Generalized Linear Mixed Model and Calibration for Gamma Random Variables: Application to Asbestos Fibers

BY

YOONSANG KIM
M.S., University of Iowa, Iowa, 2005
M.P.H., Seoul National University, Korea, 2002
B.A. Chung-Ang University, Korea, 2000

DISSERTATION

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Public Health Sciences
in the Graduate College of the
University of Illinois at Chicago, 2011

Chicago, Illinois

Defense Committee:
Dulal K. Bhaumik, Chair and Advisor
Sally Freels
Robert D. Gibbons, University of Chicago
Leslie T. Stayner
Hui Xie

This dissertation is dedicated to my mother, Jisoon Kim.

ACKNOWLEDGMENTS

I am grateful to my advisor, Dr. Dulal Bhaumik, for his guidance, kindness, and patience with me. He has so much compassion for his students and he treated me like his family. I would like to thank Dr. Robert Gibbons for introducing this topic and data to me and for giving me insights and clarity. I would also like to thank other committee members, Drs. Sally Freels, Hui Xie, and Leslie Stayner for their helpful comments. Their comments helped me think further about the topic and improve my dissertation. I also would like to thank Dr. Donald Hedeker for his help regarding SuperMix.

I am deeply indebted to Drs. Michael Berbaum and Sherry Emery in the UIC Institute for Health Research and Policy. Over the past five years, I have learned a great deal from them about becoming a biostatistician and working within multidisciplinary teams. They encouraged me, understood me, and believed in my ability better than I did.

This dissertation would not have been possible without the support of Jungwha, Sun Young, and Han. Whenever I was struggling or frustrated, they listened to me with infinite patience. I appreciate their care, understanding, and advice.

Lastly, but not least, I would like to thank my mother, Jisoon Kim, for her unconditional love.

YK

TABLE OF CONTENTS

CHAPTER **PAGE**

TABLE OF CONTENTS

CHAPTER **PAGE**

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVATIONS

AGQ	Adaptive Gaussian Quadrature
BLUP	Best Linear Unbiased Predictor
CP	Coverage Probability
CV	Coefficient of Variation
CPD	Central Posterior Density
EB	Empirical Bayes
GLMM	Generalized Linear Mixed Model
HPD	Highest Posterior Density
HPPD	Highest Posterior Predictive Density
IWLS	Iteratively re-Weighted Least Squares
IWMML	Iteratively re-Weighted Maximum Marginal Likelihood
LMM	Linear Mixed Model
MLE	Maximum Likelihood Estimator
MML	Maximum Marginal Likelihood
NYSDOH	New York State Department of Health
PCM	Phase-Contrast light Microscopy
RMSE	Root Mean Square Error
SD	Standard Deviation
TEM	Transmission Electron Microscopy
UCL	Upper Confidence Limit
UPL	Upper Prediction Limit

SUMMARY

I present statistical methods for analyzing positively skewed data, which commonly arise in environmental monitoring and assessment. Specifically, I develop relevant methods to estimate the underlying calibration curve and construct the confidence and prediction intervals based on the gamma distribution. Environmental data are subject to measurement errors, and the variance of measurement errors usually depend upon the true concentration level. In addition, when multiple laboratories are involved in analyzing monitoring samples, additional variation at the laboratory level should be incorporated in the analysis. In this dissertation, I propose a mixed-effects gamma regression model to account for the non-constant variability as well as the between-laboratory variability to estimate a calibration curve. I also explore the two-component mixed model to fit environmental data and discuss its applications. The proposed methods borrow strength from all laboratories to estimate the model parameters. I discuss how to estimate unknown true concentrations using the estimated calibration curve when an independent set of samples are obtained. I also derive the global calibration confidence interval that does not require new data from the same set of laboratories, from which background samples were collected. The performances of the estimation procedures, the calibration confidence regions, and their robustness are studied via simulation. I observe that the global calibration confidence intervals based on the gamma mixed model perform robustly in all situations considered. To illustrate the results, a real data set of amosite asbestos fibers is used.

1. INTRODUCTION

1.1 Motivation

Over the last decade, generalized linear mixed models (GLMMs) have been widely used in health sciences and social sciences. In these areas clustered observations commonly arise, for example in a multicenter randomized trial where patients are nested in clinics, and in a longitudinal study where measurements of the same subject are taken repeatedly over time. Research in analyzing clustered data has involved the development of efficient estimation methods based on maximum marginal likelihood (MML) and empirical Bayes (EB) estimation, nonparametric maximum likelihood, generalized estimating equation (GEE), or fully Bayesian approaches, etc. Since Laird and Ware (1982) developed a linear random-effects model for normally distributed data, generalization of a mixed-effects model for categorical outcomes (binary, nominal, ordinal, Poisson, etc.) has been an active area of statistical research. Furthermore, statistical packages that implement these generalized models have been developed and they are now commercially available. In spite of this fast growth and active research, surprisingly little research has been conducted for modeling non-negative skewed continuous data where clustering naturally occurs. When the data in hand are skewed, the usual approaches for analyzing continuous outcomes assuming normality, such as the linear mixed model (LMM), are questionable. In the LMM, conditional on random effects, the outcome is assumed to be normally distributed with a constant variance. For minimally skewed data, the LMM may be used, but for moderately to highly skewed data, using the LMM reduces statistical power and may result in biased parameter estimates. Further, when the sample size is small, use of the LMM, which was developed based on large sample theory, to skewed data is more problematic. With skewed data, the variance is likely to be dependent upon the mean. Using the GLMM frame, we can model skewed data more properly and allow the variance to depend on any known function of the mean outcome as well.

My interest in the mixed-effects model for skewed data stems from analysis of asbestos fiber data. In asbestos data, more broadly environmental monitoring data, the distribution of measurements is usually right skewed. In environmental monitoring problems, an instrument response is easily known but the true concentration must be estimated based on the instrument responses, which are subject to measurement errors. Using training samples, the instrument response is regressed on a true concentration, and this regression equation is then inverted to estimate the unknown true concentration of an independent data set. This regression equation is called the *calibration curve*. An important characteristic of environmental data is that heterogeneous variation commonly arises due to measurement errors, which makes fitting a regression model (i.e. calibration curve) more complicated. The magnitude of the measurement errors usually increases with the concentration levels. A plot of instrument responses versus true concentrations often has a wedge shape, indicating that the variance of the measurements increases with increasing true concentrations. For zero or near-zero concentrations, on the other hand, the measurement errors do not vanish; but the variation of errors remains approximately constant. In the context of environmental monitoring, multiple samples are taken at a site (or one sample is split into subsamples) and they are sent to several laboratories to measure concentrations. This provides a way of cross-validation of anomalous results. In this process, another type of measurement errors happens due to different protocols by different laboratories. In other words, it is expected that measurements within the same laboratory are more alike than those from different laboratories. The main issue is that the mixed-effects model assuming normality and a constant variance cannot be used for such environmental data. When analyzing this type of data, heteroscedasticity and between-laboratory variability should be incorporated as well as the skewness.

The next section describes a data set of airborne asbestos fibers to illustrate the methodological challenges mentioned above. This data set will be used throughout this dissertation, to illustrate the statistical ideas and developments.

1.2 Example: Airborne Asbestos Fibers

It is well known that inhalation of asbestos fibers increases the risk of life-threatening diseases, such as mesothelioma, asbestosis, and lung cancer (WHO-IARC, 1998; Stayner et al., 1996; Hein et al. 2007; Stayner et al., 2008; Christensen et al., 2008). In epidemiologic studies that evaluate asbestos related health hazards, the amount of exposure should be properly quantified. The amount of asbestos exposure is defined as a product of the intensity and the duration of exposure (Camus et al.,1998). The intensity is usually measured as the number of asbestos fiber counts in a unit volume (for example, fibers/m^3, fibers/mL). Recently, the size of asbestos fibers is getting attention because thinner and longer fibers appear to be more strongly associated with health hazard than shorter and thicker fibers (Stayner et al., 2008; Loomis et al., 2010). A common standard analytical technique for testing airborne fiber concentrations is Phase-Contrast light Microscopy (PCM). The standard procedure of counting fibers using PCM includes only fibers longer than $5\mu m$, and excludes fibers smaller than $0.25\mu m$ in diameter. In addition, a ratio of the length to the diameter should be at least 3 (Loomis et al., 2010). PCM is not recommended for exposure assessment of elongated mineral particles because it does not provide the detailed exposure characterization and may not detect significant quantities of asbestos fibers (Institute of Medicine, 2006). A more advanced technique, Transmission Electron Microscopy (TEM) enables all fibers, including very fine fibers, to be counted and characterized (Loomis et al., 2010); as a result, the intensity of asbestos exposure is quantified more accurately.

In spite of advanced technologies for analyzing asbestos fibers, there are still substantial measurement errors in instrument responses. The New York State Department of Health uses various types of airborne asbestos data measured by TEM for testing and laboratory assessment. The asbestos data were collected as part of the New York State Environmental Laboratory Approval Program based on the proficiency testing of laboratories analyzing airborne asbestos. The Asbestos Hazard Emergency Response Act (AHERA) criteria were used. All data are de-identified and expressed in structures$/mm^2$ of filter. *Amosite* fibers among several fibrous types are used in this dissertation.

Amosite is one of the types in amphibole group, and was often used as an insulating material (www.asbestos.com). Although the use of amosite asbestos has decreased for recent decades, they are still found in residential and commercial settings.

Asbestos samples were taken from 14 contaminated sites, and were sent to 35 laboratories to measure the concentration. Multiple samples were collected from each site, 24 to 29 samples, and each laboratory analyzed a sample from a particular site only once. Table I shows averages, standard deviations (SDs), and coefficient of variations (CVs) of amosite fiber counts by sites. Sites are organized in the order of their average concentrations. The SD of measurements increases with the average fiber count, indicating non-constant variability. The CV is the largest at the lowest mean concentration (site 2778) and roughly constant at higher mean concentrations. This shows non-vanishing measurement error at low concentrations.

TABLE I

DESCRIPTIVE STATISTICS OF AMOSITE ASBESTOS DATA

Sites	N	Mean	SD	CV
2778	26	159.00	83.61	0.53
6001	27	387.85	108.65	0.28
3739	27	768.26	227.94	0.30
187Q	25	788.20	226.57	0.29
4915	27	802.56	253.70	0.32
7420	28	849.93	221.72	0.26
5284	25	951.80	206.69	0.22
8306	27	1217.52	405.08	0.33
6482	27	1325.22	519.62	0.39
5099	25	1648.16	390.66	0.24
8214	27	1731.93	392.66	0.23
879Q	25	2137.72	489.56	0.23
5209	24	6910.17	2695.92	0.39
1987	29	8133.00	2630.82	0.32

Figure 1 displays the raw amosite asbestos counts on the vertical axis and the average counts by sites on the horizontal axis. The data exhibit considerable non-constant variability: variation increases with the average count. The mean concentrations by sites can be considered as naturally occurring concentration levels of the sites (i.e., best available estimate of the true count). I revisit these data in Chapter 2 and Chapter 5 for illustration.

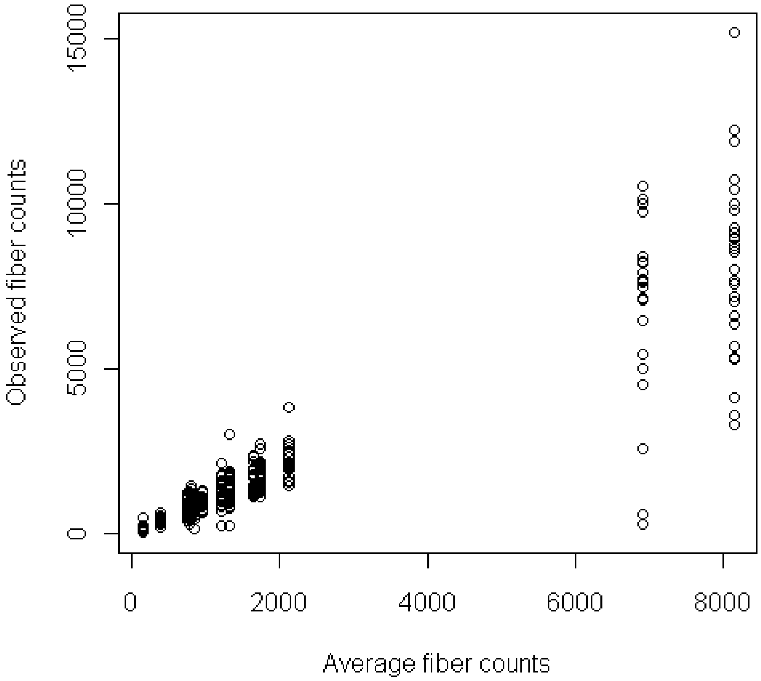

Figure 1. Heteroscedasticity in Amosite Asbestos Fibers from New York State Department of Health.

1.3 **Background**

Estimation of the calibration curve is an important subject in the environmental monitoring and assessment as well as in analytical chemistry. In this section, I begin with reviewing methodologies for calibration curve estimation. Rocke and Lorenzato (1995) proposed a two-component error model using a mixture of lognormal and normal distributed errors to describe the general calibration curve in analytical chemistry. The basic ideas were that the lognormal distribution was used to explain the skewness of data, and the normal distribution was used for the constant error variance at near-zero concentration. Gibbons and Bhaumik (2001) generalized the two-component error model for a single laboratory to the random-effects model for inter-laboratory problem, and estimated parameters using generalized least squares and maximum marginal likelihood estimation combined. They estimated the calibration function relating instrument responses to the known true concentrations for copper data (presented in Appendix) using the model. Later, Bhaumik and Gibbons (2005) estimated parameters of the extended model by using the method of moments, which was more robust and stable. To estimate the calibration curve of cadmium data, they assumed that zero (or near-zero) concentrations are available. The authors also took advantage of replicated measurements (per concentration within a laboratory). In practice, however, the environmental monitoring samples occasionally do not have such replicates, and observations at near-zero concentrations may also not be available. This is the situation occurring in the asbestos data from NYSDOH described in Section 1.2. Therefore, the estimation method of Bhaumik and Gibbons (2005) may not be applied in some situations.

The calibration curve can be estimated when a large set of background data is available, for instance when measurements are routinely obtained (often from multiple instruments and/or multiple laboratories). We can then estimate a true unknown concentration of a new measurement using all information we obtained from the background data (Bhaumik and Gibbons, 2005). Sometimes, interest is in examining a potentially contaminated area to determine if the area is contaminated or not by comparing a measurement to a regulatory standard, i.e. permissible limit. In this case, confidence and prediction limits are particularly useful. The confidence interval can be used to determine whether the data are either consistent with or higher than the standard. The U.S. EPA recommends and often requires reporting the mean concentration and its upper confidence limit (UCL) (Singh et al., 1997;

Singh et al., 2002). For example, if the UCL of asbestos measurements taken from a textile plant exceeds a permissible exposure limit, the plant may need to be remedied. The calibration confidence interval for an unknown concentration can be used in this sense. The prediction interval is useful for testing individual new samples when a large set of background samples are available. In this case interest is in determining the probability that the new sample was drawn from the distribution of the background data, which have been collected from areas considered not contaminated. A single new sample (or an average of a few new samples) from a potentially contaminated site is compared to an upper prediction limit (UPL) computed based on the background data (Bhaumik and Gibbons, 2004; Gibbons and Bhaumik, 2006). If the measured concentration exceeds the UPL, further investigation may be required to find out whether the area is an environmental concern.

Determining the distribution of data is a fundamental step in estimating the calibration curve and constructing the confidence and prediction intervals. The lognormal distribution has been widely used to analyze environmental data (Ott, 1995; Oehlert et al., 1995; Cheng, 1986; Bhaumik and Gibbons, 2004; Cheng et al., 2006). Land (1973, 1975) developed the H-statistic to compute the UCL of the lognormal mean. The U.S. EPA had recommended using the H-statistic in their guidance document in 1992. However, they recommended against its use in 1997 due to the work of Gilbert (1993), which indicates that the H-statistic is upwardly biased (Singh et al., 1997). Bhaumik and Gibbons (2004) summarized several methods for computing the UPL of a lognormal distribution and compared them via simulation. The lognormal distribution is convenient to use because one can apply the normal distribution-based approaches on log-transformed data (Finkelstein, 2008). The normal distribution-based approaches usually have good asymptotic properties. However, environmental monitoring data are typically not large enough to provide adequate power for an asymptotic test of hypothesis and interval estimation.

The gamma distribution has also been used to analyze right-skewed data in various environmental monitoring applications. For example, Bhaumik and Gibbons (2006) developed simultaneous prediction limits for gamma-distributed random variables and applied to groundwater monitoring problems. Krishnamoorthy et al. (2008) and Aryal et al. (2008) developed prediction and tolerance intervals to analyze alkalinity concentration of groundwater assuming the gamma distribution: the former used a normal approximation for the cube root of a gamma random variable, and the latter used a normal approximation for a log-transformed gamma random variable. In addition, the U.S. EPA (Singh, et al., 2002) suggested using the gamma distribution, arguing against using a lognormal distribution, as the mean (and its UCL) may be overestimated with the lognormal-based approach. Beyond environmental monitoring applications, the gamma distribution can be used to analyze insurance claims data, daily rainfall data, etc (McCullagh and Nelder, 1989). Despite the fact that asbestos data show positive skewness, it is surprising that the gamma distribution has been rarely used to characterize asbestos fiber count data.

It should be noted that a gamma distribution is more flexible than a lognormal distribution under certain conditions. The useful features of a gamma distribution are (i) as the shape parameter increases, it approaches a normal distribution, (ii) the variance is proportional to the mean which is often observed with environmental monitoring data, (iii) the cube root of a gamma distributed variable has an approximate normal distribution, and (iv) similar to the lognormal distribution, the log-transformed gamma distributed variables have an approximate normal distribution unless the shape parameter is small. It is known that the exponential distribution and chi-square distribution are special cases of the gamma distribution, and the gamma distribution looks similar to the lognormal distribution (Casella and Berger, 2002).

For both lognormal and gamma distributions, their means are functions of parameters of the distribution, which makes the construction of a confidence interval rather complicated. Krishnamoorthy and Mathew (2003) developed a test statistic using the concept of generalized confidence intervals and generalized p-values to estimate a confidence interval for the lognormal mean. The generalized confidence interval is applicable to small sample sizes. Bhaumik, Kapur, and Gibbons

(2009) used a normal approximation to the cube root of a gamma distributed variable in order to develop tests for the gamma mean. Their tests perform very well for small sample sizes.

1.4 Purpose of the Study

In this dissertation, I explore the use of gamma distributions for right skewed data; specifically, I introduce the mixed-effects model for a gamma distributed random variable to analyze environmental data, and also construct confidence and prediction intervals based on the gamma distribution model.

A calibration curve is estimated based on a large set of background samples. The challenge of estimating the parameters of the calibration curve is that skewness, heteroscedasticity, and between-laboratory variability should be incorporated. I propose a gamma mixed-effects model to explain those features of environmental data. Laboratories can be considered to be random samples from a bigger pool of laboratories; therefore, laboratory effects are treated as random. That allows estimation of the between-laboratory variation as well as laboratory-specific calibration curves. Figure 2 shows laboratory-specific calibration curves of hypothetical data. The thick solid line in Figure 2 indicates the perfect calibration line, i.e. it is obtained if laboratories provided correct concentration measurements in all circumstances. In fact, some laboratories may almost always provide higher measurements and others provide lower measurements than the true concentrations. Some laboratories may perform poorly when concentration is low; on the other hand, others give poor measurements for high concentrations. The next question is how to estimate the model parameters and use them to make inferences. Moreover, using the estimated calibration curve, how do we estimate unknown concentrations when new samples are obtained? I discuss how to address these questions in this dissertation. Since the calibration function is estimated using all laboratories from which background data are obtained, my method is borrowing strength from those laboratories to estimate unknown concentrations even if only a subset of the laboratories analyzes new samples. With these problems in mind, I organize this dissertation as follows.

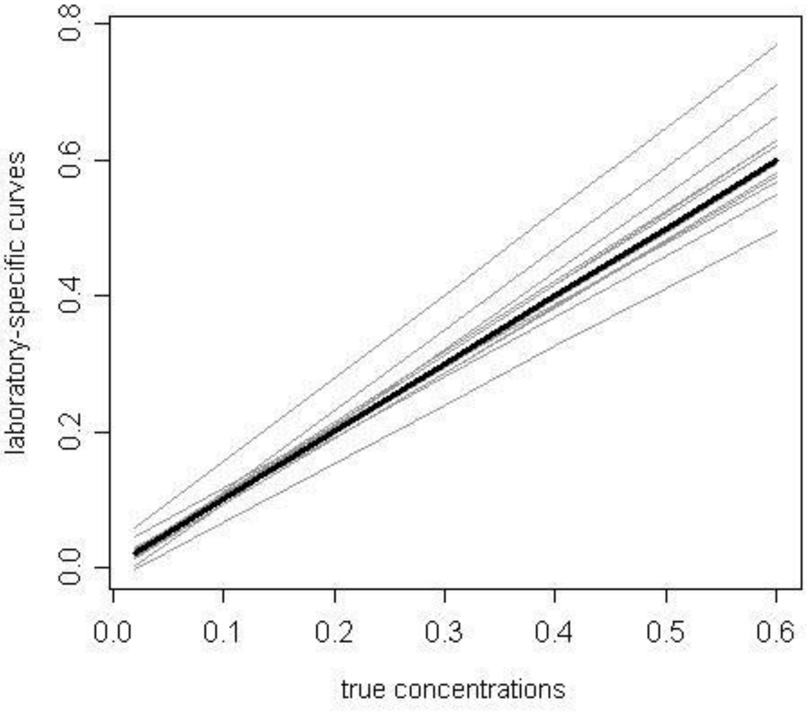

Figure 2. Laboratory-specific calibration curves of hypothetical data.

1.5 Organization of the Dissertation

In Chapter 2, using the amosite asbestos data from NYSDOH, I compute the confidence interval for the mean fiber count and the prediction interval for a single new observation assuming a lognormal distribution and a gamma distribution. I explore the generalized confidence interval method developed by Krishnamoorthy and Mathew (2003) to estimate a confidence interval of the lognormal mean. A confidence interval for the gamma mean is constructed following the approach developed by Bhaumik, Kapur, and Gibbons (2009). For a prediction interval, I use the methods proposed by Krishnamoorthy et al. (2008) and Aryal et al. (2008). I describe each method in detail and discuss their advantages and disadvantages. The Bayesian approach of calculating an interval estimate for the mean and for a single new observation is also considered.

In Chapter 3, I propose regression models to estimate the calibration function incorporating heteroscedasticity and between-laboratory variability, and explore how to estimate true concentrations. I use nonlinear regression models, in which measured concentrations are regressed on true concentrations. A two-component error model with random laboratory effects is suggested to incorporate heteroscedasticity and between-laboratories variability simultaneously. To estimate the model parameters and variance components, I modify the iteratively reweighted maximum marginal likelihood (IWMML) method, developed by Gibbons and Bhaumik (2001), to analyze data with no replicate for a particular concentration within a laboratory. I propose a gamma mixed-effects regression model that integrates both non-constant variability and between-laboratory variability. For parameter estimation, maximum marginal likelihood (MML) and empirical Bayes (EB) estimation are used. In addition, I derive the point and interval estimates for an unknown true concentration of new samples based on both of the fitted regression models, two-component mixed model and gamma mixed model, borrowing strength from all laboratories.

In Chapter 4, I evaluate the performance of estimation procedures, point estimates, and calibration confidence intervals via simulation. In Chapter 5, I illustrate the methodology by analyzing the amosite asbestos data and the experimental copper data.

2. CONFIDENCE AND PREDICTION INTERVALS

In this chapter, I examine statistical methods for computing confidence and prediction intervals that are useful in the context of environmental monitoring problem. The environmental monitoring data are often positively skewed as described in Chapter 1. I explore several methods relevant to computing confidence and prediction intervals, assuming a lognormal distribution or a gamma distribution. The distributional assumptions on the asbestos data from NYSDOH are checked, and I use these methods to compute confidence intervals for the average fiber count and prediction intervals for a single measurement. Finally, I compare the results obtained using these methods. I begin with methods relevant to a lognormal distribution.

2.1 The Lognormal Distribution

Denote a concentration measurement by X. Suppose it has a lognormal distribution with parameters μ and σ^2, i.e. $\ln X \sim N(\mu, \sigma^2)$. The pdf of X is

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\frac{1}{x}\exp\left\{-\frac{(ln\, x-\mu)^2}{2\sigma^2}\right\} I_{\{0<x<\infty\}} I_{\{0<\sigma\}}\,. \qquad (2.1)$$

The mean and variance of X are

$$E(X) = e^{\mu+\sigma^2/2}\,, \qquad V(X) = e^{2\mu+\sigma^2}\left(e^{\sigma^2}-1\right).$$

Let $\hat{\mu}$ and $\hat{\sigma}^2$ be the maximum likelihood estimates (MLEs) of μ and σ^2. These are jointly sufficient and asymptotically consistent estimators. The MLE is asymptotically normally distributed, thus it is common to construct the confidence interval for $E(X)$ based on a Wald type statistic, provided the sample size is large; otherwise we cannot obtain the correct coverage probability. In order to provide results applicable to small samples as well as large samples, I explore the idea of generalized confidence intervals introduced by Weerahandi (1993). Krishnamoorthy and Mathew (2003), and Krishnamoorthy et al. (2006) developed confidence intervals for a lognormal mean using Weerahandi's (1993) concept of generalized confidence intervals.

2.1.1 Generalized Confidence Interval for the Mean of a Lognormal Distribution

We can obtain the interval estimator of the mean of a lognormal distribution, $E(X)$, by computing the confidence interval for $\eta = \mu + \sigma^2/2$ and then taking exponents of the lower and upper bounds of η . Krishnamoorthy and Mathew (2003) applied the generalized confidence intervals and generalized p-values in order to compute the confidence interval for η. Their method is useful in this context as it does not require large sample size, it provides an exact confidence interval, and it is easy to compute. Let $Y_i = \ln(X_i)$ for $i = 1, \dots, n$, $\bar{Y}$ and S_Y^2 be the sample mean and sample variance of Y respectively. The observed realized values of $\bar{Y}$ and S_Y^2 are denoted by $\bar{y}$ and s_y^2. Krishnamoorthy and Mathew (2003) defined the generalized pivot statistic for obtaining the confidence interval of η as follows:

$$\begin{aligned} T_1 &= \bar{y} - \frac{\bar{Y} - \mu}{S_Y/\sqrt{n}} \frac{s_y}{\sqrt{n}} + \frac{1}{2} \frac{\sigma^2}{S_Y^2}\, s_y^2 \\ &= \bar{y} - \frac{Z}{U/\sqrt{n-1}} \frac{s_y}{\sqrt{n}} + \frac{1}{2} \frac{s_y^2}{U^2/(n-1)}, \end{aligned} \tag{2.2}$$

where Z is a standard normal random variable and U^2 is a chi-square random variable with degrees of freedom $n - 1$. When $\bar{Y} = \bar{y}$ and $S_Y^2 = s_y^2$, T_1 reduces to η. More details about how to construct a test for the lognormal mean can be found in Krishnamoorthy et al. (2006). The following algorithm describes how to construct the $100(1 - \alpha)\%$ generalized confidence interval for η. Suppose $(y_1, \dots, y_n)$ are log-transformed observed concentrations for a given site.

Algorithm 1.

1. Compute $\bar{y}$ and s_y^2 for a given site.
2. Generate $Z \sim N(0,1)$ and $U^2 \sim \chi_{n-1}^2$.
3. Compute T_1 in equation (2.2).

4. Repeat steps 2 and 3 m times to obtain $\left(T_{1_1}, \cdots, T_{1_m}\right)$. The two sided equal-tailed confidence interval is obtained by computing $100(\alpha/2)$ and $100(1-\alpha/2)$ percentiles of T_1. Denote these percentile points by $T_{\alpha/2}$ and $T_{1-\alpha/2}$. The interval $\left(T_{\alpha/2}, T_{1-\alpha/2}\right)$ is the generalized confidence interval for η.

Once the confidence interval for η is obtained, the confidence interval for $E(X)$ is computed by $\left(\exp(T_{\alpha/2}),\ \exp(T_{1-\alpha/2})\right)$. Generally m is a very large number (I used m =100,000). This interval has a property that after a large number of independent situations of setting $100(1-\alpha)\%$ confidence intervals for η, the true value of η will be included in the corresponding intervals $100(1-\alpha)\%$ of the time (Weerahandi 1993). The coverage probability of the confidence interval may be smaller than the pre-specified confidence level because the type I error and the power of this test may depend on unknown parameters. For this reason, Krishnamoorthy and Mathew (2006) conducted the simulation study assuming various combinations of μ and σ^2. Their simulation study shows that the coverage probabilities are very close to the pre-specified confidence level. They also compared their generalized confidence interval method to Augus's parametric bootstrap method (1994) and Land's method (1973) by estimating the upper confidence limits (UCL) of η when $\bar{y} = 1$ and $s_y = (0.1, 0.5, 5)$ with various sample sizes, $n = (3, 11, 21, 101, 501, 1001)$. It turns out that the UCL of Krishnamoorthy and Mathew (2006) are very close to Land's limit, but the parametric bootstrap results are unsatisfactory in this context (because it gives lower UCL than the other two methods) for small sample sizes, i.e. $n = (3, 11, 21)$ and large values of s_y. They also observed that their coverage probabilities obtained by Algorithm 1 are better than those obtained by the parametric bootstrap method in all scenarios considered.

I applied the generalized confidence interval to the amosite asbestos data from NYSDOH. The distributional assumption was checked using the normal probability plots (Figure 3) and Anderson-

Darling goodness-of-fit tests for log-transformed asbestos data: normal distributions fit moderately well to 9 sites (p-values>0.09) among a total of 14 sites. This result suggests moderately good fit of the data to a lognormal distribution. For each site, I simulated the coverage probability of the generalized confidence interval assuming the true value of $\eta = \hat{\mu} + \hat{\sigma}^2/2$, where $\hat{\mu}$ and $\hat{\sigma}^2$ are MLEs of μ and σ^2 for a given site. The coverage probabilities were simulated as follows: generate n lognormal random variates with parameter values equal to $\hat{\mu}$ and $\hat{\sigma}^2$ (n is the number of observations in a particular site), compute the confidence interval following Algorithm 1, and repeat this process 5,000 times. For the i^{th} repetition, denote the limits by $T^i_{\alpha/2}$ and $T^i_{1-\alpha/2}$. If the interval $\left(T^i_{\alpha/2}\ , T^i_{1-\alpha/2}\right)$ contains the true value of η, then assign the value 1, otherwise assign 0. The proportion of 1s is the simulated coverage probability. The results are presented in Section 2.1.3.

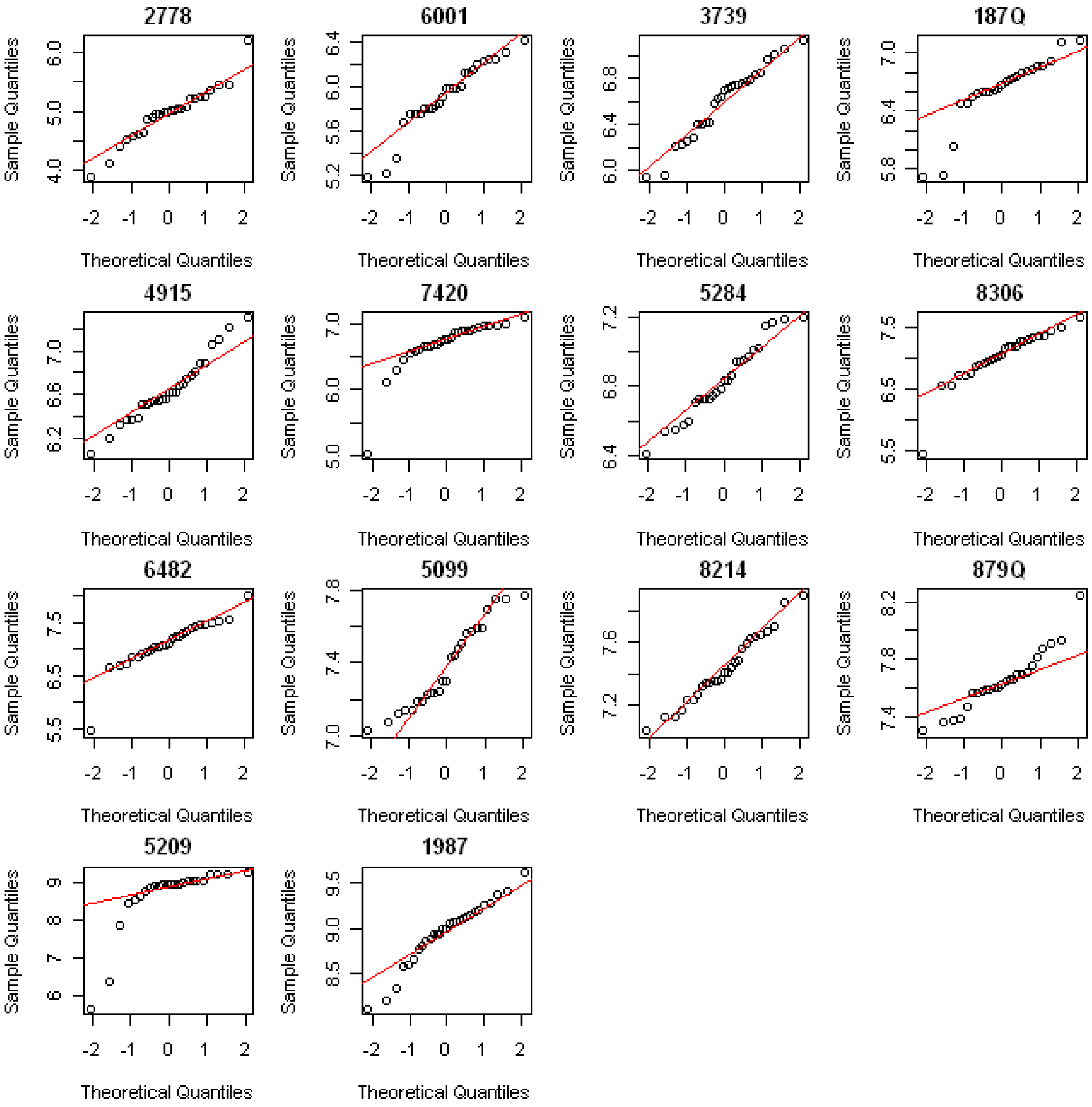

Figure 3. Normal probability plots of log-transformed data for all sites.

2.1.2 Bayesian Interval for the Mean of a Lognormal Distribution

This section explores how to obtain a Bayesian credible interval for the lognormal mean and for a single new observation. The Bayesian approach specifies prior knowledge about parameters or the quantity of interest as a prior density. The joint posterior density of (μ, σ^2) with a non-informative prior distribution $p(\mu, \sigma^2) \propto 1/\sigma^2$ is expressed as

$$\mathrm{p}(\mu, \sigma^2 | x_1, \dots, x_n) \propto (\sigma^2)^{-n/2-1} \exp\left\{-\frac{(n-1)s_y^2 + n(\bar{y}-\mu)^2}{2\sigma^2}\right\}. \tag{2.3}$$

We need the posterior density of $e^{\mu+\sigma^2/2}$ because the interest is the lognormal mean. However, it needs not to derive a functional form of the posterior density of this quantity. Instead, random draws from the joint posterior density of (μ, σ^2) are used to obtain random draws from the posterior density of $e^{\mu+\sigma^2/2}$. In fact, the posterior density of any functions of μ and σ can be obtained from random draws from the posterior density of (μ, σ^2) (Gelman et al., 2003). The steps for computing a Bayesian interval for the mean of a lognormal distribution is described in Algorithm 2.

Algorithm 2.

1. Simulate 100,000 draws from the posterior density $p(\mu, \sigma^2 | x_1, \dots, x_n)$ in (2.3) and denote them by (μ^ℓ, σ^ℓ) for ℓ =100,000.
2. Compute $exp\left\{\mu^\ell + \left(\sigma^\ell\right)^2/2\right\}$ for all drawn values of (μ^ℓ, σ^ℓ). Those values are treated as random draws from the posterior density of the lognormal mean.
3. Compute the highest posterior density (HPD) interval based on the draws from step 2.

For step 1, a plausible region of μ and σ^2 can be divided into many small grids and (μ^ℓ, σ^ℓ) are randomly sampled from each grid. I use the R package *LearnBayes* (Albert, 2007) to do step 1, and the R package *boa* (Smith, 2007) to do step 3.

The HPD interval is calculated in a way that values within the HPD interval have higher probabilities than values outside the interval. The HPD interval (L, U) satisfies the following equations,

$$p(L|x_1, \dots, x_n) = p(U|x_1, \dots, x_n) \ ,$$

$$\text{and} \qquad \int_{-\infty}^{U} p(\theta|x_1, \dots, x_n) - \int_{-\infty}^{L} p(\theta|x_1, \dots, x_n) = 1 - \alpha \ , \tag{2.4}$$

provided a mode of the posterior density is not in the boundary of domain. When the density has more than one mode, the HPD interval may consist of two distinct regions. Smith (2007) implemented Chen and Shao's algorithm (1999) in his R package *boa* to compute an HPD interval based on random draws. The equal-tailed interval, called the *central posterior density (CPD) interval*, can also be computed instead of the HPD interval. However, the length of the HPD interval is shorter than the CPD interval when the density is asymmetric (Casella and Berger, 2002).

Figure 4A shows a contour plot with random draws from the posterior density (2.3), given the asbestos data from the site 2778. The marginal distribution of μ is symmetric around the mode 4.9, but the distribution of σ^2 is right skewed as expected. The posterior distribution of the mean is shown in Figure 4B, which is slightly right skewed. The posterior median is 159.77 and the posterior mean is 161.32, so the skewness is trivial. Although the skewness in Figure 4B is small, the HPD interval for the mean (131.14, 194.71) is a little shorter and located to the left of the CPD interval (133.67, 198.51). This CPD interval of the site 2778 is very close to the corresponding generalized confidence equal-tailed interval (see Table II).

Using the asbestos data from NYSDOH, I computed the Bayesian HPD intervals for the lognormal mean over all sites. I also computed the coverage probability, although this is not exactly the Bayesian idea, to compare the performance of the Bayesian interval with the generalized confidence interval. To simulate the coverage probability, I generated n lognormal random variates with parameters fixed at $\mu = \hat{\mu}$ and $\sigma^2 = \hat{\sigma}^2$, and computed the interval following Algorithm 2. This process was repeated 5,000 times, and the proportion of times when the true mean falls in the interval was calculated. The results are presented in Table II (see Section 2.1.3).

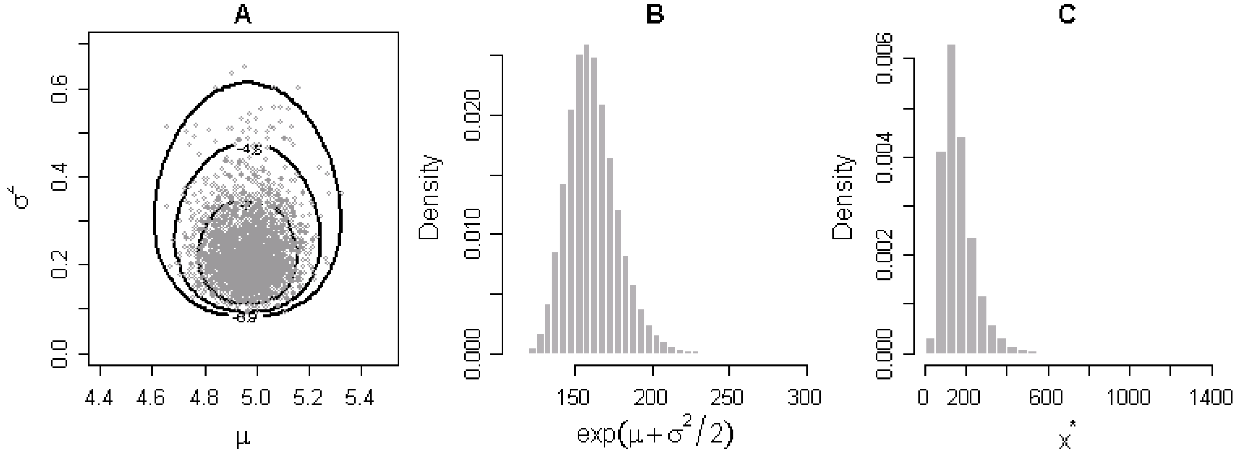

Figure 4. Posterior distributions relating to the lognormal distribution for site 2778. (A) Contour plot of the posterior density, $p(\mu, \sigma^2|\boldsymbol{x})$, and random draws from the density. (B) Histogram of the posterior density of the lognormal mean. (C) Histogram of the posterior predictive density for a new lognormal random variable given the random draws in (A).

2.1.3 Comparison of Intervals for the Lognormal Mean

In this section, using the amosite asbestos data for NYSDOH, I compare the generalized confidence interval and the HPD interval for the mean of lognormal variables with respect to their lengths, locations, and coverage probabilities. As mentioned in the previous section, the highest density interval gives generally shorter length than the equal-tailed interval when the distribution is skewed. It should be noted that the distribution of $exp(T_1)$, where T_1 is a simulated value by Algorithm 1, is right skewed (Figure 5). Therefore, its highest density interval would be shorter than the equal-tailed interval. Following this idea, I compute the lower and upper bounds of the confidence interval which satisfy conditions in (2.4), using simulated values of $exp(T_1)$, and call this interval the *generalized confidence highest density interval*.

Table II shows the generalized confidence *equal-tailed* intervals, the generalized confidence *highest density* intervals, and the Bayesian HPD intervals for the average counts of amosite fibers. Let me first compare two types of generalized confidence intervals; equal-tailed (A) and highest density intervals (B). The coverage probabilities are close to 0.95 for all sites. As expected, the generalized confidence highest density intervals (B) are shorter and located to the left of the corresponding equal-tailed intervals (A). Next, the Bayesian HPD interval (C) is compared with the generalized confidence highest density interval (B). These intervals are quite similar in terms of their locations and lengths (although the generalized confidence highest density intervals are never wider than the corresponding Bayesian HPD intervals, except for the site 5209). In addition, all coverage probabilities are close to 0.95. It appears that when the distribution of $exp(T_1)$ is skewed, the generalized confidence highest density interval can be a good alternative to the equal-tailed interval if a shorter interval is desired.

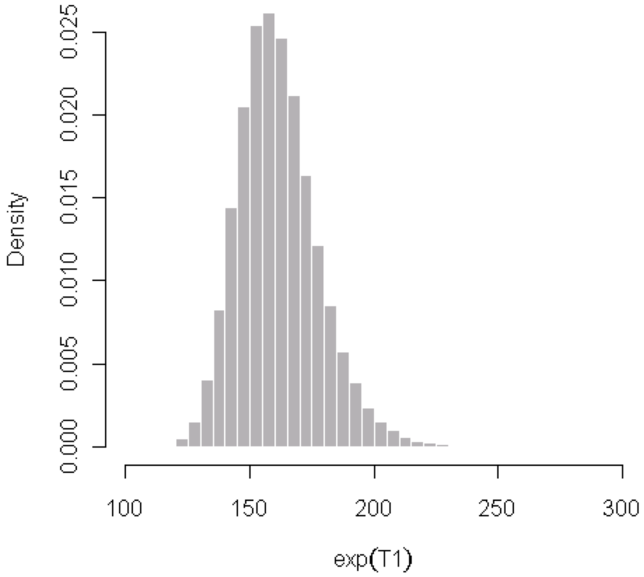

Figure 5. Distribution of $exp(T_1)$ for site 2778: $\left(T_{1_1}, \cdots, T_{1_m}\right)$ are simulated based on Algorithm 1. The median is 159.8, mean is 161.4, and maximum value is 297.1.

TABLE II

95% GENERALIZED CONFIDENCE INTERVALS AND BAYESIAN HPD INTERVALS FOR THE LOGNORMAL MEAN

	A				B				C			
	Generalized Confidence Equal-Tailed Intervals				Generalized Confidence Highest Density Intervals				Bayesian HPD Intervals			
Sites	2.5%	97.5%	Length	C.P.[b]	Lower	Upper	Length	C.P.[b]	Lower	Upper	Length	C.P.[b]
2778	133.81	198.42	64.61	0.946	131.59	194.87	63.28	0.955	131.14	194.71	63.57	0.952
6001	347.09	446.87	99.78	0.950	344.24	443.20	98.96	0.952	344.63	443.78	99.15	0.949
3739	685.52	886.42	200.90	0.957	679.00	877.93	198.93	0.964	679.28	879.25	199.97	0.949
187Q[a]	696.02	937.03	241.01	0.943	687.09	925.04	237.95	0.941	687.00	926.78	239.78	0.952
4915	718.35	913.32	194.97	0.955	713.63	907.16	193.53	0.953	714.25	909.24	194.99	0.952
7420[a]	752.73	1038.39	285.66	0.945	743.42	1025.08	281.66	0.947	741.16	1024.09	282.93	0.949
5284	874.88	1048.48	173.60	0.947	870.35	1043.17	172.82	0.942	869.34	1042.36	173.02	0.950
8306[a]	1059.76	1515.30	455.55	0.943	1044.57	1492.47	447.90	0.957	1044.13	1495.14	451.01	0.952
6482[a]	1140.64	1675.97	535.33	0.956	1117.99	1642.82	524.83	0.957	1119.88	1645.80	525.92	0.955
5099	1505.40	1829.94	324.54	0.946	1497.51	1820.00	322.49	0.952	1496.95	1821.01	324.06	0.954
8214	1594.49	1902.15	307.66	0.949	1588.72	1894.67	305.95	0.955	1589.24	1896.28	307.04	0.951
879Q	1968.74	2342.95	374.21	0.954	1960.54	2332.62	372.08	0.954	1957.94	2332.69	374.75	0.957
5209[a]	5814.95	14488.37	8673.42	0.944	5312.41	13279.67	7967.26	0.950	5340.01	13296.39	7956.38	0.952
1987	7211.36	9521.92	2310.56	0.945	7149.33	9437.32	2287.99	0.951	7145.83	9456.30	2310.47	0.947

[a] Lognormal distributions do not fit measurements from these sites ($p<0.05$).

[b] Coverage Probabilities.

2.1.4 Prediction Interval for a Single Lognormal Observation

The Bayesian prediction interval for a new single observation is obtained based on the posterior predictive distribution. Denote the posterior predictive density of a new observation x^* as $h(x^*|x_1, \dots, x_n)$, which is given by

$$h(x^*|x_1, \dots, x_n) = \int_0^\infty \int_{-\infty}^\infty f(x^*|\mu, \sigma^2)\, p(\mu, \sigma^2|x_1, \dots, x_n) d\mu d\sigma^2. \tag{2.5}$$

Dahiya and Guttman (1982) derived the functional form of $h(x^*|x_1, \dots, x_n)$ and described how to compute the shortest prediction interval based on h, which is

$$h(x^*|x_1, \dots, x_n) = \frac{\Gamma(n/2)}{\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{1}{2}\right)\sqrt{\frac{(n-1)(n+1)}{n}}\ s_y^2}\ \frac{1}{x^*}\left\{1 + \frac{n(\ln x^* - \bar{y})^2}{(n+1)s_y^2(\mathrm{n}-1)}\right\}^{-n/2}, \tag{2.6}$$

where $\bar{y} = \sum \ln x_i / n$ and s_y^2 is the sample variance of $\ln X$. According to Dahiya and Guttman (1982), the density function (2.6) is

(i) a strictly decreasing function of x^* if $s_y^2(n-1)/n^3 \geq 1/(4n+4)$, or

(ii) possesses two stationary points if $s_y^2(n-1)/n^3 < 1/(4n+4)$.

If the data satisfy condition (i), a one-sided interval $(0, x_\alpha^*)$ has the shortest length. If condition (ii) is satisfied, the shortest prediction interval could be either one-sided or two-sided depending on the value of $h(x_m^*|x_1, \dots, x_n)$, where x_m^* is the smaller of the two stationary points. That is, if $h(x_m^*|x_1, \dots, x_n) < h(x_\alpha^*|x_1, \dots, x_n)$ then the one-sided interval $(0, x_\alpha^*)$ gives the shortest length, but if $h(x_m^*|x_1, \dots, x_n) \geq h(x_\alpha^*|x_1, \dots, x_n)$ then the shortest interval consists of two distinct regions $(0,\ x_1^*)$ and $(x_2^*,\ x_3^*)$, which satisfy

$$\begin{gathered} h(x_1^*|x_1, \dots, x_n) = \ h(x_2^*|x_1, \dots, x_n) = h(x_3^*|x_1, \dots, x_n)\,, \\ 1 - \alpha = p(x^* < x_1^*) + p(x_2^* < x^* < x_3^*). \end{gathered} \tag{2.7}$$

However, the amosite asbestos data satisfy neither (i) nor (ii). The density h for asbestos data has one stationary point equal to its mode (see Figure 4C). Therefore, Dahiya and Guttman’s method does not apply to the amosite asbestos data. In addition, their prediction limits under condition (ii) and equations (2.7) are both mathematically and computationally intensive.

Following an approach discussed in Gelman et al. (2003), I simulate 100,000 random draws x^* from the posterior predictive density $h(x^*|x_1, \ldots, x_n)$ given all drawn values of (μ^ℓ, σ^ℓ) in step 1 of Algorithm 2. Then the HPD interval based on the simulated random draws of x^* is the prediction interval. For the amosite data, the posterior predictive distribution of x^* is positively skewed and unimodal (Figure 4C).

The prediction interval for a single asbestos fiber count and its coverage probability for all sites are presented in Table III. The coverage probabilities were computed as follows: generate n lognormal random variables with parameters fixed at $\mu = \hat{\mu}$ and $\sigma^2 = \hat{\sigma}^2$, and obtain random draws from the joint posterior density $p(\mu, \sigma^2|x_1, \ldots, x_n)$, then simulate draws from the posterior predictive density $h(x^*|x_1, \ldots, x_n)$ given the draws of (μ, σ^2) and compute an HPD interval based on these draws, finally generate one lognormal random variable with parameters fixed at $\hat{\mu}$ and $\hat{\sigma}^2$ and determine whether this random variate is included within the prediction interval. This process was repeated 5,000 times. The coverage probability is the proportion of times when the prediction interval includes a new lognormal variate. As shown in Table III, all coverage probabilities of the Bayesian prediction intervals are close to 0.95.

TABLE III

95% HIGHEST POSTERIOR PREDICTIVE DENSITY INTERVALS
OF A LOGNORMAL RANDOM VARIABLE

Site	Lower	Upper	Length	C.P.[b]
2778	38.81	324.55	285.74	0.953
6001	168.36	658.33	489.97	0.948
3739	323.72	1308.97	985.25	0.951
187Q[a]	302.14	1419.33	1117.19	0.947
4915	368.21	1337.21	969.00	0.949
7420[a]	267.13	1624.98	1357.85	0.949
5284	562.34	1412.74	850.40	0.945
8306[a]	332.18	2419.42	2087.24	0.951
6482[a]	326.34	2738.58	2412.24	0.944
5099	900.90	2480.27	1579.37	0.946
8214	993.65	2563.59	1569.94	0.950
879Q	1260.94	3097.85	1836.91	0.950
5209[a]	153.38	26121.67	25968.29	0.950
1987	3115.42	14659.96	11544.54	0.947

[a] Lognormal distributions do not fit data from these sites (p<0.05).

[b] Coverage Probabilities.

2.2 The Gamma Distribution

Let X be a gamma distributed random variable with parameters κ and θ, i.e. $X \sim Gamma(\kappa, \theta)$. The pdf of X is given by

$$f(x|\kappa, \theta) = \frac{1}{\Gamma(\kappa)\theta^{\kappa}} x^{\kappa-1} e^{-x/\theta} \ \ I_{\{x>0\}} I_{\{\kappa>0, \theta>0\}} \,, \tag{2.8}$$

where κ is the shape parameter and θ is the scale parameter. The mean and variance of X are

$$E(X) = \kappa\theta \,, \qquad V(X) = \kappa\theta^2.$$

The gamma distribution has a few important features that make it convenient to apply in various situations. First, its variance is proportional to the mean and it has constant coefficient of variation. We can find a gamma distribution that looks very similar to the lognormal distribution (Casella and Berger, 2002). The cube root of a gamma distributed random variable is approximately normally distributed. The log-transformed gamma random variable has also an approximate normal distribution, provided its shape parameter κ is not small. Finally, the gamma distribution converges to the normal distribution as $\kappa \to \infty$. I use a few of these features in this section to derive confidence and prediction intervals.

2.2.1 Confidence Interval for the Mean of the Gamma Distribution

Denote the maximum likelihood estimates of κ and θ by $\hat{\kappa}$ and $\hat{\theta}$. Then, the gamma mean and variance estimators are $\hat{E}(X) = \hat{\kappa}\hat{\theta}$ and $\hat{V}(X) = \hat{\kappa}\hat{\theta}^2$. Bhaumik et al. (2009) proposed a test statistic for H_0: $E(X) = \mu_0$. Let $\bar{x}$ be an arithmetic mean, $\tilde{x}$ a geometric mean, and $Y = \sum x_i$, then the test statistic proposed is

$$T_2 = \frac{9(n\mu_0)^{1/3}(n-1)(Y^{1/3} - (n\mu_0)^{1/3})^2}{2n\mu_0 \ ln(\bar{x}/\tilde{x})} \,.$$

T_2 has an approximate F-distribution with degrees of freedoms 1 and $(n-1)$. According to the simulation study conducted by Bhaumik et al. (2009), T_2 controls the type I error rate better than the test developed by Grice and Bain (1980) for testing H_0. One drawback they mention is that it is slightly conservative when the true shape parameter κ is less than 0.5. However, for the amosite asbestos data, the estimated $\hat{\kappa}$ is larger than 2 for all sites. Another advantage of T_2 is that it does not depend on any unknown parameters under the null hypothesis. The corresponding confidence interval is obtained by inverting the test statistic T_2 as follows.

$$\begin{aligned}
1-\alpha \quad &= p\left(F_{\alpha/2} < T_2 < F_{1-\alpha/2}\right) \\
&= p\left(\frac{2\,ln(\bar{x}/\tilde{x})}{9(n-1)}F_{\alpha/2} < \frac{(Y^{1/3}-(n\mu_0)^{1/3})^2}{(n\mu_0)^{2/3}} < \frac{2\,ln(\bar{x}/\tilde{x})}{9(n-1)}F_{1-\alpha/2}\right) \\
&= p\left(L < \frac{(Y^{1/3}-(n\mu_0)^{1/3})^2}{(n\mu_0)^{2/3}} < U\right) \\
&= p\,(\,L < \omega^2 - 2\omega + 1 < U)\,,
\end{aligned}$$

where $\omega = Y^{1/3}/(n\mu_0\,)^{1/3}$. If $L > 0$, then $\left(1-\sqrt{U} < \omega < 1-\sqrt{L}\right)$ and $\left(1+\sqrt{L} < \omega < 1+\sqrt{U}\right)$ satisfy the equation and lead to two separate regions. If $L \cong 0$, the equation reduces to

$$\begin{aligned}
1-\alpha \quad &= p\,(1-\sqrt{U} < \omega < 1+\sqrt{U}) \\
&= p\left(\frac{Y}{n(1+\sqrt{U})^3} < \mu_0 < \frac{Y}{n(1-\sqrt{U})^3}\right). \qquad (2.9)
\end{aligned}$$

$L \to 0$ as $n \to \infty$. I also observed $L \cong 0$ in all sites of the amosite asbestos data. Therefore, the $100(1-\alpha)\%$ confidence interval for the mean by (2.9) is

$$\left(\frac{\sum x_i}{n(1+\sqrt{U})^3}\,, \qquad \frac{\sum x_i}{n(1-\sqrt{U})^3}\right), \text{ where } U = 2\,ln\left(\frac{\bar{x}}{\tilde{x}}\right)\frac{F_{1-\,\alpha/2}}{9(n-1)}. \qquad (2.10)$$

I applied this confidence interval (2.10) to the amosite asbestos data. I first checked the distributional assumption based on the gamma quantile-quantile plots (Figure 6) and Anderson-Darling goodness-of-fit tests: gamma distributions fit 13 out of 14 sites (p-values≥0.15). The only site that the gamma distribution does not fit is 5209 which does not follow the lognormal distribution as well. These results indicate that the gamma distribution fits better to the amosite asbestos data than the routinely used lognormal distribution. I computed 95% confidence intervals (2.10) for the mean fiber counts and their coverage probabilities by sites (Table IV).

The coverage probabilities were computed as follows: generate n gamma random variables with parameters fixed at $\kappa = \hat{\kappa}$ and $\theta = \hat{\theta}$, i.e. the true mean is set to $\mu_0 = \hat{\kappa}\hat{\theta}$, and then compute $T_2(\mu_0)$ and see whether this $T_2(\mu_0)$ falls between $F_{\alpha/2}$ and $F_{1-\alpha/2}$. This process was repeated 5,000 times. The proportion of times when T_2 falls in $(F_{\alpha/2}\ ,\ F_{1-\alpha/2})$ is the simulated coverage probability. Results are presented in Table IV (see Section 2.2.3).

2.2.2 Bayesian Interval for the Mean of the Gamma Distribution

This section explores a Bayesian approach to calculate an interval estimate for the gamma mean. The Bayesian approach specifies prior knowledge about parameters or the quantity of interest as a prior density. Miller (1980) proposed a conjugate prior density for the gamma parameters. Let X be a gamma distributed variable with the shape parameter κ and the inverse scale parameter β, denoted by $X \sim Gamma(\kappa, \beta)$ following Miller's notation. Note that β is the reciprocal of θ in (2.8). The joint conjugate prior density for (κ, β) with hyperparameters (p, q, r, s) proposed by Miller is

$$\pi(\kappa, \beta) = \frac{1}{C}\frac{\beta^{\kappa s-1}}{\Gamma(\kappa)^r}\ p^{\kappa-1}\exp(-q\beta)\ ,$$

for $\kappa > 0, \beta > 0$, and C is a normalizing constant. This prior density implies past data or a hypothetical experiment with a sample size $r(= s)$, a sum of observations q, and a product of observations p. Incorporating this prior density into analysis is the same as adding r more data points.

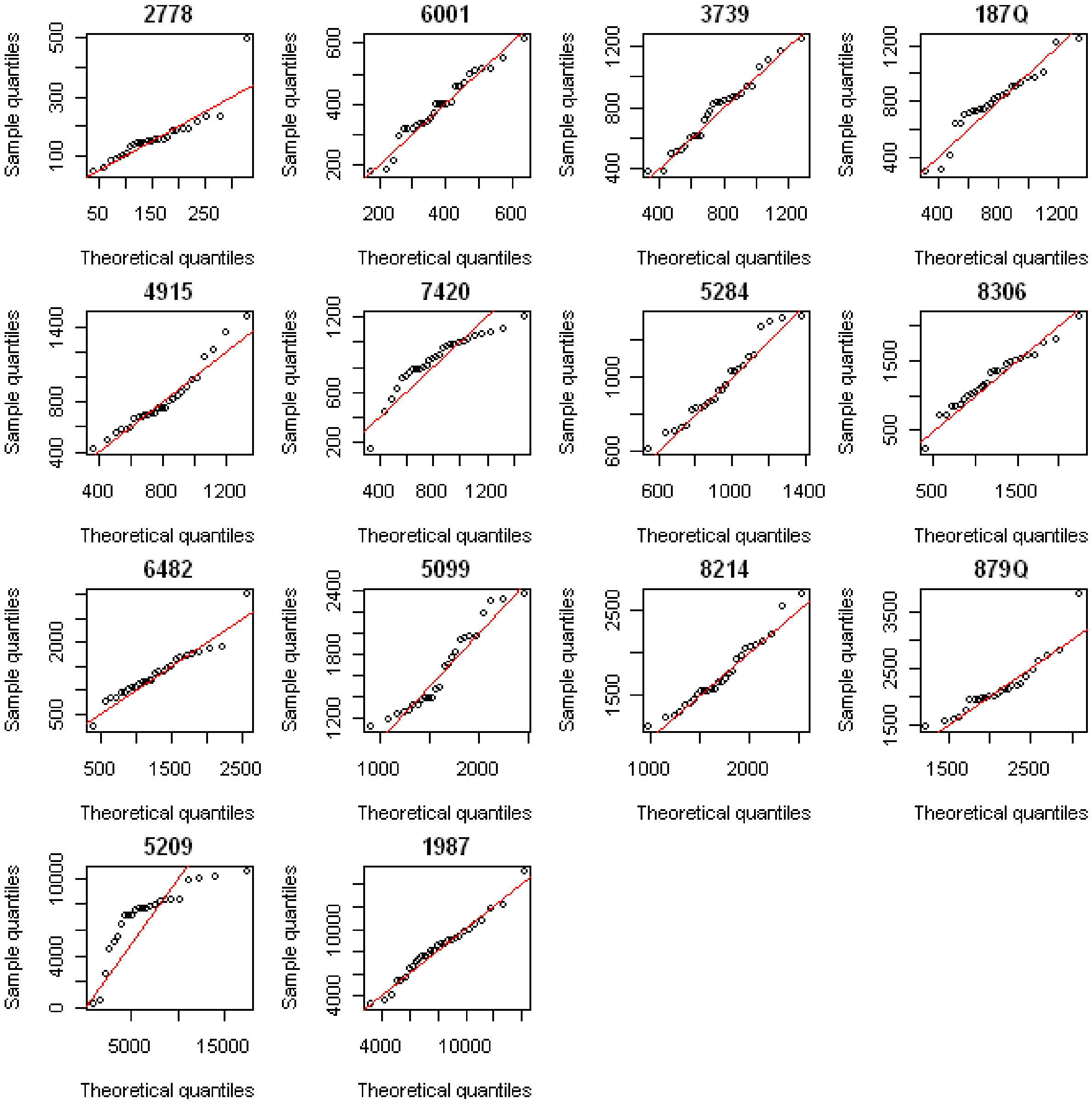

Figure 6. Gamma quantile-quantile plots for all sites.

The joint posterior density of (κ, β) with this prior is

$$p(\kappa, \beta | x_1, \ldots, x_n) \propto \frac{\beta^{\kappa(n+s)-1}}{\Gamma(\kappa)^{n+r}} \left(p \prod x_i\right)^{\kappa-1} e^{-(q+\sum x_i)\beta} .$$

If we set $r = s = q = 0$ and $p = 1$ in $\pi(\kappa, \beta)$, we obtain a non-informative prior density $\pi(\kappa, \beta) \propto 1/\beta$. Although it is an improper density, the corresponding posterior density is proper. Another non-informative prior density Miller (1980) suggested is $\pi(\kappa, \beta) \propto 1/(\kappa\beta)$. The HPD interval for the gamma mean, $\kappa\theta$, can be obtained in a similar way described in Section 2.1.2. Once the joint posterior density $p(\kappa, \beta | x_1, \ldots, x_n)$ is obtained using the specified prior density (I specified $\pi(\kappa, \beta) \propto 1/\beta$), random draws $(\kappa^{\ell}, \beta^{\ell})$ are simulated from the posterior distribution for ℓ =1, …, 100,000. The contour plot of $p(\kappa, \beta | x_1, \ldots, x_n)$ and simulated draws of $(\kappa^{\ell}, \beta^{\ell})$ of the site 2778 are given in Figure 7A. As the Figure shows, κ and β are positively correlated; the marginal posterior density of κ is right skewed as is the marginal posterior density of β. The draws from the posterior density of a gamma mean, which is the quantity of interest, can be easily obtained by computing $\kappa^{\ell}/\beta^{\ell}$ for all ℓ. Figure 7B shows the posterior distribution of the mean for the site 2778. The 95% HPD interval for the mean is (131.56, 189.53). The posterior mean is the average of simulated values $\kappa^{\ell}/\beta^{\ell}$, which is 160.34. This posterior mean should be close to the MLE, $\hat{\kappa}\hat{\theta}$, because a non-informative prior was used.

I computed 95% HPD intervals for all sites, assuming the non-informative prior $\pi(\kappa, \beta) \propto 1/\beta$. The results along with the coverage probability are presented in the next section.

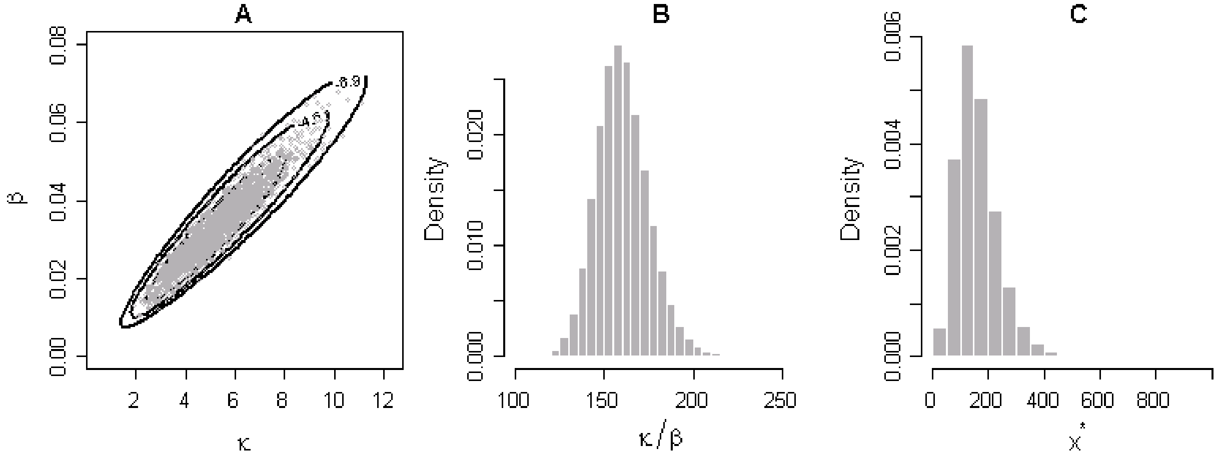

Figure 7. Posterior distributions relating to the gamma distribution for site 2778: (A) Contour plot of the joint posterior density, $p(\kappa, \beta|x_1, \dots, x_n)$, and random draws from the density. (B) Histogram of the posterior density of the gamma mean. (C) Histogram of the posterior predictive density for a new gamma random variable given the random draws in (A).

2.2.3 Comparison of Intervals for the Gamma Mean

In this section, using the amosite asbestos data, I compare the confidence interval developed by Bhaumik et al. (2009) and the HPD interval for the mean of gamma variables in terms of the lengths, locations, and coverage probabilities.

Table IV shows the two confidence intervals for all sites. The lengths of the Bayesian HPD intervals are shorter than those of the confidence interval (2.10) based on the test T_2 over all sites. The coverage probabilities of the HPD intervals are close to 0.95 (though slightly larger than 0.95 for a few sites). It is interesting that the HPD intervals have shorter length with similar coverage probabilities compared to the corresponding interval (2.10). The coverage probability of the interval (2.10) was simulated by counting the number of times that the interval $(F_{\alpha/2},\ F_{1-\alpha/2})$ included the true value of $T_2(\kappa,\theta)$, while the coverage probability of the Bayesian interval was simulated by counting the number of times that the HPD interval included the true value of the gamma mean, $\kappa\theta$.

I re-computed the coverage probability of (2.10) by checking whether the true mean μ_0 fell within the confidence interval (2.10), instead of comparing $T_2(\kappa,\theta)$ to the interval $(F_{\alpha/2},\ F_{1-\alpha/2})$. I obtained somewhat higher coverage probabilities close to 0.97 for all sites. This happened because a highly nonlinear function $T_2(\kappa,\theta)$ was inverted to construct the confidence interval and the approximation of $L \cong 0$ was used.

TABLE IV

95% CONFIDENCE INTERVALS FOR THE MEAN OF A GAMMA DISTRIBUTION

	Confidence Intervals (2.10)				Bayesian HPD intervals			
Sites	2.5%	97.5%	Length	C.P.[b]	Lower	Upper	Length	C.P.[b]
2778	128.96	199.15	70.19	0.949	131.56	189.53	57.97	0.949
6001	339.48	445.86	106.37	0.953	343.81	437.29	93.48	0.948
3739	669.63	887.25	217.62	0.949	676.98	868.05	191.06	0.960
187Q	678.22	923.30	245.07	0.955	688.74	901.49	212.75	0.956
4915	701.56	923.90	222.34	0.954	709.60	905.91	196.31	0.957
7420	733.62	992.19	258.57	0.950	742.55	969.27	226.71	0.958
5284	859.90	1057.27	197.36	0.950	861.08	1043.96	182.88	0.966
8306	1026.72	1458.67	431.94	0.951	1044.58	1418.15	373.57	0.948
6482	1100.81	1615.01	514.20	0.952	1120.53	1555.33	434.80	0.954
5099	1476.98	1846.85	369.87	0.953	1484.86	1824.87	340.02	0.963
8214	1567.77	1919.82	352.06	0.948	1573.40	1898.67	325.27	0.963
879Q	1933.66	2371.54	437.88	0.952	1941.71	2350.66	408.95	0.963
5209[a]	5118.76	9645.21	4526.45	0.954	5268.75	8913.36	3644.61	0.948
1987	7031.48	9477.14	2445.66	0.948	7126.94	9254.15	2127.21	0.956

[a] Gamma distribution does not fit the measurements from the site ($p < 0.05$).

[b] Coverage probabilities.

2.2.4 Prediction Interval for a Single New Gamma Observation

In this section I explore three methods for constructing prediction intervals for a single new observation as the following: (i) the normal distribution-based method suggested by Krishnamoorthy et al. (2008), (ii) another normal-based method by Aryal et al. (2008), and (iii) an HPD interval based on the posterior predictive distribution. These prediction intervals are applied to the amosite asbestos data.

Wilson and Hilferty (1931) provided a normal approximation to the cube root of a chi-squared variable. Krishnamoorthy et al. (2008) extended the cube root transformation technique of a chi-squared variable to a gamma variable. Let X be a gamma random variable, i.e. $X \sim Gamma(\kappa, \theta)$. An approximate distribution of $X^{1/3}$ is a normal distribution with mean and variance,

$$\mu = \frac{\theta^{1/3}\Gamma(\kappa + 1/3)}{\Gamma(\kappa)}$$

$$\sigma^2 = \frac{\theta^{2/3}\Gamma(\kappa + 2/3)}{\Gamma(\kappa)} - \mu^2.$$

Krishnamoorthy et al. (2008) compared the Wilson-Hilferty normal approximation with another normal approximation technique based on the quartic root, suggested by Hawkins and Wixley (1986). The two techniques were compared based on quantiles of gamma distributions with various shape parameters and a scale parameter of 1. Overall, the Wilson-Hilferty method is more accurate, but when the shape parameter is small $(\kappa \leq 2)$ the Hawkins and Wixley method provides more accurate values for low quantiles of gamma distributions.

Krishnamoorthy et al. (2008) developed prediction limits using this cube root transformation. Suppose the pdf of X is (2.8) and $Z = X^{1/3}$. The $(1-\alpha)100\%$ prediction interval for X is

$$\left(\bar{Z} \mp t_{1-\alpha/2,n-1}\, S_Z \sqrt{1 + \tfrac{1}{n}}\right)^3, \tag{2.11}$$

where S_Z is the standard deviation of Z, $\bar{Z}$ is the sample mean of Z, and $t_{1-\alpha/2,n-1}$ is the $(1-\alpha/2)^{th}$ percentile of a t-distribution with degrees of freedom $n-1$. This interval is easily computed and the two gamma parameters need not be estimated. A drawback of this method is that the lower limit can be negative. In that case, it is commonly set to zero.

Instead of the cube root transformation, the log transformation can be used to construct the prediction interval for a gamma distributed variable (Aryal et al., 2008). When the shape parameter κ is large, say $\kappa > 7,$ a log-transformed gamma random variable has approximately a normal distribution with mean $\mu = \psi(\kappa) + \ln(\theta)$ and variance $\sigma^2 = \psi'(\kappa)$, where $\psi(\cdot)$ is a digamma function and $\psi'(\cdot)$ is a trigamma function. κ and θ are substituted by their MLEs to estimate μ and σ^2. The prediction interval for a gamma random variable using this approximation according to Aryal et al. (2008) is

$$exp\left(\hat{\mu} \mp t_{1-\alpha/2,n-1}\hat{\sigma}\sqrt{1+\frac{1}{n}}\right). \qquad (2.12)$$

This interval can be used when the shape parameter estimate $\hat{\kappa}$ is larger than 7. Although κ and θ need be estimated to obtain the interval (2.12), it is not difficult to compute. An advantage of this interval (2.12) over interval (2.11) is that its lower limit is never negative.

The third method, Bayesian HPD interval based on the posterior predictive distribution of x^* given (κ, β) can be obtained in a similar way described in Section 2.1.3. Again, it is not necessary to derive the mathematical expression of the prediction interval, nor the posterior predictive density. Figure 7C shows the posterior predictive density for the site 2778, of which the 95% HPD interval is (31.38, 307.14).

The 95% prediction intervals (2.11), (2.12), and Bayesian HPD interval for each site and the corresponding coverage probabilities are presented in Table V. The coverage probabilities are computed by checking a newly simulated gamma variate with the prediction interval. Although $\hat{\kappa}$ is smaller than 7 for three sites (the sites 2778, 6482, 5209), coverage probabilities are close to 0.95, indicating that the interval (2.12) performs well even when κ is small. The Bayesian HPD interval for prediction has the shortest length among the three intervals, but it also has slightly lower coverage probabilities than 0.95. In addition, the Bayesian HPD prediction interval is located to the left of the others due to asymmetry. The prediction interval (2.12) is the longest among the three methods.

In summary, the gamma distribution fits the amosite asbestos data from NYSDOH better than the lognormal distribution. For the sites where both distributions fit the data, the confidence intervals relevant to the lognormal distribution give shorter length, but the gain is not substantial. For the sites where the gamma distributions fit, but the lognormal distributions do not, the gamma distribution-based approaches tend to have shorter intervals.

TABLE V

95% PREDICTION INTERVALS OF A GAMMA RANDOM VARIABLE

	A				B				C			
	Prediction Interval (2.11)[b]				Prediction Interval (2.12)[c]				Bayesian HPD interval			
Sites	2.5%	97.5%	Length	C.P.	2.5%	97.5%	Length	C.P.	Lower	Upper	Length	C.P.
2778	45.51	345.30	299.79	0.948	53.39	385.04	331.64	0.949	34.64	309.06	274.42	0.945
6001	187.44	665.01	477.57	0.950	199.97	690.80	490.83	0.949	174.62	623.35	448.73	0.944
3739	360.88	1337.18	976.30	0.949	386.93	1392.96	1006.03	0.946	326.94	1247.08	920.14	0.944
187Q	353.85	1408.01	1054.16	0.956	381.53	1473.81	1092.28	0.947	316.09	1302.81	986.72	0.939
4915	383.95	1381.44	997.48	0.943	410.81	1437.32	1026.51	0.946	357.18	1298.95	941.77	0.941
7420	372.84	1538.78	1165.94	0.949	398.79	1624.28	1225.49	0.945	332.41	1437.48	1105.07	0.941
5284	572.30	1432.67	860.37	0.954	594.23	1457.78	863.55	0.945	561.92	1382.22	820.30	0.944
8306	459.98	2367.62	1907.63	0.952	506.89	2539.02	2032.14	0.951	396.13	2180.31	1784.18	0.944
6482	448.02	2704.22	2256.20	0.955	503.75	2946.28	2442.53	0.946	359.78	2448.77	2088.98	0.941
5099	945.54	2555.55	1610.00	0.951	987.66	2610.04	1622.38	0.946	930.01	2453.73	1523.73	0.945
8214	1030.17	2623.78	1593.61	0.947	1070.14	2674.00	1603.86	0.950	1006.87	2536.36	1529.49	0.944
879Q	1294.39	3201.67	1907.28	0.950	1342.70	3257.94	1915.23	0.947	1258.25	3067.03	1808.78	0.946
5209[a]	1095.37	18640.20	17544.83	0.946	1364.46	23495.11	22130.64	0.945	356.44	15699.85	15343.41	0.945
1987	3498.03	14814.55	11316.52	0.951	3793.84	15603.38	11809.54	0.952	3173.50	13823.13	10649.63	0.944

[a] Gamma distribution does not fit the data ($p < 0.05$).

[b] Derived by Krishnamoorthy et al. (2008).

[c] Derived by Aryal et al. (2008).

3. MIXED-EFFECTS MODELS AND CALIBRATION

In this chapter, I propose two mixed-effects regression models to estimate the calibration function for environmental monitoring samples incorporating heteroscedasticity and between-laboratory variation. I describe each model and show how to estimate parameters. Then I derive a point estimate and the corresponding confidence interval for a true unknown concentration of new data analyzed by a subset of laboratories. I start with a two-component mixed model.

3.1 A Two-Component Mixed Model

3.1.1 The Model

Rocke and Lorenzato (1995) proposed a two-component error model to incorporate both constant variation at near-zero concentrations and inflated variation at larger concentrations. The proposed model is

$$y_{jk} = \beta_0 + \beta_1 \mu_j e^{\eta_{jk}} + \epsilon_{jk} \, , \tag{3.1}$$

where y_{jk} is the k^{th} observed response at the j^{th} concentration, and μ_j is the true concentration. There are two independent error terms in the model; a multiplicative error η and an additive error ϵ. The distributional assumptions of error terms are $\eta_{jk} \sim N\left(0, \sigma_\eta^2\right)$ and $\epsilon_{jk} \sim N(0, \sigma_\epsilon^2)$. Therefore, the distribution of y is a mixture of a normal and a lognormal; it is approximately a normal distribution at very low concentrations ($\mu \cong 0$), and the lognormal component, e^{η}, becomes predominant at higher concentrations, compared to the normal component. The constant variability at low concentrations is explained by ϵ and the increasing variability at higher levels is explained by μe^{η}. In terms of CV, this model explains larger CVs at very low concentrations and approximately constant CVs at higher concentrations. Application and examples of model (3.1) is described by Rocke et al. (2003).

The mean and variance of the response are

$$E(Y_{jk}) = \beta_0 + \beta_1 \mu_j \exp(\sigma_\eta^2/2)\,,$$
$$\text{and} \quad V(Y_{jk}) = (\beta_1 \mu_j)^2 e^{\sigma_\eta^2}(e^{\sigma_\eta^2} - 1) + \sigma_\epsilon^2\,. \tag{3.2}$$

The variance is a quadratic function of the true concentration. At near-zero concentrations $V(Y_{jk}) \approx \sigma_\epsilon^2$, while at large concentrations $V(Y_{jk})^{1/2}$ is roughly proportional to $E(Y_{jk})$.

Although model (3.1) explains well the non-constant variability inherent in environmental monitoring samples, it does not capture the between-laboratory variability. Gibbons and Bhaumik (2001) and Bhaumik and Gibbons (2005) extended model (3.1) to the model suitable for inter-laboratory data. When multiple laboratories are involved, it is expected that the intercept and slope of the model would vary across laboratories due to different protocols. Generally a bivariate normal distribution is assumed on $(\beta_0,\, \beta_1)$ to explain additional variability at the laboratory level. The random-effects model is

$$y_{ijk} = (\beta_0 + u_{0i}) + (\beta_1 + u_{1i})\mu_j e^{\eta_{ijk}} + \epsilon_{ijk}\,,$$
$$\text{where} \quad (u_{0i},\, u_{1i})^T \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \Sigma_u = \begin{bmatrix}\sigma_0^2 & \sigma_{01}\\ \sigma_{01} & \sigma_1^2\end{bmatrix}\right). \tag{3.3}$$

y_{ijk} is the k^{th} replicated measurement of the j^{th} concentration within the i^{th} laboratory ($i = 1, \dots, N$; $j = 1, \dots, n_i$; $k = 1, \dots, K_{ij}$). Laboratories often do not provide replicates, i.e. $K_{ij} = 1$; in that case, we can leave out the index k. β_0 and β_1 are the overall intercept and slope parameters respectively, and u_{0i} and u_{1i} are the random intercept and slope deviations for a laboratory i. The components of Σ_u explain the between-laboratory variability. σ_ϵ^2 and σ_η^2 contribute to the within-laboratory variability.

The conditional mean and variance given a laboratory i are

$$E(Y_{ijk}|\boldsymbol{u_i}) = (\beta_0 + u_{0i}) + (\beta_1 + u_{1i})\mu_j \exp(\sigma_\eta^2/2)\,,$$
$$V(Y_{ijk}|\boldsymbol{u_i}) = (\beta_1 + u_{1i})^2 \mu_j^2 e^{\sigma_\eta^2}(e^{\sigma_\eta^2} - 1) + \sigma_\epsilon^2\,,$$

where $\boldsymbol{u_i} = (u_{0i},\, u_{1i})^T$. The marginal mean is the same as in (3.2) and the marginal variance of an observation is

$$V\big(Y_{ijk}\big) = \big(2\sigma_{01}e^{\sigma_\eta^2/2}\,\big)\mu_j + \Big(\sigma_1^2 e^{2\sigma_\eta^2} + \beta_1^2 e^{\sigma_\eta^2}\big(e^{\sigma_\eta^2} - 1\big)\Big)\,\mu_j^2 + \sigma_0^2 + \sigma_\epsilon^2. \qquad (3.4)$$

The marginal variance in (3.4) consists of two parts: a between-laboratory variance $V\big[E(Y_{ijk}|\boldsymbol{u_i})\big]$ and a within-laboratory variance $E\big[V(Y_{ijk}|\boldsymbol{u_i})\big]$, each of which are

$$V\big[E(Y_{ijk}|\boldsymbol{u_i})\big] = 2\sigma_{01}e^{\sigma_\eta^2/2}\,\mu_j + \sigma_1^2 e^{\sigma_\eta^2}\mu_j^2 + \sigma_0^2\;,$$

and

$$E\big[V(Y_{ijk}|\boldsymbol{u_i})\big] = (\sigma_1^2 + \beta_1^2)\;e^{\sigma_\eta^2}\big(e^{\sigma_\eta^2} - 1\big)\;\mu_j^2 + \sigma_\epsilon^2\;.$$

If $\sigma_{01} \geq 0$, the between-laboratory variance monotonically increases with μ_j in a quadratic manner. For near-zero concentrations ($\mu_j \simeq 0$), the between- and within-laboratory variances are close to the constant values σ_0^2 and σ_ϵ^2 respectively.

Note that the true concentration μ_j is assumed to be known in model (3.3); however, it is sometimes unknown as we saw with the NYSDOH asbestos data. When it is not known, I suggest substituting the sample mean $\bar{y}_j$ for μ_j in model equation (3.3) and in mean and variance expressions as well. In addition, the total variance (3.4) should be adjusted to reflect the additional uncertainty owing to estimating μ_j.

3.1.2 Parameter Estimation

Rocke and Lorenzato (1995) estimated parameters of model (3.1) by the MML estimation. Since the multiplicative error and additive error have normal distributions with their own variances, the likelihood function using the distributional assumption is expressed as

$$L\big(\boldsymbol{\beta}, \sigma_\epsilon, \sigma_\eta\big) = \prod_j \prod_k \int \frac{1}{2\pi\sigma_\epsilon\sigma_\eta}\,\mathrm{exp}\left\{-\frac{\big(y_{jk} - \beta_0 - \beta_1\mu_j e^{\eta_{jk}}\big)^2}{2\sigma_\epsilon^2} - \frac{\eta_{jk}^2}{2\sigma_\eta^2}\right\} d\eta_{jk}\;.$$

A numerical integration or an approximation is necessary to solve the integration over η because it cannot be analytically solved.

To estimate parameters of extended model (3.3), we can use the same strategy. Let $\boldsymbol{Y_i}$ be a vector of observations nested in laboratory i, $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$, and $\boldsymbol{u_i} = (u_{0i}, u_{1i})^T$. We need to construct the marginal likelihood function of laboratory i, which is

$$\begin{aligned} L_i\big(\boldsymbol{\beta}, \sigma_\epsilon, \sigma_\eta, \Sigma_u\big) &= \int f_i\big(\boldsymbol{Y_i} | \boldsymbol{u_i}; \boldsymbol{\beta}, \sigma_\epsilon, \sigma_\eta\big) g(\boldsymbol{u_i}; \Sigma_u)\, d\boldsymbol{u_i} \\ &= \int \prod_{j,k} \int \frac{1}{2\pi\sigma_\epsilon\sigma_\eta} \exp\left\{-\frac{\big(y_{ijk} - \beta_{0i} - \beta_{1i}\mu_j e^{\eta_{ijk}}\big)^2}{2\sigma_\epsilon^2} - \frac{\eta_{ijk}^2}{2\sigma_\eta^2}\right\} d\eta_{ijk} \qquad (3.5) \\ &\qquad \frac{|\Sigma_u|^{-1/2}}{2\pi} \exp\left(-\frac{\boldsymbol{u_i^T}\Sigma_u^{-1}\boldsymbol{u_i}}{2}\right) d\boldsymbol{u_i} \end{aligned}$$

where $g(\boldsymbol{u_i}; \Sigma_u)$ is a bivariate normal distribution with mean $\boldsymbol{0}$ and variance-covariance matrix Σ_u, $\beta_{0i} = \beta_0 + u_{0i}$, and $\beta_{1i} = \beta_1 + u_{1i}$. As expressed in (3.5), another numerical integration over the random effects is required; thus, it is mathematically and computationally very complex to solve.

Gibbons and Bhaumik (2001) proposed a more feasible estimation procedure, iteratively re-weighted maximum marginal likelihood (IWMML), to estimate the parameters in model (3.3). The IWMML is a combination of the weighted least squares and the MML with updated weights at each iteration. Gibbons and Bhaumik (2001) showed that the IWMML gives satisfactory results when replicates are available, i.e. $K_{ij} > 1$. Here, I propose a modified version of the IWMML for the situation of no replicate ($K_{ij} = 1$); therefore, the index k is omitted here. Note that there are no replicates in the NYSDOH asbestos data. The estimation procedure is as follows.

First, we estimate σ_ϵ^2 and σ_η^2 by iteratively re-weighted least squares (IWLS). $\bar{y}_j$ is substituted for μ_j when it is unknown.

1. Compute weight $\omega_j = \left(\sigma_\epsilon^2 + \big(\beta_1\mu_j\big)^2 e^{\sigma_\eta^2}\big(e^{\sigma_\eta^2} - 1\big)\right)^{-1}$ with the following initial values.

$\beta_1^{(0)}$ from fitting the linear model $y_{ij} = \beta_0 + \beta_1\mu_j + \epsilon_{ij}$ using ordinary least squares,

$\sigma_\epsilon^{(0)} = SD$ of observations at the lowest concentration,

$\sigma_\eta^{(0)} = SD$ of log-transformed observations at the largest concentration.

2. Estimate β_0 and β_1 by weighted linear regression with w_j as weights.
3. Compute a mean of squared residuals and denote

$$s^2(\mu_j) = \frac{\sum_i(\hat{\beta}_0 + \hat{\beta}_1\mu_j - y_{ij})^2}{m_j},$$

where $m_j =$ the number of laboratories that analyzed samples of concentration level j.

4. Fit a variance function model

$$s^2(\mu_j) = \gamma + \delta\mu_j^2 + e_j ,$$

where the slope $\delta = \beta_1^2 e^{\sigma_\eta^2}\left(e^{\sigma_\eta^2} - 1\right)$ using weights $m_j/s^2(\mu_j)$.

5. Update $\sigma_\epsilon^2 = \gamma$ and $\sigma_\eta^2 = \ln\left[\left(1 + \sqrt{1 + 4\delta/\beta_1^2}\right)/2\right]$.
6. Iterate steps 1-5 until convergence. Obtain $\hat{\sigma}_\epsilon^2$ and $\hat{\sigma}_\eta^2$.

Second, the β_0 and β_1 are updated by the MML method, given $\hat{\sigma}_\epsilon^2$ and $\hat{\sigma}_\eta^2$.

7. For each laboratory separately, estimate β_{0i} and β_{1i} using ordinary least squares.
8. Construct weights $\omega_j = \left(\hat{\sigma}_\epsilon^2 + \left(\hat{\beta}_{1i}\mu_j\right)^2 e^{\hat{\sigma}_\eta^2}\left(e^{\hat{\sigma}_\eta^2} - 1\right)\right)^{-1}$.
9. Denote $y_{ij}^* = \sqrt{\omega_{ij}}\, y_{ij}$, $\mu_j^* = \sqrt{\omega_{ij}}\, \mu_j$, and $\epsilon_{ij}^* = \sqrt{\omega_{ij}}\, \epsilon_{ij}$. Estimate β_0, β_1, Σ_u, and the laboratory-specific parameters (β_{0i}, β_{1i}) by fitting the linear mixed model with initial values obtained in Step 7.
10. Update $\hat{\sigma}_\epsilon^2$ and $\hat{\sigma}_\eta^2$ by fitting the variance function model in step 4.

The IWLS and MML are alternated until all parameters converge. In order to study the performance of this modified version of the IWMML, a simulation study was conducted. The scenarios and results of the simulation study are described in Chapter 4.

When there are replicates per concentration level within a laboratory, the first 6 steps are carried out separately for each laboratory. Thus, $\hat{\sigma}_{\epsilon i}^2$ and $\hat{\sigma}_{\eta i}^2$ are estimated for all i and averaged to obtain $\sum_i \hat{\sigma}_{\epsilon i}^2/N$ and $\sum_i \hat{\sigma}_{\eta i}^2/N$, which are then passed to step 7.

The following sections describe how to estimate μ_j of new environmental samples and its confidence interval for the purpose of calibration.

3.1.3 Point Estimation

Consider that we have collected another set of samples that exhibit similar characteristics to background data, but the true concentration is unknown. Denote the vector of these samples by $\boldsymbol{Y}$. Also, suppose m_j samples were collected for the j^{th} level of concentration (or one sample was split into m_j subsamples) and they were sent to different laboratories to measure concentrations, and we now wish to estimate the true concentration μ_j. The estimate of μ_j inverting the model equation (Bhaumik and Gibbons, 2005) is

$$\hat{\mu}_j = \frac{1}{m_j} \sum_{i=1}^{m_j} \frac{y_{ij} - \hat{\beta}_{0i}}{\hat{\beta}_{1i} \exp(\hat{\sigma}_\eta^2/2)}. \tag{3.6}$$

$\hat{\beta}_{0i}$ and $\hat{\beta}_{1i}$ are the i^{th} laboratory-specific estimates of intercepts and slopes obtained by using the IWMML from the background data. This point estimate in (3.6) is asymptotically unbiased. Rocke et al. (2003) suggested a point estimate, $(y_{ij} - \hat{\beta}_0)/\hat{\beta}_1$, based on model equation (3.1). This estimate is positively biased because its expected value is $\mu_j e^{\sigma_\eta^2/2} > \mu_j$.

3.1.4 Interval Estimation

The confidence interval for μ of new data is called a *calibration confidence interval.* Carroll and Ruppert (1988) define the calibration confidence interval as follows. Suppose $f(x_0, \beta) = y_0$. Given the values of the response y_0, the usual estimate of x_0 is the set of all values x for which $f(x, \hat{\beta}) = y_0$. I borrow this concept to derive the calibration confidence interval for μ.

An approximate confidence interval can be constructed depending upon the level of concentrations, by properly using distributions of the corresponding observations. For example, if the concentration level is very low ($\mu_j \simeq 0$), Y_{ij} is approximately normally distributed with mean β_0 and variance $\sigma_0^2 + \sigma_\epsilon^2$. I use this normal distribution to construct the confidence interval for near-zero concentrations. Let $\bar{y}_0 = \sum_{i=1}^{m_0} y_{i0} \,/\, m_0$ be an average of near-zero observations and m_0 is the number of the observations (i.e. the number of laboratories that analyzed the near-zero concentration). A plausible estimate of μ_0 is

$$\hat{\mu}_0 = \frac{\bar{y}_0 - \hat{\beta}_0}{\hat{\beta}_1 \exp(\hat{\sigma}_\eta^2/2)}.$$

$\hat{\mu}_0$ is an asymptotically unbiased estimate of μ_0 because of the asymptotic properties of $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}_\eta^2$. By the central limit theorem, the distribution of $\bar{Y}_0$ is

$$N\left(\beta_0 + \beta_1\mu_0 \;\exp(\sigma_\eta^2/2), \qquad \frac{V(Y_{i0})}{m_0}\right),$$

where $V(Y_{i0})$ is the expression in (3.4) with $j = 0$. By Slutsky's theorem the limiting distribution of $\hat{\mu}_0$ is a normal with mean μ_0 and variance

$$V(\hat{\mu}_0) = \frac{V(Y_{i0})}{m_0\beta_1^2 \exp(\sigma_\eta^2)}.$$

Therefore, the $100(1-\alpha)\%$ calibration confidence interval for an unknown near-zero concentration is

$$\hat{\mu}_0 \pm z_{\alpha/2} \sqrt{\frac{\hat{V}(Y_{i0})}{m_0 \hat{\beta}_1^2 \exp(\hat{\sigma}_\eta^2)}}\ , \tag{3.7}$$

where $z_{\alpha/2}$ is the $100(1-\alpha/2)^{th}$ percentile of a standard normal distribution.

For a larger concentration, the magnitude of e^η becomes dominant compared to the additive error, ϵ. Therefore, we can use a lognormal distribution to derive a confidence interval. First, denote

$$T_{ij} = \ln\left(\frac{Y_{ij} - \beta_{0i}}{\beta_{1i}\, \mu_j}\right).$$

The distribution of T_{ij} is approximately normal (*approximately* because we ignore ϵ) with mean zero and unknown variance $V(T_{ij})$. The variance is estimated as follows.

The approximate distribution of $e^{T_{ij}}$ is a lognormal with mean and variance

$$E\left(e^{T_{ij}}\right) = e^{V(T_{ij})/2},$$

$$\text{and} \quad V\left(e^{T_{ij}}\right) = e^{2V(T_{ij})} - e^{V(T_{ij})}\ .$$

$V\left(e^{T_{ij}}\right)$ is also the same as

$$V\left(\frac{Y_{ij} - \beta_{0i}}{\beta_{1i}\, \mu_j}\right) = \frac{V(Y_{ij})}{\left(\beta_{1i}\, \mu_j\right)^2}\ ,$$

where $V(Y_{ij})$ is expression (3.4). Denote this equation by c_{ij} and $\nu_{ij} = e^{V(T_{ij})}$. We can obtain ν_{ij} by solving $\nu_{ij}^2 - \nu_{ij} - c_{ij} = 0$, which is

$$\nu_{ij} = \frac{1 + \sqrt{1 + 4c_{ij}}}{2}.$$

$V(T_{ij})$ is obtained by taking the log on ν_{ij}. As a result, after substituting estimates for the parameters,

$$\hat{V}(T_{ij}) = \ln\frac{1}{2}\left\{1 + \left(1 + \frac{4\hat{V}\left(Y_{ij}\right)}{\hat{\beta}_{1i}^2\, \mu_j^2}\right)^{1/2}\right\}. \tag{3.8}$$

Note that $V(T_{ij})$ varies across laboratories i and levels of concentration j. Because of $\hat{V}\left(Y_{ij}\right)$, variance of T_{ij} partially adjusts for variability at the laboratory level.

A pivot statistic for constructing the confidence region of μ_j is

$$Z(\mu_j) = \frac{1}{\sqrt{m_j}} \sum_{i=1}^{m_j} \frac{\hat{T}_{ij}}{\sqrt{\hat{V}(T_{ij})}} \ .$$

Note that T_{ij} and $V(T_{ij})$ are functions of μ_j. An approximate distribution of $Z(\mu_j)$ is a standard normal distribution because T_{ij} follows approximately $N\left(0, V(T_{ij})\right)$. Hence the $100(1-\alpha)\%$ calibration confidence region for a large concentration consists of all μ's that satisfy the following.

$$R(\mu_j) = \{\mu_j : -z_{\alpha/2} < Z(\mu_j) < z_{\alpha/2}\} \,. \tag{3.9}$$

The performance of the point and interval estimates derived in this section is examined via simulation in Chapter 4.

3.2 A Gamma Mixed Model

3.2.1 The Model

In this section, I propose a mixed-effects regression model with gamma–distributed errors to estimate the calibration curve and present how to obtain the calibration confidence interval for unknown concentrations.

First, consider a multiplicative regression model for a gamma distributed random variable,

$$y_{jk} = (\beta_0 + \beta_1 \mu_j)\epsilon_{jk} \,, \tag{3.10}$$

where μ_j is the known true concentration and y_{jk} is the k^{th} observation at the j^{th} concentration level. I assume that ϵ_{jk} is independently and identically distributed as $Gamma(\kappa, 1/\kappa)$. Note that there is restriction in the scale parameter; this makes the mean of errors to be 1. As a result, the mean and variance of Y_{jk} are

$$E\big(Y_{jk}\big) = \beta_0 + \beta_1\mu_j\ , \quad \text{and} \quad V\big(Y_{jk}\big) = \frac{\big(\beta_0 + \beta_1\mu_j\big)^2}{\kappa}\,.$$

Note that $V\big(Y_{jk}\big)$ is a quadratic function of μ_j. This property was also shown in (3.2) for the two-component error model. Let θ_j denote

$$\theta_j = \frac{\beta_0 + \beta_1\mu_j}{\kappa}\,. \tag{3.11}$$

Then, Y_{jk} has a gamma distribution with the shape parameter κ and the scale parameter θ_j, i.e.

$$f\big(y_{jk}\big) = \frac{1}{\Gamma(\kappa)\theta_j^{\kappa}}\ y_{jk}^{\kappa-1}\ e^{-y_{jk}/\theta_j}\ \ ,\ \text{for}\ y_{jk} > 0, \kappa > 0,\ \theta_j > 0\ .$$

I extend model (3.10) to a mixed-effects model by including random laboratory effects to account for the between-laboratory variability. The model is

$$E\big(Y_{ijk}|\ \boldsymbol{u_i}\big) = (\beta_0 + u_{0i}) + (\beta_1 + u_{1i})\mu_j. \tag{3.12}$$

Model (3.12) can be expressed as

$$y_{ijk} = \big\{(\beta_0 + u_{0i}) + (\beta_1 + u_{1i})\mu_j\big\}\epsilon_{ij}\ .$$

y_{ijk} is the k^{th} replicate of the j^{th} concentration nested in the i^{th} laboratory ($i = 1, \dots, N;\ j = 1, \dots, n_i;\ k = 1, \dots, K_{ij}$). When there is no replicate the index k is omitted. I assume that ϵ_{ijk} is independently and identically distributed as $Gamma(\kappa, 1/\kappa)$ within a given lab. Therefore, the distribution of Y_{ijk} conditional on laboratory i is $Gamma(\kappa, \theta_{ij})$, where

$$\theta_{ij} = \frac{(\beta_0 + u_{0i}) + (\beta_1 + u_{1i})\mu_j}{\kappa}\,. \tag{3.13}$$

μ_j is replaced with $\bar{y}_j$ when it is unknown. The assumption of a common shape parameter κ should be justified by checking the distribution of $\boldsymbol{Y_i}$ within each laboratory; this assumption appears to be reasonable with the amosite asbestos data from NYSDOH (more detail is described in the Illustration). I also assume a bivariate normal distribution with mean $\mathbf{0}$ and variance matrix Σ_u for $\boldsymbol{u_i}$.

Note that the scale parameter in (3.13) is varied across laboratories and concentrations. θ_{ij} is normally distributed with mean $(\beta_0 + \beta_1\mu_j)/\kappa$ and variance $(\sigma_0^2 + 2\sigma_{01}\mu_j + \sigma_1^2\mu_j^2)/\kappa^2$ because of the bivariate normal assumption on $\boldsymbol{u_i}$. The normality assumption on $\boldsymbol{u_i}$ can produce negative θ_{ij} whereas it should be positive being the scale parameter. The numerator of (3.13) could be negative especially when $\mu_j \cong 0$. A bivariate gamma distribution or a bivariate lognormal distribution would be an alternative to normality assumption, while the normality assumption makes estimation procedure more feasible.

The marginal mean and variance of Y_{ijk} are

$$E(Y_{ijk}) = \beta_0 + \beta_1\mu_j \ ,$$

and
$$V(Y_{ijk}) = \left(\frac{\beta_0^2 + \sigma_0^2}{\kappa} + \sigma_0^2\right) + 2\left(\frac{\beta_0\beta_1 + \sigma_{01}}{\kappa} + \sigma_{01}\right)\mu_j + \left(\frac{\beta_1^2 + \sigma_1^2}{\kappa} + \sigma_1^2\right)\mu_j^2 \ . \tag{3.14}$$

The marginal variance in (3.14) consists of two parts: the between-laboratory variance $V\left[E(Y_{ijk}|\boldsymbol{u_i})\right]$ and the within-laboratory variance $E\left[V(Y_{ijk}|\boldsymbol{u_i})\right]$, each of which are

$$V\left[E(Y_{ijk}|\boldsymbol{u_i})\right] = \sigma_0^2 + 2\sigma_{01}\mu_j + \sigma_1^2\mu_j^2 \ ,$$

$$E\left[V(Y_{ijk}|\boldsymbol{u_i})\right] = E(\kappa\theta_{ij}^2) = \kappa\left[V(\theta_{ij}) + E^2(\theta_{ij})\right]$$
$$= \left\{\beta_0^2 + \sigma_0^2 + 2(\beta_0\beta_1 + \sigma_{01})\mu_j + (\beta_1^2 + \sigma_1^2)\mu_j^2\right\}/\kappa \ .$$

Similar to the two-component mixed model described in Section 3.1.1, the between- and within-laboratory variances are quadratic functions of μ_j and they monotonically increase with the concentration levels if $\sigma_{01} \geq 0$.

3.2.2 Parameter Estimation

I propose to estimate the parameters in model (3.12) by MML. Here focus is on the situation when there is no replicate per concentration level within a laboratory. The observations within laboratory i are assumed to be conditionally independent. Let $\boldsymbol{Y_i}$ be a vector of observations nested in laboratory i, $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$, and $\boldsymbol{u_i} = (u_{0i}, u_{1i})^T$. The conditional density of $\boldsymbol{Y_i}$ given $\boldsymbol{u_i}$ is

$$f_i(\boldsymbol{Y_i}|\boldsymbol{u_i}; \boldsymbol{\beta}, \kappa) = \prod_j \frac{y_{ij}^{\kappa-1}}{\Gamma(\kappa)\theta_{ij}^{\kappa}} \exp\left(-\frac{y_{ij}}{\theta_{ij}}\right).$$

Note that θ_{ij} is a function of $\boldsymbol{\beta}$, κ, and $\boldsymbol{u_i}$. The marginal likelihood function of laboratory i is

$$\begin{aligned} L_i(\boldsymbol{\beta}, \kappa, \Sigma_u) &= \int f_i(\boldsymbol{Y_i}|\boldsymbol{u_i}; \boldsymbol{\beta}, \kappa)\, g(\boldsymbol{u_i}; \Sigma_u) d\boldsymbol{u_i} \\ &= \prod_j \frac{1}{2\pi} \int \frac{y_{ij}^{\kappa-1}}{\Gamma(\kappa)\theta_{ij}^{\kappa}} |\Sigma_u|^{-1/2} \exp\left(-\frac{y_{ij}}{\theta_{ij}} - \frac{\boldsymbol{u_i}^T \Sigma_u^{-1} \boldsymbol{u_i}}{2}\right) d\boldsymbol{u_i}\,, \end{aligned} \tag{3.15}$$

where $g(\boldsymbol{u_i}; \Sigma_u)$ is a bivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix Σ_u. The full marginal likelihood from all laboratories is formed by multiplying the marginal likelihood functions (3.15) of all laboratories. Parameter estimates of $(\boldsymbol{\beta}, \kappa, \Sigma_u)$ are obtained by maximizing this full marginal likelihood function (or its logarithm). However, equation (3.15) cannot be evaluated in a closed-form, so numerical integration or an approximation is required.

Several methods for evaluating the integral over the random effects distribution have been developed. Among those, adaptive Gauss-Hermite quadrature (AGQ) and Laplace approximation are widely used. AGQ computes the integral by direct numerical evaluation. The area under the curve (3.15) is calculated by summing over small areas; therefore, the accuracy of approximation depends on the number of these small areas represented by quadrature points and weights. The precision of AGQ increases as the number of quadrature points increases (Lessaffre and Spiessens, 2001); however, computational burden increases exponentially as well (Hedeker and Gibbons, 2006). The Laplace method can also be used to approximate the marginal likelihood function. Different from AGQ, the

Laplace method is not a direct numerical integration; but it expands the integrand around the mode $\widetilde{\boldsymbol{u}}_i$ by using Taylor series expansion, and then evaluates the integral by Laplace's method (Raudenbush et al., 2000). The accuracy of Laplace approximation depends on how far you want to expand the integrand. The magnitude of higher-order terms diminishes as n_i increases; therefore, its accuracy depends on the sample size as well. It is known that the Laplace method is computationally less intensive than AGQ; however, it may produce less accurate estimates for small n_i (Clarkson et al., 2002; Diaz, 2007).

In the context of environmental monitoring, it is often not available for a laboratory to analyze a large number of samples. In the NYSDOH amosite data, there are less than 15 observations within a laboratory, i.e. $n_i < 15$. Therefore, AGQ appears to be more appropriate for my purpose to calculate (3.15). Then, in order to find parameter estimates that maximize $\sum_i \ln L_i(\boldsymbol{\beta}, \kappa, \Sigma_u)$, I use Newton-Raphson method.

Once the MLEs of $\beta, \kappa,$ and Σ_u are obtained, we "predict" the random effects. The laboratory-specific random deviation from the overall intercept and slope, $\boldsymbol{u_i}$, is predicted by the EB estimation. The best prediction of a random effect is its conditional mean, given the available data $\boldsymbol{Y_i}$ (and therefore, $\widehat{\boldsymbol{\beta}}$ as well): i.e. $E(\boldsymbol{u_i}|\boldsymbol{Y}_i; \widehat{\boldsymbol{\beta}}, \kappa, \Sigma_u)$. This is called the Best Linear Unbiased Predictor (BLUP). Since κ and Σ_u are unknown, we replace them by their maximum likelihood estimates, $\hat{\kappa}$ and $\hat{\Sigma}_u$. Therefore, the predicted random effects for the i^{th} laboratory are

$$\widehat{\boldsymbol{u}}_{\boldsymbol{i}} = E(\boldsymbol{u_i}|\boldsymbol{Y}_i; \widehat{\boldsymbol{\beta}}, \hat{\kappa}, \hat{\Sigma}_u) .$$

This equation requires integration over the random-effects distribution. Analytic evaluation for $\widehat{\boldsymbol{u}}_{\boldsymbol{i}}$ is not available, so another numerical integration is involved (Fitzmaurice et al., 2004).

Given starting values of parameters, the solution alternates between maximizing the marginal likelihood function and the EB estimation until all parameters converge. I use AGQ technique implemented in SAS NLMIXED procedure (Pinheiro and Bates, 1995) to calculate $\sum_i \ln L_i(\boldsymbol{\beta}, \kappa, \Sigma_u)$.

Between 10 and 20 quadrature points are usually suggested, but one should monitor whether parameter estimates change as the number of quadrature points increases. The quadrature points are adaptively chosen to be centered around the EB estimates $\hat{\boldsymbol{u}}_{\boldsymbol{i}}$ at each iteration. The Hessian matrix is involved in Newton-Raphson method, and it is also used to compute standard errors of parameter estimates upon the convergence.

3.2.3 Point Estimation

In this section I propose a point estimate for the unknown true concentration using the gamma mixed model (3.12). Suppose we collected an independent set of environmental monitoring samples that exhibit similar characteristics, and the true concentration is unknown. Denote the vector of these samples by $\boldsymbol{Y}$. Again, consider that m_j samples were collected for the j^{th} level of concentration and those were sent to m_j laboratories to measure concentrations. We are interested in estimating the true concentration μ_j for these new data.

The point estimate for the true concentration level j that I propose is

$$\hat{\mu}_j = \frac{1}{m_j}\sum_{i=1}^{m_j}\frac{y_{ij} - \hat{\beta}_{0i}}{\hat{\beta}_{1i}} \ . \tag{3.16}$$

$\hat{\beta}_{0i} = \hat{\beta}_0 + \hat{u}_{0i}$ and $\hat{\beta}_{1i} = \hat{\beta}_1 + \hat{u}_{1i}$ are laboratory-specific intercept and slope estimates obtained by the MML and the empirical Bayes estimation from the background data.

3.2.4 Interval Estimation

I use the normal approximation to the cube root transformed gamma variable (Krishnamoorthy et al., 2008), described in Section 2.2.3, to construct the calibration confidence region for μ_j. By the cube root approximation, $Y_{ij}^{1/3}$ given $\boldsymbol{u}_{\boldsymbol{i}}$ has an approximate normal distribution with mean λ_{ij} and variance τ_{ij}^2, where

$$\lambda_{ij} = \frac{\theta_{ij}^{1/3}\Gamma(\kappa + 1/3)}{\Gamma(\kappa)} \ ,$$

and $$\tau_{ij}^2 = \frac{\theta_{ij}^{2/3}\Gamma(\kappa + 2/3)}{\Gamma(\kappa)} - \lambda_{ij}^2.$$

We can construct a pivot statistic using this approximate normal distribution of $Y_{ij}^{1/3}$ given $\boldsymbol{u_i}$. The pivot statistic is

$$Z_1(\mu_j) = \frac{1}{\sqrt{m_j}} \sum_{i=1}^{m_j} \left(\frac{y_{ij}^{1/3} - \hat{\lambda}_{ij}}{\hat{\tau}_{ij}} \right),$$

which has an approximate standard normal distribution. $\hat{\lambda}_{ij}$ and $\hat{\tau}_{ij}$ depend on parameter estimates and EB estimates, $\boldsymbol{\hat{u}_i}$, obtained by fitting model (3.12) to the background data. Z_1 is a nonlinear function of μ_j, for θ_{ij} is a function of μ_j. Therefore, the $100(1-\alpha)\%$ calibration confidence region of the true concentration μ_j is constructed by collecting of all μ_j's satisfying

$$R_1(\mu_j) = \left\{ \mu_j : -z_{\alpha/2} < \ Z_1(\mu_j) < z_{\alpha/2} \right\}. \tag{3.17}$$

Note that to construct $Z_1(\mu_j)$, laboratory-specific estimates $\hat{\lambda}_{ij}$ and $\hat{\tau}_{ij}$ are used. This implies that the laboratories analyzing the new data are the same set (or a subset) of the laboratories that participated in collecting the background data. If some of new environmental samples were sent to brand new laboratories, this interval (3.17) is not relevant.

With that in mind, I also derive interval estimates that do not depend on the laboratory-specific parameters, so that participation of the same set of laboratories is not required in collecting new data. I still assume that the new data exhibit similar characteristics to the existing data. The approximate distribution of $\boldsymbol{Y_i}^{1/3}$ given $\boldsymbol{u_i}$ is a n_i–variate normal owing to the cube root transformation technique, and $\boldsymbol{u_i}$ is assumed to have a bivariate normal distribution. Hence, the distribution of ($\boldsymbol{Y_i}^{1/3}$, $\boldsymbol{u_i}$) is jointly normal (Gelman et al., 2004). As a result, the marginal distribution of $\boldsymbol{Y_i}^{1/3}$ is a n_i $-$variate normal (Rohatgi and Saleh, 2001). Using this result, an approximate marginal distribution of $Y_{ij}^{1/3}$ can be obtained: a normal distribution with mean λ_j^* and variance ${\tau_j^*}^2$ given as follows.

$$\lambda_j^* = \frac{\Gamma(\kappa + 1/3)}{\Gamma(\kappa)} \left(\frac{\beta_0 + \beta_1\mu_j}{\kappa}\right)^{1/3},$$

and

$${\tau_j^*}^2 = \left(\frac{\beta_0 + \beta_1\mu_j}{\kappa}\right)^{2/3} \left[\frac{\Gamma(\kappa + 2/3)}{\Gamma(\kappa)} - \left(\frac{\Gamma(\kappa + 1/3)}{\Gamma(\kappa)}\right)^2\right] \tag{3.18}$$

$$+ \left(\frac{\Gamma(\kappa + 1/3)}{\Gamma(\kappa)}\right)^2 \frac{\sigma_0^2 + 2\sigma_{01}\mu_j + \sigma_1^2\mu_j^2}{9\kappa^{2/3}\left(\beta_0 + \beta_1\mu_j\right)^{4/3}} \ .$$

The derivation of expressions in (3.18) is given in Appendix. Now define a pivot statistic to construct the calibration confidence interval using this result:

$$Z_2\left(\mu_j\right) = \frac{1}{\sqrt{m_j}} \sum_{i=1}^{m_j} \left(\frac{y_{ij}^{1/3} - \hat{\lambda}_j^*}{\hat{\tau}_j^*}\right).$$

The statistic $Z_2\left(\mu_j\right)$ follows approximately a standard normal distribution. As a result, the $100(1 - \alpha)\%$ calibration confidence region of the true concentration level j consists of all μ_j's satisfying

$$R_2\left(\mu_j\right) = \left\{\mu_j : -z_{\alpha/2} < Z_2\left(\mu_j\right) < z_{\alpha/2}\right\}. \tag{3.19}$$

It should be noted that $Z_2\left(\mu_j\right)$ does not depend on the laboratory-specific estimates. I call this region the *global* calibration confidence region.

The statistic $Z_2\left(\mu_j\right)$ is a complicated nonlinear function of μ_j. Here I propose another global calibration confidence interval and yet simpler approach. Y_{ij} has an unknown marginal distribution with the mean and variance derived in (3.14). By the central limit theorem,

$$\bar{Y}_j = \frac{\sum_{i=1}^{m_j} Y_{ij}}{m_j} \sim N\left(\beta_0 + \beta_1\mu_j \ , \qquad \frac{V\left(Y_{ij}\right)}{m_j}\right)$$

Using this result, I define a pivot statistic $Z_3\left(\mu_j\right)$ by

$$Z_3(\mu_j) = \frac{\bar{y}_j - (\hat{\beta}_0 + \hat{\beta}_1 \mu_j)}{\sqrt{\hat{V}(Y_{ij})/m_j}} ,$$

which also follows approximately a standard normal distribution. The possible drawback of this statistic is that m_j should not be small because it is based on the central limit theorem. The $100(1-\alpha)\%$ calibration confidence region of μ_j consists of all values satisfying

$$R_3(\mu_j) = \{\mu_j : -z_{\alpha/2} < Z_3(\mu_j) < z_{\alpha/2}\} . \tag{3.20}$$

The performance of the point estimate (3.16) and calibration confidence regions (3.17), (3.19), and (3.20) are studied via simulation in Chapter 4.

3.2.5 Other Parameterizations for a Gamma Mixed Model

I used the *identity* link function to define model (3.12). It appears that the *log* link function has been more commonly chosen in a regression model for a gamma distributed random variable. For instance, a fixed-effects gamma model with the log link is

$$y_{jk} = \exp(\beta_0 + \beta_1 \mu_j)\ \epsilon_{jk} .$$

The *inverse* link has also been suggested (McCullagh and Nelder, 1989). This section reviews other parameterizations with the *log* and *inverse* link functions for the mixed-effects regression model for a gamma distributed outcome.

The proposed model (3.12) can be re-written as the following general expression,

$$E(Y_{ij}|\boldsymbol{u_i}) = \boldsymbol{X_{ij}\beta} + \boldsymbol{Z_{ij}u_i} .$$

Y_{ij} is the j^{th} observation nested in the i^{th} group ($i = 1, \dots, N;\ j = 1, \dots, n_i$). Here, the "group" can be a geographical area in which individuals are nested or a subject whose repeated measurements over time are observed. $\boldsymbol{X_{ij}}$ is the $n_i \times p$ design matrix of explanatory variables, $\boldsymbol{\beta}$ is the p-dimensional vector of regression coefficients, $\boldsymbol{Z_{ij}}$ is the $n_i \times q$ design matrix of random effects variables, and $\boldsymbol{u_i}$ is the q-dimensional vector of random effects.

Using a similar notation, a gamma mixed-effects regression model with the *log* link function is

$$\log E\left(Y_{ij}\middle|\boldsymbol{u_i}\right) = \boldsymbol{X_{ij}\beta} + \boldsymbol{Z_{ij}u_i}\ . \tag{3.21}$$

Let μ_{ij} denote

$$\mu_{ij} = E\left(Y_{ij}\middle|\boldsymbol{u_i}\right) = \exp\{\boldsymbol{X_{ij}\beta} + \boldsymbol{Z_{ij}u_i}\}.$$

Note that μ_{ij} here does not indicate the true concentration; only in this section, μ_{ij} indicates the conditional mean of Y_{ij}. Conditional on a random effects vector, $\boldsymbol{u_i}$, the distribution of Y_{ij} is a gamma with parameters ϕ and μ_{ij}, expressed by

$$f\left(y_{ij}|\boldsymbol{u_i}\right) = \frac{1}{\Gamma(1/\phi)y_{ij}}\left(\frac{y_{ij}}{\mu_{ij}\phi}\right)^{1/\phi}\exp\left(-\frac{y_{ij}}{\mu_{ij}\phi}\right)\ \text{, for } \phi > 0, \mu > 0.$$

This notation is the same to that of McCullagh and Nelder (1989), provided $\nu = 1/\phi$. The parameter ϕ determines the shape of distribution, and the limiting distribution of Y_{ij} is a normal as $\phi \to 0$. $1/\phi$ and $\mu_{ij}\phi$ correspond to κ and θ_{ij} respectively, in the notation used in Sections 3.2.1 to 3.2.4. Thus, the conditional variance of Y_{ij} is $V\left(Y_{ij}\middle|\boldsymbol{u_i}\right) = \phi\mu_{ij}^2$, and the coefficient of variation is $\sqrt{\phi}$. According to McCullagh and Nelder (1989), when ϕ is small, the variance of log-transformed Y_{ij} is close to ϕ; therefore, the *log* link is the variance-stabilizing transformation for a gamma distributed variable. Model (3.21) has been implemented in SAS PROC GLIMMIX and SuperMix.

To estimate parameters, the MML is used in SAS PROC GLIMMIX[1]. SuperMix also uses the MML to estimate most of the parameters; but for ϕ, the method of moment estimator is used, which is the Pearson chi-square statistic divided by the degrees of freedom.

[1] Either Laplace approximation or AGQ can be selected to construct the marginal likelihood function. SAS PROC GLIMMIX also provides quasi-likelihood estimation.

Another possible gamma mixed-effects model is

$$\frac{1}{\mu_{ij}} = \frac{1}{E\left(Y_{ij}\middle|\boldsymbol{u_i}\right)} = \boldsymbol{X_{ij}\beta} + \boldsymbol{Z_{ij}u_i}\ . \tag{3.22}$$

The reciprocal transformation used in this model is the canonical link. Bailey and Alimdhi (2007) implemented model (3.22) in R package *Zelig* with the following gamma density.

$$f\left(y_{ij}|\boldsymbol{u_i}\right) = \frac{1}{\Gamma(\kappa_{ij})\theta^{\kappa_{ij}}}\ y_{ij}^{\kappa_{ij}-1} \exp\left(-\frac{y_{ij}}{\theta}\right)$$

Note that in their model, the shape parameters vary with indices i and j, but the scale parameter is held constant.

4. SIMULATION

In this chapter I examine performance of the estimation procedures for the two-component mixed model and the gamma mixed model. The performance of point estimates and confidence intervals proposed in Chapter 3 is also studied via simulation. Robustness of two methods is compared as well.

4.1 Two-Component Mixed Model

4.1.1 Parameter Estimation

I generated data sets based on the two-component mixed model in (3.3) with parameter values $\beta_0 = 0, \beta_1 = 1, \sigma_\epsilon^2 = 0.0015,\ \sigma_\eta^2 = 0.050, \sigma_0^2 = 0.002,\ \sigma_{01} = -0.0015,$ and $\sigma_1^2 = 0.025$. These parameter values were chosen to make simulated data have similar properties to the real asbestos data from NYSDOH. The true concentration levels of μ were fixed at (0.06, 0.15, 0.30, 0.40, 0.50, 0.60, 0.80, 1.2, 2.8, 4.0). The number of laboratories was set to $N = (20, 30)$ and the number of replicates for each concentration level within a laboratory was set to $K = (1, 5, 10)$. I generated laboratory random effects, (u_{0i}, u_{1i}), from the bivariate normal distribution with mean **0** and variance-covariance matrix Σ_u with components $(\sigma_0^2, \sigma_{01}, \sigma_1^2) = (0.002, -0.0015, 0.025)$. The random errors η_{ijk} and ϵ_{ijk} were generated from independent normal distributions with mean 0 and variances, $\sigma_\eta^2 = 0.05$ and $\sigma_\epsilon^2 = 0.0015$, respectively. Then y_{ijk} was simulated based on model equation (3.3). A negative value of y_{ijk} was possible due to normality assumption on (u_{0i}, u_{1i}) and predetermined parameter values. However, it is usually not realistic to have negative concentrations although theoretically model (3.3) can handle negative observations, and thus such y_{ijk} was replaced with 0.

I estimated parameters by the IWMML described in Section 3.1.2. Since there is no standard package for this estimation procedure, I programmed it using R 2.10. The whole process was replicated 1,000 times with data sets generated by different random numbers. I conducted this simulation for each combination of N and K.

For each parameter, I computed the average and the root mean square error (RMSE) over 1,000 replications to study accuracy and precision of estimates. Two scenarios were considered: (a) the true concentration μ is known, and (b) μ is unknown. For scenario (b), I substituted the sample mean $\bar{y}_j$ for μ_j for concentration level j. This scenario was intended for exploring how the sample mean substitution would affect estimation of parameters and calibration confidence intervals.

There was no non-convergence in all combinations of N and K for both scenarios. Simulation results are presented in Table VI. The overall intercept β_0 and slope β_1 are well estimated in all cases. When there is no replicate ($K = 1$), estimated variance components and error variances are less accurate as expected; however, the bias and RMSE decrease as the number of replicates increases.

TABLE VI

PARAMETER ESTIMATION OF THE TWO-COMPONENT MIXED MODEL: AVERAGE (RMSE) OVER 1,000 REPLICATIONS.

	True	N/K					
Parameters	Values	30/10	30/5	30/1	20/10	20/5	20/1
				(a) μ is known.			
β_0	0	0.0040 (0.0087)	0.0036 (0.0086)	0.0035 (0.0102)	0.0037 (0.0098)	0.0032 (0.0102)	0.0034 (0.0126)
β_1	1	1.0188 (0.0356)	1.0180 (0.0357)	1.0206 (0.0412)	1.0178 (0.0407)	1.0180 (0.0417)	1.0217 (0.0492)
σ_0^2	0.002	0.0017 (0.0005)	0.0018 (0.0005)	0.0027 (0.0010)	0.0016 (0.0006)	0.0018 (0.0006)	0.0027 (0.0011)
σ_1^2	0.025	0.0262 (0.0070)	0.0274 (0.0077)	0.0357 (0.0145)	0.0258 (0.0084)	0.0276 (0.0093)	0.0356 (0.0158)
σ_{01}	-0.0015	-0.0014 (0.0013)	-0.0016 (0.0014)	-0.0032 (0.0026)	-0.0014 (0.0015)	-0.0017 (0.0016)	-0.0034 (0.0030)
σ_ϵ^2	0.0015	0.0009 (0.0007)	0.0008 (0.0009)	0.0011 (0.0009)	0.0009 (0.0007)	0.0008 (0.0008)	0.0013 (0.0010)
σ_η^2	0.050	0.0460 (0.0049)	0.0443 (0.0067)	0.0400 (0.0133)	0.0460 (0.0052)	0.0448 (0.0070)	0.0394 (0.0148)
				(b) μ is unknown.			
β_0	0	0.0006 (0.0009)	0.0006 (0.0013)	0.0001 (0.0007)	0.0006 (0.0010)	0.0006 (0.0016)	0.0001 (0.0009)
β_1	1	0.9988 (0.0019)	0.9988 (0.0026)	0.9999 (0.0014)	0.9988 (0.0021)	0.9989 (0.0031)	0.9998 (0.0018)
σ_0^2	0.002	0.0017 (0.0005)	0.0018 (0.0005)	0.0027 (0.0010)	0.0017 (0.0006)	0.0018 (0.0006)	0.0027 (0.0012)
σ_1^2	0.025	0.0253 (0.0067)	0.0265 (0.0073)	0.0344 (0.0132)	0.0249 (0.0082)	0.0267 (0.0089)	0.0342 (0.0145)
σ_{01}	-0.0015	-0.0015 (0.0013)	-0.0017 (0.0014)	-0.0033 (0.0027)	-0.0015 (0.0015)	-0.0018 (0.0017)	-0.0035 (0.0031)
σ_ϵ^2	0.0015	0.0009 (0.0007)	0.0008 (0.0009)	0.0011 (0.0009)	0.0009 (0.0007)	0.0008 (0.0008)	0.0012 (0.0010)
σ_η^2	0.050	0.0454 (0.0053)	0.0438 (0.0072)	0.0384 (0.0142)	0.0455 (0.0055)	0.0441 (0.0074)	0.0369 (0.0162)

4.1.2 Point and Interval Estimation

In this section, I study the point and interval estimation proposed in Section 3.1.3 and 3.1.4, focusing on the situation when there is no replicate and the number of laboratories is 20 (i.e., $N/K = 20/1$). Again, two scenarios are considered; (a) the true concentration μ is known, and (b) μ is unknown in background data.

For each of 1,000 replications with $N/K = 20/1$ described in Section 4.1.1, another data set was generated assuming the same set of parameter values and true concentrations. I used predetermined laboratory-specific intercepts and slopes (simulated from a bivariate normal distribution in the process of background data generation) to simulate Y_{ij}'s. This assumes that the entire set (or a subset) of laboratories participate in collecting new samples. This newly generated data set serves as "new" data, compared to the "existing" or "background" data. Using Y_{ij}'s from the new data, and parameter estimates and predicted random effects from the background data, I computed the point estimate $\hat{\mu}$ using equation (3.6) and constructed 95% calibration confidence intervals following equations (3.7) and (3.9). Equation (3.7) was used for low concentrations $\mu = (0.06, 0.15, 0.30)$, and (3.9) for higher concentrations. This process was repeated 1,000 times, and I computed the average of $\hat{\mu}$, denoted by $\bar{\hat{\mu}}$, and the coverage probability of calibration confidence intervals. I considered the situation when 5 out of 20 laboratories analyzed new samples: this reflects the AHERA criteria in which at least five samples are required to be collected. When μ_j was considered to be unknown in the background data, it was substituted by the sample mean $\bar{y}_j$ and parameters were estimated accordingly.

Simulation results are presented in Table VII. Overall, the point estimates are very accurate on average, and the coverage probabilities are close to 0.95. For the low concentration levels ($\mu = 0.06$, 0.15), the coverage probabilities are a bit higher than the pre-specified confidence level. The calibration confidence intervals in (b) cover more large values and fewer small values, compared to those in (a). Similarly, $\bar{\hat{\mu}}$'s in (b) are higher than those in (a).

TABLE VII

CALIBRATION CONFIDENCE INTERVALS USING THE TWO-COMPONENT MIXED MODEL.

	(a) μ_j is known.			(b) μ_j is unknown.		
μ	$\bar{\hat{\mu}}$	CCI[a]	C.P.[b]	$\bar{\hat{\mu}}$	CCI[a]	C.P.[b]
0.06	0.059	(0.014, 0.111)	0.982	0.064	(0.016, 0.116)	0.971
0.15	0.145	(0.088, 0.203)	0.965	0.152	(0.094, 0.210)	0.971
0.30	0.294	(0.214, 0.374)	0.944	0.304	(0.224, 0.385)	0.942
0.40	0.394	(0.300, 0.493)	0.936	0.406	(0.312, 0.505)	0.940
0.50	0.494	(0.383, 0.616)	0.958	0.508	(0.396, 0.631)	0.945
0.60	0.596	(0.465, 0.742)	0.962	0.613	(0.480, 0.758)	0.955
0.80	0.793	(0.623, 0.988)	0.943	0.814	(0.643, 1.008)	0.940
1.20	1.193	(0.939, 1.483)	0.951	1.223	(0.966, 1.513)	0.954
2.80	2.764	(2.176, 3.445)	0.946	2.831	(2.234, 3.509)	0.943
4.00	3.977	(3.123, 4.949)	0.955	4.071	(3.206, 5.040)	0.948

[a] Calibration confidence intervals (3.7) and (3.9).

[b] Coverage probability.

4.2 Gamma Mixed Model

4.2.1 Parameter Estimation

I generated data based on the gamma mixed model in (3.12) with parameter values $\beta_0 = 0,\ \beta_1 = 1,\ \kappa = 13.0,\ \sigma_0^2 = 0.0005,\ \sigma_{01} = -0.0002,$ and $\sigma_1^2 = 0.010$. These parameter values were chosen to simulate data similar to the real asbestos data from NYSDOH. The true concentration levels were fixed at $\mu = $ (0.06, 0.15, 0.30, 0.40, 0.50, 0.60, 0.80, 1.2, 2.8, 4.0). The number of laboratories was set to $N = (20, 30)$ and the number of replicates for a concentration level within a laboratory was set to $K = (1, 5, 10)$. I generated laboratory random effects, (u_{0i}, u_{1i}), from the bivariate normal distribution with mean **0** and variance-covariance matrix Σ_u with components $(\sigma_0^2,\ \sigma_{01}, \sigma_1^2) =$ (0.0005, -0.0002, 0.01). Given predetermined values of parameters, true concentrations, and simulated values of (u_{0i}, u_{1i}), θ_{ij} was determined by equation (3.13) for laboratory i and concentration level j. Then I generated a gamma random variate y_{ijk} with the shape parameter κ and the scale parameter θ_{ij} (using built-in R function *rgamma*; Ahrens and Dieter, 1982). Negative random variates of (u_{0i}, u_{1i}) were possible; as a consequence, θ_{ij} could be negative as well. Whenever this happened, the corresponding random variate y_{ijk} was not generated, and as a result some data sets became slightly unbalanced. θ_{ij} was negative in about 0.04% of simulated data mostly when the true concentration is close to zero. It may happen more or less than that depending on parameter values and true concentrations.

I estimated parameters by the MML and EB estimation as described in Section 3.2.2. The initial values were set to parameter estimates from fitting a gamma regression model without random effects. I used the adaptive Gauss-Hermite quadrature with 15 quadrature points for numerical integration and Newton-Raphon for maximization. All estimation was done using SAS PROC NLMIXED. The process was replicated 1,000 times with data sets generated by different random numbers. The simulations were conducted for each combination of N and K.

For each parameter, I computed the average and the RMSE over 1,000 replications to examine accuracy and precision of estimates. Again two scenarios were assumed; (a) the true concentration μ is known, and (b) μ is not known in background data. For scenario (b), the sample mean $\bar{y}_j$ was substituted for μ_j for concentration level j.

Simulation results are presented in Table VIII. The overall intercept β_0 and slope β_1 estimates are excellent in all combinations of N and K. The accuracy and precision of estimated variance components tend to increase (i.e. less bias and smaller RMSE) as the number of replicates and/or the number of laboratories increase. In addition, the bias and RMSE of shape parameter decrease with the number of replicates as well. It should be noted that even when there is no replicate ($N/K = 30/1$ and $20/1$), the simulation results are satisfactory.

TABLE VIII

PARAMETER ESTIMATION OF THE GAMMA MIXED MODEL:
AVERAGE (RMSE) OVER 1,000 REPLICATIONS.

	True	N/K					
Parameters	Values	30/10	30/5	30/1	20/10	20/5	20/1
(a) μ is known.							
β_0	0	0.0003	-0.0006	0.0000	0.0002	-0.0001	-0.0005
		(0.0042)	(0.0232)	(0.0039)	(0.0052)	(0.0050)	(0.0056)
β_1	1	1.0007	0.9982	0.9982	1.0009	0.9992	0.9995
		(0.0195)	(0.0443)	(0.0194)	(0.0240)	(0.0236)	(0.0263)
σ_0^2	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005
		(0.0002)	(0.0002)	(0.0003)	(0.0002)	(0.0002)	(0.0003)
σ_{01}	-0.0002	-0.0002	-0.0003	-0.0006	-0.0003	-0.0003	-0.0008
		(0.0008)	(0.0010)	(0.0018)	(0.0010)	(0.0012)	(0.0023)
σ_1^2	0.0100	0.0096	0.0097	0.0102	0.0099	0.0095	0.0140
		(0.0034)	(0.0045)	(0.0110)	(0.0042)	(0.0049)	(0.0132)
κ	13.0	12.9424	12.6762	12.3645	13.0008	12.8254	13.6421
		(0.9905)	(1.1027)	(2.6870)	(0.4654)	(0.8462)	(2.9425)
Convergence		99.9%	97.0%	98.9%	100%	97.7%	99.5%
(b) μ is unknown.							
β_0	0	-0.0002	-0.0002	-0.0002	-0.0003	-0.0002	-0.0002
		(0.0007)	(0.0007)	(0.0007)	(0.0009)	(0.0009)	(0.0011)
β_1	1	1.0004	1.0004	1.0003	0.9987	1.0005	0.9993
		(0.0015)	(0.0016)	(0.0018)	(0.0408)	(0.0019)	(0.0331)
σ_0^2	0.0005	0.0005	0.0005	0.0004	0.0005	0.0005	0.0004
		(0.0001)	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(0.0003)
σ_{01}	-0.0002	-0.0002	-0.0002	-0.0002	-0.0002	-0.0002	-0.0002
		(0.0005)	(0.0005)	(0.0008)	(0.0006)	(0.0006)	(0.0010)
σ_1^2	0.0100	0.0097	0.0097	0.0101	0.0097	0.0094	0.0094
		(0.0028)	(0.0033)	(0.0055)	(0.0035)	(0.0038)	(0.0063)
κ	13.0	13.0505	13.0725	13.0010	13.0858	13.1660	13.4874
		(0.3507)	(0.5028)	(1.7629)	(0.4326)	(0.6343)	(1.6917)
Convergence		99.4%	100%	98.1%	100%	100%	99.7%

4.2.2 Point and Interval Estimation

Next I study the point and interval estimation of the gamma mixed-effects model proposed in Section 3.2.3 and 3.2.4, focusing on the situation when there is no replicate and the number of laboratories is 20 (i.e., $N/K = 20/1$). Similar to previous sections, two scenarios are considered; (a) the true concentration μ is known, and (b) μ is unknown in background data.

For each of 1,000 replications with $N/K = 20/1$ described in Section 4.2.1, another data set was generated assuming the same parameter values and true concentrations. I used predetermined laboratory-specific random deviations of intercept and slope (simulated in the process of background data generation) to simulate new Y_{ij}'s. This assumes that new samples are collected from the entire set (or a subset) of laboratories that analyzed background data. This newly generated data set serves as "new" data compared to the "background" data. Using Y_{ij}'s from the new data and parameter estimates and predicted random effects from the background data, I computed the point estimate $\hat{\mu}$ using equation (3.16) and constructed 95% calibration confidence intervals (3.17), (3.19), and (3.20). This process was repeated 1,000 times, and I computed the average of $\hat{\mu}$ and the coverage probabilities of calibration confidence intervals. I assumed that 5 out of 20 laboratories analyzed new samples. When μ_j was assumed to be unknown in background data, it was substituted by the sample mean $\bar{y}_j$ and model parameters were estimated accordingly.

During the simulation, I experienced unusual behaviors of the interval (3.19) for the lowest concentration, $\mu = 0.06$. The lower bound of interval (3.19) was not able to be obtained on several occasions because the pivot statistic $Z_2(\mu)$ never reached the quantile value $z_{\alpha/2} = 1.96$ within the plausible range of μ. When that happened, the lower bound was set to zero. As a result, the average lower bound is lower than those of the other intervals.

Table IX shows the simulation results for the point and interval estimates of the true concentration using the gamma mixed model. $\bar{\hat{\mu}}$'s are very close to the true values in both (a) and (b). The coverage probabilities of intervals (3.17), which depend on laboratory-specific parameters, for low concentrations, ($\mu = 0.06, 0.15$), are unsatisfactorily low. The global calibration confidence intervals (3.19) and (3.20) tend to give better coverage probabilities than the interval (3.17) although its superiority is not remarkably substantial. The global confidence intervals appear to perform better as they are based on the average calibration curve instead of laboratory-specific curves. I compare these calibration confidence intervals with respect to their robustness in the next section.

4.3 Robustness of Calibration Confidence Regions

4.3.1 Model Misspecification

There are situations when we do not know the correct model and use the best plausible model we can think of. In this section I examine whether the calibration confidence regions based on both regression models perform well even though a model is misspecified or an alternative model is chosen. To study the robustness property of the confidence regions, six scenarios are considered as displayed in Table X. Data were generated according to "true models" and analyzed by both models. Scenarios A and D were already examined in Sections 4.1.2 and 4.2.2 respectively. I focus on scenarios B and C in this section, and E and F in next section.

For scenario B, I used the data generated according to the gamma mixed model with parameter values $\beta_0 = 0, \beta_1 = 1, \kappa = 13.0,\ \sigma_0^2 = 0.0005,\ \sigma_{01} = -0.0002,$ and $\sigma_1^2 = 0.010$, and with $N = 20$ and $K = 1$. This is the same data set used for examining scenario A in Section 4.2.2. Then I fit a two-component mixed model and estimated parameters by the IWMML. Another data set was generated according to the same gamma mixed model, and I computed $\hat{\mu}$ and constructed 95% calibration confidence regions (3.7) and (3.9) for these new data. I assumed that 5 out of 20 laboratories that analyzed background data also analyzed new samples. This process was replicated 1,000 times with different random numbers.

TABLE IX

CALIBRATION CONFIDENCE INTERVALS USING THE GAMMA MIXED MODEL

(a) $\boldsymbol{\mu_j}$ is known. [a]

μ	$\bar{\hat{\mu}}$	CCI (3.17)	C.P.	CCI (3.19)	C.P.	CCI (3.20)	C.P.
0.06	0.062	(0.047, 0.076)	0.748	(0.021, 0.084)	0.942	(0.039, 0.090)	0.949
0.15	0.151	(0.120, 0.194)	0.903	(0.112, 0.199)	0.948	(0.116, 0.206)	0.945
0.30	0.306	(0.243, 0.395)	0.918	(0.236, 0.402)	0.938	(0.240, 0.413)	0.930
0.40	0.410	(0.326, 0.529)	0.927	(0.317, 0.538)	0.948	(0.323, 0.552)	0.943
0.50	0.503	(0.400, 0.649)	0.930	(0.390, 0.660)	0.957	(0.397, 0.678)	0.959
0.60	0.603	(0.479, 0.778)	0.931	(0.467, 0.792)	0.948	(0.476, 0.813)	0.952
0.80	0.807	(0.641, 1.041)	0.930	(0.625, 1.059)	0.954	(0.636, 1.089)	0.949
1.20	1.210	(0.961, 1.562)	0.935	(0.938, 1.591)	0.957	(0.955, 1.634)	0.957
2.80	2.840	(2.257, 3.670)	0.923	(2.201, 3.739)	0.945	(2.240, 3.840)	0.948
4.00	4.047	(3.211, 5.223)	0.932	(3.133, 5.322)	0.953	(3.192, 5.474)	0.960

(b) $\boldsymbol{\mu_j}$ is unknown. [b]

μ	$\bar{\hat{\mu}}$	CCI (3.17)	C.P.	CCI (3.19)	C.P.	CCI (3.20)	C.P.
0.06	0.062	(0.050, 0.078)	0.798	(0.025, 0.083)	0.924	(0.040, 0.090)	0.940
0.15	0.152	(0.121, 0.195)	0.915	(0.112, 0.198)	0.931	(0.115, 0.205)	0.937
0.30	0.306	(0.244, 0.395)	0.921	(0.236, 0.399)	0.928	(0.241, 0.410)	0.925
0.40	0.409	(0.327, 0.528)	0.921	(0.318, 0.534)	0.940	(0.323, 0.548)	0.936
0.50	0.502	(0.400, 0.648)	0.939	(0.391, 0.655)	0.952	(0.397, 0.672)	0.948
0.60	0.602	(0.479, 0.776)	0.932	(0.469, 0.785)	0.942	(0.477, 0.805)	0.947
0.80	0.805	(0.641, 1.038)	0.941	(0.628, 1.050)	0.954	(0.639, 1.077)	0.951
1.20	1.208	(0.960, 1.557)	0.939	(0.943, 1.576)	0.943	(0.959, 1.616)	0.947
2.80	2.833	(2.254, 3.658)	0.929	(2.216, 3.704)	0.939	(2.252, 3.796)	0.938
4.00	4.037	(3.206, 5.204)	0.935	(3.154, 5.270)	0.948	(3.209, 5.408)	0.950

[a] based on 995 replications.

[b] based on 997 replications.

TABLE X

SIMULATION SCENARIOS FOR ROBUSTNESS COMPARISON

Scenario	True Models	Analysis and Calibration by	Denoted by
A	Gamma mixed model	Gamma mixed model	Gamma/Gamma
B		Two-component mixed model	Gamma/Two-component
C	Two-component mixed model	Gamma mixed model	Two-component/Gamma
D		Two-component mixed model	Two-component/Two-component
E	Gamma mixed model + Two-component mixed model	Gamma mixed model	Mixture/Gamma
F		Two-component mixed model	Mixture/Two-component

For scenario C, I used the data generated according to the two-component mixed model with parameter values $\beta_0 = 0, \beta_1 = 1, \sigma_\epsilon^2 = 0.0015,\ \sigma_\eta^2 = 0.050, \sigma_0^2 = 0.002,\ \sigma_{01} = -0.0015,\ \sigma_1^2 = 0.025$, and with $N = 20$ and $K = 1$. This is the same data set used for examining scenario D in Section 4.1.2. I discarded simulated y_{ij} if it was negative. Then I fit a gamma mixed model to estimate parameters by the MML using SAS PROC NLMIXED. A new data set was generated according to the same two-component mixed model, and I computed $\hat{\mu}$ and constructed 95% calibration confidence regions (3.17), (3.19), and (3.20) for this new data set. This process was replicated 1,000 times.

The simulation results for comparing robustness of models are given in Table XI. The averages of point estimates are close to the true values of concentration even though models are misspecified. For the scenario B (gamma/two-component), the convergence rate was 100%. The coverage probability is too high at $\mu = 0.06$, and approximately 0.90 at most higher concentrations. For the scenario C (two-component/gamma), the convergence rate was 98.3%. The confidence interval (3.17) has unacceptably low coverage probabilities at small concentrations, $\mu = (0.06, 0.15)$, but at larger concentrations, the coverage probabilities are higher than 0.90. The global calibration confidence intervals (3.19)-(3.20) appear to perform better in terms of the coverage probability and have longer lengths than the intervals (3.17) at all concentration levels. Overall, the performance of gamma mixed models when the true model is the two-component mixed model is moderately better than the performance of two-component mixed models when the gamma mixed model is the true model.

4.3.2 Mixture of Two Distributions

In this section, I examine the robustness of two regression models when data are distributed as a mixture of gamma and lognormal distributions. More specifically, I generated data of which the true model was 50-50% mixture of the gamma mixed model and the two-component mixed model (scenarios E and F). In the data generating process, for a given laboratory, one of the models was chosen with 50% of chance to generate y_{ij} at each concentration level. Namely, the true model changes depending on concentration levels. I analyze the data using both models and compare their robustness.

The true parameter values and simulation results for parameter estimates are shown in Table XII. The convergence rates are excellent; 99.7% and 100 % in scenarios E and F respectively. The overall intercept and slope estimates are accurate on average in both scenarios; this may be because the same true parameter values were used in both models. Variance component estimates are less accurate. This was expected because the two regression models have different amounts of variability.

TABLE XI

SIMULATION RESULTS FOR ROBUSTNESS COMPARISON: MODEL MISSPECIFICATION

B: Gamma/Two-component [a]

μ	$\bar{\hat{\mu}}$	CCI (3.7) (3.9)	C.P.
0.06	0.057	(0.024, 0.090)	0.987
0.15	0.146	(0.101, 0.191)	0.966
0.30	0.292	(0.216, 0.369)	0.927
0.40	0.388	(0.298, 0.490)	0.902
0.50	0.488	(0.377, 0.615)	0.899
0.60	0.587	(0.454, 0.738)	0.914
0.80	0.790	(0.611, 0.993)	0.917
1.20	1.189	(0.920, 1.495)	0.905
2.80	2.733	(2.114, 3.441)	0.906
4.00	3.949	(3.045, 4.961)	0.884

C: Two-component/Gamma [b]

μ	$\bar{\hat{\mu}}$	CCI (3.17)	C.P.	CCI (3.19)	C.P.	CCI (3.20)	C.P.
0.06	0.070	(0.048,0.084)	0.494	(0.007, 0.108)	0.904	(0.020, 0.101)	0.837
0.15	0.150	(0.115,0.185)	0.728	(0.083, 0.203)	0.860	(0.098, 0.216)	0.914
0.30	0.301	(0.240,0.381)	0.884	(0.219, 0.397)	0.931	(0.227, 0.413)	0.942
0.40	0.403	(0.323,0.512)	0.907	(0.303, 0.529)	0.949	(0.310, 0.546)	0.959
0.50	0.503	(0.404,0.640)	0.917	(0.383, 0.658)	0.953	(0.391, 0.677)	0.962
0.60	0.607	(0.489,0.772)	0.938	(0.466, 0.792)	0.954	(0.475, 0.814)	0.957
0.80	0.811	(0.654,1.032)	0.926	(0.628, 1.056)	0.953	(0.638, 1.083)	0.955
1.20	1.223	(0.987,1.556)	0.927	(0.953, 1.590)	0.946	(0.967, 1.628)	0.942
2.80	2.850	(2.301,3.624)	0.933	(2.230, 3.697)	0.952	(2.265, 3.786)	0.944
4.00	4.108	(3.317,5.224)	0.934	(3.219, 5.329)	0.943	(3.267, 5.454)	0.941

[a] based on 1000 replications.

[b] based on 983 replications.

TABLE XII

SIMULATION RESULTS FOR ROBUSTNESS COMPARISON: WHEN DATA ARE A MIXTURE OF TWO MODELS, AVERAGE (RMSE) OF PARAMETER ESTIMATES.

Parameters	True Values	Scenario F[a]	Scenario E[b]
β_0	0.001	0.0037 (0.0118)	0.0064 (0.0118)
β_1	1.0	1.0088 (0.0411)	1.0031 (0.0341)
σ_0^2	0.002	0.0026 (0.0011)	0.0016 (0.0014)
σ_1^2	0.020	0.0312 (0.0155)	0.0138 (0.0168)
σ_{01}	-0.001	-0.0028 (0.0027)	0.0005 (0.0032)
σ_ϵ^2	0.0015	0.0011 (0.0010)	–
σ_η^2	0.050	0.0479 (0.0120)	–
κ	13.0	–	15.2071 (6.7023)

[a] Analyzed by the two-component mixed model; based on 1000 replications.

[b] Analyzed by the gamma mixed model; based on 997 replications.

I simulated another set of data according to the same 50-50% mixture of two models for the purpose of calibration. The simulation results for point estimate and calibration confidence intervals are presented in Table XIII. In general, the calibration intervals in both scenarios perform well, and point estimates are reasonably close to the true values on average. For scenario F, the coverage probabilities of the interval (3.7) are somewhat high; and similar results were observed in Table VII. For scenario E, the two global calibration intervals are wider and their coverage probabilities are better than the interval (3.17). The coverage probabilities of the interval (3.17) are unacceptably low when the true concentration is small ($\mu = 0.06, 0.15$). When the lower bound of interval (3.19) could not be obtained, it was set to zero. That's why the average of lower bounds of the interval for $\mu = 0.06$ is much lower than that of other two intervals.

TABLE XIII

SIMULATION RESULTS FOR ROBUSTNESS COMPARISON: WHEN DATA ARE MIXTURE OF TWO MODELS, CALIBRATION CONFIDENCE INTERVALS.

F: Mixture/Two-component [a]

μ	$\bar{\hat{\mu}}$	CCI (3.7) (3.9)	C.P.
0.06	0.057	(0.012, 0.108)	0.983
0.15	0.143	(0.086, 0.201)	0.967
0.30	0.293	(0.211, 0.375)	0.947
0.40	0.393	(0.296, 0.496)	0.917
0.50	0.492	(0.377, 0.620)	0.937
0.60	0.592	(0.456, 0.743)	0.936
0.80	0.788	(0.610, 0.988)	0.937
1.20	1.188	(0.922, 1.488)	0.926
2.80	2.776	(2.153, 3.479)	0.936
4.00	3.951	(3.066, 4.957)	0.921

E: Mixture/Gamma [b]

μ	$\bar{\hat{\mu}}$	CCI (3.17)	C.P.	CCI (3.19)	C.P.	CCI (3.20)	C.P.
0.06	0.069	(0.047, 0.082)	0.559	(0.006, 0.102)	0.920	(0.031, 0.108)	0.915
0.15	0.149	(0.114, 0.186)	0.778	(0.086, 0.200)	0.902	(0.081, 0.214)	0.937
0.30	0.301	(0.239, 0.385)	0.904	(0.219, 0.397)	0.945	(0.226, 0.412)	0.958
0.40	0.403	(0.321, 0.516)	0.899	(0.301, 0.529)	0.945	(0.309, 0.547)	0.946
0.50	0.505	(0.403, 0.647)	0.921	(0.382, 0.661)	0.958	(0.390, 0.681)	0.963
0.60	0.607	(0.484, 0.778)	0.922	(0.462, 0.793)	0.956	(0.472, 0.817)	0.951
0.80	0.807	(0.644, 1.034)	0.928	(0.619, 1.052)	0.960	(0.631, 1.083)	0.964
1.20	1.217	(0.972, 1.558)	0.922	(0.938, 1.584)	0.952	(0.956, 1.629)	0.958
2.80	2.838	(2.271, 3.638)	0.928	(2.200, 3.696)	0.957	(2.237, 3.792)	0.955
4.00	4.041	(3.234, 5.179)	0.911	(3.135, 5.260)	0.949	(3.187, 5.398)	0.941

[a] based on 1000 replications.

[b] based on 997 replications.

5. ILLUSTRATION

5.1 Amosite asbestos fibers from New York State Department of Health

To illustrate the mixed-effects models proposed in Chapter 3, I return to the asbestos data from NYSDOH. As briefly described in Chapter 1, amosite asbestos samples were taken from 14 sites, and they were sent to a total of 35 laboratories to measure the fiber counts. Multiple samples were collected from each site, and each laboratory analyzed a sample from a particular site only once. That is, laboratories did not provide repeated measurements ($K = 1$). Several laboratories analyzed only a few asbestos samples (3 to 5 observations within a laboratory). I excluded the observations from these laboratories so that between- and within-laboratory variances are estimated more reliably. As a result, a total of 27 laboratories and 21 to 27 samples remained at each site. The data retain the same characteristics as shown in Section 1.2. The number of total observations is 339.

In order to illustrate the calibration curve analysis, I treat observations from site 5099 as the new independent asbestos fiber counts, and all other sites as the background data. Table I shows that the variation in measurements increases with the average fiber counts, indicating heteroscedasticity. The CV is the largest at the lowest mean fiber count and approximately constant at higher mean counts. These are the characteristics suitable for fitting a two-component error model and a gamma regression model. Average fiber counts over sites can be considered as naturally occurring concentration levels. Since the true asbestos concentrations are not known, I substitute the average fiber counts for μ in regression models.

I first analyze these data using the two-component mixed model (3.3). Based on our experience with the two-component error model, existence of near-zero measurements in the data helps us estimate the constant variance at low level concentrations, σ_ϵ^2. This is because the two-component error model is designed to explain the constant variance at near-zero concentrations as well as the inflated variances at larger concentrations (Rocke and Lorenzato, 1995; Rocke et al., 2003).

The minimum observed count in the data is $62/mm^2$ from the site 2778 that has the lowest average count, $164/mm^2$, among all sites (after excluding the laboratories with only a few observations, the average count of the site 2778 changed from 159 to 164; this is presented in Appendix). Therefore, I transformed the data by dividing observations by $164^{3/2}$. After such transformation, the average fiber counts range from 0.078 to 3.985, and SDs range from 0.040 to 1.205. Using the transformed data, I estimated parameters by the IWMML; these estimates are $\beta_0 = 0.0030\ (SE = 0.0080), \beta_1 = 0.9990 (SE = 0.0314), \sigma_0^2 = 0.0017,\ \sigma_{01} = -0.0025, \sigma_1^2 = 0.0266, \sigma_\epsilon^2 = 0.000142, \sigma_\eta^2 = 0.0481$. For the new asbestos counts of the site 5099, the point estimate of the true fiber count and its 95% calibration confidence interval is 1630.87 (1446.09, 1779.44) after back-transformation. Therefore, with 95% confidence we can say that the true amosite count of the site 5099 is between $1446/mm^2$ and $1779/mm^2$.

Before fitting the gamma mixed model (3.12), the distributional assumption needs to be checked. For observations nested within each laboratory, Anderson-Darling and Kruskal-Wallis goodness-of-fit tests of gamma distributions were conducted; all p-values are > 0.1. MLEs of shape parameters are approximately close to 1.0 for all laboratories; therefore, the assumption of a common shape parameter appears to be reasonable. Again, I transformed the data by dividing observations by $164^{3/2}$. Then I estimated parameters by the MML; estimates are $\beta_0 = -0.00038\ (SE = 0.0061),\ \beta_1 = 1.0046\ (SE = 0.0288),\ \sigma_0^2 = 0.000137,\ \sigma_{01} = 0.000925, \sigma_1^2 = 0.00709, \kappa = 10.6499\ (SE = 0.9070)$. For new observations of the site 5099, the point estimate of the true fiber count is 1658.78 after back-transformation. The 95% calibration confidence intervals are (1494.73, 1904.15) by expression (3.17), (1469.23, 1894.99) by expression (3.19), and (1456.76, 1878.82) by expression (3.20).

Based on parameter estimates from mixed-effects models, we can estimate means and standard deviations of asbestos fiber counts (Figure 8). For a two-component mixed model, expressions (3.2) and (3.4) are used, and for a gamma mixed model, expressions (3.14) is used. Note that values are in transformed scale. Both model-based means and standard deviations are in good agreement with their corresponding observed means and standard deviations (if they tallied, circles and triangles would be on top of each other on the 45 degree line). In addition, we can compute the between- and within-laboratory variances derived in Section 3.1.1 and Section 3.2.1; these are displayed in Figure 9. Both between- and within-laboratory variances increase with the mean fiber counts, indicating that measurement errors of the amosite asbestos fibers increase with concentrations within and across laboratories.

Next, I compare the confidence intervals obtained by using the mixed-effects models to those obtained by the methods described in Chapter 2. Note that both lognormal and gamma distributions fit well to the site 5099 data. First, using the lognormal distribution-based methods, the 95% generalized confidence highest density interval is (1497.51, 1820.00) and the Bayesian HPD interval is (1496.95, 1821.01). Second, using the gamma distribution-based methods, the 95% confidence interval developed by Bhaumik et al. (2009) is (1476.98, 1846.85), and the Bayesian HPD interval is (1484.86, 1824.87). All confidence intervals are displayed in Figure 10. All intervals include the average amosite fiber count (solid circle). The confidence intervals obtained by the gamma mixed model are generally wider than other intervals.

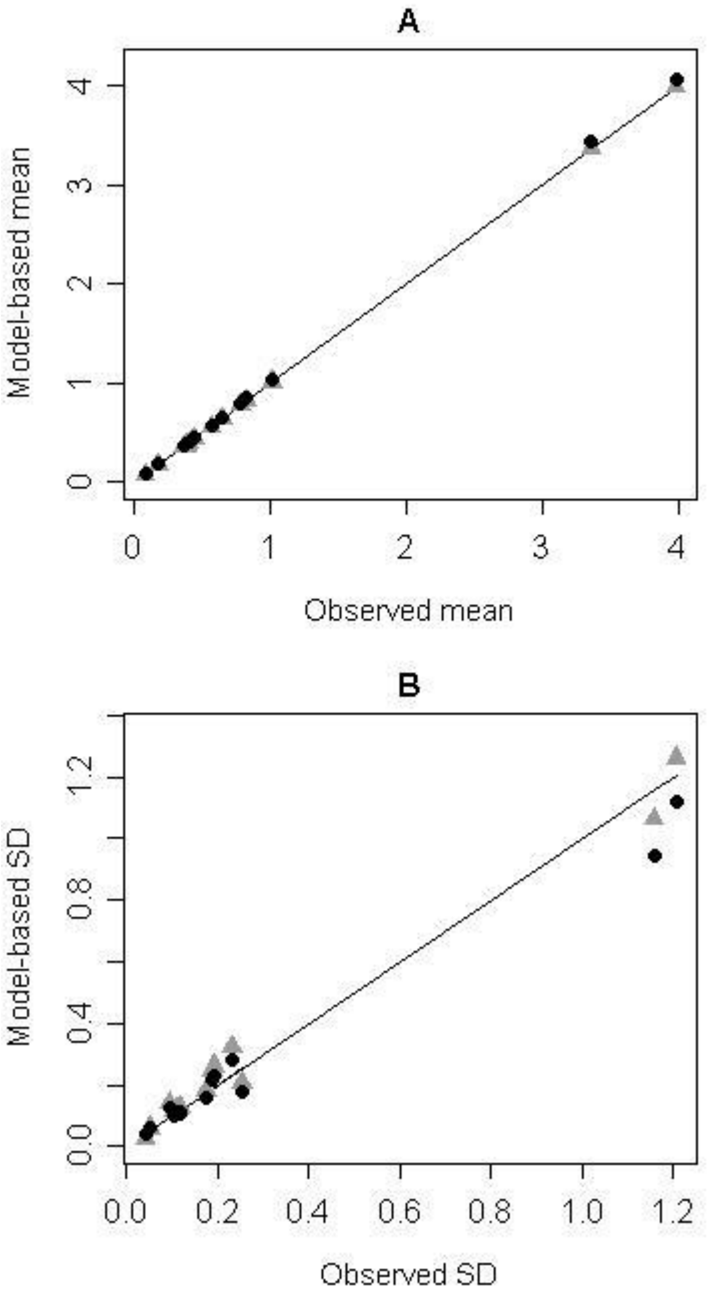

Figure 8. Model-based means and standard deviations from the two-component mixed model (●) and the gamma mixed model (▲) versus observed means and standard deviations across sites (in transformed scales).

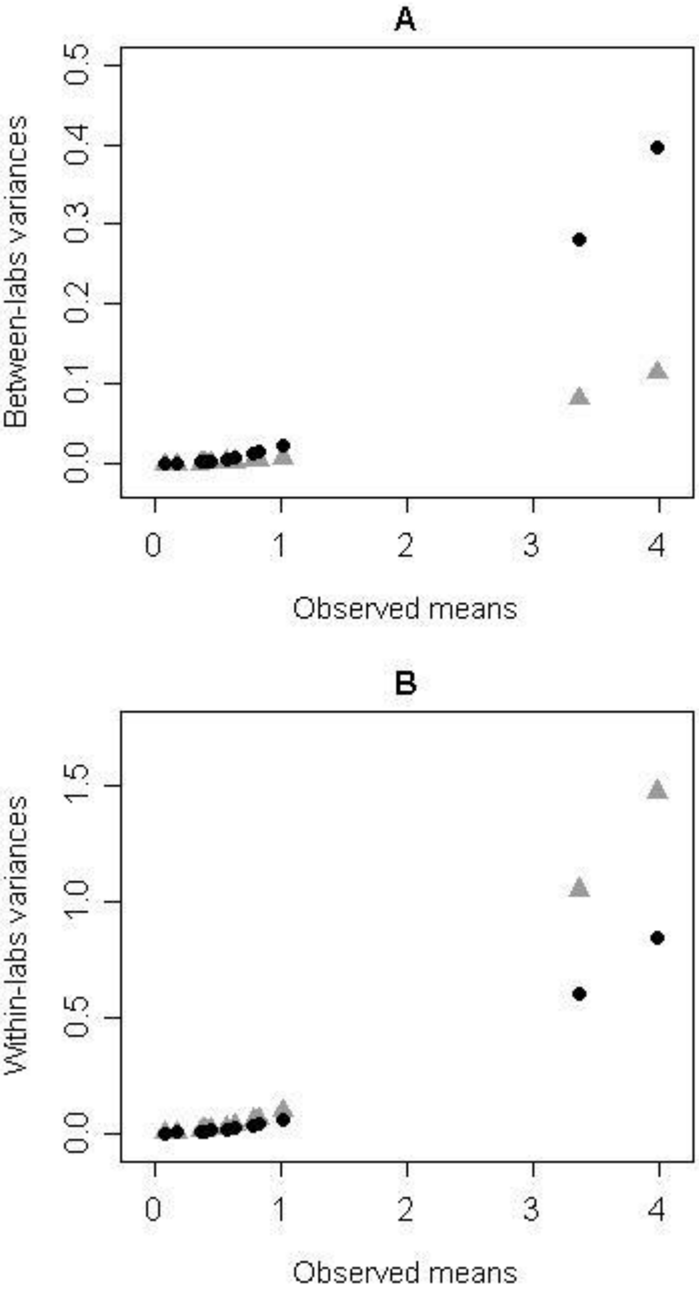

Figure 9. Between- and within-laboratory variances estimated from the two-component mixed model (●) and the gamma mixed model (▲) versus observed means across sites (in transformed scales).

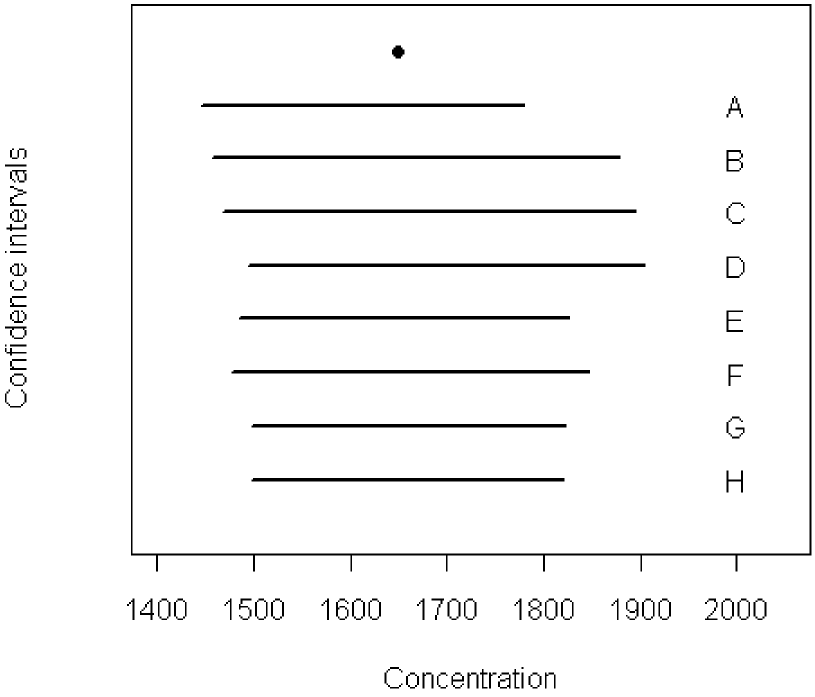

Figure 10. Confidence intervals by various methods for site 5099

● = Average amosite fiber count.

A = Two-component mixed model.

B = Interval (3.20) based on the gamma mixed model.

C = Interval (3.19) based on the gamma mixed model.

D = Interval (3.17) based on the gamma mixed model.

E = Bayesian HPD interval for the gamma mean.

F = Confidence interval (2.10) for the gamma mean.

G = Bayesian HPD interval for the lognormal mean.

H = Generalized confidence highest density interval for the lognormal mean.

5.2 Copper Data from Ford Motor Company

As another illustration, I analyze experimental data of copper concentrations. The Ford Motor Company conducted a blind inter-laboratory study of laboratories that hold Michigan State Drinking Water Certifications. This experimental data set was published in Gibbons and Bhaumik (2001) to illustrate the two-component mixed model. Copper ($\mu g/L$) samples were prepared by an independent source and weekly submitted for five weeks. In this data set, the true concentrations are known. There are five concentration levels, and five replicates per concentration within each of seven laboratories (i.e. $N = 7, n = 5,$ and $k = 5$). The data are presented in Appendix A. I exclude zero and negative measurements in the data because the gamma mixed model assumes positive dependent variables. These copper data exhibit non-constant variability as displayed in Figure 11.

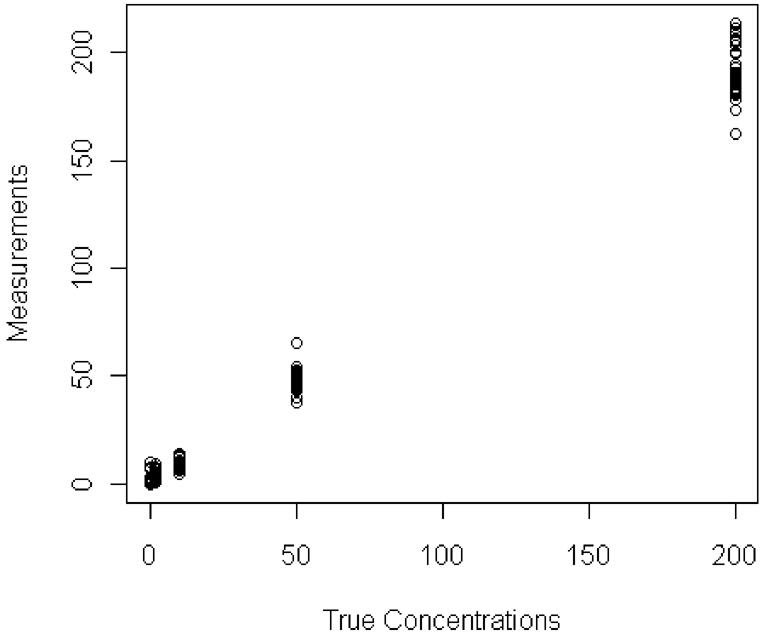

Figure 11. Copper concentration measurements versus true concentrations: variability in the measurements increases with the concentrations.

The distributional assumption about a constant shape parameter was checked: for each laboratory, the shape parameter estimates were all close to 0.20. By fitting the gamma mixed model to these data, parameter estimates are as follows: $\beta_0 = 1.0041\ (SE = 2.6965),\ \beta_1 = 0.8226\ (SE = 0.3513),\ \sigma_0^2 = 1.8323,\ \sigma_{01} = -0.0765,\ \sigma_1^2 = 0.0032,\ \kappa = 2.0100\ (SE = 0.2390)$. There is substantial between-laboratory variation, especially in the random intercepts; this was also shown in the results obtained by fitting the two-component mixed model (Gibbons and Bhaumik, 2001).

6. DISCUSSION

In this dissertation, I explored the use of the gamma distribution in environmental data analysis. I reviewed relevant methodologies for constructing the confidence interval for the mean concentration and the prediction interval for a single measurement, based on lognormal and gamma distributions. Each method can be chosen according to their advantages and disadvantages to make environmental impact decisions. For right skewed data, I observed that highest density intervals were shorter than equal-tailed intervals, and included more small values and fewer large values than equal-tailed intervals do. The generalized confidence *highest density* interval was proposed for the lognormal mean; it can be a good alternative to the generalized confidence *equal-tailed* interval if a shorter interval is desired.

For constructing the Bayesian interval for the mean concentration, I used non-informative conjugate priors on the unknown parameters. In environmental monitoring problems where the samples are routinely measured, a new sample is compared with past data to determine whether there has been any change. The past data can provide prior knowledge about the population mean and can be incorporated via setting prior density accordingly. For example, Miller's conjugate prior for a gamma distribution has hyperparameters (r, s, p, q): the past mean concentration can be expressed as q/r.

For analysis of environmental data obtained from multiple laboratories, I proposed the gamma mixed-effects model and studied the two-component mixed model. Both regression models were designed to explain heteroscedasticity and between-laboratory variability inherent in the environmental data. I explored the use of a two-component mixed model when replicated measurements per concentration within a laboratory were not available. As expected, the parameter estimates obtained by the IWMML were less accurate in such situations. However, the calibration confidence interval for a true unknown concentration still performed well. I developed the mixed-effects model for gamma distributed random variables with the identity link function. The parameter estimates obtained by the MML and EB estimation were excellent, even when replicated measurements were not available. I proposed two types of calibration confidence intervals to estimate unknown true concentrations of new data; the interval (3.17) based on laboratory-specific calibration curves, and the global calibration intervals (3.19) and (3.20) based on the average calibration curve. The interval (3.17) seems to be inappropriate for near-zero concentrations having very low coverage probability, but the global intervals performed well at all concentration levels. When only part (or none) of the laboratories analyzing new data participated in collection of background data, the global calibration intervals can be useful for estimating unknown true concentrations. The interval (3.20) would be simpler to apply in practice than the interval (3.19), but it is not recommended to use when the number of laboratories analyzing a particular concentration level of new data is just a few. On the other hand, the interval (3.19) sometimes does not provide the lower limit at near-zero concentrations.

I compared robustness of the gamma mixed model with the two-component mixed model via simulation studies. I considered situations where the alternative model was chosen (a certain kind of model misspecification) and the data were distributed as a mixture of two models. In the simulation studies, the performance of the intervals relevant to the gamma mixed model was satisfactory in both situations. For low concentrations (near-zero), the length of calibration confidence interval of the two-component mixed model tend to be somewhat longer than it should be. On the contrary, the length of interval (3.17) of the gamma mixed model tends to be shorter than it should be. Therefore, the confidence intervals obtained by the two-component mixed model appear to be more appropriate for examining near-zero concentration measurements. The global calibration confidence intervals of the gamma mixed model have better coverage probabilities than the interval (3.17) at near-zero concentrations, but they are still slightly lower than the desired level of confidence. Further study is necessary about the behavior of both of the global calibration intervals. As long as the level of concentration is not very low, the calibration confidence interval based on the gamma mixed model enables us to make more strict environmental decision than the two-component mixed model because of higher upper bound. This was shown in Illustration as well (Figure 10).

There are limitations of this study. First, I substituted the sample mean for the true concentration when it was assumed to be unknown. Additional uncertainty due to estimating the true concentration was not incorporated. The simulation study shows that the substitution of the sample mean does not seriously affect the estimation of calibration confidence intervals. Therefore, the gain by adjusting the additional uncertainty will not be substantial. Second, the calibration confidence intervals proposed in this dissertation are *approximate*, and uncertainty due to estimating parameters is not incorporated. However, I observed via simulation that their performance was still excellent. Lastly, the point estimates proposed in Chapter 3 assume that new data are collected from the same set (or a subset) of laboratories from which background data were collected. If the laboratories in new data include different laboratories, the current point estimate should be modified. An estimate combining laboratory-specific estimates and overall parameter estimates may be considered.

The proposed gamma mixed-effects model is different from previously developed models in the sense that the identity link function is used, while the log and inverse link functions have been used in previously developed gamma regression models. The identity link allows for easy and direct interpretation of regression coefficients. The disadvantage is that it restricts the possible range of regression coefficients due to the positive dependent variable.

Normality assumption on the random effects in the gamma mixed model may not be realistic in some situations because it can produce negative scale parameter, which is not allowed in a gamma distribution. Therefore, a distribution that supports non-negative values may be more appropriate to prevent this situation. A bivariate gamma distribution would be a good candidate for the random effects distribution, but mixed-effects models for continuous responses developed so far usually assume a multivariate normal distribution for random effects. There have been some studies regarding the use of gamma distribution for random effects in the analysis of correlated failure times (Hougaard, 2000). The popular prior in Bayesian approach, Dirichlet process, could be a good candidate as well because it provides very flexible distribution for random effects that is adaptively chosen by data. Further study on this area is necessary.

In addition, distributional assumptions about the random effects are difficult to assess from the data at hand, and EB estimates are sensitive to the normality assumption. However, it is known that the estimates of the fixed-effects are much less sensitive to misspecification of the random-effects distribution (Fitzmaurice et al., 2004). Hence, the "global" calibration confidence region of the gamma mixed model could be useful when the random-effects distribution is misspecified because it does not rely on the EB estimates.

Regulatory agencies can use the proposed gamma mixed model and the corresponding calibration confidence intervals to determine whether the area of interest is an environmental concern. U.S. EPA uses the UCL for the mean concentration of environmental samples to make remediation decisions, for instance at Superfund sites[2]. Usually the lognormal or gamma distribution is used to compute the UCL, and variation at the laboratory level (or at the instrument level when multiple instruments are used) is overlooked. The approach proposed in this dissertation incorporates the between-laboratory variation in obtaining the UCL. In order to obtain the calibration UCL, background data is needed. Therefore, my approach is relevant in examining the area where environmental samples are routinely collected. We can compute the $100(1-\alpha)\%$ UCL using $z_{1-\alpha}$ in the derivation of calibration confidence intervals. If the area of concern is not crucial with respect to public health or environmental perspective, the decision can be made based on the LCL. Even if the mean of samples exceeds a regulatory standard, if the LCL is lower than the standard, then we may conclude that the area is not environmental concern. The $100\alpha\%$ LCL can be computed using z_α in the derivation of calibration confidence intervals.

[2] A Superfund site is an uncontrolled or abandoned place where hazardous waste is located, possibly affecting local ecosystems or people. More details about Superfund can be found at http://www.epa.gov/superfund/sites/.

The application of the gamma mixed model proposed is not limited to calibration curve estimation. The gamma mixed model can be used when clustering arises in right skewed data. For example, the model can be useful for analyzing length of hospitalizations across different hospitals, or for evaluating the effect of a smoking cessation program on change of the amount of smoking over time, etc. Poisson mixed-effects models have been commonly used in these fields; but only integer values are available as a dependent variable in the Poisson regression model, and the variance of the response is forced to be equal to the mean response. The gamma mixed-effects model would be a good alternative approach to analyzing these types of data.

The development of TEM technique has allowed for detailed characterization of mineral particles (Institute of Medicine and National Research Council, 2009; Loomis et al., 2010). All asbestos fibers are classified into their sizes by TEM method. Dement et al. (2007, 2011) categorized the length and diameter of asbestos fibers collected at asbestos textile plants in South Carolina and North Carolina, and counted the number of fibers for each category. Thinner and longer fibers are more strongly associated with health hazard than shorter and thicker fibers (Stayner et al., 2008; Loomis et al., 2010). Therefore, it is suggested to adjust the size category in the analysis of asbestos fibers. The plausible mixed-effects model with a multiplicative gamma random error would be,

$$y_{ijk} = (\beta_0 + u_{0i} + \beta_{Lj}L_j + \beta_{Dk}D_k)\epsilon_{ijk}\ ,$$

where y_{ijk} is a fiber count in the j^{th} category of lengths and the k^{th} category of diameters within the i^{th} cluster. L_j is an indicator of the j^{th} category of lengths and D_k is an indicator of the k^{th} category of diameters, and β_{Lj} and β_{Dk} are the related fixed-effects. Data may be clustered in the sense that airborne asbestos fibers are collected at more than one textile plants or at various zones within a textile plant.

There are areas that I would like to study further. First, I would like to investigate why the calibration confidence intervals (3.19) and (3.20) have somewhat lower coverage probabilities than desired confidence level, and suitably adjust the intervals. Second, SAS NLMIXED procedure is a convenient tool to fit the gamma mixed model. However, I observed in the simulation studies that it was somewhat sensitive to initial values for parameters. One could benefit from a package that is more robust to the choice of initial values. SuperMix uses the hierarchical likelihood estimation (Lee and Nelder, 2001) to calculate initial values. This appears to be a very efficient way to find good initial values that help reduce the number of iterations (Kim, Y. et al., unpublished paper). Development of a more robust package for the proposed model will be my future work. Third, as discussed previously, the normality assumption on the random effects in the gamma mixed model may not be practical in some situations. A multivariate gamma distribution is an alternative assumption for random effects; therefore, my future work is to develop a gamma regression model with multivariate gamma-distributed random effects as well as the package to fit the model. Finally, I proposed the point and interval estimates for an unknown true concentration adjusting for heteroscedasticity and between-laboratory variation. Using an accurate exposure measurement is important when evaluating health hazard and excessive exposure to, for example, airborne asbestos fibers. Measurement errors may result in attenuated effects or biased results. I would like to extend the scope of my study by assessing the impact of using the proposed estimates for true concentrations in such epidemiologic studies.

CITED LITERATURE

Ahrens, J.H. and Dieter, U.: Generating gamma variates by a modified rejection technique. Communications of the ACM 25: 47-54, 1982.

Albert, J.: Bayesian Computation with R. Springer, 2007.

Augus, J.E.: Bootstrap one-sided confidence intervals for the lognormal mean. Statistician 43: 395-401, 1994.

Aryal, S., Bhaumik, D.K., Mathew, T., and Gibbons, R.D.: Approximate tolerance limits and prediction limits for the gamma distribution. Journal of applied statistical science 16: 103-111, 2008.

Bailey, D. and Alimdhi, F.: Gamma.mixed: Mixed effects gamma model. In: Zelig: Everyone's Statistical Software http://gking.harvard.edu/zelig, eds. Imai, K. King, G., and Lau, L. 2007.

Bhaumik, D.K. and Gibbons, R.D.: An upper prediction limit for the arithmetic mean of a lognormal random variable. Technometrics 46: 239-248, 2004.

Bhaumik, D.K. and Gibbons, R.D.: Confidence regions for random-effects calibration curves with heteroscedastic errors. Technometrics, 47: 223-230, 2005.

Bhaumik, D.K. and Gibbons, R.D.: One-sided approximate prediction intervals for at least p of m observations from a gamma population at each of r locations. Technometrics 48: 112-119, 2006.

Bhaumik, D.K., Kapur, and K., Gibbons, R.D.: Testing parameters of a gamma distribution for small samples. Technometrics 51: 326-334, 2009.

Camus, M., Siemiatycki, J., and Meek, B.: Nonoccupational exposure to Chrysotile asbestos and the risk of lung cancer. New England Journal of Medicine 338: 1565-1571, 1998.

Carroll, R.J. and Ruppert, D.: Transformation and Weighting in Regression. Champman & Hall/CRC, New York and London, 1988.

Casella, G. and Berger, R. Statistical Inference. 2nd edition, pp. 99-109, Duxbury Advanced Series. 2002.

Chen, M.H. and Shao, Q.M.: Monte Carlo estimation of Bayesian credible and HPD intervals. Journal of Computational and Graphical Statistics 8: 69-92, 1999.

Cheng, Y.S.: Bivariate lognormal distribution for characterizing asbestos fiber aerosols. Aerosol Science Technology 5: 359-368, 1986.

Cheng, Y.S., Holmes, T.D, and Fan, B.: Evaluation of respirator filters for asbestos filters. Journal of Occupational and Environmental Hygiene 3: 26-35, 2006.

Christensen, B.C., Godleski, J.J., Roelofs, C.R., Longacker, J.L., Bueno, R., Sugarbaker, D.J., Marsit, C.J., Nelson, H.H., and Kelsey, K.T.: Asbestos burden predicts survival in pleural mesothelioma. Environmental Health Perspective 116: 723-726, 2008.

Clarkson, D.B. and Zhan, Y.: Using Spherical-Radial quadruature to fit generalized linear mixed effects models. Journal of Computational and Graphical Statistics 11:639-659, 2002.

Dahiya, R.C. and Guttman, I.: Shortest confidence and prediction intervals for the log-normal. The Canadian Journal of Statistics 10: 277-291, 1982.

Diaz, R.E.: Comparison of PQL and Laplace 6 estimates of hierarchical linear models when comparing groups of small incident rates in cluster randomized trials. Computational Statistics and Data Analysis 51:2871-2888, 2007.

Finkelstein, M.M.: Asbestos fiber concentrations in the lungs of brake workers: another look. The Annals of Occupational Hygiene, 52: 455-461, 2008.

Fitzmaurice, G.M., Laird, N.M., and Ware, J.H.: Applied Longitudinal Analysis. Wiley. 2004.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B.: Bayesian Data Analysis. Champman & Hall/CRC, 2003.

Gilbert, R.O.: Comparing statistical tests for detecting soil contamination greater than background. Pacific Northwest Laboratory, Technical Report No. DE 94-005498, 1993.

Gibbons, R.D. and Bhaumik, D.K.: Weighted random-effects regression models with application to interlaboratory calibration. Technometrics 43: 192-198, 2001.

Grice, J.V. and Bain, L.J.: Inferences concerning the mean of the gamma distribution. Journal of the American Statistical Association 75:929-933, 1980.

Hawkins, D.M. and Wixley, R.A.J.: A note on the transformation of chi-squared variables to normality. The American Statistician 40: 296-298. 1986.

Hedeker, D. and Gibbons, R.D.: Longitudinal Data Analysis. Wiley, 2006.

Hein, M.J, Stayner, L.T., Lehman, E., and Dement, J.M.: Follow-up study of chrysotile textile workers: cohort mortality and exposure-response. Occupational and Environmental Medicine 64:616-625, 2007.

Hougaard, P. Analysis of Multivariate Survival Data. Springer, 2000.

Institute of Medicine: Asbestos: Selected Cancers. The National Academics Press, 2006.

Institute of Medicine and National Research Council: Counting Strategies, In: A review of the National Institute for Occupational Safety and Health (NIOSH) roadmap for research on asbestos fibers and other elongate mineral particles. Nelson, A.R., Liverman, C.T., Eide, E.A., and Abt, E. (eds.) The National Academies Press, 2009.

Krishnamoorthy, K. and Mathrew, T.: Inferences on the means of lognormal distributions using generalized p-values and generalized confidence intervals. Journal of statistical planning and inference 115: 103-121, 2003.

Krishnamoorthy, K., Mathrew, T., Ramachandran, G.: Generalized p-values and confidence intervals: a novel approach for analyzing lognormally distributed exposure data. Journal of Occupational and Environmental Hygiene 3: 642-650, 2006.

Krishnamoorthy, K., Mathew, T., and Mukherjee, S.: Normal-based methods for a gamma distribution: prediction and tolerance intervals and stress-strength reliability. Technometrics 50: 69-78, 2008.

Laird, N.M. and Ware, J.H.: Random effects models for longitudinal data. Biometrics 38:963-974, 1982.

Land, C.E.: Standard confidence limits for linear functions of the normal mean and variance. Journal of the American Statistical Association 68: 960-963, 1973.

Land C.E.: Tables of confidence limits for linear functions of the normal mean and variance. In: Selected Tables in Mathematical Statistics, vol. III, pp. 385-419. American Mathematical Society, Providence, R.I. 1975.

Lessaffre, E. and Spiessens, B.: On the effect of the number of quadrature points in a logistic random-effects model: an example. Applied Statistics 50: 325-335, 2001.

Lee, Y. and Nelder, J.A.: Hierarchical generalized linear models: a synthesis of generalized linear model, random-effect models and structured dispersions. Biometrika 88: 987-1006, 2001.

Loomis, D., Dement, J., Richardson, D., and Wolf, S.: Asbestos fiber dimensions and lung cancer mortality among workers exposed to Chrysotile. Occupational andEnvironmental Medicine 67: 580-584, 2010.

McCullagh, P. and Nelder, J.A.: Generalized Linear Models. Second edition, Champman & Hall/CRC, New York, 1989.

Miller, R.: Bayesian analysis of the two-parameter gamma distribution. Technometrics 22:65-69, 1980.

Oehlert, G.W., Lee, R.J., and Van Orden, D.: Statistical analysis of asbestos fibre counts. Environmetrics 6:115-126. 1995.

Ott, W.R.: Environmental Statistics and Data Analysis. Boca Raton, FL, CRC Press, 1995.

Pinheiro, J.C. and Bates, D.M.: Approximations to the log-likelihood function in the nonlinear mixed-effects model. Journal of Computational and Graphical Statistics 4:12-35, 1995.

Raudenbush, S.W., Yang, M-L, and Yosef, M.: Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. Journal of Computational and Graphical Statistics 9:141-157, 2000.

Rocke, D.M. and Lorenzato, S.: A two-component model for masurement error in analytical chemistry. Technometrics 37:176-183, 1995.

Rocke, D.M., Durbin, B., Wilson, M., and Kahn, H.D.: Modeling uncertainty in the measurement of low-level analytes in environmental analysis. Ecotoxicology and Environmental Safety 56:78-92, 2003.

Rohatgi, V.K. and Saleh A.K.Md.E.: An Introduction to Probability and Statistics. Second Edition. John Wiley and Sons, 2001.

Singh, A.K., Singh, A., and Engelhardt, M.: The lognormal distribution in environmental applications. U.S. EPA Technology Support Center Issue December 1997.

Singh, A., Singh, A.K., and Iaci, R.J.: Estimation of the exposure point concentration term using a gamma distribution. U.S. EPA Technology Support Center Issue October 2002.

Smith, B.J.: An R Package for MCM output convergence assessment and posterior inference. Journal of Statistical Software 21:1-35, 2007.

Stayner, L.T., Dankovic, D.A., and Lemen, R.A. Occupational exposure to Chrysotile asbestos and cancer risk: a review of the amphibole hypothesis. American Journal of Public Health 86:179-186, 1996.

Stayner, L., Kuempel, E., Gilbert, S., Hein, M., and Dement, J. An epidemiological study of the role of chrysotile asbestos fibre dimensions in determining respiratory disease risk in exposed workers. Occupational and Environmental Medicine 65:613-619, 2008.

Weerahandi, S.: Generalized confidence intervals. Journal of the American Statistical Association 88: 899-905, 1993.

Wilson, E.B. and Hilferty, M.M.: The distribution of chi-squares. Proceedings of the National Academy of Sciences 17:684-688, 1931.

World Health Organization International Agency for Research on Cancer: Asbestos, Chapter 5. Summary of data reported and evaluation. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, 14, Lyon, France. 1998.

APPENDICES

Appendix A

Copper Data

The Ford Motor Company conducted a blind inter-laboratory study of laboratories that hold Michigan State Drinking Water Certifications. This experimental data set was published in Gibbons and Bhaumik (2001) to illustrate analysis using the two-component mixed model. Copper ($\mu g/L$) samples were prepared by an independent source and weekly submitted for five weeks. There are five concentration levels, and five replicates per concentration within each of seven laboratories (i.e. balanced data with $N = 7, n = 5,$ and $k = 5$). The analysis results of this data set using the gamma mixed-effects model are presented in Chapter 5 of this dissertation.

		Concentration in $\mu g/L$				
Lab	Replicate	0	2	10	50	200
1	1	3	3	14	54	205
1	2	2	3	10	51	206
1	3	-1	5	11	52	208
1	4	1	2	12	54	211
1	5	-1	2	13	38	195
2	1	2.1	8	10	53	188.6
2	2	0.3	1.8	12.4	54.6	210
2	3	2	0.7	10.6	50	210
2	4	1.3	4	12	50.1	214
2	5	2	3	11	50	200
3	1	0.8	2.495	10.5	47.66	181.33
3	2	-0.185	2.695	10.335	45.39	173.205
3	3	0.99	2.41	9.735	44.27	180.56
3	4	0.905	1.84	10.245	46.91	183.65
3	5	0.365	2.84	10.325	47.24	181.585
4	1	1.661	3.243	12.25	48.14	205.4
4	2	1.996	3.432	13.51	54.45	200.4
4	3	0	9.246	11.16	51.01	199.7
4	4	2.993	3.39	13.44	52.86	189.6
4	5	2.042	4.109	10.47	48.72	187.7
5	1	0.09	0.86	10.03	50.06	193.4
5	2	-2.51	2.68	12.94	50.35	193.47
5	3	7.27	-0.4	8.97	49.32	203.16
5	4	7.14	4.73	9.61	49.93	190.02
5	5	0.28	5.2	9.12	48.08	191.05
6	1	7.226	4.964	4.713	48.242	191.02
6	2	-1	2	10	65	205
6	3	0	3	8	45	183
6	4	10.244	6.716	11.101	43	185
6	5	-2.177	8.844	8.249	47	182
7	1	0.018	1.323	6	45.5	162
7	2	-3	4.9	9.088	44	181
7	3	0	0	14.1	40	187
7	4	-2	0	6	43	178.3
7	5	-2	0	7	45.986	188.932

Appendix B

Derivation of Expressions in (3.18)

We derive mean and variance expressions of the approximate normal distribution of $Y_{ij}^{1/3}$. The scale parameter is normally distributed with mean $\left(\beta_0 + \beta_1\mu_j\right)/\kappa$ and variance $\left(\sigma_0^2 + 2\sigma_{01}\mu_j + \sigma_1^2\mu_j^2\right)/\kappa^2$ because of the bivariate normal assumption for $\boldsymbol{u_i}$. Assuming that replicated measurements per concentration level within a laboratory is available, the relevant replicates of θ_{ij} can be averaged and we denote the average by $\bar{\theta}_{ij}$. By using the *delta method*, $\bar{\theta}_{ij}^{1/3}$and $\bar{\theta}_{ij}^{2/3}$ are asymptotically normally distributed, given known parameters. When $K = 1$, we drop the bar and denote them by $\theta_{ij}^{1/3}$and $\theta_{ij}^{2/3}$. We can obtain the following expressions when $K = 1$. Given known parameter values,

$$E\left(\theta_{ij}^{1/3}\right) = \left(\frac{\beta_0 + \beta_1\mu_j}{\kappa}\right)^{1/3} \quad ,$$

$$E\left(\theta_{ij}^{2/3}\right) = \left(\frac{\beta_0 + \beta_1\mu_j}{\kappa}\right)^{2/3},$$

$$V\left(\theta_{ij}^{1/3}\right) = \frac{1}{9}\left(\frac{\sigma_0^2 + 2\sigma_{01}\mu_j + \sigma_1^2\mu_j^2}{\kappa^{2/3}\left(\beta_0 + \beta_1\mu_j\right)^{4/3}}\right),$$

$$E\left(\lambda_{ij}^2\right) = E\left\{\frac{\theta_{ij}^{1/3}\Gamma(\kappa + 1/3)}{\Gamma(\kappa)}\right\}^2 \approx \left(\frac{\beta_0 + \beta_1\mu_j}{\kappa}\right)^{2/3}\left\{\frac{\Gamma(\kappa + 1/3)}{\Gamma(\kappa)}\right\}^2.$$

As a result, the approximate mean and variance of the marginal distribution of $Y_{ij}^{1/3}$ are

$$\lambda_j^* = E\left[E(Y_{ij}^{1/3}|\boldsymbol{u_i})\right]$$

$$= E\left\{\frac{\theta_{ij}^{1/3}\Gamma(\kappa + 1/3)}{\Gamma(\kappa)}\right\}$$

$$\approx \frac{\Gamma(\kappa + 1/3)}{\Gamma(\kappa)}\left(\frac{\beta_0 + \beta_1\mu_j}{\kappa}\right)^{1/3}, \quad \text{and}$$

$$\tau_j^{*2} = E\left[V(Y_{ij}^{1/3}|\boldsymbol{u_i})\right] + V\left[E(Y_{ij}^{1/3}|\boldsymbol{u_i})\right]$$

$$= E\left\{\frac{\theta_{ij}^{2/3}\Gamma(\kappa+2/3)}{\Gamma(\kappa)} - \lambda_{ij}^2\right\} + V\left\{\frac{\theta_{ij}^{1/3}\Gamma(\kappa+1/3)}{\Gamma(\kappa)}\right\}$$

$$\approx \left(\frac{\beta_0+\beta_1\mu_j}{\kappa}\right)^{2/3}\left[\frac{\Gamma(\kappa+2/3)}{\Gamma(\kappa)} - \left(\frac{\Gamma(\kappa+1/3)}{\Gamma(\kappa)}\right)^2\right] + \left(\frac{\Gamma(\kappa+1/3)}{\Gamma(\kappa)}\right)^2 \frac{\sigma_0^2 + 2\sigma_{01}\mu_j + \sigma_1^2\mu_j^2}{9\kappa^{2/3}\left(\beta_0+\beta_1\mu_j\right)^{4/3}}.$$

The above expressions were derived under the assumption of replicates and applied to the situation where there is no replicate. We observed via simulation in Chapter 4 that these expressions still perform well when there is no replicate.

Appendix C

SAS PROC NLMIXED

Presented below is an example SAS code to fit the gamma mixed-effects model. The parameters are estimated by using adaptive Gauss-Hermite quadrature (`method=Gauss`) with 15 quadrature points (`qpoints=15`), and Newton-Raphson (`tech=newrap`). The maximum number of iterations is set to 500 (`maxiter=500`). Starting values are specified in `parms` statement. `a` and `b` indicate the shape and scale parameters of a gamma distribution respectively. `u0` and `u1` represent random deviations from the intercept `b0` and the slope `b1`. `lab` is an ID variable that identifies which laboratory (or cluster) an observation is nested in, i.e. the input data set must be clustered according to the `subject=` variable. Upon convergence, a SAS data set named `newEB` is created, and it contains empirical Bayes estimates of `u0` and `u1` by `lab`. The output data `parmest` contains MLEs of parameters and their standard errors. `Gconv` and `fconv` control the convergence criteria.

```
proc nlmixed data=asbestos method=Gauss qpoints=15 maxiter=500
       tech=newrap gconv=1E-8 fconv=1E-8;

/* starting values */
parms b0 0 b1 1 v00 0.001 v11 0.01 v01 0 a 10 ;

/* restriction on parameters */
bounds a>0 ,v11>=0, v00>=0;

/* model specification */
mu=(u0+b0)+(u1+b1)*x;
b= mu/a;
model y~gamma(a,b);

random u0 u1~ normal([0,0],[v00,v01,v11]) subject=lab
out=newEB;

ods output parameterestimates=parmest;
run;
```

Appendix D

Asbestos data from New York State Department of Health

TABLE XIV. THE AMOSITE ASBESTOS DATA: EXCLUDING LABORATORIES WITH ONLY A FEW OBSERVATIONS.

		Original			Transformed[d]	
Sites	N[a]	Mean	SD[b]	CV[c]	Mean	SD[b]
2778	24	164.25	83.86	0.51	0.078	0.040
6001	24	377.46	104.86	0.28	0.180	0.050
3739	24	758.71	219.60	0.29	0.361	0.105
187Q	24	790.38	241.16	0.31	0.376	0.115
4915	22	800.91	227.17	0.28	0.381	0.108
7420	27	841.63	221.47	0.26	0.401	0.105
5284	24	936.04	195.20	0.21	0.446	0.093
8306	24	1187.17	362.90	0.31	0.565	0.173
6482	25	1347.68	533.37	0.40	0.642	0.254
5099	25	1648.16	390.66	0.24	0.785	0.186
8214	26	1733.15	400.38	0.23	0.825	0.191
879Q	22	2134.73	487.50	0.23	1.016	0.232
5209	21	7065.00	2430.86	0.34	3.364	1.157
1987	27	8368.37	2531.52	0.30	3.985	1.205

[a] number of laboratories.

[b] standard deviation.

[c] coefficient of variation; same in the original and transformed scales.

[d] transformed by dividing by $164^{3/2}$.

VITA

NAME: Yoonsang Kim

EDUCATION: B.A., Applied Statistics and English Education, Chung-Ang University, Seoul, Korea, 2000

M.P.H., Public Health, Seoul National University, Seoul, Korea, 2002

M.S., Biostatistics, University of Iowa, Iowa, 2005

Ph.D., Biostatistics, University of Illinois at Chicago, Chicago, Illinois, 2011

TEACHING: Division of Epidemiology and Biostatistics, University of Illinois at Chicago: Teaching Assistant for the Introduction to Biostatistics, 2006

Inter-University Consortium for Political Social Research Summer Program, University of Michigan, Michigan: Teaching Assistant for the Longitudinal Data Analysis, 2007-2009

RESEARCH: Institute for Health Research and Policy, University of Illinois at Chicago: Research Assistant, 2006-2011

Department of Epidemiology, University of Iowa: Research Assistant and Research Scientist, 2003-2006

Department of Preventive Medicine, Ewha Women's University, Korea: Biostatistician, 2002-2003

AWARDS: Scholarship for Academic Excellence, Department of Applied Statistics, Chung-Ang University, Korea, 1997

PROFESSIONAL MEMBERSHIP: American Statistical Association
Eastern North American Region, International Biometric Society

PUBLICATIONS: Kim, H., Kim, Y., and Hong, Y.C.: The lag effect pattern in the relationship of particulate air pollution to daily mortality in Seoul, Korea. International Journal of Biometeorology 48:25-30. 2003.

Kim, H., Lee, J.T., Hong, Y.C., Yi, S.M., and Kim Y.: Evaluating the effect of daily PM_{10} variation on mortality. Inhalation Toxicology 16(suppl.1):1-4. 2004.

Smith, E.M., Wang, D., Kim, Y., Rubenstein, L.M., Lee, J.H., Haugen, T.H., and Turek, L.P.: P16INK4a expression, human papillomavirus, and survival in head and neck cancer. Oral Oncology 44:103-210, 2008.

Dennis, L.K., Kim, Y., and Lowe, J.B.: Consistency of reported tanning behaviors and sunburn history among sorority and fraternity students. Photodermatology, Photoimmunology and Photomedicine 24:191-198, 2008.

Greenberg, D., Levy, S.R., Rasher, S., Kim, Y., Carter, S.D., and Berbaum, M.L.: Testing adult basic education students for reading ability and progress: How many tests to administer? Adult Basic Education and Literacy Journal 4:96-103. 2010

Porter, K.R., McCarthy, B.J., Freels, S., Kim, Y., and Davis, F.G.: Prevalence estimates for primary brain tumors in the United States by age, gender, behavior, and histology. Neuro-Oncology 12:520-527. 2010.

Khan, M., Shah, S., Grudzien, A., Onyejekwe, N., Banskota, P., Karim, S., Jing J., Kim, Y., and Gerber, B.S.: Diabetes education multimedia in the waiting room. Diabetes Therapy In press.