Methods for processing censored data FV6

Marc Roddis

7/17/2020

Chapter 1: Introduction

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

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

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

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

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

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

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

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

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

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

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

Our workflow

We now present an eight-step overview of our workflow in this section. We will give a more detailed description of steps 1-5 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 simulation) dataset using censored (instead of simple linear) regression. We call this method `censReg0impute` because it comprises censored regression without imputation.

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

6. Discussion of all results.

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

Model design and specification

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

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

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

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

Our exploratory data analysis showed that

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

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

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

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

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

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

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

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

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

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

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

the corresponding fitted model for Y on X is

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

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

Generation of uncensored datasets

Following the aforementioned model, for every iteration we generate our uncensored datasets as follows:

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

2. For every year, we will generate the same number of observations for Y and X, we will call this number the sample size N.

3. We generate all x_i from

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

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

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

4. We generate all y_i from

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

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

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.

4

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

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

Generation of incomplete datasets

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

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.

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 LOD)

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 LOD) 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.

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

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

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

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

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

To specify this model, we first substitute

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

nto

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

which gives

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

which we can write as

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

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

The probability that y_i^* is censored equals

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

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

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

This gives us the likelihood function L

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

So the log-likelihood function is

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

which equals

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

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

We will then perform imputation as follows.

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

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

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

Variations on the `censReg1` method

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

The `censReg1naive` method

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

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

The `censReg2` method

`censReg2` uses two predictor variables X and N.

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

10

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

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

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

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

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

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

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

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

This method is designed to test our conjecture that since $|\beta_X| = 0.79$ is much greater than $|\beta_A|$, `censReg0impute` 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

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

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 $a \in \{0, 1, 2, ..., 14\}$.

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 LOD, $\frac{LOD}{\sqrt{2}}$ or $\frac{LOD}{2}$ 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 `censReg0impute`.

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

Chapter 2: Results for the estimation of $\hat{\beta}$

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.

Evaluation of methods for smaller sample sizes

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.

The MSE, squared-bias, and variance of the estimates of β from all three substitution methods `subst1, subst2, subst4`, all 3 imputation by censored regression methods `censReg1, censReg1naive, censReg2`, the `censReg0impute` 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).

```
##              mse_beta bias_beta variance_beta
## omit          673.529   626.334        47.242
## subst2        102.097    16.359        85.824
## subst1        263.096   224.608        38.527
## censReg1       67.592     0.015        67.645
## censReg2       73.239     0.016        73.296
```

```
## censReg0impute   73.307      0.020         73.360
## best              71.821      0.119         71.774
## subst4           698.697    532.509        166.354
## censReg1naive    113.477     69.481         44.040

##                 mse_beta bias_beta variance_beta
## omit             732.619   630.801       101.828
## subst2           192.885    18.133       174.770
## subst1           300.110   221.388        78.729
## censReg1         137.640     0.011       137.642
## censReg2         149.855     0.008       149.862
## censReg0impute   149.934     0.008       149.941
## best             136.221     0.120       136.114
## subst4           886.321   547.356       339.000
## censReg1naive    154.927    66.680        88.256

##                 mse_beta bias_beta variance_beta
## omit             844.831   625.923       218.930
## subst2           391.093    19.689       371.441
## subst1           385.546   217.862       167.700
## censReg1         292.659     0.028       292.661
## censReg2         317.724     0.023       317.732
## censReg0impute   317.923     0.026       317.929
## best             286.629     0.027       286.631
## subst4          1277.266   558.593       718.745
## censReg1naive    255.419    67.152       188.286

##                 mse_beta bias_beta variance_beta
## omit            1099.143   628.979       470.211
## subst2           761.409    15.935       745.548
## subst1           563.409   227.112       336.331
## censReg1         585.494     0.088       585.465
## censReg2         636.029     0.125       635.968
## censReg0impute   636.792     0.127       636.728
## best             565.520     0.163       565.414
## subst4          1973.593   531.486      1442.251
## censReg1naive    445.252    70.285       375.005
```

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

14

```
##              bias_ss50 bias_ss25 bias_ss12 bias_ss6
## omit          626.3345  630.8013  625.9227 628.9789
## subst2         16.3590   18.1328   19.6888  15.9350
## subst1        224.6076  221.3881  217.8623 227.1120
## censReg1        0.0149    0.0113    0.0275   0.0875
## censReg2        0.0157    0.0077    0.0231   0.1246
## censReg0impute  0.0204    0.0081    0.0257   0.1274
## best            0.1186    0.1202    0.0269   0.1626
## subst4        532.5090  547.3556  558.5935 531.4857
## censReg1naive   69.4810   66.6803   67.1517  70.2850
```

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

```
##              variance_ss50 variance_ss25 variance_ss12 variance_ss6
## 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
```

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.

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 `beta28year` for four sets of parameter values:

$\beta_A = -0.02$ is held fixed.

a "low" and a "high" value for each of *cprop* and σ are used. Concretely: $\{(0.1, 0.1), (0.7, 0.1), (0.1, 0.5), (0.7, 0.5)\}$ were used for $\{(cprop, \sigma)\}$ respectively.

The following four tables show the MSE, squared-bias, and variance of estimates of β from all methods, for the four sets of parameter values, respectively.

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 `best` serves as a reference value; a gold standard that we compare the other methods with.

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 expectations (ref: Helsel's book) that `omit` is a poor method, so we will not study this method further.

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.

from `censReg1`, `censReg2`, and `censReg0impute` is very low for all four parameter value sets.

from `censReg1naive` bias is significantly higher than from `censReg1`, `censReg2`, and `censReg0imp` for all four parameter sets . This illustrates the necessity of conditioning on both X value and the condition $Y \leq LOD$ by using a truncated normal distribution, and verifies the conjecture we presented in our previous chapter on mathematical theory.

Whereas `censReg1 and censReg2` all do this, in contrast `censReg1naive` conditions solely on the X value, and thus uses a (non-truncated) normal distribution; this results in significant bias because Y values can be erroneously imputed to be higher than LOD. Consequently we will not discuss `censReg1naive` any further since we have already 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 `subst4` , and the three censReg methods `censReg1, censReg2,` and `censReg0impute`. We will use `best` as our reference method throughout.

```
##                mse_beta bias_beta variance_beta
## omit            1.94573   1.50568       0.44049
## subst2          3.28387   2.47542       0.80926
## subst1          0.62231   0.19495       0.42779
## censReg1        0.48293   0.00001       0.48340
## censReg2        0.48707   0.00000       0.48756
## censReg0impute  0.48612   0.00000       0.48661
## best            0.49985   0.00002       0.50033
## subst4         14.44620  12.87532       1.57245
## censReg1naive   0.74552   0.31287       0.43308

##                mse_beta bias_beta variance_beta
## omit           14.13397  13.45675       0.67790
## subst2          1.49817   0.62596       0.87308
## subst1         12.62066  12.52761       0.09314
## censReg1        0.74680   0.00039       0.74716
## censReg2        0.96849   0.00129       0.96816
## censReg0impute  1.02894   0.00175       1.02822
## best            0.49985   0.00002       0.50033
## subst4         28.93460  26.23278       2.70452
## censReg1naive   1.03778   0.25186       0.78671
```

```
##                mse_beta bias_beta variance_beta
## omit            8.16691   2.32836       5.84439
## subst2          6.97480   0.01114       6.97063
## subst1          6.08940   0.44654       5.64852
## censReg1        6.64952   0.03803       6.61811
## censReg2        6.76917   0.03933       6.73658
## censReg0impute  6.76973   0.03983       6.73664
## best            7.76117   0.01682       7.75210
## subst4          9.06100   0.20899       8.86088
## censReg1naive   6.55698   1.37954       5.18262


##                mse_beta bias_beta variance_beta
## omit           21.18031  12.48679       8.70222
## subst2          8.37304   5.15594       3.22032
## subst1         13.19163  11.98684       1.20600
## censReg1        7.66589   0.00451       7.66905
## censReg2       11.34171   0.00386      11.34921
## censReg0impute 11.34744   0.00429      11.35451
## best            7.76117   0.01682       7.75210
## subst4          7.81905   1.16453       6.66118
## censReg1naive   9.35157   0.90675       8.45328
```

Evaluation of methods for larger absolute values of β_A

We will now focus our six chosen methods `subst1, subst2, subst4, censReg1, censReg2, and`
We will use these methods to estimate β from four simulations that each run
for 10000 iterations, using these same parameter values $\{cprop = 0.3, \sigma = 0.3\}$
as before. We will use the four parameter values $\{-0.02, -0.04, -0.08, -0.16\}$
for β_A in these four simulations, respectively. The results from these
simulations are shown in the four tables below.

We see that for these parameters value sets, censReg method 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 LOD with the LOD value itself, so the substituted
values will never be smaller than the real values. Moreover, we have chosen
to maintain a constant LOD level for all years of the same dataset. We are
also always using the fixed value $\alpha_A = -2.91$, and a variable but always
negative parameter value for β_A. This means that the Y values decrease faster
with years for larger values of $|\beta_A|$. This all means that it is an inevitable

18

consequence of our design that the bias from `subst1` increases 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 since `subst4` substitutes censored values with $\frac{LOD}{2}$, which are lower than the true values on average for low values of $|\beta_A|$ but not lower for the highest value $|\beta_A| = 0.16$. This 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 LOD and $\frac{LOD}{2}$.

Our three censored regression methods all give very similar results to one another. The values of MSE, squared-bias and variance are very similar from all these methods for both $|\beta_A| = 0.08$ and $|\beta_A| = 0.16$. However, for the lowest value $|\beta_A| = 0.02$, the variance from `censReg1` is approximately 10% lower than from `censReg2`, which is a statistically significant difference. This fits with our prior knowledge that a more complex model generally has higher variance than the corresponding less complex one. Moreover, we also expected that the estimates from `censReg2` would improve relative to those from `censReg1` as the value of $|\beta_A|$ increases, since the difference between these two methods is that `censReg2` uses A as additional predictor variable. However, the fact that `censReg0impute` has a higher variance than `censReg1` seems puzzling in this respect, so perhaps our interpretation of model complexity is wrong is this context.

Our prior expectation was that `censReg0impute` would perform relatively worse compared to the other censored regression methods for larger values of $|\beta_A|$. This is because we conjecture that the imputations from the predictor variables carry more information about Y as $|\beta_A|$ increases, and `censReg0impute` does not use imputation at all. However, these results fail to support our conjecture here too.

If we now compare the best performing models from each category, i.e. `subst2` and `censReg1`, we see that `censReg1` has much lower bias for all parameter values. However, MSE for `subst2` is lower for one of the values, $|\beta_A| = 0.08$. In conclusion, we can say that `censReg1` gives better estimates than `subst2` for most, but not necessarily all, values of $|\beta_A|$.

Evaluation of methods for other values of σ

We will now hold β_A and *cprop* fixed at their original values (-0.02 and 0.3) and investigate the effect of larger σ values, specifically: 0.1, 0.3, 0.5, and 0.7.

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 LOD 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 $\frac{LOD}{2}$ on average for this parameter value.

The large gap between the uncensored Y data and the $\frac{LOD}{2}$ value means that `subst4` gives higher variance than all other methods for all values of σ. Similarly, the smallest possible gap between LOD 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 LOD, 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. Surprisingly the results from `censReg2` and `censReg0impute` are almost identical. Is this a bug?

In conclusion, substitution methods give much higher bias than censored regression methods. Moreover, all three censored regression methods gave lower MSE than all three substitution methods for both $\sigma = 0.1$ and $\sigma = 0.3$. 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.

Further comparisons between `subst2` and `censReg1`

From our previous results, `subst2` is generally the best performing substitution method and `censReg1` is the best censReg method. In the previous section,

these methods gave 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 LOD, which is our variable of primary interest.

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 LOD is lowered, which means that a higher proportion are likely to lie closer to LOD 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 LOD 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.

Chapter 3: Results for the prediction of the annual means $E(Y|A = a)$

The MSE, squared-bias and variance of predictions of Y annual means from various censoring methods

All the graphs in this section will shows MSE, squared-bias, or variance on the y-axis and year on the x-axis for the simulated 15-year period. We begin by looking at variance of predictions from our best three substitution methods and our best three censored regression methods. We will again use `best` as our gold standard.

Variance of predictions of Y annual means from different methods

Predictions for different values of β_A

We will begin by using the same parameter values we used in our earlier section "Evaluation of methods for larger absolute values of `beta28year`". These parameters are fixed: $cprop = 0.3$, $\sigma = 0.3$, whilst β_A is given four values: -0.02, -0.04, -0.08 and -0.16 respectively.

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 β_A equal to -0.02, -0.04, -0.08, -0.16, respectively.

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.

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.

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 β_A equal to -0.02, -0.04, -0.08, -0.16, respectively.

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.

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.

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

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

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

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

Predictions for different values of sd28vs153

For all our predictions 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.

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.

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.

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

Our first set of four graphs show the MSE 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 MSE 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 MSE from the `subst2`, `censReg1` and `best` methods together on the same plot, which is displayed below.

Predictions for different values of *cprop*

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

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

Our first set of four graphs show the variance of `censReg1` and `censReg2` methods relative to `best` method for *cprop* 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 *cprop* 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.

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 *cprop* 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 *cprop* 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.

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

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

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

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

Predictions for low-low, high-low, low-high, high-high values of $\{\sigma, cprop\}$

We will now use the same sets of parameter values that we used in our earlier section "Selection of censoring methods for further study". Concretely: $\beta_A = -0.02$ is held fixed, whilst a "low" and a "high" value for each of *cprop* and σ are used. Concretely: $\{(0.1, 0.1), (0.7, 0.1), (0.1, 0.5), (0.7, 0.5)\}$ were used for $\{cprop, \sigma\}$ respectively.

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

Our first set of four graphs show the variance of `censReg1` and `censReg2` methods relative to `best` method for $(\sigma, cprop)$ equal to (0.1, 0.1), (0.1, 0.7), (0.7, 0.1) and (0.7, 0.7), respectively.

Our second set of four graphs show the variance of `subst1`, `subst2` and `subst4` methods relative to `best` method for $(\sigma, cprop)$ equal to (0.1, 0.1), (0.1, 0.7), (0.7, 0.1) and (0.7, 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.

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, *cprop*) equal to (0.1, 0.1), (0.1, 0.7), (0.7, 0.1) and (0.7, 0.7), respectively.

Our second set of four graphs show the bias of `subst1`, `subst2` and `subst4` methods relative to `best` method for (σ, *cprop*) equal to (0.1, 0.1), (0.1, 0.7), (0.7, 0.1) and (0.7, 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.

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

Our first set of four graphs show the MSE of `censReg1` and `censReg2` methods relative to `best` method for (σ, *cprop*) equal to (0.1, 0.1), (0.1, 0.7), (0.7, 0.1) and (0.7, 0.7), respectively.

Our second set of four graphs show the MSE of `subst1`, `subst2` and `subst4` methods relative to `best` method for (σ, *cprop*) equal to (0.1, 0.1), (0.1, 0.7), (0.7, 0.1) and (0.7, 0.7), respectively.

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