Stockholms universitet

Approaches for manipulating censored data

Marc Roddis

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2020:21
Matematisk statistik
September 2020

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Bachelor Thesis **2020:21**
http://www.math.su.se

Approaches for manipulating censored data

Marc Roddis*

September 2020

Abstract

Data that is too uncertain to be reported as specific numbers is reported as censored. Censored data is generally very common from chemical analyses, so the manipulation of censored data plays a key role in statistical analysis of such data. The Swedish National Monitoring Programme for Contaminants (SNMPC) manipulates censored data by substitution; our main idea is to use imputation by censored regression instead. We show that this idea is viable and has relevance to real SNMPC data. We present its mathematical basis; formulate conjectures; outline the design, purpose and implementation of our simulation experiments; and report and discuss our experimental results. In our main experiments, substitution and imputation by censored regression are used to generate distinct manipulated datasets. From these datasets, estimates and predictions and their variance, squared-bias, and MSE are computed and reported. Our main finding is that imputation by censored regression generally results in much lower squared-bias than results from substitution.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: marc.roddis@gmail.com. Supervisor: Martin Sköld.

Contents

1 Introduction

1.1 General statement of our goals

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."

Datasets collected through environmental monitoring programs such as SNMPC, invariably contain censored data. Censored data arises from censoring, which means that observed numerical measurements are changed into censored values prior to the data being reported. Here is an illustrative example: the observed value 0.0001 is reported as the censored value < 0.001. Censoring can be done for various reasons and in various contexts. The key principle is that censoring is applied to measurements that are too uncertain to be reported numerically.

SNMPC describe their approach thus "[...] concentrations reported as being below the Level Of Quantitation (LOQ). Such values are included in the analysis as if they were true observations with a value of $\frac{\text{LOQ}}{\sqrt{2}}$. Due to the arbitrariness of this procedure, any results based on series with a high rate of values below LOQ should be interpreted with caution" (Danielsson, Faxneld, and Soerensen 2020).

Such substitution methods have also 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."

Broadly speaking, the goal of this study is to establish a less arbitrary approach for manipulating censored data and evaluate it against the SNMPC's approach.

Manipulation of censored data can be done by omission, or by replacement by data that is either fabricated by substitution or imputed from a statistical model. 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." Since omission is also known to generally produce bias (Helsel 2012), our prior understanding is that imputation is generally the best approach for manipulation of censored data.

Dozens of contaminants are monitored in the SNMPC, among these are polychlorinated biphenyls (PCBs). 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. The viability of this idea, and its relevance in relation to SNMPC, is outlined in Section 1.2. The mathematical basis for imputation by censored regression is presented in Section 1.3.

Although imputation by censored regression is a general method with wide applicability, the application we focus on in this study is long-term time trends for the concentration of polychlorinated biphenyls (PCBs) in marine biota, as monitored by SNMPC. 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 one of the 12 classes of persistent organic pollutants initially included in the Stockholm Convention in 2001 (Vanden Bilcke 2002). These properties of PCBs illustrate both the importance for society of environmental monitoring programmes such as SNMPC, and our rationale for choosing to focus on this application of our statistical methods.

1.2 Exploratory Data Analysis (EDA)

We begin by performing EDA, and model fitting, from the large dataset `pcb.csv`, which was provided from SNMPC. This dataset has 5056 observations of 18 variables; these variables include: measured concentrations of seven PCBs (`CB28`, `CB52`, `CB101`, `CB118`, `CB138`, `CB153`, `CB180`); `YEAR` (1984-2017); an ID for each observation; and nine other variables such as species and age.

Our exploration of this dataset found that:

- The most recent 15-year period 2003-2017 had sufficient relevant, consistent data, so we will focus solely on this time period.
- It is reasonable to model the observed PCB concentrations as log-normal

distributed.

- The data for `CB153` had no censored values, whereas `CB28` data had the highest proportion of censored values; this proportion was 0.34. Every such censored value is known to lie within the interval $(0, \text{LOQ})$. However, more than 10 different LOQ values are used for censoring different `CB28` observations in the dataset.
- Species is clearly a confounding variable for the association between `CB153` and `CB28`, so we will focus solely on observations from 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 T from the original dataset `pcb.csv` by first removing all variables except `YEAR`, `CB28`, and `CB153`, and then removing every observation that contained at least one missing value. We then replace the concentrations of `CB28`, and `CB153` by their logarithms (with base e), which we denote throughout as $y = \log(\text{CB28})$ and $x = \log(\text{CB153})$. This means that we view every observation of y and x as having an approximately normal distribution, and the censored y values as laying in the half-open interval $(-\infty, \log(\text{LOQ}))$. We then remove all observations except those from herring species, all observations prior to 2003, and all censored observations and then re-index 2003 as "year zero". We choose to remove censored data because our EDA is only for the purpose of checking feasibility, so we are not aiming for precise estimates. The resulting dataset is our test dataset T.

From T, we fit a linear model for the regression Y on X, and note that the adjusted R-squared equals 0.96. This indicates that y and x are strongly correlated in our test dataset, which means that they are most likely strongly correlated in the SNMPC dataset.

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

$$E(X|A) = -2.91 + \hat{\beta}_A A \tag{1}$$

where $\hat{\beta}_A = -0.02$. The corresponding fitted model for the regression Y on X is

$$E(Y|X) = -3.18 + 0.79X \tag{2}$$

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

1.2.1 Designing our simulation studies from our EDA findings

In our EDA, we found that:

- `CB153` is fully observed (0 % censored), 34 % of the `CB28` values are censored, and multiple LOQ values are used in the SNMPC dataset.
- the distribution of concentrations of each PCB is approximately log-normal,
- $y = \log(\text{CB28})$ and $x = \log(\text{CB153})$ are strongly correlated; moreover, y and x show a similar rate of decrease throughout the 15-year period (results omitted).

These findings show that our method of choice (imputation by censored regression) is viable, and relevant in relation to SNMPC. The centre-piece of our work will be simulation studies because these will allow us to generate uncensored data and apply censoring, so that we obtain censored data whilst also knowing the underlying true values. Our simulations will be designed to be relevant for the SNMPC data. In particular, for all i:

- x_i and y_i will have a normal distribution.
- The association between x_i and a_i is given by (1), the association between y_i and x_i is given by (2).

However, the multiple LOQ values used in the SNMPC data raise issues beyond the scope of our study, so we will make two simplifying assumptions throughout:

- There is precisely one LLOQ value per simulated dataset. Note that we denote the censoring threshold for y as LLOQ, where $\text{LLOQ} = \log(\text{LOQ})$, which corresponds to the threshold LOQ for `CB28`. Note also that this correspondence is one-to-one.
- Type II censoring is used. This means that we determine LLOQ by censoring a fixed proportion, which we denote as `cprop`, of all observed y_i values. Concretely, LLOQ = (cprop $\times$ 100)th percentile of all y_i values, which means (for example) that we denote $\text{LLOQ} = \text{median}(y)$ as $\text{cprop} = 0.5$.

The variable parameters for our simulations will be β_A, `cprop`, and σ, where $\sigma^2 = Var(y_i|x_i)$ for all i. Further implementation details for our simulations are presented in Chapter 2.

1.3 Imputation by censored regression

This method combines the two concepts *regression imputation* and *censored regression*; these concepts are outlined in the following two sub-sections.

1.3.1 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

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

where NA represents a missing value.

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

$$y = \alpha_X + x\beta_X \tag{3}$$

for Y on X based on the c complete observations $D_c = \{(x_1, y_1), (x_2, y_2), \ldots, (x_c, y_c)\}$. We then impute each of the missing observations $\{y_{c+1}, \ldots, y_n\}$ by substituting the corresponding x value into the regression equation.

This is illustrated in Figure 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 $C = D_c \cup \{(-5.00, -7.12), (-3.50, -5.91)\}$.

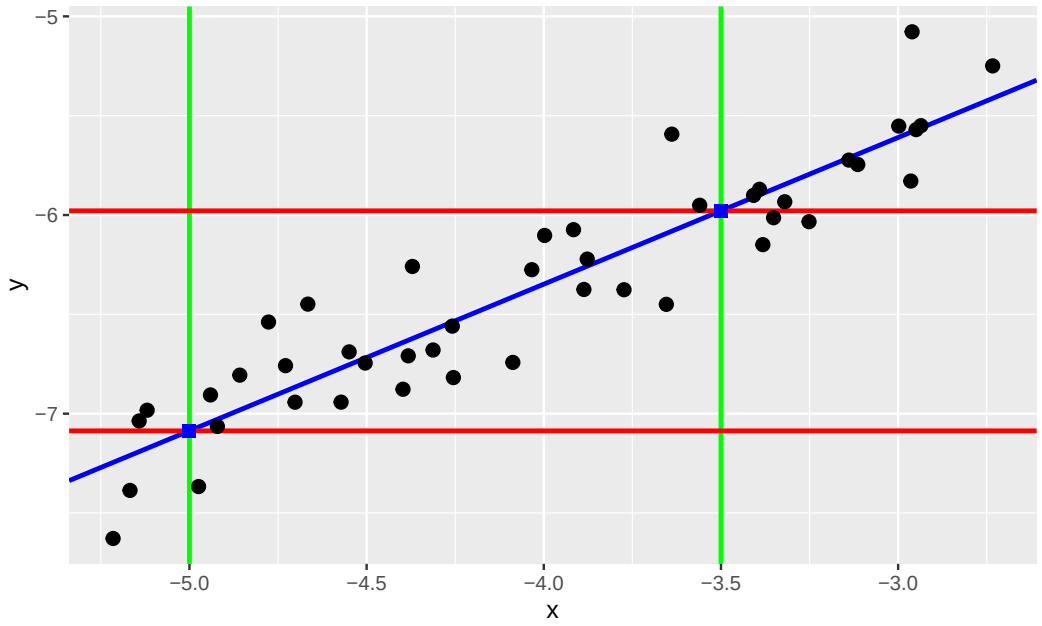

Figure 1: The green lines show incomplete observations (x is known, y is missing). The blue squares show the corresponding data points imputed from the regression line, whilst the red lines show the imputed y values

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 imputation lies outside the scope of this report.

1.3.2 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 $i \in \{1, 2, \ldots, n\}$, $y_i = y_i^*$ 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 and denote the censoring threshold as LLOQ (rather than a).

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 LLOQ. Therefore

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

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

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 LLOQ. This illustrated in Figure 2 for LLOQ $= \mu - \sigma$; the green and red regions show the assumed distribution of the uncensored observations, and of the true values for the censored observations, respectively.

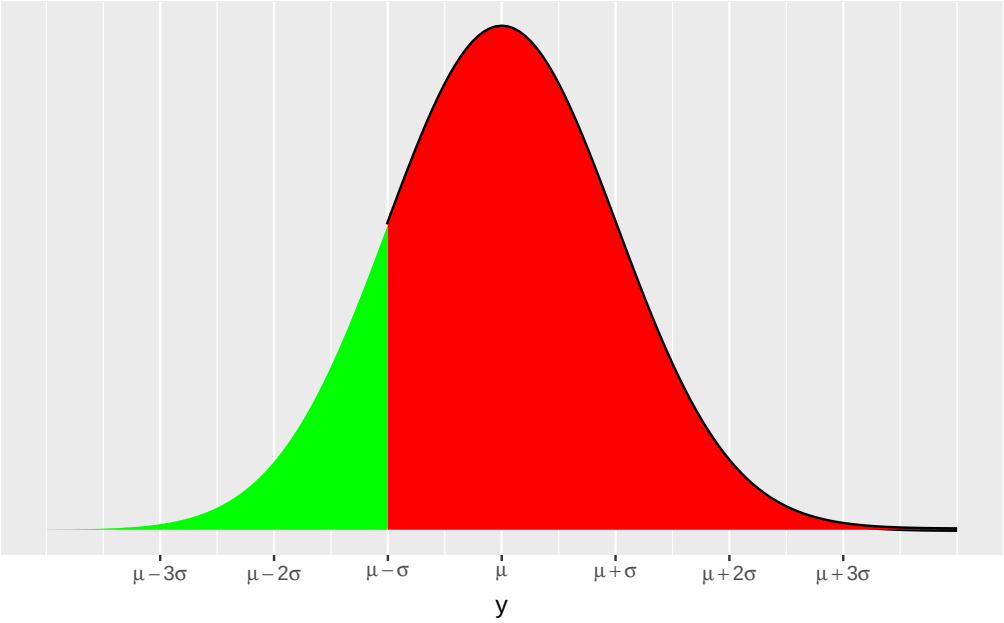

Figure 2: Truncated normal distribution

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

$$\mu_{i_X} = \alpha_X + \beta_X x_i \tag{4}$$

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

$$f(y_i^*) = \frac{\phi((y_i^* - \mu_{i_X})/\sigma)}{\sigma} \tag{5}$$

where μ_{i_X} is given by (4), and ϕ is the pdf of a normal distribution with mean = 0 and variance = 1.

The probability that y_i^* is censored equals

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

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

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

So the log-likelihood function is

$$\log(L) = \sum_{i=1}^{n} \left[(1-I_i)[\log(\phi((y_i-\mu_{i_X})/\sigma))-\log(\sigma)]+I_i\times\log[\Phi((\text{LLOQ}-\mu_{i_X})/\sigma)] \right] \tag{7}$$

We will use the `censReg()` function from the `censReg` package (Henningsen 2012) in R to maximise this log-likelihood function to obtain the maximum likehood estimates $\hat{\alpha}_X$, $\hat{\beta}_X$ and $\hat{\sigma}$. Note that the `censReg` package calls the `maxLik` package (Henningsen and Toomet 2011) to perform the likelihood maximisation step.

Recall that for regression imputation the missing values were imputed by

$$\hat{y}_i = E(Y|X = x_i) = \hat{\alpha}_X + \hat{\beta}_X x_i = \mu_{i_X} \tag{8}$$

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

$$\hat{y}_i = E(Y|X = x_i, Y < \text{LLOQ}) \tag{9}$$

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

$$\hat{y}_i = -\hat{\sigma}\frac{exp[((-1/2)(\mu_{i_X} - \text{LLOQ})/\hat{\sigma})^2]}{[1 - (\Phi(\mu_{i_X} - \text{LLOQ})/\hat{\sigma})]\sqrt{2\pi}} + \mu_{i_X} \tag{10}$$

where μ_{i_X} is given by (4).

In our practice, we use 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

$$\mu_{i_{X,A}} = \alpha_{X,A} + \beta_X x_i + \beta_A a_i \tag{11}$$

and variance σ^2.

This means that the likelihood function for the `censReg2` model is given by substitution of (11) into (6). Consequently, maximisation of the corresponding log-likelihood function gives the maximum likehood estimates $\hat{\alpha}_X$, $\hat{\beta}_X$, $\hat{\beta}_A$ and $\hat{\sigma}$. This means that the censored y_i are imputed by $\hat{y}_i$, which is obtained by substitution of (4) by (11) into (10).

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 this censoring proportion, under this assumption.

1.3.2.1 The distribution of imputed values from `censReg1` Recall that for left-censored data it is known that $y_i < \text{LLOQ}$. Suppose that we remove this condition and use (8) instead of (9) for the imputed values; we will call this naive approach `censReg1naive`.

We show (in Figure 3) the distribution of the ML estimates for y_i from illustrative dataset, using the `censReg1` and `censReg1naive` approaches. Our purpose is to illustrate our conjecture that the `censReg1naive` approach produces significantly higher squared-bias than `censReg1`. We will later verify this conjecture quantitatively through simulation studies.

For clarity, in Figure 3, we have displayed every data point of D as uncensored. However the true value is unknown for all $y_i < \text{LLOQ}$, when we view this data as censored. We choose cprop $= 0.5$. In Figure 3, the red line shows $y = E(Y|X = x)$, the green line shows $y = E(Y|X = x, Y < \text{LLOQ})$, and the blue line shows $y = \text{LLOQ}$. We have selected two observations for which $y_i \leq \text{LLOQ}$ 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 $y = E(Y|X = x, Y < \text{LLOQ})$, 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 also see that the green line stays below $y = \text{LLOQ}$ for all x, which means that $\hat{y}_i < \text{LLOQ}$ for every imputation from $y = E(Y|X = x, Y < \text{LLOQ})$. However this is not true for the red line; in fact, we see $\hat{y}_{I2} > \text{LLOQ}$ imputed from $y = E(Y|X = x, Y < \text{LLOQ})$, which contradicts the fact that $y_{I2} < \text{LLOQ}$ was observed. We have thus illustrated why it is necessary to condition on both $X = x$ and $Y < \text{LLOQ}$) and verified the plausibility of equation (10). 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.

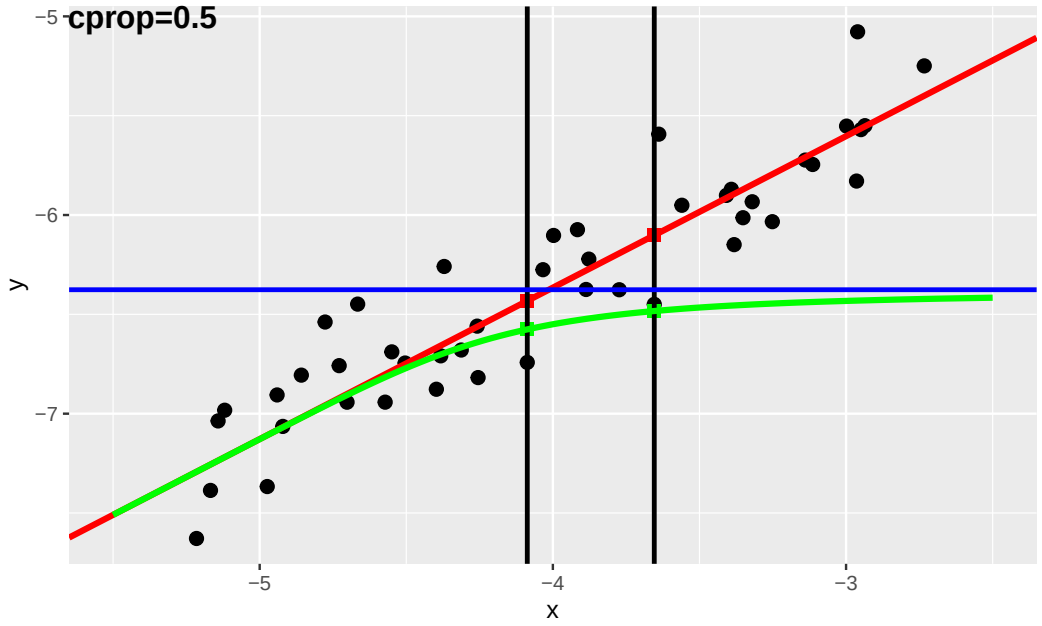

Figure 3: Imputation from `censReg1` (green) and `censReg1naive` (red)

1.4 Data manipulation

1.4.1 Overview

For each simulation study, we will first generate "uncensored datasets" with three fully observed variables Y, X, and A. From each of these, we will apply Type II left-censoring to Y to get the corresponding "censored dataset". We will then apply each data manipulation approach to the censored data to obtain the corresponding "manipulated dataset". Every approach will give a distinct manipulated dataset.

Our imputation by censored regression approaches `censReg1`, `censReg2`, and `censReg1naive` were presented in Section 1.3.2.

Recall that the approach used by SNMPC is substitution by $\frac{LOQ}{\sqrt{2}}$, which is the most commonly used substitution value in the research literature that underpins 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 the values that are observed to lie within the interval $(0, LOQ)$ are left-censored at LOQ. So we will test the three approaches that use the substitution values LOQ, $\frac{LOQ}{\sqrt{2}}$ and $\frac{LOQ}{2}$; we name these `subst1`, `subst2`, and `subst4`, respectively. We choose

these names because $LOQ = \frac{LOQ}{\sqrt{1}}$ and $\frac{LOQ}{2} = \frac{LOQ}{\sqrt{4}}$. We choose these values because $\frac{LOQ}{2} < \frac{LOQ}{\sqrt{2}} < LOQ$, so we can compare results from the substitution value that SNMPC uses with one lower and one higher value. Note that for our simulation studies we will use the corresponding logarithmised values LLOQ, $\text{LLOQ} - \log(\sqrt{2})$, and $\text{LLOQ} - \log(2)$, respectively.

We will also test the `omit` approach, in which the censored observations are omitted. Note that the number of observations is unchanged by every imputation-based or substitution-based approach, whereas it is lowered by the `omit` approach.

1.4.2 Conjectures

1.4.2.1 Bias produced by the `omit` approach Omission of censored observations is illustrated in Figure 4, in which the LLOQ is again shown as a blue horizontal line. The regression lines from the uncensored dataset, and the corresponding smaller manipulated dataset, are shown in green and red respectively. Since x and y are correlated, points that lie below the blue line are more likely to lie in the lower left of the figure and more likely to lie below the regression line. Since the `omit` approach removes all such points, the resulting red line is above the green one on the left of the figure and thus has a smaller slope. This illustrates the basis for our conjecture that the `omit` approach will generally produce relatively high squared-bias.

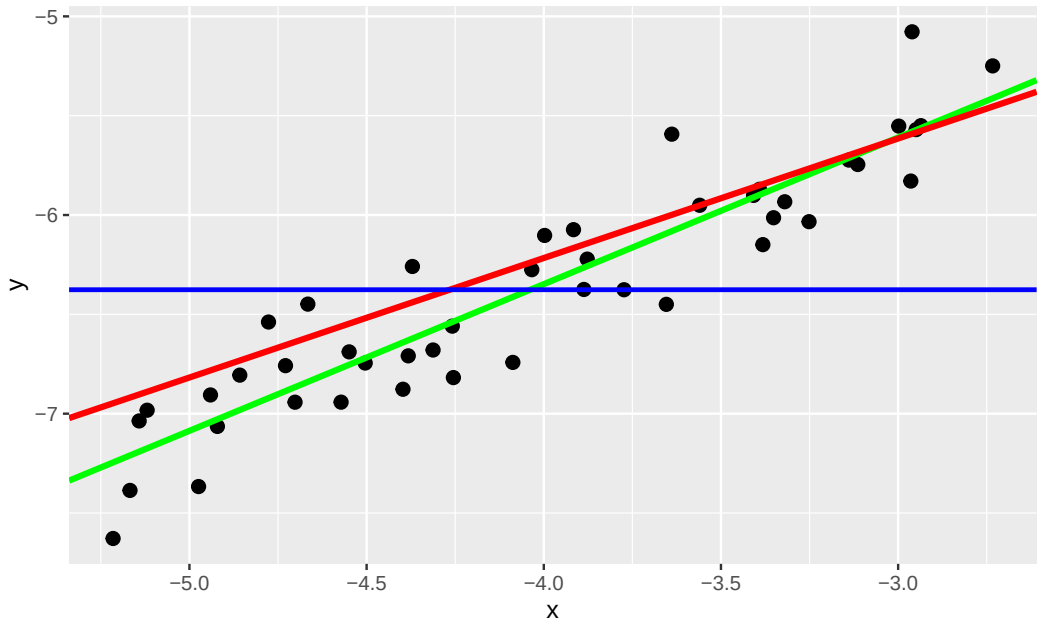

Figure 4: The regression lines for the uncensored dataset, and for the smaller manipulated dataset, are shown in green and red, respectively

1.4.2.2 The best choice of substitution value depends on the censoring proportion `cprop` The following plot is the same as Figure 3, except that it illustrates the `subst1`, `subst2`, and `subst4` approaches instead of `censReg1naive`. Moreover, whilst the right-most vertical line has been drawn through the same data point as before, a new data point was chosen for the other vertical line. For each of these two points, their values after manipulation by `subst1`, `subst2`, `subst4` and `censReg1` are shown by blue, red, yellow and green points, respectively.

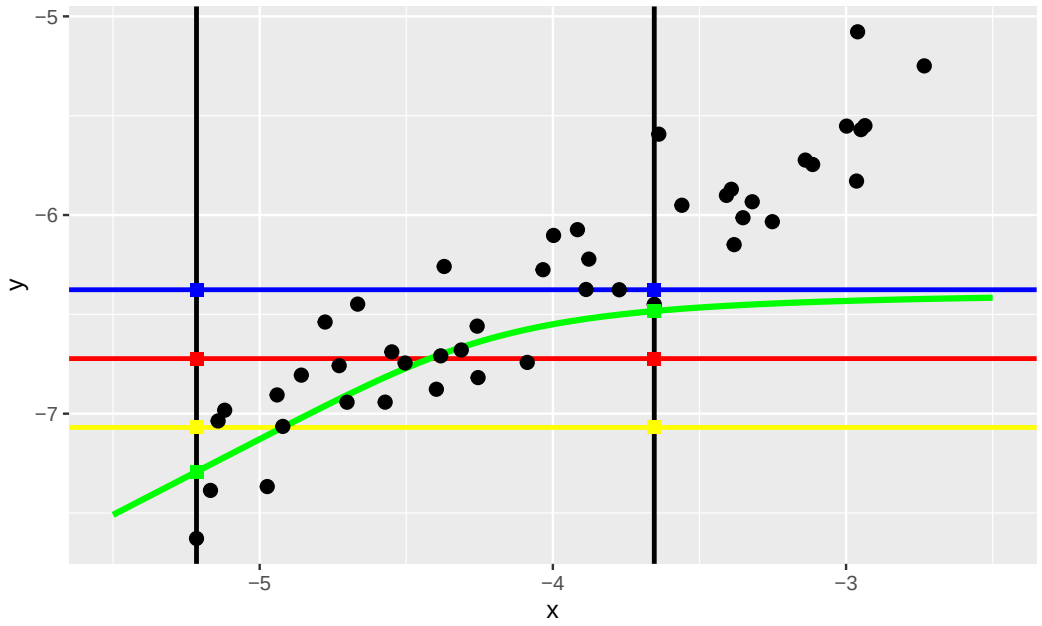

Figure 5: Data manipulation by `subst1` (blue), `subst2` (red), `subst4` (yellow), and `censReg1` (green) for cprop = 0.5

Our next illustration (Figure 6) is the same as the previous one, except that cprop = 0.2 and cprop = 0.7 are used and data manipulation is not displayed for any individual data points. We see that the censored data points lie closest on average to LLOQ for cprop = 0.2 whereas they are closest on average to LLOQ − log(2) for cprop = 0.7; moreover, we saw from the previous plot that they are closest on average to LLOQ$-\log(\sqrt{2})$ for cprop = 0.5. For this dataset, this shows that as the value of `cprop` increases, the relative amount of squared-bias from `subst1`, `subst2`, and `subst4` will be higher, intermediate, and lower, respectively. This illustrates our conjecture that fabrication by substitution generally has lower robustness than imputation by censored regression.

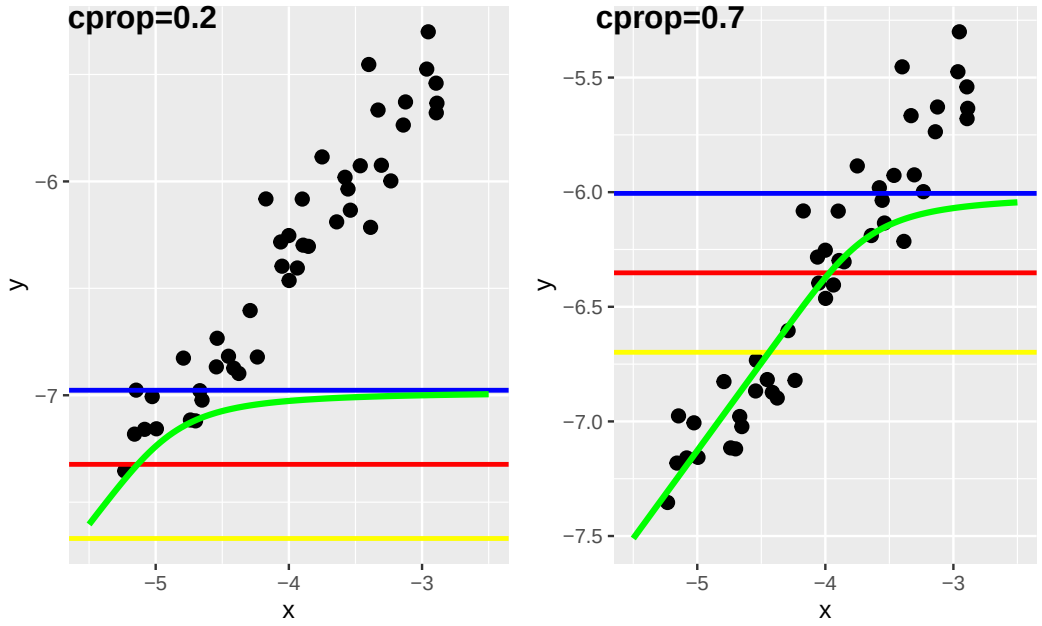

Figure 6: The effect of a higher, and a lower, value of `cprop` on data manipulation by `subst1` (blue), `subst2` (red), `subst4` (yellow), and `censReg1` (green)

1.4.2.3 The effect of the strength of correlation between Y and X

The value chosen for the variable parameter σ determines the strength of correlation for the regression Y on X. The larger the value of σ, the weaker is this correlation.

We make two conjectures regarding the value of σ.

- The performance of imputation-based approaches will be decreasingly good for increasing values of σ because the information about Y given by X is more noisy.
- The squared-bias from `subst1`, `subst2`, and `subst4` will generally be relatively lowest for small, intermediate and high relative values of σ respectively.

Our rationale is that as σ increases, the mean of the unknown true y values of the censored data decreases; this can be seen by analogy with the result (Weisstein 2020) that

$$E(Y) = \frac{\sigma\sqrt{2}}{\sqrt{\pi}}$$

for a half-normal distribution defined by $Y = |X|$ where $X \sim N(0, \sigma^2)$.

2 Implementation of our simulation study

2.1 Dataset generation

2.1.1 Generation of uncensored datasets

Guided by the findings of the EDA we presented in Section 1.2, we generate our uncensored datasets as follows (at every iteration):

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

$$x_{a_i} = -2.91 - \beta_A a_i + e_{a_i} \tag{12}$$

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

4. We generate all y_i from

$$y_i = -3.18 + 0.79x_i + \epsilon_i \tag{13}$$

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.

2.1.2 Generation of censored datasets

We generate every censored dataset from the corresponding uncensored dataset in accordance with the principles outlined in Section 1.3.2. This means that we use y_i^* instead of y_i, where y_i^* refers to the ith observation prior to it being observed. We determine $LLOQ$ 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 $15N \times (1 - \text{cprop})$ uncensored y_i observations. Recall that $LLOQ = (\text{cprop} \times 100)$th percentile of all y_i values, which means that the value of $(LLOQ|\text{cprop})$ is constant and thus

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

2.1.3 Generation of incomplete datasets

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

2.1.4 Generation of manipulated datasets

We will also obtain a distinct manipulated dataset from every data manipulation approach, as described in Section 1.4.1.

2.2 Calculation of results

2.2.1 Calculation of baseline results

We obtain baseline results, which we call the `best` results, for the estimation of β and for the prediction of $E(Y|A = a)$ for $a \in \{0, 1, 2, \ldots, 14\}$ from the uncensored dataset. Substitution from (12) into (13) gives

$$y_i = -3.18 + 0.79(-2.91 - \beta_A a_i + e_i) + \epsilon_i \tag{14}$$

$$= -3.18 + 0.79(-2.91) - 0.79\beta_A a_i + 0.79e_i + \epsilon_i \tag{15}$$

$$= \alpha + \beta a_i + \varepsilon_i \tag{16}$$

where $\alpha = -3.18 + 0.79 \times -2.91 = -5.4789$, and $\beta = 0.79\beta_A$. Also $\varepsilon_i = 0.79e_i + \epsilon_i$, where $e_i \sim N(0, 0.1^2)$ and $\epsilon_i \sim N(0, \sigma^2)$.

We obtain the parameter estimates $\hat{\alpha}$ and $\hat{\beta}$ by fitting a simple linear regression model using the `lm()` method in R. The `best` results for $\hat{\beta}$ are given by (17), (18), and (19).

$$\text{MSE}(\hat{\beta}) = E[(\beta - \hat{\beta})^2] \tag{17}$$

$$[\text{bias}(\hat{\beta})]^2 = (E[\hat{\beta}] - \beta)^2 \tag{18}$$

$$\text{Var}(\hat{\beta}) = E[(\hat{\beta} - E[\hat{\beta}])^2] \tag{19}$$

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

$$E(Y|A = a) = \widehat{\alpha} + \widehat{\beta}a \tag{20}$$

We call the MSE, squared-bias, and variance, for every such prediction, the `best` results.

2.2.2 Calculation of results from every data manipulation approach

We perform the same calculation as described in the previous section to obtain the corresponding results for every data manipulation approach. The only difference is that we start with the corresponding manipulated dataset instead of the uncensored dataset. All estimation results are presented in tables and all prediction results are displayed as graphs in Chapters 3 and 4.

2.2.3 Calculation of results from the censored dataset without data manipulation (`censReg0`)

Our `censReg0` approach differs from the `censReg1` approach in the choice of predictor variable for the model for the maximum likelihood estimation step. We have seen that the `censReg1` approach uses X_i as the predictor for this step. The `censReg0` approach uses A_i as the predictor instead for this step.

The same mathematical steps as in Section 1.3.2 yield the log-likelihood function

$$\log(L) = \sum_{i=1}^{n} \left[(1-I_i)[\log(\phi((y_i-\mu_{i_A})/\sigma))-\log(\sigma)]+I_i\times\log[\Phi((\text{LLOQ}-\mu_{i_A})/\sigma)] \right] \tag{21}$$

where $\mu_{i_A} = \alpha + \beta a_i$.

This function is the same as (7), except that μ_{i_A} is used in place of μ_{i_X}. However, the key distinction is that the parameter estimates $\hat{\alpha}$ and $\hat{\beta}$ are found directly from the maximisation of (21) without any imputation step. We then obtain the `censReg0` results from equations (17), (18), (19), and (20).

Our `censReg0` approach is designed to test our conjecture that since $|\beta_X| = 0.79$ is much greater than $|\beta_A|$, `censReg0` will result in estimates and predictions

with higher variance than from `censReg1`. The results from the `censReg0` approach are presented in Chapters 3 and 4, together with those from the other approaches.

3 Results from our screening experiments

In this chapter we describe and present results from our screening experiments, from which we determine appropriate values for the number of iterations per simulation, the sample size, and the variable parameters β, σ, and `cprop`.

We also select appropriate data manipulation approaches for our main experiments by obtaining results for the MSE, squared-bias, and variance of the estimates of β from the following data manipulation approaches:

- the three substitution approaches `subst1`, `subst2`, `subst4`.
- the three imputation by censored regression approaches `censReg1`, `censReg2`, and `censReg1naive`
- the `omit` approach.

We will also obtain our baseline results (`best`), and results from the `censReg0` approach

We will present the results from our main experiments in Chapter 4. The results from our screening experiments are presented in the remainder of this chapter in tables; these results are pre-multiplied by 10^7 and then rounded, to make them easier to read and compare. Results for MSE and variance are rounded to one decimal place, whilst squared-bias is rounded to two decimal places.

3.1 Selection of the number of iterations per simulation

Our results from preliminary experimentation (results not shown) 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 for our screening experiments, and 10000 iterations for our main experiments.

3.2 Determination of appropriate sample size

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

The parameter values $\{\text{cprop} = 0.3, \beta_A = -0.02, \sigma = 0.1\}$ are based on estimates from our test dataset (Section 1.2). This 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(sample size) ≈ 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 $\{\text{cprop} = 0.3, \beta_A = -0.02, \sigma = 0.3\}$, whilst the value of `sample size` equals 50, 25, 12, and 6, respectively.

3.2.1 Results for various sample sizes

The variance of estimates of β from all approaches, from simulations with sample sizes 50, 25, 12, and 6, is shown in the columns of the following table.

	ss=50	ss=25	ss=12	ss=6
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
censReg0	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

Table 1: Variance ($\times 10^7$) of estimates for sample sizes 50, 25, 12, and 6

The following table is the same as the previous one, except that it shows the squared-bias of the estimates of β.

	ss=50	ss=25	ss=12	ss=6
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
censReg0	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
censReg1naive	69.48	66.68	67.15	70.29

Table 2: Squared-bias ($\times 10^7$) of estimates for sample sizes 50, 25, 12, and 6

3.2.2 Our rationale for choosing sample size = 12

Allowing for random error from using only 10000 iterations, we see (from Tables 1 and 2) that the squared-bias is independent of sample size, whereas the variance is inversely proportional to sample size. Moreover since the bias-variance decomposition

$$\text{MSE} = \text{Bias}^2 + \text{Variance}$$

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

Recall that we found 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:

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

We therefore decide to use sample size = 12 for all of our subsequent experiments.

3.3 Selection of data manipulation approaches for further study

We will now use simulations with just 1000 iterations for all eight approaches (and also for our reference results `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 $\{(\text{cprop}, \sigma)\}$ respectively.

3.3.1 Variance of estimates from all approaches

The following table shows the variance of estimates from each approach 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	44.0	67.8	584.4	870.2
subst2	80.9	87.3	697.1	322.0
subst1	42.8	9.3	564.9	120.6
censReg1	48.3	74.7	661.8	766.9
censReg2	48.8	96.8	673.7	1134.9
censReg0	48.7	102.8	673.7	1135.5
best	50.0	50.0	775.2	775.2
subst4	157.2	270.5	886.1	666.1
censReg1naive	43.3	78.7	518.3	845.3

Table 3: Variance ($\times 10^7$) of estimates for low-low, high-low, low-high, and high-high combinations of values for `cprop` and σ

3.3.2 Squared-bias of estimates from all approaches

The following table is the same as the previous one, except that it shows the squared-bias of the estimates of β.

	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

Table 4: Squared-bias ($\times 10^7$) of estimates for low-low, high-low, low-high, and high-high combinations of values for `cprop` and σ

3.3.3 MSE of estimates from all approaches

The following table is the same as the previous one, except that it shows the MSE of the estimates of β.

	Low-Low	High-Low	Low-High	High-High
omit	194.6	1413.4	816.7	2118.0
subst2	328.4	149.8	697.5	837.3
subst1	62.2	1262.1	608.9	1319.2
censReg1	48.3	74.7	665.0	766.6
censReg2	48.7	96.8	676.9	1134.2
censReg0	48.6	102.9	677.0	1134.7
best	50.0	50.0	776.1	776.1
subst4	1444.6	2893.5	906.1	781.9
censReg1naive	74.6	103.8	655.7	935.2

Table 5: MSE ($\times 10^7$) of estimates for low-low, high-low, low-high, and high-high combinations of values for `cprop` and σ

3.3.4 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 approaches in the amount of squared-bias than in the amount of variance. We will therefore focus primarily on the results for squared-bias; we will use terms such as high

and low to compare the relative amount of squared-bias from our different approaches.

We see from Table 4 that the amount of squared-bias is very high from: `subst1` for $\{(\text{cprop}, \sigma)\} = \{(0.7, 0.1), (0.7, 0.5)\}$; `subst2` for $\{(\text{cprop}, \sigma)\} = \{(0.7, 0.5)\}$; `subst4` for $\{(\text{cprop}, \sigma)\} = \{(0.1, 0.1), (0.7, 0.1)\}$. However, all three substitution approaches also have low squared-bias for at least one set of parameter values. This is intriguing and merits further investigation.

For all four parameter value sets, the squared-bias from `censReg1`, `censReg2`, and `censReg0` is very low; moreover it is clearly higher from `censReg1naive`, which verifies our conjecture from Section 1.3.2.1.

The squared-bias from `omit` is generally very high for all combinations of values for `cprop` and σ.

We therefore exclude the two approaches `omit` and `censReg1naive` from our main experiments and include all other approaches. We will use `best` as our baseline results throughout.

3.4 Selection of parameter values

We will select values for the variable parameters β_A, σ and `cprop` with two goals in mind:

- Relevance for SNMPC.
- Testing the conjectures stated in Section 1.4.2.

We will begin use the parameter values $\{\text{cprop} = 0.3, \beta_A = -0.02, \sigma = 0.3\}$ that we used in Section 3.2.1. We will first hold $\{\text{cprop} = 0.3, \sigma = 0.3\}$ fixed and test the four values $\{-0.02, -0.04, -0.08, -0.16\}$ for β_A. We will then hold $\{\text{cprop} = 0.3, \beta_A = -0.02\}$ fixed and test the four values $\{0.1, 0.3, 0.5, 0.7\}$ for σ. We will then hold $\beta_A = -0.02$ fixed and test the four values $\{0.1, 0.3, 0.5, 0.7\}$ for `cprop`. However, we will hold $\sigma = 0.5$ fixed at this higher value. Our rationale is based on the conjectures we stated in Section 1.4.2.1 and Section 1.4.2.2. We reason that the relative performance of imputation-based approaches will be worse at this higher σ value, which will give substitution approaches a better chance to remain competitive at the highest values for `cprop`.

4 Results from our main experiments

We present results from our main experiments for the estimation of β and predictions of $E(Y|A = a)$ for $a \in \{0, 1, 2, \ldots, 14\}$ from our chosen data manipulation approaches, together with the corresponding `best` (baseline) results.

This chapter will have three main sections 4.1, 4.2, and 4.3, which each show the results for four chosen values for β, σ, and `cprop`, respectively. Every main section will have three sub-sections for variance, squared-bias, and MSE, respectively. Every sub-section has a table for results from estimates of β from every approach.

In Sections 4.1 and 4.2, the sub-sections for variance and squared-bias each have three figures, each with four graphs, showing these results for predictions. The three figures show results for the purpose of comparing the imputation-based approaches(`censReg1` and `censReg2`), the substitution-based approaches (`subst1`, `subst2`, and `subst4`), and the generally best performing approach of each type (`censReg1` and `subst2`), respectively. Section 4.3, and the sub-sections for MSE (4.1.3, 4.2.3, and 4.3.3), each have one figure, each with four graphs, showing results of predictions from `censReg1` and `subst2`, exclusively.

In addition, the variance and squared-bias of predictions from `censReg0` is shown in the relevant figures of Sections 4.1.1 and 4.1.2, respectively. Regrettably, as a result of a technical error, these figure legends denote `censReg0` as `censReg0impute`.

Please note that throughout this chapter:

- All results are obtained from simulations with 10000 iterations, and sample size = 12.
- All of the results for estimates are displayed pre-multiplied by 10^7 and then rounded in every table. MSE and variance are rounded to one decimal place, whilst squared-bias is rounded to two decimal places.
- All of the results for predictions are 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.
- Every figure has four graphs, which each show the results for one of the four chosen parameter values for the corresponding parameter. This presentation format is designed to allow us to more easily see the effect of each parameter.

We will discuss all of our results in Chapter 5.

4.1 Results for various values of β_A

For all our simulations in this section, these parameters are fixed: cprop = 0.3, $\sigma = 0.3$, whilst β_A is given the four values: $-0.02, -0.04, -0.08, -0.16$.

4.1.1 Variance of estimates and predictions

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

	-0.02	-0.04	-0.08	-0.16
subst1	166.7	178.3	225.0	305.9
subst2	367.3	351.3	316.7	323.1
subst4	709.9	643.8	470.5	355.2
censReg1	288.9	315.9	362.8	449.5
censReg2	314.1	325.8	364.5	449.7
censReg0	314.4	326.2	365.4	452.7
best	289.4	280.8	288.0	285.3

Table 6: Variance ($\times 10^7$) of estimates for $\beta_A = -0.02, -0.04, -0.08, -0.16$

The following three figures show the variance of predictions from imputation-based approaches, from substitution-based approaches, and from `censReg1` and `subst2`, respectively.

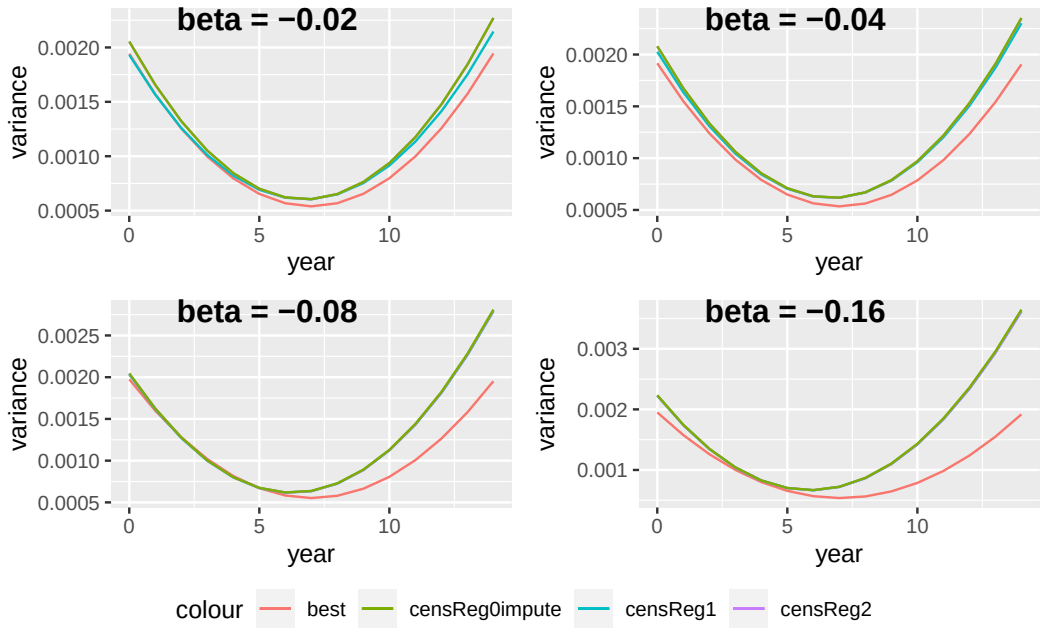

Figure 7: The effect of the value of β_A on the variance of predictions from imputation-based approaches

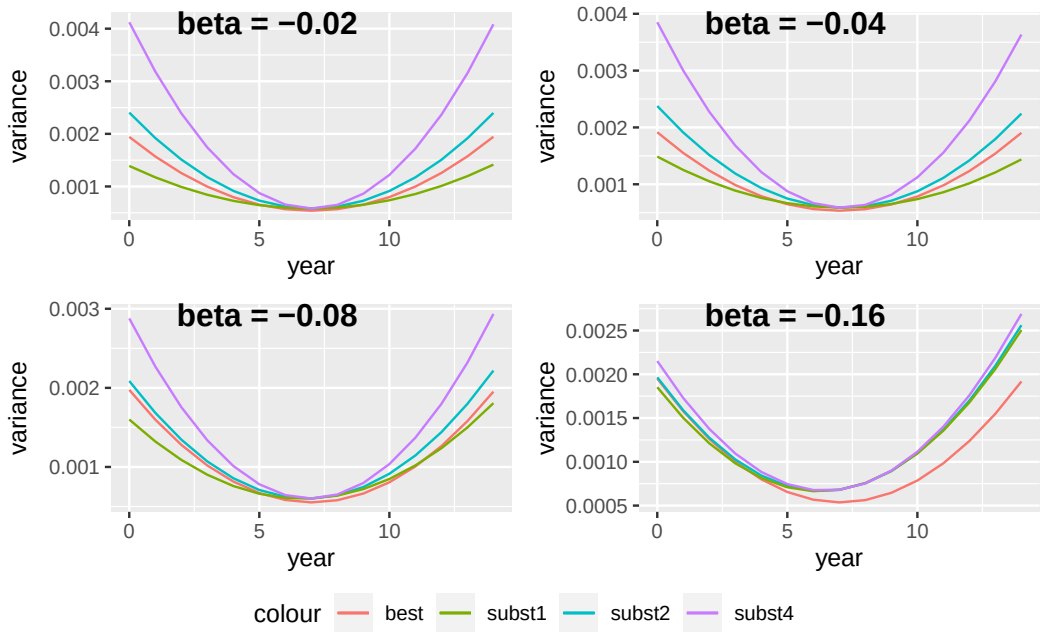

Figure 8: The effect of the value of β_A on the variance of predictions from substitution-based approaches

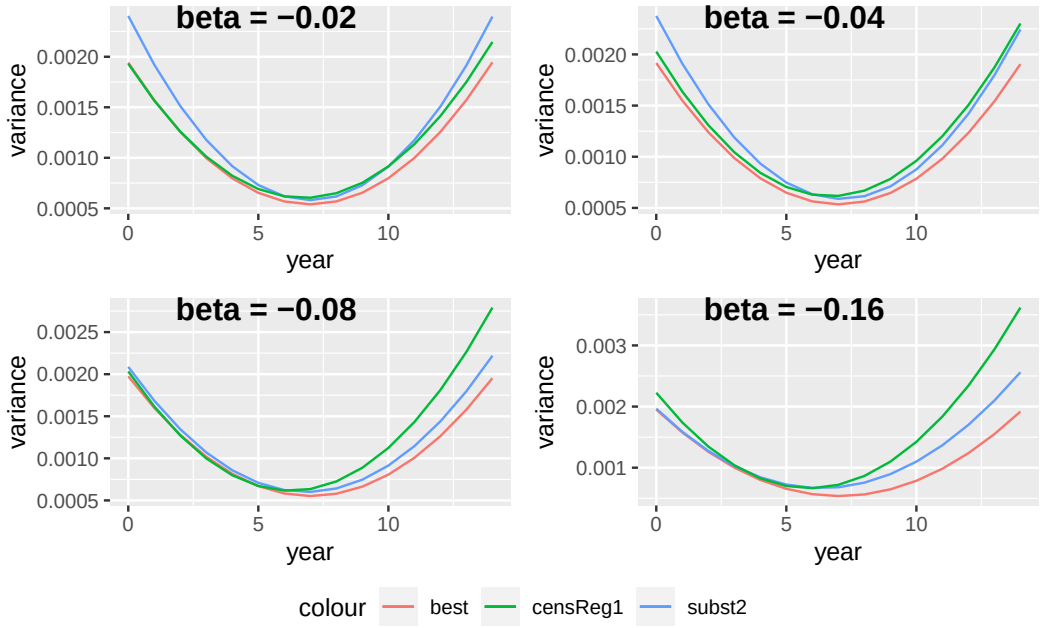

Figure 9: Comparison of the effect of the value of β_A on the variance of predictions from `censReg1` and `subst2`

4.1.2 Squared-bias of estimates and predictions

The following table shows the squared-bias of estimates of β from every approach 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

Table 7: Squared-bias ($\times 10^7$) of estimates for $\beta_A = -0.02, -0.04, -0.08, -0.16$

The following three figures show the squared-bias of predictions from imputation-based approaches, from substitution-based approaches, and from `censReg1` and `subst2`, respectively.

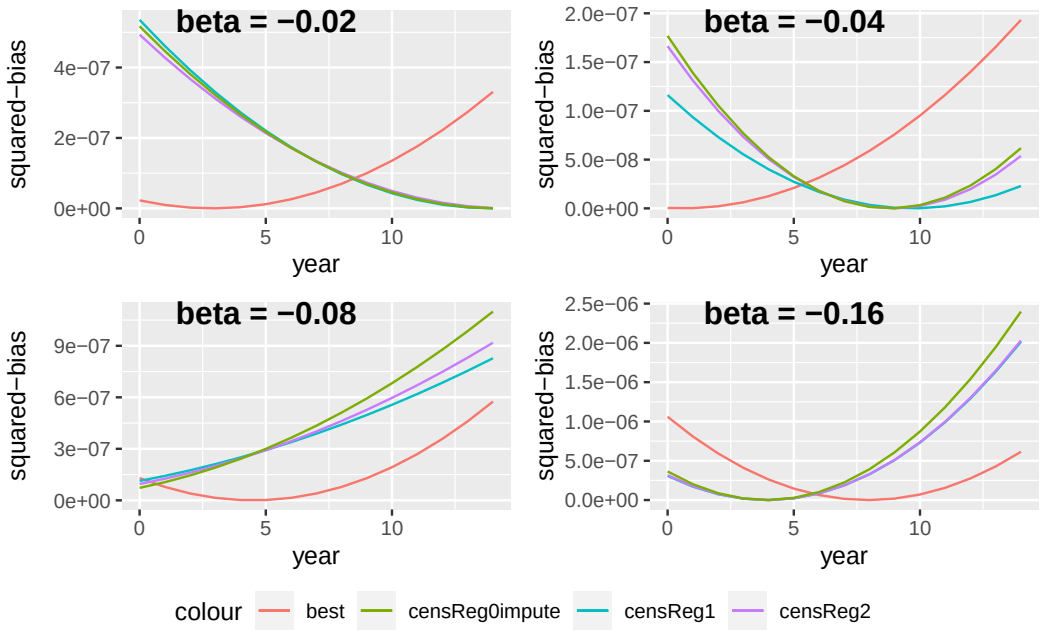

Figure 10: The effect of the value of β_A on the squared-bias of predictions from imputation-based approaches

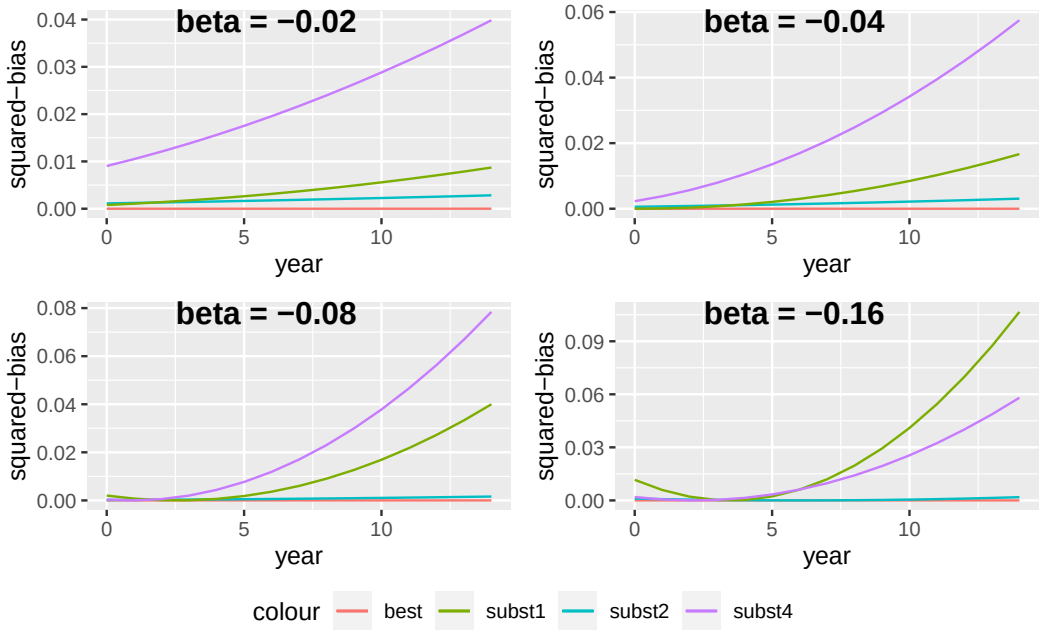

Figure 11: The effect of the value of β_A on the squared-bias of predictions from substitution-based approaches

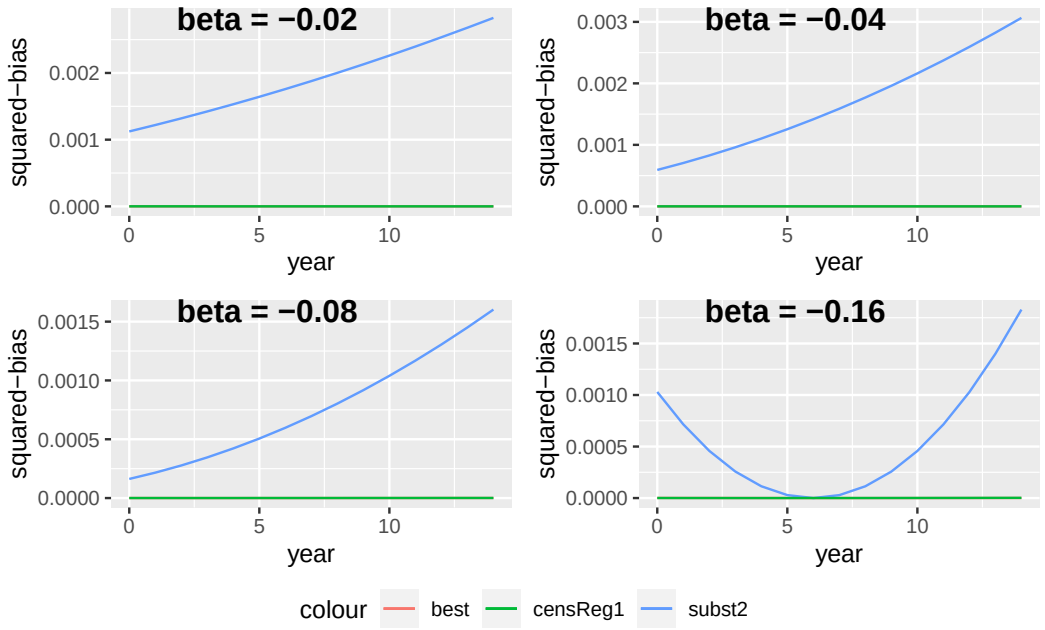

Figure 12: Comparison of the effect of the value of β_A on the squared-bias of predictions from `censReg1` and `subst2`

4.1.3 MSE of estimates and predictions

The following table shows the MSE of estimates for β_A equal to $-0.02, -0.04, -0.08, -0.16$, respectively.

	-0.02	-0.04	-0.08	-0.16
subst1	392.7	1039.1	3281.0	9932.5
subst2	383.3	400.4	354.7	609.0
subst4	1240.7	2523.5	5041.3	4489.3
censReg1	288.9	315.8	362.8	449.6
censReg2	314.1	325.8	364.4	449.9
censReg0	314.4	326.2	365.4	452.8
best	289.4	280.7	288.1	285.4

Table 8: MSE ($\times 10^7$) of estimates for $\beta_A = -0.02, -0.04, -0.08, -0.16$

The following figure shows the MSE of predictions from `censReg1` and `subst2`.

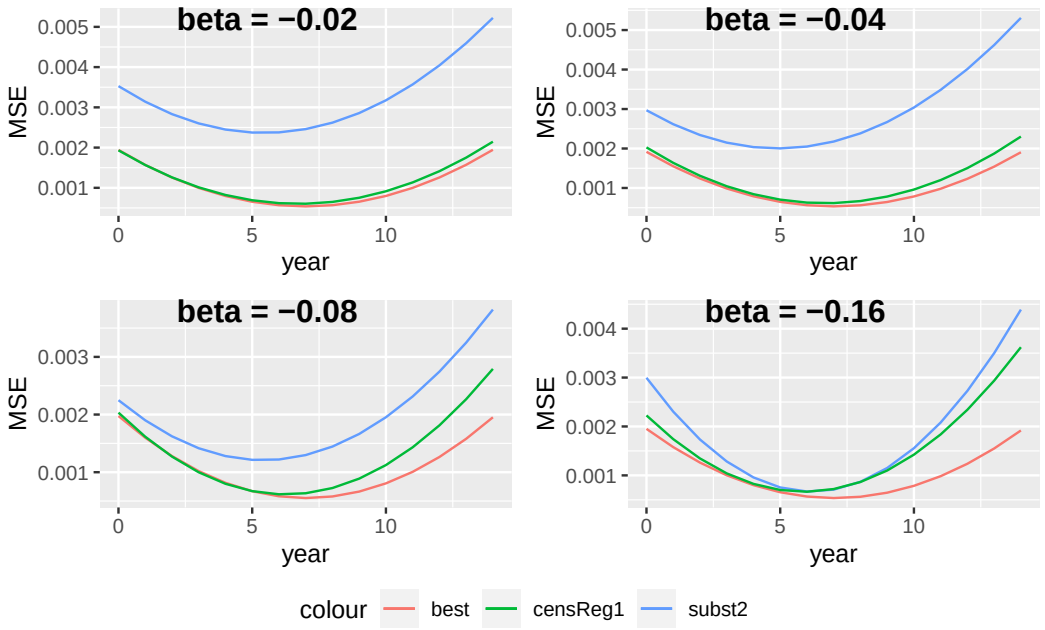

Figure 13: Comparison of the effect of the value of β_A on the MSE of predictions from `censReg1` and `subst2`

4.2 Results 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. Note that the results for $\sigma = 0.3$ appear in the previous section; they are duplicated here to allow us to see the effect of the value of σ more easily.

4.2.1 Variance of estimates and predictions

The following table shows the variance of estimates from every approach for σ equal to $0.1, 0.3, 0.5, 0.7$, respectively.

	0.1	0.3	0.5	0.7
subst1	31.5	166.7	432.7	855.6
subst2	124.2	367.3	728.3	1258.9
subst4	320.7	709.9	1172.9	1815.9
censReg1	51.7	288.9	761.3	1520.5
censReg2	54.8	314.1	831.1	1657.3
censReg0	55.7	314.4	831.0	1657.7
best	48.9	289.4	744.6	1484.1

Table 9: Variance ($\times 10^7$) of estimates for $\sigma = 0.1, 0.3, 0.5, 0.7$

The following three figures show the variance of predictions from imputation-based approaches, from substitution-based approaches, and from `censReg1` and `subst2`, respectively.

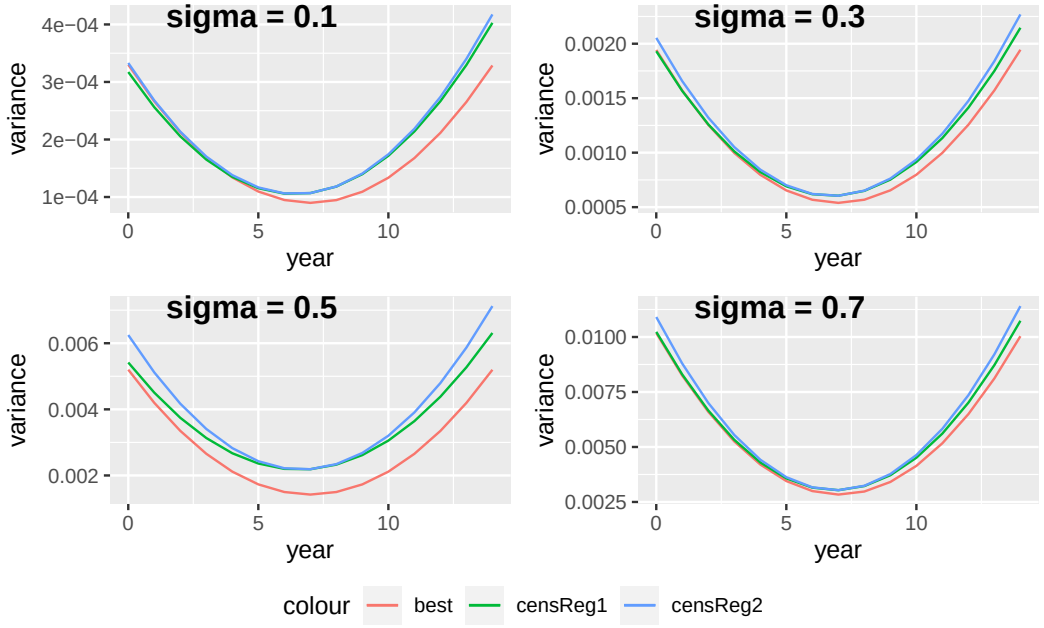

Figure 14: The effect of the value of σ on the variance of predictions from imputation-based approaches

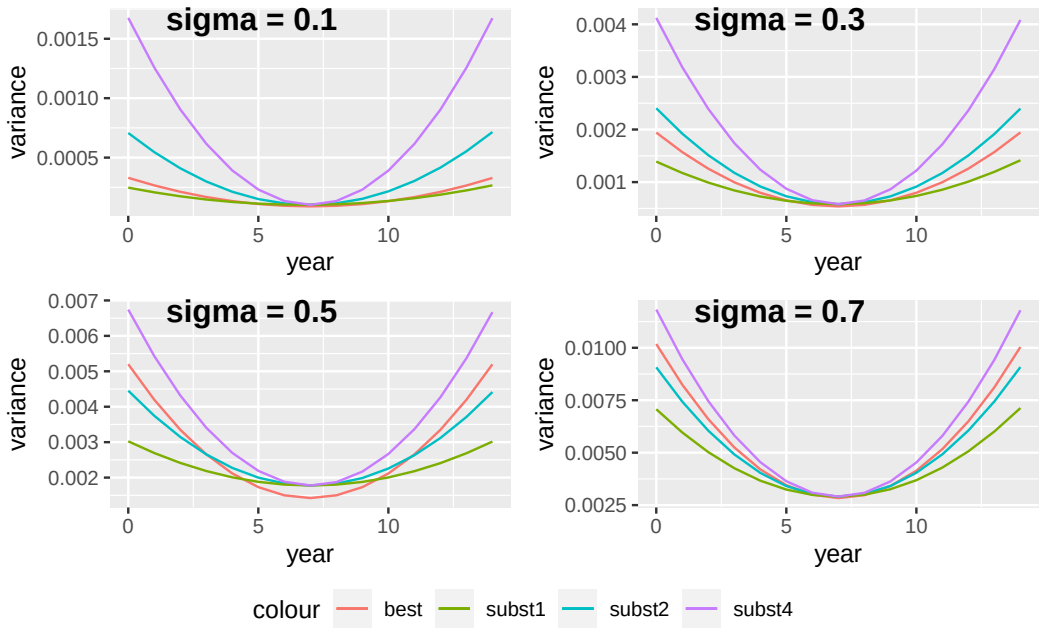

Figure 15: The effect of the value of σ on the variance of predictions from substitution-based approaches

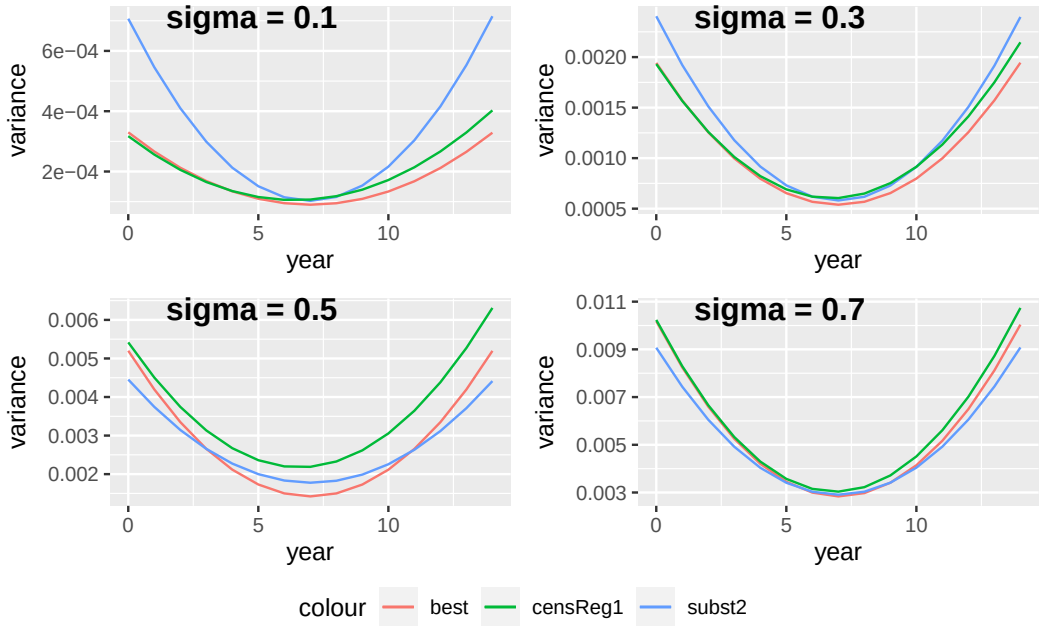

Figure 16: Comparison of the effect of the value of σ on the variance of predictions from `censReg1` and `subst2`

4.2.2 Squared-bias of estimates and predictions

The following table shows the squared-bias of estimates from every approach 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

Table 10: Squared-bias ($\times 10^7$) of estimates for $\sigma = 0.1, 0.3, 0.5, 0.7$

The following three figures show the squared-bias of predictions from imputation-based approaches, from substitution-based approaches, and from `censReg1` and `subst2`, respectively.

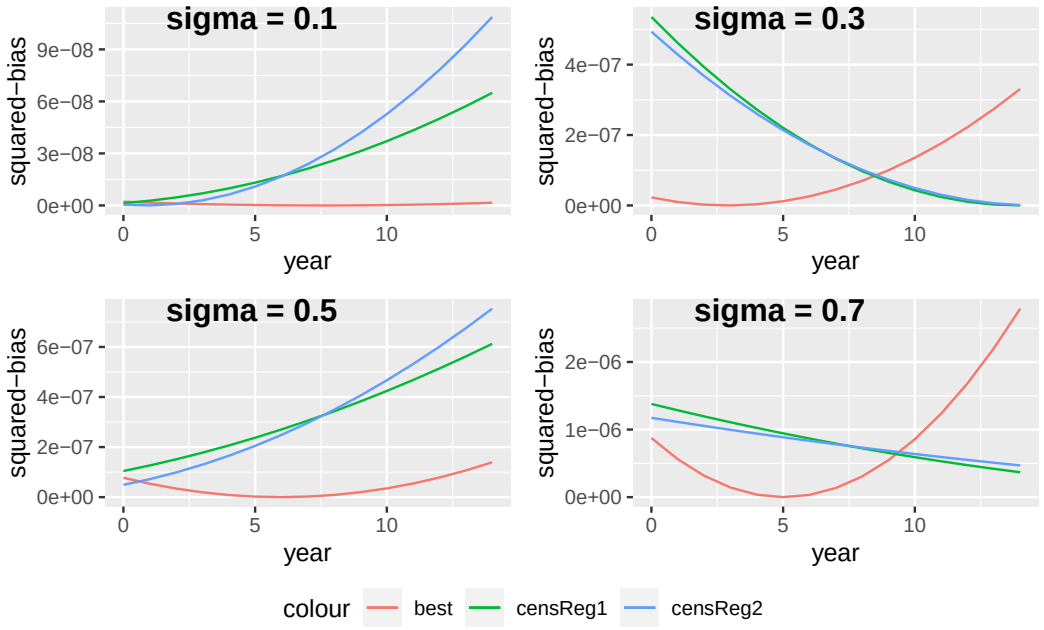

Figure 17: The effect of the value of σ on the squared-bias of predictions from imputation-based approaches

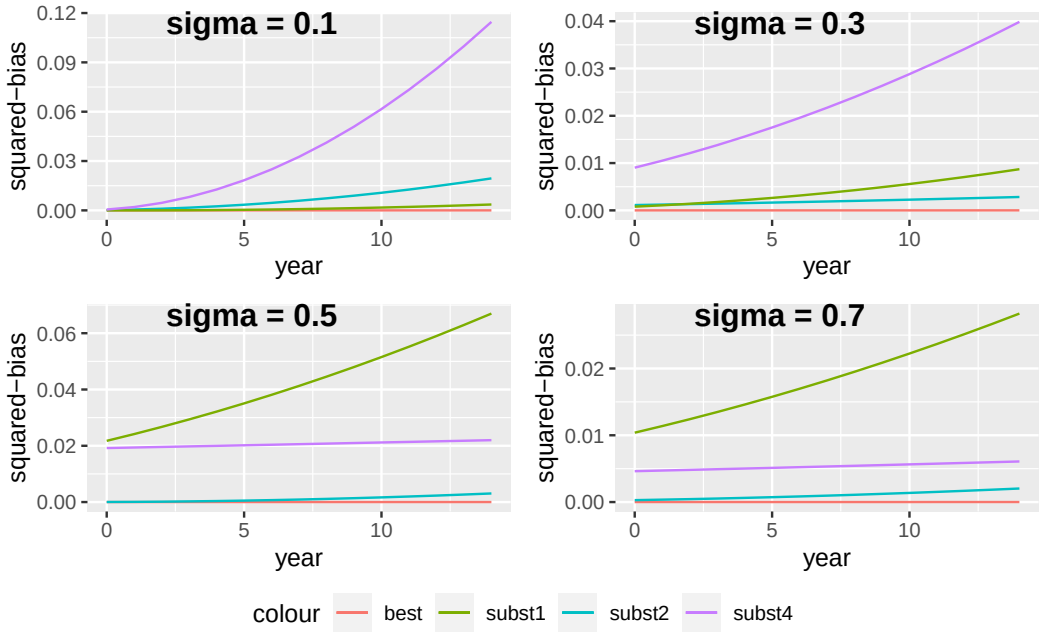

Figure 18: The effect of the value of σ on the squared-bias of predictions from substitution-based approaches

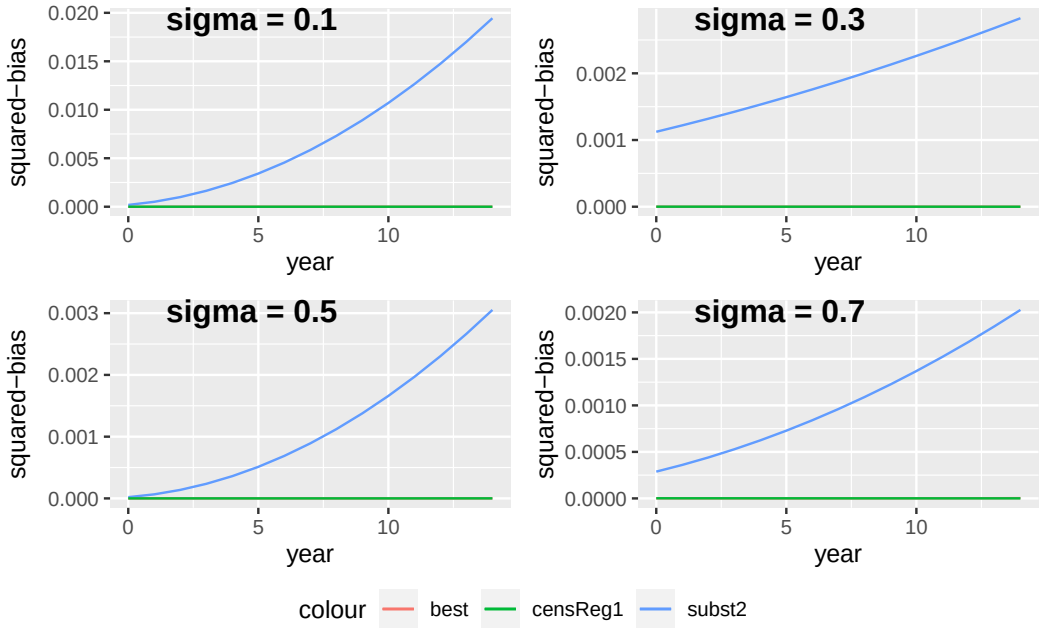

Figure 19: Comparison of the effect of the value of σ on the squared-bias of predictions from `censReg1` and `subst2`

4.2.3 MSE of estimates and predictions

The following table shows the MSE of estimates from `censReg1` and `subst2` for σ equal to $0.1, 0.3, 0.5, 0.7$, respectively.

	0.1	0.3	0.5	0.7
subst1	243.1	392.7	657.1	1078.3
subst2	933.2	383.3	738.4	1298.9
subst4	5423.9	1240.7	1247.0	1820.8
censReg1	51.7	288.9	761.3	1520.4
censReg2	54.8	314.1	831.1	1657.2
censReg0	55.7	314.4	830.9	1657.5
best	48.9	289.4	744.5	1484.3

Table 11: MSE ($\times 10^7$) of estimates for $\sigma = 0.1, 0.3, 0.5, 0.7$

The following figure shows the MSE of predictions from `censReg1` and `subst2`.

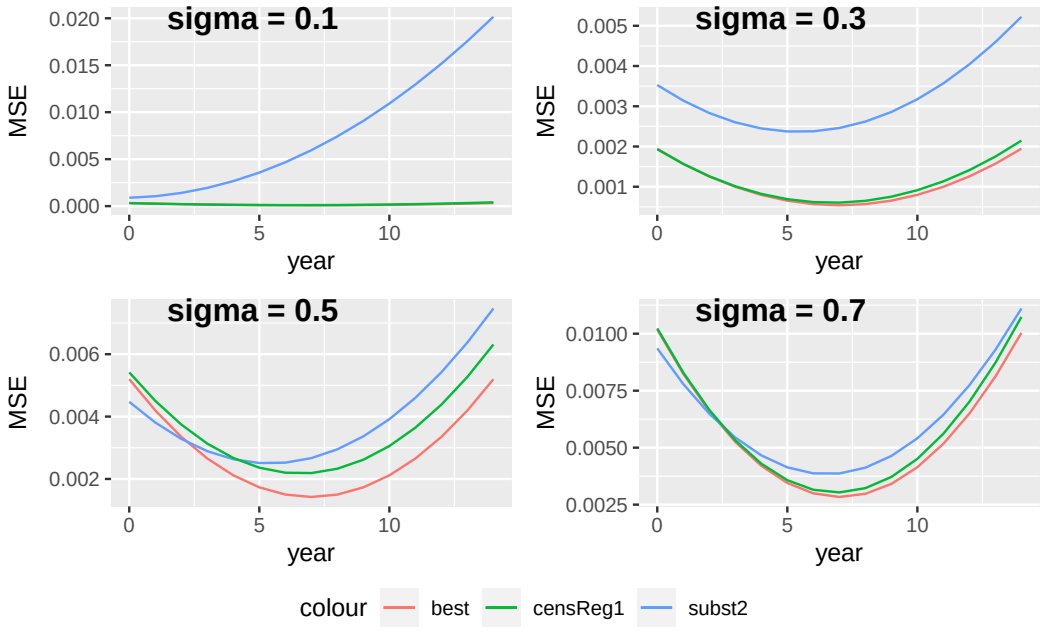

Figure 20: Comparison of the effect of the value of σ on the MSE of predictions from `censReg1` and `subst2`

4.3 Results for various values of `cprop`

For all our simulations in this section, these parameters are fixed: $\sigma = 0.3$, $\beta_A = -0.02$, whilst `cprop` is given four values: 0.1, 0.3, 0.5 and 0.7 respectively. Note that the results for `cprop` $= 0.3$ appear in the previous section; they are duplicated here to allow us to see the effect of the value of `cprop` more easily.

4.3.1 Variance of estimates and predictions

The following table shows the variance of estimates from every approach for `cprop` equal to $0.1, 0.3, 0.5, 0.7$, respectively.

	0.1	0.3	0.5	0.7
subst1	774.3	432.7	305.2	170.3
subst2	1034.0	728.3	683.1	419.9
subst4	1366.9	1172.9	1270.4	820.0
censReg1	960.5	761.3	913.2	954.5
censReg2	978.0	831.1	1157.2	1426.7
censReg0	980.6	831.0	1160.3	1444.6
best	845.0	744.5	845.0	845.0

Table 12: Variance ($\times 10^7$) of estimates for cprop $= 0.1, 0.3, 0.5, 0.7$

The following figure shows the variance of predictions from `censReg1` and `subst2`.

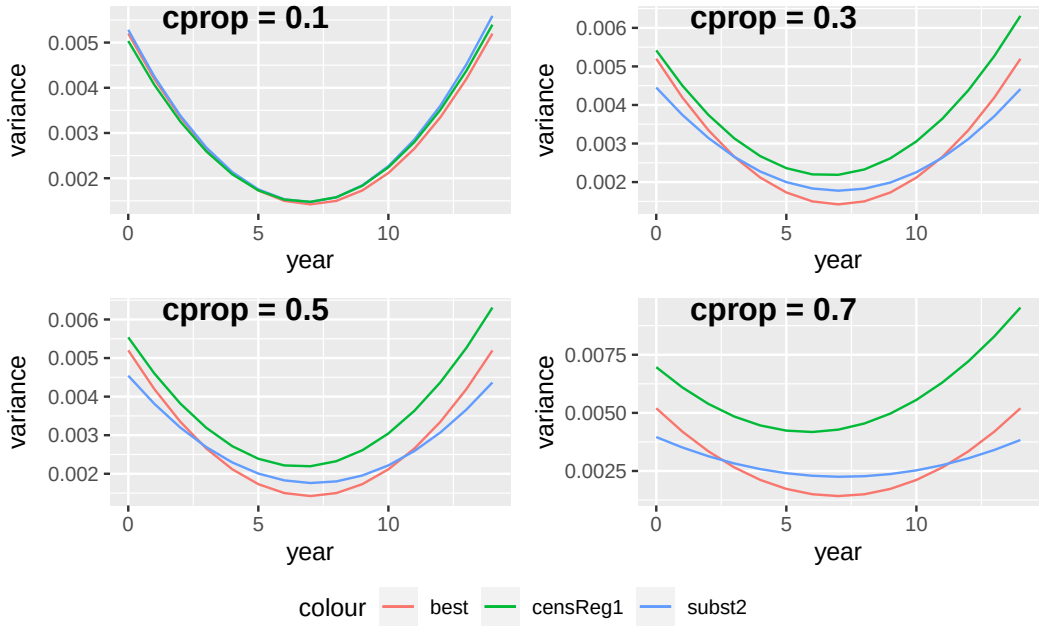

Figure 21: Comparison of the effect of the value of `cprop` on the variance of predictions from `censReg1` and `subst2`

4.3.2 Squared-bias of estimates and predictions

The following table shows the squared-bias of estimates from every approach for `cprop` equal to $0.1, 0.3, 0.5, 0.7$, respectively.

	0.1	0.3	0.5	0.7
subst1	32.87	224.49	661.37	1104.24
subst2	0.02	10.13	128.58	444.91
subst4	29.69	74.25	9.23	80.21
censReg1	1.02	0.01	0.38	14.88
censReg2	1.10	0.01	0.08	15.83
censReg0	1.16	0.01	0.04	15.57
best	0.47	0.02	0.47	0.47

Table 13: Squared-bias ($\times 10^7$) of estimates for cprop $= 0.1, 0.3, 0.5, 0.7$

The following figure shows the squared-bias of predictions from `censReg1` and `subst2`.

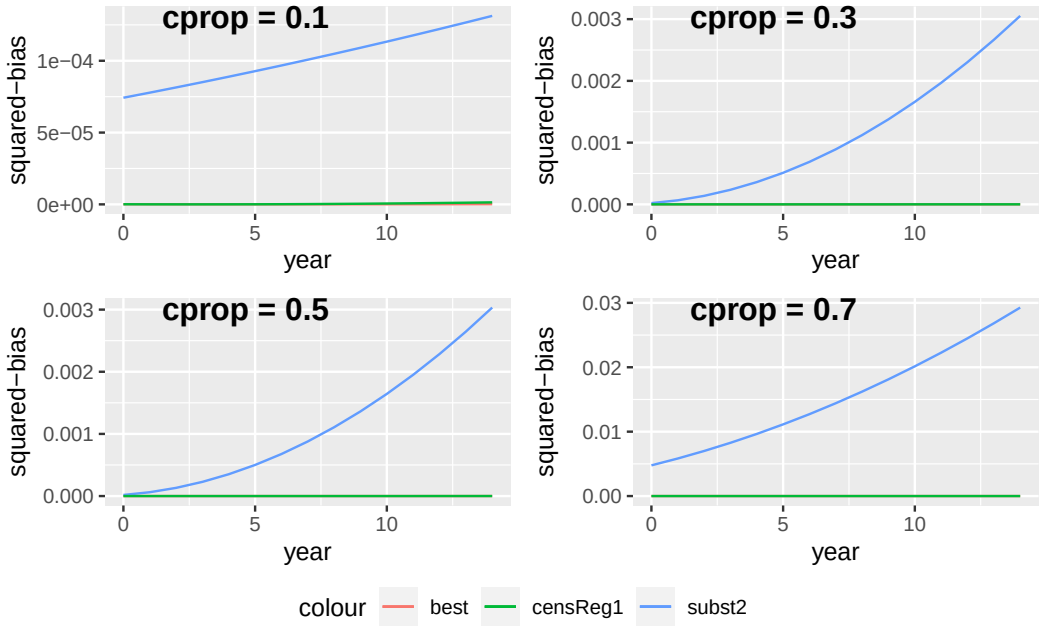

Figure 22: Comparison of the effect of the value of `cprop` on the squared-bias of predictions from `censReg1` and `subst2`

4.3.3 MSE of estimates and predictions

The following table shows the MSE of estimates from every approach for `cprop` equal to $0.1, 0.3, 0.5, 0.7$, respectively.

	0.1	0.3	0.5	0.7
subst1	799.4	657.1	963.5	1272.9
subst2	1023.7	738.4	804.8	860.6
subst4	1382.9	1247.0	1267.0	892.0
censReg1	951.9	761.3	904.5	959.9
censReg2	969.3	831.1	1145.7	1428.3
censReg0	972.0	830.9	1148.7	1445.7
best	837.0	744.5	837.0	837.0

Table 14: MSE ($\times 10^7$) of estimates for cprop $= 0.1, 0.3, 0.5, 0.7$

The following figure shows the MSE of predictions from `censReg1` and `subst2`.

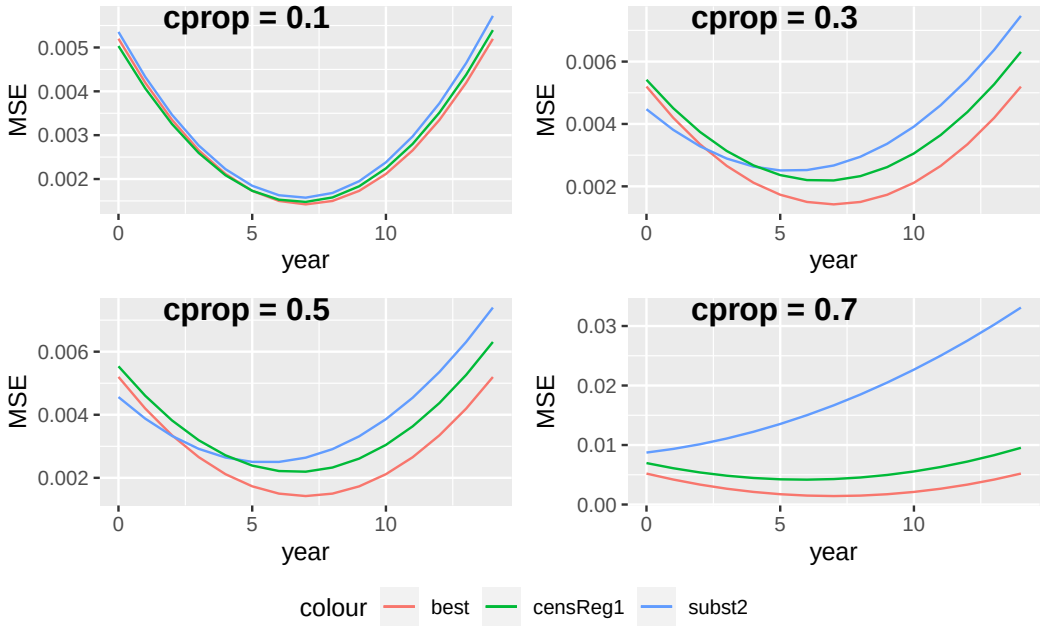

Figure 23: Comparison of the effect of the value of `cprop` on the MSE of predictions from `censReg1` and `subst2`

5 Discussion of results

In Section 5.1, we will give general comments that are true for all (or almost all) ten parameter value sets we used in our main experiments.

In Sections 5.2, 5.3, and 5.4, we will discuss the effect of β_A, σ, and `cprop` on the results, respectively.

We will give our concluding remarks in Section 5.5.

5.1 General comments

A common feature of the graphs showing the variance of predictions (Figures 7-9, 14-16, 21) is that they all 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 variance of predictions from fitted linear regression models.

It is also generally the case that the squared-bias of estimates and predictions from substitution-based approaches is much higher than from imputation-based approaches. This can be seen from Tables 7, 10, and 13, and it really jumps out from Figures 12, 19, and 22, since the curves for `censReg1` in those figures are indistinguishable from the horizontal axis.

A common feature of the graphs showing the squared-bias of predictions from substitution-based approaches (Figures 11, 18, 22) is that they almost all show increasing squared-bias as year increases. The only minor exceptions are seen in Figure 11 for β_A equal to -0.08 and -0.16.

We see from Tables 2 and 4 that the squared-bias of estimates from `omit` is generally very high, which supports our conjecture from Section 1.4.2.1 (that `omit` gives generally high squared-bias). Moreover these results tables also show that the squared-bias from `censReg1naive` is very much higher than from `censReg1`, which supports our conjecture from Section 1.3.2.1 (that `censReg1naive` gives higher squared-bias than `censReg1`).

The following statements regarding the variance of estimates are supported by every relevant table entry of Chapters 3 and 4 (see Tables 1, 3, 6, 9, and 12). The variance of estimates from

- `censReg1` is lowest, out of all imputation-based approaches.
- `censReg0` and `censReg2` is generally very similar.

- `subst1` is lowest, and is highest from `subst4`, out of all substitution-based approaches.

Thus these results support our conjecture from Section 2.2.3 (that `censReg0` gives estimates and predictions with higher variance than `censReg1`). Moreover, `censReg2` gives higher variance than `censReg1`; this makes sense because it agrees with the general principle that a model with more predictors will typically give higher variance. We reason that the information about the correlation between Y and A is carried by the correlation between Y and X encoded by the `censReg1` approach, which renders the additional predictor of `censReg2` redundant for reducing squared-bias, which means that the higher variance from `censReg2` is also reflected in higher MSE.

The last bullet point can be explained by the fact that for all uncensored data $y \geq \text{LLOQ}$, and since it is also the case that $\text{LLOQ} - \log(2) < \text{LLOQ} - \log(\sqrt{2})$, then the gap between the uncensored data and the substituted data is largest for `subst4` and smallest for `subst1`. We reason that this results in highest variance from `subst4` and lowest variance from `subst1`. We also reason that the same logic would also hold for other possible substitution values; the larger the gap between this value and LLOQ, the larger the resulting variance.

5.2 The effect of β_A on results

5.2.1 From imputation-based approaches

We see from Section 4.1.1 that the variance from `censReg1` is slightly lower than from `censReg0` and `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 approaches 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 approaches do, so we would expect the relative performance of `censReg1` to decrease as the value of $|\beta_A|$ increases.

5.2.2 From substitution-based approaches

5.2.2.1 Variance Table 6 shows that the `subst1` approach gives estimates with increasing variance as $|\beta_A|$ increases, whereas the opposite is true for `subst4`. We also see from Figure 8 that the variance of predictions is highest from `subst4` and lowest from `subst1` in general, and that the difference between these decreases as $|\beta_A|$ increases. The squared-bias from all substitution-based

approaches generally increases as $|\beta_A|$ increases, and as year increases. This all makes sense since we are using a constant $LLOQ$ 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 `subst1` gives predictions with increasing variance as year increases, whereas the opposite is true for `subst4` by the same reasoning.

5.2.2.2 Squared-bias and MSE The `subst1` approach is designed as a reference that gives biased estimates, since it substitutes y values that are observed to be below $LLOQ$ with the $LLOQ$ value itself, so the substituted values will never be smaller than the unknown true y values. From the same reasoning as is given in Section 5.2.2.1, we expect the squared-bias from `subst1` to increase as $|\beta_A|$ increases, which is precisely what these results show. In contrast, the squared-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 $\text{LOQ} - \log(2)$ is lower 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 squared-bias from `subst2` is much lower than that from `subst1` or `subst4`, which suggests that the true values of the censored data mostly lie between $LLOQ$ and $\text{LOQ} - \log(2)$ for these sets of parameter values. From the results for predictions from substitution-based approaches, we see the same trend that we saw for the estimates, i.e. the relative amount of squared-bias from `subst1` increases as $|\beta_A|$ increases, whereas the opposite is true for `subst4`.

5.3 The effect of σ on results

We first compare the squared-bias of estimates from our three substitution-based approaches (see Table 10). The squared-bias from `subst4` decreases greatly as the value of σ increases, whereas the squared-bias from `subst1` is relatively independent of the value of σ. The squared-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 increases for $\sigma = 0.7$. Our interpretation is that this can be attributed to the fact that the censored y values lie closer on average to $LLOQ$ for smaller values of σ, and further away

for larger values. The low squared-bias from `subst4` for $\sigma = 0.7$ indicates that the true y values for the censored data lie close to $\text{LOQ} - \log(2)$ on average for this parameter value. These results are consistent with our conjecture from Section 1.4.2.3.

We already know that the squared-bias from the imputation-based approaches is much lower than from substitution-based approaches. Moreover, the variance from the imputation-based approaches is much higher than the corresponding squared-bias, which means that any statements we make about variance also apply to MSE, and vice versa. We see from Table 10 that all three imputation-based approaches gave lower MSE of estimates than all three substitution-based approaches for both $\sigma = 0.1$ and $\sigma = 0.3$. However this is not the case for larger values of σ, which we attribute to the fact that the corresponding variance from imputation-based approaches increases greatly as σ increases.

Of the substitution-based approaches, `subst1` gave lowest MSE of estimates 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 of estimates than `best` for $\sigma = 0.5$ and $\sigma = 0.7$ because the lower variance from this approach more than compensated for the higher squared-bias.

Our conjecture from Section 1.4.2.3 (that imputation-based approaches will perform decreasingly well for increasing σ values) is clearly supported by the results from Sections 4.2.1 and 4.2.3. More specifically, these approaches all give estimates with low squared-bias for all σ values we used, but the variance increases greatly as σ increases.

5.4 The effect of `cprop` on results

Recall that these results for predictions were reported exclusively for the `censReg1` and `subst1` approaches, whereas for estimates the results from all approaches were reported. Moreover, we have already made general comments about variance, and about the lower squared bias from imputation-based than from substitution-based approaches in Section 5.1.

We therefore focus on discussing the results from Section 4.3 in relation to our conjecture from Section 1.4.2.2 (that the best substitution-based approach depends on the value of `cprop`). Of the substitution-based approaches, the squared-bias of estimates from `subst2` is lowest for `cprop` equal to 0.1 and 0.3, whereas it is lowest from `subst4` for larger values of `cprop`; moreover, it increases greatly as `cprop` increases, from `subst1` and `subst2`. These results

broadly follow the same pattern that is illustrated in Figure 6, and are consistent with our conjecture.

5.5 Concluding remarks

Our simulation studies were designed on the basis of our EDA from the SNMPC datasets. All of our results show that imputation-based approaches generally give much lower squared-bias than substitution-based approaches.

However, we see from Table 11 that as σ increases, which corresponds to decreasing strength of correlation between Y and X, the variance (and thus MSE) from imputation-based approaches increases steeply. This exemplifies the purpose of imputation: to use information from the correlation of X and Y to impute (estimate) censored y values from fully observed x values. Larger σ values mean that this information is more noisy, which results in higher variance.

The models fitted to our test dataset in Section 1.2, and represented by (1) and (2), both have residual standard error equal to 0.1. We therefore reason that the correlation of X and Y is sufficiently strong for imputation by censored regression to be the method of choice for manipulation of censored SNMPC data. If it would have been the case that the SNMPC datasets used a single LOQ value, then our results indicate that SNMPC could benefit from changing from their substitution-based approach to an imputation-based approach. However since the SNMPC datasets use multiple LOQ values, our findings are too limited in scope to be directly applicable for SNMPC. Our findings nonetheless constitute a solid foundation upon which a more nuanced understanding can be developed through further investigation.

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.

Henningsen, Arne. 2012. "Estimating Censored Regression Models in R Using the censReg Package." In.

Henningsen, Arne, and Ott Toomet. 2011. "maxLik: A Package for Maximum Likelihood Estimation in R." *Computational Statistics* 26 (3): 443–58. https://doi.org/10.1007/s00180-010-0217-1.

Vanden Bilcke, C. 2002. "The Stockholm Convention on Persistent Organic Pollutants." *Review of European Community & International Environmental Law* 11 (3): 328–42. https://doi.org/10.1111/1467-9388.00331.

Weisstein, Eric W. 2020. "Half-Normal Distribution." *Mathworld - A Wolfram Web Resource*, June.