Evaluation of Regression Methods for Log-Normal Data

Linear Models for Environmental Exposure and Biomarker Outcomes

Sara Gustavsson

Occupational and Environmental Medicine
Institute of Medicine
Sahlgrenska Academy at University of Gothenburg

UNIVERSITY OF GOTHENBURG

Gothenburg 2015

Cover illustration: Dolda tal by Sara Gustavsson

Evaluation of Regression Methods for Log-Normal Data

sara.gustavsson@gu.se

ISBN (printed) 978-91-628-9287-6
ISBN (e-publ.) 978-91-628-9295-1

Printed in Gothenburg, Sweden 2015
Kompendiet, Aidla Trading AB

Tamdiu discendum est, quamdiu vivas

(Så länge som man lever, så länge bör man lära)

– Lucius Annaeus Seneca

Evaluation of Regression Methods for Log-Normal Data

Linear Models for Environmental Exposure and Biomarker Outcomes

Sara Gustavsson

Occupational and Environmental Medicine, Institute of Medicine
Sahlgrenska Academy at University of Gothenburg

ABSTRACT

The identification and quantification of associations between variables is often of interest in occupational and environmental research, and regression analysis is commonly used to assess these associations. While exposures and biological data often have a positive skewness and can be approximated with the log-normal distribution, much of the inference in regression analysis is based on the normal distribution. A common approach is therefore to log-transform the data before the regression analysis. However, if the regression model contains quantitative predictors, a transformation often gives a more complex interpretation of the coefficients. A linear model in original scale (non-transformed data) estimates the additive effect of the predictor, while linear regression on a log-transformed response estimates the relative effect.

The overall aim of this thesis was to develop and evaluate a maximum likelihood method (denoted ML_{LN}) for estimating the absolute effects for the predictors in a regression model where the outcome follows a log-normal distribution. The ML_{LN} estimates were compared to estimates using common regression methods, both using large-scale simulation studies, and by applying the method to a number of real-life datasets. The method was also further developed to handle repeated measurements data. Our results show that when the association is linear and the sample size is large (> 100 observations), ML_{LN} provides basically unbiased point estimates and has accurate coverage for both confidence and predictor intervals. Our results also showed that, if the relationship is linear, log-transformation, which is the most commonly used method for regression on log-normal data, leads to erroneous point estimates, liberal prediction intervals, and erroneous confidence intervals. For independent samples, we also studied small-sample properties of the ML_{LN}-estimates; we suggest the use of bootstrap methods when samples are too small for the estimates to achieve the asymptotic properties.

Keywords: log-normal distribution, linear models, absolute effects

ISBN (printed): 978-91-628-9287-6 **ISBN (e-publ.):** 978-91-628-9295-1

SAMMANFATTNING PÅ SVENSKA

Inom miljömedicinsk forskning är det vanligt att man är intresserad av sambanden mellan olika variabler. Vid till exempel bedömning av yrkesexponering är man ofta intresserad av sambanden mellan exponeringen för ett visst ämne och arbetsuppgifter, samt vistelse i olika miljöer. Vanligen används regressionsanalys för att skatta och kvantifiera dessa samband. Den inferens som oftast används vid regressionsanalys bygger på normalfördelningsantaganden, men mycket av de data som finns inom biologi har en positiv snedfördelning. Dessa data är i regel bättre approximerade med en lognormalfördelning och det är därför vanligt att dessa data logtransformeras före en regressionsanalys. Om man har kvantitativa prediktorer och är intresserad av ett specifikt förhållande mellan respons och prediktorer, kan transformationen försvåra tolkning av sambandet. En linjär modell på originaldata skattar den additiva effekten av en prediktor medan linjär regression med en logtransformerad respons skattar den relativa effekten, det vill säga en exponentiell modell. I till exempel exponeringsbedömningar är det ofta mer rimligt att anta att den förväntade kumulativa exponeringen har ett linjärt (additivt) samband till tiden en person har tillbringat i en viss mikromiljö, än att anta ett exponentiellt samband där personen till exempel skulle kunna få en högre exponering andra timmen än första, trots att bakgrundnivåerna var detsamma.

I avhandlingen utvecklar och utvärderar vi en maximum-likelihood-metod för att skatta linjär regression med en log-normal fördelad respons (här kallad ML_{LN}). ML_{LN}-skattningarna jämförs med skattningar från andra vanliga metoder, både i storskaliga simuleringsstudier och i tillämpningar på riktigt mätdata. Simuleringsstudierna visar att för större stickprov där förhållandet mellan prediktorer och respons är linjärt så ger ML_{LN} i princip korrekta effektskattningar och korrekt täckning för både prediktions och konfidensintervall. Logtransformation av responsen kan däremot leda till för smala prediktionsintervall och felaktiga konfidensintervall. Vi studerar även hur man för oberoende observationer bör konstruera konfidensintervall för ML_{LN}-skattningarna när stickproven är små och de asymptotiska egenskaperna (som finns generellt hos ML-skattningar) inte uppnås och ger förslag på alternativa inferensmetoder som inte förlitar sig på skattningarnas asymptotiska fördelning.

LIST OF PAPERS

This thesis is based on the following studies, which are referred to in the text by their Roman numerals.

I. Gustavsson, S. M., Johannesson, S., Sallsten, G., and Andersson, E. M. (2012). Linear Maximum Likelihood Regression Analysis for Untransformed Log-Normally Distributed Data. *Open Journal of Statistics* **2**, 389-400.

II. Gustavsson, S., Fagerberg, B., Sallsten, G., and Andersson, E. (2014). Regression Models for Log-Normal Data: Comparing Different Methods for Quantifying the Association between Abdominal Adiposity and Biomarkers of Inflammation and Insulin Resistance. *International Journal of Environmental Research and Public Health* **11**, 3521-3539.

III. Gustavsson S., and Andersson E. M., Small-Sample Inference for Linear Regression on Untransformed Log-Normal Data. *Submitted for publication.*

IV. Gustavsson S., Akerstrom M., Sallsten G., and Andersson E. M., Linear Regression on Log-Normal Data with Repeated Measurements. *Submitted for publication.*

CONTENT

ABBREVIATIONS

CI	Confidence interval
LS_{exp}	Ordinary Least Squares regression with log-transformed response variable
LS_{lin}	Ordinary Least Squares regression with untransformed response variable
$Marg_{exp}$	Linear covariance pattern model with a log-transformed response.
ML	Maximum likelihood regression
MLE	Maximum likelihood estimator
ML_{LN}	Maximum Likelihood regression in which the likelihood is based on the probability density function of the log-normal distribution
MSE	Mean square error
PI	Prediction interval
se	Standard error
WLS	Weighted least squares

DEFINITIONS IN SHORT

Covariance pattern model	A marginal model in which a structure is specified for the covariance matrix in order to handle the correlation between observations.
Linear model	A model in which the expected values of the response can be written as a linear combination of the predictors.
Marginal model	A model without random effects.
Mixed model	A statistical model that can contain both fixed effects and random effects.
Residual matrix	The covariance matrix for the random error terms in a regression model.
Wald-type interval	A confidence interval for a parameter in the form $estimate \pm percentile \cdot se(estimate)$, where the percentile is selected according to a desired confidence level and a (symmetrical) reference distribution.

1 INTRODUCTION

In research there is often a need to assess and quantify associations between variables. While many statistical methods are based on the assumption that the variable of interest is normally distributed, several biological variables have a skewed distribution and are usually approximated with the log-normal distribution. A natural approach to this problem is to log-transform the variable, so that the transformed variable will have a distribution closer to the normal distribution. However, a potential problem to this approach appears when a specific model is assumed for the association. Linear regression on untransformed data produces a model where the effects are additive, while linear regression on a log-transformed variable produces a multiplicative model.

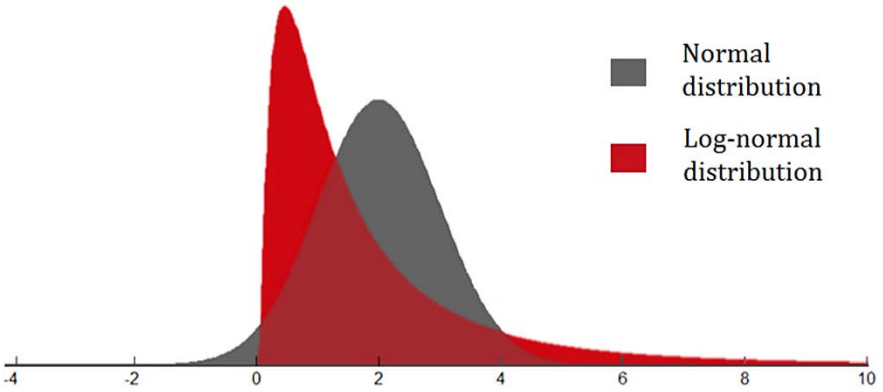

Figure 1. The probability density functions for a normal and a log-normal distribution with expected value 2 and variance 1.

1.1 The Log-Normal Distribution

Data with positively skewed distributions are very common, not least when dealing with biological data. Measurement data often have a lower limit, usually 0 or the detection limit, but no distinct upper limit. Hence below the median no observations can be further away than the lower limit, but above the median there may be values that are many times further away, and this will give a positively skewed distribution. These skewed distributions can often be approximated by the log-normal distribution; some of the theoretical reasons for this are explained elsewhere (Koch, 1966; Koch, 1969; Limpert, Stahel and Abbt, 2001). The log-normal distribution is characterized by having only positive non-zero values, a positive skewness, a non-constant variance that is proportional to the square of the mean value, and a normally distributed

natural logarithm. The probability density function for a log-normal distribution has an asymmetrical appearance, with a majority of the area below the expected value and a thinner right tail with higher values, while the probability density function for the normal distribution with the same expected value has a symmetrical bell-shaped curve (**Figure 1**). Some relationships between the characteristics of the log-normal variable Y and the log-transformed variable $Z = \ln(Y)$ are presented in **Table 1**.

Table 1. Probability density functions and characteristics of the normal and log-normal distributions.

	Normal distribution $ln(Y) = Z \sim N(\mu_Z, \sigma_Z^2)$	*Log-normal distribution* $exp(Z) = Y \sim LN(\mu_Z, \sigma_Z^2)$
Probability density function	$f_Z(z) = \frac{1}{\sqrt{2\pi\sigma_Z^2}} e^{\frac{(z-\mu_Z)^2}{2\sigma_Z^2}}$	$f_Y(y) = \frac{1}{y\sqrt{2\pi\sigma_Z^2}} e^{\frac{(ln(y)-\mu_Z)^2}{2\sigma_Z^2}}$
Expected value	$\mu_Z = ln(\mu_Y) - \frac{\sigma_Z^2}{2}$	$\mu_Y = e^{\mu_Z + \sigma_Z^2/2}$
Geometric mean	$\mu_{gZ} = \mu_Z = ln(\mu_Y) - \frac{\sigma_Z^2}{2}$	$\mu_{gY} = e^{\mu_Z}$
Variance	$\sigma_Z^2 = ln(1 + \sigma_Y^2)/\mu_Y^2$	$\sigma_Y^2 = \left(e^{\sigma_Z^2} - 1\right) e^{2\mu_Z + \sigma_Z^2}$

1.2 Exposure Assessments

The log-normal distribution is often used for modeling airborne exposures. Exposure data are non-negative by nature and often have a larger proportion of moderate sized values with a few higher values, giving the data a positive skewness. These features are shared with the log-normal distribution. There are large quantities of historical empirical results supporting a log-normal assumption for exposure data. Since the 1960s, the distribution of occupational exposures has often been approximated with the log-normal distribution (Rappaport, 1991b). Rappaport suggests that Oldham (1953) was the first to use the log-normal distribution for occupational exposures. Oldham used a normal probability plot to show that the log-transformed values of 779 randomly collected dust measurements were approximately normally distributed, and hence the untransformed values would be approximately log-normally distributed. Others have used formal test procedures to determine the distributions. For example, Kumagai et al. (1997) applied the Shapiro-Wilk W-test on the log-transformed airborne cobalt exposure concentrations obtained in a hard metal factory. None of the tests rejected a log-normal distribution on the 5% significance level. Water et al. (1991) suggested the ratio metric, which is the ratio between the direct sample mean and the maximum likelihood

estimate of the mean given a log-normal distribution, as an measure of goodness-of-fit to the log-normal distribution. They used the ratio metric to show that 15 out of 23 datasets of airborne exposures to mercury and benzene were approximately log-normally distributed. Osvoll and Woldbæk (1999) examined the distribution of 31 different occupational exposures (e.g. lead, cadmium, and welding fumes) in different elements (i.e. air, urine, or blood) and concluded that a log-normal distribution fitted well to most of the samples, or at least better than the normal distribution.

There are not just empirical but also theoretical arguments for assuming a log-normal distribution for time-dependent exposures. Kahn (1973) justified a log-normal model for air pollutants by assuming a non-constant source of error and using the law of proportionate effect, as first presented by Kepteyn (1903). The law of proportionate effect implies that the cumulative at time *t*, denoted Y_t, depends on the cumulative exposure at time *t*-1 and the proportional error *E*, and can be expressed as

$$Y_t - Y_{t-1} = Y_{t-1}E,$$

or as $Y_t = Y_{t-1}(1+E)$. In this expression the *E* is constant. However, Khan assumes a non-constant error, and Rappaport (1990) concurs that this is a reasonable assumption since the exposure over a "longer" time period (e.g. over a workday) is likely to have more than one source of error, including ventilation, mobility of the worker, and differences in work tasks. This will give an exposure that is dependent on a series of errors: $Y_t = Y_0 \prod_{i=1}^{t}(1+E_i)$. By using $(Y_i - Y_{i\text{-}1})/Y_i = E_i$ and $\sum_i^t (Y_i - Y_{i\text{-}1})/Y_i \cong \ln(Y_t) - \ln(Y_0)$, Khan showed that

$$\ln(Y_t) = \ln(Y_0) + \sum_i^t E_i,$$

Then, by the central limit theory, $\ln(Y_t)$ is asymptotically normally distributed regardless of the distribution of E_i, and so Y_t is asymptotically log-normally distributed. There are also arguments that apply to exposures other than air pollutants. Ott (1990), for example, showed that the dilution of one material into another creates non-equilibrium concentrations which are usually approximately log-normal. Ott exemplified this with comparisons to a soluble contamination in water and the release of an airborne pollutant in a ventilated room.

As already mentioned, the log-normal distribution has many characteristics in common with exposure data. It also has the desirable property of a normally distributed logarithm, and hence many traditional statistical approaches become available via a simple transformation of the data. This makes the log-normal distribution simpler to handle than many other skewed distributions. In some cases, concerns have been raised about the log-normal model being

applied to exposure data on the basis of tradition and simplicity, without the assumption ever being verified (Bencala and Seinfeld, 1976; Blackwood, 1992; Waters et al., 1991). However, Bencala and Seinfeld (1976) and Blackwood (1992) concluded that even if data are sometimes better fitted with other common statistical distributions, like the gamma distributions, in terms of accuracy of results they can usually be adequately fitted with a log-normal model.

The Arithmetic Mean and the Geometric Mean

As already stated, exposure data often have a positive skewness, and the geometric mean is less affected by skewness than the arithmetic mean. However, in risk assessment it is not really the central tendency that is of interest but rather the magnitude of an exposure, and then the arithmetic mean is often more appropriate (Parkhurst, 1998; Rappaport, 1991a). This has also been recognized by the Worker Health and Safety Branch of the California Department of Pesticide Regulation (Powell, 2003).

In risk assessment, the final aim is to determine the increased risk of damage given by an exposure. When considering the risk to an individual, a linear exposure-burden-damage model can be used (Rappaport, 1991a). In this model, the bodily damage is proportional to the bodily burden, which in turn is proportional to the cumulative exposure; that is, the product of the arithmetic mean and the time period of the exposure. For a positively skewed distribution, the geometric mean will be less or equal to the arithmetic mean, so the geometric mean would underestimate the dose. Rappaport points out that even if the effects are non-linear, a person's risk of damage (i.e. disease) would still be associated with the cumulative exposure.

In practice, exposure data may not be available for all individuals, so there is a need to evaluate the exposure on a group level. In group assessments, all individuals in a group are assumed to have the same exposure. Crump (1998) points out that grouping data can lead to biased risk assessment. However, Crump also recognizes that ungrouped data are often not available, and suggest that the group mean rather than the group median should be used if the dose-response relationship is linear, convex, or s-shaped, since it will give a less biased risk assessment. Rappaport (1991a) recommends the arithmetic mean over the geometric mean, giving the argument that an estimated group exposure will basically assign the same exposure to all group members, and if all group members are not uniformly exposed then the most common situation is for the exposure to be positively skewed, and a geometric mean would not capture the most highly exposed member. Hence, the use of a geometric mean would lead to an underestimation of the total group exposure.

The conclusion is that the arithmetic mean is usually considered more appropriate than the geometric mean when it comes to risk assessments. A consequence of this is that median regression is not an appropriate method for exposure assessments.

1.3 Linear and Log-Linear Models

Two common methods for regression analysis are ordinary least squares regression and linear mixed models. The inference in both these methods often assumes a normally distributed response, and should not be applied directly to a log-normal variable. Hence, the log-normal variable Y is often log-transformed, and the log-transformed values $Z = \ln(Y)$ will follow a normal distribution with expected value μ_Z and standard deviation σ_Z.

A linear regression analysis with Y as the response and X as the predictor will estimate the absolute effect, here denoted β, of X on Y; that is, Y is expected to increase by β units for every unit change in X. A linear regression analysis with Z as the response will estimate the relative effect, here denoted δ, of X on Y; that is, Y is expected to increase by $100 \cdot (\exp(\delta) - 1)$ percent for every unit change in X. If the relationship is indeed linear, the log-transformation results in a distortion of the relationship, as seen in the center of **Figure 2**, and the difference in expected values between the two models is illustrated on the right of **Figure 2**. Thus for a linear relationship, the arithmetic mean will be biased if log-transformation is applied.

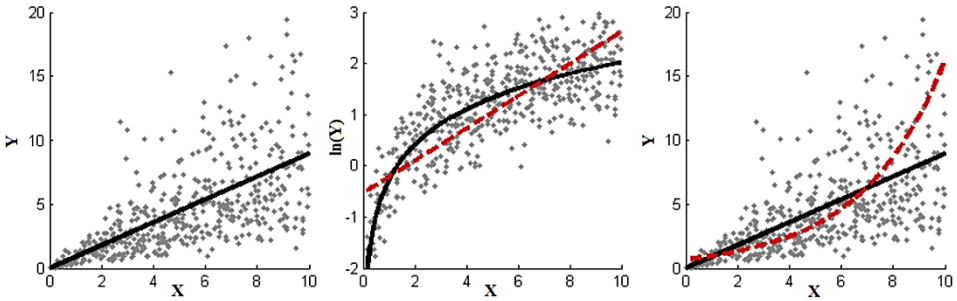

Figure 2. A linear regression where Y|X follows a log-normal distribution. The linear relationship between Y and X is represented by a black line, while an estimated exponential relationship is represented by the dached red line. The absolute effect is 0.9, (*left*)*, the log-transformation stabilizes the variance but distorts the linear relation* (*center*)*, and the estimation based on ln(Y) results in an exponential function with a relative effect of 29%* (*right*)*.* (*Paper I, Figure 1.*)

2 AIM

The overall aim of this thesis is to evaluate a method for estimating the absolute effects for the predictors in a regression model where the outcome follows a log-normal distribution, both for independent data and in repeated measurement situations.

The specific aims are:

1. Define the method for the situation with independent observations and:

    a. Evaluate the large-sample situation regarding point estimates, standard errors, and hypothesis testing.

    b. Evaluate the small-sample situation regarding point estimates, standard errors, and hypothesis testing.

    c. Compare the method to other more commonly used regression methods.

2. Adapt the method to handle repeated measurements which are not independent and:

    a. Evaluate the large-sample situation regarding point estimates, standard errors, and hypothesis testing.

    b. Compare the method to other more commonly used regression methods for repeated measurements.

3 REGRESSION ANALYSIS

We consider the situation where the response variable *Y* follows a log-normal distribution conditional on the predictors $X_1,\ldots X_k$, and where the expected values of the response, μ_Y, is a linear combination of these predictors. This can be expressed in matrix form notation as

$$\underset{(n_{tot}\times 1)}{\boldsymbol{\mu_Y}} = \underset{(n_{tot}\times p)}{\boldsymbol{X}} \cdot \underset{(p\times 1)}{\boldsymbol{\beta}}, \tag{3.1}$$

where n_{tot} is the total number of observations, p is the number of regression coefficients, $\boldsymbol{\mu_Y}$ is a vector of the expected values of **Y**, $\boldsymbol{\beta}$ is a vector of regression coefficients ("effects"), and **X** is the design matrix containing the predictor values and a first column of ones (for the intercept).

The log-transformed values Z=ln(Y) will be normally distributed and can be expressed as

$$\underset{(n_{tot}\times 1)}{\ln(\boldsymbol{Y})} = \underset{(n_{tot}\times 1)}{\boldsymbol{Z}} = \underset{(n_{tot}\times 1)}{\boldsymbol{\mu_Z}} + \underset{(n_{tot}\times 1)}{\boldsymbol{\varepsilon}}, \quad \text{with} \quad \underset{(n_{tot}\times 1)(n_{tot}\times n_{tot})}{\boldsymbol{\varepsilon} \sim N(\mathbf{0}\ ,\ \boldsymbol{\Sigma})},$$

where **Y** is a vector of response values, $\boldsymbol{\mu_Z}$ is a vector of the expected values of *Z*, $\boldsymbol{\varepsilon}$ is a vector of random terms, **0** is a constant matrix of zeros, and $\boldsymbol{\Sigma}$ is the covariance matrix for $\boldsymbol{\varepsilon}$, called the residual matrix. Using Equation (3.1) and the relationships between the mean in the log-normal distribution and the mean in the normal distribution as presented in **Table 1**, we get the following model for *Z*:

$$\ln(\boldsymbol{Y}) = \boldsymbol{Z} = ln(\boldsymbol{X\beta}) - \sigma_Z^2/2 + \boldsymbol{\varepsilon}, \text{ with } \boldsymbol{\varepsilon} \sim N(\mathbf{0}\ ,\ \boldsymbol{\Sigma}).$$

The model can also be expressed for *Y* as

$$\boldsymbol{Y} = \boldsymbol{X\beta} \cdot exp\left(-\sigma_{\boldsymbol{Z}}^{\boldsymbol{2}}/2\right) \cdot exp(\boldsymbol{\varepsilon}), \text{ with } \boldsymbol{\varepsilon} \sim N(\mathbf{0}\ ,\ \boldsymbol{\Sigma}). \tag{3.2}$$

For independent observations, $\boldsymbol{\Sigma}$ will be a constant diagonal matrix of the form $\boldsymbol{\Sigma} = \sigma_Z^2\boldsymbol{I}$, where $\boldsymbol{I}$ is the $n_{\text{tot}} \times n_{\text{tot}}$ identity matrix. For repeated measurements, $\boldsymbol{\Sigma}$ will be a block diagonal matrix defined as

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma_1} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \boldsymbol{\Sigma_2} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Sigma_3} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{3.3}$$

where **0** is a matrix block of zeros and $\boldsymbol{\Sigma}_i$ is the covariance matrix for the *i*:th individual. A model where $\boldsymbol{\Sigma}$ is defined as in (3.3) is here denoted as *a covariance pattern model*. In Paper IV, two different patterns are considered for $\boldsymbol{\Sigma}_i$; a compound symmetry pattern and a first-order autoregressive pattern.

In a compound symmetry covariance model, CS, the variance and covariance are constant; that is, $var(\varepsilon_{ij}) = \sigma^2$ for $i = 1, \dots, m, j = 1, \dots, n_i$ and $cov(\varepsilon_{ij}, \varepsilon_{ik}) = \sigma^2\rho, \forall\, j \neq k$. Thus, the covariance matrix for an arbitrary individual can be written as

$$\mathbf{\Sigma}_i = \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \cdots \\ \rho & 1 & \rho & \cdots \\ \rho & \rho & 1 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \tag{3.4}$$

The variances in a first order autoregressive model, denoted AR(1), are also all equal, but the covariance between two measurements on the same individual decreases exponentially with the distance |*j*-*k*|; $cov(\varepsilon_{ij}, \varepsilon_{ik}) = \sigma^2\rho^{|j-k|}, \forall\, j$. The covariance matrix for an arbitrary individual can be written as

$$\mathbf{\Sigma}_i = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n_i-1} \\ \rho & 1 & \rho & \dots & \rho^{n_i-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{n_i-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n_i-1} & \rho^{n_i-2} & \rho^{n_i-3} & \dots & 1 \end{bmatrix}. \tag{3.5}$$

Note that the error terms $\boldsymbol{\varepsilon}$ are added in log-scale, and so $\mathbf{\Sigma}$ specifies the covariance for the log-transformed response values ln(*Y*) = *Z*.

Table 2. Methods of regression analysis used in Papers I–IV.

Notation (Papers)	Fitting method	Response ~ Distribution	Model (matrix notation)	Σ matrix
ML_{LN} (I, II, III)	*ML*	$Y \sim$ Log-normal	$\boldsymbol{\mu}_Y = \boldsymbol{X\beta}$	$\sigma_Z^2\boldsymbol{I}$
ML_{LN} (IV)	*ML*	$Y \sim$ Log-normal	$\boldsymbol{\mu}_Y = \boldsymbol{X\beta}$	CS, AR(1)
WLS (I, II, III)	*WLS*	$Y \sim$ Log-normal	$\boldsymbol{\mu}_Y = \boldsymbol{X\beta}$	$\sigma_Z^2\boldsymbol{I}$
LS_{lin} (I, II)	*LS*	$Y \sim$ Normal	$\boldsymbol{\mu}_Y = \boldsymbol{X\beta}$	$\sigma_Y^2\boldsymbol{I}$
LS_{exp} (II)	*LS*	$\ln(Y) = Z \sim$ Normal	$\boldsymbol{\mu}_Z = \boldsymbol{X\delta}$	$\sigma_Z^2\boldsymbol{I}$
GLM_G (II)	*ML*	$Y_i \sim$ Gamma$(v, \boldsymbol{\mu_Y}/v)$	$\boldsymbol{\mu}_Y = \boldsymbol{X\beta}$	$diag\left(\frac{\boldsymbol{\mu_Y^2}}{v}\right)^*$
GLM_N (II)	*ML*	$\ln(Y) = Z \sim$ Normal	$\exp(\boldsymbol{\mu}_Z) = \boldsymbol{X\phi}$ $\boldsymbol{\beta} = \boldsymbol{\phi} \cdot exp(\sigma_Z^2/2)$ $\boldsymbol{\mu}_Y = \boldsymbol{X\beta}$	$\sigma_Z^2\boldsymbol{I}$
$Marg_{exp}$ (IV)	*ML*	$\ln(Y) = Z \sim$ Normal	$\boldsymbol{\mu}_Z = \boldsymbol{X\delta}$	CS, AR(1)

* diag(**V**) is a diagonal matrix with the elements of vector **V** in the diagonal.

3.1 Methods for Regression Analysis

A total of seven regression methods were used in the four papers, with a regression method being defined here as a combination of regression model and model fitting techniques. An overview of the methods is given in **Table 2**. All analyses, with the exception of those regarding the generalized linear model in Paper II, were performed using MATLAB® software (MATLAB R2012a). Analyses using methods based on generalized linear models (i.e. GLM_G and GLM_N) were performed using SAS (SAS, 2013).

The parameters in model (3.2) can be estimated using *maximum likelihood* based on the likelihood function of the log-normal distribution. This was suggested by Yurgens (2004) in a licentiate thesis. However, the method, here denoted $\boldsymbol{ML_{LN}}$, has to our knowledge not been published elsewhere. The likelihood function $L(\boldsymbol{\beta}, \sigma_Z^2, \boldsymbol{\Sigma}|\boldsymbol{x})$ is the joint probability density function of **X**, $f(\boldsymbol{x}|\boldsymbol{\beta}, \sigma_Z^2, \boldsymbol{\Sigma})$, but as a function of $\boldsymbol{\beta}, \sigma_Z^2$ and $\boldsymbol{\Sigma}$; that is, $L(\boldsymbol{\beta}, \sigma_Z^2, \boldsymbol{\Sigma}|\boldsymbol{x}) =$

$f(\boldsymbol{x}|\boldsymbol{\beta}, \sigma_Z^2, \boldsymbol{\Sigma})$. Generally, the maximum likelihood estimator of some parameters $\boldsymbol{\theta}$ is the value at which $L(\boldsymbol{\theta})$ attains its maximum as a function of $\boldsymbol{\theta}$. The likelihood function for model (3.2) can be expressed as

$$L(\boldsymbol{\beta}, \sigma_Z^2, \boldsymbol{\Sigma}) = \frac{\exp\left[-\frac{1}{2}\left(ln(\boldsymbol{Y}) - ln(\boldsymbol{X\beta}) + \frac{\sigma_Z^2}{2}\right)^T \boldsymbol{\Sigma}^{-1}\left(ln(\boldsymbol{Y}) - ln(\boldsymbol{X\beta}) + \frac{\sigma_Z^2}{2}\right)\right] \prod_{i=1}^{n_{tot}} \frac{1}{y_i}}{|\sigma|^{1/2}(2\pi)^{n_{tot}/2}},$$

where n_{tot} is the total number of observations. The estimates are derived using a Newton-Raphson iteration scheme. The estimated covariance matrix of $\boldsymbol{b}$, $\mathbf{S}_{\boldsymbol{b}}$, is obtained by the inverse of the observed Fisher's information matrix (the negative of the second derivative, i.e. the Hessian matrix, of the logarithm of the likelihood function). ML_{LN} is considered for independent observations (i.e. $\boldsymbol{\Sigma} = \sigma_Z^2 \boldsymbol{I}$) in Papers I–III, and for repeated measurements in Paper IV. In Paper IV, $\boldsymbol{\Sigma}$ is fitted with both a CS and an AR(1) pattern; see (3.4) and (3.5). The first and second derivatives of (3.2) are presented in Paper I (for independent observations) and the derivatives for repeated measurement are presented in the supplementary web material to Paper IV.

Ordinary least squares regression, here denoted $\boldsymbol{LS_{lin}}$, can be used to estimate the regression parameters in a linear model with independent observations; that is, $\boldsymbol{\Sigma} = \sigma_Z^2 \boldsymbol{I}$. In Paper I, we use LS_{lin} to estimate the parameters $\boldsymbol{\beta}$ in (3.2) as

$$\underset{(p \times 1)}{\boldsymbol{b}} = \left(\boldsymbol{X^T X}\right)^{-1}\left(\boldsymbol{X^T y}\right). \tag{3.6}$$

LS_{lin} does not assume the errors $\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{Xb}$ to be normally distributed. However, for the inference to be valid they are assumed to be independent and homoscedastic; that is, $var(e) = \sigma_Y^2$, see e.g. (Casella and Berger, 2001). The variance σ_Y^2 is estimated by the mean square error, MSE. The covariance matrix of $\boldsymbol{b}$, $\boldsymbol{\Sigma}_{\boldsymbol{b}}$, is then estimated by

$$\underset{(p \times p)}{\mathbf{S}_{\boldsymbol{b}}} = MSE \cdot \left(\boldsymbol{X^T X}\right)^{-1}.$$

A least squares method that can handle the heteroscedasticity in model defined in (3.2) is the method of *weighted least squares* (WLS), in which the estimates of β are determined so as to minimize the squared sum of weighted residuals. If the nature of the heteroscedasticity of the response Y is known, the weights, W, of the residuals are determined so that they are proportional to the inverse of $var(Y|\mathbf{X})$; that is, $W_i \propto \sigma_{Y|X}^{-2}$. The WLS approach were used to fit the parameters in (3.2) for independent, $\boldsymbol{\Sigma} = \sigma_Z^2 \boldsymbol{I}$, in Papers I–III. The variance parameter σ_Z^2 is the constant variance in log-scale, and using the relationships

specified in **Table 1** produces the following expression for the variance of *Y* given X:

$$\underset{(n\times 1)}{\boldsymbol{\sigma}_Y^2} = \exp(\sigma_Z^2 - 1)\,\underset{(n\times 1)}{(\boldsymbol{X\beta})^2}. \tag{3.7}$$

Here *n* denotes the total number of observations, since WLS is used on independent observations. From Equation (3.7) it can be seen that $(\boldsymbol{X_i\beta})^2 \propto \sigma_{Y|X}^2$, so the weights in WLS can be estimated by $\boldsymbol{W} = diag((\boldsymbol{Xb})^{-2})$. $\boldsymbol{W}$ is a diagonal matrix with the elements of vector $(\boldsymbol{Xb})^{-2}$ in the diagonal. The *b*-values can be obtained using LS_{lin}. The weighted mean squared error, MSE_W, will give an estimate of the constant part of the variance of *Y* and is the WLS estimate of σ_Z^2. MSE_W is given by

$$MSE_W = \underset{(1\times n)(n\times n)(1\times n)}{\boldsymbol{e}^T\ \boldsymbol{W}\ \boldsymbol{e}}\ /(n-p).$$

The WLS estimate of *b* is obtained as

$$\boldsymbol{b} = \left(\boldsymbol{X^T W X}\right)^{-1}\left(\boldsymbol{X^T W y}\right), \tag{3.8}$$

with the estimated covariance matrix of $\boldsymbol{b}$ given by

$$\boldsymbol{S_b} = MSE_W\left(\boldsymbol{X^T X}\right)^{-1}.$$

For independent observations, the parameters in (3.2) can also be estimated by a generalized linear model, GLM. Two GLMs were used in Paper II. The $\boldsymbol{GLM_N}$ method uses the exponential link, $g(\mu) = \exp(\mu)$, and assumes a normal distribution for the response $Z = \ln(Y)$. The exponential link is not frequently used in GLM, but is used in this case in an attempt to obtain the same estimates as in ML_{LN}, but with the use of a GLM. The estimations were done using PROC GLM in SAS. The model of GLM_N can be represented by

$$\boldsymbol{Z} \sim N(\boldsymbol{\mu_Z}, \sigma_Z^2), \text{ where } \boldsymbol{\mu_Z} = \ln(\boldsymbol{X\phi}).$$

The estimates of $\boldsymbol{\phi}$ are not interpreted directly but are used to obtain estimates of the absolute-effects $\boldsymbol{\beta}$. A comparison between $\boldsymbol{\mu_Z} = \ln(\boldsymbol{X\phi})$ and the relationship between the mean in a log-normal and normal distribution (**Table 1**) gives the following link between $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$:

$$\underset{(p\times 1)}{\boldsymbol{\beta}} = \underset{(p\times 1)}{\boldsymbol{\phi}} \cdot \exp(\sigma_Z^2/2). \tag{3.9}$$

The estimated covariance matrix of $\boldsymbol{b}$ is then given by

$$\underset{(p\times p)}{\boldsymbol{S_b}} = \boldsymbol{S_{\hat{\phi}}} \cdot \exp(\sigma_Z^2/2).$$

The $\boldsymbol{GLM_G}$ method uses the identity link and assumes a gamma distribution to approximate the log-normal distribution. The model for Y in GLM_G can be expressed as

$$\boldsymbol{\mu_Y} = \boldsymbol{X\beta}, \text{ with } Y_i \sim Gamma(k, \boldsymbol{X_i\beta}/v).$$

where k is the shape parameter of the gamma distribution. From this we get $\boldsymbol{\sigma}_Y^2 = diag((\boldsymbol{X\beta})/v)$ and $\sigma_Z^2 = \ln(1/v+1)$.

As mentioned earlier, the use of log-transformations is a common approach to handle a log-normally distributed response variable. A linear model is often assumed for the log-transformed response $Z = \ln(Y)$, so that

$$\ln(\boldsymbol{Y}) = \boldsymbol{Z} = \boldsymbol{X\delta} + \boldsymbol{\varepsilon}, \qquad \text{with } \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}). \tag{3.10}$$

For independent observations, the model in (3.10) can be estimated using ordinary least squares regression; in Paper II, this approach is denoted as $\boldsymbol{LS_{exp}}$. The model is also used for repeated measurements in Paper IV, assuming either an CS or an AR(1) pattern for $\boldsymbol{\Sigma}$. The repeated measurements models, denoted by $\boldsymbol{Marg_{exp}}$, are estimated using maximum likelihood for normal data. The likelihood function used in Marg_exp is

$$L(\boldsymbol{\delta}, \sigma_Z^2, \boldsymbol{\Sigma}) = \frac{\exp\left[-\frac{1}{2}(Z - \boldsymbol{X\delta})^T \boldsymbol{\Sigma}^{-1}(Z - \boldsymbol{X\delta})\right]}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}(2\pi)^{\frac{n_{tot}}{2}}},$$

where n_{tot} is the total number of observations.

Methods estimating the relative effects $\boldsymbol{\delta}$ in (3.10) are here called *relative-effects methods*, while methods estimating the absolute effects $\boldsymbol{\beta}$ in (3.2) are called *absolute-effects methods*.

3.2 Inference

Asymptotic inference is used in all four papers (see Section 3.2.1). In Paper III, the asymptotic inference is evaluated for smaller sample sizes ($n_{tot} < 30$) and an alternative inference for small samples, not based on asymptotic theory, is presented (see Section 3.2.2).

3.2.1 Asymptotic Inference

Statistical inference based on asymptotic criteria and approximations is called asymptotic inference. Maximum likelihood estimators have asymptotic normality given some regularity conditions, see e.g. (Casella and Berger, 2001). The LS_lin and WLS estimates of β are linear combinations of Y, as shown in (3.6) and (3.8), and a linear combination of independent normally

distributed variables is normally distributed, see e.g. (Kutner, Nachtsheim and Neter, 2004). Hence, if Y is independent and normally distributed, these estimates of β will also be normally distributed.

Hypothesis testing of the regression parameters

A Wald-type test (sometimes denoted z-test), based on the normal distribution, or the *t*-test, based on the *t*-distribution, is used for hypothesis tests of the parameters. In the applications, two-sided tests are used and a *p*-value less than 0.05 is considered significant. In the simulation studies, the type I error is evaluated for tests of H_0: $\beta = \beta_T$, where β_T is the parameter value specified in the simulation model.

Confidence intervals

Wald-type intervals are used to construct confidence intervals (CIs) for the regression parameters ($\boldsymbol{\delta}$ for the relative-effects methods LS_{exp} and $Marg_{exp}$, and $\boldsymbol{\beta}$ for the absolute-effects methods). A Wald-type interval can be expressed as $\hat{\theta} \pm C \cdot se(\hat{\theta})$ where C is the $100\cdot(1-\alpha/2)$ percentile of a symmetric reference distribution. In this case the reference distribution is either the normal distribution, $z_{1-\alpha/2}$, or a *t*-distribution, $t_{1-\alpha/2,df}$, where df denotes the degrees of freedom.

For variance parameters we use χ^2-distribution based intervals, and a 95% CI for σ_Z^2 is then given by $\left(df \cdot s_z^2/\chi^2_{0.025;df}\right) < \sigma^2 < \left(df \cdot s_z^2/\chi^2_{0.975;df}\right)$, where *df* are the degrees of freedom. As an approximation of the degrees of freedom, we use $\widehat{df} = 2s_Z^4/var(s_Z^2)$. When estimating σ_Z, rather than σ_Z^2, and only $var(s_Z)$ are available, the variance for s_Z^2 can be approximated by $var(s_Z^2) \cong 4s_Z^2 \cdot var(s_Z)$.

The Fisher Z transformation $\tanh^{-1}(\rho)$ is used to construct a CI for ρ. The transformation is assumed to be normally distributed, and the lower and upper limit (LL and UL) can be found as $\tanh^{-1}(r) \pm z_{1-\frac{\alpha}{2}}(1-r^2)^{-1}se(r)$. A $100\cdot(1-\alpha)$% CI for ρ is then given by $\tanh(LL) < \rho < \tanh(UL)$.

The formula for the CI for $\mu_{Y|X}$ differs between the models. For absolute-effects methods that give direct estimates of β (i.e. ML_{LN}, LS_{lin}, $Marg_{exp}$, WLS, and GLM_G), standard Wald intervals are used and a $100(1-\alpha)$% CI for $\mu_{Y|X}$ is calculated as $\hat{\mu}_{Y_i|\boldsymbol{X}_i} \pm z_{\alpha/2} \cdot se\left(\hat{\mu}_{Y_i|\boldsymbol{X}_i}\right)$. The sample-specific standard error is estimated as $se\left(\hat{\mu}_{Y|\boldsymbol{X}_i}\right) = \left(\boldsymbol{X}_i^T \boldsymbol{S}_b \boldsymbol{X}_i\right)^{1/2}$, where $\boldsymbol{X}_i$ is the predictor value for the *i*:th observation. For the GLM_N estimates, the $100(1-\alpha)$% CI for $\mu_{Y|X}$ is calculated as

$$\left(\exp\left(\hat{\mu}_{Z_i|\boldsymbol{X}_i}\right) \pm z_{\alpha/2} \cdot se\left(\hat{\mu}_{Z|\boldsymbol{X}_i}\right)\right) \exp(s_Z^2/2),$$

where s_Z^2 is the estimate of σ_Z^2, $var\left(\exp\left(\hat{\mu}_{Z_i|\boldsymbol{X}_i}\right)\right) = \boldsymbol{X}_i^T \boldsymbol{S}_{\hat{\phi}} \boldsymbol{X}_i$, and $\boldsymbol{S}_{\hat{\phi}}$ is the covariance matrix for the estimates of $\boldsymbol{\phi}$ in (3.9). The form of the 100(1-α)% CI for $\mu_{Y|X}$ for the log-linear models, estimated by LS_{exp} and $Marg_{exp}$, is based on Cox's method (Land, 1972):

$$\exp\left(\hat{\mu}_{Z|X} + s_Z^2/2 \pm z_{\alpha/2}\sqrt{var\left(\hat{\mu}_{Z_i|X}\right) + s_Z^4/(2 \cdot df)}\right),$$

where df denotes the degrees of freedom for s_Z^2 ($df = n_{tot} - p - 1$ for independent observations) and $var\left(\hat{\mu}_{Z_i|\boldsymbol{X}_i}\right) = \boldsymbol{X}_i^T \boldsymbol{S}_d \boldsymbol{X}_i$, where $\boldsymbol{S}_d$ is the covariance matrix for the estimates of $\boldsymbol{\delta}$ in (3.10).

Prediction intervals

A prediction interval for a new observation Y^* from a new individual, based on a linear model with a log-normal response, is created as a symmetric Wald-type interval in log-scale: $\hat{Z}^* \pm C \cdot se\left(\hat{Y}^*\right)$, which is then retransformed back to the original. However, since $\hat{Z}^* = \log\left(\hat{Y}^*\right) - s_Z^2/2$, we also have to take into account that s_Z is an estimate. The 100 (1 – α) % prediction intervals for Y^* can be approximated by

$$\exp\left(\log\left(\hat{Y}^*\right) - \frac{s_Z^2}{2} \pm t_{n_{tot}\text{-}p,1-\frac{\alpha}{2}}\sqrt{\frac{var(\hat{\mu}_{Y^*})}{\hat{\mu}_{Y^*}^2} + s_Z^2 var(s_Z) - \frac{2 s_Z cov(\hat{\mu}_{Y^*}, s_Z)}{\hat{\mu}_{Y^*}} + s_Z^2}\right),$$

where $\hat{Y}^* = \boldsymbol{X}^* \boldsymbol{b}$, s_Z is the estimate of σ_Z, df denotes the degrees of freedom for s_Z^2, and $cov\left(\hat{Y}^*, s_Z\right) = cov(b_0, s_Z) + \sum_{i=1}^{p-1} X_i cov(b_i, s_Z)$. The prediction interval for methods that assume a log-linear model, as in (3.10), is given by $\exp\left(\hat{Z}^* + \frac{s_Z^2}{2} \pm t_{n_{tot}\text{-}p,1-\frac{\alpha}{2}}\sqrt{var\left(\hat{Z}^*\right) + \frac{s_Z^4}{2 \cdot df} + s_Z^2}\right)$.

3.2.2 Small-Sample Inference

In practice, it is common with smaller sample sizes. Simulation studies have shown that asymptotic inference on small datasets typically leads to substantial underestimation of variance, an inflated type I error for hypothesis testing (Kenward and Roger, 1997; Munro and Wixley, 1970; Zhou, Gao and Hui, 1997), and liberal confidence intervals, see e.g. (Zhou and Gao, 1997). In Paper III, the method of bootstrapping was used to get more reliable results for small-sample confidence intervals and hypothesis testing of β.

Bootstrapping is a resampling method (Efron (1979) used to estimate the variance and bias of an estimate. The r bootstrap samples can be obtained directly by resampling with replacement from the observed data, **Y** and **X**, or through the direct estimates $\boldsymbol{b}$, $\boldsymbol{S}_b$, and s_z calculated from **Y** and **X**.

Estimates of $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_b$, and σ_z are then calculated for each of the r bootstrap samples. The j:th bootstrap estimates are denoted $\boldsymbol{b}_j^*$, $\boldsymbol{S}_b^*$, and s_Z^*, the variances of $\boldsymbol{b}_j^*$ (i.e. the diagonal elements of $\boldsymbol{S}_b^*$) are denoted $var_j(\boldsymbol{b}^*)$, and the standard errors are denoted $se_j(\boldsymbol{b}^*)$. The distribution of the r bootstrap estimates is used for testing hypotheses, estimating confidence intervals, and estimating the bias.

Bootstrapping approaches

There are different types of bootstrapping, including nonparametric, semi-parametric, and parametric bootstrapping. Paper III uses parametric and semi-parametric approaches.

In the parametric bootstrap approach, the response is assumed to have a log-normal distribution and the bootstrap samples are simulated on the basis of the direct estimates, $\boldsymbol{b}$ and s_Z, and the assumed model. That is, the observations $\mathbf{y}^*$ of the bootstrap samples are simulated according to $\ln(\boldsymbol{y}^*) = \ln(\boldsymbol{Xb}) - s_Z^2/2 + \boldsymbol{e}^*$ with $\boldsymbol{e}^* \sim N(0, s_Z^2\boldsymbol{I})$, where **X** is the observed predictor.

In the semi-parametric bootstrapping approach, the bootstrapping sampling is nonparametric, but a linear model with a log-normal response is still assumed when calculating the bootstrap estimates.

Hypothesis testing of β

In the bootstrap t-test (bt-test) of the hypothesis H_0:β=β_T, the value of the direct test statistic $t_{dir} = (b - \beta_T)/se(b)$ is compared to the percentiles of the distribution of $t_j^* = (b_j^* - b)/se_j(\boldsymbol{b}^*)$, see e.g. (Fox, 1997).

In Paper III, the behavior of the tests under H_0 is evaluated. The *α-sensitivity* of a test is defined as the probability of rejecting H_0 when the estimate b is extreme according to the true distribution under H_0 (i.e. when the β-estimate is smaller than the $100{\cdot}\alpha$ percentile or larger than the $100(1-\alpha)$ percentile). Further, we define the *α-specificity* as the probability that the test does not reject H_0 when the estimate b is moderate according to the true distribution under H_0 (i.e. when the β-estimate is between the $100{\cdot}\alpha$ and $100(1-\alpha)$ percentiles). The α-sensitivity and α-specificity are used to evaluate how well the test statistics follow the distribution of the β-estimates, given that H_0 is

true. The true distribution and percentiles of the estimates b, under H_0: $\beta=\beta_T$, are determined on the basis of 3 000 000 simulated samples.

Confidence intervals for β

There are several suggested methods for constructing a confidence interval using bootstrap methodology. In a bootstrap-*t* interval (here denoted *boot-t*) the percentiles of t_j^* are used such that

$$\left(b + t^*_{(r\cdot\alpha/2)} \cdot se^*(b^*),\ b + t^*_{(r\cdot(1-\alpha/2))} \cdot se^*(b^*)\right).$$

Unlike the asymptotic theory interval (*tdist*), the *boot-t* interval is not necessarily symmetric and can to some extent mimic a possible skewness in the distribution of b. For independent observations, the *boot-t* interval has been shown to have second-order accuracy. This means that the error of the limits is of order $O(n^{-1})$; that is, if $\mathrm{LL}_{\text{boot-t}}(\alpha/2)$ denotes the lower limit of a 100(1-α)% *boot-t* interval of β, then $\mathrm{P}[\beta \leq \mathrm{LL}_{\text{boot-t}}(\alpha/2)] \leq \alpha/2 + O(n^{-1}) = \alpha/2 + c \cdot n^{-1}$, where n is the sample size and c is some constant. However, unlike *tdist*, *boot-t* is not transformation-respecting and can also be sensitive to outliers in the data (DiCiccio and Romano, 1995; Efron and Tibshirani, 1994). The bootstrap percentile interval (here denoted *PCI*) is obtained from the percentiles of the empirical distribution of the bootstrap estimates b_j^*, see e.g. (Fox, 1997):

$$\left(b^*_{(r\cdot\alpha/2)},\ b^*_{(r\cdot(1-\alpha/2))}\right).$$

In comparison to the bootstrap-*t* interval, *PCI* is less sensitive to outliers and hence is often considered more reliable (Efron and Tibshirani, 1994). The *PCI* has second-order accuracy, and unlike *boot-t* is transformation-respecting. Heteroscedasticity might lead to a skewed distribution, and if the distribution is skewed, the *PCI* tends to be too narrow (Efron, 1982). In the bootstrap *bias-corrected accelerated percentile interval* (BC_a), the limits of the percentile interval are corrected for bias and skewness (DiCiccio and Efron, 1996; Fox, 1997). The BC_a interval uses the correction factors $\hat{z}_0$(correcting for bias) and $\hat{a}$ (correcting for skewness by allowing different variances for the estimates):

$$\left(b^*_{(r\cdot z[\alpha/2])},\ b^*_{(r\cdot[1-\alpha/2])}\right),$$

where $z[\alpha] = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0+z_\alpha}{1-\hat{a}\cdot(\hat{z}_0+z_\alpha)}\right)$. If $\hat{z}_0 = \hat{\alpha} = 0$, then $\mathrm{BC_a}$ will give the same limits as *PCI*. The BC_a interval has been shown to be transformation-respecting and to have second-order accuracy.

3.3 Evaluating Goodness of Fit

An important step of any regression analysis is to evaluate how well the fitted model fits the dataset. Suitable methods of evaluation include residual analysis and goodness of fit statistics.

Residual analysis

A major assumption in this thesis is that the response Y is log-normally distributed, given the predictor values. This implies that if the residuals, e, are given in log-scale, they should be normally distributed:

$$\boldsymbol{e} = \ln(\hat{Y}) - s_Z^2/2 - \ln(Y) \text{ and } \boldsymbol{e} \sim N(\mathbf{0}, \boldsymbol{\Sigma}).$$

A visual examination of residual plots, such as quantile-quantile plots or histograms, can give an initial test of this assumption. The use of residual plots to assess the distribution has the advantage of being easy to present, regardless of sample size, but also has the major disadvantage that assessment of the correctness of the distribution assumption becomes highly subjective.

A more objective approach to assess the distribution of e is to use a normality test. There are a number of normality tests available, but like many statistical tests, distribution tests often have low power for small datasets and high power for large sets. Hence, they rarely reject the distribution of the null hypothesis for small sample sizes but often reject it for larger sample sizes. The Shapiro-Wilk W-test usually has a high power compared to many other normality tests (Razali and Wah, 2011; Yazici and Yolacan, 2006), which can help in the case of small and moderate sample sizes, but the problem of rejecting the normality assumption for large datasets remains.

If the model fits the data, the residuals are assumed to be random and homoscedastic. One approach to checking the randomness assumption for e could be to use a residual regression model, $e_i = \beta_0' + \beta_1' \cdot \hat{Y}_i + \beta_2' \cdot \hat{Y}_i^2$. If e is random, an F-test of the residual regression model should be non-significant and all β_1' estimates should be small and also non-significant. This model can also be used to check for heteroscedasticity using the White test (White, 1980). This residual model can be fitted by LS_{lin} for non-repeated measurements and by a linear mixed model with a subject-specific random intercept for repeated measurements.

Goodness of fit measures

A popular measure of fit is the coefficient of determination, denoted R^2. In LS_{lin}, R^2 is simply the proportion of variance explained by the model: $R^2 = 1 - SS_{tot}/SS_{res}$ where $SS_{tot} = \sum_i^n (y_i - \bar{y})^2$ is the total sum of

squares, which is proportional to *var*(*Y*), and $SS_{res} = \sum_i^n (y_i - \hat{y}_i)^2$ is the residual sum of squares. Since sum of squares can be strongly affected by outliers, this R^2 is more valid as a measure of fit if the errors are homoscedastic. Hence, for a model with a log-normal response we choose to let R^2 denote the variance explained in log-scale. In a repeated measurement situation (i.e. mixed models) there are various possible R^2s, including the likelihood based R^2 and the Wald R^2 (Kramer, 2005). All these will give the same results as $R^2 = 1 - SS_{tot}/SS_{res}$ when applied to independent observations. When using mixed models with a random intercept, the proportion of variance explained can be divided into between, within, and total variance explained, see e.g. (Nakagawa and Schielzeth, 2013). Paper IV uses the total variance explained in the log-scale: $R_Z^2 = 1 - s_Z^2/s_{z0}^2$, where s_{z0}^2 is the estimated variance of the null model.

4 MATERIAL

The methods were evaluated using real data, presented in Section 4.1, and large-scale simulations using simulation models, presented in Section 4.2.

4.1 Data

The datasets were used either to create a realistic simulation model or to illustrate the properties of the methods. None of the papers were aimed at giving a complete statistical model for the response, but rather the datasets were used to compare the statistical methods presented in Section 3.1. A list of the datasets used in each paper is given in **Table 3**.

Table 3. Real-life datasets used in Papers I–IV.

Data	Paper I	Paper II	Paper III	Paper IV
Personal exposure to $PM_{2.5}$ particles	X	X	X	
Personal exposure to 1,3-butadien and NO_2	X			X
Abdominal adiposity and biomarkers of inflammation and insulin resistance		X		
Personal exposure to benzene			X	
Creatinine-adjusted cadmium in urine				X

4.1.1 Personal Exposure to $PM_{2.5}$ Particles

The simulation models in Papers I–III were based on a dataset including the concentration ($\mu g/m^3$) of personal exposure to airborne particles with a diameter of less than 2.5 μm ($PM_{2.5}$ particles) in the city of Gothenburg, Sweden. A detailed description of these data is given elsewhere (Johannesson et al., 2007). $PM_{2.5}$ particles are small enough to bypass the respiratory defenses and enter into the lungs. The dataset contains personal exposure, collected over one week in 2000, for 30 persons living in Gothenburg; 20 were randomly selected from the population register and 10 were recruited among the employees at the Department of Occupational and Environmental Medicine in Gothenburg.

4.1.2 Personal Exposure to 1,3-Butadien and NO_2

A set of data including personal exposure to certain carcinogenic substances and nitrogen dioxide (NO_2) among the general population in five Swedish

cities was used to illustrate the properties of the methods compared in Papers I and IV. The measurements were taken on a yearly basis in one of five different Swedish cities according to a rotating study plan, and each city had data available for either one year or two years: Gothenburg for 2000 and 2006, Umeå for 2001 and 2007, Stockholm for 2002, Malmö for 2003 and 2008, and Lindesberg for 2005. However, NO_2 levels were not measured in Gothenburg. Background data were collected via daily activity diaries in which the participants recorded time spent in different microenvironments: outdoors in traffic, outdoors at the workplace, outdoors elsewhere, indoors at home, indoors at the workplace, and indoors elsewhere. The participants also recorded time exposed to environmental tobacco smoke (ETS) and time spent in homes during burning of wood or pellets. The datasets are described in detail elsewhere (Hagenbjork-Gustafsson et al., 2013).

The response of interest in Paper I was 1,3-butadiene. The original dataset contained 1,3-butadiene exposures for 275 individuals. The predictors of interest were proportion of time spent in traffic, city of residence (Lindesberg, Malmö, Stockholm, Gothenburg, or Umeå), exposure to burning of wood or pellets (yes/no), and smoking habits (exposed to ETS, non-smoker, or current smoker). Of the 275 individuals, 268 had measurements on all the predictors of interest. The dataset contained repeated measurements, but only the first measurement was used in Paper I.

The response of interest in Paper IV was personal exposure to NO_2. The dataset contained NO_2 measurements from 233 individuals living in Sweden. The potential determinants were city of residence (Lindesberg, Malmö, Stockholm, or Umeå), year (2001, 2002, 2003, 2007, or 2008), occupational exposure (yes/no), proportion of the time spent in their own home, use of a gas stove (yes/no), and use of oil heating (yes/no). City and year were included as an interaction, since measurements were only performed in one city for each year. Of the 233 individuals, 209 were included in the analysis; 98 of these (47%) had two measurements per individual, while the rest had only one measurement each.

4.1.3 Personal Exposure to Benzene

Paper III used ML_{LN} and WLS to assess which determinants were associated with personal exposure to benzene ($\mu g/m^3$) in the general population in Gothenburg. Personal measurements were accumulated over a six-day period for 40 randomly selected individuals, 35 of whom had measurements for all predictors of interest in Paper III. Half of these individuals had two measurements, but only the first one was used in Paper III. The dataset is further described elsewhere (Sallsten, Ljungkvist and Barregard, 2003).

The predictors of interest in Paper III were method of heating (1 = oil, 0 = electricity or central heating), smoking status (1 = current smoker, 0 = non-smoker), home exposure (measurements from bedroom multiplied by proportion of time spent in residence), time spent in cars/buses (hours), time spent indoors but not at home (hours), and time spent outdoors (hours).

4.1.4 Abdominal Adiposity and Biomarkers of Inflammation and Insulin Resistance

Paper II used ML_{LN}, WLS, LS_{lin}, LS_{exp}, GLM_G, and GLM_N to quantify the association between abdominal adiposity and biomarkers of inflammation and insulin resistance. The data were collected in a study known as DIWA (Diabetes and Impaired glucose tolerance in Women and Atherosclerosis). The DIWA dataset is drawn from a population based cohort of 64-year-old women and living in Gothenburg at the time of the screening process. All women were 64 years old at the time of the examination (and born between 1937 and 1940). All eligible women were invited to the screening examination. Glucose tolerance was defined for each participant according to the World Health Organization (WHO) criteria. The participants were divided into three sub-groups according to metabolic status: diabetes mellitus (DM, n_{tot} = 220), impaired glucose tolerance (IGT, n_{tot} = 204), or normal glucose tolerance (NGT, n_{tot} = 188). A random sample of participants was recruited from each group for a longitudinal follow-up study. The baseline examination was performed in 2001-2004. A detailed description of the baseline examination is given elsewhere (Brohall et al., 2006).

The potential predictors in Paper III were smoking (current smoker = 1, non-smoker = 0), physical activity, waist circumference (measured in decimeters), insulin resistance (measured by a variable named HOMA-IR) and glucose tolerance (DM, IGT, or NGT).

4.1.5 Urinary-Cadmium Biomarker

Paper IV used ML_{LN} and $Marg_{exp}$ to identify important determinants for excretion of urinary cadmium (U-Cd). The dataset contained spot urine samples from 26 healthy non-smoking residents of Gothenburg (14 men and 12 women). Samples from two of the women were considered incomplete and were excluded from the study, so the final analysis used data from 24 participants (14 men and 10 women). The samples were collected at six fixed time points during a 24-hour period. The original study used measurements from six days; that is, 6×2 measurements per individual. However, Paper IV only used measurements from the first day. The determinants used were age,

gender, time point of spot sample, and urinary flow rate (UF, mL/h). The dataset are described in detail in Akerstrom et al. (2013).

4.2 Simulation Models

The simulation models were based on the real-life dataset of personal exposure to either $PM_{2.5}$ particles (described in Section 4.1.1) or benzene (described in Section 4.1.3).

In these simulation models, the outcome was personal exposure and the expected outcome was assumed to be a linear combination of one or more predictors, X_i:

$$E[Y|X] = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k.$$

The observations were then simulated according to

$$\ln(Y_i) = \ln(E[Y|X]) - \sigma_Z^2/2 + e_i,$$

where $e_i \sim N(0, \sigma_Z^2)$. The parameter values used in the simulations are presented in **Table 4**. The predictors used were smoking status (*Smoker*, 1 = Yes, 0 = No), number of cigarettes per day (*Smoke*), number of hours spent in their own home (*Home*), residential outdoor concentration of $PM_{2.5}$ (*ConcOut*, μg/m^3), and proportion of time spent in traffic (*Traffic*).

In order to facilitate interpretation and comparison without the introduction of unnecessary variation, balanced design matrices were used in the simulations. For $PM_{2.5}$, the design matrix was created so that each predictor had 3 levels (*ConcOut* = 2, 8, or 14 units; *Smoke* = 0, 7, or 14 cigarettes/day; and *Home* = 8, 16, or 24 hours/day), and every combination of predictor was equally frequent. In the Paper IV models with benzene as response, the predictors were smoking status (X_{Smoker}) and proportion of time spent in traffic ($X_{Traffic}$). These models were used to simulate repeated measurements per individual; the *Smoker* predictor was kept constant within each individual while *Traffic* varied. Half of the subjects were assumed to be smokers, and the traffic observations were simulated from a uniform distribution $X_{Traffic} \sim$ U(0,0.3).

Table 4. Models used in the simulation studies.

Paper	Y	Model	Σ_Z	ρ
I	$PM_{2.5}$	$\mu_Y = 4.803 + 0.574 \cdot ConcOut$	0.354	-
I	$PM_{2.5}$	$\mu_Y = 0.761 \cdot ConcOut + 2.092 \cdot Smoke + 0.218 \cdot Home$	0.450	-
II, III	$PM_{2.5}$	$\mu_Y = 1.564 + 0.122 \cdot Smoke + 0.075 \cdot ConcOut$	0.383	-
IV	Benzene	$\mu_Y = 1.42 + 1.15 \cdot Smoker + 9.00 \cdot Traffic$	0.707	0.36

5 RESULTS

The results are presented separately for applications (Section 5.1) and for simulation studies (Section 5.2).

5.1 Applications

5.1.1 Personal Exposure to 1,3-Butadiene

Potential determinants for personal exposure to 1,3-butadiene (μg/m^3) were assessed in Paper I using ML_{LN}, WLS, and LS_{lin}. The three methods produced similar estimates for regression coefficients, but different standard errors and hence different confidence intervals (CIs); see **Figure 3**. The only predictors shown as significant by all three methods were city of residence, with Lindesberg as reference, and smoking habits, with non-smokers that had not been exposed to ETS as reference. The expected exposure for residents of Gothenburg was about 0.5 μg/m^3 lower than that of the residents of Lindesberg, and smokers had an increased expected exposure of between 0.6−0.7 μg/m^3 compared to the non-smokers.

The ML_{LN} produced the narrowest CIs and LS_{lin} generally the widest, with the exception of the coefficient for smoking, where LS_{lin} had the narrowest CI and WLS the widest. The 95% CIs for the regression parameters did overlap between the methods, but the LS_{lin} intervals were 37–91% wider than those of ML_{LN} (except for the estimate for smoking), and the WLS intervals were 26–43% wider than those of ML_{LN}. As a consequence there was a difference in significant predictors. ML_{LN} showed a significantly lower exposure in Gothenburg, Malmö, and Umeå compared to Lindesberg, whereas WLS and LS_{lin} only showed a significant difference between Gothenburg and Lindesberg. ML_{LN} also showed a significant increase in exposure levels among current smokers and non-smokers exposed to ETS compared to non-smokers who were not exposed to ETS, whereas WLS and LS_{lin} showed significantly increased levels among smokers (but not for non-smokers exposed to ETS).

All methods indicated an increasing effect of time spent in traffic with about 0.01 μg/m^3 per percentage point, but the effect was not significant ($p = 0.123$ for WLS, $p = 0.290$ for LS_{lin}, and $p = 0.054$ for ML_{LN}).

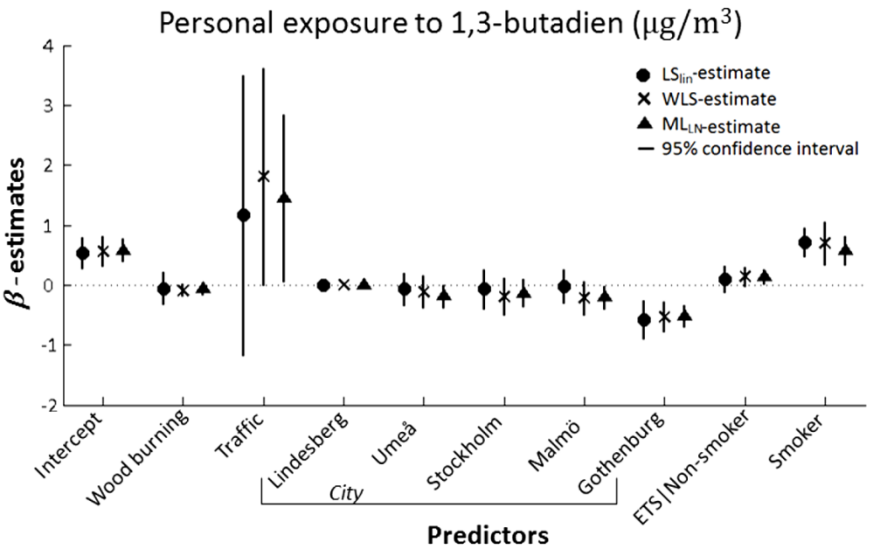

Figure 3. Estimates of β_i, and their 95% confidence intervals for regression analysis with 1,3-butadiene as the response. Estimates made using ordinary least squares regression (LS_{lin}*), least squares regression (WLS), and maximum likelihood regression with a likelihood based on the log-normal distribution (*ML_{LN}*) (adapted from Figure 3 in Paper 1).*

5.1.2 Small-Sample Data on Personal Exposure to Benzene

Potential determinants for personal exposure to benzene (µg/m^3) were assessed in Paper III, based on data from 35 persons. Because of the small sample, bootstrapping methods were used in the inference for ML_{LN} and WLS, in addition to the asymptotic inference.

The bias-corrected estimates b^*_{BC} were similar to the direct estimates *b*. There were only minor differences between ML_{LN} and WLS in point estimates of β. However, in cases where there was a difference, the WLS based *b* tended to have a larger deviation from zero; that is, $|b_{WLS}| > |b_{ML}|$. The WLS based *se*(*b*) were 27−40% larger than the ML_{LN} based *se*(*b*) and the WLS based intervals were wider than the ML_{LN} based intervals.

In comparison to the ML_{LN} based intervals, the intervals based on WLS were 23−39% wider for *tdist*, 14−33% wider for *boot-t*, and 25−33% wider for BC_a.

A comparison between the asymptotic theory based interval *tdist* and the parametric bootstrap intervals *boot-t*^{P} and BC_a^P showed that the *boot-t*^{P} intervals were the widest and the BC_a^P intervals were often the narrowest. The *t-boot* intervals were 2−16% wider than the asymptotic theory based interval *tdist*. The BC_a intervals were between 5% narrower and 7% wider than *tdist*.

The only predictors that were shown as significant by all 95% CIs were the number of hours spent in cars and buses and the estimated home exposure. The personal exposure was expected to increase by around 1.2 μg/m^3 per μg/m^3 of home exposure, and by about 0.06 μg/m^3 per hour spent in cars or buses.

The asymptotic theory based *tdist* intervals revealed that smoking habits, total time spent in car or bus, and home exposure were all significant predictors for an increased benzene exposure, both for ML_{LN} and WLS. For ML_{LN}, these three determinants were also significant based on the parametric bootstrap intervals, *boot-t*^{P} and BC_a. However, the WLS based *boot-t*^{P} interval did not show smoking as a significant predictor.

5.1.3 Repeated Measurements of Personal Exposure to NO_2

Potential predictors for personal NO_2 exposure were assessed in Paper IV using the methods ML_{LN} and $Marg_{exp}$.

The ML_{LN} and $Marg_{exp}$ estimates had the same sign and the same significant predictors. However, the ML_{LN} method estimates absolute effects, β, while the $Marg_{exp}$ estimates relative effects, δ. All predictors were significant but none of the categories of *city* and *year* were significantly different from the reference category, Umeå in the year 2001. Both methods showed *city* as a significant predictor, with higher exposure levels for Malmö in the year 2003 (7.5 μg/m^3 with ML_{LN} and 82% with $Marg_{exp}$) and Stockholm in the year 2002 (6.4 μg/m^3 with ML_{LN} and 65% with $Marg_{exp}$). Also, both methods showed a significant increase in expected NO_2 exposure in Umeå between the years 2001 and 2007; the estimated increase was 4.1 μg/m^3 with ML_{LN} and 47% with $Marg_{exp}$.

The estimated exposures for persons living in Umeå, expressed as a function of time spent in their own home, are presented in **Figure 4**. $Marg_{exp}$ and ML_{LN} gave very similar exposure estimates for persons who spent 70% of the time in their own homes (*Home* ≈ 0.7) and had no additional risk factors (i.e. no occupational exposure, no oil heater, and no gas stove). Such a person would have an expected NO_2 exposure of 8.9 μg/m^3, and the 95% CI would be [7.7; 10.1] using ML_{LN} and [7.7; 10.2] using $Marg_{exp}$. If this reference person was compared to a similar person who also had an occupational exposure, the expected difference would be 2.7 μg/m^3 using ML_{LN} and 2.2 μg/m^3 using $Marg_{exp}$. If our reference person was compared to a similar person who heated their home with oil, the expected difference would be −2.7 μg/m^3 with ML_{LN} and −2.3 μg/m^3 with $Marg_{exp}$. Finally, if the reference person was compared to a similar person using a gas stove, the expected difference in exposure would

be 6.7 μg/m^3 with ML_{LN} and 3.4 μg/m^3 with $Marg_{exp}$. To examine the estimated effects of proportion of time spent at home, we compared the estimated exposures of three persons living in Umeå in the year 2001 and who had no additional risk factors. If these three persons spent 70%, 80%, and 90% of the time in their own homes, their expected exposure would be 8.9, 8.3, and 7.6 μg NO_2/m^3 based on the ML_{LN} estimates and 8.9, 8.4, and 8.0 μg NO_2/m^3 based on the $Marg_{exp}$ estimates.

The linear model (estimated by ML_{LN}) had a slightly larger coefficient of determination than the log-linear model (estimated by $Marg_{exp}$); $R_Z^2 = 0.46$ versus $R_Z^2 = 0.44$.

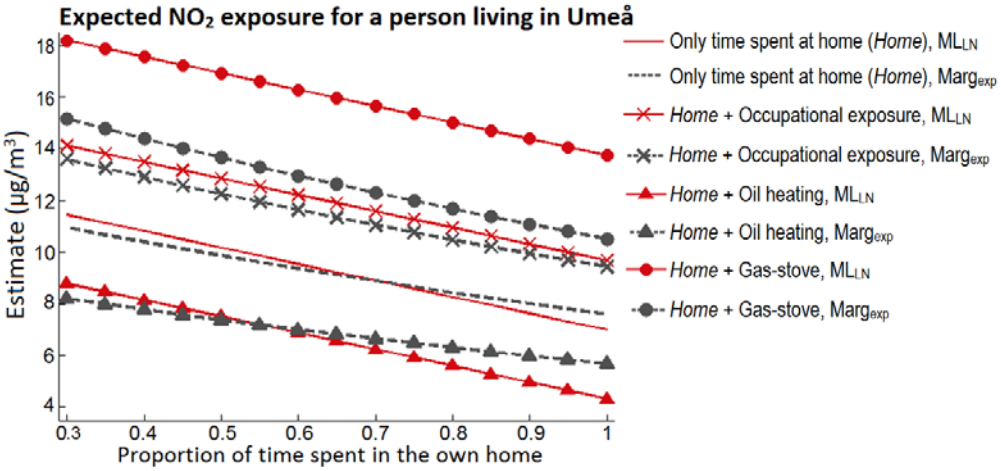

Figure 4. Expected personal exposure to NO_2 for a person living in Umeå, estimated with ML_{LN} and $Marg_{exp}$.

5.1.4 Associations between Abdominal Adiposity and Biomarkers of Inflammation and Insulin Resistance

The association between abdominal adiposity, measured by waist circumference (WC), and CRP (biomarker of inflammation) and HOMA-IR (biomarker of insulin resistance) was estimated in Paper II using five different methods: LS_{lin}, WLS, GLM_G, GLM_N, ML_{LN}, and LS_{exp}. The final models were determined by backward elimination using ML_{LN}.

Among the absolute-effects methods (i.e. all but LS_{exp}), the LS_{lin} based estimates had the widest confidence intervals for β, while the maximum likelihood methods (GLM_G, GLM_N, and ML_{LN}) had the narrowest intervals. All methods showed a significant association between CRP and WC. The β estimates of the methods WLS, GLM_G, GLM_N, and ML_{LN} were similar, while

the LS_{lin} based estimates differed. The expected absolute increase in CRP was about 1 mg/L (between 0.74 and 1.07 mg/L) for every 10 cm of WC. Using the relative-effects method, LS_{exp}, the expected increase was 49% for every 10 cm of WC. All methods also indicated a positive association between HOMA-IR and CRP. However, the association was not significant for LS_{lin} and was very high for GLM_G and WLS (0.41 and 0.42, respectively). The expected increase in CRP was between 0.12 and 0.42 mg/L for every unit increase of HOMA-IR in the absolute-effects methods and 3% per unit of HOMA-IR for LS_{lin}.

All methods found a positive association between HOMA-IR and WC in all glucose tolerance groups. Women with DM had a significantly stronger association with WC than women with NGT, and this was significant for all methods. The results also indicated a stronger association with WC for women with IGT and DM, compared to women with NGT; the interaction term WC·DM was significant for all absolute-effects methods, and the interaction term WC·IGT was significant for all absolute-effects methods except LS_{lin}. Among the absolute-effects methods, HOMA-IR was expected to increase by 0.64–1.00 per 10 cm WC for women with DM, 0.42−0.74 for women with IGT, and 0.39–0.70 for women with NGT. The relative-effects method showed an expected increase in HOMA-IR of 39% per 10 cm for women with DM, 31% for women with IGT, and 27% for women with NGT.

A comparison between the linear model for CRP, estimated by ML_{LN}, and the log-linear model for CRP, estimated by LS_{exp}, showed a slightly larger coefficient of determination for the linear model; R_Z^2 = 0.22 versus R_Z^2 = 0.21. The LS_{exp} estimate had a significant residual regression model, indicating a non-random residual. The residuals of ML_{LN} were however considered as random. A predicted versus observed plot showed that, especially for lower values of CRP, the linear model had a better fit.

A corresponding comparison between the linear model for HOMA-IR and the log-linear model for HOMA-IR showed a slightly larger coefficient of determination for the log-linear model; R_Z^2 = 0.48 versus R_Z^2 = 0.46. The ML_{LN} estimate had a significant residual regression model (i.e. non-random residual). The residual regression model was not significant for LS_{exp}. However, the residuals for both models had a skewed distribution, $\gamma < -0.7$, and showed signs of heteroscedasticity. A plot of predicted versus observed values showed that, especially for higher values of HOMA-IR, the log-linear model had a better fit.

5.1.5 Repeated Measurements of Creatinine-Adjusted Urinary Cadmium

The methods ML_{LN} and $Marg_{exp}$ were used in Paper IV to assess the possible predictors for the excretion of urinary cadmium (U-Cd). To simplify the comparison between the two methods, we defined a reference person as being a 40-year-old man with a urinary flow rate of 70 mL/h (the median values for men in our data). For this reference man, ML_{LN} estimated the expected level to be 0.10 µg cadmium/g creatinine (95% CI: 0.07; 0.13), while $Marg_{exp}$ estimated it to be 0.08 µg cadmium/g creatinine (95% CI: 0.06; 0.09).

All the ML_{LN} and $Marg_{exp}$ estimates had the same signs, but not the same significant predictors. Both ML_{LN} and $Marg_{exp}$ showed the time of day as important: both methods had the highest estimate for *time* = 9:30 and the lowest for *time* = 22:00, both significantly different from the reference category *time = ON* (overnight). Both methods showed a significant decrease for *time* = 17:30 and a non-significant increase for *time = 14:30,* when compared to *ON*. Both methods also showed the interaction between *UF* and *gender* as not significant. All other factors were significant for $Marg_{exp}$, while neither *gender* nor *time* = 12:00 were shown as significant by ML_{LN}. The correlation between observations from the same person was high (0.75 and 0.71 for ML_{LN} and $Marg_{exp}$ respectively), which could indicate that the between-subject variance was dominant.

Both methods showed the cadmium levels to increase with age; for ML_{LN}, an expected increase of 0.002 µg cadmium/g creatinine per year of age, and for $Marg_{exp}$, an increase of 3% per year of age. To compare the ML_{LN} and $Marg_{exp}$ estimates, consider an example with samples of overnight spot urine from three men aged 30, 40, and 50, each with a urinary flow of 70 mL/h. The expected cadmium levels of these three men would be 0.08, 0.09, and 0.11 µg cadmium/g creatinine when estimated with ML_{LN}, and 0.06, 0.08, and 0.10 µg cadmium/g creatinine when estimated with $Marg_{exp}$. ML_{LN} gave higher expected values for men in the interval 30-50 years of age. However, as seen in the left side of **Figure 5**, $Marg_{exp}$ gave higher values for the oldest men (at age 60).

Both methods gave higher estimates for women than for men (**Figure 5**). However, the gender difference was only significant for $Marg_{exp}$. The difference was constant for ML_{LN} estimates (0.01 µg cadmium/g creatinine), while the difference for the $Marg_{exp}$ estimates increased with increasing age (see the left side of **Figure 5**).

The log-linear model (estimated by Marg_{exp}) had a larger coefficient of determination than the linear model (estimated by ML_{LN}), R_Z^2=0.53 versus R_Z^2=0.42. A plot of observed versus predicted values with the two models also showed the log-linear model as a better fit for the data.

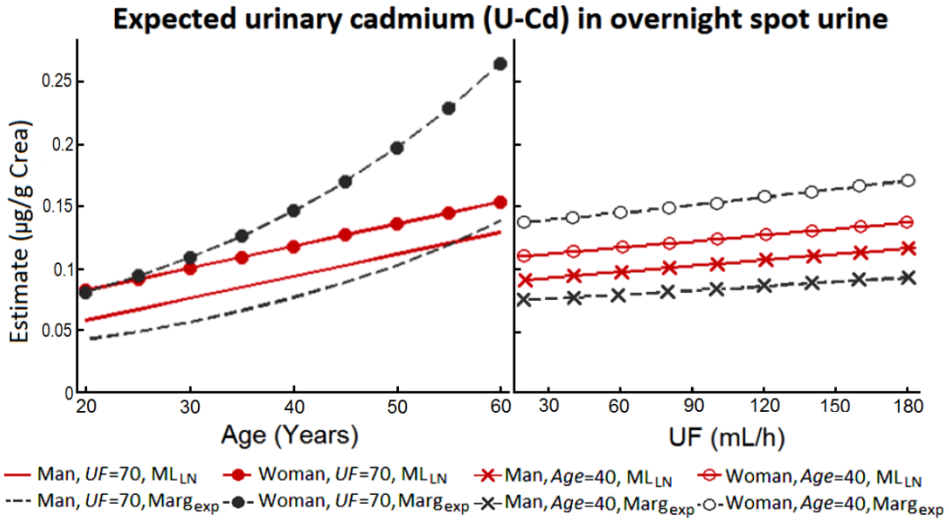

Figure 5. Expected urinary cadmium in overnight spot urine, estimated with ML_{LN} and $Marg_{exp}$. Left: The estimates as a function of age. Right: The estimates as a function of urinary flow rate, UF.

5.2 Results from Simulation Studies

The results from the simulation studies are presented separately for the large-sample (asymptotic) results obtained from Papers I, II, and IV and the small-sample results obtained from Paper III.

5.2.1 Large Samples

Larger sample sizes, with at least 100 individuals, were used in Papers I, II, and IV. In Papers I and II, all observations were independent, while Paper IV used repeated measurements; that is, correlated observations within individuals.

As in the applications, the simulations showed that all the absolute-effects methods tended to give similar estimates of β. These estimates were basically unbiased for larger samples sizes (more than 100 individuals). The largest biases were observed for ML_{LN} based estimates. However, all biases were negligible when compared to the standard deviations of the estimates, $|\text{E}[b] - \beta| < 0.2 \cdot \text{SD}(b)$. In Paper I, two different sample sizes were evaluated, n = 108 and n = 216, with independent observations; some of the results for

ML_{LN} are presented in **Table 5**. The bias of ML_{LN} based estimates decreased with increasing sample size; the relative bias was between −0.22% and 0.86% for n=108, and between −0.11% and 0.48% for n=216.

Table 5. Some results for ML_{LN} based estimates of β from Paper I.

σ_Z	n		β	E[b]	bias (b)	SD[b]	E[se]	bias [se(b)]	$\gamma(b)$	corr [b, se]
0.356	108	b_0	**4.803**	4.806	0.06%	0.430	0.424	-1.3%	0.14	0.58
		b_1	**0.574**	0.574	-0.07%	0.064	0.064	-1.2%	0.05	0.27
	216	b_0	**4.803**	4.804	0.03%	0.304	0.302	-0.6%	0.10	0.58
		b_1	**0.574**	0.574	-0.04%	0.045	0.045	-0.6%	0.04	0.28
0.450	108	b_1	**2.092**	2,087	-0.22%	0.195	0.19	-2.9%	3.94	0.46
		b_2	**0.761**	0.760	-0.10%	0.136	0.132	-2.9%	1.51	0.39
		b_3	**0.218**	0.220	0.86%	0.061	0.059	-2.8%	0.33	0.50
	216	b_0	**2.092**	2.09	-0.11%	0.141	0.14	-0.8%	0.09	0.51
		b_1	**0.761**	0.761	-0.07%	0.099	0.098	-0.8%	0.09	0.45
		b_2	**0.218**	0.219	0.48%	0.046	0.045	-0.9%	0.18	0.70

Some positive skewness could be seen for the distribution of all the β estimates, where $\beta \neq 0$, independent of the method used. However, the skewness was minor in most cases, $\gamma(b) \ll 0.5$, and the results from independent observations in Paper I showed that the skewness also decreased with increasing sample size. The only result with any notable skewness was for the estimates of the regression coefficients in the no-intercept model in Paper I, where two of the estimates had $\gamma(b_1)$=3.9 and $\gamma(b_2)$=1.5, respectively, for n=108. However, in both cases the skewness decreased to $\gamma(b)$=0.09 when the sample size increased to n=216. There was a noticeable positive correlation between the β estimates and $se(b)$. This correlation was between 0.25 and 0.86, and a comparison between the different sample sizes used in Paper I showed that the correlation did not decrease when the sample size increased.

As in the applications, the simulations showed that LS_{lin} tended to give larger $se(b)$ than the other absolute-effects methods. The smallest $se(b)$ was seen for ML_{LN}, GLM_N, and GLM_G. The simulation approach allowed a comparison between the estimated deviation, $se(b)$, and the actual deviation, SD(b); this showed that, for all methods, $se(b)$ tended to be an underestimation of SD(b). This bias was usually minor, |bias| < 3%, for the absolute-effects methods that handled the heteroscedasticity of Y (i.e. ML_{LN}, WLS, GLM_G, and GLM_N). The

comparison of different sample sizes of independent observations in Paper I showed that the bias decreased from between −2.9% and −0.7% for n=108 to between −1.2% and −0.2% for n=216. However, the bias of the LS_{lin} based $se(b)$ was between −8% and 80%, and there was no noticeable decrease in bias for increased sample sizes; the relative bias in Paper I was between −8.1% and 79% for n=108, and between −7.4% and 80% for n=216.

The narrowest CIs for β were seen for the maximum likelihood based methods, while LS_{lin} tended to have the widest intervals; this is in accordance with the sizes of their $se(b)$. For a sample with 108 independent observations, the coverage probability for a 95% CI for β was between 0.943 and 0.948 for ML_{LN}, GLM_N, GLM_G, and WLS, and between 0.928 and 0.974 for LS_{lin}. For repeated measurements (100 individuals, four measurements per individual), the coverage probability of a ML_{LN} based 95% CI for β was between 0.947 and 0.950, given a correctly specified covariance pattern.

Both the relative methods and the absolute-effects methods, with the exception of LS_{lin}, could be used to estimate σ_Z. ML_{LN}, $Marg_{exp}$, GLM_N, and WLS all tended to give a slight underestimation of σ_Z, while GLM_G gave a larger underestimation of σ_Z. The bias for ML_{LN}, $Marg_{exp}$, GLM_N, and WLS was usually small, between −1.8% and −1% for the independent observations methods in Papers I and II and between −0.7% and −0.1% for the repeated measurements methods (i.e. ML_{LN} and $Marg_{exp}$) in Paper IV. In Paper II, the GLM_G based σ_Z estimate, in which a gamma distribution is assumed instead of a log-normal distribution, had a much larger bias at about −6.5%.

In Paper IV, repeated measurements were used and the correlation parameter ρ was estimated. Both the absolute effects method ML_{LN} and the relative-effects method $Marg_{exp}$ tended to underestimate ρ. The ML_{LN} based estimates had a bias between −2.2% and −1.7%, while the $Marg_{exp}$ estimates had a bias between −3.1% and −2.2%.

Some consequences of using a wrongly specified covariance pattern were also investigated in Paper IV. Data were simulated according to one of the patterns, AR(1) or CS, but the other pattern was assumed in the estimation. Using the wrong covariance pattern in the estimation did not affect the point estimates of β, but the underestimation of SD(b) increased; that is, |E[se(b)]−SD(b)| was larger for models with misspecified patterns, compared to correctly specified models. The bias for $se(b)$ based on a wrongly assumed AR(1) pattern was between −12% and −6% for the coefficient of a predictor that was constant within the individual, while it was only about −2% for a predictor that varied within the individual. The bias of $se(b)$ based on a wrongly assumed CS

pattern was about −2%, both for the varying and for the constant-within-individual predictor.

In the estimation of the expected value μ_Y, the absolute-effects methods gave similar, and basically unbiased, estimates. The consequences of estimating a log-linear model for a linear relationship could be seen in the μ_Y estimates of the relative-effects methods, LS_{exp} and $Marg_{exp}$, which overestimated the μ_Y values in the highest and lowest quantile and underestimated values near the median. For samples with at least 100 individuals, the coverage of 95% CIs of μ_Y was around 0.95 for ML_{LN} (both for independent and repeated measurements) and WLS, between 0.92 and 0.99 for LS_{lin}, around 0.93 for GLM_G, around 0.94 for GLM_N, and between 0.83 and 0.95 for LS_{exp} and $Marg_{exp}$.

5.2.2 Small Samples with Independent Observations

Smaller sample sizes were investigated in Paper III using ML_{LN} and WLS. Only results for $n = 18$ are presented in the paper, but sample size $n = 27$ was also investigated and the results were similar.

The direct ML_{LN} estimates of β had only a small bias of less than 2%, and the WLS estimates of β were unbiased. The parametric and semi-parametric bias-corrected bootstrap estimate reduced the bias of the ML_{LN} estimates to less than 0.5%.

As with the applications described in Section 5.1.2, the ML_{LN} based $se(b)$ were smaller and the ML_{LN} based intervals narrower than the WLS based $se(b)$ and intervals. For a sample size of $n = 18$, the bias of $se(b)$ was about −10% for the ML_{LN} based estimates and between −8% and −6% for the WLS based estimates.

The ML_{LN} based intervals were narrower than the WLS based intervals, in the same way as with the applications. Both intervals had roughly the same coverage probability. Again as with the applications, the bootstrap-*t* intervals were wider than the Wald-type and BC_a intervals (the Wald-type intervals used were based on the *t*-distribution). However, for the effects (β_1 and β_2), the bootstrap-*t* intervals were the only intervals with a correct coverage; either equal to the nominal 0.95 or a small over-coverage (between 0.95 and 0.96). The 95% Wald-type interval had a coverage probability of about 0.93, and the BC_a intervals had a substantial under-coverage at about 0.90 for the parametric bootstrap and between 0.90 and 0.93 for the semi-parametric bootstrap. The

bootstrap percentile intervals had about the same coverage as the BC_a intervals.

The actual type I error, when testing H_0: $\beta = \beta_T$, where β_T is the parameter value specified in the simulation model, was compared to the nominal type I error. The Wald-test and *t*-test (based on asymptotic assumptions) had inflated α_T, up to twice the size of the nominal value. Also, α_T was unequal for one-sided tests; α_T for H_1: $\beta < \beta_T$ was larger than α_T for H_1: $\beta > \beta_T$. For the effects (β_1 and β_2), the bootstrap *t*-tests (*bt*-tests) were the only test with an α_T close to α, as well as similar α_T for the two one-sided tests. The *bt*-test based on the bias-adjusted estimate (only performed for ML_{LN}) did not give better results than the *bt*-test based on the direct estimate. The actual α_T was overall closer to α for the *bt*-test compared to asymptotic theory based tests.

The α-specificity for tests of H_0: $\beta_i = \beta_T$, was between 91 and 100%. However, the α-sensitivity was low, and varied both between the test statistics and depending on α level. For both the parametric and semi-parametric *bt*-tests, the α-sensitivity increased with increasing α-values, while no such pattern could be seen for the regular *t*-test. The α-sensitivity was higher for the parametric *bt*-test than for the semi-parametric *bt*-test. The average difference in α-sensitivity between parametric and semi-parametric *bt*-tests was, for the effects, 9.8 percentage points for ML_{LN} and 14.2 for WLS. When comparing ML_{LN} and WLS for the parametric *bt*-test, they had similar α-sensitivity; for the effects the average difference was 0.9 percentage points in favor of ML_{LN}. The differences were larger for the semi-parametric *bt*-test, averaging 5.3 percentage points when only the effects were considered. The asymptotic theory test (the standard *t*-test) had a very low α-sensitivity; an average of 1.1% using ML_{LN} and 0.6% using WLS (for the effects in both cases).

6 DISCUSSION

The focus in this thesis is on linear regression models for an untransformed log-normally distributed response.

6.1 Parameter Estimates and Standard Errors

Point estimates

The ML_{LN} estimates of β showed a small bias both for independent observations and for repeated measurements. This was not surprising, since many maximum likelihood estimators are biased, see e.g. (Cordeiro and McCullagh, 1991; Diggle, 1988). However, maximum likelihood estimators are also known to be consistent and asymptotically efficient under some regularity conditions (see e.g. (Casella and Berger, 2001)), and so the bias decreases with increasing sample size and the estimates usually have a small variance (Casella and Berger, 2001). The ML_{LN} based β estimates followed the pattern of a consistent estimator, and the bias decreased with increasing sample size. For independent observations we found an absolute bias of $< 1.7\%$ for $n = 18$, $< 1\%$ for $n = 108$, and $< 0.5\%$ for $n = 216$. For repeated measurements with $\rho = 0.36$, the absolute bias was $< 0.4\%$ for 100 individuals and 4 measurements per individual (Paper IV); this decreased to $<0.2\%$ for 100 individuals and 5 measurements per individual. We found the bias to be negligible in comparison to the standard deviation; $\text{bias}(b) \leq 0.03 \cdot \text{SD}(b)$. In most situations, this bias will probably be too small to have clinical relevance.

Almost all the β estimates, both for independent observations and for repeated measurements, had a slightly positively skewed distribution; but the skewness was in most cases moderate ($\gamma < 0.5$) and decreased with increasing sample size. For example, the skewness of ML_{LN} based estimates of $\beta_2 = 0.122$ in the model $\mu_Y = 1.564 + 0.122 \cdot X_1 + 0.075 \cdot X_2$ (from Papers II and III) decreased from $\gamma(\beta_2) = 0.119$ for $n = 18$ to $\gamma(\beta_2) = 0.022$ for $n = 108$.

Standard errors for b

The ML_{LN} based standard errors for b, $se(b)$, were usually a slight underestimation of the standard deviation; $0.97 \cdot \text{SD}(b) < \text{E}[\text{se}(b)] \leq \text{SD}(b)$. In maximum likelihood estimation, the estimated covariance matrix for b is given by the observed Fisher information. For independent observations, this is known to have bias of order n^{-1}, hence $se(b)$ should have bias of order $n^{-1/2}$, see e.g. (Efron and Hinkley, 1978).

Confidence intervals for β

Wald-type intervals for β were evaluated in all of the papers (I–IV). For larger sample sizes with more than 100 observations, the ML_{LN} based Wald-type intervals had a coverage probability close to the nominal confidence level.

For independent observations we also investigated small sample sizes (Paper III). In samples with less than 30 observations the Wald-type intervals became liberal; that is, the coverage probability was less than the nominal confidence level. We found that intervals with a more accurate coverage probability could be achieved with bootstrapping (see Section 3.2.2). Since small samples can produce estimates with bias and skewness, much of the literature suggests that the bootstrap bias corrected accelerated percentile interval, BC_a, should be used for constructing confidence intervals for β, rather than the bootstrap *t*-statistics based interval, *boot-t*, see (Efron, 1987). However, in our results the *boot-t* intervals outperformed BC_a when it came to coverage probability. BC_a gave narrower intervals and had less variance in interval length, but also a severe under-coverage. For example, with the ML_{LN} based estimates, the coverage probability for the parametric BC_a was 0.90−0.91 for a 95% CI for the effect estimates, while *boot-t* had a coverage probability of 0.95-0.96 for the same estimates.

As expected, the underestimation by $se(b)$ was more noticeable for smaller sample size, about −10% for ML_{LN} and about −7% for WLS. However, this underestimation had little effect on the coverage of the bootstrap confidence intervals. The bootstrap percentile interval and the BC_a interval are not based on $se(b)$, so their liberal coverage was not due to the underestimation by $se(b)$. On the other hand, *boot-t* is based on estimated standard errors. Nevertheless, even though the underestimation for $se(b)$ was larger for ML_{LN}, the ML_{LN} and WLS based intervals had about the same coverage for the effect. The *boot-t* interval is based on both the standard errors for the parameter, $se(b)$, and the bootstrap standard errors, $se^*(b^*)$, and underestimation of these estimates will have adverse effects; if $se(b)$ is an underestimation this will narrow the interval and increase the risk of under-coverage, while if the bootstrap estimates $se^*(b^*)$ are underestimations this will widen the *boot-t* interval, since it is based on the percentiles of $t^* = (b^* - b)/se^*(b^*)$.

Since the ML_{LN} estimates are maximum likelihood estimates, there might be arguments for the likelihood ratio interval being a better choice (see e.g. (Pawitan, 2001). The reason for using Wald-type intervals was that these can be calculated for all the methods compared, which we felt would lead to a fairer comparison between the methods. The same reasoning was behind the

choice of using z-tests or *t*-tests instead of a likelihood ratio test for hypothesis testing of β.

Hypothesis testing of β

For larger samples with more than 100 participants, the two-sided *t*-test (i.e. tests of H_0: $\beta \neq \beta_T$, where β_T was the value specified in the simulation model) gave accurate results with actual α-levels, α_T, close to the nominal for ML_{LN} based estimates. However, the one-sided *t*-tests had a slightly uneven type I errors; α_T for tests of H_0: $\beta < \beta_T$ tended to be smaller than α_T for tests of H_0: $\beta > \beta_T$. This discrepancy between the one-sided tests is probably due to a correlation between b and $se(b)$. Consider a β-estimate $b_H > \beta_T$ and a β-estimate $b_L < \beta_T$, such that $b_H - \beta_T = \beta_T - b_L > 0$. We get $t_H = (b_H - \beta_T)/se(b_H) > 0$ and $t_L = (b_L - \beta_T)/se(b_L) < 0$. The positive correlation will probably give us $se(b_H) > se(b_L)$, hence we will get $|t_H| < |t_L|$. So even if the β-estimates had a symmetric distribution, the test statistic will have a negative skewness. This correlation between b and $se(b)$ is not present when the response is normally distributed, and is most likely a result of the increasing variance for the log-normal distribution. The difference in α_T between one-sided t-test were about 0.015, for nominal $\alpha = 0.05$. A conclusion draw from this is that for balanced datasets with more than 100 participants, *t*-test of H_0: $\beta \neq \beta_T$ can be used for ML_{LN} with less than four predictors, but that one-sided *t*-test should be avoided if possible.

From our simulations with smaller datasets we draw the conclusion that for ML_{LN}, with more than one predictor, a sample size of 30 observations is too small to obtain valid test results for *t*-test of H_0: $\beta \neq \beta_T$, but that bootstrap-t test can be used instead. For smaller samples, the *t*-test had inflated type I errors; that is, α_T was larger than the nominal α. The bootstrap *t*-test however had for the effects α_T values close to α. As mentioned earlier, the ML_{LN} based β estimates were slightly biased, and this bias could be decreased by using the bootstrap bias corrected estimates. There was however no gain in using a bias corrected bootstrap *t*-test, compared to the regular bootstrap *t*-test, with regard to type I error or symmetry between one-sided tests.

Confidence intervals for μ_Y and prediction intervals

For larger samples, with at least 100 participants, the ML_{LN} based confidence and prediction intervals had a nearly accurate coverage probability that did not depend on the value of μ_Y. The 95% ML_{LN} confidence intervals in Papers II and IV all had coverage probabilities between 0.94 and 0.95. For small samples (Paper III), the Wald-type confidence intervals for μ_Y became too

liberal for samples with less than 30 participants, similarly to the confidence intervals for β.

For ML_{LN} confidence intervals for μ_Y and prediction intervals, we draw similar conclusions about Wald-type intervals as we did regarding *t*-test of H_0: $\beta \neq \beta_T$; more than 30 observations are needed for the Wald-type interval to be valid for ML_{LN} with more than one predictor, but that for balanced data with more than 100 participants, the Wald-type intervals give adequate coverage for ML_{LN} with less than four predictors.

Sample size

Since we only investigated a limited number of sample sizes, we are not able to give an absolute guideline regarding the minimum number of observations needed to use asymptotic inference in multivariate regression. The literature contains a number of guidelines, such as $n \geq 10 \cdot k$ (Kleinbaum et al., 1998; Maxwell, 2000; Wampold and Freund, 1987) and $n \geq 104 + (p-1)$ or $n \geq 50 + 8 \cdot (p-1)$ for $(p-1) < 7$ (Green, 1991), where p is the number of regression coefficients in the model. However, these limits are based on expected effect sizes and power, and do not take into account the sample size needed to achieve the asymptotic properties. In our situation, with balanced and slightly unbalanced data and at least two predictors, our results show that a sample size of 108 was sufficient to achieve the asymptotic properties. In the application on benzene exposures in Paper III, we had $n = 40$ and $p = 6$, and could still see indications that the asymptotic distributions were not achieved.

Other models for estimating absolute effects

In the case with independent observations, the results of ML_{LN} were compared to the results of other methods estimating the absolute effects, including weighted least squares (WLS) regression and a generalized linear model (GLM_N).

WLS produces unbiased estimates, as seen in our results where the bias for *b* was < 0.02% for $n>100$. The WLS based covariance matrix for *b* is based on the assumption that σ_Z^2 and the weights $\boldsymbol{W}$ are known and not estimated, and therefore a small downward bias is expected; this could also be seen in our results from Papers I–III, where $0.98 \cdot SD(b) \leq E[se(b)] \leq SD(b)$. This bias was usually slightly smaller than the bias for the ML_{LN} based standard errors, but at the same time the $SD(b)$ were usually larger for the WLS based estimates than for the ML_{LN} based estimates. Hence, the difference in actual bias, $SD(b) - E[se(b)]$, was negligible. The difference in standard deviation can be the difference between a predictor being significant or not at a 0.05

significance level, as in the 1,3-butadiene exposure application in Paper I and the small-sample benzene exposure example in Paper III.

To obtain estimates of β, and $se(b)$, the direct GLM_N estimates were transformed using the term $\exp(\sigma_Z^2/2)$, or rather the estimate $\exp(s_Z^2/2)$. The GLM_N estimates in Paper II were similar to the ML_{LN} estimates both in point estimates and standard errors. Like ML_{LN}, GLM_N is a maximum likelihood method and hence the same downward bias is expected for $se(b)$, which also could be seen in the results from Paper II; $0.98 \cdot SD(b) \leq E[se(b)] \leq SD(b)$.

Like ML_{LN}, WLS and GLM_N provided correct coverage probabilities for the confidence interval for $\mu_{Y|X}$. This coverage probability did not depend on $\mu_{Y|X,}$ and was close to the nominal confidence level. However, GLM_N had a slight under-coverage, with a coverage probability of about 0.94 for the 95% CIs. The WLS CIs tended to be wider than the ML_{LN} intervals, which is not surprising since the WLS estimates had larger standard errors. GLM_N had narrower but as earlier mentioned also slightly liberal intervals. The under-coverage of the GLM_N based intervals was a result of using the estimate s_Z^2, and thus not including the stochastic variation of s_Z^2. An approximate CI taking into account this stochastic variation could be derived using, for example, the delta method, see e.g. (Casella and Berger, 2001).

6.2 Models for Repeated Measurements

We used a mixed model approach to adapt the ML_{LN} method for handling repeated measurements. There are different ways in which mixed models can handle repeated measurements. In this thesis, we used a covariance pattern model which we define as a marginal model; that is, a model without random effects, in which a structure is specified for the covariance matrix in order to handle the correlation between observations. One reason to use a covariance pattern model as a first approach is that in environmental medicine measurements are often taken over a time period, and covariance pattern models offer a simple way to allow us to specify a correlation that changes over time. Like Brown and Prescott (2006), we define covariance structure within the residual matrix; that is, the covariance matrix for the random errors. If instead a random effects model were used, it would be possible to also specify the covariance pattern within the covariance matrix for the random effects. However, that approach is beyond the scope of this thesis.

A relatively simple alternative to a covariance pattern model, which also allows for correlation between measurements, is to use a random intercept model. However, this assumes a constant correlation between all measurements on the same individual. A random intercept model will in most

cases give similar estimates to a compound symmetry (CS) covariance pattern. The relations between the σ_Z and ρ estimates for CS and the between-individual variance σ_B^2 and within-individual variance σ_W^2 estimated by the random intercept model can then be written as $\sigma_Z^2 = \sigma_b^2 + \sigma_w^2$ and $\sigma_B^2 = \rho\sigma_Z^2$. However, theoretically, there are some situations where a random intercept model and a CS covariance pattern model would give completely different results and where the random intercept model would be preferred. Such a situation might be when the trend for the individuals is the opposite of the trend for the mean values of the individuals. In a covariance pattern model, the intercept will be estimated on the basis of the overall mean values, and so the trend will go in the opposite direction of the individual trends. This is basically a version of Simpson's paradox (Simpson, 1951). A random intercept model, however, will be based on the individual mean values and the slope will follow the individual trends. This particular situation might be detected by using a plot of expected versus observed values, in which marks indicates from which individual the observation belong to.

In a situation with equal variance in all subgroups, Brown and Prescott (2006) recommend the use of a random intercept rather than a covariance pattern model when the response over time is of interest; that is, if time is included as a predictor.

6.3 Model Misspecification

Consequences of assuming other distributions for the response

In Papers I and II, the effects of ignoring the skewness and heteroscedasticity of a log-normally distributed response were demonstrated by using ordinary least squares regression on untransformed log-normal data. This approach is denoted by LS_{lin}. The use of LS_{lin} often resulted in large standard errors for the estimated effects and large bias for the standard errors. The LS_{lin} estimates also tended to have inflated type I errors for tests of the effects, and hence erroneous coverage for CIs of the effects. This approach also produced erroneous coverage for CIs of the expected response $\hat{\mu}_Y$. LS_{lin} produced intervals with over-coverage for smaller values and under-coverage for higher expected values, which could be expected since this approach ignores the increasing variance in the data.

It has been argued that statistical analysis based on assumptions of log-normal distribution and gamma distribution will provide basically the same results (Atkinson, 1982; Firth, 1988; McCullagh and Nelder, 1989). However, these examples have assumed multiplicative models. Some studies have shown

discrepancies between gamma and log-normal assumptions even for these multiplicative models (Das and Park, 2012; Wiens, 1999). In Paper II we used a generalized linear model with gamma distribution and the identity link, denoted GLM_G, to estimate the linear model and the constant variance term σ_Z. Our results showed that ML_{LN} and GLM_G gave similar β and σ_Z estimates. However, while ML_{LN} provided the nominal coverage for the confidence intervals for μ_Y, the GLM_G based intervals for μ_Y showed a constant under-coverage; for example, the coverage was 93% for a 95% CI. The reason for approximating the log-normal distribution with the gamma distribution is that the generalized linear model is a common procedure and hence a built-in option in statistical software such as SPSS® (IBM, 2013), SAS® (SAS, 2013), and MATLAB® (MATLAB R2014b). In our simulations, we know the distribution to be log-normal, but as discussed by Atkinson (1982) and Wien (1999) it might in practice be hard to distinguish between a gamma and a log-normal distribution.

Consequences of log-transforming a linear relationship

Since skewed data are common, many approaches have been developed to enable regression analysis on skewed data. Often a transformation is used to stabilize the variance, such as ordinary least squares regression or linear mixed models on log-transformed data. This log-transformation implies an exponential rather than a linear relationship between response and predictors. Another approach is to use generalized linear models or their mixed model extension, generalized linear mixed models. These generalized linear models also use a transformation, the link function. Unless the link function is the identity link, this will also imply a nonlinear relationship between response and predictors. In Papers II and IV, the consequences of assuming an exponential relationship for a linear relationship were investigated by applying ordinary least squares regression or linear mixed models on a log-transformed response, denoted LS_{exp} and $Marg_{exp}$, respectively. These approaches fit an exponential model to a linear relationship, and therefore it is not surprising that they severely overestimated both lower and higher μ_Y values. They also produced under-coverage for the 95% confidence intervals of μ_Y. The size of the under-coverage varied for different μ_Y, but tended to be more severe for the highest and lowest μ_Y values. In Paper IV, we also investigated the prediction intervals of a new observation Y^* for $Marg_{exp}$ and found that the length of the interval was severely underestimated, especially for higher μ_Y values.

The choice between a linear and log-linear model is probably less crucial in situations where all predictors are categorical. In group comparisons, for example where the only predictor is workgroup, a log-linear and linear model

will provide the same estimates for μ_Y. In situations where there is more than one categorical predictor, for example if both workgroup and smoking habits are predictors, the two models will asymptotically give the same expected values for the estimates for μ_Y if interaction terms are included. A log-transformation might however affect the statistical significance of the predictors, and hence the conclusions drawn from the results.

Consequences of a misspecified covariance pattern

Two types of covariance patterns were used for repeated measurement in Paper IV; a first order autoregressive (AR(1)) pattern and a compound symmetry (CS) pattern. The methods' robustness against misspecifications of the covariance patters was investigated. A misspecified covariance pattern means that if simulations were done according to an AR(1) pattern, estimations were done assuming a CS pattern, and vice versa.

A misspecification of the covariance pattern did not affect the point estimate of the regression coefficients, which was expected since the regression coefficients β were the same for all covariance patterns. It did however affect the standard errors $se(b)$ and in some cases the point estimates of the covariance parameters ρ and the standard error of the σ_Z estimate.

A slight underestimation could be seen for $se(b)$, even for the correctly specified models; $E[se(b)] \approx 0.99 \cdot SD(b)$. This underestimation increased when the covariance pattern was misspecified; $0.88 \cdot SD(b) \leq E[se(b)] \leq 0.98 \cdot SD(b)$. The effect of the misspecification was smaller for β_2, the coefficient associated with the varying-within-individual predictor; $E[se(b)] \approx 0.98 \cdot SD(b)$. The largest underestimations were produced by a wrongly assumed AR(1) for coefficients associated with the constant within-individual predictors, $se(b_0)$ and $se(b_1)$; $E[se(b_0)] = 0.94 \cdot SD(b_1)$ and $E[se(b_1)] = 0.88 \cdot SD(b_2)$. However, the point estimates of both ρ and σ_Z were about the same as for a correctly specified model. The unbiased estimate of ρ is most likely a consequence of the AR(1) based ρ estimate only depending on the correlation between the residuals of sequential observations, $corr(residual_i, residual_{i+1})$, and for both a CS and AR(1) pattern we have $corr(residual_i, residual_{i+1}) = \rho$. A wrongly assumed CS pattern had less problems with underestimation of $se(b)$; $E[se(b)] = 0.98 \cdot SD(b)$. However, it severely underestimated ρ, probably because the maximum likelihood estimate for ρ under the assumption of a CS covariance pattern is basically the mean of all within-individual pairwise correlations. Since the pairwise correlation for an AR(1) pattern decreases with distance, the mean of the correlations will be less than ρ, and so the ρ estimate of a wrongly assumed CS pattern will be an underestimation.

6.4 Linear Models for Biological Data

In this thesis, we applied ML_{LN} to two different kinds of data; personal airborne exposures in Papers I, III, and IV, biomarkers in blood in Paper II and in urine in Paper IV. The results were compared to those of log-linear model approaches (i.e. linear models for log-transformed response data).

It is not always obvious whether a linear or log-linear model is more appropriate. To assess the fit of the models, we used the total variance explained by the model in log-scale (R_Z^2), compared observed and estimated values in scatter plots, and estimated residual regression models to assess trends among the residuals.

In our applications, we saw an overall better fit for the linear exposure models, while most of the biomarkers were better fitted by a log-linear model. It might be reasonable to assume that the cumulative direct exposure has a linear relationship to the exposure time. Larson (1969) showed that the median exposure was proportional to the average exposure time; this would indicate a log-linear relationship between time and exposure, since the median in original scale is $\exp(\mu_Z)$. However, Larsen did not compare the mean and median exposures. One reason for preferring a linear over a log-linear model in cases when the two have a similarly good fit is that no additional transform is needed before interpreting the estimated effects, which makes the model easier to understand and present.

Biomarkers can be measured in different elements, including blood as in Paper II and urine as in Paper IV. Studies have shown that the relationship between exposure and biomarker is often better fitted by more complex, non-linear, models (Johnson et al., 2005; Johnson and Rappaport, 2007). Biomarkers often reflect reactions within the body, and can therefore sometimes be seen as a continuous health outcome. The relationship between health outcome and exposure is usually not seen as a linear model, but rather as a model on the log-scale, see e.g. (Rappaport and Kupper, 2008). However, partly due to the complexity of the human body and the range of different exposures and biomarkers, the relationship between exposure and biomarker relationship should not be generalized in advance but must be assessed for each exposure and biomarker combination. For example, in our applications, the log-linear model showed a clearly better fit than the linear model when modeling urinary-cadmium (Paper IV) and a slightly better fit for HOMA-IR (Paper III), but the linear model had a better fit when modeling CRP, the biomarker for inflammation (Paper II).

7 CONCLUSIONS

For linear regression with a log-normal response, the suggested maximum likelihood method ML_{LN} can be used to obtain estimates of the absolute effects, for both independent and repeated measurement data, while taking into account the heteroscedasticity of the data.

The ML_{LN} based β estimates have small biases in both the independent and the repeated data situation. The bias will decrease further with increasing sample size.

In the independent data situation, ML_{LN} estimates tend to have narrower confidence intervals than weighted least squares based estimates. However, the differences are minor, so weighted least squares can be a valid alternative to ML_{LN}.

For inference in the small-sample situation, asymptotic theory methods should not be used, but bootstrapping can be used to obtain valid values. We found that for ML_{LN} estimates the often-recommended BC_a intervals had a severe under-coverage and that the bootstrap-*t* intervals should be used instead.

ML_{LN} can handle repeated measurements by specifying a non-diagonal covariance pattern within the residual matrix. Our results showed that misspecification of the covariance pattern had little effect on the point estimates of the regression parameter. However, the underestimation by the standard errors was more severe when a first order auto regressive pattern was wrongly assumed for data with a compound symmetry structure, than vice versa.

If the relationship is linear, log-transformation of the response leads to liberal predictor intervals with a severe under-coverage. The ML_{LN} based estimates give wider and more accurate intervals.

8 FUTURE PERSPECTIVES

As previously mentioned, there are situations with repeated measurements data in which a random effects model would be preferred over a covariance pattern model. A future perspective could be to include random effects in a linear model with a log-normally distributed response.

In this thesis, we investigated small-sample properties for ML_{LN} for independent observations. A next step could be to investigate small-sample inference for repeated measurements data. Bootstrapping could also be a valid option here. A suggested nonparametric resampling could be to resample individuals, and not specific observations, in order to preserve the within-individual correlations.

Our results revealed a minor bias for the ML_{LN} estimates of the covariance parameters ρ and σ_Z. A suggested way to obtain unbiased estimates of the variance and covariance parameters in $\mathbf{\Sigma}$ is to use residual maximum likelihood. Future work could establish a way to use this method to estimate the parameters in $\mathbf{\Sigma}$, with the aim of reducing the bias observed for the ML_{LN} estimates of ρ and σ_Z.

ACKNOWLEDGEMENTS

I want to thank several people for their contribution to this thesis, both through their direct input on the work itself and for providing encouragement and support.

I want to thank several people for their contribution to this thesis, both through their direct input on the work itself and for providing encouragement and support.

My supervisor *Eva M. Andersson*, for giving me this opportunity and for sharing your experience and knowledge. I have learned a lot from you during these past years and I am sincerely grateful for the deep commitment you have shown in your role as supervisor.

Warm thanks to my co-supervisor *Gerd Sällsten*, for being there with advice and support when I needed it.

All the members of my research group, *Toxicology and Environmental Health*, my other colleagues at *AMM*, and my fellow *PhD students*. Special thanks go to:

Agneta Agge, for being there and giving me extra encouragement when I needed it.

Cina Holmer and *Ann-Sofie Liljeskog Hill* for your patience with all my (not always 100% relevant) questions and all your help with all things administrative.

Lars Barregård, *Björn Fagerberg*, *Oscar Hammar*, *Christopher Pickering* and *Linus Schiöler* for sharing your expertise and for taking the time to help me to improve my work.

All the people at Health Metrics that I have had the pleasure to work with: *Max Petzold*, *Anna Ekman*, *Madeleine Glücksman Retamales*, and *Catrin Wessman*. It has been both fun and educational to work with you.

Those of you who have enriched my life both in and outside work: *Christofer Andersson, Sofie Fredriksson, Jaana Gustavsson, Sandra Johannesson, Anna Levinsson, Peter Molnár, Fredrik Petersson, Mike Smith, Leo Stockfelt, Lisa Svedbom, Mia Söderberg*, and *Magnus Åkerström*. Thank you for all great discussion, for good company at various conference trips and for having been such excellent colleagues and friends. Extra thanks to *Sofie* for always being so supportive (and for your persistent encouragement that got me to start dancing), to *Jaana* for being a terrific office-roommate, to *Fredrik* for all the IT support (and for an excellent poker face when I came to you with a less-intelligent IT question or a glued keyboard), and to *Magnus* for all the wasted time and frustration your orderly mind and helpful nature have saved me from.

Jag vill även tacka mina vänner utanför arbetet som har funnits där och även stått ut med mig under perioder av total förvirring och enkelspårigt tänkande. Särskilt tack till:

Christin Edblom, *Linda Sängstuvall*, *Johanna Östling* och för att ni alltid finns där för mig. Mitt liv hade varit klart fattigare utan er.

Min kusin och vän *Marie Appelquist*, för ditt stöd och uppmuntran, och för alla våra långa diskussioner om allt och inget.

Slutligen så vill jag tacka min familj:

Mina fina syskon, *Lina*, *Jens*, *Anna* och *Tim*, och min far *Egil* och min bonusmamma *Ingegärd*. Tack för allt ert stöd och all uppmuntran.

Ett extra varmt tack till mina underbara systerdöttrar *Lovis* och *Molly* för att ni påminner mig om det viktiga i livet.

Min faster och farbror, *Ruth* och *Pär Appelquist*. Ni har alltid funnits där för mig när jag behövt er, och var ett ovärderligt stöd under min gymnasie- och universitetstid. Utan er hade jag aldrig kommit så här långt. Tack.

Även om hon inte längre är med oss vill jag även tacka min älskade mamma *Christina*, som jag bär med mig i mitt hjärta.

REFERENCES

Akerstrom, M., Barregard, L., Lundh, T., and Sallsten, G. (2013). Variability of urinary cadmium excretion in spot urine samples, first morning voids, and 24 h urine in a healthy non-smoking population: Implications for study design. *Journal of Exposure Science and Environmental Epidemiology* **24**, 171-179.

Atkinson, A. C. (1982). Regression Diagnostics, Transformations and Constructed Variables. *Journal of the Royal Statistical Society. Series B* (*Methodological*) **44**, 1-36.

Bencala, K. E., and Seinfeld, J. H. (1976). On frequency distributions of air pollutant concentrations. *Atmospheric Environment* (*1967*) **10**, 941-950.

Blackwood, L. G. (1992). The lognormal distribution, environmental data, and radiological monitoring. *Environmental Monitoring and Assessment* **21**, 193-210.

Brohall, G., Behre, C.-J., Hulthe, J., Wikstrand, J., and Fagerberg, B. (2006). Prevalence of Diabetes and Impaired Glucose Tolerance in 64-Year-Old Swedish Women. *Diabetes Care* **29**, 363-367.

Brown, H., and Prescott, R. (2006). *Applied mixed models in medicine*, 2 edition: John Wiley & Sons.

Casella, G., and Berger, R. L. (2001). *Statistical inference*: Duxbury Press.

Cordeiro, G. M., and McCullagh, P. (1991). Bias Correction in Generalized Linear Models. *Journal of the Royal Statistical Society. Series B* (*Methodological*) **53**, 629-643.

Crump, K. S. (1998). On Summarizing Group Exposures in Risk Assessment: Is an Arithmetic Mean or a Geometric Mean More Appropriate? *Risk Analysis* **18**, 293-297.

Das, R. N., and Park, J.-S. (2012). Discrepancy in regression estimates between log-normal and gamma: some case studies. *Journal of Applied Statistics* **39**, 97-111.

DiCiccio, T. J., and Efron, B. (1996). Bootstrap Confidence Intervals. *Statistical Science* **11**, 189-212.

DiCiccio, T. J., and Romano, J. P. (1995). On bootstrap procedures for second-order accurate confidence limits in parametric models. *Statistica Sinica* **5**, 141-160.

Diggle, P. J. (1988). An approach to the analysis of repeated measurements. *Biometrics* **44**, 959-971.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics* **7**, 1-26.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*: SIAM.

Efron, B. (1987). Better Bootstrap Confidence Intervals. *Journal of the American Statistical Association* **82**, 171-185.

Efron, B., and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika* **65**, 457-483.

Efron, B., and Tibshirani, R. J. (1994). *An introduction to the bootstrap*: Chapman & Hall/CRC.

Firth, D. (1988). Multiplicative Errors: Log-Normal or Gamma? *Journal of the Royal Statistical Society. Series B* (*Methodological*) **50**, 266-268.

Fox, J. (1997). Bootstrapping Regression Models. In *Applied Regression analysis, Linear models, and Related methods*, 20: Sage Publications, Incorporated.

Green, S. B. (1991). How many subjects does it take to do a regression analysis. *Multivariate Behavioral Research* **26**, 499-510.

Hagenbjork-Gustafsson, A., Tornevi, A., Andersson, E. M., *et al.* (2013). Determinants of personal exposure to some carcinogenic substances and nitrogen dioxide among the general population in five Swedish cities. *J Expo Sci Environ Epidemiol*.

IBM (2013). IBM SPSS Statistics 20 Algorithms.

Johannesson, S., Gustafson, P., Molnar, P., Barregard, L., and Sallsten, G. (2007). Exposure to fine particles (PM2.5 and PM1) and black smoke in the general population: personal, indoor, and outdoor levels. *J Expos Sci Environ Epidemiol* **17**, 613-624.

Johnson, B. A., Kupper, L. L., Taylor, D. J., and Rappaport, S. M. (2005). Modeling Exposure-Biomarker Relationships: Applications of Linear and Nonlinear Toxicokinetics. *Journal of Agricultural, Biological, and Environmental Statistics* **10**, 440-459.

Johnson, B. A., and Rappaport, S. M. (2007). On modelling metabolism-based biomarkers of exposure: a comparative analysis of nonlinear models with few repeated measurements. *Statistics in Medicine* **26**, 1901-1919.

Kahn, H. D. (1973). Note on the distribution of air pollutants. *Journal of the Air Pollution Control Association* **23**, 973-973.

Kenward, M. G., and Roger, J. H. (1997). Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood. *Biometrics* **53**, 983-997.

Kepteyn, J. (1903). Skew frequency curves in biology, and statistics: Astronomical Laboratory. *Groningen: Noordhoff*.

Kleinbaum, D. G., Kupper, L. L., Muller, K. E., and Nizam, A. (1998). Applied regression analysis and other multivariable methods. 798: CengageBrain. com.

Koch, A. L. (1966). The logarithm in biology 1. Mechanisms generating the log-normal distribution exactly. *Journal of Theoretical Biology* **12**, 276-290.

Koch, A. L. (1969). The logarithm in biology: II. Distributions simulating the log-normal. *Journal of Theoretical Biology* **23**, 251-268.

Kramer, M. (2005). R2 statistics for mixed models. In *Proceedings of the Conference on Applied Statistics in Agriculture*, 148-160.

Kumagai, S., Kusaka, Y., and Goto, S. (1997). Log-normality of distribution of occupational exposure concentrations to cobalt. *Annals of Occupational Hygiene* **41**, 281-286.

Kutner, M. H., Nachtsheim, C. J., and Neter, J. (2004). *Applied linear regression models*.

Land, C. E. (1972). An Evaluation of Approximate Confidence Interval Estimation Methods for Lognormal Means. *Technometrics* **14**, 145-158.

Larsen, R. I. (1969). A New Mathematical Model of Air Pollutant Concentration Averaging Time and Frequency. *Journal of the Air Pollution Control Association* **19**, 24-30.

Limpert, E., Stahel, W. A., and Abbt, M. (2001). Log-normal distributions across the sciences: keys and clues. *BioScience* **51**, 341-352.

Maxwell, S. E. (2000). Sample size and multiple regression analysis. *Psychological Methods* **5**, 434-458.

McCullagh, P., and Nelder, J. A. (1989). Generalized linear models, pp. 286 & 293. London: Chapman and Hall.

Munro, A. H., and Wixley, R. A. J. (1970). Estimators Based on Order Statistics of Small Samples from a Three-Parameter Lognormal Distribution. *Journal of the American Statistical Association* **65**, 212-225.

Nakagawa, S., and Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution* **4**, 133-142.

Oldham, P. (1953). The nature of the variability of dust concentrations at the coal face. *British Journal of Industrial Medicine* **10**, 227.

Osvoll, P. O., and Woldbæk, T. (1999). Distribution and skewness of occupational exposure sets of measurements in the Norwegian industry. *Annals of Occupational Hygiene* **43**, 421-428.

Ott, W. R. (1990). A Physical Explanation of the Lognormality of Pollutant Concentrations. *Journal of the Air & Waste Management Association* **40**, 1378-1383.

Parkhurst, D. F. (1998). Peer reviewed: arithmetic versus geometric means for environmental concentration data. *Environmental science & technology* **32**, 92A-98A.

Pawitan, Y. (2001). *In all likelihood: statistical modelling and inference using likelihood*: Oxford University Press, USA.

Powell, S. (2003). Why worker health and safety branch uses arithmetic means in exposure assessment. [Memorandum]. California: Environmental Protection Agency, Department of Pesticide Regulation.

Rappaport, S. (1991a). Selection of the measures of exposure for epidemiology studies. *Applied Occupational and Environmental Hygiene* **6**, 448-457.

Rappaport, S., and Kupper, L. (2008). Exposure Measurment Errors. In *Quantitative Exposure Assessment*, 111-126: Stephen Rappaport.

Rappaport, S. M. (1991b). Assessment of long-term exposures to toxic substances in air. *Annals of Occupational Hygiene* **35**, 61-121.

Razali, N. M., and Wah, Y. B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics* **2**, 21-33.

Sallsten, G., Ljungkvist, G., and Barregard, L. (2003). [in Swedish] Allmänbefolkningens exponering för bensen, toluen och xylen – personlig exponering, individrelaterade stationära mätningar och bakgrundsmätningar i Göteborg. Goteborg: Sahlgrenska akademin, Göteborgs universitet.

SAS (2013). SAS/STAT® User's Guide. In *The GENMOD Procedure*: SAS Institute Cary, NC.

Simpson, E. H. (1951). The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society. Series B* (*Methodological*) **13**, 238-241.

Wampold, B. E., and Freund, R. D. (1987). Use of multiple regression in counseling psychology research: A flexible data-analytic strategy. *Journal of Counseling Psychology* **34**, 372-382.

Waters, M. A., Selvin, S., and Rappaport, S. M. (1991). A measure of goodness-of-fit for the lognormal model applied to occupational exposures. *Am Ind Hyg Assoc J* **52**, 493-502.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica* **48**, 817-838.

Wiens, B. L. (1999). When Log-Normal and Gamma Models Give Different Results: A Case Study. *The American Statistician* **53**, 89-93.

Yazici, B., and Yolacan, S. (2006). A comparison of various tests of normality. *Journal of Statistical Computation and Simulation* **77**, 175-183.

Yurgens, Y. (2004). Quantifying environmental impact by log-normal regression modelling of accumulated exposure. Licentiate of Engineering, Chalmers University of technology and Göteborg University, Gothenburg.

Zhou, X.-H., Gao, S., and Hui, S. L. (1997). Methods for Comparing the Means of Two Independent Log-Normal Samples. *Biometrics* **53**, 1129-1135.

Zhou, X. H., and Gao, S. (1997). Confidence intervals for the log-normal mean. *Statistics in Medicine* **16**, 783-790.

ERRATA

This document lists errors found in the published version of **Paper II**. Only errors that matter from a computational or interpretational viewpoint are listed.

Location	Original text	Correction
p3524, line 6 Equation (5)	$2x_{p-1}x_p \cdot \mathrm{cov}(\hat{\beta}_{p-1}, \hat{\beta}_1)$	$2x_{p-1}x_p \cdot \mathrm{cov}(\hat{\beta}_{p-1}, \hat{\beta}_p)$
p3524, line 9	$\left(\exp(\mu_{Y\|X}) \pm \sqrt{\mathrm{var}(\exp(\hat{\mu}_{Y\|X}))}\right)$	$\left(\exp(\hat{\mu}_{Z\|X}) \pm \sqrt{\mathrm{var}(\exp(\hat{\mu}_{Z\|X}))}\right)$
p3524, line 11 Equation (6)	$2x_{p-1}x_p \cdot \mathrm{cov}(\hat{\phi}_{p-1}, \hat{\phi}_1)$	$2x_{p-1}x_p \cdot \mathrm{cov}(\hat{\phi}_{p-1}, \hat{\phi}_p)$
p3524, line 13	$var(\hat{\mu}_{Z\|X})^2$	$var(\hat{\mu}_{Z\|X})$
p3524, line 15 Equation (7)	$2x_{p-1}x_p \cdot \mathrm{cov}(\hat{\delta}_{p-1}, \hat{\delta}_1)$	$2x_{p-1}x_p \cdot \mathrm{cov}(\hat{\delta}_{p-1}, \hat{\delta}_p)$
p3527, line 4 Table caption	95% confidence interval for $\hat{\mu}_Y$,	95% confidence interval for μ_Y,