Contents

Title: Substitution or maximum likelihood methods dealing with censored environmental data: A simulation study

Abstract: Environmental data is often censored due to the limitation of measuring equipment. At the Swedish Museum of Natural History, a substitution method is used, substituting the non-detects with a fraction of the limit of quantification to analyze time trends of the concentration levels of various metals in different wildlife in Swedish water. A variation of the Tobit model, implementing random effects is instead proposed, approximating the maximum likelihood estimates using the EM-algorithm. A simulation study comparing the two methods is carried out using different levels of censoring, regression slope value, and standard deviation of the error terms. It is shown that for data with small noise, the proposed model produced unbiased estimates in contrast to the substitution method. It is also shown that when using the substitution method, as the proportion of censoring and the regression slope increase, the estimate of the regression slope centers around a value further away from the truth while the proposed model keeps centered at the correct value. The substitution method also failed to produce confidence intervals with reasonable coverage as opposed to the Tobit variant model having coverage at the confidence level. The two methods were applied to real data from the Swedish Museum of Natural History showing the possibility of producing inaccurate conclusions.

1 Introduction

In some studies, it's difficult to achieve exact values in an experiment. It could, for example, be the reason that patients stop coming back to the hospital, making it impossible to know the exact length of survival after an operation or as in the case of this thesis, the instruments not being able to detect small enough concentrations of toxins in samples. These types of data are called censored. One of the most common errors being made when analyzing censored data is *fabrication*. The analysts substitute the non-detects with a fraction (often one half) of the quantification- or detection limit. A study was made by Helsel (2006) showcasing that this method produces poor estimates of statistics and have the potential to not only overlook patterns in the data but also impose its own fabricated patterns. This could cause a government investing millions to clean a lake of toxins after a report displaying an increase in concentrations of a certain metal in fish when in fact, there was no such pattern to begin with. The reverse is even more terrifying, obtaining a report showing no significant increase in concentration, when indeed the concentration of said metal has been increasing for years. Causes of an increase in concentration have been missed, remedies go undone and the health of humans and the ecosystem is unnecessarily endangered. There are plenty more mistakes commonly being made when handling censored data, including misinterpreting an improvement in measuring technique for a decrease in censored data. However, this will not be discussed in this thesis.

This thesis will study the methods used by the Swedish Museum of Natural History to analyze censored data and compare these with a Linear Mixed-effects model which takes censoring into consideration more than using substitution. The thesis begins by describing some background and data sampling. A theory section follows, describing briefly the methods of linear regression, Mixed-effect models, Maximum Likelihood estimation, and the Tobit model before deriving the model which will be used to compare with the current methods. A simulation study is made, comparing the two methods performance dependant on different factors. Two different R functions for analyzing the derived model are studied looking at when to pick what function. Thereafter the methods are applied to real data from the Swedish Museum of Natural History and results are analyzed.

1.1 Background

At the Swedish Museum of Natural History, the Department of Environmental Research and Monitoring in a joint effort with other departments conducts monitoring of environmental toxicants as part of the National Swedish Contaminant Programme in marine biota. One of the programs conducted regards analyzing long-term time trends of several toxins in Swedish waters and to estimate the rate of change. The models used to analyze such time trends are at the moment elemental and disregards much of the data collected. One of the more common, but crucial oversights, concerns building models and drawing conclusions from fabricated data due to data being censored.

1.2 Data

The report from Bignert et al (2017) explains much of the data sampling. The data comes from several sampling areas regarded as locally uncontaminated. Several species of fish, as well as guillemot eggs and blue mussels, are collected from different sampling areas each year. When collected, a constant number of 10-12 specimens independent of each other are analyzed for a large number of toxins. For some species, the analysis is done for pooled samples containing several specimens in each pool. To reduce the between-years variation, each sampling area tries to analyze specimens of the same sex and age. However, the variation can not be reduced to zero and other parameters affect the variation such as fat content and local discharges as an example. The concentration between each fish will also contain noise, hence the data sampled will have variation between years as well as within years.

1.3 Censoring

As a result of test equipment not being able to detect small enough quantities of toxins, there is a *Limit of Detection (LOD)* of which any concentration level under the LOD is reported as zero by the equipment.

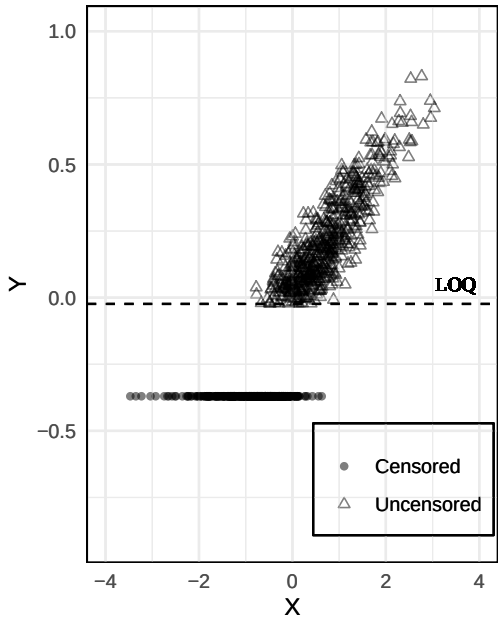


Figure 1a: Example of 1000 observations following a log−normal distribution. Censored observations are replaced with the LOQ divided by the squareroot of 2

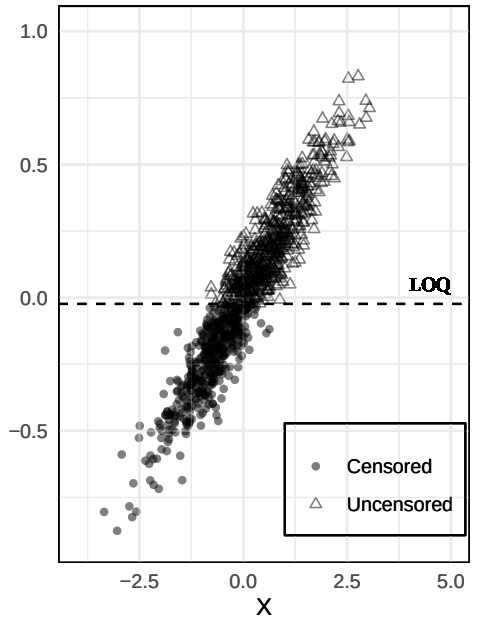


Figure 1b: The same data of 1000 observations is plotted without censoring any of the observations.

It is also possible that whenever the equipment reports a value in an interval between the LOD and another higher value called the *Limit of Quantification (LOQ)* it's accuracy is

questioned. In the case of this thesis, a portion of the data is reported as *below the limit of quantification (LOQ)*. It's also possible that there are several different limits of both sorts due to for example the instruments improving over the years as is the case for the data analyzed in Section 4. This portion of the data is reported as the negative LOQ and later when analyzed is used by taking the absolute value of the reported value divided by the square root of 2. This type of censoring is called left-censoring and is quite common in environmental studies. An example of what data looks like when censored can be seen in Figure 1a while Figure 1b illustrates what the data would truly look like given the scenario that every data point is observable.

Due to biological properties such as size and fat tissues being able to affect the concentration of toxins and these attributes being effected by the sampling site, this thesis will analyze sampling areas individually.

2 Theory

When working with censored data, the censored data points, also known as the non-detects before being censored, can not be looked at as having a specific value. Instead, a combination based on information on the proportion of non-detects with the numerical values of the uncensored observations gives a better understanding of the data. Assuming a distribution for the data above and below the reported limit in combination with the above-mentioned information gives a foundation to work with Maximum Likelihood Estimates (MLE). In a study of Chung (1990) regarding regression analysis of geochemical data with non-detects, it was shown that MLE gave much better estimates of the true value of the slope coefficient than any of the substitution values (0, 0.1, ..., 1 times the detection limit). Regression analysis for censored data is being used in many fields, including but not limited to, medical statistics as used by Lee and Go (1997) and in economics where Chay and Honore (1998) used MLE regression on right-censored data to model incomes. However, for left-censored data where the residuals are assumed to follow a normal distribution, the MLE regression is sometimes mentioned as Tobit analysis after the famous economist James Tobin. For the particular data from the Museum of Natural History, the use of Tobit regression models can serve useful to handle the censoring while the use of a Linear Mixed-effect Model (LMM) will deal with the fact that data contains variation both within and between years.

2.1 Linear Regression Models

Consider a normal simple linear regression model

$$Y_i = \beta_0 + x_i\beta_1 + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2)$$

where Y_i is the stochastic response variable with y_i being an observation from Y_i, x_i the explanatory variable, β_0 is the intercept, β_1 an effect parameter and ϵ_i the error term.

It's then easy to find the cumulative distribution function (CDF) for this model.

$$F(y_i) = P(x_i\beta + \epsilon_i \leq y_i) = P(\frac{\epsilon_i}{\sigma} \leq \frac{1}{\sigma}(y_i - x_i\beta)) = \Phi[\frac{1}{\sigma}(y_i - x_i\beta)]$$

where $\Phi(\cdot)$ is the CDF for a standard normal variable. The probability density function (PDF) is further given by $f(y_i) = \frac{dF(y_i)}{dy_i}$.

2.2 Linear mixed-effects model

Mixed models are an extension of normal models where random effects are integrated. A linear mixed model is further an extension of mixed models where both the fixed and random effects take place linearly in the model. The random effects can be observed as additional error terms in the model. Following the notation of Pinheiro and Bates (2000) the linear mixed model for a single level of grouping, as described by Laird and Ware (1982), can be expressed as

$$\mathbf{y_i} = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \epsilon_i$$

for $i = 1, ..., M$. Here, $\mathbf{y}_i$ is the $n_i \times 1$ dimension response vector for group i, β the $p \times 1$ dimensional vector of fixed-effect parameters, $\mathbf{b}_i$ the $q \times 1$ dimensional vector of random-effects, $\mathbf{X}_i$ a matrix with covariates of size $n_i \times p$, $\mathbf{Z}_i$ a design matrix of size $n_i \times q$ linking $\mathbf{b}_i$ to $\mathbf{y}_i$ and ϵ_i an $n_i \times 1$ dimension vector of error terms independent of each other within group i with $\mathbf{b}_i \sim N(0, \Sigma)$, Σ being the symmetrical, positive semi-definite $q \times q$ dimension covariance matrix and $\epsilon_i \sim N(0, \sigma^2 I)$, I being the $n_i \times n_i$ dimension identity matrix.

2.3 Maximum Likelihood Estimation

One of the most interesting analyzes to be made within regression analysis is what effect each covariate has on the response variable. This is represented by the unknown effect parameter vector β, and thus something of great importance to be able to estimate. This is often done using Maximum Likelihood Estimation. Let θ be the vector containing all parameters, often unknown, of which the function $f(\mathbf{y}; \theta)$ depends on excluding the realisation $\mathbf{y}$ from $\mathbf{Y}$. For the case in Section 2.2, this vector contains the effect parameter vector β as well as the variance parameters for $\mathbf{b}$ and ϵ. For a response variable $\mathbf{Y}$ with observations $\mathbf{Y} = \mathbf{y}$ having a probability mass or density function $f(\mathbf{y}; \theta)$, depending on the observations $\mathbf{y}$ and $\theta \in \mathbf{\Theta}$ being the often unknown parameter vector taking values in the parameter space $\mathbf{\Theta}$, the Likelihood Function is given by $L(\theta; \mathbf{y}) = f(\mathbf{y}; \theta)$. Using the definition of Held and Bové (2014), the likelihood function is the probability mass or density function of the observed data $\mathbf{y}$ viewed as a function of the parameter vector θ. The maximum likelihood estimate of θ denoted as $\hat{\theta}_{MLE}$ is then given as the parameter vector maximizing the likelihood function.

2.4 Tobit Model

The Tobit model is characterized by the latent regression equation

$$y_i^* = \mathbf{x}_i\beta + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2)$$

where y_i^* is the laten dependent variable, $\mathbf{x}_i$ is a vector of covariates, β a vector of effect parameters and ϵ_i is the error term. Given this, the observed dependent variable can be specified as:

$$\begin{cases} y_i = y_i^*, & y_i^* > y_L \\ y_i = y_L, & otherwise \end{cases}$$

with y_L being the reporting limit. This leads us to a function describing the Tobit model:

$$f(y_i|\mathbf{x_i}) = \begin{cases} f(y_i|\mathbf{x_i}) = 0, & y_i < y_L \\ f(y_L|\mathbf{x_i}) = P(y_i^* \leq y_L|\mathbf{x_i}), & y_i = y_L \\ f(y_i|\mathbf{x_i}) = f(y_i^*|\mathbf{x_i}), & y_i > y_L \end{cases}$$

Using the same method as for a normal simple linear regression model, it's further deduced that

$$f(y_i|x_i) = \begin{cases} 0, & y_i < y_L \\ \Phi\left(\frac{y_L - \mathbf{x_i}\beta}{\sigma}\right), & y_i = y_L \\ \frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{x_i}\beta}{\sigma}\right), & y_i > y_L \end{cases}$$

where $\phi(\cdot)$ is the PDF of a standard normal distribution. Hence, the likelihood function for the Tobit model is:

$$L = \prod_{y_i = y_L} \Phi\left(\frac{y_L - \mathbf{x}_i\beta}{\sigma}\right) \cdot \prod_{y_i > y_L} \frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{x}_i\beta}{\sigma}\right)$$

2.5 Linear Mixed-effects With Censored Response

Now, in the case of the analysis conducted by the Swedish Museum of Natural History, a Linear Mixed-effects Model with Censored Response (LMMC) could be implemented regarding each year as a separate group t having n_t specimens. The between-year variance is the same for each specimen in the same group while the within-year variance is the same for every specimen through each year.

Hence, the model, would there be no censored data, is

$$\log(\mathbf{y}_t) = \mathbf{x}_t\beta + \mathbf{z}_i\mathbf{b}_t + \epsilon$$

where $\mathbf{y}_t$ is the $n_t \times 1$ dimension response vector containing the measured concentration of a certain toxin, $\mathbf{x_t}$ a matrix of dimension $n_t \times 2$ having a column of ones for the intercept and a column of the year of sampling, β the 2×1 dimensional vector of fixed effect parameters including the intercept, $\mathbf{z}_t = \mathbf{z}$ a $n_t \times n_t$ identity matrix (due to the fact that there is only one random effect per observation) which links b_t to y_t, $\mathbf{b}_t$ an $n_t \times 1$ dimensional vector of the random effect (between-years) b_t and ϵ the $n_t \times 1$ dimensional vector with the within-years variance for each specimen $\epsilon_i, i = 1, 2, ..., n_t$ where all ϵ_i are independent of each other. The identity matrix $\mathbf{z}$ can be omitted and is only mentioned for clarity in comparing the model with the Lairde and Ware (1982) notations. Further more, since $b_t \sim N(0, \psi_t^2)$ and $\epsilon \sim N(0, \delta^2)$, the distribution of $\log(\mathbf{y}_t)$ follows

$$\log(\mathbf{y}_t) \sim N_{n_t}(\mathbf{x}_t\beta, \mathbf{\Sigma})$$

with $\mathbf{\Sigma} = (a_{ij}) \in \mathbb{R}^{n_t \times n_t}$ the covariance matrix where $(a_{ij}) = Cov(b + \epsilon_i, b + \epsilon_j)$. Further calculations of the covariance gives

$$Cov(b + \epsilon_i, b + \epsilon_j) = E[(b + \epsilon_i)(b + \epsilon_j)] - E[b + \epsilon_i]E[b + \epsilon_j] = E[(b + \epsilon_i)(b + \epsilon_j)]$$

for all i, j such that $i \neq j$ since $E[e] = E[\epsilon_k] = 0$ for all k. Using the fact that the specimens are independent of eachother, it follows that

$$E[(b + \epsilon_i)(b + \epsilon_j)] = E[b^2] = Var(b) - E[b]^2 = \psi_t^2$$

In addition, $(a_{ij}) = Var(b + \epsilon_i) = \psi^2 + \delta^2$ when $i = j$.

Following the method in Section 2.1, the CDF of the model in question can also be derived. First of all, the fact that observations can be censored must be taken into consideration. This is done by partitioning the data into censored and non-censored components

$$\mathbf{y}_t = \begin{bmatrix} \mathbf{y}_t^o \\ \mathbf{y}_t^c \end{bmatrix}, \mathbf{x}_t = \begin{bmatrix} \mathbf{x}_t^o \\ \mathbf{x}_t^c \end{bmatrix}, \mathbf{\Sigma}_t = \begin{bmatrix} \mathbf{\Sigma}_t^{oo} & \mathbf{\Sigma}_t^{oc} \\ \mathbf{\Sigma}_t^{oc^T} & \mathbf{\Sigma}_t^{cc} \end{bmatrix}$$

where $\mathbf{y}_t^o$ is the $n_t^o \times 1$ dimension vector of all the observed, non-censored values and $\mathbf{y}_t^c$ the $n_t^c \times 1$ dimension vector of all censored observations before being censored, the same following for $\mathbf{x}_t$ being partioned in to a $n_t^o \times 2$ matrix and a $n_t^c \times 2$ matrix while $\mathbf{\Sigma}_t^{oo}$ and $\mathbf{\Sigma}_t^{cc}$ are the matrices of variances and covariances between all observed values and censored values respectively and $\mathbf{\Sigma}_t^{oc} = \mathbf{\Sigma}_t^{co^T}$ being the matrix of covariances between non-censored and censored observations. It follows that $\mathbf{y}_t^o$ has a multivariate normal distribution with mean vector $\mathbf{X}_t^o\beta$ and covariance matrix $\mathbf{\Sigma}_t^{oo}$. Using the properties of the multivariate normal distribution, following Eaton (1983), the conditional distribution of $\mathbf{y}_t^c|\mathbf{y}_t^o$ is also multivariate normally distributed with mean vector and covariance matrix as follows

$$\mu_t^{c|o} = \mathbf{x}_t^c\beta + \mathbf{\Sigma}_t^{co}\mathbf{\Sigma}_t^{oo^{-1}}(\mathbf{y}_t^o - \mathbf{x}_t^o\beta), \quad \mathbf{\Sigma}_t^{c|o} = \mathbf{\Sigma}_t^{cc} - \mathbf{\Sigma}_t^{co}\mathbf{\Sigma}_t^{oo^{-1}}\mathbf{\Sigma}_t^{co^T}$$

here $\mathbf{\Sigma}_t^{oo^{-1}}$ is the inverse of $\mathbf{\Sigma}_t^{oo}$. Denote $\Phi_t^{c|o}(\cdot)$ as the conditional distribution function of $\mathbf{y}_t^c$ given $\mathbf{y}_t^o$ and $\mathbf{c}_t$ the n_t^c vector where c_{tj} is the censoring threshold for the j^{th} censored outcome. Now, since all $\mathbf{y}_t$ are independent, using the methods of previous sections and the definition of the conditional probability density function (Held and Bové, p.321), the likelihood function can be written as

$$L(\beta; \mathbf{y}_t) = \prod_t \phi_{\mathbf{y}_t}^o(\mathbf{y}_t^o|\beta) \cdot \Phi_t^{c|o}(\mathbf{c}_t|\beta)$$

which given the PDF of a multivariate normal distributed variable gives

$$L(\beta; \mathbf{y}_t) = \prod_t \frac{1}{\sqrt{(2\pi)^{n_t^o}|\mathbf{\Sigma}_t^{oo}|}} \cdot \exp\left\{ -\frac{1}{2}(\mathbf{y}_t^o - \mathbf{x}_t^o\beta)^T \mathbf{\Sigma}_t^{oo^{-1}}(\mathbf{y}_t^o - \mathbf{x}_t^o\beta)\right\} \cdot$$
$$\int_{-\infty}^{c_{t1}}\int_{-\infty}^{c_{t2}} \cdots \int_{-\infty}^{c_{tn_t^c}} \frac{1}{\sqrt{(2\pi)^{n_t^c}|\mathbf{\Sigma}_t^{c|o}|}} \cdot \exp\left\{ -\frac{1}{2}(\mathbf{z} - \mu^{c|o})^T \mathbf{\Sigma}_t^{c|o^{-1}}(\mathbf{z} - \mu^{c|o})\right\} d\mathbf{z}$$

Considering the museum is working on analyzing time trends and estimating the rate of change, what is of interest now is just that, to estimate the rate of change or in other words, to find the estimate for the parameter vector β. This is more often than not done by finding the root to the *score function* $S(\theta) = \frac{d}{d\theta} l(\theta)$ for each parameter and making sure that the solution is a global maxima where $l(\cdot) = \log[L(\cdot)]$ is the log-likelihood function.

Now, since the likelihood function acquired from the model of the museum being so complex whilst having censored observations, the maximum likelihood estimate is difficult, if not impossible, to find analytically. Therefore, a numerical approach is suggested as also suggested by Dempster, Laird and Rubin (1977), namely, the Expectation-Maximization algorithm, also called the EM-algorithm.

2.6 EM-Algorithm

The EM algorithm is an iterative method for estimating the MLE when the complete dataset is $Z = (X, Y)$ where X is observed data while Y is unobserved. The algorithm contains two steps, the Expectation-step and the Maximizing step, hence it's name. For each iteration, the algorithm produces an estimate $\theta^{(i)}$ resulting in a sequence of estimates $\theta^{(0)}, \theta^{(1)}, ..., \theta^{(p)}$ converging towards $\hat{\theta}_{MLE}$, the MLE estimate of the parameter vector in question as p tends towards infinity (Dempster et al., 1977). Although, it's not correct to say that the algorithm produces the same estimation as the MLE considering the fact that the algorithm will stop, either after some number of iterations decided beforehand or when $|\theta^{(i)} - \theta^{(i-1)}| < \epsilon$ for some determined $\epsilon > 0$. Using the definition of the conditional probability density function, it's possible to write the joint pdf of X and Y as

$$f(\mathbf{x}, \mathbf{y}) = f(\mathbf{y}|\mathbf{x})f(\mathbf{x})$$

and so following the derivations of Held and Bové (2014, pp.35) the log-likelihood can be expressed as,

$$l(\theta; \mathbf{x}, \mathbf{Y}) = l(\theta; \mathbf{Y}|\mathbf{x}) + l(\theta; \mathbf{x})$$

where $\mathbf{y}$ is unobserved and hence exchanged by the random variable $\mathbf{Y}$. Now taking the expectation of this equation with regards to the complete dataset $\mathbf{Z}$ conditioned on the observed data $\mathbf{X}$ and the i:th estimate $\theta^{(i)}$ it follows that

$$E_{\mathbf{Z}}[l(\theta; \mathbf{x}, \mathbf{Y}); \theta^{(i)}] = E_{\mathbf{Z}}[l(\theta; \mathbf{Y}|\mathbf{x}); \theta^{(i)}] + l(\theta; \mathbf{x})$$

where the left hand side is denoted as $Q(\theta, \theta^{(i)})$. The fact that $l(\theta; \mathbf{x})$ is left unchanged is due to it not depending on $\mathbf{Y}$. Knowing this, the EM-algorithm can now be explained in 3 steps:

1. Let $i = 0$ and $\theta^{(i)}$ be the initial guess of the estimate and compute $Q(\theta, \theta^{(i)})$ called the E-step.
2. Maximize $Q(\theta, \theta^{(i)})$ with respect to θ which yields $\theta^{(i+1)}$, called the M-step.
3. Iterate step 1 and 2 by exhanging $\theta^{(i)}$ with $\theta^{(i+1)}$ in step 1 until one of the mentioned stopping criteria of the algorithm has been reached.

Wu (1983) made a study regarding the convergence properties of the EM-algorithm. It was shown that the $L(\theta^{(i)})$ converge monotonically towards some value L^* is the EM sequences bounded. In other words, it always holds that $L(\theta^{(i+1)}) \geq L(\theta^{(i)})$ which is a great property when trying to find the values of θ maximizing the likelihood function. However, there is no guarantee that L^* is the global maximum for the likelihood function over the parameter space Θ. There is a chance that the EM-algorithm converge towards a local maximum or even a saddle point. The convergence to either type of point is decided by the initial values of the starting point.

3 Simulation

The existence of bias for estimates where fabricated data were used has been evaluated by many others, see for example Thompson and Nelson (2003). El-Shaarawi and Esterby (1992) further showed that it's impossible to get unbiased estimates of the mean and standard deviation when using a single value replacing the censored observations while also showing that the bias is independent of sample size, and so what effects the bias is the proportion of censored values and the attributes for the distribution of the data. What is left to investigate is under what conditions one model is better than the other. A simulation study is therefore applied, trying to mimic the environmental setting of the museum as well as possible.

3.1 Simulation Structure

When looking at what methodology to use for a certain dataset in this environment one of the described models needs to be chosen where one is more advanced and time-consuming when doing many analyzes (LMMC) and the other is working with fabricated

data (Substitution). For the LMMC model there are two functions which could be used in R, both trying to maximize the same likelihood function but getting different numerical results, the *lmec* function from the package with the same name produced by Vaida and Liu (2009) or the *mixcens* function, constructed by Martin Sköld at the Swedish Museum of National History. Hence, a smaller analysis looking at the squared bias and variance of the estimated slope using both the LMMC model with the two functions as well as the substitution method for a fixed dataset were made. Considering a choice of which function to use on the dataset in Section 4, the factor levels chosen were those most similar to that of the mentioned dataset. First, the squared bias and variance were analyzed as a function of censoring proportion for all three methodologies. Afterward, a more in-depth simulation study is done using the substitution method and one of the functions for the LMMC model where different factors are altered.

3.1.1 Model Description

A mixed linear model containing one centered covariate X representing years ranging between -5 and 5, and two error terms, ϵ and b, the former representing the noise for each specimen and the latter representing noise between years, was used to investigate different conditions. The between-years variance is different for each year but otherwise independent and the intercept was set to 0. The sample size was set to $n_i = 12$ samples for every year, the same as most of the studies used by the museum. Consequently, the model assumed for the simulation was:

$$\log(Y_{ij}) = X_i\beta + b_i + \epsilon_{ij},\ i = 1, ..., 11,\ j = 1, ..., 12$$

with i being the index denoting the corresponding year and j denoting the individual specimen for that year. Both error terms following a normal distribution with mean 0 and different variances for different scenarios.

3.1.2 Scenario Design

There are countless scenarios to consider but this simulation study takes a closer look at three factors, namely

1. The proportion of censored data varied between 30% and 60% with all data being left-censored.
2. The slope of the regression line alternating between a yearly increase and decrease of 1% and 5% on the original scale
3. The two error terms ϵ and b changed between small, medium, and large for the noise of the individual specimen and between small and large for the between-year noise.

resulting in 48 different scenarios. The limit of quantification was put at the value representing 30% and 60% censored data when both the intercept and slope were put to 0. Hence the proportion of censored data is affected by the slope. The exact values of the standard deviation for the error terms were calculated using the methods of Helsel (2005)

and the *NADA* packages in *R*, namely the *cenmle* function used on the data retrieved from the Swedish Museum of Natural History to calculate the standard deviations on an individual and yearly level. The *cenmle* function with suitable input assumes a log-normal distribution of the data. It thereafter produces values for a probability plot for the data, using the subset of censored data to produce correct percentiles for the uncensored data. The statistics (mean and variance) of the log-normal distribution with the highest probability to have produced the dataset is then given.

Three levels for the within-years noise were chosen to be included in the study while two levels, one low and one high, were chosen for the between-years variance. For the between-years noise, the lower values represent some of the most common levels of noise found while the larger values represent some of the more extreme cases. For the noise of each specimen this resulted in a standard deviation of 0.05, 0.5 and 1.4. Most of the locations in the data had standard deviations on an individual level at somewhere between 0.05 and 0.5. The standard deviation for the noise between years were given by calculating the noise for each year separately, choosing some of the lowest and highest value for each year resulting in the standard deviation ranging between 0.0007 and 0.05417 on the lower scale and between 1.044 and 4.069 for the larger scale.

The number of simulations to run for each scenario should preferably be as large as possible while taking into consideration the time investment required. The larger the number of simulations the lower the Monte Carlo standard error for estimates and performance measurements. Burton et. al. (2006) compared 58 articles which showed that the most common number of simulation used was 500, 1000, or 10000. In this thesis, the R functions used to evaluate the models are highly time-consuming and thus the simulation size had to be lower than preferred.

3.1.3 R Functions

For each of these scenarios, 100 simulation were made each having a sample size of 132 observations in which the method of substituting censored observations with a fraction of the limit of quantification (in this case using the entity of $\mathrm{LOQ}/\sqrt{2}$ to continue mimicking the museum) and the maximum likelihood method were both used. The datasets were simulated using the model described in Section 3.1.1 for each scenario in Section 3.1.2 and the R function *rlnorm* was used to simulate the error terms. The results from the model using fabricated data were retrieved using the base R function *lm* while the results for the maximum likelihood method were calculated using the *lmec* function and the *mixcens* function.

The *lmec* function determines the likelihood function for a linear mixed-effect model as done in Section 2.5 and acquire an approximation of the maximum likelihood estimates using the EM-algorithm. For the *lmec* function, the vector of unobserved data as denoted by $\mathbf{Y}$ in Section 2.6 is, for each different year the vector $\mathbf{y}_t^c$ from Section 2.5. Consequently the observed data as denoted by X in Section 2.6 is for each year the vector $\mathbf{y}_t^o$. Further, the likelihood function in Section 2.5 for each t is the product of the likelihood function of the observed data and the conditional likelihood function of the unobserved data conditioned on knowing the value of the observed data. The complete likelihood function is the product over each t. Hence, $Q(\theta, \theta^{(i)})$ from Section 2.6 in the case of the

lmec function is acquired by taking the expectation of the logarithm of the likelihood function from Section 2.5 conditioned on knowing the values of $\mathbf{y}_t^o$ for each t as well as the i^{th} estimate of the parameter vector θ where $\theta = (\beta^T, \psi, \delta)^T$. To clarify, whenever the expectation is taken over an observed value, the term is left unchanged keeping θ as it is. However, would the expectation be taken over any unobserved value, the parameter vector now used when taking the expectation will be $\theta^{(i)}$ would the term contain any θ to begin with. The term left after taking this expectation of the entire log-likelihood function is maximized with regard to any θ still left in the term. The θ maximizing this will be used in the next iteration as $\theta^{(i+1)}$. For the EM-algorithm used in the *lmec* function to estimate effect parameters a maximum of 20 iterations was decided due to the immense time effort needed for the *lmec* function when using a dataset with a high proportion of censored data.

The *mixcens* function uses the r package *mnormt* to create the likelihood function for the LMMC model. The maximum likelihood estimate is thereafter found by using the function *optim* in R. This function uses the theory by Nelder and Mead (1965) for optimization, however, this theory goes beyond the scope of this thesis.

3.2 Simulation Results

An example of the data obtained from one of the simulations can be seen in Figure 1. Figure 2 shows one of the simulated datasets when computing at a censoring level of 60% with a slope implying a 5% yearly increase, where the x-axis is to represent different years and the y-axis serve as an illustration of the logarithmic value of the concentration of an arbitrary metal found in a specimen. Seeing next to no correlation between the covariate and the response is to be expected in an environment where changes happen slowly. However, small changes over a longer period of time might still have a huge impact as illuminated by Bignert et. al. (2017, pp. 46).

3.2.1 Function Comparison

Figures 3-5 show the results when comparing the three methods *lm* with substitution, the LMMC model with the function *lmec*, and the LMMC model with the function *mixcens*. For a fixed dataset, acquired by the same methods as described in Section 3.1.1-3, each method was used to estimate the regression slope while altering the censoring proportions, increasing it from 0% to 99% by one percent unit at a time. The scenario of a fully censored dataset is of no interest since both the *lmec* and the *lm* functions produce estimates and variances of the estimates at zero and the *mixcens* function is unable to calculate the inverse of the numerically estimated hessian due to numerical errors. For larger censoring proportion the EM-algorithm used in the *lmec* function is particularly slow in contrast to the *mixcens* function which is something to take into consideration. However, the *mixcens* function has numerical problems when the standard deviation of the error terms are low, sometimes producing squared bias and variances which increase by a factor of over one hundred when increasing the proportion of censoring by a single percent unit rendering any graphical analysis useless. Therefore, when the standard deviation is set to 0.05 only the *lmec* function is compared against the substitution method. The squared-bias and

variance for each model were plotted as a function of censoring proportion. Since this analysis is made to find which function to use on the dataset in Section 4, the levels of the factors used are those found in that particular dataset. Therefore, the standard deviation for individual specimen was set to 0.05 and 0.5. The between-year noise set to the lower scale. The slope was altered using the values implying a 1, 5, and 10% yearly increase on the original scale.

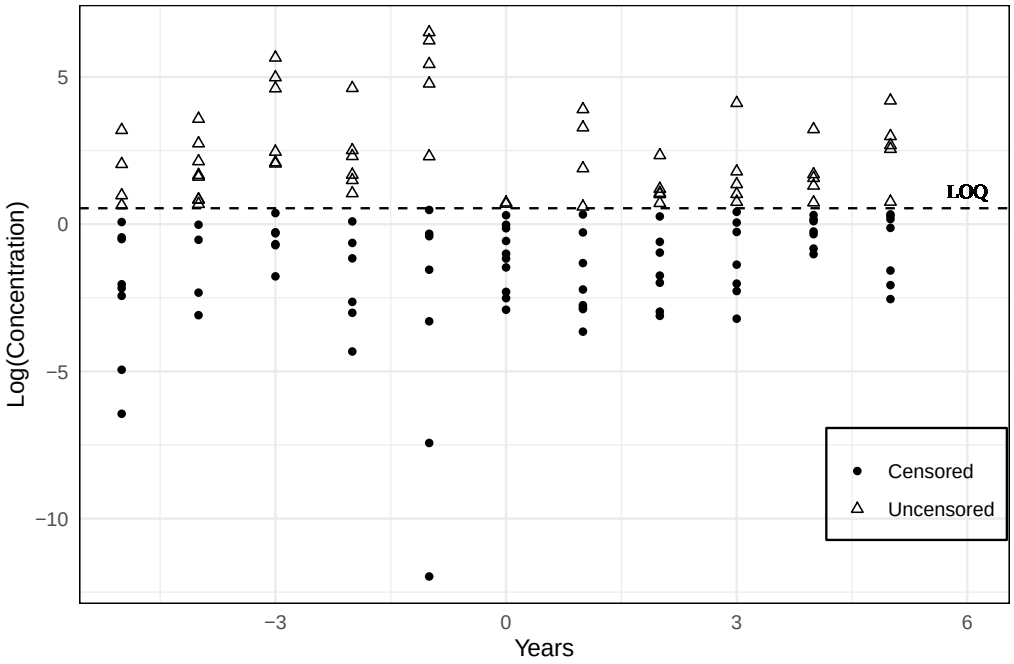


Figure 2: Simulated data with 60% censoring, large variations and slope representing 5% yearly increase.

In Figure 3-4 it's possible to see that for a standard deviation at 0.05 the *lmec* function produces unbiased, or close to, estimates. Figure 5 shows that the same holds for a more steep inclination of the slope when the proportion of censoring is larger. At a censoring of around $25-50\%$ the substitution method would seem to produce more accurate estimates. The variance of the estimated slope for both methods seems to steadily increase as the censoring increase except for in Figure 3 where the variance fluctuate for the substitution method. The variance is always higher when using substitution than when using the LMMC model.

When there is a standard deviation of 0.5 for the error term representing noise within years, it's possible to see that when the inclination of the slope is close to zero (see Figure 3) it appears the methods barely differ for data with a low proportion of censoring. Had the proportion of data been around $30-80\%$, the substitution method is to be preferred. Higher proportion of censoring than that and it's a decision whether or not the squared bias or variance is of more importance. Both the squared bias and the variance increase for the *mixcens* function while they decrease for the *lmec* function. The variance also decreases for the substitution method while the squared bias is kept low. The decision is between the substitution method having a lower variance or the *lmec* function having a

lower squared bias. The *mixcens* function seems to work as well as the *lmec* function up until the highest proportions of censoring where the variance and squared bias deviates.

When the inclination of the slope increases to log(1.05), implying a 5% yearly increase on the original scale (see Figure 4), the same analysis as for the less steeper inclination holds up until a censoring proportion of around 65%. Higher than this and the variance behaves as it did for the lower slope value while the squared bias differs. It increases for the *lmec* function, fast for the substitution method, and a bit slower for the *lmec* function, while decreasing for the *mixcens* function. Once again it's a choice between favoring lower squared bias (*mixcens*) versus lower variance (*lmec*).

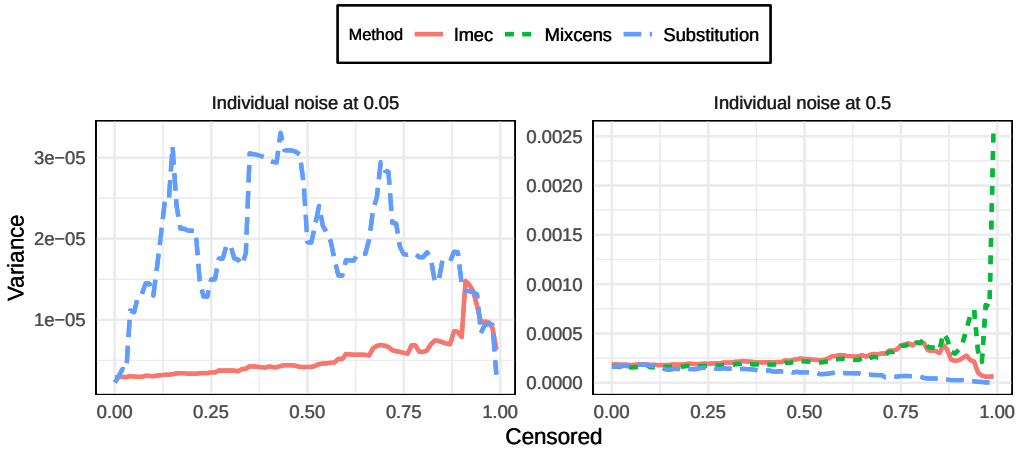


Figure 3a: Standard error of the estimated regression slope as a function of censoring proportion. The true value of the slope being log(1.01).

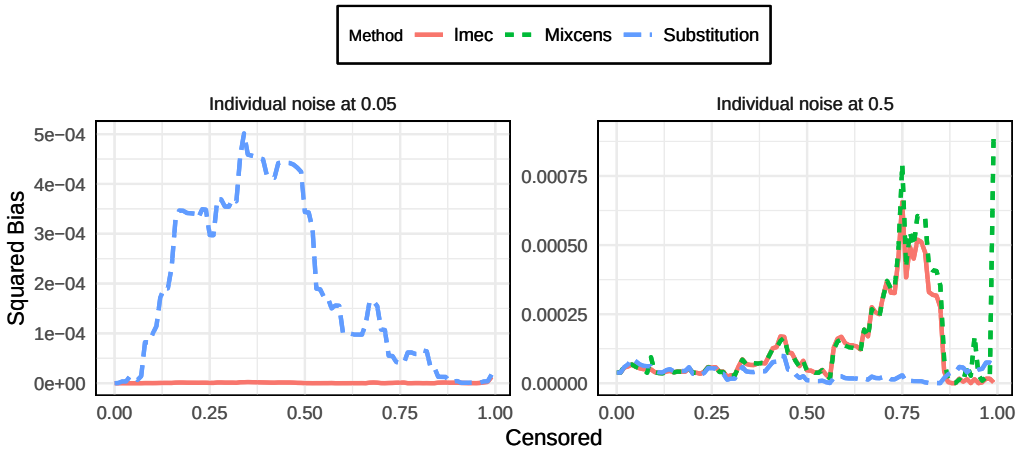


Figure 3b: Squared Bias of the estimated regression slope as a function of censoring proportion. The true value of the slope being log(1.01).

As the value of the slope continues to increase, this time to a value of log(1.1) implying a yearly increase of 10% on the original scale (Figure 5), a pattern emerges. All methods are close to equal up to a censoring proportion of around 30%. Between $30-50\%$ still implies the use of the substitution method. Between $50-85\%$ results in a higher squared

bias but lower variance for the substitution method while the *lmec* and *mixcens* function perform about the same. After over 85%, the substitution method continues it's behavior while the variance of the *lmec* function reduces in contrast to it's squared bias which increases while the reverse holds for the *mixcens* model.

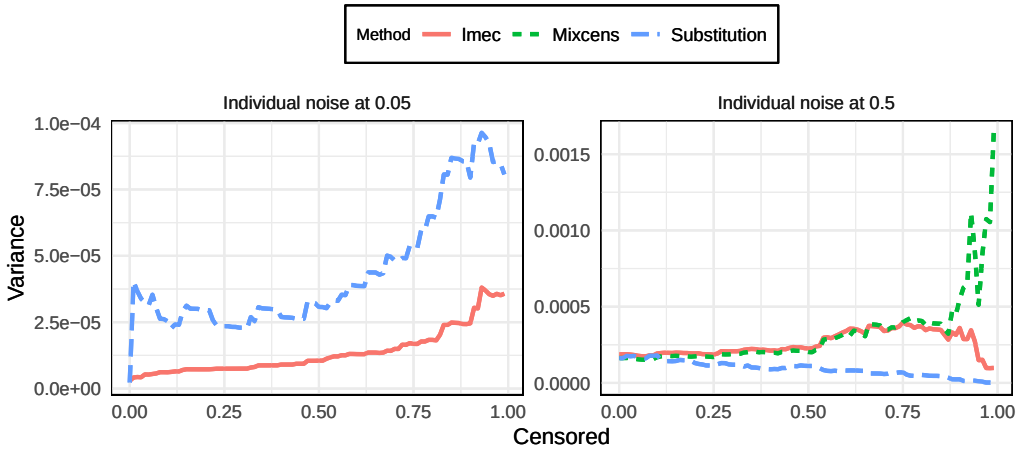


Figure 4a: Standard error of the estimated regression slope as a function of censoring proportion. The true value of the slope being log(1.05).

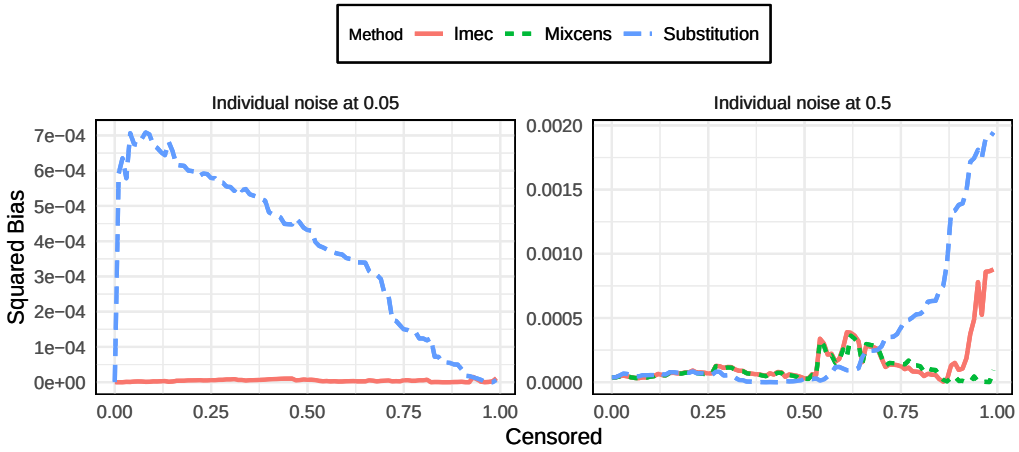


Figure 4b: Squared Bias of the estimated regression slope as a function of censoring proportion. The true value of the slope being log(1.05).

To summarise, when the standard deviation of the error term within years is at 0.05, both the variance and the squared bias is for the most part lower for the method using the *lmec* function. When the standard deviation increase to 0.5 the variance of the estimated slope for each method seems to stay the same up until a 30% censoring. Afterward there is a downwards trend for the substitution method while the two functions using the LMMC model have a similar upwards trend until the larger censoring proportion where the *mixcens* function spikes.

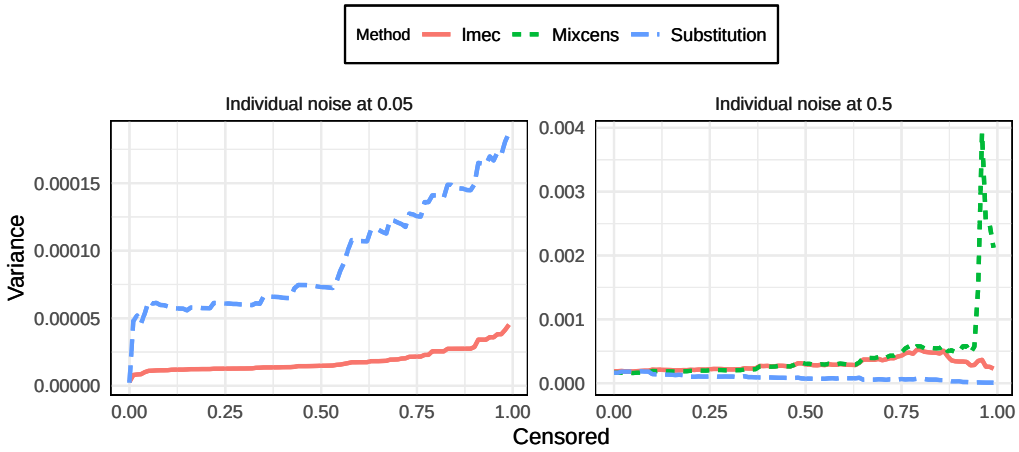


Figure 5a: Standard error of the estimated regression slope as a function of censoring proportion. The true value of the slope being log(1.1).

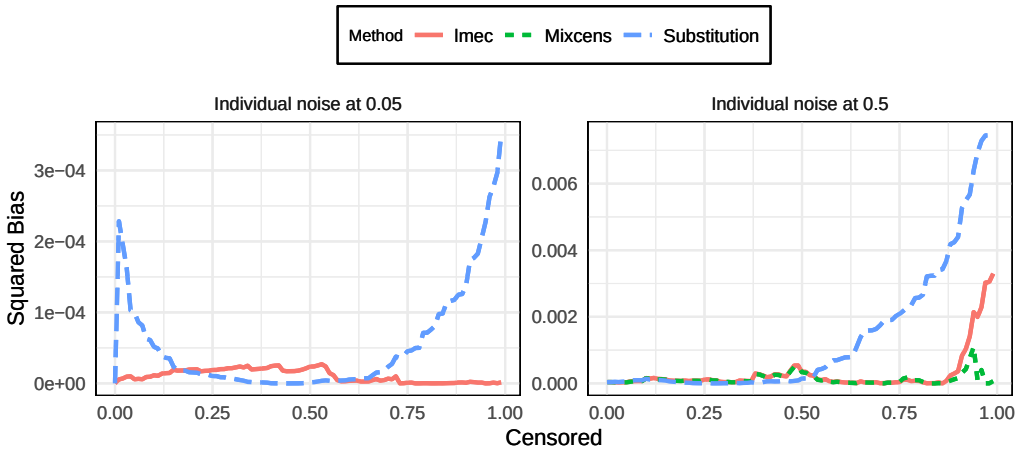


Figure 5b: Squared Bias of the estimated regression slope as a function of censoring proportion. The true value of the slope being log(1.1).

The squared bias for both methods stay around the same up until 30% censoring. For larger censoring the squared bias increase for the substitution method except for at the lowest value of the slope. The *lmec* and *mixcens* functions have an interval past 30% of where the squared bias increase with the interval size decreasing as the value of the slope increase. Continuing to increase the censoring leads to the squared bias decreasing for both functions up until the largest censoring proportions where the *lmec* function starts increasing again. One possible action to decide on the method is to look at the mean squared error (MSE) which is defined as $E[(\hat{\theta}-\theta)^2]$ or equivalently the sum of the squared bias and the variance.

3.2.2 Model Analysis

For comparing the two models, the function *lmec* will be used to analyze the LMMC model in this section due to it being more numerically stable than the *mixcens* function.

Tables 1-4 show a summary of the simulation grouped by the proportion of censored values and the true value of the slope. The squared bias for each estimator was estimated using Monte Carlo methods. For each scenario, the squared bias was calculated as taking the square of the mean of the bias defined as $(\hat{\beta} - \beta)$ over all 100 simulations.

To clarify the notations in the table, the $\text{Var}(\hat{\beta})$ column is the variance of the estimator estimating β, representing the spread of the estimates calculated as $\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\beta}_i - \bar{\beta})^2$ where $\bar{\beta}$ is the mean of all the estimates. On the other hand, the $\text{Se}((\hat{\beta} - \beta)^2)$ column shows the Monte Carlo standard error of the estimate of the bias. For exact calculations of the Monte Carlo standard error, See Appendix. Lastly, the $\text{Se}(\text{Var}(\hat{\beta}))$ column is the Monte Carlo standard error of the empirical variance of the estimates of the regression slope (see Appendix for details).

The first thing that stands out is the fact that whenever at least one of the error terms are not set to a lower value, the LMMC model produced estimates with a larger squared bias than when using the method of substitution. However, when the error terms have less influence, the LMMC model produced unbiased estimates while the substitution method, as shown by El-Shaarawi and Esterby (1992), still fails to produce unbiased estimates.

Another conclusion is the fact that when using substitution, the bias increase as the slope increase while the reverse seems to be true for the LMMC model except for the special case of holding both error terms and the proportion of censored values at a high level (compare Table 3 & 4). When alternating the proportion of censoring the LMMC model gives more biased estimates at higher proportions as to be expected considering there is less information while the substitution method does the same for the larger slope value and at the same time the reverse for a lower value of the slope. The LMMC model does however still give unbiased estimates at small noise for both of the proportion of censoring.

Table 1: Summary statistics of simulations at 30% censored data and a 1% yearly increase. Statistics except coverage is multiplied by 1000.

Individual Sd	Random Effect	Method	$(E[\hat{\beta} - \beta])^2$	Coverage	$\text{Var}(\hat{\beta})$	MSE	$\text{Se}((E[\hat{\beta} - \beta])^2)$	$\text{Se}(Var(\hat{\beta}))$
0.05	High	Substitution	1.6577	0.95	1.3997	3.0574	0.0124	0.1989
		LMMC	2.5258	0.99	2.4872	5.0130	0.0252	0.3535
	Low	Substitution	0.2917	0.08	0.0214	0.3130	0.0003	0.0030
		LMMC	0.0037	0.96	0.0034	0.0071	0.0000	0.0005
0.50	High	Substitution	1.3750	0.97	1.2694	2.6443	0.0098	0.1804
		LMMC	2.0830	0.99	2.0984	4.1813	0.0191	0.2983
	Low	Substitution	0.1818	0.96	0.1780	0.3598	0.0005	0.0253
		LMMC	0.2092	0.97	0.2011	0.4103	0.0006	0.0286
1.40	High	Substitution	2.4242	0.96	2.3546	4.7788	0.0235	0.3347
		LMMC	4.2574	0.97	4.2982	8.5556	0.0558	0.6109
	Low	Substitution	0.9621	0.94	0.9602	1.9223	0.0060	0.1365
		LMMC	1.5630	0.96	1.5293	3.0923	0.0122	0.2174

Table 2: Summary statistics of simulations at 30% censored data and a 5% yearly increase. Statistics except coverage is multiplied by 1000.

Individual Sd	Random Effect	Method	$(E[\hat{\beta}-\beta])^2$	Coverage	Var($\hat{\beta}$)	MSE	Se($(E[\hat{\beta}-\beta])^2$)	Se($Var(\hat{\beta})$)
0.05	High	Substitution	2.0649	0.94	1.2124	3.2773	0.0144	0.1723
		LMMC	2.2694	0.97	2.1116	4.3810	0.0209	0.3001
	Low	Substitution	0.5751	0.00	0.0020	0.5771	0.0002	0.0003
		LMMC	0.0073	0.97	0.0069	0.0142	0.0000	0.0010
0.50	High	Substitution	1.9327	0.96	1.3283	3.2609	0.0141	0.1888
		LMMC	2.3893	0.98	2.3594	4.7487	0.0232	0.3354
	Low	Substitution	0.2268	0.93	0.2187	0.4455	0.0007	0.0311
		LMMC	0.2574	0.94	0.2599	0.5173	0.0008	0.0369
1.40	High	Substitution	2.6944	0.96	2.1502	4.8446	0.0250	0.3056
		LMMC	3.8570	0.98	3.8221	7.6791	0.0477	0.5432
	Low	Substitution	1.2554	0.91	1.1715	2.4269	0.0086	0.1665
		LMMC	1.7484	0.95	1.7655	3.5139	0.0147	0.2509

In the report of Bignert et. al (2017), confidence intervals are often used to give an indication of the true value of the slope. Therefore, it seems reasonable to investigate the coverage for these confidence intervals. In this simulation study, a Wald confidence interval is used defined as $\hat{\beta} \pm 1.96 \cdot se(\hat{\beta})$ where $se(\hat{\beta})$ is the standard error of the estimated slope. The number 1.96 is the 97.5% quantile of a standard normal distribution, putting the confidence level at 95%. For the LMMC model, using maximum likelihood estimates, the standard deviation of the i^{th} parameter in the parameter vector θ is calculated as

$$se(\hat{\theta}_i) = \sqrt{[I(\hat{\theta})]^{-1}_{ii}}$$

where $I(\hat{\theta})$ is the observed Fisher information matrix (Held and Bové, 2014, pp.128). The Fisher information matrix has elements $[I(\hat{\theta})]_{ij}$ being $-\frac{d^2}{d\theta_i d\theta_j} l(\theta)$ where $l(\theta)$ is the log-likelihood function. Since the likelihood function in Section 2.5 is too complex, the derivative is hard, if not impossible to find. Hence, the observed Fisher information matrix is also numerically approximated. The methods for the numerical approximation of the matrix is beyond the scope of this thesis and left out.

The method of substitution however does not estimate using maximum likelihood but instead uses the method of least-squares. Nonetheless, under the assumption of having data following a normaldistribution, the two methods are equivalent. The method of least-squares however is in words, a method of where a regression line is fitted to best fit the data. This is done by minimizing the sum of the distance between the fitted regression line and the observed data. This implies, for the model in Section 3.1.1, minimizing the expression

$$\sum_{i=1}^{11} \sum_{j=1}^{12} (y_{ij} - X_i \beta)^2$$

The value of β which minimizes the expression is the method of least-squares estimate of the regression slope. The standard error of this estimate is thus calculated as

$$se(\hat{\beta}) = \sqrt{\frac{1}{n_{obs}-2}\frac{\sum_{i=1}^{11}\sum_{j=1}^{12}(y_{ij}-x_i\hat{\beta}_{LS})^2}{\sum_{i=1}^{11}\sum_{j=1}^{12}x_i^2}}$$

with n_{obs} being the number of observations for a simulation and $\hat{\beta}_{LS}$ is the estimate of β using the method of least-squares (Sundberg, 2016, pp53). Here the covariates of the model are already centered and thus does not need to be centered in the calculations of the standard error. The coverage was then calculated by looking at the proportion of simulations obtaining a confidence interval containing the true value of the slope. In other words, the coverage was calculated as $\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(\hat{\beta}_{lower,i} \leq \beta \leq \hat{\beta}_{upper,i})$ where $\hat{\beta}_{lower,i}$ is the lower bound of the confidence interval of the i^{th} estimated β in a simulation and when indexed with upper referring to the upper bound of that same confidence interval. The indicator function $\mathbf{1}$ returns a value 1 if the true value of the slope is within the confidence interval and 0 otherwise. When taking a closer look at the coverage of both methods the LMMC model has coverage around 95% in all cases. On the other hand, even though using substitution might have produced less bias in most cases, the coverage is far from 95% for all scenarios. Anytime the error terms are held at a low level, the coverage is close to zero.

Table 3: Summary statistics of simulations at 60% censored data and a 1% yearly increase. Statistics except coverage is multiplied by 1000.

Individual Sd	Random Effect	Method	$(E[\hat{\beta}-\beta])^2$	Coverage	Var($\hat{\beta}$)	MSE	Se($(E[\hat{\beta}-\beta])^2$)	Se($Var(\hat{\beta})$)
0.05	High	Substitution	0.9754	0.99	0.6089	1.5843	0.0048	0.0865
		LMMC	4.1363	0.99	3.9130	8.0493	0.0517	0.5562
	Low	Substitution	0.2954	0.04	0.0184	0.3139	0.0003	0.0026
		LMMC	0.0056	0.98	0.0048	0.0103	0.0000	0.0007
0.50	High	Substitution	1.2397	0.95	0.6064	1.8461	0.0061	0.0862
		LMMC	4.5340	0.97	3.6706	8.2046	0.0549	0.5217
	Low	Substitution	0.1284	0.94	0.1198	0.2482	0.0003	0.0170
		LMMC	0.2612	0.96	0.2637	0.5249	0.0008	0.0375
1.40	High	Substitution	1.3486	0.97	0.8664	2.2150	0.0079	0.1231
		LMMC	4.8886	0.99	4.2765	9.1651	0.0639	0.6078
	Low	Substitution	0.5656	0.94	0.5270	1.0926	0.0026	0.0749
		LMMC	2.1211	0.95	2.1401	4.2612	0.0196	0.3042

The variance of the estimator for the model using fabricated data is to no surprise much lower than that of the LMMC model considering the method of substitution. The variance of the estimators are affected by the level of the error terms. However, the effect is much clearer for the LMMC model than it is for the method of substitution. The inclination of the slope seems to have no major effect on the variance except for once again one special case for the LMMC model, when all factors are set to a high level (compare Table 3 & 4).

What might be of more interest is the effect censoring has on the estimator. For the LMMC model, there is a clear increase in variance whenever the proportion of censoring is larger while the reverse, to no surprise, holds when substituting values. Whenever the error terms are set to low, or the censoring level in combination with the slope both being

high, the variance for the method of substitution is low, resulting in too small confidence intervals.

Table 4: Summary statistics of simulations at 60% censored data and a 5% yearly increase. Statistics except coverage is multiplied by 1000.

Individual Sd	Random Effect	Method	$(E[\hat{\beta}-\beta])^2$	Coverage	$\text{Var}(\hat{\beta})$	MSE	$\text{Se}((E[\hat{\beta}-\beta])^2)$	$\text{Se}(Var(\hat{\beta}))$
0.05	High	Substitution	2.4496	0.78	0.5736	3.0232	0.0117	0.0815
		LMMC	3.4847	0.99	3.2690	6.7537	0.0398	0.4646
	Low	Substitution	0.3438	0.00	0.0023	0.3461	0.0001	0.0003
		LMMC	0.0155	0.91	0.0138	0.0293	0.0000	0.0020
0.50	High	Substitution	2.3765	0.85	0.6095	2.9859	0.0117	0.0866
		LMMC	3.4403	0.99	3.2972	6.7375	0.0395	0.4686
	Low	Substitution	0.3132	0.76	0.1038	0.4170	0.0006	0.0148
		LMMC	0.2454	0.96	0.2432	0.4886	0.0008	0.0346
1.40	High	Substitution	3.0828	0.76	1.1370	4.2199	0.0208	0.1616
		LMMC	6.3173	0.97	5.9581	12.2754	0.0975	0.8468
	Low	Substitution	1.1126	0.79	0.5713	1.6839	0.0053	0.0812
		LMMC	2.1734	0.96	2.1756	4.3490	0.0203	0.3092

The precision of the Monte Carlo estimates of the squared bias is good, as demonstrated by the standard error of the squared bias being close to zero for each scenario.

Figure 9-11 in the Appendix shows each simulated estimate of the slope for both methods plotted against each other with Figure 9 treating each scenario when the standard deviation of the individual noise is set to low, Figure 10 the case of a medium noise and Figure 11 when it is set to high. The first thing that jumps out is how big of an influence both error terms have separately since altering just one of them from low to a higher value instantly results in much more biased estimates for both models.

The figures also show that for each scenario, the estimates of the LMMC model seem to center around the true value of the slope, having around the same proportion of estimates under the true value as over. However the same can not be said for the substitution model. Especially when the error terms have low effect and either the inclination of the slope or the proportion of censored data is larger, not a single unbiased estimate is produced by the model. It's also possible to see that when there is more noise in the data, for most scenarios, especially when the proportion of censoring is larger, the majority of the estimation for the substitution method underestimates the slope, giving a result skewed towards lower values. At the same time, while the substitution method underestimates the value of the slope more often than the LMMC model, the latter does miss by a lot more at some times. This is most likely one of the reasons for the higher bias of the LMMC model.

Further it's shown that the proportion of censoring seems to have a big impact on the correlation between the estimates of the two models. More specifically, whenever the proportion of censoring increases, the reverse goes for the correlation between the estimates. This is most likely an effect from the previously stated fact that the LMMC model keeps centering around the true slope value while the substitution model skews towards lower estimates.

An identical simulation study was made looking at a yearly decrease instead of increase. The results were very similar and will therefore not be analyzed in detail here. One

detail from that study to take into consideration is the fact that when using substitution, the estimates are for most scenarios with higher noise and censoring proportion, centered around a value higher than the true slope. It is also possible to see that the LMMC model in the case of a yearly decrease at some scenarios (see for example Figure 13 in Appendix with 60% censoring) would likewise be centered around another value, however in contrast to the substitution model, a lower value than the actual value of the slope. The tables and figures for the simulation study of a yearly decrease is shown in the appendix.

4 Application

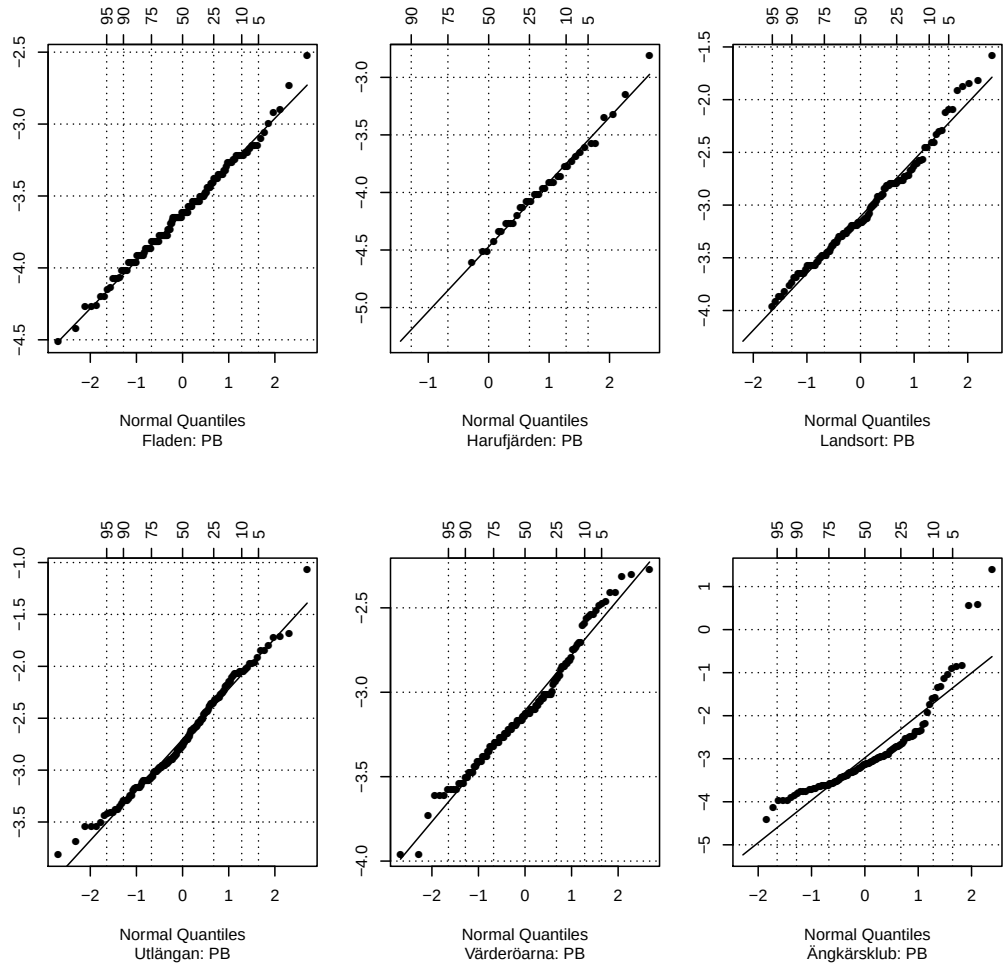
4.1 Probability plots and distribution assumptions

In the dataset used by the museum for their analysis of environmental toxins a large number of different metals were analyzed. Three of which tends to have a rather large proportion of censored values and are the three metals of most interest to analyze in this thesis. These metals are nickel (NI), lead (PB), and chromium (CR). Due to the large difference in concentration levels depending on locations, this study performs one analysis for each location. In the interest of keeping the study on a moderate level and to get the most reliable analyzes, only the locations using at least 10 specimens each year for more than 10 years between 2007-2018 were used resulting in the 6 locations of Fladen, Harufjärden, Landsort, Utlängan, Väderöarna and Ängskärsklubb. For the same reasons, only the concentration level in Herring was considered.

To justify the use of the LMMC model for the dataset, an analysis concerning whether or not the log-normal distribution holds for the concentrations has to be made. For this purpose, plotting the result from the *cenros* function in the *NADA* packages, as used by Helsel (2005) is one way to go. Since there is no information regarding an exact position for a non-detect, only the uncensored data is plotted. Using substitution for the censored data points is of no use since this will result in a different shape of the probability plot dependent on the chosen substitution point. Instead, the proportion of data below each reported limit is calculated and used to fit the uncensored data to the correct percentiles when using a distribution plot. As a result, the uncensored data above the highest reporting limit will have the same positions on the plot as they would have if all data were uncensored. The uncensored values between limits will however be affected by the censored values between these limits, as they should be. Just as bad would be to simply delete the censored values from the dataset, using only the uncensored data when plotting a probability plot considering this would skew the percentiles and the distribution will be incorrect. This will also only show the distribution of the uncensored data, not the entire dataset. For this thesis, the distribution plot will take the logarithmic values of the concentration and plot against the quantiles of a normal distribution. As can be seen in Figures 6-8, in many of the distribution plots, the first point starts around the median or even further to the right. This is the effect of the censored values not being plotted, but at the same time having an impact on the position of the uncensored values. The *cenros* function by default performs a log-normal transformation prior to operations over the data (Lee, 2017) and thereafter performs the reverse transformation. Here, to clearly show that concentrations are on the log-scale, both of these transformations are set to

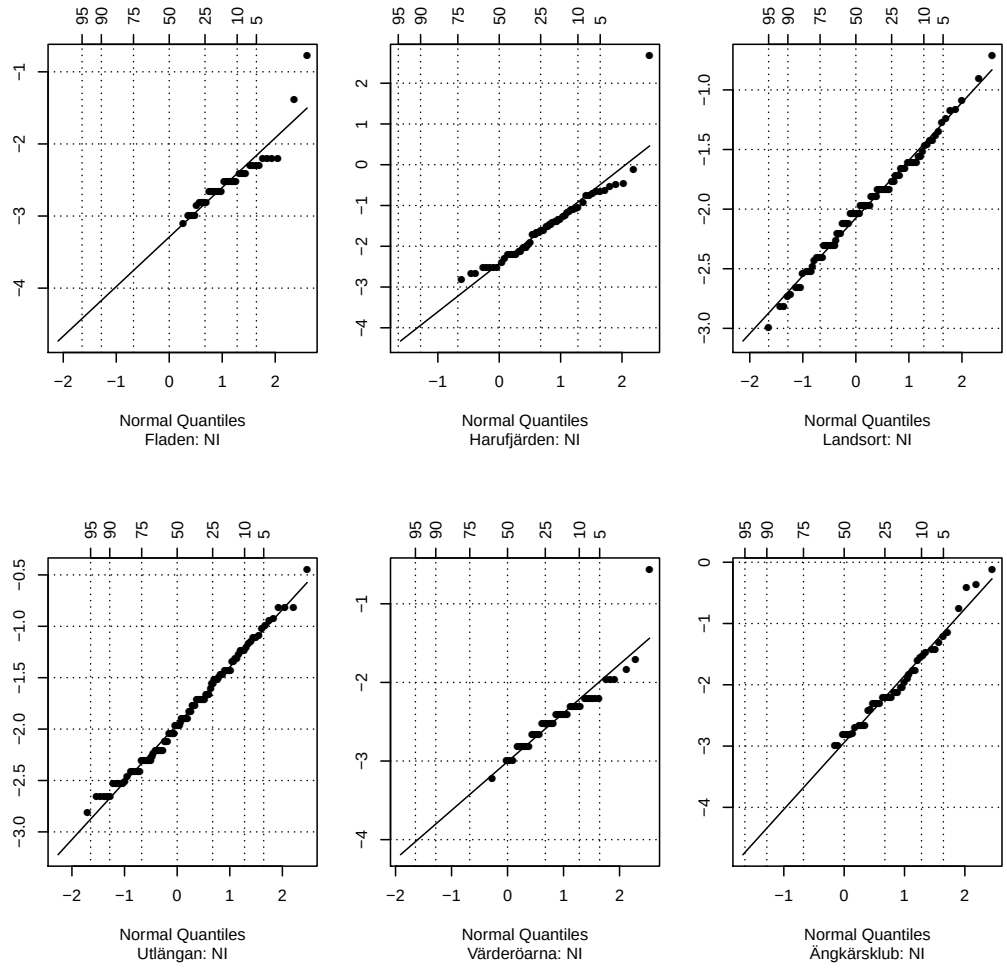
NULL, and instead the logarithmic values for the concentrations are used. Hence, the *cenros* function assumes a log-normal distribution, and so when using the *cenros* results in a distribution plot, a log-normal distribution assumption is tested.

Figure 6: Distribution plot testing the lead concentrations against a log-normal distribution.



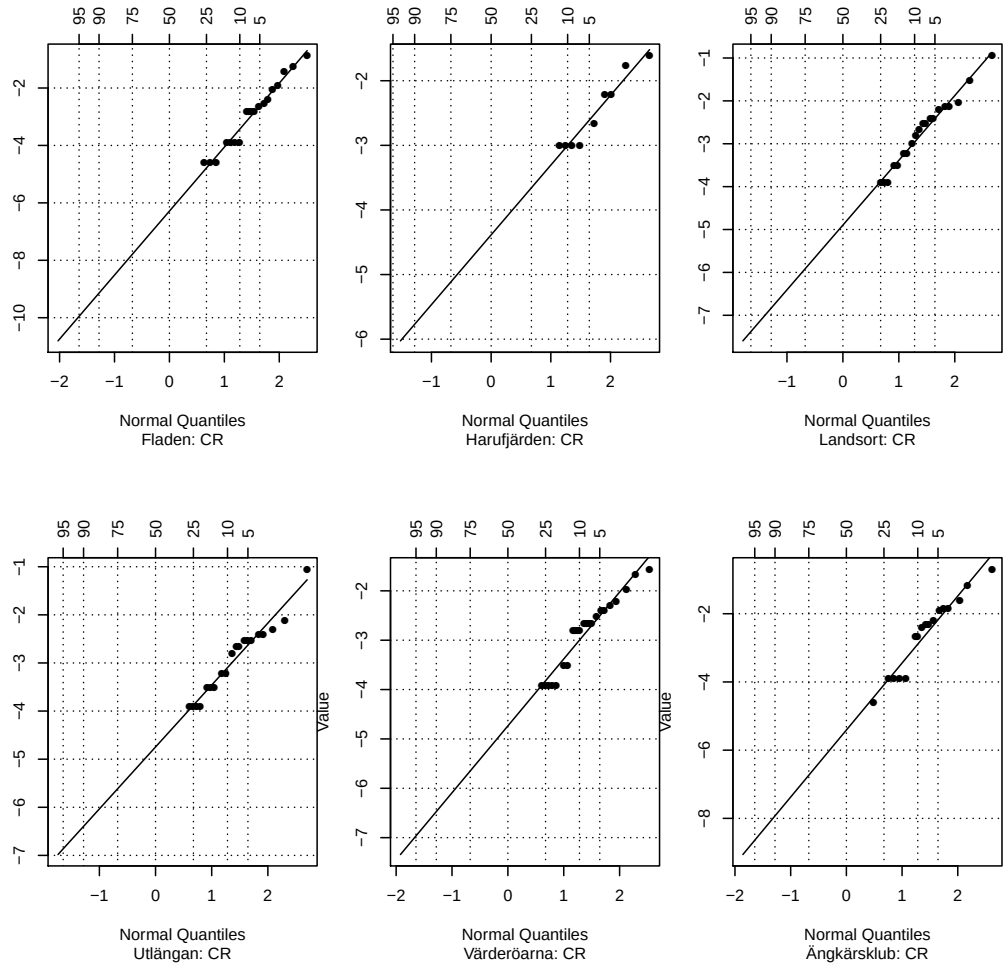
To estimate the percentiles for the uncensored data, regression on ordered statistics (ROS) is used by the *cenros* function. ROS is favorable over MLE when the proportion of censored data is too high (Helsel, 2005, pp. 86) which is the case in some locations for each metal (see Table 5-7), and for every location for chromium. For each combination of metal and location, data points are first given a rank i ranking the data point with the smallest value as $i = 1$. The ranks are then converted to percentiles by giving each point a plotting position p. For the *cenros* function, the position p is given using the Weibull formula $p = i/(n + 1)$ where n is the sample size.

Figure 7: Distribution plot testing the nickel concentrations against a log-normal distribution.



Even though most commercial statistical software use the formula $p = i/n$ (Helsel, 2005, pp.48), the Weibull formula is to be preferred when only using a sample which is a part of the total population. This because when using the formula $p = i/n$, it is stated that the largest value has a zero percent chance of being exceeded. This would be the case if the entire population was used but not with a smaller sample of a population. When points have the same values and therefore ties in ranks, as is the case for censored data, each point is given its own rank.

Figure 8: Distribution plot testing the chromium concentrations against a log-normal distribution.



When the percentiles are calculated, they are fitted against the quantiles of a normal distribution. The uncensored data is used to calculate the slope and intercept for the linear regression between the logarithmic values of the data and the normal quantiles and thus, fitting this line is fitting a log-normal distribution to the observed data (Helsel, 2005, pp.80). Now, looking at Figure 6-8, it seems like a reasonable assumption that the data follows a log-normal distribution seeing that they more or less follow a straight line.

4.2 Applying the LMMC model

In this section, the theory and results of the simulation study is used to compare the results of using the substitution method and the LMMC model on data used by the Swedish Museum of Natural History to analyze time trends of concentration levels of different metals in fish.

4.2.1 Data and Methods

The data material was acquired from Martin Sköld from the Swedish Museum of Natural History. The analysis for metals was prior to 2007 performed by the Department of Environmental Assessment at the Swedish University of Agricultural Sciences (SLU). However, from 2007, this was carried out by the Department of Environmental Science and Analytical Chemistry (ACES), at Stockholm University (SU) (see Section 6, Bignert. et. al. 2017). Therefore, the data prior to 2007 and the data from 2007 and forward are not deemed optimal to analyze together. Hence, only the dataset from 2007-2018 is analyzed in this thesis in contrast to the report by Bignert. et. al (2017) where both an analysis using all data were used as well as an analysis for the most recent ten years (2007-2017 in the case of the report) for the longer time series (Section 7, Bignert. et. al, 2017). Hence, an analysis using the LMMC model as well as the substitution method is made instead of using the result from the report. The dataset contains several variables other than year, location, and metal concentration, for example, length and weight of the specimens, but only the first three mentioned variables are analyzed to follow the line of the published report making it easier to do a good comparison between the two methods.

The logarithmic values of the reported concentrations, in combination with the coherent year and a vector of censoring indicators, indicating whether or not the observations are censored is used with the *lmec* function of the *NADA* package. The proportion of censored values and the standard deviation of the estimated slope is also calculated as well as the standard deviation between years and for individuals at each location. The standard deviation between years was calculated using the same method as in Section 3.1 with the *cenmle* function from the *NADA* packages. For lead, the standard deviation of each location, each year, was under 0.1 except for in 2017 at Ängkärsklubben where it was around 1.2. For nickel, the yearly standard deviation was around $0.05 - 0.2$ except for a couple of instances, Harufjärden having a year with a standard deviation of around 1.4 and Ängkärsklubben having one year at around 0.7. For chromium, the censoring proportion is too high to get a good estimation of the standard deviation. One of the best ways to get an approximation of the standard deviation is however to use the same method which resulted in the standard deviation for each year being at around $0.05 - 0.15$ for the most part, having a couple of year and location combinations with a small increase. One that stood out was Utlängan in 2008 which had a standard deviation of over 5. The standard deviation of the individual specimens for each location can be seen in Table 5-7.

Since that the data as used in Bignert. et. al. (2017) and the data for this thesis differs, an analysis is also made using the exact method as would have been used in the report would this dataset have been available, namely a simple linear regression analysis using the substitution method with a substitution value of $1/\sqrt{2}$ times the limit of quantification for the censored values. The results for both methods, including the proportion of censoring and the standard deviation of the location, can be seen in Tables 5-7. A glance at the tables shows that the proportion of censored values are for the most part different for the three metals. Low to none for all but one location when looking at lead, with the anomalous proportion at Harufjärden most likely since the concentration levels in that location were much lower than in the other locations. For nickel, the proportion is either quite low at 15% censoring or higher at around 50% censoring while chromium has above 80% censoring for all locations.

4.2.2 Analysis

Starting the analysis by taking a look at lead (see Table 5) it's possible to see that for each location having a low proportion of censored data both models produce similar estimates for the slope as to be expected. However, for Harufjärden having a proportion of censoring at around 70% the same can not be said. The LMMC model produces an estimate of -0.0215 on the log-scale implying a yearly decrease in concentration by 2.1% on the original scale. The substitution model gives an estimate of 0.0301 implying an increase of 3% per year. From the simulation study it is shown that the LMMC model produces unbiased estimates when the noise is low as in this case. Figure 3-4 shows the higher squared bias of the substitution model looking at around 0.7 on the x-axis. The same figures show the variance as being much larger for the substitution model when the implied slope value is around 0.01 ($\log(1.01)$) aswell as when it's around 0.05. This would suggest that the substitution model would wrongly estimate the yearly concentration development severely. Following the information in Section 7 of Bignert et. al (2017), a yearly increase of 3% would imply a doubled concentration level in 24 years which could commence some sort of action being taken to stop the increase when in fact, a decrease by 2% mean the concentration level would be halved in 35 years.

When looking at nickel (Table 6) the standard deviation for individual specimen increase a small bit, landing somewhere around $0.04 - 0.13$ except for in Harufjärden. As in the case of analyzing lead, both models produced similar estimates for the slope for all locations but Harufjärden. The LMMC model produced an estimate representing an increase of 5.7% per year on the original scale while substitution showed a yearly decrease by 2.2%. The noise for individual specimens being close to 0.3 and the between-year noise being around 0.1 for all but one year in combination with 53% censoring indicates on how to judge the difference in estimation for the two models. Figure 3-4 shows a similar squared bias for both models and a slightly lower variance for the substitution model when the censoring proportion is around 50% and the slope implies a yearly increase by 5%. For the lower slope value however, the variance is larger for the substitution model. Looking at Figure 10 in the Appendix, it's possible to see that for a scenario like this, if the true value of the slope were to be around 0.05 on the logarithmic scale as estimated by the LMMC model, the majority of the estimates when using the substitution method will be lower than the true value of the slope. The figure also shows that for a higher yearly variance, the amount the substitution method undervalues the true value increase by a lot. The yearly variance in this scenario may not be as high as in the simulation for all years, but it is higher than the lower level of the study. To summarise, there are indications that if the estimate produced by the LMMC model is correct, it is plausible that the substitution method could give an estimate as low as it did. At the same time, assuming the truth is closer to the estimate acquired by substitution, a yearly decrease by around 2%, the LMMC model infrequently overestimates the value of the slope but indeed more often underestimate the value of the slope (see Figure 9 & 10 in the Appendix). It is also possible to see that implying a 1% decrease never resulted in the LMMC model obtaining an estimation as high as 0.05 whenever the noise was held at a moderate level. This indicates that the reality most likely resemble that of the estimate given by the LMMC model. Once again following the line of the report by Bignert et. al. (2017, Section 7), in a scenario like this using the substitution method one could conclude that no actions need to be taken considering a 2% yearly decrease imply a halved concentration level after 35

years. Nevertheless, the truth is that a close to 6% increase would result in a doubled concentration level in almost 10 years.

The data for chromium has a large proportion of censored data. However, the censoring is done, as for most if not all metals, with several different limit of quantifications as can be seen by the distribution plots (see Figure 8) by the fact that uncensored data is plotted around the 25th percentile even though the censoring is at over 80%. Therefore, it is still possible to get a decent regression analysis of the data and estimate of the slope. The differences in the estimates between the two methods for chromium is clear. Due to the large amount of censoring for this dataset the standard deviations should not be looked at as an exact measurement however the chromium concentrations range between possible values around $0-0.5$ which gives some indication of how large the standard deviation can be. The largest differences between the estimates are around 0.2. The most extreme cases being both Fladen and Utlängan which have reported a yearly increase of roughly 6- and 4% respectively while the LMMC model produces estimates implying a yearly decrease of $11-12\%$.

Figure 3-5 showcase the fact that the LMMC model using the *lmec* function produce more or less unbiased estimates when the censoring proportion is over 80% and the standard deviation for the error term for individual specimens is 0.05. For the same scenario, the substitution method does produce estimates with some bias. The variance of the estimated regression slope is larger for the LMMC model when there is an underlying yearly increase of 5 or 10%.

Since there is extensive censoring for the chromium data, a smaller, more targeted simulation study was applied. This study uses the same structure as described in Section 3.1 with the change of focusing the censoring at 80%. As a result of the immense time investment needed to run simulations at these levels of censoring the study was confined to looking at slope values implying a 5 and 10% yearly increase in combination with the lowest level of the standard deviation for the between-years error term and a standard deviation of 0.05 and 0.5 for the within-years error term. In other words, the study looks at the factor levels found in the chromium data. The study could also be done using a yearly decrease but as shown in the simulation study in Section 3, the results are very similar. Table 12 in the Appendix shows the summary statistics and performance measurements for this study. These results show that the substitution method produces estimates with higher squared bias and lower coverage while the LMMC model has a larger spread of the estimated regression slope. Figure 15 in the Appendix confirms the result of the LMMC method producing more or less unbiased estimates when the standard deviation of the within-years error term is set to 0.05. It's also possible to see that the substitution method fails to produce a single unbiased estimate of the slope. Especially when the standard deviation is increased to 0.5, the estimates produced by the substitution method is centered around a much lower value than the true inclination of the slope. The LMMC method however is still centered around the true value. It was also shown in the study looking at a yearly decrease that the result holds with the difference being that the substitution method was centered around a higher value than the true value of the slope. Also shown in the larger study was the fact that whenever the standard deviation of the between-years error term increased, so did the difference between the true value of the slope and the value of which the estimates by the substitution method were centered around. In the smaller study the lowest level of between-year variance was used and thus

the true variance is most likely larger than the used value. In conclusion, combining all of these results, it would seem more likely that the truth lay closer to the estimates produced by the LMMC model. However, it's impossible to say anything with complete certainty.

Table 5: Results of application of both models on data for lead concentrations in herring (Estimated slope on Log-Scale)

Location	$\hat{\beta}_{LMMC}$	sd($\hat{\beta}_{LMMC}$)	$\hat{\beta}_{Sub}$	sd($\hat{\beta}_{Sub}$)	% Censored	Sd of Location
Fladen	-0.0015	0.0138	-0.0328	0.0075	4.3103	0.0157
Harufjärden	-0.0215	0.0367	0.0301	0.0120	70.5426	0.0079
Landsort	0.0191	0.0188	0.0191	0.0131	3.0534	0.0289
Utlängan	0.0090	0.0190	0.0090	0.0118	0.0000	0.0391
Väderöarna	0.0042	0.0190	0.0045	0.0089	0.0000	0.0161
Ängskärsklubb	0.0662	0.0483	0.0723	0.0272	6.1947	0.1179

Table 6: Results of application of both models on data for nickel concentrations in herring (Estimated slope on Log-Scale)

Location	$\hat{\beta}_{LMMC}$	sd($\hat{\beta}_{LMMC}$)	$\hat{\beta}_{Sub}$	sd($\hat{\beta}_{Sub}$)	% Censored	Sd of Location
Fladen	-0.0418	0.0414	-0.0550	0.0105	53.0172	0.0855
Harufjärden	0.0559	0.0971	-0.0235	0.0221	53.4884	0.2879
Landsort	-0.0096	0.0172	-0.0115	0.0125	14.5038	0.0721
Utlängan	-0.0007	0.0199	-0.0029	0.0131	17.3611	0.0961
Väderöarna	-0.0188	0.0528	-0.0152	0.0120	61.1940	0.0412
Ängskärsklubb	-0.1105	0.0469	-0.0849	0.0185	56.6372	0.1327

Table 7: Results of application of both models on data for chromium concentrations in herring (Estimated slope on Log-Scale)

Location	$\hat{\beta}_{LMMC}$	sd($\hat{\beta}_{LMMC}$)	$\hat{\beta}_{Sub}$	sd($\hat{\beta}_{Sub}$)	% Censored	Sd of Location *
Fladen	-0.1158	0.0871	0.0653	0.0181	84.4828	0.0431
Harufjärden	0.0395	0.0866	0.1202	0.0141	93.0233	0.0446
Landsort	-0.0427	0.1023	0.0329	0.0203	84.7328	0.0856
Utlängan	-0.1241	0.1085	0.0408	0.0176	86.1111	0.0431
Väderöarna	0.1078	0.0955	0.1403	0.0160	82.8358	0.0650
Ängskärsklubb	-0.0384	0.1323	0.0286	0.0230	84.9558	0.1641

* Due to high proportion of censoring the standard deviations are unreliable.

5 Discussion

In this thesis the model for a linear mixed-effect model with censored response variables having only one random effect and one covariate was determined. The likelihood function was derived and the maximum likelihood estimate of the slope was then examined using the LMMC model and the estimates using the method of least-square as used by the substitution method. The two R functions *mixcens* and *lmec* produce estimates using the LMMC model and the two functions together with the *lm* function used for the substitution method were compared. It was shown that the *mixcens* function had numerical problems when the data contained little noise and thus only the *lmec* and *lm* function with substitution were compared for lower level of noise. At this level the *lmec* function produced more or less unbiased estimates but for larger values of the slope had a higher variance than the substitution method. When the noise was increased all three functions were compared. For a larger proportion of censoring, the LMMC model produced less biased estimates compared to the substitution method while the variance of the estimate was lower for the substitution method. At lower censoring proportions all three functions seemed to be equal. For the higher values on the slope combined with the higher levels of censoring proportion, the *mixcens* produced estimates with lower bias than the *lmec* function but with higher variance.

After comparing the functions, the substitution method using the *lm* function and the LMMC model using the *lmec* function was studied in a simulation study. The study demonstrated the inability of producing unbiased results when fabricated data is used to calculate the estimate of the slope from a linear regression model. It further showed the reverse for the LMMC model, producing unbiased estimates when not having much noise in the data no matter if the value of the slope represents a yearly increase or decrease by either 1 or 5% or the proportion of censored data were 30 or 60%. From the study it was also possible to see that for every scenario, the estimates by the LMMC model almost always centered around the true value of the slope. In opposition, using substitution results in estimates centered around a much lower (or higher in the case of a yearly decrease) value when the error terms have more of an impact, especially when the proportion of censored data is larger. This effect increased as the inclination of the regression slope increased. At the same time, when the error terms were larger, using substitution on average produced less biased estimates.

It was shown that for small variance in the error terms, the method of substitution produces too small confidence intervals ending up in coverage below the chosen confidence level and in some cases even coverage of 0%, rendering confidence intervals for this method useless. The confidence intervals for the LMMC model at the same time had coverage at, or close to, the confidence level for all scenarios.

It was possible to see that properties of the estimates for the LMMC model seemed to follow what was expected when altering the levels of the factors, and the reverse following for the substitution method. For example, as the proportion of censoring increased, introducing more uncertainty into the dataset, the mean squared error of the estimates from the LMMC model increased, while the reverse ended up being true when using fabricated data.

The precision of the squared bias in the simulation study, as seen by the monte carlo

standard errors, is quite high having a standard error close to zero at all times. Although, granted that there was more time, more iterations for each scenario should be made, increasing the strength of the results of the simulation study. Simultaneously, more levels of the factors should be tested, especially for the between-years variance. The maximum number of iterations for the EM-algorithm should also be increased, alternatively a fixed stopping criterion could be established. Considering the EM-algorithm does not always converge towards a global maximum, sometimes converging towards a local maximum or a saddle point, different starting values should be tested when using the algorithm and then choosing the parameter values resulting in the largest likelihood.

The two models were applied to data used by the Swedish Museum of Natural History which showed there is a chance of acquiring estimates much higher or lower than the truth when using the substitution method. It was shown that with large probability, the yearly change in concentration of both lead and nickel at Harufjärden was misjudged using the substitution method. This method showed a yearly increase of lead at around 3% and yearly decrease of nickel at around 2% when in fact the LMMC model showed a yearly decrease by over 2% in lead and a yearly increase of nickel by almost 6%. It was also shown that for the chromium concentrations the two models produced estimates with a large difference for almost all locations. The biggest differences being at over 20 percent units. A specified analysis was made, mimicking the dataset for the chromium concentrations which showed that the LMMC model more likely showed the truth than the substitution model. However, no certainties can be given. A more in-depth study is suggested, focusing strictly on the scenario with over 80% censoring.

6 Appendix

6.1 Theory

This part is used to find the Monte Carlo standard error (MCSE) for the squared bias using the delta method. Definitions are learned from Held and Bové (2014).

6.1.1 Definitions

Definition 1: A real-valued statistic $T_n = h(X_{1:n})$ based on a random sample from $X_{1:n}$ with a probability mass or density function $f(x;\theta)$ for some unknown parameter θ computed as to make inference on θ is called an estimator.

Definition 2: Let $X_1, X_2, ...$ be a sequence of random variables. Convergence in probability towards X denoted as $X_n \xrightarrow{p} X$ is defined as

$$P(|X_n - X| > \epsilon) \to 0 \qquad \text{as n} \to \infty \;\; \forall \epsilon > 0$$

Convergence in distribution denoted as $X_n \xrightarrow{d} X$ is true if

$$P(X_n \leq x) \to P(X \leq x) \qquad \text{as n} \to \infty$$

for all x of which the distribution function is continuous.

Definition 3: An estimator T_n is asymptotic unbiased for θ if

$$E[T_n] = \theta \qquad \text{as n} \to \infty$$

for all θ in the parameter space Θ. Further, from Held and Bové (2014, pp. 98), the maximum likelihood estimate is asymptotic unbiased.

6.1.2 Slutsky's Theorem

To prove the Delta Method, Slutsky's theorem is needed.

Theorem 1: If $W_n \to W$ in distribution and $Z_n \to a$ in probability where a is a constant then

$$\begin{cases} W_n + Z_n \xrightarrow{d} W + a \\ W_n Z_n \xrightarrow{d} aW \end{cases}$$

The proof is left out.

6.1.3 Delta Method

Theorem 2: Suppose X_n is a sequence of random variables X_i for $i = 1, 2, ..., n$. If this sequence satisfy $\sqrt{n}(X_n - \tau) \xrightarrow{d} N(0, \psi^2)$, then for a function $g(\cdot)$, continiuously differentiable (at least in a neighborhood of τ) and a specific value of τ and $g'(\tau) \neq 0$ it holds that

$$\sqrt{n}(g(X_n) - g(\tau)) \xrightarrow{d} N(0, \psi^2 g'(\tau)^2)$$

Proof: The taylor expansion of $g(X_n)$ around $X_n = \tau$ is

$$g(X_n) \approx g(\tau) + g'(\tau)(X_n - \tau) + C$$

where C is a remainder term going towards zero as X_n tends towards τ. It was assumed that X_n satisfy the central limit theorem, and thus, $X_n \xrightarrow{p} \tau$ and so it follows that $C \xrightarrow{p} 0$. Now, rewriting gives

$$\sqrt{n}(g(X_n) - g(\tau)) = \sqrt{n}g'(\tau)(X_n - \tau) + C$$

Now, as the assumption where made that $\sqrt{n}(X_n - \tau) \xrightarrow{d} N(0, \psi)$, then $\sqrt{n}g'(\tau)(X_n - \tau) \xrightarrow{d} N(0, \psi^2 g'(\tau)^2)$ once again by Held and Bové (2014, pp. 337). Now using slutsky's theorem with $W_n = \sqrt{n}g'(\tau)(X_n - \tau)$, $Z_n = C$ and $a = 0$ it holds that the right hand side converge to $N(0, \psi^2 g'(\tau)^2)$ as desired.

It follows from the delta method then that if $\hat{\theta}$ is an estimator of θ, and $h(\theta)$ is some transformation of θ with properties as in Theorem 2

$$Se(h(\hat{\theta})) = Se(\hat{\theta})\big|\frac{d}{d\theta}h(\hat{\theta})\big|$$

6.1.4 MCSE for the squared bias and variance

First, define the bias as $E[\hat{\theta} - \theta]$ and the estimate of the bias as

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i - \theta$$

where $\hat{\theta}_i$ is the estimate of θ in the i^{th} simulation and are all independent and identically distributed.

Now, using the central limit theorem and the fact that $E[\hat{\theta}] = \theta$ when n is large enough.

$$\sqrt{\frac{n}{\text{Var}(\hat{\theta})}} \cdot \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i - \theta \overset{a}{\sim} N(0, 1)$$

and thus by Held and Bové (2014, pp. 337)

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i - \theta \overset{a}{\sim} N(0, \frac{Var(\hat{\theta}_1)}{n})$$

Now, the estimated empirical variance for the estimator is

$$Var(\hat{\theta}_1) = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{\theta}_i - \bar{\theta})^2$$

where $\bar{\theta} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i)$ is the sample mean of the estimates. And thus, the Monte Carlo standard error is

$$\text{Se}(\hat{\text{bias}}) = \sqrt{\frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\theta}_i - \bar{\theta})^2}$$

where $\hat{\text{bias}}$ is the estimate of the bias and n is the number of simulations.

Now, applying the delta method for the Monte Carlo standard error with the function $g(x) = x^2$ gives the Monte Carlo standard error for the squared bias.

$$\text{Se}(g(\hat{\text{bias}})) = \text{Se}(\hat{\text{bias}})\left|2 \cdot \hat{\text{bias}}\right|$$

The Monte Carlo standard error of the empirical standard error of the regression slope as calculated by Morris, White and Crowther (2019) is

$$\text{Se}(\sqrt{Var(\hat{\theta})}) = \frac{\sqrt{Var(\hat{\theta}_1)}}{\sqrt{2(n-1)}}$$

and thus, using the deltha method, the Monte Carlo standard error of the empirical variance of the estimated regression slope is

$$\text{Se}(\sqrt{Var(\hat{\theta})})\left|2 \cdot \sqrt{Var(\hat{\theta})}\right|$$

6.2 Tables and Figures

6.2.1 Estimation Plots for the Yearly Increase Simulation Study

In this section, the plots where each estimated regression slope for both the LMMC model and the substitution method are plotted against each other is found. These are the plots for the simulation study with an implied yearly increase. The plots are divided into the three different levels of the standard deviation for the within-years error term and contain the true value of the slope visualized as the vertical and horizontal lines.

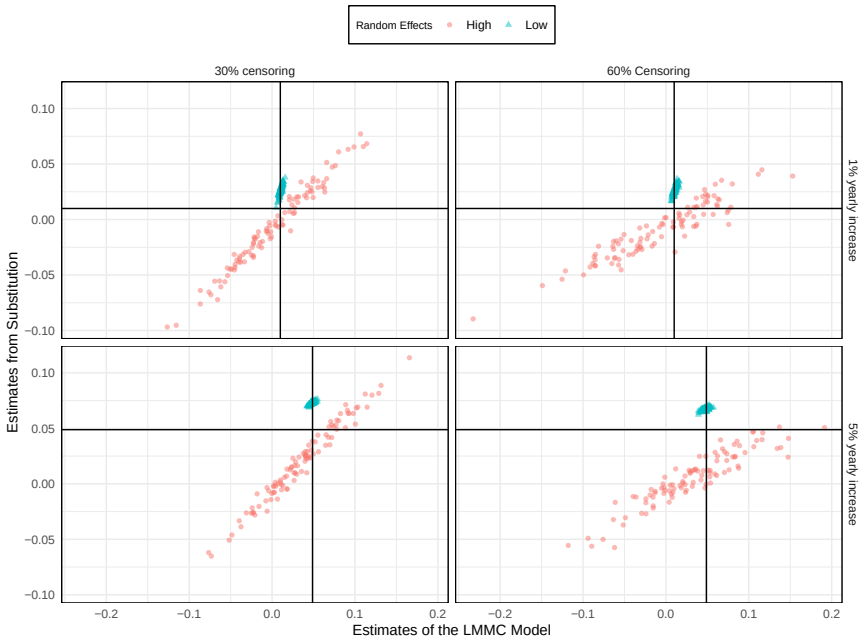


Figure 9: Plotting the estimated slopes for the substitution method and the LMMC model against eachother having the variance of individual specimens set to low. The vertical and horizontal lines correspond to the true value of the slope.

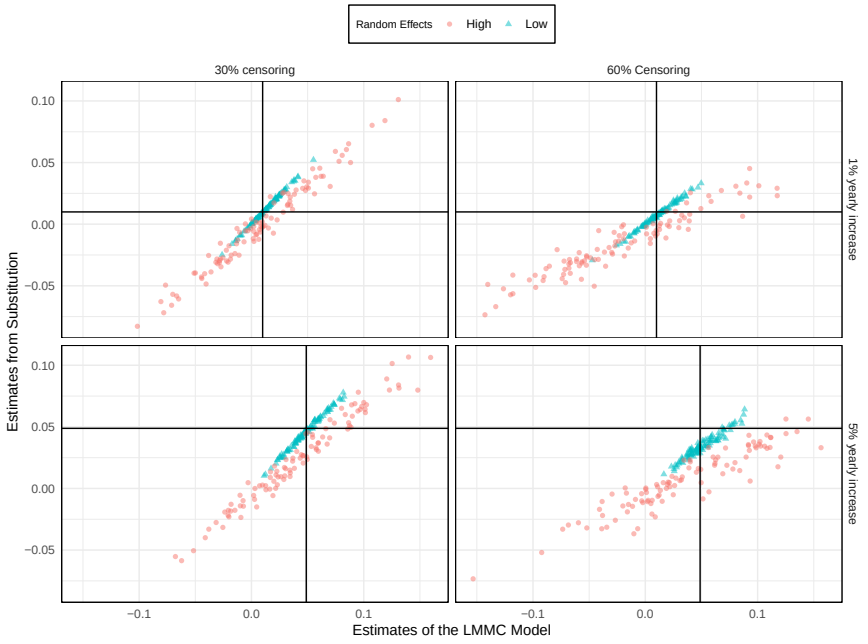


Figure 10: Plotting the estimated slopes for the substitution method and the LMMC model against eachother having the variance of individual specimens set to medium. The vertical and horizontal lines correspond to the true value of the slope.

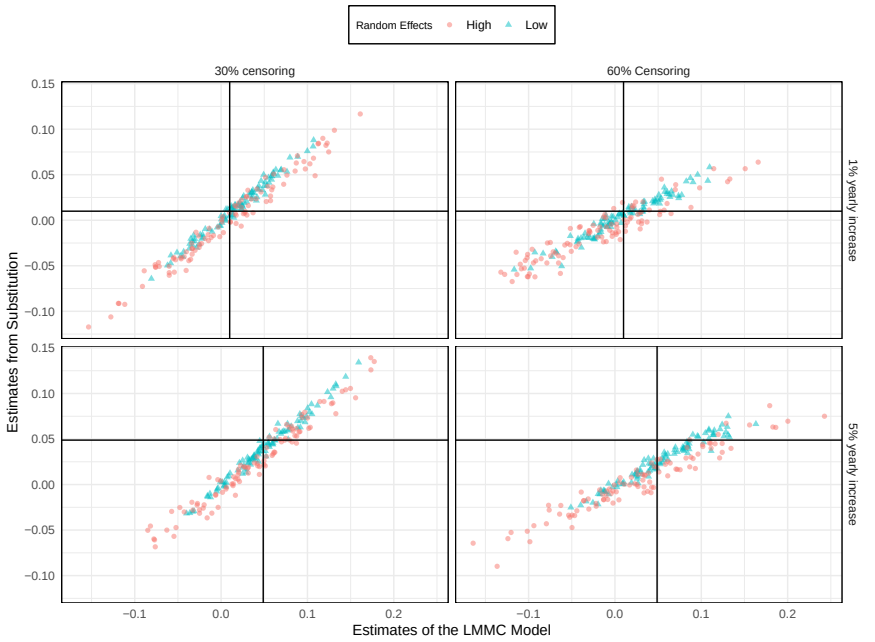


Figure 11: Plotting the estimated slopes for the substitution method and the LMMC model against eachother having the variance of individual specimens set to high. The vertical and horizontal lines correspond to the true value of the slope.

6.2.2 Yearly Decrease Simulation Study

This part of the Appendix contains results from the simulation study done working with an implied yearly decrease. The simulation study was made in the same manner as presented in Section 3 with the only difference of having a yearly decrease instead of a yearly increase.

Table 8: Summary statistics of simulations at 30% censored data and a 1% yearly decrease. Statistics except coverage is multiplied by 1000.

Individual Sd	Random Effect	Method	$(E[\hat{\beta}-\beta])^2$	Coverage	$\text{Var}(\hat{\beta})$	MSE	$\text{Se}((E[\hat{\beta}-\beta])^2)$	$\text{Se}(Var(\hat{\beta}))$
0.05	High	Substitution	1.4065	0.97	1.3538	2.7603	0.0104	0.1924
		LMMC	2.4301	0.98	2.4287	4.8588	0.0240	0.3452
	Low	Substitution	0.2972	0.08	0.0234	0.3206	0.0003	0.0033
		LMMC	0.0035	0.93	0.0036	0.0071	0.0000	0.0005
0.50	High	Substitution	1.5015	0.95	1.2695	2.7710	0.0107	0.1804
		LMMC	2.4319	0.97	2.1655	4.5974	0.0226	0.3078
	Low	Substitution	0.1443	0.95	0.1457	0.2900	0.0003	0.0207
		LMMC	0.1617	0.96	0.1627	0.3244	0.0004	0.0231
1.40	High	Substitution	1.6963	0.99	1.5824	3.2787	0.0135	0.2249
		LMMC	2.9367	0.99	2.8219	5.7587	0.0312	0.4011
	Low	Substitution	0.6541	0.98	0.6469	1.3010	0.0033	0.0919
		LMMC	1.0254	0.99	1.0302	2.0555	0.0066	0.1464

Table 9: Summary statistics of simulations at 60% censored data and a 1% yearly decrease. Statistics except coverage is multiplied by 1000.

Individual Sd	Random Effect	Method	$(E[\hat{\beta}-\beta])^2$	Coverage	$\text{Var}(\hat{\beta})$	MSE	$\text{Se}((E[\hat{\beta}-\beta])^2)$	$\text{Se}(Var(\hat{\beta}))$
0.05	High	Substitution	0.8553	0.97	0.5245	1.3798	0.0039	0.0745
		LMMC	5.2978	0.96	3.2722	8.5699	0.0606	0.4651
	Low	Substitution	0.2878	0.06	0.0164	0.3042	0.0002	0.0023
		LMMC	0.0040	0.98	0.0040	0.0080	0.0000	0.0006
0.50	High	Substitution	0.7000	0.98	0.5429	1.2428	0.0033	0.0772
		LMMC	4.3646	0.98	3.3436	7.7082	0.0505	0.4752
	Low	Substitution	0.1160	0.93	0.1093	0.2253	0.0002	0.0155
		LMMC	0.2452	0.98	0.2466	0.4918	0.0008	0.0351
1.40	High	Substitution	0.9486	0.98	0.8598	1.8083	0.0056	0.1222
		LMMC	4.8994	0.99	4.3182	9.2176	0.0644	0.6138
	Low	Substitution	0.4705	0.93	0.4611	0.9316	0.0020	0.0655
		LMMC	1.7618	0.96	1.7715	3.5333	0.0148	0.2518

Table 10: Summary statistics of simulations at 30% censored data and a 5% yearly decrease. Statistics except coverage is multiplied by 1000.

Individual Sd	Random Effect	Method	$(E[\hat{\beta}-\beta])^2$	Coverage	$\text{Var}(\hat{\beta})$	MSE	$\text{Se}((E[\hat{\beta}-\beta])^2)$	$\text{Se}(Var(\hat{\beta}))$
0.05	High	Substitution	1.2467	0.97	1.1854	2.4320	0.0086	0.1685
		LMMC	2.7119	0.97	2.3016	5.0134	0.0260	0.3271
	Low	Substitution	0.5447	0.00	0.0023	0.5470	0.0002	0.0003
		LMMC	0.0093	0.97	0.0090	0.0183	0.0000	0.0013
0.50	High	Substitution	1.6463	0.99	1.6271	3.2734	0.0133	0.2313
		LMMC	3.3051	0.99	2.9677	6.2729	0.0360	0.4218
	Low	Substitution	0.2042	0.91	0.2021	0.4063	0.0006	0.0287
		LMMC	0.2452	0.95	0.2459	0.4911	0.0008	0.0350
1.40	High	Substitution	2.2340	0.97	2.1427	4.3767	0.0207	0.3046
		LMMC	3.7165	0.98	3.7492	7.4657	0.0455	0.5329
	Low	Substitution	1.2882	0.91	1.2343	2.5225	0.0091	0.1754
		LMMC	1.9899	0.95	1.9867	3.9766	0.0177	0.2824

Table 11: Summary statistics of simulations at 60% censored data and a 5% yearly decrease. Statistics except coverage is multiplied by 1000.

Individual Sd	Random Effect	Method	$(E[\hat{\beta}-\beta])^2$	Coverage	$\text{Var}(\hat{\beta})$	MSE	$\text{Se}((E[\hat{\beta}-\beta])^2)$	$\text{Se}(Var(\hat{\beta}))$
0.05	High	Substitution	0.7156	0.98	0.7197	1.4353	0.0038	0.1023
		LMMC	8.0507	0.91	4.5347	12.5854	0.1084	0.6445
	Low	Substitution	0.3137	0.00	0.0026	0.3163	0.0001	0.0004
		LMMC	0.0129	0.96	0.0130	0.0259	0.0000	0.0018
0.50	High	Substitution	0.7497	0.99	0.6871	1.4368	0.0039	0.0977
		LMMC	5.7689	0.98	3.9854	9.7543	0.0728	0.5665
	Low	Substitution	0.3228	0.75	0.1138	0.4366	0.0007	0.0162
		LMMC	0.3105	0.97	0.3010	0.6114	0.0011	0.0428
1.40	High	Substitution	0.9600	0.99	0.7258	1.6858	0.0052	0.1032
		LMMC	4.1230	0.99	3.6437	7.7667	0.0498	0.5179
	Low	Substitution	1.2237	0.74	0.6046	1.8284	0.0060	0.0859
		LMMC	2.2110	0.97	2.2281	4.4391	0.0209	0.3167

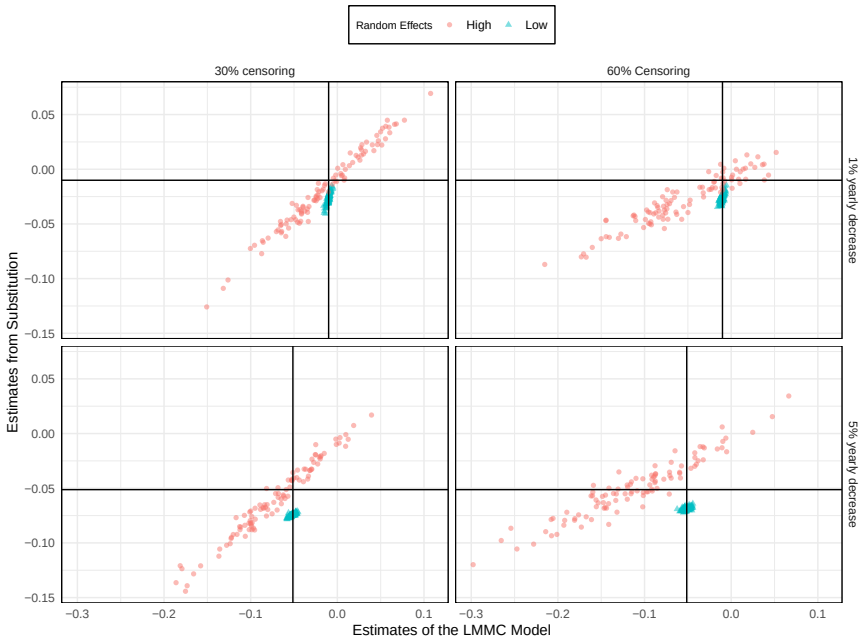


Figure 12: Plotting the estimated slopes for the substitution method and the LMMC modelagainst eachother having the variance of individual specimens set to low. The vertical and horizontal lines correspond to the true value of the slope.

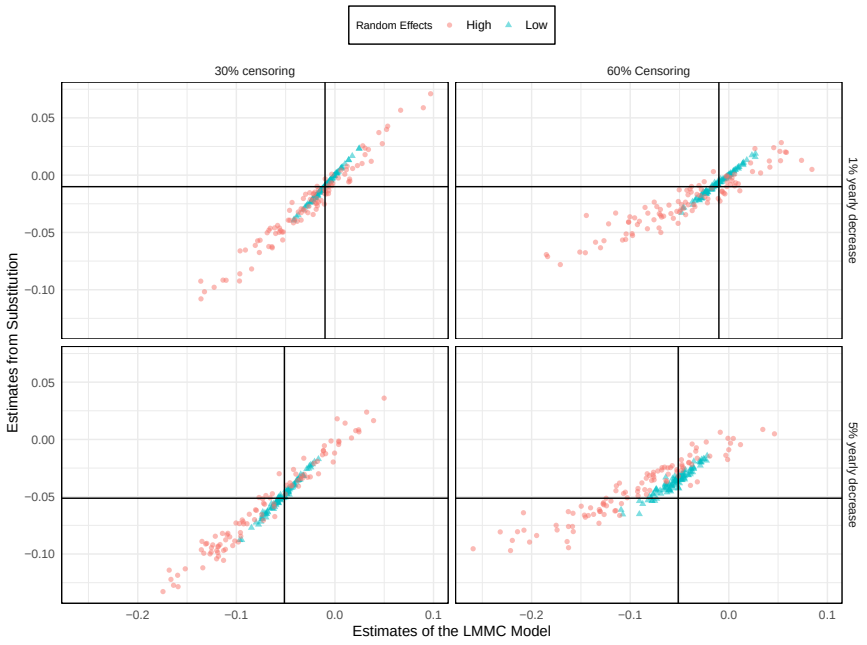


Figure 13: Plotting the estimated slopes for the substitution method and the LMMC model against eachother having the variance of individual specimens set to medium. The vertical and horizontal lines correspond to the true value of the slope.

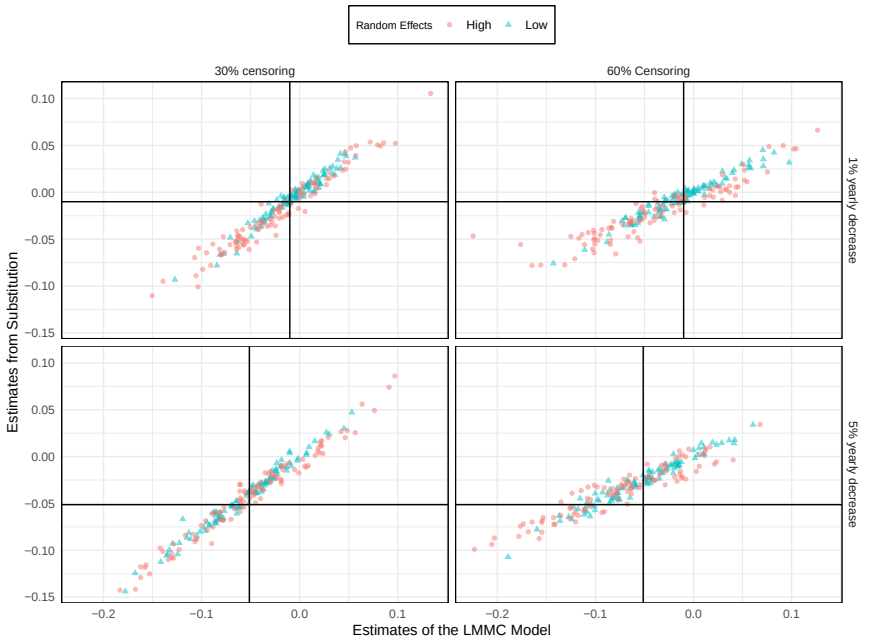


Figure 14: Plotting the estimated slopes for the substitution method and the LMMC model against eachother having the variance of individual specimens set to high. The vertical and horizontal lines correspond to the true value of the slope.

6.2.3 Focused Simulation Study for Chromium Data

This is the result of the aimed simulation study performed to get a better understanding of the results of the chromium data in Section 4.2.2.

Table 12: Summary statistics of simulations at 80% censored data. Statistics except coverage is multiplied by 1000

Individual Sd	Random Effect	Method	$(E[\hat{\beta}-\beta])^2$	Coverage	$\text{Var}(\hat{\beta})$	MSE	$\text{Se}((E[\hat{\beta}-\beta])^2)$	$\text{Se}(Var(\hat{\beta}))$
5 % Yearly Increase	0.05	lmec	0.0268	0.93	0.0180	0.0448	0.0000	0.0026
		Substitution	0.0582	0.38	0.0017	0.0599	0.0000	0.0002
	0.50	lmec	0.4474	0.95	0.4518	0.8992	0.0019	0.0642
		Substitution	2.9537	0.00	0.0664	3.0201	0.0048	0.0094
10 % Yearly Increase	0.05	lmec	0.0183	0.94	0.0157	0.0340	0.0000	0.0022
		Substitution	0.1772	0.01	0.0043	0.1815	0.0001	0.0006
	0.50	lmec	0.4220	0.95	0.4213	0.8433	0.0017	0.0599
		Substitution	0.9089	0.07	0.0640	0.9729	0.0015	0.0091

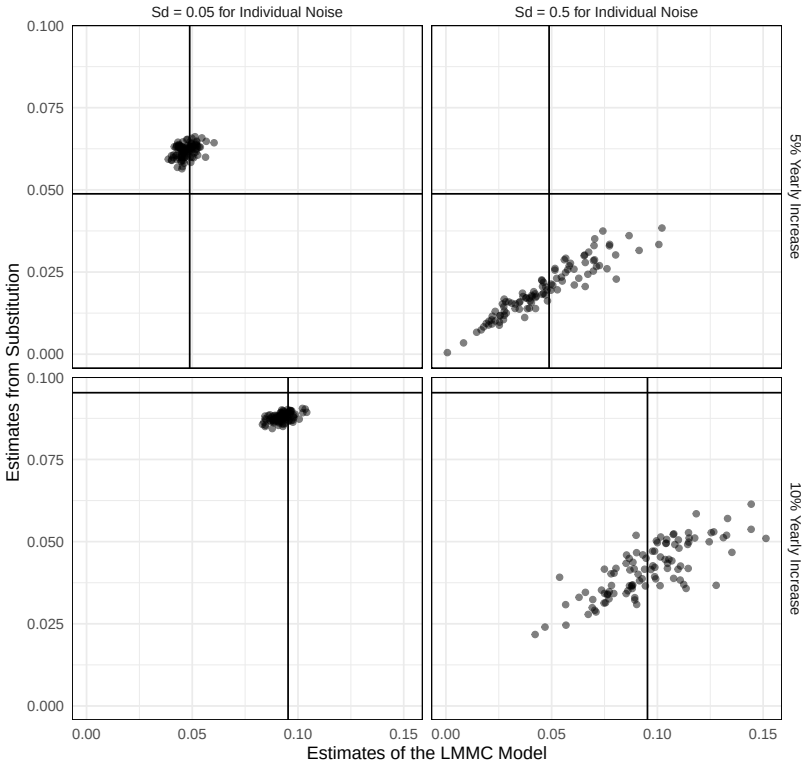


Figure 15: Plotting the estimated slopes for the substitution method and the LMMC model against eachother. The vertical and horizontal lines correspond to the true value of the slope. The data have a censoring proportion of 80%

7 References

1) Helsel D.R., (2006), Fabricating data: how substituting values for censored observations can ruin results, and what can be done about it. *Chemosphere*, Vol. 65, (No. 11), pp. 2434–2439, doi: https://doi.org/10.1016/j.chemosphere.2006.04.051

2) Chung C.F., (1990), Regression analysis of geochemical data with observations below detection limits, *Computer Applications in Resource Estimation*, in G. Gaal and D. F. Merriam (Eds.), Pergamon Press, London, pp. 421–433, doi: https://doi.org/10.1016/B978-0-08-037245-7.50032-9

3) Lee T.L and Go O.T, (1997), Survival Analysis, *Annual Review of Public Health*, vol.18, (No. 1), pp. 105-134, doi: https://doi.org/10.1146/annurev.publhealth.18.1.105

4) Chay K.Y. and Honore B.E., (1998), Estimation of censored semiparametric regression models: an application to changes in Black–White earnings inequality during the 1960s. *Journal of Human Resources*, Vol.33, (No. 1), pp. 4–38, doi: 10.2307/146313

5) Pinheiro J.C and Bates D.M, (2000), Mixed-Effects Models in S and S-PLUS (1. ed.), New York: Springer

6) Laird N. M. and Ware J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, Vol 38. (No. 4), pp. 963–974., DOI: 10.2307/2529876

7) Bignert A., Danielsson S., Faxneld S., Ek C., Nyberg E. (2017). Comments Concerning the National Swedish Contaminant Monitoring Programme in Marine Biota, 2017, 4:2017, Swedish Museum of Natural History, Stockholm, Sweden, Retrieved from the website of the Museum of Natural History: http://nrm.diva-portal.org/smash/get/diva2:1090746/FULLTEXT01.pdf

8) Eaton M. L. (1983). Multivariate Statistics: a Vector Space Approach. John Wiley and Sons. pp. 116–117. ISBN 978-0-471-02776-8

9) Held L, Bové D.S, (2014), Applied Statistical Inference (1. ed.), New York: Springer

10) Dempster A. P., Laird N. M., Rubin, D. B. (1977), Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 39, (No. 1), pp. 1-38, Retrieved from the website jstor: https://www.jstor.org/stable/2984875?seq=1

11) Thompson M .L. and Nelson K. P., (2003), Linear regression with Type I interval- and leftcensored response data. *Environmental and Ecological Statistics* Vol. 10, 221–230. Retrieved from the website of the University of Washington: http://faculty.washington.edu/mlt/Thompson%202003b.pdf

12) El-Shaarawi A. H., Esterby S.R.(1992), Replacement of censored observations by a constant: An evaluation. *Water Research*, Vol 26. (No. 6), pp. 835-844, doi: https://doi.org/10.1016/0043-1354(92)90015-V

13) Helsel D.R., (2005), STATISTICS FOR CENSORED ENVIRONMENTAL DATA USING MINITAB AND R (2. ed.), Hoboken, New Jersey: John Wiley & Sons, pp. 62-69, Inc., ISBN 978-0-470-47988-9

14) Vaida F., Liu L. (2009), Fast Implementation for Normal Mixed Effects Models With Censored Response. *Journal of Computational and Graphical Statistics* Vol 18. (No. 4), 2009 - Issue 4, doi: https://doi.org/10.1198/jcgs.2009.07130

15) Lee L (2017). NADA: Nondetects and Data Analysis for Environmental Data. R package version 1.6-1. https://CRAN.R-project.org/package=NADA

16) Nelder J. A. and Mead R. (1965). A simplex algorithm for function minimization. *Computer Journal*, Vol. 7, (No. 4), pp. 308-313. doi: 10.1093/comjnl/7.4.308

17) Morris T. P, White I. R. and Crowther M. J. (2019), Using simulation studies to evaluate statistical methods, *Statistics in Medicine*, Vol 38, (No. 11), pp. 2074-2102, doi: https://doi.org/10.1002/sim.8086

18) Burton A., Altman D. G., Royston P., Holder R. L. (2006), The design of simulation studies in medical statistics. *Statistics in Medicine*. Vol 25, (No 24), pp. 4279–4292, doi: https://doi.org/10.1002/sim.2673

19) Sundberg R., (2016), Kompendium i Lineära Statistiska Modeller, Department of Mathematics, Stockholm University

20) Wu J.C.F, (1983), On the Convergence Properties of the EM Algorithm, *Annals of Statistics*, Vol. 11, (No. 1), pp. 95-103, doi: Retrieved from the website jstor: https://www.jstor.org/stable/2240463?seq=1