CHAPTER 1

Introduction

The quintessential statistical task is to understand or 'learn' the relationship between an outcome variable Y and an explanatory variable X, having observed both variables simultaneously for a number of study units. Often in biostatistical and epidemiological investigations the study units are people, the outcome variable is health-related, and the explanatory variable reflects an *exposure* of some kind. This book is concerned with consequences and remedies if in fact the explanatory variable cannot be measured very precisely. Specifically, sometimes a rough or *surrogate* variable X^* is recorded, rather than X itself. Whereas the scientific goal is to understand the relationship between X and Y, the available data consist of measurements of X^* and Y (and usually other variables as well). This book focusses on how misleading it is to ignore this 'disconnect' by analyzing the data as if they were precisely measured, and on how to generate statistical inferences that reflect or adjust for the mismeasurement at play. We start by giving some examples where it is necessary to rely on surrogate explanatory variables. While such contexts abound in many subject areas where statistical methods are used, we focus on scenarios from medical research.

1.1 Examples of Mismeasurement

Many variables of interest to medical researchers as possible culprits in the development of disease are difficult to measure on individuals. Intakes of various foods and drugs are prime examples, as are exposures to ambient entities such as airborne pollutants, radiation, and magnetic fields. To give some specific examples we mention three scenarios taken from the recent epidemiological literature.

Brown, Kreiger, Darlington and Sloan (2001) investigate the common scenario where X is average daily caffeine intake, but X^* is based on the self-reported average cups of coffee consumed. In particular, X^* is the average number of cups per day reported by a subject on a questionnaire, multiplied by the amount of caffeine in a 'standard' cup of coffee, i.e., a cup of typical size and strength. On biological grounds the relationship between an outcome Y and X is of much more interest than the relationship between Y and X^*. Brown *et al.* conduct a study to assess how closely X^* and X tend to agree, and give some implications for statistical analysis in hypothetical scenarios with a strong association between X and Y. In particular, study participants are questioned about typical intake of a variety of caffeine-containing products,

leading to a demonstration that X^* based on coffee alone often vastly underestimates X. The authors then recommend that in practice questionnaires include queries about tea and cola consumption, as well as coffee consumption. It should be noted, however, that this is far from the end of the story. In this scenario mismeasurement also arises because participants have imperfect recall when completing questionnaires, and because some subjects will have a systematic tendency to consume cups of coffee with caffeine content higher or lower than that of a standard cup. Thus even very detailed questionnaires are likely to yield substantial mismeasurement of caffeine intake.

More generally, nutritional intakes are exposure variables of considerable interest in medical studies, while also being difficult to measure. In studies with large numbers of subjects, resource limitations might only allow for use of a *food-frequency questionnaire* (FFQ), on which subjects self-report the frequency with which they consume specific foods. Sometimes, however, it is possible to make more precise measurements on a small subset of study subjects. For instance, Rosner and Gore (2001) consider the Nurses' Health Study, where a FFQ was administered to about 89,000 subjects, with additional detailed measurements obtained for about 170 of these subjects. The subset of subjects for whom both measurements are made is referred to as the *validation sample*, while those for whom only the rougher measurement is made constitute the main sample. In the present scenario the more detailed measurements are based on a weighed diet record, whereby subjects weigh and record what they eat in real-time over a short period. We can regard $X = (X_1, \ldots, X_J)$ as the intake of a J different foods for a given subject according to the weighed diet record, while $X^* = (X_1^*, \ldots, X_J^*)$ are the analogous intakes as determined from the FFQ. Rosner and Gore demonstrate that the precision with which X_j can be predicted from X^* varies considerably across j, i.e., the FFQ works better for some foods than others. A very important point is that even a small validation sample gives some basis for guessing the unobserved values of X from the observed values of X^* in main sample subjects. Thus one sees opportunity to do better than an analysis which simply proceeds by treating X^* as if it were X.

Another general class of variables that are subject to considerable mismeasurement are biochemical variables based on laboratory analysis of human-subject samples. As one example, Bashir, Duffy and Qizilbash (1997) describe a study where the outcome is incidence of minor ischaemic stroke, while the explanatory variables of interest are levels of four haemostatic factors—agents which retard the flow of blood in blood vessels. The study follows a case–control design, comparing a group of subjects who had a minor stroke to a group that did not. Initial blood samples were assayed twice, giving two measurements of the factor levels. Also, one-year follow-up blood samples were obtained for some of the control subjects, and these were also assayed twice. As is typical for biochemical variables, mismeasurement is at play here in two ways. First, there is pure laboratory error which leads to two different numerical measurements for the same blood sample. Second, levels of the haemostatic factors vary somewhat from day-to-day within a given subject. Operationally

it makes sense to define the explanatory variable as the subject's average level, but measurement error arises because of the day-to-day fluctuations. Even in the absence of pure laboratory error then, two blood samples taken on different days are not likely to give identical measurements. Bashir *et al.* discuss the implications of both sources of mismeasurement for inference about the relationship between the haemostatic factors and incidence of stroke.

Examples such as these speak to the difficulty in measuring many explanatory variables of interest to medical researchers. Indeed, exposure assessment itself is a very important area of Epidemiology (see, for instance, the book by Armstrong, White and Saracci 1992). In many circumstances, however, substantial mismeasurement will remain no matter how much care and expense is devoted to measuring the variable in question. This behooves us to (i) consider what effect mismeasurement might have on a statistical analysis which simply pretends the surrogate variable X^* is the real explanatory variable X, and (ii) contemplate more refined statistical analysis which acknowledges the difference between the observable X^* and the unobservable X, while retaining an inferential focus on the relationship between X and Y.

1.2 The Mismeasurement Phenomenon

Henceforth an analysis which pretends the surrogate variable is the same as the variable of interest is referred to as *naive*, since it pretends mismeasurement is not manifested. One can glean a sense of how well naive analysis performs with some simple computer-simulated examples. Before doing so, however, we acknowledge an asymmetry in our development. We focus on situations where the explanatory variable X is subject to mismeasurement, and ignore equally realistic scenarios where the outcome variable Y is subject to mismeasurement. The following simulations speak to this seemingly arbitrary choice of focus.

Consider the scatterplot of simulated data displayed in the upper-left panel of Figure 1.1. These data consist of 500 (X, Y) pairs. For completeness we note that the data were simulated from a bivariate normal distribution with both means equal to zero, both variances equal to one, and a correlation coefficient of 0.5. Fitting a standard linear regression to these data gives $0.012 + 0.514X$ as the estimated conditional expectation of Y given X. This fitted regression line is superimposed on the plotted points. In the usual manner we interpret the estimated slope as describing a 0.514 unit change in the average value of Y per one unit change in X. The standard error associated with the slope estimate is 0.037.

Now say that in fact the outcome variable is subject to measurement error, so that Y^* is observed in lieu of Y, where Y^* is obtained by adding 'noise' to Y, without any regard to the X variable. In the present example the noise is taken to have a normal distribution with mean zero and variance one. A scatterplot of the (X, Y^*) pairs appears in the lower-left panel of Figure 1.1, with the fitted linear regression line superimposed. The slope of this line is 0.475, with a standard error of 0.060. It comes as no surprise that the

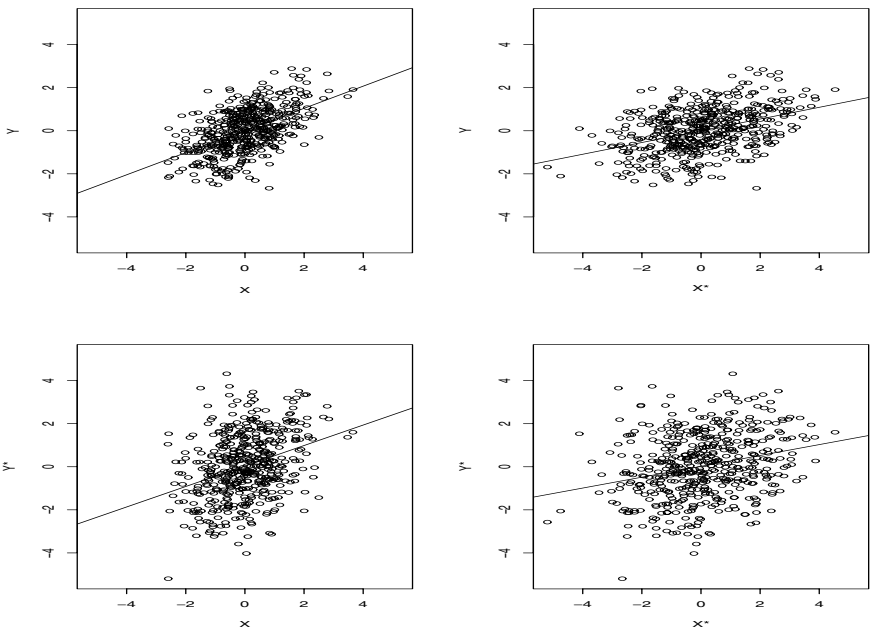

Figure 1.1 *Scatterplots for the example of Section 1.2. Starting at the upper-left and proceeding clockwise we have scatterplots of* (X, Y), (X^*, Y), (X^*, Y^*), *and* (X, Y^*).

addition of noise to Y does not change the slope estimate greatly, but it does increase the the reported uncertainty in the estimate, i.e., the standard error. Mathematically it is simple to verify that if Y^* is obtained by adding noise to Y, then the conditional expectation of Y^* given X is identically the conditional expectation of Y given X, provided the noise is uncorrelated with X. Thus adding noise to Y does not shift the estimated regression relationship in a systematic manner, though it does increase the inferential uncertainty about the (X, Y) relationship.

In contrast, say X rather than Y is subject to measurement error. The upper-right panel of Figure 1.1 gives a scatterplot of (X^*, Y), where X^* is obtained from X by adding normal noise which is independent of Y. Again the noise itself has mean zero and variance one. Visually the fitted regression line is flatter than that for (X, Y), with an estimated slope of 0.272 and a standard error of 0.027. Moreover it is hard to attribute this difference to chance in any sense. An approximate 95% confidence interval for the population slope of (X, Y) is 0.51 ± 0.07, while a similar interval from the (X^*, Y) regression is 0.27 ± 0.05. Indeed, it is not hard to verify that the conditional expectation of Y given $X^* = a$ and the conditional expectation of Y given $X = a$ are *different* functions of a. Put succinctly, if we are interested in estimating the conditional expectation of Y given X, then adding noise to X imparts a bias,

whereas adding noise to Y does not. This explains the common emphasis on mismeasurement of explanatory variables rather than mismeasurement of outcome variables.

For completeness the lower-right panel of Figure 1.1 considers the case of mismeasurement in both variables, i.e., Y^* is regressed on X^*. As might be expected the estimated regression line is close to that obtained when Y is regressed on X^*. Again, adding noise to Y does not impart a systematic change in the estimated regression function.

On the face of it a bias resulting from the mismeasurement described above might seem surprising, as the nature of the mismeasurement itself seems free from bias. In particular, the particular mismeasurement mechanism might be described as both 'blind' and 'fair.' It is blind in the sense of not depending on the outcome variable (formally the distribution of X^* given X and Y does not depend on Y), and it is fair in the sense of not systematically favouring under or over measurement (formally the distribution of X^* given X is symmetric about X). However, in looking at Figure 1.1 one can gain some geometric insight into why an inferential bias results. The impact of adding noise to the X variable is to spread out the point cloud in the horizontal direction while leaving it fixed in the vertical direction. This leads directly to a flatter slope for the regression line. On the other hand, adding noise to Y does not have an analogous effect. Since the regression line is based on minimizing the vertical variation of the points from the line, adding noise to Y does not exert a systematic 'pull' on the line.

In fact, a flattening or *attenuation* in the apparent strength of association between explanatory and outcome variables induced by mismeasurement of the explanatory variable is quite a general phenomenon. To give a simple demonstration that it carries over to categorical variables, consider a binary outcome variable Y and a binary explanatory variable V (throughout the book X and V are reserved as labels for continuous and categorical explanatory variables respectively). Say that (V, Y) values for a sample of 500 subjects are as given on the left side of Table 1.1. In particular, the empirical odds-ratio for this table is 2.68, indicating a strong positive association betwen V and Y. Now say that V cannot be measured precisely, and rather V^* is recorded, where V^* is a noisy version of V. In particular, say that each subject has a 20% chance of V^* differing from V, regardless of the value of Y. A simulated realization of V^* values yields the (V^*, Y) table on the right side of Table 1.1. The empirical odds-ratio for this table is only 2.00. Again the impact of mismeasurement is a weaker estimated relationship between the explanatory and outcome variables. Note that again attenuation is manifested even though the mismeasurement is blind to the outcome variable, and fair in the sense that the chance of mismeasurement is the same whether $V = 0$ or $V = 1$.

Thus far we have suggested that many explanatory variables of interest cannot be measured very well, and we have shown that doing the analysis by pretending the surrogate is actually the variable of interest can lead to biased inferences, not just less precise inferences. These two findings provide a rationale for what is discussed and developed in the remainder of the book.

	$Y=0$	$Y=1$		$Y=0$	$Y=1$
$V=0$	378	39	$V^*=0$	331	34
$V=1$	65	18	$V^*=1$	112	23

Table 1.1 *Simulated binary data for* 500 *subjects. Both* (V, Y) *and* (V^*, Y) *tables are given.*

1.3 What is Ahead?

The basic road-map for what is covered in this book is quite simple. Chapters 2 and 3 investigate the properties and performance of naive estimation for mismeasured continuous and categorical variables respectively. In particular, the magnitude of the attenuating bias seen in Figure 1.1 and Table 1.1 is quantified in a variety of realistic scenarios. Much of the literature gives cursory coverage of how well naive estimation performs before moving on to statistical methods which try to reverse the impact of mismeasurement. In contrast, Chapters 2 and 3 take a more detailed and leisurely look at the circumstances which make the attenuation slight or substantial. This material can be regarded as useful for developing intuition about when it is important to try to reverse the impact of mismeasurement.

Having studied the quality of naive inference, Chapters 4 and 5 turn to statistical analysis which aims to 'undo' the impact of mismeasurement. Often this might be referred to as *adjusting*, *accounting*, or *correcting* for mismeasurement. Again the separate chapters deal with continuous and categorical variables respectively. Examples of such analysis applied to both simulated datasets and real datasets are provided, and a variety of modelling strategies and issues are discussed. A central focus in these chapters is discussion of how much must be assumed or known about the mismeasurement process in order to implement a reasonable analysis which accounts for the mismeasurement.

Chapter 6 concludes the book with a look at three more specialized topics surrounding the impact of mismeasurement and attempts to ameliorate this impact. While the topics are chosen idiosyncratically, collectively they give some idea of the surprising and intricate 'twists and turns' surrounding the mismeasurement of explanatory variables. First, scenarios where a mismeasured binary variable is created from a mismeasured continuous variable are studied. It turns out that such scenarios are quite realistic, and the impact of mismeasurement in this context differs greatly from that of 'ordinary' misclassification. Second, the interplay between model misspecification and mismeasurement is scrutinized. Biased inference may arise because the postulated model for the outcome Y given the unobservable explanatory variable X is incorrect. Interesting connections and contrasts between this bias and the bias due to mismeasurement are seen to arise. Third, close attention is given to the question of how much must be known and understood about the mismeasurement process for a mismeasurement adjustment to be worthwhile. This investigation is based on quantifying the performance of estimation procedures when formally the statistical model at hand is *nonidentifiable*.

As implied by the book's subtitle, the work in Chapters 4 and 5 on adjusting for mismeasurement is undertaken via the Bayesian paradigm for statistical inference. In contrast, other books and review articles on adjusting for mismeasured variables focus almost exclusively on non-Bayesian techniques (Fuller 1987, Walter and Irwig 1987, Chen 1989, Willett 1989, Thomas, Stram and Dwyer 1993, Carroll, Ruppert and Stefanski 1995 Bashir and Duffy 1997). The present author's premise is that the Bayesian handling of unobserved quantities yields a very natural treatment of mismeasurement problems. In particular, the analysis involves aggregating over plausible values of the correct but unobserved variable. While the Bayesian approach has long been recognized as conceptually appealing, computational difficulties have been its historic Achilles' heel. Essentially, high-dimensional mathematical integration is required. The Bayesian approach has become popular in the last decade or so with the advent of new computational algorithms that tackle this integration in an indirect way. In fact, these algorithms originated much earlier in the statistical physics literature, but sustained technology-transfer to the statistical community did not commence until the work of Gelfand and Smith (1990). They introduced the *Gibbs sampling* algorithm to the statistical community, which in turn is a special instance of the class of *Markov chain Monte Carlo* (MCMC) algorithms. Most of the analyses given in Chapters 4 and 5 are implemented via MCMC algorithms. For the sake of readers unfamiliar with the Bayesian paradigm, an Appendix gives a self-contained, skeletal description of Bayesian inference and MCMC algorithms in general.

This book is intended to be neither strictly applied nor strictly theoretical in nature. The example analyses provided are meant to demonstrate what can be accomplished with the general modelling strategies presented. They are not intended as methodologies which can be taken 'off-the-shelf' and applied without modification in new problems. Mismeasurement scenarios tend to be specialized and nuanced, and usually some refinements are required to adapt a given methodology for a particular problem. On the other hand, the book is written with the idea of focussing on issues of general relevance to applied work. The overall tone of the book reflects the author's penchant for topics and issues involving an interplay between statistical theory and application. As well, the text tends to move back and forth between expounding established concepts and delving into areas of current research. This is done with the aim of focussing on the most interesting and relevant ideas, be they old or new.

The computed examples in this book are implemented in the R environment (www.r-project.org), which is standard software in the academic statistics community. Very roughly put, the R environment is an Open Source variant of the S-Plus environment (www.insightful.com). The computer code for the examples is available on the author's website (www.stat.ubc.ca/people/gustaf). Implementing MCMC algorithms in R is relatively easy given the underlying mathematical and statistical functionality of the software. Another good alternative for implementing MCMC which is not pursued here is the BUGS software (www.mrc-bsu.cam.ac.uk/bugs). BUGS has developed into the primary MCMC-specific software package. For problems without unusual fea-

tures BUGS is ideal, as the user need only specify the modelling details of the problem at hand. The software then determines and runs an appropriate MCMC algorithm. Some of the examples in the book could be readily implemented via BUGS. On the other hand, some of the examples involve 'tricky' application of MCMC, where it is necessary to exert direct control over the algorithmic details. This prompts the choice of R as the implementation platform.