Open Journal of Statistics, **2012, 2, 389-400**
http://dx.doi.org/10.4236/ojs.2012.24047 Published Online October 2012 (http://www.SciRP.org/journal/ojs)

Linear Maximum Likelihood Regression Analysis for Untransformed Log-Normally Distributed Data

Sara M. Gustavsson, Sandra Johannesson, Gerd Sallsten, Eva M. Andersson
Department of Occupational and Environmental Medicine, Sahlgrenska University Hospital,
Academy at University of Gothenburg, Gothenburg, Sweden
Email: sara.gustavsson@amm.gu.se

Received June 25, 2012; revised July 27, 2012; accepted August 10, 2012

ABSTRACT

Medical research data are often skewed and heteroscedastic. It has therefore become practice to log-transform data in regression analysis, in order to stabilize the variance. Regression analysis on log-transformed data estimates the relative effect, whereas it is often the absolute effect of a predictor that is of interest. We propose a maximum likelihood (ML)-based approach to estimate a linear regression model on log-normal, heteroscedastic data. The new method was evaluated with a large simulation study. Log-normal observations were generated according to the simulation models and parameters were estimated using the new ML method, ordinary least-squares regression (LS) and weighed least-squares regression (WLS). All three methods produced unbiased estimates of parameters and expected response, and ML and WLS yielded smaller standard errors than LS. The approximate normality of the Wald statistic, used for tests of the ML estimates, in most situations produced correct type I error risk. Only ML and WLS produced correct confidence intervals for the estimated expected value. ML had the highest power for tests regarding β_1.

Keywords: Heteroscedasticity; Maximum Likelihood Estimation; Linear Regression Model; Log-Normal Distribution; Weighed Least-Squares Regression

1. Introduction

Measurements in occupational and environmental research, e.g. exposure and biomarkers, often have a skewed distribution with a median smaller than the mean and only positive values. It is also common with hetero-scedasticity where the variance increases with the expected value. Such data can often be described by a log-normal or quasi-log-normal distribution [1].

Associations (for example between exposure and health effects/biomarkers or between personal exposure and background variables) are often analyzed using regression models. Ordinary least squares (LS) regression analysis is a commonly used regression method, and it is based on the assumption of a constant variance and a normal distribution for the stochastic term. One way of handling non-normal data is with nonparametric median regression (or Least-Absolute-Value regression), see e.g. [2] where no assumptions are made about the distribution of the response variable. The parameter estimates are then found by minimizing the sum of absolute value of the residuals (whereas LS minimizes the sum of squared residuals). However, as a nonparametric method it requires larger samples and it may have multiple solutions [3].

In situations where the response variable has a skewed distribution and an increasing variance, it has become the practice to log-transform the response variable. Regression analysis on log-transformed data have for instance been used to establish reference models [4], to find suitable biomarkers [5], to determine suitable surrogates for exposure [6,7], and to estimate the cost function in health economics [8]. A model in which the response variable is log-transformed, $\ln(Y)$, will estimate the relative effect of each predictor, whereas in many cases it is the absolute effect that is desired [9]. It must also be considered that e.g. a t-test for comparing the expected values of two groups based on the mean of $\ln(Y)$ is not equal to a test based on the mean of the original log-normal data Y, since the expected value of Y is a function of both μ and σ, whereas the expected value of $\ln(Y)$ is a function of only μ. If the two σ-parameters are not equal, a t-test based on $\ln(Y)$ may not give the correct type I-error regarding the difference between $E[Y_1]$ and $E[Y_2]$, [10,11].

In many cases a linear relationship between the response and the predictor (e.g. between personal exposure and background variables) is a reasonable assumption; for example it is realistic that the personal exposure increases linearly with time spent in a certain environment (e.g. time spent in traffic). This linearity will be lost in a

log-transformation. On the other hand, if the log-normal distribution is ignored in order to preserve the linearity, tests based on the assumption of a constant variance may give misleading results, [12].

There is a need for methods that handle log-normally distributed data in linear regression models, based on moderate sample sizes. In order to estimate the linear association (and the absolute effect), but still take into account the log-normal distribution with a non-constant variance, we propose a maximum likelihood (ML) based method for regression analysis. In this paper we have evaluated this new method using large scale simulations, which allowed us to analyze the bias, variance and distribution of the regression coefficients resulting from the new method, as well as comparing it to LS- and weighted-least-squares (WLS) regression analysis. A data set on personal exposure to 1,3-butadiene in five Swedish cities was used to illustrate the three methods.

2. Data

We considered the situation where the response variable Y is assumed to follow a log-normal distribution (*i.e.* $\ln(Y)$ is normally distributed) and where the expected value is assumed to be a linear function of the predictors, $\mu_Y = E[Y] = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p$. In regression analysis, some X-variables can be included because of known (or suspected) association with Y, in order to decrease the variance or lower the risk of confounder effects [5]. Since Y follows a log-normal distribution, $\ln(Y)$ can be expressed using the following model

$$\ln(Y_i) = \ln\left(\beta_0 + X_1\beta_1 + \cdots + X_p\beta_p\right) - \sigma_Z^2/2 + e_i,$$

where $e_i \sim N\left(0, \sigma_Z^2\right)$. The $\hat{\beta}_k$ is an estimate of the absolute effect of X_k on Y. The log-transformation results in a distortion of the linearity, as can be seen in the model above. Linear regression analysis on the log-transformed data would yield an estimate of the relative effect of X_k, rather than the absolute effect. This is illustrated in **Figure 1**.

A log-transformation is not suitable in situations when the aim is to estimate the absolute effect of X on Y (rather than the relative one). In a licentiate thesis [13] it was suggested that the linear regression should be estimated from untransformed log-normal data using maximum likelihood methods. In this article, the properties of these maximum likelihood estimates will be evaluated.

2.1. Simulation Study

The properties of a new method for statistical analysis can be derived theoretically and/or by simulations and examples. In a simulation study, a model for the variable of interest (here personal exposure) is used to generate samples of observations, $y_1, \cdots, y_n$ and then the parameters under investigation (here regression parameters) are estimated from each sample. This is repeated to obtain a distinctive distribution for the estimates. Our simulation study allowed us to assess the bias and standard errors of the parameters estimated resulting from the new maximum likelihood-based regression analysis, and compare these results to those of LS and WLS.

We used simulation models in which the response, personal exposure to Particulate Matter smaller than 2.5 μm ($PM_{2.5}$), was assumed to be a linear function of background variables (residential outdoor level of $PM_{2.5}$, smoking and time spent in home). Two different simulation models were used to generate data. **Model A** had only one predictor, the personal exposure to $PM_{2.5}$-particles (μg/m^3), Y, was assumed to be a linear function of the residential outdoor concentration of $PM_{2.5}$ (μg/m^3), *ConcOut*. **Model B** had three predictors and no intercept,

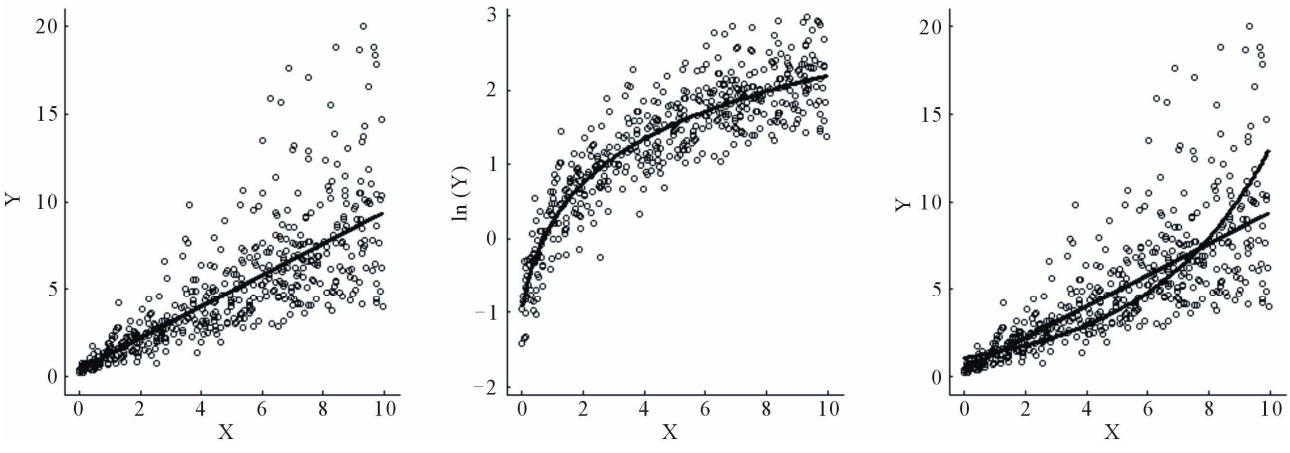

Figure 1. A linear regression where $Y|X$ follows a log-normal distribution. The absolute effect is 0.9 (left), the log-transformation stabilizes the variance but distort the linear relation (middle) and the estimation based on log(Y) result in an exponential function with a relative effect of 29% (right).

Y was a linear function of the number of cigarettes per day, *Smoke*, number of hours spent in their own home, *Home*, and residential outdoor concentration of $PM_{2.5}$ ($\mu g/m^3$), *ConcOut*. Since *ConcOut* > 0, its regression coefficient can be interpreted as a stochastic intercept. Datasets were simulated by generating normally distributed observations from **Model A**:

$$\ln(Y_i) = \ln(4.803 + 0.574 \cdot ConcOut) - 0.354^2/2 + e_i,$$

where $e_i \sim N(0, \sigma_Z^2 = 0.354^2)$. Samples from **Model B** were simulated according to

$$\ln(Y_i) = \ln(2.092 \cdot Smoke + 0.761 \cdot ConcOut + 0.218 \cdot Home) - 0.450^2/2 + e_i,$$

where $e_i \sim N(0, \sigma_Z^2 = 0.450^2)$. The parameters in the simulation models, $\{\beta_0, \beta_1, \sigma_Z\}$ and $\{\beta_1, \beta_2, \beta_3, \sigma_Z\}$, were estimated from real measurement data (Johannesson *et al*. [14]).

The number of repetitions needed in the simulation study was estimated. In order to obtain a 95% confidence interval for $E[\hat{\beta}]$ that is smaller than 2·0.0005, 4 million samples were needed. For the predictors, discrete values were used: *ConcOut* = {2, 8, 14}, *Smoke* = {0, 7, 14} and *Home* = {8, 16, 24}. Sample sizes n = 108 and n = 216 were used in the simulations and the data sets were balanced with regard to the predictors. For Model A a second set is also created with a data structure similar to the observed one in the original dataset [14], which was slightly unbalanced.

2.2. Application to Exposure Data

The properties of the three regression methods (LS, WLS and ML) were illustrated using a set of data on personal exposure of 1,3-butadiene from five Swedish cities. 1,3-butadiene is an alkene and has been listed as a known carcinogen by the International Agency for Research on Cancer (IARC). Traffic and exposure to tobacco smoke are considered to be two sources for personal exposure to 1,3-butadiene [15,16]. Wood burning has also been showed to increase personal exposure [17]. The dataset was collected in a study of exposure to carcinogens in urban air in five Swedish cities; Gothenburg, Umea, Malmo, Stockholm and Lindesberg, see [18], and consisted of 268 measurements of personal 1,3-butadiene exposures. Background data were collected by a questionnaire.

3. Methods

In our investigation, the outcome variable Y was assumed to be log-normal with an expected value that was a linear function of the predictors;

$$\mu_{Y|X_1,\cdots,X_p} = \beta_0 + x_{1,i}\beta_1 + \cdots + x_{1,p}\beta_p$$

From the simulated data, the parameters of the regression model were estimated using ordinary least-squares and weighted-least-squares estimation as well as maximum-likelihood estimation, as described below.

3.1. Least-Squares Estimates

As mentioned before, the inference in ordinary least-squares regression method, LS, is made under the assumption that $Y_i | X \sim \text{iid } N(\mu_{Yi}, \sigma_{Y,i}^2)$ where $\sigma_{Y,i}^2 = \sigma_Y^2$. The standard deviation is assumed constant for all Y and $\hat{\sigma}_Y^2 = \text{MSE}$. The covariance matrix for the vector $\hat{\boldsymbol{\beta}}_{LS}$ is estimated as $\text{cov}(\hat{\boldsymbol{\beta}}_{LS}) = \text{MSE}(\boldsymbol{X'X})^{-1}$, where $\boldsymbol{X}$ is a $n \times (p + 1)$ matrix. The standard errors, $SE(\hat{\boldsymbol{\beta}}_{LS})$, are estimated from the diagonal elements of $\text{cov}(\hat{\boldsymbol{\beta}}_{LS})$.

3.2. Weighted-Least-Squares Estimation

For data that cannot be considered homoscedastic, there are several estimation procedures, for instance the White heteroscedasticity consistent estimator (see e.g. [19], p. 199). Since Y is assumed to follow a log-normal distribution, the nature of the heteroscedasticity is known; $\sigma_{Y,i}^2 = (e^{\sigma_Z^2} - 1) \cdot \mu_{Y_i}^2$. The variance is a function of the expected value. Thus weighted-least-squares regression analysis, WLS, is appropriate, in which each observation is weighted using $W_i \propto \sigma_{Y_i}^{-2}$. The weights can be estimated from the LS regression; $W_i = \hat{Y}_{i,LS}^{-2} \tilde{\propto} \sigma_{Y_i}^{-2}$. The covariance matrix for the vector $\hat{\boldsymbol{\beta}}_{\text{WLS}}$ is estimated as $\text{cov}(\hat{\boldsymbol{\beta}}_{\text{WLS}}) = \text{MSE}_W (\boldsymbol{X'WX})^{-1}$.

3.3. Maximum Likelihood Estimation in Regression

The maximum likelihood estimator (MLE) of some parameter θ is the value at which the likelihood function $L(\theta)$ attains its maximum as a function of θ, with the sample $y_1, \cdots, y_n$ held fixed. The log likelihood is often more convenient to work with and since the logarithm transformation is monotone, the function log $L(\theta|y)$ can be maximized instead. For a continuous variable the likelihood function is the same as the probability density function, pdf. The MLE of θ can be found by differentiating log $L(\theta|y)$ with respect to θ. When θ is a scalar it is enough that the likelihood function is differentiable to get a direct estimate of θ. In a regression analysis, the aim is to estimate the parameter vector

$$\boldsymbol{\theta} = [\beta_0, \beta_1, \cdots, \beta_p, \sigma_Z].$$

Let $y_1, y_2, \cdots, y_n$ be a log-normal sample where

$$E\left[Y_i\right] = \beta_0 + x_i \cdot \beta_1 .$$

Then $\ln\left(Y_i\right) \sim N\left(\mu_{z_i}, \sigma_Z^2\right)$ where

$$\mu_{z_i} = \ln\left(\beta_0 + x_i \cdot \beta_1\right) - \sigma_Z^2/2$$

and the log likelihood function is

$$\ln L = -n\ln\left(\sigma_Z\right) - \frac{n}{2}\ln\left(2\pi\right)$$
$$-\frac{1}{2\sigma_Z^2}\sum_i\left(\ln y_i - \ln\left(\beta_0 + x_i \cdot \beta_1\right) - \sigma_Z^2/2\right)^2 - \sum_i \ln y_i .$$

The derivatives were previously calculated by Yurgens [13] and are given below with some corrections:

$$\frac{\partial \ln L}{\partial \beta_k} = \frac{1}{\sigma_Z^2}\sum_i \frac{x_{ik}}{\sum_j x_{ij}\beta_j}\left(\ln\left(y_i\right) - \ln\left(\sum_j x_{ij}\beta_j\right) + \frac{\sigma_Z^2}{2}\right),$$

$$\frac{\partial^2 \ln L}{\partial \beta_m \partial \beta_k}$$
$$= -\frac{1}{\sigma_Z^2}\sum_i\left[\frac{x_{ik}x_{im}}{\left(\sum_j x_{ij}\beta_j\right)^2}\left(1 - \ln\left(\sum_j x_{ij}\beta_j\right) + \ln\left(\mathrm{y}_i\right) + \frac{\sigma_Z^2}{2}\right)\right],$$

$$\frac{\partial \ln L}{\partial \sigma_Z} = -\frac{n}{\sigma_Z} + \frac{1}{\sigma_Z^3}\sum_i\left(\ln\left(y_i\right) - \ln\left(\sum_j x_{ij}\beta_j\right)\right)^2 - \frac{n\sigma_Z}{4},$$

$$\frac{\partial^2 \ln L}{\partial \sigma_Z \partial \beta_k} = -\frac{2}{\sigma_Z^3}\sum_i \frac{x_{ik}}{\sum_j x_{ij}\beta_j}\left(\ln\left(y_i\right) - \ln\left(\sum_l x_{il}\beta_l\right)\right),$$

$$\frac{\partial^2 \ln L}{\partial \sigma_Z^2} = \frac{n}{\sigma_Z^2} + \frac{3}{\sigma_Z^4}\sum_i\left(\ln\left(y_i\right) - \ln\left(\sum_j x_{ij}\beta_j\right)\right)^2 - \frac{n}{4}.$$

The maximum likelihood estimates of $\boldsymbol{\theta} = \left[\beta_0, \beta_1, \cdots, \beta_p, \sigma_Z\right]$ can be found by iterations, for example the Newton-Raphson algorithm, see [20]. The covariance matrix for the vector $\hat{\boldsymbol{\theta}} = \left[\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\sigma}_Z\right]$ is estimated by the inverse of the observed Fisher information matrix (see e.g, [21]), where the elements are the negative second derivative of the log-likelihood. One of the known properties of MLE is, under some regularity conditions, its asymptotic normality; when the sample size increases the MLE of θ tend to a normal distribution with expected value θ and a covariance matrix equal to the inverse of the Fisher information matrix (see e.g, [21]).

3.4. Descriptive Statistics and Inference

In the simulation study, samples of log-normal observations were generated and three different methods were used to estimate the regression coefficients. The results from the three methods were compared using expected value (mean), standard error, bias, skewness, the 95% central range and correlation of the regression estimates. The *standard error* of $\hat{\beta}_i$ was denoted $SE\left[\hat{\beta}_i\right]$ and the *sample specific standard error* for $\hat{\beta}_i$, $se\left(\hat{\beta}_i\right)$, was the estimate of $SE\left[\hat{\beta}_i\right]$. The *bias* of an estimator, $E\left[\hat{\beta}_i\right] - \beta_i$, was used as a measure of the systematic error. The *skewness* of an estimator was estimated as $\gamma = E\left[\left(\left(\hat{\beta} - E\left[\hat{\beta}\right]\right)/SE\left[\hat{\beta}\right]\right)^3\right]$. For log-normal data, the skewness is $\gamma = \left(\mathrm{e}^{\sigma^2} + 2\right)\sqrt{\mathrm{e}^{\sigma^2} - 1}$, while a symmetric distribution has $\gamma = 0$. It has been suggested that a sample with skewness less than 0.5 should be considered as approximately symmetric while skewness above 1 may be considered highly skewed. The 95% *central range*, CR_{95}, was defined as the difference between the 2.5th and 97.5th percentiles. The sample-specific standard error of $\hat{\mu}_{Y|X}$ is

$$se\left(\hat{\mu}_{Y|X}\right) = \sqrt{\sum_{i=1}^{p} x_i^2 \cdot se\left(\hat{\beta}_i\right)^2 + \sum_{i \neq j} x_i x_j \cdot \mathrm{cov}\left(\hat{\beta}_i, \hat{\beta}_j\right)},$$

where $x_0 = 1$ and $\mathrm{cov}\left(\hat{\beta}_i, \hat{\beta}_j\right)$ is the sample specific estimation of $\mathrm{cov}\left[\hat{\beta}_i, \hat{\beta}_j\right]$.

The inference properties of the three methods were evaluated by comparing the results from tests of the null hypothesis H_0:$\beta = \beta^*$, where $\beta^* = \left\{0, \beta_T\right\}$ in which β_T is the value specified in the simulation model. LS and WLS estimates was tested with the t-test, $t\left(\hat{\beta}\right) = \left(\hat{\beta} - \beta^*\right)/se\left(\hat{\beta}\right)$, which follows a Student's t-distribution. The ML estimates were tested using a Wald-type of test statistic, which utilized the large sample normality of the ML-estimator and the observed Fisher information. Under H_0, the Wald statistic $W\left(\hat{\beta}\right) = \left(\hat{\beta} - \beta^*\right)/se\left(\hat{\beta}\right)$ asymptotically follows a $N(0, 1)$ distribution and $W\left(\hat{\beta}\right)^2$ asymptotically follows a chi-square distribution, see e.g. [21], Section 9.4. Other possible large-sample test procedures for ML estimates (not used in this study) are the score statistic and the full likelihood ratio statistic, see e.g. [21]. We chose the Wald statistic because of its computational advantages.

The properties of the test statistics above, when used on log-normal data, were evaluated by the risk of type I error and the power. The true probability of a type I error for a specified nominal α-value (denoted α^*) was estimated as the proportion of test statistic values beyond the respective critical value. The power results were based on simulations in which the tested parameter (e.g. β_0) was varied according to H_1 whereas the other parameter (e.g. β_1) was held constant.

4. Results

The results are presented in four sections; the distribution

of $\hat{\beta}_i$ and $\hat{\sigma}_Z$, predictions $\left(\hat{\mu}_{Y/X}\right)$, inference and an application to exposure data for 1,3-butadiene.

4.1. The Distribution of the Estimates

For data generated according to Model A, all three methods produced unbiased estimates of the β-coefficients (the absolute effect of the X-variables), see **Table 1**. Of the three methods, ML gave the smallest $SE\left[\hat{\beta}_i\right]$, 11% smaller than that of LS. For LS, $se\left(\hat{\beta}_0\right)$ overestimated $SE\left[\hat{\beta}_0\right]$ by over 35% for the balanced data and about 18% for the unbalanced. For all three methods the standard error increased when unbalanced data were used. LS had the largest CR_{95}, while ML has the smallest (the differences were around 12% between LS and ML and

Table 1. The expected value (E[–]), standard errors (SE[–], E[se(–)]) and skewness (γ[–]) for the estimates of β, and the expected value and standard errors for the estimate of σ_Z[a].

		n = 108			n = 216		
		LS	WLS	ML	LS	WLS	ML
Balanced data							
β_0 = 4.803	$E\left[\hat{\beta}\right]$	4.803	4.803	4.806	4.803	4.803	4.804
	$SE\left[\hat{\beta}\right]$	0.486	0.441	0.430	0.343	0.312	0.304
	$E\left[se\left(\hat{\beta}\right)\right]$	0.656	0.438	0.424	0.465	0.311	0.302
	$\gamma\left[\hat{\beta}\right]$	0.063	0.151	0.143	0.045	0.107	0.100
	$CR_{95}\left[\hat{\beta}\right]$	1.904	1.730	1.685	1.346	1.223	1.190
β_1 = 0.574	$E\left[\hat{\beta}\right]$	0.574	0.574	0.574	0.574	0.574	0.574
	$SE\left[\hat{\beta}\right]$	0.072	0.066	0.064	0.051	0.047	0.045
	$E\left[se\left(\hat{\beta}\right)\right]$	0.070	0.066	0.064	0.050	0.047	0.045
	$\gamma\left[\hat{\beta}\right]$	0.125	0.054	0.053	0.089	0.039	0.039
	$CR_{95}\left[\hat{\beta}\right]$	0.282	0.260	0.252	0.199	0.183	0.178
σ_Z = 0.354	$E\left[\hat{\sigma}_Z\right]$	-	0.351	0.350	-	0.353	0.352
	$SE\left[\hat{\sigma}_Z\right]$	-	0.028	0.024	-	0.020	0.017
Unbalanced							
β_0 = 4.803	$E\left[\hat{\beta}\right]$	4.803	4.803	4.807	4.803	4.803	4.805
	$SE\left[\hat{\beta}\right]$	0.564	0.488	0.475	0.399	0.345	0.335
	$E\left[se\left(\hat{\beta}\right)\right]$	0.665	0.484	0.469	0.472	0.344	0.334
	$\gamma\left[\hat{\beta}\right]$	−0.030	0.156	0.149	−0.020	0.110	0.105
	$CR_{95}\left[\hat{\beta}\right]$	2.215	1.911	1.860	1.565	1.352	1.315
β_1 = 0.574	$E\left[\hat{\beta}\right]$	0.574	0.574	0.573	0.574	0.574	0.574
	$SE\left[\hat{\beta}\right]$	0.089	0.078	0.075	0.063	0.055	0.053
	$E\left[se\left(\hat{\beta}\right)\right]$	0.082	0.077	0.075	0.058	0.055	0.053
	$\gamma\left[\hat{\beta}\right]$	0.184	0.024	0.026	0.130	0.017	0.018
	$CR_{95}\left[\hat{\beta}\right]$	0.347	0.305	0.296	0.246	0.215	0.209
σ_Z = 0.354	$E\left[\hat{\beta}\right]$	-	0.351	0.350	-	0.353	0.352
	$SE\left[\hat{\beta}\right]$	-	0.028	0.024	-	0.020	0.017

[a]Data were generated from Model A, 4 million iterations. For reference, the skewness of the normal distribution is γ = 0, whereas the log-normal data $y_1 \cdots y_n$ from Model A has skewness γ = 1.15.

3% between WLS and ML). There was a strong association between the LS- and ML-estimates; the correlation between the estimates was 0.88 and 0.90 for $\hat{\beta}_0$ and $\hat{\beta}_1$ respectively. The association was weaker for high values. The WLS- and ML-estimates were even more similar, with a correlation of 0.97 for both the β_0- and β_1-estimates, and with no weakening of association for higher values. For σ_Z, both WLS and ML showed only small biases that decreased with increasing sample size, **Table 1**. ML gave the smallest standard error and seemed robust to unbalanced data. Since there is no generally accepted method for estimating σ_Z with LS, no such results were presented.

Data were also generated from versions of Model A in which one of the β-parameters was set to 0. All three methods produced unbiased estimates of β and the standard errors were much smaller than the results shown in **Table 1** (both for $SE\left[\hat{\beta}\right]$ and $E\left[se\left(\hat{\beta}\right)\right]$); for the zero-parameter the standard error was between 45 and 76% smaller than the corresponding value in **Table 1**, and for the other parameter the standard error was between 26% and 52% smaller. For a situation where the intercept is zero ($\beta_0 = 0$), ML gave the smallest $SE\left[\hat{\beta}\right]$ for both parameters, 50% smaller than that of LS. For both ML and WLS, $se\left(\hat{\beta}_0\right)$ was a good estimator for $SE\left[\hat{\beta}_0\right]$, but for LS $se\left(\hat{\beta}_0\right)$ overestimated $SE\left[\hat{\beta}_0\right]$ by about 80%. For a situation where the predictor X has no effect on the response Y ($\beta_1 = 0$), all three methods produced approximately the same $SE\left[\hat{\beta}_1\right]$ and $se\left(\hat{\beta}_1\right)$ was a good estimator for $SE\left[\hat{\beta}_1\right]$. The σ_Z-estimates (and their standard error) were independent of the values of β_0 and β_1 and had the same values as in **Table 1**.

For data with a large variation (large σ_Z) we detected the occasional estimation problem for ML. The ML method is based on iterations whereas LS and WLS have analytical expressions for the parameter estimates, and ML was more sensitive to large outliers. Situations in which ML produces unreasonable results can sometimes be avoided by excluding the extreme observations, but all three methods will then tend to underestimate both the intercept and the standard errors.

Data were also generated according to Model B. ML provided the smallest variation ($SE\left[\hat{\beta}_i\right]$ was between 18% and 44% smaller than that of LS). There was a very small underestimation of σ_Z, but the bias decreases with increasing sample size, as expected according to the properties of MLE. As before, ML had the smaller $SE\left[\hat{\sigma}_Z\right]$, **Table 2**. For both ML and WLS, $se\left(\hat{\beta}\right)$ as a good estimator for $SE\left[\hat{\beta}\right]$. For LS, $se\left(\hat{\beta}_1\right)$ underestimated $SE\left[\hat{\beta}_1\right]$ by 7% - 8% while $se\left(\hat{\beta}_3\right)$ did overestimate $SE\left[\hat{\beta}_i\right]$ by 12% - 13%, **Table 2**. The $SE\left[\hat{\beta}_i\right]$ depends on the value of β_i, the range and values of the X-variables and the value of σ_Z. A separate simulation was conducted in which $\beta_0 = 0$, $\beta_1 = \beta_2 = \beta_3 = 1$, $\sigma_Z = 0.450$ and where all predictors (*ConcOut*, *Smoke* and *Home*) hade values between 2 and 14 and the results showed that for this situation all the standard errors were the same;

$$SE_{\mathrm{LS}}\left(\hat{\beta}_1\right) = SE_{\mathrm{LS}}\left(\hat{\beta}_2\right) = SE_{\mathrm{LS}}\left(\hat{\beta}_3\right) = 0.196,$$

$$SE_{\mathrm{WLS}}\left(\hat{\beta}_1\right) = SE_{\mathrm{WLS}}\left(\hat{\beta}_2\right) = SE_{\mathrm{WLS}}\left(\hat{\beta}_3\right) = 0.169$$

and

$$SE_{\mathrm{ML}}\left(\hat{\beta}_1\right) = SE_{\mathrm{ML}}\left(\hat{\beta}_2\right) = SE_{\mathrm{ML}}\left(\hat{\beta}_3\right) = 0.161.$$

4.2. Predictions

The results in **Tables 1** and **2** illustrated that the $\hat{\beta}$-values were approximately symmetrical and hence the same would apply to $\hat{\mu}_{Y|X}$ for fixed x. All three methods provided unbiased estimates of the β-parameters and thus the point estimates of $\mu_{Y|X}$ will be unbiased. For ML and WLS, $se\left(\hat{\mu}_{Y|X}\right)$ did adequately estimate $SE\left[\hat{\mu}_{Y|X}\right]$, while LS produced too large standard errors for $x < \overline{x}$ and too small for $x > \overline{x}$. This will for most values of x result in erroneous confidence intervals for LS. ML did produce slightly narrower confidence intervals than WLS, see **Figure 2**.

For a simple regression model (with intercept and one X-variable) it can be shown that $SE\left[\hat{\mu}_{Y|X}\right]$ has a minimum at $x = -\rho_{0,1} \cdot SE\left[\hat{\beta}_0\right] \Big/ SE\left[\hat{\beta}_1\right]$, in **Figure 2** at $x \approx 5\left(\rho_{0,1} = \mathrm{Corr}\left[\hat{\beta}_0, \hat{\beta}_1\right]\right)$. As mentioned above, this minimum was well estimated with both ML and WLS, while LS indicated a minimum at $x = 8$. The incorrect $se\left(\hat{\mu}_{Y|X}\right)$ for LS is a result of the underestimated correlation ($r_{0,1} < \rho_{0,1}$) and overestimating the standard error ($E\left[se\left(\hat{\beta}_0\right)\right] > SE\left[\hat{\beta}_0\right]$). For a regression model with no intercept and several X-variables, $SE\left[\hat{\mu}_{Y|X}\right]$ is always minimized for $\left\{x_1 = x_2 = \cdots = x_p = 0\right\}$. For a multiple regression model with an intercept, the minimum $SE\left[\hat{\mu}_{Y|X}\right]$ can be found by solving the equation system

$$\frac{\partial \mathrm{Var}\left[\hat{\mu}_{Y|X_1=x_1,\cdots,X_p=x_p}\right]}{\partial x_i} = 0, \forall i = \{1, \cdots, p\}.$$

4.3. Inference

Hypothesis testing regarding separate regression parameters using the t-test (LS and WLS) or the Wald test (ML) was evaluated. Data were generated according to versions of Model A where one parameter was set to zero ($\beta_i = 0$, $I = \{0, 1\}$). The null hypothesis H_0:$\beta_i = 0$ was tested against both one- and two-sided alternatives in a situation where H_0 was true, **Table 3**.

Table 2. The expected value (E[–]), standard errors (SE[–], E[se(–)]) , skewness (γ[–]) and CR_{95} for the estimates of β[a].

		n = 108			n = 216		
		LS	WLS	ML	LS	WLS	ML
$\beta_1 = 2.092$	$E\left[\hat{\beta}\right]$	2.092	2.092	2.087	2.092	2.092	2.090
	$SE\left[\hat{\beta}\right]$	0.243	0.208	0.200	0.172	0.147	0.141
	$E\left[se\left(\hat{\beta}\right)\right]$	0.224	0.205	0.196	0.159	0.146	0.140
	$\gamma\left[\hat{\beta}\right]$	0.213	0.125	0.121	0.150	0.090	0.085
	$CR_{95}\left[\hat{\beta}\right]$	0.953	0.816	0.783	0.674	0.577	0.554
$\beta_2 = 0.761$	$E\left[\hat{\beta}\right]$	0.761	0.761	0.760	0.761	0.761	0.760
	$SE\left[\hat{\beta}\right]$	0.206	0.146	0.139	0.146	0.103	0.099
	$E\left[se\left(\hat{\beta}\right)\right]$	0.205	0.144	0.137	0.146	0.103	0.098
	$\gamma\left[\hat{\beta}\right]$	0.064	0.133	0.124	0.043	0.094	0.087
	$CR_{95}\left[\hat{\beta}\right]$	0.812	0.574	0.547	0.573	0.406	0.386
$\beta_3 = 0.218$	$E\left[\hat{\beta}\right]$	0.218	0.218	0.220	0.218	0.218	0.219
	$SE\left[\hat{\beta}\right]$	0.116	0.068	0.065	0.082	0.048	0.046
	$E\left[se\left(\hat{\beta}\right)\right]$	0.131	0.066	0.063	0.093	0.047	0.045
	$\gamma\left[\hat{\beta}\right]$	−0.102	0.282	0.252	−0.074	0.198	0.177
	$CR_{95}\left[\hat{\beta}\right]$	0.458	0.265	0.253	0.323	0.187	0.179
$\sigma_Z = 0.450$	$E\left[\hat{\sigma}_Z\right]$	-	0.444	0.443	-	0.447	0.446
	$SE\left[\hat{\sigma}_Z\right]$	-	0.038	0.031	-	0.028	0.022

[a]Data were generated from Model B, 4 million iterations. For reference, the skewness of the normal distribution is $\gamma = 0$, whereas the log-normal data $y_1 \cdots y_n$ from Model B has skewness $\gamma = 1.53$.

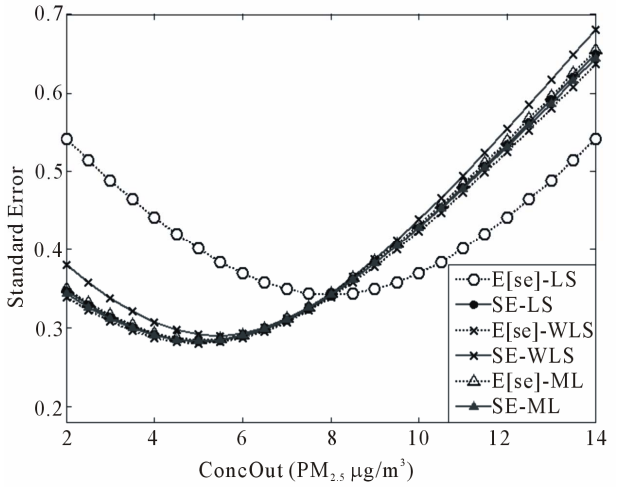

Figure 2. The $SE\left[\hat{\mu}\middle|X = x_0\right]$ and $E\left[se\left(\hat{\mu}\middle|X = x_0\right)\right]$ for the three methods, applied to data generated from Model A. Results were based on 4 million iterations with sample size n = 108.

For a model with no intercept ($\beta_0 = 0$), tests of the LS-estimates produced an α^* which was much smaller than α ($\alpha^* < 0.2\alpha$, regardless of H_1). This was a result of the overestimation of $SE\left[\hat{\beta}_0\right]$ by $se\left(\hat{\beta}_0\right)$. For tests of WLS- and ML-estimates, α^* was approximately equal to α (within 10% of the true value). A slight skewness could be observed; $\alpha^* \le \alpha$ for H_1: $\beta_0 > 0$ and $\alpha^* \ge \alpha$ for H_1:$\beta_0 < 0$, while α^* for H_1:$\beta_0 \ne 0$ was slightly higher or the same as α. The source of this skewness was a positive correlation between $\hat{\beta}_0$ and $se\left(\hat{\beta}_0\right)$; small values of $\hat{\beta}_0$ gave more extreme (negative) test-scores. For a model where X has no effect on Y ($\beta_1 = 0$), all three tests produced α^*-values close to α (within 20% of the true value). For ML, α^* was slightly higher than for LS and WLS. The risk of type I error was approximately the same even when the sample size was doubled (n = 216).

Data were generated from Model A and the hypothesis H_0:$\beta_i = \beta_T$ was tested (β_T was the parameter value specified in Model A). The results for WLS and ML (data not shown) were that α^* was within 30% and 16% of α, re-

spectively for nominal size 0.05 and 0.10. For nominal size 0.01 the relative deviation could be up to 70%. As before, tests of the LS-estimates of β_0 produced an α^* much too small ($\alpha^* \leq 0.1\alpha$). Both WLS- and ML-produced a slight skewness regarding tests of both β_0 and β_1, similar to that in **Table 3**.

Data were also generated according to Model B (data not shown) and the hypothesis H_0:$\beta_i = \beta_T$ was tested. For all three methods, α^* was closest to α for nominal size 0.10. The relative deviation was largest for nominal size 0.01; $0.7\alpha \leq \alpha^* \leq 2.2\alpha$. Again, a skewness (as above) could be seen in the tests of all three parameters, caused by the negative skewness of the test statistics. For tests of the LS-estimate of β_1, α^* was consistently too large ($1.04\alpha \leq \alpha^* \leq 2.20\alpha$), whereas α^* was consistently too small in tests of β_3 ($0.30\alpha \leq \alpha^* \leq 0.76\alpha$). These erroneous risks were, again, the result of under- and overestimation, respectively, of $SE\left[\hat{\beta}_i\right]$, see **Table 2**. Tests regarding the LS-estimates of β_2 followed the same pattern as the tests of all the WLS- and ML-estimates. The risk of type I error was approximately the same even when the sample size was doubled ($n = 216$).

The power of the tests, under the alternative hypothesis H_1:$\beta_1 > 0$, was estimated for both $\alpha = 0.05$ and $\alpha^* = 0.05$, see **Table 4**. When $\alpha = 0.05$, ML had the highest power while LS had the lowest. For LS the type I error

Table 3. Risk for type I error (α^*) for two-sided and one-sided tests regarding β_0 and β_1[a].

				$\alpha = 0.010$			$\alpha = 0.050$			$\alpha = 0.100$		
Alternative hypothesis H_1:				$\beta_i \neq 0$	$\beta_i < 0$	$\beta_i > 0$	$\beta_i \neq 0$	$\beta_i < 0$	$\beta_i > 0$	$\beta_i \neq 0$	$\beta_i < 0$	$\beta_i > 0$
β_i	n		Test									
β_0	108	LS	t-test	0.000	0.000	0.000	0.001	0.001	0.003	0.004	0.007	0.016
	108	WLS	t-test	0.010	0.012	0.008	0.051	0.055	0.047	0.102	0.106	0.096
	108	ML	Wald	0.012	0.013	0.010	0.054	0.056	0.050	0.106	0.106	0.100
	216	LS	t-test	0.000	0.000	0.000	0.001	0.001	0.003	0.003	0.007	0.014
	216	WLS	t-test	0.010	0.011	0.009	0.050	0.053	0.047	0.100	0.104	0.097
	216	ML	Wald	0.011	0.012	0.009	0.052	0.053	0.049	0.102	0.103	0.099
β_1	108	LS	t-test	0.010	0.010	0.010	0.050	0.050	0.050	0.100	0.101	0.100
	108	WLS	t-test	0.011	0.011	0.010	0.052	0.051	0.051	0.102	0.102	0.101
	108	ML	Wald	0.012	0.012	0.011	0.055	0.053	0.053	0.106	0.104	0.103
	216	LS	t-test	0.010	0.010	0.010	0.050	0.050	0.050	0.100	0.100	0.100
	216	WLS	t-test	0.010	0.010	0.010	0.051	0.050	0.051	0.101	0.100	0.101
	216	ML	Wald	0.011	0.011	0.011	0.052	0.051	0.052	0.103	0.101	0.102

[a]Data were generated from Model A where $\beta_i = 0$ and estimations are based on 4 million iterations with balanced data.

Table 4. The power for tests with alternative hypothesis H_1:$\beta_1 > 0$ where the risk of type I error is set to 0.05[a].

	Nominal:			True:		
Critical value:	1.659	1.659	1.645	1.659	1.670	1.673
β_1	LS (t-test)	WLS (t-test)	ML (Wald)	LS (t-test)	WLS (t-test)	ML (Wald)
0.00	0.050	0.052	0.055	0.050	0.050	0.050
0.02	0.150	0.151	0.160	0.150	0.149	0.154
0.04	0.321	0.324	0.344	0.321	0.320	0.334
0.06	0.530	0.533	0.562	0.530	0.528	0.551
0.08	0.721	0.723	0.753	0.721	0.720	0.744
0.10	0.857	0.859	0.882	0.857	0.856	0.876
0.12	0.936	0.938	0.952	0.936	0.936	0.949
0.14	0.975	0.976	0.983	0.975	0.975	0.982
0.16	0.991	0.992	0.995	0.991	0.991	0.994
0.18	0.997	0.997	0.998	0.997	0.997	0.998
0.20	0.999	0.999	1.000	0.999	0.999	1.000
0.22	1.000	1.000	1.000	1.000	1.000	1.000

[a]Data were generated from Model A with different β_1 values but constant $\beta_0 = 4.804$. Estimations are based on 1 million iterations with balanced data and sample size $n = 108$.

risk was correct but for WLS and ML the critical values were adjusted to give $\alpha^* = 0.05$. After adjustment ML still had the highest power while WLS now had a lower power compared to LS. The relative difference between the nominal and true power was largest for $\beta_1 = 0$ and decreases with β_1.

4.4. Application: 1,3-Butadiene Exposure in Five Swedish Cities

The data on 1,3-butadiene were found to be highly skewed ($\gamma = 2.764$) whereas the log-transformed values were approximately symmetrical ($\gamma = 0.279$). The median exposure was similar to the geometric mean exposure (0.345 against 0.386). Thus the data can be considered log-normal. Five predictors were included in the model: "Have you lit a wood-fire or been in a residence heated with wood burning?" (Wood burning, yes/no), "Are you a smoker?" (Smoker, yes/no), "Have you been in an indoor environment where people where smoking?" (ETS, yes/no), "Proportion of time spent in traffic?" (Traffic), and "City of residence": Umeå (Ume), Stockholm (Sthlm), Malmö (Malmo), Gothenburg (Gbg), and reference category Lindersberg. These predictors had been shown to be significant in previous studies on other datasets [15-17].

Neither *Wood burning* nor *Traffic* were significant in any of the regression models (for ML *Traffic* was borderline significant, p = 0.059), **Figure 3**. All three methods showed a significant difference between the cities; in the LS- and WLS-regression models only Gothenburg differed from Lindesberg, but the ML-regression model also showed a significant difference between Malmo and Lindesberg (p = 0.041). Only the ML-regression model showed a significantly lower exposure for non-smokers with no ETS (p = 0.013). There were clear differences in range of the confidence intervals; with one exception ML had the narrowest intervals and LS the widest.

A second analysis with only the non-smokers ($n = 225$) was performed and then *Traffic* became significant in the ML-regression model (p = 0.039). Otherwise the same predictors were significant in all three analyses, as for the full dataset.

5. Discussion

In this study a new maximum likelihood-based method for estimating a linear regression model for log-normal heteroscedastic data was evaluated and compared with two least squares methods. For log-normal data, the log-transformation is often used to stabilize the variance, but this distorts the linear relationship and does not give an estimate of the absolute effect of a predictor.

Our simulation study demonstrated that the new maximum likelihood method (ML) provides unbiased estimates of the regression parameters β_i, and so do the least squares method (LS) and the weighted least squares method (WLS).

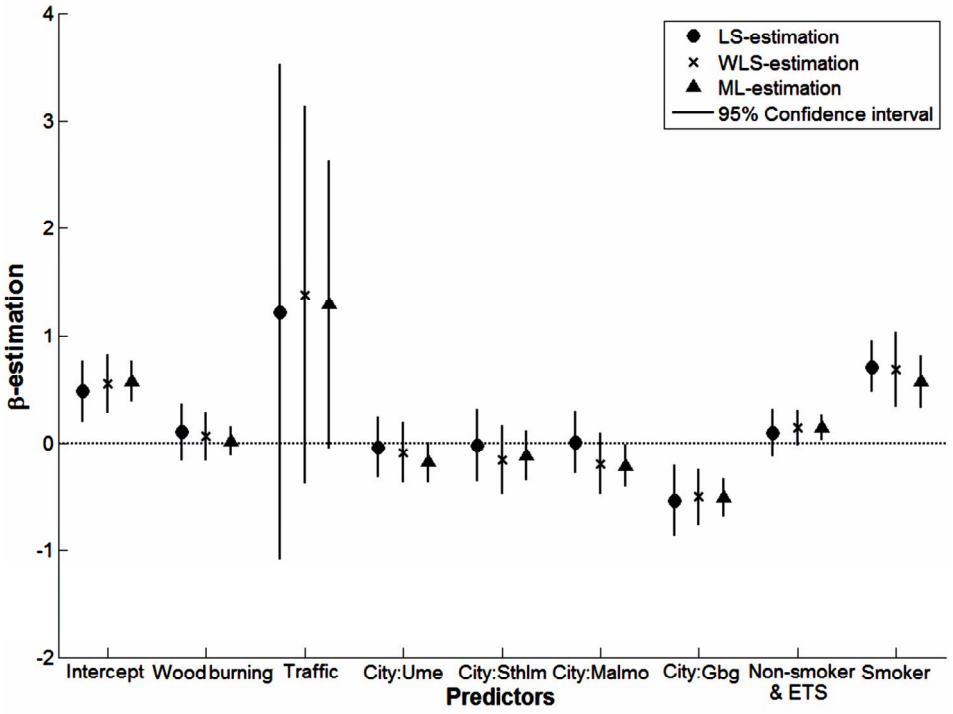

Figure 3. The LS-, WLS- and ML-estimations of the predictor effects, β_i, and their 95% confidence intervals (CI) for regression analysis with 1,3-butadiene as the response.

One reason for proposing a new regression method is the need for a method to estimate a linear regression in the presence of heteroscedasticity. The effect of ignoring the increasing variance is demonstrated in the use of LS, where the standard error, $SE\left[\hat{\beta}_i\right]$, was up to 78% larger than that of ML, for the examples that were investigated. Since the heteroscedasticity of the data was ignored when using LS, the confidence interval was too wide. The results of our example with 1,3-butadiene data were consistent with those in the simulation study; LS overall had the widest confidence intervals and ML the narrowest for the predictor effects.

For all three methods, estimates of the expected response, $\hat{\mu}_{Y|X}$, were unbiased. The confidence interval for $\hat{\mu}_{Y|X}$ is based on $se\left(\hat{\mu}_{Y|X}\right)$, which depends on the value of the predictor, x_0. For a model with one predictor, both ML and WLS produced an almost correct confidence interval, with the narrowest interval at $x_0 = -\rho_{0,1} \cdot SE\left[\hat{\beta}_0\right] \Big/ SE\left[\hat{\beta}_1\right]$, whereas LS had its narrowest confidence interval at $x_0 = \bar{X}$ which therefore does produce erroneous confidence intervals.

For ML and WLS the sample-specific standard error, $se\left(\hat{\beta}_i\right)$, was a good estimator of the true standard error, whereas for LS, $se\left(\hat{\beta}_0\right)$ greatly overestimated $SE\left[\hat{\beta}_0\right]$. When investigating the true risk of type I error, α^*, for tests regarding the regression parameters, the t-tests of the LS-estimates of the intercepts produced an α^* much smaller than the nominal α. This was a consequence of the too large $se\left(\hat{\beta}_0\right)$, causing the distribution of the t-statistics to have fewer observations in the tails than expected. For both the Wald-test (ML) and the t-test used for the WLS-estimates, the α^* was approximately equal to α (for nominal size {0.01, 0.05, 0.10} the largest relative deviations were α^* = {0.021, 0.072, 0.125}). For all three methods, the relative deviation (α^*/α) was largest at nominal size 0.01, indicating a skewness of the test statistics in the tails. Regarding the power (in tests of β_1), ML was superior to LS, both regarding the true and nominal power.

The precision of an estimated expected value ($\hat{\mu}_{Y|X}$) depends both on the regression method and on the variation of the predictors $X_1, \cdots, X_p$. If a predictor has a very small variation, then the estimated effect of that predictor will have a large standard error. In some situations a better estimation of the exposure might be achieve with a model that excludes predictors with very small variation, as has been investigated in [22] for a situation where land-use regression is used to estimate the exposure. Apart from effect estimation, regression analysis can be used to estimate the expected value of the response variable given certain values of the background factors (e.g. to estimate the personal exposure as a function of easily measured background factors). Regression models can also be used to estimate the exposure for a whole population, provided that the distribution of the background factors is known.

When investigating an exposure-disease associations, the measured exposure sometimes include a stochastic error ($exposure_{\text{measured}} = exposure_{\text{true}} + error_1$), which is unrelated to the stochastic error of the (linear) exposure-response model ($response = \delta_0 + \delta_1 \cdot exposure + error_2$). The issue with regression analysis where the explanatory variable has measurement errors is raised in [23]; given the true exposure, the measured exposure has a stochastic variation and different approaches to estimating the true exposure are compared. In situations with measurement errors, the individual- and group-based exposure assessment approaches are often compared, see e.g. [24,25]. In an individual-based approach, each person's measured exposure is used which often leads to a bias of δ_1 towards null. In a group-based approach, each person is assigned an exposure based on group affiliation, which is the same as estimating the exposure from a regression model with only one explanatory variable, *Group*. If several relevant explanatory variables are used to estimate the exposure (rather than only *Group*) the standard error will decrease and, thus, be more similar to an individual-based approach. A group-based design often leads to errors in the exposure which are of Berkson-type ($exposure_{\text{true}} = exposure_{\text{group}} + error_3$), or approximately Berkson-type. As discussed in [24], in a true Berkson-error-model, *the* $error_3$ is independent of $exposure_{\text{group}}$, whereas the approximate Berkson-error may depend of the group size. The approximate Berkson-error is, however, independent of $error_2$. A group-based approach often leads to less bias in the δ_1 estimate but a larger standard error, see e.g. [24,26]. A group-based approach with a log-linear exposure-response model can have a substantial bias in the estimated exposure-response effect [27].

The evaluation of the new ML method, and the comparison to other regression methods, was conducted by a simulation study and exemplified on a dataset. The simulation study allowed us to assess, with great precision, the bias, standard error and the distribution of the parameter estimates, as well as the risk of type I error and the power of the methods. Because of the large number of iterations, our results on bias, standard error and the inference can be considered to be valid, for log-normal data where the relationship is linear. The simulation study allowed us to compare the parameters of the simulation model ("true β values") to the estimates from each method and also to see how these estimates can vary. This would not have been possible had we only used empirical data sets, in which the "true" values are seldom known. Also, to get a reliable estimate of the distribution of the estimated parameters, a large number

of data sets would be needed.

In the simulation study, data were generated from a log-normal heteroscedastic distribution with certain parameter values. However, empirical data might be only approximately log-normal and thus exhibit a larger variance than the one in a perfect log-normal distribution. Thus the results from the simulation study might not hold completely for real data sets of e.g. exposure data. Also, in the simulation study the predictors were assumed to be measured without measurement errors which might be unrealistic for some predictors. Therefore, evaluation of the new ML method should also be made on several empirical data sets.

Two specific simulation models were used in this study where the parameters were estimated from an empirical dataset and therefore the simulation models can be considered to be fairly realistic. The results on bias and standard error are likely to be valid also in models with other parameter values. The parameter estimates are unbiased both for a simple model (with intercept and one explanatory variable) and for a model with three explanatory variables and no intercept, and in both situations ML gives the smallest standard errors. This suggests that the results can be generalized to other datasets.

The ML estimates are derived true iterations and in some situations (sometimes caused by unsuitable start values or large variance in the data) the iterations don't converges. This may lead to unrealistic estimations. In those cases, censoring data larger than the 95% quantile will stabilize the results but will cause underestimations of the intercept and standard errors. In those situations WLS will generally give the best result.

In this study the methods have been evaluated for relatively large samples (n = {108, 216}) and the conclusion was that the asymptotic properties of maximum likelyhood estimates are valid for these sample sizes. However, for smaller samples, further evaluation is needed, especially regarding the distribution of the estimates. In other studies regarding smaller samples of untransformed log-normal data, likelihood-based approaches has been suggested for constructing confidence intervals for the mean and mean response [28,29].

This study illustrated that the new maximum likelyhood-based regression method has good properties and can be used for linear regression on log-normal heteroscedastic data. It provides an estimate of the absolute effect of a predictor, while taking the increasing variance into account in an optimal way. The study also demonstrated that the results from the weighted least squares regression (where the increasing variance is accounted for) were very similar to those from the ML method. Although ML gave slightly narrower confidence intervals and had higher power regarding β_1, both WLS and ML can be recommended for linear regression models for log-normal heteroscedastic data.

6. Acknowledgements

We would like to thank Urban Hjort for sharing his MATLAB codes for the regression analysis.

REFERENCES

[1] P. O. Osvoll and T. Woldbæk, "Distribution and Skewness of Occupational Exposure Sets of Measurements in the Norwegian Industry," *Annals of Occupational Hygiene*, Vol. 43, No. 6, 1999, pp. 421-428.

[2] K. M. McGreevy, S. R. Lipsitz, J. A. Linder, E. Rimm and D. G. Hoel, "Using Median Regression to Obtain Adjusted Estimates of Central Tendency for Skewed Laboratory and Epidemiologic Data," *Clinical Chemistry*, Vol. 55, No. 1, 2009, pp. 165-169. doi:10.1373/clinchem.2008.106260

[3] R. Branham Jr, "Alternatives to Least Squares," *The Astronomical Journal*, Vol. 87, 1982, pp. 928-937. doi:10.1086/113176

[4] A. C. Olin, B. Bake and K. Toren, "Fraction of Exhaled Nitric Oxide at 50 mL/s—Reference Values for Adult Lifelong Never-Smokers," *Chest*, Vol. 131, No. 6, 2007, pp. 1852-1856. doi:10.1378/chest.06-2928

[5] J. O. Ahn and J. H. Ku, "Relationship between Serum Prostate-Specific Antigen Levels and Body Mass Index in Healthy Younger Men," *Urology*, Vol. 68, No. 3, 2006, pp. 570-574. doi:10.1016/j.urology.2006.03.021

[6] L. Preller, H. Kromhout, D. Heederik and M. J. Tielen, "Modeling Long-Term Average Exposure in Occupational Exposure-Response Analysis," *Scandinavian Journal of Work, Environment & Health*, Vol. 21, No. 6, 1995, p. 8. doi:10.5271/sjweh.67

[7] M. Watt, D. Godden, J. Cherrie and A. Seaton, "Individual Exposure to Particulate Air Pollution and Its Relevance to Thresholds For Health Effects: A Study of Traffic Wardens," *Occupational and Environmental Medicine*, Vol. 52, No. 12, 1995, pp. 790-792. doi:10.1136/oem.52.12.790

[8] B. Dickey, W. Fisher, C. Siegel, F. Altaffer and H. Azeni, "The Cost and Outcomes of Community-Based Care for the Seriously Mentally Ill," *Health Services Research*, Vol. 32, No. 5, 1997, p. 599.

[9] R. Kilian, H. Matschinger, W. Loffler, C. Roick and M. C. Angermeyer, "A Comparison of Methods to Handle Skew Distributed Cost Variables in the Analysis of the Resource Consumption in Schizophrenia Treatment," *Journal of Mental Health Policy and Economics*, Vol. 5, No. 1, 2002, pp. 21-32.

[10] J. P. T. Higgins, I. R. White and J. Anzures-Cabrera, "Meta-Analysis of Skewed Data: Combining Results Reported on Log-Transformed or Raw Scales," *Statistics in Medicine*, Vol. 27, No. 29, 2008, pp. 6072-6092. doi:10.1002/sim.3427

[11] X.-H. Zhou, S. Gao and S. L. Hui, "Methods for Comparing the Means of Two Independent Log-Normal Sam-

ples," *Biometrics*, Vol. 53, No. 3, 1997, pp. 1129-1135. doi:10.2307/2533570

[12] T. H. Wonnacott and R. J. Wonnacott, "Regression: A Second Course in Statistics," Wiley, New York, 1981.

[13] Y. Yurgens, "Quantifying Environmental Impact by Log-Normal Regression Modelling of Accumulated Exposure," Licentiate of Engineering, Chalmers University of technology and Göteborg University, Gothenburg, 2004.

[14] S. Johannesson, P. Gustafson, P. Molnar, L. Barregard and G. Sallsten, "Exposure to Fine Particles (PM2.5 and PM1) and Black Smoke in the General Population: Personal, Indoor, and Outdoor Levels," *Journal of Exposure Science and Environmental Epidemiology*, Vol. 17, No. 7, 2007, pp. 613-624. doi:10.1038/sj.jes.7500562

[15] G. J. Dollard, C. J. Dore and M. E. Jenkin, "Ambient Concentrations of 1,3-Butadiene in the UK," *Chemico-Biological Interactions*, Vol. 135-136, 2001, pp. 177-206. doi:10.1016/S0009-2797(01)00190-9

[16] Y. M. Kim, S. Harrad and R. M. Harrison, "Concentrations and Sources of VOCs in Urban Domestic and Public Microenvironments," *Environmental Science & Technology*, Vol. 35, No. 6, 2001, pp. 997-1004. doi:10.1021/es000192y

[17] P. Gustafson, L. Barregard, B. Strandberg and G. Sallsten, "The Impact of Domestic Wood Burning on Personal, Indoor and Outdoor Levels of 1,3-Butadiene, Benzene, Formaldehyde and Acetaldehyde," *Journal of Environmental Monitoring*, Vol. 9, No. 1, 2007, pp. 23-32. doi:10.1039/b614142k

[18] U. Bergendorf, K. Friman and H. Tinnerberg, "Cancer-Framkallande ämnen i Tätortsluft-Personlig Exponering och Bakgrundsmätningar i Malmö 2008," Report to the Swedish Environmental Protection Agency Department of Occupational and Environmental Medicine Malmö, 2010.

[19] W. H. Greene and C. Zhang, "Econometric Analysis," Vol. 5, Prentice Hall, Upper Saddle River, 2003.

[20] J. Hass, M. D. Weir and G. B. Thomas, "University Calculus," Pearson Addison-Wesley, 2008.

[21] Y. Pawitan, "In All Likelihood: Statistical Modelling and Inference Using Likelihood," Oxford University Press, Oxford, 2001.

[22] A. A. Szpiro, C. J. Paciorek and L. Sheppard, "Does More Accurate Exposure Prediction Necessarily Improve Health Effect Estimates? [Miscellaneous Article]," *Epidemiology*, Vol. 22, No. 5, 2011, pp. 680-685. doi:10.1097/EDE.0b013e3182254cc6

[23] Y. Guo and R. J. Little, "Regression Analysis with Covariates That Have Heteroscedastic Measurement Error," *Statistics in Medicine*, Vol. 30, No. 18, 2011, pp. 2278-2294. doi:10.1002/sim.4261

[24] H. G. M. I. Kim, D. Richardson, D. Loomis, M. Van Tongeren and I. Burstyn, "Bias in the Estimation of Exposure Effects with Individual- or Group-Based Exposure Assessment," *Journal of Exposure Science and Environmental Epidemiology*, Vol. 21, No. 2, 2011, pp. 212-221. doi:10.1038/jes.2009.74

[25] S. Rappaport and L. Kupper, "Quantitative Exposure Assessment," Stephen Rappaport, 2008.

[26] E. Tielemans, L. L. Kupper, H. Kromhout, D. Heederik and R. Houba, "Individual-Based and Group-Based Occupational Exposure Assessment: Some Equations to Evaluate Different Strategies," *The Annals of Occupational Hygiene*, Vol. 42, No. 2, 1998, pp. 115-119.

[27] K. Steenland, J. A. Deddens and S. Zhao, "Biases in Estimating the Effect of Cumulative Exposure in Log-Linear Models When Estimated Exposure Levels Are Assigned," *Scandinavian Journal of Work Environment & Health*, Vol. 26, No. 1, 2000, pp. 37-43. doi:10.5271/sjweh.508

[28] J. Wu, A. C. M. Wong and G. Jiang, "Likelihood-Based Confidence Intervals for a Log-Normal Mean," *Statistics in Medicine*, Vol. 22, No. 11, 2003, pp. 1849-1860. doi:10.1002/sim.1381

[29] J. Wu, A. C. M. Wong and W. Wei, "Interval Estimation of the Mean Response in a Log-Regression Model," *Statistics in Medicine*, Vol. 25, No. 12, 2006, pp. 2125-2135. doi:10.1002/sim.2329

