Kandidat Skrivandet

Anton Holm

2020-02-24

Abstrakt

Introduction

- Maybe add graph to show how censored data looks like? So 2 graphs, one how it actually is, one how it's reported.
- Show that the museum data is log-linear
- Explain censored data

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 surprisingly 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.

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.

As a result of test equipments not being able to detect small enough quantities of toxins, a portion of the data is reported as *below the limit of quantification (LOQ)*. This portion of the data is reported as the LOQ divided by the square root of 2.

Due to biological properies 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.

Bignert at al (2017) uses log-linear regression analysis, hence the data is assumed to follow a log-linear distribution.

Common Errors

One of the most common error being made when analysing censored data is fabricating. The analysists simply substitute the non-detects with a fraction (often one half) of the quantitive- or detetection limit. A simulation were made by Helsel (2006) showcasing that this method produces lousy 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 goverment 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, remediations 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 non-detects. However, this will not be discussed in detail in this thesis.

Theory

When working with censored data, the non-detects can't be looked at as having a specific value. Instead, a combination of the 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 a much better estimation for 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.

CDF of a linear regression model

Consider a normal simple linear regression model

$$y_i = x_i\beta + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2)$$

were y_i is the response variable, x_i the explanatory variable, β 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 $f(y_i) = \frac{dF(y_i)}{dy_i}$.

Linear mixed-effects model

Can aggregate data. Take mean of each group => the avg data points are now independent: Less noise but disregard a lot of data Can do regression on each group => a lot of noise but takes all data LMM somewhere in between

Mixed models are an extension of normal models where random effects are integrated. A linear mixed model is 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

$$\mathbf{y_i} = \mathbf{X_i}\beta + \mathbf{Z_i}\mathbf{b_i} + \epsilon_\mathbf{i}$$

for $i = 1, ..., M$. Here, $\mathbf{y_i}$ is the n_i dimension respons 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 x p, $\mathbf{Z_i}$ a design matrix of size n_i x q linking $\mathbf{b_i}$ to $\mathbf{y_i}$ and $\epsilon_\mathbf{i}$ 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 n_i x n_i dimension covariance matrix and $\epsilon_\mathbf{i} \sim N(0, \sigma^2 I)$, I being the n_i dimension vector of ones.

Maximum Likelihood Estimation

One of the most interesting analysis to be made within regression analysis is what effect each covariate has on the response variable. This is represented by the unknown effect parametervector θ (β in the model above), and thus something of great importance to be able to estimate. This is often done using Maximum Likelihood Estimation. For a response variable $\mathbf{Y}$ with observations $\mathbf{Y} = \mathbf{y}$ having a probobility mass or density function $f(\mathbf{y}; \theta)$, depending on the observations $\mathbf{y}$ and $\theta \in \mathbf{\Theta}$ being the often unknown parametervector taking values in the parameterspace $\mathbf{\Theta}$, 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 probobility mass or density function of the observed data $\mathbf{y}$ viewed as a function of the paramatervector θ. The maximum likelihood estimate of θ denoted as $\hat{\theta}_{MLE}$ is then given as the parametervector maximising the likelihood function.

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 the PDF of 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, we further deduce

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

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

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

Case of the museum (Multivariate Normal-distribution)

Now, in the case of the analysis conducted by the Swedish Museum of Natural History, a Linear Mixed Tobit Model 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 is

$$\log(\mathbf{y_t}) = \mathbf{x_t} \cdot \beta + \mathbf{z} \cdot \mathbf{e_t} + \epsilon$$

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 x 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, $\mathbf{z}$ an n_t dimensional row vector of ones, $\mathbf{e_t}$ an n_t dimensional vector of the random effect e_t and ϵ the n_t dimensional vector with the within-years variance for each speciemen $\epsilon_i, i = 1, 2, ..., n_i$. Further more, since $e_t \sim N(0, \sigma_t^2)$ and $\epsilon \sim N(0, \delta^2)$, the distribution of $\log(\mathbf{y_t})$ follows

$$\log(\mathbf{y_t}) \sim N_{n_t}(\mathbf{x_t} \cdot \beta, \mathbf{\Sigma})$$

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

$$Cov(e + \epsilon_i, e + \epsilon_j) = E[(e + \epsilon_i)(e + \epsilon_j)] - E[e + \epsilon_i]E[e + \epsilon_j] = E[\epsilon^2] = \delta^2$$

for all i, j such that $i \neq j$ since $E[e] = E[\epsilon_k] = 0$ for all k. In addition, $(a_{ij}) = Var(e + \epsilon_i) = \sigma^2 + \delta^2$ when $i = j$.

Following the method above used to derive the CDF of a linear regression model, 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, the same following for $\mathbf{x_t}$ being partioned into a n_t^o x 2 matrix and a n_t^c x 2 matrix while $\mathbf{\Sigma_t^{oo}}$ and $\mathbf{\Sigma_t^{cc}}$ is the matrix 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 PDF $f_{\mathbf{y_i}^o}$. 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 and variance 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 $\Sigma_t^{oo^{-1}}$ is the inverse of Σ_t^{oo}. Denote $\phi_t^{c|o}(\cdot)$ as the PDF of 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 proboility density function (Held and Bové, p.321), the likelihood function can be written as

$$L(\beta; \mathbf{y_t}) = \prod_t f_{\mathbf{y_t}^o}(\mathbf{y_t}^o|\beta) \cdot \phi_t^{c|o}(\mathbf{c_t}|\beta)$$

which given the PDF of a multivariate normal distributed variable gives

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

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 parametervector β. This is more often than not done by finding the root to the *score equation* $S(\beta) = \frac{d}{d\beta}L(\beta)$ 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 parametervector maximising $l(\beta)$ is the same parametervector 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 observation, 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.

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 probobility 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 unchanges 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.

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 simulationstudy was therefore applied, trying to mimic the environmental setting of the museum as well as possible. Thus, 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 othervise 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:

$$\log(Y_{ij}) = X_i\beta + Zb_i + \epsilon,\ i = 1, ..., 11,\ j = 1, ..., n_i$$

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. There are countless of scenarios to consider but this simulationstudy 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 of 1% and 5% on the original scale
3. The two error terms ϵ and b changed between large and small

resulting in 16 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 proportions chosen are directly taken from the proportion of censored data of the two metals with the most censoring of those examined by the museum. 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 lowest and highest values were chosen to be included in the simulationstudy. For the noise of each individual specimen this resulted in a standard deviation of 0.05 and 1.4. 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. For most cases, the between-year variances were around the lower scale, however, considering some of the cases had higher variances, using a larger scale is also of interest.

For each of these scenarios, 100 simulation were made in which the method of substituting censored observations with a fraction of the limit of quantification (in this case using the entity of LOQ/$\sqrt{2}$ to continue mimicing the museum) and the maximum likelihood method were both used. The data-sets were simulated using the model describe 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). For the EM-algorithm used in the lmec function to estimate effect parameters a cap 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. An example of the data obtained from one of the simulations can be seen in Figure 1.

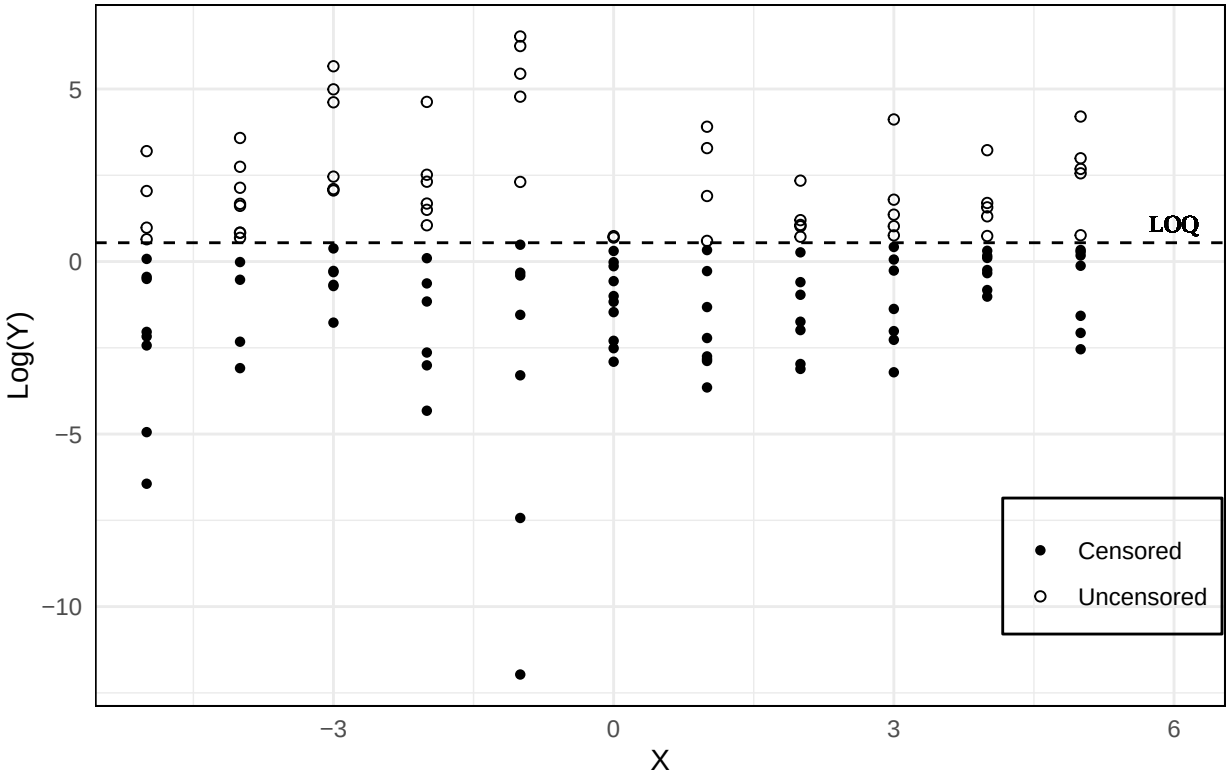


Figure 1: Simulated data with 60% censoring, large variations and slope representing 5% yearly increase hold both error terms at a high variance level.

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).

Tables 1-4 shows a summary of the simulation grouped by proportion of censored values and the true value of the slope. The first thing that stands out, which might come as a surprice is the fact that whenever at least one of the errorterms are set to a high value, the Tobit model produce more biased estimates than in the case of using substitution. However, when the errorterms have less influence, the Tobit model produced unbiased estimates while the substitution method, as shown by El-Shaarawi and Esterby (1992), still fails to produce unbiased estimates. Another interesting conclusion is in that when using substitution, the bias increase as the slope increase while the reverse seems to be true for the Tobit model except for the special case of holding both error terms at a high level (see Table 2 & 4). When alterning the proportion of censoring the Tobit model gives more biased estimates at higher proportions as to be expected 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 Tobit model does however still give unbiased estimates at small noise at both of the proportion of censoring.

Förklara Vilket konfidensintervall!! Wald-CI: se(beta)^2 = (squared bias)/(simulations-1)

When taking a closer look at the coverage of both methods the Tobit model clearly has a coverage over 95%

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

Std	Random Effects	Method	bias	Coverage	Variance	MSE
0.05	High	Museum	0.0017	0.95	0.0014	0.0031
		Tobit	0.0025	0.99	0.0025	0.0050
	Low	Museum	0.0003	0.08	0.0000	0.0003
		Tobit	0.0000	0.96	0.0000	0.0000
1.40	HIGH	Museum	0.0024	0.96	0.0024	0.0048
		Tobit	0.0043	0.97	0.0043	0.0086
	Low	Museum	0.0010	0.94	0.0010	0.0019
		Tobit	0.0016	0.96	0.0015	0.0031

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

Std	Random Effects	Method	bias	Coverage	Variance	MSE
0.05	High	Museum	0.0021	0.94	0.0012	0.0033
		Tobit	0.0023	0.97	0.0021	0.0044
	Low	Museum	0.0006	0.00	0.0000	0.0006
		Tobit	0.0000	0.97	0.0000	0.0000
1.40	High	Museum	0.0027	0.96	0.0022	0.0048
		Tobit	0.0039	0.98	0.0038	0.0077
	Low	Museum	0.0013	0.91	0.0012	0.0024
		Tobit	0.0017	0.95	0.0018	0.0035

in all cases but one (see Table 4). On the other hand, even though using substitution might have produced less bias in most cases, the coverage is not even close to the promised 95%. Anytime the errorterms are held at a low level, the coverage is close to zero. There are at the same time no clear patterns for any of the factors affecting the coverage of the Tobit model except for a slight decrease in coverage when the between-year errors are set to low over high.

The variance of the estimator for the model of the museum is to no surprise much lower than that of the Tobit 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 Tobit 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 Tobit model, when all factors are set to a high level (see Table 4). What might be of more interest is the effect censoring has on the estimator. For the Tobit 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 the museum is too low, resulting in too small confidence intervals.

Figure 2 & 3 shows each simulated estimate of the slope for both methods plotted against eachother with Figure 2 treating each scenario when the individual noise is set to low and Figure 3 when it is set to high. The most interesting part is how big of an influence both errorterms have separately since altering just one of them from low to high instantly results in much more biased estimates for both models.

Another interesting fact is that for each and every scenario, the estimates of the Tobit model seem to center around the true value of the slope, however the same can not be said for the substitution model. Especially when the inclination of the slope is larger it seems that the substitution method results in very few unbiased estimation. When combining this with having a larger proportion of censored data, this method actually fails to produce a single unbiased estimation which is in line with the results found by El-Shaarawi and

Table 3: Summary statistics of simulations at 60% censored data and a 1% yearly increase.

Std	Random Effects	Method	bias	Coverage	Variance	MSE
0.05	High	Museum	0.0010	0.99	0.0006	0.0016
		Tobit	0.0041	0.99	0.0039	0.0080
	Low	Museum	0.0003	0.04	0.0000	0.0003
		Tobit	0.0000	0.98	0.0000	0.0000
1.40	HIGH	Museum	0.0013	0.97	0.0009	0.0022
		Tobit	0.0049	0.99	0.0043	0.0092
	Low	Museum	0.0006	0.94	0.0005	0.0011
		Tobit	0.0021	0.95	0.0021	0.0043

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

Std	Random Effects	Method	bias	Coverage	Variance	MSE
0.05	High	Museum	0.0024	0.78	0.0006	0.0030
		Tobit	0.0035	0.99	0.0033	0.0068
	Low	Museum	0.0003	0.00	0.0000	0.0003
		Tobit	0.0000	0.91	0.0000	0.0000
1.40	High	Museum	0.0031	0.76	0.0011	0.0042
		Tobit	0.0063	0.97	0.0060	0.0123
	Low	Museum	0.0011	0.79	0.0006	0.0017
		Tobit	0.0022	0.96	0.0022	0.0043

Esterby (1992).

Yet another sriking 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.

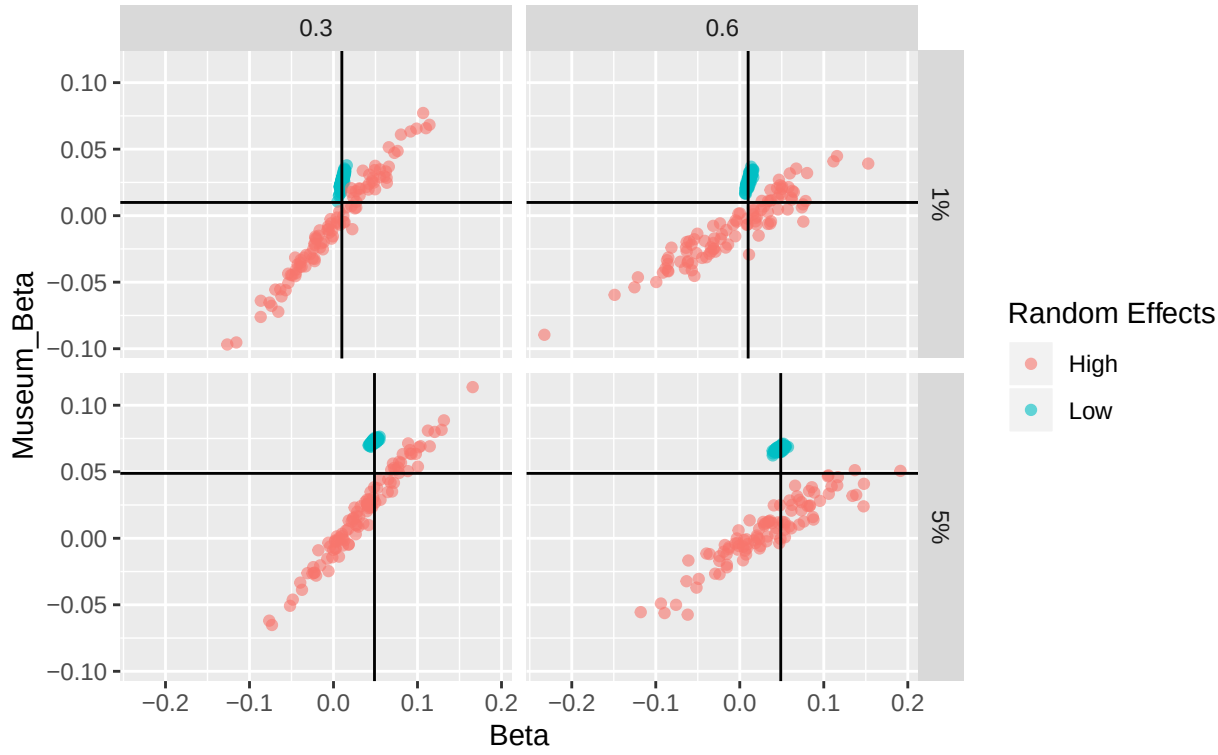


Figure 2: Plotting the estimated slopes for the Museum model
d Tobit model against eachother having the variance of individual specimens set to low.
The vertical and horizontal lines correspond to the true value of the slope.

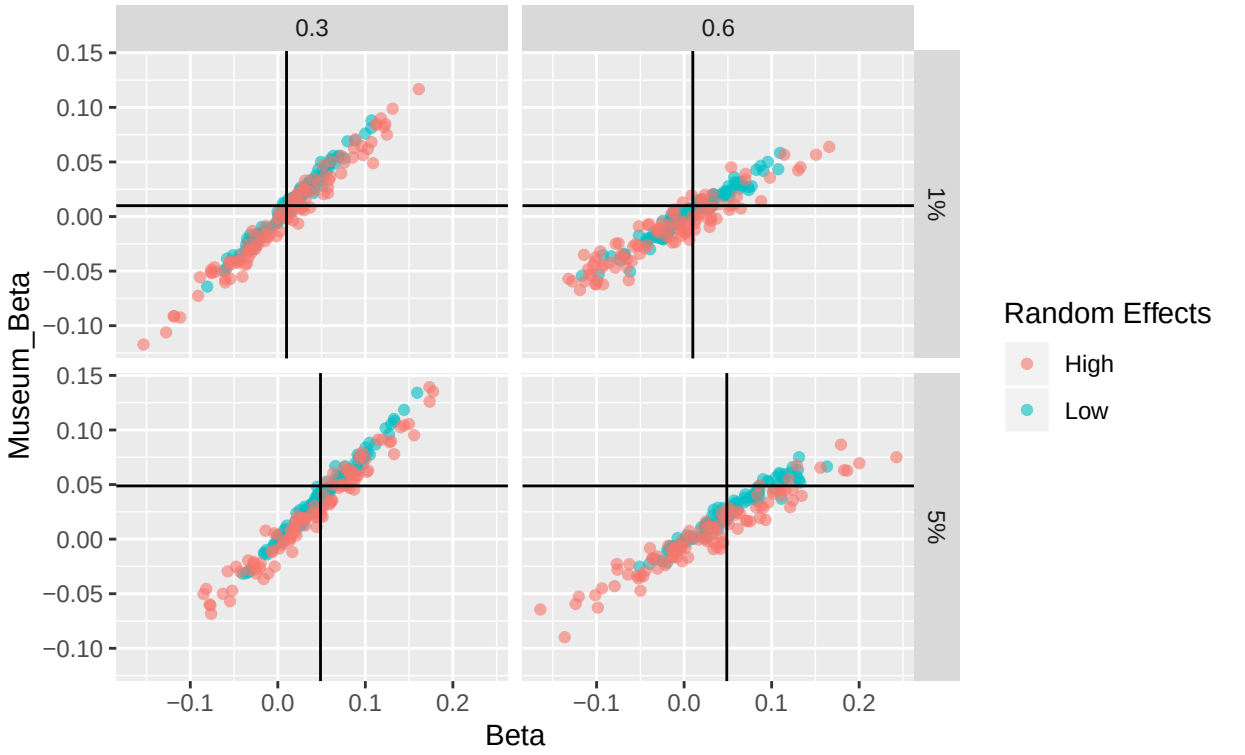


Figure 3: Plotting the estimated slopes for the Museum model Tobit model against eachother having the variance of individual specimens set to high. The vertical and horizontal lines correspond to the true value of the slope.

Application

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 analysis perform one analysis for each location. In the interest of keeping the analysis 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 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 Tobit model for the dataset of the museum, an analysis concering whether or not the log-normal distribution holds for the data 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 censored data, 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 quantiles 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 **Write number of the 3 figures**, 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). Furthermore, the transformation back, after operations, is set to *NULL* in order to stop the reverse transformation. Hence, the *cenros* function, with the reverse transformation set to *NULL* assumes a log-normal distribution, and so when using the *cenros* results in a distribution plot, a log-normal distribution assumption is tested.

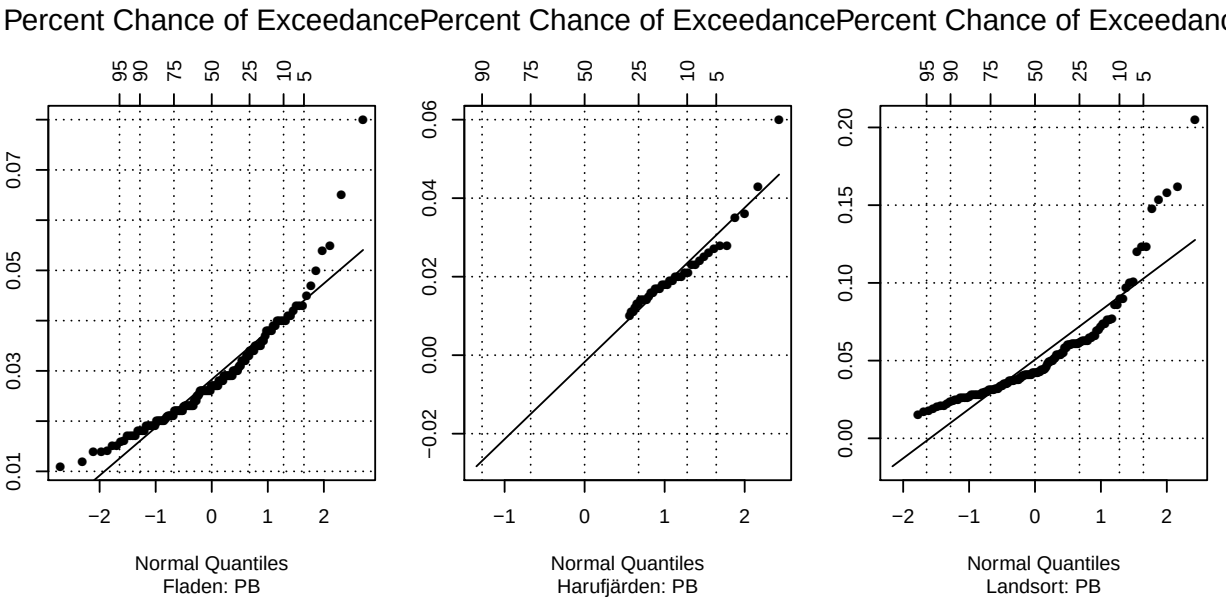
In order to estimate the percentiles for the uncensored data, regression on ordered statistics (ROS) is used by the *cenros* function. ROS is favorable over MLE when the proportion of censored data are too high (Helsel, 2005, pp. 86) which is the case in some locations for each metal (see Table 5-7), and for every location for chromium. Each data point is first given a rank i ranking the data point with the smallest values as $i = 1$. The ranks are then converted to procentiles 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, which is rarely the case for environmental studies. When points have the same values and therefor ties in the ranks, as is the case for censored data, each point gets it's own rank.

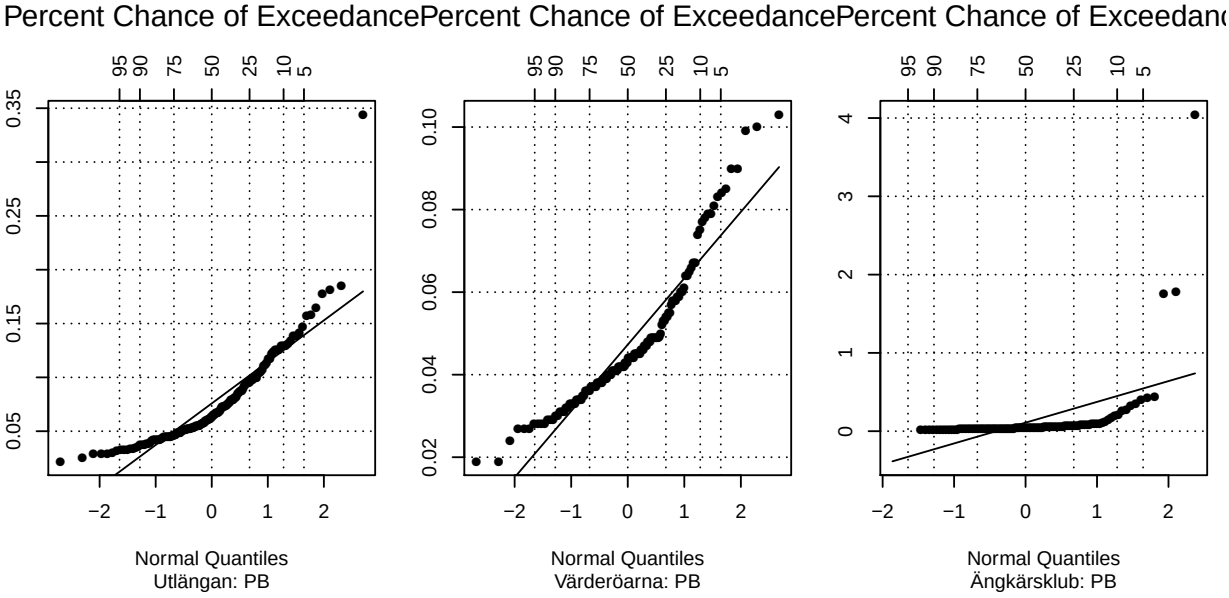
When the percentiles are calculated, they are fitted against the quantiles of a normal distribution. The uncensored data is used to calculated the slope and intercept for the linear regression between the logaritmic values of the data and the normal quantiles and thus, fitting this line is fitting a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

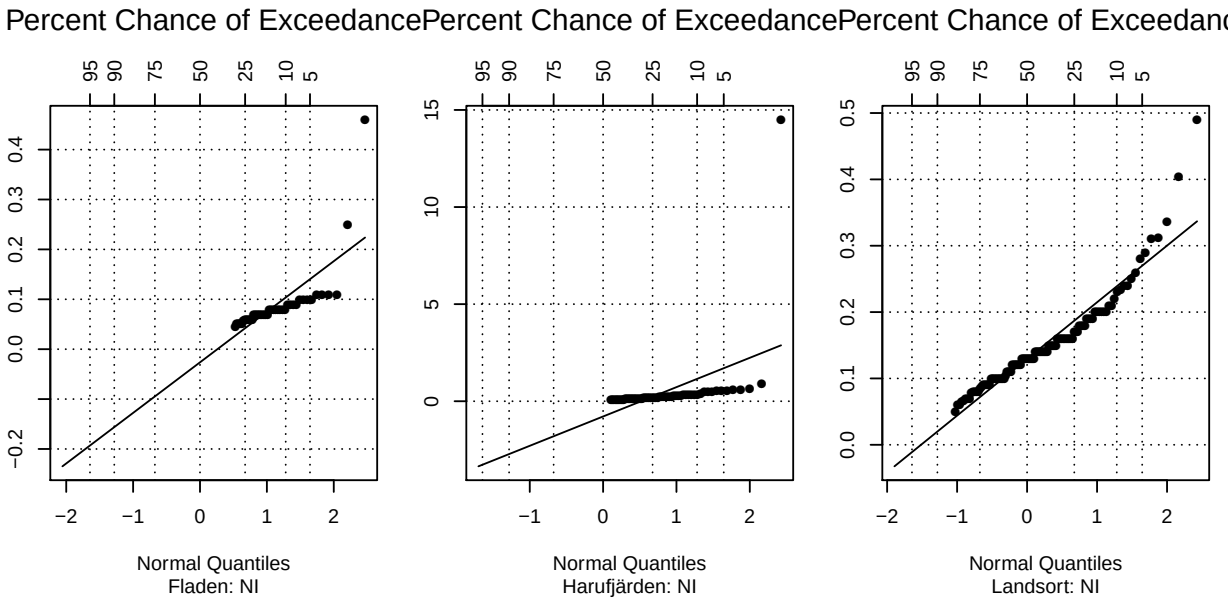
Applying Tobit model

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 or not 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 the standard deviation for individuals at each location. The standard deviation between-years was calculated using the same method as before with the *cenmle* function from the *NADA* packages. For lead, the standard deviation of each location, each year, had a standard deviation under 0.1 except for in 2017 at Ängkärsklubben where the standard deviation was around 1.2. For nickel, the yearly standard deviation was around $0.05 - 0.2$ through out except for a couple of instances, Harufjärden having a year with standard deviation around 1.4 and Ängkärsklubben having one year with a standard deviation at around 0.7. For chromium, the censoring proportion is too high to get a sensible estimation of the standard deviation, however, still using the same method, the standard deviation for each year lied around $0.05 - 0.15$ for the most part, having a couple of year and location combinations with a small increase in standard deviation. One that stood out was the standard deviation at Utlängan in 2008 had a standard deviation over 5

Considering the simulation study showing that the method of substitution being the most unreliable when having low noise for both between-years and for individual specimens and the Tobit model at the same time being the most precise in the very same scenario, studying the scenarios when having low standard deviation also seems the most interesting when comparing the new method with an already used one.


Percent Chance of Exceedance
Normal Quantiles
Fladen: PB
Normal Quantiles
Harufjärden: PB
Normal Quantiles
Landsort: PB


Percent Chance of Exceedance
Normal Quantiles
Utlängan: PB
Normal Quantiles
Värderöarna: PB
Normal Quantiles
Ängkärsklub: PB


Percent Chance of Exceedance
Normal Quantiles
Fladen: NI
Normal Quantiles
Harufjärden: NI
Normal Quantiles
Landsort: NI

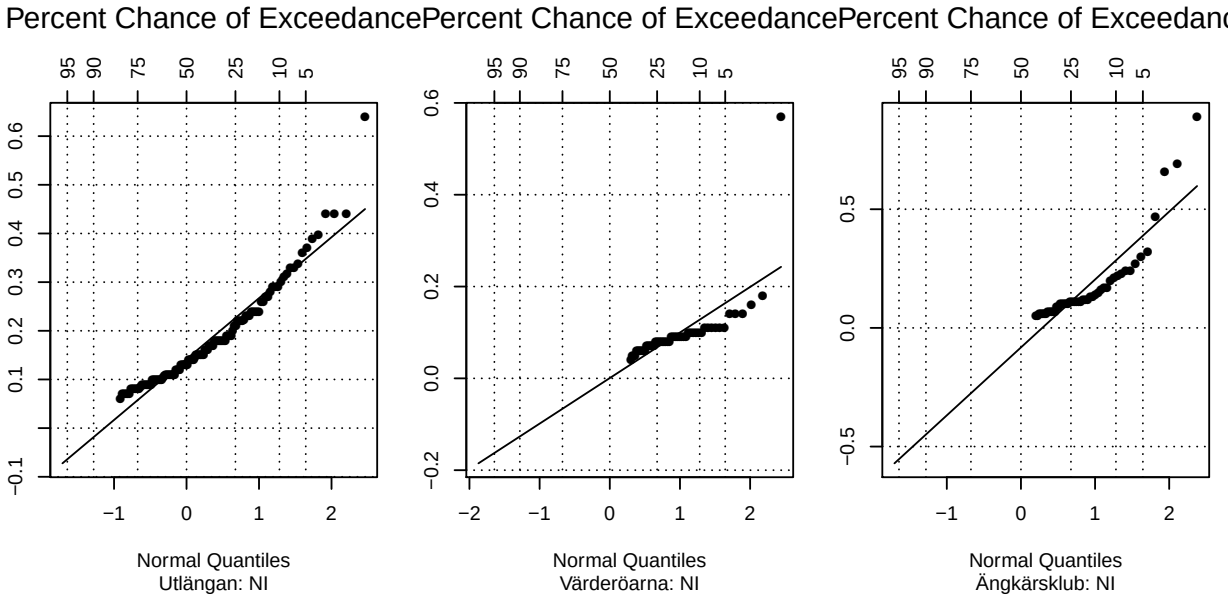
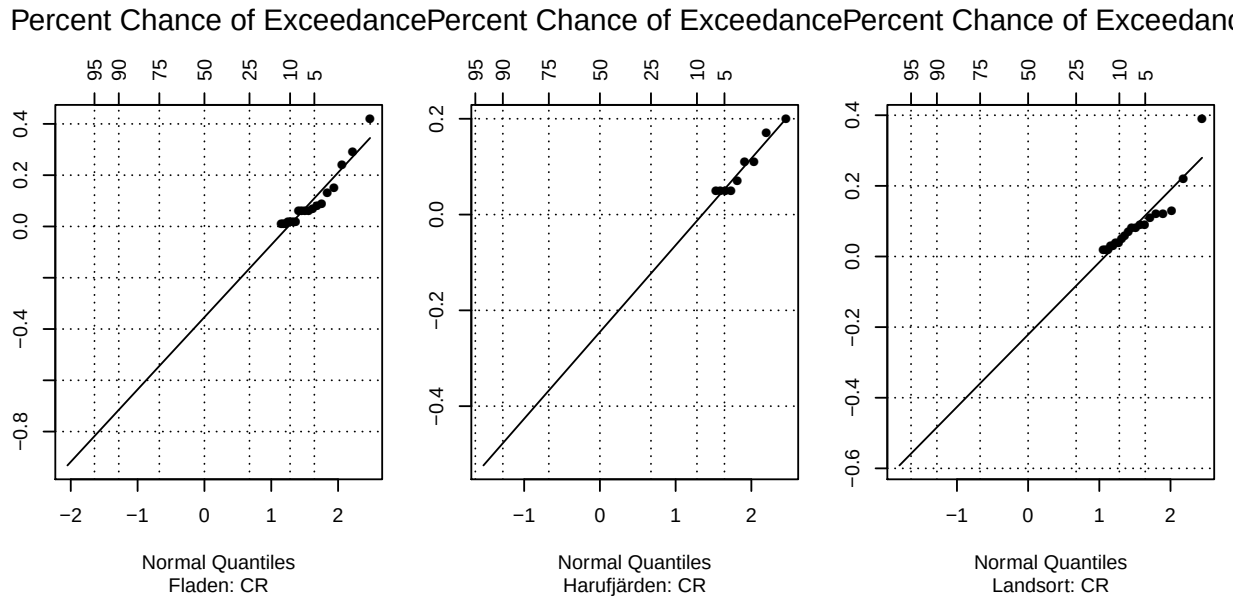
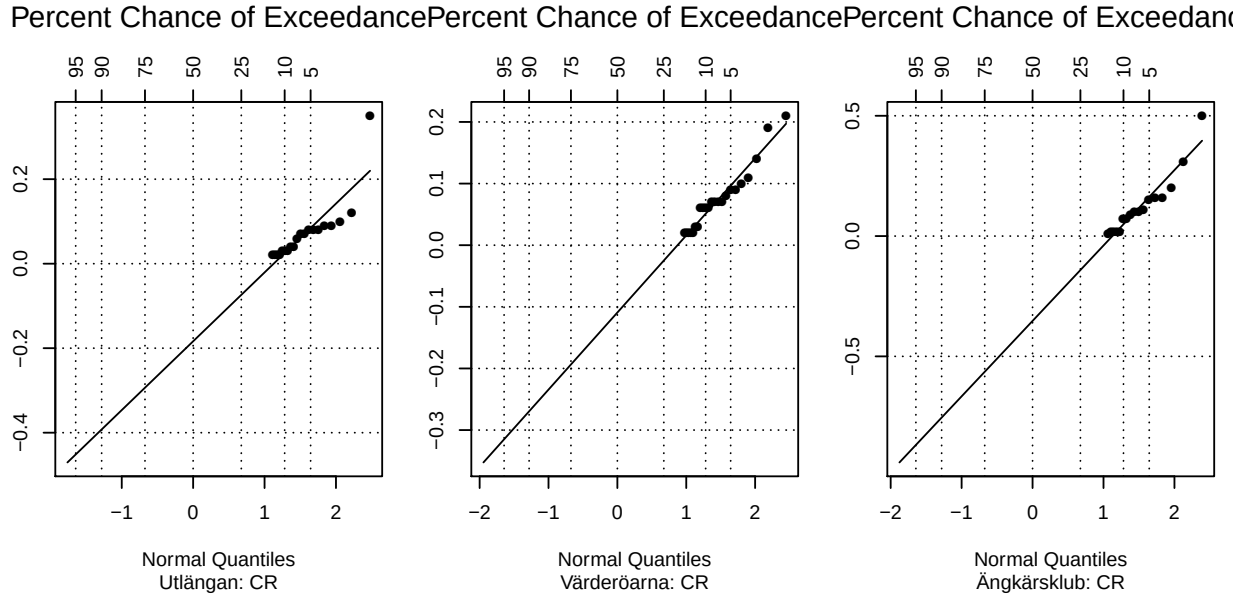

Percent Chance of Exceedance
Normal Quantiles
Utlängan: NI
Normal Quantiles
Värderöarna: NI
Normal Quantiles
Ängkärsklub: NI

Table 5: Result for Tobit model on data for PB in Herring (Log-Scale)

Location	Beta	Var(Beta)	%Censored	# of Years	# of Observations	Std of Location
Fladen	-0.0015375	0.000191563126344312	4.310345	12	232	0.1250827
Harufjärden	-0.0214947	0.00134959325998428	70.542636	11	129	0.0888893
Landsort	0.0191061	0.000354453808326119	3.053435	11	131	0.1699230
Utlängan	0.0090050	0.000360424523986072	0.000000	12	144	0.1976623
Väderöarna	0.0041565	0.000361767094022832	0.000000	12	134	0.1266632
Ängskärsklubb	0.0661702	0.00233012656289598	6.194690	11	113	0.3432503





Ta fram variansen för specifika platser när du testar Testa bara Herring? ty ingen annan har många observationer

Table 6: Result for Tobit model on data for NI in Herring (Log-Scale)

Location	Beta	Var(Beta)	%Censored	# of Years	# of Observations	Std of Location
Fladen	-0.0417653	0.0017139480308245	53.01724	12	232	0.2924028
Harufjärden	0.0558913	0.00941985046383763	53.48837	11	129	0.5365813
Landsort	-0.0096339	0.000294938060678381	14.50382	11	131	0.2685938
Utlängan	-0.0006600	0.000397713815068446	17.36111	12	144	0.3099659
Väderöarna	-0.0187921	0.00279162482605775	61.19403	12	134	0.2031083
Ängskärsklubb	-0.1104987	0.00219559024335967	56.63717	11	113	0.3643164

Table 7: Result for Tobit model on data for PB in Herring (Log-Scale)

Location	Beta	Var(Beta)	%Censored	# of Years	# of Observations	Std of Location
Fladen	-0.1157873	0.00757837280181469	84.48276	12	232	0.2075987
Harufjärden	0.0395350	0.00749512237240437	93.02326	11	129	0.2112722
Landsort	-0.0426512	0.0104651742907098	84.73282	11	131	0.2925520
Utlängan	-0.1241104	0.011782827039487	86.11111	12	144	0.2075850
Väderöarna	0.1078462	0.00911864629005241	82.83582	12	134	0.2549880
Ängskärsklubb	-0.0384072	0.0175056151154999	84.95575	11	113	0.4050687

Result

Conclusion

- EM-algorithm can stop at local maxima or sadle points, at sadle point the LH-fkn grows without bound.
- Can, and should, try with higher (or no) limit of iterations if not so time heavy.
- Test explicit starting values (based on what?)
- Should have more levels of each factor
- Tobit model seems more consistent in what factors affects estimates in what ways.

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