Independent Study of Robust estimation in accelerated failure time models

Korede Adegboye

400014008

STATS 756 - Topics in Biostatistics

December 10, 2021

Introduction

When considering lifetime data with covariates, information (characteristics) is given by these covariates that affect an individual's lifetime. The interest of a subject's survival time may include characteristics such as age, sex, the severity of disease, smoking status, blood test results, etc. Analogous to other statistical models, covariates can be quantitative or categorical. With the covariates introduced to the statistical model, it is expected that lifetimes can be produced effectively, and thus, efficient analysis of the study can take place. In most cases, the goal is to introduce covariates to parametric models. Where through analysis, it can be verified that the assumed distribution(s) are reliable for the set of covariates given. For example, the time for a subject to get stage 4 cancer (event) given their covariates could be of interest

Incorporating covariates into the traditional linear regression model is one of several ways. Here we have that the residuals are normally distributed with mean zero and variance σ^2 and that the response variable is normally distributed with mean μ and variance σ^2. Considering the Weibull distribution, there are 2 forms in which a regression model can be formed. Start by noting the pdf of the Weibull distribution as

$$S(t) = e^{-(\frac{t}{\alpha})^{\eta}} \tag{1}$$

The first form derived from the Weibull distribution has that scale parameter α depending on x. Other considerations stem from introducing the covariates such that it acts on different characteristics of a model,

say η. As a result, it produces a different model.

$$S(t) = e^{-(\frac{t}{\alpha(x)})^{\eta}}$$

To get the second form, let $\alpha(x) = e^{\beta' x}$. Then we have the survival and hazard function, respectively as

$$S(t;x) = e^{-(te^{-\beta' x})^{\eta}} \tag{2}$$

$$h(t;x) = \eta t^{\eta-1} e^{(\eta\beta)' x} \tag{3}$$

Many of the robust developments and the one popular for survival analysis among those in biomedical sciences make use of the semi-parametric Cox's proportional hazards model. In equation (1), the perspective used resulted in proportional hazards. Now of consideration is that the defining property is proportional hazards. This affects how the covariates are also introduced. Thus, this gives a regression model for $lnh.$. Instead of the proportional hazards function, use the proportional odds model to get a linear model for log odds. In other words, the result is logistic regression. Hence, the effect of the covariates, x is to accelerate or decelerate the time scale t as $(te^{\beta' x})$, giving the accelerated failure lifetime model.

$$S(t;x) = e^{-(te^{\beta' x})^{\eta}}$$

When the goal of the study is to produce an analysis based on regression models for lifetime data with covariates, we can look to apply the accelerated lifetime model. Although the basis of using the accelerated lifetime model may be satisfied, formal documentation may be needed to access if the proposed distributions fit the data well. The accelerated failure time model describes a situation where the subject's history of an event is accelerated. This subject can be of a biological or mechanical nature. It is popular in practice because it supports various survival models. Some of those being, from the Weibull, log-normal and log-logistic distributions. It is usually chosen in contrast to proportional hazards models where a distribution may be useful to the model. Its use requires the assumption of a probability distribution, and to assume such an application one must access which distributions fit the data best. With that it requires precise specifications, otherwise, an insufficient model can be expected. With covariates being the main component of accelerated lifetime models, accelerated lifetime regresses the logarithm of these covariates. Many find its use to be effective when analyzing censored survival data.

The motivation of this paper stems from the sensitivity of outliers when modeling. Like traditional

regression, outliers are of importance to survival analysis. They tend to be more sensitive to the most influential observations. With that, the goal of the paper is to construct a method where outliers are rejected or down-weighted.

The main problem that is being raised in this study is that outliers in Cox regression models result in sensitive estimators. Note outliers tend to tell truth about the data, and it may be necessary to investigate what may have happened when the data was recorded. Nevertheless, investigations could be lengthy or inconclusive so proceeding with robust estimation via a robust maximum likelihood estimator should be feasible so long as its use gives a model where outliers have been detected and are given less weight. In the past 30 years, there has been extensive development regarding robust methods that utilized the semi-parametric Cox proportional hazards model. Generally speaking, these methods were robust against outliers in survival times but not against potential outliers in covariates. Dr. Sinha proposes the use of a parametric accelerated failure time model. To produce a parametric accelerated failure time model that can handle outliers in 2 settings – survival outcome variable and covariates. His proposed solution involves robust estimation of the parameters of interest using a Huber-type psi function and sandwich-like variance.

To build on this motivation, an analogy can be provided. There exists photographer(s) that are interested in capturing (studying) the nature of penguins. Penguins have their own goals and live in such a way that they aim to prolong their survival. With photographers around, they are easily distracted and no longer attend to their necessary routines. In such a way, this is how outliers are influencing the analysis of failure lifetime models. Humans, foreign to the environment are heavily influencing the penguins. It is this study's goal to find a way for the penguins to continue their everyday activities without viewing humans, loud noises, and flashes of light as a reason to reduce their efforts of survival. Note, this analogy does not encompass the entirety of the issue(s) that are being faced with influential outliers but brings the problem into terms that can be easily understood by most.

Model, notation, and method

The log-linear form of the accelerated failiure time model

Accelerated failure time models include exponential, Weibull, log-normal, and log-logistic distributions. The purpose for using an accelerated lifetime model stems from the assumption that the covariates on a subject are accelerated (tends to multiply) in regards to time rather than hazards (proportional hazards model). In other words, accelerated lifetimes can be applied when interested in analyzing the speed of progression of a disease over time. Equation (1) gives the general accelerated failure time model. Here we start by

understanding the survival and hazard functions for the ith individual. Note, $i = 1, \ldots, n$ individuals and $j = 1, \ldots p$ covariates for the ith individual, the baseline survivor function as $S_0 * (t)$, and baseline hazard function at time t can be notated as $h_0 * (t)$. Hence, we have

$$S_i(t) = S_0(\frac{t}{\alpha_1 x_{1i} + \cdots + \alpha_p x_{pi}}) = S(t; x) = e^{-(te^{-\beta' x})^{\eta}}$$

$$h_i(t) = e^{\alpha_1 x_{1i} + \cdots + \alpha_p x_{pi}} h_0(\frac{t}{\alpha_1 x_{1i} + \cdots + \alpha_p x_{pi}}) = h(t; x) = \eta t^{\eta - 1} e^{(\eta\beta)' x}$$

As previously discussed, the accelerated lifetime models can be parameterized in various ways. Here we consider a log-linear form, which results in a form similar to that of linear regression. It can be denoted.

$$lnt_i = \mu + \beta' x + \sigma\epsilon_i \tag{4}$$

Producing the respective survivor and hazard functions

$$S_i(t) = S_{\epsilon_i}(\frac{lnt - \mu - \eta_i}{\sigma})$$

$$h_i(t) = \frac{1}{\sigma t} h_{\epsilon_i}(\frac{lnt - \mu - \eta_i}{\sigma})$$

It follows that the accelerated lifetime model has a double property as Weibill also follows a proportional hazards distribution. The other models that can be considered come from log-normal and log-logistic distributions.

Robust Estimation

Robust estimation can be defined as a method that is not influenced by slight differences in the proposed assumptions (ie., probability distribution). As a consequence, it optimizes the algorithm's search for a robust estimator. In other words, the results obtained are deemed to be reliable given these slight variations. In general, maximum likelihood estimates, linear combinations, and statistical rank tests can aid in achieving robust estimation. Of interest is estimating the model parameters μ, the row vector of scale $\alpha = \alpha_1, \ldots \alpha_p$, and σ. To do this suitable robust method via maximum likelihood estimation is required. One that is resistant to potential outliers in both survival times and covariates in the data.

Estimations of parameters can be derived via the standard maximum likelihood method. The i-th sigma denotes the censoring status of the i-th survival time. From here, inference can proceed as follows from the

maximum likelihood theory. It is the likelihood score that is a function of error terms and covariates that are unbounded for both the survival time ti and covariates x_i. It has been discovered that this construction is not robust to outliers (sensitive to outliers). Therefore the motivation of this literature comes from this issue. Dr. Sinha proposes that the solution for this includes the Huber-type monotone psi function on the error terms to bound the influence of outliers. (Huber PJ 1981) has developed such a function that is also encompassed by M-estimators. Where these are estimators under the maximum likelihood framework. It is a technique in regression modeling that is robust to outliers. It has a redescending property that enables it to reject unusual observations. Redescending can be interpreted as never decreasing towards the origin and descending towards zero.

Detecting outliers is of interest before the proposed robust method is applied. It will be discussed later, that if outliers do not exist it is best to use the traditional maximum likelihood method. (Pinto JD, Carvalho AM, Vinga S 2015) have researched the detection of outliers in Cox proportional hazard models based on the concordance c-index and named the method Dual Bootstrap Hypothesis Testing (DBHT). The c-index is analogous to the receiver operating characteristic (ROC) curve, where interest is in the sensitivity and specificity of a model. When outliers are present the c-index lowers and thus suggests the model is sensitive.

Dual Bootstrap Hypothesis Testing is an improvement on Bootstrap Hypothesis Testing (BHT). BHT is a method that removes a single observation from the data set and analyzes the c-index to signal outliers (and inliers) in a Cox proportional hazards model. It has been discovered that BHT increases the number of outliers, so DBHT aims to provide a solution. Note that, bootstrapping is a resampling technique that aims to uncover the true distribution given by the data (rephrase a little). DBHT takes two bootstrap samples, where the 'poison' sample comes from the original data set and the 'antidote' is taken where one observation is removed from the original data set. Both samples are then compared via p-values. This is a test to see if an observation is "poison" (outlier). The test between the two samples gives a null hypothesis that the expected c-index of the "antidote" sample is greater than that of the "poison" sample. Whereas the alternative hypothesis would be that the expected c-index of the "antidote" sample is less than or equal to that of the "poison" sample. The p-values are calculated via the two-sample t-test (Welch's t-test) for unequal variances. Such a method requires high computing power.

```
library("survBootOutliers")
```

Example - Dual Bootstrap Hypothesis Testing

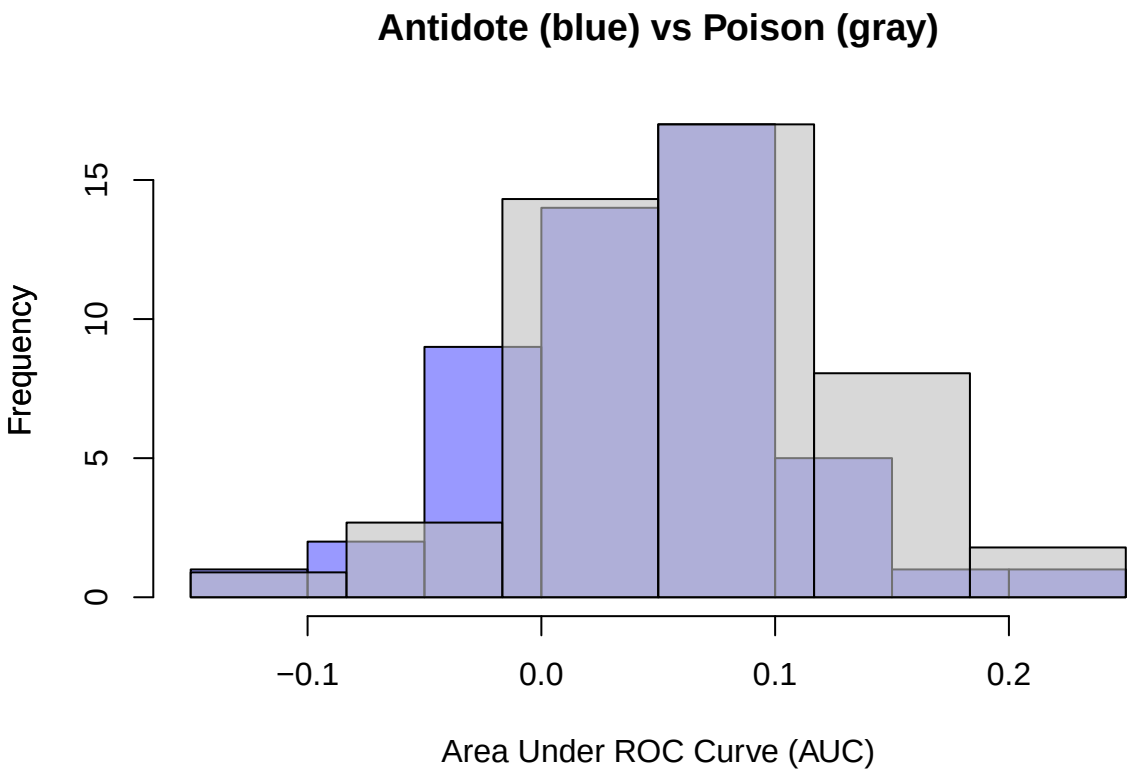
```
## Loading required package: survival
```

```r
## Dual Bootstrap Hypothesis Test "dbht" with 50 bootstrap samples,
## each with 50 individuals and running on all available cores.
whas <- get.whas100.dataset()
outliers_dbht <- survBootOutliers(
   surv.object=Surv(time = whas$times,event = whas$status ),
   covariate.data = whas[,2:5],
   sod.method = "dbht",
   B = 50,
   B.N = 50,
   parallel.param = BiocParallel::SnowParam()
)
```

```r
outliers_dbht$outlier_set[1:10,]
```

```
##       obs_id      pvalue
##  [1,]     51 0.000298115
##  [2,]     56 0.001955906
##  [3,]     67 0.006525616
##  [4,]     91 0.006612412
##  [5,]      1 0.009865626
##  [6,]     69 0.017169312
##  [7,]     93 0.021902428
##  [8,]     90 0.025564527
##  [9,]     32 0.031048894
## [10,]     94 0.038458610
```

```r
hist(outliers_dbht$histograms[[1]][[1]], main="Antidote (blue) vs Poison (gray)",
     col = alpha('blue', 0.4),
     xlab = "Area Under ROC Curve (AUC)")
par(new = TRUE)
hist(outliers_dbht$histograms[[1]][[2]],
     main = "", axes = FALSE,
     xlab='',col = alpha('gray', 0.6))
```



Although both samples do not fit the Cox proportional hazards well, the "poison" sample indicates that it tends to have data that is most relevant to that of the assumed model. Note the higher the AUC, the better the data fits the model. Here the data point removed could potentially be an influential outlier. The dataset used of the Worcester Heart Attack dataset with 100 individuals.

Asymptotics

The author provides a sketch of the asymptotic properties needed for estimating the parameters of interest. Recall, the term asymptotic leads to determining the properties that are assumed to correct when dealing with a large sample (infinitely large). The Huber-type monotone psi function is expressed in terms of asymptotic properties. (Casella, G., & Berger, R. L 1990) gives an account of such properties. The estimator derived by (Huber PJ 1981) considered an estimators understanding of both the mean and median. The mean is square, (much-like mean squared error) that happens to be sensitive to outliers. In other words, the outliers (tails) are overweighted. The median is given by an absolute value that is not sensitive to big or small outliers. Note, it has a constant k that can be tuned in such a fashion where the values look to be more median-like to mean-like. But in all the values will tend to do vary between the two central measures. The estimator acts as

a minimizer of the following equation

$$\sum_{i=1}^{n} \rho(x_i - a),$$

where

$$\rho(x) = \begin{cases} \frac{1}{2}x^2, & \text{if} |x| \leq k \\ k|x| - \frac{1}{2}k^2, & \text{if} \geq k \end{cases}$$

Note, minimization is of interest, analogous to traditional maximum likelihood, we solve for the derivative (when possible) and solve for the zeros. We then have, the most common expression of the Huber-type estimator,

$$\sum_{i=1}^{n} \rho'(x_i - a) = \sum_{i=1}^{n} \psi(x_i - a)$$

By the Weak law of large numbers, the estimator can be found to be asymptotically normal, logistic, double exponential, etc, by the necessary assumptions. Although much is gained with this robust estimator, there exists the drawback of being less efficient than that of the traditional MLE. In the context of this study, the expectation is obtained using the mean value theorem. Normality if the estimator properly behaves, then the central limit properly can be applied. The asymptotic covariance matrix under RML estimation can also be obtained. Recall, mean value theorem refers to if a continuous function and differentiable on an interval, there exists a point in the interval, such that the first derivative function of c, gives the average rate of change over the interval. Concerning there being a tangent in the interval.

Robust estimation in Weibull accelerated failure time models

Recall equation (4), the log-linear form of the accelerated failure time model. With the baseline hazard function concerning the Weibull distribution, the scale and shape parameters are constructed differently. More specifically, with log-linear form, the error terms can be visualized as the Gumbel distribution. The Gumbel can be denoted as an extreme value and is a member of the exponential family. Traditionally its purpose lies with modeling the distribution of the maximum number of samples required of various distributions. With the Weibull, it is the log-Weibull distribution. When latent (inferred) variables are considered it has then been discovered by Gumbel that these are the error terms of the log-linear model.

Robust maximum likelihood is used to numerically solve the estimating equations. The central idea behind iterative methods is the use of an initial value and subsequent approximations that improve at each iteration until a specified error bound is met. Here other iteration methods could have been considered. For instance, the secant method, newtons method, or Bisection method. Some characteristics distinguish the Newton-Raphson method from the others. The first is that it deals with extrapolation, in which the goal of

this survival analysis is to predict values that fall outside the range of data. This is clear by definition of an accelerated failure time model, where interest is in analyzing the effects of covariates on a subject's lifetime (unknown) (ie., how much longer they will live given qualitative or quantitative characteristics?). Next, the Newton-Raphson method is fast and suitable when the first and second derivatives are convenient to find. Such a method is called a root-finding algorithm where interest is in finding a value such that it takes the function to be zero.

By finding the maximum likelihood as done traditionally, the estimators obtained are unbounded. This is where the Huber-type psi function gives a bound on the influence of outliers. It bounds the error terms and covariates. And next, a weight function is chosen such that it utilizes Mahalanobis distance. Mahalanobis distance is most commonly used to find multivariate outliers, and it accounts for how correlated variables are amongst each other. In more formal terms, it is a measure that calculates the distance between a point and a distribution, discovered by (P. Mahalanobis 1936). An Indian Statistician with work well-known in sampling surveys and anthropometry. Anthropometry is a field in which the measurement of the size and proportions of the human body are of interest. It has a solid application to injury prevention, where a subject's state of movement may pose some issues. For example, its study is important for ones sitting posture (ergonomics). There are parts of the body that are exposed to injuries and it could be important to access, how much it may vary between humans. Nevertheless, the Mahalanobis distance, amongst others is a choice for clustering algorithms. The proportional and survivor functions are expressed with bound on the errors. Note the RML, denotes the uber-type function as z_i. The minimization of the ML method and RML for estimating mu, the location(intercept) is given respectively,

$$\sum_{i=1}^{n}\left[\delta_i\frac{\partial\log\{h_{\epsilon_i}(\epsilon_i)\}}{\partial\epsilon_i}+\frac{\partial\log\{S_{\epsilon_i}(\epsilon_i)\}}{\partial\epsilon_i}\right]=0$$

$$\sum_{i=1}^{n}\left[\left\{\delta_i\frac{\partial\log\{h_{\epsilon_i}(z_i)\}}{\partial z_i}+\frac{\partial\log\{S_{\epsilon_i}(z_i)\}}{\partial z_i}\right\}w(\mathbf{x}_i)-a_\mu\right]=0$$

The initial estimates are derived from first-order Taylor series expansion which is used to get the initial estimates. Therefore the iterative equations until convergence can be obtained. As a consequence, a $(p+2)\ddot{\mathrm{O}}(p+2)$ matrix is produced for the objective function. It is the equations given by the components, where the Huber psi function is imposed. Note that it can be tuned. Approximate variances were obtained from the sandwich-type variance-covariance matrix. ie,

$$V(\hat{\theta})=M^{-1}QM^{-1}$$

. , where $\hat{\theta}$ is an estimate of $(\mu, \alpha, \sigma)'$.

Example 2 - Mahalanobis distance

```
# https://www.r-bloggers.com/2021/08/how-to-calculate-mahalanobis-distance-in-r/
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x readr::col_factor() masks scales::col_factor()
## x purrr::discard()    masks scales::discard()
## x dplyr::filter()     masks stats::filter()
## x dplyr::lag()        masks stats::lag()
```

```
# process breast cancer data
data <- read_tsv('data/GSE9893_clinicalData.txt')
```

```
## Rows: 155 Columns: 22

## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (13): Tumor sample, Status, Adjuvant therapy, Histological  type, pT (TN...
## dbl  (9): Patient age (years), SBR Grade, Tumor size (mm), N+, N dich+, Foll...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# data <- data %>% select(6, 7, 8, 13, 14)
# data <- data %>% select(6, 7)
```

```r
# looks like there is some cleaning that needs to be done, before it can be used..
# Referring back to the WHAS dataset as used in Example 1
data <- whas[, 2:5]
data$mahalanobis <- mahalanobis(whas[,2:5], colMeans(whas[,2:5]), cov(whas[,2:5]))
# Identify distances that are signficant by
# chi-square statistic with k-1 df,
# k = number of variables = 5
data$pvalue <- pchisq(data$mahalanobis, df=4, lower.tail=FALSE)
whas.order<-data[order(data$pvalue),]
whas.order[1:10,]
```

```
##    los age gender      bmi mahalanobis       pvalue
## 8   56  81      1 28.27676   70.220618 2.039002e-14
## 97   3  32      1 39.93835   14.952884 4.799942e-03
## 30   4  85      1 36.71647    9.499527 4.975698e-02
## 52   4  43      1 25.33148    7.722589 1.022852e-01
## 10   9  40      0 21.78971    7.683838 1.038708e-01
## 22   8  69      1 37.60097    7.569530 1.086821e-01
## 98   8  86      1 14.91878    7.396866 1.163440e-01
## 45  18  76      1 32.41986    6.976342 1.371437e-01
## 50   6  80      0 36.02333    6.572966 1.602506e-01
## 7    3  66      1 35.71147    6.162955 1.873047e-01
```

```r
# p-values determined to be less than 0.001 are considered
# to be outliers.
# There exists 2 outliers in this dataset
```

Robust estimation in log-normal accelerated failure time models

If $T\ log-normal$, then accelerated failure time models has associated location $(\mu + \eta_i)$ and scale parameter (σ). The error terms follow a standard normal distribution. The standard maximum likelihood estimators are displayed, then the robust approach. Which thereafter can be solved numerically using the Newton-Raphson method. Expectations and covariance are in respect to the standard normal distribution. Note the general

form of log-normal is

$$\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{lnx-\mu}{2\sigma^2}}$$

Simulation study

Two sets of data are compared against each other. A set in which no outliers were considered in the data. Another set, where the data is altered and polluted with outliers (additional noise). Next of consideration is the performance of these sets under ML and RML. An informal way of detecting outliers is by a residual plot. Both residual plots clearly show four outliers are present. The plots suggest RML is not as influenced by these outliers as much as that for estimation under ML. The actual values tend to be more than the predicted values (positive trend).

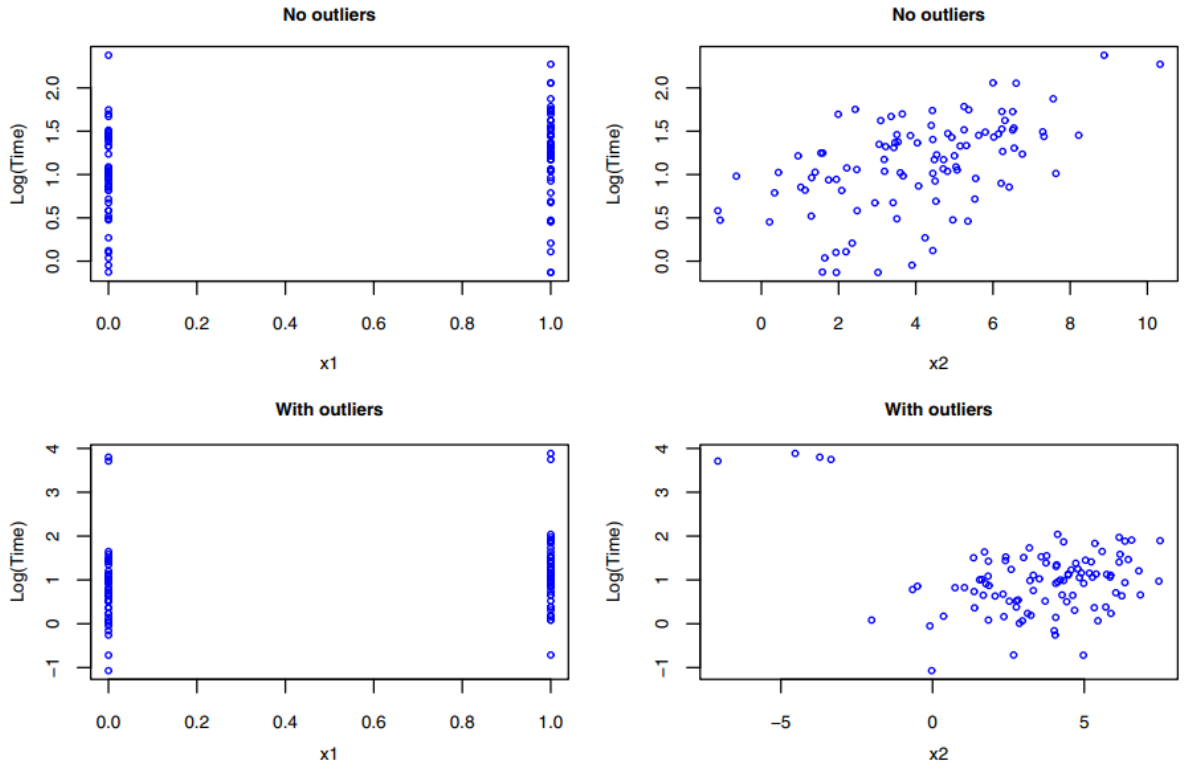


Fig. 1 Plots of survival time against covariates x_1 and x_2 for two representative data sets from Weibull distribution. Top two panels are for uncontaminated data; bottom two panels are for data contaminated with outliers

Figure 1: Figure 1

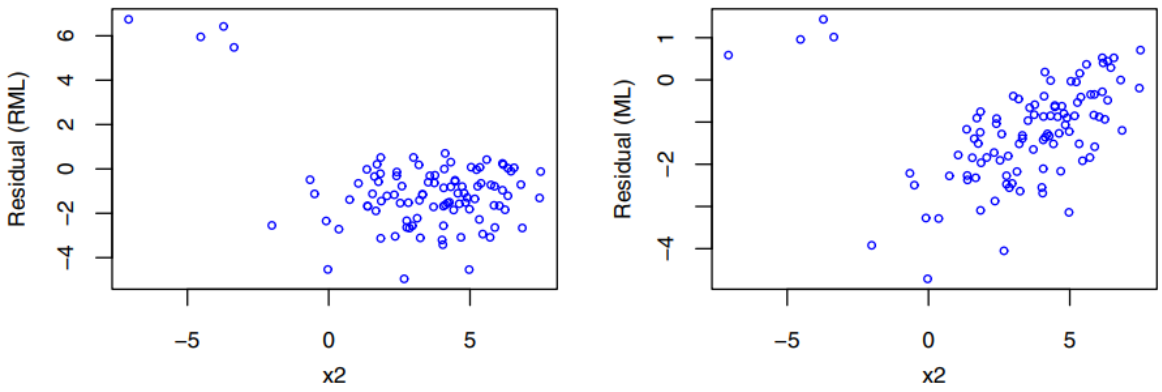


Fig. 2 Plots of residuals $\hat{\epsilon}_i = (\log t_i - \hat{\mu} - \hat{\alpha}_1 x_{1i} - \hat{\alpha}_2 x_{2i})/\hat{\sigma}$ from RML and ML fits against covariate x_2. Residuals are obtained from fitting Weibull accelerated failure time model to the contaminated data shown in Fig. 1. Four large residuals in each panel correspond to the outliers considered in the data

To formally access whether this interpretation is feasible, empirical results such as biases, mean squared errors, coverage probabilities, and average lengths of 95% confidence intervals are closely examined using Monte Carlo simulations. Note, coverage probabilities refer to the proportion of random regions (intervals) in regards to the time that will contain the true value.

First, the empirical results are analyzed in Weibull models. When there are no outliers, the maximum likelihood (ML) estimator has a lower mean-squared error than the robust maximum likelihood (RML). This is okay (not a cause for concern) since the focus is on data with outliers – expected in practice – (see Table 1 in Sinha, S.K. 2019). In contrast, one can argue that this is a cause for concern, which implies the importance of conducting such a comparison between the two methods. As it should not be overlooked that there exists and a number of data that where ML outperforms RML by larger margins. Adding a single outlier clearly suggests that ML begins to produce larger mean squared errors, larger biases, and lower coverage probabilities than the estimators produced by robust maximum likelihood (RML). Now with a 4% (of n = 100) outlier, RML has some biases but it is much more severe in the case of ML. When the sample size is increased, it can be shown that the respective estimators do much better under the weak law of large numbers (WLLN) but RML generally outperforms the ML method. Recall, WLLN refers to the average of a sequence (large samples) of independent and identically distributed (iid) random variables with a common mean and variance that converges in probability to the true value (see Table 3, Table 4 in Sinha, S.K. 2019).

In regards to log-normal models, it is again determined that when outliers are assumed to present in the data, then the RML outperforms the ML method (see Table 5 and 6 in Sinha, S.K. 2019).

Considering the log-logistic models, the conclusions mentioned for the other two models are consistent. More specifically, the ML method of estimating the first regression parameter is not as sensitive to outliers. When the sample size is increased, although the bias and MSE decrease at a constant rate for both estimators,

the RML exceeds the ML.

A reviewer of this study provided feedback that distinguished between moderate and extreme outliers. This insight can be interpreted as the question; how well does the robust method in this paper perform under extreme outliers. Real-life data can contain any number of outliers. More plausible would be several moderate outliers and a few extreme outliers. Knowing the amount of moderate or extreme outliers could help in the construction of the RML. To address this concern, Dr. Sinha suggests that the robust method would remain more reliable than that of the standard ML.

```
# TODO: finish simulation
# create covariates
x1 <- sample(c(0,1),100,replace=T)
x2 <- rnorm(100, mean = 4, sd = 2)
# data
data_weibull <- rweibull(46, shape = 1, scale = 1)
```

Application: breast cancer data

The data was presented in a previous study of survival analysis. A residual analysis by the assumed parametric distribution resulted in all models performing similarly. Although, due to the Weibull model having the double property of accelerated lifetimes and proportional hazards it was chosen. We then have the log-linear form as

$$\log(\text{TIME}) = \mu + \alpha_1 \text{SBR} + \alpha_2 \text{Node} + \alpha_3 \text{Tumor} + \alpha_4 \text{ER} + \alpha_5 \text{PR} + \sigma\epsilon$$

where SBR is a important prognostic factor in breast cancer, and can be denoted as a binary indicator. Nodes refers to Lymph nodes, tumor size, estrogen receptor and progesterone receptor. Concerning the data quality, the author suggests by the residual plots that outliers in regards to covariates exist and by the baseline probability survival there exist extreme outliers in regards to recurrence times. Hence, the RML method is valid to use, and conducted analyses are more so valid than ML. When comparing the two methods further, the location parameters are similar, but RML provides a smaller standard error. With the scale parameter, the results are slightly different. Similarly for the estimates of the coefficients. But the RML method produces smaller standard errors. (See figure 3, figure 4 in Sinha, S.K. 2019)

Discussion

Dr. Sinha, remarks that the RML method may be extended to frailty models, where the survival times and frailty distributions may be modified.

In reviewing this paper it has been noticed that there was not much comparison between the previous research of robust methods, and of that being proposed. It would be interesting to evaluate the other robust methods that have been "discovered" to not handle outliers in covariates sufficiently. This could be a potential extension of studying this paper. A useful method for studying the process of simulations concerning the research paper of interest is to conduct it oneself. Although this involves knowledge of statistical computing, it is a process that requires attention to detail. As seen above, another practical component of this report was the double bootstrap hypothesis test. Where the understanding and application of modern techniques were demonstrated.

The following remarks are courtesy of Professor Balakrishnan. Throughout this report, it has been outlined that the study differentiated which likelihood method of the model parameters was best given that outliers exist or not. When extending this paper for use by a medical practitioner the covariates are more of interest than that of the location μ. With that, it is the significance of the respective covariates that the medical researcher is interested in, as it can be used to make decisions (ie., say a subject's treatment plan). For instance, if the SBR grade is significant, then its coefficient would be away from zero. Meaning it would have a weight that suggests that it is of value when estimating the accelerated lifetime (log time). In such a fashion it is of interest to conduct a test of significance and access whether the robust method leads to more efficient results or not. It gives a reputable response to how often the covariates are expected to be significant. Another extension of this study comes from reviewing the impact of robustness when there is a departure from the assumed model. Then, the standard maximum likelihood method and the robust method can be compared in such a way.

In all, the reported study conveys the introduction of robust estimation in accelerated failure time models. Up to date review of literature of accelerated lifetime models assisted the author in investigating a potential research problem and solution. The complications in such studies seem to stem from the methodology. Each problem varies and as such properties of a method are expected to be facilitated in a trust-worthy. So that if there are any routes of improvement, one that is interested can further research. Dr. Sinha had outlined formally asymptotic properties that can be used, then proceeds by applying them to the distributions of interest. Note that, such derivations are also important for when one wants to apply this method to their data. Mainly, because it gives and explains the distributions or constants that can be tuned to fit one's data best.

References

Casella, G., & Berger, R. L. 1990. *Statistical Inference*. Pacific Grove. https://books.google.ca/books/about/Statistical_Inference.html?id=0x_vAAAAMAAJ.

Collett D. 2014. *Modelling Survival Data in Medical Research, 3rd Edn.* Chapman; Hall/CRC, New York. https://www.routledge.com/Modelling-Survival-Data-in-Medical-Research/Collett/p/book/9781439856789.

Huber PJ. 1981. *Robust Statistics*. Lifetime Data Anal 25, 52–78. https://link.springer.com/referenceworkentry/10.1007%2F978-3-642-04898-2_594.

Lin DY, Wei LJ. 1989. *The Robust Inference for the Cox Proportional Hazards Model*. J Am Stat Assoc 84:1074–1078. https://www.tandfonline.com/doi/abs/10.1080/01621459.1989.10478874.

P. Mahalanobis. 1936. *On the Generalized Distance in Statistics*. Proceedings of the National Institute of Sciences (Calcutta). http://library.isical.ac.in:8080/jspui/bitstream/10263/6765/1/Vol02_1936_1_Art05-pcm.pdf.

Pinto JD, Carvalho AM, Vinga S. 2015. *Outlier Detection in Cox Proportional Hazards Models Based on the Concordance c-Index*. Machine learning, optimization,; big data: lecture notes in computer science, pp 252–256. https://link.springer.com/chapter/10.1007/978-3-319-27926-8_22.

Sinha, S.K. 2019. *Robust Estimation in Accelerated Failure Time Models*. Lifetime Data Anal 25, 52–78. https://doi.org/10.1007/s10985-018-9421-z.

Sinha SK, Rao JNK. 2009. *Robust Small Area Estimation*. Can J Stat 37:381–399. https://onlinelibrary.wiley.com/doi/10.1002/cjs.10029.

Susana Vinga. 2017. *Outlier Detection in Survival Analysis (All Techniques)*. Instituto Superior Técnico. http://web.ist.utl.pt/~susanavinga/outlierRP/Dataset-myeloma/myeloma_OD_All_Techniques_.html.