STATISTICS IN MEDICINE
Statist. Med. 2002; **21**:1089–1101 (DOI: 10.1002/sim.1062)

Estimation of an errors-in-variables regression model when the variances of the measurement errors vary between the observations

S. B. Kulathinal[1,*,†], Kari Kuulasmaa[1] and Dario Gasbarra[2]

[1]*Department of Epidemiology and Health Promotion, National Public Health Institute, Mannerheimintie 166, 00300 Helsinki, Finland*
[2]*Rolf Nevanlinna Institute, University of Helsinki, P.L.4 00014 Helsinki, Finland*

SUMMARY

It is common in the analysis of aggregate data in epidemiology that the variances of the aggregate observations are available. The analysis of such data leads to a measurement error situation, where the known variances of the measurement errors vary between the observations. Assuming multivariate normal distribution for the 'true' observations and normal distributions for the measurement errors, we derive a simple EM algorithm for obtaining maximum likelihood estimates of the parameters of the multivariate normal distributions. The results also facilitate the estimation of regression parameters between the variables as well as the 'true' values of the observations. The approach is applied to re-estimate recent results of the WHO MONICA Project on cardiovascular disease and its risk factors, where the original estimation of the regression coefficients did not adjust for the regression attenuation caused by the measurement errors. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: bivariate normal distribution; EM algorithm; maximum likelihood estimation; measurement errors; MONICA; regression model

1. INTRODUCTION

The estimation of the parameters of a linear regression model where the variables were measured with error have been studied extensively over the last decade. We refer to references [1, 2] and references therein. In Chapter 1 of reference [1], the moment estimators of regression parameters, when the explanatory variable is assumed to be normally distributed but observed in the presence of measurement error, are given. Further, the measurement error variances are assumed to be equal and known for all the observations. In reference [3], approximate maximum likelihood estimators for the regression parameters, based on the conditional

*Correspondence to: S. B. Kulathinal, Department of Epidemiology and Health Promotion, National Public Health Institute, Mannerheimintie 166, 00300 Helsinki, Finland
†E-mail: sangita.kulathinal@ktl.fi

Received October 2000
Accepted July 2001

distribution of the observed variables given the true variables, in the presence of measurement errors in both variables are given. The measurement error variances are allowed to vary with observations and the true variables are considered to be fixed but unknown quantities. Here, we will consider the case where both the true variables are random.

More specifically, we will assume that a random sample of size n is available on a pair of random variables of interest, (x_i, y_i), which has a bivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ. Each of these variables are observed with measurement error, the variance of which is known but can vary with i. Let (X_i, Y_i) be the corresponding variables obtained by adding the measurement errors to the original true variables. The data consist of (X_i, Y_i), $i = 1, 2, \ldots, n$. We propose using the EM algorithm to obtain maximum likelihood estimates of the parameters to allow for measurement errors. Other quantities, such as correlation and regression coefficients, can easily be derived using the estimates of $\boldsymbol{\mu}$ and Σ. As an alternative, we show briefly how the results given in reference [1] can be generalized in the present case and we give the moment estimators for the regression parameters. Asymptotic theory of the moment estimators can be developed using the central limit theorem for independently distributed random variables. It is to be noted that the conditional distribution of (X_i, Y_i) given (x_i, y_i) is similar to the distributional assumptions given on page 2889 of reference [3].

In the last section of the paper, we give examples to illustrate our approach of handling measurement errors. Simulated data, when the variances of the measurement errors do not vary with i, facilitates assessment and confirmation of our approach with the existing methods. In reference [2], the correlation coefficients are estimated assuming a model with varying variances of measurement error, using a concentrated maximum likelihood approach. Those results will be compared with ours in the second example.

The motivation for the present paper comes from epidemiological studies using aggregate data, where the sampling variances of the estimators are known. It is usually reasonable to consider the sampling variances as known fixed values in further analyses. A specific example of such a study is the WHO MONICA Project. It is a population-based monitoring study of cardiovascular diseases and their risk factors, involving about 40 populations in 21 countries. One of its main objectives was assessment of the extent to which the incidence changes in the population can be explained by changes in the population mean values of the major risk factors. The project has analysed the data by fitting a linear regression model using an iterative reweighting approach [4]. The regression error term was weighted according to the variances of the trend estimates from each population. However, as mentioned in the paper, the approach does not adjust the regression coefficients for the expected attenuation of regression due to the measurement error in the explanatory variable. Therefore, we reanalyse these data using our approach in the third example.

Although the theory is developed for the bivariate normal distribution, it has direct generalization to the multivariate normal distribution and hence to the multiple linear regression model.

2. MODEL AND EM ALGORITHM

We will describe the bivariate normal model in the general set-up first and then show how the EM algorithm works in the presence of measurement errors.

2.1. Bivariate normal model (BVNM)

Let $\boldsymbol{u}_i = (x_i, y_i)^\top$ be independent and identically distributed variables of interest, $i = 1, 2, \ldots, n$, and assume that each of them has bivariate normal distribution with mean $\boldsymbol{\mu}$ and variance-covariance matrix Σ. These variables are observed in the presence of measurement errors. Let $\boldsymbol{v}_i = (X_i, Y_i)^\top$ denote the observations where $\boldsymbol{v}_i = \boldsymbol{u}_i + \boldsymbol{\eta}_i$, and $\boldsymbol{\eta}_i = (\eta_{xi}, \eta_{yi})^\top$ are the measurement errors for variables x_i and y_i. We assume that η_{xi} and η_{yi} are independent of each other and are normally distributed with mean zero and known variances τ_{xi} and τ_{yi}, respectively. Also, $\boldsymbol{\eta}_i$ and $\boldsymbol{u}_i$ are assumed to be independent of each other. These assumptions imply that $\boldsymbol{v}_i$ has the bivariate normal distribution with mean $\boldsymbol{\mu}$ and variance-covariance matrix $S_i = \Sigma + \tau_i$, where τ_i is a 2×2 diagonal matrix with diagonal elements (τ_{xi}, τ_{yi}), $i = 1, 2, \ldots, n$. The covariance between $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ is Σ. Let $C = (\boldsymbol{u}_i, i = 1, 2, \ldots, n)$ and data $= (\boldsymbol{v}_i, i = 1, 2, \ldots, n)$ denote the true variables and the observed variables, respectively.

Our interest is in obtaining maximum likelihood estimators for $\boldsymbol{\mu}$ and Σ given the data. The log-likelihood contribution from $\boldsymbol{v}_i$, $i = 1, 2 \ldots, n$ is

$$l_i(\boldsymbol{\mu}, \Sigma; \boldsymbol{v}_i) = -\log 2\pi - \tfrac{1}{2} \log |S_i| - \tfrac{1}{2} (\boldsymbol{v}_i - \boldsymbol{\mu})^\top S_i^{-1} (\boldsymbol{v}_i - \boldsymbol{\mu}) \tag{1}$$

and the log-likelihood based on the data is

$$\begin{aligned} l(\boldsymbol{\mu}, \Sigma; \text{data}) &= \sum_{i=1}^{n} l_i(\mu, \Sigma; \boldsymbol{v}_i) \\ &= -n \log 2\pi - \frac{1}{2} \sum_{i=1}^{n} \log |S_i| - \frac{1}{2} \sum_{i=1}^{n} (\boldsymbol{v}_i - \boldsymbol{\mu})^\top S_i^{-1} (\boldsymbol{v}_i - \boldsymbol{\mu}) \end{aligned} \tag{2}$$

We need to maximize the above log-likelihood with respect to $\boldsymbol{\mu}$ and Σ. Even though the likelihood has nice form, numerical techniques are required to obtain $\boldsymbol{\mu}$ and Σ which will maximize $l(\boldsymbol{\mu}, \Sigma; \text{data})$ because of τ_i. We adopt the EM algorithm to obtain the maximum likelihood estimators.

Before we describe the EM algorithm, we will see how the moment estimators given in Chapter 1 of reference [1] could be generalized in the present situation. Rewriting the above bivariate normal model in terms of the model described in reference [1], the conditional distribution of y_i given x_i is normal with mean $\alpha + \beta x_i$ and variance σ^2, and x_i is normally distributed with mean μ_x and variance σ_x^2. This implies that (X_i, Y_i) has the bivariate normal distribution with mean $(\mu_x, \alpha + \beta\mu_x)$ and variance-covariance matrix

$$\begin{bmatrix} \sigma_x^2 + \tau_{xi} & \beta\sigma_x^2 \\ \beta\sigma_x^2 & \sigma^2 + \beta^2\sigma_x^2 + \tau_{yi} \end{bmatrix}$$

It is easy to see that the distribution of $(X_i - \bar{X}, Y_i - \bar{Y})$ is bivariate normal with mean zero and variance-covariance matrix

$$\begin{bmatrix} \frac{(n-1)}{n}\sigma_x^2 + \frac{(n-2)}{n}\tau_{xi} + \frac{1}{n}\bar{\tau}_x & \frac{(n-1)}{n}\beta\sigma_x^2 \\ \frac{(n-1)}{n}\beta\sigma_x^2 & \frac{(n-1)}{n}\beta^2\sigma_x^2 + \frac{(n-1)}{n}\sigma^2 + \frac{(n-2)}{n}\tau_{yi} + \frac{1}{n}\bar{\tau}_y \end{bmatrix}$$

where $\bar{\tau}_x$ and $\bar{\tau}_y$ are averages of τ_{xi}'s and τ_{yi}'s, respectively.

The moment estimators of $(\mu_x, \alpha, \beta, \sigma_x^2, \sigma^2)$ can be obtained by equating sample moments with the population moments. Let m_{XY}, m_{XX} and m_{YY} denote the sample variance-covariances obtained using (X_i, Y_i), $i = 1, 2, \ldots, n$, for example $m_{XY} = [1/(n-1)] \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$ and so on. The moment estimators are $\hat{\mu}_x = \bar{X}$, $\hat{\sigma}_x^2 = m_{XX} - \bar{\tau}_x$, $\hat{\beta} = m_{XY}/(m_{XX} - \bar{\tau}_x)$, $\hat{\alpha} = \bar{Y} - \beta\bar{X}$, $\hat{\sigma}^2 = m_{YY} - \beta m_{XY} - \bar{\tau}_y$, provided this difference is positive. An estimator of the correlation coefficient between (x, y) is $\hat{\beta}\hat{\sigma}_x^2 / \sqrt{\{\hat{\sigma}_x^2(\hat{\sigma}^2 + \hat{\beta}^2\hat{\sigma}_x^2)\}}$. Note that the common measurement error variance appearing in the moment estimators of reference [1] is replaced by the average of measurement error variances over the sample. These moment estimators will reduce to the one given in reference [1] in the case of common measurement error variances for x observations and no measurement error in y observations.

2.2. EM algorithm

The log-likelihood of C is

$$l_0(\boldsymbol{\mu}, \Sigma; C) = -n \log 2\pi - \frac{n}{2} \log |\Sigma| - \frac{1}{2} \sum_{i=1}^{n} (\boldsymbol{u}_i - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{u}_i - \boldsymbol{\mu}) \tag{3}$$

The E-step of the EM algorithm involves expectation of $l_0(\boldsymbol{\mu}, \Sigma; C)$ given $(\text{data}, \boldsymbol{\mu}_0, \Sigma_0)$ and the M-step involves maximization of this expectation with respect to $(\boldsymbol{\mu}, \Sigma)$ (see reference [5]). The conditional distribution of $\boldsymbol{u}_i$ given $(\text{data}, \boldsymbol{\mu}_0, \Sigma_0)$, P_i is the bivariate normal distribution with mean

$$\boldsymbol{m}_i = \boldsymbol{\mu}_0 + \Sigma_0 S_{i0}^{-1} (\boldsymbol{v}_i - \boldsymbol{\mu}_0) \tag{4}$$

and the variance-covariance matrix

$$\Lambda_i = \Sigma_0 - \Sigma_0 S_{i0}^{-1} \Sigma_0 \tag{5}$$

The maximization of the expectation of log-likelihood given $(\text{data}, \boldsymbol{\mu}_0, \Sigma_0)$ with respect to $(\boldsymbol{\mu}, \Sigma)$ is essentially minimizing

$$\text{constant} + \frac{n}{2} \log |\Sigma| + \frac{1}{2} \sum_{i=1}^{n} E_{P_i} (\boldsymbol{u}_i - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{u}_i - \boldsymbol{\mu})$$

Note that this expression which needs to be minimized is the sum of the Kullback–Leibler informations

$$\sum_{i=1}^{n} I(P_i | Q) = \sum_{i=1}^{n} E_{P_i} \log \frac{\mathrm{d}P_i}{\mathrm{d}Q}$$

where Q has a bivariate normal distribution with mean $\boldsymbol{\mu}$ and variance-covariance matrix Σ. The following theorem shows that such $\boldsymbol{\mu}$ and Σ are available in the closed form. We will prove a general result for d dimensions:

Theorem 2.1

Let $P_1, P_2, \ldots, P_n$ be d-dimensional multivariate normal distributions with respective mean and covariance matrices $(\boldsymbol{m}_i, \Lambda_i)$. Then a normal distribution Q which minimizes the sum of the

Kullback–Leibler informations $\sum_{i=1}^{n} I(P_i|Q)$ has mean and variance-covariance matrix

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{m}_i \tag{6}$$

$$\hat{\Sigma} = \frac{1}{n}\left[\hat{M}\hat{M}^{\top} + \sum_{i=1}^{n} \Lambda_i\right] \tag{7}$$

respectively, where $\hat{M}_i = \boldsymbol{m}_i - \hat{\boldsymbol{\mu}}$ and $\hat{M} = (\hat{M}_1, \ldots, \hat{M}_n)$.

Proof
Let $P_1, \ldots, P_n$ be d-dimensional multivariate normal distributions with mean $\boldsymbol{m}_i$ and covariance matrices Λ_i. Then the question is to find a normal distribution Q parameterized by its mean $\boldsymbol{\mu}$ and covariance matrix Σ such that the following expression is minimized:

$$\sum_{i=1}^{n} I(P_i|Q) = \sum_{i=1}^{n} E_{P_i} \log \frac{\mathrm{d}P_i}{\mathrm{d}Q} = \text{constant} + \frac{n}{2}\log|\Sigma| + \frac{1}{2}\sum_{i=1}^{n} E_{P_i}(X_i - \boldsymbol{\mu})^{\top}\Sigma^{-1}(X_i - \boldsymbol{\mu}) \tag{8}$$

where we have introduced the Kullback–Leibler information $I(P_i|Q)$ and X_i has distribution P_i. Applying the standard techniques of multivariate analysis (see reference [6]), the expression (8) to be minimized over $\boldsymbol{\mu}$ and Σ can be reformulated without a constant term as

$$\begin{aligned}
&-\frac{n}{2}\log|\Sigma^{-1}| + \frac{1}{2}\sum_{i=1}^{n}\{\mathrm{Tr}(\Lambda_i\Sigma^{-1}) + (\boldsymbol{\mu} - \boldsymbol{m}_i)^{\top}\Sigma^{-1}(\boldsymbol{\mu} - \boldsymbol{m}_i)\} \\
&= \frac{1}{2}\mathrm{Tr}\left(\Sigma^{-1}\sum_{i=1}^{n}\Lambda_i\right) - \frac{n}{2}\log|\Sigma^{-1}| + \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{\mu} - \boldsymbol{m}_i)^{\top}\Sigma^{-1}(\boldsymbol{\mu} - \boldsymbol{m}_i) \\
&= \frac{1}{2}\mathrm{Tr}\left(\Sigma^{-1}\sum_{i=1}^{n}\Lambda_i\right) - \frac{n}{2}\log|\Sigma^{-1}| + \frac{1}{2}\mathrm{Tr}(M^{\top}\Sigma^{-1}M) \\
&= \frac{1}{2}\mathrm{Tr}\left(\Sigma^{-1}\sum_{i=1}^{n}\Lambda_i\right) - \frac{n}{2}\log|\Sigma^{-1}| + \frac{1}{2}\mathrm{Tr}(\Sigma^{-1}MM^{\top}) \\
&= \frac{1}{2}\mathrm{Tr}\left(\Sigma^{-1}\left[MM^{\top} + \sum_{i=1}^{n}\Lambda_i\right]\right) - \frac{n}{2}\log|\Sigma^{-1}|
\end{aligned}$$

where $M = (M_1, \ldots, M_n)$ is a $d \times n$ matrix with $M_i = \boldsymbol{m}_i - \boldsymbol{\mu}$. Hence the theorem.

Note that $\hat{\mu}$ is the average of $\boldsymbol{m}_i$'*s* and covariance matrix $\hat{\Sigma}$ is the average of Λ_i'*s* plus the empirical covariance matrix of $\boldsymbol{m}_i$'*s*.

We now describe the EM algorithm:

Step 1. Start with some initial values $\boldsymbol{\mu}_0$ and Σ_0.
Step 2. Evaluate $\boldsymbol{m}_i$ and Λ_i given above in equations (4) and (5) using $\boldsymbol{\mu}_0$ and Σ_0 for each i.

Step 3. Update $\boldsymbol{\mu}$ and Σ by

$$\boldsymbol{\mu}_1 = \bar{m} = \frac{1}{n}\sum_{i=1}^{n} m_i \tag{9}$$

$$\Sigma_1 = \frac{1}{n}\sum_{i=1}^{n} \Lambda_i + \frac{1}{n}\sum_{i=1}^{n} (m_i - \bar{m})(m_i - \bar{m})^\top \tag{10}$$

Step 4. Repeat steps 2 and 3 using $\boldsymbol{\mu}_1$ and Σ_1 and obtain $\boldsymbol{\mu}_2$ and Σ_2. Continue until the sequences $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_i, \ldots\}$ and $\{\Sigma_1, \Sigma_2, \ldots, \Sigma_i, \ldots\}$ converge.

Let $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ be the maximum likelihood estimators obtained using the EM algorithm. The variance-covariance matrix of these estimators can be obtained by evaluating their information matrix. It is important to note that the observations $\boldsymbol{v}_i$'*s* are independent but not identically distributed. Hence, the above estimators need not be consistent in all the situations. The general condition under which consistency will hold is that the average of the overall information matrices tends to a positive definite matrix as the number of observations goes to infinity. This in turn puts some condition on X_i, $i = 1, 2, \ldots, n$ and also on measurement error variances. We will see that the expression for overall information is quite complex. We refer to Section 6 of Chapter 6 of reference [7] for the detailed discussion.

The regression line of y on x defined as $y = \alpha + \beta x + \varepsilon$ can be obtained easily. The maximum likelihood estimators of the correlation coefficient (ρ), intercept (α) and the slope (β) are given by

$$\hat{\rho} = \frac{\hat{\Sigma}_{12}}{\sqrt{(\hat{\Sigma}_{11}\hat{\Sigma}_{22})}} \tag{11}$$

$$\hat{\alpha} = \hat{\mu}_2 - \hat{\beta}\hat{\mu}_1 \tag{12}$$

$$\hat{\beta} = \frac{\hat{\Sigma}_{12}}{\hat{\Sigma}_{11}} \tag{13}$$

These parameters are continuous functions of $\boldsymbol{\mu}$ and Σ and using the standard theory of maximum likelihood estimators, the asymptotic distribution can be obtained. The information matrix $I_i(\theta_i)$ for $\theta_i = (\Sigma_{11} + \tau_{xi}, \Sigma_{22} + \tau_{yi}, \rho_i)$ obtained using $\boldsymbol{v}_i$ is a 3×3 matrix with elements

$$I_i(\theta_i) = \begin{bmatrix} \frac{2-\rho_i^2}{4(1-\rho_i^2)(\Sigma_{11}+\tau_{xi})^2} & \frac{-\rho_i^2}{4(1-\rho_i^2)(\Sigma_{11}+\tau_{xi})(\Sigma_{22}+\tau_{yi})} & \frac{-\rho_i}{2(1-\rho_i^2)(\Sigma_{11}+\tau_{xi})} \\ & \frac{2-\rho_i^2}{4(1-\rho_i^2)(\Sigma_{22}+\tau_{yi})^2} & \frac{-\rho_i}{2(1-\rho_i^2)(\Sigma_{22}+\tau_{yi})} \\ & & \frac{1+\rho_i^2}{(1-\rho_i^2)^2} \end{bmatrix} \tag{14}$$

where ρ_i is the correlation coefficient between (X_i, Y_i) and is given by

$$\rho_i = \rho \frac{\sqrt{(\Sigma_{11}\Sigma_{22})}}{\sqrt{\{(\Sigma_{11} + \tau_{xi})(\Sigma_{22} + \tau_{yi})\}}}$$

Note that τ_{xi} and τ_{yi} are constants. The information about the parameters $(\Sigma_{11}, \Sigma_{22}, \rho)$ available in the distribution of v_i can then be easily obtained by using $I_i(\theta_i)$ and the Jacobian of transformation J_i which is a 3×3 matrix with elements

$$J_i = \begin{bmatrix} 1 & 0 & \frac{\rho\sqrt{\Sigma_{22}}}{\sqrt{\{2(\Sigma_{11}+\tau_{xi})(\Sigma_{22}+\tau_{yi})\}}}\left[\frac{1}{\sqrt{\Sigma_{11}}} - \frac{\sqrt{\Sigma_{11}}}{(\Sigma_{11}+\tau_{xi})}\right] \\ 0 & 1 & \frac{\rho\sqrt{\Sigma_{11}}}{\sqrt{\{2(\Sigma_{11}+\tau_{xi})(\Sigma_{22}+\tau_{yi})\}}}\left[\frac{1}{\sqrt{\Sigma_{22}}} - \frac{\sqrt{\Sigma_{22}}}{(\Sigma_{22}+\tau_{yi})}\right] \\ 0 & 0 & \frac{\sqrt{(\Sigma_{11}\Sigma_{22})}}{\sqrt{\{(\Sigma_{11}+\tau_{xi})(\Sigma_{22}+\tau_{yi})\}}} \end{bmatrix}$$

The information matrix $I_i(\Sigma_{11}, \Sigma_{22}, \rho)$ is given by

$$I_i(\Sigma_{11}, \Sigma_{22}, \rho) = J_i I_i(\theta_i) J_i' \tag{15}$$

The overall information about Σ is the sum of the information available from each v_i and is denoted by $I(\Sigma_{11}, \Sigma_{22}, \rho) = \sum_{i=1}^{n} I_i(\Sigma_{11}, \Sigma_{22}, \rho)$. The variance-covariance matrix of the maximum likelihood estimators of the parameters is the inverse of the information matrix. The regression coefficient is a function of elements of Σ and using the Jacobian of transformation from $(\Sigma_{11}, \Sigma_{22}, \rho)$ to β, the information of β can be obtained and hence the variance of the maximum likelihood estimator of β.

From the above results, a naive estimator of the variables of interest, $\boldsymbol{u}_i$ can be given by the conditional expectation of $\boldsymbol{u}_i$ given $\boldsymbol{v}_i$ and replacing $\boldsymbol{\mu}$ and Σ by their maximum likelihood estimates and the standard error of this estimate by the conditional standard deviation of $\boldsymbol{u}_i$ given $\boldsymbol{v}_i$. The following expression gives an estimate of $\boldsymbol{u}_i$ along with the square of its standard error

$$\hat{\boldsymbol{u}}_i = \hat{\boldsymbol{\mu}} + \hat{\Sigma}(\hat{\Sigma} + \tau_i)^{-1}(\boldsymbol{v}_i - \hat{\boldsymbol{\mu}}) \tag{16}$$

$$\mathrm{SE}^2(\hat{\boldsymbol{u}}_i) = \hat{\Sigma} - \hat{\Sigma}(\hat{\Sigma} + \tau_i)^{-1}\hat{\Sigma} \tag{17}$$

$\hat{\boldsymbol{u}}_i$ is the best linear predictor given the data but it is biased towards $\boldsymbol{\mu}$. The amount of bias depends on the size of the measurement error variances. In the next section, we will apply the algorithm to several data sets.

3. ILLUSTRATIONS

We illustrate the EM algorithm using simulated data in the absence and presence of measurement errors. To allow comparison with the existing standard methods, we assume that the measurement error variance does not vary with i. Secondly, we evaluate the maximum likelihood estimates of correlation coefficients using the data given in reference [2] on the estimates of the rates of coronary heart disease and estimated changes in risk factors for coronary heart disease and compare the results. We refer to reference [2] for details regarding the data. Finally, we apply EM algorithm to the data from WHO MONICA Project described in Section 1. MATLAB and SAS codes were written to carry out the analyses. In all the examples, the EM algorithm converged in a fairly small number of iterations.

Table I. Estimates of regression parameters in various cases. Case 1: $\tau_{xi} = \tau_{yi} = 0$; case 2: $\tau_{xi} = 0, \tau_{yi} = 1$; case 3: $\tau_{xi} = 1, \tau_{yi} = 0$; case 4: $\tau_{xi} = \tau_{yi} = 1$.

Parameter	Method	Case 1	Case 2	Case 3	Case 4
α	WME	0.32	0.32	−0.01	−0.01
	BVNM	0.32	0.32	0.35	0.34
	MOMENT	0.32	0.32	0.34	0.34
	True	0.33	0.33	0.33	0.33
β	WME	0.66	0.65	0.49	0.49
	BVNM	0.66	0.65	0.64	0.64
	MOMENT	0.66	0.65	0.64	0.64
	True	0.67	0.67	0.67	0.67
σ^2	WME	2.77	3.81	3.12	4.16
	BVNM	2.77	2.81	2.80	2.85
	MOMENT	2.77	2.85	2.84	2.89
	True	2.67	2.67	2.67	2.67

We will use WME to indicate the ordinary least squares estimates obtained ignoring measurement errors, BVNM to denote the results obtained using the above EM algorithm under bivariate normality and MOMENT to denote moment estimates of the parameters accounting for measurement errors.

3.1. Example 3.1: Simulated data

It is assumed that (x, y) have bivariate normal distribution with $\boldsymbol{\mu} = (-2, -1)'$, $\Sigma_{11} = 3$, $\Sigma_{22} = 4$, and $\Sigma_{12} = 2$. We generated random samples of size 100 and applied the EM algorithm to find maximum likelihood estimates of $\boldsymbol{\mu}$ and Σ. In the following cases, where it is assumed that the measurement error variance does not vary with i, the maximum likelihood estimate of $\boldsymbol{\mu}$ is simply the sample mean and that of Σ is the sample variance-covariance matrix minus τ. Then the coefficient of attenuation, defined as the ratio of the variance of x to the variance of X is $\Sigma_{11}/(\Sigma_{11} + \tau_x)$, which is also equal to the ratio of the regression coefficient of Y on X to the regression coefficient of y on x (see reference [1] for details). An estimate of this coefficient is obtained by using the respective estimates of the unknown quantities. We only report the regression parameters here.

We consider four different cases in Table I. Case 1 refers to the situation of no measurement error and hence the variances of measurement errors were taken as 0. In case 2, there is no measurement error in x observations but the variance of measurement error in y observations was taken as 1. Similarly, case 3 allows measurement error in x observations with variance 1 and no measurement error in y while case 4 allows measurement error in both x and y observations with variance 1.

The estimate of the regression coefficient is the same under all the methods in cases 1 and 2 but when the EM algorithm is used, the estimate of σ^2 reduces by an amount of measurement error variance of y in case 2. Under this situation, it is easy to verify that $\sigma^2(\text{MOMENT}) = (100/99)(\sigma^2(\text{BVNM}) + 1) - 1 = (100/99)\sigma^2(\text{WME}) - 1$. A clear increase in the β is observed in cases 3 and 4 when measurement errors were taken into account but there is very little difference between the estimate under BVNM and MOMENT. The true

Table II. Estimates of correlation coefficients using data in reference [2].

CHD with	corr(WME)	corr(Dear)	corr(BVNM)	CI(Dear)	CI(BVNM)
Cholesterol	0.04	0.53	0.53	(−0.25, 0.93)	(−0.03, 1.0)
MRFIT	0.67	0.96	0.96	(0.49,1)	(0.66, 1.0)
SBP	0.39	0.27	0.27	(−0.6, 0.88)	(−0.54, 1.0)
Smoking	0.53	1.0	0.99	(0.67, 1.0)	(0.61, 1.0)

coefficients of attenuation are 1, 1, 0.75 and 0.75 and estimates under BVNM are 1, 1, 0.76 and 0.76 for cases 1, 2, 3 and 4, respectively.

The increase in the estimate of β when measurement errors were taken into account and small differences between the true and estimated values suggest that the maximum likelihood and moment estimators are satisfactory. The empirical probabilities that the true β is contained in a 95 per cent confidence interval under BVNM, obtained through 1000 repetitions, were between 91 per cent to 93 per cent in all the four cases.

3.2. Example 3.2: Data used in reference [2]

In Table II, corr(Dear) refers to the correlation coefficients estimated in reference [2], in the case of varying variances of measurement error, using a concentrated maximum likelihood approach. Table II gives the estimates of the correlation coefficients along with their 95 per cent confidence intervals (CI) between trends in coronary heart disease (CHD) rates and population trends in known risk factors: cholesterol; systolic blood pressure (SBP); prevalence of smoking, and a linear combination of these (MRFIT).

Whenever the upper confidence limit under BVNM was greater than one, it was replaced by 1. The estimates of correlation coefficients obtained without taking into account the measurement errors are smaller compared to the estimates obtained using the EM algorithm and using the approach of reference [2]. There is no difference between the estimates of the correlation coefficients obtained by later two methods. It is apparent that the EM algorithm is simpler to apply.

3.3. Example 3.3: MONICA data, reference [4]

The WHO MONICA Project, which includes a large number of populations, is a monitoring study of cardiovascular diseases. The data which we will be analysing are the estimates of trends in coronary event rates and trends in the mean value of risk scores for men and women in each population. The risk score was defined as a linear combination of smoking status, systolic blood pressure, body mass index and total cholesterol. Its coefficients were derived from a follow-up study using the proportional hazards model. We refer to reference [4] for details. The sampling errors of the trend estimates are considered as measurement errors. These vary from population to population and we will consider the sampling variances determined from the estimate of the trends as fixed values.

Table III gives estimates of regression parameters and error variance, using the EM algorithm and also without assuming measurement errors, for men and women. It also gives the standard error of the estimator of regression coefficient β, SE(β) and 95 per cent confidence intervals for β, CI(β). The results of the analysis reported in reference [4] are referred to as K(2000).

Table III. Estimates of regression parameters using MONICA data.

	α	β	σ^2	SE(β)	CI(β)	R^2 (%)
(*a*) *Men*						
WME	−1.88	0.31	7.11	0.20	(−0.08, 0.71)	6
K(2000)	−2.15	0.43	4.49	0.22	(−0.01, 0.87)	19
BVNM	−2.08	0.47	4.89	0.23	(0.02, 0.92)	16
(*b*) *Women*						
WME	−0.47	0.51	17.20	0.34	(−0.16, 1.18)	6
K(2000)	−0.21	0.57	10.59	0.33	(−0.10, 1.24)	12
BVNM	0.03	0.68	11.08	0.24	(0.21, 1.15)	13

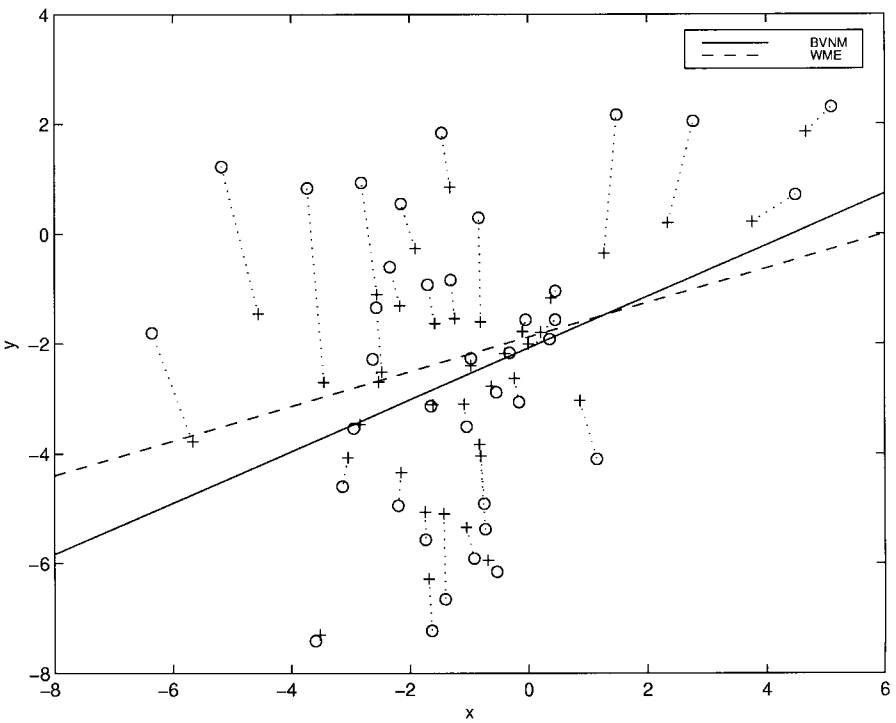

Figure 1. Fitted regression lines using MONICA data for men: ○ observed; + estimated under BVNM.

There is a noticeable increase in the estimates of β and decrease in the error variance (σ^2) for both men and women when measurement errors were taken into account. Under the BVNM, the estimates of $\boldsymbol{\mu}$ and ρ are $(-1.09, -2.59)$ and 0.4033 for men and are $(-2.07, -1.37)$ and 0.362 for women.

The parameter of main interest in reference [4] was the percentage of variation in the trend in event rates explained by the trend in risk factors. In the last column, R^2 (%) of Table III, this percentage is estimated by the square of the correlation coefficient of the observed values (WME), by a statistic defined in reference [4] and by the square of the estimate of the

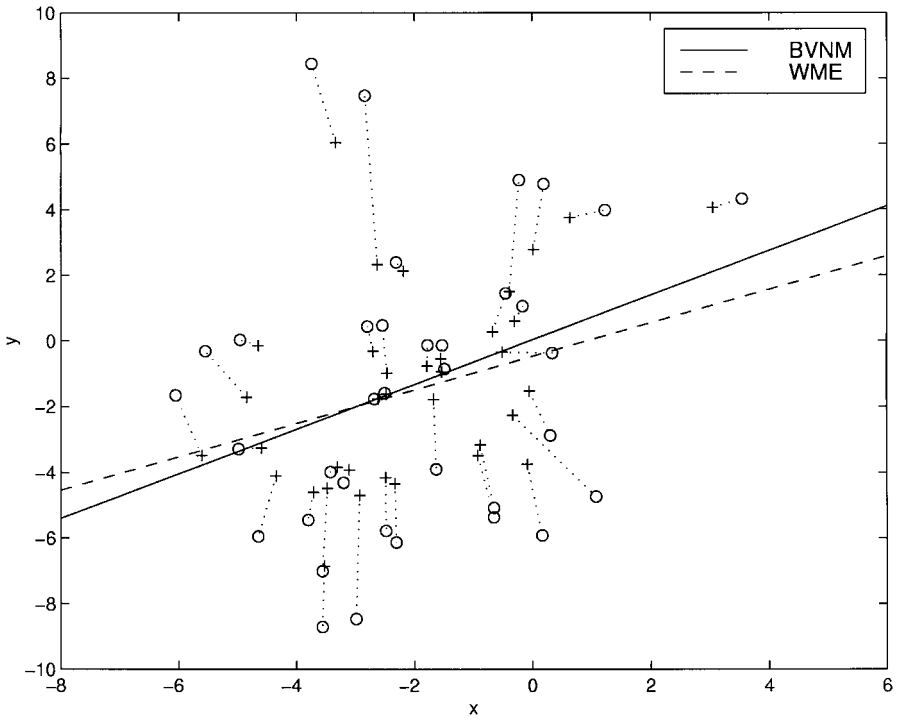

Figure 2. Fitted regression lines using MONICA data for women: ○ observed; + estimated under BVNM.

correlation coefficient between the true variables (x, y) using the EM algorithm. As expected, R^2(WME) also estimates the square of the correlation coefficient between the true variables but without taking into account the measurement error and is much smaller compared to R^2(BVNM). The estimate of β under BVNM is higher than that under K(2000) but there is little effect on R^2.

Figures 1 and 2 compare the regression lines obtained using ordinary least squares and using bivariate normal model for men and women, respectively. Also, the observed values (X_i, Y_i) and the estimated values of true (x_i, y_i) for $i = 1, 2, \ldots, n$ are given. The figures clearly illustrate the nature of the effect of measurement errors. The shift in the estimated (x_i, y_i) depends upon the size of the measurement error variances corresponding to both the variables. Note, however, that the bias in the estimated y observations towards the mean may be substantial because of large measurement error variances in the y observations.

4. DISCUSSION

The EM algorithm provides a convenient approach to handling the measurement errors of unequal variances in x and y variables, and the existence of a closed form solution makes it easy to program. The number of iterations required for the convergence of the EM algorithm was small in all the above examples. The assumption of normality of the variables is reasonable

in a variety of applications. Our approach, however, does not cover the situations when this assumption is not valid.

Using the maximum likelihood estimates of the mean and covariance matrix, one can obtain maximum likelihood estimates of the regression parameters, which are asymptotically unbiased. One can also get estimates of the true values of the variables which are the best linear predictors but which are biased towards the mean. This means that the estimate of the true variable is usually closer to the mean than the true variable. The bias is severe if the measurement error variance is high. This can be seen from Figures 1 and 2 where, especially, the estimate of y is biased towards its mean because the measurement error variances were large (ranging from 0.14 to 17.33). When the estimate of ρ is close to 1, the estimates of true values will fall on the fitted regression line since it is the best predictor of y given x.

Example 3.2 shows that the EM algorithm gives correlation coefficients similar to those given in reference [2], which are based on a concentrated maximum likelihood approach. However, the EM algorithm uses the full likelihood and is easier to use. In this example, most of the measurement errors were large compared to the scatter of the observed values. As a consequence, the attenuation of most of the correlation coefficients also were large, when the measurement errors were not taken into account.

In the MONICA data of example 3.3, the measurement errors were large and varied substantially between the observations. In this example, the regression attenuation was partly corrected in the analysis in reference [4], which weighted the observations properly, but did not involve direct adjustment for the attenuation. In their approach, it was not possible to estimate the traditional R^2 for the true values which defines the percentage of variation explained by the model, but they used an intuitive statistic similar to the one proposed by reference [8]. The EM algorithm enables one to estimate the traditional R^2 for the true values. It was interesting to see that in the MONICA data, the estimate of the percentage explained by the model using the EM algorithm did not differ much from the estimates given in reference [4].

The data for example 3.3 were estimates of trends in different populations. Although the trend estimates incorporate both the statistical error of the estimation of the trends and the imprecision of the data used for the estimation, it can be considered as measurement error of the data used for the analysis of example 3.3. Yet another source of the measurement error is the possible bias in the data used for estimation of the trends. This component of the measurement error was ignored because it could not be quantified. In this regard, the measurement error variance considered in example 3.3 is smaller than what it could have been.

This paper introduces an expected method for epidemiology but a similar problem may also arise in other fields. The proposed method has a direct generalization to the multivariate normal distribution. It is based on a conditional argument of given measurement error variances. The assumption of known measurement error variances seems reasonable, since it simplifies the analysis considerably, even though only their estimates are usually available in practice. Extending the proposed model to incorporate sampling variations in the estimates of the measurement error variances is a problem for further research.

ACKNOWLEDGEMENTS

The authors thank Professor Annette Dobson (University of Queensland) and Patrick McElduff (University of Newcastle) for helpful comments on this paper.

REFERENCES

1. Fuller WA. *Measurement Error Models*. Wiley: New York, 1987.
2. Dear KBG, Puterman ML, Dobson AJ. Estimating correlations from epidemiological data in the presence of measurement error. *Statistics in Medicine* 1997; **16**:2177–2189.
3. Walter SD. Variation in baseline risk as an explanation of heterogeneity in meta-analysis. *Statistics in Medicine* 1997; **16**:2883–2900.
4. Kuulasmaa K, Tunstall-Pedoe H, Dobson A, Fortmann S, Tolonen H, Evans A, Ferrario M, Tuomilehto J for the WHO MONICA project. Estimation of contribution of changes in classic risk factors to trends in coronary-event rates across the WHO MONICA Project populations. *Lancet* 2000; **355**:675–687.
5. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM-algorithm. *Journal of the Royal Statistical Society*, *Series B* 1977; **39**:1–38.
6. Kshirsagar AM. *Multivariate Analysis*. Marcel Dekker: New York, 1972.
7. Lehmann EL. *Theory of Point Estimation*. Wiley: New York, 1983.
8. Pocock SJ, Cook DG, Beresford SAA. Regression of area mortality rates on explanatory variables: what weighting is appropriate? *Applied Statistics* 1981; **30**:286–295.

