SOME ASPECTS OF MEASUREMENT ERROR IN EXPLANATORY VARIABLES FOR CONTINUOUS AND BINARY REGRESSION MODELS

G. K. REEVES1*, D. R. COX2, S. C. DARBY1 AND E. WHITLEY1

1Imperial Cancer Research Fund Cancer Epidemiology Unit, University of Oxford, Oxford, U.K.
2Nuffield College and Department of Statistics, Oxford, U.K.

SUMMARY

A simple form of measurement error model for explanatory variables is studied incorporating classical and Berkson cases as particular forms, and allowing for either additive or multiplicative errors. The work is motivated by epidemiological problems, and therefore consideration is given not only to continuous response variables but also to logistic regression models. The possibility that different individuals in a study have errors of different types is also considered. The relatively simple estimation procedures proposed for use with cohort data and case-control data are checked by simulation, under the assumption of various error structures. The results show that even in situations where conventional analysis yields slope estimates that are on average attenuated by a factor of approximately 50 per cent, estimates obtained using the proposed amended likelihood functions are within 5 per cent of their true values. The work was carried out to provide a method for the analysis of lung cancer risk following residential radon exposure, but it should be applicable to a wide variety of situations. © 1998 John Wiley & Sons, Ltd.

1. INTRODUCTION

This paper is about errors in explanatory variables in studies of the dependence of a response variable on those explanatory variables. Its immediate motivation comes from the need to allow for errors in the assessment of radon gas concentrations in case-control studies of the effect of residential radon on the risk of lung cancer. In such studies, the exposure of interest is the average radon concentration in an individual’s home during, say, the past 30 years. If an individual’s estimated exposure is based on just one home, then for each individual one of two types of observation may arise. First, if it has been possible to make a direct measurement in the home, the observation will be subject to random error in the sense that repeat measurements taken in a single home at different times will vary. Secondly, in homes where no measurement is possible, for example because they have been demolished or because the current occupant refuses, the
radon concentrations can be estimated only from data from other homes in the same locality. The first type of observation is an example of simple measurement error, but the second type of error will typically be larger and different in structure. This latter type of error also arises in other types of analyses in which missing values of explanatory variables are indirectly estimated, perhaps by a suitable regression equation. The aim of this paper, therefore, is to develop relatively simple methods for incorporating observations with either of these two types of error into standard regression models. Since our motivation comes from an epidemiological study of disease incidence, emphasis is placed on binary regression models.

There is a large body of work on measurement errors in linear regression, an early discussion with some epidemiological motivation being by Cochran. More recently, there have been substantial studies of more general problems including linear logistic regression. In all such studies, the basic problem consists of using information about the relationship between the true covariate, \( X \), and the observed surrogate, \( X_s \), to provide a tractable form for the disease model in terms of \( X \). The range of methods in the literature reflects different approaches to this problem and, in particular, different assumptions about the relationship between \( X \) and \( X_s \). One particularly simple approach due to Rosner et al. involves replacing \( X \) in the exposure–disease model by \( E(X_i \mid X_s) \), the latter being derived empirically from a linear regression of \( X_i \) on \( X_s \) using data from a validation or reproducibility study. Although the resulting parameter estimates are approximately unbiased, their variance estimates must still be adjusted to take account of the variability in the estimate of \( E(X_i \mid X_s) \). This method can be viewed as a first-order approximation to the true logistic model and Kuha shows that the corresponding second-order extension gives similar results. Several authors have considered broadly similar approaches. Other general approaches to estimation of the true parameter estimates include many of the above methods as special cases and allow for a more general error structure. The problem has also been considered from a Bayesian viewpoint. Considerably less attention has been paid to the question of errors in discrete variables although this is important in epidemiological studies; Bashir and Duffy provide a brief overview of existing approaches to this problem. For a comprehensive review of methods for non-linear models, including much of the above work, see Carroll et al.; an earlier review by Thomas et al. focuses on epidemiological applications. Many of the methods discussed above involve rather mild assumptions and quite elaborate computations. By contrast, in the present paper we make rather stronger assumptions with the aim of obtaining relatively simple, directly interpretable, results.

Often, measurement error has the effect of attenuating the true regression, though there may also be a change of shape. Of course, if the purpose of the regression analysis is to construct empirical prediction formulae based on values of explanatory variables with the same distribution as the data under analysis no special discussion is needed. Regression on the surrogate variable will, however, in general need some adjustment if interest lies in regression on the notional true value of the explanatory variable. This will often be the case in epidemiology, where, for example, if public health intervention is successful it will change the true rather than the measured level of exposure. In multiple regression there is also a tendency for dependence on component explanatory variables measured with substantial error to be shifted towards correlated components measured with small error; it is then important to rectify this by proper adjustment. In what follows, we assume that the error structure is the same across the whole range of the explanatory variable. Selective biases in measuring the explanatory variable, that is biases that are different in different parts of the range, will, if present, distort the conclusions systematically; this possibility is not discussed here.
In Section 2 we outline the types of error structure considered. Section 3 presents basic results for the regression of a continuous variable on one or more covariates, including discrete covariates, and discusses possible approaches to estimation. Section 4 presents the results for binary response models. Estimation is discussed separately for prospective (that is, cohort) data and retrospective (that is, case-control) data and the properties of the resulting parameter estimates are assessed in Section 5 using simulated data. The paper concludes in Section 6 with an application of the method to a subset of the radon study which considers the appropriateness of the distributional assumptions. The limitations and advantages of the method in relation to existing approaches are also discussed.

### 2. SPECIFICATION OF THE ERROR STRUCTURE

We start by specifying the relationship between the surrogate value, \( X_s \), and the true value, \( X_t \), of the explanatory variable. If errors can be assumed to act additively, observations based on a direct measurement in the example in Section 1 correspond to a simple extreme case, namely

\[
X_s = X_t + \varepsilon_{s,t}
\]  

(1)

where \( \varepsilon_{s,t} \) is independent of \( X_t \) with \( E(\varepsilon_{s,t}) = 0 \), \( \text{var}(\varepsilon_{s,t}) = \sigma^2_{e_{s,t}} \); we call this the classical error model. A second simple extreme case can be represented as

\[
X_t = X_s + \varepsilon_{t,s}
\]  

(2)

where \( \varepsilon_{t,s} \) is independent of \( X_s \) with \( E(\varepsilon_{t,s}) = 0 \), \( \text{var}(\varepsilon_{t,s}) = \sigma^2_{e_{t,s}} \); this is known as the Berkson model. In the radon example, observations which use an indirect surrogate value of \( X \) such as the local area mean would correspond to (2) if that surrogate value was known without error. In practice, however, only an estimate of this local area mean is available and we therefore consider a generalization of (2) which includes (1) as a special case, namely

\[
X_s = \bar{X} + \varepsilon_s, \quad X_t = \bar{X} + \varepsilon_t
\]  

(3)

where, for any individual, \( (\bar{X}, \varepsilon_s, \varepsilon_t) \) are independent random variables with mean \((\mu, 0, 0)\) and variance \((\hat{\sigma}^2, \sigma^2_s, \sigma^2_t)\). Representation (3) has been used by Tosteson and Tsiatis\(^{15}\) to encapsulate classical and Berkson models. The special cases (1) and (2) correspond, respectively, to \( \sigma_t = 0 \) and to \( \sigma_s = 0 \). We may specify lack of dependence between \( (\bar{X}, \varepsilon_s, \varepsilon_t) \) in various ways of which the most extreme are that:

- (i) \( (\bar{X}, \varepsilon_s, \varepsilon_t) \) are independently normally distributed;
- (ii) \( (\bar{X}, \varepsilon_s, \varepsilon_t) \) are mutually uncorrelated.

The mild assumption (ii) is enough to justify many of the general conclusions that follow whereas the strong assumption (i) is used to obtain log-likelihood functions.

Although the more general error structure (3) encompasses both types of observation discussed in the example in Section 1, \( (\bar{X}, \varepsilon_s, \varepsilon_t) \) has a different interpretation in each case. For the first type of observation, involving only simple measurement error, \( \varepsilon_s \) is zero, \( \bar{X} \) corresponds to \( X_t \), \( \hat{\sigma}^2 \) is the variability in \( X_s \) within the same home and \( \hat{\sigma}^2 \) is the variability in \( X_t \) between the homes of the study subjects. For the second type of observation, if missing measurements are replaced by, say, the average of the available measurements in the same local area, \( \bar{X} \) represents the true area mean, \( X_s \) is an estimate of that true area mean with error \( \varepsilon_s \), and \( X_t \) is some random deviation, \( \varepsilon_t \), of an
individual's true value away from that area mean. For such observations, $\sigma_i^2$ is the within-area variability in $X_i$, $\sigma_e^2$ is the variance of the error associated with estimation of the true area mean and $\hat{\sigma}^2$ is the variability between the true area means.

Quite often when the explanatory variable represents dose or exposure level, logarithmic values are used in the resulting linear regression equation. The assumption of additive errors on log dose translates into proportional errors on dose and this is frequently a very reasonable starting assumption. In other situations, however, and specifically in the context of the radon data, it is more appropriate to consider a linear relationship between exposure and response even though empirical data point to proportional measurement errors. In such cases we need to consider the effect of proportional measurement errors on a linear regression. For this we write $Z_t = \log X_t$ and use the previous representation for the error structure of the log transformed variables, that is

$$Z_t = \bar{Z} + \epsilon_t, \quad Z_s = \bar{Z} + \epsilon_s$$

where the lack of dependence between $(\bar{Z}, \epsilon_s, \epsilon_i)$ can be specified in either of the ways described above. In what follows we concentrate on deriving results for the simple formulations (3) and (4) but generalizations of this form follow in a straightforward manner.

3. CONTINUOUS RESPONSES

3.1. Basic results

Consider the regression of a continuous response $Y$ on a single ‘true’ unobserved variable $X_t$, for which the surrogate value, $X_s$, is subject to error. We start with the simplest case of linear regression

$$Y|X_t = \alpha_t + \beta_t X_t + \xi$$

where $\xi$ is the usual regression equation error, independent of $X_t$, and such that $E(\xi) = 0$ and $\text{var}(\xi) = \sigma_\xi^2$.

To apply the representation (3) we shall need the conditional distribution of $X_t$ given $X_s$. As noted by, for example, Cox and Hinkley,\textsuperscript{16} we can write

$$X_t = \mu + \gamma_{t,s}(X_s - \mu) + X_{t,s}$$

where

$$\mu = E(X_t) = E(\bar{X}) = E(X_s), \quad \gamma_{t,s} = \frac{\text{cov}(X_t, X_s)}{\text{var}(X_s)} = \frac{\hat{\sigma}^2}{(\hat{\sigma}^2 + \sigma_i^2)}$$

and $X_{t,s}$ is an error term of zero mean and variance

$$\sigma_{t,s}^2 = \text{var}(X_t) - \frac{[\text{cov}(X_t, X_s)]^2}{\text{var}(X_s)} = \sigma_i^2 + \hat{\sigma}^2 - \gamma_{t,s}^2(\hat{\sigma}^2 + \sigma_e^2).$$

Under the normal theory assumption (i) of Section 2, (6) is a normal linear regression with $X_{t,s}$ independent of $X_s$ and normally distributed. Under the milder assumption (ii), (6) is a linear least squares regression equation with

$$E(X_{t,s}) = 0, \quad \text{cov}(X_{t,s}, X_s) = 0.$$
Substitution of (6) into the linear regression model (5) gives
\[ Y|X_s = \beta + \beta_1(X_s - \mu) + (\zeta + \beta_2X_t, s) \]
where \( \alpha = \alpha + \beta_2 \), and the regression coefficient on \( X_s \) is attenuated unless \( \gamma_{t,s} = 1 \), that is \( \sigma^2 = 0 \), the Berkson case. Furthermore, the variance of the error term is given by \( \sigma^2 + \beta_1^2 \sigma^2 \).
Thus heteroscedasticity is introduced if \( \sigma^2 \) varies between observations, and this will occur if there are observations with different types of error. On the other hand, if we consider polynomial regressions up to order three
\[ Y|X_t = \alpha_t + \beta_1(X_t - \mu) + \beta_2(X_t - \mu)^2 + \beta_3(X_t - \mu)^3 + \zeta \]
then, under the normal assumptions (i) in Section 2, \( X_{t,s} \) is independent of \( X_s \), and
\[ E(Y|X_s) = \alpha + (\beta_1 + 3\beta_3 \sigma^2)(X_s - \mu) + \beta_2\sigma_2^2(X_s - \mu)^2 + \beta_3\sigma_2^3(X_s - \mu)^3 \]
where \( \alpha = \alpha + \beta_2 \). Thus, even in the Berkson case, \( \gamma_{t,s} = 1 \), the regression is distorted unless \( \beta_3 = 0 \), that is, unless the relation is actually quadratic. Also, the variance of the modified error term now depends on \( X_s \) as well as \( \sigma^2 \), and so heteroscedasticity is inevitably induced even when the same type of error is associated with all observations. Note that although the milder assumptions (ii) in Section 2 imply \( E((X_s - \mu)X_{t,s}) = 0 \), they do not imply that \( E((X_s - \mu)^2X_{t,s}) = 0 \) and a non-zero value for the latter term would induce an additional distortion in the regression equation, corresponding to non-linear regression of \( X_t \) on \( X_s \). Other forms of non-linear regression can be investigated similarly. For example, sinusoidal regression of \( Y \) on \( X_t \) has in general its phase and amplitude affected.

3.2. Multiple regression
The above results generalize fairly directly to multiple regression although a qualitatively new aspect of interpretation arises, namely the transfer of apparent dependence from one explanatory variable to another.

The error structure is again given by (3) where, for each observation, the variables \( \tilde{X}_t, \varepsilon_t \) and \( \varepsilon_t \) now represent \( q \times 1 \) vectors with covariance matrices \( \tilde{\Sigma}, \Sigma_n \) and \( \Sigma_t \), respectively. Corresponding to (2) we can write the regression of \( X_t \) on \( X_s \) in the form
\[ X_t = \mu + \Gamma_{t,s}(X_s - \mu) + E_{t,s} \]
where \( \Gamma_{t,s} \) is the matrix of regression coefficients in which each row corresponds to a particular component of \( X_t \). It follows from Section 3.1 that
\[ \Gamma_{t,s} = \tilde{\Sigma}(\tilde{\Sigma} + \Sigma_s)^{-1} = \tilde{\Sigma}(I + \tilde{\Sigma}^{-1}\Sigma_s)^{-1}\Sigma^{-1} \]
and that the covariance matrix of the residual term \( E_{t,s} \) is given by
\[ \text{cov}(E_{t,s}) = (\tilde{\Sigma} + \Sigma_n) - \tilde{\Sigma}(\tilde{\Sigma} + \Sigma_n)^{-1}\tilde{\Sigma} = \tilde{\Sigma}(I - (I + \tilde{\Sigma}^{-1}\Sigma_n)^{-1}) + \Sigma_n \]
Conditional upon \( X_s \), the multiple regression equation can, therefore, be written
\[ Y|X_s = \alpha + \beta_1(X_s - \mu) + \Gamma_{t,s}(X_s - \mu) + E_{t,s} + \zeta \]
which is equivalent to linear regression of $Y$ on $X_s$ with regression coefficients $\beta^T$, where

$$\beta^T = \beta^T I_{s,s} = \beta^T \bar{\Sigma}(I + \Sigma^{-1}\Sigma_s)^{-1}\Sigma^{-1}. $$

If the measurement error variance is relatively small we can make the expansion

$$(I + \bar{\Sigma}^{-1}\Sigma_s)^{-1} \approx I - \bar{\Sigma}^{-1}\Sigma_s $$

leading to

$$\beta^T_s \approx \beta^T_t - \beta^T_t\Sigma_s\Sigma^{-1}. $$

It throws some light on the multivariate case to suppose that the errors are in just one of the explanatory variables, say the first. Then $\Sigma_s$ is zero except for the leading element. Explicit calculation of $(I + \bar{\Sigma}^{-1}\Sigma_s)^{-1}$ is now possible. If $\sigma^2_{rs}$ denotes the $(r, s)$ element of $\bar{\Sigma}^{-1}$, it follows from the properties of inverse covariance matrices that $\sigma^2_{1,rem}$, the conditional variance of $\tilde{X}_1$, the first component of $\tilde{X}$, about its linear regression on all other components, is equal to $1/\sigma^2_{11}$ and $\tilde{\beta}_{1,rem}$, the regression coefficient of $\tilde{X}_1$ on $\tilde{X}_j$, adjusting for all other components, is equal to $-\sigma^2_{11}/\sigma^2_{11}$. It follows after evaluating the relevant matrix product that

$$\beta_{s1} = \beta_{t1}\sigma^2_{1,rem}/(\sigma^2_{1,rem} + \sigma^2_{s1}) $$

and that for $j \neq 1$

$$\beta_{s1} = \beta_{t1} + \beta_{t1}\tilde{\beta}_{1,rem}\sigma^2_{s1}/(\sigma^2_{1,rem} + \sigma^2_{s1}). $$

This shows that the regression coefficient on $\tilde{X}_1$ is attenuated in the same way as in the scalar case and that the effect of errors in it are spread across the other explanatory variables in a quite simple way.

If some explanatory variables are categorical and some continuous, no new considerations are involved if the categorical variables are recorded without error and the errors in the continuous variables are independent of the categorical variables. However, where the categorical variables are strongly correlated with $X_1$, then account should be taken of this in the estimation procedure, see Section 3.4.

Suppose now that a categorical variable is subject to measurement error, that is, to potential misclassification. We consider for simplicity a binary variable $W$ taking values 0 and 1 and recorded with error, combined with a vector, $X$, of continuous explanatory variables also recorded with error, but such that the errors in $X$ and $W$ are mutually independent and independent of the true values of $X$. For simplicity we also make the strong assumption that $W_1$ and $X_1$ are independent. Then for continuous responses, we start from the model

$$Y|X, W = \mu + \beta^T X + \delta_1 W + \zeta $$

where $\zeta$ is the usual regression equation error. As before, $X_1$ can be replaced by its regression on $X_s$, and the effect on the regression coefficient is given by the results above. Under the analogue for binary $W_1$ of the classical error model, we can write

$$\Pr(W_1 = 1) = \pi_{11}, \Pr(W_1 = 0) = \pi_{00} = 1 - \pi_{11}$$
and introduce the error probabilities $\rho_1, \rho_0$ such that

$$\Pr(W_s = 0|W_t = 1) = \rho_1, \ Pr(W_s = 1|W_t = 0) = \rho_0$$

so that

$$\pi_{s1} = \Pr(W_s = 1) = \pi_{t1}(1 - \rho_1) + \pi_{t0}\rho_0$$

and

$$\pi_{s0} = \Pr(W_s = 0) = \pi_{t0}(1 - \rho_0) + \pi_{t1}\rho_1.$$  

Now it follows, on calculating the least squares regression of $W_t$ on $W_s$, that we can write

$$W_t = \pi_{t1} + \gamma_{t,s}(W_s - \pi_{s1}) + W_{t,s}$$

where

$$\gamma_{t,s} = \frac{\text{cov}(W_s, W_t)}{\text{var}(W_s)} = \frac{\pi_{t1}\pi_{t0}(1 - \rho_1 - \rho_0)}{\pi_{t1}\pi_{t0}(1 - \rho_1 - \rho_0) + \pi_{t1}\rho_1 + \pi_{t0}\rho_0 - (\rho_1\pi_{t1} - \rho_0\pi_{t0})^2}$$

and $W_{t,s}$ is a residual of zero mean, uncorrelated with $W_s$, and having for each $W_s$ a two-point distribution. Further, the variance of $W_{t,s}$ is given by

$$\frac{\pi_{t1}\pi_{t0}}{\pi_{s1}\pi_{s0}} \{\pi_{t1}\pi_{t0}(1 - \rho_1 - \rho_0)(1 - \pi_{t1}\pi_{t0}(1 - \rho_1 - \rho_0)) + \rho_1\pi_{t1}^2 + \rho_0\pi_{t0}^2 - (\rho_1\pi_{t1} - \rho_0\pi_{t0})^2\}.$$  

On substitution into (10) we have that the effect of the binary explanatory variable is attenuated to

$$\hat{\delta}_s = \delta_{\gamma_{t,s}}$$

and that the regression equation error from the regression equation for $Y$ on $(X_s, W_t)$ has a distribution that is the convolution of the distribution of $\zeta$ in (10), for example a normal distribution, with the distribution of $X_{t,s}$ and the two-point distribution for $W_{t,s}$. Thus in extreme cases its distribution can be seen to be a mixture of two components.

If the errors in $X$ and $W$ are not independent, we replace $\rho_0, \rho_1$ by their conditional values given $X_s$, obtained in principle via the corresponding values given $X_t$. The result is an interaction term, to a first approximation linear by linear, between $X_s$ and $W_s$ in the regression of $Y$ on $(X_s, W_s)$.

If there is an interaction between the effect of $W_t$ and one or more of the continuous explanatory variables, the interaction is attenuated by both sources of measurement error. When, as would typically be the case, $X_t$ and $W_t$ are not independent, so that their conditional distributions given $X_s$ and $W_s$ both depend on both arguments, the approach can be extended but the results are complicated and will not be given here.

### 3.3. Proportional errors

If $Z_t = \log X_t$, $Z_s = \log X_s$ and the underlying error structure is given by (4), the conditional distribution of $Z_t$ given $Z_s$ is of the form

$$Z_t = v + \gamma_{t,s}(Z_s - v) + Z_{t,s}$$

© 1998 John Wiley & Sons, Ltd.  
where $v = E(Z_t) = E(\tilde{Z}) = E(Z_s)$ and $Z_{t,s}$ is an error of zero mean and variance $\sigma^2_{t,s}$, defined previously. Thus, given $X_s$

$$X_t = \bar{v}^{1-\gamma_s}X_{t,s}e^{Z_t}$$

where $\bar{v} = e^\nu$. Substitution of (11) into the original linear regression equation (5) gives

$$Y = \alpha_t + \beta_t(\bar{v}^{1-\gamma_s}X_{t,s}e^{Z_t}) + \zeta.$$  \hfill (12)

It can be shown that, under normality of $Z_{t,s}$, and approximately more generally if $\sigma^2_{t,s}$ is small

$$E(e^{Z_t} | Z_s) = e^{\sigma^2_{t,s}/2}, \quad \text{var}(e^{Z_t} | Z_s) = e^{\sigma^2_{t,s}}(e^{\sigma^2_{t,s}} - 1).$$  

Hence a simple linear regression of the form (5) is converted into a non-linear regression of the form

$$Y = \alpha_t + \beta_tX_{t,s}^\gamma + \zeta*$$

where $\beta_t = \beta_t\bar{v}^{1-\gamma_s}e^{\sigma^2_{t,s}/2}$ and $\zeta*$ is an error term of zero mean and variance $\sigma^2_{t,s} + \beta_t^2(e^{\sigma^2_{t,s}} - 1)X_{t,s}^\gamma$.

A qualitative conclusion is that an observed concave dependence of response on the measured explanatory variable might sometimes be accounted for by proportional measurement errors in that explanatory variable taken with a linear dose–response relation.

### 3.4. Estimation

The full likelihood of the observed data is a product of two factors, namely the conditional distribution, $f(Y|X_s)$, and the marginal distribution of the $X_s$, $f(X_s)$. The precise form of $f(Y|X_s)$ and $f(X_s)$ for a given observation will depend on the type of error associated with that observation but the product can generally be expressed in terms of the regression parameters $\alpha_t$ and $\beta_t$, and up to four distinct nuisance parameters. Let $\sigma_m^2$ denote the component of variance due to measurement error of the classical type and, if the local areas referred to in Section 2 are considered generally to be groups, let $\sigma_{wg}^2$ be the within-group variability in $X_1$, $\sigma_{bg}^2$ the variability between the true group means and $\mu$ the population mean of $X_1$. Since information about $\sigma_2^2$ can be obtained only from auxiliary data it is assumed here that this parameter is effectively known. This requires either duplicate observations on a subsample of individuals, preferably randomly selected, or more detailed analysis of the method used to produce the surrogate values. Under the assumption of additive errors, the value of $f(Y|X_s)$ for individuals with a direct measurement of $X$ is derived from (9) in which

$$\gamma_{1,s} = (\sigma_{wg}^2 + \sigma_{bg}^2)/(\sigma_{wg}^2 + \sigma_{bg}^2 + \sigma_m^2), \quad \gamma_{1,s} = \sigma_{wg}^2 + \sigma_{bg}^2 - \gamma_{1,s}(\sigma_{wg}^2 + \sigma_{bg}^2 + \sigma_m^2)$$

and $X_s$ is taken to be $N(\mu, \sigma_{wg}^2 + \sigma_{bg}^2 + \sigma_m^2)$. The corresponding value of $f(Y|X_s)$ for individuals whose exposure is estimated by their group mean, is also derived from (9) but with

$$\gamma_{1,s} = \sigma_{bg}^2/(\sigma_{bg}^2 + (\sigma_{wg}^2 + \sigma_m^2)/n), \quad \gamma_{1,s} = \sigma_{bg}^2 + \sigma_{wg}^2 - \gamma_{1,s}(\sigma_{bg}^2 + (\sigma_{wg}^2 + \sigma_m^2)/n)$$

where $n$ denotes the number of direct measurements in the individual's group; in this case $X_s$ is taken to be $N(\mu, \sigma_{bg}^2)$. In the case of proportional errors, the definitions are exactly analogous but refer to $Z_t$ rather than to $X_t$.

Although a full maximum likelihood treatment based on the above definitions of $f(Y|X_s)$ and $f(X_s)$ is possible, it is likely to be cumbersome and in the present context a much simpler, but
virtually equivalent, analysis is possible. This simplification stems from the fact that information about the nuisance parameters, \( \mu, \sigma_{w}^2 \) and \( \sigma_{\epsilon}^2 \), is essentially restricted to the marginal distribution, \( f(X_s) \), while information about the regression coefficients, \( x_i \) and \( \beta_i \), is contained only in the conditional part of the likelihood, \( f(Y|X_s) \). It is, therefore, feasible to base estimation of \( x_i \) and \( \beta_i \) on a likelihood function consisting solely of \( f(Y|X_s) \) in which the unknown nuisance parameters \( \mu, \sigma_{w}^2 \) and \( \sigma_{\epsilon}^2 \) have all been replaced by suitable estimates based on the sample distribution of \( X_s \) and the known value of \( \sigma_{m}^2 \). Although this approach is intuitively appealing, the following gives a more formal justification which considers the effect on the estimated standard errors of the regression coefficients.

The full likelihood can be expressed as the product
\[
f(Y|X_s; \theta) f(X_s; \phi) \tag{13}
\]
where \( \theta \) is a set of parameters sufficient to determine the first factor and \( \phi \) determines the marginal distribution of \( X_s \). It is difficult to specify \( \theta \) explicitly but this is fortunately not necessary. Suppose that the parameters of interest are \( \beta = \hat{\beta}(\theta, \phi) \). If we regard \( \phi \) as known, \( \phi_0 \), say, then \( \hat{\beta} \) is estimated as \( \beta = \hat{\beta}(\hat{\theta}, \phi_0) \) directly by maximum likelihood applied to the first factor. If now we estimate \( \phi \) by any reasonable method, possibly but not necessarily maximum likelihood, applied to the marginal distribution of \( X_s \), giving an estimate of \( \phi \) with asymptotic covariance matrix \( \text{cov}(\hat{\phi}) \), then locally, because of the factorization of the likelihood
\[
\text{cov}(\beta(\hat{\theta}, \hat{\phi})) = \text{cov}(\beta(\hat{\theta}, \phi_0)) + (\nabla_\phi \beta)^T \text{cov}(\hat{\phi}) (\nabla_\phi \beta^T)
\tag{14}
\]
where the second term is the inflation of the covariance matrix of \( \hat{\beta} \), arising from errors in \( \hat{\phi} \), and \( (\nabla_\phi \beta^T) \) is the matrix of gradients of \( \beta \) with respect to \( \phi \) and can be estimated numerically.

In all the examples considered in Section 5, the second term in (14) is negligible. In particular, the estimate of the slope parameter, \( \hat{\beta}_i \), is remarkably insensitive to the assumed value of the mean, \( \mu \), over a wide range and, despite its rather greater dependence on the ratios of variance components, the effect on the precision of \( \hat{\beta}_i \) is still very small. This is confirmed in Section 5 by comparison with standard errors based on maximization of the full likelihood (13).

In practice, the appropriate choice of estimation procedure will depend on the problem in hand and, in particular, the nature of the information regarding the variability in repeat measurements, \( \sigma_{\epsilon}^2 \), which we have assumed known. For example, if this assumption was unreasonable, or if the repeat measurement data could be viewed as containing information about the marginal distribution of \( X_s \) in the study data, then it would be feasible to include an additional term in (13) to describe the likelihood of these auxiliary data and maximize the resulting likelihood function with respect to all unknown parameters.

One minor complication is that although observations on different individuals are independent given the true values of the explanatory variable, the error structure may induce some degree of correlation between certain individuals given their observed values. For example, in the application given in Section 1, within a given local area, observations with no direct measurement will be assigned the same estimate of the local area mean and are therefore correlated through their common value of \( \epsilon \). In addition, if the local area means are based on direct measurements made as part of the study, there will be correlations between the estimated and directly measured observations within the same local area. In practice these correlations are likely to be negligible, especially if the number of direct measurements within a group is large. Otherwise, the appropriate covariance matrix would in principle need to be specified in terms of the various variance components, including the unknown \( \sigma_{\epsilon}^2 \), and incorporated into the estimation procedure.
The same general approach can be applied to estimation in multiple regression involving several covariates, each of which may or may not have measurement error and can be extended to accommodate regression on categorical variables with error. It is worth noting, however, that in the case where categorical variables are measured without error but are strongly correlated with \( X \), then it may be more appropriate to calculate estimates of \( \sigma_{w}^{2} \) and \( \sigma_{b}^{2} \) conditional on the values of the covariates; such estimates may then be pooled to provide a more stable estimate of the relevant component of variance. When such conditioning is carried out it may also be necessary to consider \( \mu \) specific to the values of the covariates.

4. BINARY RESPONSES

4.1. Basic results

Consider now a single binary response variable, \( Y \), and a single continuous explanatory variable, \( X \), with the additive error structure described in (3). Since our motivation for this work comes from a case-control study we assume throughout that interest lies in the logistic regression of \( Y \) on \( X \), that is

\[
Pr(Y = 1|X_i) = \Lambda(x_i + \beta_i X_i) \tag{15}
\]

where as before, \( x_i \) and \( \beta_i \) are unknown parameters to be determined, \( X_i \) is the true, but unobserved, value of the explanatory variable, and \( \Lambda \) denotes the logistic function \( \Lambda(t) = e^t/(1 + e^t) \). The close relationship between the logit and probit form,\(^{17,18} \) namely \( \Lambda(t) \approx \Phi(kt) \), where \( k \approx 0.588 \), allows us to rewrite (15) in the form

\[
Pr(Y = 1|X_i) \approx \Phi \left[ k(x_i + \beta_i X_i) \right] = Pr \{ U \leq k(x_i + \beta_i X_i) \}
\]

where \( U \) is a standardized normal random variable. To express this probability conditionally upon \( X_s \), we replace \( X_i \) by its linear regression on \( X_s \) derived in (6) to give

\[
Pr(Y = 1|X_s) \approx Pr \{ U - k \beta_i X_{i,s} \leq k \{ x_i + \beta_i \mu (1 - \gamma_{i,s}) + \beta_{i,s} X_s \} \} \tag{16}
\]

where \( \mu \) and \( \gamma_{i,s} \) are as defined in (7). Under the normal theory assumptions (i) in Section 2, the random component in (16) has expectation zero and variance \( 1 + k^2 \beta_i^2 \sigma_{i,s}^2 \), where \( \sigma_{i,s}^2 \) is the variance of \( X_{i,s} \), defined in (8). Hence

\[
Pr(Y = 1|X_s) \approx \Phi \left( \frac{k \{ x_i + \beta_i \mu (1 - \gamma_{i,s}) + \beta_{i,s} X_s \}}{1 + k^2 \beta_i^2 \sigma_{i,s}^2} \right) \approx \Lambda \left( \frac{x_s + \beta_{i,s} X_s}{1 + k^2 \beta_i^2 \sigma_{i,s}^2} \right) \tag{17}
\]

where \( x_s = x_i + \beta_i \mu (1 - \gamma_{i,s}) \). Under these assumptions, therefore, the logistic form is preserved but with regression coefficient \( \beta_s \), which, in terms of the true coefficient, \( \beta_i \), is given by

\[
\beta_s = \frac{\beta_{i,s}}{1 + k^2 \beta_i^2 \sigma_{i,s}^2} \tag{18}
\]

This illustrates that for binary responses, unlike continuous responses, there is still some degree of attenuation even in the case of the simple Berkson model\(^{12} \) (\( \gamma_{i,s} = 1, \sigma_s = 0 \)). The effects of measurement error on the results of probit regression are discussed by Carroll et al.\(^{19} \) and, in the particular case of Berkson error, by Burr.\(^{20} \) Duffy et al.\(^{21} \) have considered errors in the
case-control setting emphasizing misclassification rather than measurement errors in continuous explanatory variables.

This approach is easily extended to accommodate more than one continuous covariate simply by replacing the components of (17) by their vector forms, defined in Section 3.2. In some applications it may be found that the denominator of (18) is effectively unity, with a resulting simplification in both interpretation and computation.

4.2. Proportional errors

Suppose now that the measurement errors in $X_i$ are proportional to $X_i$, with structure as given in (4). If we assume the same basic model as before, namely (5), and replace $X_i$ by the expression given in (11) we have

$$
\Pr(Y = 1 | X_s) \approx \Pr(U - k\beta_i X_i^* \leq k z_i).
$$

(19)

Under the assumption of normality of $Z_i$, and conditional upon $X_i$, the random part in (19) is approximately normal with mean $- k\beta_i X_i^*$ and variance $1 + k^2 \beta_i^2 (v^*)^2 (e^{\sigma^2} - 1) X_i^2$, where $v^* = v^* - e^{\sigma^2}$. Thus

$$
\Pr(Y = 1 | X_s) \approx \Phi \left( \frac{k(z_i + \beta_i X_i^*)}{\{1 + k^2 \beta_i^2 (v^*)^2 (e^{\sigma^2} - 1) X_i^2\}^{1/2}} \right) \approx \Lambda \left( \frac{z_i + \beta_i X_i^*}{\{1 + k^2 \beta_i^2 (v^*)^2 (e^{\sigma^2} - 1) X_i^2\}^{1/2}} \right).
$$

(20)

If the denominator is close to unity, this represents a logistic regression modified in a way similar to the continuous case. For larger values of $X_s$ there will be a further flattening.

4.3. Estimation

4.3.1. Prospective data

If the observations can be assumed to be approximately independent given $X_s$, the likelihood for the observed data is derived from (17) or (20), depending on whether the errors are additive or multiplicative. Thus estimation of $z_i$ and $\beta_i$ can be based on maximization of this likelihood, treating parameters other than $z_i$ or $\beta_i$ as described in Section 3.4.

As for the continuous case, however, if Berkson type errors are involved, some degree of correlation will necessarily exist between certain individuals in the same group. If these are large enough to be taken into account, one approach might be to approximate the joint distribution of the binary responses in a particular group by an appropriate quadratic exponential binary model.\(^\text{22}\)

4.3.2. Retrospective data

It has been shown that maximum likelihood estimates of odds ratios and their asymptotic covariance matrices can be obtained from case-control data by applying the usual logistic model,\(^\text{23}\) although the intercept term obtained from such data does not have the same interpretation as for prospective data. It follows then that for unmatched case-control data with errors in $X$, an unconditional analysis can be based directly on (17) or (20), depending on whether the errors are additive or multiplicative. It is worth noting that, as in Section 3.4 for covariates, when the data have been matched with respect to certain variables, however broadly, the precise choice of estimates for parameters such as $\sigma_{wg}$ and $\sigma_{bg}$ may depend on the matching criteria.
When a matched design has been used it is often more appropriate to carry out a conditional logistic regression analysis, thus avoiding the need to estimate a separate intercept term, $\alpha$, for each matched set. Suppose that there are $M$ matched sets, comprising $n_{i(i)}$ cases and $n_{0(i)}$ controls, $i = 1, \ldots, M$. If the relationship between $Y$ and $X_i$ is given by (15), then in the absence of measurement error the contribution to the conditional likelihood from the $i$th matched set, with intercept $\alpha_{i(i)}$, is given by

$$
\frac{\prod_{j=1}^{n_{i(i)}} \Pr(Y = 1 | X_{i(j)})}{\sum_{l=1}^{M} \prod_{j=1}^{n_{i(l)}} \Pr(Y = 1 | X_{i(l)})} = \prod_{j=1}^{n_{i(i)}} \frac{\exp(\alpha_{i(i)} + \beta_i X_{i(j)})}{\sum_{l=1}^{M} \prod_{j=1}^{n_{i(l)}} \exp(\alpha_{i(l)} + \beta_i X_{i(l)})}
$$

where summation over $l$ is over all possible choices of $n_{i(l)}$ cases out of $n_{i(i)} + n_{0(i)}$ controls. In the presence of, say, additive errors, then $\Pr(Y = 1 | X_i)$ would be substituted by $\Pr(Y = 1 | X_s)$, as given in (17), to give

$$
\prod_{j=1}^{n_{i(i)}} \frac{\exp(\alpha_{i(i)} + \beta_i (1 - \gamma_{i,s}) \mu + \beta_i \gamma_{i,s} X_{i(j)})}{\sum_{l=1}^{M} \prod_{j=1}^{n_{i(l)}} \exp(\alpha_{i(l)} + \beta_i (1 - \gamma_{i,s}) \mu + \beta_i \gamma_{i,s} X_{i(l)})}
$$

In situations where $\sigma^2_{i,s}$, $\mu$ and $\gamma_{i,s}$ are identical for all members of the matched set, for example when all members of the set are subject only to simple measurement error, then (21) simplifies to

$$
\prod_{j=1}^{n_{i(i)}} \frac{\exp(\beta_i \gamma_{i,s} X_{i(j)})}{\sum_{l=1}^{M} \prod_{j=1}^{n_{i(l)}} \exp(\beta_i \gamma_{i,s} X_{i(l)})}
$$

In general, however, not all the $\alpha_{i(l)}$ will be eliminated from the conditional likelihood and it is likely that the contribution from some matched sets will involve a term $\alpha_{i(l)}$ and others will not, depending on whether individuals within the matched set differ with respect to the type of error associated with them.

In the special case of simple measurement error, it is straightforward to estimate $\beta_i$ via maximization of the likelihood function based on (22), treating parameters other than $\beta_i$ as described in Section 3.4. In the more general situation (21) one possible approach is to assume that, since conditioning should effectively eliminate dependence on the individual $\alpha_{i(i)}$, estimation of $\beta_i$ can, approximately, be based on the simplified form

$$
\prod_{j=1}^{n_{i(i)}} \frac{\exp(\beta_i (1 - \gamma_{i,s}) \mu + \beta_i \gamma_{i,s} X_{i(j)})}{\sum_{l=1}^{M} \prod_{j=1}^{n_{i(l)}} \exp(\beta_i (1 - \gamma_{i,s}) \mu + \beta_i \gamma_{i,s} X_{i(l)})}
$$

using one of the approaches described in Section 3.4. As mentioned previously, if the matching variables are substantially correlated with $X_i$, then this will need to be taken into account when estimating parameters relating to the sample variance of $X_i$. 

Similar arguments can be used to derive an appropriate conditional likelihood for the case of proportional errors. The approach is illustrated in Section 5.2 using simulated case-control data with additive errors. The example in Section 6 illustrates the results of applying the method to case-control data with proportional errors.

5. SIMULATION STUDIES

We illustrate the methods on simulated binary response data with error structure reflecting the radon study mentioned in Section 1. For most individuals the explanatory variable is measured with error but for some it is missing and is imputed from the average of a suitable subgroup. We deal with both additive and proportional errors and with prospective and matched case-control data.

To simulate prospective data, we took 10 groups of 50 individuals each. For additive errors the true covariate values were generated from normal distributions, $\mathcal{N}(\mu_i, \sigma^2_{\text{true}})$, where $\mu_i$ has the distribution $\mathcal{N}(0, \sigma^2_{\text{me}})$. For each individual the surrogate value was found by adding to $X_i$, a random term, $\mathcal{N}(0, \sigma^2_{\text{me}})$. To achieve a small number of individuals with Berkson type error, the last five values of $X_i$ in each group were replaced by the mean of the other 45 values. For proportional errors the same procedure was followed for $\log X_i$. Finally for each individual a binary variable, $y$, was formed with probability of positive response $\Lambda(x_i + \beta_i X_i)$, where $x_i$ was chosen so that $\Lambda(x_i) = 0.1$.

To generate matched pair case-control data, we matched with respect to a single variable $\psi$ representing age at five levels and assumed known without error and independent of $X_i$. Here $X_i$ and $\psi$ are random and disease status $y$ is fixed. The marginal distribution of $X_i$ is specified as for the prospective data. It is also assumed that each age-group has probability $1/5$ in the population.

Finally it was assumed that

$$\Pr(Y = y | X_i, V) = \Lambda\{(x_i + \beta_i X_i + \delta) y\}$$

where $\delta$ represents the effect of age taking values $-0.5, -0.25, 0, 0.25, 0.5$ and where $x_i$ is chosen to give the probability of a positive response of $0.1$ at $X_i=0, V=3$.

With this specification it is possible to find the conditional distribution of $X_i$ and $V$ given disease status and to generate synthetic matched case-control data as follows. For the case, values of $X_i$ and $V$ were generated from their marginal distributions and $y$ then generated. If it was one, it was assigned to a case, if not the procedure was repeated using new values of $X_i$ and $V$. To generate a matching control, the same process was carried out until a value of $Y = 0$ was achieved for which the value of $V$ was the same as that for the corresponding case. Each simulated set contained 250 matched pairs. Then measurement errors were added to $X_i$ as previously.

In all the analyses $\sigma^2_{\text{me}}$ was assumed known, the other variance parameters and means being estimated from the data and then regarded as known in the subsequent maximum likelihood analysis for $\beta_i$. Some more detailed simulation studies confirmed the argument of Section 3 that the effect of ignoring errors of estimation in these parameters was negligible.

The results of one of the sets of simulations are reported in two different forms. Tables I and II show average estimates from 1000 simulations together with standard errors of estimation, the first table for a prospective study and the second for a matched case-control study. In Table I the results for both additive and proportional errors are shown for a range of values of $\sigma^2_{\text{me}}$. The conclusions, confirmed in the other configurations examined, are similar in the two types of
Table I. Summary of estimates of the true slope parameter ($\beta_i$) from simulated data from a prospective study with either additive errors with $\sigma_{\epsilon_i}^2 = 1-0$ and $\sigma_{\epsilon_i}^2 = 1-5$ or proportional errors with $\sigma_{\epsilon_i}^2 = 0-75$ and $\sigma_{\epsilon_i}^2 = 1-0$. Entries are the average estimated value of $\beta_i$ (and average estimated standard error*) from 1000 simulations of size 500 using (a) the amended likelihood and (b) unadjusted logistic regression

(i) Additive errors

<table>
<thead>
<tr>
<th>$\beta_i$</th>
<th>Measurement error variance ($\sigma_{\epsilon_i}^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0-25 (a)</td>
</tr>
<tr>
<td>0-25</td>
<td>0-25 (0-11)</td>
</tr>
<tr>
<td>0-5</td>
<td>0-50 (0-12)</td>
</tr>
<tr>
<td>0-75</td>
<td>0-76 (0-13)</td>
</tr>
<tr>
<td>1-0</td>
<td>1-01 (0-14)</td>
</tr>
</tbody>
</table>

(ii) Proportional errors

<table>
<thead>
<tr>
<th>$\beta_i$</th>
<th>Coefficient of variation of $X_i \sqrt{(e^{\epsilon_i^2} - 1)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0-25 (a)</td>
</tr>
<tr>
<td>0-25</td>
<td>0-25 (0-05)</td>
</tr>
<tr>
<td>0-5</td>
<td>0-49 (0-07)</td>
</tr>
<tr>
<td>0-75</td>
<td>0-76 (0-10)</td>
</tr>
<tr>
<td>1-0</td>
<td>1-02 (0-13)</td>
</tr>
</tbody>
</table>

* Estimated as the square root of the average of the individual variance estimates

Table II. Summary of estimates of the true slope parameter ($\beta_i$) from simulated data from a case-control study with additive errors with $\sigma_{\epsilon_i}^2 = 1-0$ and $\sigma_{\epsilon_i}^2 = 1-5$. Entries are the average estimated value of $\beta_i$ (and average estimated standard error*) from 1000 simulated data sets containing 250 matched pairs using: (a) an amended likelihood based on (23) and (b) unadjusted conditional logistic regression

<table>
<thead>
<tr>
<th>$\beta_i$</th>
<th>Measurement error variance ($\sigma_{\epsilon_i}^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0-25 (a)</td>
</tr>
<tr>
<td>0-25</td>
<td>0-25 (0-07)</td>
</tr>
<tr>
<td>0-5</td>
<td>0-50 (0-08)</td>
</tr>
<tr>
<td>0-75</td>
<td>0-75 (0-11)</td>
</tr>
<tr>
<td>1-0</td>
<td>1-01 (0-14)</td>
</tr>
</tbody>
</table>

* Estimated as the square root of the average of the individual variance estimates

study. In particular, the substantial attenuation biases which arise at the higher slopes and larger $\sigma_{\epsilon_i}^2$ are successfully removed by the amended approach.

The second analysis, summarized in Figures 1 and 2, examines the distribution of the standardized score statistics for $\beta_i$ as derived from the point estimate and estimated standard
Figure 1. Plot of observed score statistics versus expected normal order statistics from 200 simulations of data from a prospective study subject to additive errors with $\sigma^2_{\epsilon} = 1.0$, $\sigma^2_{\eta} = 1.5$, $\sigma^2_{m} = 0.75$, and true slope parameter $\beta$ equal to (a) 0.25, (b) 0.5, (c) 0.75, (d) 1.0.
Figure 2. Plot of observed score statistics versus expected normal order statistics from 200 simulations of data from a prospective study subject to proportional errors with $\sigma_2^2 = 0.75, \sigma_3^2 = 1.0, \sigma_4^2 = 0.75$, and true slope parameter $\beta_i$ equal to (a) 0.25, (b) 0.5, (c) 0.75, (d) 1.0
error computed from the inverse of the observed information matrix. Agreement with asymptotic theory is shown by a straight line of unit slope in the normal probability plots; except for a suggestion of some skewness in the more extreme case-control situations, the agreement is very good, showing that standard confidence intervals for $\beta_i$ can be easily derived via the proposed method.

6. EXAMPLE AND DISCUSSION

In this section, we illustrate the application of the method to a simple subset of data from a case-control study of radon and lung cancer in Southwest England. In particular, we consider how to assess the most appropriate choice of error structure and whether the distributional assumptions required by the method are met. The data consist of all available cases who reported being current smokers of 15–24 cigarettes per day for whom there was a control in the same age-sex category with similar reported smoking habits. A total of 196 such matched pairs were available. For simplicity, the exposure of interest, $X$, is the radon concentration in the residence occupied 5 years prior to interview. Data collection was still under way at the time of this analysis and for 92 of these 392 individuals no direct measurement of $X$ available. Maps of residential radon concentration in the U.K. based on surveys by the National Radiological Protection Board, were used to divide the study area into five local areas, with similar levels of residential radon. The relationship between disease and true exposure is assumed to be given by the logistic model $\text{Pr}(\text{disease} \mid X_i) = \Lambda(z_i + \beta_i X_i)$, where interest lies in the value of $\beta_i$.

Before applying the methods of Sections 3 and 4 to these data, we must first establish what form the error structure is likely to take. External data were available from 217 residences in the study area from a survey carried out using similar detectors and a similar measurement protocol to those used for the study. For each residence, a pair of measurements was available, separated by an interval of up to 10 years. Figure 3(a) shows a plot of the within-residence variability in $X$, according to the average value of $X$ within residences. The variance of the measurement error clearly increases with $X$, indicating that the measurement error is proportional to $X$. The corresponding plot (Figure 3(b)) of the variability in log $X$ confirms that the measurement error is approximately additive on a log scale. The estimate of the within-residence variance in log $X$ was 0.23, corresponding to a coefficient of variation on the original scale of about 0.5.

Under the assumption of proportional errors, the methods of Section 4 require that both log $X$ and log $X$ are normally distributed. A normal order plot for the 300 direct measurements within the study data is shown in Figure 4 and can be seen to agree closely with normality. Although it is impossible to assess directly the distribution of the unobserved log $X$ within the study data, one possibility is to examine the distribution of the mean value of log $X$ within residences in the repeated measurement data. A plot of the latter variable was also in line with normality. On the basis of these results, proportional errors were assumed and for those residences without a direct measurement, $X$ was taken to be the geometric mean of the direct measurements within the same local area.

Although the error structure is additive on a log scale, interest lies in the regression on $X_i$ rather than log $X_i$ thus estimation of the true slope parameter $\beta_i$ is based on the form (20) in which $\sigma^2_m = 0.23$. First, however, the remaining nuisance parameters relating to the marginal distribution of log $X$ must be estimated as described in Section 3.4. This was done by assuming a random-effects model for those values of log $X$ arising from a direct measurement, in which the random term due to area has variance $\sigma^2_{\text{area}}$ and the error term has variance $\sigma^2_{\text{error}} + \sigma^2_m$, where $\sigma^2_m$ is...
Figure 3. Plot of within residence variability in $X_4$, according to the average value of $X_4$ within residences: (a) using original scale in Bq m$^{-3}$; (b) using a log transformation. Data from Lomas and Green$^{24}$.

assumed to be 0.23. In this case the number of direct measurements varied substantially by area and so all the nuisance parameters were estimated from the marginal distribution of $X_4$ using restricted maximum likelihood. Substituting these estimates into the amended conditional likelihood described in Section 4, and maximizing with respect to $\beta_1$ gives a slope estimate of 0.15 per 100 Bq m$^{-3}$ with estimated standard error 0.25. By comparison, an unadjusted analysis of these data using conventional conditional logistic regression, ignoring all sources of error in the assessment of radon concentrations, and estimating the missing radon measurements by the geometric means of all the available measurements in the relevant local area gave an estimated slope parameter of 0.09 with standard error 0.16. If the missing radon measurements were estimated by the arithmetic means of all the available measurements in the relevant local areas, the estimated slope parameter from the unadjusted analysis was 0.08 with standard error 0.15.
Figure 4. Plot of standardized values of log $X_4$ versus expected normal order statistics for 300 measurements of residential radon concentration using data from the Southwest England case-control study.

Figure 5. Plot of $E(X_5 | X_4)$ versus $X_4$ in Bq m$^{-3}$ for data based on the Southwest England case-control study. The circles correspond to the 300 direct measurements that were available. The five triangles correspond to the geometric mean values of the direct measurements within the five local areas and were used to estimate values for the 92 observations where no measurement was available. The line $E(X_5 | X_4) = X_4$ is also shown for comparison.

The substantial increase in the estimated slope that occurs when the amended likelihood is used is in line with what one would expect from the simulation results in Section 5. The effect of adjustment for measurement error is also illustrated in Figure 5, where it can be seen that for measurements above around 200 Bq m$^{-3}$ the expected value of the true radon concentration ($X_i$) is substantially less than the measured value. By contrast, estimated values are close to their expected value for all five local areas.
The above example considers the application of the methods developed in this paper to a simplified subset of data from the radon case-control study. However, to provide a practical solution to the problem discussed in Section 1, two simple generalizations are necessary. In reality the quantity of interest is the average radon concentration in an individual’s home during, say, the past 30 years, which, for individuals who have moved during this period, is a weighted average of several values, each subject to one or other type of error. In this situation estimation of missing values is essential unless the analysis is confined strictly to individuals with a complete exposure history. In addition, it would not in practice be possible to match for all the relevant covariates in the design of the study and so they would have to be included explicitly in the analysis. Equation (20) can readily be extended to cover both these generalizations. If the additional covariates can be expressed in terms of $k$ categorical variables, $x_i$ in (20) can be replaced by $z_i + \sum \eta_k z_k$, and $n^* X_{s}^{(o)}$ by $\sum w_j n^{*} X_{j}^{(o)}$, where the $z_k$ are appropriate dummy variables with $\eta_k$ their corresponding regression coefficients, and for each individual the index $j$ runs over all the homes of interest, while $w_j$ is a weight denoting the proportion of the period of interest for which the individual lived in the $j$th home.

The methods described in this paper provide a relatively simple approach to adapting various standard models to take account of random errors in explanatory variables. The main requirement is that the distribution of both the true explanatory variable and the measurement error can be reasonably approximated by a normal distribution; the method does not require that the magnitude of the measurement error is small and, because estimation is based on maximum likelihood, the standard errors of the parameters of interest are obtained directly from the observed matrix of second derivatives. In cases where the explanatory variable is clearly skew, multiplicative errors are often indicated and a log transformation may well achieve both normality and additivity of errors. Furthermore, if appropriate, regression may still be carried out with respect to $X$ rather than log $X$, as illustrated above. By comparison, the method of Rosner et al. requires that either (i) the measurement error is moderate and/or the association between $Y$ and $X$ is not strong, or (ii) the outcome is rare and the distribution of $X$ given its surrogate is normal. Thus, although use of the latter method is not restricted to normally distributed variables, in the absence of normality, some limit is placed on the magnitude of the errors and/or the association between $X$ and $Y$.

An important aspect of the approach presented here is that it may be used for data which is subject not only to classical measurement error but also to errors of imputation, both of which are likely to occur in epidemiological studies. In addition, the method is easy to apply in practice and for the data considered here the estimate of the regression parameter could be easily computed via iterative application of the logistic command in a standard computer package. Because this work was originally motivated by a specific problem, namely the radon study described in Section 1, the method has been illustrated in the context of that example. It should, however, be possible to generalize the approach to other settings and in particular to cases where alternative types of estimates are used for missing values or to cases where the level of variability in repeat measurements differs between individuals.

**ACKNOWLEDGEMENTS**

We acknowledge the role of our colleagues in the case-control study of radon and lung cancer in Southwest England, including Sir Richard Doll, Dr. Paul Silcocks, staff of the Imperial Cancer Research Fund and the staff of the National Radiological Protection Board. We thank the medical and other staff, hospital patients, and members of the public in Devon and Cornwall who
MEASUREMENT ERROR IN REGRESSION MODELS

helped in obtaining the data given in the example in Section 6, and the referees and Dr. Tom Fearn for helpful comments. This work was supported by the Imperial Cancer Research Fund and by grants from the Department of Health, the Department of the Environment and the European Commission, and DRC is grateful for a Leverhulme Fellowship.

REFERENCES
