MEASUREMENT ERROR IN EPIDEMIOLOGIC STUDIES

Raymond J. Carroll

Department of Statistics

Texas A&M University

College Station TX 77843–3143

email: carroll@stat.tamu.edu

Phone: 409-845-3141

Fax: 409-845-3144

Sabbatical Address: 8206 Fenway Road, Bethesda MD 20817

Sabbatical Phone: 301-365-7783

Sabbatical Fax: 301-365-7785

Acknowledgment

INTRODUCTION

This article is concerned with relating a response or outcome to an exposure and confounders in the presence of measurement error in one or more of the variables. We focus almost entirely on measurement error in a continuous or measured variable. When categorical variables (exposed or not exposed, case or control, quintiles of fat) are measured with error, they are said to be misclassified [*see* Misclassification Error]. There are also many links in this topic with methods for handling missing data and with validation studies [*see* Missing data in Epidemiologic Studies; Validation Study]. For further details and a general overview of the topic, see Carroll, et al. [19]; Fuller [30] should be consulted for the linear model.

Before describing the problem, it is useful to first consider a number of specific examples that have had an impact on the development of the field.

- Measurement error has long been a concern in relating error–prone predictors such a systolic blood pressure (SBP) to the development of coronary heart disease (CHD). That SBP is measured with error is well–known, and estimates (Carroll, et al. [20]) suggest that approximately 1/3 of its observed variability is due to measurement error. The Framingham Heart Study is perhaps the best known cohort study in which the role of measurement error in SBP has been a concern for many years. MacMahon, et al. [44] describe the important public health implications of properly accounting for the measurement error inherent in SBP. In an as yet unpublished paper, David Yanez, Richard Kronmal and Lynn Shemanski have discovered an example also in the CHD context where the failure to account properly for measurement error leads to misleading conclusions based on falsely detected statistical significance.

- In measuring nutrient intake, measurement error has been a long–term concern, as has the impact of this error on the ability to detect nutritional factors leading to cancer, especially breast and colon cancer. Typical cohort studies measure diet by means of food frequency questionnaires which, while related to long–term diet, are known to have biases and measurement errors. Other instruments are in use in this field, including food records (essentially diaries), 24–hour recalls and (for a limited number of variables such as total caloric intake) biomarkers. Measurement error in nutrient instruments can be very large, for example because of the daily and seasonal variability of an individual's diet, and the biases in and loss of power to detect nutrient–cancer relationships can be profound. There is still considerable controversy in this field, see Hunter, et al. [37], Prentice [54] and Li, et al. [41]. Because of the cost of cohort studies in nutrition, case–control studies are of considerable interest. However, nutrient intakes in case–control studies are measured after the development of disease in cases,

and this might cause differential measurement error, a topic we discuss in some detail below.

- There are a number of on–going prospective and case–control studies of disease and serum hormone levels, and this is an area on considerable potential. Measurement error is a major concern here, due to within–individual variation of hormones, as well as various laboratory errors.

- In measuring environmental risk factors, measurement error is a common problem. For example, measuring household lead levels is an error–prone process, not only because of laboratory and device error, but also because lead levels are inhomogeneous in both space and time, while measurement methods tend to be in fixed locations at fixed times. Because lead exposure has many possible media (air, dust, soil) with possibly correlated errors, the effects of measurement error can be large and complex.

Outline

This article consists of a series of major sections, as follows.

1. We will first outline the basic concepts of measurement error modeling, making particular distinction between differential and non–differential measurement error. We will also describe the ideas of functional and structural modeling, as well as indicate how the measurement error problem can be treated as a missing data problem.

2. Following the introductory concepts, we will discuss the problem of measurement error as it pertains to the linear regression model. Here we will introduce the idea of attenuation of regression coefficients, and the biases in parameter estimates caused by measurement error. We will also discuss hypothesis testing, In the simplest cases, measurement error causes an often large decrease in the power to detect significant effects, while as indicated above, and as exhibited through the analysis of covariance, in an observational study, measurement error in a confounder can cause misleading inferences about exposure effects.

3. Having described the effects of measurement error on estimation and hypothesis testing, we turn to correcting for the effects due to measurement error. We first describe the two most common methods, known as regression calibration and SIMEX, and also a group of techniques called corrected score methods. We also describe the use of instrumental variables.

2

4. Maximum likelihood and Bayesian estimation form an important component of the measurement error problem, and are described in some detail. We define the likelihood function, and show the crucial difference in the likelihood function between the non–differential and differential measurement error cases, see equations (13) and (14).

5. While most of the article is based on measurement error in predictors, there is an important literature on response error, which we also review.

6. Case–control studies are important in epidemiology. A distinguishing feature of case-control studies is that the measurement error may be differential. In the differential measurement error case, we indicate that a specific type of data is required, the validation data sets, in which the true predictor can be observed for a subset of the study participants. If the measurement error is non–differential, matters are much easier, and the famous result of Prentice & Pyke [55] on the analysis of case–control studies is shown to have an analogue in the measurement error context.

7. There is a significant and developing literature on measurement error in survival analysis, and we indicate two possible approaches to the problem.

Measurement error models have a common structure. We illustrate the terms using the breast cancer and nutrition example

- An underlying model for a response in terms of predictors, e.g., linear, logistic, non-linear regression. This is the model we would fit if all variables were observed without error. In what follows, we will call $\mathbf{Y}$ the response. For example, in the breast cancer and nutrition example, $\mathbf{Y}$ is breast cancer incidence fit to covariables using logistic regression.

- A variable which is measured subject to error. This could be an exposure or a confounder. We will call this variable $\mathbf{X}$. It is often called the *error–prone* predictor or the *latent predictor*. In the breast cancer example, $\mathbf{X}$ is long–term nutrient intake.

- The observed value of the mismeasured variable. We will call this $\mathbf{W}$, e.g., nutrient intake measured from a food frequency questionnaire.

- Those predictors which for all practical purposes are measured without error, which we will call $\mathbf{Z}$, e.g., age, body mass index.

- We are interested in relating the response $\mathbf{Y}$ to the true predictors $(\mathbf{Z}, \mathbf{X})$. One method, often called the *naive* method, simply replaces the error-prone predictor $\mathbf{X}$ with its

3

measured version **W**. This substitution typically leads to biases in parameter estimates and can lead to misleading inferences.

- The goal of measurement error modeling is to obtain nearly unbiased estimates of exposure effects and valid inferences. Attainment of this goal requires careful analysis. Substituting **W** for **X**, but making no adjustments in the usual fitting methods for this substitution, leads to estimates that are biased, sometimes seriously. In assessing measurement error, careful attention needs to be given to the type and nature of the error, and the sources of data which allow modeling of this error.

It should be obvious that one should design studies and instruments in such a way as to best lessen or eliminate measurement error. In this article, we demonstrate some of the impacts of ignoring measurement error, ranging from bias in parameter estimates (Figure 1), to loss of power requiring therefore much larger sample sizes to detect effects (Figure 2) to cases where the Type I errors occur at higher rates than the usual 5% (Figure 3).

COMPUTER PROGRAMS

Splus and SAS computer programs (on SPARC architecture SunOS versions 4 and 5 and for Windows on PC's) which implement many of the methods described in this article (for major generalized linear models such as linear, logistic, Probit, Poisson and gamma regression) are available at no cost on the world wide web at http://stat.tamu.edu/qvf/qvf.html. Bootstrap standard errors are available. They have been developed by Raymond Carroll, Henrik Schmiedieche and H. Joseph Newton.

A set of programs for logistic regression (in SAS and FORTRAN) is available from Professor Donna Spiegelman (e-mail stdls@gauss.bwh.harvard.edu). Interested readers should contact her for information concerning extension of these programs to proportional hazards and linear regression.

Iowa State University (Department of Statistics, Iowa State University, Ames IA 50011) distributes programs called EV–CARP for linear measurement error models at a cost of $300.

MODELS FOR MEASUREMENT ERROR

A fundamental prerequisite for analyzing a measurement error problem is specification of a model for the measurement error process. The *classical error model*, in its simplest form, is appropriate when an attempt is made to determine **X** directly, but one is unable to do so because of various errors in measurement. For example, consider systolic blood pressure (SBP), which is known to have strong daily and seasonal variations. In trying to measure SBP, the various sources of error include simple machine recording error, administration

4

error, time of day and season of the year. In such a circumstance, it sometimes makes sense to hypothesize an unbiased additive error model which we write as

$$\text{(the classical model) } \mathbf{W} = \mathbf{X} + \mathbf{U}, \tag{1}$$

where $\mathbf{U}$, the error, is assumed independent of $\mathbf{X}$. An alternative model, the *controlled variable or Berkson model* [6] is especially applicable to laboratory studies. As an example, consider the herbicide study of Rudemo, et al. [61]. In that study, a nominal measured amount $\mathbf{W}$ of herbicide was applied to a plant. However, the actual amount $\mathbf{X}$ absorbed by the plant differed from $\mathbf{W}$, e.g., because of potential errors in application. In this case,

$$\text{(the Berkson model) } \mathbf{X} = \mathbf{W} + \mathbf{U}, \tag{2}$$

where $\mathbf{U}$, the error, is assumed independent of $\mathbf{W}$.

Determining an appropriate error model to use in the data analysis depends upon the circumstances and the available data. For example, in the herbicide study, the measured concentration $\mathbf{W}$ is fixed by design and the true concentration $\mathbf{X}$ varies due to error, so that model (2) is appropriate. On the other hand, in the measurement of long-term systolic blood pressure, it is the true long-term blood pressure which is fixed for an individual, and the measured value which is perturbed by error, so model (1) should be used. Estimation and inference procedures have been developed both for error and controlled-variable models.

This hardly exhausts the possibly error models. See Carroll, et al. [19] and Freedman, et al. [29] for more details and further examples with more complex structure.

Sources of Data

In order to perform a measurement error analysis, one needs information about the error structure. These data sources can be broken up into two main categories:
- *Internal* subsets of the primary data;
- *External* or independent studies.

Within each of these broad categories, there are three types of data, all of which might be available only in a random subsample of the data set in question:
- *Validation* data in which $\mathbf{X}$ is observable directly.
- *Replication* data, in which replicates of $\mathbf{W}$ are available.
- *Instrumental* data, in which another variable $\mathbf{T}$ is observable in addition to $\mathbf{W}$.

An internal validation data set is the ideal, because it can be used with all known analytical techniques, permits direct examination of the error structure and tests of critical error model assumptions, typically leads to much greater precision of estimation and inference, and has strong links to the well–developed theory of missing data analysis (see below). We

cannot express too forcefully that if at all possible, one should obtain an internal validation data set.

With external validation data, one must assume that the error structure in those data also applies to the primary data (see below).

Replication data are used when it is impossible to measure **X** exactly, as, for example, when **X** represents long-term systolic average blood pressure or long-term average nutrient intake. Usually, one would make replicate measurements if there were good reason to believe that the replicated mean is a better estimate of **X** than a single observation, i.e., the classical error model is the target. In the classical error model (1), replication data can be used to estimate the variance of the measurement error, **U**.

Internal instrumental data sets containing a second measure **T** are useful for instrumental variable analysis, discussed briefly later in this article.

Transportability of Models and Parameters

In some studies, the measurement error process is not assessed directly, but instead is estimated from external data sets. We say that parameters of a model can be transported from one study to another if the model holds with the same parameter values in both studies. Typically, in applications only a subset of the model parameters need be transportable.

In many instances, approximately the same classical error model holds across different populations. For example, consider systolic blood pressure at two different clinical centers. Assuming similar levels of training for technicians making the measurements and a similar measurement protocol, it is reasonable to expect that the distribution of the error in the recorded measure is independent of the clinical center one enters, the technician making the measurement, and the value of **X** being measured. Thus, in classical error models it is often reasonable to assume that the error distribution is the same across different populations, i.e., transportable.

A common mistake is to transport a correction for measurement error from one study to the next. Such transportation is almost never appropriate. For instance, while the properties of errors of measurement may be reasonably transportable, the distribution of the true (or latent) predictor **X** is rarely transportable, since it depends so heavily on the population being sampled. Problems arise because corrections for measurement error involve not only the measurement error process but also the distribution of **X**. For example, systolic blood pressure measurements in the MRFIT study and the Framingham Heart Study may well have the same measurement error variance, but the distribution of true blood pressure **X** appears to differ substantially in the two studies, and the "correction for attenuation" described below cannot be transported from Framingham to MRFIT. See Carroll & Stefanski [22] for further details.

Is There an "Exact" Predictor?

We have based our discussion on the existence of an exact predictor $\mathbf{X}$ and measurement error models that provide information about this predictor. However, in practice, it is often the case that the definition of "exact" needs to be carefully considered prior to discussion of error models. In the measurement error literature the term "gold standard" is often used for the operationally defined exact predictor, though sometimes this term is used for an exact predictor that cannot be operationally defined. Using an operational definition for an "exact" predictor is often reasonable and justifiable on the grounds that it is the best one could ever possibly hope to accomplish. However, such definitions may be controversial. For example, consider the problem of relating breast cancer risk to the dietary intake of fat. One way to determine whether decreasing one's fat intake lowers the risk of developing breast cancer is to conduct a clinical trial in which members of the treatment group are encouraged to reduce fat intakes. If instead one uses observational prospective data, along with an operational definition of long-term intake, one should be aware that the results of a measurement error analysis could be invalid if true long-term intake and operational long-term intake differ in subtle ways.

Differential and Non–Differential Error

It is important to make a distinction between *differential* and *non–differential* measurement error. Non–Differential measurement error occurs in a broad sense when one would not even bother with $\mathbf{W}$ if $\mathbf{X}$ were available, i.e., $\mathbf{W}$ has no information about the response other than what is available in $\mathbf{X}$. Non–Differential measurement error typically holds in <u>COHORT STUDIES</u>[*], but is often a suspect assumption in <u>CASE-CONTROL STUDIES</u>[*].

Technically, measurement error is non–differential if the distribution of $\mathbf{Y}$ given $(\mathbf{X}, \mathbf{Z}, \mathbf{W})$ depends only on $(\mathbf{X}, \mathbf{Z})$. In this case $\mathbf{W}$ is said to be a *surrogate*. Measurement error is *differential* otherwise.

For instance, consider the Framingham example. The predictor of major interest is long-term systolic blood pressure ($\mathbf{X}$), but we can only observe blood pressure on a single day ($\mathbf{W}$). It seems plausible that a single day's blood pressure contributes essentially no information over and above that given by true long-term blood pressure, and hence that measurement error is non–differential. The same remarks apply to the nutrition examples: measuring diet on a single day should not contribute information not already available in long-term diet.

Many problems can plausibly be analyzed assuming non–differential measurement error, especially when the covariate measurements occur at a fixed point in time, and the response is measured at a later time, as is typical in cohort studies.

There are two exceptions that need to be kept in mind. First, in case-control studies, the disease response is obtained first, and then one measures antecedent exposures and other covariates. In nutrition studies, this ordering of measurement may well cause differential measurement error. For instance, here the true predictor would be long-term dietary intake before diagnosis, but the dietary interview data are obtained only after diagnosis. A woman who develops breast cancer may exaggerate her estimated fat intake, thus introducing recall bias * [*see* Bias in Case–Control Studies]. In such circumstances, estimated fat intake will be associated with disease status even after conditioning on true long-term diet before diagnosis.

When measurement error is non–differential, one can estimate parameters in models for responses given true covariates even when the true covariates are not observable. This is not true when measurement error is differential, except for the linear model. With differential error, one must obtain a validation subsample in which both true covariate measurements and surrogate measurements are available. Most of this article focuses on non–differential measurement error models. Differential models with a validation study are typically best analyzed by techniques for handling missing data [*see* Validation Study; Missing Data in Epidemiologic Studies; Misclassification Error].

Prediction

Prediction of a response is different from estimation and inference for parameters. If a predictor $\mathbf{X}$ is measured with error, and one wants to predict a response *based on the error-prone version* $\mathbf{W}$ *of* $\mathbf{X}$, then except for an important case discussed below, it makes little sense to worry about measurement error. The reason for this is quite simple. If one has an original set of data $(\mathbf{Y}, \mathbf{W})$, one can fit a convenient model to $\mathbf{Y}$ as a function of $\mathbf{W}$. Predicting $\mathbf{Y}$ from $\mathbf{W}$ is merely a matter of using this model for prediction. There is no need then for measurement error to play a role in the problem.

The one situation requiring that we model the measurement error occurs when we develop a prediction model using data from one population but we wish to predict in another population. A naive prediction model that ignores measurement error may not be transportable. This context often becomes quite complex, requiring a combination of missing data and measurement error techniques, and to the best of our knowledge has not been investigated in detail in the literature. One exception is Ganse, et al. [31].

Is Bias Always Towards the Null?

It is commonly thought that the effect of measurement error is to bias estimates of exposure effects "towards the null". Hence, one could ignore measurement error when testing the null hypothesis of no exposure effect, and one could assume that non–null estimates, if anything,

underestimate the effect of exposure. This lovely and appealing folklore is sometimes true but, unfortunately, often wrong. We discuss this point in detail below. A numerical example has recently been provided to us by David Yanez, Richard Kronmal and Lynn Shemanski in a heart disease context with seven covariates and a baseline variable. They found that while an analysis ignoring measurement error showed highly statistically significant effects in all variables, none of the effects were even close to being statistically significant when the analysis takes measurement error into account.

Functional and Structural Models

The words *functional* and *structural* have important places in the area of measurement error models. They act as a short-hand terminology for the basic approach one uses to solve the problem. In *functional modeling* nothing is assumed about the $\mathbf{X}$'s; they could be fixed constants (the usual definition) or random variables. In *structural modeling*, $\mathbf{X}$ is assumed to be random, and a parametric distribution (usually the normal) is assumed. There has traditionally been considerable concern in the measurement error literature about the robustness of estimation and inferences based upon structural models for unobservable variates. Fuller [30], page 263, discusses this issue briefly in the classical nonlinear regression problem, and basically concludes that the results of structural modeling "may depend heavily on the (assumed) form of the $\mathbf{X}$ distribution". In probit regression, Carroll, et al. [20] report that if one assumes that $\mathbf{X}$ is normally distributed, and it really follows a chi-squared distribution with one degree of freedom, then the effect on the likelihood estimate is "markedly negative", see also Schafer [63]. Essentially all research workers in the measurement error field come to a common conclusion: likelihood methods can be of considerable value, but the possible non-robustness of inference due to model misspecification is a vexing and difficult problem.

The issue of model robustness is hardly limited to measurement error modeling. Indeed, it pervades statistics, and has led to the rise of a variety of semiparametric and nonparametric techniques. From this general point of view, *functional modeling* may be thought of as a group of semiparametric techniques. Functional modeling uses parametric models for the response, but makes no assumptions about the distribution of the unobserved covariate.

There is no agreement in the statistical literature as to whether functional or structural modeling is more appropriate. Many researchers believe that one should make as few model assumptions as possible and favor functional modeling. The argument is that any extra efficiency gained by structural modeling is more than offset by the need to perform careful and often time-consuming sensitivity analyses. Other researchers believe that appropriate statistical analysis requires one to do one's best to model every feature of the data, and thus favor structural modeling.

We take a somewhat more relaxed view of these issues. There are many problems, e.g., linear and logistic regression with additive measurement error, where functional techniques are easily computed and fairly efficient, and we have a strong bias in such circumstances towards functional modeling. In other problems, for example the segmented regression problem (Küchenhoff & Carroll [38]), structural modeling clearly has an important role to play, and should not be neglected.

Measurement Error as a Missing Data Problem

From one perspective, measurement error models are special kinds of missing data problems, because the **X**'s, being mostly and often entirely unobservable, are obviously missing as well. Readers who are already familiar with linear measurement error models and functional modeling will be struck by the fact that most of the recent missing data literature has pursued likelihood and Bayesian methods, i.e., structural modeling approaches. Readers familiar with missing data analysis will also be interested to know that, in large part, the measurement error model literature has pursued functional modeling approaches. We feel that both functional and structural modeling approaches are useful in the measurement error context, and this article will pursue both strategies.

The usual interpretation of the classical missing data problem (Little & Rubin [42]) is that the values of some of the variables of interest may not be observable for all study participants. For example, a variable may be observed for 80% of the study participants, but unobserved for the other 20%. The techniques for analyzing missing data are continually evolving, but it is fair to say that most of the recent advances (multiple imputation, data augmentation, etc.) have been based on likelihood (and Bayesian) methods.

The classical measurement error problem discussed to this point is one in which one set of variables, which we call **X**, is *never* observable, i.e., always missing. As such, the classical measurement error model is an extreme form of a missing data problem, but with *supplemental information* about **X** in the form of a surrogate, which we call **W**. Part of the art in measurement error modeling concerns how the supplemental information is related to the unobservable covariate.

Because there is a formal connection between the two fields, and because missing data analysis has become increasingly parametric, it is important to consider likelihood and Bayesian analysis of measurement error models, topics taken up later in this article.

LINEAR REGRESSION AND THE EFFECTS OF MEASUREMENT ERROR

A comprehensive account of linear measurement error models can be found in Fuller [30].

Many textbooks contain a brief description of measurement error in linear regression, usually focusing on simple linear regression and arriving at the conclusion that the effect of measurement error is to bias the slope estimate in the direction of 0. Bias of this nature is commonly referred to as *attenuation* or *attenuation to the null.*

In fact though, even this simple conclusion has to be qualified, because it depends on the relationship between the measurement, $\mathbf{W}$, and the true predictor, $\mathbf{X}$, and possibly other variables in the regression model as well. In particular, the effect of measurement error depends upon the model under consideration and on the joint distribution of the measurement error and the other variables. In multiple linear regression, the effects of measurement error vary depending on: (i) the regression model, be it additive or multiple regression; (ii) whether or not the predictor measured with error is univariate or multivariate; and (iii) the presence of bias in the measurement. The effects can range from the simple attenuation described above to situations where: (i) real effects are hidden; (ii) observed data exhibit relationships that are not present in the error-free data; and (iii) even the signs of estimated coefficients are reversed relative to the case with no measurement error.

The key point is that the measurement error distribution determines the effects of measurement error, and thus appropriate methods for correcting for the effects of measurement error depend on the measurement error distribution.

Simple Linear Regression with Additive Error: Regression to the Mean

We start with the simple linear regression model $\mathbf{Y} = \beta_0 + \beta_x \mathbf{X} + \epsilon$, where the scalar $\mathbf{X}$ has mean μ_x and variance σ_x^2, and the error in the equation ϵ is independent of $\mathbf{X}$, has mean zero and variance σ_ϵ^2. The error model is additive as in (1). In this classical additive measurement error model, it is well-known that an ordinary least squares regression ignoring measurement error produces an estimate not of β_x, but instead of $\beta_{x*} = \lambda \beta_x$, where

$$\lambda = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} < 1. \tag{3}$$

Thus ordinary least squares regression of $\mathbf{Y}$ on $\mathbf{W}$ produces an estimator that is attenuated to 0. The attenuating factor, λ, is called the *reliability ratio.*

One would expect that because $\mathbf{W}$ is an error-prone predictor, it has a weaker relationship with the response than does $\mathbf{X}$. This can be seen both by the attenuation, and also by the fact that the residual variance of this regression is increased, being not σ_ϵ^2 but instead

$$\text{var}(\mathbf{Y}|\mathbf{W}) = \text{ residual variance of observed data } = \sigma_\epsilon^2 + \lambda \beta_x^2 \sigma_u^2.$$
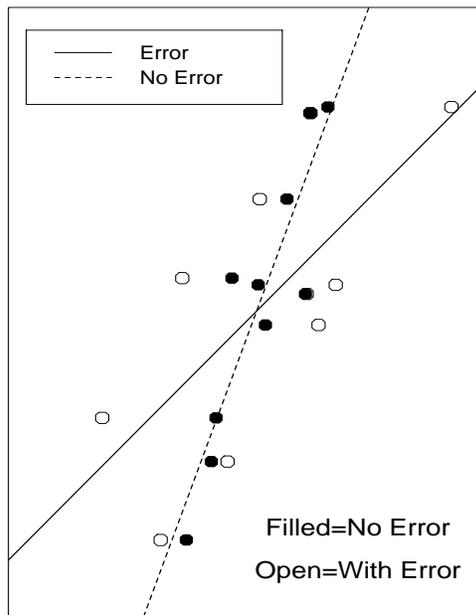
Figure 1: *Illustration of additive measurement error model. The filled circles are the true* $(\mathbf{Y}, \mathbf{X})$ *data and the dashed (steeper) line is the least squares fit to these data. The open circles and solid (attenuated) line are the observed* $(\mathbf{Y}, \mathbf{W})$ *data and the associated least squares regression line. For these data* $\sigma_x^2 = \sigma_u^2 = 1$, $(\beta_0, \beta_x) = (0, 1)$ *and* $\sigma_\epsilon^2 = .25$.

This facet of the problem is often ignored, but it is important. <u>Measurement error</u> <u>causes</u> <u>a double-whammy</u> : not only is the slope attenuated, but the data are more noisy, with an increased error about the line.

To illustrate the attenuation associated with the classical additive measurement error, the results of a small simulation are displayed in Figure 1.

Ten observations were generated with $\sigma_x^2 = \sigma_u^2 = 1$, $(\beta_0, \beta_x) = (0, 1)$ and $\sigma_\epsilon^2 = .25$. The filled circles and steeper line depict the true but unobservable data $(\mathbf{Y}, \mathbf{X})$ and the regression line of $\mathbf{Y}$ on $\mathbf{X}$. The empty circles and attenuated line depict the observed $(\mathbf{Y}, \mathbf{W})$ data and the linear regression of $\mathbf{Y}$ on $\mathbf{W}$.

Figure 1 is indicative of a phenomenon called *Regression to the Mean*. Intuitively, what this means is that the extremes in the observed $(\mathbf{W})$ data are *too* extreme, and that the true $\mathbf{X}$ is closer to the mean of the data. In fact, in normally distributed data, if $\mathbf{X}$ has a population mean μ_x, then having observed the fallible instrument, the best prediction of $\mathbf{X}$
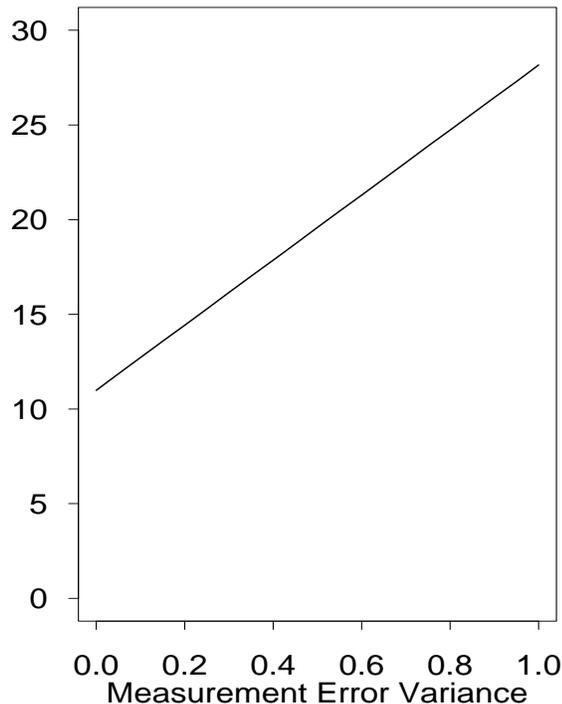
Figure 2: *Sample size for 80% power in a one–sided test of level 5% in linear regression, as a function of the measurement error variance. Here the true slope = 0.75, the true variance of* **X** *is 1.0, the true variance about the line is* 1.0.

is $\mu_x(1 - \lambda) + \lambda\mathbf{W}$, where $\lambda < 1$ is defined in (3). The net effect is that the best (linear) predictor of $\mathbf{X}$ is always closer to the overall mean than any observed $\mathbf{W}$.

The foregoing is one facet of regression to the mean. A more common definition is complementary. In a study participant with an an unusually large observed $\mathbf{W}$, if one repeats the measurement and obtained a second (replicated) measure, this replicate is generally less (and often much less) than the original extreme value.

For instance, in a study of true long-term fat intake ($\mathbf{X}$) using a 24-hour recall instrument ($\mathbf{W}$), if one focuses on the person with the highest reported fat intake, then (i) that person's true fat intake is most likely less than the the observed intake; and (ii) if one repeats the 24-hour recall instrument, the new reported fat intake is likely to be less than the original reported fat intake.

The second part of the "double–whammy" is a loss of power. The following example is meant to illustrate this loss of power, and it is easiest to do this illustration in the special case that all variances are known. Suppose that one wants to test the null hypothesis $H_0 : \beta_x = 0$

of zero slope against the one–sided alternative $H_1 : \beta_x > 0$, using a test with 5% level (Type I error) which has power 80% to detect that the slope $\beta_x = 0.75$. With known variances, in the absence of measurement error, the required sample size is

$$n = \frac{(z_{.95} + z_{.80})^2 \sigma_\epsilon^2}{\sigma_x^2 \beta_x^2},$$

where z_α is the usual α percentile of the normal distribution. With measurement error, the same formula applies, except that with $\beta_x = 0.75$, one replaces σ_ϵ^2 by $\sigma_\epsilon^2 + \lambda \beta_x^2 \sigma_u^2$, σ_x^2 by $\sigma_x^2 + \sigma_u^2$ and β_x by $\lambda \beta_x$. In Figure 2, we plot the sample sizes as a function of the measurement error variance in the case that $\mathbf{X}$ has variance $\sigma_x^2 = 1$ and the error about the line has variance $\sigma_\epsilon^2 = 1$. In the absence of measurement error, approximately 10 observations are required to obtain the desired power. However, if the measurement error variance $\sigma_u^2 = 1$ and thus the reliability $= 1/2$, approximately 30 observations are required. Thus, measurement error causes a loss of power. In planning a study with large measurement error in a covariate, one will typically require a much larger sample size to meet power goals than if there were no measurement error.

It is a common belief that the effect of measurement error is always to attenuate the slope of the regression line, but in fact attenuation depends critically on the assumed classical additive measurement error model. Very different results are obtained if measurement errors are differential. One example where this problem may arise is in dietary calibration studies. In a typical dietary calibration study, one is interested in the relationship between a self-administered food frequency questionnaire (FFQ, the value of $\mathbf{Y}$) and usual (or long-term) dietary intake (the value of $\mathbf{X}$) as measures of, for example, the percentage of calories from fat in a person's diet. FFQ's are thought to be biased for usual intake, and in a calibration study researchers will obtain a second measure (the value of $\mathbf{W}$), typically from a food diary, a 24-hour recall or a short-term biomarker. In this context, it is often assumed that the diary, recall or biomarker is unbiased for usual intake. If, as sometimes occurs, the FFQ and the diary/recall are given very nearly contemporaneously, it is unreasonable to assume that the error in the relationship between the FFQ and usual intake is uncorrelated with the error in the relationship between a diary–recall–biomarker and usual intake. This correlation has been demonstrated (Freedman, et al. [29]), and gives rise to differential error. It can be shown [19] that if there is significant correlation between the measurement error and the error about the true line, then the regression of $\mathbf{Y}$ on $\mathbf{W}$ can have a slope biased away from the null. Thus, correction for bias induced by measurement error clearly depends on the nature, as well as the extent of the measurement error.

Multiple Regression: Single Covariate Measured with Error

14

In multiple linear regression the effects of measurement error are more complicated, even for the classical additive error model.

We now consider the case where $\mathbf{X}$ is scalar, but there are additional covariates $\mathbf{Z}$ measured without error. In the linear model the mean is $\beta_0 + \beta_x\mathbf{X} + \beta_z\mathbf{Z}$. Under the usual conditions of independence of errors, the least squares regression estimator of the coefficient of $\mathbf{W}$ consistently estimates $\lambda_1\beta_x$, where

$$\lambda_1 = \frac{\sigma^2_{x|z}}{\sigma^2_{w|z}} = \frac{\sigma^2_{x|z}}{\sigma^2_{x|z} + \sigma^2_u}, \tag{4}$$

and $\sigma^2_{w|z}$ and $\sigma^2_{x|z}$ are the (residual) variances of the regressions of $\mathbf{W}$ on $\mathbf{Z}$ and $\mathbf{X}$ on $\mathbf{Z}$, respectively. Note that λ_1 is equal to the simple linear regression attenuation $\lambda = \sigma^2_x/(\sigma^2_x+\sigma^2_u)$ only when $\mathbf{X}$ and $\mathbf{Z}$ are uncorrelated. *The basic point is that the attenuation depends on the relationships among the covariates.*

The problem of measurement error-induced bias is not restricted to the regression coefficient of $\mathbf{X}$. The coefficient of $\mathbf{Z}$ is also biased in general, unless $\mathbf{Z}$ is independent of $\mathbf{X}$ (Carroll et al. [15]). In fact, naive ordinary least squares estimates not β_z but rather

$$\beta_{z*} = \beta_z + \beta_x(1 - \lambda_1)\gamma_z, \tag{5}$$

where γ_z is the coefficient of $\mathbf{Z}$ in the regression of $\mathbf{X}$ on $\mathbf{Z}$.

This result has important consequences in epidemiology when interest centers on the effects of covariates measured without error. For example, consider the case that $\mathbf{Z}$ is a binary exposure variable (exposed or not) which is classified correctly, and $\mathbf{X}$ is an important confounder measured with significant error. Then Carroll, et al. [15] show that ignoring measurement error produces a consistent estimate of the exposure effect only if the design is balanced, i.e., $\mathbf{X}$ has the same mean in both groups and is independent of treatment. With considerable imbalance, the naive analysis may lead to the conclusion that: (i) there is a treatment effect when none actually exists; and (ii) the effects are negative when they are actually positive, and vice-versa. In most observational studies the confounder and the exposure are correlated, see Greenland [34] and Greenland & Robins [35]. Errors in measuring the confounders can produce very misleading results.

Multiple Covariates Measured with Error

If multiple covariates are measured with error, then the direction of the bias induced by this error does not follow any simple pattern. One may have attenuation, reverse-attenuation, changes of sign, or an observed positive effect even at a true null model. This is especially the case when the predictors measured with error are correlated or their errors are correlated. In such problem, there really seems to be no substitute to a careful measurement error analysis.

Correcting for Bias

As we have just seen, the ordinary least squares estimator is typically biased under measurement error, and the direction and magnitude of the bias depends on the regression model and the measurement error distribution. We next describe two commonly used methods for eliminating bias.

In simple linear regression with the classical additive error model, we have seen in (3) that ordinary least squares is an estimate of $\lambda \beta_x$; recall that λ is called the reliability ratio. If the reliability ratio were known, then one could obtain a proper estimate of β_x simply by dividing the ordinary least squares slope by the reliability ratio.

Of course, the reliability ratio is rarely known in practice, and one has to estimate it. If $\widehat{\sigma}_u^2$ is an estimate of the measurement error variance (this is discussed below), and if $\widehat{\sigma}_w^2$ is the sample variance of the $\mathbf{W}$'s, then a consistent estimate of the reliability ratio is $\widehat{\lambda} = (\widehat{\sigma}_w^2 - \widehat{\sigma}_u^2)/\widehat{\sigma}_w^2$. The resulting estimate is $\beta_{x*}/\widehat{\lambda}$. In small samples the sampling distribution of this estimate is highly skewed, and in such cases a modified version of the method-of-moments estimator is recommended (Fuller [30]).

The algorithm described above is called the *method-of-moments* estimator. The terminology is apt, because ordinary least squares and the reliability ratio depend only on moments of the observed data.

The method-of-moments estimator can be constructed for the general linear model, as well as for simple linear regression. Consult the book by Fuller [30], especially Chapter 2.

Another well publicized method for linear regression in the presence of measurement error is *orthogonal regression*. It is fairly rare in epidemiological situations that the model underlying orthogonal regression holds (Carroll & Ruppert [18]), and we will not discuss the method any further.

Bias Versus Variance

Estimates that do not account for measurement error are typically biased. Unfortunately, correcting for this bias often has a price. In particular, the resulting corrected estimator will be more variable than the biased estimator, and wider confidence intervals result. For example, Rosner, et al. [60] describe a problem in logistic regression, where the response is the development of breast cancer, and the predictor measured with error is daily saturated fat intake. Ignoring measurement error, they obtained an estimated odds ratio for saturated fat of 0.92, with a 95% confidence interval from 0.80 to 1.05. The corrected estimated odds ratio was 0.83 with a confidence interval from 0.61 to 1.12, which is twice as wide as the previous interval.

Attenuation in General Problems

We have already seen that with multiple covariates, even in linear regression the effects of measurement error are complex, and not easily described. In this section, we provide a brief overview of what happens in nonlinear models.

Consider a scalar covariate $\mathbf{X}$ measured with error, and suppose that there are no other covariates. In the classical error model for simple linear regression we have seen that the bias caused by measurement error is always in the form of attenuation, so that ordinary least squares preserves the sign of the regression coefficient asymptotically, but is biased towards zero. Attenuation is a consequence then of (i) the simple linear regression model; and (ii) the classical additive error model. Without (i)–(ii), the effects of measurement error are more complex; we have already seen that attenuation may not hold if (ii) is violated.

In logistic regression when $\mathbf{X}$ is measured with additive error, attenuation does not always occur, but it is typical and generally much like that of linear regression.

Dosemeci, et al. [28] give an example of <u>MISCLASSIFICATION ERROR</u>* that shows that trends are not always preserved under non–differential measurement error. Suppose that 1,348 subjects are exposed at no ($\mathbf{X} = 0$), low ($\mathbf{X} = 1$) and high ($\mathbf{X} = 2$) levels to a harmful substance. Suppose the chance of an adverse outcome is 1/2, 2/3 and 6/7 for no, low and high exposures, while the chances of the exposures themselves are .0059347, .8902077 and .1038576, respectively. If true exposure could be ascertained, the expected outcomes would be as in the section of Table 1 labeled **TRUE**. If we were to regress $\mathbf{Y}$ on the dummy variables $\mathbf{X}_1$ indicating low exposure ($\mathbf{X}_1 = 1$), and $\mathbf{X}_2$ indicating high exposure ($\mathbf{X}_2 = 1$), then the true logistic regression parameters for $\mathbf{X}_1$ and $\mathbf{X}_2$ would be $\log(2) = .69$ and $\log(6) = 1.79$, respectively, indicating that the two higher exposure levels have response rates higher than the response rate associated with the no-exposure level. The true data clearly indicate a harmful effect due to exposure.

Now suppose, however, that measurement error (in this case misclassification) occurs, so that 40% of those truly at high exposure are misclassified into the no exposure group, and 40% of those truly at low exposure are misclassified into the high exposure group. Let $\mathbf{W}$ be the resulting variable taking on the three observed levels of exposure, with corresponding dummy variables $\mathbf{W}_1$ and $\mathbf{W}_2$. This is a theoretical example, of course, and one can criticize it for not being particularly realistic, but it is an example of non–differential measurement error. The observed data we expect to see using the surrogates $\mathbf{W}_1$ and $\mathbf{W}_2$ are also given in Table 1.

The observed logistic regression parameters for $\mathbf{W}_1$ and $\mathbf{W}_2$ are $\log(.46) = -.78$ and $\log(.53) = -.63$, respectively, indicating that the two higher exposure levels have response

Disease Status	Exposure = None	Exposure = Low	Exposure = High
True			
$\mathbf{Y} = 1$	4	800	120
$\mathbf{Y} = 0$	4	400	20
Observed			
$\mathbf{Y} = 1$	52	480	392
$\mathbf{Y} = 0$	12	240	172

Table 1: *A hypothetical logistic regression example with non–differential measurement error. The entries are the expected counts. The true logistic parameters for dummy variables low and high exposure are log(2) and log(6), respectively, while the observed coefficients for the error prone data are log(.46) and log(.53), respectively.*

rates lower than the response rate associated with the no-exposure level. The observed data suggest a beneficial effect due to exposure, even though the exposure is harmful!

HYPOTHESIS TESTING

In this section, we discuss hypothesis tests concerning regression parameters. To keep the exposition simple, we will focus on linear regression. However, the results hold in some generality, especially for logistic and Poisson regression. We assume non–differential measurement error and the classical additive error model.

The simplest approach to hypothesis testing calculates the required test statistic from the parameter estimates obtained from a measurement error analysis and their estimated standard errors. Such tests are justified whenever the estimators themselves are justified. However, this approach to testing is only possible when the indicated methods of estimation are possible, and thus requires either knowledge of the measurement error variance, or the presence of validation data, or replicate measurements, or instrumental variables.

There are certain situations in which naive hypothesis tests are justified and thus can be performed without additional data or information of any kind. Here "naive" means that we ignore measurement error and substitute $\mathbf{W}$ for $\mathbf{X}$ in a test that is valid when $\mathbf{X}$ is observed. This section studies naive tests, describing when they are and are not acceptable.

We use the criterion of asymptotic validity to distinguish between acceptable and non–acceptable tests. We say a test is asymptotically valid if its Type I error rate approaches its

nominal level as the sample size increases. Asymptotic validity, which we shorten to validity, of a test is a minimal requirement for acceptability.

The main results on the validity of naive tests under non–differential measurement error are as follows.

- The naive test of no effects due to $\mathbf{X}$ is valid. This means that if one wants to test whether *all* components of $\mathbf{X}$ together have no effect, then it is valid to ignore non–differential measurement error. Thus, for example, if $\mathbf{X}$ is the exposure, then a valid test of the null hypothesis for $\mathbf{X}$ is obtained by ignoring measurement error and performing the standard test for the problem at hand.

- The naive test described above is also fully efficient if $\mathbf{X}$ is linearly related to $\mathbf{W}$ and $\mathbf{Z}$, but not otherwise (Tosteson & Tsiatis [79]). Thus, while in principle one can obtain additional power by a measurement error analysis, many times in practice the naive test of the null hypothesis for $\mathbf{X}$ is reasonably efficient.

- In many problems, more than one covariate is measured with error. For example, suppose that the exposure and one of the confounders are measured with error. Generally, the naive test of the null hypothesis for the exposure is invalid, except under special circumstances, e.g., the exposure and confounder are statistically independent, as are their measurement errors.

- In general, naive tests for $\mathbf{Z}$ are invalid, except possibly if $\mathbf{Z}$ is uncorrelated with $\mathbf{X}$. Thus, if $\mathbf{X}$ is the exposure and $\mathbf{Z}$ is a confounder, then naive tests for significance of the exposure are valid but they are not valid for testing the significance of the confounder. Somewhat more troubling though, is the case that $\mathbf{X}$ is a confounder related to the exposure $\mathbf{Z}$; here the naive test for the exposure is generally invalid, *even if exposure is measured without error*. We already have mentioned this example previously in the case that the exposure is binary, see Carroll, Gallo & Gleser[15].

The last point can be demonstrated in the analysis of covariance, in which $\mathbf{Z}$ is a binary exposure variable, and $\mathbf{X}$ is a confounder with strong predictive ability. In the analysis of covariance, the model is

$$Y = \beta_0 + \beta_z \mathbf{Z} + \beta_x X + \epsilon,$$

where ϵ is the error–about–the–line, with variance σ_ϵ^2. The binary indicator $\mathbf{Z}$ takes on the values ± 1, with 50% of the data being unexposed ($\mathbf{Z} = -1$) and 50% of the data being exposed ($\mathbf{Z} = 1$). Within the unexposed group, $\mathbf{X}$ has mean $-\theta/2$ and variance σ_x^2, while

within the exposed group, $\mathbf{X}$ has mean $\theta/2$ and variance σ_x^2. The difference between the means for $\mathbf{X}$ in the two groups is θ. In a randomized clinical trial, one would expect that $\theta = 0$, since randomization insures that the population means of $\mathbf{X}$ are the same in the exposed and unexposed groups. In non–randomized studies, one would expect that $\theta \neq 0$. Thus, the larger the value of θ, the more unbalanced is the study. In Figure 3, we plot the level (Type I error) of the test for exposure effect which ignores measurement error, as a function of the difference in group means θ. This calculation is done for the case that $n = 20$ (10 exposed and 10 unexposed), $\sigma_\epsilon^2 = 1$, $\sigma_x^2 = 1$ and $\beta_x = 1$, for reliability ratios $\lambda = 1/2$ and $= 2/3$. The graph shows that if the means of the confounders are sufficiently different, then instead of a Type I error of 5%, the test for exposure effect which ignores measurement error in the *confounder* can have Type I error rates higher than 10%, even for such small sample sizes.

REGRESSION CALIBRATION AND SIMEX

We now describe two simple, generally applicable approaches to non–differential measurement error analysis, regression calibration and simulation extrapolation (SIMEX).

Regression Calibration

The basis of regression calibration is the replacement of $\mathbf{X}$ by the regression of $\mathbf{X}$ on $(\mathbf{Z}, \mathbf{W})$. After this approximation, one performs a standard analysis. This *regression calibration* algorithm was suggested as a general approach by Carroll & Stefanski [21] and Gleser [33]. Prentice [52] pioneered the idea for the proportional hazard model, and a modification of it has been suggested for this topic by Clayton [25]; see below. Armstrong [4] suggests regression calibration for generalized linear models, and Fuller [30], pages 261–262, briefly mentions the idea. Rosner, et al. [59], [60] have developed the idea for logistic regression into a workable and popular methodology, complete with a good computer program. Because of the importance of their contribution to epidemiological applications, regression calibration is often referred to as "Rosner's Method". Other interesting and important applications and methodology related to regression calibration include work by Whittemore [83], Pierce, et al. [51], Liu & Liang [43] and Kuha [39]. In some special cases, regression calibration is equivalent to the classical method of moments bias correction.

The main justifications of the regression calibration approximation are that for some models, e.g., loglinear mean models and linear regression, the regression calibration approximation is often exact except for a change in the intercept parameter. For logistic regression, in many cases the approximation is almost exact.

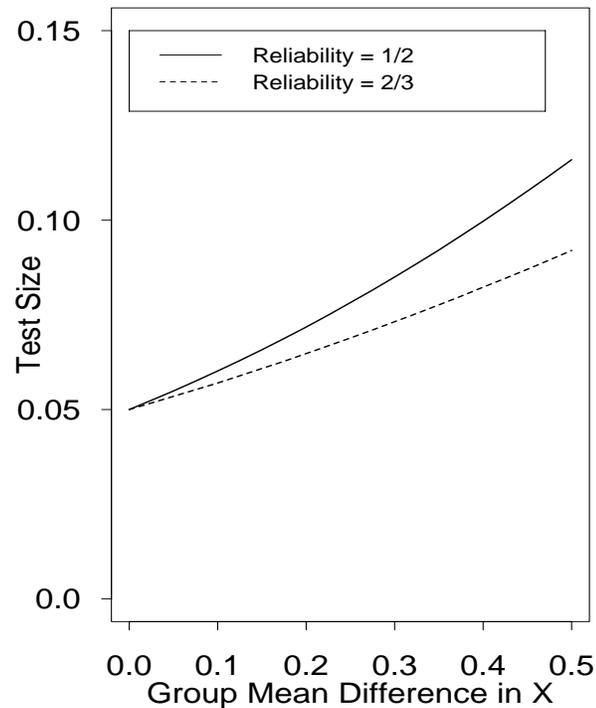### The Regression Calibration Algorithm

Figure 3: *The actual level of a test for exposure effect with a highly predictive covariate measured with error, based on a sample of size $n = 20$. Here the true slope for the covariate* $\mathbf{X} = 1.0$, *the true variance of* $\mathbf{X}$ *is* 1.0, *the true variance about the line is* 1.0, *and the reliability is either* $2/3$ *(dashed line) or* $1/2$ *(solid line). The term "Group Mean Difference in X" is the difference in the mean of* $\mathbf{X}$ *in the exposure group minus the mean of* $\mathbf{X}$ *in the control group.*

The regression calibration algorithm is what Pierce, et al. [51] call a "replacement method".

- Using replication, validation or instrumental data, estimate the regression of $\mathbf{X}$ on $(\mathbf{Z}, \mathbf{W})$ (see below). This is called the *calibration function*.

- Replace the unobserved $\mathbf{X}$ by its estimate from the regression model, and then run a standard analysis to obtain parameter estimates.

- Adjust the resulting standard errors to account for the estimation at the first step, using either the bootstrap or asymptotic methods (Carroll, et al. [19]).

The simplest form of regression calibration is the "correction for attenuation" used in linear regression. It is easiest to describe in the following situation:

(i) $\mathbf{X}$ is a scalar;

(ii) The measurement error is additive with estimated error variance $\widehat{\sigma}_u^2$.

For estimating the effect of $\mathbf{X}$, the regression calibration estimator is formed by three steps: (a) Form the naive estimator by ignoring measurement error; (b) Let $\widehat{\sigma}_{w|z}^2$ be the regression mean squared error from a linear regression of $\mathbf{W}$ on $\mathbf{Z}$ (this is the sample variance of the $\mathbf{W}$'s if there are no other covariates $\mathbf{Z}$); (c) The regression calibration estimator is defined by multiplying the naive estimator by $\widehat{\sigma}_{w|z}^2/(\widehat{\sigma}_{w|z}^2 - \widehat{\sigma}_u^2)$.

Estimating the Calibration Function Parameters

With *internal validation data*, the simplest approach is to regress $\mathbf{X}$ on the other co-variates $(\mathbf{Z}, \mathbf{W})$ in the validation data. While linear regression will be typical, it is not required.

In some problems, an *unbiased second instrument* $\mathbf{T}$ is available for a subset of the study participants. For instance, one might be interested in $\mathbf{X} =$ caloric intake over a year, but have available only $\mathbf{T} =$ the result of a biomarker experiment using a technique known as doubly-labeled water over a two-week period, which does not equal $\mathbf{X}$ because it does not take into account the variability of diet over a year. In this case one uses the regression of $\mathbf{T}$ on $(\mathbf{Z}, \mathbf{W})$ as the calibration function. This is the method used by Rosner, et al. [59] in their analysis of the Nurses' Health Study.

Finally, in the classical additive error model, one often has merely a second measurement (a replicate) for a subset of the study population. One could treat this replicate as an unbiased second instrument and apply the method described in the previous paragraph. If the $\mathbf{W}$'s are not too far from normally distributed, a more efficient method is to use the so–called best linear approximation to the calibration function, see Carroll, et al. [19], pages

47–48. This takes into account that some of the study participants do have a replicated $\mathbf{W}$, and hence uses the data in a reasonably efficient fashion.

Suppose there are k_i replicate measurements of $\mathbf{X}_i$, and $\overline{\mathbf{W}}_{i\cdot}$ is their mean. Replication enables us to estimate the measurement error covariance matrix σ_u^2 by the usual components of variance analysis, as follows:

$$\widehat{\sigma}_u^2 = \frac{\sum_{i=1}^n \sum_{j=1}^{k_i} \left(\mathbf{W}_{ij} - \overline{\mathbf{W}}_{i\cdot}\right)^2}{\sum_{i=1}^n (k_i - 1)}. \tag{6}$$

The calibration function is defined as follows. Suppose the observations are $(\mathbf{Z}_i, \overline{\mathbf{W}}_{i\cdot})$, where $\overline{\mathbf{W}}_{i\cdot}$ is the mean of k_i replicates. We use analysis of variance formulae. Let

$$\widehat{\mu}_x = \widehat{\mu}_w = \sum_{i=1}^n k_i \overline{\mathbf{W}}_{i\cdot} / \sum_{i=1}^n k_i; \quad \widehat{\mu}_z = \overline{\mathbf{Z}}_{\cdot};$$

$$\nu = \sum_{i=1}^n k_i - \sum_{i=1}^n k_i^2 / \sum_{i=1}^n k_i;$$

$$\widehat{\sigma}_z^2 = (n-1)^{-1} \sum_{i=1}^n \left(\mathbf{Z}_i - \overline{\mathbf{Z}}_{\cdot}\right)^2;$$

$$\widehat{\sigma}_{xz} = \sum_{i=1}^n k_i \left(\overline{\mathbf{W}}_{i\cdot} - \widehat{\mu}_w\right) \left(\mathbf{Z}_i - \overline{\mathbf{Z}}_{\cdot}\right) / \nu;$$

$$\widehat{\sigma}_x^2 = \left[\left\{\sum_{i=1}^n k_i \left(\overline{\mathbf{W}}_{i\cdot} - \widehat{\mu}_w\right)^2\right\} - (n-1)\widehat{\sigma}_u^2\right] / \nu.$$

The resulting estimated calibration function which is used to replace $\mathbf{X}$ in the standard analysis is

$$\widehat{\mu}_w + (\widehat{\sigma}_x^2, \widehat{\sigma}_{xz}) \begin{bmatrix} \widehat{\sigma}_x^2 + \widehat{\sigma}_u^2/k_i & \widehat{\sigma}_{xz} \\ \widehat{\sigma}_{xz} & \widehat{\sigma}_z^2 \end{bmatrix}^{-1} \begin{pmatrix} \overline{\mathbf{W}}_{i\cdot} - \widehat{\mu}_w \\ \mathbf{Z}_i - \overline{\mathbf{Z}}_{\cdot} \end{pmatrix}. \tag{7}$$

Expanded Regression Calibration Models

Rudemo, et al. [61], Carroll & Stefanski [21] and Carroll, et al. [19] all describe refinements to the regression calibration algorithm. Rudemo, et al. [61] describe a bioassay problem with a heteroscedastic Berkson error model. Racine-Poon, et al. [56] describe a similar problem.

There is a long history of approximately consistent estimates in nonlinear problems, of which regression calibration and the SIMEX method are the most recent such methods. Readers should also consult Stefanski & Carroll [70], Stefanski [67], Amemiya & Fuller [3], and Whittemore & Keller [85] for other approaches.

THE SIMEX METHOD

We now describe a method that shares the simplicity of regression calibration and is well suited to problems with additive or multiplicative measurement error. Simulation extrapolation (SIMEX) is a simulation-based method of estimating and reducing bias due to measurement error. SIMEX estimates are obtained by adding additional measurement error to the data in a resampling-like stage, establishing a trend of measurement error-induced bias versus the variance of the added measurement error, and extrapolating this trend back to the case of no measurement error. The technique was proposed by Cook & Stefanski [26], and further developed by Carroll, et al. [16] and Stefanski & Cook [73].

An integral component of SIMEX is a self-contained simulation study resulting in graphical displays that illustrate the effect of measurement error on parameter estimates and the need for bias correction. The graphical displays are especially useful when it is necessary to motivate or explain a measurement error model analysis.

This section describes the basic idea of SIMEX, focusing on linear regression with additive measurement error. For this simple model the effect of measurement error on the least squares estimator is easily determined mathematically,, as we have shown. *The key idea underlying SIMEX is the fact that the effect of measurement error on an estimator can also be determined experimentally via simulation.* If we regard measurement error as a factor whose influence on an estimator is to be determined, we are naturally led to consider simulation experiments in which the level of the measurement error, i.e., its variance, is intentionally varied.

The SIMEX Algorithm

Suppose that in addition to the original data used to calculate the naive estimate $\widehat{\beta}_{x,\text{naive}}$, there are $M - 1$ additional data sets available, each with successively larger measurement error variances, say $(1 + \zeta_m)\sigma_u^2$, where $0 = \zeta_1 < \zeta_2 < \cdots, < \zeta_M$. Of course, the least squares estimate of slope from the mth data set ignoring measurement error, $\widehat{\beta}_{x,m}$, consistently estimates $\beta_x \sigma_x^2 / \{\sigma_x^2 + (1 + \zeta_m)\sigma_u^2\}$.

We can think of this problem as a nonlinear regression model, with dependent variable $\widehat{\beta}_{x,m}$ and independent variable ζ_m, having a mean function of the form

$$\mathcal{G}(\zeta) = \frac{\beta_x \sigma_x^2}{\sigma_x^2 + (1 + \zeta)\sigma_u^2}, \qquad \zeta \geq 0.$$

The parameter of interest, β_x, is obtained from $\mathcal{G}(\zeta)$ by extrapolation to $\zeta = -1$. We describe the process schematically in Figure 4.

SIMEX imitates the procedure just described. In the *simulation step* additional independent measurement errors with variance $\zeta_m \sigma_u^2$ are generated and added to the original data, thereby creating data sets with successively larger measurement error variances. For the mth data set, the total measurement error variance is $\sigma_u^2 + \zeta_m \sigma_u^2 = (1 + \zeta_m)\sigma_u^2$. Next, estimates are obtained from each of the resulting contaminated data sets. The simulation
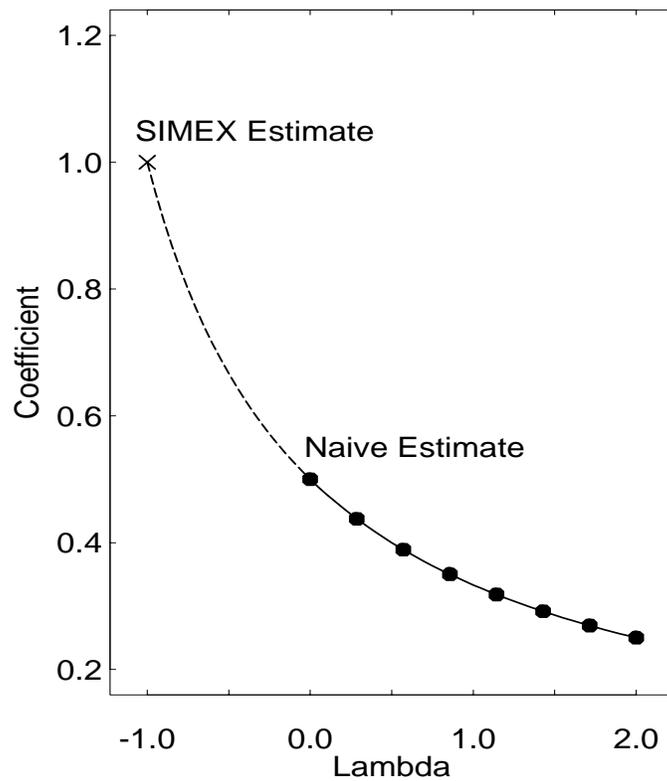
Figure 4: *A generic plot of the effect of measurement error of size* $(1 + \zeta)\sigma_u^2$ *on parameter estimates. The value of* ζ *is on the x-axis, while the value of the estimated coefficient is on the y-axis. The SIMEX estimate is an extrapolation to* $\zeta = -1$. *The naive estimate occurs at* $\zeta = 0$.

and reestimation step is repeated a large number of times (to remove simulation variability) and the average value of the estimate for each level of contamination is calculated. These averages are plotted against the ζ values and regression techniques are used to fit an extrapolant function to the averaged, error-contaminated estimates. Extrapolation back to the ideal case of no measurement error ($\zeta = -1$) yields the SIMEX estimate.

The first part of the algorithm is the simulation step. As described above, this involves using simulation to create additional datasets with increasingly large measurement error $(1 + \zeta)\sigma_u^2$. For any $\zeta \geq 0$, define

$$\mathbf{W}_{b,i}(\zeta) = \mathbf{W}_i + \sqrt{\zeta}\, \mathbf{U}_{b,i}, \quad i = 1, \ldots, n, \quad b = 1, \ldots, B, \tag{8}$$

where the computer-generated *pseudo errors*, $\{\mathbf{U}_{b,i}\}_{i=1}^n$, are mutually independent, independent of all the observed data and identically distributed, normal random variables with mean 0 and variance σ_u^2.

Having generated the new predictors, we compute the resulting naive estimates, component by component. For each ζ, do this B times ($B = 100$ usually works fine) and compute their average, $\widehat{\beta}(\zeta)$. It is the points $\{\widehat{\beta}(\zeta_m), \zeta_m\}_1^M$ that are plotted as filled circles in Figure 4. This is the simulation component of SIMEX.

The extrapolation step of the proposal entails modeling each of the components of $\widehat{\beta}(\zeta)$ as functions of ζ for $\zeta \geq 0$, and extrapolating the fitted models back to $\zeta = -1$. In Figure 4 the extrapolation is indicated by the dashed line and the SIMEX estimate is plotted as a cross. Carroll, et al. [19] describe practical modifications of the algorithm, and how to estimate variances of parameters. Inference for SIMEX estimators can also be performed via the bootstrap. Because of the computational burden of the SIMEX estimator, the bootstrap requires considerably more computing time than do other methods. Without efficient implementation of the estimation scheme at each step, the SIMEX bootstrap may take an inconveniently long time to compute. On our computing system for measurement error models the implementation is efficient, and most bootstrap applications take little time.

We have described the SIMEX algorithm in terms of the additive measurement error model. However, SIMEX applies more generally.

For example, consider multiplicative error. Taking logarithms transforms the multiplicative model to the additive model. SIMEX works naturally here, in that one performs the simulation step (8) on the logarithms of the $\mathbf{W}$'s, and not on the $\mathbf{W}$'s themselves.

With replicates, one can also investigate the appropriateness of different transformations. For example, after transformation the standard deviation of the intra-individual replicates should be uncorrelated with their mean, and one can find the transformation (logarithm, square root, etc.) which makes the two uncorrelated.

Example

To illustrate SIMEX, we use data from the Framingham Heart Study, correcting for bias due to measurement error in systolic blood pressure measurements. The Framingham study consists of a series of exams taken two years apart. We use Exam #3 as the baseline. There are 1,615 men aged 31–65 in this data set, with the outcome, $\mathbf{Y}$, indicating the occurrence of coronary heart disease (CHD) within an eight-year period following Exam #3; there were 128 such cases of CHD. Predictors employed in this example are the patient's age at Exam #2, smoking status at Exam #1 and serum cholesterol at Exams #2 and #3, in addition to systolic blood pressure (SBP) at Exam #3, the latter being the average of two measurements taken by different examiners during the same visit. In addition to the measurement error in SBP measurements, there also is measurement error in the cholesterol measurements. However, for this example we ignore the latter source of measurement error and illustrate the methods under the assumption that only SBP is measured with error.

The covariates measured without error, $\mathbf{Z}$, are age, smoking status and serum cholesterol, with $\mathbf{W} = \log(\text{SBP} - 50)$. Implicitly, we are defining $\mathbf{X}$ as the long-term average of $\mathbf{W}$. We illustrate the analyses for the case where $\mathbf{W}$ is the mean of the two transformed SBP's, and σ_u^2 is estimated using (6). The estimated linear model correction for attenuation, or inverse of the reliability ratio, is 1.16; if only one SBP measurement were used, the correction would be 1.33.

Figure 5 contains plots of the logistic regression coefficients $\widehat{\Theta}(\zeta)$ for eight equally spaced values of ζ spanning $[0, 2]$ (solid circles). For this example $B = 2000$. The points plotted at $\zeta = 0$ are the naive estimates $\widehat{\Theta}_{\text{naive}}$. The nonlinear least-squares fits of $\mathcal{G}_{\text{RL}}(\lambda,,)$ to the components of $\{\widehat{\Theta}(\zeta_m), \zeta_m\}_1^8$ (solid curves) are extrapolated to $\zeta = -1$ (dashed curves) resulting in the SIMEX estimators (crosses). The open circles are the SIMEX estimators that result from fitting quadratic extrapolants. To preserve clarity the quadratic extrapolants were not plotted. Note that the quadratic-extrapolant estimates are conservative relative to the rational linear-extrapolant estimates in the sense that they fall between the rational linear-extrapolant estimates and the naive estimates.

We have stated previously that the SIMEX plot displays the effect of measurement error on parameter estimates. This is especially noticeable in Figure 5. In each of the four graphs in Figure 5, the range of the ordinate corresponds to a one-standard error confidence interval for the naive estimate constructed using the information standard errors. Thus Figure 5 illustrates the effect of measurement error relative to the variability in the naive estimate. It is apparent that the effect of measurement error is of practical importance only on the coefficient of $\log(\text{SBP} - 50)$.
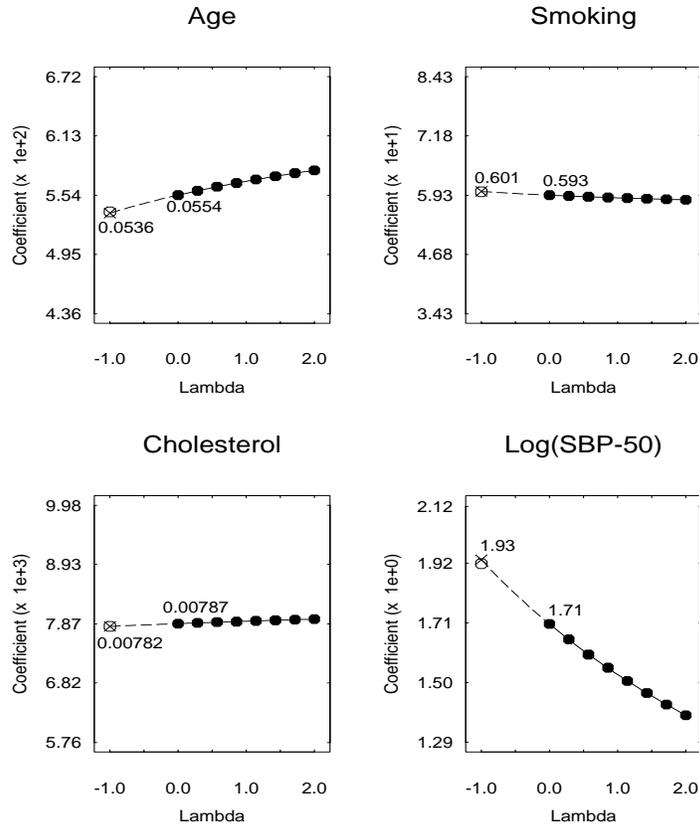
Figure 5: *Coefficient extrapolation functions for the Framingham logistic regression modeling. The simulated estimates* $\{\widehat{\Theta}(\zeta_m),\ \zeta_m\}_1^8$ *are plotted (solid circles) and the fitted rational linear extrapolant (solid line) is extrapolated to* $\zeta = -1$ *(dashed line) resulting in the SIMEX estimate (cross). Open circles indicate SIMEX estimates obtained with the quadratic extrapolant.*

CONDITIONAL AND CORRECTED SCORES FOR FUNCTIONAL MODELING

Regression calibration and SIMEX are easily applied general methods for non–differential error.. Although the resulting estimators are consistent in important special cases such as linear regression and loglinear mean models, they are only approximately consistent in general.

For certain generalized linear models and measurement error distributions there are easily applied functional methods that are fully (and not just approximately) consistent, and make no assumptions about the distribution of $\mathbf{X}$.

We focus on the case of additive normally distributed measurement error with measurement error variance σ_u^2. Although the problem has this parametric error assumption, it also has a nonparametric component: no assumptions are made about the true predictors $\mathbf{X}$.

Suppose for the sake of discussion that the measurement error variance σ_u^2 is known. In the functional model, the unobservable $\mathbf{X}$'s are fixed constants, and hence the unknown parameters include the $\mathbf{X}$'s. With additive normally distributed measurement error, one strategy is to maximize the joint density of the observed data with respect to all of the unknown parameters including the $\mathbf{X}$'s. While this works for linear regression (Gleser [32]), it fails for more complex models such as logistic regression. Indeed, the logistic regression functional maximum likelihood estimator is both inconsistent and difficult to compute (Stefanski & Carroll [70]). An alternative approach is to change to the structural model and apply likelihood techniques (see below).

In this section, we consider two functional methods, the conditional-score and corrected-score methods. We start with logistic and gamma-loglinear modeling as important examples for which these techniques apply. The conditional methods exploit special structures in important models such as linear, logistic, Poisson loglinear and gamma-inverse, and then use a traditional statistical device, conditioning on sufficient statistics, to obtain estimators. The corrected-score method effectively estimates the estimator one would use if there were no measurement error.

First consider the multiple linear regression model with mean $\beta_0 + \beta_x \mathbf{X} + \beta_z \mathbf{Z}$, and write the unknown regression parameter as $\Theta = (\beta_0, \beta_x, \beta_z)$. When the measurement error is additive with non–differential measurement error variance Σ_{uu}, the usual method-of-moments regression estimator can be derived as the solution to the equation

$$\sum_{i=1}^{n} \psi_*(\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{W}_i, \Theta, \Sigma_{uu}) = 0, \qquad (9)$$

where

$$\psi_*(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, \Theta, \Sigma_{uu}) = \left(\mathbf{Y} - \beta_0 - \beta_x^t \mathbf{X} - \beta_z^t \mathbf{Z}\right) \begin{pmatrix} 1 \\ \mathbf{Z} \\ \mathbf{W} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \Sigma_{uu}\beta_x \end{pmatrix}$$

is the *corrected score* for linear regression. If Σ_{uu} is unknown, one substitutes an estimate of it into (9) and solves for the regression parameters.

The key point to note here is that in solving (9), we need know nothing about the $\mathbf{X}$'s. This feature is common to all the methods in this section.

Equation (9) is an example of an estimating equation approach for estimating a set of unknown parameters. The reader can consult the appendix of Carroll, et al. [19] for an overview of estimating equations, although this is unnecessary for the purpose of using the methods. Asymptotic standard errors for the estimators can be derived using either the bootstrap or the sandwich formula.

Logistic regression is best handled using the conditional-score method. For example, consider the usual linear-logistic model, where $\mathbf{Y}$ is binary and has success probability following the logistic model $H(\beta_0 + \beta_x \mathbf{X} + \beta_z \mathbf{Z})$. The conditional score is

$$\psi_*(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, \Theta, \sigma_u^2) = \left[\mathbf{Y} - H\left\{\beta_0 - \beta_x^t \Delta(\cdot) - .5\beta_x^t \sigma_u^2 \beta_x - \beta_z^t \mathbf{Z}\right\}\right] \begin{bmatrix} 1 \\ \mathbf{Z} \\ \Delta(\cdot) \end{bmatrix}, \qquad (10)$$

where $\Delta(\cdot) = \Delta(\mathbf{Y}, \mathbf{W}, \beta_x, \sigma_u^2) = \mathbf{W} + \mathbf{Y}\sigma_u^2\beta_x$. Equation (10) is substituted into (9), and the resulting equation is solved numerically.

When $\mathbf{Y}$ has a gamma distribution with loglinear mean $\exp(\beta_0 + \beta_x \mathbf{X} + \beta_z \mathbf{Z})$, it has variance which is ϕ times the square of the mean. For this important example, the corrected-score estimator is obtained from the corrected score

$$\psi_*(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, \Theta, \sigma_u^2) = \begin{pmatrix} 1 \\ \mathbf{Z} \\ \mathbf{W} \end{pmatrix} - \exp\left\{\Delta(\mathbf{Z}, \mathbf{W}, \Theta, \sigma_u^2)\right\} \begin{pmatrix} \mathbf{Y} \\ \mathbf{ZY} \\ \mathbf{Y}(\mathbf{W} + .5\sigma_u^2\beta_x) \end{pmatrix}, \quad (11)$$

where $\Delta(\mathbf{Z}, \mathbf{W}, \Theta, \sigma_u^2) = -\beta_0 - \beta_x^t \mathbf{W} - \beta_z^t \mathbf{Z} - .5\beta_x^t \sigma_u^2 \beta_x$.

Unbiased Score Functions via Conditioning

The conditional estimators of Stefanski & Carroll [71] are discussed in detail in Carroll, et al. [19], Chapter 6. They apply to linear, logistic, Poisson loglinear and Gamma inverse regression (the mean is $1/(\beta_0 + \beta_x \mathbf{X} + \beta_z \mathbf{Z})$). There methods have simple formulae for standard errors, although of course as usual the bootstrap applies.

Exact Corrected Estimating Equations

Suppose that it is possible to find a function of the observed data, say $\psi_*(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, \Theta)$, having the property that

$$E\left\{\psi_*(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, \Theta) \mid \mathbf{Y}, \mathbf{Z}, \mathbf{X}\right\} = \psi(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \Theta), \tag{12}$$

for all $\mathbf{Y}$, $\mathbf{Z}$, $\mathbf{X}$ and Θ. Then corrected score function estimators simply replace ψ by ψ_*. Corrected score functions satisfying (12) do not always exist and finding them when they do is not always easy.

One useful class of models that admit corrected functions contains those models with log-likelihoods of the form

$$\log\left\{f(y|z, x, \Theta)\right\} = \sum_{k=0}^{2}\left\{c_k(y, z, \Theta)(\beta_x^t x)^k\right\} + c_3(y, z, \Theta)\exp(\beta_x^t x),$$

see the examples given below. Then, using normal distribution moment generating function identities, the required function is

$$\psi_*(y, z, w, \Theta, \sigma_u^2) = \frac{\partial}{\partial \Theta^t}\left[\sum_{k=0}^{2}\left\{c_k(y, z, \Theta)(\beta_x^t w)^k\right\}\right.$$

$$-c_2(y, z, \Theta)\beta_x^t\sigma_u^2\beta_x + c_3(y, z, \Theta)\exp(\beta_x^t w - .5\beta_x^t\sigma_u^2\beta_x)\bigg].$$

Regression models in this class include:

- Normal linear with mean $= \eta$, variance $= \phi$, $c_0 = -(y - \beta_0 - \beta_z^t z)^2/(2\phi) - \log(\sqrt{\phi}\,)$, $c_1 = (y - \beta_0 - \beta_z^t z)/\phi$, $c_2 = -(2\phi)^{-1}$, $c_3 = 0$;

- Poisson with mean $= \exp(\eta)$, variance $= \exp(\eta)$, $c_0 = y(\beta_0 + \beta_z^t z) - \log(y!)$, $c_1 = y$, $c_2 = 0$, $c_3 = -\exp(\beta_0 + \beta_z^t z)$;

- Gamma with mean $= \exp(\eta)$, variance $= \phi\exp(2\eta)$, $c_0 = -\phi^{-1}(\beta_0 + \beta_z^t z) + (\phi^{-1} - 1)\log(y) + \phi^{-1}\log(\phi^{-1}) - \log\left\{, (\phi^{-1})\right\}$, $c_1 = \phi^{-1}$, $c_2 = 0$, $c_3 = -\phi^{-1}y\exp(-\beta_0 - \beta_z^t z)$.

Comparison of Methods

The methods are applicable at the same time only in linear regression (where they are identical) and Poisson regression. For Poisson regression the corrected estimating equations are more convenient because they are explicit, whereas the conditional estimator involves numerical summation. For Poisson regression the conditional-score estimator is more efficient than the corrected-score estimator in some practical cases.

INSTRUMENTAL VARIABLES

We have assumed that it was possible to estimate the measurement error variance, say with replicate measurements or validation data. However, it is not always possible to obtain replicates or validation data and thus direct estimation of the measurement error variance is sometimes impossible. In the absence of information about the measurement error variance, estimation of the regression model parameters is still possible provided the data contain an *instrumental variable* $\mathbf{T}$, in addition to the unbiased measurement $\mathbf{W} = \mathbf{X} + \mathbf{U}$.

There are three basic requirements that an instrumental variable must satisfy: (i) it must be correlated with $\mathbf{X}$; (ii) it must be independent of $\mathbf{W} - \mathbf{X}$; (iii) it must be a surrogate, i.e., subject to non–differential measurement error

One possible source of an instrumental variable is a second measurement of $\mathbf{X}$ obtained by an independent method. This second measurement need not be unbiased for $\mathbf{X}$. Thus the assumption that a variable is an instrument is weaker than the assumption that it follows the classical additive error model.

Instrumental variable estimation in linear models is covered in depth Fuller [30]. The work described here outside the linear model is based on that of Carroll & Stefanski [22] and Stefanski & Buzas [69]. Other references include Amemiya [1], [2] and Buzas & Stefanski [13].

We have found that instrumental variables require a slightly different notation. For example, $\beta_{Y|\underline{1}ZX}$ is the coefficient of $\mathbf{1}$, i.e., the intercept, in the regression of $\mathbf{Y}$ on $\mathbf{1}$, $\mathbf{Z}$ and $\mathbf{X}$; $\beta_{Y|1\underline{Z}X}$ is the coefficient of $\mathbf{Z}$ in the regression of $\mathbf{Y}$ on $\mathbf{1}$, $\mathbf{Z}$ and $\mathbf{X}$. This notation allows representation of subsets of coefficient vectors, e.g., $\beta_{Y|\underline{1}ZX} = (\beta_{Y|\underline{1}ZX}, \ \beta_{Y|1\underline{Z}X})$ and $\beta_{X|\underline{1}ZT} = (\beta_{X|\underline{1}ZT}, \ \beta_{X|1\underline{Z}T}, \ \beta_{X|1Z\underline{T}})$.

Our analysis is based upon regression calibration in generalized linear models, e.g., linear, logistic and Poisson regression. It might be useful to simply think of this section as dealing with a class of important models, the details of fitting of which are standard in many computer programs.

The approximate models and estimation algorithms are best described in terms of the composite vectors

$$\widetilde{\mathbf{X}} = (\mathbf{1}, \mathbf{Z}, \mathbf{X}), \quad \widetilde{\mathbf{W}} = (\mathbf{1}, \mathbf{Z}, \mathbf{W}) \text{ and } \widetilde{\mathbf{T}} = (\mathbf{1}, \mathbf{Z}, \mathbf{T}).$$

Define $\beta_{Y|\tilde{X}} = (\beta_{Y|\underline{1}ZX}, \beta_{Y|1\underline{Z}X}, \beta_{Y|1Z\underline{X}})$.

We note here that in addition to the assumptions stated previously, we will also assume that the regression of $\mathbf{X}$ on $(\mathbf{Z}, \mathbf{T}, \mathbf{W})$ is approximately linear. This restricts the applicability of our methods somewhat, but is sufficiently general to encompass many potential applications.

The simplest instrumental variables estimator starts with a (possibly multivariate) re-

gression of $\widetilde{\mathbf{W}}$ on $\widetilde{\mathbf{T}}$ to obtain $\widehat{\beta}_{\widetilde{W}|\underline{\widetilde{T}}}$. Then $\mathbf{Y}$ is regressed on the predicted values $\widehat{\beta}_{\widetilde{W}|\underline{\widetilde{T}}}\widetilde{\mathbf{T}}$, which results in an estimator of $\beta_{Y|\underline{\tilde{X}}}$.

This estimator is easily computed as it requires only linear regressions of the components of $\widetilde{\mathbf{W}}$ on $\widetilde{\mathbf{T}}$, and then the use of standard regression programs to regress $\mathbf{Y}$ on the "predictors" $\widehat{\beta}_{\widetilde{W}|\underline{\widetilde{T}}}\widetilde{\mathbf{T}}$.

Carroll, et al. [19] describe somewhat more elaborate methods of instrumental variable estimation, which can be more efficient than this simple method, especially if the number of components of T differs from the number of components of W.

LIKELIHOOD AND BAYESIAN STRUCTURAL METHODS

This section describes the use of likelihood methods in measurement error models. There have been a few examples in the literature based on likelihood. See Carroll, et al. [20] and Schafer [63], [64] and Tosteson, et al. [78] for probit regression, Whittemore & Gong [84] in a Poisson model, Crouch & Spiegelman [27], Satten & Kupper [62] and Wang, et al. [81] in logistic regression, and Küchenhoff & Carroll [38] in a change point problem. The relatively small literature belies the importance of the topic and the potential for further applications.

There are number of important differences between likelihood methods and the methods described in previous sections.

- The previous methods are based on additive or multiplicative measurement error models, possibly after a transformation. Typically, few if any distributional assumptions are required. Likelihood methods require stronger distributional assumptions, but they can be applied to more general problems, including those with discrete covariates subject to <u>MISCLASSIFICATION ERROR</u>*.

- The likelihood for a fully specified parametric model can be used to obtain likelihood ratio confidence intervals. In methods not based on likelihoods, inference is based on bootstrapping or on normal approximations. In highly nonlinear problems, likelihood-based confidence intervals are generally more reliable than those derived from normal approximations.

- Likelihood methods are often computationally more demanding, whereas the previous methods require little more than the use of standard statistical packages.

- Robustness to modeling assumptions is a concern for both approaches, but generally more difficult to understand with likelihood methods.

- There is a belief that the simpler methods described previously perform just as well as likelihood methods for many statistical models, including the most common generalized linear models. There is little documentation as to whether the folklore is realistic. The only evidence that we know of is given for logistic regression by Stefanski & Carroll [72], who contrast the maximum likelihood estimate and a particular functional estimate. They find that the functional estimate is fairly efficient relative to the maximum likelihood estimate unless the measurement error is "large" or the logistic coefficient is "large." One should be aware, however, that their calculations indicate that there are situations where *properly parameterized* maximum likelihood estimates are considerably more efficient than estimates derived from functional modeling.

Likelihood Specification: Differential and Non–Differential Error

We consider here only the simplest problem in which $\mathbf{X}$ is not observable for all subjects, but there are sufficient data, either internal or external, to characterize the distribution of $\mathbf{W}$ given $(\mathbf{X}, \mathbf{Z})$ (with validation data, we are in the realm of missing data). To perform a likelihood analysis, one must specify a parametric model for every component the data. Likelihood analysis starts with a model for the distribution of the response given the true predictors. The likelihood (density or mass) function of $\mathbf{Y}$ given $(\mathbf{Z}, \mathbf{X})$ will be called $f_{Y|Z,X}(y|z,x,\mathcal{B})$ here, and interest lies in estimating $\mathcal{B}$. For example, if $\mathbf{Y}$ is normally distributed with mean $\beta_0 + \beta_x \mathbf{X} + \beta_z \mathbf{Z}$ and variance σ^2, then $\mathcal{B} = (\beta_0, \beta_x, \beta_z, \sigma^2)$ and

$$f_{Y|Z,X}(y|z,x,\mathcal{B}) = \sigma^{-1}\phi\left\{(y - \beta_0 + \beta_x x + \beta_z z)/\sigma\right\},$$

where $\phi(v) = (2\pi)^{-1/2}\exp(-.5v^2)$ is the standard normal density function. If $\mathbf{Y}$ follows a logistic regression model with mean $H(\beta_0 + \beta_x \mathbf{X} + \beta_z \mathbf{Z})$, then $\mathcal{B} = (\beta_0, \beta_x, \beta_z)$ and

$$f_{Y|Z,X}(y|z,x,\mathcal{B}) = H^y(\beta_0 + \beta_x x + \beta_z z)\left\{1 - H(\beta_0 + \beta_x x + \beta_z z)\right\}^{1-y}.$$

A likelihood analysis starts with determination of the joint distribution of $\mathbf{Y}$ and $\mathbf{W}$ given $\mathbf{Z}$, as these are the observed variates. There are three components required:

- A model relating the response to the "true" covariates, see just above.

- An error model, here called $f_{W|Z,X}(w|z,x,\tilde{\alpha}_1)$. In many applications, the error model does not depend on $\mathbf{Z}$. For example, in the classical additive measurement error model (1) with normally distributed measurement error, σ_u^2 is the only component of $\tilde{\alpha}_1$, and the error model density is $\sigma_u^{-1}\phi\left\{(w-x)/\sigma_u\right\}$, where $\phi(\cdot)$ is the standard normal density function. In the classical error model with independent replicates, $\mathbf{W}$ consists

of the k replicates, and $f_{W|Z,X}$ is the k-variate normal density function with mean zero, common variance σ_u^2, and zero correlation. A generalization of this error model that allows for correlations among the replicates has been studied (Wang, et al. [81]). In some application areas, error model structures are studied independently of their role in measurement error modeling, and one can use this research to estimate error models for the problem at hand.

- A model for the distribution of the latent variable, here called $f_{X|Z}(x|z, \tilde{\alpha}_2)$. Specifying a model for the distribution of the true covariate $\mathbf{X}$ given all the other covariates $\mathbf{Z}$ is more difficult. Difficulties arise because: (a) the distribution is usually not transportable, so that different studies yield very different models; and (b) $\mathbf{X}$ is not observed.

Having hypothesized the various models, the likelihood of the observed data under non–differential measurement error is

$$f_{Y,W|Z}(y, w|z, \mathcal{B}, \tilde{\alpha}_1, \tilde{\alpha}_2) = \int f_{Y|Z,X}(y|z, x, \mathcal{B}) f_{W|Z,X}(w|z, x, \tilde{\alpha}_1) f_{X|Z}(x|z, \tilde{\alpha}_2) d\mu(x). \quad (13)$$

The notation $d\mu(x)$ indicates that the integrals are sums if $\mathbf{X}$ is discrete and integrals if $\mathbf{X}$ is continuous. The likelihood for the problem is just the product over the sample of these terms.

There is a significant difference between the likelihood function in the differential and non–differential cases. This can be expressed in various ways, but the simplest is as follows. In general, and dropping parameters, the likelihood of the observed data is

$$f_{Y,W|Z}(y, w|z) = \int f_{Y,W,X|Z}(y, w, x|z) d\mu(x).$$

Using standard conditioning arguments, this becomes

$$
\begin{aligned}
f_{Y,W|Z}(y, w|z) &= \int f_{W|Y,Z,X}(w|y, z, x) f_{Y|Z,X}(y|z, x) f_{X|Z}(x|z) d\mu(x) \\
&= \int f_{Y|Z,X}(y|z, x) f_{W|Y,Z,X}(w|y, z, x) f_{X|Z}(x|z) d\mu(x). \quad (14)
\end{aligned}
$$

Note that the only difference between (13) and (14) is in the error term. In the former, under non–differential measurement error, $\mathbf{W}$ and $\mathbf{Y}$ are independent, so that $f_{W|Y,Z,X}(w|y, z, x)$ $= f_{W|Z,X}(w|z, x)$.

What makes differential error so difficult is that, under differential measurement error, we must ascertain the distribution of $\mathbf{W}$ given the other covariates *and the response* $\mathbf{Y}$. This is essentially impossible to do in practice unless one has a subset of the data in which all of $(\mathbf{Y}, \mathbf{Z}, \mathbf{X}, \mathbf{W})$ are observed, i.e., a *validation* data set [*see* <u>MISSING DATA IN EPIDEMIOLOGIC STUDIE</u> and <u>VALIDATION STUDY</u>[*]].

Numerical Computation of Likelihoods

Typically one maximizes the logarithm of the overall likelihood in the unknown parameters. There are two ways one can maximize the likelihood function. The most direct is to compute the likelihood function itself, and then use numerical optimization techniques to maximize the likelihood. Below we provide a few details about computing the likelihood function. The second general approach is to view the problem as a missing data problem, and then use missing data techniques, see for example Little & Rubin [42] and Tanner [75].

Computing the likelihood analytically is easy if $\mathbf{X}$ is discrete, as the conditional expectations are simply sums of terms. Likelihoods in which $\mathbf{X}$ has some continuous components can be computed using a number of different approaches. In some problems the loglikelihood can be computed or very well approximated analytically. In most problems that we have encountered, $\mathbf{X}$ is a scalar or a 2×1 vector. In these cases, standard numerical methods such as Gaussian quadrature can be applied, although they are not always very good. When sufficient computing resources are available, the likelihood can be computed using Monte-Carlo techniques.

Bayesian Methods

Bayesian estimation and inference in the measurement error problem is a promising approach under active development [*see* Bayesian Methods].. Examples of this approach are given by Schmid & Rosner [65], Richardson & Gilks [57], Stephens & Dellaportas [74], Müller & Roeder [47], Mallick & Gelfand [45] and Kuha [40].

Bayesian analysis of parametric models requires specifying a likelihood (as described above) and a prior distribution for the parameters, the latter representing knowledge about the parameters prior to data collection. The product of the prior and likelihood is the joint density of the data and the parameters. Using Bayes' Theorem, one can in principle obtain the posterior density, i.e., the conditional density of the parameters given the data. The posterior summarizes all of the information about the values of the parameters and is the basis for all Bayesian inference. For example, the mean, median, or mode of the posterior density are all suitable point estimators. A region with probability $(1-\alpha)$ under the posterior is called a "credible set," and is a Bayesian analog to a confidence region.

Computing the posterior distribution is often a non-trivial problem, because it usually requires high-dimensional numerical integration. This computational problem is the subject of much recent research, with many major advances. The method currently receiving the most attention in the literature is the Gibbs sampler, see Smith & Gelfand [66] and Casella & George [24]. Also, see Tanner [75] for a book-length introduction to modern methods for computing posteriors distributions.

In the Bayesian approach with Gibbs sampling, the $\mathbf{X}$'s are treated as "missing data" (they just happen to be missing for all study subjects unless there is a validation study!). The approach for the classical additive error model is:

- Assuming non–differential error, write the likelihood of $\mathbf{Y}$ given $(\mathbf{X}, \mathbf{Z})$. the likelihood of $\mathbf{W}$ given $(\mathbf{X}, \mathbf{Z})$, and the likelihood of $\mathbf{X}$ given $\mathbf{Z}$ depending on parameters, just as in a regular likelihood problem.

- If $\mathbf{X}$ were observable, the likelihood would be the product of the three terms given above.

- Select a starting value for the parameters, e.g., from SIMEX.

- Use a simulation approach to fill in the "missing" $\mathbf{X}$'s, i.e., from the posterior distribution of $\mathbf{X}$ given the observed data and the current values of the parameters. In this step, it is rare that the posterior distribution is known exactly, and so one has to use a device such as the Metropolis–Hastings algorithm.

- Now one has complete data, with $\mathbf{X}$'s all filled in, and one uses simulation to draw a sample of parameters from the posterior distribution of the parameters given the observed data and the current $\mathbf{X}$'s.

- Repeat the process of generating $\mathbf{X}$ and the parameters. These multiple samples of parameters are used to evaluate features of the posterior distribution.

While the procedure is easy to write down, the computations may be difficult.

More importantly though is the need to consider the distribution of $\mathbf{X}$ given $\mathbf{Z}$. As we emphasized above, the simplest structural approach assumes that $\mathbf{X}$ is normally distributed, but this is often a strong assumption. The popularity of functional methods lies in the fact that such methods require no distributional assumptions about the $\mathbf{X}$'s. There is considerable current effort being made to get around the problem of model robustness by specifying a flexible distribution for $\mathbf{X}$.

Mixture Modeling

When there are no covariates measured without error, the nonlinear measurement error problem can be viewed as a special case of what are called mixture problems; see Titterington, et al. [77]. The idea is to pretend that $\mathbf{X}$ has a distribution, but to estimate this distribution nonparametrically. Applications of nonparametric mixture methods to nonlinear measurement error models have only recently been described by Thomas, et al. [76] and Roeder, et al. [58].

An alternative formulation is to let **X** have a flexible distribution, which covers a wide range of possibilities including the normal distribution. The simplest such model is the mixture of normals, which has been applied by Wang, et al. [81] and by Küchenhoff & Carroll [38].

RESPONSE ERROR

In preceding sections we have focused exclusively on problems associated with measurement error in predictor variables. Here we consider problems that arise when a true response is measured with error. For example, in a study of factors affecting dietary intake of fat, e.g., sex, race, age, socioeconomic status, etc., true long-term dietary intake is impossible to determine and instead it is necessary to use error-prone measures of long-term dietary intake. Wittes, et al. [86] describe another example in which damage to the heart muscle caused by a myocardial infarction can be assessed accurately, but the procedure is expensive and invasive, and instead it is common practice to use peak cardiac enzyme level in the bloodstream as a proxy for the true response.

When the response is binary (case or control say), see <u>MISCLASSIFICATION ERROR</u>*.

The exclusive attention paid to predictor measurement error earlier in this article is explained by the fact that predictor measurement error is seldom ignorable, by which is meant that the usual method of analysis is statistically valid, whereas response measurement error is often ignorable when the response is continuous. Here "ignorable" means that the model holding for the true response holds also for the proxy response with parameters unchanged, except that a measurement error variance component is added to the response variance. For example, in linear regression models with simple types of response measurement error, the response measurement error is confounded with equation error and the effect is simply to increase the variability of parameter estimates. Thus, response error is ignorable in these cases, although of course power will be lost. However, in more complicated regression models, certain types of response error are not ignorable and it is important to explicitly account for the response error in the regression analysis.

Although the details differ between methods for predictor error and response error, many of the basic ideas are similar. Throughout this section, the response proxy is denoted by **S**. We consider only the case of measurement error in the response, and not the more complex problem where both the response and some of the predictors are measured with error.

We first consider the analysis of the observed data when the response is subject to in-dependent additive or multiplicative measurement error. Suppose that the proxy response **S** is unbiased for the true response. Then in either case, the proxy response has the same mean (as a function of exposure and confounders) as the true response, although the variance

structure differs. In models such as linear regression, or more generally for quasilikelihood estimation [*see* QUASILIKELIHOOD FUNCTIONS], this means that the parameter estimates are consistent, but inferences may be affected. For example, in linear regression, additive, unbiased response error does not change the mean and simply increases the variance by a constant, so that there is no effect of measurement error other than loss of power. However, for multiplicative, unbiased response error, while the mean remains unchanged, the variances now are no longer constant, and hence inferences which pretend that the variances are constant would be affected. The usual solution is to use a robust covariance estimator, also known as the sandwich estimator, [*see* GENERALIZED ESTIMATING EQUATIONS].

If the proxy response **S** is not unbiased for the true response, then a validation study is required to understand the nature of the bias and to correct for it. In a series of papers, Buonaccorsi [8], [9], [11] and Buonaccorsi & Tosteson [12] discuss the use of adjustments for a biased response. See Carroll, et al. [19] for further details.

We will call **S** a *surrogate response* if its distribution depends only on the true response and not otherwise on the covariates, i.e., the information about the surrogate response contained in the true response is the same no matter what the values of the covariates. In symbols, if $f_{\mathbf{S}|\mathbf{Y},\mathbf{Z},\mathbf{X}}(s|y,z,x,\gamma)$ denotes the density or mass function for **S** given $(\mathbf{Y},\mathbf{Z},\mathbf{X})$, then $f_{\mathbf{S}|\mathbf{Y},\mathbf{Z},\mathbf{X}}(s|y,z,x,\gamma) = f_{\mathbf{S}|\mathbf{Y}}(s|y,\gamma)$. In both the additive and multiplicative error models, **S** is a surrogate. This definition of a surrogate response is the natural counterpart to a surrogate predictor, because it implies that all the information in the relationship between **S** and the predictors is explained by the underlying response. See Prentice [53] and Carroll, et al. [19] for further details.

In general, i.e., for a possibly nonsurrogate response, the likelihood function for the observed response is

$$f_{\mathbf{S}|\mathbf{Z},\mathbf{X}}(s|z,x,\mathcal{B},\gamma) = \int f_{\mathbf{Y}|\mathbf{Z},\mathbf{X}}(y|z,x,\mathcal{B})f_{\mathbf{S}|\mathbf{Y},\mathbf{Z},\mathbf{X}}(s|y,z,x,\gamma)d\mu(y). \tag{15}$$

There are a number of implications of this formula.

- If **S** is a surrogate, and if there is no relationship between the true response and the predictors, then neither is there one between the observed response and the predictors. Hence, if interest lies in determining whether *any of the predictors* contain any information about the response, one can use naive hypothesis tests and ignore response error. The resulting tests have asymptotically correct level, but decreased power relative to tests derived from true response data. This property of a surrogate is important in clinical trials; see Prentice [53].

- If **S** is *not* a surrogate, then there may be no relationship between the true response

and the covariates, but the observed response may be related to the predictors. Hence, naive tests will not be valid in general if **S** is not a surrogate.

Note that one implication of (15) is that a likelihood analysis with mismeasured responses requires a model for the distribution of response error. Except for additive and multiplicative error, understanding such a model requires a validation study.

CASE-CONTROL STUDIES*

A *case–control study* is one in which sampling is conditioned on the disease response; it is useful to think that the response is first observed and only later are the predictors observed. A similar design, *choice-based sampling* is used in in econometrics. We will use case–control former terminology and concentrate on logistic regression models. A distinguishing feature of case-control studies is that the measurement error may be differential.

Two-stage case–control designs, where **X** is observed on a subset of the data have been studied by Breslow & Cain [7], Zhao & Lipsitz [87]. Tosteson & Ware [80], and Carroll, Gail & Lubin [14], among others. These designs are significant because the validation, if done on both cases and controls, frees us from the non–differential error assumption.

We assume that the data follow a logistic model in the underlying source population, although the results apply equally well to the more general models described by Weinberg & Wacholder [82]. For such models, Prentice & Pyke [55] and Weinberg & Wacholder [82] show that when analyzing a classical case-control study one can ignore the case-control sampling scheme entirely, at least for the purpose of estimating relative risk. Furthermore, these authors show that if one *ignores the case-control sampling scheme and runs an ordinary logistic regression*, then the resulting relative risk estimates are consistent and the standard errors are asymptotically correct.

The effect of measurement error in logistic case-control studies is to bias the estimates. Carroll, et al. [23] show that for many problems, one can ignore the case-control study design and proceed to correct for the bias from measurement error as if one were analyzing a random sample from the source population. With non–differential measurement error, this result applies to the methods we have described previously for prospective studies. Regression calibration needs a slight modification, namely that the regression calibration function should be estimated using the controls only.

Michalek & Tripathi [46], Armstrong, et al. [5] and Buonaccorsi [10] consider the normal discriminant model. Satten & Kupper [62] have an interesting example of likelihood analysis for non–differential error validation studies when the validation sampling is in the controls.

SURVIVAL ANALYSIS

One of the earliest applications of the regression calibration method was discussed by Prentice [52] in the context of survival analysis. Further results in survival analysis were obtained by Pepe, et al. [50], Clayton [25], Nakamura [49] and Hughes [36]. While the details differ in substantive ways, the ideas are the same as put forward in the rest of this article, and here we provide only a very brief overview in the case of covariates which do not depend on time.

Suppose that the instantaneous risk that the time $\mathbf{T}$ of an event equals t conditional on no events prior to time t and conditional on the true covariate $\mathbf{X}$ is denoted by

$$\psi\left(t, \mathbf{X}\right) = \psi_0(t)\exp\left(\beta_x \mathbf{X}\right), \tag{16}$$

where $\psi_0(t)$ is the baseline hazard function. When the baseline hazard is not specified, (16) is commonly called the proportional hazards assumption, [see PROPORTIONAL HAZARDS OVERVIEW]. When $\mathbf{X}$ is observable, it is well-known that estimation of β_x is possible without specifying the form of the baseline hazard function.

If $\mathbf{X}$ is unobservable and instead we observe a surrogate $\mathbf{W}$, the induced hazard function is

$$\psi^*\left(t, \mathbf{W}, \beta_x\right) = \psi_0(t)E\left\{\exp\left(\beta_x \mathbf{X}\right)|T \geq t, \mathbf{W}\right\}. \tag{17}$$

The difficulty is that the expectation in (17) for the observed data depends upon the unknown baseline hazard function ψ_0. Thus, the hazard function does not factor into a product of an arbitrary baseline hazard times a term that depends only on observed data and an unknown parameter, and the technology for proportional hazards regression cannot be applied without modification.

The problem simplifies when then event is rare, so that $T \geq t$ occurs with high probability for all t under consideration. As shown by Prentice [53] and others, under certain circumstances this leads to the regression calibration algorithm. The rare event assumption allows the hazard of the observed data to be approximated by

$$\psi^*\left(t, \mathbf{W}, \beta_x\right) = \psi_0(t)E\left\{\exp\left(\beta_x \mathbf{X}\right)|\mathbf{W}\right\}. \tag{18}$$

The hazard function (18) requires a regression calibration formulation! If one specifies a model for the distribution of $\mathbf{X}$ given $\mathbf{W}$, then (18) is in the form of a proportional hazards model (16), but with $\beta_x \mathbf{X}$ replaced by $\log\left(E\left\{\exp\left(\beta_x \mathbf{X}\right)|\mathbf{W}\right\}\right)$. An important special case leads directly to the standard regression calibration model, namely when $\mathbf{X}$ given $\mathbf{W}$ is normally distributed.

Clayton [25] proposed a modification of regression calibration which does not require events to be rare. At each time t_i, $i = 1, ..., k$, for which an event occurs, define the risk set $\mathcal{R}_i \subseteq \{1, ..., n\}$ as the case numbers of those members of the study cohort for whom

an event has not occurred and who were still under study just prior to t_i. If the $\mathbf{X}$'s were observable, and if $\mathbf{X}_i$ is the covariate associated with the ith event, in the absence of ties the usual proportional hazards regression would maximize

$$\prod_{i=1}^{k} \frac{\exp(\beta_x \mathbf{X}_i)}{\sum_{j \in \mathcal{R}_i} \exp(\beta_x \mathbf{X}_j)}.$$

Clayton basically suggests using regression calibration within each risk set. He assumes that the true values $\mathbf{X}$ within the ith risk set are normally distributed with mean μ_i and variance σ_x^2, and that within this risk set $\mathbf{W} = \mathbf{X} + \mathbf{U}$, where $\mathbf{U}$ is normally distributed with mean zero and variance σ_u^2. Neither σ_x^2 nor σ_u^2 depend upon the risk set in his formulation.

Given an estimate $\widehat{\sigma}_u^2$, one applies the usual regression calibration calculations to construct an estimate of $\widehat{\sigma}_x^2$.

Clayton modifies regression calibration by using it within each risk set. Within each risk set, he applies the formula (7) for the best unbiased estimate of the $\mathbf{X}$'s. Specifically, in the absence of replication, for any member of the ith risk set, the estimate of the true covariate $\mathbf{X}$ from an observed covariate $\mathbf{W}$ is

$$\widehat{\mathbf{X}} = \widehat{\mu}_i + \frac{\widehat{\sigma}_x^2}{\widehat{\sigma}_x^2 + \widehat{\sigma}_u^2} \left(\mathbf{W} - \widehat{\mu}_i \right),$$

where $\widehat{\mu}_i$ is the sample mean of the $\mathbf{W}$'s in the ith risk set.

As with regression calibration in general, the advantage of Clayton's method is that no new software need be developed, other than to calculate the means within risk sets.

References

[1] Amemiya, Y. (1985). Instrumental variable estimator for the nonlinear errors in variables model. *Journal of Econometrics*, 28, 273-289.

[2] Amemiya, Y. (1990a). Instrumental variable estimation of the nonlinear measurement error model. In *Statistical Analysis of Measurement Error Models and Application*, P. J. Brown and W. A. Fuller, editors. American Mathematics Society, Providence.

[3] Amemiya, Y. & Fuller, W. A. (1988). Estimation for the nonlinear functional relationship. *Annals of Statistics*, 16, 147-160.

[4] Armstrong, B. (1985). Measurement error in generalized linear models. *Communications in Statistics, Series B*, 14, 529-544.

[5] Armstrong, B. G., Whittemore, A. S., & Howe, G. R. (1989). Analysis of case-control data with covariate measurement error: application to diet and colon cancer. *Statistics in Medicine*, 8, 1151–1163.

[6] Berkson, J. (1950). Are there two regressions? *Journal of the American Statistical Association*, 45, 164-180.

[7] Breslow, N. E. & Cain, K. C. (1988). Logistic regression for two–stage case–control data. *Biometrika*, 75, 11–20.

[8] Buonaccorsi, J. P. (1988). Errors in variables with systematic biases. *Communications in Statistics, Theory and Methods*, 18(3), 1001- 1021.

[9] Buonaccorsi, J. P. (1990a). Double sampling for exact values in some multivariate measurement error problems. *Journal of the American Statistical Association*, 85, 1075–1082.

[10] Buonaccorsi, J. P. (1990b). Double sampling for exact values in the normal discriminant model with application to binary regression. *Communications in Statistics, Series A*, 19, 4569–4586.

[11] Buonaccorsi, J. P. (1991) Measurement error, linear calibration and inferences for means. *Computational Statistics and Data Analysis*, 11, 239-257.

[12] Buonaccorsi, J. P. & Tosterson, T. (1993). Correcting for nonlinear measurement error in the dependent variable in the general linear model. *Communications in Statistics, Theory & Methods*, 22, 2687–2702.

[13] Buzas, J. S. & Stefanski, L. A. (1995). A note on corrected score estimation. *Statistics & Probability Letters*, 28, 1–8.

[14] Carroll, R. J., Gail, M. H., & Lubin, J. H. (1993). Case–control studies with errors in predictors. *Journal of the American Statistical Association*, 88, 177–191.

[15] Carroll, R. J., Gallo, P. P. & Gleser, L. J. (1985). Comparison of least squares and errors-in-variables regression, with special reference to randomized analysis of covariance. *Journal of the American Statistical Association, 80*, 929–932.

[16] Carroll, R. J., Küchenhoff, H., Lombard, F., & Stefanski, L. A. (1996). Asymptotics for the SIMEX estimator in structural measurement error models. *Journal of the American Statistical Association*, 91, 242–250.

[17] arroll, R. J. & Ruppert, D. (1988). *Transformation and Weighting in Regression*. Chapman & Hall, London.

[18] Carroll, R. J. & Ruppert, D. (1996). The use and misuse of orthogonal regression in measurement error models. *American Statistician*, 50, 1–6.

[19] Carroll, R. J., Ruppert, D. & Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models*, Chapman & Hall, London.

[20] Carroll, R. J., Spiegelman, C., Lan, K. K., Bailey, K. T. & Abbott, R. D. (1984). On errors-in-variables for binary regression models. *Biometrika, 71*, 19–26.

[21] Carroll, R. J. & Stefanski, L. A. (1990) Approximate quasilikelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association*, 85, 652–663.

[22] Carroll, R. J. & Stefanski, L. A. (1994). Measurement error, instrumental variables and corrections for attenuation with applications to meta-analyses. *Statistics in Medicine*, 13, 1265–1282.

[23] Carroll, R. J., Wang, S. & Wang, C. Y. (1995). Asymptotics for prospective analysis of stratified logistic case–control studies. *Journal of the American Statistical Association*, 90, 157–169.

[24] Casella, G. & George, E. I. (1992). Explaining the Gibbs sampler. *American Staistician*, 46, 167–174.

[25] Clayton, D. G. (1991). Models for the analysis of cohort and case-control studies with inaccurately measured exposures. In *Statistical Models for Longitudinal Studies of Health*, Dwyer, J. H., Feinleib, M., Lipsert, P., et al., editors, pages 301-331. Oxford University Press, New York.

[26] Cook, J. & Stefanski, L. A. (1995). A simulation extrapolation method for parametric measurement error models. *Journal of the American Statistical Association*, 89, 1314–1328.

[27] Crouch, E. A. & Spiegelman, D. (1990). The evaluation of integrals of the form $\int_{-\infty}^{\infty} f(t)\exp(-t^2)dt$: applications to logistic–normal models. *Journal of the American Statistical Association*, 85, 464–467.

[28] Dosemeci, M., Wacholder, S. & Lubin, J. H. (1990). Does non–differential misclassification of exposure always bias a true effect towards the null value? *American Journal of Epidemiology*, 132, 746–748.

[29] Freedman, L. S., Carroll, R. J. & Wax, Y. (1991). Estimating the relationship between dietary intake obtained from a food frequency questionnaire and true average intake. *American Journal of Epidemiology*, 134, 510–520.

[30] Fuller, W. A. (1987). *Measurement Error Models*. John Wiley & Sons, New York.

[31] Ganse, R. A., Amemiya, Y. & Fuller, W. A. (1983). Prediction when both variables are subject to error, with application to earthquake magnitude. *Journal of the American Statistical Association*, 78, 761-765.

[32] Gleser, L. J. (1981). Estimation in a multivariate errors in variables regression model: large sample results. *Annals of Statistics*, 9, 24–44.

[33] Gleser, L. J. (1990). Improvements of the naive approach to estimation in nonlinear errors-in-variables regression models. In *Statistical Analysis of Measurement Error Models and Application*, P. J. Brown and W. A. Fuller, editors. American Mathematics Society, Providence.

[34] Greenland, S. (1980). The effect of misclassification in the presence of covariates. *American Journal of Epidemiology*, 112, 564–569.

[35] Greenland, S. & Robins, J. M. (1985). Confounding and misclassification. *American Journal of Epidemiology*, 122, 495–506.

[36] Hughes, M. D. (1993). Regression dilution in the proportional hazards model. *Biometrics*, 49, 1056–1066.

[37] Hunter, D. J., Spiegelman, D., Adami, H.-O., Beeson, L., van der Brandt, P. A., Folsom, A. R., Fraser, G. E., Goldbohm, A., Graham, S., Howe, G. R., Kushi, L. H., Marshall, J. R., McDermott, A., Miller, A. B., Speizer, F. E., Wolk, A., Yaun, S. S. & Willett, W. (1996). Cohort studies of fat intake and the risk of breast cancer – a pooled analysis. *New England Journal of Medicine*, 334, 356–361.

[38] Küchenhoff, H. & Carroll, R. J. (1997). Biases in segmented regression with errors in predictors. *Statistics in Medicine*, in press.

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

[40] Kuha, J. (1997). Estimation by data augmentation in regression models with continuous and discrete covariates measured with error'. *Statistics in Medicine*, in press.

[41] Li, L., Freedman, L., Kipnis, V. and Carroll, R. J. (1997). Effects of bias and correlated measurement errors in the validation of food frequency questionnaires. Preprint.

[42] Little, R. J. A., & Rubin, D. B. (1987). *Statistical Analysis with Missing Data* John Wiley & Sons, New York.

[43] Liu, X. & Liang, K. Y. (1992). Efficacy of repeated measures in regression models with measurement error. *Biometrics*, 48, 645–654.

[44] MacMahon, S., Peto, R., Cutler, J., Collins, R., Sorlie, P., Neaton, J., Abbott, R., Godwin, J., Dyer, A., & Stamler, J. (1990). Blood pressure, stroke and coronary heart disease: Part 1, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. *Lancet*, 335, 765–774.

[45] Mallick, B. K. & Gelfand, A. E. (1996). Semiparametric errors-in-variables models: a Bayesian approach. *Journal of Statistical Planning and Inference*, 52, 307–322.

[46] Michalek, J. E. & Tripathi, R. C. (1980). The effect of errors in diagnosis and measurement on the probability of an event. *Journal of the American Statistical Association*, 75, 713–721.

[47] Müller, P. & Roeder, K. (1997). A Bayesian semiparametric model for case-control studies with errors in variables. *Biometrika*, to appear.

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

[49] Nakamura, T. (1992). Proportional hazards models with covariates subject to measurement error. *Biometrics*, 48, 829–838.

[50] Pepe, M. S., Self, S. G. & Prentice, R. L. (1989). Further results in covariate measurement errors in cohort studies with time to response data. *Statistics in Medicine*, 8, 1167–1178.

[51] Pierce, D. A., Stram, D. O., Vaeth, M., Schafer, D. (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.

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

[53] Prentice, R. L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine*, 8, 431–440.

[54] Prentice, R. L. (1996). Dietary fat and breast cancer: measurement error and results from analytic epidemiology. *Journal of the National Cancer Institute*, 88, 1738–1747.

[55] Prentice, R. L. & Pyke, R. (1979). Logistic disease incidence models and case–control studies. *Biometrika*, 66, 403–411.

[56] Racine–Poon, A., Weihs, C. & Smith, A. F. M. (1991). Estimation of relative potency with sequential dilution errors in radioimmunoassay. *Biometrics*, 47, 1235–1246.

[57] Richardson, S. & Gilks, W. R. (1993). A Bayesian approach to measurement error problems in epidemiology using conditional independence models. *American Journal of Epidemiology*, 138, 430–442.

[58] Roeder, K., Carroll, R. J. & Lindsay, B. G. (1996). A nonparametric mixture approach to case-control studies with errors in covariables. *Journal of the American Statistical Association*, 91, 722–732.

[59] Rosner, B., Spiegelman, D. & Willett, 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.

[60] Rosner, B., Willett, W. C. & Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine*, 8, 1051-1070.

[61] Rudemo, M., Ruppert, D. & Streibig, J. C. (1989). Random effect models in nonlinear regression with applications to bioassay. *Biometrics*, 45, 349-362.

[62] Satten, G. A. & Kupper, L. L. (1993). Inferences about exposure–disease association using probability of exposure information. *Journal of the American Statistical Association*, 88, 200–208.

[63] Schafer, D. (1987). Covariate measurement error in generalized linear models. *Biometrika*, 74, 385-391.

[64] Schafer, D. (1993). Likelihood analysis for probit regression with measurement errors. *Biometrika*, 80, 899–904.

[65] Schmid, C. H. & Rosner, B. (1993). A Bayesian approach to logistic regression models having measurement error following a mixture distribution. *Statistics in Medicine*, 12, 1141–1153.

[66] Smith, A. F. M. & Gelfand, A. E. (1992) Bayesian statistics without tears: a sampling–resampling perspective. *American Statistician*, 46, 84–88.

[67] Stefanski, L. A. (1985). The effects of measurement error on parameter estimation. *Biometrika*, 72, 583-592.

[68] Stefanski, L. A. (1989). Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models. *Communications in Statistics, Series A*, 18, 4335–4358.

[69] Stefanski, L. A. & Buzas, J. S. (1995). Instrumental variable estimation in binary regression measurement error models. *Journal of the American Statistical Association*, 90, 541–549.

[70] Stefanski, L. A. & Carroll, R. J. (1985). Covariate measurement error in logistic regression. *Annals of Statistics*, 13, 1335–1351.

[71] Stefanski, L. A. & Carroll, R. J. (1987). Conditional scores and optimal scores in generalized linear measurement error models. *Biometrika*, 74, 703-716.

[72] Stefanski, L. A. & Carroll, R. J. (1990b). Structural logistic regression measurement error models. *Proceedings of the Conference on Measurement Error Models*, P. J. Brown & W. A. Fuller, editors.

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

[74] Stephens, D. A. & Dellaportas, P. (1992). Bayesian Analysis of Generalized Linear Models with Covariate Measurement Error. In *Bayesian Statistics 4*, Bernado, J. M., Berger, J. O., Dawid, A. p. and Smith, A. F. M. (Eds.), pp 813–820, Oxford University Press.

[75] Tanner, M. A. (1993). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions, second edition.* Springer–Verlag, New York.

[76] Thomas, D., Stram, D. & Dwyer, J. (1993). Exposure measurement error: influence on exposure-disease relationships and methods of correction. *Annu. Rev. Publ. Health*, 14, 69–93.

[77] Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions.* John Wiley & Sons, New York.

[78] Tosteson, T., Stefanski, L. A. & Schafer D.W. (1989). A measurement error model for binary and ordinal regression. *Statistics in Medicine*, 8, 1139-1147.

[79] Tosteson, T. & Tsiatis, A. (1988). The asymptotic relative efficiency of score tests in a generalized linear model with surrogate covariates. *Biometrika*, 75, 507-514.

[80] Tosteson, T. D. & Ware, J. H. (1990). Designing a logistic regression study using surrogate measures of exposure and outcome. *Biometrika*, 77, 11–20.

[81] Wang, N., Carroll, R. J. & Liang, K. Y. (1996). Quasilikelihood and variance functions in measurement error models with replicates. *Biometrics*, 52, 401–411.

[82] Weinberg, C. R. & Wacholder, S. (1993). Prospective analysis of case–control data under general multiplicative–intercept models. *Biometrika*, 80, 461–465.

[83] Whittemore, A. S. (1989). Errors in variables regression using Stein estimates. *American Statistician*, 43, 226–228.

[84] Whittemore, A. S. & Gong, G. (1991). Poisson regression with misclassified counts: application to cervical cancer mortality rates. *Applied Statistics*, 40, 81–93.

[85] Whittemore, A. S. & Keller, J. B. (1988). Approximations for regression with covariate measurement error. *Journal of the American Statistical Association*, 83, 1057-1066.

[86] Wittes, J., Lakatos, E. & Probstfield, J. (1989). Surrogate endpoints in clinical trials: cardiovascular trials. *Statistics in Medicine*, 8, 415–425.

[87] Zhao, L. P. & Lipsitz, S. (1992). Designs and analysis of two–stage studies. *Statistics in Medicine*, 11, 769–782.