STATISTICS IN MEDICINE
Statist. Med. (in press)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2845


Paper Celebrating the 25th Anniversary of *Statistics in Medicine*

On modelling metabolism-based biomarkers of exposure: A comparative analysis of nonlinear models with few repeated measurements

Brent A. Johnson[1,*,†] and Stephen M. Rappaport[2]

[1]*Department of Biostatistics, Rollins School of Public Health, Emory University, Atlanta, GA 30322, U.S.A.*
[2]*Department of Environmental Sciences and Engineering, School of Public Health, University of North Carolina, Chapel Hill, NC 27599, U.S.A.*

SUMMARY

Establishing and characterizing exposure–biomarker relationships is an important problem in molecular epidemiology. The problem is difficult due to several complicating features, namely, the biomarker response is a nonlinear function of exposure and unknown parameters; variation in exposure and biomarker levels occurs both within and between subjects; and errors tend to be heteroscedastic. To overcome some of the statistical challenges in analysing such data, it is common for the investigator to make several assumptions about the data structure. For example, it is common to assume that the natural logarithm of right-skewed, biomarker measurements lead to homoscedasicity and normality so the effect of outliers is minimized and Gauss–Markov theory is applicable. In this paper, we compare a lognormal maximum likelihood estimator (MLE) to generalized estimating equations (GEE) for drawing statistical inference in a nonlinear model of a benzene biomarker (benzene oxide–albumin adducts) as a function of benzene exposure. We explore the characteristic properties of the lognormal MLE under a certain type of model misspecification and compare its small sample performance to the estimating equation approach in simulation studies. We show that the multiplicative lognormal model can lead to severe biases for modest deviations from the true outcome (biomarker) distribution. Furthermore, the lognormal MLE can exhibit very poor small sample properties even under the true model. All methods are applied in a novel data analysis from a study of benzene-exposed workers in China. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: carcinogensis; environmental epidemiology; multiplicative lognormal; semiparametric inference; transform-both-sides

*Correspondence to: Brent A. Johnson, Department of Biostatistics, Rollins School of Public Health, Emory University, 1518 Clifton Rd NE, 3rd fl., Atlanta, GA 30322, U.S.A.
†E-mail: bajohn3@emory.edu

Contract/grant sponsor: American Chemistry Council; contract/grant number: MTH0311-01
Contract/grant sponsor: National Institute for Environmental Health Sciences; contract/grant numbers: T32ES07018, P42ES05948, P30ES10126

Received 23 August 2006
Accepted 10 January 2007

1. INTRODUCTION

Nonlinear models are popular in environmental health, particularly in kinetics. For example, consider the situation where one wishes to relate a local source of exposure (in the community or at the workplace) to levels of some biomarker in the exposed population. Rather than posit a (generalized) linear model relating biomarker levels directly to exposure, environmental scientists often relate the two through a series of biological and environmental processes. The partial differential equations that formulate these stochastic processes often result in statistical models that are nonlinear in the parameters and possess heteroscedastic errors. In this article, we compare two popular strategies for analysing such data and apply the methods to a study of benzene-exposed workers.

When a person is exposed to a chemical in the air, the rate of uptake into the body is determined by the physical and chemical properties of the contaminant and by the physiological characteristics of the subject. The subject eliminates the chemical from the body through several different processes, including passive clearance, e.g. in breath and metabolism. Passive clearance processes tend to follow first-order linear kinetics over a wide range of exposure levels while metabolism tends to follow nonlinear (saturable) Michaelis–Menten-like kinetics [1]. While the adverse health effects of benzene exposure are well documented [2, 3], relatively little is known about the mechanism of toxic action in humans other than the fact that toxicity requires metabolism to other products [4]. Since it is believed that all benzene metabolism proceeds through the initial metabolite, benzene oxide, adducts of benzene oxide and human serum albumin (BO–Alb) have been recommended as biomarkers of benzene metabolism [1]. In a study of factory workers in China, biomarker outcomes (levels of BO–Alb) and benzene exposure levels (inhaled air concentration over an 8-h workday) were collected. The attempts at modelling these Chinese data motivated the scientific investigation in this paper.

Nonlinear regression has been reviewed extensively in the statistics literature [5–8], with a key component being the handling of non-constant variance in the observed outcomes. For example, in the analysis of metabolism-based biomarkers, we often see that the variance of the biomarker outcome increases as the mean of the biomarker increases. In this paper, we consider two different approaches which model the observed biomarker outcome Y as a function of the mean biomarker outcome μ in different ways. One approach posits a multiplicative model, $Y = \mu\varepsilon^*$ where ε^* is a random variable with mean one, while the second approach writes the observed biomarker outcome as the sum of the mean plus error, i.e. $Y = \mu + \varepsilon$, where ε is a random variable with mean zero. An illustration of data generated from these two models is given in Figure 1.

Figure 1 illustrates the fact that for any given data set, it may be difficult to determine which error model, if any, generated the observed data. One may never be able to completely rule out the possibility that the additive error model generated the observed data in Figure 1(a) and *vice versa* for the multiplicative model and panel Figure 1(b). One goal of this paper is to illustrate the sensitivity of the multiplicative error model under modest deviations from the true model. In addition, we show *via* simulation that moment-based approaches using an additive error model will often fit data from a multiplicative error model reasonably well; however, we cannot make a similar statement about the converse based on our small sample studies.

By taking the natural logarithm of both sides of our multiplicative error model, we have an additive error model on the logarithmic scale—$\log Y = \log\mu + \log\varepsilon^*$. This transformation leads to our first of two methods for modelling the relationship between a biomarker outcome and exposure, namely, a parametric lognormal model. In practice, one assumes the natural logarithm of the biomarker outcomes are normally distributed and proceeds with standard linear model packages (e.g. SAS `PROC GLM`, R/S `glm`). For more complicated models with measurement error in the

DOI: 10.1002/sim

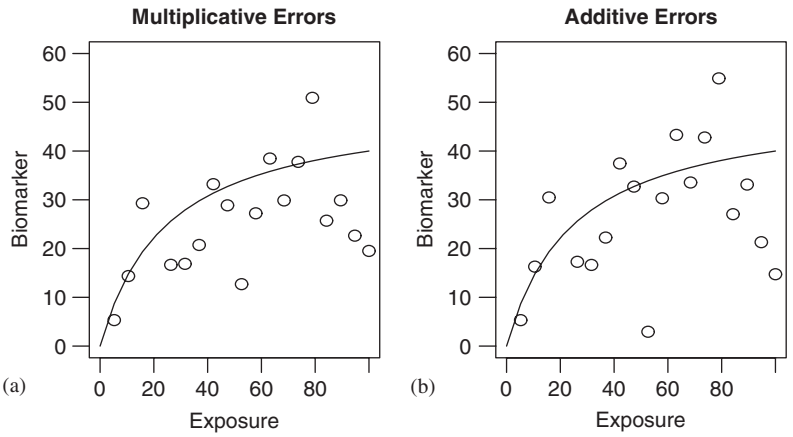

Figure 1. Illustration of data generated from multiplicative *versus* additive error model. The simulated biomarkers in panel (a) are lognormally distributed whereas the biomarkers in panel (b) are normally distributed with variance increasing as a power of the mean. The biomarker data in both panels have constant coefficient of variation.

exposure and/or limits of detection in the biomarker outcomes, one may proceed in a likelihood framework where the biomarker outcomes are assumed to follow a lognormal distribution. Our second method is a method-of-moments estimator, where one posits models for the first and second moment (that is, the mean and variance) of the biomarker outcome. The latter approach is often called the generalized estimating equations (GEE) [8, 9] method in the biostatistical literature. The fundamental difference between the two methods is GEE assumes the analyst 'knows' or models correctly the first moment (that is, the mean model), whereas the lognormal approach assumes the analyst knows every moment, thereby yielding the entire distribution function. An important feature of GEE methodology is the robust variance (sandwich) estimator, as opposed to the model-based estimator, for estimating the asymptotic covariance of the regression parameter estimates. Traditionally, the statistical mantra has been to expect the robust standard error estimates (SEEs) to be wider than their model-based counterparts, on average, in exchange for protection against misspecified variance or correlation models. Recent investigations, including our own simulation studies in Section 6, suggest that the robust sandwich estimate is not a panacea to variance misspecification [10].

The paper is organized as follows. First, we summarize the scientific model for benzene metabolism in Section 2. Then we describe two approaches for estimating the parameters of interest in Section 3 and describe potential bias in the maximum likelihood estimator (MLE) under certain model misspecification in Section 4. We present a novel analysis of the data from the Chinese study in Section 5, then present simulation studies in Section 6, and conclude with a discussion in Section 7.

2. A PHYSICAL MODEL RELATING BENZENE EXPOSURE TO ITS PRODUCTS

Recently, Rappaport *et al.* [11] applied a physiologically based toxicokinetic (PBT) model for volatile organic compounds (VOCs), including benzene, which is summarized in Figure 2. The

DOI: 10.1002/sim

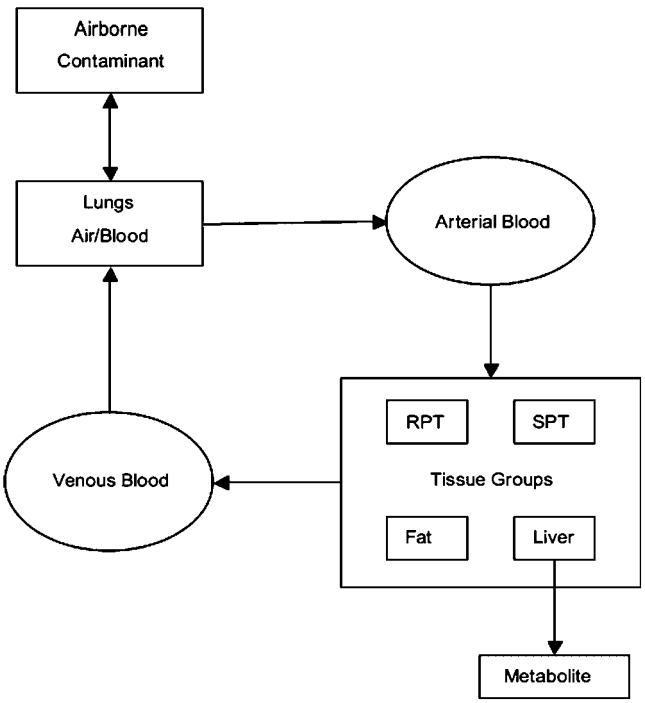

Figure 2. Physiologically based toxicokinetic model for volatile organic compounds. The four tissue groups include rapidly perfused tissues (RPT), slowly perfused tissues (SPT), fat, and liver.

model is accepted as a reasonable approximation for the uptake and elimination of VOCs in mammals. The input to the model is the airborne contaminant, which enters the lung at the alveolar ventilation rate. The chemical is absorbed into the arterial blood according to its blood–air partition coefficient and is subsequently distributed to four parallel tissue groups—rapidly perfused tissues (RPT), slowly perfused tissues (SPT), fat, and liver (the only metabolizing tissue)—at tissue-specific perfusion rates. The chemical exits the body passively, by exhalation, or by metabolism in the liver. In the case of benzene, the initial metabolite is benzene oxide. The rate (R) at which the chemical is metabolized in the liver is governed by Michaelis–Menten kinetics, i.e.

$$R = \frac{V_{\max}(C_{\mathrm{L}}/\lambda_{\mathrm{L}})}{K_{\mathrm{m}} + (C_{\mathrm{L}}/\lambda_{\mathrm{L}})} \tag{1}$$

where the liver-blood chemical concentration $(C_{\mathrm{L}}/\lambda_{\mathrm{L}})$ is expressed as the ratio of the concentration of the chemical in the liver to the tissue–blood partition coefficient, λ_{L}, $V_{\max}$ is the maximum rate of metabolism (mg/h), and K_{m} (mg/l) is the liver-blood concentration at which $R = V_{\max}/2$.

Under certain regularity conditions on the metabolic process, e.g. steady-state assumptions, a three-parameter Michaelis–Menten-like model provides an adequate summary of the contaminant–metabolite or exposure–biomarker relationship in the PBT model in Figure 2. For $i = 1, \ldots, m$, let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$ denote an $n_i \times 1$ vector of biomarker outcome measurements for the ith subject and X_i denote the (scalar) observed exposure level. Then, writing the conditional mean

DOI: 10.1002/sim

biomarker response for the ith subject, $E(\mathbf{Y}_i|X_i)=\mu_i(\boldsymbol{\beta})$, we assume the mean outcome vector is related to chemical exposure through the following model:

$$\mu_i(\boldsymbol{\beta})=\beta_0+\frac{\beta_1 X_i}{\beta_2+X_i} \tag{2}$$

where $\boldsymbol{\beta}=(\beta_0,\beta_1,\beta_2)'$ is a trivariate vector of population parameters. Different versions of (2) have been used in wide-ranging applications [12, 13]. In our application, we interpret β_1 approximately as the population mean of $V_{\max}$ across subjects and β_2 a 'surrogate' for the population mean of K_{m} across subjects. Under additional assumptions on body weight, tissue volumes, partition coefficients, and respiratory rate, it is possible to derive an unbiased estimator for the population mean $V_{\max}$ and K_{m} from β_1 and β_2, respectively [11]. The intercept, β_0, represents the background level of the mean biomarker for the ith subject, is an important parameter and cannot be assumed to be negligible. For this reason, we term $\mu_i(\boldsymbol{\beta})$ a 'Michaelis–Menten-like' model. When the intercept is exactly zero, then the mean outcome model $\mu_i(\boldsymbol{\beta})$ is precisely the Michaelis–Menten or Beverton–Holt model.

3. METHODS

3.1. A multiplicative lognormal approach

A common strategy for modelling outcomes with long-right tails is to assume that the natural logarithm of the outcome is normally distributed. Using the abbreviation μ_i for $\mu_i(\boldsymbol{\beta})$, we write the following multiplicative model [14–16]:

$$Y_{ij}=\mu_i\varepsilon^*_{ij},\quad j=1,\ldots,n_i \tag{3}$$

where $E(\varepsilon^*_{ij})=1$ for all j. It is well known that if W is a Gaussian random variable with mean a and variance b^2, then $E(e^W)=e^{a+b^2/2}$. Consequently, an alternative model formulation to (3) with additive errors [17–22] and possibly correlated outcomes is

$$\log\mathbf{Y}_i=\left(\log\mu_i-\frac{\tau^2}{2}\right)\mathbf{1}_{n_i}+\tau\boldsymbol{\Omega}_i^{1/2}(\boldsymbol{\alpha}^\dagger)\boldsymbol{\varepsilon}_i \tag{4}$$

for τ a scalar, $\boldsymbol{\varepsilon}_i$ an n_i-dimensional standard Gaussian random vector, and $\boldsymbol{\Omega}_i$ a n_i-dimensional square correlation matrix, and $\boldsymbol{\alpha}^\dagger$ are correlation parameters. The model in (4) adds a correction factor (i.e. $\tau^2/2$) to preserve the conditional mean model on the original scale, that is, $E(Y_{ij}|X_i)=\mu_i$. Also, for τ small,

$$(4)\approx(\log\mu_i)\mathbf{1}_{n_i}+\tau\boldsymbol{\Omega}_i^{1/2}(\boldsymbol{\alpha}^\dagger)\boldsymbol{\varepsilon}_i \tag{5}$$

Hence, the approximate model in (5) is a special case of the 'transform-both-sides' (TBS) methodology [23]. In model (5), the analyst assumes a particular *a priori* transformation of the original biomarker outcomes $\mathbf{Y}_i$ and mean μ_i which leads to residuals that achieve both normality and homoscedasicity and does not estimate the transformation *via* the Box–Cox family of transformations [24].

DOI: 10.1002/sim

3.2. An estimating equation approach

A different estimation strategy avoids transformations altogether, places no distributional assumptions on the errors, but requires a model for the conditional covariance $\text{cov}(\mathbf{Y}_i|X_i)$ as well as the conditional mean $E(\mathbf{Y}_i|X_i)$; this is referred to as the quasilikelihood and variance function method [9, 25]. For Y_{ij}, the jth biomarker outcome on subject i, we write

$$\text{var}(Y_{ij}|X_i) = \sigma^2 g^2(\mu_{ij}, A_i, \vartheta)$$

for a user-specified variance function $g(\cdot)$, fixed auxiliary variables A_i, and a finite number of variance parameters ϑ. In our analysis of the Chinese data in Section 5, we consider several variance functions, including the common power-of-the-mean model, $g_1(\mu_{ij}, A_i, \vartheta) = \mu_{ij}^{\vartheta}$, of which constant coefficient of variation (CV) models, such as gamma and lognormal, are special cases when the power $\vartheta = 1$. In addition, we let A_i be the indicator for unexposed subjects in the study, i.e. $A_i = 1$ for workers from the food-processing or flour plant and zero otherwise. To model the correlation among replicate biomarkers within a subject, one writes the multivariate conditional covariance $\text{cov}(\mathbf{Y}_i|X_i)$ as

$$\boldsymbol{\Sigma}_i = G_i^{1/2}(\boldsymbol{\beta}, \vartheta)\boldsymbol{\Omega}_i(\boldsymbol{\alpha})G_i^{1/2}(\boldsymbol{\beta}, \vartheta)$$

where $G_i(\boldsymbol{\beta}, \vartheta) = \text{diag}\{g^2(\mu_{i1}, A_i, \vartheta), \ldots, g^2(\mu_{n_i}, A_i, \vartheta)\}$, $\boldsymbol{\Omega}_i(\boldsymbol{\alpha})$ is an $n_i \times n_i$ square correlation matrix [8]. Note, the correlation parameters $\boldsymbol{\alpha}$ (no 'dagger') represent the correlation parameters of the biomarker outcomes on the original scale whereas $\boldsymbol{\alpha}^{\dagger}$ represent correlation of the transformed biomarker outcomes.

Let $\boldsymbol{\gamma}$ denote a q-vector of variance parameters, i.e. $\boldsymbol{\gamma} = (\sigma, \vartheta, \boldsymbol{\alpha})$. Then, the mean parameters $\boldsymbol{\beta}$ may be estimated by solving the following trivariate vector of estimating equations:

$$\mathbf{U}_{\beta}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{m} \mathbf{D}_i' \boldsymbol{\Sigma}_i^{-1}(\mathbf{Y}_i - \mu_i) = 0$$

where $\mathbf{D}_i = (\partial/\partial\boldsymbol{\beta})\mu_i$, the p-dimensional gradient of μ_i with respect to $\boldsymbol{\beta}$. Using a 'pseudo-likelihood' approach to variance function estimation [26–28], our variance parameters $\boldsymbol{\gamma}$ minimize the objective function

$$\text{PL} = \sum_{i=1}^{m} \text{PL}_i(\sigma, \vartheta, \boldsymbol{\alpha}) = \sum_{i=1}^{m} \log|\sigma^2\boldsymbol{\Sigma}_i| + \sigma^{-2}(\mathbf{Y}_i - \mu_i)'\boldsymbol{\Sigma}_i^{-1}(\mathbf{Y}_i - \mu_i) \tag{6}$$

Hence, our variance parameters simultaneously solve the estimating equations

$$\mathbf{U}_{\gamma}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{m} \left(\frac{\partial}{\partial\boldsymbol{\gamma}}\right) \text{PL}_i(\sigma, \vartheta, \boldsymbol{\alpha}) = 0$$

where the exact expressions depend upon user-specified variance functions and assumed correlation model for replicate biomarker outcomes within subjects. In practice, it is unusual to solve the whole system of estimating equations directly, i.e.

$$\begin{pmatrix} \mathbf{U}_{\beta}(\boldsymbol{\beta}, \boldsymbol{\gamma}) \\ \mathbf{U}_{\gamma}(\boldsymbol{\beta}, \boldsymbol{\gamma}) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

DOI: 10.1002/sim

Instead, a multi-stage estimation procedure is used [7, 29]. For example, in this paper, we iteratively solve the estimating equations $\mathbf{U}_\beta(\boldsymbol{\beta}, \boldsymbol{\gamma}) = 0$ and then minimize PL using numerical minimization techniques that do not require derivatives. (An exemplary step-by-step algorithm is given by Davidian and Giltinan [8, Chapter 2, pp. 34–36]) When the variance function $g(\cdot)$ and the correlation structure $\boldsymbol{\Omega}_i(\boldsymbol{\alpha})$ are specified correctly, $m^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges in distribution to a mean-zero, Gaussian random vector with covariance $\mathbf{V}$. A consistent estimator for $\mathbf{V}$ is $\widehat{\mathbf{V}}_{\mathrm{GLS}}$,

$$\mathbf{V}_{\mathrm{GLS}} = \hat{\sigma}^2 \left(\sum_{i=1}^{m} \mathbf{D}_i' \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_i \right)^{-1}$$

where $\hat{\sigma}$ is the first element in the solution vector to the estimating equation: $\mathbf{U}_\gamma(\boldsymbol{\beta}, \boldsymbol{\gamma}) = 0$ [6, 8]. To guard against possible misspecification of the variance and correlation models, a robust covariance estimator for $\mathbf{V}$ is $\widehat{\mathbf{V}}_{\mathrm{R}}$, where

$$\mathbf{V}_{\mathrm{R}} = \left(\sum_{i=1}^{m} \mathbf{D}_i' \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_i \right)^{-1} \left[\sum_{i=1}^{m} \mathbf{D}_i' \boldsymbol{\Sigma}_i^{-1} (\mathbf{Y}_i - \mu_i)(\mathbf{Y}_i - \mu_i)' \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_i \right] \left(\sum_{i=1}^{m} \mathbf{D}_i' \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_i \right)^{-1}$$

Under suitable regularity conditions, it is well known that $m^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges in distribution to a mean-zero, Gaussian random vector with variance $\mathbf{V}$ even if $\mathrm{cov}(\mathbf{Y}_i | X_i) \neq \sigma^2 \boldsymbol{\Sigma}_i$ [9]. Thus, we may obtain valid statistical inference for the mean parameters $\boldsymbol{\beta}$, assuming only that the mean model $\mu_i(\boldsymbol{\beta})$ is specified correctly.

4. SENSITIVITY OF THE LOGNORMAL MAXIMUM LIKELIHOOD

In this section, we conduct a sensitivity study of the lognormal MLE for the one-sample problem of estimating the mean $E(Y) = \mu$ under a constant CV model and then consider the regression problem. The purpose of this investigation is to understand how robust the parametric estimator for the mean $E(Y)$ is to incorrect distributional assumptions. An illustration of residual patterns consistent with the variance increasing like a power of the mean, i.e. μ_{ij}^{ϑ}, is given in Figure 3.

In Figure 3(a), the residual pattern is simulated from normally distributed data with a constant CV (i.e. $\vartheta = 1$). The residual pattern displayed in Figure 3(b) has power $\vartheta = 0.85$ although the pattern looks almost identical to that in panel (a). Here, we only wish to emphasize the point that residual patterns may appear to be consistent with a constant CV model when the true data generating process is one with non-constant CV. Estimating the mean parameters $\boldsymbol{\beta}$ with an incorrect variance function $g(\mu_{ij}, A_i, \vartheta)$ will generally result in a less precise estimate. Our MLE is based on a multiplicative lognormal model, not the additive normal MLE following the generalized nonlinear model framework [6]. The comparison considered here is germane to the analysis below and helpful to researchers where lognormal assumptions are often made (e.g. environmental health).

With some abuse of notation, let $Y_i, \ldots, Y_m$ be lognormally distributed, where

$$\log Y_i = \log \mu - \frac{\tau^2}{2} + \tau \varepsilon_i$$

where ε_i are iid mean-zero, Gaussian random variables and τ known. Assuming τ known, the MLE for μ is

$$\hat{\mu}_{\mathrm{LN}} = \exp(\tilde{T} + \tau^2/2) \tag{7}$$

DOI: 10.1002/sim

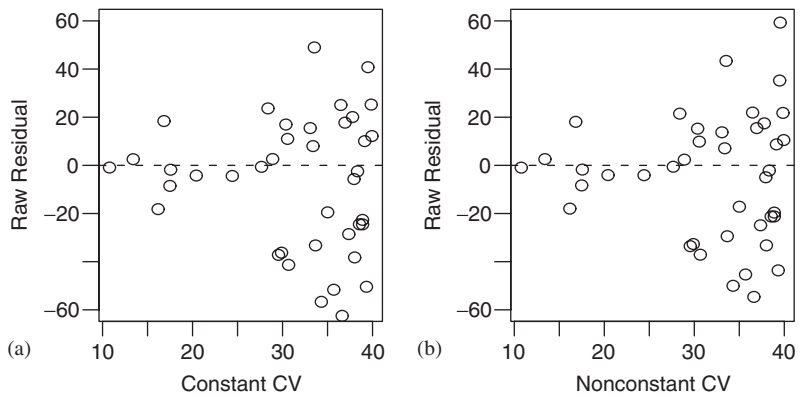

Figure 3. Illustration of residual patterns for simulated data generated from variance models with constant coefficient of variation (CV) *versus* non-constant CV. In both panels, the residuals are simulated as mean zero Gaussian random variables with variance $\{\sigma\mu_i(\boldsymbol{\beta})^{\vartheta}\}^2$, where $\mu_i(\boldsymbol{\beta})$ is the Michaelis–Menten model (i.e. $\beta_0 = 0$, $\beta_1 = 50$, $\beta_2 = 25$). In panel (a), we let $\sigma = 1$ and $\vartheta = 1$, whereas $\sigma = 1.5$ and $\vartheta = 0.85$ in panel (b). Incorrectly modelling the residuals in panel (b) using any constant CV model, either normally or lognormally distributed, will result generally in a loss of efficiency for the mean parameter estimates, $\hat{\boldsymbol{\beta}}$.

where $\tilde{T} = n^{-1}\sum_{i=1}^{n} \log Y_i$. By the law of large numbers and the continuity theorem, $\hat{\mu}_{\mathrm{LN}}$ converges in probability to

$$\mu_0 = \exp\{E(\log Y) + \tau^2/2\} \tag{8}$$

A direct calculation of $E(\hat{\mu}_{\mathrm{LN}})$, assuming the correct lognormal density, shows that $\mathrm{bias}(\hat{\mu}_{\mathrm{LN}}) = \exp(\tau^2/2m)$ or trivially that $\mu_0 = \mu$ in (8).

Remark

A closely related MLE to estimators considered in this paper is one given in Carroll and Ruppert [6, Chapter 2]. Assume that Y_i are iid normal with mean μ and variance $\sigma^2\mu^2$, σ known. Then the CV is constant and the MLE for μ is

$$\hat{\mu}_{\mathrm{N}} = \frac{T_1^2 + 4\sigma^2 T_2 - T_1}{2\sigma^2}$$

where $T_1 = m^{-1}\sum_{i=1}^{m} Y_i$ and $T_2 = m^{-1}\sum_{i=1}^{m} Y_i^2$ [6]. We note that these two MLEs, $\hat{\mu}_{\mathrm{LN}}$ and $\hat{\mu}_{\mathrm{N}}$ are very different, and we do not discuss $\hat{\mu}_{\mathrm{N}}$ in the sequel.

To consider how $\hat{\mu}_{\mathrm{LN}}$ might behave under slightly different models, we wish to express μ_0 in the moments and cumulants of Y. Define μ_k, $k = 2, \ldots,$ as the central moments of Y and κ_j, $j = 1, \ldots, n$ as the cumulants and rewrite τ^2 in (7) as $\mathrm{var}(\log Y)$. So, $\mu_0 = \mu B$ where

$$B = \exp\left\{\sum_{k=2}^{\infty} \frac{(-1)^{(k-1)}\mu_k}{k\mu^k} + \mathrm{var}(\log Y)/2\right\} \tag{9}$$

DOI: 10.1002/sim

Using a first-order Taylor-series approximation, $\mathrm{var}(\log Y) \approx \kappa_2/\kappa_1^2$; then, define

$$B^\dagger = \exp\left\{-\frac{\kappa_2}{2\kappa_1^2} + \frac{\kappa_3}{3\kappa_1^3} - \frac{\kappa_4 + 3\kappa_2^2}{4\kappa_1^4} + \frac{\kappa_2}{2\kappa_1^2}\right\}$$

$$= \exp\left\{\frac{\kappa_2^{3/2}\gamma_1}{3\kappa_1^3} - \frac{(\gamma_2+3)\kappa_2^2}{4\kappa_1^4}\right\} \tag{10}$$

where γ_1, γ_2 are the skewness and kurtosis excess coefficients, respectively. In the generalized nonlinear model with variance proportional to some power of the mean, say, $\kappa_2 = \sigma^2\kappa_1^{2(1+\vartheta)}$, and replacing κ_1 with μ, then

$$B^\dagger = \exp\left[\frac{\sigma^3\mu^{3\vartheta}}{12}\{4\gamma_1 - 3(\gamma_2+3)\sigma\mu^\vartheta\}\right] \tag{11}$$

Here, we see that $B^\dagger$ is a complicated function of the mean μ, variance parameters σ and ϑ, skewness γ_1 and kurtosis excess γ_2. It is not obvious that $B^\dagger$ will be 'close' to one so that $\hat{\mu}_{\mathrm{LN}}$ will have small bias under the general power-of-the-mean model. For the multiplicative lognormal model under which we derived the MLE, $\hat{\mu}_{\mathrm{LN}}$, B is exactly one and $B^\dagger$ is 'close' to one (see Appendix). Also, if $Y_1, \ldots, Y_m$ are iid normally distributed with constant CV (i.e. $\vartheta = 0$, $\gamma_1 = \gamma_2 = 0$), then $B^\dagger = \exp(-3\sigma^4/4) \approx 1 - 3\sigma^4/4$. Hence, we might expect the lognormal MLE to fit well normally distributed data with small, constant CV but to be biased when the variance of Y does not behave like a constant power of μ.

We now turn to the three-parameter regression setting. In the regression setting under the multiplicative lognormal model in (3)–(4), the score vector for $\boldsymbol{\beta}$ is

$$\widetilde{\mathbf{U}}_\beta(\boldsymbol{\beta}) = \sum_{i=1}^m \left(\frac{\partial}{\partial\boldsymbol{\beta}}\log\mu_i\right)\boldsymbol{\Omega}_i^{-1}(\boldsymbol{\alpha}^\dagger)\left(\log\mathbf{Y}_i - \log\mu_i + \frac{\tau^2}{2}\right)$$

Observe that $\widetilde{\mathbf{U}}_\beta(\boldsymbol{\beta})$ will not, in general, have mean zero under the truth if $E\mathbf{Y}_i = \mu_i$ as we noted above. However, in a constant CV model, further simplification of $\widetilde{\mathbf{U}}_\beta(\boldsymbol{\beta})$ is revealing. Using notation from Section 3.2,

$$\widetilde{\mathbf{U}}_\beta(\boldsymbol{\beta}) = \sum_{i=1}^m \mathbf{D}_i\mathbf{M}_i^{-1/2}\boldsymbol{\Omega}_i^{-1}(\boldsymbol{\alpha}^\dagger)\left(\log\mathbf{Y}_i - \log\mu_i + \frac{\tau^2}{2}\right)$$

$$= \sum_{i=1}^m \mathbf{D}_i\mathbf{M}_i^{-1/2}\boldsymbol{\Omega}_i^{-1}(\boldsymbol{\alpha}^\dagger)\mathbf{M}_i^{-1/2}\mathbf{Z}_i^*$$

$$= \sum_{i=1}^m \mathbf{D}_i\boldsymbol{\Gamma}_i^{-1}(\boldsymbol{\alpha}^\dagger)\mathbf{Z}_i^*$$

where $\mathbf{M}_i = \mathrm{diag}(\mu_{i1}^2, \ldots, \mu_{in_i}^2)$, $\boldsymbol{\Gamma}_i(\boldsymbol{\alpha}^\dagger) = \mathbf{M}_i^{1/2}\boldsymbol{\Omega}_i(\boldsymbol{\alpha}^\dagger)\mathbf{M}_i^{1/2}$, and $\mathbf{Z}_i^* = \mathbf{M}_i^{1/2}(\log\mathbf{Y}_i - \log\mu_i + \tau^2/2)$. For power-of-the-mean variance function with $\vartheta = 1$, $g(\mu_{ij}, A_i, \boldsymbol{\vartheta}) = \mu_{ij}$ and $\mathbf{G}_i = \mathbf{M}_i$. Thus, $\widetilde{\mathbf{U}}_\beta(\boldsymbol{\beta})$ and $\mathbf{U}_\beta(\boldsymbol{\beta}, \boldsymbol{\gamma})$ have the same weight, $\mathbf{D}_i\boldsymbol{\Sigma}_i^{-1}$ (or equivalently, $\mathbf{D}_i\boldsymbol{\Gamma}_i^{-1}$) on each contribution to the score. Continuing in a heuristic manner, when $\mathbf{Y}_i$ and μ_i are small, then $(\log\mathbf{Y}_i - \log\mu_i + \tau^2/2)$

DOI: 10.1002/sim

is approximately $c(\mathbf{Y}_i - \mu_i)$, c a constant. In the other extreme, when $\mathbf{Y}_i$ and μ_i are large, the weight $\mathbf{D}_i\boldsymbol{\Sigma}_i^{-1}$ is small due to large variance (*via* $\boldsymbol{\Sigma}_i$). Consequently, one may expect the ith contributions to the score functions to be on the same order despite modest differences in the raw residuals, $(\log \mathbf{Y}_i - \log \boldsymbol{\mu}_i)$ and $(\mathbf{Y}_i - \boldsymbol{\mu}_i)$, respectively. Hence, on this heuristic level, one might expect the two score functions $\widetilde{\mathbf{U}}_\beta(\boldsymbol{\beta})$ and $\mathbf{U}_\beta(\boldsymbol{\beta}, \boldsymbol{\gamma})$ to yield similar point estimates under any constant CV model, even, say, under an additive error model (see Section 6) rather than a multiplicative one (3) as long as the CV is small. A formal study of bias in $\widetilde{\mathbf{U}}_\beta(\boldsymbol{\beta})$ under different misspecified models involves calculating $E_{F_0}\{\widetilde{\mathbf{U}}_\beta(\boldsymbol{\beta})\}$, where F_0 is the true outcome distribution (or conditional distribution given the exposure X_i) and E_{F_0} the expectation under F_0. In Section 6, we explore the bias of $\widetilde{\mathbf{U}}_\beta(\boldsymbol{\beta})$ under misspecified models through simulation studies.

5. ANALYSIS OF BENZENE DATA

Understanding the relationships between environmental exposures and subsequent biomarkers or health-related outcomes continues to challenge scientists. Benzene is one of several contaminants of keen interest because it is a known carcinogen and is ubiquitous in the environment. Although scientists have studied benzene for more than a century, little is actually known about the exposure–disease relationship. A better understanding of benzene metabolism will, in part, lead to a better understanding of the mechanism which drives the exposure–disease relationship. The main goal of the Chinese study was to investigate how the body metabolizes benzene. For this, a total of 133 exposed and 51 control factory workers in Tianjin, China, were surveyed for five weeks. The exposed subjects worked at a glue, shoe, or sporting goods factory while the control subjects worked at a nearby food-processing or flour plant. All exposed subjects wore personal monitors to measure the individual exposure levels on several days prior to blood collection (at weekly intervals). Between four and six air measurements (that is, exposure measurements) per individual were collected. For our data analysis in this paper, we use the average of the air measurements to denote a subject's shift-long exposure. A thorough description of the study appears elsewhere [1].

While the investigators collected multiple surrogate exposure measurements from all exposed subjects, only a fraction of them were included in a substudy where replicate biomarker measurements were also collected on three subsequent Mondays. Specifically, 11 of 133 exposed workers were included in the substudy and three replicate biomarker outcomes were observed for these workers. One biomarker outcome was collected for every other worker in the study. A scatterplot of the data is given in Figure 4 and analytic results using methods described in Section 3 are displayed in Tables I and II.

Our first analysis of the Chinese data assumes that $\mu_i(\boldsymbol{\beta})$ follows the three-parameter Michaelis–Menten model in (2) and that the power-of-the-mean variance function, $g(\mu_{ij}, A_i, \boldsymbol{\vartheta}) = \mu_{ij}^{\vartheta}$. We compare parameter estimates using GEE to estimates using the lognormal MLE and consider three different correlation models for replicate biomarker outcomes within a subject: uncorrelated; exchangeable; and autoregressive with lag one (AR1). As shown in Table I, parameter estimates are similar whether one uses the multiplicative lognormal MLE or GEE. We also see that the MLE yields smaller SEEs than the GLS SEEs for GEE. Interestingly, the robust SEEs are about 20% smaller than the GLS SEEs for $\hat{\beta}_1$ and $\hat{\beta}_2$. Finally, as we move across Table I from left to right, although the mean parameter estimates $\hat{\boldsymbol{\beta}}$ change very little, we find the correlation among replicate observations to be very strong. In fact, because the replicate biomarker outcomes

DOI: 10.1002/sim

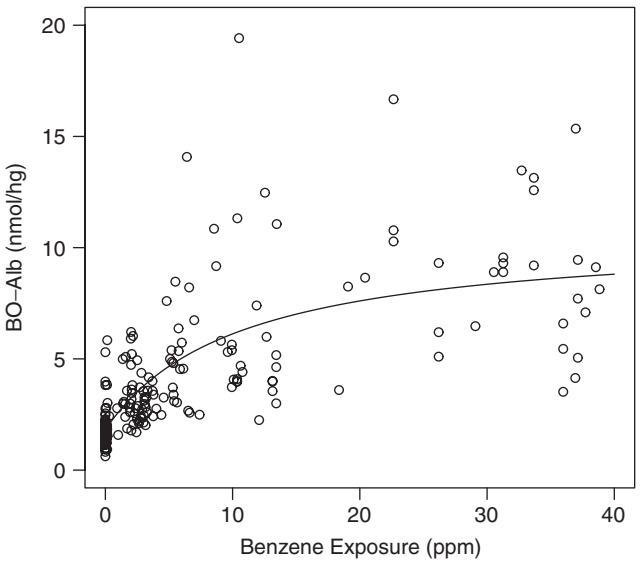

Figure 4. Scatter plot of observed benzene albumin (BO–Alb, nmol/hg) adduct levels as function of shift-long benzene exposure (ppm) in Chinese study. The Chinese study includes data on 184 factory workers, 51 of whom are regarded as controls. A total of 11 out of 133 exposed workers had three replicate biomarker outcomes for a grand total of 206 observations.

Table I. Parameter estimates assuming simple three-parameter Michaelis–Menten-like model.

	Independent		Exchangeable		Autoregressive	
Method	$\beta_0, \beta_1, \beta_2$	σ, ϑ	$\beta_0, \beta_1, \beta_2$	$\sigma, \vartheta, \alpha$	$\beta_0, \beta_1, \beta_2$	$\sigma, \vartheta, \alpha$
MLE	0.62, 2.18, 2.43	0.41†, —	0.62, 2.21, 2.43	0.41†, —, 0.67	0.62, 2.21, 2.43	0.41†, —, 0.48
	(0.05, 0.15, 0.28)		(0.05, 0.17, 0.30)		(0.05, 0.17, 0.30)	
GEE[a]	0.62, 2.21, 2.46	0.47, 1.0*	0.62, 2.21, 2.46	0.47, 1.0*, 0.78	0.62, 2.21, 2.46	0.47, 1.0*, 0.84
GLS	(0.06, 0.17, 0.31)		(0.06, 0.20, 0.35)		(0.06, 0.20, 0.35)	
Robust	(0.06, 0.17, 0.31)		(0.06, 0.17, 0.28)		(0.06, 0.17, 0.28)	
GEE[b]	0.62, 2.21, 2.46	0.45, 1.02	0.62, 2.21, 2.46	0.43, 1.06, 0.80	0.62, 2.21, 2.46	0.44, 1.05, 0.85
GLS	(0.06, 0.17, 0.32)		(0.06, 0.21, 0.35)		(0.06, 0.21, 0.36)	
Robust	(0.06, 0.17, 0.30)		(0.06, 0.17, 0.28)		(0.06, 0.17, 0.28)	

*ϑ fixed and set equal to 1.0.

†Parameter estimate for τ as in (4).

Note: Parameter estimates *via* lognormal maximum likelihood (MLE) and generalized estimating equations (GEE) with power-of-the-mean variance function are presented. GEE[a] makes the constant coefficient of variation model assumption (i.e. $\vartheta = 1$) whereas GEE[b] does not make that assumption and treats ϑ as an unknown parameter. We interpret β_0 as the ambient level of mean biomarker and β_1, β_2 approximately as surrogates for the population mean $V_{\max}$ and K_{m}, respectively. Three different assumptions about the correlation among repeated observations for a given subject in the Chinese substudy are given across the top of the table.

were observed over three subsequent Mondays and BO–Alb is known to have a half-life of three weeks, then it is sensible to assume that the correlated adduct levels follow an AR(1) correlation model.

DOI: 10.1002/sim

Table II. Parameter estimates assuming three-parameter Michaelis–Menten-like mean model and alternative variance functions.

	Independent				Autoregressive			
$g(\cdot)$	$\beta_0, \beta_1, \beta_2$	σ	ϑ	PL	$\beta_0, \beta_1, \beta_2$	σ, α	ϑ	PL
g_1	0.62, 2.21, 2.46	0.45	1.02	412.65	0.62, 2.21, 2.46	0.44, 0.85	1.05	395.14
GLS	(0.06, 0.17, 0.32)				(0.06, 0.21, 0.36)			
Robust	(0.06, 0.17, 0.30)				(0.06, 0.17, 0.28)			
g_2	0.61, 2.21, 2.45	0.50	1.62, 0.96	412.18	0.61, 2.26, 2.48	0.48, 0.84	1.70, 1.00	394.91
GLS	(0.06, 0.17, 0.31)				(0.06, 0.21, 0.35)			
Robust	(0.06, 0.17, 0.31)				(0.06, 0.17, 0.28)			
g_3	0.64, 2.27, 2.72	1.98	0.53, 0.28	405.08	0.64, 2.31, 2.75	1.36, 0.84	0.77, 0.51	389.11
GLS	(0.07, 0.17, 0.30)				(0.07, 0.22, 0.35)			
Robust	(0.06, 0.19, 0.31)				(0.06, 0.19, 0.31)			
g_4	0.65, 2.27, 2.72	2.75	0.38, 0.03	405.40	0.64, 2.32, 2.75	2.31, 0.83	0.45, 0.06	389.77
GLS	(0.07, 0.17, 0.30)				(0.07, 0.22, 0.34)			
Robust	(0.06, 0.19, 0.31)				(0.06, 0.19, 0.31)			

Note: A generalized estimating equations (GEE) strategy is employed with two different assumptions about the correlation among repeated observations. Different variance functions $g(\mu_{ij}, A_i, \vartheta)$ are given in (12)–(15).

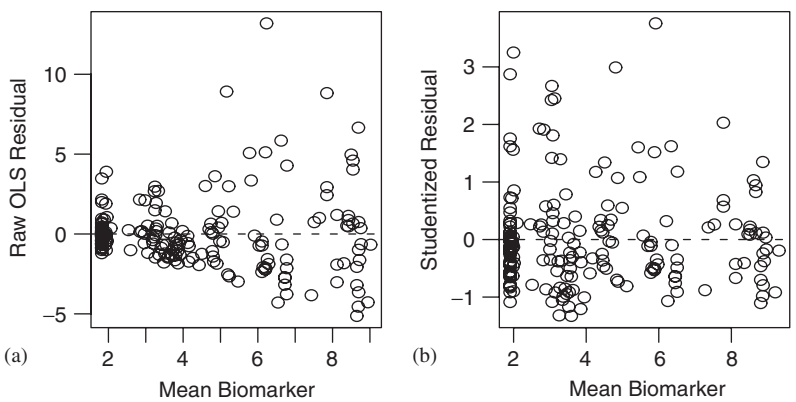

Figure 5. Residual plots from ordinary (OLS) and generalized least-squares (GLS) analyses. In panel (a), we plot the raw fitted residual $r_{ij} = Y_{ij} - \hat{\mu}_{ij}$ when $\hat{\mu}_{ij} = \mu_{ij}(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}})$ and $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ is the OLS estimate fit to the Chinese data. Note the residual pattern in panel (a) appears to be constant for mean biomarker levels μ_{ij} less than about 4 ppm and then increase as a function of μ_{ij}. We plot the weighted or 'studentized' residual, $r_{ij}/g_3(\hat{\mu}_{ij}, A_i, \vartheta)$, in panel (b), where $\hat{\mu}_{ij} = \mu_{ij}(\hat{\boldsymbol{\beta}}_{\mathrm{GLS}})$ for the generalized least-squares estimate using variance function $g_3(\cdot)$ defined in Section 5.

Next, we consider more general variance functions motivated by inspection of the residual plot from an ordinary least-squares fit in Figure 5(a). We note that for predicted mean biomarker levels, the error variance remains relatively constant up to some level, say L, but then exhibits a 'megaphone' effect, for predicted values larger than L. Initially, we assume that the level L is an effect of the study design, namely, the presence of a control group and then try related variance

DOI: 10.1002/sim

functions. We repeat our analysis of the Chinese data using the Michaelis–Menten conditional mean model in (2) but now consider the following four variance functions, including the one-parameter, power-of-the-mean variance function, listed as $g_1(\cdot)$:

$$g_1(\mu_{ij}, A_i, \vartheta) = \mu_{ij}^{\vartheta} \tag{12}$$

$$g_2(\mu_{ij}, A_i, \vartheta) = A_i\vartheta_1 + (1 - A_i)\mu_{ij}^{\vartheta_2} \tag{13}$$

$$g_3(\mu_{ij}, A_i, \vartheta) = I(\mu_{ij} < L)\vartheta_1 + I(\mu_{ij} \geqslant L)\mu_{ij}^{\vartheta_2} \tag{14}$$

$$g_4(\mu_{ij}, A_i, \vartheta) = I(\mu_{ij} < L)\vartheta_1 + I(\mu_{ij} \geqslant L)\exp(\vartheta_2\mu_{ij}) \tag{15}$$

where $I(\cdot)$ is an indicator function. Since $\mu_{ij}^{\vartheta_2} = \exp(\vartheta_2 \log \mu_{ij})$, then $g_4(\cdot)$ may be regarded as $g_3(\cdot)$ with ϑ_2 simply estimated on a different scale. In any given data set, there may be small sample differences between the two variance functions or substantive reasons to choose one variance function over another. The results of our extended analyses using the new variance functions are reported in Table II. Using the pseudo-likelihood (PL) as a criterion for model fit (smaller is better), we find that the variance function which assumes a constant variance of the biomarker outcomes in the control subjects and then increases like the power-of-the-mean in exposed subjects yields no significant advantage over the simpler one-parameter variance function $g_1(\cdot)$. However, using either $g_3(\cdot)$ or $g_4(\cdot)$ with $L = 4$ does show a substantial decrease (about a seven unit decrease for an assumed independent correlation model and about a six unit decrease for an assumed autoregressive correlation model) in the PL objective function. While a decrease in objective function PL does not constitute a formal hypothesis test of model improvement, we interpret this decrease in the PL as empirical evidence that $g_3(\cdot)$ and $g_4(\cdot)$ offer a better summary of the residual variance than $g_1(\cdot)$.

6. SIMULATION STUDIES

In this section, we conduct several simulation studies to examine the small sample properties of the lognormal MLE compared with the GEE approach under different response distributions. We use two statistical models in our investigations: a multiplicative error model in (3) and an additive error model:

$$\mathbf{Y}_i = \mu_i(\boldsymbol{\beta})\mathbf{1}_{n_i} + \sigma\boldsymbol{\Sigma}_i^{1/2}\boldsymbol{\varepsilon}_i \tag{16}$$

where $\mu_i(\boldsymbol{\beta})$ is the three-parameter Michaelis–Menten-like model (2), $\boldsymbol{\Sigma}_i = \mathbf{G}_i^{1/2}\mathbf{R}_i(\boldsymbol{\alpha})\mathbf{G}_i^{1/2}$ with $g(\mu_{ij}, A_i, \vartheta) = \mu_{ij}^{\vartheta}$, $E(\boldsymbol{\varepsilon}_i) = 0$ and $E(\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i') = I_{n_i}$. We begin our investigations with the multiplicative lognormal model in (3).

We begin by simulating exposure X_i, $i = 1, \ldots, m_e = 50$, from a uniform(1, 100) distribution. Then, in the case of lognormal data, we generate $n_i = 3$ iid normal errors, ε_{i1}, ε_{i2}, ε_{i3}, and calculate the outcome according to the model

$$\log \mathbf{Y}_i = \left[\log\{\mu_i(\boldsymbol{\beta})\} - \frac{\tau^2}{2}\right]\mathbf{1}_{n_i} + \tau\mathbf{R}_i^{1/2}(\boldsymbol{\alpha})\boldsymbol{\varepsilon}_i$$

DOI: 10.1002/sim

where $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3})'$, $\mathbf{R}_i(\boldsymbol{\alpha}) = (r_{jk})$, $r_{jk} = \alpha^{|j-k|}$ and $\tau = \{\log(1 + \mathrm{CV}^2)\}^{1/2}$ for a user-defined CV. In addition to lognormal outcomes, we also consider outcomes following model (16) with normally and exponentially distributed errors. For exponential errors, one initially generates $\varepsilon_{ij}^{\dagger} \sim \mathrm{Exp}(1)$ and then defines $\varepsilon_{ij} = \varepsilon_{ij}^{\dagger} - 1$. These new *translated exponential* errors will satisfy the conditions of model (16) but will have non-zero skewness ($\gamma_1 = 2$) as opposed to normal errors which have skewness coefficient equal to zero. For model (16), the CV for variance function $g_1(\cdot)$ is $\sigma\mu_{ij}^{\vartheta-1}$. When $\vartheta = 1$, we return to outcomes with constant CV, such as lognormal or gamma distributed outcomes.

Remark

In order to apply the lognormal MLE to our simulated data from model (16), some care must be taken to ensure that the normally distributed outcomes are positive random variables. Let $F_Y(t) = \Pr(Y_{ij} \leqslant t)$, then in model (16) we have

$$F_Y(0) = \Pr\left\{\varepsilon_{ij} \leqslant -\frac{\mu_{ij}}{\sigma g(\mu_{ij}, A_i, \mathbf{Z}_i)}\right\} = \Pr(\varepsilon_{ij} \leqslant -\sigma^{-1}\mu_{ij}^{1-\vartheta}) = \Phi(-\sigma^{-1}\mu_{ij}^{1-\vartheta})$$

where the second-to-last equality follows from $g(\mu_{ij}, A_i, \mathbf{Z}_i) = g_1(\cdot)$ and the last equality follows when the errors ε_{ij} are normally distributed. For $\vartheta = 1$, if σ is small enough, then few if any observations in a Monte Carlo data set will have negative response (e.g. $\sigma = 0.3$ implies $F_Y(0) = 0.00043$). In practice, if a simulated response was negative, we set it equal to a small positive number.

With no control subjects, β_0 is weakly identified for modest sample sizes. Hence, we also generate responses for $m_c = 25$ subjects with $n_i = 1$ replicate under the same conditional mean model, which agrees with the design of the Chinese study. Adding the control subjects to our simulation implies that we have a total of $N = 175$ observations from $m_c = 25$ control and $m_e = 50$ exposed subjects.

Our simulation study results are summarized in Tables III and IV. These results include summary statistics from a 3×3 factorial design, that is, three estimators—the MLE, GEE with constant CV (i.e. $\vartheta = 1$ fixed) and GEE with variance function $g_1(\cdot)$—and three working correlation matrices, $\boldsymbol{\Omega}_i(\boldsymbol{\alpha})$, for replicate outcomes within a subject—independence, exchangeable ($\omega_{jk} = \alpha, \ j \neq k$) and autoregressive ($\omega_{jk} = \alpha^{|j-k|}$). For brevity and clarity, we only present results which show meaningful differences between summary statistics. The Monte Carlo bias of parameter estimates is presented in Table III. Here, we find that for lognormal data following the multiplicative model (3), the Monte Carlo bias in the MLE is small when the CV is small. When the CV is large, the MLE exhibits small sample bias. Interestingly, the sample size is sufficiently large for GEE to obtain small bias in $\hat{\beta}_1$ and $\hat{\beta}_2$ regardless of the magnitude of CV in the multiplicative lognormal model. However, the GEE estimates of $\hat{\beta}_0$ tend to be too small on average. For normally distributed data following model (16), the lognormal MLE exhibits small bias even when the CV is non-constant. When the outcomes are exponential, the MLE is biased while GEE still provides consistent parameter estimates.

In Table IV, we report the Monte Carlo standard deviation of parameter estimates (SE) and the average of SEEs. When the CV is constant, GEE produces results that are very similar, whether or not a constant CV is assumed. The only simulations where we present both types of GEE estimators is when the true CV is not constant. First, note that when the outcomes follow the

DOI: 10.1002/sim

Table III. Monte Carlo estimates of parameter bias in three-parameter Michaelis–Menten-like model.

			β_0		β_1		β_2	
$F_Y(\cdot)$	CV	$\mathbf{\Omega}_i(\boldsymbol{\alpha})$	MLE	GEE	MLE	GEE	MLE	GEE
LN	0.3	Ind	0.004	0.003	0.018	0.021	0.032	0.040
		AR1	−0.012	−0.015	0.014	0.017	0.023	0.030
LN	1.0	Ind	0.038	−0.241	0.157	0.055	0.123	−0.024
		AR1	0.033	−0.235	0.126	0.023	0.096	−0.045
N	0.3	Ind	0.105	−0.004	0.038	0.014	0.033	0.025
		AR1	0.103	−0.005	0.038	0.015	0.031	0.025
N	$0.3\mu_{ij}^{-2/5}$	Ind	−0.042	−0.003	0.013	0.006	0.017	0.014
		AR1	−0.040	−0.003	0.014	0.007	0.016	0.014
E	$2.5\mu_{ij}^{-1/2}$	Ind	−1.872	0.172	0.481	0.031	−0.165	−0.017
		AR1	−1.723	0.154	0.532	0.024	−0.147	−0.024

Note: Three different distributions for the errors, $F_Y(\cdot)$, are lognormal (LN), normal (N), and translated exponential (E) with two different working correlation models: independence (Ind) and autoregressive with lag one (AR1) for equally spaced intervals.

lognormal multiplicative model (3) with small CV, the variance in parameter estimates between the MLE and GEE is about the same. However, for large CV, the lognormal MLE is more efficient than GEE. For the constant CV model with normal errors, GEE standard errors are 25–33 per cent smaller. Unfortunately, attempts to increase this constant CV model with normal errors are limited by the increase in the probability of a negative response. When the CV is small but non-constant, the difference between the standard estimates for MLE *vis-a-vis* GEE is dramatically reduced. When the outcomes follow our exponential distribution, the lognormal MLE has small standard error but is an inconsistent estimator for the truth (see Table III).

Next, we wish to compare the average of model-based standard error estimates (SEEs[c]) and robust standard error estimates (SEEs[d]). Note that in our simulation study when the true CV is non-constant and one assumes an uncorrelated working correlation model for replicate outcomes within a subject, GEE[b] uses the correct variance function but incorrect correlation model while GEE[a] is wrong on both counts. Hence, we can gain some insight into how well the robust standard error estimates behave; not only when the correlation is misspecified, but also under the potentially more severe problem of when the variance function *and* correlation structure are both misspecified. For normal errors with $\text{CV} = 0.3\mu_{ij}^{-2/5}$, using the robust SEEs in GEE[a] does seem to behave as one would hope; that is, to increase the SEE from 0.17 to 0.21 (for β_0) and thereby match the true standard error very closely. We see an even more dramatic average increase from model-based to robust standards, 0.67–1.03, when the errors are exponentially distributed but both the variance function and correlation structure are misspecified. For β_1 and β_2, the robust error estimates still appear to be too small, on average. However, in many simulation scenarios, the average robust SEE with incorrect variance function is similar to the average robust SEE with correct variance function.

DOI: 10.1002/sim

Table IV. Standard error estimates in simulation study of three-parameter Michaelis–Menten-like model.

				β_0			β_1			β_2		
$F_Y(\cdot)$	CV	$\boldsymbol{\Omega}_i(\boldsymbol{\alpha})$	Method	SE	SEE*	SEE†	SE	SEE*	SEE†	SE	SEE*	SEE†
LN	0.3	Ind	MLE	0.30	0.28		0.13	0.08		0.34	0.25	
			GEE‡	0.31	0.29	0.29	0.13	0.09	0.11	0.35	0.25	0.32
		AR1	MLE	0.31	0.28		0.13	0.11		0.35	0.32	
			GEE‡	0.31	0.29	0.28	0.13	0.11	0.11	0.35	0.31	0.31
LN	1.0	Ind	MLE	0.87	0.81		0.56	0.46		1.20	0.99	
			GEE‡	1.37	0.86	0.83	0.81	0.47	0.59	1.36	1.15	1.59
		AR1	MLE	0.87	0.82		0.46	0.45		1.08	1.12	
			GEE‡	1.37	0.87	0.83	0.76	0.46	0.46	1.27	1.22	1.29
N	0.3	Ind	MLE	0.45	0.37		0.16	0.11		0.44	0.32	
			GEE‡	0.32	0.29	0.29	0.12	0.08	0.11	0.34	0.25	0.32
		AR1	MLE	0.43	0.37		0.16	0.14		0.43	0.41	
			GEE‡	0.32	0.29	0.29	0.12	0.11	0.11	0.34	0.32	0.32
N	$0.3\mu_{ij}^{-2/5}$	Ind	MLE	0.24	0.18		0.07	0.05		0.21	0.15	
			GEE‡	0.23	0.17	0.21	0.07	0.05	0.06	0.20	0.15	0.19
			GEE§	0.23	0.21	0.21	0.07	0.05	0.06	0.20	0.15	0.19
		AR1	MLE	0.24	0.18		0.07	0.07		0.21	0.21	
			GEE‡	0.23	0.17	0.21	0.07	0.06	0.06	0.20	0.18	0.19
			GEE§	0.23	0.21	0.21	0.07	0.05	0.06	0.20	0.17	0.19
E	$2.5\mu_{ij}^{-1/2}$	Ind	MLE	0.92	0.56		0.30	0.22		0.82	0.64	
			GEE‡	1.16	0.67	1.03	0.41	0.22	0.26	1.01	0.70	0.87
			GEE§	1.16	1.02	1.06	0.37	0.21	0.28	0.91	0.65	0.84
		AR1	MLE	0.97	0.63		0.29	0.34		0.79	1.00	
			GEE‡	1.15	0.71	1.04	0.41	0.32	0.26	0.96	0.93	0.84
			GEE§	1.11	1.00	1.05	0.34	0.27	0.26	0.87	0.82	0.80

*Average of standard error estimates from $\hat{\mathbf{V}}_{\mathrm{GLS}}$.

†Average of robust standard error estimates from $\hat{\mathbf{V}}_{\mathrm{R}}$.

‡GEE with constant CV.

§GEE with ϑ unknown.

Note: True responses follow one of three different distributions, $F_Y(\cdot)$, lognormal (LN), normal (N), and translated exponential (E) with autoregressive errors. Table entries are derived from 500 Monte Carlo data sets using the lognormal MLE and GEE under two different assumed correlation structures for replicate outcomes within a subject: independence and autoregressive with lag one.

7. DISCUSSION

In this article, we presented a nonlinear Michaelis–Menten-like regression model for relating chemical exposure to a metabolism-related biomarker. We then applied this model to relate benzene

DOI: 10.1002/sim

exposure to benzene oxide adducts in a Chinese population. We use both parametric and semiparametric methods for drawing statistical inference in the setting where we have a general nonlinear model with few repeated biomarker measurements. In the Chinese data, we find that a simple conditional mean model (2) with variance function $g_3(\cdot)$ fits well. The chosen variance function is defined piecewise: constant for mean biomarker levels less than a limit $L = 400$ pmol/g and then like a power-of-the mean for mean biomarker levels greater than L.

We conducted small sample studies to investigate two questions of interest: sensitivity of the multiplicative lognormal MLE to model misspecification and sensitivity of the robust SEEs to misspecifying the variance function $g(\cdot)$ and/or correlation model $\mathbf{\Omega}_i(\boldsymbol{\alpha})$. When the response vectors follow the generalized nonlinear model in (16) with constant, small CV, or even mild deviations from the constant CV model, we find the bias in the lognormal MLE to be small. However, when the CV is not constant, parameter estimates may be severely biased. Our simulation studies suggest that even when the multiplicative lognormal model is correct, the generalized nonlinear model (16) will have better small sample properties when the CV is relatively large! Thus, based on our investigations, even investigators who believe strongly that their data follow the multiplicative lognormal model should consider the generalized nonlinear model (16) in small samples. As for misspecifying variance and/or correlation models in nonlinear models *via* GEE, we conclude that the robust SEEs occasionally correct for covariance misspecification in small samples. Certainly, more technical work into the behaviour of the robust variance estimate is needed.

In an earlier analysis of the Chinese data, Johnson *et al.* [30] used a nonlinear, mixed effects model with a different conditional mean model $\mu_i(\beta)$. Our experience with exposure–biomarker modelling suggests that the Michaelis–Menten-like model (2) is more suitable than the one proposed earlier for population-based studies of a contaminant and its relationship to subsequent biomarkers. Moreover, the parameter estimates from the Michaelis–Menten-like model have an approximate physical interpretation of interest to environmental epidemiologists. The mixed model framework allowed Johnson *et al.* to estimate subject-specific trajectories at low doses for 11 exposed subjects with replicate biomarker outcomes. Here, we focus on estimating the population-averaged exposure–biomarker relationship rather than on individual trajectories. Had multiple biomarkers been collected over a range of benzene exposure levels, then several methods using nonlinear mixed models would have been available. Given the nature and timing of the observed data, the population-averaged approach seemed most appropriate without making strong assumptions about the underlying data generating process. On the whole, we believe the data analysis in this paper is a better summary of the exposure–biomarker relationship in the Chinese population, particularly for subjects at low doses.

APPENDIX A: ADDITIONAL ALGEBRA FOR THE LOGNORMAL MLE

If $\log Y$ is normal with mean $\log \mu - \tau^2/2$ and variance τ^2, then the central moments are given by

$$\mu_2 = \mu^2(e^{\tau^2} - 1)$$

$$\mu_3 = \mu^3(e^{\tau^2} - 1)^2(e^{\tau^2} + 2)$$

$$\mu_4 = \mu^4(e^{\tau^2} - 1)^2(e^{4\tau^2} + 2e^{3\tau^2} + 3e^{2\tau^2} - 3)$$

DOI: 10.1002/sim

The coefficient of variation is given by

$$\mathrm{CV}(\text{lognormal}) = \sqrt{e^{\tau^2} - 1}$$

It then follows that the cumulants are given by $\kappa_1 = \mu$, $\kappa_2 = \mu_2$, $\kappa_3 = \mu_3$, and

$$\kappa_4 = \mu^4(e^{\tau^2} - 1)^2(e^{4\tau^2} + 2e^{3\tau^2} + 3e^{2\tau^2} - 6)$$

In addition, the skewness and kurtosis ($\mu_4/\mu_2^2 - 3$) are given by

$$\gamma_1 = (e^{\tau^2 - 1})^{1/2}(e^{\tau^2} + 2)$$

$$\gamma_2 = (e^{4\tau^2} + 2e^{3\tau^2} + 3e^{2\tau^2} - 6)$$

A second-order Taylor-series expansion for $\mathrm{var}(\log Y)$ where Y follows the above lognormal distribution is given by the following expression:

$$\mathrm{var}(\log Y) = (e^{\tau^2} - 1)[1 + \tfrac{1}{4}(e^{\tau^2} - 1)(e^{4\tau^2} + 2e^{3\tau^2} + 3e^{2\tau^2} + e^{\tau^2} - 2)] \tag{A1}$$

ACKNOWLEDGEMENTS

This research was supported, in part, by American Chemistry Council grant MTH0311-01 and National Institute for Environmental Health Sciences grants T32ES07018, P42ES05948, and P30ES10126. Data for the study of Chinese workers was obtained under contract HEI-96-5 from the Health Effects Institute. We acknowledge the assistance of our coinvestigators and colleagues—Suramya Waidyanatha, Qingshan Qu, Roy Shore, Ximei Jin, Beverly Cohen, Lung-Chi Chen, Assieh A. Melikian, Guilan Li, Songnian Yin, Huifang Yan, Bohong Xu, Ruidong Mu, Yuying Li, Xiaoling Zhang, and Keqi Li—who collected and analysed the exposure and biomarker samples from the Chinese study. We also acknowledge the affiliated institutions who made this research possible: School of Public Health, University of North Carolina, Chapel Hill, NC 27599-7400 (S. W.); Nelson Institute of Environmental Medicine, New York University School of Medicine, Tuxedo, NY 10987 (Q. Q., R. S., X. J., B. C., L-C. C.); American Health Foundation, Valhalla, NY 10595 (A. A. M.); Chinese Academy of Preventative Medicine, Institute for Occupational Medicine, Beijing 100050, China (G. L., S. Y., H. Y., B. X.); Tianjin Institute of Industrial Health and Occupational Medicine, Tianjin 300204, China (R. M., Y. L.); Tianjin Hebei District Antiepidemic Station, Tianjin 300400, China (X. Z.); and Wuqing County Antiepidemic Station, Tianjin 301700, China (K. L.).

REFERENCES

1. Rappaport SM, Waidyanatha S, Qu Q, Shore R, Jin X, Cohen B, Chen L-C, Melikian AA, Yin S, Yan H, Xu B, Mu R, Li Y, Zhang X, Li K. Albumin adducts of benzene oxide and 1,4-benzoquinone as measures of benzene metabolism. *Cancer Research* 2002; **62**:1330–1337.
2. Savitz DA, Andrews KW. Risk of myelogenous leukaemia and multiple myeloma in workers exposed to benzene. *Occupational and Environmental Medicine* 1996; **53**:357–358.
3. Snyder R. Benzene and leukemia. *Critical Review of Toxicology* 2002; **32**:155–210.
4. Golding BT, Watson WP. Possible mechanisms of carcinogenesis after exposure to benzene. *Exocyclic Nucleic Acid Adducts in Carcinogenesis and Mutagenesis*. IARC Scientific Publications: Lyon, France, 1999; 75–88.
5. Bates DM, Watts DG. *Nonlinear Regression Analysis and its Applications*. Wiley: New York, 1988.
6. Carroll RJ, Ruppert D. *Transformations and Weighting in Regression*. Chapman & Hall/CRC Press: New York, Boca Raton, FL, 1995.
7. Seber GAF, Wild CJ. *Nonlinear Regression*. Wiley: New York, 1989.
8. Davidian M, Giltinan DM. *Nonlinear Models for Repeated Measurement Data*. Chapman & Hall/CRC Press: New York, Boca Raton, FL, 1995.

DOI: 10.1002/sim

9. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.
10. O'Brien LM, Fitzmaurice GM. Analysis of longitudinal multiple source binary data using generalized estimating equations. *Journal of the Royal Statistical Society*, *Series C* 2004; **53**:177–193.
11. Rappaport SM, Kupper LL, Lin YS. On the importance of exposure variability to the doses of volatile organic compounds. *Toxicological Sciences* 2005; **83**:224–236.
12. Beverton RJH, Holt SJ. *On the Dynamics of Exploited Fish Populations*. Her Majesty's Stationery Office: London, 1957.
13. Ruppert D, Carroll RJ. In *Data Transformations in Regression with Applications to Stock Recruitment Relationships. Resource Management*, Mangel M (ed.). Lecture Notes in Biomathematics, vol. 61. Springer: New York, 1985.
14. Aitchison J, Brown JAC. *The Lognormal Distribution, with Special Reference to its Use in Economics*. Cambridge University Press: New York, 1957.
15. Crow EL, Shimizu K. *Lognormal Distributions*: *Theory and Applications*. Kluwer: Amsterdam, Netherlands, 1988.
16. Leech D. Testing the error specification in nonlinear regression. *Econometrika* 1975; **43**:719–725.
17. Lyles RH, Kupper LL. A detailed evaluation of adjustment methods for multiplicative measurement error in linear regression with applications in occupational epidemiology. *Biometrics* 1997; **53**:1008–1025.
18. Lyles RH, Kupper LL, Rappaport SM. A lognormal distribution-based exposure assessment method for unbalanced data. *Annals of Occupational Hygiene* 1997; **41**:63–76.
19. Lyles RH, Kupper LL, Rappaport SM. Assessing regulatory compliance of occupational exposures *via* the balanced one-way random effects anova model. *Journal of Agricultural*, *Biological*, *and Environmental Statistics* 1997; **2**:64–86.
20. Lyles RH, Kupper LL, Rappaport SM. On prediction of lognormal-scale mean exposure levels in epidemiologic studies. *Journal of Agricultural*, *Biological*, *and Environmental Statistics* 1997; **2**:417–439.
21. Taylor DJ, Kupper LL, Rappaport SM, Lyles RH. A mixture model for occupational exposure mean testing with a limit of detection. *Biometrics* 2001; **57**:681–688.
22. Taylor DJ, Kupper LL, Rappaport SM. Statistical methods for evaluating exposure–biomarker relationships. *Symposium on Biostatistical and Biomathematical Problems in Environmental Health*. National Institute of Environmental Health Sciences, Research Triangle Park, NC, U.S.A., 2002.
23. Carroll RJ, Ruppert D. Power transformations when fitting theoretical models to data. *Journal of the American Statistical Association* 1984; **79**:321–328.
24. Box GEP, Cox DR. An analysis of transformations. *Journal of the Royal Statistical Society*, *Series B* 1964; **26**:211–246.
25. McCullagh P. Quasi-likelihood functions. *The Annals of Statistics* 1983; **11**:59–67.
26. Davidian M, Carroll RJ. Variance function estimation. *Journal of the American Statistical Association* 1987; **82**:1079–1091.
27. Davidian M, Carroll RJ. A note on extended quasilikelihood. *Journal of the Royal Statistical Society*, *Series B* 1988; **50**:74–82.
28. Beal SL, Sheiner LB. Heteroscedastic nonlinear regression. *Technometrics* 1988; **30**:327–338.
29. Davidian M, Giltinan DM. Some general estimation methods for nonlinear mixed models. *Journal of Biopharmaceutical Statistics* 1993; **3**:23–55.
30. Johnson BA, Kupper LL, Taylor DJ, Rappaport SM. Modeling exposure–biomarker relationships: applications of linear and nonlinear toxicokinetics. *Journal of Agricultural*, *Biological and Environmental Statistics* 2005; **10**:440–459.

DOI: 10.1002/sim