelihood for β_1

The profile likelihood for β_1 is obtained by first solving equations (D.2) and (D.4)-(D.9) as a function of β_1, and then substituting in (D.1). For simplicity, it will be assumed that there are no additional covariates $\mathbf{z_i}$. Solving those equations, the estimator $\hat{\alpha}_1(\beta_1)$ is obtained from the following cubic equation

$$\begin{aligned}
&-\left[(n+m)\beta_1^2 SS_{W_cW_c} SS_{W_pW_p}\right]\hat{\alpha}_1^3 \\
&+\left[(n+2m)\beta_1 SS_{W_cW_c} SS_{YW_p} + (2n+m)\beta_1^2 SS_{XW_c} SS_{W_pW_p}\right]\hat{\alpha}_1^2 \\
&-\left[n\beta_1^2 S_{XX} SS_{W_pW_p} + 2(n+m)\beta_1 SS_{XW_c} SS_{YW_p} + mSS_{YY} SS_{W_cW_c}\right]\hat{\alpha}_1 \\
&+\left[n\beta_1 S_{XX} SS_{YW_p} + mSS_{YY} SS_{XW_c}\right] = 0
\end{aligned} \tag{F.1}$$

Then, the estimators of the remaining parameters, as a function of β_1 are

$$\hat{\alpha}_0 = \overline{x}_c - \hat{\alpha}_1 \overline{w}_c \tag{F.2}$$

$$\hat{\beta}_0 = \overline{y}_p - \beta_1 \overline{x}_c - \beta_1 \hat{\alpha}_1 \left(\overline{w}_p - \overline{w}_c\right) \tag{F.3}$$

$$\hat{\sigma}_{X|W}^2 = \frac{1}{m}\sum_{j=1}^{m}\left(x_{c,j} - \hat{\alpha}_0 - \hat{\alpha}_1 w_{c,j}\right)^2 \tag{F.4}$$

$$\hat{\sigma}_{Y|X}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_{p,i} - \hat{\beta}_0 - \beta_1\hat{\alpha}_0 - \beta_1\hat{\alpha}_1 w_{p,i}\right)^2 - \beta_1^2\hat{\sigma}_{X|W}^2 \tag{F.5}$$

Substituting in (D.1), the unrestricted profile likelihood for β_1 is

$$\begin{aligned}
2pl(\beta_1) \propto &-n\log\left(SS_{YY} + \beta_1^2\hat{\alpha}_1^2 SS_{W_pW_p} - 2\beta_1\hat{\alpha}_1 SS_{YW_p}\right) \\
&-m\log\left(S_{XX} + \hat{\alpha}_1^2 SS_{W_cW_c} - 2\hat{\alpha}_1 SS_{XW_c}\right)
\end{aligned} \tag{F.6}$$

For some values of β_1, eq. (F.5) may yield a negative estimate of $\sigma_{Y|X}^2$. Then, $\sigma_{Y|X}^2$ can be set to 0 and the resulting likelihood equations solved to yield

$$\hat{\alpha}_1 = \left(\frac{1}{\beta_1} SS_{YWp} + SS_{WcX}\right) \Big/ \left(SS_{WpWp} + SS_{WcWc}\right)$$

$$\hat{\alpha}_0 = \overline{x}_c - \hat{\alpha}_1 \overline{w}_c$$

$$\hat{\beta}_0 = \overline{y}_p - \beta_1 \hat{\alpha}_0 - \beta_1 \hat{\alpha}_1 \overline{w}_p$$

$$\hat{\sigma}^2_{X|W} = \frac{1}{m+n}\left[\frac{1}{\beta_1^2} SS_{YY} + SS_{XX} - \left(\frac{1}{\beta_1} SS_{YWp} + SS_{XWc}\right)^2 \Big/ \left(SS_{WpWp} + SS_{WcWc}\right)\right]$$

Substituting in (D.19), the profile likelihood becomes

$$2pl(\beta_1) = -n\log\left(\beta_1^2\right) - (m+n)\log\left(\hat{\sigma}^2_{X|W}\right) \propto$$

$$m\log\beta_1^2 - (n+m)\log\left\{\begin{aligned}&\left(-SS^2{}_{XWc} + SS_{XX}SS_{WpWp} + SS_{XX}SS_{WcWc}\right)\beta_1^2 + 2SS_{YWp}SS_{XWc}\beta_1\\ &-SS^2{}_{YWp} + SS_{YY}SS_{WpWp} + SS_{YY}SS_{WcWc}\end{aligned}\right\}$$

Appendix G. Simulation details from the external calibration study

The focus of this study is the effect of the uncertainty in estimating *E(X | W)* on regression calibration inference. In section 2.3, $\tau_\alpha = \alpha_1 / Var\left(\hat{\alpha}_1\right)$ was identified as a key factor influencing the performance of the RC estimator. τ_α depends on the size of the calibration sample, the values of α_1 and $\sigma^2_{X|W}$, and the sample variance of *W* in the calibration study. The simulations explore the effect of the size of the calibration sample and $\sigma^2_{X|W}$, while holding the other two parameters constant.

Since the main interest is in the calibration component of the study, the parameters of the distribution of *Y* given *X* were chosen to reflect a strong (large β_1) and tight (small $\sigma^2_{Y|X}$) relationship between those variables. For all the situations examined, the intercept of the regression of *Y* on *X* was set to $\beta_0 = 1$, the variance of *Y* given *X* to $\sigma^2_{Y|X} = 0.6^2$, and the primary sample size to $n = 1000$. The intercept and slope of the regression of *X* on *W* were set to $\alpha_0 = 0$ and $\alpha_1 = 1$, respectively. Inference is conditional on the values of *W*, so it was chosen to be regularly spaced in (0, 1) for all scenarios. Although this is a rather artificial distributional choice, it can be thought as a most favorable scenario, because the performance of the RC estimator decreases as the sample variance of *W* in the calibration study increases. If the distribution of *W* is skewed, a more realistic scenario, the sample variance of *W* would be greater than if it is regularly distributed.

The first set of simulations fixed the slope of the regression of *Y* on *X* at $\beta_1 = 2$ and examined a combination of three values of the variance of *X* given *W*, $\sigma^2_{X|W} = \left(0.25^2, 0.5^2, 0.75^2\right)$ and six calibration sample sizes, *m* = 50, 100, 150, 300, 500 or 1000. The values of $\sigma^2_{X|W}$ were chosen so that the correlations between *X* and *W* cover the range typically observed between reference and surrogate variables in epidemiologic studies of diet-disease associations (Table G.1). A second set of simulations examined the effect of the value of β_1, which may affect the probability of negative estimates of the variance of $Y \mid X$. The slope of the regression of Y on X was set to $\beta_1 = 1/2$, the variance of $X \mid W$ at $\sigma^2_{X|W} = 0.75^2$, and the calibration sample sizes were maintained at m = 50, 100, 150, 300, 500 or 1000. A summary of the parameters of the simulation scenarios is included in table G.1. The parameters of the sampling distribution of the regression calibration estimator for each simulation situation are included in Table G.2.

For each of the 24 settings described, 2500 simulations were run. Primary and calibration samples were generated from normal distributions with the set parameters. From each sample, we calculated the RC and ML estimators and several types of confidence intervals. The performance of different estimators is typically compared in terms of bias and root mean square error (RMSE). However, given the lack of moments of the sampling distribution of the RC estimator, those measures of performance may behave erratically and lack a direct interpretation, especially for small sample sizes. As the calibration sample size increases, the sampling

Table G.1. Summary of the parameters and features of the simulation scenarios. The parameters not shown in the table are $\left(\alpha_0, \alpha_1, \beta_0, \sigma^2_{Y|X}\right) = \left(0, 1, 1, 0.6^2\right)$, n = 1000 and m = (50, 100, 150, 300, 500, 1000).

		Corr(X, W)	Corr(Y, X)	(1)	$P_{m\to\infty}\left(\hat{\sigma}^2_{Y|X} < 0\right)$
	$\sigma^2_{X|W} = 0.25^2$	0.76	0.79	2.00	0 (approx.)
$\beta_1 = 2$	$\sigma^2_{X|W} = 0.50^2$	0.50	0.88	2.96	0.01
	$\sigma^2_{X|W} = 0.75^2$	0.36	0.94	4.12	0.20
$\beta_1 = 1/2$	$\sigma^2_{X|W} = 0.75^2$	0.36	0.55	1.03	0 (approx)

(1) Difference between $E\left(Y \mid X\right)$ at the 90th and 10th percentile of the distribution of X.

Table G.2. Parameters of the sampling distribution of the RC estimator. The three parameters are $\tau_\gamma = \gamma_1 \big/ \sqrt{Var(\hat{\gamma}_1)}$, $\tau_\alpha = \alpha_1 \big/ \sqrt{Var(\hat{\alpha}_1)}$ and the scale parameter $\eta = \sqrt{Var(\hat{\alpha}_1) / Var(\hat{\gamma}_1)}$. $(\hat{\gamma}_1, \hat{\alpha}_1)$ are the least squares estimators of the regression of *Y* on *W* and *X* on *W*, respectively.

	$\beta_1 = 2$						$\beta_1 = 1/2$	
	$\sigma^2_{X\|W} = 0.25^2$		$\sigma^2_{X\|W} = 0.50^2$		$\sigma^2_{X\|W} = 0.75^2$		$\sigma^2_{X\|W} = 0.75^2$	
τ_γ	23.38		15.66		11.30		6.45	
M	τ_α	η	τ_α	η	τ_α	η	τ_α	η
50	8.16	1.54	4.08	1.92	2.72	2.08	2.72	4.74
100	11.55	1.01	5.77	1.36	3.85	1.47	3.85	3.35
150	14.14	0.83	7.07	1.11	4.71	1.20	4.71	2.74
300	20.00	0.58	10.00	0.78	6.67	0.85	6.67	1.94
500	25.82	0.45	12.91	0.61	8.61	0.66	8.61	1.50
1000	36.51	0.32	18.26	0.43	12.17	0.46	12.17	1.06

distribution of the RC estimator converges to a normal distribution and the mean and RMSE become more stable. Thus, we report the median and 2.5 and 97.5 percentiles of the distribution of the estimator, and the mean and RMSE when they stabilized.

The performance of the 95% confidence intervals was measured in terms of total coverage, left- and right-side coverage, and length. As with the estimator itself, the expected value of the length of the RC confidence intervals is not defined, so we report the median and the 2.5 and 97.5 percentiles of its distribution, and the mean and RMSE when appropriate. We computed the following confidence intervals:

(1) RC-Wald, based on the asymptotic normality of the sampling distribution of the RC estimator and the asymptotic variance of this distribution, calculated with the Delta method (eq. 2.5).

(2) RC-bootstrap: we computed two types of bootstrap confidence intervals for the RC estimator, based on the percentiles on the bootstrapped distribution and the

BCa method (Efron and Tibshirani 1993). The confidence intervals were based on a total of 2500 bootstrapped samples.

(3) LRT: obtained by inverting a χ^2 likelihood ratio test. The likelihood ratio test was calculated on a grid of values around the MLE. The set of values not rejected by this test defined the confidence interval. Grid spacing was 0.01 units.

Tables G3-G9 show detailed summaries of the results of the simulations.

Table G.3. Summary statistics of the Monte Carlo sampling distributions of the RC estimator of the slope in a simple linear regression with measurement error. The distributions *Y*|*X* and *X*|*W* are normal, with a primary sample size of n = 1000 and a calibration sample size of m. The results are based on 2500 simulated primary and calibration samples.

			Percentiles				
m	Average	RMSE	Min.	2.5%	Median	97.5%	Max.
			$\beta_1 = 2,\ \sigma^2_{X\mid W} = 0.25^2$				
50	2.0298	0.28	1.38	1.59	1.9964	2.66	4.61
100	2.0158	0.20	1.48	1.68	2.0034	2.44	2.80
150	2.0129	0.17	1.50	1.71	2.0061	2.37	2.76
300	2.0105	0.13	1.61	1.78	2.0035	2.28	2.49
500	2.0062	0.12	1.64	1.79	2.0017	2.24	2.42
1000	2.0015	0.10	1.71	1.81	1.9969	2.21	2.36
			$\beta_1 = 2,\ \sigma^2_{X\mid W} = 0.50^2$				
50	2.1667	1.66	1.06	1.32	1.9819	3.77	77.37
100	2.0623	0.42	1.12	1.46	1.9928	3.07	4.70
150	2.0539	0.34	1.24	1.52	2.0160	2.83	3.61
300	2.0181	0.25	1.37	1.59	1.9973	2.54	3.09
500	2.0098	0.20	1.34	1.65	1.9978	2.43	2.75
1000	2.0090	0.17	1.40	1.70	2.0012	2.38	2.63
			$\beta_1 = 2,\ \sigma^2_{X\mid W} = 0.75^2$				
50	2.6738	8.07	-55.32	1.14	1.9733	6.52	277.87
100	2.1822	0.78	0.92	1.26	2.0148	4.09	8.15
150	2.1002	0.59	1.01	1.32	1.9933	3.47	9.10
300	2.0433	0.38	1.18	1.44	1.9984	2.90	4.24
500	2.0297	0.30	1.10	1.52	1.9989	2.68	3.49
1000	2.0073	0.25	1.29	1.55	2.0005	2.55	2.98
			$\beta_1 = 0.5,\ \sigma^2_{X\mid W} = 0.75^2$				
50	0.3555	12.16	-600.65	0.26	0.4922	1.78	24.99
100	0.5570	0.27	0.19	0.29	0.5076	1.12	5.99
150	0.5295	0.16	0.21	0.30	0.5009	0.92	1.71
300	0.5139	0.12	0.22	0.32	0.5001	0.77	1.56
500	0.5096	0.10	0.20	0.34	0.5020	0.72	0.93
1000	0.5024	0.09	0.17	0.34	0.5002	0.68	0.82

Table G.4. Summary statistics of the Monte Carlo sampling distributions of the MLE of the slope in a simple linear regression with measurement error. The distributions of $Y|X$ and $X|W$ are normal, with a primary sample size of n = 1000 and a calibration sample size of m. The results are based on 2500 simulated primary and calibration samples.

			Percentiles				
m	Mean	RMSE	Min.	2.5%	Median	97.5%	Max.
			$\beta_1 = 2,\ \sigma^2_{X\|W} = 0.25^2$				
50	2.0289	0.27	1.38	1.59	1.9963	2.66	3.54
100	2.0158	0.20	1.48	1.68	2.0033	2.44	2.80
150	2.0129	0.17	1.50	1.71	2.0061	2.37	2.76
300	2.0105	0.13	1.61	1.78	2.0035	2.28	2.49
500	2.0062	0.12	1.64	1.79	2.0017	2.24	2.42
1000	2.0015	0.10	1.71	1.81	1.9969	2.21	2.36
			$\beta_1 = 2,\ \sigma^2_{X\|W} = 0.50^2$				
50	1.9832	0.37	1.06	1.32	1.9749	2.69	3.53
100	1.9933	0.29	1.12	1.46	1.9906	2.54	2.87
150	2.0109	0.26	1.24	1.52	2.0155	2.48	3.02
300	2.0011	0.22	1.37	1.59	1.9972	2.40	2.63
500	2.0023	0.19	1.34	1.65	1.9978	2.35	2.49
1000	2.0056	0.17	1.40	1.70	2.0011	2.33	2.51
			$\beta_1 = 2,\ \sigma^2_{X\|W} = 0.75^2$				
50	1.8788	0.41	0.88	1.14	1.9148	2.58	2.94
100	1.9211	0.32	0.92	1.26	1.9677	2.42	2.87
150	1.9237	0.29	1.01	1.32	1.9717	2.37	2.61
300	1.9430	0.24	1.18	1.44	1.9852	2.29	2.48
500	1.9601	0.21	1.10	1.52	1.9942	2.27	2.45
1000	1.9629	0.19	1.29	1.55	1.9993	2.24	2.39
			$\beta_1 = 0.5,\ \sigma^2_{X\|W} = 0.75^2$				
50	0.5492	0.2164	0.1679	0.2644	0.4958	1.0288	1.2590
100	0.5401	0.1769	0.1917	0.2891	0.5076	0.9494	1.2173
150	0.5261	0.1488	0.2050	0.2953	0.5009	0.8896	1.1081
300	0.5134	0.1162	0.2249	0.3191	0.5001	0.7727	1.0197
500	0.5096	0.0989	0.2016	0.3350	0.5020	0.7183	0.9284
1000	0.5024	0.0888	0.1739	0.3357	0.5002	0.6814	0.8166

Table G.5. Summary statistics of the Monte Carlo sampling distributions of the length of a 95% Wald confidence interval of the slope in a simple linear regression with measurement error. The interval is based on the asymptotic normality of the sampling distribution of the RC estimator. The distributions $Y|X$ and $X|W$ are normal, the primary sample size is n = 1000 and the calibration sample size is m. The results are based on 2500 simulated primary and calibration samples. (**** indicates that the value is >1000.)

			Percentiles				
m	Mean	RMSE	Min.	2.5%	Median	97.5%	Max.
			$\beta_1 = 2,\ \sigma^2_{X\|W} = 0.25^2$				
50	1.06	0.292	0.51	0.64	1.01	1.76	4.77
100	0.77	0.136	0.46	0.56	0.76	1.08	1.44
150	0.66	0.089	0.39	0.51	0.65	0.86	1.09
300	0.52	0.045	0.40	0.44	0.52	0.62	0.75
500	0.45	0.028	0.37	0.40	0.45	0.51	0.56
1000	0.40	0.016	0.35	0.37	0.40	0.43	0.46
			$\beta_1 = 2,\ \sigma^2_{X\|W} = 0.50^2$				
50	3.57	56.014	0.56	0.91	1.95	6.68	****
100	1.58	0.646	0.55	0.83	1.44	3.24	7.86
150	1.30	0.392	0.53	0.77	1.23	2.23	4.04
300	0.95	0.182	0.60	0.68	0.93	1.36	1.89
500	0.80	0.106	0.53	0.62	0.78	1.03	1.29
1000	0.67	0.058	0.51	0.57	0.66	0.79	0.88
			$\beta_1 = 2,\ \sigma^2_{X\|W} = 0.75^2$				
50	51.45	****	0.69	1.04	2.85	31.58	****
100	2.82	2.507	0.63	0.97	2.18	8.69	39.48
150	2.10	1.453	0.67	0.93	1.79	4.97	38.06
300	1.45	0.459	0.71	0.86	1.36	2.59	5.16
500	1.19	0.252	0.60	0.82	1.15	1.75	2.84
1000	0.96	0.126	0.61	0.75	0.94	1.26	1.56
			$\beta_1 = 0.5,\ \sigma^2_{X\|W} = 0.75^2$				
50	479.89	****	0.20	0.29	0.76	10.43	****
100	0.83	1.650	0.18	0.29	0.61	2.42	62.76
150	0.59	0.320	0.20	0.29	0.51	1.41	5.12
300	0.45	0.126	0.24	0.29	0.42	0.74	2.24
500	0.39	0.071	0.23	0.29	0.38	0.55	0.84
1000	0.35	0.037	0.25	0.28	0.34	0.43	0.50

Table G.6. Summary statistics of the Monte Carlo sampling distributions of the length of a 95% percentile bootstrap confidence interval of the slope in a simple linear regression with measurement error. The interval is based on the percentile bootstrap method applied to the RC estimator from bootstrap sample of size 2500. The distributions of $Y|X$ and $X|W$ are normal, the primary sample size is n = 1000 and the calibration sample size is m. The results are based on 2500 simulated primary and calibration samples. (**** indicates that the value is >1000.)

			Percentiles				
m	Mean	RMSE	Min.	2.5%	Median	97.5%	Max.
				$\beta_1 = 2,\ \sigma^2_{X\|W} = 0.25^2$			
50	1.12	0.361	0.52	0.64	1.05	1.95	6.93
100	0.79	0.153	0.44	0.55	0.77	1.15	1.51
150	0.67	0.098	0.37	0.51	0.65	0.88	1.10
300	0.52	0.049	0.39	0.44	0.52	0.63	0.77
500	0.46	0.031	0.36	0.40	0.45	0.52	0.59
1000	0.40	0.019	0.34	0.36	0.40	0.44	0.48
				$\beta_1 = 2,\ \sigma^2_{X\|W} = 0.50^2$			
50	4.67	12.564	0.58	0.94	2.47	17.56	390.61
100	1.84	1.012	0.56	0.86	1.60	4.34	16.49
150	1.42	0.512	0.53	0.79	1.31	2.62	6.25
300	0.98	0.202	0.56	0.68	0.96	1.46	2.13
500	0.81	0.115	0.52	0.62	0.80	1.06	1.38
1000	0.67	0.061	0.50	0.56	0.66	0.80	0.90
				$\beta_1 = 2,\ \sigma^2_{X\|W} = 0.75^2$			
50	17.95	32.100	0.74	1.18	5.05	120.30	271.95
100	5.87	12.921	0.67	1.06	2.88	32.97	175.49
150	3.05	6.204	0.66	1.00	2.10	9.68	185.99
300	1.60	0.612	0.70	0.88	1.47	3.11	7.66
500	1.24	0.289	0.62	0.84	1.19	1.92	3.49
1000	0.97	0.134	0.60	0.76	0.96	1.29	1.67
				$\beta_1 = 0.5,\ \sigma^2_{X\|W} = 0.75^2$			
50	4.82	8.579	0.20	0.33	1.31	33.29	60.134
100	1.73	4.271	0.20	0.31	0.78	11.42	72.98
150	0.83	1.305	0.21	0.30	0.59	2.37	33.65
300	0.49	0.171	0.24	0.30	0.45	0.85	3.57
500	0.41	0.080	0.23	0.29	0.39	0.60	0.98
1000	0.35	0.040	0.25	0.28	0.35	0.43	0.52

Table G.7. Summary statistics of the Monte Carlo sampling distributions of the length of a 95% BCa bootstrap confidence interval of the slope in a simple linear regression with measurement error. The interval is based on the BCa bootstrap method applied to the RC estimator from bootstrap sample of size 2500. The distributions of $Y|X$ and $X|W$ are normal, the primary sample size is n = 1000 and the calibration sample size is m. The results are based on 2500 simulated primary and calibration samples. (**** indicates that the value is >1000.)

			Percentiles				
m	Mean	RMSE	Min.	2.5%	Median	97.5%	Max.
			$\beta_1 = 2,\ \sigma^2_{X\|W} = 0.25^2$				
50	1.11	0.362	0.52	0.64	1.05	1.94	6.93
100	0.79	0.152	0.44	0.55	0.77	1.14	1.50
150	0.67	0.098	0.37	0.50	0.65	0.88	1.10
300	0.52	0.049	0.39	0.44	0.52	0.63	0.76
500	0.46	0.031	0.36	0.40	0.45	0.52	0.59
1000	0.40	0.019	0.33	0.36	0.40	0.44	0.47
			$\beta_1 = 2,\ \sigma^2_{X\|W} = 0.50^2$				
50	47.69	****	0.59	0.94	2.41	19.59	****
100	1.83	1.014	0.60	0.86	1.59	4.28	18.53
150	1.42	0.510	0.53	0.79	1.31	2.61	6.00
300	0.98	0.203	0.58	0.68	0.95	1.46	2.09
500	0.81	0.115	0.52	0.63	0.80	1.07	1.37
1000	0.70	0.061	0.50	0.56	0.66	0.80	0.90
			$\beta_1 = 2,\ \sigma^2_{X\|W} = 0.75^2$				
50	****	****	0.72	1.18	4.84	465.39	****
100	6.92	26.032	0.67	1.05	2.83	32.69	687.81
150	3.28	12.951	0.66	1.00	2.10	9.32	513.83
300	1.60	0.611	0.70	0.88	1.47	3.09	7.50
500	1.24	0.288	0.62	0.84	1.20	1.92	3.34
1000	0.97	0.134	0.60	0.75	0.96	1.29	1.67
			$\beta_1 = 0.5,\ \sigma^2_{X\|W} = 0.75^2$				
50	66.93	****	0.20	0.32	1.27	150.92	****
100	4.07	68.058	0.20	0.31	0.78	12.04	****
150	0.86	2.160	0.21	0.30	0.59	2.38	70.60
300	0.49	0.170	0.24	0.30	0.45	0.86	3.62
500	0.41	0.081	0.23	0.29	0.39	0.60	0.98
1000	0.35	0.040	0.25	0.28	0.35	0.44	0.53

Table G.8. Summary statistics of the Monte Carlo sampling distributions of the length of a 95% likelihood-ratio confidence interval of the slope in a simple linear regression with measurement error. The interval is calculated by inverting the likelihood ratio test. The distributions of $Y|X$ and $X|W$ are normal, the primary sample size is n = 1000 and the calibration sample size is m. The results are based on 2500 simulated primary and calibration samples.

			Percentiles				
m	Mean	RMSE	Min.	2.5%	Median	97.5%	Max.
$\beta_1 = 2,\ \sigma^2_{X\|W} = 0.25^2$							
50	1.07	0.236	0.52	0.66	1.06	1.52	1.83
100	0.79	0.137	0.47	0.57	0.77	1.09	1.31
150	0.67	0.092	0.39	0.52	0.65	0.88	1.04
300	0.52	0.047	0.40	0.44	0.52	0.62	0.75
500	0.46	0.029	0.37	0.41	0.45	0.52	0.56
1000	0.40	0.017	0.34	0.36	0.40	0.43	0.46
$\beta_1 = 2,\ \sigma^2_{X\|W} = 0.50^2$							
50	1.34	0.197	0.58	0.94	1.35	1.73	1.96
100	1.07	0.143	0.56	0.77	1.09	1.32	1.47
150	0.95	0.123	0.55	0.68	0.96	1.16	1.33
300	0.79	0.096	0.42	0.56	0.80	0.94	1.02
500	0.70	0.079	0.39	0.50	0.71	0.83	0.90
1000	0.61	0.063	0.31	0.44	0.62	0.70	0.77
$\beta_1 = 2,\ \sigma^2_{X\|W} = 0.75^2$							
50	1.43	0.226	0.74	0.99	1.43	1.88	2.30
100	1.15	0.182	0.61	0.80	1.16	1.48	1.67
150	1.02	0.171	0.54	0.68	1.04	1.32	1.48
300	0.85	0.156	0.43	0.54	0.87	1.11	1.31
500	0.76	0.148	0.36	0.46	0.79	1.00	1.10
1000	0.67	0.138	0.27	0.39	0.70	0.89	0.96
$\beta_1 = 0.5,\ \sigma^2_{X\|W} = 0.75^2$							
50	0.73	0.163	0.23	0.33	0.75	1.00	1.33
100	0.61	0.134	0.19	0.32	0.64	0.82	0.93
150	0.54	0.116	0.21	0.30	0.56	0.73	0.85
300	0.45	0.089	0.24	0.30	0.45	0.62	0.74
500	0.40	0.065	0.24	0.29	0.39	0.54	0.64
1000	0.35	0.038	0.26	0.29	0.35	0.43	0.51

Table G.9. Error rate of a 95% confidence interval for the slope in a simple linear regression with measurement error. *m* denotes calibration sample size ; "L" the percent of samples for which the true slope was greater than the upper limit of a 95% CI; "R" the percent for which the true slope was smaller than the lower limit of a 95% CI; and "Tot" the total error rate. The confidence intervals are based on the asymptotic normality of the sampling distribution of the RC estimator (Wald), the percentile BCa bootstrap methods, and inverting a LRT. The results are based on 2500 simulated primary and calibration samples.

	Wald			Boots. Percentile			Bootstrap BCa			LRT		
m	L	R	Tot	L	R	Tot	L	R	Tot	L	R	Tot
$\beta_1 = 2,\ \sigma^2_{X\|W} = 0.25^2$												
50	5.36	0.36	5.72	2.76	3.00	5.76	3.08	2.88	5.96	2.72	2.72	5.44
100	3.92	0.72	4.64	2.52	2.76	5.28	2.48	2.60	5.08	2.32	2.32	4.64
150	3.96	1.36	5.32	2.48	2.72	5.20	2.60	2.68	5.28	2.52	2.76	5.28
300	2.48	1.56	4.04	1.88	2.76	4.64	1.72	2.80	4.52	1.64	2.68	4.32
500	2.80	2.04	4.84	2.48	2.72	5.20	2.44	2.60	5.04	2.52	2.80	5.32
1000	2.68	2.28	4.96	2.44	2.64	5.08	2.48	2.56	5.04	2.48	2.60	5.08
$\beta_1 = 2,\ \sigma^2_{X\|W} = 0.50^2$												
50	6.80	0.00	6.80	2.92	1.08	4.00	3.12	0.96	4.08	2.76	0.56	3.32
100	6.12	0.00	6.12	2.68	2.84	5.52	2.84	2.56	5.40	2.36	0.88	3.24
150	4.76	0.08	4.84	2.56	2.80	5.36	2.60	2.88	5.48	2.44	1.24	3.68
300	4.96	0.72	5.68	3.16	2.48	5.64	3.28	2.48	5.76	3.24	1.68	4.92
500	3.36	0.72	4.08	2.52	2.24	4.76	2.40	2.36	4.76	2.36	1.76	4.12
1000	3.08	2.00	5.08	2.44	2.92	5.36	2.40	2.72	5.12	2.20	2.80	5.00
$\beta_1 = 2,\ \sigma^2_{X\|W} = 0.75^2$												
50	8.88	0.00	8.88	2.76	0.00	2.76	3.04	1.08	4.12	2.72	0.04	2.76
100	6.24	0.00	6.24	2.84	0.64	3.48	2.84	0.88	3.72	3.00	0.12	3.12
150	6.92	0.00	6.92	3.16	1.48	4.64	3.00	1.48	4.48	3.08	0.00	3.08
300	5.48	0.00	5.48	3.16	2.24	5.40	3.00	2.24	5.24	2.96	0.04	3.00
500	4.12	0.44	4.56	2.28	2.28	4.56	2.28	2.36	4.64	2.32	0.08	2.40
1000	4.60	1.48	6.08	3.24	2.84	6.08	3.16	2.80	5.96	2.92	0.40	3.32
$\beta_1 = 0.5,\ \sigma^2_{X\|W} = 0.75^2$												
50	8.08	0.00	8.08	3.00	0.00	3.00	3.40	1.28	4.68	2.52	2.00	4.52
100	6.76	0.00	6.76	2.88	0.96	3.84	2.80	1.00	3.80	2.60	2.88	5.48
150	6.12	0.00	6.12	3.36	2.28	5.64	3.52	2.28	5.80	3.20	2.80	6.00
300	3.96	0.24	4.20	2.40	2.76	5.16	2.44	2.72	5.16	2.44	2.68	5.12
500	3.36	0.76	4.12	2.12	2.52	4.64	2.28	2.68	4.96	2.20	2.52	4.72
1000	3.24	1.40	4.64	2.76	2.64	5.40	2.80	2.64	5.44	2.72	2.52	5.24

Appendix H. Asymptotic variance of the regression calibration estimators for designs with internal calibration data (from Thurston et al., 2005)

The following results are provided by Thurston et al., 2005:

$$Var\left(\hat{\beta}_{1,EXT}\right)=\frac{\sigma_Y^2}{n\sigma_X^2}\left[\frac{\rho_{YX}^2\left(1-\rho_{XW}^2\right)\left(1+\frac{n}{m}\right)+\left(1-\rho_{YX}^2\right)}{\rho_{XW}^2}\right]$$

$$Var\left(\hat{\beta}_{1,CRS}\right)=\frac{\sigma_Y^2}{n\sigma_X^2}\left[\frac{\rho_{YX}^2\rho_{XW}^2\left(1-\rho_{XW}^2\right)\left(1+\frac{n}{m}\right)+\left(1-\rho_{YX}^2\right)\left(\rho_{XW}^2+\frac{m}{n}\right)}{\left(\rho_{XW}^2+\frac{m}{n}\right)^2}\right]$$

$$Var\left(\hat{\beta}_{1,SCK}\right)=\frac{\sigma_Y^2}{n\sigma_X^2}\left[\frac{\rho_{YX}^2\left(1-\rho_{XW}^2\right)\left(1+\frac{n}{m}\right)+\left(1-\rho_{YX}^2\right)}{\rho_{XW}^2+\frac{m}{n}+\frac{\rho_{YX}^2\left(1-\rho_{XW}^2\right)}{\left(1-\rho_{YX}^2\right)}\left(1+\frac{m}{n}\right)}\right]$$

where the subscripts in σ^2 and ρ^2 denote the variance and correlation between the subscripted variables, respectively, and *n* and *m* the sample size of the primary and calibration datasets.

Appendix I. Simulation details from the internal calibration study

Table I.1. Relative bias and RMSE the RC and ML estimators of the slope in a simple linear regression with measurement error and an internal calibration study. The true value of the parameter is 1. The distributions $Y|X$ and $X|W$ are normal, with a primary sample size of n = 1000 and an internal calibration sample size of m. The results are based on 2000 simulated primary and calibration samples. (**** indicates that the value is >100.)

.

	Internal		External		CRS		SCK		MLE	
m	Rel. Bias (%)	rmse	Rel. Bias (%)	rmse	Rel. Bias (%)	rmse	Rel. Bias (%)	rmse	Rel. Bias (%)	rmse
$Corr(Y,X)=0.75,\ Corr(X,W)=0.75$										
10	0.88	0.32	12.16	1.09	9.19	0.51	-1.28	0.23	6.23	0.24
25	-0.79	0.19	3.29	0.20	2.98	0.19	-0.86	0.13	2.12	0.13
50	-0.26	0.13	1.68	0.14	1.41	0.13	-0.43	0.09	1.13	0.09
100	-0.14	0.09	0.46	0.10	0.28	0.09	-0.44	0.07	0.39	0.06
200	0.32	0.07	0.59	0.08	0.45	0.06	0.15	0.05	0.51	0.05
500	-0.09	0.04	0.10	0.06	-0.02	0.04	-0.12	0.03	-0.01	0.03
$Corr(Y,X)=0.75,\ Corr(X,W)=0.36$										
10	-0.65	0.32	****	****	-11.5	0.98	-10.1	0.33	7.48	0.30
25	0.14	0.18	25.3	17.56	3.95	0.45	-4.40	0.18	3.09	0.14
50	-0.08	0.13	10.9	4.93	2.27	0.26	-2.69	0.13	1.53	0.09
100	0.11	0.09	10.2	0.57	0.94	0.15	-1.04	0.09	0.84	0.07
200	0.02	0.06	3.40	0.25	-0.25	0.09	-0.61	0.06	0.31	0.05
500	-0.10	0.04	1.55	0.17	-0.16	0.05	-0.27	0.04	0.06	0.03
$Corr(Y,X)=0.36,\ Corr(X,W)=0.75$										
10	-1.25	0.95	16.2	2.85	8.39	0.45	1.49	0.35	9.88	0.43
25	-2.34	0.53	2.87	0.23	2.51	0.22	0.62	0.20	2.68	0.21
50	0.62	0.38	1.80	0.18	1.60	0.17	0.89	0.16	1.83	0.16
100	0.09	0.26	1.11	0.15	0.86	0.13	0.45	0.13	0.96	0.13
200	0.03	0.19	0.23	0.13	0.11	0.11	-0.10	0.11	0.25	0.10
500	-0.02	0.12	0.21	0.12	0.06	0.08	-0.02	0.08	0.13	0.08
$Corr(Y,X)=0.37,\ Corr(X,W)=0.36$										
10	-3.58	0.97	****	****	-11.4	1.05	-18.0	0.69	7.95	0.88
25	2.52	0.54	****	****	4.22	0.54	-6.35	0.41	6.71	0.41
50	-0.23	0.37	44.9	8.47	2.74	0.33	-5.59	0.29	3.04	0.27
100	-0.26	0.25	7.07	0.42	0.22	0.22	-2.98	0.20	1.46	0.19
200	-0.20	0.18	4.81	0.34	0.12	0.16	-1.35	0.16	0.60	0.15
500	-0.37	0.11	1.84	0.27	-0.33	0.10	-0.70	0.10	0.09	0.10

Table I.2. 2.5 and 97.5 percentiles of the Monte Carlo sampling distribution of the RC and ML estimators of the slope in a simple linear regression with measurement error and an internal calibration study. The true value of the parameter is 1. The distributions $Y|X$ and $X|W$ are normal, with a primary sample size of n = 1000 and an internal calibration sample size of m. The results are based on 2000 simulated primary and calibration samples. (**** indicates that the value is >100.)
.

	Internal		External		CRS		SCK		MLE	
m	2.5	97.5	2.5	97.5	2.5	97.5	2.5	97.5	2.5	97.5
$Corr(Y,X)=0.75,\ Corr(X,W)=0.75$										
10	0.36	1.63	0.63	2.18	0.64	2.13	0.62	1.53	0.70	1.59
25	0.62	1.35	0.74	1.53	0.75	1.49	0.76	1.29	0.80	1.30
50	0.75	1.26	0.79	1.33	0.80	1.29	0.84	1.20	0.85	1.19
100	0.82	1.18	0.82	1.23	0.84	1.19	0.87	1.13	0.88	1.13
200	0.88	1.13	0.86	1.17	0.90	1.13	0.91	1.10	0.92	1.10
500	0.92	1.08	0.89	1.13	0.93	1.07	0.93	1.07	0.94	1.06
$Corr(Y,X)=0.75,\ Corr(X,W)=0.36$										
10	0.36	1.64	-6.15	9.99	-1.67	2.91	0.37	1.52	0.57	1.66
25	0.65	1.36	0.30	6.91	0.46	2.00	0.62	1.32	0.77	1.33
50	0.75	1.25	0.53	2.99	0.60	1.55	0.73	1.22	0.85	1.21
100	0.83	1.18	0.63	2.08	0.73	1.32	0.82	1.17	0.89	1.15
200	0.88	1.12	0.67	1.59	0.82	1.18	0.88	1.11	0.91	1.10
500	0.92	1.07	0.72	1.39	0.91	1.09	0.92	1.07	0.94	1.06
$Corr(Y,X)=0.36,\ Corr(X,W)=0.75$										
10	-0.75	2.96	0.60	2.13	0.60	2.12	0.56	1.84	0.60	2.26
25	-0.07	2.04	0.69	1.57	0.70	1.53	0.71	1.47	0.71	1.52
50	0.26	1.76	0.73	1.42	0.74	1.38	0.74	1.35	0.75	1.37
100	0.49	1.52	0.75	1.32	0.77	1.28	0.78	1.27	0.78	1.28
200	0.63	1.37	0.75	1.27	0.80	1.22	0.80	1.21	0.81	1.21
500	0.77	1.24	0.77	1.25	0.84	1.17	0.84	1.17	0.84	1.17
$Corr(Y,X)=0.37,\ Corr(X,W)=0.36$										
10	-0.93	2.84	-10.3	10.94	-1.65	3.10	-0.44	2.31	-1.29	2.72
25	-0.10	2.06	-3.91	7.11	0.28	2.18	0.26	1.83	0.45	1.90
50	0.28	1.73	0.43	3.37	0.51	1.77	0.47	1.59	0.56	1.59
100	0.51	1.52	0.50	2.12	0.62	1.47	0.60	1.39	0.66	1.40
200	0.65	1.35	0.48	1.78	0.69	1.33	0.68	1.31	0.71	1.30
500	0.78	1.21	0.54	1.63	0.80	1.21	0.80	1.20	0.82	1.19

Table I.3. Average length of a 95% confidence interval of the slope in a simple linear regression with measurement error and an internal calibration study. The distributions $Y|X$ and $X|W$ are normal, the primary sample size is n = 1000 and the calibration sample size is *m*. The results are based on 2000 simulated primary and calibration samples. (**** indicates that the value is >100.)

		Wald			Bootstrap			
	Int.	Ext.	CRS	SCK	Ext	CRS	SCK	LRT
				$Corr(Y,X)=0.75,\ Corr(X,W)=0.75$				
10	1.35	2.36	1.64	0.77	2.42	3.96	0.99	0.69
25	0.75	0.77	0.74	0.50	0.87	0.87	0.52	0.45
50	0.51	0.54	0.49	0.36	0.57	0.51	0.36	0.33
100	0.36	0.39	0.34	0.26	0.40	0.34	0.26	0.24
200	0.25	0.30	0.23	0.19	0.31	0.23	0.19	0.18
500	0.16	0.24	0.15	0.13	0.24	0.15	0.13	0.12
				$Corr(Y,X)=0.75,\ Corr(X,W)=0.36$				
10	1.39	****	9.12	1.00	18.00	4.20	1.37	0.87
25	0.75	****	2.91	0.64	16.36	1.97	0.72	0.53
50	0.51	40.67	1.31	0.46	9.02	1.01	0.50	0.36
100	0.35	1.61	0.68	0.33	3.07	0.57	0.34	0.26
200	0.24	0.93	0.37	0.23	1.11	0.34	0.24	0.19
500	0.15	0.66	0.18	0.15	0.68	0.18	0.15	0.13
				$Corr(Y,X)=0.36,\ Corr(X,W)=0.75$				
10	3.99	12.11	1.65	1.23	3.84	2.48	1.78	1.28
25	2.21	0.87	0.84	0.79	1.01	0.93	0.83	0.81
50	1.50	0.68	0.63	0.61	0.71	0.65	0.62	0.62
100	1.04	0.57	0.51	0.50	0.58	0.51	0.50	0.50
200	0.73	0.51	0.42	0.41	0.51	0.42	0.41	0.41
500	0.46	0.47	0.33	0.33	0.47	0.33	0.33	0.33
				$Corr(Y,X)=0.37,\ Corr(X,W)=0.36$				
10	3.97	****	9.70	2.23	4.61	18.35	2.87	2.03
25	2.15	****	3.18	1.48	9.63	9.44	1.63	1.24
50	1.44	67.02	1.53	1.07	9.54	1.32	1.12	0.95
100	1.00	1.65	0.92	0.79	2.96	0.86	0.81	0.73
200	0.70	1.27	0.63	0.60	1.47	0.62	0.60	0.55
500	0.44	1.05	0.41	0.40	1.09	0.40	0.40	0.38

Table I.4. Error rate of a 95% confidence interval for the slope in a simple linear regression with measurement error and an internal calibration study. *m* denotes calibration sample size ; "L" the percent of samples for which the true slope was greater than the upper limit of a 95% CI; "R" the percent for which the true slope was smaller than the lower limit of a 95% CI; and "Tot" the total error rate. The results are based on 2000 simulated primary and calibration samples.

				Wald intervals								
	Internal			External			CRS			SCK		
m	L	R	Tot	L	R	Tot	L	R	Tot	L	R	Tot
$Corr(Y,X)=0.75,\ Corr(X,W)=0.75$												
10	2.9	2.8	5.6	10.1	0.0	10.1	10.1	0.0	10.1	11.2	3.2	14.3
25	3.3	2.4	5.6	5.1	0.1	5.1	5.2	0.1	5.2	5.9	1.9	7.8
50	2.2	2.4	4.6	4.6	0.4	4.9	4.6	0.4	5.0	4.5	2.3	6.8
100	2.9	2.7	5.6	5.3	0.9	6.2	5.4	1.0	6.4	4.8	2.0	6.7
200	2.4	3.3	5.7	3.9	1.4	5.3	3.1	1.5	4.6	2.7	2.4	5.1
500	2.8	2.5	5.3	3.0	1.8	4.8	3.0	2.3	5.3	3.2	2.5	5.7
$Corr(Y,X)=0.75,\ Corr(X,W)=0.36$												
10	2.9	2.2	5.0	17.6	0.0	17.6	17.6	0.0	17.6	17.2	2.3	19.5
25	2.3	2.5	4.8	11.3	0.0	11.3	11.0	0.0	11.0	9.4	2.2	11.5
50	2.8	2.2	5.0	9.5	0.0	9.5	9.5	0.0	9.5	7.5	1.8	9.3
100	2.7	2.7	5.4	6.3	0.0	6.3	6.0	0.0	6.0	5.0	2.2	7.2
200	2.4	2.4	4.8	6.2	0.0	6.2	5.1	0.1	5.2	3.8	1.7	5.5
500	2.5	2.0	4.5	4.4	0.6	5.0	3.7	1.9	5.6	3.5	2.0	5.5
$Corr(Y,X)=0.36,\ Corr(X,W)=0.75$												
10	2.6	2.8	5.4	8.9	0.1	8.9	8.8	0.1	8.9	9.3	1.3	10.6
25	2.6	2.3	4.9	5.0	0.1	5.1	5.2	0.1	5.2	5.5	0.7	6.2
50	2.3	2.9	5.2	3.8	0.7	4.4	4.0	0.8	4.8	3.9	1.2	5.1
100	2.3	2.3	4.6	3.6	1.7	5.2	3.1	1.7	4.8	3.1	1.9	5.0
200	2.6	2.5	5.1	3.4	2.2	5.5	3.0	2.0	5.0	2.9	1.9	4.8
500	2.9	3.2	6.0	2.5	2.7	5.2	2.4	2.7	5.1	2.3	2.5	4.7
$Corr(Y,X)=0.37,\ Corr(X,W)=0.36$												
10	2.9	2.7	5.5	16.6	0.0	16.6	16.3	0.0	16.3	17.4	1.8	19.2
25	2.8	2.6	5.4	10.0	0.0	10.0	9.4	0.0	9.4	10.3	1.5	11.8
50	2.8	2.8	5.6	8.6	0.0	8.6	8.1	0.0	8.1	7.6	1.8	9.4
100	2.3	2.7	5.0	5.3	0.0	5.3	4.4	0.1	4.4	4.6	1.5	6.0
200	2.2	2.6	4.8	4.8	0.1	4.8	4.1	1.2	5.3	4.2	2.0	6.2
500	2.8	2.1	4.9	3.2	1.6	4.8	2.7	2.3	5.0	3.0	2.2	5.2

Table I.4 (Cont).

	Bootstrap percentile intervals											
	External			CRS			SCK			LRT		
m	L	R	Tot	L	R	Tot	L	R	Tot	L	R	Tot
$Corr(Y,X)=0.75,\ Corr(X,W)=0.75$												
10	5.2	1.9	7.1	5.1	2.0	7.0	5.4	1.6	6.9	3.2	7.2	10.4
25	2.7	3.2	5.9	2.8	3.1	5.9	4.6	2.1	6.6	2.4	4.0	6.4
50	2.9	3.3	6.2	2.9	3.1	6.0	3.2	2.3	5.4	2.6	3.7	6.2
100	3.9	2.7	6.6	3.9	2.4	6.3	4.0	2.2	6.2	2.9	3.0	5.8
200	2.9	2.9	5.8	2.2	2.9	5.1	2.6	2.9	5.5	2.3	3.0	5.3
500	2.7	2.6	5.3	2.9	2.6	5.4	3.0	2.8	5.7	2.6	2.6	5.1
$Corr(Y,X)=0.75,\ Corr(X,W)=0.36$												
10	5.5	0.0	5.5	5.4	0.0	5.4	10.6	0.1	10.7	2.2	6.2	8.4
25	3.7	0.0	3.7	3.9	0.0	3.9	6.5	0.6	7.1	1.6	5.0	6.5
50	3.7	0.0	3.7	3.5	0.1	3.5	6.0	1.3	7.3	2.0	3.6	5.6
100	2.2	0.3	2.5	2.6	0.7	3.3	4.8	1.8	6.6	2.2	3.7	5.9
200	2.9	2.6	5.5	3.2	2.3	5.5	4.3	1.4	5.6	2.4	3.0	5.4
500	2.7	3.0	5.6	3.2	2.8	5.9	4.1	1.7	5.8	2.6	2.6	5.2
$Corr(Y,X)=0.36,\ Corr(X,W)=0.75$												
10	4.8	1.4	6.2	4.7	1.4	6.1	5.4	1.2	6.6	4.4	4.1	8.4
25	2.8	3.0	5.8	2.8	3.0	5.8	3.3	1.3	4.6	2.6	3.0	5.6
50	2.1	3.2	5.3	2.1	3.0	5.1	2.6	2.2	4.8	2.3	3.0	5.2
100	2.4	3.4	5.8	2.5	3.2	5.7	2.8	3.1	5.9	2.5	3.2	5.7
200	3.0	2.9	5.9	2.9	2.4	5.3	2.7	2.2	4.9	2.5	2.5	5.0
500	2.6	3.0	5.5	2.3	2.9	5.2	2.3	2.9	5.2	2.2	2.6	4.8
$Corr(Y,X)=0.37,\ Corr(X,W)=0.36$												
10	5.7	0.0	5.7	5.4	0.0	5.4	11.4	0.0	11.4	5.4	6.5	11.8
25	3.5	0.0	3.5	3.8	0.0	3.8	7.5	0.3	7.8	2.9	4.0	6.9
50	3.2	0.1	3.2	3.3	0.3	3.6	6.5	1.0	7.5	2.5	3.6	6.1
100	2.4	1.2	3.6	2.6	1.3	3.9	4.1	1.0	5.1	1.8	3.1	4.9
200	3.0	2.5	5.5	3.4	2.2	5.5	4.2	1.6	5.8	2.9	3.3	6.1
500	2.2	3.0	5.2	2.5	2.4	4.9	3.1	2.0	5.1	2.1	2.6	4.7

Table I.5. Robustness study: relative bias and RMSE of the RC and ML estimators of the slope in a simple linear regression with measurement error. The true value of the parameter is 1. The distributions $Y|X$ is normal, and the distribution of the regression error of $X|W$ is either a t-distribution with 5 d.f. or a lognormal distribution. The primary sample size is $n = 1000$ and the internal calibration sample size is m. The results are based on 2000 simulated primary and calibration samples.

	Internal		External		CRS		SCK		MLE	
m	Rel. Bias (%)	rmse	Rel. Bias (%).	rmse	Rel. Bias (%)	rmse	Rel. Bias (%)	rmse	Rel. Bias (%)	rmse
$t-distribution, Corr(Y,X)=0.75,\ Corr(X,W)=0.75$										
25	0.25	0.19	3.94	0.22	3.60	0.20	-0.07	0.13	2.40	0.13
50	0.43	0.13	1.84	0.14	1.60	0.13	0.03	0.09	1.25	0.09
100	0.35	0.09	1.09	0.10	0.89	0.09	0.17	0.07	0.79	0.06
200	-0.16	0.06	0.19	0.08	0.03	0.06	-0.28	0.05	0.04	0.05
500	0.00	0.04	0.13	0.06	0.03	0.04	-0.06	0.03	0.10	0.03
$t-distribution, Corr(Y,X)=0.36,\ Corr(X,W)=0.35$										
25	0.81	0.55	-14.7	11.3	3.87	0.51	-7.35	0.40	8.24	0.39
50	-0.50	0.38	19.3	1.34	2.31	0.32	-5.60	0.29	2.70	0.27
100	-0.64	0.25	6.99	0.50	-0.63	0.22	-3.67	0.21	0.73	0.19
200	-0.59	0.18	3.92	0.33	-0.37	0.16	-1.79	0.15	0.47	0.14
500	0.08	0.11	1.74	0.27	-0.01	0.10	-0.35	0.10	0.14	0.10
$lognormal\ distribution, Corr(Y,X)=0.75,\ Corr(X,W)=0.75$										
25	-1.06	0.18	3.58	0.21	3.22	0.20	-1.02	0.13	1.87	0.13
50	0.61	0.13	1.67	0.14	1.45	0.13	-0.02	0.09	1.27	0.09
100	-0.40	0.09	0.78	0.10	0.52	0.09	-0.42	0.07	0.32	0.06
200	0.08	0.06	0.49	0.08	0.32	0.06	-0.01	0.05	0.34	0.05
500	-0.06	0.04	0.04	0.06	-0.03	0.04	-0.11	0.03	0.04	0.03
$lognormal\ distribution, Corr(Y,X)=0.36,\ Corr(X,W)=0.35$										
25	-2.29	0.53	6.41	16.1	2.55	0.50	-11.4	0.40	5.12	0.39
50	-1.55	0.36	27.6	4.09	2.44	0.33	-6.11	0.28	2.51	0.27
100	0.31	0.25	5.54	0.42	-0.10	0.21	-2.86	0.20	1.31	0.19
200	0.17	0.18	3.85	0.34	-0.04	0.16	-1.36	0.15	0.80	0.15
500	-0.31	0.11	0.70	0.26	-0.48	0.10	-0.79	0.10	-0.13	0.09

Table I.6. Robustness study: average length of a 95% confidence interval of the slope in a simple linear regression with measurement error. The distributions $Y|X$ is normal, and the distribution of the regression error of $X|W$ is either a t-distribution with 5 d.f. or a lognormal distribution. The primary sample size is n = 1000 and the internal calibration sample size is m. The results are based on 2000 simulated primary and calibration samples. (**** indicates that the value is >100.)

		Wald			Bootstrap			
	Int.	Ext.	CRS	SCK	Ext	CRS	SCK	LRT
$t-distribution, Corr(Y,X)=0.75,\ Corr(X,W)=0.75$								
25	0.76	0.78	0.74	0.49	1.01	0.81	0.51	0.45
50	0.52	0.53	0.49	0.35	0.58	0.51	0.36	0.33
100	0.36	0.39	0.34	0.26	0.40	0.34	0.26	0.24
200	0.25	0.30	0.23	0.19	0.31	0.23	0.19	0.18
500	0.16	0.24	0.15	0.13	0.24	0.15	0.13	0.12
$t-distribution, Corr(Y,X)=0.36,\ Corr(X,W)=0.35$								
25	2.23	****	2.96	1.47	15.09	2.28	1.61	1.27
50	1.46	4.89	1.52	1.07	9.11	1.31	1.11	0.95
100	1.01	1.73	0.91	0.79	3.19	0.86	0.81	0.74
200	0.70	1.26	0.63	0.60	1.48	0.62	0.60	0.56
500	0.44	1.05	0.41	0.40	1.08	0.41	0.40	0.38
$lognormal\ distribution, Corr(Y,X)=0.75,\ Corr(X,W)=0.75$								
25	0.75	0.77	0.74	0.49	0.96	0.82	0.51	0.45
50	0.51	0.54	0.49	0.36	0.57	0.51	0.36	0.33
100	0.36	0.39	0.34	0.26	0.40	0.34	0.26	0.24
200	0.25	0.30	0.23	0.19	0.31	0.23	0.19	0.18
500	0.16	0.24	0.15	0.13	0.24	0.15	0.13	0.12
$lognormal\ distribution, Corr(Y,X)=0.36,\ Corr(X,W)=0.35$								
25	2.16	****	2.98	1.44	15.40	2.25	1.58	1.25
50	1.46	34.24	1.51	1.07	8.93	1.30	1.11	0.96
100	1.00	1.61	0.91	0.79	2.89	0.86	0.81	0.73
200	0.70	1.28	0.63	0.60	1.50	0.62	0.60	0.56
500	0.44	1.05	0.41	0.40	1.08	0.40	0.40	0.38

Table I.7. Robustness study: error rate of a 95% confidence interval for the slope in a simple linear regression with measurement error. The distributions $Y|X$ is normal, and the distribution of the regression error of $X|W$ is either a t-distribution with 5 d.f. or a lognormal distribution. Error rate of a 95% confidence interval for the slope in a simple linear regression with measurement error. *m* denotes calibration sample size ; "L" the percent of samples for which the true slope was greater than the upper limit of a 95% CI; "R" the percent for which the true slope was smaller than the lower limit of a 95% CI; and "Tot" the total error rate. The results are based on 2000 simulated primary and calibration samples.

				Wald intervals								
	Internal			External			CRS			SCK		
m	L	R	Tot	L	R	Tot	L	R	Tot	L	R	Tot
$t-distribution, Corr(Y,X)=0.75,\ Corr(X,W)=0.75$												
25	1.7	3.0	4.6	6.2	0.2	6.4	6.2	0.2	6.4	5.8	2.6	8.4
50	2.4	2.5	4.8	4.5	0.4	4.9	4.4	0.4	4.7	4.0	2.2	6.2
100	2.5	2.4	4.9	2.9	0.7	3.6	3.2	0.8	3.9	3.0	2.5	5.5
200	2.5	2.5	5.0	3.6	1.3	4.8	3.7	1.0	4.7	3.3	1.7	4.9
500	2.8	2.4	5.2	2.9	1.7	4.6	2.8	1.5	4.3	2.6	2.0	4.6
$t-distribution, Corr(Y,X)=0.36,\ Corr(X,W)=0.35$												
25	2.3	2.2	4.4	9.9	0.0	9.9	9.5	0.0	9.5	9.6	1.3	10.9
50	2.3	2.2	4.5	8.2	0.0	8.2	7.9	0.0	7.9	8.3	1.6	9.9
100	2.8	2.7	5.5	6.9	0.0	6.9	5.7	0.4	6.0	5.8	1.6	7.4
200	2.6	2.4	5.0	4.6	0.2	4.7	3.3	1.1	4.4	3.8	1.5	5.2
500	2.5	2.8	5.3	3.1	1.1	4.2	3.0	2.5	5.5	3.4	2.3	5.7
$lognormal\ distribution, Corr(Y,X)=0.75,\ Corr(X,W)=0.75$												
25	3.1	1.8	4.8	6.2	0.2	6.3	6.1	0.2	6.3	6.4	1.6	8.0
50	2.1	2.6	4.7	4.8	0.4	5.2	4.6	0.4	5.0	4.1	2.1	6.1
100	3.1	2.4	5.4	4.7	0.9	5.6	4.4	0.9	5.3	4.3	1.7	6.0
200	2.3	3.2	5.5	4.0	1.5	5.4	3.7	1.4	5.1	3.6	2.7	6.3
500	2.3	2.3	4.6	3.0	1.7	4.7	3.2	1.8	5.0	3.1	2.3	5.4
$lognormal\ distribution, Corr(Y,X)=0.36,\ Corr(X,W)=0.35$												
25	3.0	1.9	4.8	13.0	0.0	13.0	12.6	0.0	12.6	12.8	1.1	13.9
50	2.8	1.9	4.6	8.5	0.0	8.5	7.8	0.0	7.8	8.5	1.5	10.0
100	2.1	2.5	4.6	4.7	0.0	4.7	4.3	0.2	4.5	5.0	1.1	6.1
200	2.4	2.6	5.0	4.2	0.1	4.3	3.1	1.5	4.6	3.6	2.1	5.6
500	2.3	2.2	4.5	2.9	1.1	4.1	2.2	1.7	3.9	2.3	1.7	4.0

Table I.7 (Cont).

	Bootstrap percentile intervals											
	External			CRS			SCK			LRT		
M	L	R	Tot	L	R	Tot	L	R	Tot	L	R	Tot
$t-distribution, Corr(Y,X)=0.75,\ Corr(X,W)=0.75$												
25	3.5	4.0	7.4	3.4	3.9	7.3	4.2	2.6	6.8	3.5	5.1	8.6
50	3.0	3.5	6.5	2.9	3.5	6.4	2.8	2.5	5.3	3.0	4.1	7.0
100	1.9	2.6	4.4	1.9	2.7	4.5	2.7	3.0	5.6	2.5	3.7	6.2
200	2.5	2.4	4.9	2.5	2.2	4.7	2.7	2.0	4.7	3.0	2.9	5.8
500	2.5	2.5	5.0	2.5	1.8	4.3	2.6	2.1	4.7	2.7	2.6	5.3
$t-distribution, Corr(Y,X)=0.36,\ Corr(X,W)=0.35$												
25	2.5	0.0	2.5	2.6	0.0	2.6	5.9	0.1	6.0	2.2	3.3	5.5
50	3.4	0.2	3.5	2.8	0.5	3.3	6.3	0.9	7.2	2.6	3.6	6.2
100	2.7	1.0	3.6	3.3	1.4	4.7	5.1	0.9	6.0	2.3	2.7	5.0
200	3.1	2.2	5.3	2.8	2.0	4.8	3.7	1.4	5.1	2.1	2.9	5.0
500	2.4	2.2	4.5	2.9	2.6	5.4	3.5	2.1	5.6	2.9	2.8	5.7
$lognormal\ distribution, Corr(Y,X)=0.75,\ Corr(X,W)=0.75$												
25	3.4	3.6	7.0	3.5	3.3	6.8	5.3	1.9	7.1	3.0	4.0	7.0
50	3.3	3.0	6.3	3.2	3.0	6.1	3.3	2.4	5.7	2.4	3.4	5.8
100	2.9	2.7	5.6	3.1	2.8	5.9	3.8	1.9	5.7	3.4	3.4	6.8
200	2.6	2.4	5.0	2.8	2.7	5.5	2.9	2.7	5.6	2.5	3.3	5.7
500	2.5	2.7	5.2	3.0	2.1	5.1	2.9	2.4	5.3	2.9	2.4	5.3
$lognormal\ distribution, Corr(Y,X)=0.36,\ Corr(X,W)=0.35$												
25	4.7	0.0	4.7	5.3	0.0	5.3	9.0	0.3	9.3	3.5	3.5	7.0
50	3.1	0.2	3.2	3.9	0.3	4.2	6.6	0.7	7.3	2.6	3.5	6.1
100	2.3	0.9	3.2	2.8	1.1	3.9	4.4	0.6	4.9	2.3	2.4	4.6
200	2.4	2.9	5.3	2.4	2.4	4.7	3.9	1.5	5.4	1.9	3.2	5.1
500	2.4	2.1	4.5	2.4	2.1	4.5	2.7	1.8	4.5	2.3	2.3	4.6