Independent Study of Robust estimation in accelerated failure time models

Korede Adegboye

400014008

STATS 756 - Topics in Biostatistics

December 3, 2021

Introduction

When considering lifetime data with covariates, information (characteristics) is given by these covariates that have an effect on an individual's lifetime. The interest of a subject is survival time, may include characteristics such as age, sex, the severity of disease, smoking status, results of blood tests, and other laboratory data. In all, with the covariates introduced to the statistical model, it is expected that better production of lifetimes and thus, efficient analysis of the study can take place. Analogous to other statistical models, covariates can be quantitative or categorical. 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.

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 the pdf of the Weibull distribution as

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

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

The first form derived from Weibull distribution has that scale parameter α depending on x.

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

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

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

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

Most of the robust developments and the one popular for analysis among those in biomedical sciences make use of the semi-parametric Cox's proportional hazards model. In eq 1, the perspective used resulted in proportional hazard, now of consideration is that the defining property is proportional hazards. This has an effect on how the covariates are also introduced. Thus, this gives a regression model for lnh.

Other considerations stem from introducing the covariates such that it acts on different characteristics of a model that can produce a different model. 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. The effect of the covariates, x is to accelerate or decelerate the time scale t as $(te^{\beta' x})$.

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

When the goal of the study is to produce an analysis based on regression models for lifetime data with covariates we can apply the accelerated lifetime model 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.

Popular in practice because it supports various survival models. 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. To assume such an application, one must access which distributions fit the data best. With that it requires precise specifications, otherwise, it is expected for it to be an insufficient model. With covariates being the main component of accelerated lifetime models, accelerated lifetime regresses the logarithm of these covariates. Where many find it useful 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 outlier observations

are rejected or down-weighted.

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

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

```
## 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()
)



# outliers_dbht <- survBootOutliers(
#    surv.object=Surv(time = whas$times,event = whas$status ),
#    covariate.data = whas[,2:5],
#    sod.method = "dbht",
#    B = 1000,
#    B.N = 50,
#    parallel.param = BiocParallel::MulticoreParam()
# )
```

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

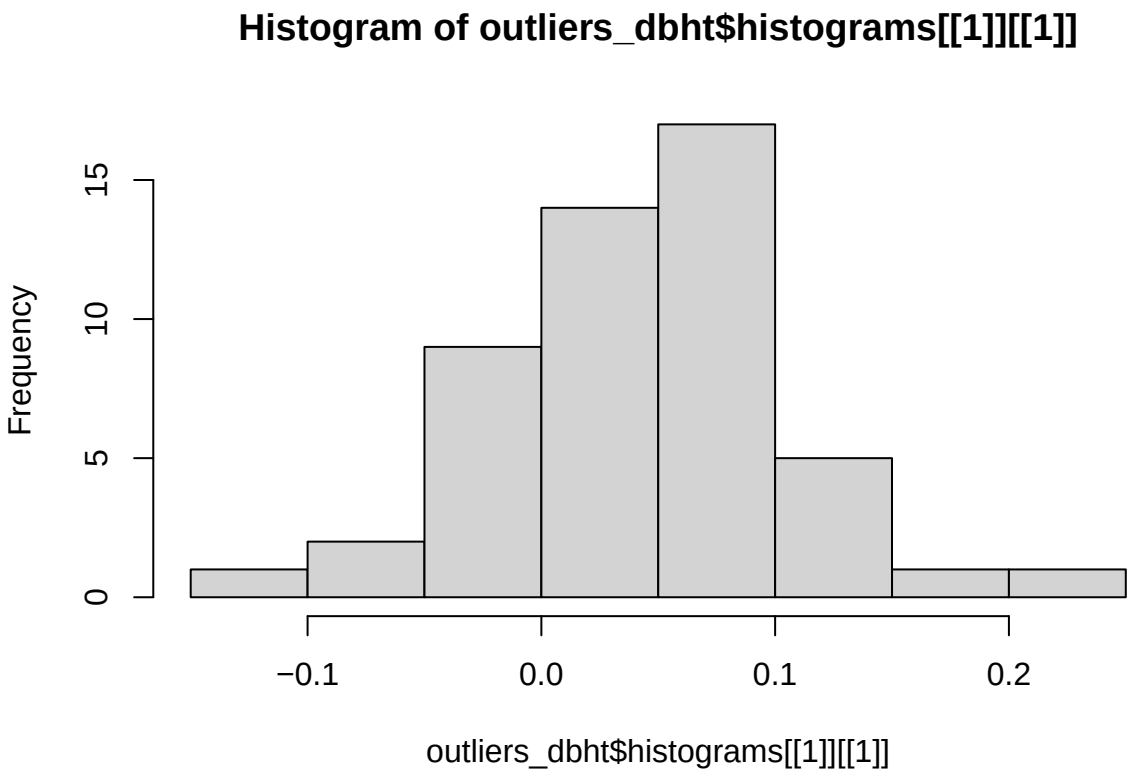
```
##       obs_id      pvalue
##  [1,]     51 0.000298115
##  [2,]     56 0.001955906
##  [3,]     67 0.006525616
##  [4,]     91 0.006612412
```

```
##  [5,]      1 0.009865626
##  [6,]     69 0.017169312
##  [7,]     93 0.021902428
##  [8,]     90 0.025564527
##  [9,]     32 0.031048894
## [10,]     94 0.038458610
```

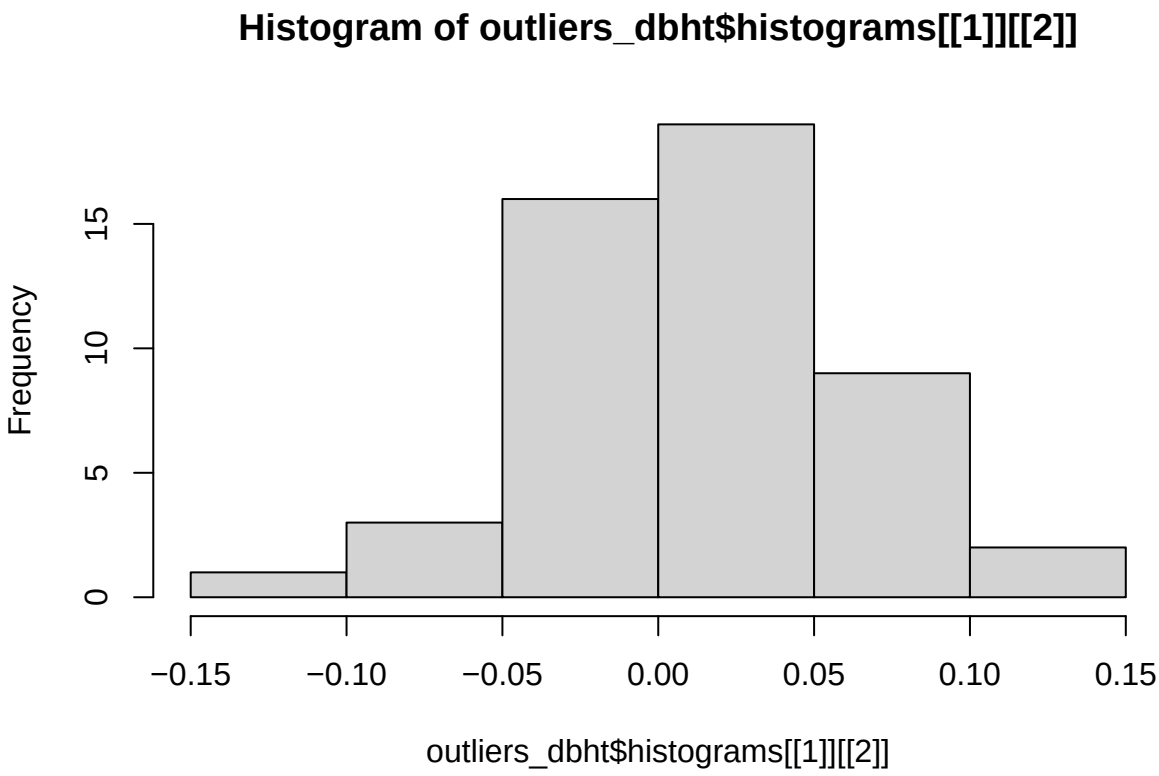
```
dbht<-as.data.frame(outliers_dbht$outlier_set)# convert to a dataframe (more easy to handle)
names(dbht)[1]="id" # change the name of the 1st) column
dbht.order<-dbht[order(dbht$id),]
dbht.order[1:10,]
```

```
##    id      pvalue
## 5   1 0.009865626
## 96  2 0.976264699
## 13  3 0.068336088
## 93  4 0.933845395
## 53  5 0.535617261
## 57  6 0.584105898
## 48  7 0.472572446
## 28  8 0.240662362
## 62  9 0.665766065
## 88 10 0.905488454
```

```
hist(outliers_dbht$histograms[[1]][[1]])
```




```
hist(outliers_dbht$histograms[[1]][[2]])
```



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. With the aim to produce a parametric accelerated failure time model that can handle outliers in 2 settings – survival outcome variable and covariates.

(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). 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.

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 life time 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 interest in analyzing the speed of progression of a disease over time. Equation () above gives the general accelerated failure time model. To derive the log-linear form, consider the parametric accelerated failure time model. It can be denoted by the following...

Interest lies in estimating the model parameters given by the parametric distribution in such a fashion that it is resistant to outlier. - Show formulas...

Robust Estimation

Robust estimation can be defined as a method that is not influenced by slight differences of the proposed assumptions (ie., probability distribution). As a consequences it optimizes the algorithms 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. Here, of interest is estimating the

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. Explain more about Huber-type monotone psi function, (Huber PJ (1981))

Here our goal is to estimate the model parameters , and by a suitable robust method which is resistant to potential outliers in both survival times and covariates in the data

Asymptotics

The author provides a sketch of the asymptotic properties needed for estimating the parameters of interest. Recall, by the term asymptotic leads to determine the properties that are assumed to correct when dealing with a large sample. Expectation using the mean value theorem. Normality if the estimator is properly behaved, then the central limit properly can be applied. Covariance

Robust estimation in Weibull accelerated failure time models

Review of 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, we can visualize the error terms as the Gumbel distribution. The Gumbel can be denoted as an extreme value (type 1) and is a member of the exponential family. Traditionally its purpose lies with modeling the distribution of the maximum number of samples 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 (ie., how much longer they will live given qualitative or quantitative characteristics?). Next, the Newton-Raphson method is fast and a standard when the first and second derivatives are convenient to find and compute. Such a method is called a root-finding algorithm where interest is finding a value such that it takes the function to be zero. This is consistent with the motivation behind accelerated lifetime models. (rephrase some parts)

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 $(p+2)\ddot{O}(p+2)$ 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 ...

Robust estimation in log-normal accelerated failure time models

Under log-normal settings there are some differences. Under log-normal settings, the accelerated failure time models have an associated shape and scale parameter. 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 - Generalized M (GM) estimating equations for ordinary linear regression???

Robust estimation in log-logistic accelerated failure time models

- Eq 9 used reference in all considerations of distributions and the equation is solved (in relation to the distribution)
- What does one need to understand in order to grasp the use of RML

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).

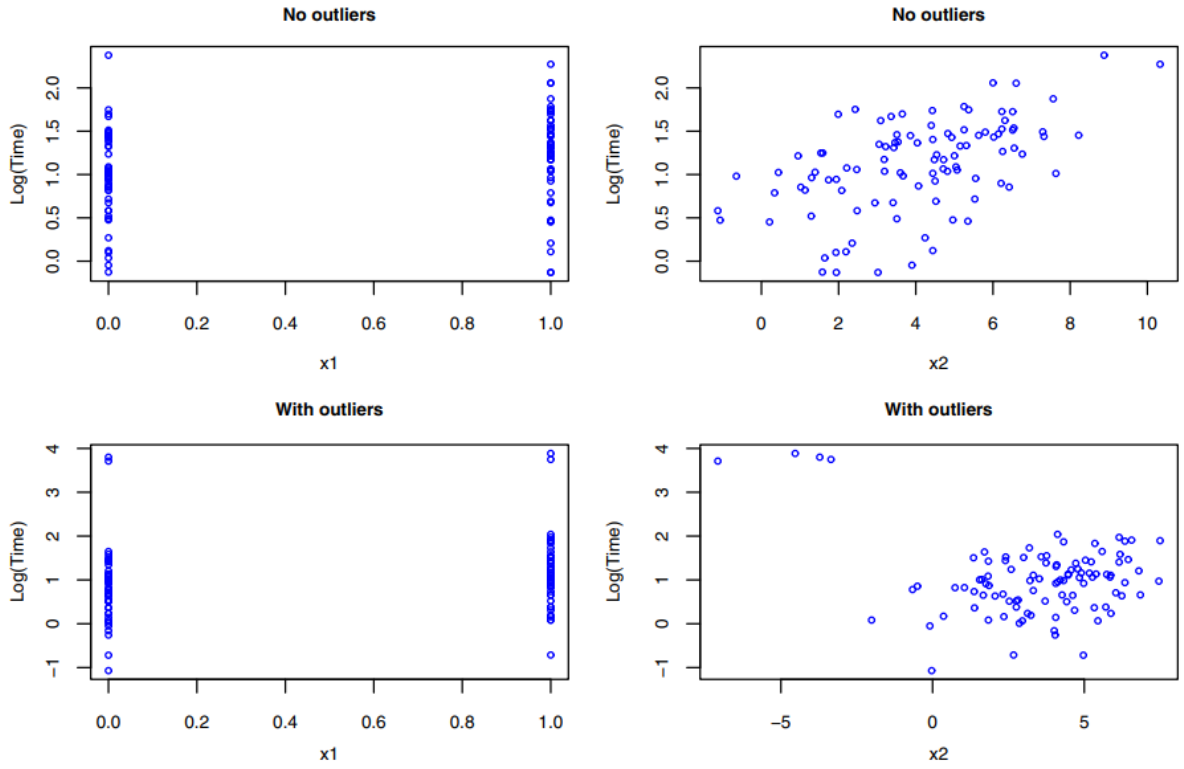


Fig. 1 Plots of survival time against covariates x_1 and x_2 for two representative data sets from Weibull distribution. Top two panels are for uncontaminated data; bottom two panels are for data contaminated with outliers

Figure 1: Figure 1

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

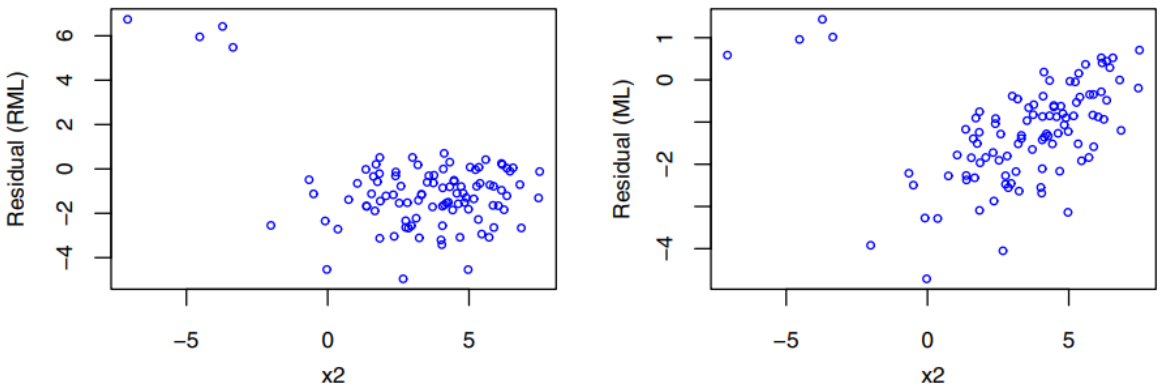


Fig. 2 Plots of residuals $\hat{\epsilon}_i = (\log t_i - \hat{\mu} - \hat{\alpha}_1 x_{1i} - \hat{\alpha}_2 x_{2i})/\hat{\sigma}$ from RML and ML fits against covariate x_2. Residuals are obtained from fitting Weibull accelerated failure time model to the contaminated data shown in Fig. 1. Four large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

Figure 2: Figure 2

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 a number of 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.

```r
# 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

In the data that was given for example, there exists ... for quantitative and ... for categorical - If I were to use this data. What else would I need to know. Or can I provide a review on how I used it given the information

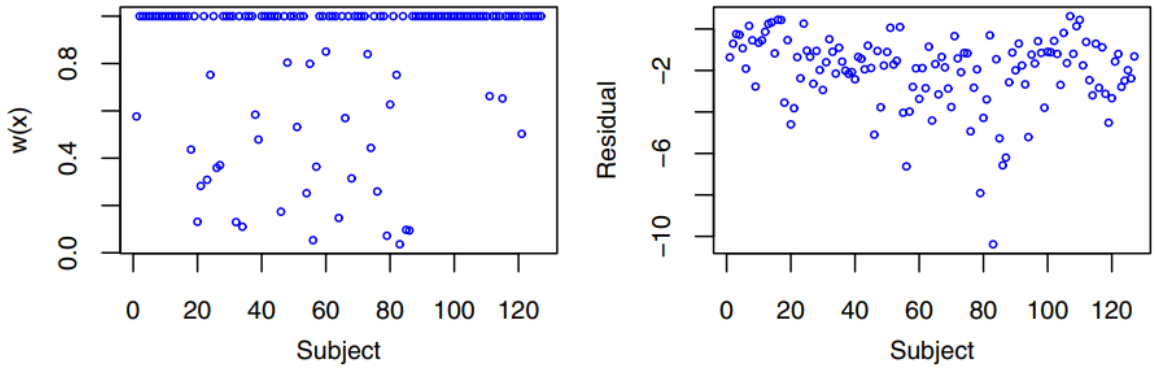


Fig. 3 Plots of weights $w(\mathbf{x}_i)$ and residuals $\hat{\epsilon}_i = (\log t_i - \hat{\mu} - \hat{\eta}_i)/\hat{\sigma}$ from robust (RML) fit to the primary breast cancer data

Figure 3: Figure 3

Discussion

When choosing between which models how is residual analysis done with applied data - What other determining factors can be used to determine the model? - The RML was compared against ML. But why not compared against the RML proposed by Locatelli et al ? Wouldn't this be the comparison that should be discussed in future ? - I can start thinking myself as to how this robust estimation can be extended to frailty models

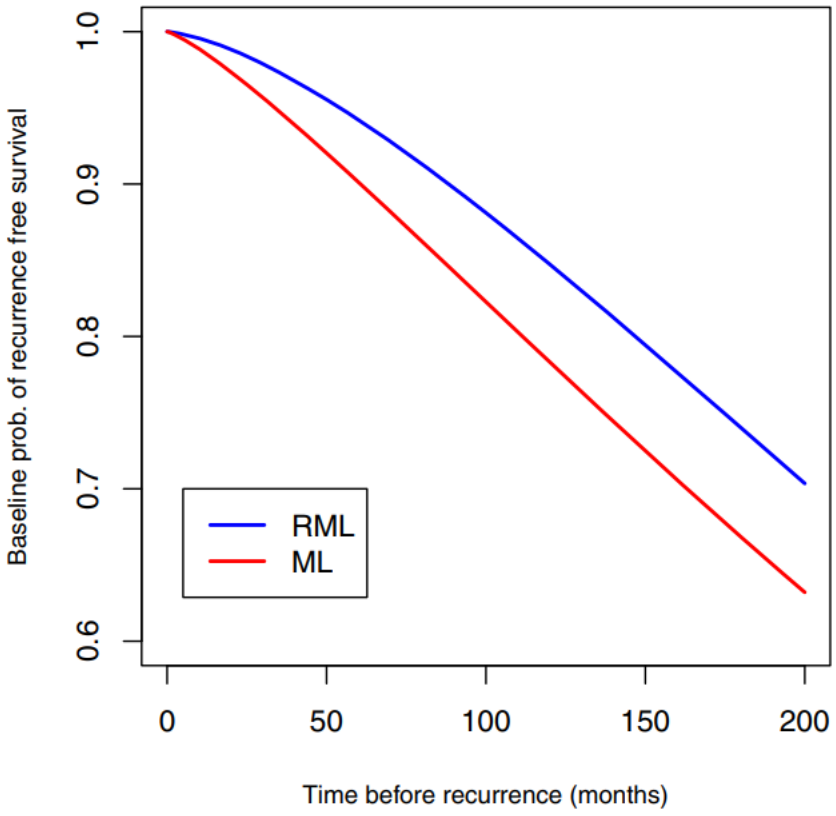


Fig. 4 Baseline probability of recurrence free survival estimated by RML and ML methods

Figure 4: Figure 4

References

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.

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.