Mathematical theory v1

Marc Roddis

7/15/2020

Parameter values are based on estimates from real data

We will use parameters values estimated from real data for our first simulation to study the effectiveness of various methods of dealing with censored data. In subsequent studies, we will investigate how generally applicable these methods are for various other possible choices of parameter values. We will use logarithmised concentrations for CB28 and CB153 and refer to these as Y and X respectively throughout.

100 values for cb153 per year, for 15 years, were generated and denoted as `cb153` from

$$CB153 = -2.91 - 0.02 * YEAR$$

with added noise (modeled with normal distribution with mean = 0 and sd = 0.1).

From every such CB153 value, the corresponding value for CB28 was generated from

$$CB28 = -3.18 + 0.79 * CB153$$

, again with added noise (modeled with normal distribution with mean = 0 and sd = 0.1). From these equations, we deduce that `true_beta28year` = 0.79 * -0.02; we will use this as the “true” value against which we evaluate the estimates for this parameter from applying various methods to censored data.

From real data for the 15 year period 2003-2017, 34 % of the cb28 values were censored, so we will use the parameter value `cprop=0.34` in this first simulation study. Values of cb28 below the value below the level of detection (LOD) were then censored. The LOD was calculated from the cprop*100th percentile of the simulated data at each iteration.

Specification of our mathematical model

We will use logarithmised concentrations for CB28 and CB153 and denote these as Y and X respectively throughout.

Our model covers a 15-year period; we will use $A \in \{0, 1, 2, \ldots, 14\}$ to denote year.

We first use $x_i = -2.91 - \beta_N a_i + \epsilon_i$ where $i \in \{1, 2, \ldots, N\}$ denotes the ith observation.

We will use the same distribution $\epsilon_i \sim N(0, 0.1^2)$ for the noise term of this model throughout this study.

However, we will use various values of β_X in this study.

We then use $y_i^* = -3.18 + \beta_X x_i + \varepsilon_i$ where y_i^* refers to the ith observation prior to it being observed. We will use $\beta_X = 0.79$, which is based on estimates from real data, for all our simulations in this study.

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 use vary the amount of noise $\varepsilon_i \sim N(0, \sigma^2)$ in this model by using various values for σ.

Throughout this study, we will determine LOD by censoring a proportion, which we denote as $cprop$, of all observed y_i values.

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

In summary, $cprop$, β_X and σ are our variable parameters of primary interest for this study.

Theoretical basis for imputation by censored regression

Each observation y_i^* is from a normal distribution with mean $\mu_i = -3.18 + \beta_X x_i$ and variance σ^2, which has pdf

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

which we can write as

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

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

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

So the log-likelihood function is

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

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

We then impute for censored values by replacing them with the expected value of a truncated normal distribution. This distribution is best viewed as originating from a normal distribution with $ mean = }$, $variance = \sigma^2$, and truncation at $y = LOD $. To do this, we used the `etruncnorm()` function from the `truncnorm` R package, from which every censored value was subsituted with the corresponding imputed value.

The fact that it is necessary to use a truncated normal distribution rather than the corresponding non-truncated distribution will be demonstrated by our simulation results. We will use truncation for our main method `censReg1naive`, and show that using the same method but without truncation `censReg1naive` results in significantly higher bias in estimates of $\beta_X\beta_N$.

Specification of our censoring methods

Our main goal will be to generate results using various censoring methods, from various parameters values and to interpret our results mathematically.

We will use the two methodologies to process the censored data: substitution, and imputation from censored regression.

Substitution 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, which give values that are systematically higher than the true values.

We will denote substitution by these values as SUBST2, SUBST4, and SUBST1, respectively.

This notation is based on the fact that $LOD = \frac{LOD}{1}$, and that $2 = \sqrt{(4)}$ and $1 = \sqrt{(1)}$, respectively.

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

We will also experiment with the following three variations on our main **`censReg1`** method:

`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.

`censReg0impute` is based on a censored regression model in which the predictor variable is N (whereas **`censReg1`** uses X). As its name suggests, no imputation is done in this method. We conjecture that since $|\beta_X|$ is much greater than $|\beta_N|$, **`censReg0impute`** will result in estimates with higher variance than from **`censReg1`**.

`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`**.