Available online at www.sciencedirect.com
ScienceDirect
Journal of Statistical Planning and Inference 138 (2008) 552–567
journal of statistical planning and inference
www.elsevier.com/locate/jspi

Efficient mean estimation in log-normal linear models

Haipeng Shen, Zhengyuan Zhu*

Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Received 6 April 2006; received in revised form 28 September 2006; accepted 13 October 2006
Available online 14 March 2007

Abstract

Log-normal linear models are widely used in applications, and many times it is of interest to predict the response variable or to estimate the mean of the response variable at the original scale for a new set of covariate values. In this paper we consider the problem of efficient estimation of the conditional mean of the response variable at the original scale for log-normal linear models. Several existing estimators are reviewed first, including the maximum likelihood (ML) estimator, the restricted ML (REML) estimator, the uniformly minimum variance unbiased (UMVU) estimator, and a bias-corrected REML estimator. We then propose two estimators that minimize the asymptotic mean squared error and the asymptotic bias, respectively. A parametric bootstrap procedure is also described to obtain confidence intervals for the proposed estimators. Both the new estimators and the bootstrap procedure are very easy to implement. Comparisons of the estimators using simulation studies suggest that our estimators perform better than the existing ones, and the bootstrap procedure yields confidence intervals with good coverage properties. A real application of estimating the mean sediment discharge is used to illustrate the methodology.

Keywords: Maximum likelihood; Parametric bootstrap; Mean squared error; Uniformly minimum variance unbiased; Sediment discharge

1. Introduction

The prevalence of log-normality has been reported in a wide range of applications from mining (Marcotte and Groleau, 1997), insurance reserves estimation (Doray, 1996), water quality control (Gilliom and Helsel, 1986), to air pollution concentration monitoring (Holland et al., 2000) and sediment discharge estimation (Cohn, 1995; Elliott and Anders, 2004), to name just a few. Log-normal linear models are often used in these applications, in which linear models are fitted to logarithmic transformed response variables. To fix ideas, let $Z = (Z_1, \ldots, Z_n)^{\mathrm{T}}$ be the log-normal response vector, and $x_i = (1, x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ be the covariate vector for observation i. A log-normal linear model assumes that

$$Y = \log(Z) = X\beta + \varepsilon, \tag{1}$$

where $X = (x_1, \ldots, x_n)^{\mathrm{T}}$, $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathrm{T}}$, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$ with $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$.

In many cases, for a new set of covariate values x_0, one is interested in predicting the response variable at the original scale,

$$Z_0 = \exp(x_0^{\mathrm{T}}\beta + \varepsilon_0),$$

* Corresponding author. Tel./fax: +1 919 843 2431.
E-mail address: zhuz@email.unc.edu (Z. Zhu).

0378-3758/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2006.10.016

or estimating the conditional mean of the response variable,

$$\mu(x_0) = \mathrm{E}(Z_0|x_0) = \exp(x_0^{\mathrm{T}}\beta + \tfrac{1}{2}\sigma^2), \tag{2}$$

where ε_0 is the corresponding normal error with mean zero and variance σ^2. As a motivating example, a sediment discharge study is reported in Section 5, where log-normal linear models are used to develop sediment-transport curves for estimating mean sediment discharge levels. If β and σ^2 are both known, it is easy to show that the best predictor of Z_0 under the squared error loss is in fact $\mu(x_0)$. However, in practice, both β and σ^2 are usually unknown, and need to be estimated in order to obtain an estimate of $\mu(x_0)$.

In this paper we consider the problem of efficient estimation of $\mu(x_0)$. A number of authors have considered this problem in the literature. Finney (1941) developed the uniformly minimum variance unbiased (UMVU) estimators of parameters of the log-normal distribution, which are extended by Heien (1968) to simple log-normal linear models. Bradu and Mundlak (1970) derived the UMVU estimator and its variance for general log-normal linear models. Maximum likelihood (ML) and restricted maximum likelihood (REML) methods have also been used in practice. A general discussion can be found in, for example, Lawless (1982, Chapter 6). More recently, El-shaarawi and Viveros (1997) proposed a bias-corrected REML estimator, which we term the EV estimator. We review all these estimators in Section 2 and compare their performance using a simulation study in Section 4.1.

A common measure of the quality of an estimator in the statistical literature is the *mean squared error* (MSE). Suppose $\hat{\mu}(x_0)$ is an estimator for $\mu(x_0)$, then its MSE is defined as

$$\mathrm{MSE}[\widehat{\mu}(x_0)] = \mathrm{E}[\widehat{\mu}(x_0) - \mu(x_0)]^2 = \mathrm{Var}[\widehat{\mu}(x_0)] + \mathrm{Bias}^2[\widehat{\mu}(x_0)], \tag{3}$$

where $\mathrm{Bias}[\widehat{\mu}(x_0)] = \mathrm{E}[\widehat{\mu}(x_0)] - \mu(x_0)$. In terms of MSE, the UMVU estimator is the best estimator among all unbiased estimators of $\mu(x_0)$. However, the UMVU estimator can only be expressed as the sum of an infinite series of Hypergeometric functions (Seaborn, 1991), which is not very convenient to use for practitioners. Furthermore, one can find better estimators in terms of MSE if one can tolerate a small bias. The EV estimator, for example, is biased; however its MSE is smaller than the UMVU estimator in many cases. See Section 4 for more details.

In this paper, we investigate a class of estimators obtained by plugging in the ML estimator for β and a degree of freedom (d.f.) adjusted "ML" estimator for σ^2 in (2). Both the ML and REML estimators of $\mu(x_0)$ belong to this class. We propose two new estimators of $\mu(x_0)$ from this class, the minimum MSE (MM) estimator and the minimum bias (MB) estimator, which minimize the asymptotic MSE and bias, respectively. In practice, one may use either estimator depending on the desirable tradeoff between bias and variance for a particular application. We also describe a parametric bootstrap method to derive confidence intervals for our estimators, which are shown to have nice coverage properties.

A direct comparison with all the estimators mentioned above indicates that our estimators have superior performance in terms of MSE or bias, and for small sample sizes, the improvement is rather substantial. Our results also show that the d.f. used for estimating σ^2 should be between the d.f.s used for the MM and MB estimators if small bias and MSE is desirable, no matter what kind of tradeoff one wants to achieve between bias and variance. Hence, the REML estimator without bias correction should not be used for estimating $\mu(x_0)$ under any circumstances.

The rest of the paper is organized as follows. Section 2 reviews the existing estimators for $\mu(x_0)$ using log-normal linear models, including their bias and MSE expressions. In Section 3 we derive the MM and MB estimators with their biases and MSEs. A parametric bootstrap procedure is also described to derive the corresponding confidence intervals. Various estimators are compared in Section 4 in terms of MSE and bias. A wide range of sample sizes and σ^2 values are considered that cover most practical situations. Empirical coverage properties of the bootstrap confidence intervals are also investigated. In Section 5 we analyze a sediment discharge data set collected from the Upper Colorado River Basin and present a practical comparison of the different estimators. All the technical details are relegated to the Appendix.

2. Existing estimators

In this section we review several existing estimators for $\mu(x_0)$ under the model assumption (1), and derive their biases and MSEs. To better facilitate the derivation, the following two propositions first give some well-known results about the distributions for the ordinary least squares (OLS) estimator for β and the corresponding residual sum of squares (RSS).

Proposition 1. *The OLS estimator for* β *is*

$$\widehat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y \sim \mathrm{N}(\beta, \sigma^2(X^{\mathrm{T}}X)^{-1}).$$

As a result, $x_0^{\mathrm{T}}\widehat{\beta} \sim \mathrm{N}(x_0^{\mathrm{T}}\beta, \sigma^2 v_0)$ *where* $v_0 = x_0^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_0$. *In addition, if* $X^{\mathrm{T}}X = \mathrm{O}(n)$, *then* $v_0 = \mathrm{O}(1/n)$.

Proposition 2. *Let* $m = n - (p+1)$. *The residual sum of squares is*

$$\mathrm{RSS} = Y^{\mathrm{T}}[I - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}]Y \sim \sigma^2\chi^2_m.$$

Thus, the moment generating functions (MGF) of RSS is

$$\mathrm{E}[\exp(c\mathrm{RSS})] = (1 - 2c\sigma^2)^{-m/2} \quad \textit{for } c < \frac{1}{2\sigma^2}.$$

Note that the OLS, ML, and REML estimators for β are identical. Furthermore, the OLS and REML estimators for σ^2 are the same, which is $\widehat{\sigma}^2_{\mathrm{REML}} = \mathrm{RSS}/m$. The ML estimator for σ^2 is $\widehat{\sigma}^2_{\mathrm{ML}} = \mathrm{RSS}/n$.

2.1. The naive back-transform estimator

Since the UMVUE of $\log(Z_0)$ is $x_0^{\mathrm{T}}\hat{\beta}$, it may seem reasonable to estimate $\mu(x_0)$ by $\widehat{\mu}_{\mathrm{BT}}(x_0) = \exp(x_0^{\mathrm{T}}\hat{\beta})$, the back-transform (BT) estimator. This is also a commonly used estimator in practice. However, by comparing it to (2), it is easy to show that the BT estimator is not even consistent, with an asymptotic multiplicative bias of $\exp(-\sigma^2/2)$, which is always less than one. As a result, the BT estimator underestimates $\mu(x_0)$, and the bias is large when σ^2 is large. In our study, it appears that the BT estimator performs much worse than the other estimators. Thus, we do not include it in our comparison below in Section 4. Actually, the BT estimator is more suitable for estimating the median of Z_0, which is $\exp(x_0^{\mathrm{T}}\beta)$ in this case.

2.2. The ML/REML estimators

The ML and REML estimators of $\mu(x_0)$ are given by

$$\widehat{\mu}_{\mathrm{ML}}(x_0) = \exp(x_0^{\mathrm{T}}\hat{\beta} + \widehat{\sigma}^2_{\mathrm{ML}}/2) \quad \text{and} \quad \widehat{\mu}_{\mathrm{REML}}(x_0) = \exp(x_0^{\mathrm{T}}\hat{\beta} + \widehat{\sigma}^2_{\mathrm{REML}}/2),$$

respectively. The MSE and bias of $\widehat{\mu}_{\mathrm{ML}}(x_0)$ can be derived using the results in Propositions 1 and 2 as follows:

$$\mathrm{MSE}[\widehat{\mu}_{\mathrm{ML}}(x_0)] = \mu^2(x_0)[\mathrm{e}^{(2v_0-1)\sigma^2}(1 - 2\sigma^2/n)^{-m/2} - 2\mathrm{e}^{(v_0-1)\sigma^2/2}(1 - \sigma^2/n)^{-m/2} + 1] \tag{4}$$

and

$$\mathrm{Bias}[\widehat{\mu}_{\mathrm{ML}}(x_0)] = \mu(x_0)[\mathrm{e}^{1/2(v_0-1)\sigma^2}(1 - \sigma^2/n)^{-m/2} - 1]. \tag{5}$$

One can obtain the MSE and bias of $\widehat{\mu}_{\mathrm{REML}}(x_0)$ by replacing n in (4) and (5) with m.

2.3. The UMVU estimator

Bradu and Mundlak (1970) derived the UMVU estimator of $\mu(x_0)$ using the fact that $\hat{\beta}$ and $\hat{\sigma}^2_{\mathrm{REML}}$ are complete sufficient statistics for β and σ^2, and any unbiased function of the complete sufficient statistics is the UMVU estimator of the mean of that function (Lehmann and Casella, 1998). To obtain the UMVU estimator of $\mu(x_0)$, one only needs to find $f(x)$ such that

$$\mathrm{E}[\mathrm{e}^{x_0^{\mathrm{T}}\hat{\beta}} f(\hat{\sigma}^2_{\mathrm{REML}})] = \exp(x_0^{\mathrm{T}}\beta + \tfrac{1}{2}\sigma^2),$$

which leads to

$$f(\hat{\sigma}^2_{\mathrm{REML}}) = \sum_{i=0}^{\infty} \frac{\Gamma(m/2)}{i!\Gamma(m/2+i)} \left[\frac{m(1-v_0)}{4}\hat{\sigma}^2_{\mathrm{REML}}\right]^i = {}_0F_1\left(\frac{m}{2}; \frac{m(1-v_0)}{4}\hat{\sigma}^2_{\mathrm{REML}}\right),$$

where ${}_0F_1(\alpha; z)$ is the Hypergeometric function (Seaborn, 1991). The UMVU estimator and its variance are then given by

$$\widehat{\mu}_{\mathrm{UMVU}}(x_0) = \mathrm{e}^{x_0^{\mathrm{T}}\hat{\beta}}\, {}_0F_1\left(\frac{m}{2}; \frac{m(1-v_0)}{4}\hat{\sigma}^2_{\mathrm{REML}}\right) \tag{6}$$

and

$$\mathrm{Var}[\widehat{\mu}_{\mathrm{UMVU}}(x_0)] = \mu^2(x_0)\left[\mathrm{e}^{v_0\sigma^2}\, {}_0F_1\left(\frac{m}{2}; \frac{(1-v_0)}{4}\sigma^4\right) - 1\right]. \tag{7}$$

More details can be found in Finney (1941) and Bradu and Mundlak (1970). Note that the UMVU estimator is an unbiased estimator; hence its variance is the same as its MSE. The UMVU estimator has the smallest MSE among all unbiased estimators.

2.4. The bias-corrected REML estimator

It is well known that the REML estimator exhibits some bias, and El-shaarawi and Viveros (1997) proposed to correct this bias using the leading terms of the Taylor expansion of $\mathrm{Bias}[\widehat{\mu}_{\mathrm{REML}}(x_0)]$ with respect to σ^2/m, and obtained the following bias-corrected REML estimator which we refer to as the EV estimator:

$$\widehat{\mu}_{\mathrm{EV}}(x_0) = \exp\left[x_0^{\mathrm{T}}\hat{\beta} + \frac{(1-v_0)}{2}\hat{\sigma}^2_{\mathrm{REML}} - \frac{1}{4m}\hat{\sigma}^4_R - \frac{1}{6m}\hat{\sigma}^6_R\right]. \tag{8}$$

Its MSE and bias are given by

$$\mathrm{MSE}[\widehat{\mu}_{\mathrm{EV}}(x_0)] = \mu^2(x_0)\{\mathrm{e}^{(2v_0-1)\sigma^2}\mathrm{E}[f^2_{\mathrm{EV}}(\hat{\sigma}^2_{\mathrm{REML}})] - 2\mathrm{e}^{(v_0-1)\sigma^2/2}\mathrm{E}[f_{\mathrm{EV}}(\hat{\sigma}^2_{\mathrm{REML}})] + 1\}$$

and

$$\mathrm{Bias}[\widehat{\mu}_{\mathrm{EV}}(x_0)] = \mu(x_0)\{\mathrm{e}^{(v_0-1)/2\sigma^2}\mathrm{E}[f_{\mathrm{EV}}(\hat{\sigma}^2_{\mathrm{REML}})] - 1\},$$

where

$$f_{\mathrm{EV}}(x) = \exp\left[\frac{(1-v_0)}{2}x - \frac{1}{4m}x^2 - \frac{1}{6m}x^3\right].$$

The two expectations have to be evaluated using numerical integration.

3. Two new estimators and their confidence intervals

In this section, we propose two new estimators from the following class of estimators:

$$\left\{\widehat{\mu}_c(x_0) : \widehat{\mu}_c(x_0) = \exp(x_0^{\mathrm{T}}\widehat{\beta} + c\mathrm{RSS}/2),\ c = \frac{1}{n-a},\ a < n\right\}. \tag{9}$$

This estimator class is motivated by the special relationship (2). The first estimator minimizes the MSE (3) approximately and is defined as

$$\widehat{\mu}_{\mathrm{MM}}(x_0) = \exp\left[x_0^{\mathrm{T}}\widehat{\beta} + \frac{m\mathrm{RSS}}{2(n-p+1+3nv_0)m + 3\mathrm{RSS}}\right],$$

where $m = n - (p + 1)$. The second estimator, on the other hand, minimizes the bias considerably and is defined as

$$\widehat{\mu}_{\mathrm{MB}}(x_0) = \exp\left[x_0^{\mathrm{T}}\widehat{\beta} + \frac{m\mathrm{RSS}}{2(n - p - 1 + nv_0)m + \mathrm{RSS}}\right].$$

The proposed estimators can be viewed as degree-of-freedom-adjusted ML estimators. In practice, it is very easy to obtain these estimators, because $\widehat{\beta}$ and RSS can be readily calculated. Below we describe how the estimators are derived in Section 3.1 and discuss their connection with the ML/REML estimators in Section 3.2. We also propose a parametric bootstrap procedure in Section 3.3 to generate confidence intervals.

3.1. Derivation of the estimators

As one can see, the estimators in the class (9) can be described as plug-in estimators of $\mu(x_0)$ based on the basic formula (2) with $\widehat{\beta}$ and $c\mathrm{RSS} = \mathrm{RSS}/(n - a)$ serving as the estimators of β and σ^2, respectively. These estimators are asymptotically equivalent and efficient, because the ML estimator belongs to the class with $a = 0$.

Our goal is to find estimators from this class that can asymptotically minimize the MSE or the bias and have better or comparable finite-sample performances as the existing estimators reviewed in Section 2.

Lemma 1. *Under the condition that $c < 1/2\sigma^2$, the MSE of $\widehat{\mu}_c(x_0)$ is*

$$\begin{aligned}\mathrm{MSE}[\widehat{\mu}_c(x_0)] &= \mathrm{E}[\widehat{\mu}_c(x_0) - \mu(x_0)]^2 \\ &= \mu^2(x_0)[\mathrm{e}^{(2v_0-1)\sigma^2}(1 - 2c\sigma^2)^{-m/2} - 2\mathrm{e}^{1/2(v_0-1)\sigma^2}(1 - c\sigma^2)^{-m/2} + 1].\end{aligned}$$

In addition, the bias of $\widehat{\mu}_c(x_0)$ is

$$\mathrm{Bias}[\widehat{\mu}_c(x_0)] = \mathrm{E}[\widehat{\mu}_c(x_0) - \mu(x_0)] = \mu(x_0)[\mathrm{e}^{1/2(v_0-1)\sigma^2}(1 - c\sigma^2)^{-m/2} - 1].$$

Lemma 1 can be proved by making use of the MGF of RSS as stated in Proposition 2. The lemma suggests that a direct minimization of $\mathrm{MSE}[\widehat{\mu}_c(x_0)]$ seems implausible due to the rather complicated expression. As an alternative, we propose to look at the second order asymptotics to find a constant c that can asymptotically minimize the MSE. A similar approach is employed in Shen et al. (2006) to derive an efficient estimator for one-population log-normal means.

Note the following standard expansion:

$$c = \frac{1}{n - a} = \frac{1}{n} + \frac{a}{n^2} + \mathrm{o}\left(\frac{1}{n^2}\right),$$

which leads us to consider estimators of the form $\widehat{\mu}_c(x_0)$ with c having the above expression.

Theorem 1. *Suppose $c = 1/(n - a) = 1/n + a/n^2 + \mathrm{o}(1/n^2)$. Then,*

$$\begin{aligned}&\mathrm{MSE}[\widehat{\mu}_c(x_0)] \\ &\quad= \mu^2(x_0)\frac{\sigma^2}{n}\left\{1 + \frac{\sigma^2}{2} + \frac{\sigma^2}{4n}[a^2 + (2 - 2p + 6nv_0 + 3\sigma^2)a + f(p, n, \sigma^2, v_0)]\right\} \\ &\qquad+ \mathrm{o}\left(\frac{1}{n^2}\right),\end{aligned}$$

where $f(p, n, \sigma^2, v_0) = -1 + p^2 - 6nv_0(p + 1) + 7n^2v_0^2 + (1 - 3p + 7nv_0)\sigma^2 + 7\sigma^4/4$;

$$\mathrm{Bias}[\widehat{\mu}_c(x_0)] = \mu(x_0)\frac{\sigma^2}{2n}\left(nv_0 + a - p - 1 + \frac{\sigma^2}{2}\right) + \mathrm{o}\left(\frac{1}{n}\right).$$

The proof of Theorem 1 is provided in the Appendix along with some discussion.

Suppose one wants to find a constant c that can minimize the MSE up to the order of $1/n^2$. Theorem 1 suggests to find a to minimize

$$a^2 + (2 - 2p + 6nv_0 + 3\sigma^2)a.$$

According to the quadratic form, the minimizer depends on σ^2 and is

$$-(1 - p + 3nv_0 + 3\sigma^2/2).$$

Thus, the constant c which minimizes the approximate MSE should be of the order of $1/(n + 1 - p + 3nv_0 + 3\sigma^2/2)$. However, in real applications, the true variance σ^2 is usually unknown. We propose to use an "adaptive" estimator by replacing σ^2 with its consistent estimator, $\widehat{\sigma}^2_{\mathrm{REML}} = \mathrm{RSS}/m$. As a result, our proposed estimator is

$$\widehat{\mu}_{\mathrm{MM}}(x_0) = \exp\left[x_0^{\mathrm{T}}\widehat{\beta} + \frac{m\mathrm{RSS}}{2(n - p + 1 + 3nv_0)m + 3\mathrm{RSS}}\right].$$

On the other hand, to reduce the bias to the order of $1/n$, Theorem 1 suggests that it suffices to find a to satisfy

$$nv_0 + a - p - 1 + \frac{\sigma^2}{2} = 0,$$

which leads to

$$a = p + 1 - nv_0 - \sigma^2/2.$$

Hence the constant c that minimizes the approximate bias should be of the order of $1/(n - p - 1 + nv_0 + \sigma^2/2)$. Similarly, we propose to use an "adaptive" estimator by replacing σ^2 with its consistent estimator $\hat{\sigma}^2_{\mathrm{REML}}$. As a result, our proposed estimator is

$$\widehat{\mu}_{\mathrm{MB}}(x_0) = \exp\left[x_0^{\mathrm{T}}\widehat{\beta} + \frac{m\mathrm{RSS}}{2(n - p - 1 + nv_0)m + \mathrm{RSS}}\right].$$

The *exact* MSE and bias of the two proposed estimators are summarized in the following corollary. Numerical methods are needed in order to evaluate them. In Section 4 we will compare the MSE and bias of our estimators $\widehat{\mu}_{\mathrm{MM}}(x_0)$ and $\widehat{\mu}_{\mathrm{MB}}(x_0)$ with the existing estimators described in Section 2.

Corollary 1. *Let*

$$f_{\mathrm{MM}}(\mathrm{RSS}) = \exp\left[\frac{m\mathrm{RSS}}{2(n - p + 1 + 3nv_0)m + 3\mathrm{RSS}}\right]$$

and

$$f_{\mathrm{MB}}(\mathrm{RSS}) = \exp\left[\frac{m\mathrm{RSS}}{2(n - p + 1 + nv_0)m + \mathrm{RSS}}\right].$$

Then,

$$\begin{aligned}&\mathrm{MSE}[\widehat{\mu}_{\mathrm{MM}}(x_0)]\\ &\quad = \mu^2(x_0)[\mathrm{e}^{(2v_0-1)\sigma^2}\mathrm{E}(f^2_{\mathrm{MM}}(\mathrm{RSS})) - 2\mathrm{e}^{1/2(v_0-1)\sigma^2}\mathrm{E}(f_{\mathrm{MM}}(\mathrm{RSS})) + 1],\end{aligned}$$

$$\mathrm{Bias}[\widehat{\mu}_{\mathrm{MM}}(x_0)] = \mu(x_0)[\mathrm{e}^{1/2(v_0-1)\sigma^2}\mathrm{E}(f_{\mathrm{MM}}(\mathrm{RSS})) - 1];$$

$$\begin{aligned}&\mathrm{MSE}[\widehat{\mu}_{\mathrm{MB}}(x_0)]\\ &\quad = \mu^2(x_0)[\mathrm{e}^{(2v_0-1)\sigma^2}\mathrm{E}(f^2_{\mathrm{MB}}(\mathrm{RSS})) - 2\mathrm{e}^{1/2(v_0-1)\sigma^2}\mathrm{E}(f_{\mathrm{MB}}(\mathrm{RSS})) + 1],\end{aligned}$$

$$\mathrm{Bias}[\hat{\mu}_{\mathrm{MB}}(x_0)] = \mu(x_0)[\mathrm{e}^{1/2(v_0-1)\sigma^2}\mathrm{E}(f_{\mathrm{MB}}(\mathrm{RSS})) - 1].$$

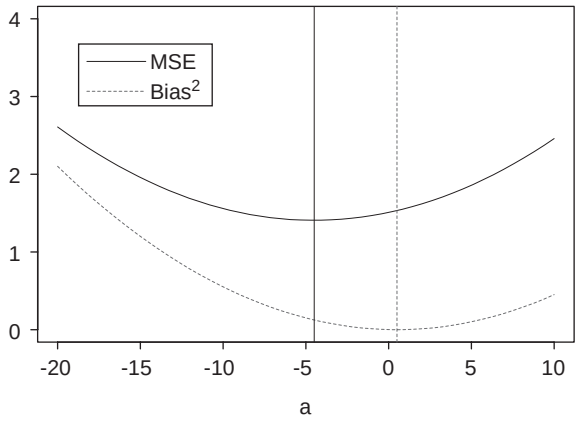

Fig. 1. MSE and Bias^2 as functions of a.

3.2. Discussion about the ML/REML estimators

Four aforementioned estimators (ML, REML, MM, and MB) belong to the class of estimators defined in (9), with

$$a_{\mathrm{ML}} = 0,$$
$$a_{\mathrm{REML}} = p + 1,$$
$$a_{\mathrm{MM}} = p - 1 - 3nv_0 - 3\mathrm{RSS}/(2m),$$
$$a_{\mathrm{MB}} = p + 1 - nv_0 - \mathrm{RSS}/(2m).$$

According to Theorem 1, under model (1), both the MSE and the squared bias of $\widehat{\mu}_c(x_0)$ are asymptotically quadratic functions of a when ignoring the $o(1/n^2)$ terms. Fig. 1 plots them as functions of a for $p = 1$, $n = 50$, $x_i = (1, i/50)^{\mathrm{T}}$, $\beta = (1, 1)^{\mathrm{T}}$, $\sigma^2 = 1$, and $x_0 = (1, 0.5)^{\mathrm{T}}$. The constants a_{MM} and a_{MB} are marked by the solid and dotted vertical lines, respectively, which correspond to the MM and MB estimators. In practice, one could have different tradeoffs between the bias and variance, but an estimator with both larger MSE and larger bias would certainly be undesirable. Fig. 1 suggests that, in principle, one should always choose a value of a between a_{MM} and a_{MB} to construct an estimator, because any value of a outside the interval of $[a_{\mathrm{MM}}, a_{\mathrm{MB}}]$ is worse than either the MM or the MB estimator in terms of both MSE and Bias^2. Since $v_0 \geqslant 0$, it is easy to show that $a_{\mathrm{REML}} > a_{\mathrm{MM}}$ and $a_{\mathrm{REML}} > a_{\mathrm{MB}}$. Consequently, we conclude that the REML estimator should be avoided in practice. The numerical studies in Section 4 also confirm this conclusion. The constant a_{ML}, on the other hand, may fall between a_{MM} and a_{MB} for some combination of X, x_0, σ^2, n, and p.

To define the MB estimator, we choose a such that the bias term is of the order of $o(1/n)$. We can actually do better than that by considering $c = 1/n + a/n^2 + b/n^3 + o(1/n^3)$ and choosing a and b such that $\mathrm{Bias}[\hat{\mu}_c(x_0)] = o(1/n^2)$. See (18) in the Appendix for details. The bias of the estimator obtained is then of higher order than $1/n^2$, and we will refer to it as the super minimum bias (SMB) estimator. Numerical studies indicate that the SMB estimator indeed has smaller bias than the MB estimator, but the difference is negligible for all practical purposes. Thus we do not present its MSE and bias here.

3.3. Parametric bootstrap confidence intervals

For statistical inference purpose, it makes sense to investigate confidence intervals for the log-normal mean, $\mu(x_0)$. Relation (2) suggests that confidence intervals for $\mu(x_0)$ can be derived by exponentiating confidence intervals for $\tau(x_0) = \log[\mu(x_0)] = x_0^{\mathrm{T}}\beta + \frac{1}{2}\sigma^2$. In this section, we first propose a general procedure to derive parametric bootstrap confidence intervals for $\tau(x_0)$ around an arbitrary estimator of the following form:

$$\widehat{\tau}(x_0) = \log[\widehat{\mu}(x_0)] = x_0^{\mathrm{T}}\widehat{\beta} + g(\mathrm{RSS}),$$

which then leads to a confidence interval for $\mu(x_0)$ around $\widehat{\mu}(x_0)$. Then, confidence intervals for $\widehat{\mu}_{\mathrm{MM}}(x_0)$ and $\widehat{\mu}_{\mathrm{MB}}(x_0)$ can be derived as special cases. Our simulation study in Section 4 suggests that the proposed confidence intervals have nice coverage properties.

We know that $x_0^{\mathrm{T}}\widehat{\beta} \sim \mathrm{N}(x_0^{\mathrm{T}}\beta, v_0\sigma^2)$ and $\mathrm{RSS} \sim \sigma^2\chi^2_m$. Then, using the Delta method, we can obtain the following approximate expression for the variance of $\widehat{\tau}(x_0)$:

$$\mathrm{Var}[\widehat{\tau}(x_0)] \approx v_0\sigma^2 + 2m\sigma^4 g'^2(m\sigma^2).$$

Note that σ^2 can be estimated using RSS/m. Define the following statistic:

$$K[\widehat{\tau}(x_0)] = \frac{\widehat{\tau}(x_0) - \tau(x_0)}{\sqrt{\widehat{\mathrm{Var}}\,[\widehat{\tau}(x_0)]}} = \frac{x_0^{\mathrm{T}}\widehat{\beta} + g(\mathrm{RSS}) - \tau(x_0)}{\sqrt{v_0\mathrm{RSS}/m + 2mg'^2(m\mathrm{RSS}/m)(\mathrm{RSS}/m)^2}}. \tag{10}$$

For a significance level α, let t_1 and t_2 be the $\alpha/2$ and $1-\alpha/2$ percentiles of $K[\widehat{\tau}(x_0)]$, respectively. Then, one can obtain a $1-\alpha$ confidence interval for $\tau(x_0)$ as

$$\left[\widehat{\tau}(x_0) - t_2\sqrt{\widehat{\mathrm{Var}}\,[\widehat{\tau}(x_0)]}, \widehat{\tau}(x_0) - t_1\sqrt{\widehat{\mathrm{Var}}[\widehat{\tau}(x_0)]}\right].$$

To estimate the two percentiles, we observe from (10) that $K[\widehat{\tau}(x_0)]$ has the same distribution as

$$T(\sigma) = \frac{N + (\sigma/2\sqrt{v_0})[(2/\sigma^2)g(m\sigma^2 C_m) - 1]}{\sqrt{C_m + (2m/v_0\sigma^2)g'^2(m\sigma^2 C_m)(\sigma^2 C_m)^2}}, \tag{11}$$

where $N \sim \mathrm{N}(0,1)$, $C_m \sim \chi^2_m/m$ and they are independent. Thus we propose the following parametric bootstrap procedure to estimate t_1 and t_2:

(1) generate $N_i \sim \mathrm{N}(0,1)$ and $C_i \sim \chi^2_m/m$ independently for $i = 1, \ldots, B$;
(2) calculate T_i according to (11) with N, C, and σ replaced with N_i, C_i, and $\sqrt{\mathrm{RSS}/m}$;
(3) estimate t_1 by $\hat{t}_1$, the $\alpha/2$ percentile of $\{T_i : i = 1, \ldots, B\}$, and t_2 by $\hat{t}_2$, the $1-\alpha/2$ percentile of the T_i's.

As a result, we obtain a $1-\alpha$ parametric bootstrap confidence interval for $\tau(x_0)$ as

$$\left[\widehat{\tau}(x_0) - \hat{t}_2\sqrt{\widehat{\mathrm{Var}}[\widehat{\tau}(x_0)]}, \widehat{\tau}(x_0) - \hat{t}_1\sqrt{\widehat{\mathrm{Var}}[\widehat{\tau}(x_0)]}\right]. \tag{12}$$

Then, the corresponding $1-\alpha$ bootstrap confidence interval for $\mu(x_0)$ is

$$\exp\left[\widehat{\tau}(x_0) - \hat{t}_2\sqrt{\widehat{\mathrm{Var}}[\widehat{\tau}(x_0)]}, \widehat{\tau}(x_0) - \hat{t}_1\sqrt{\widehat{\mathrm{Var}}[\widehat{\tau}(x_0)]}\right]. \tag{13}$$

In particular, our two estimators correspond to the following g functions:

$$g_{\mathrm{MM}}(\mathrm{RSS}) = \log[f_{\mathrm{MM}}(\mathrm{RSS})] = \frac{m\mathrm{RSS}}{2(n-p+1+3nv_0)m + 3\mathrm{RSS}}$$

and

$$g_{\mathrm{MB}}(\mathrm{RSS}) = \log[f_{\mathrm{MB}}(\mathrm{RSS})] = \frac{m\mathrm{RSS}}{2(n-p-1+nv_0)m + \mathrm{RSS}}.$$

We can use the above general parametric bootstrap procedure to generate the corresponding confidence intervals for our proposed estimators.

Alternatively, one can assume that $\widehat{\tau}(x_0)$ is normally distributed and derive an approximate $1-\alpha$ variance confidence interval for $\mu(x_0)$ as

$$\exp\left[\widehat{\tau}(x_0) - z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}[\widehat{\tau}(x_0)]}, \widehat{\tau}(x_0) + z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}[\widehat{\tau}(x_0)]}\right], \tag{14}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-percentile of the standard normal distribution. In Section 4.2, the coverage properties of the two confidence intervals are investigated via a simulation study.

4. Simulation studies

4.1. Numerical comparison of the estimators

In this section we numerically compute the MSE and bias of the MM and MB estimators for some combinations of n and σ^2 which are typical in many applications, and compare them with the other estimators mentioned in Section 2 to illustrate how different estimators perform under different scenarios. We assume that the data can be modeled as in (1), with only one covariate x, taking values that are uniformly distributed between 0 and 1. The regression coefficient vector $\beta = (\beta_0, \beta_1)^{\mathrm{T}}$ is taken to be $(1, 1)^{\mathrm{T}}$. Extension to scenarios with multiple covariates is simple. We present our results for $\sigma^2 \in \{0.25, 0.5, 1\}$ and sample size $n \in \{10, 50\}$, where σ^2 is the variance of ε_i. σ^2 larger than one is unlikely to occur in real applications, and sample sizes larger than 50 yield qualitatively similar results as the case of $n = 50$; hence those results are not reported here. We consider the estimation of $\mu(x_0)$ for $x_0 \in \{0, 0.1, \ldots, 1.2\}$.

For different combinations of n and σ^2, Fig. 2 plots the relative MSE of the ML, REML, UMVU, EV, MB, and MM estimators, which is the ratio between the MSEs of a particular estimator and the MM estimator. The relative MSE of the MM estimator is always one by construction, and the values for the other estimators represent their MSEs as percentages of the MSE of the MM estimator. The following observations are made from the plots:

1. The MM estimator has the smallest MSE among all the estimators for all the cases we considered, and the difference in some cases is rather substantial. For example, for a small sample size ($n = 10$), the MSE of the UMVU estimator is about 8% larger for $\sigma^2 = 0.25$, and about 40% larger for $\sigma^2 = 1$. Even when the sample size is relatively large ($n = 50$), the MSE of the UMVU estimator is still more than 10% larger when $\sigma^2 = 1$. These results show that the MM estimator can be much more efficient in terms of MSE, especially when the sample size is not large.

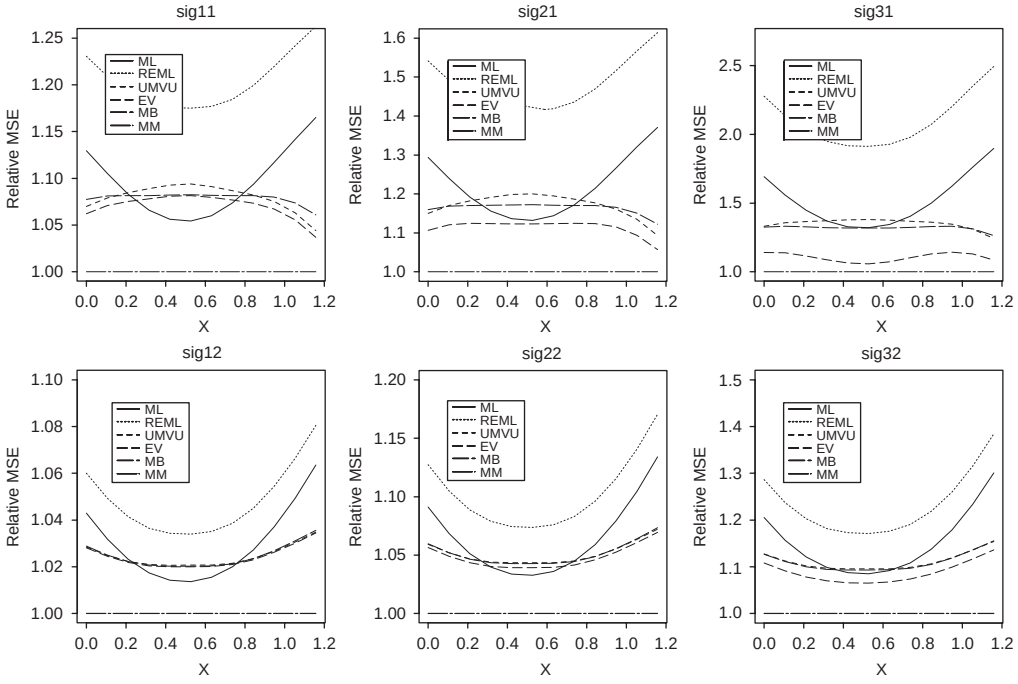

Fig. 2. Relative MSE for different estimators, which is obtained by dividing the MSE of each estimator by the MSE of the MM estimator. The MM estimator has the smallest MSE.

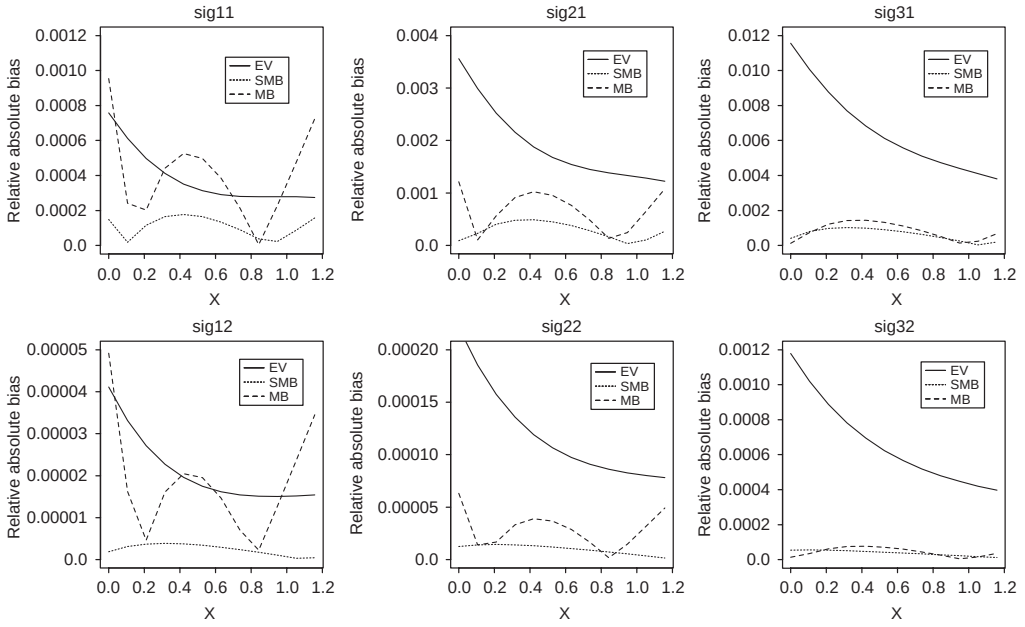

Fig. 3. Relative absolute bias for different estimators, which is obtained by dividing the absolute bias by $\mu(x_0)$. The MB/SMB estimators perform the best in most cases.

2. The MSE of the MB estimator is very close to that of the UMVU estimator, especially when n is large. Since the bias of the MB estimator is also very small (see Fig. 3 and its comment 2), in practice one can use the MB estimator as a surrogate for the UMVU estimator, which is much harder to compute.
3. The REML estimator is worse than all the other estimators across the whole range of x_0 under all the considered scenarios, and the difference is quite large for a small sample size ($n=10$) and a large σ^2 ($\sigma^2=1$). This is consistent with what we found in Section 3, and reaffirmed our claim that the REML estimator should not be used.
4. The MSE of the ML estimator is comparable to the UMVU, EV, and MB estimators when x_0 is close to the center of the data used to fit the regression model, but it increases much faster than the other estimators when x_0 departs from the center. Thus one needs to be extra cautious when using the ML estimator for extrapolation.

Below we compare the biases for the three estimators constructed to reduce bias, the EV, MB, and SMB estimators. In Fig. 3, the relative absolute bias of these three estimators are plotted. The relative absolute bias is defined as the absolute bias of an estimator of $\mu(x_0)$ divided by $\mu(x_0)$. Thus it has the nice interpretation as being the bias in terms of the percentage of the estimand. The following comments can be made about the plots:

1. All three estimators have rather small relative bias. In the worst case ($n=10$, $\sigma^2=1$, and $x_0=0$), the largest relative bias for the EV estimator is barely larger than 1%. In all the cases we considered, the relative bias of the MB and SMB estimators are below 0.2%.
2. Except for the case when $\sigma^2=0.25$, the MB estimator has smaller bias than the EV estimator. The SMB estimator has the smallest bias over most of the range of x_0. However, the difference between the MB estimator and the SMB estimator is too small to justify the use of the more complicated SMB estimator.

4.2. *Coverage properties of the bootstrap and variance confidence intervals*

In this section, we use the same setup as in Section 4.1 to investigate the coverage properties of the bootstrap and variance confidence intervals for the MM and MB estimators. The two confidence intervals are defined in (13) and (14), respectively. For each combination of sample size and σ^2, 5000 confidence intervals are derived for each estimator at each x_0. The empirical coverage probability is calculated as the proportion of these confidence intervals that cover the

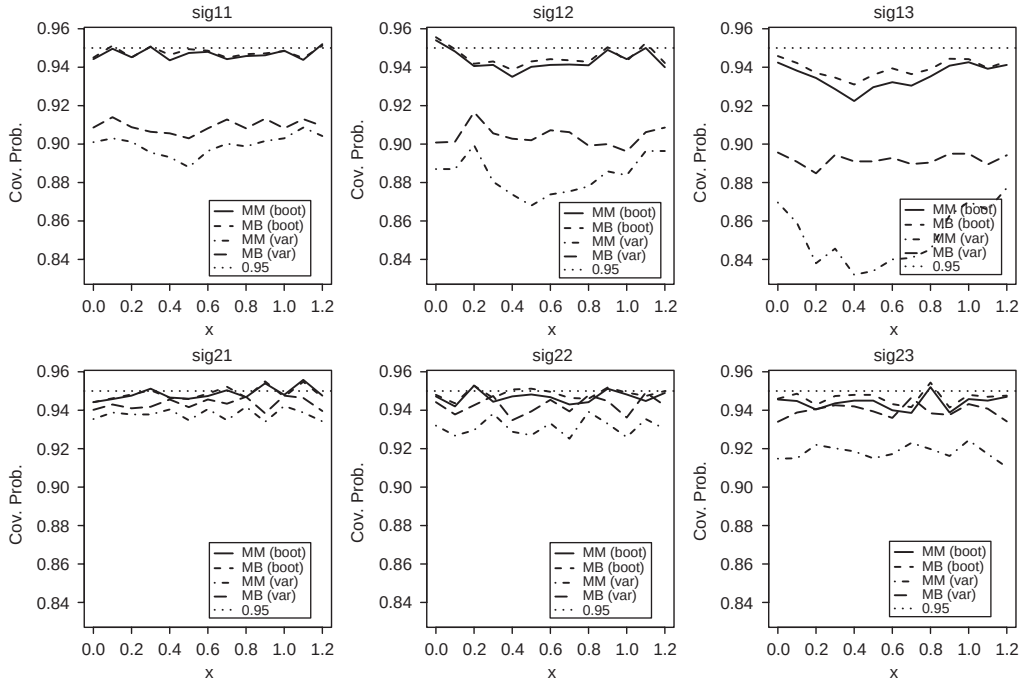

Fig. 4. Coverage probabilities for the bootstrap and variance confidence intervals of the MM and MB estimators. The bootstrap intervals have nice coverage properties and perform better than the variance intervals.

true value. To derive a bootstrap confidence interval, we set B to be 5000 as well when using the parametric bootstrap procedure of Section 3.3.

Fig. 4 compares the empirical coverage probabilities for the confidence intervals corresponding to the two estimators. The nominal coverage level (95%) is also superimposed as the benchmark. A number of findings emerge from the plots:

1. In general, the bootstrap confidence intervals have better coverage properties than the variance confidence intervals. The bootstrap confidence intervals perform reasonably well with coverage probabilities ranging between 92% and 96%. On the other hand, the variance intervals have coverage probabilities ranging between 83% and 95%, and undercover considerably for a small sample size, especially when σ^2 is large.
2. The bootstrap confidence intervals for the two estimators have comparable coverage properties. The average coverage probabilities across the x range are 94.4% and 94.6%, respectively. We also compare the widths of the confidence intervals and find that the MM estimator leads to shorter intervals in all cases. The average width ratio ranges between 96.0% and 99.7%.
3. The variance confidence intervals for the MB estimator have higher coverage probabilities than those for the MM estimator (90.4% versus 92.7%).

Thus, we recommend to use the bootstrap confidence intervals, especially in scenarios where a small sample size and a large σ^2 might occur. The MM bootstrap confidence interval is preferred for its shorter width.

5. Application to a sediment discharge study

The study of sediment transport is of interest to many people because of its environmental impacts (Bollman, 1992; Cohn, 1995). Elliott and Anders (2004) gave a summary of a sediment data set from the Yampa River and Upper Green River Basin, Colorado. The water-resource development at the Upper Colorado River Basin has significantly changed the sediment delivery. This may have potential effects on the habitat for endangered fishes. It is important to understand

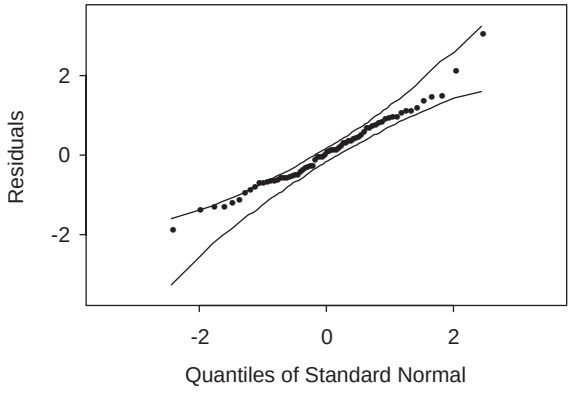

Fig. 5. Normal quantile plot with confidence band for the residuals of the sediment data.

the timing and mechanics of sediment delivery at a specific location, such as the spawning bars for razorback sucker in the Green River, so that one can create a sediment budget. To derive the sediment-transport equation, data were collected at five sediment-sampling sites within the watershed. Sediment loads were calculated using instantaneous measurements of streamflow, suspended-sediment concentration, and bedload. The calculation of sediment loads was performed for both suspended load and bedload using the methods described in Guy and Norman (1970) and Edwards and Glysson (1999). Each sediment measurement was given a seasonal label of *normal*, *rising-limb*, or *falling-limb* based on when the sample was collected relative to the date of the annual peak discharge. To predict suspended-sediment loads, separate sediment-transport curves were developed for each type using the REML estimator under the log-linear model (1). Troutman and Williams (1987) pointed out that such a modeling approach is appropriate when the objective is to predict the dependent variable.

In this section, we use the data from the Little Snake River near Lily, Colorado, to illustrate the practical performance of our estimators. The same methodology can be applied to other locations. The gage is located at latitude 40.32.50, longitude 108.25.25. Detailed site description can be found in Elliott and Anders (2004). We focus on the suspended-sediments measured between 1994 and 2002. The sediment-transport equations derived in Elliott and Anders (2004) are

$$\log(L) = 2.122 + 0.854\log(Q) \tag{15}$$

for rising limb and

$$\log(L) = -4.653 + 1.706\log(Q) \tag{16}$$

for falling limb, where L is the sediment discharge in tons per day, and Q is the water discharge in the unit of m^3/s. The error variances are estimated to be $\hat{\sigma}^2_{\mathrm{REML}} = 0.742$ and 0.925, respectively.

There are 68 observations available at this location with either rising limb or falling limb. We fit a log-normal linear model with $\log(Q)$ as one covariate, and season (rising limb/falling limb) as an indicator variable. The fitted regression lines are identical to those obtained in Elliott and Anders (2004), and the estimated error variance is $\hat{\sigma}^2_{\mathrm{REML}} = 0.903$ with $R^2 = 0.69$. The normal quantile plot of the residuals (Fig. 5) shows no significant deviation from the normal assumption, as none of the observations are outside the 95% confidence band. The confidence band is computed using simulations from the same regression model with errors randomly drawn from a normal distribution with mean 0 and variance 0.903. The ACF plot of the residuals also shows no sign of time-dependency.

Table 1 compares various estimators obtained from the model. The first two columns give the average relative bias and MSE for each estimator. The MB estimator has the smallest bias (0.1%) among all the estimators other than the UMVU estimator, which has no bias by construction. The MM estimator has the largest bias. The average relative MSE of the MB estimator is very close to that of the UMVU estimator. Since the MB estimator can be easily computed using the OLS estimators of β and σ^2, it can be used as a surrogate of the UMVU estimator in practice when no software is handy to evaluate the Hypergeometric function. The MM estimator has the smallest average relative MSE among all

Table 1
Comparison of estimators for the sediment data

	Bias	MSE	RMSPE
MLE	0.021	0.0629	7710
REML	0.036	0.0677	7754
UMVU	0.000	0.0598	7677
EV	0.002	0.0594	7674
MM	0.052	0.0546	7600
MB	0.001	0.0599	7681

Notes: Bias and MSE are average relative bias and MSE, respectively. RMSPE is the square-root of the average of delete-one squared prediction errors.

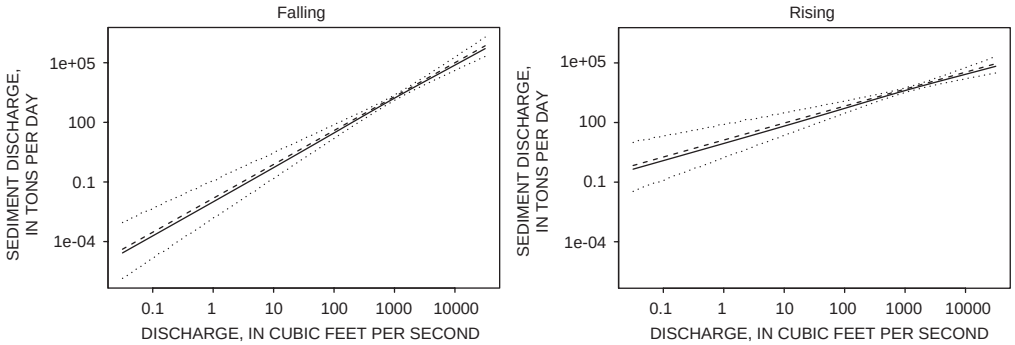

Fig. 6. The MM estimator for the sediment data (the solid line) with its bootstrap confidence band (the dotted lines). The REML estimator is also plotted as a reference (the broken line).

estimators. It is about 8% better than the EV estimator, the second best in terms of MSE, and about 19% better than the REML estimator, the one used in Elliott and Anders (2004).

We also use cross-validation to compare the performance of the different estimators. More specifically, we use the "delete-one" procedure, in which for each i, we delete the ith observation Y_i, re-estimate the regression parameters $\hat{\beta}_{-i}$ and $\hat{\sigma}^2_{-i}$ without Y_i, and estimate $\mu(x_i)$ using the estimators. For each estimator, we compute the *root mean squared prediction error* (RMSPE) as

$$\text{RMSPE} = \left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\mu}(x_i))^2\right)^{1/2},$$

and the results are reported in the third column of Table 1. The MM estimator is again better than the other estimators with a smaller margin. Note that our estimators are optimized for estimating the mean responses instead of predicting the new observations. We believe this is more relevant in the current context as one is mostly interested in estimating the long-term average sediment discharge for a given set of streamflow characteristics. Nevertheless, the MM estimator still gives the best performance in terms of predicting new observations, as is demonstrated by cross-validation.

For *falling* and *rising* flows separately, Fig. 6 plots the MM estimator of the mean sediment discharge as a function of water discharge (the solid line), along with its bootstrap confidence band. The REML estimator is also plotted for comparison purpose (the broken line). The MM estimator appears uniformly smaller than the REML estimator, and it is not a linear function of the water discharge on the log scale. Our theoretical results as well as the simulation studies indicate that the MM estimator has a smaller MSE for estimating the mean sediment discharge at the original scale, and it is as easy to compute as the REML estimator. Its use is thus recommended if one is interested in accurate estimation of the mean sediment discharge.

6. Conclusion

In this paper, we have proposed two new estimators for the conditional mean of the response variable in the original scale for log-normal linear models, the MM estimator and the MB estimator. Both estimators can be viewed as degree-of-freedom-adjusted "ML" estimators, which are simple functions of the ML estimator of the regression coefficient β and the RSS from the log-normal linear models, and hence are very easy to compute. Comparisons with the existing estimators show that the MB estimator has almost identical performance as the UMVU estimator, while is much easier to compute. It thus can be used as a simple alternative to the UMVU estimator. The MM estimator has smaller MSE than the other estimators for the scenarios considered, and the gain in efficiency can be quite substantial when the sample size is small and σ^2 is moderately large. Its use is recommended if one wants to minimize the squared error risk. A parametric bootstrap procedure is also proposed to produce confidence intervals for our estimators, which have nice coverage properties. Among the existing estimators, we conclude that the usual REML estimator should never be used, and our simulation studies suggest that the EV estimator (El-shaarawi and Viveros, 1997), i.e. the bias-corrected REML estimator, often achieves a reasonable balance between bias and variance, even though its bias is slightly larger than our MB estimator, and its MSE is larger than our MM estimator.

The current study assumes that the errors in the log-normal linear models are independent and homoscedastic. In practice, both assumptions may be violated. It is relatively straight forward to adapt our approach to take into account possible inhomogeneity in the error variance. However, it is more difficult to adjust for correlated errors, which are often present in time series and spatial data sets, and deserve further investigation. We intend to address this issue in a separate manuscript.

Acknowledgments

The authors thank the AE and two referees whose comments improved the paper. The authors are partially supported by the US NSF Grants DMS-0606577 and DMS-0605434 and SAMSI.

Appendix

Proof of Theorem 1. We first consider estimators of the form $\widehat{\mu}_c(x_0)$ with the following expansion of c:

$$c = \frac{1}{n} + \frac{a}{n^2} + \frac{b}{n^3} + \mathrm{o}\left(\frac{1}{n^3}\right).$$

For $c = 1/(n-a)$, $b = a^2$.

Notice the following Taylor series expansion:

$$\log(1-t) = -\sum_{i=1}^{\infty} \frac{t^i}{i}.$$

Define $V_1 = \exp[(2v_0-1)\sigma^2](1-2c\sigma^2)^{-m/2}$ and $V_2 = \exp[\frac{1}{2}(v_0-1)\sigma^2](1-c\sigma^2)^{-m/2}$. Below we expand V_1 and V_2 using the above Taylor expansion.

$$\begin{aligned}
V_1 &= \mathrm{e}^{[(2v_0-1)\sigma^2 - \frac{m}{2}\log(1-2c\sigma^2)]} \\
&= \mathrm{e}^{[(2v_0-1)\sigma^2 + \frac{m}{2}(2c\sigma^2+2c^2\sigma^4+\frac{8}{3}c^3\sigma^6+o(\frac{1}{n^3}))]} \\
&= \mathrm{e}^{[\frac{2n^2v_0+n(a-p-1)+b-a(p+1)}{n^2}\sigma^2+(\frac{1}{n}+\frac{2a-1}{n^2})\sigma^4+\frac{4}{3n^2}\sigma^6+\mathrm{o}(\frac{1}{n^2})]} \\
&= 1 + \left[2v_0 + \frac{a-p-1}{n} + \frac{b-a(p+1)}{n^2}\right]\sigma^2 + \left(\frac{1}{n} + \frac{2a-p-1}{n^2}\right)\sigma^4 + \frac{4}{3n^2}\sigma^6
\end{aligned}$$

$$
\begin{aligned}
&+\frac{(2nv_0+a-p-1)^2}{2n^2}\sigma^4+\frac{1}{2n^2}\sigma^8+\frac{2nv_0+a-p-1}{n^2}\sigma^6+\mathrm{o}\left(\frac{1}{n^2}\right)\\
&=1+(2nv_0+a-p-1+\sigma^2)\frac{\sigma^2}{n}+[b-a(p+1)]\frac{\sigma^2}{n^2}\\
&\quad+\left[2a-p-1+2n^2v_0^2+2nv_0(a-p-1)+\frac{1}{2}(a-p-1)^2\right]\frac{\sigma^4}{n^2}\\
&\quad+\left(\frac{4}{3}+2nv_0+a-p-1\right)\frac{\sigma^6}{n^2}+\frac{1}{2n^2}\sigma^8+\mathrm{o}\left(\frac{1}{n^2}\right).
\end{aligned}
$$

$$
\begin{aligned}
V_2&=\mathrm{e}^{[\frac{1}{2}(v_0-1)\sigma^2-\frac{m}{2}\log(1-c\sigma^2)]}\\
&=\mathrm{e}^{[\frac{1}{2}(v_0-1)\sigma^2+\frac{m}{2}(c\sigma^2+\frac{c^2\sigma^4}{2}+\frac{c^3\sigma^6}{3}+\mathrm{o}(\frac{1}{n^3}))]}\\
&=\mathrm{e}^{[\frac{n^2v_0+n(a-p-1)+b-a(p+1)}{2n^2}\sigma^2+\frac{n+2a-p-1}{4n^2}\sigma^4+\frac{1}{6n^2}\sigma^6+\mathrm{o}(\frac{1}{n^2})]}\\
&=1+\frac{n^2v_0+n(a-p-1)+b-a(p+1)}{2n^2}\sigma^2+\frac{n+2a-p-1}{4n^2}\sigma^4+\frac{1}{6n^2}\sigma^6\\
&\quad+\frac{(nv_0+a-p-1)^2}{8n^2}\sigma^4+\frac{1}{32n^2}\sigma^8+\frac{nv_0+a-p-1}{8n^2}\sigma^6+\mathrm{o}\left(\frac{1}{n^2}\right)\\
&=1+\left(nv_0+a-p-1+\frac{\sigma^2}{2}\right)\frac{\sigma^2}{2n}+[b-a(p+1)]\frac{\sigma^2}{2n^2}\\
&\quad+\left[a-\frac{p+1}{2}+\frac{n^2v_0^2}{4}+\frac{nv_0(a-p-1)}{2}+\frac{(a-p-1)^2}{4}\right]\frac{\sigma^4}{2n^2}\\
&\quad+\left(\frac{1}{3}+\frac{nv_0}{4}+\frac{a-p-1}{4}\right)\frac{\sigma^6}{2n^2}+\frac{\sigma^8}{32n^2}+\mathrm{o}\left(\frac{1}{n^2}\right).
\end{aligned}
\tag{17}
$$

According to Lemma 1, we know that

$$\mathrm{MSE}[\widehat{\mu}_c(x_0)]=\mu^2(x_0)(V_1-2V_2+1).$$

Thus, incorporating the above expressions for V_1 and V_2, we obtain

$$\mathrm{MSE}[\widehat{\mu}_c(x_0)]=\mu^2(x_0)\frac{\sigma^2}{n}\left\{1+\frac{\sigma^2}{2}+\frac{\sigma^2}{4n}[a^2+(2-2p+6nv_0+3\sigma^2)a+f(p,n,\sigma^2,v_0)]\right\}+\mathrm{o}\left(\frac{1}{n^2}\right),$$

where $f(p,n,\sigma^2,v_0)=-1+p^2-6nv_0(p+1)+7n^2v_0^2+(1-3p+7nv_0)\sigma^2+7\sigma^4/4$.

From Lemma 1, we also know that

$$\mathrm{Bias}[\widehat{\mu}_c(x_0)]=\mu(x_0)(V_2-1).$$

Thus, incorporating the above expression for V_2, we obtain

$$\mathrm{Bias}[\widehat{\mu}_c(x_0)]=\mu(x_0)\frac{\sigma^2}{2n}\left(nv_0+a-p-1+\frac{\sigma^2}{2}\right)+\mathrm{o}\left(\frac{1}{n}\right).$$

Note that the above asymptotic expressions of MSE and Bias do not depend on the second constant b in the expansion of c. This means that we can drop the third term of the expansion and consider $c=1/(n-a)=1/n+a/n^2+\mathrm{o}(1/n^2)$. This proves Theorem 1.

In addition, if we keep one more term for $\mathrm{Bias}[\widehat{\mu}_c(x_0)]$, then

$$\mathrm{Bias}\left[\widehat{\mu}_c(x_0)\right]=\mu(x_0)\left(\frac{\sigma^2}{2n}c_1+\frac{\sigma^2}{2n^2}c_2\right)+\mathrm{o}\left(\frac{1}{n^2}\right),\tag{18}$$

where

$$c_1=nv_0+a-p-1+\frac{1}{2}\sigma^2$$

and

$$c_2 = b - a(p+1) + \left[a - \frac{p+1}{2} + \frac{n^2 v_0^2}{4} + \frac{n v_0 (a-p-1)}{2} + \frac{(a-p-1)^2}{4} \right] \sigma^2 + \left(\frac{1}{3} + \frac{n v_0}{4} + \frac{a-p-1}{4} \right) \sigma^4 + \frac{1}{16} \sigma^6.$$

Thus, one can define an estimator by choosing a and b such that c_1 and c_2 equal to 0. See a relevant discussion at the end of Section 3.2. □

References

Bollman, F.H., 1992. The socio-economic perspective, context and implications of sediment monitoring, erosion and sediment monitoring programmes in river basins. Poster Contributions. IAHS, Oslo, Norway. pp. 24–30.

Bradu, D., Mundlak, Y., 1970. Estimation in lognormal linear models. J. Amer. Statist. Assoc. 65, 198–211.

Cohn, T.A., 1995. Recent advances in statistical methods for the estimation of sediment and nutrient transport in rivers. Rev. Geophys. 33 (S1), 1117–1123.

Doray, L.G., 1996. UMVUE of the IBNR reserve in a lognormal linear regression model. Insur. Math. Econom. 18, 43–57.

Edwards, T., Glysson, G., 1999. Field methods for measurement of fluvial sediment (revised). In: Techniques of Water-Resources Investigations of the United States Geological Survey.p. 89.vol. 3, (Chapter C2).

El-shaarawi, A.H., Viveros, R., 1997. Inference about the mean in log-regression with environmental applications. Environmetrics 8, 569–582.

Elliott, J.G., Anders, S.P., 2004. Summary of sediment data from the Yampa River and Upper Green River Basins, Colorado and Utah, 1993–2002. Technical Report 5242, USGS Scientific Investigations Report, ⟨http://pubs.usgs.gov/sir/2004/5242/⟩.

Finney, D.J., 1941. On the distribution of a variate whose logarithm is normally distributed. J. Roy. Statist. Soc. 7 Suppl., 144–161.

Gilliom, R.J., Helsel, D.R., 1986. Estimation of distributional parameters for censored trace level water quality data 1. Estimation techniques. Water Resour. Res. 22, 135–146.

Guy, H.P., Norman, V.W., 1970. Field methods for measurement of fluvial sediment. In: Techniques of Water Resources Investigations of the United States Geological Survey.p. 59.vol. 3, (Chapter C2).

Heien, D.M., 1968. A note on log-linear regression. J. Amer. Statist. Assoc. 63, 1034–1038.

Holland, D.M., De Oliveira, V., Cox, L.H., Smith, R.L., 2000. Estimation of regional trends in sulfur dioxide over the eastern United States. Environmetrics 11 (4), 373–393.

Lawless, J.F., 1982. Statistical Models and Methods for Lifetime Data. Wiley, New York.

Lehmann, E.L., Casella, G., 1998. Theory of Point Estimation. second ed. Springer, New York.

Marcotte, D., Groleau, P., NOV 1997. A simple and robust lognormal estimator. Math. Geol. 29 (8), 993–1009.

Seaborn, J.B., 1991. Hypergeometric Functions and Their Applications. Springer, New York.

Shen, H., Brown, L.D., Zhi, H., 2006. Efficient estimation of log-normal means with application to pharmacokinetic data. Statist. Med. 25, 3023–3038.

Troutman, B.M., Williams, G.P., 1987. Fitting straight lines in the earth sciences. In: Size, W.B. (Ed.), International Association for Mathematical Geology Studies in Mathematical Geology—Use and Abuse of Statistical Methods in the Earth Sciences. Oxford University Press, Oxford, pp. 107–128.