Kandidat Skrivandet

Anton Holm

2020-02-24

Contents

1 Introduction

- Maybe explain more about how *cenmle* calculates the sd.

In some studies, it's difficult to achive exact values in an experiment. It could for example be the reason that patients stops coming back to the hospital, making it impossible to estimate the length of survival after an operation or as in the case of this thesis, the instruments not being able to detect too small concentrations of toxins in samples. These type of data is called censored. One of the most common error being made when analysing censored data is fabrication. The analysists simply substitute the non-detects with a fraction (often one half) of the quantification- or detetection 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 it's own fabricated patterns. This could result in 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 were 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 have been increasing for years. Causes of an increase in concentration have been missed, remedies goes 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 analyse 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 going on to 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. Thereafter the methods are applied on real data from the Swedish Museum of Natural History and results are analysed.

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 statistical research of environmental toxicants as part of the National Swedish Contaminant Programme in marine biota. One of the programs conducted regards analysing long term time trends of several toxins in Swedish waters and to estimate the rate of change. The models used to analyse such time trends are at the moment elemental and disregards much of the data collected. One of the more common, but nonetheless crucial oversights, concerns building models and drawing conclusions from fabricated data due to data being censored.

1.2 Data

The report from Bignert at 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 speciemens independent of each other are analysed for a large number of toxins. For some species, the analysis is done for pooled samples containing a number of speciemens in each pool. To reduce the between-years variation, each sampling area tries to analyse speciemens of the same sex and age. However, the variation can not be reduced to zero and other parameters effects 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 equipments not being able to detect small enough quantities of toxins, a *Limit of Detection (LOD)* is set of which any concentration level under the LOD is reported as zero by the equpment. 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)*. This portion of the data is reported as the negative LOQ and later when analysed 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 effect the concentration of toxins and these attributes being effected by sampling site, this thesis will analyse sampling areas individually.

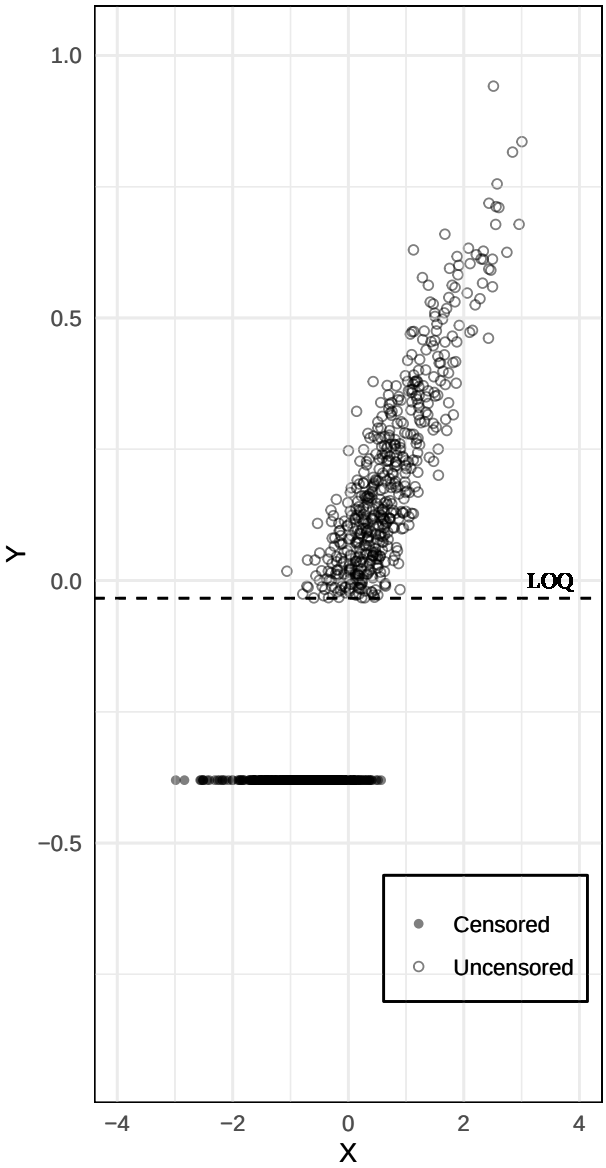


Figure 1a: Example of 1000 log−normal distributed observations censored at LOQ

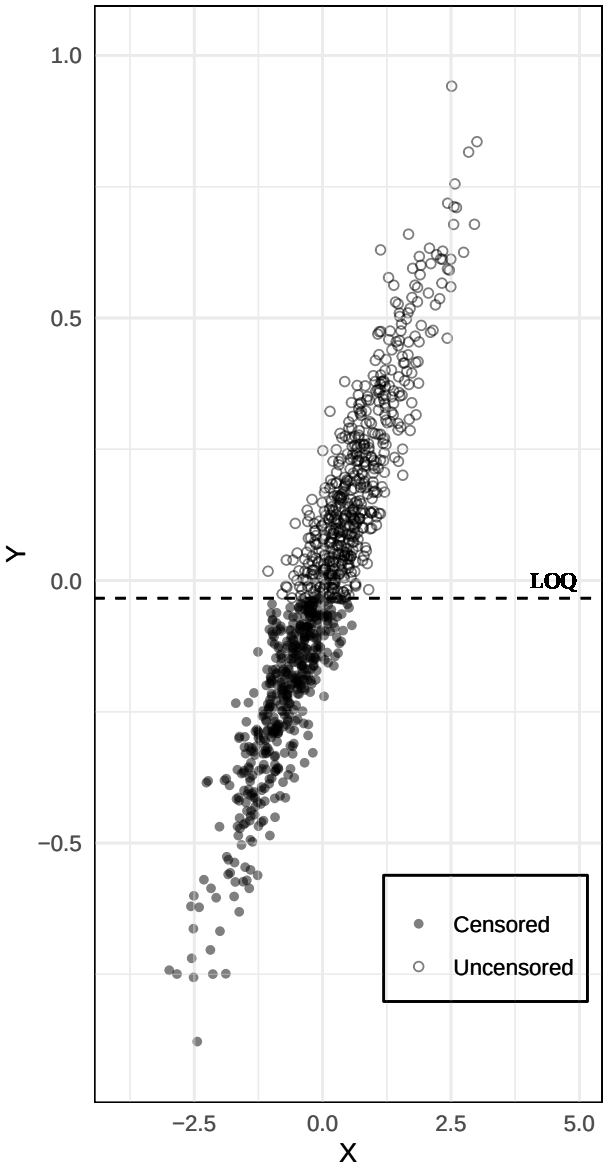


Figure 1b: Example of the same 1000 log−normal distributed observations without censoring

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 of 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 is 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 dimension response vector for group i, β the p dimensional vector of fixed-effect parameters, $\mathbf{b}_i$ the q 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 dimension vector of error terms 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 analyses 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 paramaters, often unknown, of which the function $f(\mathbf{x};\theta)$depends on excluding the realisation $\mathbf{x}$ from $\mathbf{X}$. In the case above, this vector contains the effect parameter vector β aswell as the variance parameters for $\mathbf{b}$ and ϵ. For a response variable $\mathbf{X}$ with observations $\mathbf{X} = \mathbf{x}$ having a probability mass or density function $f(\mathbf{x};\theta)$, depending on the observations $\mathbf{x}$ and $\theta \in \mathbf{\Theta}$ being the often unknown parameter vector taking values in the paramater space $\mathbf{\Theta}$, the Likelihood Function is given by $L(\theta;\mathbf{x}) = f(\mathbf{x};\theta)$. Using the definition of Held and Bové (2014), the likelihood function is the probability mass or density function of the observed data $\mathbf{x}$ viewed as a function of the paramater vector θ. The maximum likelihood estimate of θ denoted as $\hat{\theta}_{MLE}$ is then given as the parameter vector maximising the likelihood function.

2.4 Tobit Model

The Tobit model is characterized by the latent regression equation

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

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:

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

(Note that this function does not integrate to 1 and thus is not a probability function). Using the same method as for a normal simple linear regression model, we further deduce

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

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 seperate group t having n_t speciemens. The between-year variance is the same for each speciemen in the same group while the within-year variance is the same for every speciemen 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 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 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 dimensional vector of the random effect (between-years) b_t and ϵ the n_t dimensional vector with the within-years variance for each speciemen $\epsilon_i, i = 1, 2, ..., n_t$. 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 $\mathbf{\Sigma} = (a_{ij}) \in \mathbb{R}^{n_t \mathrm{x} n_t}$ 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 partioning 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 vector of all the observed, non-censored values and $\mathbf{y}_t^c$ the n_t^c vector of all censored observations before being censored, the same following for $\mathbf{x}_t$ being partioned into 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 $y_t^c|y_t^o$ is also multivariate normally distributed with mean vector and covariance matrix as follows

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

here $\Sigma_t^{oo^{-1}}$ is the inverse of Σ_t^{oo}. Denote $\Phi_t^{c|o}(\cdot)$ as the conditional distribution function of y_t^c given 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

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

Considering the museum is working on analysing timetrends 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 paramater 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. To simplify the calculations, the *log-likelihood function* $l(\beta) = \log[L(\beta)]$ is often used instead of the likelihood function. In light of the fact that the natural logarithm is a monotone and injective function, the paramater vector maximising $l(\beta)$ is the same paramater vector maximising $L(\beta)$.

Now, due to the fact that 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. Therefor, 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 data-set 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 produce 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 produce the same estimation as the MLE considering the fact that the algorithm will stop, either after some number of iterations decided before hand or when $|\theta^{(i)} - \theta^{(i-1)}| < \epsilon$ for some determined $\epsilon > 0$. Once again using the definition of the conditional probability density function, we can 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) 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 data-set $\mathbf{Z}$ conditioned on the observed data $\mathbf{X}$ and the i:th estimate $\theta^{(i)}$ we get

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

where we denote the left hand side 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 untill one of the mentioned reason to stop the algorithm has been reached.

3 Simulation

The existence of bias for estimates where fabricated data were used have 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 was 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 data-set there are two things to look at. Firstly, one of the models needs to be picked where one is more advanced and time consuming when working with many analyses (LMMC) and the other is working with fabricated data (Substitution). Secondly, if the former is chosen, there are two functions which could be used in R, the *lmec* function from the package with the same name or the *mixcens* function, constructed by Martin Sköld at the Swedish Museum of National History. This study will analyse both.

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 individual specimen, the latter representing noise between-years, was used to investigate different conditions. The between-years variance are 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 normaldistribution with mean 0 and different variances for different scenarios.

3.1.2 Scenario Design

There are countless of scenarios to consider but this simulation study takes a closer look on four 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 betweel small and large for the between-year noise.

resulting in 48 different scenarios. The limit of quantification was in other words put at the value representing 30% and 60% censored data if both the intercept and slope were to be put to 0. Hence the proportion of censored data are affected by the slope. The exact values of 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 individual and yearly level. One of the lower and higher noise of the individual specimen for a location was chosen aswell as the standard deviation between these two were chosen to be included in the simulation study. For the between-years noise, the lower values represents most common levels of noise found while the larger values represents some of the more extreme cases. For the noise of each individual 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 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 seperately, 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.

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 $\text{LOQ}/\sqrt{2}$ to continue mimicing the museum) and the maximum likelihood method were both used. The data-sets were simulated using the model described above and the R function *rlnorm* 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 from the package with the same name produced by Vaida and Liu (2009). The *lmec* package determine 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 $\mathbf{y}_t^c$. Consequently the observed data as denoted by X in section 2.6 is for each year the vector $\mathbf{y}_t^o$ from section 2.5. Further, the likelihood function in section 2.5 for each t is simply the product of the likelihood function of the observed data and the likelihood function of the unobserved data. The complete likelihood function is simply 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 paramater 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 paramater 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 as defined in section 2.6 is maximized with regard to any θ still left in the term. Keep in mind that θ is a random vector while $\theta^{(i)}$ is a vector of scalars. 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 were decided due to the immense time effort needed for the lmec function when using a data-set with high proportion of censored data.

The *mixcens* function uses the r package *mnormt* in order 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.

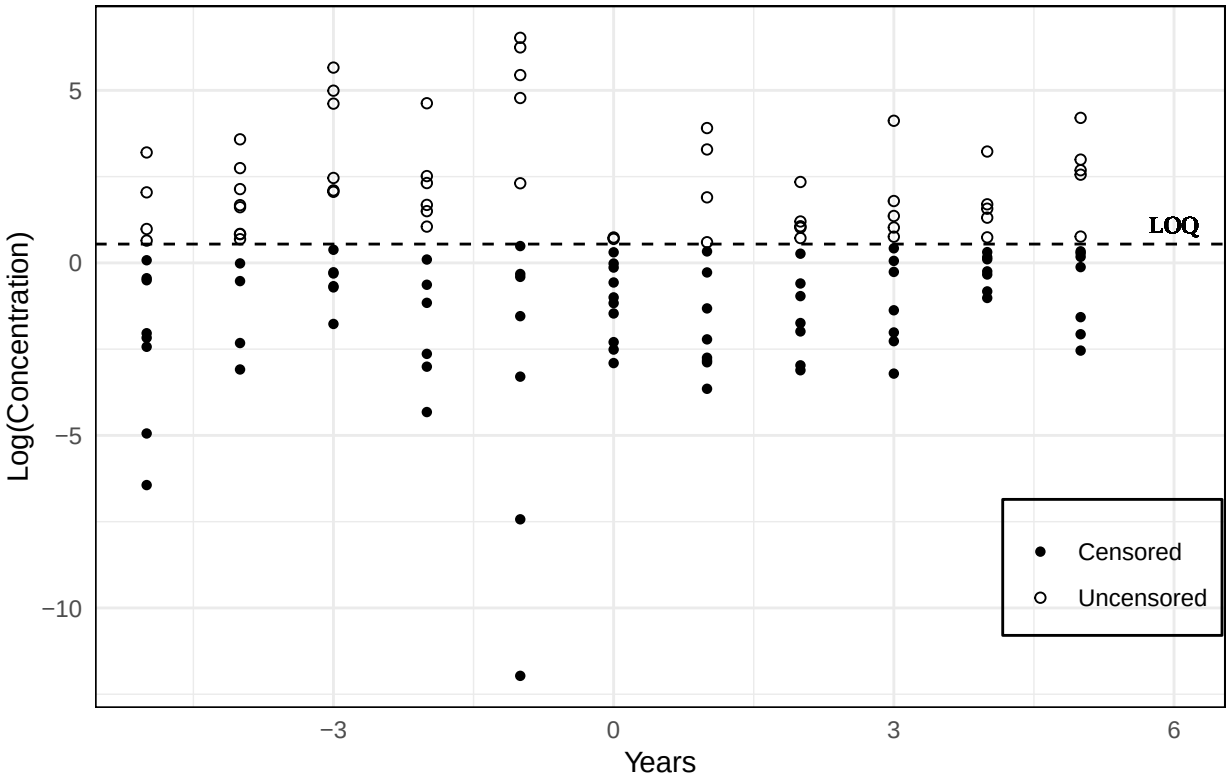


Figure 2: Simulated data with 60% censoring, large variations and slope representing 5% yearly increase holding both error terms at a high variance level.

Figure 1 shows one of the simulated data-sets when computing at a censoring level of 60% with a slope implying a 5% yearly increase, where X is to represent different years and the y-axis serve as an illustration of the logarithmic value of concentration of an arbritary metal found in a specimen. Seing next to no correlation between the covariate and the response is to be expected in an environment where changes happens 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 **nummer figur** show the result when comparing the three methods *lm* with substitution, LMMC model with the function *lmec* and the LMMC model with the function *mixcens*. For a fixt data set, 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 data-set is of no interest due to the fact that both the *lmec* and the *lm* functions produced estimates and standard errors at zero and the *mixcens* function is unable to calculate the inverse of the hessian due to numerical errors The squared-bias and standard errors for each model were then plotted as a function of censoring proportion. The noise were set to what has shown to be the most common, a standard deviation for individual specimen set at 0.5 and the between-year noise set to the lower scale. The slope were altered using the values implying a 1, 5 and 10% yearly increase on the original scale. Figure **nummer 1% ökning** gives some interesting results. First of all, when the inclination of the slope is close to zero it appears the methods barely differs for data with low proportion of censoring. Had the proportion of data been around $30 - 80\%$, the substitution method is to be prefered. Higher proportion of censoring than

that and it's a decision whether or not the squared bias or standard error is of more importance. However, then the decision is between the substitution method having a lower standard error or the *lmec* function having a lower squared bias. The *mixedcens* function seem to work as well as the *lmec* function up until the highest proportions of censoring where the standard error deviates. When the inclination of the slope increases to log(1.05) (see Figure *nummer 5%*), the same analysis as for the smaller inclination holds up until a censoring proportion of around 65%. Higher than this and the standard error behaves as it did for the previous slope value while the squared bias differs. It increases fast for the substitution method while decreasing for the *mixcens* function. Once again it's a choice between favoring lower squared bias (*mixcens*) versus lower standard error (*lmec*). As the value of the slope continuous to increase, this time to a value of log(1.1), a pattern emerge. 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 standard error for the substitution method while the *lmec* and *mixcens* function perform about the same. After over 85%, the substitution method continue it's behavior while the standard error of the *lmec* function reduces in contrast to it's squared bias which increases while the reverse holds true for the *mixcens* model. One possible action to decide on the method is to look at the mean squared error (MSE) which is defined as the sum of the squared bias and the standard error squared.

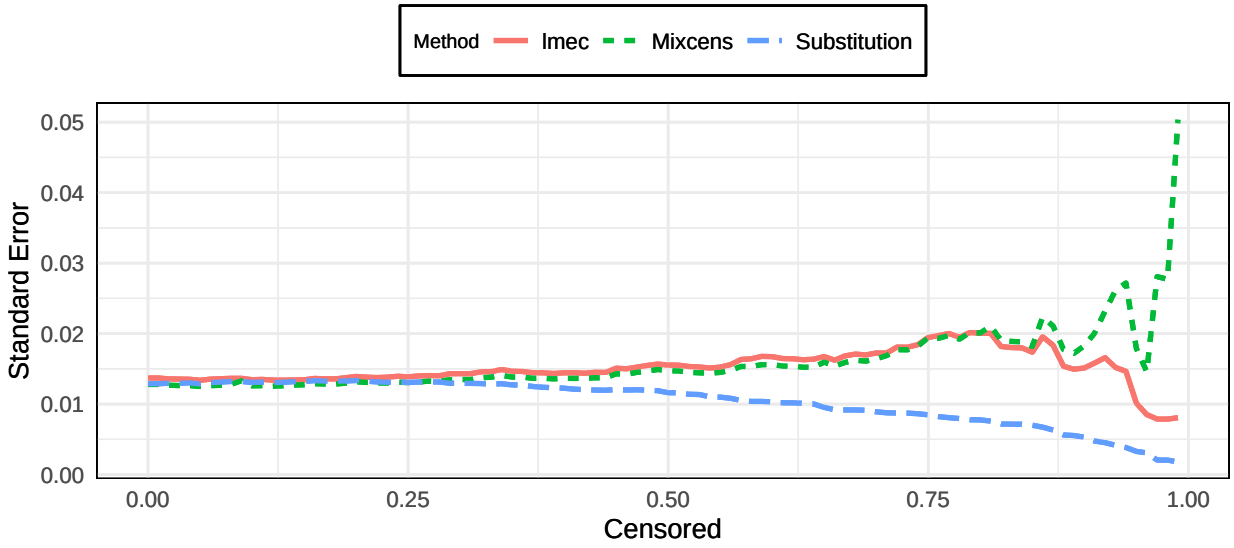


Figure 6a: Standard error of the estimated regression slope as a function of censoring proportion. The true value of the slope being log(1.01).

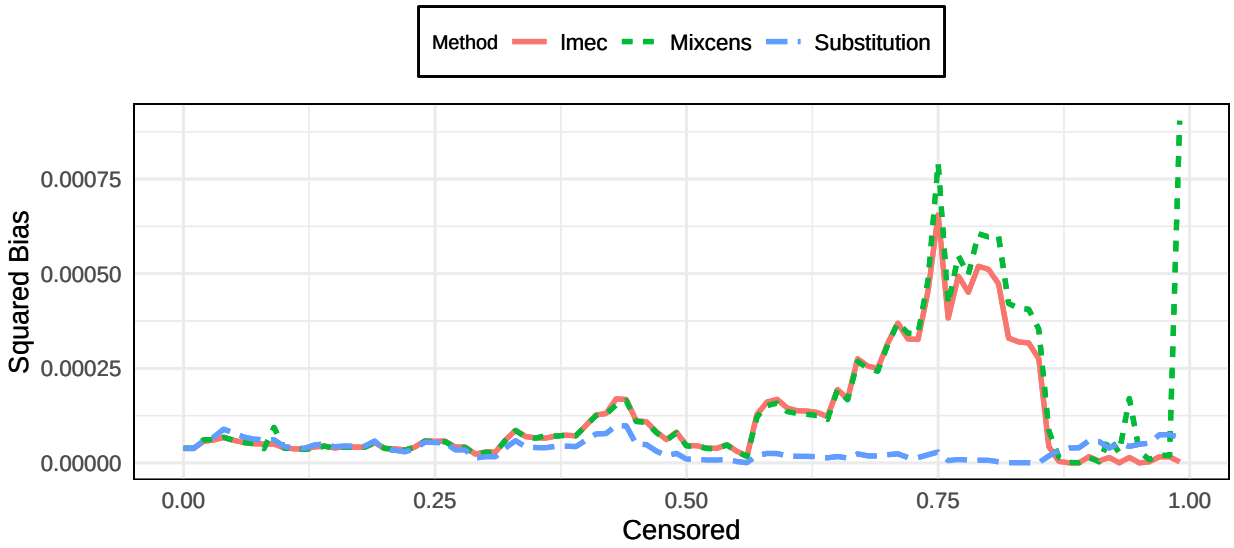


Figure 6b: Squared Bias of the estimated regression slope as a function of censoring proportion. The true value of the slope being log(1.01).

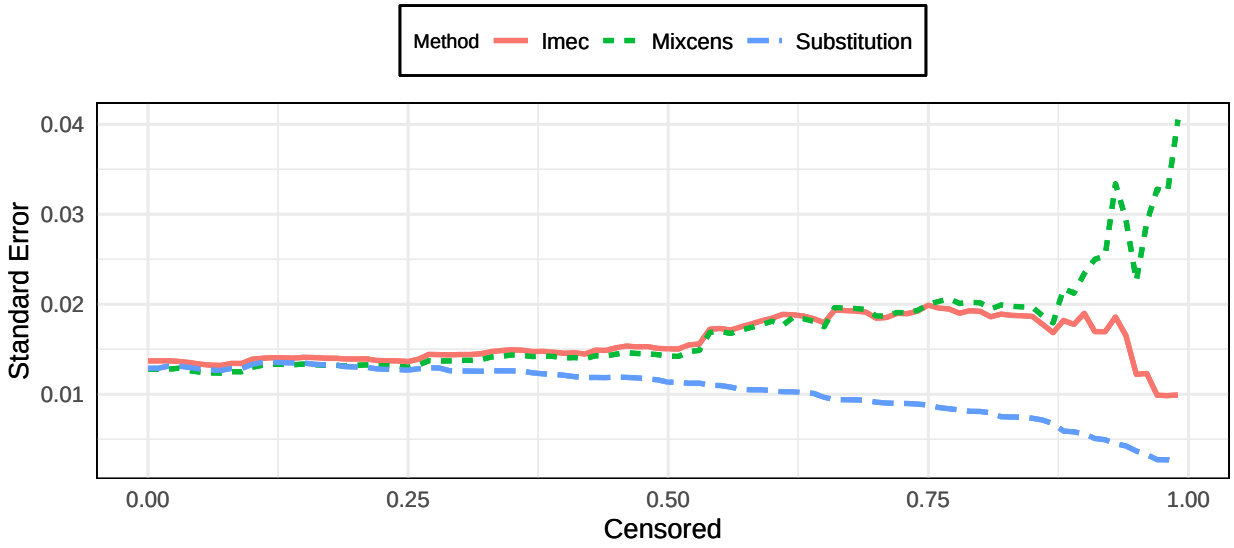


Figure 7a: Standard error of the estimated regression slope as a function of censoring proportion. The true value of the slope being log(1.05).

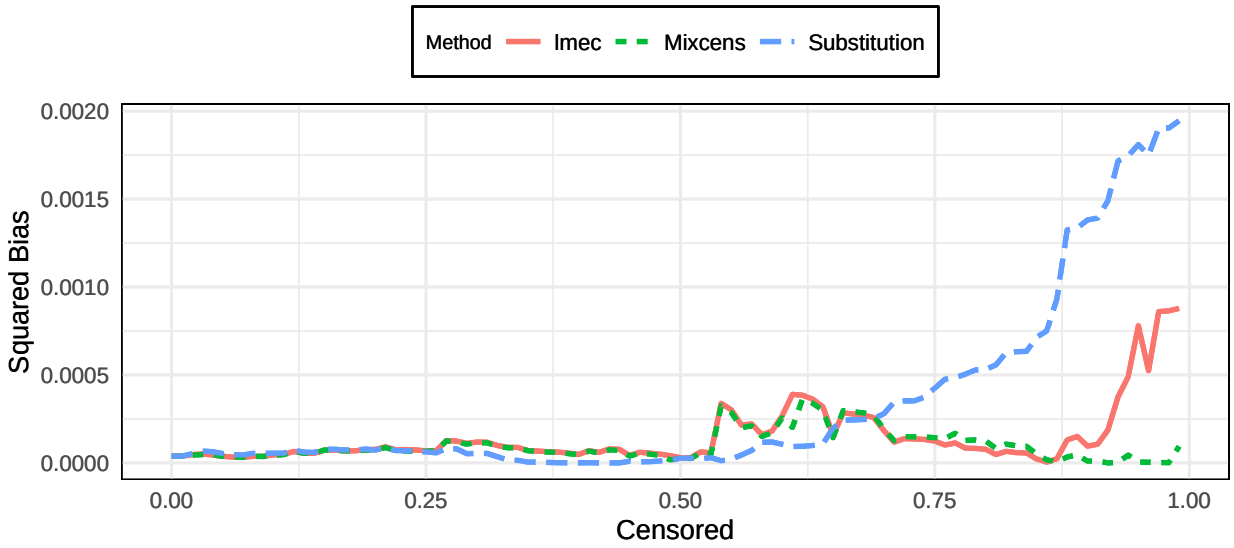


Figure 7b: Squared Bias of the estimated regression slope as a function of censoring proportion. The true value of the slope being log(1.05).

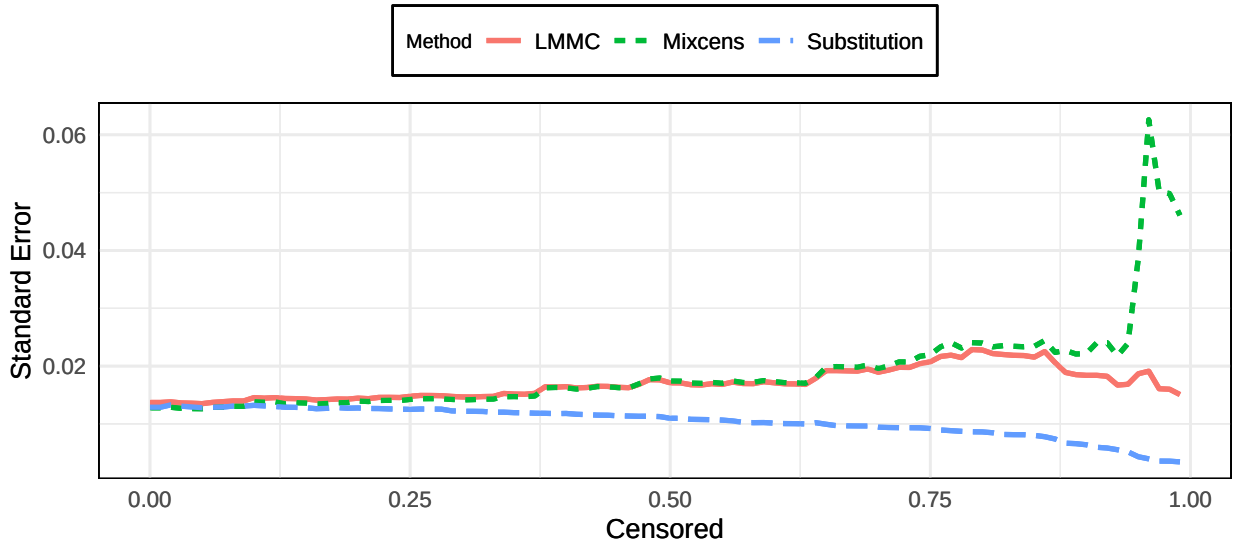


Figure 8a: Standard error of the estimated regression slope as a function of censoring proportion. The true value of the slope being log(1.1).

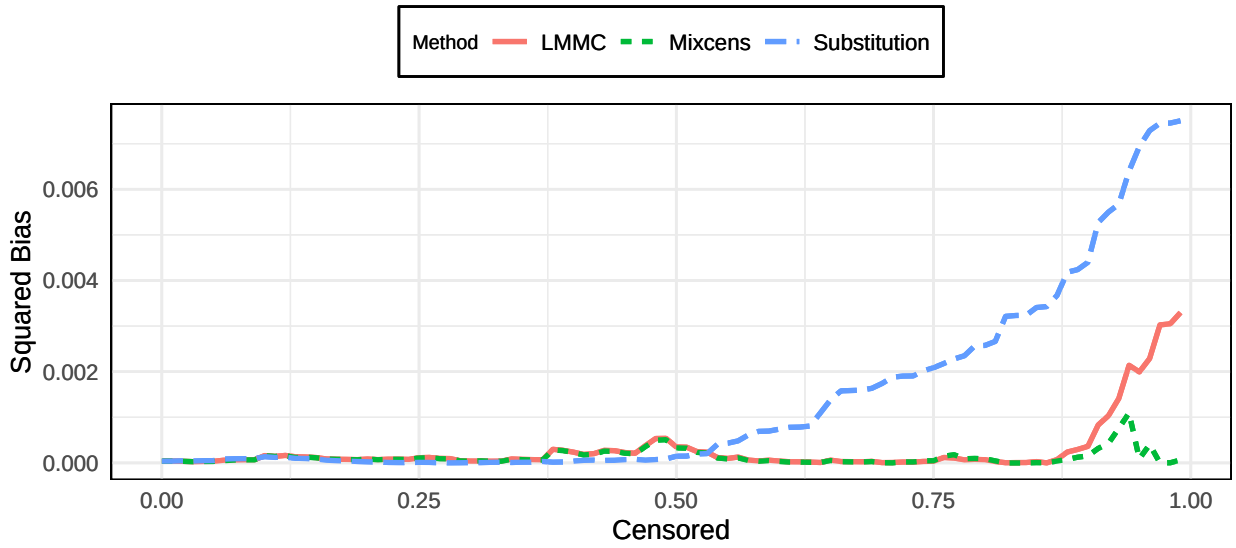


Figure 8b: Squared Bias of the estimated regression slope as a function of censoring proportion. The true value of the slope being log(1.1).

3.2.2 Model Analysis

Based on the analysis in Section 3.2.1, the function *lmec* will be used to analyse the LMMC model in this section. Tables 1-4 show a summary of the simulation grouped by 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 simulation $1, 2, ..., 100$, the squared bias was calculated as $(\hat{\beta} - \beta)^2$ and the mean of this squared bias over all 100 simulations were determined for every scenario. The precision of the monte carlo method were computed using the standard error of the estimated bias defined as the standard deviation of the bias divided by the sqaure root of the number of simulation, in this case $\sqrt{100}$.

Table 1: Summary statistics of simulations at 30% censored data and a 1% yearly increase.

Individual Sd	Random Effect	Method	$(\hat{\beta}-\beta)^2$	Coverage	$\text{Var}(\hat{\beta})$	MSE	$\text{Se}(\hat{\beta})$
0.05	High	Substitution	0.0017	0.95	0.0014	0.0031	0.0037
		LMMC	0.0025	0.99	0.0025	0.0050	0.0050
	Low	Substitution	0.0003	0.08	0.0000	0.0003	0.0000
		LMMC	0.0000	0.96	0.0000	0.0000	0.0000
0.50	High	Substitution	0.0014	0.97	0.0013	0.0026	0.0036
		LMMC	0.0021	0.99	0.0021	0.0042	0.0046
	Low	Substitution	0.0002	0.96	0.0002	0.0004	0.0014
		LMMC	0.0002	0.97	0.0002	0.0004	0.0014
1.40	High	Substitution	0.0024	0.96	0.0024	0.0048	0.0049
		LMMC	0.0043	0.97	0.0043	0.0086	0.0066
	Low	Substitution	0.0010	0.94	0.0010	0.0019	0.0032
		LMMC	0.0016	0.96	0.0015	0.0031	0.0039

Table 2: Summary statistics of simulations at 30% censored data and a 5% yearly increase.

Individual Sd	Random Effect	Method	$(\hat{\beta}-\beta)^2$	Coverage	$\text{Var}(\hat{\beta})$	MSE	$\text{Se}(\hat{\beta})$
0.05	High	Substitution	0.0021	0.94	0.0012	0.0033	0.0035
		LMMC	0.0023	0.97	0.0021	0.0044	0.0046
	Low	Substitution	0.0006	0.00	0.0000	0.0006	0.0000
		LMMC	0.0000	0.97	0.0000	0.0000	0.0000
0.50	High	Substitution	0.0019	0.96	0.0013	0.0033	0.0036
		LMMC	0.0024	0.98	0.0024	0.0047	0.0049
	Low	Substitution	0.0002	0.93	0.0002	0.0004	0.0014
		LMMC	0.0003	0.94	0.0003	0.0005	0.0017
1.40	High	Substitution	0.0027	0.96	0.0022	0.0048	0.0047
		LMMC	0.0039	0.98	0.0038	0.0077	0.0062
	Low	Substitution	0.0013	0.91	0.0012	0.0024	0.0035
		LMMC	0.0017	0.95	0.0018	0.0035	0.0042

The first thing that stands out is the fact that whenever at least one of the errorterms are not set to a lower value, the LMMC model produce more or close to equally biased estimates than in the case of using substitution. However, when the errorterms 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 alterning the proportion of censoring the LMMC model gives more biased estimates at higher proportions as to be expected considering there are 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 at both of the proportion of censoring.

In the report of Bignert et. al (2017), confidence intervals are often used to give an indication of the true value of the slope. Therefor, 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 defined as

$$se(\hat{\beta}) = \sqrt{\frac{1}{n-2}\frac{\sum(y_i - x_i\beta)^2}{\sum x_i^2}}$$

(considering the covariate is already centered) with n being the number of observations for a simulation. The coverage was then calculated by looking at the proportion of simulations obtaining a confidence interval containing the true value of the slope. When taking a closer look at the coverage of both methods the LMMC model clearly has a 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 errorterms 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.

Individual Sd	Random Effect	Method	$(\hat{\beta}-\beta)^2$	Coverage	$\text{Var}(\hat{\beta})$	MSE	$\text{Se}(\hat{\beta})$
0.05	High	Substitution	0.0010	0.99	0.0006	0.0016	0.0024
		LMMC	0.0041	0.99	0.0039	0.0080	0.0062
	Low	Substitution	0.0003	0.04	0.0000	0.0003	0.0000
		LMMC	0.0000	0.98	0.0000	0.0000	0.0000
0.50	High	Substitution	0.0012	0.95	0.0006	0.0018	0.0024
		LMMC	0.0045	0.97	0.0037	0.0082	0.0061
	Low	Substitution	0.0001	0.94	0.0001	0.0002	0.0010
		LMMC	0.0003	0.96	0.0003	0.0005	0.0017
1.40	High	Substitution	0.0013	0.97	0.0009	0.0022	0.0030
		LMMC	0.0049	0.99	0.0043	0.0092	0.0066
	Low	Substitution	0.0006	0.94	0.0005	0.0011	0.0022
		LMMC	0.0021	0.95	0.0021	0.0043	0.0046

Table 4: Summary statistics of simulations at 60% censored data and a 5% yearly increase.

Individual Sd	Random Effect	Method	$(\hat{\beta}-\beta)^2$	Coverage	$\text{Var}(\hat{\beta})$	MSE	$\text{Se}(\hat{\beta})$
0.05	High	Substitution	0.0024	0.78	0.0006	0.0030	0.0024
		LMMC	0.0035	0.99	0.0033	0.0068	0.0057
	Low	Substitution	0.0003	0.00	0.0000	0.0003	0.0000
		LMMC	0.0000	0.91	0.0000	0.0000	0.0000
0.50	High	Substitution	0.0024	0.85	0.0006	0.0030	0.0024
		LMMC	0.0034	0.99	0.0033	0.0067	0.0057
	Low	Substitution	0.0003	0.76	0.0001	0.0004	0.0010
		LMMC	0.0002	0.96	0.0002	0.0005	0.0014
1.40	High	Substitution	0.0031	0.76	0.0011	0.0042	0.0033
		LMMC	0.0063	0.97	0.0060	0.0123	0.0077
	Low	Substitution	0.0011	0.79	0.0006	0.0017	0.0024
		LMMC	0.0022	0.96	0.0022	0.0043	0.0047

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 obviously affected by the level of the errorterms. 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 true 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 too low, resulting in too small confidence intervals.

The precision of the monte carlo estimates of the bias, as demonstrated by the standard error of the bias, is good when the noise is low. The standard error increase, meaning the precision decrease, when the noise gets more noticable. The precision for the LMMC model seem to decrease when the proportion of censoring increases which is to be expected, while the reverse is true for the substitution model.

Figure 3-5 shows each simulated estimate of the slope for both methods plotted against eachother with Figure 3 treating each scenario when the individual noise is set to low, Figure 4 the case of a medium noise and Figure 5 when it is set to high. The first thing that jumps out is how big of an influence both errorterms have separately since altering just one of them from low to a higher value instantly results in much more biased estimates for both models.

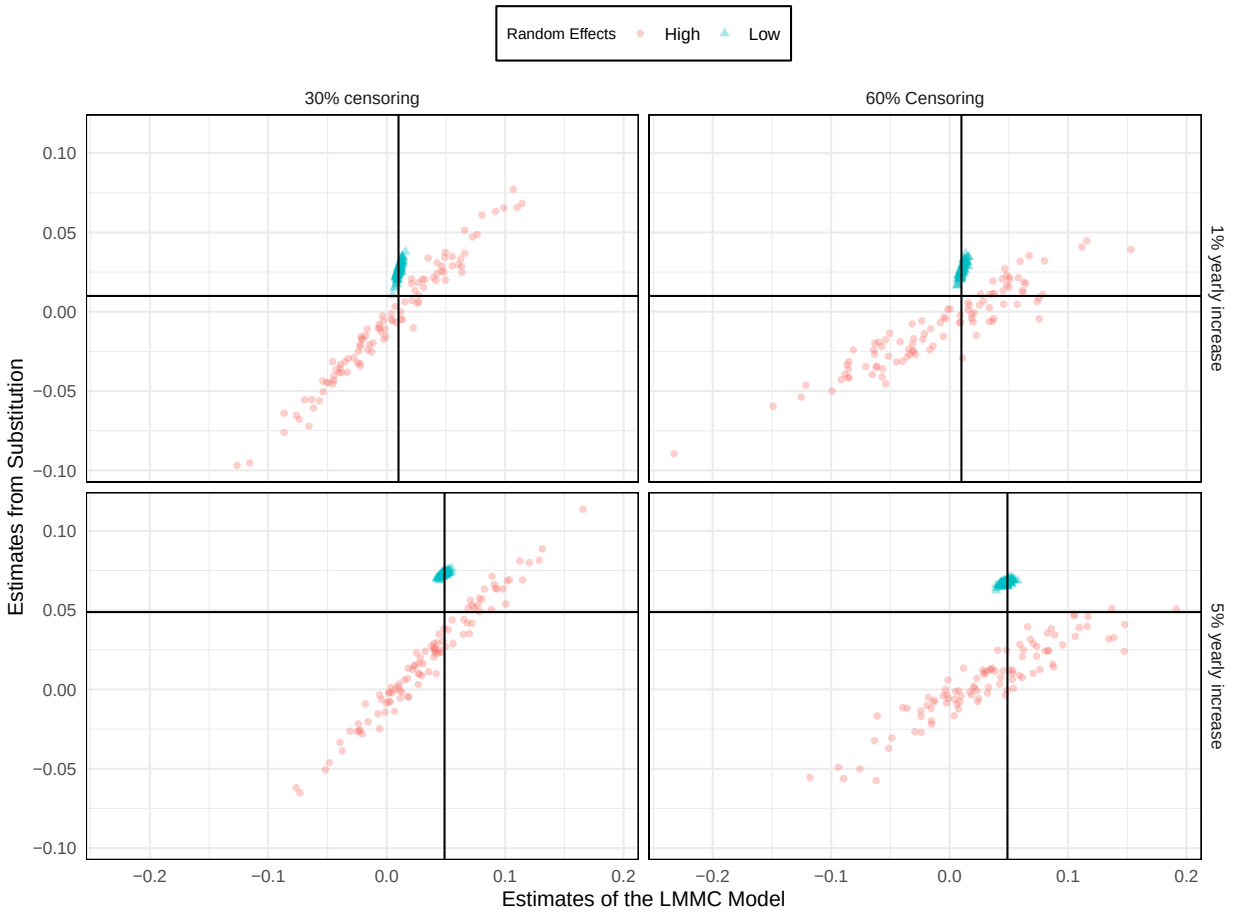


Figure 3: Plotting the estimated slopes for the Museum model and Tobit model against eachother having the variance of individual specimens set to low. The vertical and horizontal lines correspond to the true value of the slope.

The figures also show that for each and every 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. Explicitly 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 actually underestimate 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 might be one of the reasons for the higher bias of the LMMC model.

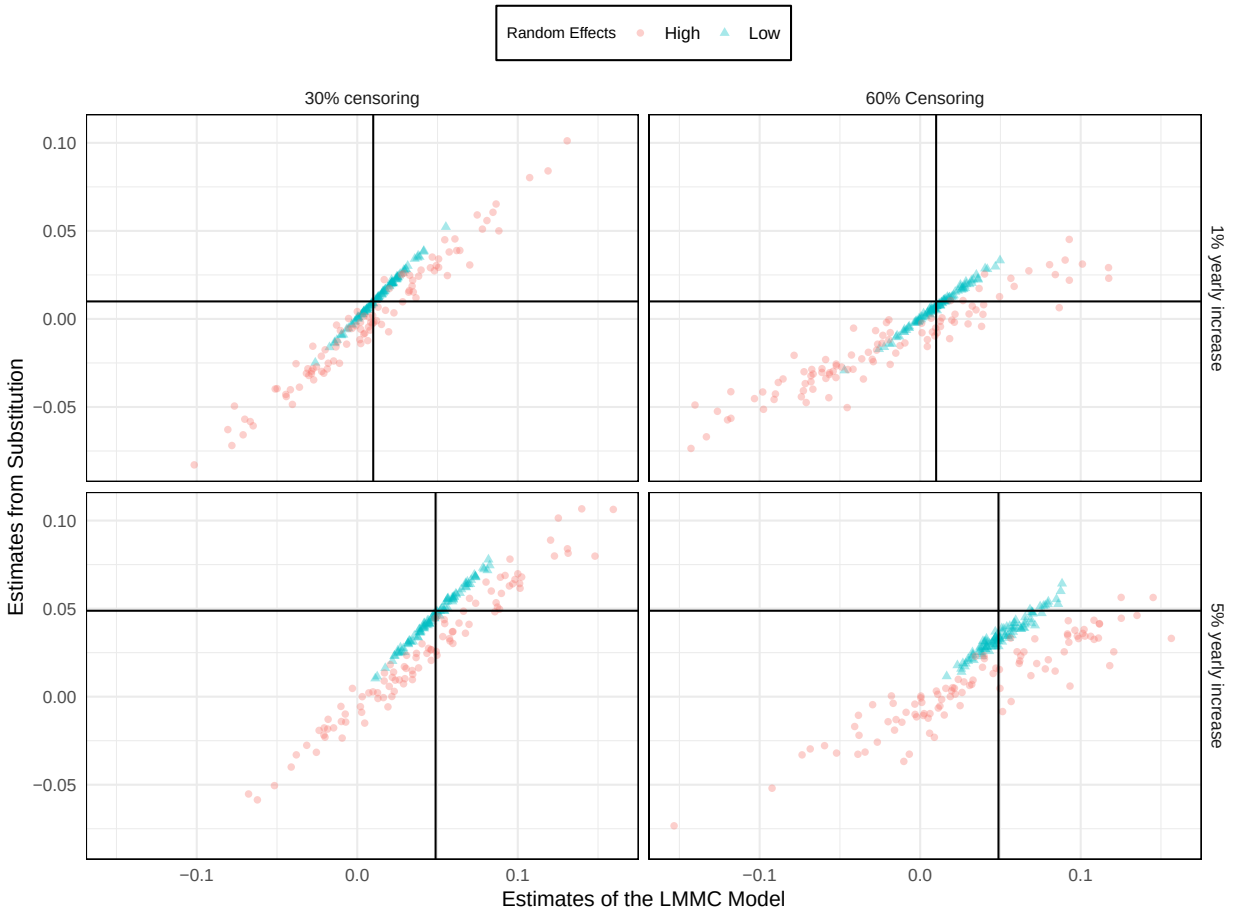


Figure 4: Plotting the estimated slopes for the Museum model and Tobit model against eachother having the variance of individual specimens set to medium. The vertical and horizontal lines correspond to the true value of the slope.

Yet another result is the fact that the proportion of censoring seems to have a big impact on the correlation between the estimates of the two models. More specific, whenever the proportion of censoring increase, the reverse goes for the correlation between the estimates. This is most likely an effect from the 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 therefor not be analysed in detail here. One detail to take into consideration is the fact that when using substitution, the estimates are now 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 7 with 60% censoring) would likewise be centered around another value, however in contrast to the substitutttion 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.

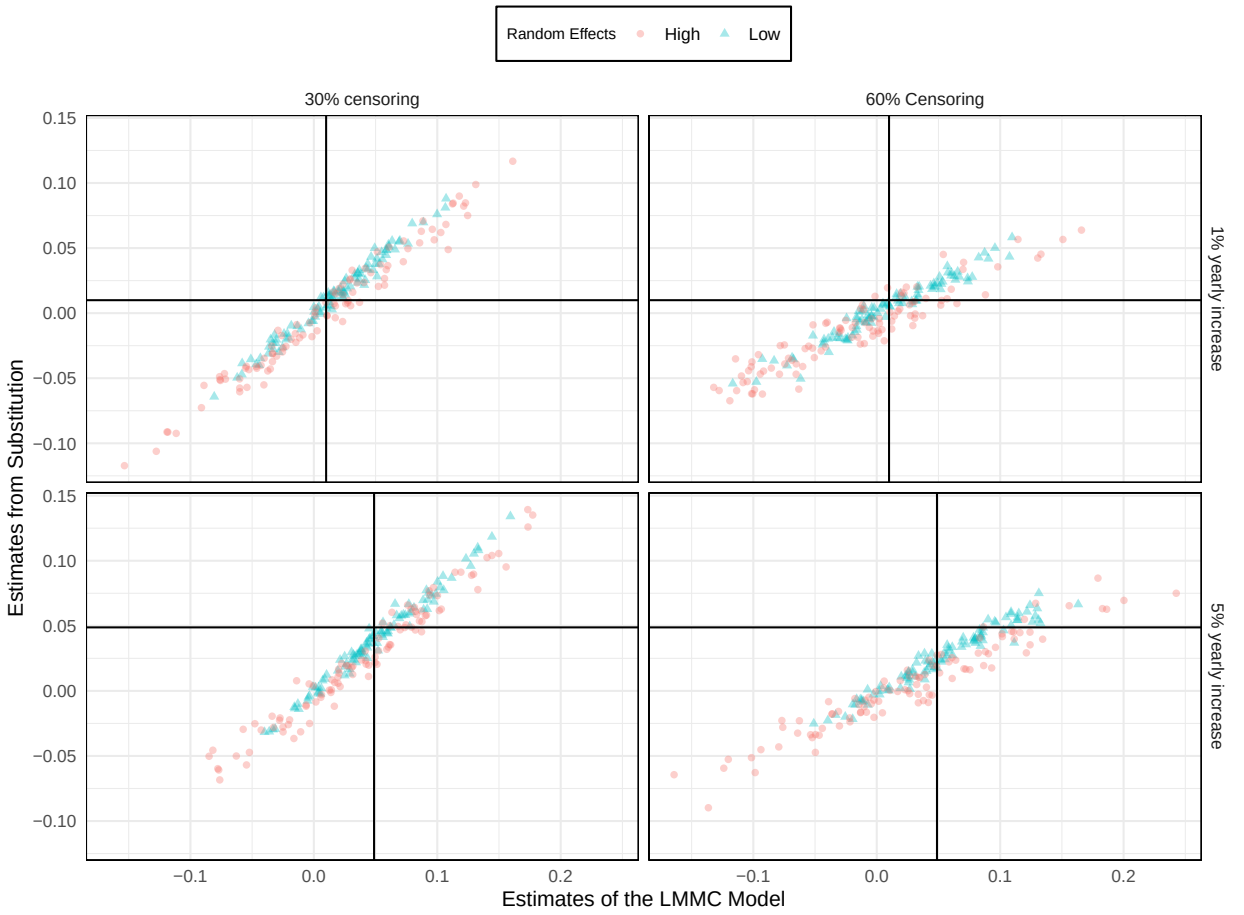


Figure 5: Plotting the estimated slopes for the substitution method and the LMMC model against eachother having the variance of individual specimens set to high. The vertical and horizontal lines correspond to the true value of the slope.

4 Application

4.1 Probabiliy plots and distribution assumptions

In the dataset used by the museum for their analysis of environmental toxins a large number of different metals were analysed. Three of which has the tendency of having a rather large proportion of censored values and are the three metals of most interest to analyse 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 perform one analysis for each location. In the interest of keeping the study on a moderate level and to get the most reliable analyses, 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 were considered.

In order to justify the use of the LMMC model for the dataset, an analysis concering 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 data set, 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 data set. 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 4-6, 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 uncensored values position. The *cenros* function by default performs a log-normal transformation prior to operations over the data (Lee, 2017) and thereafter perform the reverse transformation. Here, to clearly show that concentrations are on the log-scale, both of these transformations is 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.

In order 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 are 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 is 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. Even though most commercial statistical softwares use the formula $p = i/n$ (Helsel, 2005, pp.48), the Weibull formula is to be prefered when only using a sample which is a part of the total population. This due to the fact that 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. When points have the same values and therefor ties in ranks, as is the case for censored data, each point is given its own rank.

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 logaritmic 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 4-6, it seems like a reasonable assumption that the data follows a log-normal distribution seing that they more or less follows a straight line.

4.2 Applying the LMMC model

The data material was acquired from Martin Sköld from the Swedish Museum of Natural History. The analysis for metals were 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 Department of Environmental Science and Analytical Chemistry (ACES), at Stockholm University (SU) (see Section 6, Bignert. et. al. 2017). Therefor, the data prior to 2007 and the data from 2007 and forward are not deemed optimal to analyse together. Hence, only the data-set from 2007-2018 is analysed in this thesis in contrast to the report by Bignert. et. al (2017) where both an analysis using all data were used aswell 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 aswell as the substitution method is made instead of using the result from the report. The data-set 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 analysed to follow the line of the published report making it easier to do a good comparison between the two methods.

The logaritmic values of the reported concentrations, in combination with the coherent year and a vector of censoring indicators, indicating whether or not the observations is 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, were 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 standard deviation 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 over 5. The standard deviation of the individual specimens for each locations can be seen in Table 5-7.

Due to the fact 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 data-set have been available, namely a simple linear regression analysis using the substitution method with a substitution value of $1/\sqrt(2)$ 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 quick glance of the tables show 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 due to the fact that 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 lies above 80% censoring for all locations.

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 produce 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 produce unbiased estimates when the noise are low as in this case suggesting 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 $0.2-0.35$ except for in Harufjärden. As in the case of analysing lead, both models produced similar estimates for the slope for all locations but Harufjärden. The LMMC model produced an estimate resulting in 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 0.5 and the between-year noise being around 0.1 for all but one year in combination with 53% censoring gives an indication on how to judge the difference in estimation for the two models. Looking at Figure 4, 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 undervaluate 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 being closer to the estimate acquired by substitution, a yearly decrease by around 2%, the LMMC model infrequently overestimate the value of the slope but indeed more often underestimate the value of the slope (see Figure 6 & 7). 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. This gives an indication 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 come to the conclusion that no actions need to be taken considering a 2% yearly decrease imply a halved concentration level after 35 years. Nevertheless, the truth as a matter of fact being that a close to 6% increase would result in a doubled concentration level in almost 10 years.

The data for chromium clearly have 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 6), due to the fact that uncensored data is plotted around the 25th percentile even though the censoring is at over 80%. Therefor, it is still possible to get a decent regression analysis of the data and estimate of the slope. The differences of the estimates between the two methods for chromium is clear. The simulation study showed that, whenever the noise gets lower, the more precise the LMMC model is in contrast to the substitution method.

"“Higher Inclination => Overestimate (decrease) or underestimate (increase) by substitution”" "“Gör en simulering, specialfall för just detta scenario? Svårt göra 10% increase/decrease för alla scenarion, tar för lång tid”" "“**Outlier analys?**”" ##

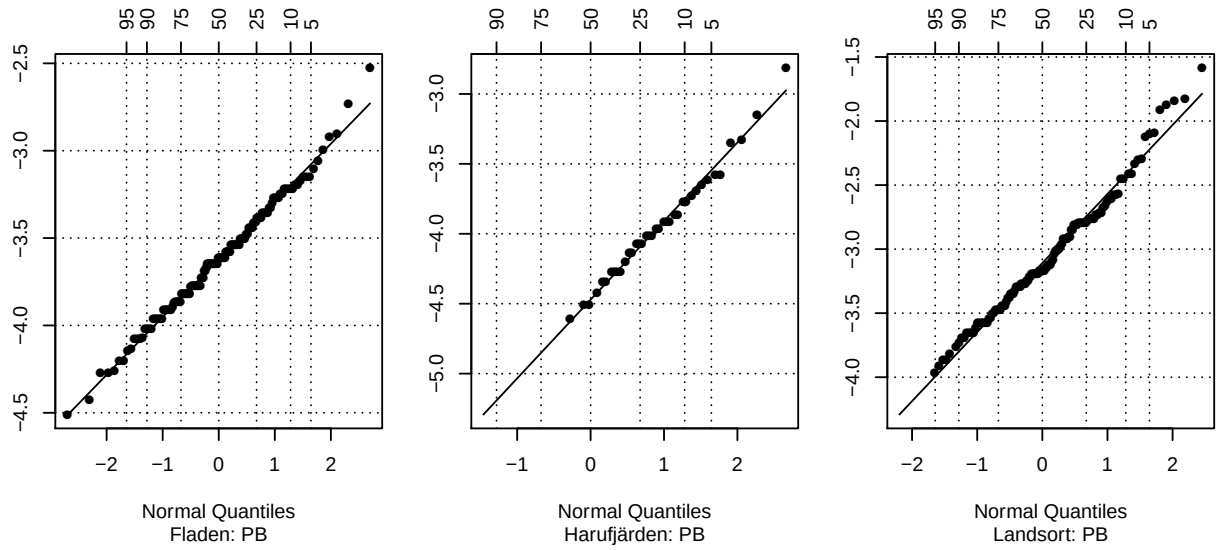
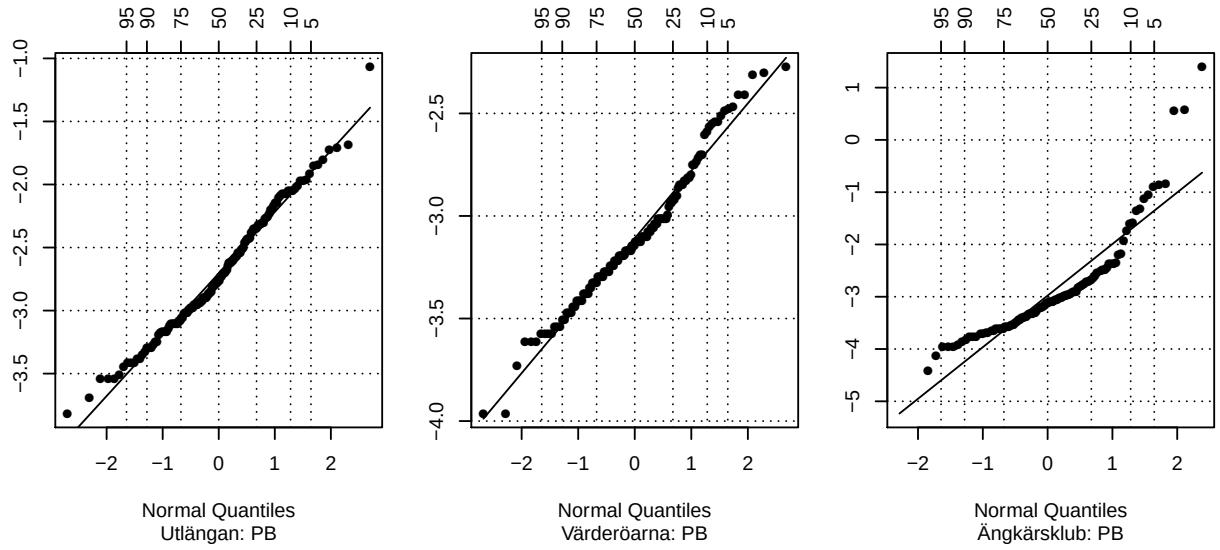


Figure 4: Distribution plot testing the lead concentrations for each location against a log-normal distribution.



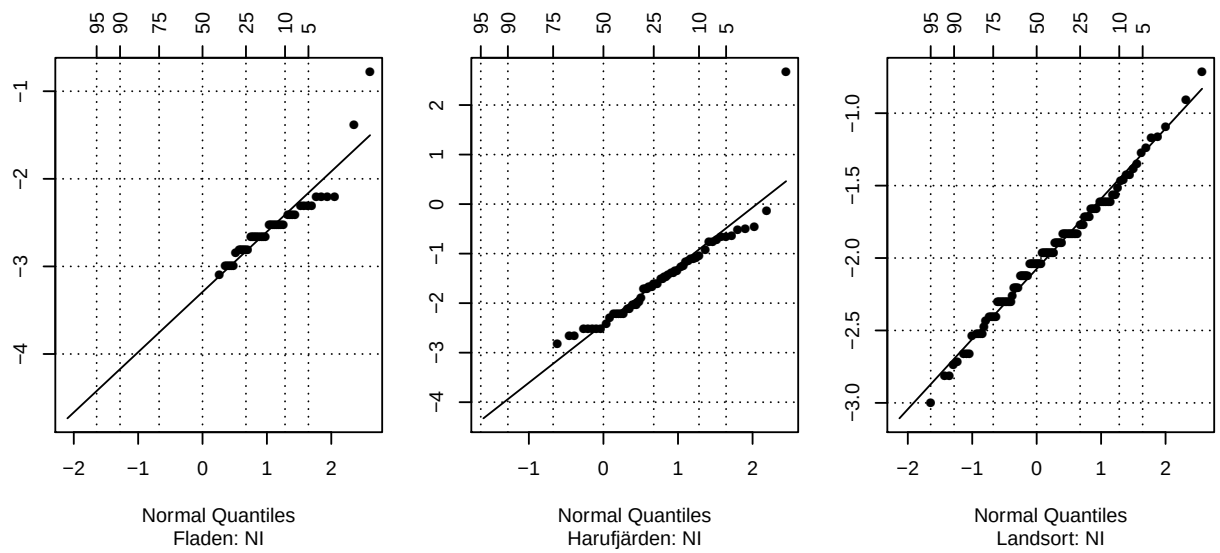
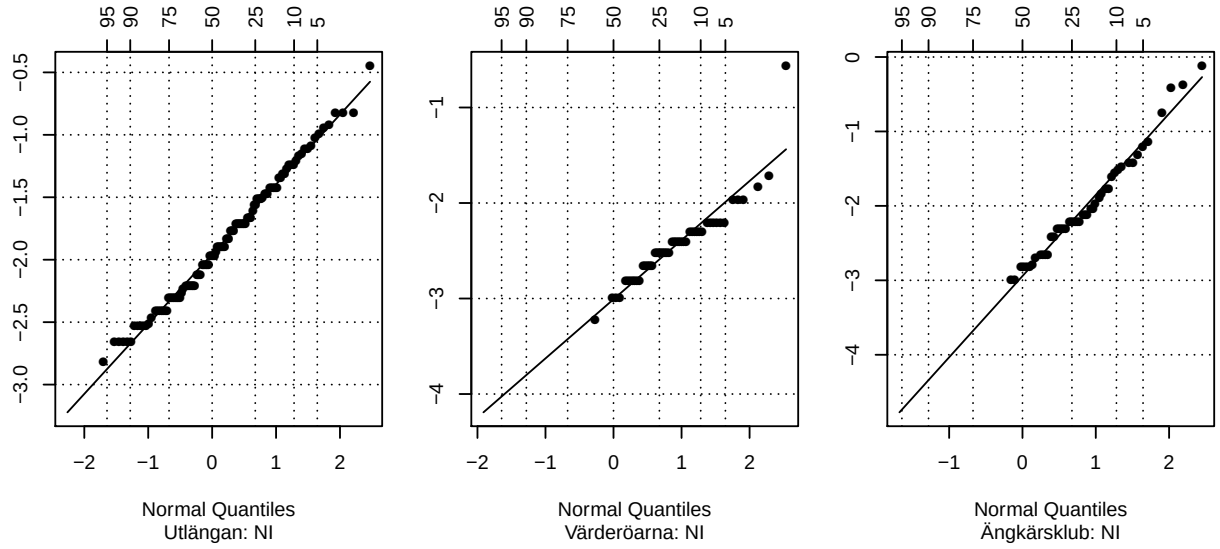


Figure 5: Distribution plot testing the nickel concentrations for each location against a log-normal distribution.



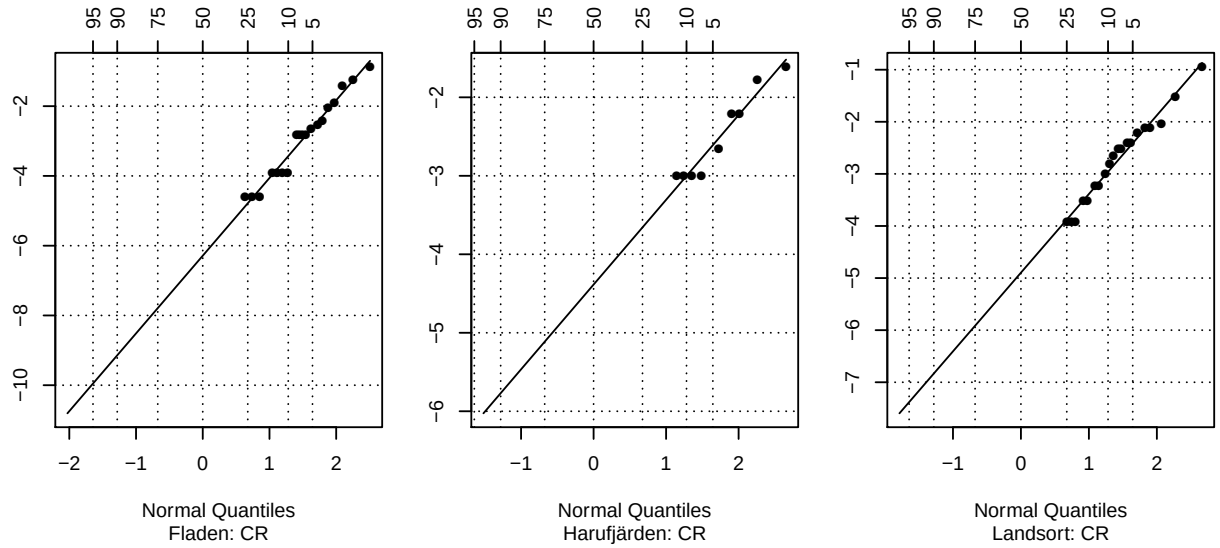


Figure 6: Distrubution plot testing the chromium concentrations for each location against a log-normal distribution

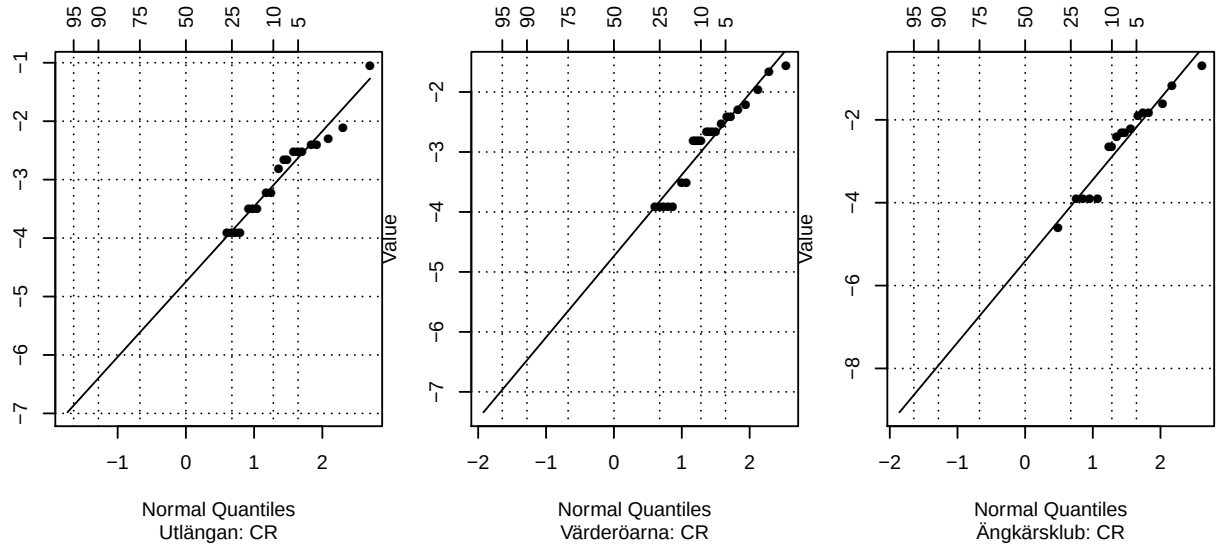


Table 5: Results of application of both models on data for lead concentrations in herring (Log-Scale)

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	0.1251
Harufjärden	-0.0215	0.0367	0.0301	0.0120	70.5426	0.0889
Landsort	0.0191	0.0188	0.0191	0.0131	3.0534	0.1699
Utlängan	0.0090	0.0190	0.0090	0.0118	0.0000	0.1977
Väderöarna	0.0042	0.0190	0.0045	0.0089	0.0000	0.1267
Ängskärsklubb	0.0662	0.0483	0.0723	0.0272	6.1947	0.3433

Table 6: Results of application of both models on data for nickel concentrations in herring (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	0.2924
Harufjärden	0.0559	0.0971	-0.0235	0.0221	53.4884	0.5366
Landsort	-0.0096	0.0172	-0.0115	0.0125	14.5038	0.2686
Utlängan	-0.0007	0.0199	-0.0029	0.0131	17.3611	0.3100
Väderöarna	-0.0188	0.0528	-0.0152	0.0120	61.1940	0.2031
Ängskärsklubb	-0.1105	0.0469	-0.0849	0.0185	56.6372	0.3643

Table 7: Results of application of both models on data for chromium concentrations in herring (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	0.2076
Harufjärden	0.0395	0.0866	0.1202	0.0141	93.0233	0.2113
Landsort	-0.0427	0.1023	0.0329	0.0203	84.7328	0.2926
Utlängan	-0.1241	0.1085	0.0408	0.0176	86.1111	0.2076
Väderöarna	0.1078	0.0955	0.1403	0.0160	82.8358	0.2550
Ängskärsklubb	-0.0384	0.1323	0.0286	0.0230	84.9558	0.4051

5 Discussion

In this thesis the likelihood model for a linear mixed-effect model with censored responsvariables having only one random effect and one covariate is determined. The maximum likelihood estimate of the slope is then examined using the LMMC model aswell as by using substitution together with a linear regression model in a simulation study. The study demonstrated the inability of producing unbiased results when fabricated data is used to calculate the maximum likelihood estimates. It further showed the reverse for the LMMC model, producing unbiased estimates when not having much noise in the data no matter the inclination of the slope and proportion of censored data. 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. At the same time, when the error terms were larger, using substitution on average produced less biased estimates. This as a result of the LMMC model on a few occasions generating estimates with large bias.

It was shown that for small errorterms, 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 data-set, the mean sqaured error of the estimates from the Tobit model increased, while the reverse ended up being true when using fabricated data.

The precision of the simulation study, as seen by the standard errors, is quite high never having a standard error exceeding 0.008. Although, granted that there were 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 can 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 maxima, sometimes converging towards a local maxima or a sadle point, resulting in the likelihood function growing without bound, perhaps more methods of numerical estimations can be tested. Another way to avoid this complication would be to try different starting values when using the EM-algorithm.

Applying the two models on data used by the Swedish Museum of Natural History showed there are a chance of acquiring estimates much higher or lower than the truth. Especially in Harufjärden, both of these scenarios has an implication of being true. **Förmodligen lite text om krom**.

6 Appendix

The Appendix contains results from the simulation study done working with an implied yearly decrease. The simulation study was made in the exact same matter 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

Individual Sd	Random Effect	Method	$(\hat{\beta}-\beta)^2$	Coverage	$\text{Var}(\hat{\beta})$	MSE	$\text{Se}(\hat{\beta})$
0.05	High	Substitution	0.0014	0.97	0.0014	0.0028	0.0037
		LMMC	0.0024	0.98	0.0024	0.0049	0.0049
	Low	Substitution	0.0003	0.08	0.0000	0.0003	0.0000
		LMMC	0.0000	0.93	0.0000	0.0000	0.0000
0.50	High	Substitution	0.0015	0.95	0.0013	0.0028	0.0036
		LMMC	0.0024	0.97	0.0022	0.0046	0.0047
	Low	Substitution	0.0001	0.95	0.0001	0.0003	0.0010
		LMMC	0.0002	0.96	0.0002	0.0003	0.0014
1.40	High	Substitution	0.0017	0.99	0.0016	0.0033	0.0040
		LMMC	0.0029	0.99	0.0028	0.0058	0.0053
	Low	Substitution	0.0007	0.98	0.0006	0.0013	0.0024
		LMMC	0.0010	0.99	0.0010	0.0021	0.0032

Table 9: Summary statistics of simulations at 60% censored data and a 1% yearly decrease

Individual Sd	Random Effect	Method	$(\hat{\beta}-\beta)^2$	Coverage	$\text{Var}(\hat{\beta})$	MSE	$\text{Se}(\hat{\beta})$
0.05	High	Substitution	0.0009	0.97	0.0005	0.0014	0.0022
		LMMC	0.0053	0.96	0.0033	0.0086	0.0057
	Low	Substitution	0.0003	0.06	0.0000	0.0003	0.0000
		LMMC	0.0000	0.98	0.0000	0.0000	0.0000
0.50	High	Substitution	0.0007	0.98	0.0005	0.0012	0.0022
		LMMC	0.0044	0.98	0.0033	0.0077	0.0057
	Low	Substitution	0.0001	0.93	0.0001	0.0002	0.0010
		LMMC	0.0002	0.98	0.0002	0.0005	0.0014
1.40	High	Substitution	0.0009	0.98	0.0009	0.0018	0.0030
		LMMC	0.0049	0.99	0.0043	0.0092	0.0066
	Low	Substitution	0.0005	0.93	0.0005	0.0009	0.0022
		LMMC	0.0018	0.96	0.0018	0.0035	0.0042

Table 10: Summary statistics of simulations at 30% censored data and a 5% yearly decrease

Individual Sd	Random Effect	Method	$(\hat{\beta}-\beta)^2$	Coverage	$\text{Var}(\hat{\beta})$	MSE	$\text{Se}(\hat{\beta})$
0.05	High	Substitution	0.0012	0.97	0.0012	0.0024	0.0035
		LMMC	0.0027	0.97	0.0023	0.0050	0.0048
	Low	Substitution	0.0005	0.00	0.0000	0.0005	0.0000
		LMMC	0.0000	0.97	0.0000	0.0000	0.0000
0.50	High	Substitution	0.0016	0.99	0.0016	0.0033	0.0040
		LMMC	0.0033	0.99	0.0030	0.0063	0.0055
	Low	Substitution	0.0002	0.91	0.0002	0.0004	0.0014
		LMMC	0.0002	0.95	0.0002	0.0005	0.0014
1.40	High	Substitution	0.0022	0.97	0.0021	0.0044	0.0046
		LMMC	0.0037	0.98	0.0037	0.0075	0.0061
	Low	Substitution	0.0013	0.91	0.0012	0.0025	0.0035
		LMMC	0.0020	0.95	0.0020	0.0040	0.0045

Table 11: Summary statistics of simulations at 60% censored data and a 5% yearly decrease

Individual Sd	Random Effect	Method	$(\hat{\beta}-\beta)^2$	Coverage	$\text{Var}(\hat{\beta})$	MSE	$\text{Se}(\hat{\beta})$
0.05	High	Substitution	0.0007	0.98	0.0007	0.0014	0.0026
		LMMC	0.0081	0.91	0.0045	0.0126	0.0067
	Low	Substitution	0.0003	0.00	0.0000	0.0003	0.0000
		LMMC	0.0000	0.96	0.0000	0.0000	0.0000
0.50	High	Substitution	0.0007	0.99	0.0007	0.0014	0.0026
		LMMC	0.0058	0.98	0.0040	0.0098	0.0063
	Low	Substitution	0.0003	0.75	0.0001	0.0004	0.0010
		LMMC	0.0003	0.97	0.0003	0.0006	0.0017
1.40	High	Substitution	0.0010	0.99	0.0007	0.0017	0.0026
		LMMC	0.0041	0.99	0.0036	0.0078	0.0060
	Low	Substitution	0.0012	0.74	0.0006	0.0018	0.0024
		LMMC	0.0022	0.97	0.0022	0.0044	0.0047

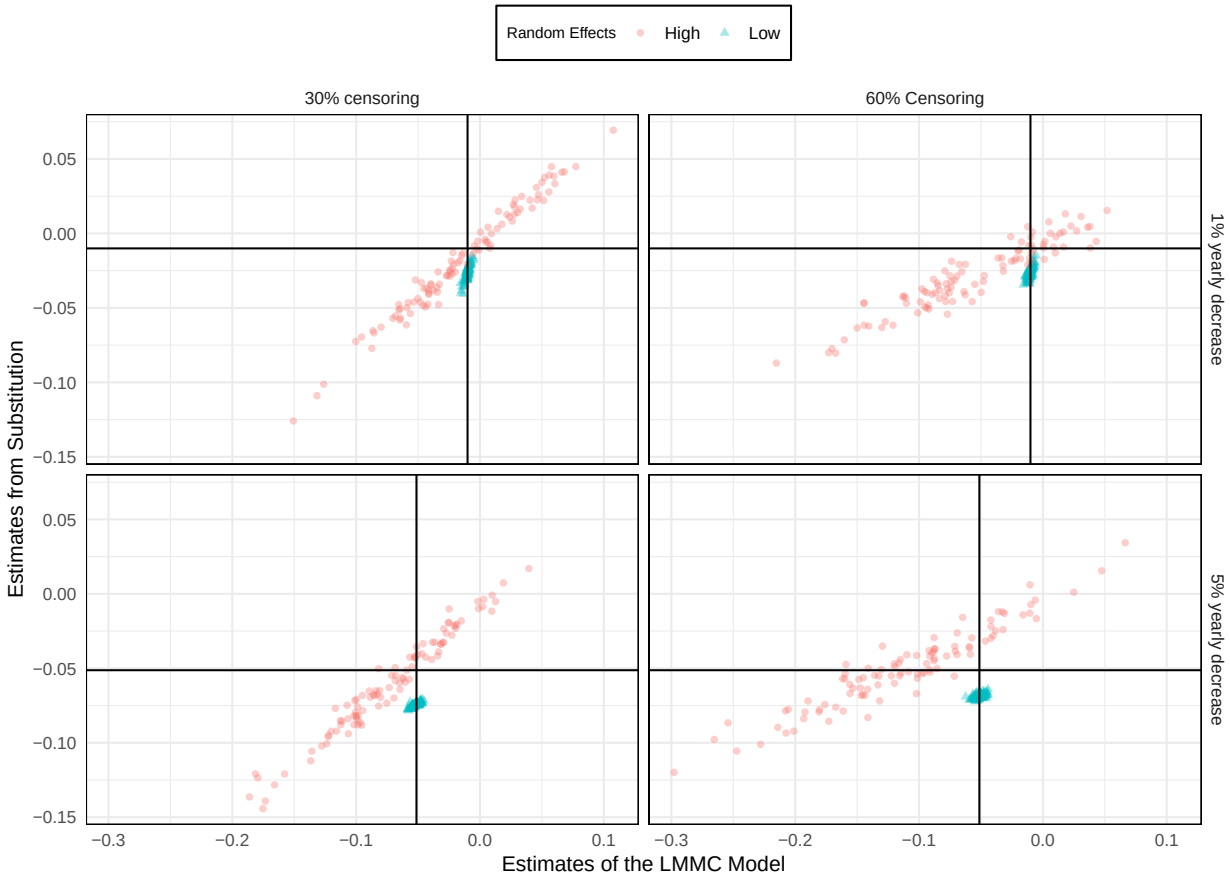


Figure 6: Plotting the estimated slopes for the substitution method and the LMMC modelagainst eachother having the variance of individual specimens set to low. The vertical and horizontal lines correspond to the true value of the slope.

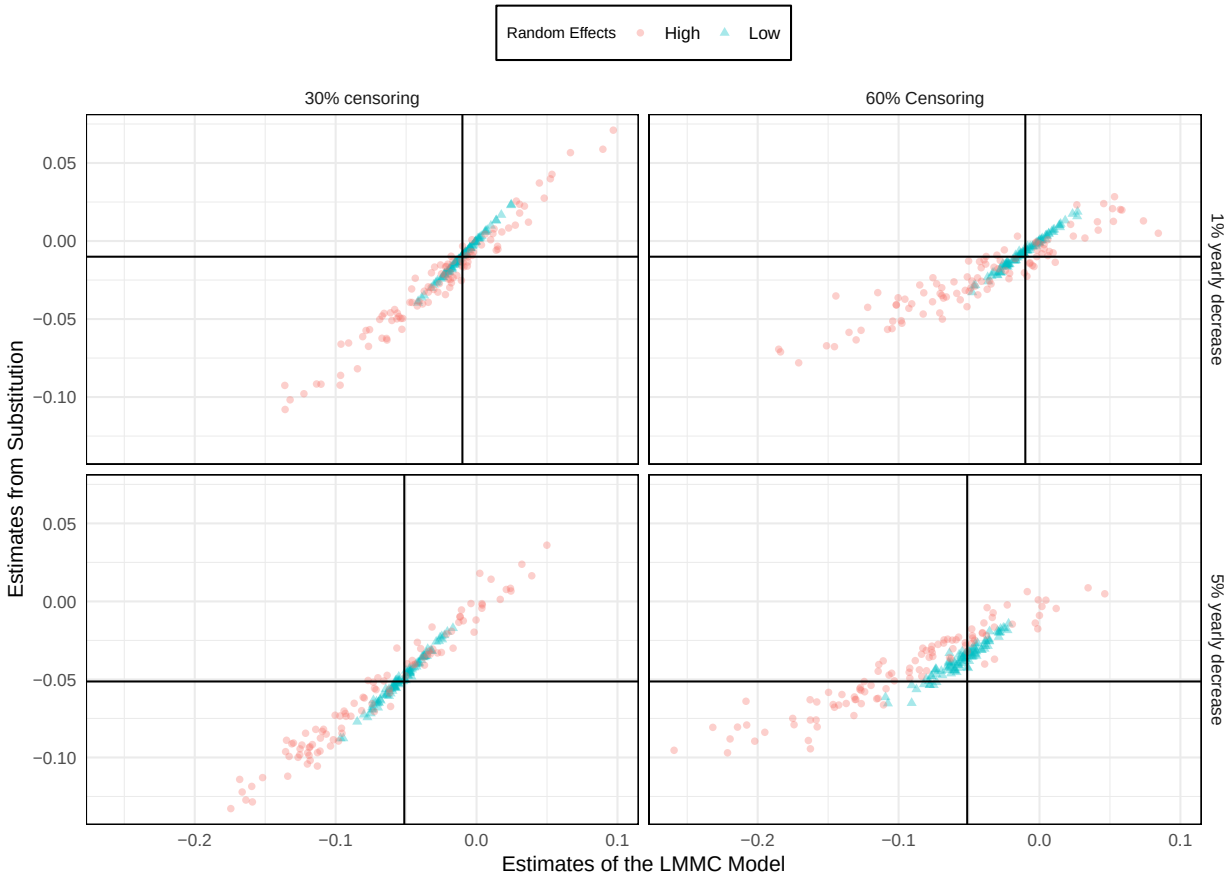


Figure 7: Plotting the estimated slopes for the substitution method and the LMMC model against eachother having the variance of individual specimens set to medium. The vertical and horizontal lines correspond to the true value of the slope.

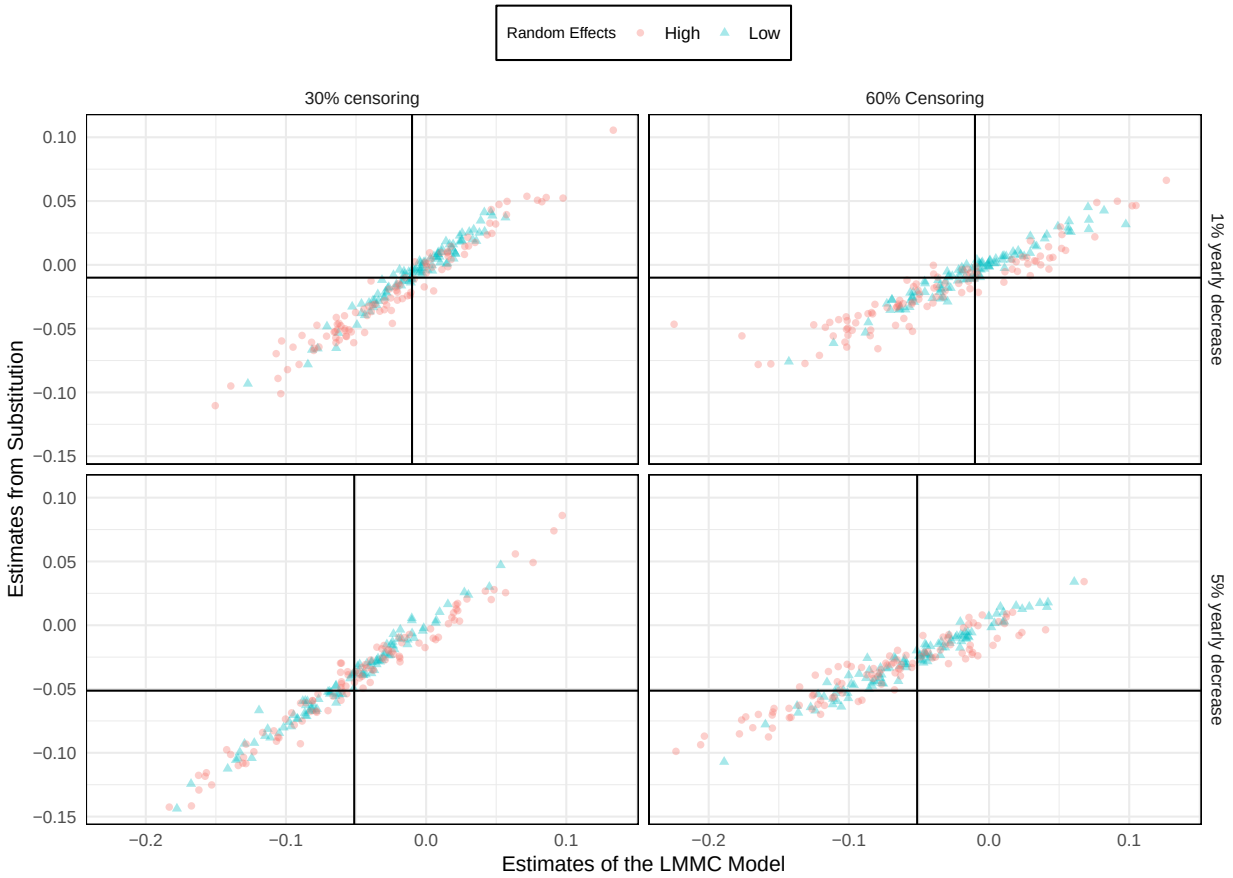


Figure 8: Plotting the estimated slopes for the substitution method and the LMMC model against eachother having the variance of individual specimens set to high. The vertical and horizontal lines correspond to the true value of the slope.

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 65, 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, in G. Gaal and D.F. Merriam,eds., Computer Applications in Resource Estimation. Pergammon Press, New York, 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 in Public Health Research, vol.18, 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, 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 Historys: 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(cloth)

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