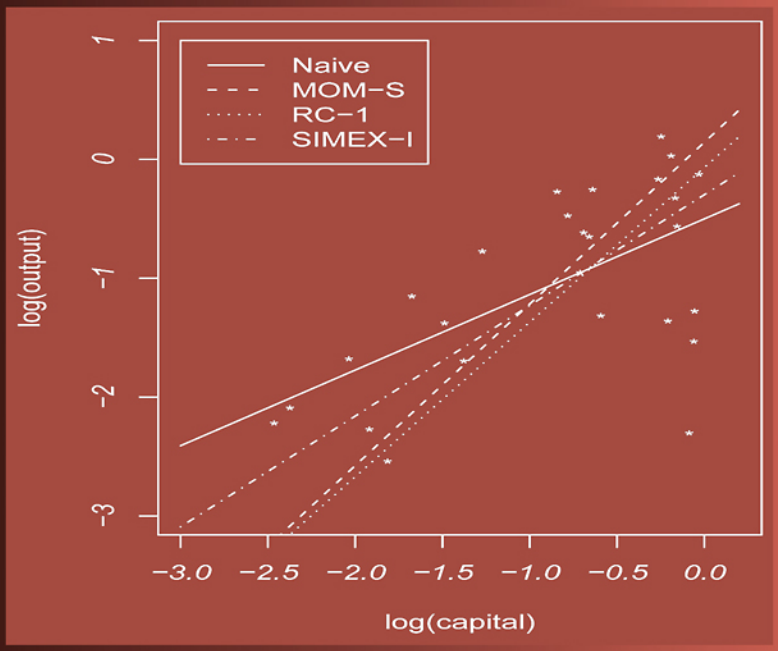
Chapman & Hall/CRC
Interdisciplinary Statistics Series

MEASUREMENT ERROR

MODELS, METHODS, and APPLICATIONS



John P. Buonaccorsi

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

MEASUREMENT ERROR

MODELS, METHODS, and APPLICATIONS

CHAPMAN & HALL/CRC

Interdisciplinary Statistics Series

Series editors: N. Keiding, B.J.T. Morgan, C.K. Wikle, P. van der Heijden

Published titles

AN INVARIANT APPROACH TO STATISTICAL ANALYSIS OF SHAPES	S. Lele and J. Richtsmeier
ASTROSTATISTICS	G. Babu and E. Feigelson
BAYESIAN ANALYSIS FOR POPULATION ECOLOGY	Ruth King, Byron J.T. Morgan, Olivier Gimenez, and Stephen P. Brooks
BAYESIAN DISEASE MAPPING: HIERARCHICAL MODELING IN SPATIAL EPIDEMIOLOGY	Andrew B. Lawson
BIOEQUIVALENCE AND STATISTICS IN CLINICAL PHARMACOLOGY	S. Patterson and B. Jones
CLINICAL TRIALS IN ONCOLOGY, SECOND EDITION	J. Crowley, S. Green, and J. Benedetti
CLUSTER RANDOMISED TRIALS	R.J. Hayes and L.H. Moulton
CORRESPONDENCE ANALYSIS IN PRACTICE, SECOND EDITION	M. Greenacre
DESIGN AND ANALYSIS OF QUALITY OF LIFE STUDIES IN CLINICAL TRIALS, SECOND EDITION	D.L. Fairclough
DYNAMICAL SEARCH	L. Pronzato, H. Wynn, and A. Zhigljavsky
GENERALIZED LATENT VARIABLE MODELING: MULTILEVEL, LONGITUDINAL, AND STRUCTURAL EQUATION MODELS	A. Skrondal and S. Rabe-Hesketh
GRAPHICAL ANALYSIS OF MULTI-RESPONSE DATA	K. Basford and J. Tukey
INTRODUCTION TO COMPUTATIONAL BIOLOGY: MAPS, SEQUENCES, AND GENOMES	M. Waterman

Published titles

MARKOV CHAIN MONTE CARLO IN PRACTICE	W. Gilks, S. Richardson, and D. Spiegelhalter
MEASUREMENT ERROR AND MISCLASSIFICATION IN STATISTICS AND EPIDEMIOLOGY: IMPACTS AND BAYESIAN ADJUSTMENTS	P. Gustafson
MEASUREMENT ERROR: MODELS, METHODS, AND APPLICATIONS	J. P. Buonaccorsi
META-ANALYSIS OF BINARY DATA USING PROFILE LIKELIHOOD	D. Böhning, R. Kuhnert, and S. Rattanasiri
STATISTICAL ANALYSIS OF GENE EXPRESSION MICROARRAY DATA	T. Speed
STATISTICAL AND COMPUTATIONAL PHARMACOGENOMICS	R. Wu and M. Lin
STATISTICS IN MUSICOLOGY	J. Beran
STATISTICAL CONCEPTS AND APPLICATIONS IN CLINICAL MEDICINE	J. Aitchison, J.W. Kay, and I.J. Lauder
STATISTICAL AND PROBABILISTIC METHODS IN ACTUARIAL SCIENCE	P.J. Boland
STATISTICAL DETECTION AND SURVEILLANCE OF GEOGRAPHIC CLUSTERS	P. Rogerson and I. Yamada
STATISTICS FOR ENVIRONMENTAL BIOLOGY AND TOXICOLOGY	A. Bailer and W. Piegorsch
STATISTICS FOR FISSION TRACK ANALYSIS	R.F. Galbraith
VISUALIZING DATA PATTERNS WITH MICROMAPS	D.B. Carr and L.W. Pickle

Chapman & Hall/CRC
Interdisciplinary Statistics Series

MEASUREMENT ERROR

MODELS, METHODS, and APPLICATIONS

John P. Buonaccorsi

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20140514

International Standard Book Number-13: 978-1-4200-6658-6 (eBook - PDF)

DEDICATION

In memory of my parents, Eugene and Jeanne, and to my wife, Elaine, and my children, Jessie and Gene.

Contents

Preface

This book is about measurement error, which includes misclassification. This occurs when some variables in a statistical model of interest cannot be observed exactly, usually due to instrument or sampling error. It describes the impacts of these errors on "naive" analyses that ignore them and presents ways to correct for them across a variety of statistical models, ranging from the simple (one-sample problems) to the more complex (some mixed and time series models) with a heavy focus on regression models in between.

The consequences of ignoring measurement error, many of which have been known for some time, can range from the nonexistent to the rather dramatic. Throughout the book attention is given to the effects of measurement error on analyses that ignore it. This is mainly because the majority of researchers do not account for measurement error, even if they are aware of its presence and potential impact. In part, this is due to a historic lack of software, but also because the information or extra data needed to correct for measurement error may not be available. Results on the bias of naive estimators often provide the added bonus of suggesting a correction method.

The more dominant thread of the book is a description of methods to correct for measurement error. Some of these methods have been in use for a while, while others are fairly new. Both misclassification and the so-called "errors-in-variables" problem in regression have a fairly long history, spanning statistical and other (e.g., econometrics and epidemiology) literature. Over the last 20 years comprehensive strategies for treating measurement error in more complex models and accounting for the use of extra data to estimate measurement error parameters have emerged. This book provides an overview of some of the main techniques and illustrates their application across a variety of models. Correction methods based on the use of known measurement error parameters, replication, internal or external validation data, or (in the case of linear models) instrumental variables are described. The emphasis in the book is on the use of some relatively simple methods: moment corrections, regression calibration, SIMEX, and modified estimating equation methods. Likelihood techniques are described in general, but only implemented in some of the examples.

There are a number of excellent books on measurement error, including

Fuller's seminal 1987 book, mainly focused on linear regression models, and the second edition of Carroll et al. (2006). The latter provides a comprehensive treatment of many topics, some of which overlap with the coverage in this book. Other, more specialized books include Cheng and Van Ness (1999) and Gustafson (2004). The latter provides Bayesian approaches for dealing with measurement error and misclassification in numerous settings. This book (i.e., the one in your hands), which is all non-Bayesian, differs from these others both in the choice of topics and in being more applied. The goal is to provide descriptions of the basic models and methods, and associated terminology, and illustrate their use. The hope is that this will be a book that is accessible to a broad audience, including applied statisticians and researchers from other disciplines with prior exposure to statistical techniques, including regression analysis. There are a few places (e.g., the last two chapters and parts of Chapters 5 and 6) where the presentation is a bit more advanced out of necessity.

Except for Chapter 6, the book is structured around the model for the true values. In brief, the strategy is to first discuss some simpler problems including misclassification in estimating a proportion and in two-way tables, as well as additive measurement error in simple and multiple regression. We then move to a broad treatment of measurement error in Chapter 6, and return to specific models for true values in the later chapters. A more detailed overview of the chapters and the organization of the book is given in Section 1.5. The layout is designed to allow a reader to easily locate a specific problem of interest. A limited number of mathematical developments, based on relatively basic theory, are isolated into separate sections for interested readers. However, I have made no attempt to duplicate the excellent theoretical coverage of certain topics in the aforementioned texts.

The measurement error literature has been growing at a tremendous rate recently. I have read a large number of other papers that have influenced my thinking but do not make their way explicitly into this book. I have tried to be careful to provide a direct reference whenever I use a result from another source. Still, given the more applied focus of the book, the desire to limit the coverage in certain areas, and size restrictions, means that many important papers are left unrecognized. I apologize beforehand to anyone who may feel slighted in any way by my omission of their work. In addition, while some very recent work is referenced in certain places, since parts of the book were completed up to two years before publication date, I have not tried to incorporate recent developments associated with certain topics. I also recognize that in the later stages of the book the coverage is based heavily on my own work (with colleagues). This is especially true in the last two chapters of the book. This is largely a matter of writing about what I know best, but it also happens to coincide with coverage of some of the more basic settings within some broader topics. When I set out, I had planned to provide more discussion about the con-

nections between measurement error models and structural equation and latent variable models, as well as related issues with missing data and omitted covariates. There are a few side comments on this, but a fuller discussion was omitted for reasons of both time and space.

In a way this book began in the spring of 1989, when I was on my first sabbatical. In the years right before that I had wandered into the "errors-in-variables", or measurement error, world through an interest in calibration and ratios. Being a bit frustrated by the disconnect between much of the traditional statistics literature and that from other areas, especially econometrics, I began to build a systematic review and a bibliography. Fortuitously, around this time Jim Ware invited me to attend a conference held at NIH in the fall of 1989 on measurement error problems in epidemiology. I also had the opportunity to attend the IMS conference on measurement error at Humboldt State University the following summer and listen to many of the leading experts in the field. This period marked the beginning of an explosion of new work on measurement error that continues to this day. I am extremely grateful to Jim for the invitation to the NIH conference, getting me involved with the measurement error working group at Harvard and supporting some of my early work in this area through an EPA-SIMS (Environmental Protection Agency and the Societal Institute of the Mathematical Sciences) grant. I benefited greatly from interactions with some of the core members of the Harvard working group including Ernst Linder, Bernie Rosner, Donna Spiegelman and Tor Tosteson. This led to an extensive and fruitful collaboration with Tor and Eugene Demidenko (and many pleasant rides up the Connecticut River Valley to Dartmouth to work with them, and sometimes Jeff Buzas from the University of Vermont).

By the early 90s I had a preliminary set of notes and programs, portions of which were used over a number of years as part of a topics in regression course at the University of Massachusetts. The broader set became the basis for a short course taught for the Nordic Region of the Biometrics Society in Ås, Norway in 1998. These were updated and expanded for a short course taught in 2005 at the University of Oslo Medical School, and then a semester long course at the University of Massachusetts in the fall of 2007. I appreciate the feedback from the students in those various courses. The material from those courses made up the core of about half of the book. I am extremely grateful to Petter Laake for helping to arrange support for my many visits to the Medical School at the University of Olso. I have enjoyed my collaborations with him, Magne Thoresen, Marit Veirod and Ingvild Dalen and greatly appreciate the hospitality shown by them and the others in the biostatistics department there.

I'm also grateful to Ray Carroll for sharing some of the Framingham Heart Study data with me, Walt Willet for generously allowing me access to the the external validation data from the Nurses Health Study, Rob Greenberg for use of the beta-carotene data, Sandy Liebhold for the egg mass defoliation data, Joe

Elkinton for the mice density data, Paul Godfrey and the Acid Rain monitoring project for the pH and alkalinity data, and Matthew DiFranco for the urinary neopterin/HIV data.

A number of other people contributed to this book, directly or indirectly. Besides sharing some data, Ray Carroll has always been willing over the years to answer questions and he also looked over a portion of Chapter 10. My measurement error education owes a lot to reading the works of, listening to many talks by, and having conversations with Ray, Len Stefanski, and David Ruppert. After years of a lot of traveling to collaborate with others, I certainly appreciate having John Staudenmayer, and his original thinking, here at the University of Massachusetts. Our joint work forms the basis of much of Chapter 12. I can't thank Meng-Shiou Shieh enough for her reading of the manuscript, double checking many of the calculations, picking up notational inconsistencies and helping me clean up the bibliography. Magne Thoresen generously provided comments on the first draft of Chapter 7 and Donna Spiegelman advised on the use of validation data. I am grateful to Graham Dunn for his reviews of Chapters 2–5 and Gene Buonaccorsi for his assistance with creating the index. Of course, none of those mentioned are in any way responsible for remaining shortcomings or errors. I want to thank Hari Iyer for being my adviser, mentor, collaborator, and friend over the years. On the lighter side, the Coffee Cake Club helped keep me sane (self report) over the years with the Sunday long runs.

Over the years I have been fortunate to receive support for some of my measurement error work from a number of sources. In addition to EPA-SIMS, mentioned earlier, this includes the National Cancer Institute, the U.S. Department of Agriculture and the Division of Mathematical Sciences at the NSF. The University of Massachusetts, in general, and the Department of Mathematics and Statistics in particular, has provided continuing support over the years, including in the computational area. I especially appreciate the assistance of the staff in the departmental RCF (research computing facility). Finally, I am extremely grateful to the Norwegian Research Council for supporting my many visits to collaborate with my colleagues in the Medical School at the University of Oslo.

I appreciate the guidance and patience of Rob Calver of Taylor & Francis throughout this project as well as the production assistance from Karen Simon and Marsha Pronin. I am grateful to John Wiley and Sons, the American Statistical Association, the Biometrics Society, Blackwell Publishing, Oxford University Press, Elsevier Publishing and Taylor & Francis for permissions to use tables and figures from earlier publications. Finally, I want to thank my wife, Elaine Puleo, for her constant support, editing Chapter 1 and serving as a sounding board for my statistical ideas throughout the process of writing of the book.

Computing

Computing was occasionally an obstacle. While there are a number of programs, described below, which handle certain models and types of measurement errors, there were many techniques that I wanted to implement in the examples for which programs were not available. While I had built up some SAS programs to handle measurement error over the years, many of the examples in the book required new, and extensive, programming. There are still examples where I would have liked to have illustrated more methods, but I needed to stop somewhere.

The software used in the book, and some associated programs, include:

- STATA routines rcal (regression calibration) and simex. These, along with the related qvf routine, are available through `http://www.stata.com/merror/`.
- STATA gllamm routines based on Skrondal and Rabe-Hesketh (2004). Available at `http://www.gllamm.org/`. This includes the cme command for parametric maximum likelihood used in a couple of places in the book, as well as some other measurement error related programs.
- SAS macro Blinplus from Spiegelman and colleagues at Harvard University. Available at `http://www.hsph.harvard.edu/faculty/spiegelman/blinplus.html`. This corrects for measurement error using external and internal validation data, based on a linear Berkson model.
- SAS macro from Spiegelman and colleagues at Harvard University which correct for measurement error under additive error with the use of replicate values. This is available at `http://www.hsph.harvard.edu/faculty/spiegelman/relibpls8.html`. This was not used in the book, since the rcal command in STATA was used instead.

The remainder of the computing was carried out using my own SAS programs. Any of these available for public use can be found at `http://www.math.umass.edu/~johnpb/meprog.html`.

Although not spelled out in any real detail in the book (except for a few comments here and there), there are other programs within the standard software packages that will handle certain analyses involving measurement errors. This includes programs that treat structural equation or latent variables models, instrumental variables and mixed models.

John Buonaccorsi
Amherst, Massachusetts
November, 2009

List of Examples

Other applications not in worked examples (in order of appearance)

CHAPTER 1

Introduction

1.1 What is measurement error?

This is a book about measurement error in statistical analyses; what it is, how to model it, what the effects of ignoring it are and how to correct for it. In some sense, all statistical problems involve measurement error. For the purposes here, measurement error occurs whenever we cannot exactly observe one or more of the variables that enter into a model of interest. There are many reasons such errors occur, the most common ones being instrument error and sampling error. In this chapter we provide a collection of examples of potentially mismeasured variables, followed by a brief overview of the general structure and objectives in a measurement error problem, along with some basic terminology that appears throughout the book. Finally, we provide a road map for the rest of the book.

1.2 Some examples

Measurement error occurs in nearly every discipline. A number of examples are given here to illustrate the variety of contexts where measurement error can be a concern. Some of these examples appear later in the book. Any of the variables described below could play the role of either an outcome or a predictor, but there is no need for notational distinctions at this point. Where any notation is used here, the true value is denoted x and the variable observed in place of x, by w. The latter can go by many names, including the observed or measured value, the error-prone measurement, a proxy variable or a surrogate measure. The term surrogate will carry a more precise meaning later. When the true and observed values are both categorical then measurement error is more specifically referred to as **misclassification**. Our first set of examples fall under this heading.

Misclassification examples

- Disease status. In epidemiology, the outcome variable is often presence or absence of a disease, such as AIDS, breast cancer, hepatitis, etc. This is often assessed through an imperfect diagnostic procedure, such as a blood test or an imaging technique, which can lead to either false positives or false negatives.
- Exposure. The other primary variable of interest in epidemiologic studies is "exposure," used in a broad sense. This is typically measured with self report or in some other manner which is subject to error. Potentially misclassified categorical values of this type include antibiotic use during pregnancy or not, whether an individual is a heavy smoker or not, a categorized measure of physical activity, dietary intake, tanning activity level (used in skin cancer studies), drug use, adherence to prescribed medication, etc.
- Acceptance Sampling. In quality control, a product can be classified by whether it meets a certain standard or tolerance. Obtaining an error-free determination can be an expensive proposition and cheaper/faster tests, subject to inspection errors, may be used instead.
- Auditing can involve assessing whether a transaction/claim is fraudulent or contains a mistake of some type. The item of interest may be complex and misclassification can occur from making a decision on partial information (e.g., sampling) or simply from errors of judgment in the evaluation.
- Satellite images are now routinely used for categorizing land usage or habitat type. These categorizations are often prone to some errors, but with some validation data available based on determining the truth on the ground.

Mismeasurement of a quantitative variable

- Measuring water or blood chemistry values typically involves some error. Laboratories routinely calibrate their measuring instruments using standards and values are returned based on the use of the resulting calibration curve. These will still contain some errors unless there is a deterministic relationship between the true and measured values, rarely the case. Examples in Chapters 8 and 10 use some of the data from the acid rain monitoring project (Godfrey et al., 1985). This involved water samples from approximately 1800 water bodies with analyses for ph, alkalinity and other quantities carried out by 73 laboratories. In this case, each lab here was calibrated by being sent blind samples with "known" values.
  Another application of this type is with immunoassays which measure concentrations of antigens or antibodies. The measured value is a radioactive count, or standardized count. Similar to the water chemistry examples, the assay is calibrated using standards and the resulting curve here is usually

nonlinear. One of the examples in Chapter 10 utilizes a four-parameter logistic model for measuring urinary neopterin with a radioimmunoassay.

- The Harvard Six Cities Study (see Section 7.3.1) examined the impact of exposure to certain pollutants on a child's respiratory status. An individual's actual exposure was difficult to measure. In its place observations were made in the home, in different rooms at different times of year. These values served as surrogates for exposure. The measures were validated in studies where individuals wore a lapel monitor that measures exposure continuously.
- Instead of categorized exposure, as discussed in the misclassification examples, quantitative measures of "exposure" are often preferred. The notion of true exposure can be a bit elusive here, especially when it is a dietary intake, but could be defined as a total or average intake over a certain time period. In that case, the measurement error can come from two sources: sampling over time and/or from the use of a fallible instrument, such as a food frequency questionnaire. The same comments apply in measuring physical activity and other quantities.
- In many designed experiments, there is some target value that an experimenter wants to apply to a "unit," but the actual delivered "dose" may differ from the target value. Examples include temperature and pressure in industrial experiments, fertilizer or watering levels in agricultural settings, protein levels in a diet in balance/intake studies to determine nutritional requirements, speed on a tread mill, etc. In this case the "measured" value is the fixed target dose, but the true value of interest is random.
- Many ecological problems strive to model the relationship among variables, over different spatial locations. Costs often make it impossible to measure the variables of interest exactly and instead some spatial sampling is used to obtain estimated values. An example of this appears in Chapters 4 and 5, where gypsy moth egg mass densities and defoliation rates are estimated over stands 60 hectares in size, based on sampling a small fraction of the total area.
- Many financial and economic variables, whether measured at an individual, company or other aggregate level (e.g., state, country, etc.), are subject to measurement error. The error may be due to self-reporting, sampling (similar to the ecological problems in the previous item), the construction of variables from other data, etc.
- The U.S. and many other governments carry out continuous surveys to measure unemployment, wages, and other labor and health variables. These are often then modeled over time. The measurement error here is sampling error arising from estimation using a subsample of individuals, households, etc. An interesting feature of these problems is that the measurement errors

can be correlated as a result of block resampling, in which individuals or households stay in the survey for some amount of time and are then rotated out.

- Another important ecological problem is the modeling of population abundance over time and exploring the relationship of abundance to other factors such as weather or food abundance. It is impossible to ever obtain an exact abundance, or population density, but instead it is estimated. This is well known to be a challenging problem, especially for mobile populations, and can involve the use of capture/recapture methods or other techniques, such as aerial surveillance. Once again the measurement error here is sampling error with the added complication that the accuracy and precision of the estimate may be changing over time as a result of changes in the population being sampled and/or changes in the sampling effort or technique. These same issues arise in modeling temperature and its relationship to human activities. In this latter application the quality of the data (i.e., the nature of the measurement error) has changed considerably over time, with both changes in instruments and enlargement of the grid of monitoring stations.
- As a final illustration, we consider an example discussed by Tosteson et al. (2005). The goal was to use vascular area, obtained from an image, to classify breast disease. One measure of the effectiveness of a classification technique is the receiver operating characteristic (ROC) curve. This depends on the distribution of the variable in each of the disease and nondisease groups or, under normality, the mean and variance. There is measurement error at the individual level however as the image has to be spatially subsampled and the vascular area for the whole image estimated. The induced measurement error will distort the estimated ROC curve.

1.3 The main ingredients

There are typically three main ingredients in a measurement error problem.

1. A Model for the True Values.
   This can be essentially any statistical model. See the overview below for where the emphasis is in this book.
2. A Measurement Error Model.
   This involves specification of the relationship between the true and observed values. This can be done in a couple of different ways as outlined in the next section and described in fuller detail in later chapters.
3. Extra data, information or assumptions that may be needed to correct for measurement error. This may not always be available, in which case one has to be satisfied with an assessment of the impacts of the measurement error. This extra "information" is typically:

(a) Knowledge about some of the measurement error parameters or functions of them.

(b) Replicate values. This is used mainly, but not exclusively, for additive error in quantitative variables.

(c) Estimated standard errors attached to the error prone variables.

(d) Validation data (internal or external) in which both true and mismeasured values are obtained on a set of units.

(e) Instrumental variables.

Some discussion of validation data and replication appear throughout Chapters 2 to 5, and instrumental variables are discussed in Section 5.7 in the context of multiple linear regression. A broader discussion of extra data is given in Section 6.5.

There are two general objectives in a measurement error problem.

- What are the consequences for **naive analyses** which ignore the measurement error?
- How, if at all, can we correct for measurement error?
  With some exceptions (see parts of Chapters 11 and 12), correcting for measurement error requires information or data as laid out in item 3 above. Myriad approaches to carrying out corrections for measurement error have emerged, a number of which will be described in this book. These include direct bias corrections, moment based approaches, likelihood based techniques, "regression calibration," SIMEX and techniques based on modifying estimating equations.

1.4 Some terminology

This section provides a brief introduction to measurement error models and some associated terminology. These models are revisited and expanded on in later chapters. Chapters 2 and 3 spell out misclassification models while additive error models arise in treating linear regression problems in Chapters 4 and 5. Section 6.4 then provides a more comprehensive and in depth look at these and other models. Still, a preliminary overview is helpful at this point. For convenience the discussion here is in terms of a univariate true value and its mismeasured version, but this can, and will be, extended to handle multivariate measures later.

In some places it is necessary to distinguish a random variable from its realized value. When necessary, we follow the usual convention of a capital letter (e.g., X) for the random variable and a small letter (e.g., x) for the realized

value. Throughout the notation $|x$ is a shorthand for the more precise $|X = x$ which is read "given X is equal to x."

When the x is a predictor in a regression type problem a distinction is made between the **functional case** where the x's are treated as fixed values and the **structural case** where X's are random. This distinction, and its implications, is spelled out carefully in Chapters 4 and 5, with that discussion carrying over into later chapters. Some authors use structural to refer just to the case where a distribution is specified for a random X. The term functional model has also sometimes been used to refer to the case where there is a deterministic relationship among true variables. Our usage of the terminology will be as defined above. We also note that a combination of fixed and random predictors is possible. Chapter 5 illustrates this and adopts a strategy that handles these mixed cases.

1.4.1 Measurement versus Berkson error models

A fundamental issue in specifying a measurement error model is whether we make an assumption about the distribution of the observed values given the true values or vice versa. The **classical measurement error** model specifies the former, while the latter will be referred to as **Berkson error** model. The latter was initially introduced by Berkson (1950) in cases where an experimenter is trying to achieve a target value w but the true value achieved is X. Examples of this were presented earlier. In this case there is no random variable W but just a targeted fixed value w. In other problems X and W can both be random and the phrase "Berkson error model" has been broadened to include referring to the conditional distribution of $X|w$.

If the x's are fixed or we condition on them, the model for $W|x$ should be used. If the observed w is fixed then the Berkson model applies. With random X and random W given x, then a choice can sometimes be made between which model to use. This choice will depend in part on the nature of the main study (are the X's a random sample?) and the nature of any validation data. See Heid et al. (2004) for an example discussing the use of classical and Berkson error models in the measurement of radon exposure.

A special case is when both X and W are categorical. For classical measurement error, the model is given by the probability function $P(W = w|x)$ and these quantities are referred to as **misclassification probabilities** or misclassification rates. The Berkson error model, on the other hand, specifies $P(X = x|w)$, which we refer to as **reclassification probabilities** or reclassification rates. Both of these are discussed in Chapters 2 and 3.

1.4.2 Measurement error models for quantitative values

For a quantitative variable, any type of regression model can be used for how W and X are related. Historically, the most commonly used model is the **additive measurement error** model where

$$W|x = x + u \tag{1.1}$$

where u is a random variable with $E(u|x) = 0$. Equivalently $E(W|x) = x$, so W is unbiased for the unobserved x.

Nonadditive measurement error includes everything else, where $E(W|x) = g(\boldsymbol{\theta}, x) \neq x$. This means there is some form of bias present. Examples include constant bias, $E(W|x) = \theta + x$ and linear measurement error, $E(W|x) = \theta_0 + \theta_1 x$, while with nonlinear measurement error models $g(x, \boldsymbol{\theta})$ is nonlinear in the θ's. Nonadditive models are particularly prevalent when the measurement error is instrument error.

A **constant variance/homoscedastic** measurement error model refers to the case where the variance of W given x is constant. A **heteroscedastic** model allows the measurement error variance to possibly change. While historically, constant measurement error variance has been used, there are many situations where changing variances need to be allowed. This is discussed in some detail in Section 6.4.5. Heteroscedastic measurement error models are accommodated throughout many parts of this book.

1.4.3 Nondifferential/differential measurement error, conditional independence and surrogacy

To this point the discussion has been limited to how a single measured W relates to a true value x. More generally, we need to accommodate the fact that the measurement error may depend on other variables, which themselves may or may not be measured with error. General developments of this type are accommodated in Chapter 6. Here we introduce a few important concepts where W (measured) and X (true) are both univariate as is a third variable Y, which is, usually, a response variable. That designation is not essential to the discussion below, nor is the fact that W and X are univariate.

- The measurement error model for W given x is said to be **nondifferential** (with respect to Y) if the distribution of $W|x, y$ (W given $X = x$ and $Y = y$) equals the distribution of $W|x$. That is, the measurement error model does not depend on the value y.
  The measurement error model is said to be **differential** if the distribution of $W|x, y$ changes with y.

- W is a **surrogate** for X (with respect to Y) if the distribution of $Y|x, w$ equals that of $Y|x$. This says that given x, W contains no information about the distribution of Y beyond that which is contained in x.
- **Conditional independence** states that Y and W are independent given $X = x$. Mathematically, $f(y, w|x) = f(y|x)f(w|x)$, where we have used the convenient, but obviously imprecise, device of denoting the distribution of interest by the symbols in the argument. So $f(y|x)$ is the conditional density or mass function of Y given $X = x$, $f(w|x)$ is that of W given $X = x$, and $f(y, w|x)$ is the joint density or mass function of Y and W given $X = x$.

An interesting, and very useful, result is the following:

The three concepts of surrogacy, conditional independence and nondifferential measurement error are equivalent. That is, any one implies the other two.

1.5 A look ahead

The guiding principles in laying out the book were 1) have it be introductory and cover some of the more basic statistical models in detail and 2) to allow the reader who is interested in a particular type of problem to jump into the book at various places and be able to at least get a feel for how to handle that particular problem. The second emphasis means the book is based primarily around the model for the true values, rather than around the measurement error correction technique. The exception to this is the somewhat monolithic Chapter 6. This chapter provides a fairly wide ranging overview and details on the various components of the problem: models for true values, various measurement error models, types of additional data and correction techniques. Rather than jump into this first, which would make for fairly boring reading without a lot of excursions into particular problems, some simple problems are discussed first.

Misclassification of categorical data is treated first in the contexts of estimating a single proportion and in two-way tables, Chapters 2 and 3, respectively. This provides a good look in particular at the various uses of validation data. We then consider additive error, in the predictor(s) and/or the response, for simple and multiple linear regression. This is in Chapters 4 and 5. Once again, we follow the principle of looking at a simple setting first (simple linear models) in which to illustrate a number of basic ideas in explicit detail before moving on to the more complex setting of multiple predictors. To a large extent, these four chapters go over somewhat old ground, but hopefully in a way that provides a

unified summary in an applied manner. The multiple linear regression chapter has a few useful features, including allowing error in either predictors or response with possibly changing measurement error variances or (if applicable) covariances, from the beginning, and a formulation within which a number of results can be expressed which cover random or fixed predictors, or a combination thereof.

After this is the aforementioned Chapter 6. Some readers may chose a selective read through this depending on what they are after, and return to it later as needed. It does contain all of the key concepts and methods used elsewhere in the book. At the least, it is suggested that the reader look over the descriptions of measurement error models, the use of extra data, the descriptions of regression calibration, SIMEX and modified estimating equation correction methods (which are used in most examples), as well as the overviews on using validation data and bootstrapping.

The subsequent chapters cover specific models. Chapter 7 illustrates most of the key elements of Chapter 6 in the context of binary regression. This provides an important setting in which to examine many of the correction techniques for nonlinear models in general, and generalized linear models, in particular. Chapter 8 then returns to linear models but now with nonadditive error in certain predictors or functions of them. Among other things, this chapter includes how to handle quadratic and interaction terms as well as the use of external validation data for measurement error models that have systematic biases. Chapter 9, which is fairly short, reinforces a few ideas from the binary chapter and expands a bit on handling nonlinear models that are not generalized linear models. Chapter 10 isolates some additional questions when dealing with error in the response only. Among other things, this chapter includes some discussion of nonparametric estimation of a distribution in the presence of measurement error and fitting mean-variance relationships. Finally, the last two chapters cover mixed/longitudinal and time-series models. These are a bit different than the earlier chapters in a couple of ways. They both provide fairly general surveys and then focus on a few specific problems within the much larger class of problems. They also assume more prior knowledge of the topics than some of the earlier chapters.

See the preface for additional comments about coverage, emphasis and computational aspects.

CHAPTER 2

Misclassification in Estimating a Proportion

This chapter treats the problem of estimating a proportion when the true value is subject to misclassification. After some motivating examples, Sections 2.2 - 2.4 examine the problem under the assumption that the true values are a random sample in the sense of having independent observations. We first examine bias in the naive estimator which ignore the misclassification and then show how to correct for the misclassification using both internal and external validation data or with known misclassification probabilities. As part of this discussion, Section 2.3 delineates the difference between misclassification and reclassification probabilities. Section 2.5 extends the results to allow for sampling from a finite population, while Section 2.6 provides some brief comments on how to account for misclassification with repeated or multiple measures without validation data. Finally, Section 2.7 provides an overview of the case where there are more than two categories for the true values and Section 2.8 provides a few mathematical developments.

2.1 Motivating examples

Example 1. Knowing the prevalence of the HIV virus in a population is critical from a public health perspective. This is of special interest for the population of women of child bearing age. For example, Mali et al. (1995) estimated the prevalence of women with HIV-1 attending a family planning clinic in Nairobi to be 4.9% from a sample of 4404 women. Assays that test for the presence or absence of the virus are often subject to misclassification. What effect do the misclassification errors have on the estimated prevalence of 4.9%? If we had estimates of the misclassification rates how can we correct for the misclassification? As an example of validation data, Table 2.1 shows some data, based on Weiss et al. (1985) and also used by Gastwirth (1987), for an HIV test on 385 individuals. The error-prone w is based on ELISA (enzyme-linked immunoabsorbent assay). There are two false negatives. If what Weiss et al.

(1985) considered borderline are classified as negative then there are only 4 false positives, while if they are considered positive there are 22 false positives. The latter case is given by the parenthetical values in the first line of the table. Later, for illustration, we use this data as external validation data in analyzing the data from Mali et al. (1995).

Table 2.1 *External validation data for the ELISA test for presence of HIV. x is the true value, with indicating being HIV positive, and w is the error prone measure from the assay.*

		ELISA(w) 0	1	
Truth(x)	0	293 (275)	4 (22)	297
	1	2	86	88

Example 2. Here we use an example from Fleiss (1981, p. 202) where the goal was to estimate the proportion of heavy smokers in a population. The "truth" is determined by a blood test while the error-prone measure w is based on a self assessment by the individual. Some level of misclassification is often present with self reporting. Table 2.2 shows data for a random sample of 200 individuals of which 88 reported themselves to be heavy smokers. A random subsample of 50 from the 200 were subjected to a blood test which determined the true x. The values of w and x for these 50 subjects are given in the top portion of the table. The other 150 have a self reported value (w), but the truth is unknown. This differs from the first example in that the individuals for which we obtain true values are a subset of the main study. This is an example of internal validation.

Table 2.2 *Smoking example: x = true status with w = self assessment. From Fleiss (1981). Used with permission of John Wiley and Sons.*

		w 0	1	
x =	0	24	2	26
	1	6	18	24
		30	20	50
	?	82	68	150
		112	88	200

Example 3. Lara et al. (2004) use what is known as a randomized response technique to estimate the proportion of induced abortions in Mexico. In their implementation, with probability 1/2 the woman answered the question "Did you ever interrupt a pregnancy?" and with probability 1/2 answered the question "Were you born in April?" Only the respondent knew which question they answered. Of 370 women interviewed, 56 answered yes. The objective is to estimate the proportion of women who interrupted a pregnancy. The randomized response technique (RRT) intentionally introduces misclassification but with known misclassification probabilities. In this example, the sensitivity and specificity (defined in the next section) are 13/24 and 23/24, respectively. See Lara et al. (2004) and van den Hout and van der Heijden (2002) for further discussion of the RRT method. The latter paper also describes the situation where misclassification is intentionally introduced to protect privacy, again with known probabilities.

Example 4. The use of satellite imagery to classify habitat types has become widespread. Table 2.3 shows validation data taken from an online document at the Canada Centre for Remote Sensing for five classes of land-use/land-cover maps produced from a Landsat Thematic Mapper image of a rural area in southern Canada. In this case, each unit is a pixel. The goal would be to use this validation data to obtain corrected estimates for the proportions for an area on which only the Landsat classifications are available. This is an example with multiple categories, which is touched on in Section 2.7.

Table 2.3 *Validation data for remote sensing.*

	LANDSAT				
	Water	BareGround	Deciduous	Coniferous	Urban
TRUTH					
Water	367	2	4	3	6
BareGround	2	418	8	9	17
Deciduous	3	14	329	24	25
Coniferous	12	5	26	294	23
Urban	16	26	29	43	422

There are two questions considered in the following sections.

- What if we ignore the misclassification?
- How do we correct for the bias induced by misclassification and obtain corrected standard error and/or confidence intervals for the proportion of interest?

2.2 A model for the true values

We begin by assuming that we have a "random sample" of size n where each observation falls into one of two categories. In terms of random variables, $X_1, \ldots, X_n$ are assumed to be *independent and identically distributed* (i.i.d.), where X_i = 1 if the ith observation is a "success" and = 0 if a "failure," and

$$\pi = P(X_i = 1).$$

The objective is estimation of π, which can be the proportion of successes in a finite population or, as above, the probability of success on repeated trials. The assumption of independence is only approximately true when sampling from a finite population, but generally acceptable unless the sample size is "large" relative to the population size. Explicit modifications for sampling from a finite population appear in Section 2.5.

With no measurement error, $T = \sum_{i=1}^{n} X_i$ = number of successes in the sample is distributed Binomial(n, π) and

$$p = T/n,$$

the proportion of successes in the sample, is an unbiased estimator of π, with $E(p) = \pi$ and $V(p) = \pi(1 - \pi)/n$. The estimated standard error of p is $SE(p) = [p(1-p)/n]^{1/2}$. An approximate large sample confidence interval for π is given by $p \pm z_{\alpha/2} SE(p)$ and approximate large sample tests of hypotheses about π, such as $H_0 : \pi = \pi_0$, are based on either $Z = (p - \pi_0)/SE(p)$ or $(p-\pi_0)/(\pi_0(1-\pi_0)/n)^{1/2}$. These large sample inferences for a proportion can be found in most introductory statistics books. For small to moderate samples, "exact" confidence intervals and test of hypotheses are available based on the Binomial distribution; see for example Agresti and Coull (1998) or Casella and Berger (2002, Exercise 9.21).

2.3 Misclassification models and naive analyses

The fallible/error-prone measure is W, which is assumed to also be binary. It is the outcome of W, denoted by w, that is obtained on the n observations making up the main study. The measurement error model, which specifies the behavior of W given $X = x$ ($x = 0$ or 1), is specified by the misclassification probabilities,

$$\theta_{w|x} = P(W = w|X = x),$$

with

sensitivity: $P(W = 1|X = 1) = \theta_{1|1}$

specificity: $P(W = 0|X = 0) = \theta_{0|0}$.

We only need to specify two probabilities since $P(W = 0|X = 1) = \theta_{0|1} = 1 - \theta_{1|1}$ (the probability of a false negative) and $P(W = 1|X = 0) = \theta_{1|0} = 1 - \theta_{0|0}$ (the probability of a false positive).

Since X is random, if we specify the distribution of $W|x$ we can also derive a model for the distribution of X given w. This is given by

$$\lambda_{x|w} = P(X = x|W = w).$$

We refer to these as **reclassification probabilities**, but they are also known as **predictive probabilities**. Again, we need just two probabilities

$$\lambda_{1|1} = P(X = 1|W = 1) \text{ and } \lambda_{0|0} = P(X = 0|W = 0).$$

The reclassification model is a special case of a Berkson model (see Chapter 1) in that it models the truth given the observed. The reclassification and misclassification rates are related via $\lambda_{x|w} = \theta_{w|x}P(X = x)/(\theta_{w0}(1 - \pi) + \theta_{w1}\pi)$.

Naive inferences are based on use of the W's rather than the possibly unobservable X's. The naive estimator of π is

$$p_W = \text{the proportion of the sample with } W = 1.$$

The W_i's are a random sample with (see Section 2.8)

$$\pi_W = P(W_i = 1) = \pi(\theta_{1|1} + \theta_{0|0} - 1) + 1 - \theta_{0|0}. \tag{2.1}$$

We can also reverse the roles of X and W, leading to

$$\pi = \pi_W(\lambda_{1|1} + \lambda_{0|0} - 1) + 1 - \lambda_{0|0}. \tag{2.2}$$

Since $E(p_W) = \pi_W$ rather than π, the naive estimator has a bias of

$$BIAS(p_w) = \pi_W - \pi = \pi(\theta_{1|1} + \theta_{0|0} - 2) + (1 - \theta_{0|0}).$$

For a simple problem the nature of the bias can be surprisingly complex. Figure 2.3 illustrates the bias for π = .01 and .5, displaying the bias as a function of sensitivity for different levels of specificity. Notice that the absolute bias can actually increase as the sensitivity or specificity increases with the other held fixed. For example, consider the case with $\pi = .5$ and specificity .90. At sensitivity .9, the bias is 0 with the bias then increasing as sensitivity increases with a bias of .05 at sensitivity = 1. With a rarer event ($\pi = .01$), the bias is insensitive to the level of sensitivity, but heavily sensitive to specificity with severe relative bias even at specificity of .95 where the bias is approximately .05, when estimating a true value of .01.

We could also evaluate the naive performance through the behavior of the naive confidence interval. Using the standard normal based confidence interval

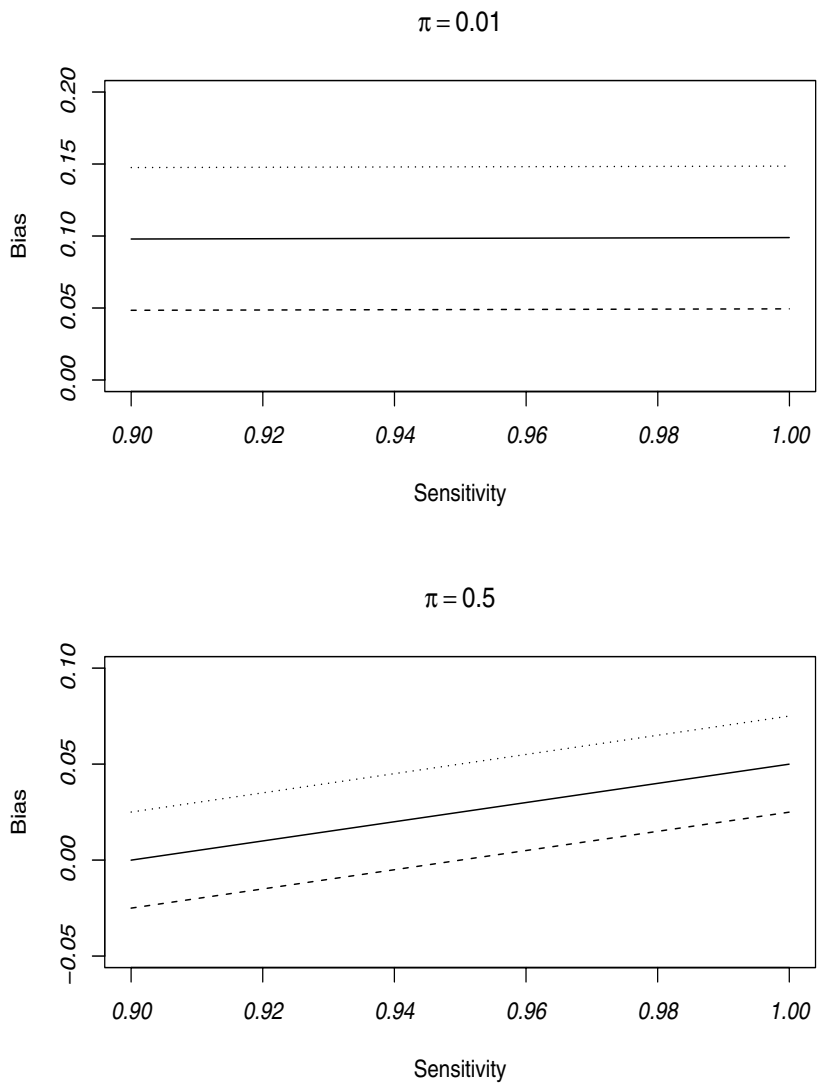


Figure 2.1 *Plot of bias in naive proportion plotted versus sensitivity for different values of specificity (dotted line = .85; solid line = .90; dashed line = .95).*

the probability that the naive interval contains an arbitrary value c is approximately

$$P\left[\frac{c-\pi_w}{(\pi_W(1-\pi_W)/n)^{1/2}} - z_{\alpha/2} \le Z \le \frac{c-\pi_W}{(\pi_W(1-\pi_W)/n)^{1/2}} + z_{\alpha/2}\right],$$

where Z is distributed as a standard normal. With $c = \pi$ this gives the coverage rate of the naive interval.

2.4 Correcting for misclassification

Equation (2.1) can be used to correct for misclassification using known or estimated misclassification rates. For convenience these are denoted $\widehat{\theta}_{0|0}$ and $\widehat{\theta}_{1|1}$, even if they are known. Inverting (2.1) leads to the corrected estimator

$$\hat{\pi} = \frac{p_W - (1-\widehat{\theta}_{0|0})}{\widehat{\theta}_{0|0} + \widehat{\theta}_{1|1} - 1}. \tag{2.3}$$

It is possible for this estimate to be less than 0 or greater than 1. If it is, then from the point estimation perspective, the estimate would be set to 0 or 1, respectively. Notice that the estimated specificity is particularly important in this regard since if the estimated probability of a false positive, $1 - \widehat{\theta}_{0|0}$, is bigger than p_w then the corrected estimate is negative (assuming $\widehat{\theta}_{0|0} + \widehat{\theta}_{1|1} - 1$ is positive, which is essentially always the case). Certain of the intervals presented later bypass this problem since they don't use a point estimate.

If the sensitivity and specificity are known, then $\widehat{\pi}$ is unbiased for π. Otherwise the corrected estimator is consistent but biased with decreasing bias as the size of the validation sample increases. In practice the bias of the corrected estimator could be an issue with small validation studies. This can be investigated through the bootstrap.

2.4.1 Ignoring uncertainty in the misclassification rates

As noted in the introduction, there are cases where the misclassification rates are known exactly. This happens with randomized response techniques and with intentional misclassification to protect privacy. In other cases, previously reported sensitivity and specificity may be treated as known. In this case, the estimated standard error of $\hat{\pi}$ is

$$SE(\hat{\pi}) = \left[\frac{p_W(1-p_W)}{n(\theta_{1|1} + \theta_{0|0} - 1)^2}\right]^{1/2}$$

and an approximate large sample confidence interval for π is

$$\hat{\pi} \pm z_{\alpha/2} SE(\hat{\pi}). \tag{2.4}$$

As an alternative to the Wald interval in (2.4), an exact confidence interval can be obtained. First, an exact confidence interval (L_W, U_W) can be obtained for π_W using the error-prone values. This would be based on the Binomial distribution. Then, assuming $\widehat{\theta}_{1|1} + \widehat{\theta}_{0|0} - 1$ is positive (it usually is) and using the fact that π is a monotonic function of π_W, a confidence interval for π is given by

$$[L, U] = \left[\frac{(L_W - (1 - \widehat{\theta}_{0|0}))}{(\widehat{\theta}_{1|1} + \widehat{\theta}_{0|0} - 1)}, \frac{(U_W - (1 - \widehat{\theta}_{0|0}))}{(\widehat{\theta}_{1|1} + \widehat{\theta}_{0|0} - 1)} \right]. \tag{2.5}$$

If $\widehat{\theta}_{1|1} + \widehat{\theta}_{0|0} - 1$ is not positive then some modification is needed, but this would almost never occur. The value of this approach is that it allows one to handle small sample problems exactly, while at the same time bypassing any potential problems with point estimates of π that are outside of the interval $[0, 1]$. The same approach could be carried out by constructing the interval (L_W, U_W) with the method proposed by Agresti and Coull (1998) as an alternative to the binomial based interval.

Hypothesis tests about π can be carried out using confidence intervals for π. Alternatively, an hypothesis about π can be expressed in terms of π_W and tested directly based on the W data.

Example 3 (continued). Table 2.4 provides the analysis for the third example in the introduction, estimating the proportion of interrupted unwanted pregnancies, using a randomized response technique, which has known sensitivity $\theta_{1|1} = 13/24$ and specificity $\theta_{0|0} = 23/24$. Lara et al. (2004) were interested in comparing the RRT method to other techniques, rather than in providing inferences for the population proportion, as we do here. The naive p_W, which is the proportion that answered yes to whichever question they received, is approximately .15. The corrected estimate is $\widehat{\pi} = (.22 - (1/24))/((36/24) - 1) = .22$ with standard errors and confidence intervals calculated as described above. The correction is about 50 percent of the original estimate and the corrected confidence intervals are wider and shifted towards larger values, with fairly close agreement between the Wald and exact intervals.

Table 2.4 *Analysis of abortion example*

Method	Estimate	SE	Wald Interval	Exact Interval
Naive	.1514	.0186	(.1148, .1878)	(.1161, .1920)
Corrected	.2194	.0373	(.1463, .2924)	(.1488, .3007)

2.4.2 Using external validation data and misclassification rates

Here we account for uncertainty from estimating the misclassification rates assuming those estimates come from external validation data, represented generically in Table 2.5.

Table 2.5 *Representation of external validation data*

		W = 0	W = 1	
X	0	n_{V00}	n_{V01}	$n_{V0.}$
	1	n_{V10}	n_{V11}	$n_{V1.}$
		$n_{V.0}$	$n_{V.1}$	n_V

There are n_V observations for which W and X are both observed. The . notation indicates summation over the missing index; e.g., $n_{V0.} = n_{V00} + n_{V01}$. These data are assumed to be independent of the main body of data. The W's entering into the external validation study also must be random (so we can estimate the probabilities for W given X) and therefore the validation study can't be obtained using stratification based on W. To illustrate this point, suppose we are validating reported smoking status versus true smoking status. If this external validation design used stratified sampling from reported smokers and non-smokers then this data only allows estimation of reclassification rates and so could not be used in the manner described in this section. The other important assumption is that the misclassification model is assumed to be the same for this data as the main data; that is, the measurement error model is exportable from the validation data to the main study. It is less common to use the reclassification rates from external validation data. If, however, this is done then the estimation would proceed as when using internal validation; see (2.7) and the discussion following it.

The validation data yields estimated specificity and sensitivity

$$\widehat{\theta}_{0|0} = n_{V00}/n_{V0.} \qquad \text{and} \qquad \widehat{\theta}_{1|1} = n_{V11}/n_{V1.}$$

leading to $\widehat{\pi}$ as in (2.3). If we condition on the x values in the external validation data, this is in fact the maximum likelihood estimator (MLE) of π, as long as $0 \leq \hat{\pi} \leq 1$.

Delta and Fieller intervals.

The estimate $\widehat{\pi}$ is a ratio Z_1/Z_2, where $Z_1 = p_W - (1 - \widehat{\theta}_{0|0})$ and $Z_2 = \widehat{\theta}_{0|0} + \widehat{\theta}_{1|1} - 1$. The estimated variances of Z_1 and Z_2 and their covariance are

given by

$$\hat{V}_1 = \frac{p_W(1-p_W)}{n} + \frac{\widehat{\theta}_{0|0}(1-\widehat{\theta}_{0|0})}{n_{V0.}}, \qquad \hat{V}_2 = \frac{\widehat{\theta}_{0|0}(1-\widehat{\theta}_{0|0})}{n_{V0.}} + \frac{\widehat{\theta}_{1|1}(1-\widehat{\theta}_{1|1})}{n_{V1.}}$$

and $\hat{V}_{12} = \widehat{\theta}_{0|0}(1-\widehat{\theta}_{0|0})/n_{V0.}$, respectively.

Using the approximate variance of a ratio and estimating it (see Section 2.8) leads to an estimated standard error $SE(\widehat{\pi}) = \hat{V}(\hat{\pi})^{1/2}$, where

$$\hat{V}(\hat{\pi}) = \frac{1}{(\widehat{\theta}_{0|0} + \widehat{\theta}_{1|1} - 1)^2}\left[\hat{V}_1 - 2\widehat{\pi}\hat{V}_{12} + \hat{\pi}^2\hat{V}_2\right] \tag{2.6}$$

$$= \frac{p_W(1-p_W)}{n(\widehat{\theta}_{0|0} + \widehat{\theta}_{1|1} - 1)^2} + \frac{1}{(\widehat{\theta}_{0|0} + \widehat{\theta}_{1|1} - 1)^2}\left[\frac{\widehat{\theta}_{0|0}(1-\widehat{\theta}_{0|0})}{n_{V0.}} - 2\widehat{\pi}\hat{V}_{12} + \hat{\pi}^2\hat{V}_2\right].$$

The expression for the estimated variance following (2.6) is written as the variance if the misclassification parameters are assumed known plus a piece which accounts for estimation from the validation data. Notice that as the validation sample sizes ($n_{V0.}$ and $n_{V1.}$) increase the second term goes to 0.

An approximate Wald confidence interval for π, based on the delta method approximation to the variance of $\widehat{\pi}$, is given by $\hat{\pi} \pm z_{\alpha/2} SE_{EV}(\hat{\pi})$.

A better option when working with ratios is to use Fieller's method, which is discussed in Section 13.5. The use of this method in the current context was recently investigated by Shieh (2009) and Shieh and Staudenmayer (2009). With the definitions above the Fieller interval for π can be calculated using (13.1) in Section 13.5. This interval holds as long as Z-test of $\theta_{0|0}+\theta_{1|1}-1=0$ is signficant at size α. This is almost always the case unless the error-prone measure is particularly bad. When this test is highly significant, the Fieller and Wald intervals are very similar.

Bootstrapping.

Another approach to inference is to use the bootstrap, discussed in a bit more detail in Section 6.16. Here, this is implemented as follows:

For the bth bootstrap sample, $b = 1, \ldots, B$, where B is large:

1. Generate $p_{wb} = T_b/n$, $\widehat{\theta}_{0|0b} = n_{V00b}/n_{V0.}$ and $\widehat{\theta}_{1|1b} = n_{V11b}/n_{V1.}$, where T_b, n_{V00b} and n_{V11b} are generated independently as Binomial(n, p_w), Binomial$(n_{V0.}, \widehat{\theta}_{0|0})$ and Binomial $(n_{V1.}, \widehat{\theta}_{1|1})$, respectively.
2. Use the generated quantities to obtain $\widehat{\pi}_b = (p_{wb} - (1 - \widehat{\theta}_{0|0b}))/(\widehat{\theta}_{0|0b} + \widehat{\theta}_{1|1b} - 1)$ (truncated to 0 or 1 if necessary).