Methods for processing censored data
FVFVFV6

Marc Roddis

8/3/2020

Contents

1 Introduction

The Swedish National Monitoring Programme for Contaminants (SNMPC) (Danielsson, Faxneld, and Soerensen 2020) in freshwater biota has various goals and large scope.

The SNMPC goals that are most relevant for this study are:

"To estimate the current levels and normal variation of various contaminants in marine biota [. . .]"

"To monitor long-term time trends and estimate the rate of changes found."

Contaminant concentrations reported as being below the Level Of Quantification (LOQ) are termed left-censored data; such data are very common from environmental monitoring programs (such as SNMPC).

SNMPC includes such values in the analysis as if they were true observations with a value of $\frac{\text{LOQ}}{\sqrt{2}}$.

However, such substitution methods have been criticised in the research literature.

For example, (Helsel 2006) found that substitution methods have low robustness (i.e their performance is highly situational); they wrote:
"Substituting values for nondetects should be used rarely, and should generally be considered unacceptable in scientific research. There are better ways."

Our main goal in this study is to explore alternative methodologies for the manipulation of censored data and to evaluate these alternatives against the methodology used by SNMPC, with the purpose of finding a "better way".

At the outset, we limit the scope of our study by choosing to view all concentrations as either uncensored (i.e having a known value which is greater than the LOQ), or censored (i.e the value is not quantified but is known to be greater than 0 and less than the LOQ).

This means that considerations such as multiple reporting limits, or missing data, lie outside the scope of our study.

We will also focus specifically on the estimation of long-term time trends for the concentration of polychlorinated biphenyls (PCBs) in biological samples.

PCBs are synthetic chemicals used in manufacturing processes, especially as plasticizers, insulators and fire retardants.

PCBs are widely distributed in the environment, degrade very slowly, bioaccumulate in biota to high concentrations, and can be harmful to human health.

PCBs are included in the Stockholm Convention; they have been banned since 1978.

Since PCBs are widely distributed, and have similar chemical and physical properties, we conjecture that the concentrations of different PCBs are correlated.

Let us denote a PCB with censored data as C and a fully observed PCB as F; our main idea is to use the censored regression C on F to impute the censored data.

We will verify the viability of this idea, and demonstrate its relevance in relation to SNMPC, in the following section on exploratory data analysis (EDA).

We will then perform simulation studies to obtain results that allow us to evaluate imputation by censored regression against the substitution methodology currently used by SNMPC.

The distinction between substitution and imputation has been made clear (Helsel 2012):

"Substitution is NOT imputation, which implies using a model such as the relationship with a correlated variable to impute (estimate) values. Substitution is fabrication."

We will conclude by discussing our results in the context of our stated goals.

1.1 Exploratory Data Analysis (EDA)

We begin by performing EDA, 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

$$E(X|A) = -2.91 - 0.02A \tag{1}$$

the corresponding fitted model for the regression Y on X is

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

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

We will use these parameter estimates and findings to guide our implementation of our simulation studies (see Chapter 3).

In summary, we found that X is fully observed whereas 34 % of the data for Y is censored, and that the concentrations of X and Y are strongly correlated, and show a very similar rate of decrease.

These findings establish that our idea to impute censored Y values from the corresponding uncensored observations for X from the censored regression Y on X is relevant in the context of SNMPC data.

2 Statistical methods

2.1 Overview and terminology

In our simulation study, we will first generate "uncensored datasets" with three variables Y, X, and A.

From each of these, we will apply Type II left-censoring to Y to get the corresponding "censored dataset".

The censored data will then be replaced by values that are either fabricated by substitution or imputed by censored regression to obtain the corresponding "completed dataset".

In some cases, we will also omit the censored observations to get the corresponding "incomplete dataset".

We will then fit linear models for the regression Y on A to the resulting completed (and in some cases, incomplete) datasets, and use these to estimate the regression coefficient σ, and predict $E(Y|A = a)$ for $a \in \{0, 1, 2, ..., 14\}$.

We will calculate the MSE, squared-bias, and variance for these estimates and predictions.

We will then evaluate these results against the corresponding results obtained by simple linear regression on the corresponding uncensored dataset.

2.2 Imputation by censored regression

This method combines regression imputation and censored regression.

We will describe these two concepts in mathematical terms in the following two sub-sections.

2.2.1 Regression imputation

Suppose we have a dataset D with n observations and two associated variables x and y, of which c of the observations are complete.

Suppose also that the other $n - c$ observations of D are incomplete since whilst x is fully observed, y is missing.

Therefore

$$D = \{(x_1, y_1), (x_2, y_2), ..., (x_c, y_c), (x_{c+1}, \text{NA}), ..., (x_n, \text{NA})\}$$

where NA represents a missing value.

Performing regression imputation for dataset D means that we first find the regression equation

$$y = \alpha_X + x\beta_X \tag{2}$$

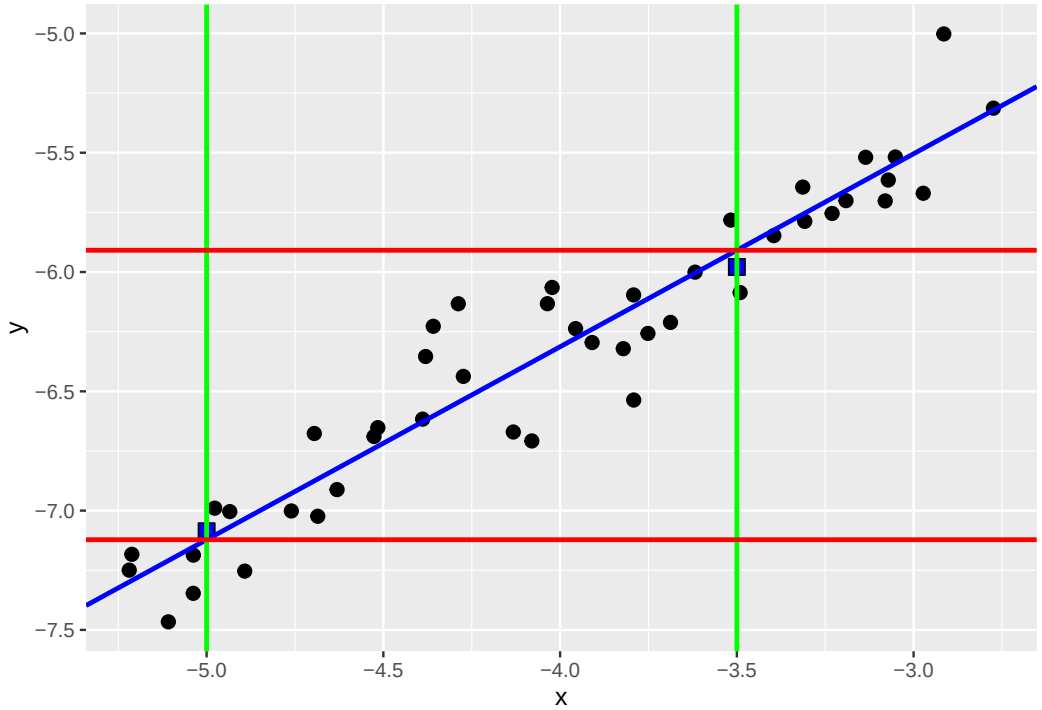
for y on x based on the c complete observations $D_c = \{(x_1, y_1), (x_2, y_2), ...(x_c, y_c)\}$.

We then impute each of the missing observations $\{y_{c+1}, ..., y_n\}$ by substituting the corresponding x value into the regression equation.

This is illustrated in the Figure 2.1 below for the case where there are 45 complete observations shown as black dots, and two incomplete observations $(x_{46}, y_{46}) = (-5.00, \text{NA})$ and $(x_{47}, y_{47}) = (-3.50, \text{NA})$, shown as green vertical lines.

The imputed values $y_{46} = \alpha_X - 5.00\beta_X = -7.12$ and $y_{47} = \alpha_X - 3.50\beta_X = -5.91$ are shown as red horizontal lines.

Thus the completed dataset is $C = D_c \cup \{(-5.00, -7.12), (-3.50, -5.91)\}$.



In general, the main advantage of regression imputation is that it uses information known about the association between x and y to impute information about y.

The main disadvantage is that the imputed values all lie on the regression line so the resulting completed dataset has unrealistically low variance.

Further discussion of the pros and cons of regression imputations lies outside the scope of this report.

2.2.2 Censored regression

To say that an observation is censored means that its value is known to lie within a certain closed or half-open interval.

Let y_i^* denote the ith observation prior to it being observed.

If for all $i \in \{1, 2, ..., n\}$, $y_i = y_i^*$ for $y_i^* > a$, and $y_i = a$ for $y_i^* \leq a$, we say that y is left-censored at a.

Similarly if, $y_i = y_i^*$ for $y_i^* \geq b$, and $y_i = b$ for $y_i^* \geq b$, we say that y is right-censored at b.

We will focus solely on left-censored data throughout this study.

We will denote the threshold for censoring as LOQ (rather than a), where LOQ denotes the limit of detection, as defined in the previous chapter.

Suppose we have a dataset D with n observations and two associated variables x and y, and that there are no missing values.

Suppose also that whilst all n observations of x are fully observed, only f of the observations of y are fully observed, and the remainder are left-censored at LOQ.

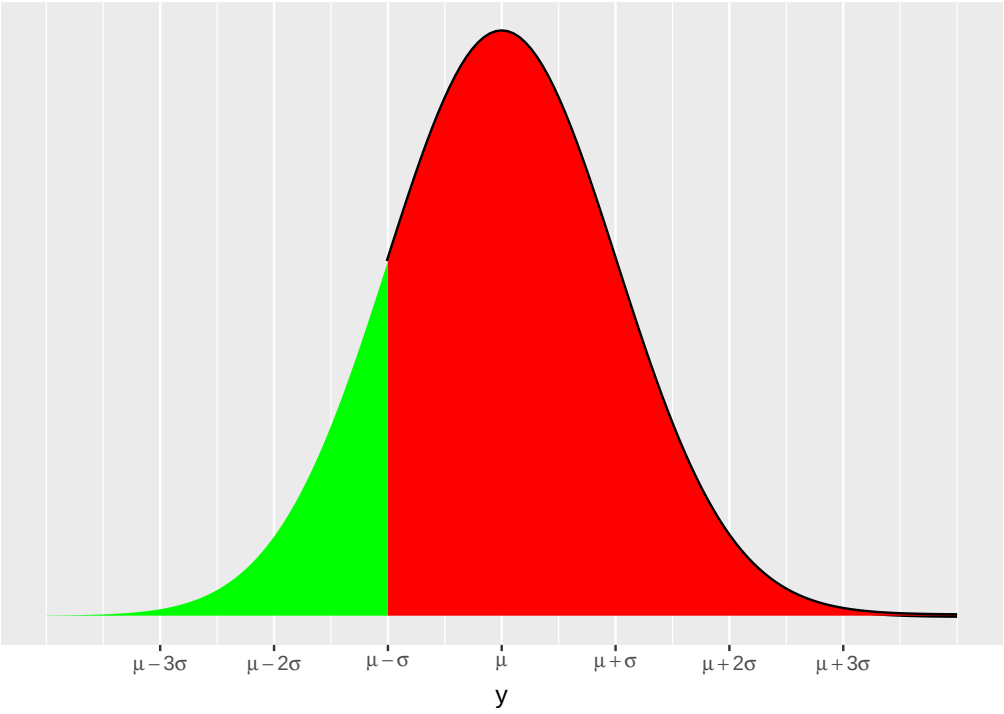
Therefore

$$D = \{(x_1, y_1), (x_2, y_2), ..., (x_f, y_f), (x_{f+1}, \text{LOQ}), ..., (x_n, \text{LOQ})\}$$

where the f full observations have been assigned the subscripts $1, 2, ..., f$.

If we assume that the y_i^* each have a normal distribution with mean $= \mu$ and variance $= \sigma^2$, then the y_i have the corresponding distribution, truncated at LOQ.

This illustrated in Figure 2.2 for LOQ $= \mu - \sigma$; the green and red regions show the assumed distribution of the uncensored observations, and of the true values for the censored observations, respectively.



In the censored regression context, each observation y_i^* is from a normal distribution with mean

$$\mu_{i_X} = \alpha_X + \beta_X x_i \tag{3}$$

and variance σ^2, where α_X and β_X are the intercept and slope parameters for the regression y on x.

The corresponding probability density function is

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

which we can write as

$$f(y_i^*) = \frac{\phi((y_i^* - \mu_{i_X})/\sigma)}{\sigma} \tag{4}$$

where μ_{i_X} is given by (3), and ϕ is the pdf of a normal distribution with mean = 0 and variance = 1.

The probability that y_i^* is censored equals

$$P(y_i^* \leq \text{LOQ}) = \Phi((\text{LOQ} - \mu_{i_X})/\sigma)$$

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

Every y_i^* is either censored or not.

We will use the indicator variable, where $I_i = 1$ and $I_i = 0$ denote that y_i is censored, and not censored, respectively.

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

$$L = \prod_{i=1}^{n} \left[[(1/\sigma)\phi((y_i - \mu_{i_X})/\sigma)]^{1-I_i} \times \Phi((\text{LOQ} - \mu_{i_X})/\sigma)^{I_i} \right] \tag{5}$$

So the log-likelihood function is

$$\log(L) = \sum_{i=1}^{n} \left[(1-I_i)[\log(\phi((y_i-\mu_{i_X})/\sigma))-\log(\sigma)]+I_i\times\log[\Phi((\text{LOQ}-\mu_{i_X})/\sigma)] \right]$$

which equals

$$\log(L) = \sum_{i=1}^{n} \left[(1-I_i)[\log(\phi((y_i-\mu_{i_X})/\sigma))-\log(\sigma)]+I_i\times\log[\Phi((\text{LOQ}-\mu_{i_X})/\sigma)] \right] \tag{6}$$

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

Recall that for regression imputation the missing values were imputed by

$$\hat{y}_i = E(Y|X = x_i) = \hat{\alpha}_X + \hat{\beta}_X x_i = \mu_{i_X}$$

where $\widehat{Var}(\hat{y}_i) = \hat{\sigma}^2$ and $\hat{y}_i \sim N(\hat{\alpha}_X + \hat{\beta}_X x_i, \hat{\sigma}^2)$.

To perform imputation by censored regression, we substitute every censored observation (x_i, y_i) by its imputed value $(x_i, \hat{y}_i)$, where

$$\hat{y}_i = E(Y|X = x_i, Y < \text{LOQ}) \tag{7}$$

using equation (8) from (Donald R. Barr and E. Todd Sherrill 1999) gives

$$\hat{y}_i = -\hat{\sigma}\frac{exp[((-1/2)(\mu_{i_X} - \text{LOQ})/\hat{\sigma})^2]}{[1 - (\Phi(\mu_{i_X} - \text{LOQ})/\hat{\sigma})]\sqrt{2\pi}} + \mu_{i_X} \tag{8}$$

where $\mu_{i_X} = \hat{\alpha}_X + \hat{\beta}_X x_i$.

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

We will refer to this imputation model as `censReg1`, since it is based on censored regression with one predictor variable.

We will also study imputation from the censored regression of y on the two predictors x and a, which we will denote as `censReg2`.

The mathematical formulation for `censReg2` corresponds to that presented above for `censReg1`, except that we model each observation y_i^* as from a normal distribution with mean

$$\mu_{i_{X,A}} = \alpha_{X,A} + \beta_X x_i + \beta_A a_i \tag{9}$$

and variance σ^2.

This means that the likelihood function for the `censReg2` model is given by substitution of (9) into (5).

Consequently, maximisation of the corresponding log-likelihood function gives the maximum likehood estimates $\widehat{\alpha}_X$, $\widehat{\beta}_X$, $\widehat{\beta}_A$ and $\hat{\sigma}$.

This means that the censored y_i are imputed by $\hat{y}_i$, which is obtained by substitution of (3) by (9) in (8).

In summary, our MLE calculations use the values of the uncensored data, the censoring proportion, and the formula for the assumed distribution.

The resulting parameter estimates have the maximum likelihood of giving these values and censoring proportion under this assumption.

2.3 Graphical illustration of our data manipulation methods

2.3.1 Imputations based on the `censReg1` model

We now show (in Figure 2.3) the distribution of the ML estimates for y_i based on the `censReg1` model for an illustrative example.

This example uses the same dataset D as in the previous section, except that D now has no missing values.

We have instead censored all y_i such that $y_i \leq \text{LOQ}$ values, where LOQ has been chosen as $\text{LOQ} = \text{median}(y)$.

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

Concretely, $LOQ = (cprop \times 100)$th percentile of all y_i values, which means that we denote $\text{LOQ} = \text{median}(y)$ as $\text{cprop} = 0.5$

In Figure 2.3, we have displayed the true value of every data point of D as a black dot, regardless of whether or not it is censored, for the purpose of illustrating our point.

In this figure, the red line shows $y = E(Y|X = x)$, the green line shows $y = E(Y|X = x, Y < \text{LOQ})$, and the blue line shows $y = \text{LOQ}$.

We have selected two observations for which $y_i \leq \text{LOQ}$ and drawn black vertical lines through them.

We denote these data points as (x_{I1}, y_{I1}) and (x_{I2}, y_{I2}),

For these points, the green squares show the $(x_{I1}, \hat{y}_{I1})$ and $(x_{I2}, \hat{y}_{I2})$ imputed from $y = E(Y|X = x, Y < \text{LOQ})$, whilst the red squares show the corresponding imputations from $y = E(Y|X = x)$.

We see that the green and red squares lie at the intersection of each black line with the green and red curves, respectively.

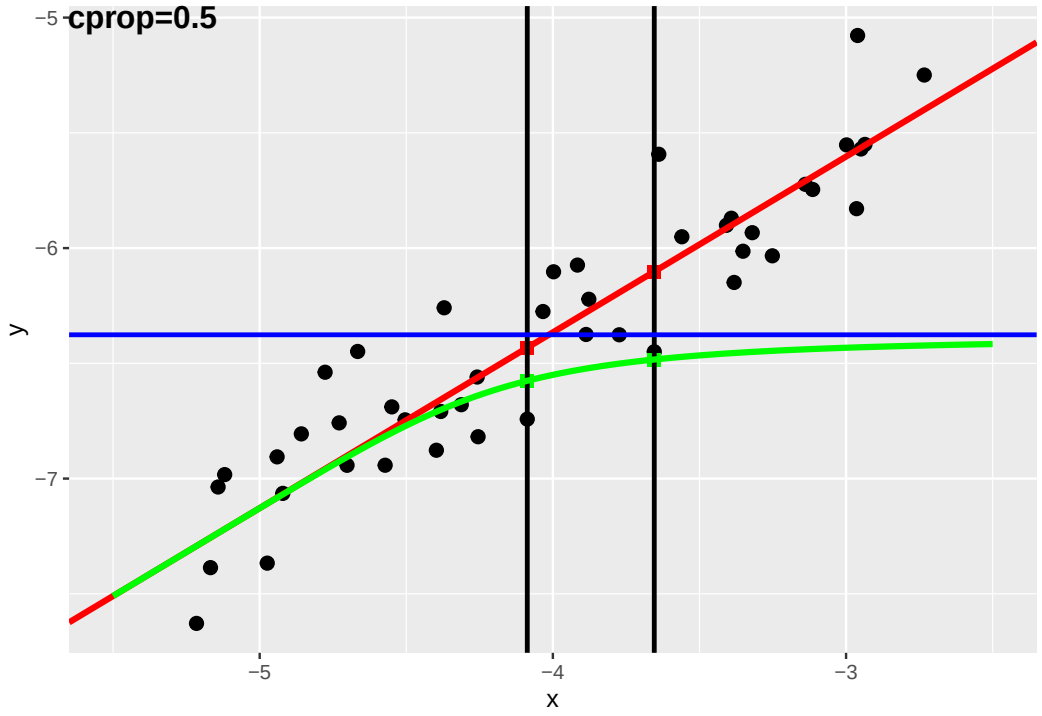
We see that the green line stays below $y = \text{LOQ}$ for all x, which means that $\hat{y}_i < \text{LOQ}$ for every imputation from $y = E(Y|X = x, Y < \text{LOQ})$.

However this is not true for the red line; in fact, we see $\hat{y}_{I2} >$ LOQ imputed from $y = E(Y|X = x, Y <$ LOQ), which contradicts the fact that $y_{I2} <$ LOQ was observed.

We have thus illustrated why it is necessary to impute to condition on both $X = x$ and $Y <$ LOQ) by using the truncated normal distribution as specified in equation XX above.

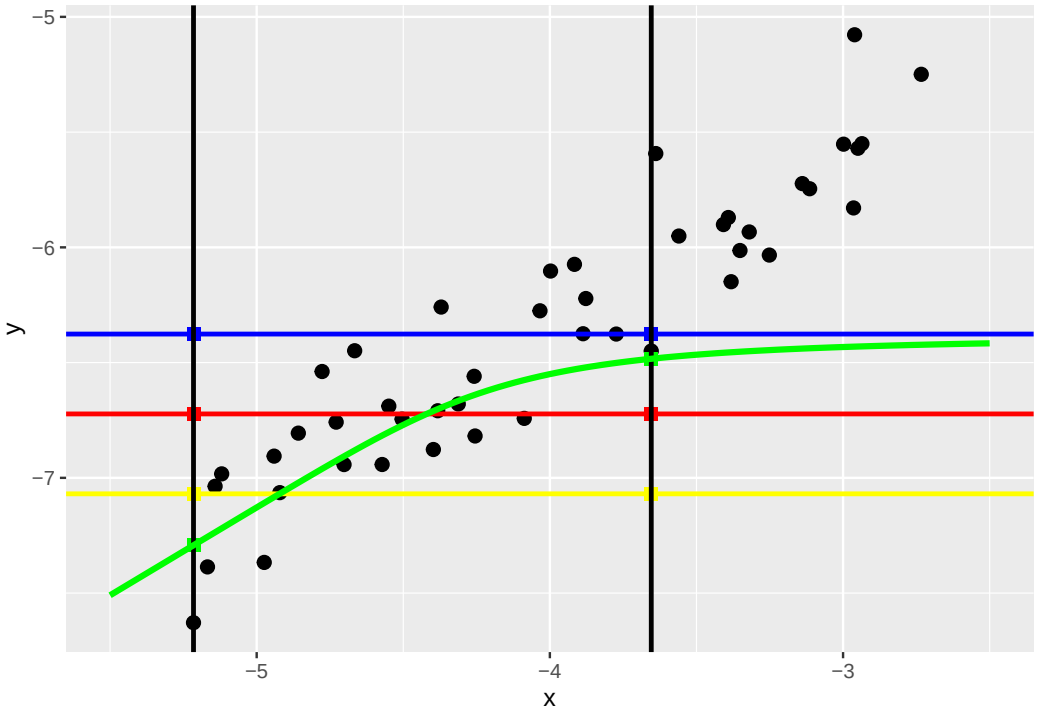
We will later verify this conjecture quantitatively through simulation studies.

We will use `censReg1naive` to denote such naive imputations from $y = E(Y|X = x)$ from now onwards.



2.3.2 Comparison of imputation by censored regression with fabrication by substitution

The following plot shows green points imputed by censored regression, blue points created from substitution by LOQ, red points created from substitution by $\frac{\text{LOQ}}{\sqrt{2}}$, and yellow points created from substitution by $\frac{\text{LOQ}}{2}$; the original parameter values, 0.3 and 0.2 were used for *cprop* and σ respectively.

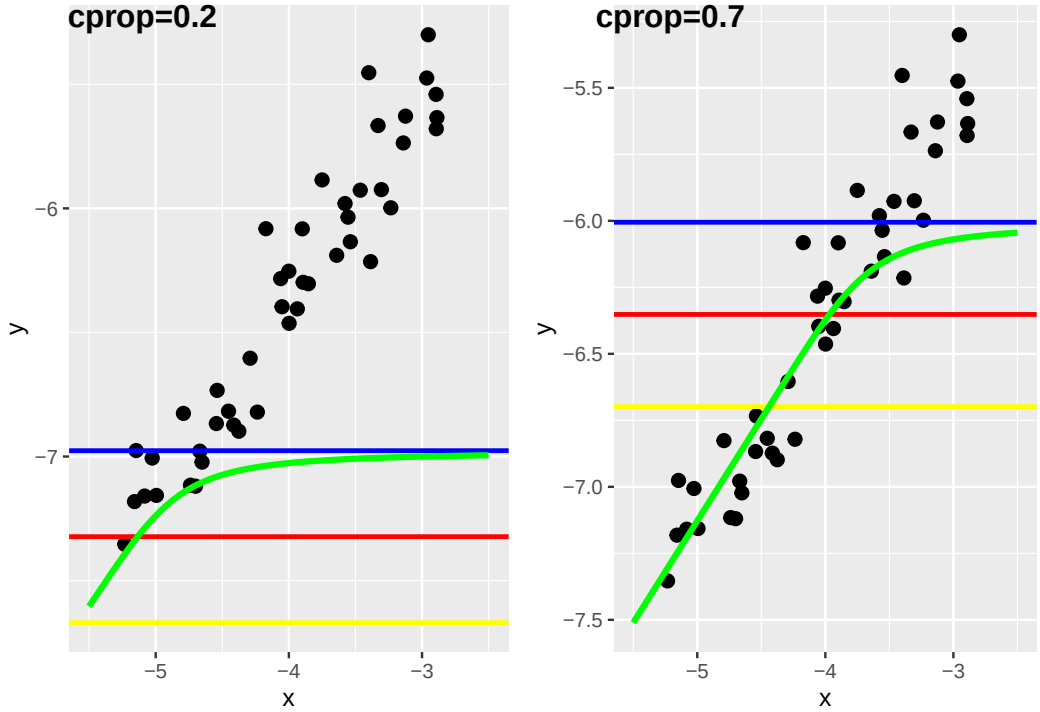


Let's repeat the previous visualisation, except using the different censoring proportions 0.2 and 0.7 respectively.

We see that the imputation by censored regression (represented by the green curve) appears to be more robust than substitution to changes in the censoring proportion.

In particular, for this dataset, we see that the censored data points lie closest on average to LOQ for $cprop = 0.2$ whereas they are closest on average to $\frac{\text{LOQ}}{2}$ for $cprop = 0.7$; moreover, we saw from the previous plot that they are closest on average to $\frac{\text{LOQ}}{\sqrt{2}}$ for $cprop = 0.5$.

This observation agrees with our expectation (as stated in the previous chapter) that fabrication by substitution has lower robustness than imputation by censored regression.



We are now familiar with the key concepts, and have visualised fabrication by substitution and imputation by censored regression in this general setting.

So we are ready to begin our quantitative investigation using simulated datasets, which we will design based on our EDA of the SNMPC datasets.

2.3.3 Illustration of bias produced by omission of censored data

Omission of censored observations is illustrated in Figure 2.3b, in which the LOQ is again shown as a blue horizontal line.

The regression lines from the uncensored dataset, and the corresponding incomplete dataset, are shown in green and red respectively.

Omission is well-known to produce bias (Helsel 2012).

This plot illustrates this since the slope of the yellow line is smaller.

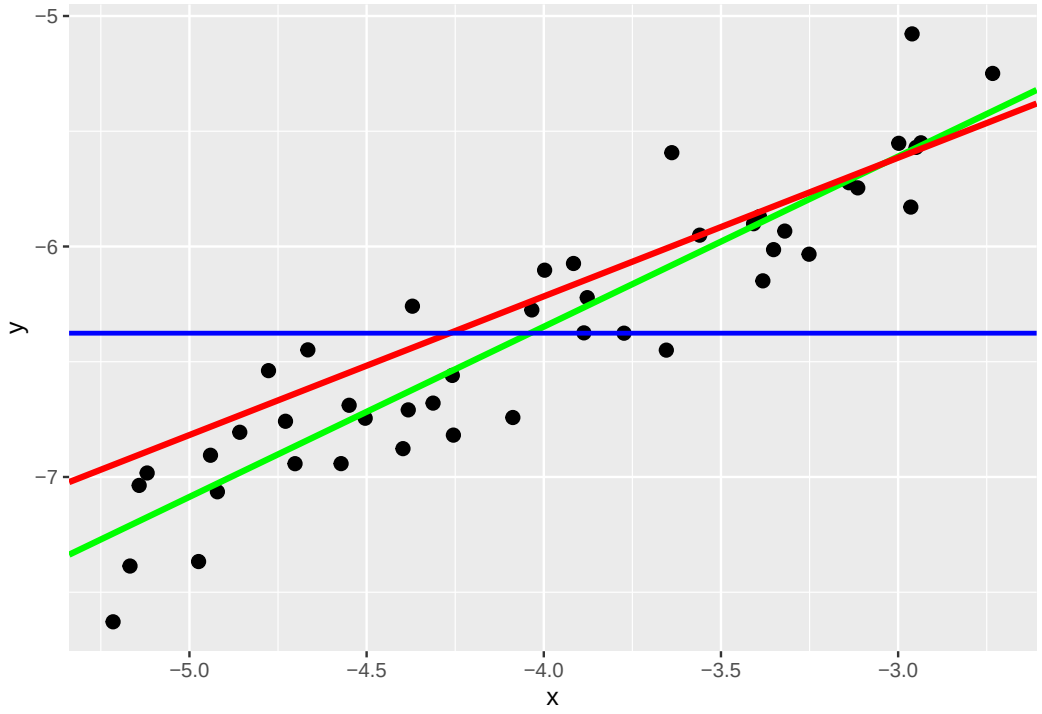
This is because whether or not an observation is omitted is based solely on its y value, so the range of y values is more likely to decrease more than the range of x values.

Alternatively, we could say that every censored y value is more likely than not

to lie below the true regression line because the smallest y values are precisely the ones that are censored.

This results in the regression line being biased towards larger y when x is smaller, which is what we see from the figure below.

The yellow line in Figure 2.3b shows the censored regression of y on a from the censored dataset using the model



3 Simulation study

We now present an eight-step overview of the design and implementation of our simulation study, and our plans for the subsequent analysis.

We will give a more detailed description of steps 1-5 in the subsequent five sections.

1. Selection of a set of parameter values for the set of variable parameters $\{\beta_A, \sigma, cprop\}$, and generation of the corresponding "uncensored dataset", "censored dataset", and "incomplete dataset".

2. Estimation of β, and prediction of $E(Y|A = a)$ for $a \in \{0, 1, 2, ..., 14\}$ from the uncensored dataset by simple linear regression.

Calculation of the MSE, squared-bias, and variance, of these estimates and predictions.

Tabular presentation of these results for the estimates and graphical presentation of these results for the predictions.

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

We will use two families of methods, in which the replacement is done by "direct substitution" and "imputation by censored regression" respectively.

The completed datasets created by different methods will be distinct.

Step 2 is then done for every completed dataset.

Step 2 is also done for the incomplete dataset for the purpose of comparison.

4. Do as in step 2, except from the censored (instead of the uncensored) dataset using censored (instead of simple linear) regression.

We call this method `censReg0` because it comprises censored regression without imputation.

5. Repetition of steps 1-4 for various selections of parameter values.

6. Discussion of all results.

3.1 Design and implementation for generation of our uncensored, censored, and incomplete datasets

3.1.1 Generation of uncensored datasets

Guided by the findings of the EDA we presented in Section XX, we generate our uncensored datasets as follows (at every iteration):

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.
3. We generate all x_i from

$$x_{a_i} = -2.91 - \beta_A a_i + e_{a_i}$$

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

Every resulting uncensored dataset has N observations for X and Y for every year, so in total there are $15N$ observations, where each observation is for the three variables Y, X, and A.

3.1.2 Generation of censored datasets

We generate every censored dataset from the corresponding uncensored dataset by using y_i^* instead of y_i, where y_i^* refers to the ith observation prior to it being observed.

We determine LOQ by censoring a fixed proportion, which we denote as $cprop$, of all observed y_i values.

This means every censored dataset that we generate has $15N \times cprop$ censored y_i, and $15N \times (1 - cprop)$ uncensored y_i observations.

Concretely, $LOQ = (cprop \times 100)$th percentile of all y_i values, which means that the value of $LOQ|cprop$ is constant and thus independent of A.

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

3.1.3 Generation of incomplete datasets

We then generate the incomplete dataset by removing all y_i such that $y_i = LOQ$ from the corresponding uncensored dataset.

3.1.4 Selection of sample size and number of iterations

Our results from preliminary experimentation indicated that the percentage error of estimates of β was inversely proportional to the square root of the number of iterations used to generate the datasets.

This percentage error was approximately 2% and 0.7% for 1000 and 10000 iterations, respectively.

We will therefore use 1000 iterations for simulation runs where approximate results suffice, and 10000 iterations for all our other runs.

Our estimates from our test dataset (which is a subset of the SNMPC dataset) for our variable parameters are $\{\beta_A = -0.02, \sigma = 0.1, cprop = 0.34, \text{sample size} \approx 100\}$.

We will select related parameter values for our simulations, so that our results have real real-world relevance.

We will use the parameter values $\{cprop = 0.3, \beta_A = -0.02, \sigma = 0.3\}$ for our first four simulation runs, for which the value of `sample size` will equal 50, 25, 12, and 6, respectively.

We will use these results to inform our selection of a value for `sample size`; we will use this fixed value for all our subsequent runs.

3.1.5 Selection of values for the variable parameters

We will restrict our selection of values for β_A to negative values $\beta_A < 0$, which means that X follows a decreasing trend over time (i.e as A increases).

This means that Y also decreases with time since we use the fixed proportionality constant 0.79 (i.e $\beta = 0.79\beta_A$) for generation all of our uncensored datasets.

The purpose of this restriction is convenience and clarity.

The value of σ determines the strength of correlation for the regression Y on X.

The lower the value of σ, the stronger is this correlation.

We conjecture the estimates and predictions from methods using censored regression methods will be decreasingly good for increasing values of σ.

Moreover as σ increases, the mean of the true values of the censored data decreases.

This is shown in Figure 1.

(Add a figure showing a thin and a fat normal distribution each truncated at LOQ)

For $cprop = 0$, a normal distribution is a perfect fit for the y_i^*.

As *cprop* increases, this fit is decreasingly good, whereas a truncated normal distribution (truncated at LOQ) is a perfect fit.

We therefore conjecture that censored regression followed by imputation using the expectation of the corresponding truncated normal, will give results that are increasingly good relative to those from other methods, as *cprop* increases.

3.2 Estimation of β and prediction of $E(Y|A = a)$ for $a \in \{0, 1, 2, ..., 14\}$ from the uncensored 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

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

into

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

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

We obtain the parameter estimates $\{\widehat{\alpha}, \widehat{\beta}\}$ by fitting a linear model using the `lm()` method in R.

We obtain the prediction of $E(Y|A = a)$ for $a \in \{0, 1, 2, ..., 14\}$ from

$$E(Y|A = a) = \widehat{\alpha} + \widehat{\beta} a$$

We calculate and present the MSE, squared-bias, and variance of our estimates and annual predictions.

3.3 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 $[0, LOQ]$.

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 LOQ.

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

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

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

Our three substitution methods will use substitution by either LOQ, $\frac{LOQ}{\sqrt{2}}$ or $\frac{LOQ}{2}$; we name them `subst1`, `subst2`, and `subst4`, respectively.

Our notation is based on the fact that $LOQ = \frac{LOQ}{1}$, and that $2 = \sqrt{4}$ and $1 = \sqrt{1}$, respectively.

Our rationale for choosing these three methods is that since $\frac{LOQ}{2} < \frac{LOQ}{\sqrt{2}} < LOQ$ 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.

3.3.1 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`.

3.3.1.1 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 LOQ, whereas the true value is known to be not higher than LOQ.

`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 LOQ, despite the fact that the censored value are below LOQ.

3.3.1.2 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 $\widehat{\alpha}_X$, $\widehat{\beta}_X$, $\widehat{\beta}_A$ 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 $y = LOQ$.

3.4 Estimation of β directly from the censored dataset by censored regression by the `censReg0` 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 `censReg0` method uses A_i as the predictor instead for this step.

This method is designed to test our conjecture that since $|\beta_X| = 0.79$ is much greater than $|\beta_A|$, `censReg0` will result in estimates with higher variance than from `censReg1`.

This method models each observation y_i^* as from a normal distribution with mean

$$\mu_{i_A} = \alpha_A + \beta_A a_i$$

and variance σ^2.

The same mathematical steps as in Section XXX yield the corresponding log-likelihood function

$$\log(L) = \sum_{i=1}^{N} \Bigg[(1 - I_i)[\log(\phi((y_i - (\alpha_A + \beta_A a_i))/\sigma)) - \log(\sigma)]$$

$$+ I_i \times \log[\Phi((LOQ - (\alpha_A + \beta_A a_i))/\sigma)] \Bigg]$$

Thus the parameter estimates $\{\widehat{\alpha}_A, \widehat{\beta}_A\}$ are found directly from the maximisation of the corresponding log-likelihood function, without any imputation step.

We obtain the prediction of $E(Y|A = a)$ for $a \in \{0, 1, 2, ..., 14\}$ from

$$E(Y|A = a) = \widehat{\alpha} + \widehat{\beta} a$$

We calculate and present the MSE, squared-bias, and variance of the estimates of $\widehat{\beta}$ and the annual predictions $E(Y|A = a)$ for $a \in \{0, 1, 2, ..., 14\}$.

3.5 Summary

We will generate an uncensored dataset, a censored dataset, and an incomplete dataset.

Finding estimates and predictions from the uncensored dataset will serve as the benchmark.

We call this method `best`; the corresponding method from the incomplete dataset is called `omit`.

All of our other methods will be applied to the censored dataset:

The methods in which censored y_i are replaced by LOQ, $\frac{LOQ}{\sqrt{2}}$ or $\frac{LOQ}{2}$ are `subst1`, `subst2`, and `subst4`, respectively.

The methods in which censored y_i are imputed by the censored regressions Y on X, or Y on both X and A are called `censReg1` and `censReg2`, respectively.

The variant of `censReg1` which imputes using the expectation of a non-truncated instead of a truncated normal distribution is called `censReg1naive`.

The method in which censored y_i are estimated directly (without imputation) from the censored regression Y on A is called `censReg0`.

In the following two chapters we present the results from using these methods to estimate β, and to predict the annual means $E(Y|A = a)$.

4 Results for the determination of an appropriate sample size and selection of data manipulation methods for further study

4.1 Terminology for results from our data manipulation methods

At every iteration, we generate an uncensored dataset, a censored dataset, and an incomplete dataset.

Finding estimates and predictions from the uncensored dataset will serve as the benchmark.

We call this method `best`; the corresponding method from the incomplete dataset is called `omit`.

All of our other methods will be applied to the censored dataset:

The methods in which censored y_i are replaced by LOQ, $\frac{LOQ}{\sqrt{2}}$ or $\frac{LOQ}{2}$ are `subst1`, `subst2`, and `subst4`, respectively.

The methods in which censored y_i are imputed by the censored regressions Y on X, or Y on both X and A are called `censReg1` and `censReg2`, respectively.

The variant of `censReg1` which imputes using the expectation of a non-truncated instead of a truncated normal distribution is called `censReg1naive`.

The method in which censored y_i are estimated directly (without imputation) from the censored regression Y on A is called `censReg0`.

We will now present our results from using these methods to estimate β, and to predict the annual means $E(Y|A = a)$.

4.2 Determination of appropriate sample size

We got the parameter values $\{cprop = 0.3, \beta_A = -0.02, \sigma = 0.1\}$ based on our estimates from our exploratory data analysis on our test dataset.

We will first obtain results from datasets with different sample sizes in order to decide an appropriate sample size for all our subsequent work.

Our real dataset has approximately 100 observations per year for Y and X from herring in years 2003-2017.

However these observations are from various locations and have differences for various other variables such as age, fat-percentage etc., which means that any statistical analysis which controls for such variables would have a smaller sample size.

We will test sample sizes that differ by a factor of 2: we do this by generating datasets by simulation using 10000 iterations, with sample sizes 50, 25, 12 and 6 respectively.

Since our estimate $\sigma = 0.1$ is from the test dataset for which $mean(samplesize) \approx 100$ and we wish to simulate smaller samples sizes, we choose a higher value for σ whilst leaving the other parameter values unchanged.

This means that we will perform four simulations, all of which run for 10000 iterations and use parameters $\{cprop = 0.3, \beta_A = -0.02, \sigma = 0.3\}$, whilst the value of `sample size` equals 50, 25, 12, and 6 for simulations 1-4, respectively.

4.2.1 Results from estimation of β for various sample sizes

The MSE, squared-bias, and variance of the estimates of β from all three substitution methods `subst1, subst2, subst4`, all three imputation by censored regression methods `censReg1, censReg1naive, censReg2`, the `censReg0` method for these four simulations are shown in the four tables below.

We also show in each table the result from our `best` method which we use as our benchmark.

We display all values as 10^7 times bigger than the actual values (to make them easier to read and compare).

The bias from simulations with sample sizes 50, 25, 12, and 6 are shown in the columns of the following table.

	ss50Bias	ss25Bias	ss12Bias	ss6Bias
omit	626.33	630.80	625.92	628.98
subst2	16.36	18.13	19.69	15.94
subst1	224.61	221.39	217.86	227.11
censReg1	0.01	0.01	0.03	0.09
censReg2	0.02	0.01	0.02	0.12
censReg0impute	0.02	0.01	0.03	0.13
best	0.12	0.12	0.03	0.16
subst4	532.51	547.36	558.59	531.49

	ss50Bias	ss25Bias	ss12Bias	ss6Bias
censReg1naive	69.48	66.68	67.15	70.29

The following table below is the same as the previous one, except that it shows the variance of the estimates of $\widehat{\beta}$

	ss50Var	ss25Var	ss12Var	ss6Var
omit	47.2	101.8	218.9	470.2
subst2	85.8	174.8	371.4	745.5
subst1	38.5	78.7	167.7	336.3
censReg1	67.6	137.6	292.7	585.5
censReg2	73.3	149.9	317.7	636.0
censReg0impute	73.4	149.9	317.9	636.7
best	71.8	136.1	286.6	565.4
subst4	166.4	339.0	718.7	1442.3
censReg1naive	44.0	88.3	188.3	375.0

4.2.2 Our rationale for choosing $samplesize = 12$

Allowing for random error from using only 10000 iterations, we can conclude that the squared-bias is independent of sample size, whereas the variance is inversely proportional to sample size.

Moreover since the bias_variance decomposition

$$MSE = Bias^2 + Variance$$

always holds, we need not look at the MSE values for the purpose of choosing sample size.

We find in additional experiments (details not shown) that the standard error of the estimates is inversely proportional to the square root of the number of simulation iterations, so we have three factors to balance:

1. We want our results to be potentially applicable for real data.
2. We want sample size to be sufficiently large to avoid MSE being dominated by variance alone.
3. We want the number of iterations to be sufficiently large that our estimates have sufficiently low standard error.

We therefore decide to use $samplesize = 12$ for all of our subsequent experiments.

4.3 Selection of censoring methods for further study

We will now use simulations with just 1000 iterations for all eight methods (and also for our reference method `best`) to estimate β for four sets of parameter values.

We will hold $\beta_A = -0.02$ fixed.

We will use "low" and a "high" value for each of *cprop* and σ.

Concretely, $\{(0.1, 0.1), (0.7, 0.1), (0.1, 0.5), (0.7, 0.5)\}$ will be used for $\{(cprop, \sigma)\}$ respectively.

4.3.1 Variance of estimates from all methods

The following table shows the variance of estimates from each method for our low-low, high-low, low-high, and high-high combinations of values for *cprop* and σ, respectively.

	Low-Low	High-Low	Low-High	High-High
omit	44.05	67.79	584.44	870.22
subst2	80.93	87.31	697.06	322.03
subst1	42.78	9.31	564.85	120.60
censReg1	48.34	74.72	661.81	766.91
censReg2	48.76	96.82	673.66	1134.92
censReg0	48.66	102.82	673.66	1135.45
best	50.03	50.03	775.21	775.21
subst4	157.25	270.45	886.09	666.12
censReg1naive	43.31	78.67	518.26	845.33

4.3.2 Bias of estimates from all methods

The following table shows the bias of estimates from each method for our low-low, high-low, low-high, and high-high combinations of values for *cprop* and σ, respectively.

	Low-Low	High-Low	Low-High	High-High
omit	150.57	1345.67	232.84	1248.68
subst2	247.54	62.60	1.11	515.59
subst1	19.49	1252.76	44.65	1198.68
censReg1	0.00	0.04	3.80	0.45
censReg2	0.00	0.13	3.93	0.39
censReg0	0.00	0.17	3.98	0.43
best	0.00	0.00	1.68	1.68
subst4	1287.53	2623.28	20.90	116.45
censReg1naive	31.29	25.19	137.95	90.67

4.3.3 MSE of estimates and predictions from all methods

The following table shows the MSE of estimates from each method for our low-low, high-low, low-high, and high-high combinations of values for *cprop* and σ, respectively.

	Low-Low	High-Low	Low-High	High-High
omit	194.57	1413.40	816.69	2118.03
subst2	328.39	149.82	697.48	837.30
subst1	62.23	1262.07	608.94	1319.16
censReg1	48.29	74.68	664.95	766.59
censReg2	48.71	96.85	676.92	1134.17
censReg0	48.61	102.89	676.97	1134.74
best	49.98	49.98	776.12	776.12
subst4	1444.62	2893.46	906.10	781.91
censReg1naive	74.55	103.78	655.70	935.16

4.3.4 Our rationale for selecting `subst1`, `subst2`, `subst2`, `censReg1`, `censReg2`, and `censReg0` for further study

We see that there is a much bigger difference between different methods in the amount of bias than in the amount of variance.

We will therefore focus primarily on the results for bias; we will use terms such as high and low to compare the relative amount of bias from our different methods.

We see that the amount of bias from `omit` is high for $\{(0.1, 0.1), (0.1, 0.5)\}$, and is very high for $\{(0.7, 0.1), (0.7, 0.5)\}$.

It makes sense that there is higher bias with higher proportion of censored values since a higher proportion of the data has been omitted.

Moreover, these generally high values are commensurate with our prior expectation that `omit` is a poor method, so we will not study this method further.

The amount of bias is very high from: `subst1` for $\{(0.7, 0.1), (0.7, 0.5)\}$; `subst2` for $\{(0.7, 0.5)\}$; `subst4` for $\{(0.1, 0.1), (0.7, 0.1)\}$.

However, all three substitution methods also have low bias for at least one set of parameter values.

This is intriguing and merits further investigation.

For all four parameter value sets, the bias from `censReg1`, `censReg2`, and `censReg0` is very low; moreover it is clearly higher from `censReg1naive`, which verifies our conjecture from the previous chapter.

Consequently we will exclude `censReg1naive` from further investigation, since we have now verified our conjecture.

In summary, we have rejected the two methods `omit, censReg1naive`, so we will limit our attention to six methods for all our subsequent work: the three substitution methods `subst1`, `subst2`, and `subst2`, and the three censReg methods `censReg1`, `censReg2`, and `censReg0`.

We will use `best` as our reference method throughout.

5 Evaluation of methods for various values of β_A

5.1 Estimation of β and predictions of annual $mean(Y)$ for every year

We will now focus our six chosen methods `subst1`, `subst2`, `subst4`, `censReg1`, `censReg2`, and `censReg0`.

We will use these methods to estimate β and predict $E(Y|A = a)$ for $a \in \{0, 1, 2, ..., 14\}$.

We will do four simulations that each run for 10000 iterations, using the same parameter values $\{cprop = 0.3, \sigma = 0.3\}$ as we used in the previous section.

We will use the four parameter values $\{-0.02, -0.04, -0.08, -0.16\}$ for β_A in these four simulations, respectively.

We will again use `best` as our benchmark.

For the estimates and predictions from our selected methods, we will report the variance, squared-bias (which we refer to as "bias" throughout for conciseness), and MSE in the following three sections, respectively.

These results for estimates will be presented as tables; the results for predictions will be shown as graphs with MSE, squared-bias, or variance on the y-axis and year on the x-axis for the simulated 15-year period.

5.1.1 Variance of estimates and predictions

The following table shows the variance of estimates of β from each method for β_A equal to $-0.02, -0.04, -0.08, -0.16$, respectively.

	-0.02	-0.04	-0.08	-0.16
subst1	166.67	178.33	224.97	305.91
subst2	367.32	351.30	316.72	323.13
subst4	709.86	643.79	470.55	355.19
censReg1	288.91	315.86	362.78	449.45
censReg2	314.09	325.80	364.46	449.72
censReg0	314.41	326.22	365.40	452.65
best	289.41	280.76	288.05	285.28

We will now show graphs of the variance of the predictions of annual means from our chosen censoring methods.

A common feature of all these graphs is that they typically have an approximately parabolic "U" shape, with higher variance at each end of the time period than in the middle of the period.

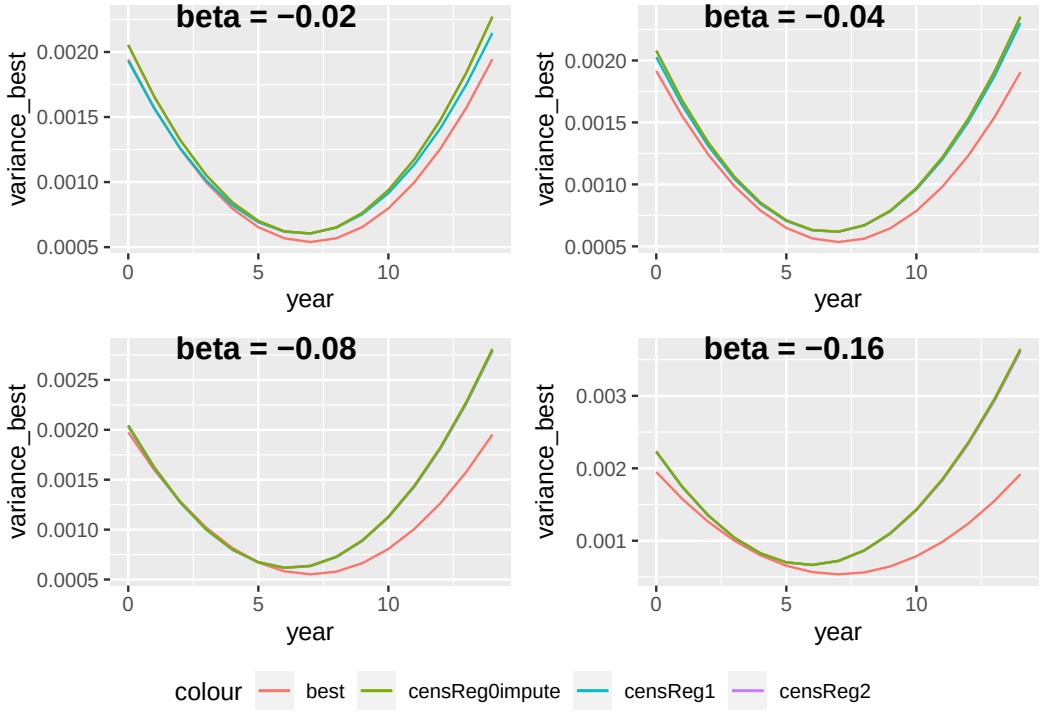
This is in accordance with our prior expectations because this is generally the case for the residual variance from fitted linear regression models.

Our first set of four graphs show the variance of `censReg1` and `censReg2` methods relative to `best` method for β_A equal to -0.02, -0.04, -0.08, -0.16, respectively.

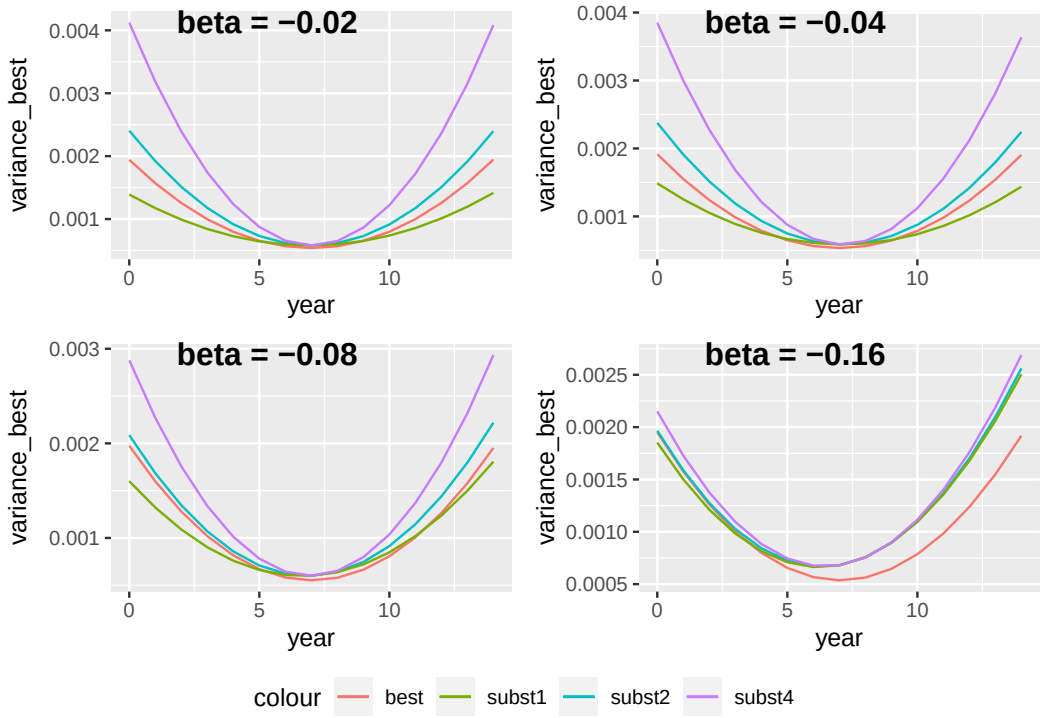
We see that the variance from `censReg1` is clearly lower than from `censReg0` for `censReg2` for the lowest value of β_A and at the beginning and end of the 15-year period.

Moreover, there is no visible difference in the variance from these three censored regression methods for higher values of β_A and/or for years in the middle of the 15-year period.

This makes sense since `censReg1` does not use A as a predictor variable, whereas these other two methods do, so we would expect the relative performance of `censReg1` to decrease as the value of $|\beta_A|$ increases.



Our second set of four graphs show the variance of `subst1`, `subst2` and `subst4` methods relative to `best` method for β_A equal to -0.02, -0.04, -0.08, -0.16, respectively.



We see that the variance is highest from `subst4` and lowest from `subst1` in general, and that the difference between these decreases as $|\beta_A|$ increases.

Moreover the variance from all three methods (relative to `best` method) increases as $|\beta_A|$ increases.

This all makes sense since we are using a constant LOQ value for the whole 15-year period for every dataset.

We are also always using the fixed value $\alpha_A = -2.91$, and a variable but always negative parameter value for β_A.

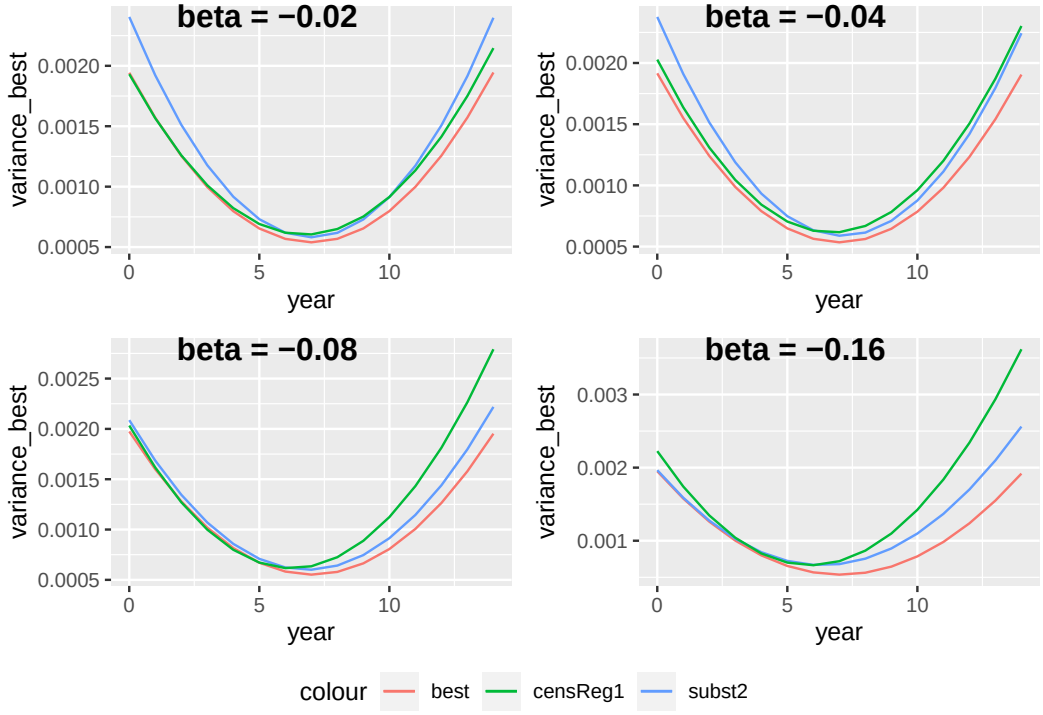
Moreover, the definition of $|\beta_A|$ tells us that the rate of decrease of $E(Y|A = a)$ as a increases is larger for larger values of $|\beta_A|$.

This all means that the proportion of y_i that are censored each year increases with year, and that this rate of increase increases as $|\beta_A|$ increases; moreover the mean of the true values of the censored data also decreases with year at an increasing rate as $|\beta_A|$ increases.

This explains why substitution by the highest value LOQ gives predictions with increasing variance as year increases, whereas the opposite is true for substitution by the lowest value $\frac{LOQ}{2}$ by the same reasoning.

However, the following set of four graphs show that the variance from substitution by $\frac{LOQ}{\sqrt{2}}$ decreases relative to that from `censReg1` as $|\beta_A|$ increases.

Moreover, we see that the strength of this trend increases as year increases.



5.1.2 Bias of estimates and predictions

The following table shows the bias of estimates of β from each method for β_A equal to $-0.02, -0.04, -0.08, -0.16$, respectively.

	-0.02	-0.04	-0.08	-0.16
subst1	226.07	860.84	3056.08	9626.60
subst2	16.03	49.12	37.98	285.91
subst4	530.96	1879.80	4570.78	4134.17
censReg1	0.03	0.01	0.02	0.20
censReg2	0.06	0.02	0.02	0.20
censReg0	0.06	0.02	0.03	0.24
best	0.01	0.01	0.06	0.17

We see that for these parameters value sets, censored regression methods give

estimates that have much lower bias, in general.

The `subst1` method is designed as a reference that gives biased estimates, since it substitutes Y values that are observed to be below LOQ with the LOQ value itself, so the substituted values will never be smaller than the real values.

For the reasons given in the previous section, we expect the bias from `subst1` to increase as $|\beta_A|$ increases, which is precisely what these results show.

In contrast, the bias from `subst4` first increases from $|\beta_A| = 0.02$ to $|\beta_A| = 0.08$ and then decreases for $|\beta_A| = 0.16$.

This suggests that the substitution value $\frac{LOQ}{2}$ is than the true values on average for $|\beta_A| = 0.02$ but not lower for the highest value $|\beta_A| = 0.16$.

This conclusion is also supported by the fact that the bias from `subst2` is much lower than that from `subst1` or `subst4`, which suggests that the real values of the censored data mostly lie between LOQ and $\frac{LOQ}{2}$ for these sets of parameter values.

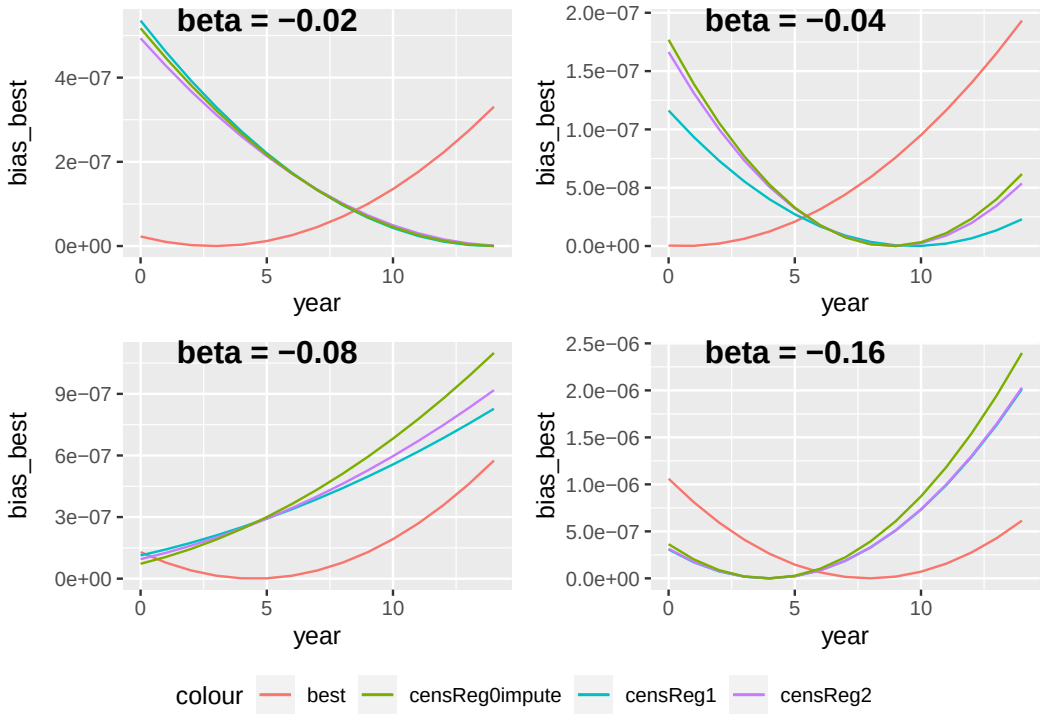
We will now show graphs of the bias of predictions of $E(Y|A = a)$ from our chosen censoring methods for β_A equal to -0.02, -0.04, -0.08, -0.16, respectively.

Our first set of four graphs show the bias from the three censored regression methods.

We see that the bias from these methods is lower than from `best` at the start of the 15-year period, whereas it is higher at the end.

This makes sense since the proportion of censored data increases as year increases.

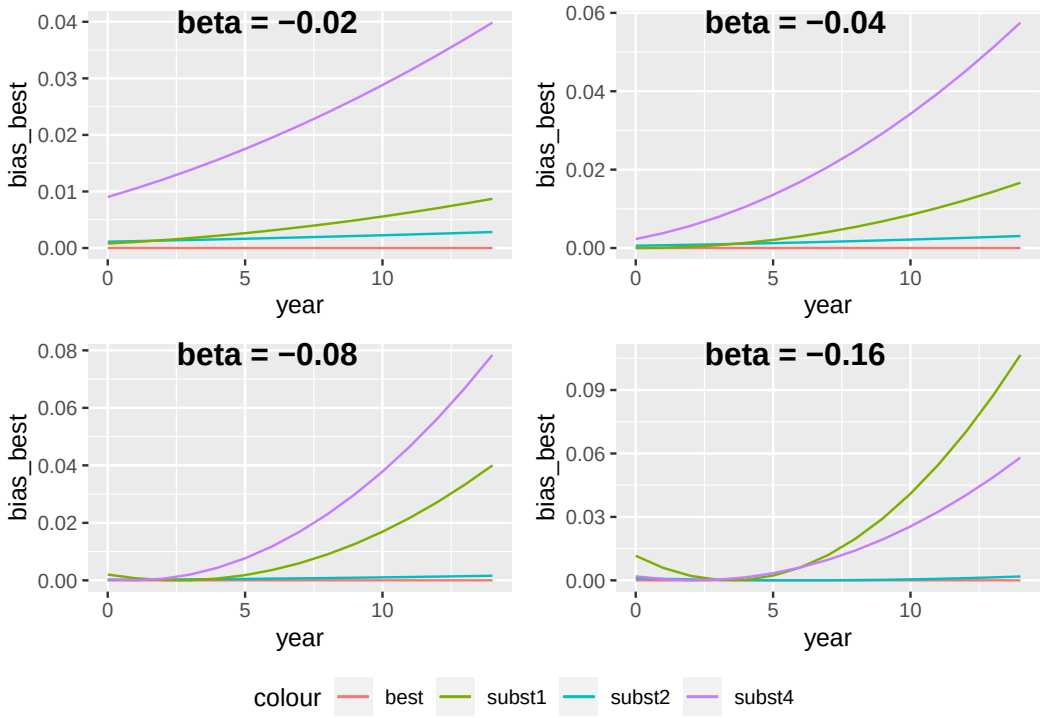
We also see that the bias from these three methods is very similar; the only notable difference is that `censReg1` has lowest bias for $\beta_A = -0.04$.



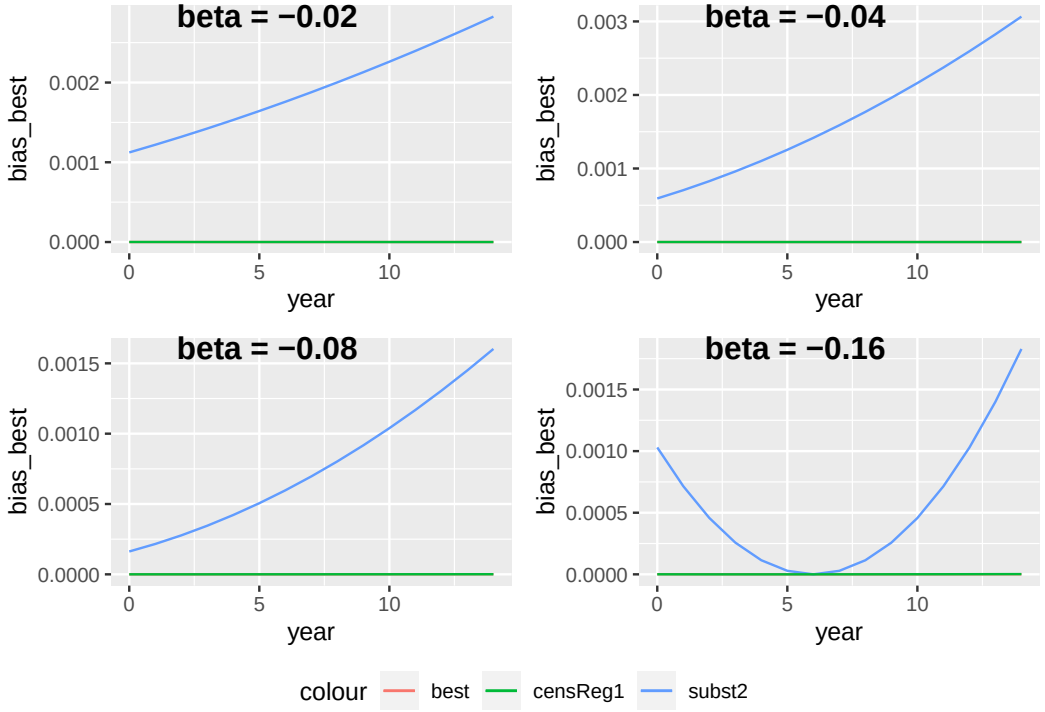
Our second set of four graphs show the bias of `subst1`, `subst2` and `subst4` methods relative to `best` method for β_A equal to -0.02, -0.04, -0.08, -0.16, respectively.

We see the same trend for the predictions that we saw for the estimates, i.e. the relative amount of bias from `subst1` increases as $|\beta_A|$ increases, whereas the opposite is true for `subst1`.

We also see a general trend for all methods that bias increases as year increases (since a higher proportion of data is censored).



Although the bias from `subst2` appears low relative to the other substitution methods it is very high relative to that from censored regression methods, which is shown by the following set of graphs.



5.1.3 MSE of estimates and predictions

The following table shows the MSE from the best performing method of each type (`censReg1` and `subst2`) for β_A equal to $-0.02, -0.04, -0.08, -0.16$, respectively.

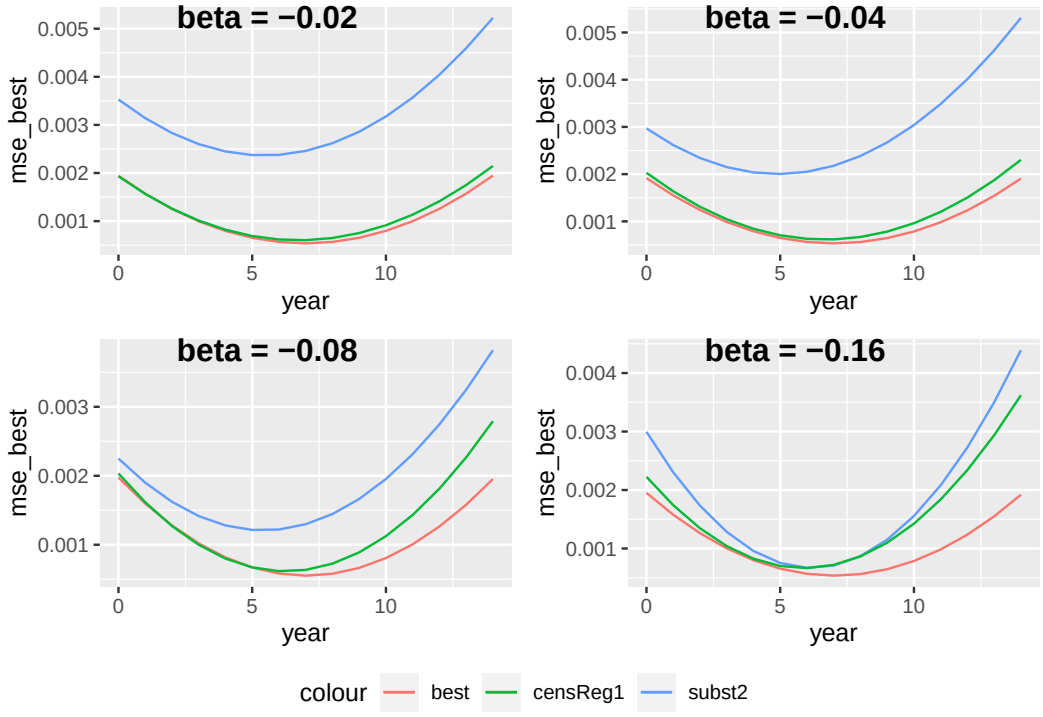
	-0.02	-0.04	-0.08	-0.16
subst2	383.31	400.38	354.67	609.01
censReg1	288.92	315.84	362.76	449.61
best	289.39	280.74	288.08	285.42

As expected the MSE of the estimates from both methods increases relative to `best` as $|\beta_A|$ increases (since the proportion of censored data increases).

The MSE from `subst2` is slightly lower than that from `censReg1` for $\beta_A = -0.08$, whereas it is much higher for the other values of β_A.

The set of graphs showing the MSE of the corresponding predictions is shown below.

We see that the MSE of the predictions from `censReg1` is generally lower than that from `subst2` and that this gap decreases as $|\beta_A|$ increases.



5.1.4 General comments

The estimates and predictions from censored regression methods have lower bias than those from substitution methods.

The MSE was also lower in the vast majority of cases, but not in every case.

Each censored regression method was also more robust to variation of parameter values, whereas the relative performance of each substitution method was more sensitive.

6 Estimation of β and predictions of $E(Y|A)$ from our selected methods for various values of σ

For all our simulations in this section, these parameters are fixed: $cprop = 0.3$, $\beta_A = -0.02$, whilst σ is given four values: 0.1, 0.3, 0.5 and 0.7 respectively.

6.1 Variance of estimates and predictions from our selected methods

The large gap between the uncensored Y data and the $\frac{LOQ}{2}$ value means that `subst4` gives higher variance than all other methods for all values of σ.

Similarly, the smallest possible gap between LOQ and the uncensored Y data explains the fact that `subst1` always gives the lowest variance.

We conjecture that the same logic would also hold for other possible substitution values; the larger the gap between this value and LOQ, the larger the resulting variance.

Again, we see that the variance from `censReg1` is approximately 10% lower than that from `censReg2` for all four values of σ.

However, the variance from censored regression methods increases faster than from substitution methods as σ increases; in fact for higher values of σ, `subst1` and `subst2` gave the lowest and second lowest MSE values, respectively.

This relative failure of censored regression methods for relatively high values of σ makes sense, here is our explanation:

A higher σ value means that the correlation between Y and X is weaker, which results in less accurate imputation by `censReg1` and `censReg2`, since the accuracy of imputation by these methods relies on the strength of correlation between Y and X.

The following table shows the variance from each method for σ equal to $0.1, 0.3, 0.5, 0.7$, respectively.

	0.1	0.3	0.5	0.7
subst1	31.50	166.67	432.69	855.64
subst2	124.16	367.32	728.31	1258.87
subst4	320.75	709.86	1172.87	1815.87

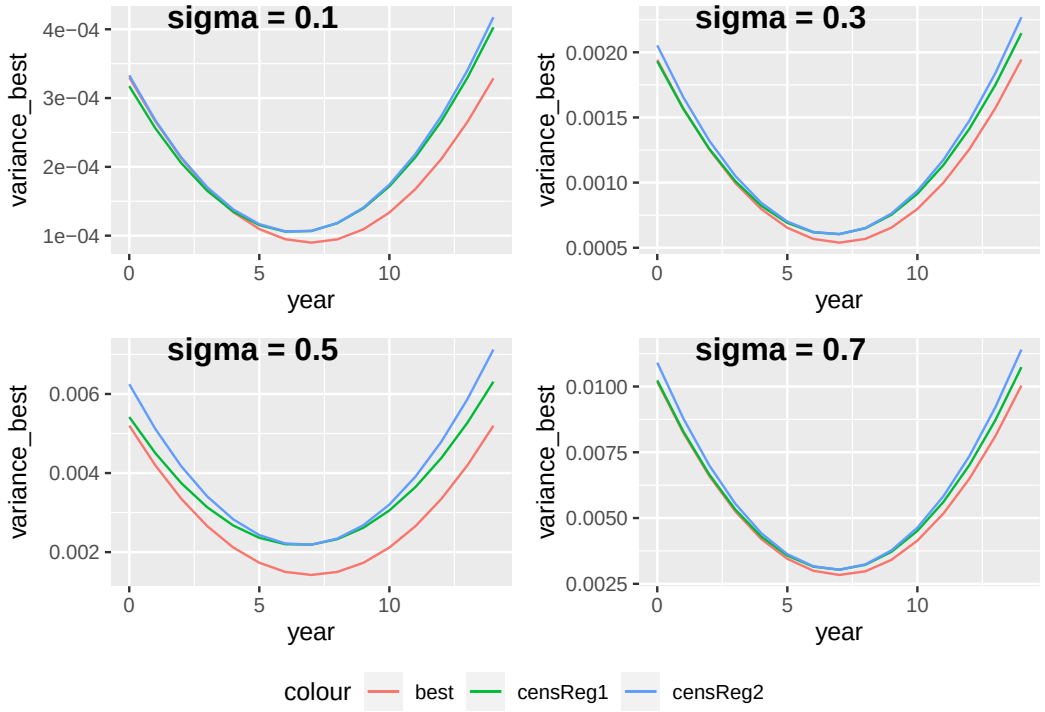
	0.1	0.3	0.5	0.7
censReg1	51.67	288.91	761.33	1520.52
censReg2	54.78	314.09	831.14	1657.35
censReg0	55.74	314.41	830.95	1657.67
best	48.87	289.41	744.55	1484.11

We begin by showing graphs of the variance of predictions of Y annual means from our chosen censoring methods.

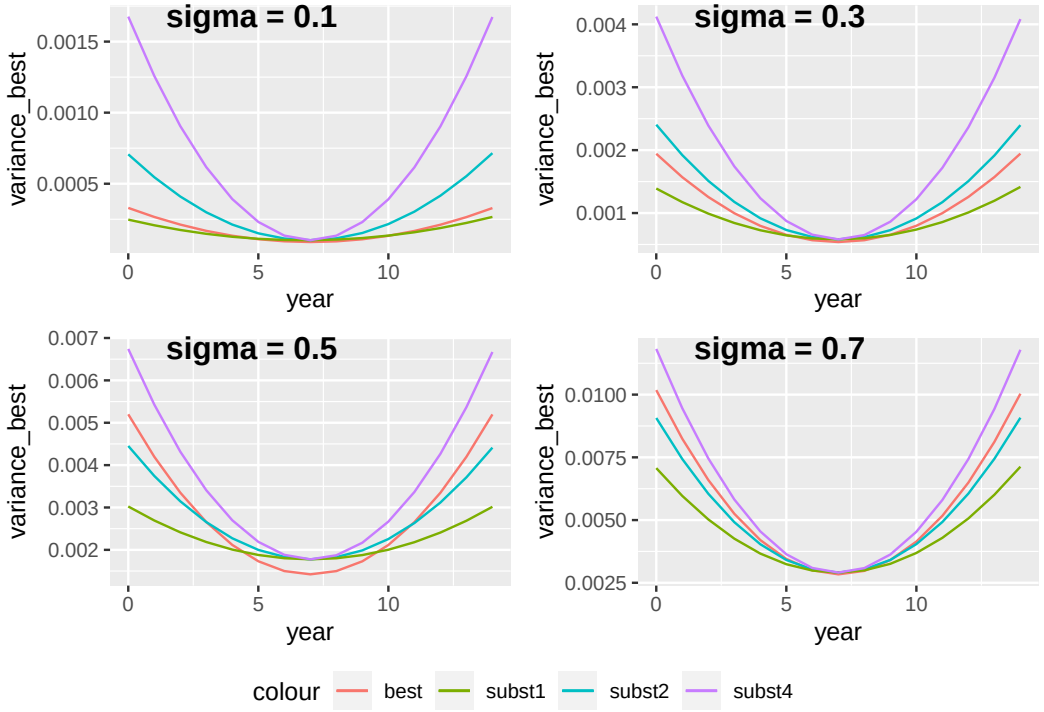
A common feature of all these graphs is that they typically have an approximately parabolic "U" shape, with higher variance at each end of the time period than in the middle of the period.

This is in accordance with our prior expectations because this is generally the case.

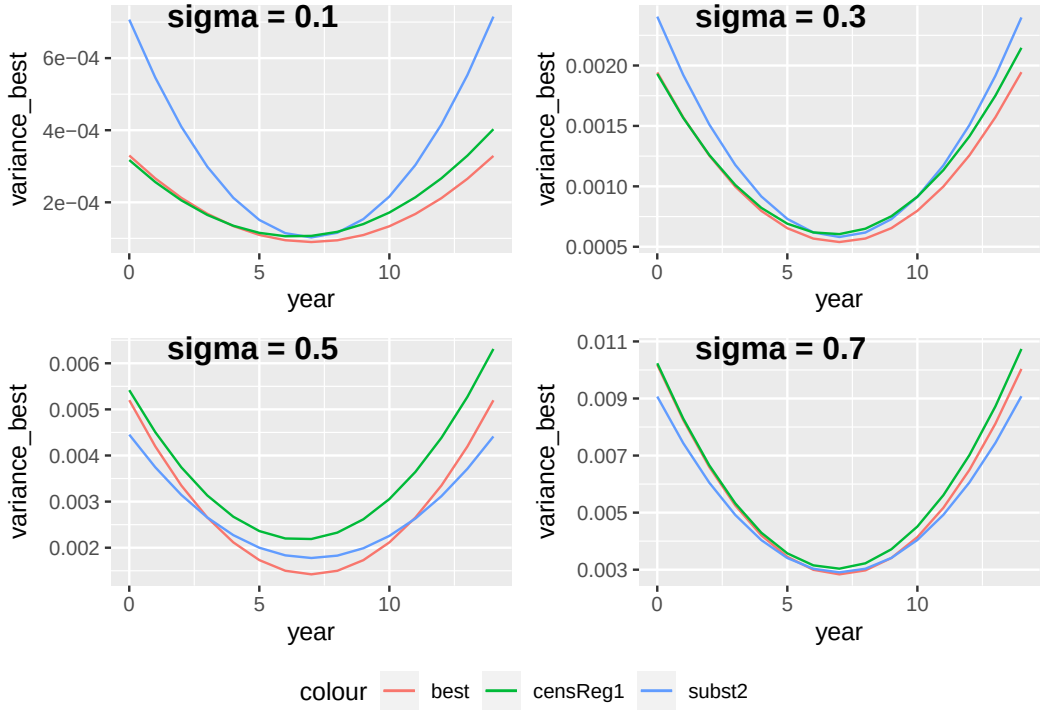
Our first set of four graphs show the variance of `censReg1` and `censReg2` methods relative to `best` method for σ equal to 0.1, 0.3, 0.5, 0.7, respectively.



Our second set of four graphs show the variance of `subst1`, `subst2` and `subst4` methods relative to `best` method for σ equal to 0.1, 0.3, 0.5, 0.7, respectively.



Our third set of four graphs simply displays the results from the `subst2`, `censReg1` and `best` methods together on the same plot, which is displayed below.



6.2 Bias of estimates and predictions from our selected methods

The following table shows the bias from each method for σ equal to $0.1, 0.3, 0.5, 0.7$, respectively.

	0.1	0.3	0.5	0.7
subst1	211.64	226.07	224.49	222.70
subst2	809.08	16.03	10.13	40.14
subst4	5103.18	530.96	74.25	5.07
censReg1	0.00	0.03	0.01	0.02
censReg2	0.01	0.06	0.01	0.01
censReg0	0.01	0.06	0.01	0.01
best	0.00	0.01	0.02	0.35

We see again that for these parameters value sets, all three of our censored regression methods give estimates that have very much lower bias, in general.

Since these three methods all give very similar results to one another and very

different results from the three substitution methods, we will again begin by interpreting the results for these two method categories separately.

The bias from `subst4` decreases greatly as the value of σ increases, whereas the bias from `subst1` is relatively independent of the value of σ.

The bias from `subst2` again follows a trend intermediate between that of `subst1` and `subst4`, since it decreases from $\sigma = 0.1$ to $\sigma = 0.5$ and then decreases for $\sigma = 0.7$.

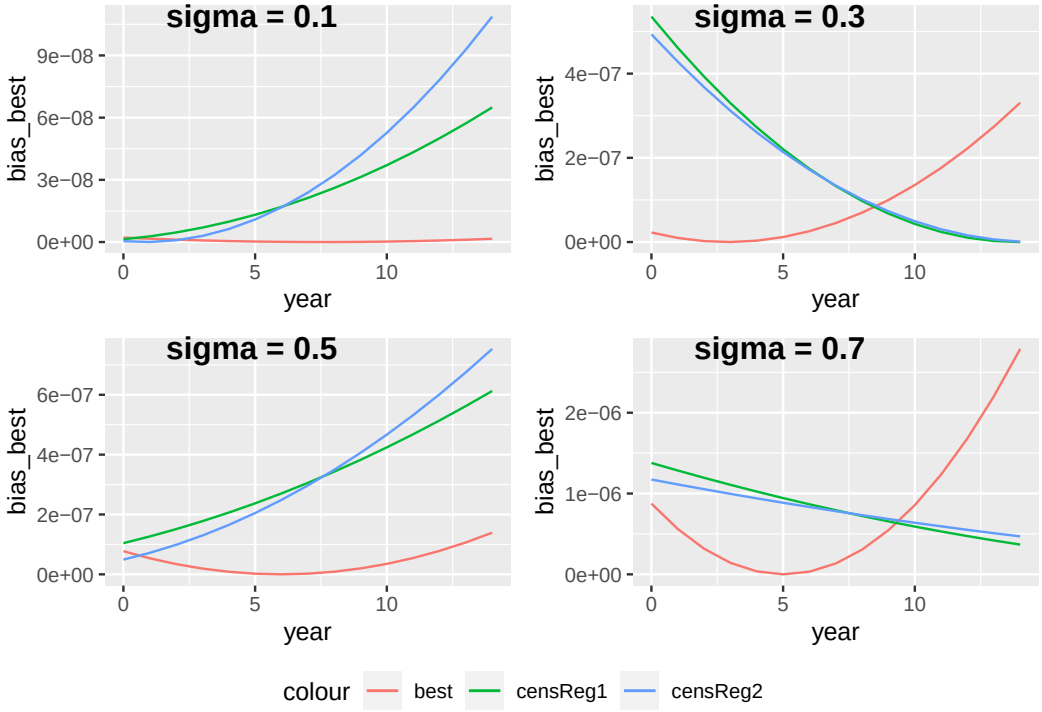
Our interpretation is that since the censored Y values lie closer on average to LOQ for smaller values of σ, and further away for larger values.

The low bias from `subst4` for $\sigma = 0.7$ indicates that the real values for the censored data lie close to $\frac{LOQ}{2}$ on average for this parameter value.

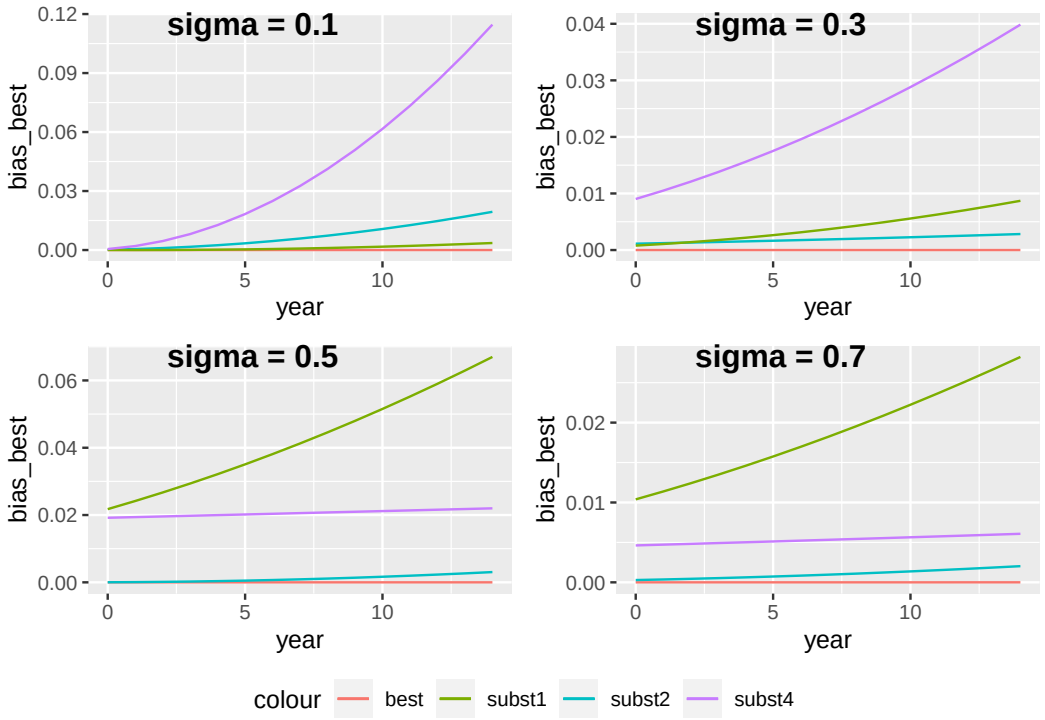
In conclusion, substitution methods give much higher bias than censored regression methods.

We will now show graphs of the bias of predictions of Y annual means from our chosen censoring methods.

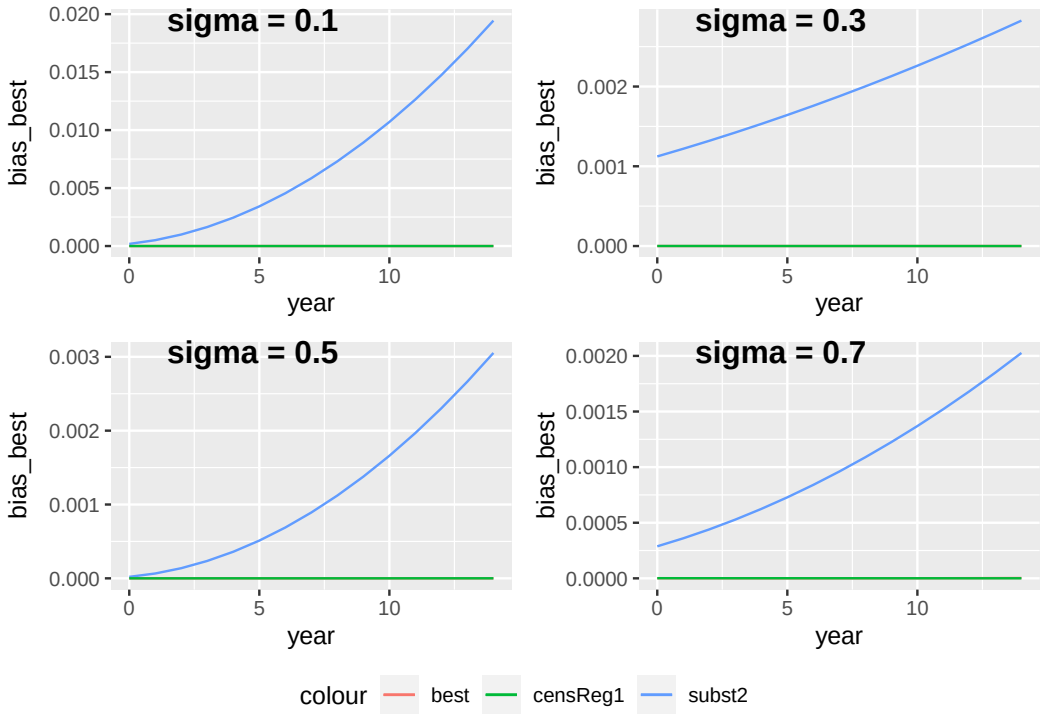
Our first set of four graphs show the bias of `censReg1` and `censReg2` methods relative to `best` method for σ equal to 0.1, 0.3, 0.5, 0.7, respectively.



Our second set of four graphs show the bias of `subst1`, `subst2` and `subst4` methods relative to `best` method for σ equal to 0.1, 0.3, 0.5, 0.7, respectively.



Our third set of four graphs simply displays the results from the `subst2`, `censReg1` and `best` methods together on the same plot, which is displayed below.



6.3 MSE of estimates and predictions from our selected methods

The following table shows the MSE of estimates from our selected methods for σ equal to σ equal to 0.1, 0.3, 0.5, 0.7, respectively.

We see that all three censored regression methods gave lower MSE than all three substitution methods for both $\sigma = 0.1$ and $\sigma = 0.3$.

Of the censored regression methods, `censReg1` gave lowest MSE for all values of σ.

Of the censored regression methods, `subst1` gave lowest MSE for all values of σ, with the sole exception that `subst2` gave slightly lower for $\sigma = 0.3$.

In fact, `subst1` even gave lower MSE than `best` for $\sigma = 0.5$ and $\sigma = 0.7$ because the lower variance more than compensated for the higher bias from this method.

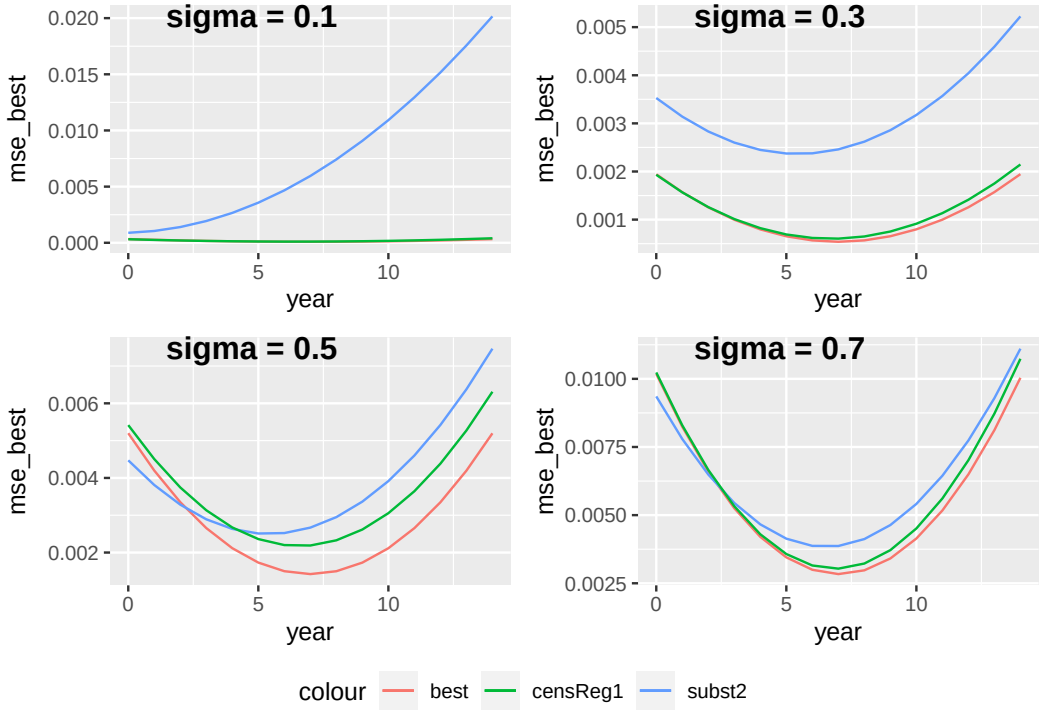
	0.1	0.3	0.5	0.7
subst1	243.14	392.72	657.13	1078.25

	0.1	0.3	0.5	0.7
subst2	933.23	383.31	738.37	1298.88
subst4	5423.90	1240.75	1247.01	1820.77
censReg1	51.66	288.92	761.26	1520.39
censReg2	54.78	314.12	831.07	1657.19
censReg0	55.75	314.44	830.88	1657.51
best	48.87	289.39	744.50	1484.31

The following graphs show the MSE of predictions of Y annual means from `censReg1` and `subst2` for σ equal to 0.1, 0.3, 0.5, 0.7, respectively.

We see that `censReg1` gives much lower MSE for $\sigma = 0.1$ and $\sigma = 0.3$, whereas these methods give overall similar MSE for larger values of σ;

Recall that `susbst2` gives relatively higher bias as year increases, this trend is also evident for MSE.



7 Estimation of β and predictions of $E(Y|A)$ from our selected methods for various values of *cprop*

From our previous results, `subst2` is generally the best performing substitution method and `censReg1` is the best censReg method, so we will focus solely on these methods in this chapter.

In the previous section, these methods gave estimates and predictions with similar MSE values for $\sigma = 0.5$, so we will fix this parameter at this value and investigate these estimation methods for four values of *cprop*: $0.1, 0.3, 0.5, 0.7$.

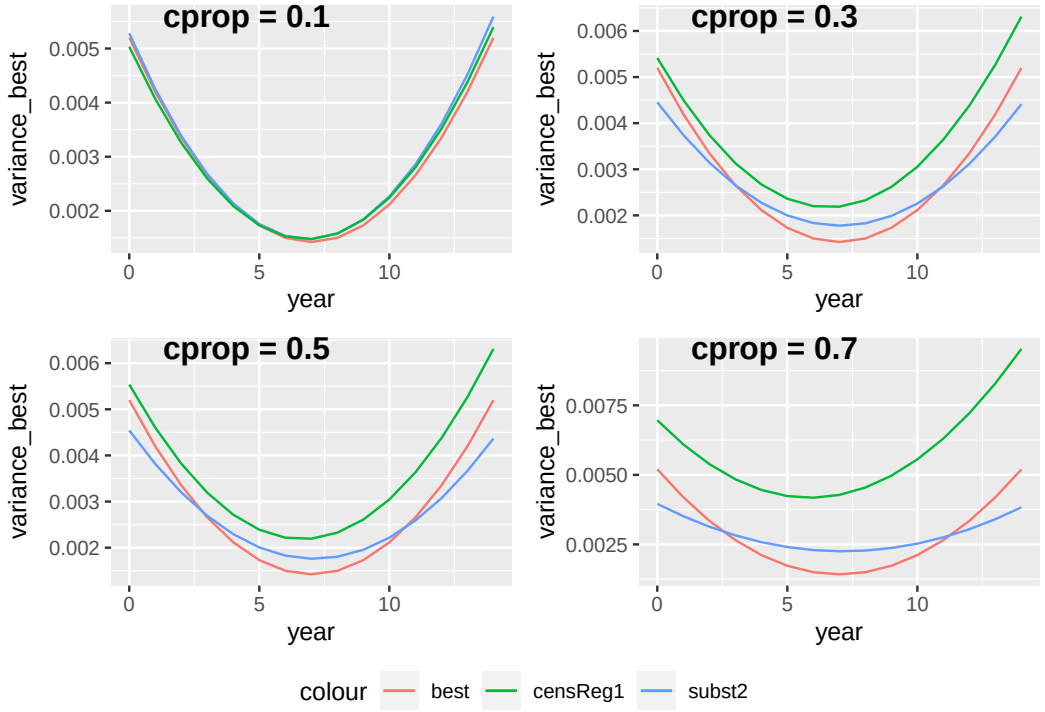
These *cprop* values correspond to censoring 10%, 30%, 50%, and 70% of the data respectively, so they correspond to decreasing values of LOQ, which is our variable of primary interest.

7.1 Variance of estimates and predictions

The following table shows the variance of the estimates from these methods for *cprop* equal to $0.1, 0.3, 0.5, 0.7$, respectively.

	0.1	0.3	0.5	0.7
subst2	808.5	683.1	723.8	395.6
censReg1	762.6	913.2	984.1	983.5
best	770.7	845.0	845.0	845.0

The following set of four graphs shows the variance of the predictions from these methods for *cprop* equal to $0.1, 0.3, 0.5, 0.7$, respectively.



7.2 Bias of estimates and predictions

We see that `censReg1` gives estimates with very low bias for all values of *cprop*, whereas the bias from `subst2` increases greatly as *cprop* increases.

We interpret this as meaning that the real Y values are unchanged when LOQ is lowered, which means that a higher proportion are likely to lie closer to LOQ for larger values of *cprop* which means that substituted values are increasingly biased towards being too small as *cprop* increases.

Since `censReg1` fits a model to all the data (censored and uncensored) it maintains low bias as the *LOQ* decreases, whilst the variance remains approximately constant.

However, as a greater proportion of values are substituted for the same constant value by the `subst2` method, the variance decreases because a higher proportion of the data values are identical.

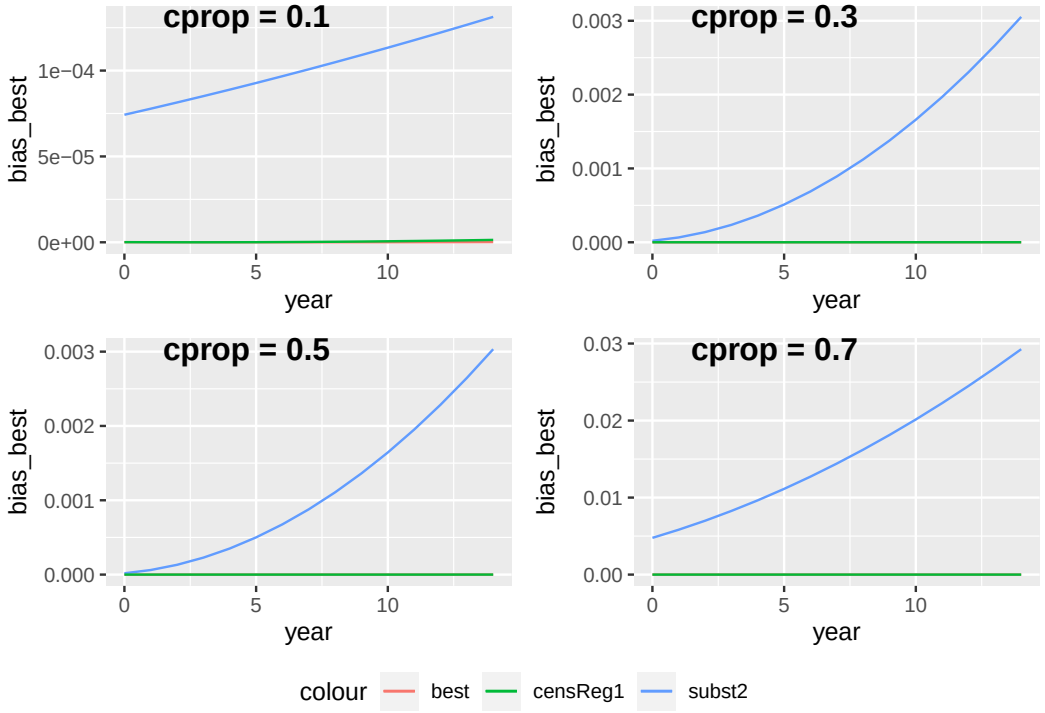
In conclusion, `censReg1` gives similar bias and variance for different values of *cprop* whereas `subst2` does not.

From `subst2` the bias increases and the variance decreases as *cprop* increases.

The following table shows the bias of the estimates from these methods for *cprop* equal to $0.1, 0.3, 0.5, 0.7$, respectively.

	0.1	0.3	0.5	0.7
subst2	0.41	128.58	90.04	595.96
censReg1	0.11	0.38	8.01	1.05
best	0.02	0.47	0.47	0.47

The following set of four graphs shows the bias of the predictions from these methods for *cprop* equal to $0.1, 0.3, 0.5, 0.7$, respectively.



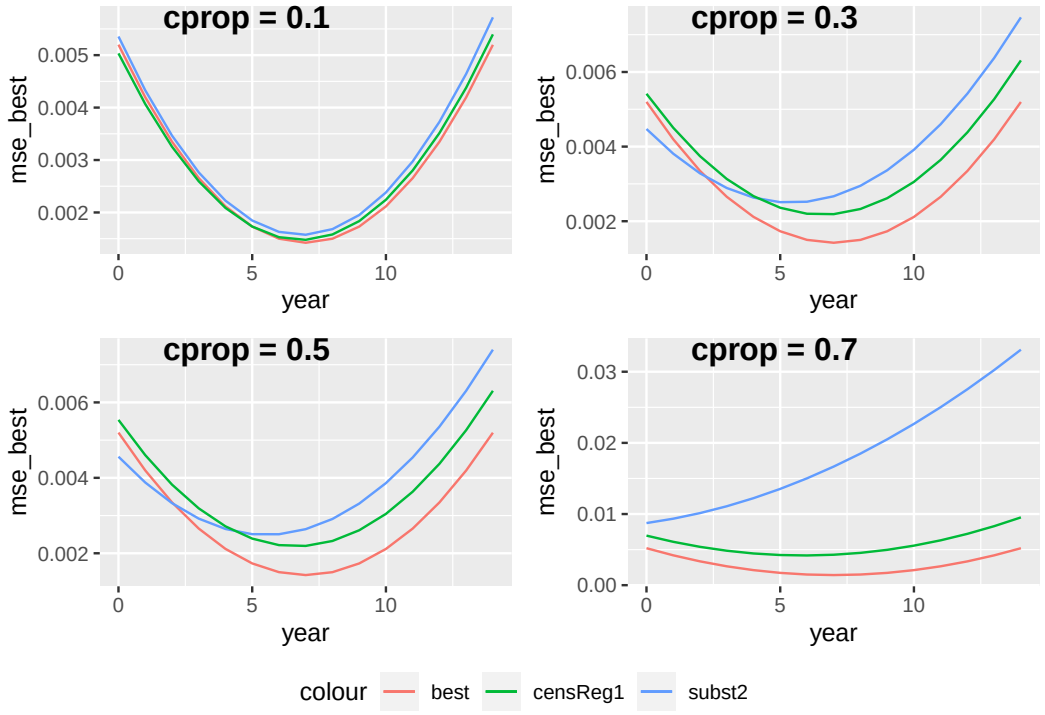
7.3 MSE of estimates and predictions

The following table shows the MSE of the estimates from these methods for *cprop* equal to $0.1, 0.3, 0.5, 0.7$, respectively.

	0.1	0.3	0.5	0.7
subst2	808.79	804.84	806.61	987.64
censReg1	762.65	904.49	982.25	974.67

	0.1	0.3	0.5	0.7
best	770.61	837.01	837.01	837.01

The following set of four graphs shows the MSE of the predictions from these methods for *cprop* equal to $0.1, 0.3, 0.5, 0.7$, respectively.



8 Excluded content

8.0.1 An alternative approach, `censReg0`

We will also evaluate an alternative approach, which we will call `censReg0`.

The `censReg0` model differs from `censReg1` in that it uses A_i instead of X_i as the predictor variable.

This means that each observation y_i^* is assumed to have a normal distribution with mean

$$\mu_{i_A} = \alpha_A + \beta_A a_i$$

and variance σ^2.

The same mathematical steps as shown above yield the corresponding log-likelihood function

$$\log(L) = \sum_{i=1}^{N} \Bigg[(1 - I_i)[\log(\phi((y_i - (\alpha_A + \beta_A a_i))/\sigma)) - \log(\sigma)]$$

$$+ I_i \times \log[\Phi((LOQ - (\alpha_A + \beta_A a_i))/\sigma)] \Bigg]$$

Thus the parameter estimates $\{\widehat{\alpha}_A, \widehat{\beta}_A\}$ are found directly from the maximisation of the corresponding log-likelihood function, without any imputation step; hence the name `censReg0`.

We obtain the prediction of $E(Y|A = a)$ for $a \in \{0, 1, 2, ..., 14\}$ from

$$E(Y|A = a) = \widehat{\alpha} + \widehat{\beta} a$$

We calculate and present the MSE, squared-bias, and variance of the estimates of $\widehat{\beta}$ and the annual predictions $E(Y|A = a)$ for $a \in \{0, 1, 2, ..., 14\}$.

References

Danielsson, Sara, Suzanne Faxneld, and Anne L. Soerensen. 2020. *The Swedish National Monitoring Programme for Contaminants in Marine Biota (Until 2018 Year's Data) - Temporal Trends and Spatial Variations.*

Donald R. Barr, and E. Todd Sherrill. 1999. "Mean and Variance of Truncated Normal Distributions." *The American Statistician* 53 (4): 357. https://doi.org/10.2307/2686057.

Helsel, Dennis R. 2012. *Statistics for Censored Environmental Data Using Minitab and R.* Hoboken, N.J.: Wiley.

———. 2006. "Fabricating Data: How Substituting Values for Nondetects Can Ruin Results, and What Can Be Done About It." *Chemosphere*, Environmental Chemistry, 65 (11): 2434–9. https://doi.org/10.1016/j.chemosphere.2006.04.051.