Methods for processing censored data FV2

Marc Roddis

7/17/2020

Introduction

The Swedish National Monitoring Programme for Contaminants (SNMPC) in freshwater biota has various goals and large scope (citation needed).

Our main goal in this study was to explore the viability of alternative methodologies for parameter estimation from censored data and to compare these alternatives with the methodology used by SNMPC.

Censored data is very common in environmental chemistry so our research area has been researched extensively by others. We will select well-regarded methods and apply these according to best practice, according to the cited works (citation needed).

At the outset, we limited the scope of our study by choosing to focus on the estimation of long-term time trends for the concentration of polychlorinated biphenyls (PCBs) in biological samples.

Our main idea was that since PCBs have similar chemical and physical properties their concentrations may be correlated such that censored measurements can be imputed using censored regression.

Our idea is supported by the SNMPC dataset since it has no censored data for CB153, whereas 34 % of the data for CB28 is censored.

More importantly, our exploratory data analysis of SNMPC data showed that CB153 and CB28 concentrations are strongly correlated.

Moreover, CB153 and CB28 also show a very similar rate of decrease over the time period 2003-2017.

Concretely, our idea is to impute censored CB28 values from the corresponding uncensored observations for CB153.

The resulting "imputed datasets" could then be used to obtain better parameter estimates than from the methodology currently used by SNMPC, which uses "substituted datasets" instead.

Specifically, SNMPC substitute all censored data by $\frac{LOD}{\sqrt{2}}$.

Our workflow

We now present an eight-step overview of our workflow in this section. We will give a more detailed description of steps 1-7 in the subsequent sections.

1. Selection of a set of parameter values, and generation of the simulation dataset accordingly.
2. Estimation of β from the simulation dataset by simple linear regression to get a benchmark to compare other methods against.
3. Selection of the proportion `cprop` of CB28 values to censor, and generation of the censored dataset accordingly.
4. Creation of a "completed dataset" by replacing censored data using some or all of our six methods. The completed datasets created by different methods will be distinct.
5. Estimation of β from each completed dataset by simple linear regression (by the same procedure as Step 2).
6. Estimation of β directly from the censored dataset by censored regression.
7. Presentation of the MSE, squared-bias, and variance, of the estimates from each method.
8. Repetition of steps 1-7 for various selections of parameter values.
9. Discussion of all results.

Model selection for generation of our simulation datasets

We begin by performing exploratory data analysis and model fitting from datasets from the SNMPC.

We do this in order to design our simulation studies to have real real-world relevance.

A large dataset `pcb.csv` was provided from SNMPC.

This dataset has 5056 observations of 18 variables; these variables include: measured concentrations of seven PCBs (CB28, CB53, CB101, CB118, CB138, CB153, CB180); year (1984-2017); an ID for each observation; and nine other variables such as species and age.

Our exploratory data analysis showed that

1. The most recent 15-year period 2003-2017 had sufficient relevant data, so we will focus solely on this time period.

2. It is reasonable to model the observed pcb concentrations as log-normal distributed.

3. The data for CB153 had no censored values, whereas CB28 data had the highest proportion of censored values. This proportion was 0.34.

4. Species is clearly a confounding variable for the association between CB153 and CB28, so we will focus solely on herring (since this was the species for which there were most observations). No other variable showed clear evidence for confounding.

From this basis, we create our test dataset from the original dataset `pcb.csv` by omitting all missing values of `CB28` and `CB153`, removing all observations except those from herring species, removing all observations prior to 2003, re-indexing 2003 as "year zero", removing all variables except `YEAR`, `CB28` and `CB153`, and omitting all censored observations.

We fit linear regression models to our test dataset for $y = CB28, x = CB153$ and for $y = log(CB28), x = log(CB153)$; the adjusted R-squared values were 0.93 and 0.96 respectively.

Based on this, we decide we will use logarithmised concentrations throughout, which we will model as normally distributed.

We have three variables $log(CB28)$, $log(CB153)$, and $YEAR$; we will denote these as Y, X and A respectively, throughout the rest of our work.

We make the key observation that Y and X are strongly correlated in our test dataset, which means that our key idea of using censored regression for Y on X to make imputations for censored Y values is plausible.

We also fit a model to our test dataset for the regression X on A, which gave

$$E(X|A) = -2.91 - 0.02A$$

the corresponding fitted model for Y on X is

$$E(Y|X) = -3.18 + 0.79X$$

the residual standard error was equal to 0.1 from both models.

From this basis, we will generate our simulation datasets as follows:

1. We will also always simulate a 15-year period; we will use $A \in \{0, 1, 2, ..., 14\}$ to denote year.

2. For every year, we will generate the same number of observations for Y and X, we will call this number the sample size N. For our first simulation we will use `sample size` $= 100$ because there are typically 100 observations on herring each year by the Monitoring Program.

3. We generate all x_i from

$$x_i = -2.91 - \beta_A a_i + e_i$$

where $i \in \{1, 2, ..., N\}$ denotes the ith observation, and the noise is modeled as normally distributed with $mean = 0$ and $variance = 0.1^2$, i.e. $e_i \sim N(0, 0.1^2)$.

We will be interested in evaluating our methods for various values of β_A, so this will be a variable parameter for our simulations.

4. We generate all y_i from

$$y_i = -3.18 + 0.79x_i + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$.

We will be interested in evaluating our methods for various values of β_A and σ^2, so these will be the two variable parameters for our simulation datasets.

Estimation of β from the simulation dataset by simple linear regression

The main body of our work will be to evaluate various methods for the estimation of the regression coefficient $\widehat{\beta}$ for datasets containing censored values.

In this section, we will instead assume that there are no censored values, which allows us to find estimates by simple linear regression.

We will later use these estimates as the benchmark for evaluating the methods we use in the main body of our work.

Our primary goal is to find $\widehat{\beta}$, which means we will find estimates for β where $Y_i = \alpha + \beta a_i + \varepsilon_i$.

To specify this model, we first substitute

$$x_i = -2.91 - \beta_A a_i + e_i$$

nto

$$y_i^* = -3.18 + 0.79 x_i + \epsilon_i$$

which gives

$$y_i^* = -3.18 + 0.79(-2.91 - \beta_A a_i + e_i) + \epsilon_i$$

$$= -3.18 + 0.79 - 2.91 - 0.79\beta_A a_i + 0.79 e_i + \epsilon_i$$

$$= \alpha + \beta a_i + \varepsilon_i$$

where $\alpha = -3.18 + 0.79 \times -2.91 = -5.4789$, and $\beta = 0.79\beta_A$.

Also $\varepsilon_i = 0.79 e_i + \epsilon_i$, where $e_i \sim N(0, 0.1^2)$ and $\epsilon_i \sim N(0, \sigma^2)$.

Creation of datasets with censored values

For our simple linear regression of Y on X we used $y_i = -3.18 + 0.79 x_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$.

We now use y_i^* instead of y_i, where y_i^* refers to the ith observation prior to it being observed.

This means that after y_i has been observed and left-censoring at LOD has been applied, we have $y_i = y_i^*$ if $y_i^* > LOD$ and $y_i = LOD$ if $y_i^* \leq LOD$.

We will determine LOD by censoring a fixed proportion, which we denote as $cprop$, of all observed y_i values, for each of our simulations.

We will be interested in evaluating our methods for various values of $cprop$, which is our variable parameter of primary interest.

Moreover since $LOD = cprop * 100$th percentile of all y_i values, the value of $LOD|cprop$ is constant and thus independent of A.

Our primary goal is to find $\widehat{\beta}$, which means we will find estimates for β where $Y_i = \alpha + \beta a_i + \varepsilon_i$.

Creation of a "completed dataset" by replacing censored data using some or all of our six methods

We view every censored observation as having a true but unknown value within the interval $[0, LOD]$.

Our goal is to replace all such unknown values with a known value such that the resulting values are as close to the true values as possible.

The most straightforward way to this is by substitution, which means that all censored values from a censored dataset are substituted by the same fixed value, which is a fraction of LOD.

The monitoring program that motivates our work uses substitution by $\frac{LOD}{\sqrt{2}}$, which is the most commonly used value based in the research literature cited in this report.

The second most commonly used value is $\frac{LOD}{2}$.

The largest possible value that can be used for substitution is LOD, since all of the censored values are known to lie within the interval $[0, LOD]$.

Our three substitution methods will use substitution by either LOD, $\frac{LOD}{\sqrt{2}}$ or $\frac{LOD}{2}$; we name them `subst1`, `subst2`, and `subst4`, respectively.
Our notation is based on the fact that $LOD = \frac{LOD}{1}$, and that $2 = \sqrt{4}$ and $1 = \sqrt{1}$, respectively.

Our rationale for choosing these three methods is that since $\frac{LOD}{2} < \frac{LOD}{\sqrt{2}} < LOD$ we can compare results from the `subst2` method that SNMPC uses with two alternative substitution methods, which use substitution by lower and higher values, respectively.

However, such substitution methods are limited since they do not use observations from other variables of the dataset.

Our conjecture is that we can use censored regression to impute censored y_i values from the corresponding uncensored x_i values, thus leveraging the strong correlation between the Y and X.

We call our main imputation by censored regression method `censReg1` , because it is based on a censored regression model with 1 predictor variable X, as described in the following section.

Creation of completed datasets by censored regression by our main method `censReg1`

Each observation y_i^* is from a normal distribution with mean $\mu_{i_X} = \alpha_X + \beta_X x_i$ and variance σ^2, which has pdf

$$f(y_i^*) = \frac{exp[(-1/2)((y_i^* - \mu_{i_X})/\sigma)^2]}{\sigma\sqrt{2\pi}}$$

which we can write as

$$f(y_i^*) = \frac{\phi((y_i^* - \mu_{i_X})/\sigma)}{\sigma}$$

where ϕ is the pdf of a normal distribution with $mean = 0$ and $variance = 1$.

The probability that y_i^* is censored equals

$$P(y_i^* \leq LOD) = \Phi((LOD - \mu)/\sigma)$$

where Φ is the cdf of a normal distribution with $mean = 0$ and $variance = 1$.

Every y_i^* is either censored or not, so we will use the indicator variable $I = 1$ for censored, and $I = 0$ for not censored. Moreover, we assume that y_i are all independent, which means that the joint likelihood over all observations is the product of the density functions for all y_i.

This gives us the likelihood function L

$$L = \prod_{i=1}^{A}[[(1/\sigma)\phi((y_i - \mu_{i_X})/\sigma)]^{1-I} \times \Phi((LOD - \mu_{i_X})/\sigma)^I]$$

So the log-likelihood function is

$$\log(L) = \sum_{i=1}^{A}[(1-I)[\log(\phi((y_i - \mu_{i_X})/\sigma)) - \log(\sigma)] + I \times \log[\Phi((LOD - \mu_{i_X})/\sigma)]]$$

which equals

$$\log(L) = \sum_{i=1}^{A}[(1-I)[\log(\phi((y_i-(\alpha_X+\beta_X x_i))/\sigma))-\log(\sigma)]+I\times\log[\Phi((LOD-(\alpha_X+\beta_X x_i))/\sigma)]]$$

We will use the `censReg()` function from the censReg package in R to maximise this log-likelihood function to obtain the maximum likehood estimates $\widehat{\alpha_X}$, $\widehat{\beta_X}$ and $\hat{\sigma}$.

We will then perform imputation as follows.

Every censored value y_i is substituted by the expected value of a truncated normal distribution, which we describe as originating from a normal distribution with $mean = \widehat{\mu}_{i_X} = \widehat{\alpha}_X + \widehat{\beta}_X x_i$, $variance = \sigma^2$, and with truncation at $y = LOD$.

In our practice, we used the `etruncnorm()` function from the `truncnorm` R package to calculate every such expected value.

We will use the term “completed dataset” for every dataset that results from imputation or substitution.

Variations on the `censReg1` method

We will also use two methods that are closely related to `censReg1` for the purpose of comparison; we call these methods `censReg1naive` and `censReg2`.

The `censReg1naive` method

The only difference from `censReg1` in our `censReg1naive` method is that the latter uses the corresponding non-truncated normal distribution rather than the truncated one. Our conjecture is that estimates of β from `censReg1naive` will have significantly higher bias than the corresponding estimates from `censReg1`. Our rationale is that the censored y_i values could be substituted by values that are higher than LOD, whereas the true value is known to be not higher than LOD.

`censReg1naive` is the same as `censReg1` except that a non-truncated normal distribution is used in the imputation step. We conjecture that this will result in estimates with higher bias than from `censReg1`. This was done to check that we get a more biased estimate because it is possible that the imputed values are above LOD, despite the fact that the censored value are below LOD.

The `censReg2` method

`censReg2` uses two predictor variables X and N.

We conjecture that using one additional redundant predictor variable will result in estimates with higher variance than from `censReg1`.

The mathematical formulation for this method corresponds to that presented above for `censReg1`, except that we model each observation y_i^* as from a normal distribution with mean $\mu_{i_{X,A}} = \alpha_{X,A} + \beta_X x_i + \beta_A a_i$ and variance σ^2.

This means that the likelihood function for `censReg2` will have the same form as that for `censReg1`; it will differ only in having $\mu_{i_{X,A}}$ in place of μ_{i_X}.

Consequently, maximisation of the corresponding log-likelihood function gives the maximum likehood estimates $\widehat{\alpha_X}$, $\widehat{\beta_X}$, $\widehat{\beta_X}$ and $\hat{\sigma}$.

Therefore, every censored value y_i is substituted by the expected value of a truncated normal distribution, which we describe as originating from a normal distribution with $mean = \widehat{\mu}_{i_{X,A}} = \widehat{\alpha}_X + \widehat{\beta}_X x_i + \widehat{\beta}_A a_i$, $variance = \sigma^2$, and with truncation at $y = LOD$.

Estimation of β directly from the censored dataset by censored regression by the `censReg0impute` method

This method differs from `censReg1` in the choice of predictor variable for the model for the maximum likelihood estimation step.

We have seen that the `censReg1` method uses X_i as the predictor for this step.

The `censReg0impute` method uses A_i as the predictor instead for this step.

Thus the estimate $\widehat{\beta_A}$ is found directly from the maximisation of the corresponding log-likelihood function, without any imputation step.

We conjecture that since $|\beta_X| = 0.79$ is much greater than $|\beta_A|$, `censReg0impute` will result in estimates with higher variance than from `censReg1`.