Research Article

Statistics in Medicine

Received 14 January 2010, Accepted 25 June 2010 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4030

Slope estimation in structural line-segment heteroscedastic measurement error models

Michael P. McAssey and Fushing Hsieh*, †

This paper extends the line-segment parametrization of the structural measurement error (ME) model to situations in which the error variance on both variables is not constant over all observations. Under these conditions, we develop a method-of-moments estimate of the slope, and derive its asymptotic variance. We further derive an accurate estimator of the variability of the slope estimate based on sample data in a rather general setting. We perform simulations that validate our results and demonstrate that our estimates are more precise than estimates under a different model when the ME variance is not small. Finally, we illustrate our estimation approach using real data involving heteroscedastic ME, and compare its performance with that of earlier models. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: delta method; heteroscedasticity; measurement error; method of moments; slope estimation

1. Introduction

Consider an observational study in which information on two variables is collected from random samples of n distinct groups within a population. Suppose that a researcher is given only a set of summary statistics on the observed variables for each sample, along with the corresponding sampling errors. He wishes to determine whether there is a significant linear association between the two variables and, if so, to model that association with an accurate slope estimate. The sampling error involved with each variable eliminates the applicability of the simple linear regression approach and calls for the implementation of a measurement error (ME) model. Moreover, the variability of the error on each observation for each variable is likely to be different, due to different sample sizes and other influences, so that a model that accounts for heteroscedastic ME is required.

Heteroscedastic ME models have been developed for estimating the slope in such scenarios. Kulathinal *et al.* apply this approach in [1] to the data collected in the World Health Organization (WHO) Multinational MONItoring of trends and determinants of CArdiovascular disease (MONICA) Project (2000) on cardiovascular disease and its risk factors. Patriota *et al.* also examine these data in [2], as well as astronomical data obtained from the *Chandra* observatory, accounting for the heteroscedastic ME that characterizes each data set. Under this approach, each observation (x_i, y_i) is modeled as

$$x_i = \chi_i + \varepsilon_i, \quad y_i = \varphi_i + v_i \quad \text{with } \varphi_i = \alpha + \beta\chi_i + \gamma_i,$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ and $v_i \sim \mathcal{N}(0, \tau_i^2)$ are the independently distributed heteroscedastic MEs, $\gamma_i \sim \mathcal{N}(0, \zeta^2)$ is the independently distributed equation error, and all three errors are mutually independent. To assure identifiability, it is assumed that the ME variances (σ_i, τ_i), $i = 1, \ldots, n$, are known or can be estimated independently. Under the structural model the χ_i are independent and distributed as $\mathcal{N}(\mu_\chi, \sigma_\chi^2)$, while the φ_i are also independent and distributed as $\mathcal{N}(\mu_\varphi, \sigma_\varphi^2)$. Both maximum-likelihood and method-of-moments estimators of the slope have been derived under the structural heteroscedastic ME model, as presented in [1–3] using this conventional equation error model. However, as will be demonstrated in an analysis of the WHO MONICA data, this model is not robust against misspecification, so that the inclusion of the equation error component makes it susceptible to underestimation of the slope when the unknown data-generation mechanism does not actually have the equation-error structure.

Department of Statistics, University of California at Davis, MSB 4118, 1 Shields Ave., Davis, CA 95616, U.S.A.
**Correspondence to: Fushing Hsieh, Department of Statistics, University of California at Davis, MSB 4232, 1 Shields Ave., Davis, CA 95616, U.S.A.*
†E-mail: fushing@wald.ucdavis.edu

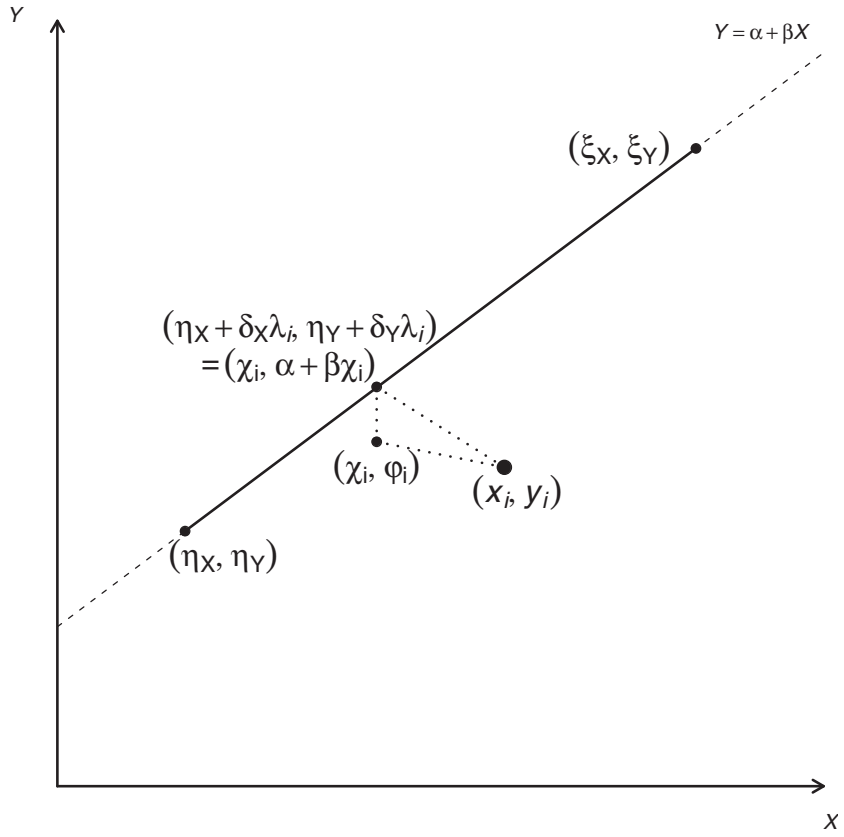

Figure 1. Diagram of the equation-error and line-segment structural models with heteroscedastic measurement error.

In this paper, we present an extension of the structural line-segment model for homoscedastic ME, as introduced by Davidov in [4], to this heteroscedastic ME scenario. This alternative approach omits the equation error component and is symmetric in the two variables. These features make the line-segment model more robust against misspecification, and thus preferable when the data-generation mechanism is completely unknown. We assume that the points $(\mathscr{E}X_i, \mathscr{E}Y_i)$ are randomly distributed on a line segment having latent endpoints at (η_X, η_Y) and (ξ_X, ξ_Y), with $\eta_X \leqslant \xi_X$. Let $\delta_X = \xi_X - \eta_X \geqslant 0$ and $\delta_Y = \xi_Y - \eta_Y$. We then have the model

$$x_i = \eta_X + \lambda_i\delta_X + \varepsilon_i \quad \text{and} \quad y_i = \eta_Y + \lambda_i\delta_Y + v_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\lambda_1, \ldots, \lambda_n$ are independently drawn from a common distribution G and take values in the unit interval, and $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$, while $v_i \sim \mathcal{N}(0, \tau_i^2)$ for each i. We also assume that $\sigma_1, \ldots, \sigma_n$ are independently drawn from a common distribution, and likewise for $\tau_1, \ldots, \tau_n$. Again, to assure identifiability, we assume that these error variances are known for both variables. In practice, the error variance is rarely known, but may be accurately estimated using repeated measurements, or, when the observations are statistics based on samples, by using sampling errors. Let μ_λ and σ_λ^2 denote the mean and variance of λ_i, respectively. Since λ_i is bounded, these and all higher moments must be finite. Finally, assume that λ_i, σ_i and τ_i are mutually independent for all i.

Figure 1 provides a diagram of these two structural models. In this diagram, (x_i, y_i) is the observed data pair, while the linear association between X and Y is represented by the dashed trendline $Y = \alpha + \beta X$, upon which the line segment having endpoints at (η_X, η_Y) and (ξ_X, ξ_Y) rests. The conditional expectation of (x_i, y_i), given χ_i, onto the line/segment is the point $(\chi_i, \alpha + \beta\chi_i)$ under the equation error model. This point is equivalent to the point $(\eta_X + \delta_X\lambda_i, \eta_Y + \delta_Y\lambda_i)$, which is the conditional expectation of (x_i, y_i), given λ_i, under the line-segment model. This point lies on the dotted line between (x_i, y_i) and the line segment from the perspective of the line-segment model. But under the equation error model, it lies on the vertical dotted line through the unobserved point (χ_i, φ_i).

Trivially, the path from the observed pair (x_i, y_i) to the line segment, corresponding to the line-segment model, is always smaller than the path that goes from (x_i, y_i) to (χ_i, φ_i), and then to the line, corresponding to the equation error model. When there is a strong correlation between the two variables, this difference in path lengths is almost negligible when using either model to compute an estimate of the slope, since the errors are small. In such cases, the dominating influence on the variance of the slope estimate is the overall dispersion in X—the larger the spread, the smaller the

variance. Since the equation error is a vertical displacement, its inclusion in the model attenuates the effect of the dispersion in X, giving it an advantage over the line-segment model. However, when the errors become larger, the path difference becomes the dominating influence on the variance of the slope estimate among the two models. Hence, in any scenario where the variances of the MEs are not small, the line-segment model will provide a more precise estimate of the slope. This improvement is demonstrated using simulation studies.

The line-segment model provides an additional benefit, whether or not there is ME or heteroscedasticity. The equation-error model assumes a specific structure for the underlying data-generation mechanism, a structure that may not properly explain the association between the two variables. In this case, points that lie far from the line and toward the horizontal extremes exert an excessive influence on the slope estimate in equation-error models due to the effect of including the equation errors. But in the line-segment model, no such effect occurs. Consequently the influence of such points on the slope estimate is attenuated. Therefore, the line-segment model is more robust against outliers at the horizontal extremes, and is better able to explain the association between the two variables in situations in which the unknown data-generation mechanism does not have the equation-error structure. This benefit will be illustrated in the application of our model to the same WHO MONICA data, which contain influential points that have caused incorrect slope estimates under the equation-error models.

In the next section we derive a method-of-moments estimate of the slope under our extension of the line-segment model, and obtain its asymptotic variance in Section 3 using the delta method. We derive a large-sample estimate of the variance of our slope estimate in Section 4, which may be used in real data analysis. In Section 5 we perform simulation studies which verify the accuracy and precision of our estimates when the data-generation mechanism conforms to the line-segment model, and we show that the precision of our slope estimate is superior in this setting to that of a method-of-moments estimate obtained through the equation-error model, using assorted ranges of the ME variances. In Section 6 we illustrate the application of our slope estimation to real data, and compare our estimates with those derived using several equation-error methods, and using the line-segment model when homoscedastic errors are naïvely assumed. We summarize and discuss our results briefly in Section 7.

2. Point estimation of the slope

Given the structural line-segment model described above, we have

$$\mathrm{Var}(X_i)=\delta_X^2\sigma_\lambda^2+\sigma_i^2,\quad \mathrm{Var}(Y_i)=\delta_Y^2\sigma_\lambda^2+\tau_i^2\quad\text{and}\quad \mathrm{Cov}(X_i,Y_i)=\delta_X\delta_Y\sigma_\lambda^2,$$

so that

$$\mathrm{Var}\left(\frac{X_i}{\sigma_i}\right)=\frac{\delta_X^2\sigma_\lambda^2}{\sigma_i^2}+1\quad\text{and}\quad \mathrm{Var}\left(\frac{Y_i}{\tau_i}\right)=\frac{\delta_Y^2\sigma_\lambda^2}{\tau_i^2}+1.$$

Hence

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}\left(\frac{X_i}{\sigma_i}\right)=\delta_X^2\sigma_\lambda^2\sigma_n^\star+1,\quad \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}\left(\frac{Y_i}{\tau_i}\right)=\delta_Y^2\sigma_\lambda^2\tau_n^\star+1$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{Cov}(X_i,Y_i)=\delta_X\delta_Y\sigma_\lambda^2,$$

where

$$\sigma_n^\star=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sigma_i^2}\quad\text{and}\quad \tau_n^\star=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\tau_i^2}.$$

We then solve for δ_X (which we take to be nonnegative) and δ_Y, and use the sign of the mean covariance to determine the sign of δ_Y, to obtain

$$\delta_X=\sigma_\lambda^{-1}\sqrt{\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}\left(\frac{X_i}{\sigma_i}\right)-1\right]_+\Big/\sigma_n^\star}$$

and

$$\delta_Y = \operatorname{sgn}\left(\frac{1}{n}\sum_{i=1}^{n}\operatorname{Cov}(X_i, Y_i)\right)\sigma_\lambda^{-1}\sqrt{\left[\frac{1}{n}\sum_{i=1}^{n}\operatorname{Var}\left(\frac{Y_i}{\tau_i}\right)-1\right]_+}\Big/\sqrt{\tau_n^\star},$$

where $[\cdot]_+ = \max(0, \cdot)$. Then the slope of the line segment is $\beta = \delta_Y/\delta_X$, provided $\delta_X > 0$. Note that in the ratio δ_Y/δ_X the common σ_λ^{-1} factor drops out, so that knowing the moments of λ is unnecessary for slope estimation. However, estimates of

$$\sum_{i=1}^{n}\operatorname{Var}(X_i/\sigma_i)/n, \quad \sum_{i=1}^{n}\operatorname{Var}(Y_i/\tau_i)/n \quad \text{and} \quad \sum_{i=1}^{n}\operatorname{Cov}(X_i, Y_i)/n$$

are required.

Following the approach described in [4, 5], we employ method-of-moments estimators, by equating sample moments with theoretical moments:

$$\frac{1}{n}\sum_{i=1}^{n}\operatorname{Var}\left(\frac{X_i}{\sigma_i}\right) \stackrel{\text{set}}{=} \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{\sigma_i}\right)^2 = S_{xx}^*,$$

$$\frac{1}{n}\sum_{i=1}^{n}\operatorname{Var}\left(\frac{Y_i}{\tau_i}\right) \stackrel{\text{set}}{=} \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\overline{y}}{\tau_i}\right)^2 = S_{yy}^* \quad \text{and}$$

$$\frac{1}{n}\sum_{i=1}^{n}\operatorname{Cov}(X_i, Y_i) \stackrel{\text{set}}{=} \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})y_i = S_{xy}.$$

Hence

$$\widehat{\delta}_X = \sigma_\lambda^{-1}\sqrt{\left[S_{xx}^*-1\right]_+/\sigma_n^\star} \quad \text{and} \quad \widehat{\delta}_Y = \operatorname{sgn}(S_{xy})\sigma_\lambda^{-1}\sqrt{[S_{yy}^*-1]_+/\tau_n^\star}, \tag{2}$$

so that the estimated slope of the line segment is

$$\widehat{\beta} = \widehat{\delta}_Y/\widehat{\delta}_X = \operatorname{sgn}(S_{xy})\sqrt{\frac{\sigma_n^\star[S_{yy}^*-1]_+}{\tau_n^\star[S_{xx}^*-1]_+}}. \tag{3}$$

When the MEs are homoscedastic, (3) agrees with (2.12) in [4].

3. Variance of the slope estimate

We continue to follow [4] in our derivation of the variance of our estimate. Since the slope estimate $\widehat{\beta}$ is location invariant, we may set, without loss of generality, $\eta_X = \eta_Y = 0$ in our derivation of $\operatorname{Var}(\widehat{\beta})$. We define

$$Z_i = \left(X_i, \frac{X_i}{\sigma_i^2}, \frac{X_i^2}{\sigma_i^2}, Y_i, \frac{Y_i}{\tau_i^2}, \frac{Y_i^2}{\tau_i^2}, X_iY_i\right)' = (z_1, z_2, z_3, z_4, z_5, z_6, z_7)'$$

and let

$$T_n = \frac{1}{n}\sum_{i=1}^{n}Z_i = \left(\overline{X}, \frac{1}{n}\sum_{i=1}^{n}\frac{X_i}{\sigma_i^2}, \frac{1}{n}\sum_{i=1}^{n}\frac{X_i^2}{\sigma_i^2}, \overline{Y}, \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{\tau_i^2}, \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{\tau_i^2}, \frac{1}{n}\sum_{i=1}^{n}X_iY_i\right)'$$
$$= (t_1, t_2, t_3, t_4, t_5, t_6, t_7).$$

This structure on T_n is more complex than its five-dimensional counterpart in the homoscedastic case. Nevertheless, with $\phi_\lambda = \mathscr{E}(\lambda_i^2) = \sigma_\lambda^2 + \mu_\lambda^2$, it is straightforward to show that

$$\boldsymbol{\mu} = \mathscr{E}(T_n) = (\delta_X\mu_\lambda, \delta_X\mu_\lambda\sigma^\star, \delta_X^2\phi_\lambda\sigma^\star + 1, \delta_Y\mu_\lambda, \delta_Y\mu_\lambda\tau^\star, \delta_Y^2\phi_\lambda\tau^\star + 1, \delta_X\delta_Y\phi_\lambda)'$$

based on the model assumptions in (1), and the assumed existence of $\sigma^\star = \mathscr{E}(\sigma_i^{-2}) = \lim_{n\to\infty}\sigma_n^\star$ and $\tau^\star = \mathscr{E}(\tau_i^{-2}) = \lim_{n\to\infty}\tau_n^\star$. The central limit theorem then assures that the distribution of $\sqrt{n}(T_n - \boldsymbol{\mu})$ converges to the $\mathscr{N}(0, \Sigma)$ distribution as $n \to \infty$, where Σ is the 7×7 covariance matrix of Z_i with entries $\Sigma_{ij} = \operatorname{Cov}(z_i, z_j)$, $1 \leqslant i, j \leqslant 7$.

Now define $S_n=(S_{xx}^*, S_{yy}^*, S_{xy})$. Expanding each of these components and applying some algebraic manipulation, we find that S_n may be expressed as a function H of $T_n=(t_1,t_2,t_3,t_4,t_5,t_6,t_7)$, namely,

$$S_n=H(T_n)=(t_3-2t_1t_2+t_1^2\sigma_n^\star, t_6-2t_4t_5+t_4^2\tau_n^\star, t_7-t_1t_4)',$$

so that, when H is applied to $\mathscr{E}(T_n)=\boldsymbol{\mu}$, we have

$$\boldsymbol{\phi}=\mathscr{E}(S_n)=H(\boldsymbol{\mu})=(\delta_X^2\sigma_\lambda^2\sigma^\star+1, \delta_Y^2\sigma_\lambda^2\tau^\star+1, \delta_X\delta_Y\sigma_\lambda^2)'.$$

Applying the delta method and Slutzky's theorem to S_n, we have

$$\sqrt{n}(S_n-\boldsymbol{\phi})\xrightarrow{\mathscr{L}}\mathscr{N}(0, M\Sigma M'),$$

where

$$M=\left.\frac{\partial H(T_n)}{\partial(T_n)}\right|_{T_n=\boldsymbol{\mu}}=\begin{bmatrix}0 & -2\mu_\lambda\delta_X & 1 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & -2\mu_\lambda\delta_Y & 1 & 0\\ -\mu_\lambda\delta_Y & 0 & 0 & -\mu_\lambda\delta_X & 0 & 0 & 1\end{bmatrix}.$$

The law of large numbers assures that $S_n\xrightarrow{P}\boldsymbol{\phi}$, and the continuous mapping theorem then implies that $\widehat{\beta}=\widehat{\beta}(S_n)\xrightarrow{P}\widehat{\beta}(\boldsymbol{\phi})=\delta_Y/\delta_X=\beta$, so that $\widehat{\beta}$ is a consistent estimator of β. A second application of the delta method then gives us

$$\sqrt{n}(\widehat{\beta}-\beta)\xrightarrow{\mathscr{L}}\mathscr{N}(0, B'M\Sigma M'B)=\mathscr{N}(0,\omega),$$

where

$$B=\left.\frac{\partial\widehat{\beta}(S_n)}{\partial(S_n)}\right|_{S_n=\boldsymbol{\phi}}=\left[\frac{-\delta_Y}{2\delta_X^3\sigma_\lambda^2\sigma^\star}, \quad \frac{1}{2\delta_X\delta_Y\sigma_\lambda^2\tau^\star}, \quad 0\right]'.$$

Hence

$$\begin{aligned}\omega=&\frac{\delta_Y^2}{4\delta_X^6\sigma_\lambda^4(\sigma^\star)^2}(4\delta_X^2\mu_\lambda^2\Sigma_{22}+\Sigma_{33}-4\delta_x\mu_\lambda\Sigma_{23})\\&+\frac{1}{4\delta_x^2\delta_y^2\sigma_\lambda^4(\tau^\star)^2}(4\delta_Y^2\mu_\lambda^2\Sigma_{55}+\Sigma_{66}-4\delta_Y\mu_\lambda\Sigma_{56})\\&-\frac{1}{2\delta_X^4\sigma_\lambda^4\sigma^\star\tau^\star}(4\delta_X\delta_Y\mu_\lambda^2\Sigma_{25}-2\delta_X\mu_\lambda\Sigma_{26}-2\delta_Y\mu_\lambda\Sigma_{35}+\Sigma_{36})\end{aligned}$$

is the desired asymptotic variance of $\sqrt{n}(\widehat{\beta}-\beta)$.

We substitute

$$\begin{aligned}\Sigma_{22}&=\delta_X^2\sigma_\lambda^2\sigma^{\star\star}+\sigma^\star, & \Sigma_{33}&=\delta_X^4(\kappa_\lambda-\phi_\lambda^2)\sigma^{\star\star}+4\delta_X^2\phi_\lambda\sigma^\star+2,\\ \Sigma_{23}&=\delta_X^3(\gamma_\lambda-\mu_\lambda\phi_\lambda)\sigma^{\star\star}+2\delta_X\mu_\lambda\sigma^\star, & \Sigma_{55}&=\delta_Y^2\sigma_\lambda^2\tau^{\star\star}+\tau^\star,\\ \Sigma_{66}&=\delta_Y^4(\kappa_\lambda-\phi_\lambda^2)\tau^{\star\star}+4\delta_Y^2\phi_\lambda\tau^\star+2, & \Sigma_{56}&=\delta_Y^3(\gamma_\lambda-\mu_\lambda\phi_\lambda)\tau^{\star\star}+2\delta_Y\mu_\lambda\tau^\star,\\ \Sigma_{25}&=\delta_X\delta_Y\sigma_\lambda^2(\sigma\tau)^\star, & \Sigma_{26}&=\delta_X\delta_Y^2(\gamma_\lambda-\mu_\lambda\phi_\lambda)(\sigma\tau)^\star,\\ \Sigma_{35}&=\delta_X^2\delta_Y(\gamma_\lambda-\mu_\lambda\phi_\lambda)(\sigma\tau)^\star, & \Sigma_{36}&=\delta_X^2\delta_Y^2(\kappa_\lambda-\phi_\lambda^2)(\sigma\tau)^\star\end{aligned}$$

into this expression, where $\gamma_\lambda=\mathscr{E}(\lambda_i^3)$, $\kappa_\lambda=\mathscr{E}(\lambda_i^4)$, and assuming the existence of each limit, we define

$$\sigma^{\star\star}=\mathscr{E}(\sigma_i^{-4})=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n\frac{1}{\sigma_i^4}=\lim_{n\to\infty}\sigma_n^{\star\star}, \quad \tau^{\star\star}=\mathscr{E}(\tau_i^{-4})=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n\frac{1}{\tau_i^4}=\lim_{n\to\infty}\tau_n^{\star\star},$$

and $(\sigma\tau)^\star=\mathscr{E}(\sigma_i^{-2}\tau_i^{-2})$. Note that $(\sigma\tau)^\star=\sigma^\star\tau^\star$ by the independence of σ_i and τ_i for all i.

After much labor, we obtain

$$\omega = \frac{1}{4\delta_X^6\delta_Y^2\sigma_\lambda^4}\left\{2\frac{\delta_X^4}{(\tau^\star)^2}+4\delta_X^2\delta_Y^2\sigma_\lambda^2\left(\frac{\delta_X^2}{\tau^\star}+\frac{\delta_Y^2}{\sigma^\star}\right)+2\frac{\delta_Y^4}{(\sigma^\star)^2}\right.$$
$$\left.+\,\delta_X^4\delta_Y^4\Gamma_\lambda\left[\frac{\sigma^{\star\star}}{(\sigma^\star)^2}+\frac{\tau^{\star\star}}{(\tau^\star)^2}-2\frac{(\sigma\tau)^\star}{\sigma^\star\tau^\star}\right]\right\}, \tag{4}$$

where $\Gamma_\lambda=\kappa_\lambda-4\mu_\lambda^4+8\mu_\lambda^2\phi_\lambda-4\mu_\lambda\gamma_\lambda-\phi_\lambda^2$. The Cauchy–Schwartz inequality guarantees that $\sigma^{\star\star}\geqslant(\sigma^\star)^2$ and $\tau^{\star\star}\geqslant(\tau^\star)^2$, so that the expression $\sigma^{\star\star}/(\sigma^\star)^2+\tau^{\star\star}/(\tau^\star)^2-2(\sigma\tau)^\star/\sigma^\star\tau^\star$ will be nonnegative, and will vanish in the case of homoscedasticity. Although the ratio $(\sigma\tau)^\star/\sigma^\star\tau^\star$ equals one, we keep it in the form of (4) because it will need to be estimated in the next section. Therefore, the variance of the estimated slope based on n independent observed pairs (x_i, y_i) with heteroscedastic ME is $\mathrm{Var}(\widehat{\beta})=\omega/n$.

We note that in the case of homoscedastic error variance, (4) reduces to (3.3) given in [4].

4. Estimating the variance of the slope estimate

To obtain an estimate of $\mathrm{Var}(\widehat{\beta})$, we replace each occurrence of δ_X and δ_Y in (4) with the corresponding estimates given in (2). We also replace $\sigma^\star$, $\tau^\star$, $\sigma^{\star\star}$, $\tau^{\star\star}$ and $(\sigma\tau)^\star$ with $\sigma_n^\star$, $\tau_n^\star$, $\sigma_n^{\star\star}$, $\tau_n^{\star\star}$ and $(\sigma\tau)_n^\star$, respectively. After some simplification we obtain

$$\widehat{\mathrm{Var}}(\widehat{\beta}) = \frac{\sigma_n^\star/\tau_n^\star}{n(S_{xx}^*-1)}\left\{1+\frac{1}{2(S_{yy}^*-1)}+\frac{S_{yy}^*-1}{S_{xx}^*-1}+\frac{S_{yy}^*-1}{2(S_{xx}^*-1)^2}\right.$$
$$\left.+\,\frac{S_{yy}^*-1}{4}\left[\frac{\sigma_n^{\star\star}}{(\sigma_n^\star)^2}+\frac{\tau_n^{\star\star}}{(\tau_n^\star)^2}-2\frac{(\sigma\tau)_n^\star}{\sigma_n^\star\tau_n^\star}\right]\sigma_\lambda^{-4}\widehat{\Gamma}\right\}. \tag{5}$$

To derive the estimate $\sigma_\lambda^{-4}\widehat{\Gamma}$, we first require an estimate of λ_i for $i=1,\ldots,n$. Suppose we observe (x_i, y_i), while $\mathscr{E}(X_i, Y_i)=(\eta_X+\lambda_i\delta_X, \eta_Y+\lambda_i\delta_Y)$. In our model, λ_i must be the unique value in $[0, 1]$ that minimizes the Euclidean distance between (x_i, y_i) and the line segment whose endpoints are at (η_X, η_Y) and (ξ_X, ξ_Y). Simple calculus gives us

$$\lambda_i=\frac{\delta_X(x_i-\eta_X)+\delta_Y(y_i-\eta_Y)}{\delta_X^2+\delta_Y^2},$$

so that

$$\widehat{\lambda}_i=\frac{\widehat{\delta}_X(x_i-\widehat{\eta}_X)+\widehat{\delta}_Y(y_i-\widehat{\eta}_Y)}{\widehat{\delta}_X^2+\widehat{\delta}_Y^2}$$
$$=\left[\frac{(x_i-\widehat{\eta}_X)\sqrt{[S_{xx}^*-1]_+/\sigma_n^\star}+(y_i-\widehat{\eta}_Y)\,\mathrm{sgn}(S_{xy})\sqrt{S_{yy}^*-1]_+/\tau_n^\star}}{(S_{xx}^*-1)/\sigma_n^\star+(S_{yy}^*-1)/\tau_n^\star}\right]\sigma_\lambda$$
$$=\psi(x_i, y_i)\sigma_\lambda=\psi_i\sigma_\lambda.$$

We then have

$$\widehat{\mu}_\lambda=\sigma_\lambda\left(\frac{1}{n}\sum_{i=1}^n\psi_i\right),\quad \widehat{\phi}_\lambda=\sigma_\lambda^2\left(\frac{1}{n}\sum_{i=1}^n\psi_i^2\right),\quad \widehat{\gamma}_\lambda=\sigma_\lambda^3\left(\frac{1}{n}\sum_{i=1}^n\psi_i^3\right),$$

and

$$\widehat{\kappa}_\lambda=\sigma_\lambda^4\left(\frac{1}{n}\sum_{i=1}^n\psi_i^4\right).$$

Thus

$$\begin{aligned}\widehat{\Gamma}_\lambda &= \widehat{\kappa}_\lambda - 4\widehat{\mu}_\lambda^4 + 8\widehat{\mu}_\lambda^2\widehat{\phi}_\lambda - 4\widehat{\mu}_\lambda\widehat{\gamma}_\lambda - \widehat{\phi}_\lambda^2 \\ &= \sigma_\lambda^4\left(\frac{1}{n}\sum_{i=1}^n \psi_i^4\right) - 4\sigma_\lambda^4\left(\frac{1}{n}\sum_{i=1}^n \psi_i\right)^4 + 8\sigma_\lambda^4\left(\frac{1}{n}\sum_{i=1}^n \psi_i\right)^2\left(\frac{1}{n}\sum_{i=1}^n \psi_i^2\right) \\ &\quad - 4\sigma_\lambda^4\left(\frac{1}{n}\sum_{i=1}^n \psi_i\right)\left(\frac{1}{n}\sum_{i=1}^n \psi_i^3\right) - \sigma_\lambda^4\left(\frac{1}{n}\sum_{i=1}^n \psi_i^2\right)^2,\end{aligned}$$

so

$$\begin{aligned}\sigma_\lambda^{-4}\widehat{\Gamma}_\lambda &= \left(\frac{1}{n}\sum_{i=1}^n \psi_i^4\right) - 4\left(\frac{1}{n}\sum_{i=1}^n \psi_i\right)^4 + 8\left(\frac{1}{n}\sum_{i=1}^n \psi_i\right)^2\left(\frac{1}{n}\sum_{i=1}^n \psi_i^2\right) \\ &\quad - 4\left(\frac{1}{n}\sum_{i=1}^n \psi_i\right)\left(\frac{1}{n}\sum_{i=1}^n \psi_i^3\right) - \left(\frac{1}{n}\sum_{i=1}^n \psi_i^2\right)^2. \qquad (6)\end{aligned}$$

We may then substitute (6) into (5) to compute $\widehat{\mathrm{Var}}(\widehat{\beta})$.

However, we have sidestepped the issue of estimating (η_X, η_Y), which is necessary for computing ψ_i, $i = 1, \ldots, n$. One remote possibility is that η_X and η_Y are known. Another approach, given in [5] under the assumption that μ_λ and σ_λ^2 are known, uses the Method of Moments to obtain $\widehat{\eta}_X = \overline{x} - \mu_\lambda\widehat{\delta}_X$ and $\widehat{\eta}_Y = \overline{y} - \mu_\lambda\widehat{\delta}_Y$. Third, one may construct consistent estimators of η_X and η_Y using nonparametric estimation of G, as discussed in [5]. When information about G cannot be determined, we suggest a heuristic alternative.

The line segment in our model rests on the line having slope β and passing through the point $(\eta_X + \mu_\lambda\delta_X, \eta_Y + \mu_\lambda\delta_Y)$. We estimate the location of this segment with a line having slope $\widehat{\beta}$ and passing through the point $(\overline{x}, \overline{y})$, whose equation is thus $y = \widehat{\beta}(x - \overline{x}) + \overline{y}$. Now, for $i = 1, \ldots, n$, consider the perpendicular line having slope $-1/\widehat{\beta}$ which passes through the point (x_i, y_i), whose equation is thus $y = -1/\widehat{\beta}(x - x_i) + y_i$. Then both lines contain the orthogonal projection (x_i^*, y_i^*) of (x_i, y_i) onto the estimated line segment. We substitute (x_i^*, y_i^*) into each equation in place of (x, y), and set their right-hand sides equal, to get $\widehat{\beta}(x_i^* - \overline{x}) + \overline{y} = -1/\widehat{\beta}(x_i^* - x_i) + y_i$. Solving for x_i^* and simplifying gives us

$$x_i^* = \frac{\widehat{\beta}(y_i - \overline{y} + \widehat{\beta}\overline{x}) + x_i}{\widehat{\beta}^2 + 1}.$$

This value can then be used to compute $y_i^* = -1/\widehat{\beta}(x_i^* - x_i) + y_i$. If we let $x_0^* = \min\{x_1^*, \ldots, x_n^*\}$, we have $x_0^* \to \eta_X$ as $n \to \infty$. If $\widehat{\beta} \geqslant 0$, we let $y_0^* = \min\{y_1^*, \ldots, y_n^*\}$. Otherwise, we let $y_0^* = \max\{y_1^*, \ldots, y_n^*\}$. In either case, $y_0^* \to \eta_Y$ as $n \to \infty$. Hence, (x_0^*, y_0^*) is a biased but consistent estimator of (η_X, η_Y), hence, we propose the use of $(\widehat{\eta}_X, \widehat{\eta}_Y) = (x_0^*, y_0^*)$ in the computation of ψ_i, $i = 1, \ldots, n$ when nothing is known about the distribution of λ_i.

5. Simulation study

To confirm the accuracy of these estimates, we generate 50 data pairs using (1) with $\eta_X = 0$, $\eta_Y = 0$, $\delta_X = 9$, and $\delta_Y = 6$, so that $\beta = \frac{2}{3}$. For $i = 1, \ldots, n$, λ_i is drawn from a beta distribution with both parameters equal to 2, σ_i is drawn from a uniform distribution on $(a, 2a)$, and τ_i is drawn from a uniform distribution on $(b, 2b)$, where a and b are fixed but arbitrary positive numbers. Hence, Γ_λ can be computed from the known moments of a beta distribution, while $\sigma^\star$, $\tau^\star$, $\sigma^{\star\star}$ and $\tau^{\star\star}$ are derived from the known moments of a uniform distribution. We choose a range of values for a and b, then use (3), (4), (5) and (6) to compute $\widehat{\beta}$, its variance $\mathrm{Var}(\widehat{\beta})$, and the estimate $\widehat{\mathrm{Var}}(\widehat{\beta})$ of this variance. For this first simulation we assume that η_X and η_Y are known. We repeat 500 times, and record the estimates at each iteration.

For each run of the simulation, Table I presents the selected values of a and b, which control the magnitude of the error variances, in the first two columns. Column 3 gives the median value of $\widehat{\beta}$, computed using (3), over 500 iterations, along with the sample variance of this estimate in column 4. Column 5 provides the expected value of $\mathrm{Var}(\widehat{\beta})$, based on (4) divided by n. Column 6 provides the median value of $\widehat{\mathrm{Var}}(\widehat{\beta})$, computed using (5), over 500 iterations. Ideally, the value in column 3 matches the true value of the slope, i.e. $\frac{2}{3}$, and the values in columns 4 and 6 are both close to the value in column 5 for each run. For all choices of a and b, our estimate of the line-segment slope is centered very near the true value of 0.667. Moreover, the observed variance among 500 computed values of $\widehat{\beta}$ (in column 4) is consistently close to the expected variance of $\widehat{\beta}$ (in column 5), and the median estimate of the variance of $\widehat{\beta}$ over 500 iterations (in

Table I. Simulation results for several choices of a and b, based on 500 iterations, with (η_X, η_Y) known and $\beta = \frac{2}{3}$. (S.V. = sample variance).

a	b	Median $(\widehat{\beta})$	(S.V.)	Var$(\widehat{\beta})$	Median $(\widehat{\text{Var}}(\widehat{\beta}))$
0.05	0.03	0.669	(0.00086)	0.00087	0.00083
0.35	0.25	0.662	(0.00224)	0.00204	0.00219
0.50	0.55	0.663	(0.00564)	0.00550	0.00610
0.75	0.70	0.666	(0.01077)	0.00982	0.01085
0.85	0.90	0.657	(0.02037)	0.01618	0.01801

Table II. Simulation results for several choices of a and b, based on 500 iterations, with (η_X, η_Y) unknown and $\beta = -\frac{2}{3}$. (S.V. = sample variance).

a	b	Median $(\widehat{\beta})$	(S.V.)	Var$(\widehat{\beta})$	Median $(\widehat{\text{Var}}(\widehat{\beta}))$
0.05	0.04	−0.667	(0.00086)	0.00087	0.00085
0.30	0.33	−0.665	(0.00228)	0.00239	0.00257
0.46	0.52	−0.660	(0.00581)	0.00490	0.00528
0.71	0.65	−0.673	(0.00976)	0.00849	0.00968
0.98	0.93	−0.664	(0.02160)	0.01871	0.02178

column 6) also proves to be quite accurate, with a gradual loss of accuracy as the error variances grow. To the extent that our estimates of the variance are off-target, they are consistently a bit high, and hence give us more conservative estimates. The precision of our estimates even as the error variance grows is remarkable given our need to estimate many parameters. Hence, our first simulation confirms the reliability of our estimates when η_X and η_Y are known.

In our second simulation, we set $\eta_X = 2$, $\eta_Y = 3$, $\delta_X = 9$, and $\delta_Y = -6$, so that $\beta = -\frac{2}{3}$. This time we assume that η_X and η_Y are unknown and employ our heuristic estimates. All other conditions are the same as above. Table II, which has the same structure as Table I, displays a summary of our results. Despite our need to use biased estimates of η_X and η_Y to compute the values in column 6, these estimates of the variance of $\widehat{\beta}$ are centered at values only slightly larger than the expected values listed in column 5. Our sample variance of $\widehat{\beta}$ over 500 iterations, given in column 4, also corresponds well with the expected variance in every run given in column 5, although it becomes inflated as the error variances grow. Moreover, the sample median of $\widehat{\beta}$ is consistently accurate even as the variance of the MEs increases. Hence, our derivations are strongly validated by these simulations.

We then compare the performance of our slope estimate under the line-segment model to the performance of the method-of-moments slope estimate proposed recently in [2], based on the equation-error structural heteroscedastic ME model described in the Introduction. An EM algorithm to compute maximum-likelihood estimates for these model parameters was proposed in [1], and additional estimation methods were presented in [3]. In [2], Patriota *et al.* derive both method-of-moments and maximum-likelihood estimates for the parameters and provide performance comparisons with the earlier approaches. We will focus on their method-of-moments estimate, which we designate here as MM-P, as a contrast to ours, since that estimate is provided in closed form, and the authors found its performance to be superior to that of alternate equation-error models when the Gaussian error assumption is valid.

Using our notational equivalents, the MM-P slope estimate is $\widehat{\beta} = S_{xy}/(S_{xx} - \sigma_n^\star)$, where

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2/(n-1), \quad S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})/(n-1),$$

and (needed below)

$$S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2/(n-1).$$

The asymptotic variance corresponding to the MM-P slope estimate is $[2\beta^2(\sigma^{\star\star} - \sigma_\chi^4) + \pi]/\sigma_\chi^4$, where $\pi = \beta^2\sigma_\chi^2\sigma^\star + \zeta^2\sigma_\chi^2 + (\sigma\tau)^\star + \zeta^2\sigma^\star + \sigma_\chi^2\tau^\star + 2\beta^2\sigma_\chi^4$. For our line-segment model, $\sigma_\chi^2 = \delta_X^2\sigma_\lambda^2$ and the equation error ζ^2 equals zero. Hence, we may compute the expected large-sample variance of $\widehat{\beta}$ under MM-P for our simulation scenario using the known moments of the beta and uniform distributions, once a and b are specified. Moreover, using the proposed estimates of these parameters, the MM-P estimate of Var$(\widehat{\beta})$ is

Table III. Simulation variance estimates for several choices of a and b, based on 500 iterations, using the MM-P model and our model, with $\beta=\frac{2}{3}$.

		MM-P model		Our model	
a	b	$\text{Var}(\widehat{\beta})$	Median $(\widehat{\text{Var}}(\widehat{\beta}))$	$\text{Var}(\widehat{\beta})$	Median $(\widehat{\text{Var}}(\widehat{\beta}))$
0.05	0.03	0.00002	0.00002	0.00087	0.00083
0.35	0.25	0.00150	0.00151	0.00204	0.00219
0.50	0.55	0.00569	0.00555	0.00550	0.00610
0.75	0.70	0.01248	0.01259	0.00982	0.01085
0.85	0.90	0.02043	0.02003	0.01618	0.01801

Table IV. Simulation variance estimates for several choices of a and b, based on 500 iterations, using the MM-P model and our model, with $\beta=-\frac{2}{3}$.

		MM-P model		Our model	
a	b	$\text{Var}(\widehat{\beta})$	Median $(\widehat{\text{Var}}(\widehat{\beta}))$	$\text{Var}(\widehat{\beta})$	Median $(\widehat{\text{Var}}(\widehat{\beta}))$
0.05	0.04	0.00003	0.00003	0.00087	0.00085
0.30	0.33	0.00184	0.00179	0.00239	0.00257
0.46	0.52	0.00488	0.00472	0.00490	0.00528
0.71	0.65	0.01057	0.01102	0.00849	0.00968
0.98	0.93	0.02660	0.02685	0.01871	0.02178

$$\widehat{\text{Var}}(\widehat{\beta})=\frac{2S_{xy}^2[\sigma_n^{\star\star}-(S_{xx}-\sigma_n^{\star})^2]}{n(S_{xx}-\sigma_n^{\star})^4}+\frac{S_{xy}^2+S_{xx}S_{yy}+(\sigma\tau)_n^{\star}-\sigma_n^{\star}\tau_n^{\star}}{n(S_{xx}-\sigma_n^{\star})^2}. \tag{7}$$

When we repeat the first simulation using the MM-P estimates, with the same choices for a and b, the median of the 500 point estimates of β is consistently close to the true value of $\frac{2}{3}$. The first two columns of Table III show the expected value of $\text{Var}(\widehat{\beta})$ under the MM-P model for each pair (a,b), along with the corresponding median of the 500 estimates of $\text{Var}(\widehat{\beta})$ computed using (7). As the magnitude of the error variance grows, the difference between these two values grows, but remains within reason. But note that when we compare these values with those in the last two columns, which are the corresponding columns imported from Table I, we find that the MM-P expected and estimated variance of $\widehat{\beta}$ is much smaller than ours when a and b are small (first two rows), about the same when a and b are close to 0.5 (middle row), and much larger when a and b are larger than 0.5 (last two rows). In other words, the MM-P estimates are more precise when the error variances are quite small, but our estimates have greater precision when the error variances become appreciable. Hence we have evidence that the line-segment model will provide a better estimate of the slope when the ME variance is not small, as is frequently the case.

Of course, our estimates from Table I are based on the assumption that η_X and η_Y are known. However, even in the second simulation scenario, in which we must estimate (η_X,η_Y), our estimate of $\text{Var}(\widehat{\beta})$ still returns a smaller value than the MM-P version once the error variances become appreciable, as Table IV shows. In this scenario we observe that the variance estimate is indeed smaller under the MM-P model when the σ_i and the τ_i are restricted to small values, but when the σ_i are drawn from the interval $(0.46, 0.92)$ and the τ_i are drawn from $(0.52, 1.04)$, the estimates of the variance of $\widehat{\beta}$ are approximately the same under both methods. Then, when the σ_i are drawn from the interval $(0.71, 1.42)$ and the τ_i are drawn from $(0.65, 1.30)$, the variance estimate under our method is significantly smaller, and when the σ_i are drawn from the interval $(0.98, 1.96)$, while the τ_i are drawn from $(0.93, 1.86)$, our estimate of the variance of $\widehat{\beta}$ is almost half of the corresponding estimate under the MM-P model. Hence, we recommend our approach to slope estimation when the heteroscedastic ME is not small, as our method will yield narrower confidence intervals for the slope.

It should be noted that in this simulation the data-generation mechanism was specified according to the line-segment model, whereas the MM-P procedure is based on the equation-error model. Thus, our estimation procedure had a built-in advantage. When the data-generation mechanism is unknown, one must take care in choosing a model, as the results from disparate models can be quite different. The next section demonstrates this issue.

6. Data application

The WHO established the MONICA during the 1980s to study the association between known risk factors, such as smoking and obesity, and trends in cardiovascular disease. The linear association between data on the average annual

change in the observed risk score (X) and the average annual change in event rate (Y), both given as percentages, was modeled in [1] and [2] using equation-error ME procedures, with the sampling error in the trend estimates taken as the heteroscedastic ME. Figure 2 displays the data for males ($N=38$) and for females ($N=36$) separately, with the magnitude of the MEs indicated by the crosshairs.

Table V displays the estimated slope and its estimated standard error for each gender computed under several different models. The ordinary least-squares (OLS) method disregards ME. K-2000 and K-2002 represent maximum likelihood estimates provided in [1], whereas MM-P and ML-P represent the method-of-moments and maximum-likelihood estimates reported in [2], using ME models. Finally, the MM-LS estimates are those obtained using our line-segment model. We add the lines having the estimated slopes under our model and under the ML-P and MM-P models to the plots in Figure 2. For the MM-LS results, we pass the lines through the means $(\overline{X}, \overline{Y})$ for each gender, whereas for the ML-P and MM-P results we use the estimated intercepts provided in that paper.

It is quite surprising that the estimated slopes under the line-segment model are dramatically steeper than those obtained under all the other models—more than three times the size—while the standard errors on the estimates are about the same for all models. While this initially makes the MM-LS results appear to be in error, inspection of the data scatter in each plot of Figure 2 reveals that these results are more consistent with the observable trend. Indeed, the steeper slopes send an even stronger message to the public about the urgency of maintaining cardiovascular health. While it is difficult to ascertain which model is most appropriate for the latent WHO MONICA data-generation mechanism, many tools are available for assessing model fitness.

For example, when we apply diagnostic tools designed for identification of influential points under the OLS framework,

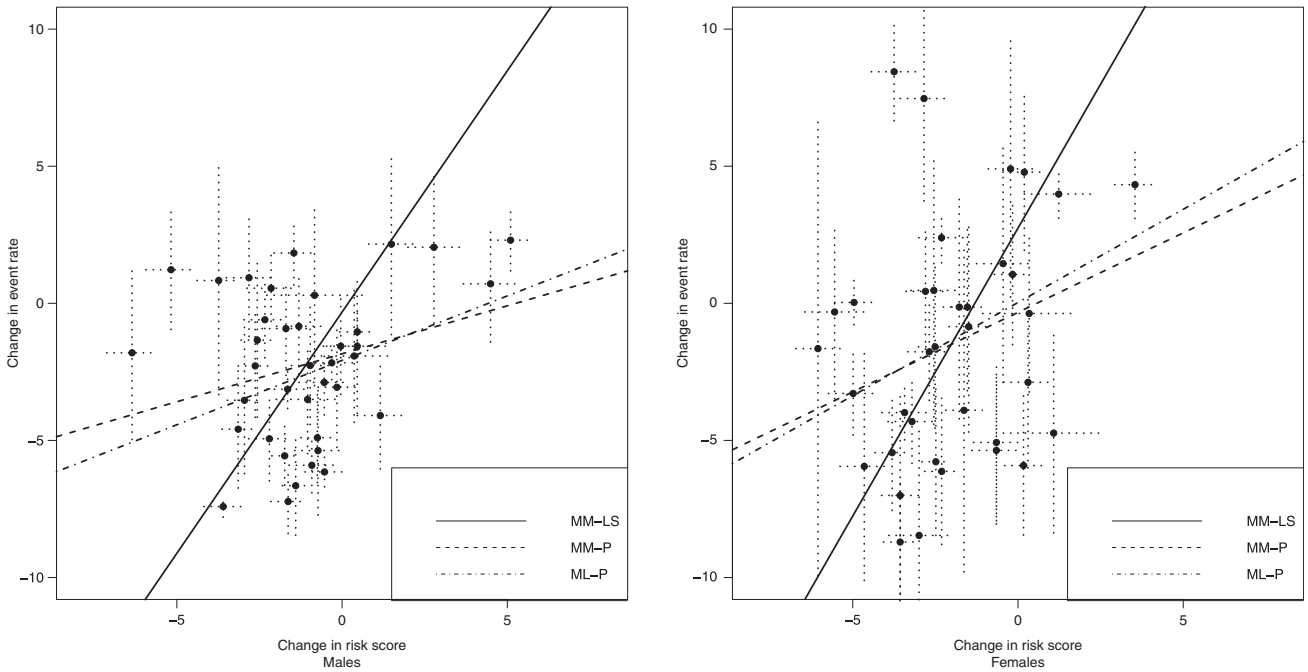

Figure 2. Scatterplot of change in event rate versus change in risk score, with standard errors, from WHO MONICA project, and lines having estimated slopes under three models, for males and females.

Table V. Estimates of the slope and standard errors of the estimates for the WHO MONICA data on males and females, based on seven models.

	Males		Females	
Model	$\widehat{\beta}$	$\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta})}$	$\widehat{\beta}$	$\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta})}$
OLS	0.31	0.20	0.51	0.33
K-2000	0.43	0.22	0.57	0.33
K-2002	0.47	0.23	0.68	0.24
MM-P	0.35	0.22	0.58	0.38
ML-P	0.47	0.23	0.68	0.41
MM-LS	1.76	0.23	2.10	0.38

Table VI. Estimates of the slope and standard errors of the estimates for the WHO MONICA data on males and females, based on two models, when no points are deleted, and when individual influential points are deleted.

Model	Full data	Point 1	Point 2	Point 3	Point 4
Males					
MM-P	0.35(0.22)	0.45(0.24)	0.31(0.25)	0.23(0.25)	
MM-LS	1.76(0.23)	1.85(0.25)	1.86(0.26)	2.01(0.27)	
Females					
MM-P	0.58(0.38)	0.72(0.35)	0.65(0.41)	0.63(0.35)	0.44(0.44)
MM-LS	2.10(0.38)	1.93(0.38)	2.22(0.41)	2.10(0.37)	2.39(0.50)

such as Cook's distance, DFFITS, DFBETA and the covariance ratio, we discover that three points are designated as influential among the male data, and four points among the female data. While these diagnostic tools are not intended for ME models, they imply that methods based on models which incorporate equation error are very vulnerable to the influence of outliers at the horizontal extremes. Under models that rely on such methods, the combined effect of these influential points is to rotate the line in the clockwise direction, resulting in a flatter slope. In contrast, the line-segment model does not involve equation error and is thus robust against such influences. This illustrates the effect of using equation error in ME models when support for this assumption is lacking, as discussed in the introduction. In a case such as that considered here, when the latent mechanism which has generated the data is unknown, it is safer to rely on a model which, like the line-segment model, is symmetric and less rigid.

We may demonstrate the robustness of the line-segment model by deleting each of the identified influential points one at a time and computing the slope estimate on each subset of the data. We do likewise for the MM-P model, and display the results in Table VI. Deletion of influential points alters the slope under our line-segment model by 5 to 14 per cent for males and by -8 to 14 per cent for females. But under the MM-P model the slope is altered by -34 to 29 per cent for males and by -24 to 24 per cent for females. This illustrates the robustness of the line-segment model to influential points and helps explain the disparity between the slope estimates from the two models. Given that the precision of the slope estimates is equivalent between the two models, this robustness property recommends the line-segment ME model over the others considered.

7. Discussion

Many scientific investigations involve assessing the association between the two variables when measurements recorded on both variables are subject to random error. When the variance of this error differs from one subject to another, a heteroscedastic ME model is appropriate. Conventional ME models incorporate an equation error component, which involves making potentially insupportable assumptions about the unknown data-generation mechanism. Misspecification of the chosen model can lead to significant underestimation of the slope, regardless of the estimation procedure used.

In this paper, we have provided an alternate heteroscedastic ME model based on a line-segment parameterization. This model does not incorporate equation error and is symmetric in both variables. For any setting in which this model corresponds well with the underlying data-generation mechanism, we have provided an accurate estimate of the linear association between the two variables, signified by the slope of the line segment, along with a reliable estimate of its precision. We have demonstrated through simulations that, under conditions when the line-segment model is properly specified, the corresponding variance estimate will yield smaller confidence intervals than will equation-error models when the error variances are not small. This novel estimation procedure enables an investigator to make precise inferences about the slope when heteroscedastic ME models are applied to scientific data, and thereby draw conclusions that will generally be more trustworthy than those derived using other ME models. Moreover, because the line-segment parameterization is robust against influential points which may plague equation-error models when their implementation is misspecified, the advantages of our model are further reinforced.

Acknowledgements

The authors thank Kari Kuulasmaa for making the WHO MONICA data available, and Alexandre Patriota for sharing the data.

This research is supported in part by the NSF under Grant No: BCS 0527766 and HSD 0826844, and by the NIH under Grant no: 1R01AG025218-01A2.

References

1. Kulathinal SB, Kuulasmaa K, Gasbarra D. Estimation of an errors-in-variables regression model when the variances of the measurement errors vary between the observations. *Statistics in Medicine* 2002; **21**:1089–1101. DOI: 10.1002/sim.1062.
2. Patriota AG, Bolfarine H, de Castro M. A heteroscedastic structural errors-in-variables model with equation error. *Statistical Methodology* 2009; **6**:408–423. DOI: 10.1016/j.stamet.2009.02.003.
3. Cheng CL, Riu J. On estimating linear relationships when both variables are subject to heteroscedastic measurement errors. *Technometrics* 2006; **48**:511–519. DOI: 10.1198/004017006000000237.
4. Davidov O. Estimating the slope in measurement error models—a different perspective. *Statistics and Probability Letters* 2005; **71**:215–223. DOI: 10.1016/j.spl.2004.11.011.
5. Davidov O, Goldenshluger A. Fitting a line segment to noisy data. *Journal of Statistical Planning and Inference* 2004; **119**:191–206. DOI: S0378-3758(02)00409-3.