NTNU – Trondheim
Norwegian University of
Science and Technology

Analysis of Life Regression Models for Censored Data using Pseudo Observations

Ida Hakavik Braarud

Master of Science in Physics and Mathematics
Submission date: June 2013
Supervisor: Bo Henry Lindqvist, MATH

Norwegian University of Science and Technology
Department of Mathematical Sciences

Problem Description

- Give an introduction to parametric life regression modelling.
- Study methods for estimation and residual analysis using pseudo observations.
- Illustrate methods by simulation and analyses of real data.

Assignment given: January 16, 2013.
Supervisor: Bo Henry Lindqvist.

Preface

This thesis concludes my Master of Science in Applied Physics and Mathematics, at the Norwegian University of Science and Technology (NTNU). The work was carried out at the Department of Mathematical Sciences during the spring semester of 2013.

I would like to thank my supervisor, Professor Bo Lindqvist for motivation and excellent guidance. I have enjoyed working with you and greatly appreciate having had you as my supervisor.

I would also like to thank my friends and classmates for five good years together at NTNU. Thanks to Helene Eide Wiik and Jeanett Gunnekleiv Støtvig for proofreading. To Ove Frisnes for always cheering me up, and to my family for their support and encouragement.

Trondheim, 2013-06

Ida Hakavik Braarud

Abstract

Censoring is a common form for missing data in survival analysis. When a data set is censored, there is only partial knowledge of the survival time of some of the study units. To compensate for this, special techniques and adjusted residuals may be used in analyses. An alternative to this is to obtain new data sets through pseudo observations from jackknife theory. These new data sets can then be treated as uncensored data sets, and ordinary regression methods can be applied. This master's thesis studies methods for obtaining pseudo observations based the Kaplan-Meier estimator and modelling by accelerated failure time models (AFT models). Three methods are presented, one parametric and two non-parametric. How well the three methods preform under different levels of censoring and true distributions are studied, and some recommendations on when they are appropriate to use are made. Pseudo observations are also studied for Cox-Snell and standardized residuals for AFT models, and also here we arrive at some recommendations regarding their use. Both pseudo observations and pseudo residuals are then used in residual analysis and model checking. Methods are illustrated with simulated and real data sets.

Sammendrag

I levetidsanalyse er sensurering en vanlig form for manglende data. Når et datasett er sensurert, vil man kun ha delvis kjennskap til levetiden til enkelte studieenheter. For å kompensere for dette, kan spesielle teknikker og justerte residualer brukes til å analysere dataene. Et alternativ til dette er å lage nye datasett bestående av pseudo observasjoner fra jackknife teori. Disse nye datasettene kan behandles som usensurerte datasett, og vanlige regresjonsmetoder kan anvendes. Denne masteroppgaven studerer metoder for å finne pseudo observasjoner basert på Kaplan-Meier estimatoren og modellering av "accelerated failure time" modeller (AFT modeller). Tre metoder presenteres, en parametrisk og to ikke-parametriske. Hvor godt de tre metodene presterer under ulike nivåer av sensurering og ulike sanne fordelinger blir studert, og noen anbefalinger rundt bruken av de blir gjort. Pseudo observasjoner blir også studert for Cox-Snell og standardiserte residualer for AFT modeller, og også her er kommer vi med noen anbefalinger angående anvendelser. Både pseudo observasjoner og pseudo residualer blir så brukt i residualanalyse og modellsjekking. Metodene er illustrert med simulerte og reelle datasett.

Contents

1 Introduction **1**

2 Basic Concepts in Survival Analysis **3**
2.1 Survival time 3
2.2 Censoring 4
2.2.1 Types of censoring 4
2.3 Survival and hazard functions 5
2.4 Mean survival time 6
2.5 Likelihood function for right censored data 7

3 Kaplan-Meier Estimator **9**
3.1 Kaplan-Meier estimator of the survival function for T 9
3.1.1 Mean and restricted mean 10
3.1.2 Example, Kaplan Meier 10
3.2 Kaplan-Meier estimator of the survival function for $log(T)$ . . 12
3.2.1 Mean and restricted mean 12
3.2.2 Example, Kaplan-Meier continued 12

4 Accelerated Failure Time Models **15**
4.1 Regression models 15
4.2 General accelerated failure time models 15
4.3 Log-linear representation 16
4.4 Estimation of log-linear regression models 17
4.4.1 Estimation with R 18
4.5 Residuals 19
4.5.1 Standardized residuals 19
4.5.2 Cox-Snell residuals 19
4.5.3 Adjusted Cox-Snell residuals 20
4.6 Weibull accelerated failure time models 21
4.6.1 Residuals 21
4.6.2 The distribution and likelihood of Y 21
4.7 Lognormal accelerated failure time models 23
4.7.1 Residuals 23
4.7.2 Likelihood function for Y 23
4.8 Simulating survival data from accelerated failure time models 24

5 Pseudo Observations **25**
5.1 Introduction to pseudo observations 25
5.1.1 Jackknife and pseudo observations 25
5.1.2 Pseudo observations in survival analysis 26
5.2 Pseudo observations based on Kaplan-Meier for T 28
5.2.1 Uncensored observations 28
5.2.2 Censored observations 29
5.3 Pseudo observations based on Kaplan-Meier for log(T) 41
5.3.1 Uncensored observations 41
5.3.2 Censored observations 41
5.4 Parametric pseudo observations 47
5.4.1 Uncensored data sets 47
5.4.2 Censored data sets 47
5.5 Estimation with pseudo observations 59
5.6 Prostatic cancer example 60
5.6.1 Nonparametric methods 62
5.6.2 Parametric pseudo observations 66

6 Pseudo Residuals **71**
6.1 Introduction to Pseudo Residuals 71
6.1.1 Pseudo residuals for data set W1 72
6.2 KM, KMlog and parametric Cox-Snell pseudo residuals 75
6.3 Pseudo standardized residuals 75

7 Model Checking and Functional Form **77**
7.1 Analysis of residuals 77
7.1.1 Example, Nelson's superalloy data 77
7.2 Functional form for covariates 81
7.2.1 Simulated Weibull example 83

8 Concluding Remarks **93**

Bibliography **96**

Appendices **99**

A Distributions **101**

B Data sets **105**

C Figures **111**

D R codes **117**

List of Figures

2.1 Example of survival experiences for ten patients. Right censored observations are represented by circles and observed events are represented by dots 4

3.1 Figure illustrating ordering of event times for the Kaplan-Meier estimator. D indicates observed events and C indicates censored events 9

3.2 Plot of Kaplan-Meier estimator and 95% confidence interval for the survival times in the example. The plot is made using the *survfit* function from the *survival package* in *R*. 11

3.3 Plot of Kaplan-Meier estimated survival curves for log(T), for the same survival times as in figure 3.2. Circles indicates event and censoring times. 13

5.1 Plot of KM pseudo observations (*), real survival times (blue ∘) and observed survival times (red •) for 6 different data sets, all with 30% censoring and $\sigma = 2/3$. 30

5.2 Pseudo observation (*), Real times (blue ∘) and observed times (red •) for the simulated W1 data set sorted after ascending observed survival time. 31

5.3 Figures showing Kaplan-Meier curves for data set W1: A) The full data set, B) Jackknife replication 2, C) Jackknife replication 3, D) Jackknife replication 5, E) Jackknife replication 10, F) Jackknife replication 11, G) Jackknife replication 15, H) Jackknife replication 20. 33

5.4 Pseudo observations from the original W1 data set, where observation 16 are censored, plotted against pseudo observations if observation 16 was uncensored. 35

5.5 Pseudo observations from the original W1 data set, where observation 16 are censored, plotted against pseudo observations for when we treat observation 16 as censored at time 8. 35

5.6 Three plots of the same data set with 300 observations. Pseudo observations (*), observed survival times (red •) and real survival times (blue ∘). Figures show: left: A plot where the observed times are unsorted. Middle: A figure where survival times are sorted in ascending order. Right: A figure where pseudo observations observed and real survival times are sorted individually in ascending order 37
5.7 Top row: Figure showing pseudo observations (*), real survival times (blue ∘) and observed survival times (red •) for the same Weibull distributed data set, but different levels of censoring. Bottom row: Real survival times plotted against pseudo observations 38
5.8 Weibull probability density functions for different shape parameters ($1/\sigma$), all with scale=1. 39
5.9 Figures showing KM pseudo observations plotted against real survival times for different values of σ. The data set have the same covariates, $\boldsymbol{\beta}$=(0,0.8,0.8,0.3) and 30% level censoring. . . 40
5.10 Figure showing KM pseudo observations (*), KMlog pseudo observations (+) real survival times (blue ∘) and observed survival times (red •) for the same 6 different data sets as in figure 5.1. Note: KMlog for observation 20 in plot A and D are outside the range of the plot. 42
5.11 Figure showing real survival (red •) and pseudo observations from KMlog (+), for $log(T)$ and T. The data set are the same as in plot D in figure 5.10. 43
5.12 For the same data set with 300 observations as in figure 5.6, three figures are plotted with pseudo observations (+), real survival times (blue ∘) and observed survival times (red •). From the left: Unsorted data. Middle: Sorted after increasing observed survival time. Right: sorted individually, no link between pseudo, real and observed observations. 44
5.13 Figures showing plots for different levels of censoring of the same data set as in figure 5.1. Top row: Real survival times (blue ∘), observed survival times (red •) and KMlog pseudo observations (+). Sorted after increasing observed time. Bottom row: Plot of KMlog pseudo observations against real survival times. The line shows where point will be if real survival times are equal to pseudo observations. OBS: Several points are not included for high censoring, so this plot may be misleading. . . 45
5.14 Plots of KMlog pseudo observations against real survival times for different values of σ. The data set have the same covariates, $\boldsymbol{\beta}$=(0,0.8,0.8,0.3) and 30% level censoring, and is the same as in figure 5.9. 46
5.15 Plots for a simulated uncensored data set from a Weibull AFT model with β=(0,0.5,1,0.3) and 30% censoring. Observed and event time (blue and red •), Parametric pseudo observation (△). 48

5.16 Parametric pseudo observations (△), KM pseudo observations (*), KMlog pseudo observations (+), real survival time (blue ∘) and observed survival time (red •). The data sets are the same as for figure 5.1 and figure 5.10. 49
5.17 Pseudo observations (△), real survival times (blue ∘) and observed survival times (red •) for the same data set as in figure 5.6 and 5.12. Left: Unsorted. Middle: Sorted observations in increasing time. Right: sorted individually, no link between pseudo, real and observed observations. 50
5.18 Pseudo observations (△), real survival times (blue ∘) and observed survival times (red •) for censored and uncensored observations. The data set is the same as in figure 5.17. 51
5.19 Pseudo observations (△), Observed survival times (red •) and real survival times (blue ∘) for the W1 data set. 53
5.20 Original pseudo observations from the W1 data set plotted against pseudo observations when we changed observation 16. 54
5.21 Top row: Parametric pseudo observations (△), real survival times (blue ∘) and observed survival times (red •) for different levels of censoring of the same data set as in figure 5.7 and 5.13. Bottom row: Real survival times plotted against parametric pseudo observations. 55
5.22 Plot of parametric pseudo observations (△), real survival times (blue ∘) and observed survival times (red •) for different values of σ. The data set is the same as in figure 5.9 and 5.14. 57
5.23 PDF for lognormal distributions with different values of σ. . . 58
5.24 Plot of parametric pseudo observations for Lognormal and Weibull distribution against real Weibull survival times. $\boldsymbol{\beta} = (0, 0.7, 0.6, 0.2)$, $\sigma = 0.2$ and there is 30% censoring. 58
5.25 Histogram of survival times for prostatic cancer trial. 60
5.26 Plot of survival times versus covariates for the prostatic cancer data set. 61
5.27 Plot of survival times for both treatments in the prostatic cancer data set. 62
5.28 KM/KMlog pseudo observations (black) and observed survival times (red). Observed times (•) and censored times (∘) for the prostatic cancer data set 63
5.29 Histogram for KM pseudo observations for the prostetic cancer data set 63
5.30 Plot of KM pseudo observations against covariates for the prostatic cancer data set. 64
5.31 Plot of survival times for both treatments in the prostatic cancer data set. 64
5.32 For the prostatic cancer data set: Plot of survival times estimated with *survreg* (black) and observed (red) for Weibull and lognormal distribution. Uncensored survival times (•) and censored survival times (∘). 67

5.33 Parametric pseudo observations (black) and observed survival times (red) for Weibull and lognormal distribution. Censored survival times are (∘) and uncensored survival times are (•). Obs: high pseudo observations are not included. The data set is the Prostatic cancer data set. 68

6.1 Pseudo observations for unit exponentially distributed survival times. 72
6.2 Cox-Snell residuals and Pseudo Cox-Snell residuals for W1 data set. Red ∘ indicate censored observations, and black ∘ indicate uncensored observations. 73
6.3 Pseudo residuals and Cox-Snell residuals for the W1 data set plotted against the value of the Cox-Snell residual. 74
6.4 Pseudo residuals for W1 plotted against Cox-Snell residual. KM (*), KMlog (+) and Parametric (△). Censored observations are red, and uncensored observations are black. Lines show where the points would be if the difference were one (red) or zero (black). 76
6.5 Standardized residuals (circle) and pseudo standardized residual for W1 dataset. Red observations are censored and black observations are uncensored. 76

7.1 Nelsons superally data set. Number of cycles (in thousands) plotted against the logarithm of the pseudostress (measured in ksi). Censored observations are ∘, uncensored are •. 78
7.2 Logarithm of Cox-Snell residuals for the original superalloy data set. Censored observations are circles, uncensored are dots 79
7.3 KM pseudo observations of th number of cycles in Nelson's superalloy data, plotted against log(pseudostress). Censored observations are circles, uncensored are dots 79
7.4 Left: Logarithm of Cox-Snell residuals for pseudo observations vs. logarithm of pseudostress. Residuals from censored observations are circles, and uncensored are dots. Right: Exponential probability plot for Cox-Snell residuals from pseudo observations. 80
7.5 Left: Logarithm of pseudo Cox-Snell residuals vs. logarithm of pseudostress. Residuals from censored observations are circles, and uncensored are dots. Right: Exponential probability plot for pseudo Cox-Snell residuals. 81
7.6 Pseudo observations (black) for four different levels of censoring plotted with the observed survival times (red). NB: Some pseudo observations are to big to be included in the plot . . . 85
7.7 Left: Standardized residuals from pseudo observations. Right: Estimated covariate function for X 86
7.8 Log(x) for values of x in the simulated data set 86
7.9 Left: Standardized residuals from pseudo observations. Right: Estimated covariate function for X 87

7.10 Left: Standardized residuals from pseudo observations. Right: Estimated covariate function for X 88
7.11 Left: Standardized residuals from pseudo observations. Right: Estimated covariate function for X 88
7.12 Left: Pseudo standardized residuals. Right: Estimated covariate function for X 90
7.13 Left: Pseudo standardized residuals. Right: Estimated covariate function for X 90
7.14 Left: Pseudo standardized residuals. Right: Estimated covariate function for X 91
7.15 Left: Pseudo standardized residuals. Right: Estimated covariate function for X 91

C.1 Left: Standardized residuals plotted against X. Right: Estimated covariate function for X, 0% censoring 111
C.2 Left: Standardized residuals plotted against X. Right: Estimated covariate function for X, 25% censoring 112
C.3 Left:Standardized residuals plotted against X. Right: Estimated covariate function for X, 50% censoring 112
C.4 Left: Standardized residuals plotted against X. Right: Estimated covariate function for X, 75% censoring 113
C.5 Left: 1-adjusted Cox-Snell residuals. Right: Estimated covariate function for X, 0% censoring 113
C.6 Left: 1-adjusted Cox-Snell residuals. Right: Estimated covariate function for X, 25% censoring 114
C.7 Left: 1-adjusted Cox-Snell residuals. Right: Estimated covariate function for X, 50% censoring 114
C.8 Left: 1-adjusted Cox-Snell residuals. Right: Estimated covariate function for X, 75% censoring 115

List of Tables

3.1 Kaplan-Meier survival function for survival times in the Kaplan-Meier example. 11

5.1 Selected estimated expected survival times and pseudo observations from the W1 data set. Note that the values are rounded off. 32
5.2 Estimates of parameters and expected value of covariates, for the full and some reduced data sets from the W1 data set. Along with pseudo observations for $log(T)$ $(\hat{\theta}_i)$, and pseudo observations for T. 52
5.3 Estimated parameters from the original simulated data set with 10 000 observations and three pseudo observation data sets for four levels of censoring. 59

7.1 Theoretical parameter values minimizing the Kullback-Leiber distance between model 7.8 and 7.9 for four levels of censoring. 83
7.2 Parameters in model 7.9 estimated from pseudo observations . 84
7.3 Parameters in model 7.9 estimated from with *survreg*. 89

B.1 The W1 data set. 106
B.2 The prostetic cancer data set. 107
B.3 Data and pseudo observations for the prostetic cancer data set. 108
B.4 Nelson's superalloy data set. 109

Chapter 1

Introduction

In survival analysis the focus is on examining the time until a specific event or endpoint. Examples are time to Aids for HIV patients, time to a light bulb stops working or the number laps around a track a person can run. The variable that we measure, T, is called the survival time, event time or failure time. In some occasions however, we do not observe the event for all individuals or items that we study the survival time for. The HIV patient may drop out of the study, we may not observe the exact time the light bulb stops working or the time the track is available to the runner may run out. The real event time will then be unknown and we say that the survival time is censored. This aspect of survival data makes standard methods inappropriate, and several methods have been developed to handle censoring.

Without censored observations, the survival data can be analysed with standard regression models, and standard graphical methods for assessing the fit of the model can be used. An alternative to the methods developed for censored observations is therefore to obtain a synthetic data set that behaves like a uncensored version of the original data set. A method for finding such data sets can be found in pseudo observations from jackknife theory. The idea, proposed by Andersen et al. [4], is that pseudo observations of different functions of T can be found with equation

$$\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{-i},$$

where $\hat{\theta}$ is an estimator of the function for the whole data set, and $\hat{\theta}_{-1}$ is an estimator for the data set when we remove observation i. The function can for instance be the expected value of the survival time, the survival function or residuals.

This thesis starts with some introduction chapters. In Chapter 2, basic concepts like survival and hazard functions, mean time to failure and likelihood are presented. Chapter 3 introduces the Kaplan-Meier estimator for the survival time and the logarithm of the survival time. This will be needed in later chapters when we need an estimator for the expected survival time or the expected value of a residual. In chapter 4, a parametric model for survival data is presented. This parametric model is called the accelerated

failure time model, or AFT for short. On log-linear form it can be expressed as

$$Y = log(T) = \mu + \boldsymbol{\beta}'\mathbf{X} + \sigma\epsilon.$$

where μ and σ are intercept and scale parameter, and ϵ is the error term. Different properties for log-linear AFT models are then discussed, in particular for Weibull and lognormal distributed survival times.

Chapter 5 and chapter 6 are the main chapters in this thesis. In chapter 5 three methods for finding pseudo observations for the survival time is studied. The two first are non-parametric and based on the Kaplan-Meier estimator. The second is parametric and based on the specific AFT model we assume for our data. For the three methods, we look at how the degree of censoring and the value of sigma affects the pseudo observations. Chapter 6 studies pseudo observations for residuals. Because censored survival times lead to censored residuals, we can obtain pseudo residuals the same way as pseudo survival times.

Then chapter 7 is dedicated to the use of pseudo observations and pseudo residuals in residual analysis and assessing the functional form of a covariate. Followed by concluding remarks are given in chapter 8.

In appendix A, the distributions used in this thesis is presented. Some of the data sets we look at is included in appendix B. Figures not included in chapter 7 can be found in appendix C, and R-codes for creating data sets and pseudo observations/residuals are in appendix D.

Main sources of information have been Colletts book *Modelling Survival Data in Medical Reaseach* [7] and articles *Regression Analysis of Restricted Mean Survival Time Based on Pseudo-Observations* [2] by Andersen, P.K., Hansen, M.G. & Klein, J.P., and *Residuals and Functional Form in Accelerated Life Regression Models* [15] by Lindqvist, B.H., Aaserud, S. & Kvaløy, J.T.

Methods are illustrated with simulated and real data sets. Implementations are made in R.

Chapter 2

Basic Concepts in Survival Analysis

2.1 Survival time

Survival analysis is a branch of statistical methodology that focuses on examining data representing the time between a well-defined start point and the time that we observe a specific event or endpoint. For example how long time an item in a machine is functioning, time until a child takes its first steps or the time to recovery after illness. Its applications can be found in many fields, like medical statistics or reliability theory. In medical statistics we might want to use survival models to study expected time to recovery given different treatments or the probability of surviving longer than a given time. In reliability theory we could be interested in modelling time to failure for a system or a component, or how the probability of instant failure changes with time.

The random variable T that we measure is called the failure time, lifetime or survival time. T does not have to be calendar time like the examples above, but can also stand for the number of kilometers driven by a car, the number of times we use a machine, or the number of times a switch have been switched.

If we let X be a state variable such that

$$\mathrm{X(t)} = \begin{cases} 0 & \text{Event has not been observed} \\ 1 & \text{Event has been observed} \end{cases},$$

we can define the survival time as

$$T = \min_t \{t | X(t) = 1\}.$$

2.2 Censoring

In many situations we do not observe the event for all individuals included in a study. The exact survival time will then be unknown and we say that the observation is censored. Whether an observation i is an event time or a censoring time can be denoted by the event indicator δ_i. If we observe an event we have $\delta_i = 1$ and if we observe a censoring time we have $\delta_i = 0$.

2.2.1 Types of censoring

Censoring can be right, left or interval censoring depending on the range where the survival time is known to lie.

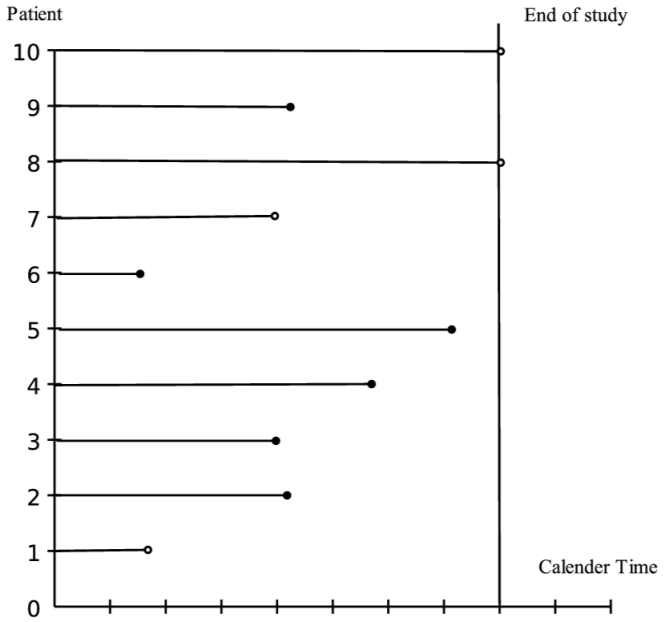

Figure 2.1: Example of survival experiences for ten patients. Right censored observations are represented by circles and observed events are represented by dots

Right censoring is the most common type of censoring in survival analysis, and occurs when the event happens after we stop observing the individual. The censored time will therefore be smaller than the actual survival time. As an example of right censoring we can look at a study of divorces. The couples that are still married when the study ends, or drop out of the study for other reasons than divorce, will be right censored. The censored observations in figure 2.1 are right censored.

A survival time is left censored if the event of interest has happened before we start observing. The true survival time will then be smaller than the observed time. An example is observations of childhood milestones, such

as learning to read. If we observe a group of primary school children some may already have learned to read before starting school and they will be left censored.

The third type, interval censoring, is typical for events that have to be tested. If we find that the event has happened we will not know the exact time of the event, only that it has happened since the last time we tested. An example of this is going to the dentist. Given that a dentist discovers all cavities, we will know that a cavity has occurred between the last time the patient went to the dentist and now, but not the exact time.

An important assumption we will make regarding our survival data is that censoring times are random and independent from the actual survival times. This means that none of the variables that has an effect on the survival will have an effect on the censoring. Assuming this makes the censored times representative for the individuals still at risk (page 4. in Collett [7]).

2.3 Survival and hazard functions

The cumulative distribution function for the survival time T gives us the probability that the event has occurred before time t,

$$F(t) = P(T \leq t) = \int_0^t f(t)dt.$$

Two other functions of particular interest in survival analysis are the survival and hazard functions.

The survival function, $S(t)$, is the probability of not experiencing the event before time t, and will therefore be a right-continuous, non-increasing function of t, with S(0) = 1.

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t). \tag{2.1}$$

In reliability theory the survival function is also known as the reliability function (page 17 in Rausand & Høyland [17]).

The hazard function is the instantaneous probability of death at time t given survival up to time t:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{S(t)} = \frac{-S'(t)}{S(t)}. \tag{2.2}$$

It must be non-negative but can take any shape. Other names for the hazard function is instantaneous death rate, the intensity rate, or the force of mortality (page 11. in Collett [7]).

2.4 Mean survival time

Mean survival time or mean time to failure for an item is defined by

$$MTTF = E[T] = \int_0^\infty tf(t)dt.$$

Because $f(t) = -S'(t)$, we have

$$E[T] = -\int_0^\infty tS'(t)dt.$$

Using partial integration we find

$$E[T] = -[tS'(t)]_0^\infty + \int_0^\infty S(t)dt.$$

It can be shown that if MTTF is finite, then $[tS'(t)]_0^\infty$=0, which gives us

$$\boxed{E[T] = \int_0^\infty S(t)dt.} \tag{2.3}$$

In other situations we are interested in the expected survival time up until a given time τ. This is called the restricted mean and can be found by

$$E[T] = \int_0^\tau S(t)dt. \tag{2.4}$$

Equations 2.3 and 2.4 assume that $T \geq 0$. If survival times are allowed to be negative, we need to find a more general expression for the mean survival time. Let Y be a random variable representing survival times that can be negative. The expected value of Y can be found by:

$$E(Y) = \int_{-\infty}^\infty yf_Y(y)dy = \int_{-\infty}^a yf_Y(y)dy + \int_a^\infty yf_Y(y)dy. \tag{2.5}$$

Each part can then be parted further by partial integration

$$\int_{-\infty}^a yf_Y(y)dy = \Big|_{-\infty}^a yF_Y(y) - \int_{-\infty}^a F(y)dy. \tag{2.6}$$

$$\int_a^\infty yf_Y(y)dy = \Big|_a^\infty - yS_Y(y) - \int_a^\infty -S_Y(y)dy. \tag{2.7}$$

Setting equation 2.6 and 2.7 into equation 2.5 gives

$$E(Y) = \Big|_{-\infty}^a yF_Y(y) - \int_{-\infty}^a F_Y(y)dy - \Big|_a^\infty yS_Y(y) + \int_a^\infty S_Y(y)dy,$$

$$E(Y) = aF_Y(a) + \infty F_Y(-\infty) - \int_{-\infty}^a F_Y(y)dy - \infty S_Y(\infty) + aS_Y(a) + \int_a^\infty S_Y(y)dy.$$

It can be shown that $\infty F_Y(-\infty)$ and $\infty S_Y(\infty)$ equals zero. Using this and that $F_Y(a) + S_Y(a) = 1$, we find an expression for the expectation.

$$\boxed{E(Y) = a - \int_{-\infty}^a F_Y(y)dy + \int_a^\infty S_Y(y)dy.} \tag{2.8}$$

Integrating to τ will give a corresponding restricted mean,

$$E(Y) = a - \int_{-\infty}^a F_Y(y)dy + \int_a^\tau S_Y(y)dy. \tag{2.9}$$

2.5 Likelihood function for right censored data

The likelihood function can be used to find estimators for coefficients in parametric models and will for n uncensored data points have its standard form

$$L = \prod_{i=1}^{n} f(t_i).$$

For censored data sets we need to be more careful. The data sets that we will study will consist of independent and random right censored survival times. We will therefore only consider the likelihood function for such data.

Suppose that r out of n observed survival times are uncensored. The contribution of those r survival times to the likelihood function can be expressed as

$$\prod_{j=1}^{r} f(t_j).$$

For the remaining $n - r$ censored survival times we know that the event time is at least as big as the observed time. Using that the probability for surviving past time t_l is $P(T \geq t_l) = S(t_l)$, we find that the contribution to the likelihood for censored survival times will be

$$\prod_{l=1}^{n-r} S(t_l).$$

Using the event indicator δ_i, we can write the likelihood function as

$$L = \prod_{i}^{n} f(t_i)^{\delta_i} S(t_i)^{(1-\delta_i)}. \tag{2.10}$$

More detailed information on likelihood can be found in section 3.5 in Klein and Moeschberger [14] and appendix B in Collett [7].

Chapter 3

Kaplan-Meier Estimator

3.1 Kaplan-Meier estimator of the survival function for T

The Kaplan-Meier estimator, also called the product limit estimator, is a much used non-parametric method to estimate the survival function for complete and right censored data sets. It is named after Edward L. Kaplan and Paul Meier, who introduced the estimator in their joint paper *Non-parametric Estimation from Incomplete Observations* [12] in 1958.

Based on the explanation in chapter 2 in Collett [7], we will now describe the general idea behind the estimator and illustrate the method with a simple example.

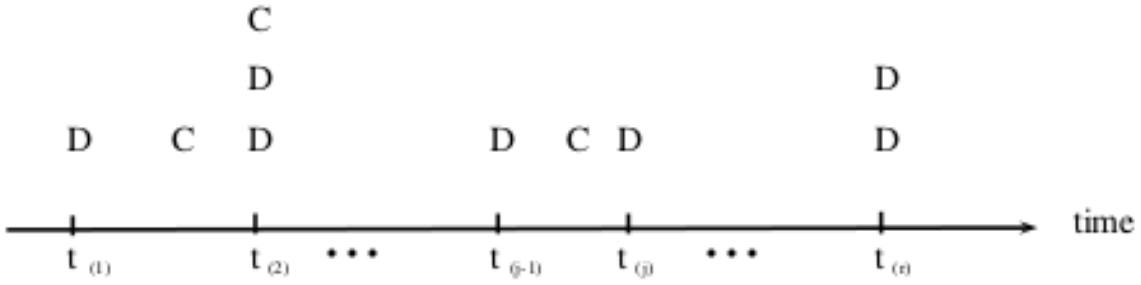

Figure 3.1: Figure illustrating ordering of event times for the Kaplan-Meier estimator. D indicates observed events and C indicates censored events

Suppose that we have n ordered survival times $t_1, t_2 \cdots, t_n$ where some of them may be right censored, and some may be equal. Among these times there are $r \leq n$ distinct event times. We start by arranging the r event times such that $t_{(1)} < t_{(2)} \cdots < t_{(r)}$, denoting the j-th event time as $t_{(j)}$ (see figure 3.1). If there are censored observations at the same time as an event time we will treat the censored time as if it occurred right after the event time. Denote the number of individuals at risk just before time $t_{(j)}$ as n_j, and let

d_j be the number of events at time $t_{(j)}$. The probability of surviving time interval $[t_{(j)}, t_{(j+1)}]$ can then be estimated as $(n_j - d_j)/n_j$. Assuming that events occur independently of each other, we find the probability of surviving intervals $1, 2, \cdots, k$ by $\prod_{j=1}^{k} \frac{n_j - d_j}{n_j}$. This gives us the Kaplan-Meier estimator of the survival function,

$$\hat{S}_{KM}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{j=1}^{k} \frac{n_j - d_j}{n_j} & \text{if } t_{(k)} \leq t < t_{(k+1)} \end{cases}.$$

If the last survival time, t_n, is a censored time, the survival function will be undefined for $t > t_n$. On the other hand, if the largest observed survival time is uncensored, then $t_n = t_{(r)}$, $d_r = n_r$, and the survival function will be zero for $t > t_{(r)}$.

3.1.1 Mean and restricted mean

The Kaplan-Meier estimator can be used to find an estimate of the mean survival time. Inserting the Kaplan-Meier estimate in equation 2.3 gives us

$$\hat{\mu} = \int_0^\infty \hat{S}_{KM}(t)dt. \tag{3.1}$$

When the last observed time t_n is uncensored, $\hat{S}_{KM}(t)$ will be zero for $t > t_n$, and the integral will be finite. On the other hand, if t_n is uncensored the integral will be infinite. To make sure that the expected survival time is finite for both cases, we can use the restricted mean given by equation 2.4, and set $\tau = t_n$ to find,

$$\hat{\mu} = \int_0^{t_n} \hat{S}_{KM}(t)dt. \tag{3.2}$$

3.1.2 Example, Kaplan Meier

In a study of recovery after surgery we get the following observed survival times: $1, 2, 2+, 3, 5, 5, 6, 8+, 9+, 10$, where $+$ indicates right censoring. The uncensored survival times $1, 2, 3, 5, 5, 6, 10$ give us the intervals that we can estimate the survival function for.

In this case equation 3.1 and 3.2 will give the same estimate, $\hat{\mu} = 5.9$, because the last observation is a uncensored time. Plot of the Kaplan-Meier estimator is shown in figure 3.2. In R we can use the *survfit* function from the *survival package* to find Kaplan-Meier estimates. For instructions, see [19] or the *R*-help-guide

Table 3.1: Kaplan-Meier survival function for survival times in the Kaplan-Meier example.

j	t_j	n_j	d_j	$\hat{S}(t)$
1	$0 \leq t < 1$	10	0	1
2	$1 \leq t < 2$	10	1	0.9
3	$2 \leq t < 3$	9	1	0.8
4	$3 \leq t < 5$	7	1	0.686
5	$5 \leq t < 6$	6	2	0.457
6	$6 \leq t < 10$	4	1	0.343
5	$t \geq 10$	1	1	0

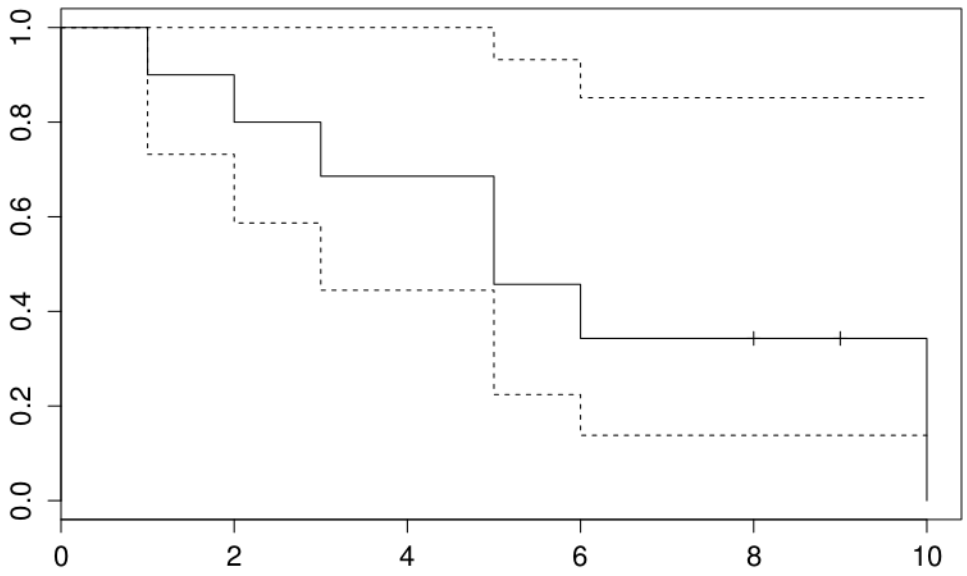

Figure 3.2: Plot of Kaplan-Meier estimator and 95% confidence interval for the survival times in the example. The plot is made using the *survfit* function from the *survival package* in *R*.

3.2 Kaplan-Meier estimator of the survival function for $log(T)$

In some parts of this thesis we will work with the logarithm of the survival times. We will therefore need a Kaplan-Meier estimator for $Y = log(T)$. It can be proven that this estimator will be equal to the Kaplan-Meier estimator for T:

$$S_Y(y) = P(Y > y) = P(log(T) > y) = P(T > e^y) = S_T(e^y).$$

Using the Kaplan-Meier estimator we find that $\hat{S}_{KM,Y}(y) = \hat{S}_{KM,T}(e^y)$. Therefore we have

$$\hat{S}_{KM,Y}(y) = \begin{cases} 1 & \text{if } y < y_{(1)} \\ \prod_{j=1}^{k} \frac{n_j - d_j}{n_j} & \text{if } y_{(k)} \leq y < y_{(k+1)} \end{cases}.$$

3.2.1 Mean and restricted mean

In order to find an estimator for the expected value of $Y = log(T)$ we can use equation 2.8 from chapter 2,

$$E(Y) = a - \int_{-\infty}^{a} F_Y(y)dy + \int_{a}^{\infty} S_Y(y)dy.$$

If we use Kaplan-Meier estimates for $S_Y(y)$ and $F_Y(y)$, and choose $a = log(t_{(1)})$, we find that

$$\hat{\mu} = log(t_{(1)}) + \int_{log(t_{(1)})}^{\infty} \hat{S}_{KM,Y}(y)dy, \tag{3.3}$$

or

$$\hat{\mu} = log(t_{(1)}) + \int_{log(t_{(1)})}^{t_n} \hat{S}_{KM,Y}(y)dy. \tag{3.4}$$

3.2.2 Example, Kaplan-Meier continued

For the same data set as before, we get Kaplan-Meier estimates for $log(T)$ as shown in figure 3.3. We see that the number of intervals and the height of each interval is equal to the the number of intervals and heights in figure 3.2, but the length of each interval have changed.

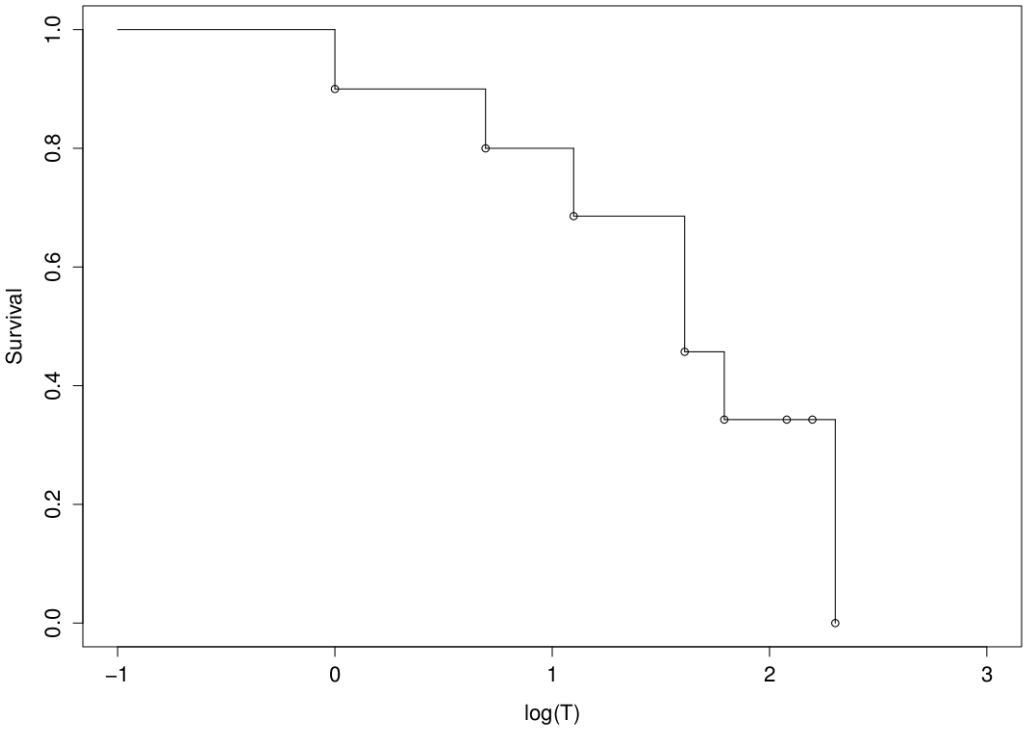

Figure 3.3: Plot of Kaplan-Meier estimated survival curves for log(T), for the same survival times as in figure 3.2. Circles indicates event and censoring times.

Chapter 4

Accelerated Failure Time Models

4.1 Regression models

The parametric distributions introduced in appendix A and non-parametric methods like Kaplan-Meier (chapter 3) provide straightforward methods for modelling survival experiences for homogeneous populations and comparing two or more groups. However, in many situations when we are modelling survival data, we are interested in obtaining survival and hazard functions that accounts for additional information like gender, age and length. This information is often refereed to as covariates, explanatory variables or independent variables. The covariates can be considered as dependent or independent of time. They can be categorical like treatment and ethnicity or continuous like blood pressure or height.

Popular choices of regression models to incorporate the covariates are the proportional hazards model, the additive hazards model, the proportional odds model and the accelerated failure time model. This thesis will focus on accelerated failure time models.

Most of the theory in this chapter is from chapter 6 in Collett [7]. Other books that cover regression models are for example Klein & Moeschberger [14], Meeker & Escobar [16] and Kalbfleisch & Prentice [11].

4.2 General accelerated failure time models

In accelerated failure time models [AFT] we assume that the effect of the covariates will be a multiplication of the expected survival time. A general formulation for the AFT hazard for an individual i with p covariates summarized in vector $\mathbf{x}_i$ is

$$h_i(t) = e^{-\eta i} h_0(t/e^{\eta_i}), \tag{4.1}$$

where

$$\eta_i = \boldsymbol{\alpha}'\mathbf{x_i} = \alpha_1 x_{1i} + \alpha_2 x_{2i} \cdots + \alpha_p x_{pi}.$$

and h_0 is the hazard function for an individual where $\mathbf{x}_i$=0. This h_0 is also called the baseline hazard.

Corresponding survival function will be

$$S_i(t) = S_0(t/exp(\eta_i)) \tag{4.2}$$

where S_0 is the baseline survival function.

4.3 Log-linear representation

For AFT models it is common to use the log-linear representation

$$Y = log(T) = \mu + \boldsymbol{\beta}'\mathbf{X} + \sigma\epsilon. \tag{4.3}$$

where μ and σ are intercept and scale parameter and ϵ is the error term. The $\boldsymbol{\beta}$ are unknown regression coefficients reflecting the effect that each explanatory variable have on the survival time. Positive β_j means that if covariate x_{ji} increases, then the expected survival time increases, and if β_j is negative, an increasing x_{ji} will lead to a decreasing expected survival time.

Survival function for the log-linear representation can be found by using that the baseline survival function S_0 is the survival function for $t = exp(\mu + \sigma\epsilon)$.

$$\begin{aligned} S_i(t) &= P(T_i > t) \\ &= P(Y_i > ln(t)) \\ &= P(\mu + \boldsymbol{\beta}'\mathbf{x_i} + \sigma\epsilon_i > ln(t)) \\ &= P(exp(\mu + \sigma\epsilon_i) > (t/exp(\boldsymbol{\beta}'\mathbf{x_i}))) \\ &= S_0(t/exp(\boldsymbol{\beta}'\mathbf{x_i})). \end{aligned}$$

The hazard function can be found by using equation 2.2, resulting in

$$h_i = exp(-\boldsymbol{\beta}'\mathbf{x_i})h_0\{t/exp(\boldsymbol{\beta}'\mathbf{x_i})\}. \tag{4.4}$$

We see that by setting $\eta_i = \boldsymbol{\beta}'\mathbf{x_i}$ we get hazard and survival functions as in equation 4.1 and 4.2.

The distribution of the error term in equation 4.3 is assumed to be known, and determines the distribution of T and vice versa. Taking $S_{\epsilon_i}(\epsilon)$ as the survival function for the error we find that

$$\begin{aligned} S_i(t) &= P\Big(\mu + \boldsymbol{\beta}'\mathbf{x_i} + \sigma\epsilon_i \geq log(t)\Big) \\ &= P\Big(\epsilon_i \geq \frac{log(t) - \mu - \boldsymbol{\beta}'\mathbf{x_i}}{\sigma}\Big) \\ &= S_{\epsilon_i}\Big(\frac{log(t) - \mu - \boldsymbol{\beta}'\mathbf{x_i}}{\sigma}\Big). \end{aligned} \tag{4.5}$$

4.4 Estimation of log-linear regression models

Estimation of the unknown parameters in the log-linear regression models can be done using maximum likelihood. If we know the distribution of the error we can use it to find the distribution of Y, and hence the likelihood function for Y. To make the derivation easier we set $x_0 = 1$ and let $\beta_0 = \mu$.

Let $f_\epsilon(\epsilon)$ be the PDF of the error term, then the PDF of Y can then be written as

$$f(y) = \sigma^{-1} f_\epsilon(\epsilon).$$

From equation 2.10 we find

$$L(\boldsymbol{\beta}, \sigma) = \prod_{i=1}^{n} [\sigma^{-1} f_\epsilon(\epsilon_i)]^{\delta_i} S_\epsilon(\epsilon_i)^{1-\delta_i}, \tag{4.6}$$

where $\epsilon_i = \frac{log(t) - \boldsymbol{\beta}' \mathbf{X}}{\sigma}$.

The log-likelihood will be

$$l(\boldsymbol{\beta}, \sigma) = \sum_{i,\delta=1} log(f_\epsilon(\epsilon_i)) - \sum_{i,\delta=1} log(\sigma) + \sum_{i,\delta=0} log(S_\epsilon(\epsilon)). \tag{4.7}$$

The partial derivative of the log-likelihood is also called the score statistic U, and will for AFT models be:

$$U_j(\boldsymbol{\beta}, \sigma) = \frac{\partial log(L(\boldsymbol{\beta}, \sigma))}{\partial \beta_j} = \sigma^{-1} \sum_{i=1}^{n} x_{ji} a_i \quad j = 0, 1, \cdots, p$$

$$U_{p+1}(\boldsymbol{\beta}, \sigma) = \frac{\partial log(L(\boldsymbol{\beta}, \sigma))}{\partial \sigma} = \sigma^{-1} \sum_{1}^{n} (\epsilon_i ai - \delta_i)$$

where

$$a_i = -\left[\delta_i \frac{dlog(f_\epsilon(\epsilon_i))}{d\epsilon_i} + (1 - \delta_i) \frac{dlog(S_\epsilon(\epsilon))}{d\epsilon_i} \right]$$

If $l(\boldsymbol{\beta}, \sigma)$ is twice differentiable with respect to all parameters we can also find the observed information matrix I, consisting of the negative second derivatives:

$$-\frac{\partial^2 log(L(\boldsymbol{\beta}, \sigma))}{\partial \beta_j \partial \beta_k} = -\sigma^{-2} \sum_{i=1}^{n} x_{ji} x_{ki} \frac{da_i}{d\epsilon_i}$$

$$-\frac{\partial^2 log(L(\boldsymbol{\beta}, \sigma))}{\partial \beta_j \partial \sigma} = \sigma^{-2} \sum_{i=1}^{n} x_{ji} \epsilon_i \frac{da_i}{d\epsilon_i} + \sigma^{-1} U_j(\boldsymbol{\beta}, \sigma)$$

$$-\frac{\partial^2 log(L(\boldsymbol{\beta}, \sigma))}{\partial \sigma^2} = \sigma^{-2} \sum_{i=1}^{n} (\epsilon_i^2 \frac{da_i}{d\epsilon_i} + \delta_i) + 2\sigma^{-1} U_{p+1}(\boldsymbol{\beta}, \sigma)$$

Now that we have the score and observed information matrix, the maximum likelihood estimates is found by setting U=0 and $I > 0$. A method for solving this numerical is the Newton-Raphson method:

$$\boldsymbol{\beta}^{(m)} = \boldsymbol{\beta}^{(m-1)} + \frac{U^{m-1}}{I^{m-1}}$$

The theory in this section is mainly from section 3.6 in Kalbfleisch & Prentice [11]. More on estimation of parameters and numerical methods can be found in appendix A in Klein & Moeschberger [14] and chapter 4 in Dobson & Barnett [8].

4.4.1 Estimation with R

In *R*, the *survreg* function from the *survival* package fits a log-linear accelerated failure time model using maximum likelihood and the Newton-Raphson method. Section 5.7 in Therneau's paper [18] gives a detailed description of the implementation and the theory behind *survreg*, while the *R*-help-guide give as brief introduction on how to use the *survreg* function.

With *survreg* we can fit Weibull (and hence exponential and Rayleigh), loglogistic and lognormal survival data. We will now look at an example where we try to fit a Weibull distribution to

$$Y = log(T) = \mu + \beta_1 x_1 + \beta_2 x_2 + \sigma\epsilon$$

The *R*-code for this will be:

```
survreg(Surv(times,events)~x_1+x_2,dist="weibull")
```

and the output gives us

```
Call:
survreg(formula = Surv(Data$Time, Data$Status) ~ Data$x1 + Data$x2 +
    Data$x3, data = Data, dist = "weibull")
              Value Std. Error      z        p
(Intercept)  2.0136      0.190 10.612 2.62e-26
Data$x1      0.0557      0.189  0.295 7.68e-01
Data$x2     -0.3560      0.206 -1.727 8.42e-02
Data$x3      0.2188      0.298  0.735 4.62e-01
Log(scale)  -0.6048      0.188 -3.211 1.32e-03

Scale= 0.546

Weibull distribution
Loglik(model)= -46.1   Loglik(intercept only)= -47.5
```

```
Chisq= 2.85 on 3 degrees of freedom, p= 0.41
Number of Newton-Raphson Iterations: 5
n= 20
```

Another possibility is to use the likelihood or log-likelihood function from equation 4.6 or 4.7, and use the *optim* function from the *stats* package to maximize and find the estimated parameters. *optim*s default setting is to minimize the function it's given as input, so we must multiply the likelihood with minus one in order to get maximum. The initial guesses should be carefully chosen, and the result should be treated with some scepticism because the result might be a local maximum instead of a global.

4.5 Residuals

Residuals and residual plots are widely used tools when evaluating models. We will now look at two residuals used for accelerated failure time models, the standardized and Cox-Snell residual. When working with survival models we must be aware of the fact that censored survival times will lead to censored residuals.

4.5.1 Standardized residuals

Standardized residuals come from solving equation 4.3 with respect to ϵ, and are therefore defined as

$$R_s = \frac{log(T) - \mu - \boldsymbol{\beta}'\mathbf{X}}{\sigma}. \tag{4.8}$$

When we have estimated parameters $\hat{\mu}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$, we find standardized residuals as

$$\hat{r}_{s,i} = \frac{log(t_i) - \hat{\mu} - \hat{\boldsymbol{\beta}}'\mathbf{x_i}}{\hat{\sigma}}. \tag{4.9}$$

These residual are assumed to have the same distribution, Φ_ϵ, as ϵ in equation 4.3 if the estimated model is appropriate. For example, if T is Weibull distributed the standardized residuals should be Gumbel distributed.

When the data set is right censored, some $log(t_i)$ will be smaller than the logarithm of the actual survival time and hence, some standardized residuals will be too small.

4.5.2 Cox-Snell residuals

Cox-Snell residuals are based on the random variable $-log(F(T))$, which will be unit exponentially distributed for all Φ_ϵ (section 3 in [15]). Noting that

$$F(t|\mathbf{X}) = 1 - \Phi_\epsilon\Big(\frac{log(t) - \mu - \boldsymbol{\beta}'\mathbf{X}}{\sigma}\Big),$$

implies that

$$-log(F(t|\mathbf{X})) = -log\Big(1 - \Phi_\epsilon\Big(\frac{log(t) - \mu - \boldsymbol{\beta}'\mathbf{X}}{\sigma}\Big)\Big).$$

This lead us to the Cox-Snell residuals

$$\hat{r}_{c,i} = -log\Big(1 - \Phi_\epsilon\Big(\frac{log(t_i) - \hat{\mu} - \hat{\boldsymbol{\beta}}'\mathbf{x_i}}{\hat{\sigma}}\Big)\Big), \tag{4.10}$$

which should behave like a censored sample of unit exponential survival times. An equivalent expression is

$$\hat{r}_{c,i} = -log\Big(S_\epsilon\Big(\frac{log(t_i) - \hat{\mu} - \hat{\boldsymbol{\beta}}'\mathbf{x_i}}{\hat{\sigma}}\Big)\Big),$$

and expressed by the standardized residuals we have

$$\hat{r}_{c,i} = -log(S_\epsilon(\hat{r}_{s,i})).$$

From this we find that assessing whether the standardized residuals has a certain distribution will be the same as assessing if the Cox-Snell residuals have an unit exponential distribution. More on the connection between the two residuals can be found in appendix A in Lindqvist et al. [15].

4.5.3 Adjusted Cox-Snell residuals

As for the standardized residuals, a censored Cox-Snell residual will be smaller than the "actual" residual. To compensate for this, one can add a given value to the censored residuals. Examples of that are the 1- and log(2)-adjusted Cox-Snell residuals.

The 1-adjusted Cox-Snell residual is found simply by adding 1 to the censored residuals. This is done because of the "memoryless" property of the exponential distribution and the fact that the unit exponential distribution has expectation 1. Using the event indicator this can be written as

$$r_{adj,i} = r_i + (1 - \delta_i).$$

The log2-adjusted Cox-Snell residuals are found by adding the median value of the exponential distribution to the censored residuals, giving us

$$r_{adj2,i} = r_i + (1 - \delta_i)log(2).$$

4.6 Weibull accelerated failure time models

Weibull is one of the most used distributions in survival analysis. It is the only distribution that can be expressed as both an accelerated failure time model and a proportional hazard model (see section 6.5 in Collett [7]).

When a survival time, T, is Weibull distributed, we know that ϵ is Gumbel distributed. This can be shown using equation 4.5 and the Gumbel survival function from equation A.3.

$$\begin{aligned} S_i(t) &= S_{\epsilon_i}\left(\frac{log(t) - \mu - \mathbf{x_i}'\boldsymbol{\beta}}{\sigma}\right) \\ &= exp\Bigg[- exp\left(\frac{log(t) - \mu - \mathbf{x_i}'\boldsymbol{\beta}}{\sigma}\right)\Bigg] \\ &= exp\Big(- \lambda_i t^{\frac{1}{\sigma}}\Big), \end{aligned}$$

where $\lambda_i = exp\Big(\frac{-\mu - \mathbf{x_i}'\beta}{\sigma}\Big)$. We see that this is a Weibull distribution with scale parameter λ_i and shape parameter σ^{-1}. Using this and equation A.2 we find that the Weibull AFT hazard function will be

$$h_i(t) = \lambda_i \sigma^{-1} t^{\sigma^{-1}-1}.$$

4.6.1 Residuals

Because ϵ is Gumbel distributed we know that the standardized residuals given in equation 4.8 will behave as a (censored) Gumbel distributed sample if the model is appropriate.

The Cox-Snell residuals will according to equation 4.10 and A.3 be

$$\hat{r}_{c,i} = -log\Big(S_{\epsilon_i}\Big(\frac{log(t_i) - \hat{\mu} - \mathbf{x_i}'\hat{\boldsymbol{\beta}}}{\hat{\sigma}}\Big)\Big) = exp\Big(\frac{log(t_i) - \hat{\mu} - \mathbf{x_i}'\hat{\boldsymbol{\beta}}}{\hat{\sigma}}\Big) = exp(\hat{r}_{s,i})$$

and they will be unit exponential distributed if the model is appropriate.

4.6.2 The distribution and likelihood of Y

We will now show that a Weibull distributed T leads to a Gumbel distributed Y, and derive an expression for the likelihood for Y.

From appendix A we know that the probability density function for a Weibull distributed survival time is

$$f_T(t_i) = \lambda_i \gamma t_i^{\gamma-1} exp(-\lambda t_i^{\gamma}),$$

where $\lambda_i = \lambda(x_{i.})$ depends on the covariates. Using transformation $T = e^Y$ we can find the distribution for Y:

$$\begin{aligned} F_Y(y_i) &= P(Y \leq y_i) = P(log(T) \leq y_i) = P(T \leq e^{y_i}) \\ &= \int_0^{e^{y_i}} f(t_i)dt = \int_0^{e^{y_i}} \lambda_i^{\gamma-1} exp(-\lambda t_i^{\gamma} dt \end{aligned}$$

.

Substituting with $u = \lambda_i t_i^{\gamma}$ gives us:

$$\begin{aligned} F_Y(y_i) &= \int_{t_i=0}^{e^{y_i}} e^{-u}du = \Big|_{t_i=0}^{e^{y_i}} - e^{-u} = \Big|_0^{e^{y_i}} - e^{\lambda_i t_i^{\gamma-1}} \\ &= 1 - e^{-\lambda_i e^{y_i \alpha}} = 1 - e^{-e^{\frac{y_i + (1/\gamma) log(\lambda_i)}{1/\gamma}}}. \end{aligned}$$

Letting $\sigma = 1/\gamma$ and $-(1/\gamma)log(\lambda_i) = \boldsymbol{\beta}'\mathbf{x_i}$ we see that Y is Gumbel distributed with

$$F_Y(y_i) = 1 - exp\Big[- exp\Big(\frac{y_i - \mathbf{x_i}'\boldsymbol{\beta}}{\sigma}\Big)\Big].$$

giving us survival function

$$S_Y(y_i) = exp\Big[- exp\Big(\frac{y_i - \mathbf{x_i}'\boldsymbol{\beta}}{\sigma}\Big)\Big],$$

and density function

$$f_Y(y_i) = \frac{1}{\sigma} exp\Big[\frac{y_i - \mathbf{x_i}'\boldsymbol{\beta}}{\sigma} - exp\Big(\frac{y_i - \mathbf{x_i}'\boldsymbol{\beta}}{\sigma}\Big)\Big].$$

The likelihood function for Y can then, according to equation 2.10, be written on the form:

$$L(\boldsymbol{\beta}, \sigma) = \prod_{i,\delta_i=1} \frac{1}{\sigma} exp\Big[\frac{y_i - \mathbf{x_i}'\boldsymbol{\beta}}{\sigma} - exp\Big(\frac{y_i - \mathbf{x_i}'\boldsymbol{\beta}}{\sigma}\Big)\Big] \prod_{i,\delta_i=0} exp\Big[-exp\Big(\frac{y_i - \mathbf{x_i}'\boldsymbol{\beta}}{\sigma}\Big)\Big],$$

and log-likelihood:

$$\ell(\boldsymbol{\beta}, \sigma) = log(L(\boldsymbol{\beta}, \sigma)).$$

$$\ell(\boldsymbol{\beta}, \sigma) = \sum_{i,\delta_i=1} \Big[\frac{y_i - \mathbf{x_i}'\boldsymbol{\beta}}{\sigma} - exp\Big(\frac{y_i - \mathbf{x_i}'\boldsymbol{\beta}}{\sigma}\Big)\Big] - \sum_{i,\delta_i=1} log(\sigma) - \sum_{i,\delta_i=0} exp\Big(\frac{y_i - \mathbf{x_i}'\boldsymbol{\beta}}{\sigma}\Big).$$

This result could also be found by equation 4.6 and 4.7.

4.7 Lognormal accelerated failure time models

From the definition of a lognormal distribution we know that if T is lognormal, then $Y = log(T)$ is normal distributed. We will now show that a normal distributed ϵ will lead to a lognormal T:

$$\begin{aligned} S_i(t) &= S_{\epsilon_i}\left(\frac{log(t) - \mu - \boldsymbol{\beta}'\mathbf{x_i}}{\sigma}\right) \\ &= 1 - \Phi\left(\frac{log(t) - \mu - \boldsymbol{\beta}'\mathbf{x_i}}{\sigma}\right) \end{aligned}$$

,

which we see from equation A.4 is the survival function for a lognormal distributed survival time with parameters $\mu + \boldsymbol{\beta}'\mathbf{x_i}$ and σ.

4.7.1 Residuals

If the model is appropriate, the standardized residuals (see equation 4.8) should behave as a (censored) sample from a normal distribution.

Cox-Snell residuals will be

$$\hat{r}_{c,i} = -log(1 - \Phi(\hat{r}_{s,i})),$$

and they will behave as from a (censored) unit exponential distribution if the model is appropriate.

4.7.2 Likelihood function for Y

The distribution of ϵ, is as mentioned above, normal. This gives us

$$f_\epsilon(\epsilon) = \frac{1}{2\pi}e^{-\frac{1}{2}\epsilon^2}$$

$$S_\epsilon(\epsilon) = 1 - \Phi(\epsilon)$$

From equation 4.6 and 4.7 we then get

$$L(\boldsymbol{\beta}, \sigma) = \prod_{i=1}^{n}\left[\sigma_i^{-1}\frac{1}{2\pi}e^{-\frac{1}{2}\epsilon_i^2}\right]^{\delta_i}\left[1 - \Phi(\epsilon_i)\right]^{1-\delta_i}.$$

The log-likelihood will be

$$l(\boldsymbol{\beta}, \sigma) = \sum_{i,\delta=1} log\Big(\frac{1}{2\pi}e^{-\frac{1}{2}\epsilon_i^2}\Big) - \sum_{i,\delta=1} log(\sigma_i) + \sum_{i,\delta=0} log\Big(1 - \Phi(\epsilon_i)\Big).$$

Inserting $\epsilon_i = \frac{log(t_i) - \boldsymbol{\beta}'\mathbf{X}_i}{\sigma_i}$ gives us

$$L(\boldsymbol{\beta}, \sigma) = \prod_{i=1}^{n}\left[\sigma_i^{-1}\frac{1}{2\pi}e^{-\frac{1}{2}\left(\frac{log(t_i) - \boldsymbol{\beta}'\mathbf{X}_i}{\sigma_i}\right)^2}\right]^{\delta_i}\left[1 - \Phi\Big(\frac{log(t_i) - \boldsymbol{\beta}'\mathbf{X}_i}{\sigma_i}\Big)\right]^{1-\delta_i}.$$

and

$$l(\boldsymbol{\beta},\sigma) = \sum_{i,\delta=1} log\Big(\frac{1}{2\pi}e^{-\frac{1}{2}\left(\frac{log(t)-\boldsymbol{\beta}'\mathbf{x}}{\sigma}\right)^2}\Big) - \sum_{i,\delta=1} log(\sigma_i) + \sum_{i,\delta=0} log\Big(1-\Phi\Big(\frac{log(t_i)-\boldsymbol{\beta}'\mathbf{X}_i}{\sigma_i}\Big)\Big).$$

4.8 Simulating survival data from accelerated failure time models

Simulated data sets from AFT models will be used frequently in chapters 5, 6 and 7. We will therefore look at how we can use R to generate survival data from AFT models.

Unless stated otherwise, data sets in this thesis will be based on the AFT model

$$log(T) = \mu + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \sigma\epsilon, \tag{4.11}$$

where x_1 is binary with $P(x_1 = 0) = P(x_1 = 1) = 0.5$, $x_2 \sim unif[-1, 1]$ and $x_3 \sim N(0, 1)$. Parameters μ, β_1, β_2, β_3 and σ will be known, and ϵ depends on the distribution of T.

Finding covariates for n observations in R can be done using the following code:

```
x1 <-round(runif(n,0,1))
x2 <-runif(n,-1,1)
x3 <-rnorm(n)
```

The n values of ϵ must also be simulated. If ϵ is Gumbel distributed, Aaserud [1] warns us not to use the $rgumbel(n)$ function, but instead implement our own simulating procedure. This is done by letting $u \sim unif[0, 1]$ and use that $\epsilon = ln[-ln[u]]$ will be Gumbel distributed.

When we have values for all covariates, parameters and ϵ, we can find simulated survival times T_{real} by inserting the values in model 4.11.

Censoring is found by generating n censoring times and letting the observed survival time be

$$T = min(T_{real}, C).$$

We have used exponential censoring times C, found by setting $v \sim unif[0, 1]$ and $C = (-1/\lambda)log(v)$. By adjusting the value of λ we can decide the level of censoring.

Code for simulating Weibull and lognormal data sets can be found in appendix D.

Chapter 5

Pseudo Observations

5.1 Introduction to pseudo observations

Although there are several good methods and tools available to treat censored data sets, there have been studies of methods that allow us to treat censored data as if they where complete. One such method is to create a synthetic data set using pseudo observations known from jackknife theory. This new synthetic data set can then be treated as an uncensored version of the original data set, and ordinary regression methods can be used to estimate regression coefficients. Further more, standard methods for assessing the fit of the model can be used, including standard residuals and plots.

5.1.1 Jackknife and pseudo observations

Let $\mathbf{t} = (t_1, t_2, \cdots t_n)$ be a sample of observations of the random variable T, and $\hat{\theta} = \hat{\theta}(T)$ be an approximately unbiased estimator of $\theta = E[f(T)]$. By removing one observation at the time we get n jackknife samples

$$\mathbf{t}_i = (t_1, t_2, \cdots t_{i-1}, t_{i+1}, \cdots t_n),$$

for i= 1,2,$\cdots$ n.

Using these jackknife samples we can find n jackknife replications, or "leave-one-out" estimators, of $\hat{\theta}(T)$,

$$\hat{\theta}_{-i} = \hat{\theta}(\mathbf{t}_i).$$

These jackknife replications can then be used to find the n independent data values, $\hat{\theta}_i$, that we call pseudo observations of $f(T)$,

$$\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{-i}. \tag{5.1}$$

This set of pseudo observations is our new synthetic data set, and the average of the new values is what is called the jackknife estimator of θ.

$$\hat{\theta}_J = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_i.$$

For more theory on jackknife see for example Efron [9] or Garthwaite et al. [10].

5.1.2 Pseudo observations in survival analysis

In survival analysis censored data sets will consist of some incomplete observations of the survival times. This will lead to incomplete observations of any function $f(T)$ of the survival time. The idea is to replace all observations of $f(T)$ with the pseudo observations $\hat{\theta}_i$. $\hat{\theta}_i$ can then be used as an outcome variable in a generalized linear model or in model assessing by computing residuals and scatter plots.

Andersen and Perme's article *Pseudo-observations in survival analysis* [4], serves as a good introduction to different aspects of pseudo observations in survival analysis. It presents a review of recent work on applications of pseudo observations, including examples of applications in regression models and numerical and graphical methods for assessing goodness of fit. As examples of applications in survival analysis, Andersen and Perme look at the survival function, restricted mean survival time, competing risks cumulative incidences and the illness-death model. A more detailed article on restricted mean and pseudo observations is Andersen, Hansen and Klein's paper *Regression analysis of restricted mean survival time based on pseudo observations* [2]. This chapter of the thesis will start out based on this article, and look at three methods for finding pseudo observations of survival times in AFT models. First we will give a short introduction to the methods, and in the three next sections study them in detail.

To find pseudo observations for the survival time we set $f(T) = T$ and

$$\theta = E(f(T)) = E(T) = \mu.$$

As an estimator $\hat{\theta}$, we can use the Kaplan-Meier estimate for restricted mean with $\tau = t_n$ as given in equation 3.2:

$$\hat{\mu} = \int_0^{t_n} \hat{S}_{KM}(t)dt,$$

where t_n is the largest observation (censored or uncensored). Pseudo values can then be found using equation 5.1. This method will be referenced to as the KM-method.

The pseudo values that we get with the KM-method can then be used as uncensored observations in estimation of the parameters in AFT model 4.3. A problem with this method is that we might get negative pseudo observations, and most of the software made for survival analysis require positive survival times. A solution to this is to find pseudo observations for $log(T)$ and then exponentiate them to find pseudo observations for T. In [2], Andersen et al. suggest using an integrated Kaplan-Meier for $log(T)$. This estimator is derived in section 3.2 resulting in equation 3.4,

$$E(\widehat{log(T)}) = log(t_{(1)}) + \int_{log(t_{(1)})}^{t_n} \hat{S}_{log(T)}(v)dv.$$

The pseudo observations we get will now be observations of $log(T)$. Taking the exponential of this gives us pseudo observations for T, which will be positive. We will refer to this method as KMlog.

KM and KMlog are both based on non-parametric estimation, and can therefore be used on data sets where the distribution and model is unknown. However, if we assume a certain distribution and model for the survival times T, we can use the parametric distribution of the accelerated failure time model (see equation 4.3) to find pseudo observations. Setting

$$\theta = E(log(T)) = \mu + \beta_1 E(x_1) + \cdots + \beta_p E(x_p) + \sigma E(\epsilon_i),$$

we can find an estimator for θ as

$$\hat{\theta} = E(\widehat{log(T)}) = \hat{\mu} + \hat{\beta}_1 \widehat{E(x_1)} + \cdots + \hat{\beta}_p \widehat{E(x_p)} + \hat{\sigma} \widehat{E(\epsilon_i)}.$$

Covariates for all n observations are assumed to be known, so $\widehat{E(x_j)} = \frac{1}{n}\sum_{j=1}^{n} x_{ji}$ can be used for $j = 1, 2, ..., p$. Expected value for ϵ is known given the distribution of the survival times, and estimators for the parameters can be found using maximum likelihood or other estimation methods. This method for finding pseudo observations will be referred to as the parametric method. Again we find pseudo observations for log(T) by equation 5.1 and exponentiate them to get pseudo observations for T. A drawback with this method is that for each jackknife replication $\hat{\theta}_{-i}$, we need to find new estimators for x_i, σ, μ and β_i, and the method must be adjusted to each different AFT model.

5.2 Pseudo observations based on Kaplan-Meier for T

Pseudo observations based on Kaplan-Meier can be found using the following procedure:

- Find an estimator, $\hat{\theta}$, for the expected survival time for the full data set using equation 3.2.
- For each observation i: Remove observation i from the original data set and find the "leave-one-out" estimator, $\hat{\theta}_{-i}$, for the expected survival time using equation 3.2.
- Find pseudo observations, $\hat{\theta}_i$, using equation 5.1.

5.2.1 Uncensored observations

For uncensored observations with no ties it can be shown that pseudo observations based on Kaplan-Meier will be equal to the survival times.

Assume that we have a set of n observations sorted in increasing order $t_1 < t_2 < \cdots < t_n$. Because we assume no tied observations we know that only one individual will experience the event at each time, so $d_i = 1$ for all intervals i. The number of individuals at risk n_i will decrease with one for each interval because there is no censoring. Hence

$$\hat{S}(t_j) = \prod_{i=1}^{j} \frac{n-j}{n-j+1} = \frac{n-j}{n-j+1}\frac{n-j+1}{n-j+2}\cdots\frac{n-1}{n} = \frac{n-j}{n}$$

Survival for times $t < t_1 = 1$ as usual. An estimate of the mean can be found by

$$\begin{aligned}\hat{\theta} =& t_1 + (t_2 - t_1)\frac{(n-1)}{n} + (t_3 - t_2)\frac{(n-2)}{n} + \cdots + (t_j - t_{j-1})\frac{(n-j)}{n} \\ &+ (t_j - t_{j+1})\frac{(n-(j+1))}{n} + (t_{j+1} - t_{j+2})\frac{(n-(j+2))}{n} + \cdots \\ &+ (t_{n-1} - t_{n-2})\frac{(n-(n-1))}{n} + (t_n - t_{n-1})\frac{1}{n}.\end{aligned}$$

This gives us

$$n\hat{\theta} = t_1 + t_2 + \cdots + t_{j-1} + t_j + t_{j+1} + \cdots + t_{n-1} + t_n.$$

For the leave-one-out estimators $\hat{\theta}_{-i}$ we have $n = n-1$. For the j-th estimator we have

$$\begin{aligned}\hat{\theta}_{-j} =& t_1 + (t_2 - t_1)\frac{(n-2)}{n-1} + (t_3 - t_2)\frac{(n-3)}{n-1} + \cdots + (t_{j+1} - t_{j-1})\frac{(n-j)}{n-1} \\ &+ (t_{j+2} - t_{j+1})\frac{(n-(j+1))}{n} + (t_{j+3} - t_{j+2})\frac{(n-(j+2))}{n-1} + \cdots \\ &+ (t_{n-1} - t_{n-2})\frac{(n-(n-1))}{n-1} + (t_n - t_{n-1})\frac{1}{n-1}.\end{aligned}$$

This gives us

$$(n-1)\hat{\theta}_{-j} = t_1 + t_2 + \cdots + t_{j-1} + t_{j+1} + \cdots + t_{n-1} + t_n,$$

and pseudo observations

$$\hat{\theta}_j = n\hat{\theta} - (n-1)\hat{\theta}_{-j} = t_j.$$

This also holds for $t = t_1$ and $t = t_n$, so the conclusion is that pseudo observations based on uncensored data sets will be identical to the original data set.

5.2.2 Censored observations

When we use the KM-method we can not utilize any knowledge about covariates or the distribution of the survival times to find pseudo observations. Pseudo observations will therefore rely entirely on the set of observed survival times and censoring status (t_i, δ_i) for $i = 1 \cdots, n$. In order to be a good replicate of the real survival times we want pseudo observations for uncensored observations to be approximately the same as the observed time, and pseudo observations for censored observations to be higher than the observed time. To obtain an impression of how pseudo observations from the KM-method works, pseudo observations for six Weibull distributed data sets are plotted in figure 5.1. They are sorted after increasing observed survival times because it makes it easier to see how pseudo observations depend on the observed time. From the figure we see that censored observations almost always give pseudo observations that are larger than the observed survival times, which is what we want them to be. For uncensored survival times, pseudo observations are both larger and smaller than the observed times. In most cases they are smaller for small observed survival times and larger for large observed survival times.

Another thing that we notice, is that all uncensored pseudo observations smaller than the smallest censored observations seem to be equal to the observed survival time. Also, when there are two or more censored observations between two successive uncensored observations, the censored observations have the same pseudo observation value.

The R code used to find these pseudo observations are included in appendix D. One aspect of the code that one should notice is that the restricted mean is taken to be the largest observed survival time in the data set. This is the case for each jackknife sample, so the n-th jackknife replication ,θ_{-n}, will be found by integration to t_{n-1} instead of t_n. This differs from the R code in the *Pseudo*-package created by Klein et al. [13], where they always integrate to t_n. Therefore results found by using those R-codes may differ from the ones found with R-codes from the appendix. In particular, Andersen en al., found that all pseudo observations corresponding to uncensored observations are smaller than the observed time (page 340 in [2]), while figure 5.1 shows that they can be larger.

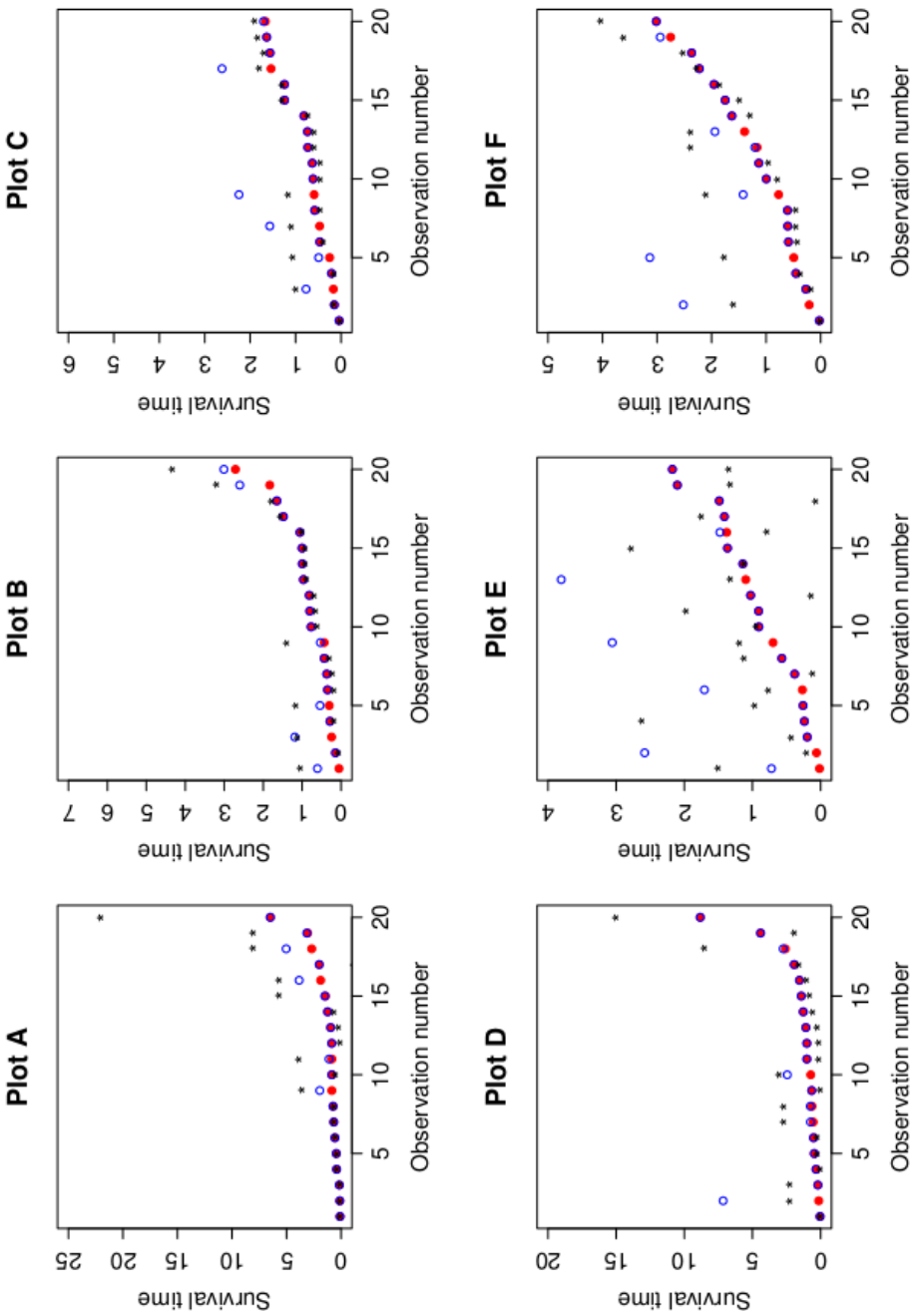

Figure 5.1: Plot of KM pseudo observations (*), real survival times (blue ○) and observed survival times (red ●) for 6 different data sets, all with 30% censoring and $\sigma = 2/3$.

5.2.2.1 A more detailed study of some KM pseudo observation

In the previous paragraph, we made several observations concerning the behaviour of pseudo observations found with the KM-method. To study these observations in detail we now turn our attention to a small data set with 20 observations from a Weibull distribution. This data set will be referenced to as W1, and covariates and results for this data set can be found in appendix B. This data have similar properties to the data sets we studied in figure

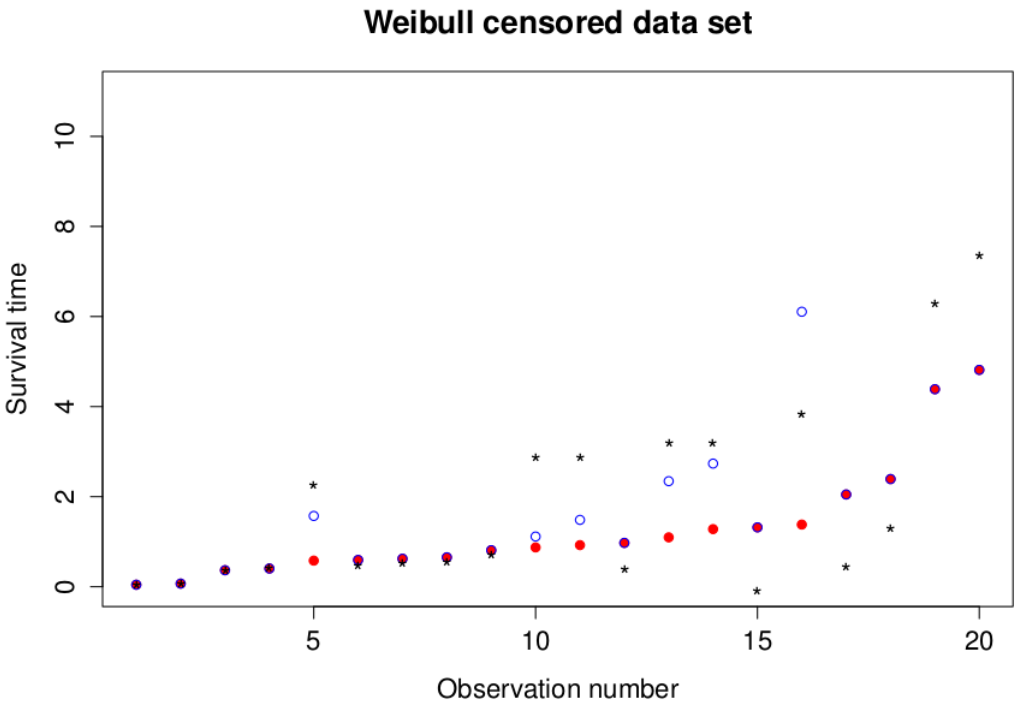

Figure 5.2: Pseudo observation (*), Real times (blue ∘) and observed times (red •) for the simulated W1 data set sorted after ascending observed survival time.

5.1. The first four observations are uncensored and have pseudo observations equal to the observed times. Then the uncensored observations have smaller pseudo observations than observed survival time, even one negative, and then they are larger than the observed times for the last two uncensored observations. All censored survival times have pseudo observations that are larger than the observed time. When two censored observations are between two successive uncensored times, they have the same pseudo observation value. The reason for this lies in the Kaplan-Meier survival function. Figure 5.3 shows the Kaplan-Meier survival curve for the full data set and some selected jackknife samples.

The equation for finding pseudo observations, (5.1), says

$$\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{-i}$$

where $\hat{\theta}$ and $\hat{\theta}_{-i}$ are estimates of the expected survival times for the full data set, and the data set where we have removed observation i. For the KM-method $\hat{\theta}$ is the area under the Kaplan-Meier curve for the full data set, and $\hat{\theta}_{-i}$ is the area under the curve for the reduced data set.

Among the first four observations, observation 2 and 3 are chosen as examples. The Kaplan-Meier curves for the corresponding jackknife samples are plot B and C in 5.3, and plot A is the full data set. There is a small difference between the area under the curve for plot A and B and plot A and C. First of all, the number of steps in curve A and B is reduced by one compared to A because we remove an uncensored observation. The height of the steps are also different because the estimated survival at the intervals change. For the first four observations the difference between $n\hat{\theta}$ and $(n-1)\hat{\theta}_{-i}$ will be t_i. Comparing pseudo observations and observed survival times in table B.1 in appendix B, we see that for the first four observations, KM pseudo observations are equal to the observed survival times.

In table 5.1 we find $\hat{\theta}$, $\hat{\theta}_{-i}$ and pseudo observation for the same selected jackknife samples from the W1 data set as shown in figure 5.3.

Table 5.1: Selected estimated expected survival times and pseudo observations from the W1 data set. Note that the values are rounded off.

Data set	Estimated expected survival	Pseudo observation
A (full)	1.85	
B (2)	1.94	0.07
C (3)	1.93	0.37
D (10)	1.83	2.25
E (11)	1.79	2.87
F (12)	1.79	2.87
G (15)	1.95	-0.09
H (20)	1.56	7.35

Observation 5 is a censored observation, so removing it will not change the number of steps in the Kaplan-Meier curve. It will, however, change the height of the steps because removing an observation, censored or uncensored, always changes the estimated survival function. Looking at the plots in figure 5.3 we see that the height of the levels in plot A are higher than in plot D. This makes $\hat{\theta}$ larger than $\hat{\theta}_{-5}$ and hence,

$$n\hat{\theta} - (n-1)\hat{\theta}_{-5} = \hat{\theta} + (n-1)(\hat{\theta} - \hat{\theta}_{-5}) > \hat{\theta}.$$

This is why almost all pseudo observations for censored times are larger than the observed survival time.

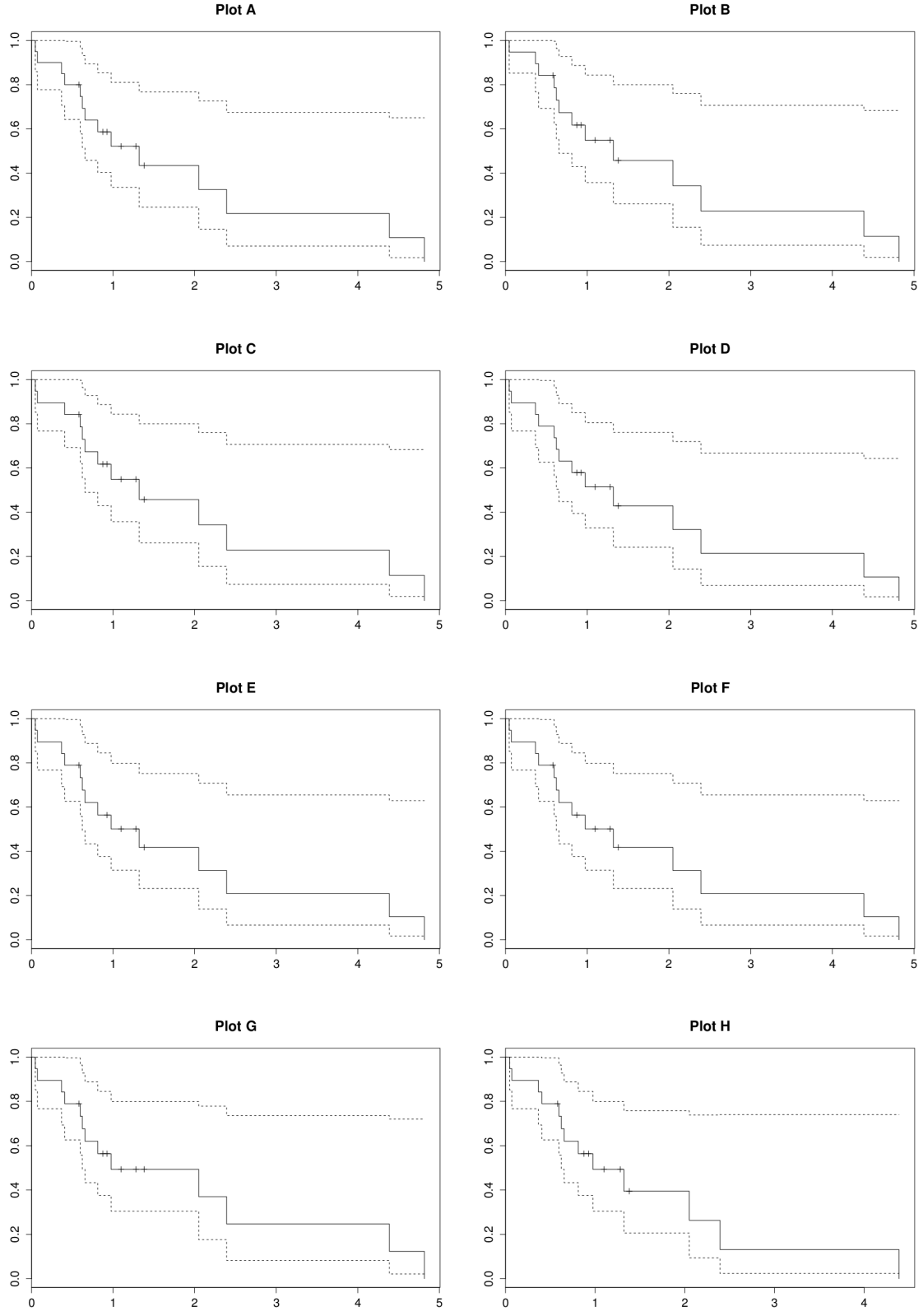

Figure 5.3: Figures showing Kaplan-Meier curves for data set W1: A) The full data set, B) Jackknife replication 2, C) Jackknife replication 3, D) Jackknife replication 5, E) Jackknife replication 10, F) Jackknife replication 11, G) Jackknife replication 15, H) Jackknife replication 20.

For observation 15, which is a uncensored observation, we see that both the number of steps, and the height of each step changes. This is because it counts as an event time. The area under curve G is larger than the area under curve A. This will lead to small pseudo observations, even negative when the difference is big enough, which it is for observation 15. By looking at equation 5.1 we see that in fact all pseudo observations where

$$\frac{n}{n-1}\hat{\theta} < \hat{\theta}_{-i}$$

will lead to negative pseudo observations.

Removing data points with high observed survival time can have a large impact on the estimated expected survival, especially for the last observation. From figure 5.3 we see that removing observation 20 changes the length of the curve. This makes $\hat{\theta}_{-20}$ small and hence, we get a large $\hat{\theta}_{20}$.

From figure 5.2 we also see that observation 10 and 11 and observation 13 and 14 lead to the same pseudo observation values. This will always be the case when we have two or more censored observations in a row between two succesive uncensored observations. Removing a censored observation will not influence the number of steps in the KM curve, and because the censored observations are in a row, removing either one of them will have the same effect on the probability of survival. In table 5.1 we see that both observation 10 and observation 11 give $\hat{\theta}_{-i}$= 1.79, and therefore a pseudo observation of 2.87.

5.2.2.2 Changes in observations

Occasionally when collecting data, some of the observations might be wrong. For example, one could mistake a censoring time for an observed time, the other way around, or the censoring time that is observed may be too small or too large for several reasons. A mistake in one value will influence the value of the pseudo observations for other points, as well as the one that is mistaken. This is because we use all observations to calculate $\hat{\theta}$, and all but one to calculate $\hat{\theta}_i$.

In figure 5.4 and 5.5 we see plots of the original pseudo observations against pseudo observations where observation 16 have changed. The first is for when we think observation 16 is uncensored at the observed time, and the second is for when observation 16 is treated as censored at time 8. We see that in both cases some pseudo observations change. The pseudo observation value that changes the most is observation 16, which is natural since it is the one we change.

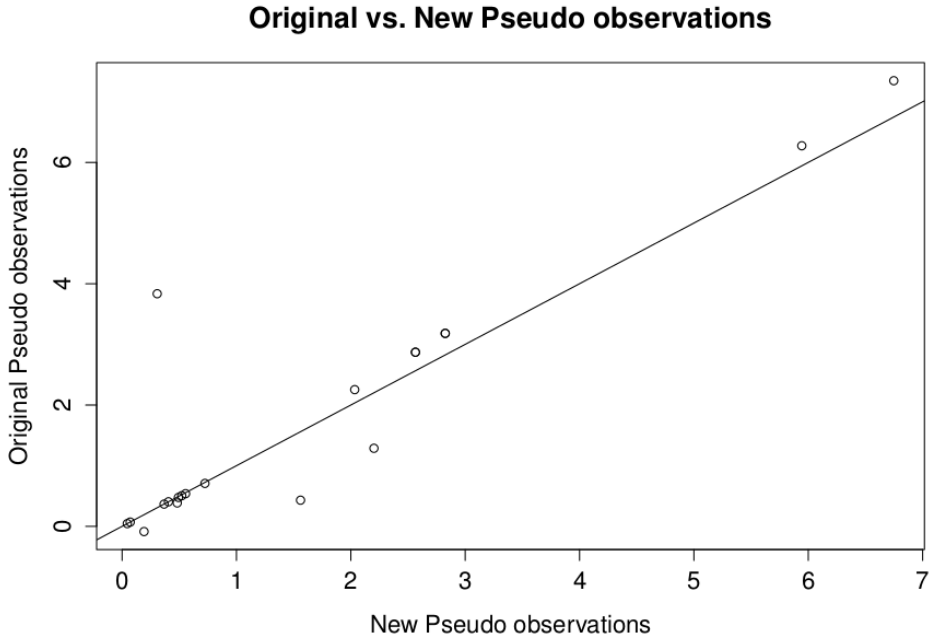

Figure 5.4: Pseudo observations from the original W1 data set, where observation 16 are censored, plotted against pseudo observations if observation 16 was uncensored.

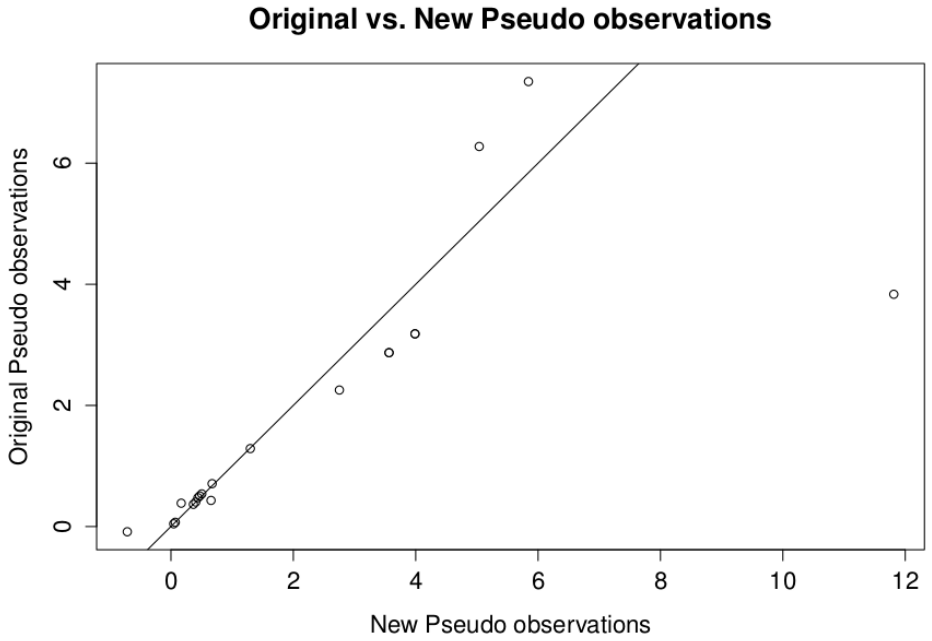

Figure 5.5: Pseudo observations from the original W1 data set, where observation 16 are censored, plotted against pseudo observations for when we treat observation 16 as censored at time 8.

5.2.2.3 Study of a censored data set with 300 observations

Figure 5.6 shows three figures made for a data set with 300 observations, simulated from a Weibull distributed AFT model as described in section 4.8. There is 30% censoring, σ=2/3 and $\boldsymbol{\beta}$=(0,0.5,0.5,0.1).

The left of the three plots show unsorted observations of real survival time, observed survival time and KM pseudo observations. From this we see that it is hard to draw many conclusions when the data is unsorted. Still, we notice that there are some high observations, but none higher than 8.

The middle plot shows the same observations, but sorted after increasing observed survival time. Notice that pseudo observations are divided in two lines. The highest line is pseudo observations corresponding to censored observations and the lower line corresponds to uncensored observations. This pattern occurs because pseudo observations for censored observations usually are higher than the observed times, and pseudo observations for uncensored survival times remain around the observed time.

In the third plot, pseudo observations, real and observed survival times are sorted individually in ascending order, and plotted. If we view the three types of observations as individual data sets we see that the set consisting of pseudo observations are closer to the real survival times than the observed times. This indicates that the set of pseudo observations have the same distribution as the real survival times, which is a desirable property. Because KM is non-parametric any knowledge of covariates will be unused. High pseudo observations does not always correspond to high real survival times. So unfortunately, this does not necessarily guarantee a good estimated AFT model. However, estimation of mean and variance might improve.

5.2.2.4 Performance under different levels of censoring

Censoring status is one of the properties the KM-method use to find pseudo observations. If there are fewer observed events, there will be fewer accurate observations to build the pseudo observations on. In figure 5.7 we see figures from another simulated data set with 20 observations. The same data set is used for all figured, but with different levels of censoring. This data set is also on the form discussed in section 4.8, with $\beta_0 = 0, \beta_1 = 0.8, \beta_2 = 0.6, \beta_3 = 0.2$ and $\sigma = 2/3$.

From the figures we see that for high levels of censoring, pseudo observations tend to flatten out. This is reasonable since the observed times for high levels of censoring are small, and several censored observations in a row leads to equal pseudo observations. A consequence of this is that the difference between pseudo observations and simulated real survival times gets bigger, as we see from the lower row of plots in figure 5.7.

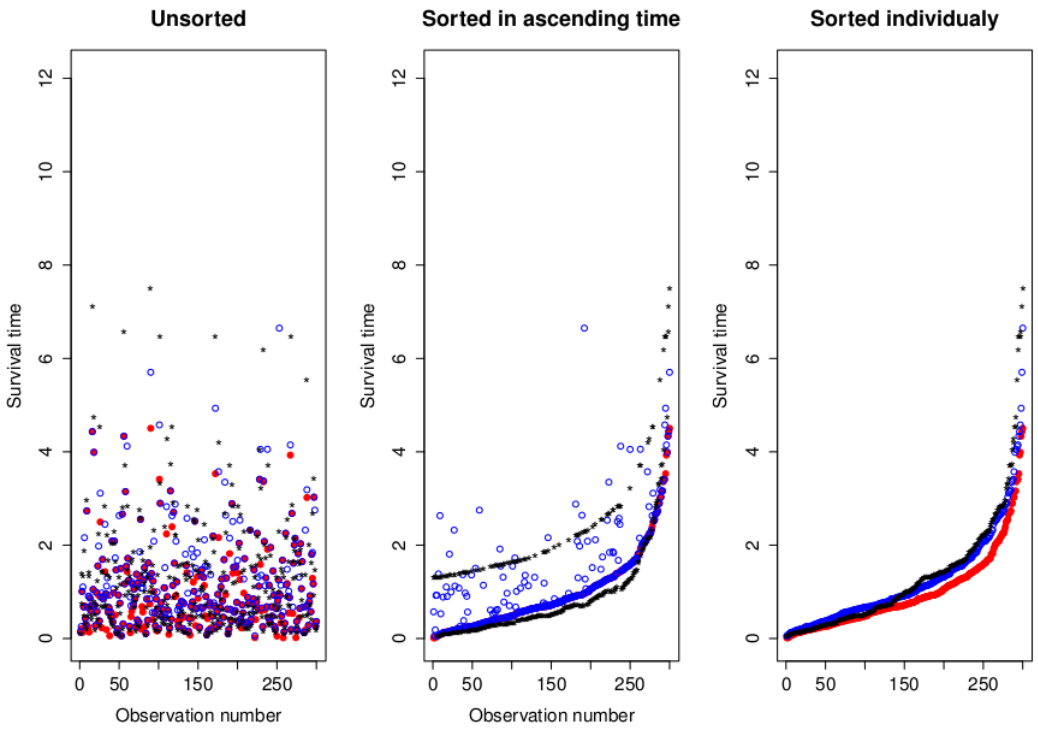

Figure 5.6: Three plots of the same data set with 300 observations. Pseudo observations (*), observed survival times (red •) and real survival times (blue ∘). Figures show: left: A plot where the observed times are unsorted. Middle: A figure where survival times are sorted in ascending order. Right: A figure where pseudo observations observed and real survival times are sorted individually in ascending order

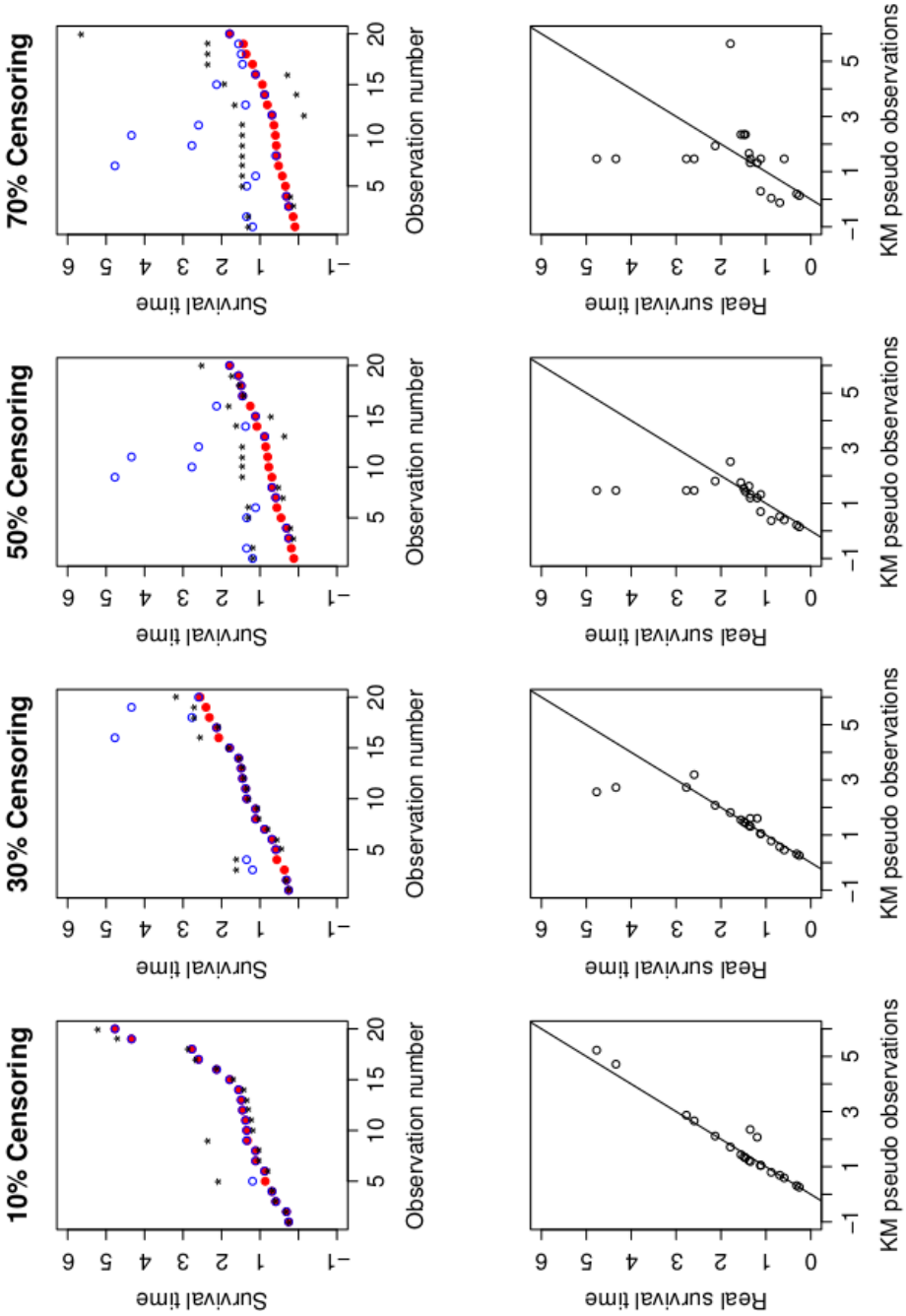

Figure 5.7: Top row: Figure showing pseudo observations (*), real survival times (blue ○) and observed survival times (red ●) for the same Weibull distributed data set, but different levels of censoring. Bottom row: Real survival times plotted against pseudo observations

5.2.2.5 Performance for different values of σ

In figure 5.8 we see plots of the probability density function for different values of σ, but with the same scale parameter. All the data sets that we have looked at so far have been Weibull distributed with $\sigma = 2/3$, which corresponds to a shape parameter $\gamma = 1/\sigma = 1.5$.

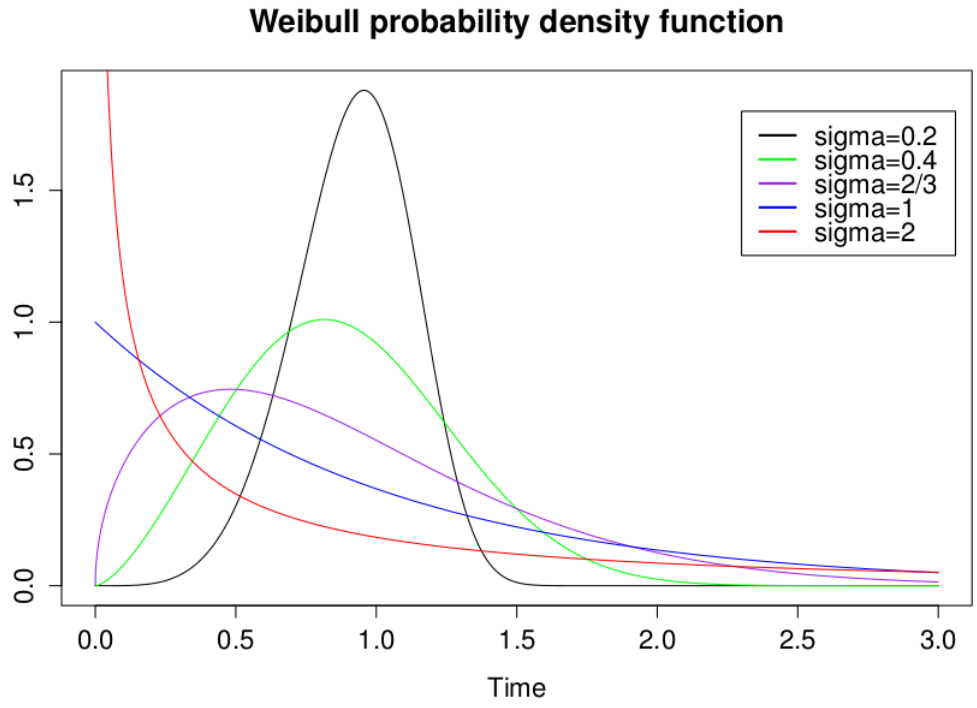

Figure 5.8: Weibull probability density functions for different shape parameters ($1/\sigma$), all with scale=1.

The distribution of the real survival times may influence how well our pseudo observations fit the real survival times. Figure 5.9 shows plots for data sets with the same covariates and level of censoring, but for different values σ, plotted against the real simulated survival times. From this we see that pseudo observations for small values of σ are closer to the observed times than pseudo observations for large values of σ. We also see that pseudo observations corresponding to large real survival times tend to be too small. Looking at figure 5.8 we see that for large values of σ there should be some large survival times. When we have censoring, these high values may be censored, and KM may not be able to give high enough values to match the true distribution.

It does not show very well in the figure, but we also get more and more negative pseudo observations the higher the value of σ.

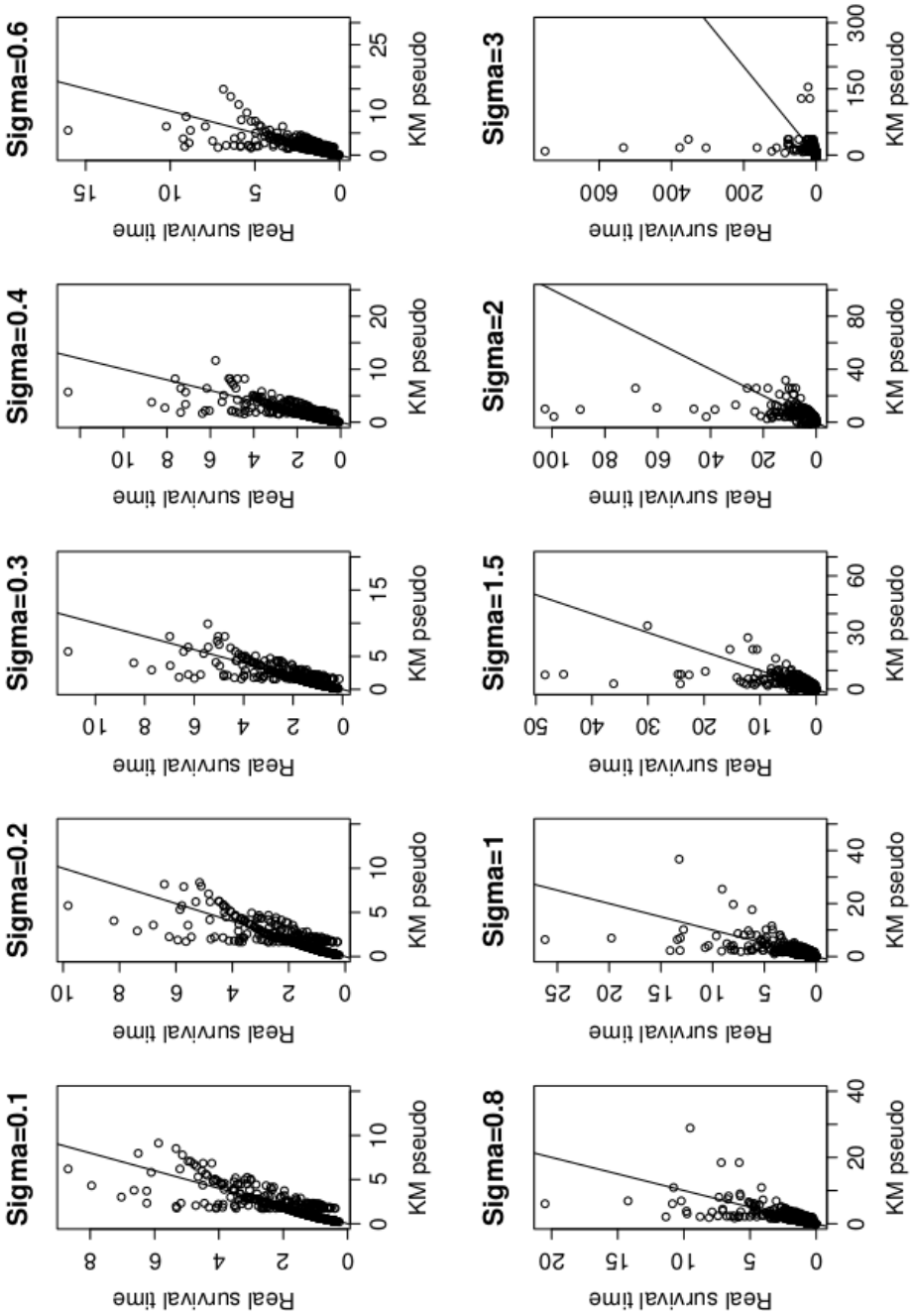

Figure 5.9: Figures showing KM pseudo observations plotted against real survival times for different values of σ. The data set have the same covariates, $\boldsymbol{\beta}$=(0,0.8,0.8,0.3) and 30% level censoring.

5.3 Pseudo observations based on Kaplan-Meier for log(T)

KM can, as we have seen, give negative pseudo observations. Depending on what we want to use the pseudo observations for, this can be a problem. An alternative non-parametric method is to find pseudo observations for $log(T)$ using Kaplan-Meier, and then exponentiate them to find positive pseudo observations for T. The method for finding pseudo observations based on $log(T)$ with Kaplan-Meier will be referred to as KMlog in the following. The procedure for KMlog is:

- Find an estimator, $\hat{\theta}_{log}$, for the expected value of $log(T)$ for the full data set using equation 3.4.
- For all observations i: Remove the observation i from the original data set and find the "leave-one-out" estimate $\hat{\theta}_{log,-i}$ by equation 3.4.
- Find pseudo observations, $\hat{\theta}_{log,i}$, for $log(T)$ by equation 5.1.
- Exponentiate $\hat{\theta}_{log,i}$ to find pseudo observations, $\hat{\theta}_i$, for the survival time T.

Because KMlog is very similar to KM, we will not look as detailed into this method, but focus at the differences between them.

5.3.1 Uncensored observations

In section 5.2.1 we proved that the Kaplan-Meier estimated survival function for $log(T)$ is similar to the one for T. Therefore, by following the same derivation as for KM, we can prove that pseudo observations from KMlog for uncensored data sets with no ties, also will be identical to the survival times in the original data set.

5.3.2 Censored observations

Figure 5.10 shows plots of KM pseudo observations, KMlog pseudo observations and observed and real survival times for the same data sets we looked at in i figure 5.7. Pseudo observations for KM and KMlog are very similar, which is natural since the procedures are very similar. Still there are some differences. For small survival times we see that KM pseudo observations tend to be further away from the observed times than KMlog, and conversely for larger observed times. To see why this may be we will study plot D more closely.

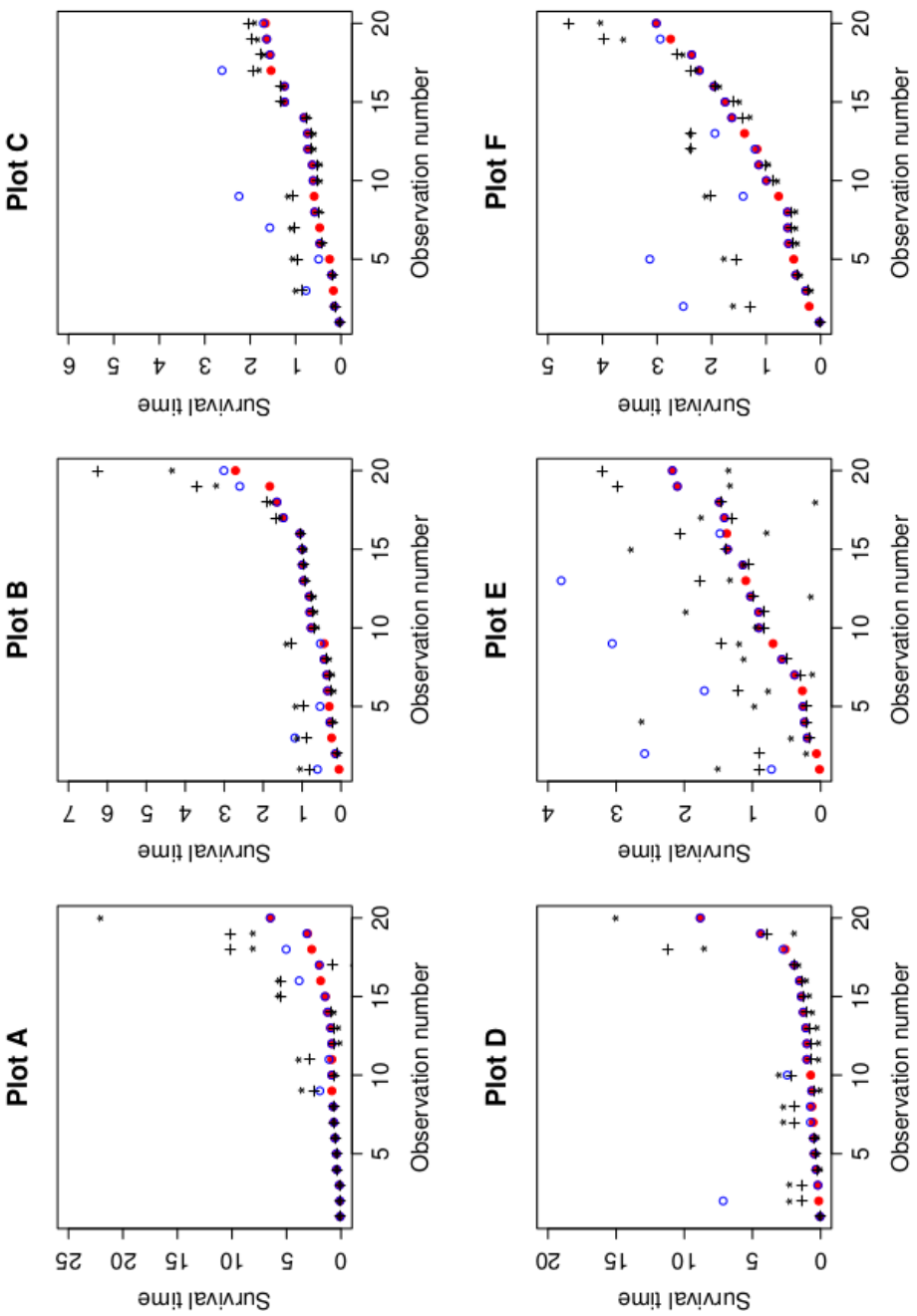

Figure 5.10: Figure showing KM pseudo observations (*), KMlog pseudo observations (+) real survival times (blue ○) and observed survival times (red ●) for the same 6 different data sets as in figure 5.1. Note: KMlog for observation 20 in plot A and D are outside the range of the plot.

KMlog pseudo observations are first found for $log(T)$ and then exponentiated. In figure 5.11, pseudo observations and observed survival times in plot D, are plotted for $log(T)$ and T.

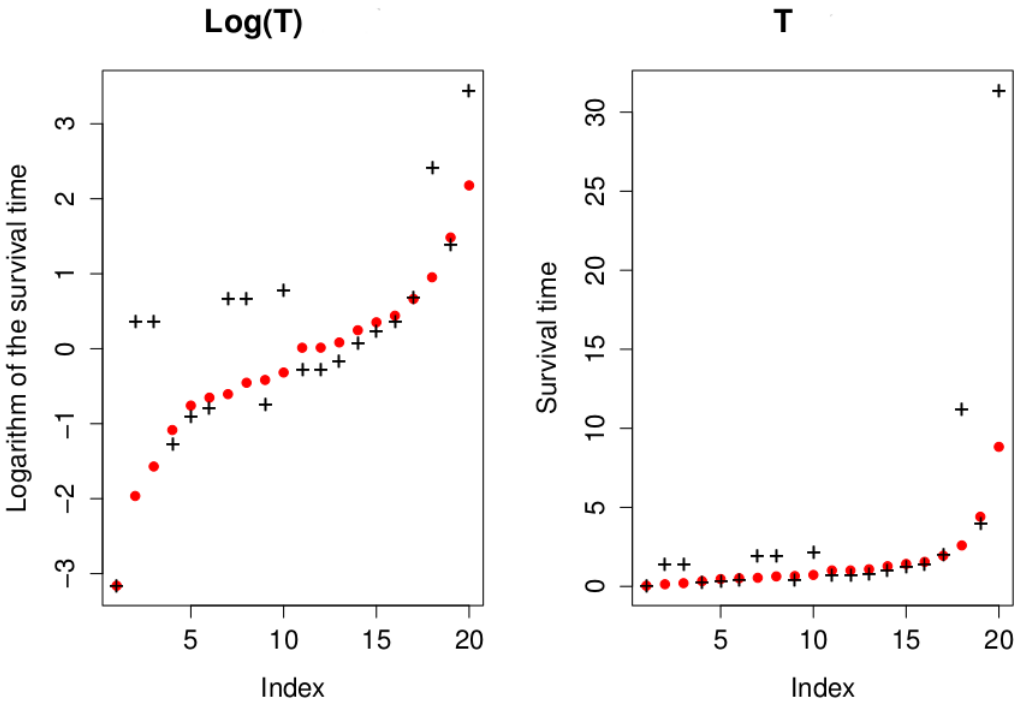

Figure 5.11: Figure showing real survival (red •) and pseudo observations from KMlog (+), for $log(T)$ and T. The data set are the same as in plot D in figure 5.10.

From this plot we see that there is a difference in how the differences between pseudo observations and observed times of $log(T)$ transfers to differences in T. Large differences in $log(T)$ will be small differences in T if the survival time is small, and for large survival times a small difference in $log(T)$ can be big for T. This may be an explanation to why the largest KMlog observations are so big compared to KM and observed survival time.

Plots for KMlog corresponding to figure 5.6, 5.7 and 5.9 are found on the next pages. Be aware that very high pseudo observations are outside the range of some of the plots in order to have the same scale as the figures for KM.

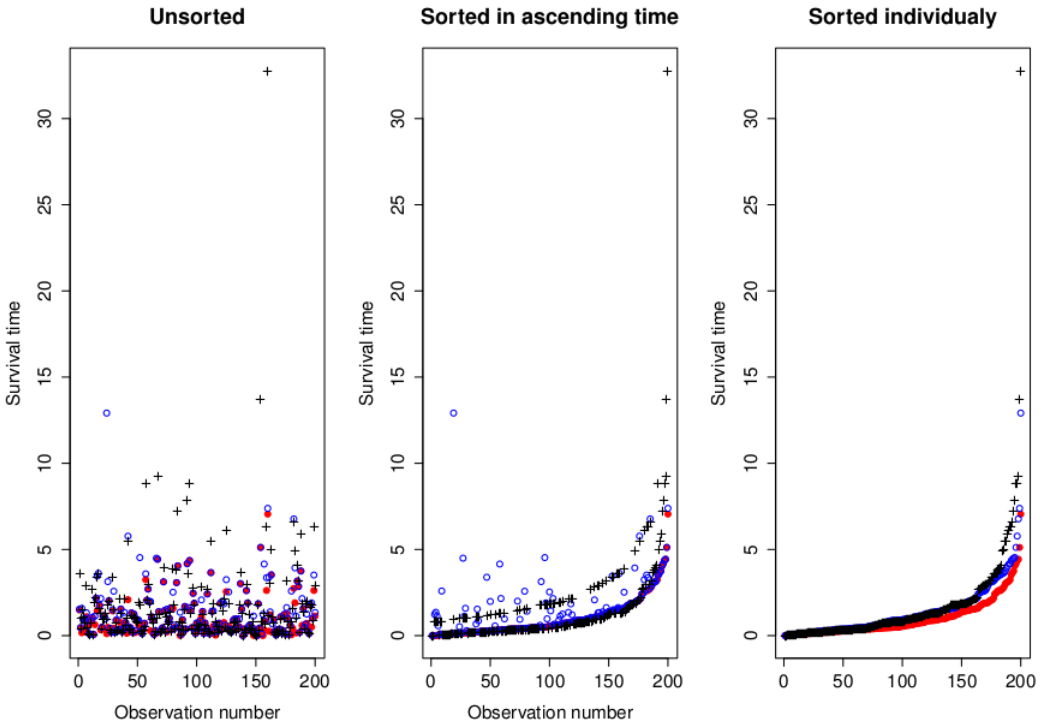

Figure 5.12: For the same data set with 300 observations as in figure 5.6, three figures are plotted with pseudo observations (+), real survival times (blue ∘) and observed survival times (red •). From the left: Unsorted data. Middle: Sorted after increasing observed survival time. Right: sorted individually, no link between pseudo, real and observed observations.

Figure 5.12 shows that there are several high observations. In line with our previous observations we see that the highest pseudo observations originate from the highest observed survival times. The right figure shows observed survival times, real survival times and pseudo observations sorted individually in increasing order. From this we see that the data set consisting of KMlog pseudo observations gives a good approximation to the real data set for small survival times, but the large KMlog values at the end makes it not as good as observations from the KM-method (see figure 5.1).

Figures 5.13 and 5.14 shows us the same trends as we had for KM. The difference is that there are more high pseudo observation values for KMlog. In some situations this can be a good thing, but most of the time it makes the difference between pseudo observations and real survival time larger than it is for KM.

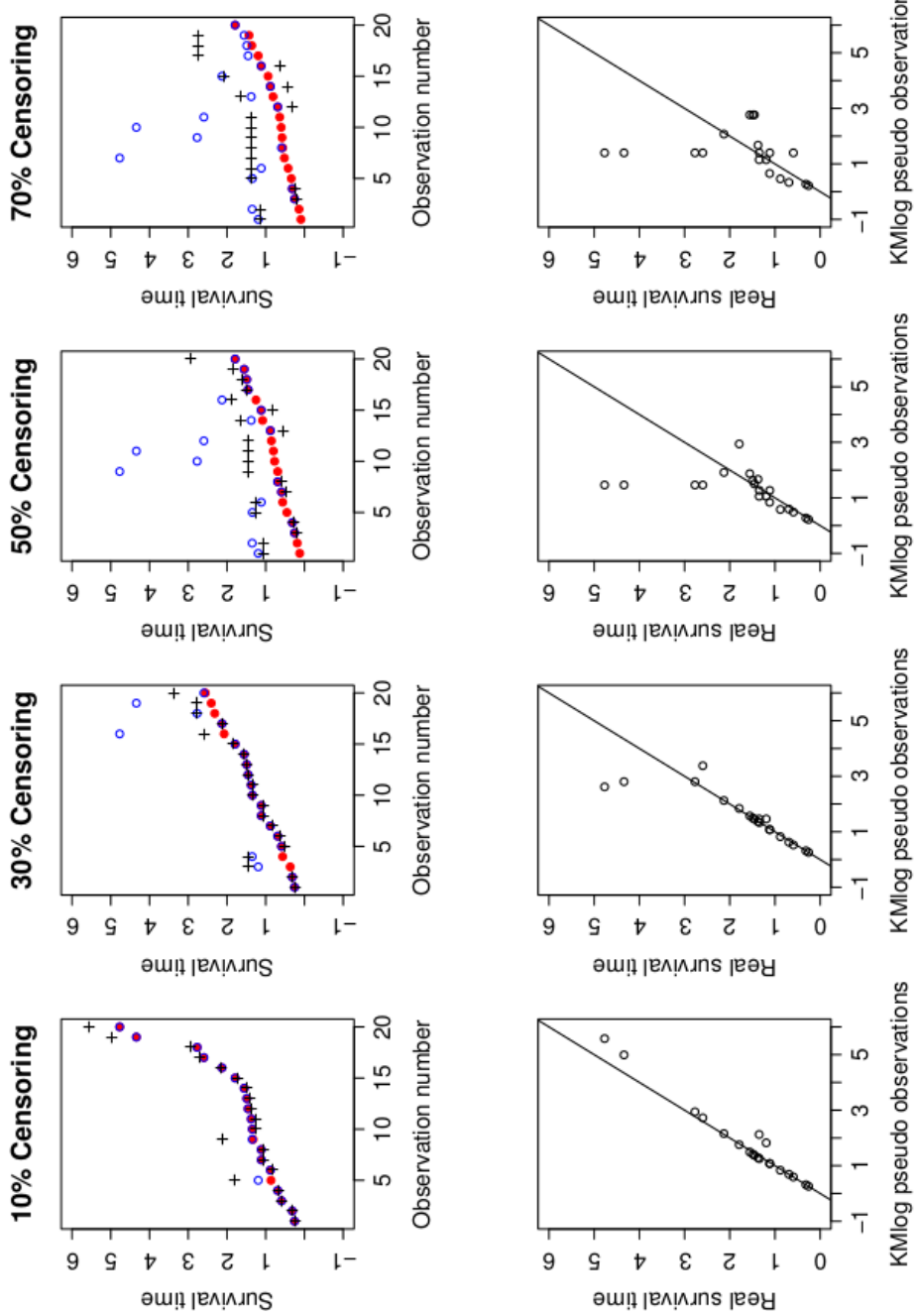

Figure 5.13: Figures showing plots for different levels of censoring of the same data set as in figure 5.1. Top row: Real survival times (blue ○), observed survival times (red ●) and KMlog pseudo observations (+). Sorted after increasing observed time. Bottom row: Plot of KMlog pseudo observations against real survival times. The line shows where point will be if real survival times are equal to pseudo observations. OBS: Several points are not included for high censoring, so this plot may be misleading.

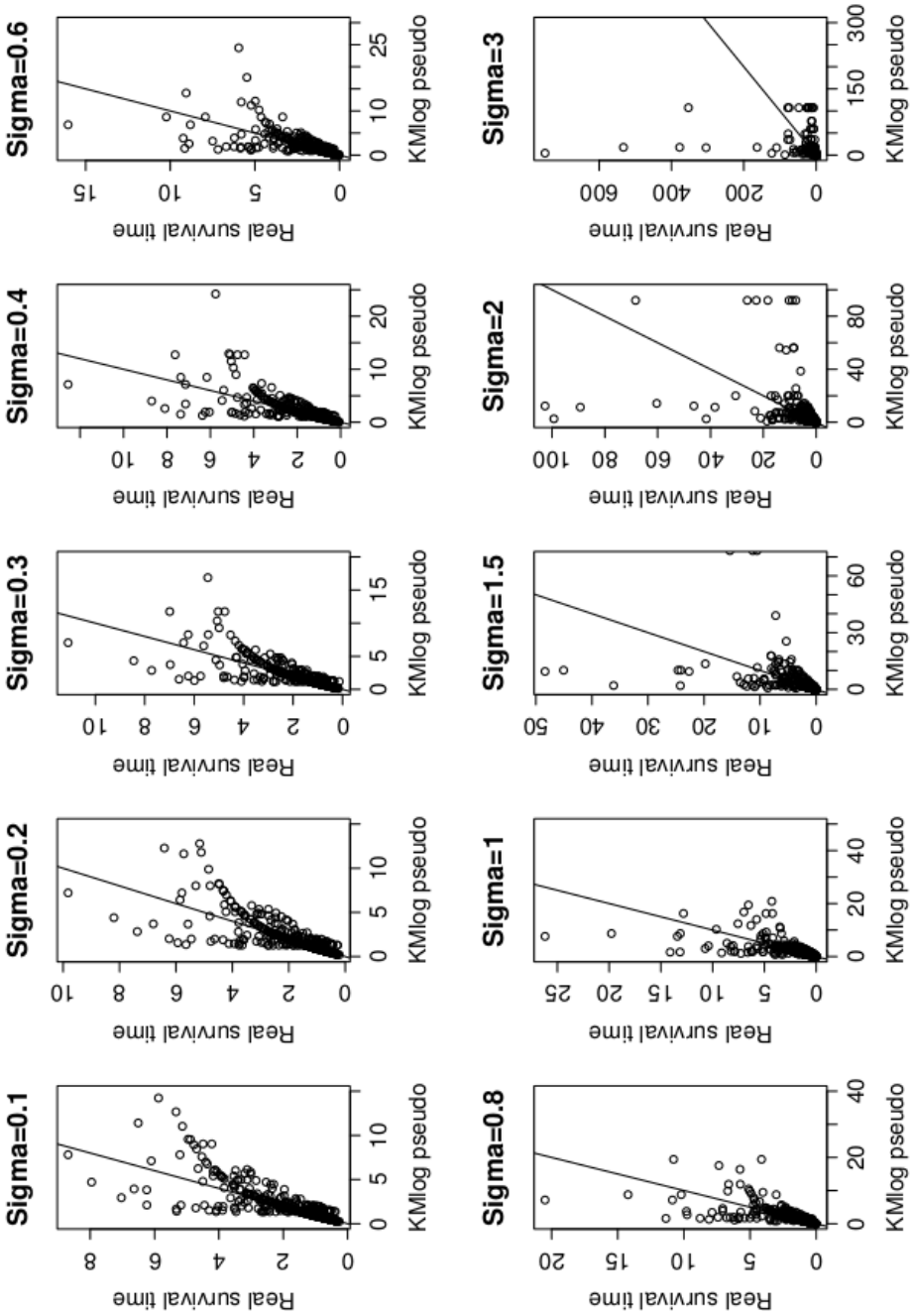

Figure 5.14: Plots of KMlog pseudo observations against real survival times for different values of σ. The data set have the same covariates, $\boldsymbol{\beta}$=(0,0.8,0.8,0.3) and 30% level censoring, and is the same as in figure 5.9.

5.4 Parametric pseudo observations

When we know the distribution of our survival times, we can use the measured covariates to find parametric pseudo observations. The procedure for finding such pseudo observations is:

- Express an AFT model for the data set.
- Find estimates for parameters β_j, μ and σ, possibly using *surveg*. In addition, find an estimate for the expected value of the covariates x_j (using $\frac{1}{n}\sum_{j=1}^{n} x_{ij}$), and the expected value of the error term ϵ. The later will be -0.5722 for Weibull, 0 for lognormal.
- Use this to find the estimated expected log-survival time, $\hat{\theta}$, by setting $\hat{\theta} = \hat{\mu} + \hat{\beta}_1\widehat{E(x_1)} + \cdots + \hat{\beta}_p\widehat{E(x_p)} + \hat{\sigma}E(\epsilon)$.

Then for each observation i:

- Remove observation i from the original data set and find "leave-one-out" estimates for covariates and parameters
- Find "leave-one-out" estimate for the expected log survival time using $\hat{\theta}_{-i} = \hat{\mu}_{-i} + \hat{\beta}_{-i,1}\widehat{E(x_{-i,1})} + \cdots + \hat{\beta}_{-i,p}\widehat{E(x_{-i,p})} + \hat{\sigma}_{-i}E(\epsilon)$.
- Then find pseudo observations for $log(T)$ using equation 5.1, and find pseudo observations for the survival time, T, by exponentiating the pseudo observations found for $log(T)$.

5.4.1 Uncensored data sets

For KM and KMlog we found that uncensored observations without ties result in pseudo observations that were equal to the observed survival times. This will not necessarily be the case for parametric pseudo observations. In figure 5.15 we see plots of parametric pseudo observation and observed survival times for an uncensored data set. Most of the pseudo observations stay very close to the observed times, but some of them are slightly above or below the observed survival times.

5.4.2 Censored data sets

In figure 5.16 we see plots for the same six censored data sets that we studied for KM and KMlog. Parametric pseudo observations for these data sets are sometimes better than those based on KM and KMlog, and sometimes worse. They also tend to be too big for the largest observations, for the same reason as KMlog, but seem to be smaller than KMlog.

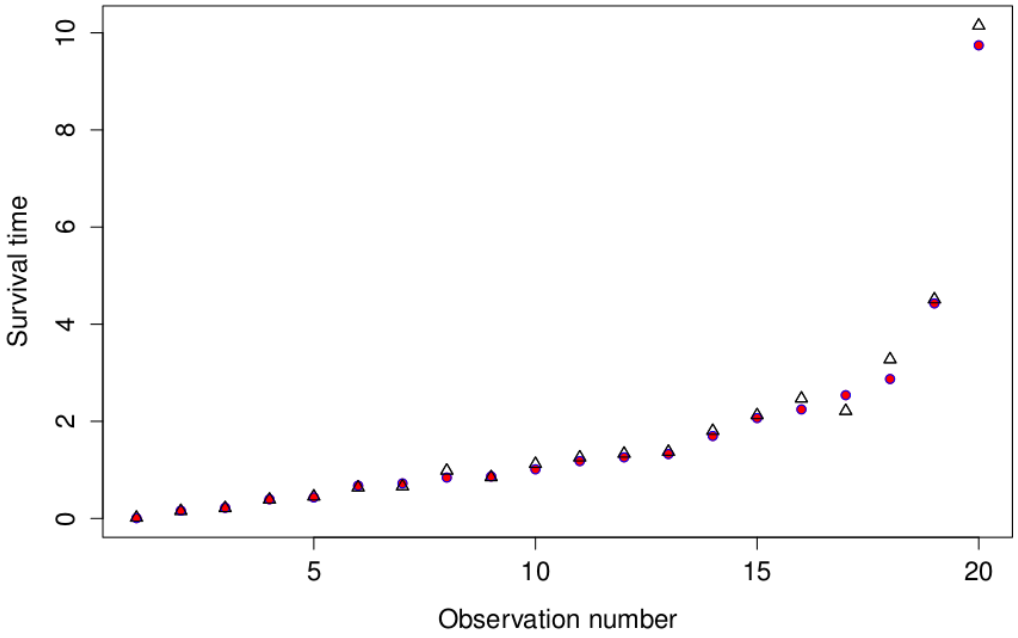

Figure 5.15: Plots for a simulated uncensored data set from a Weibull AFT model with β=(0,0.5,1,0.3) and 30% censoring. Observed and event time (blue and red •), Parametric pseudo observation (△).

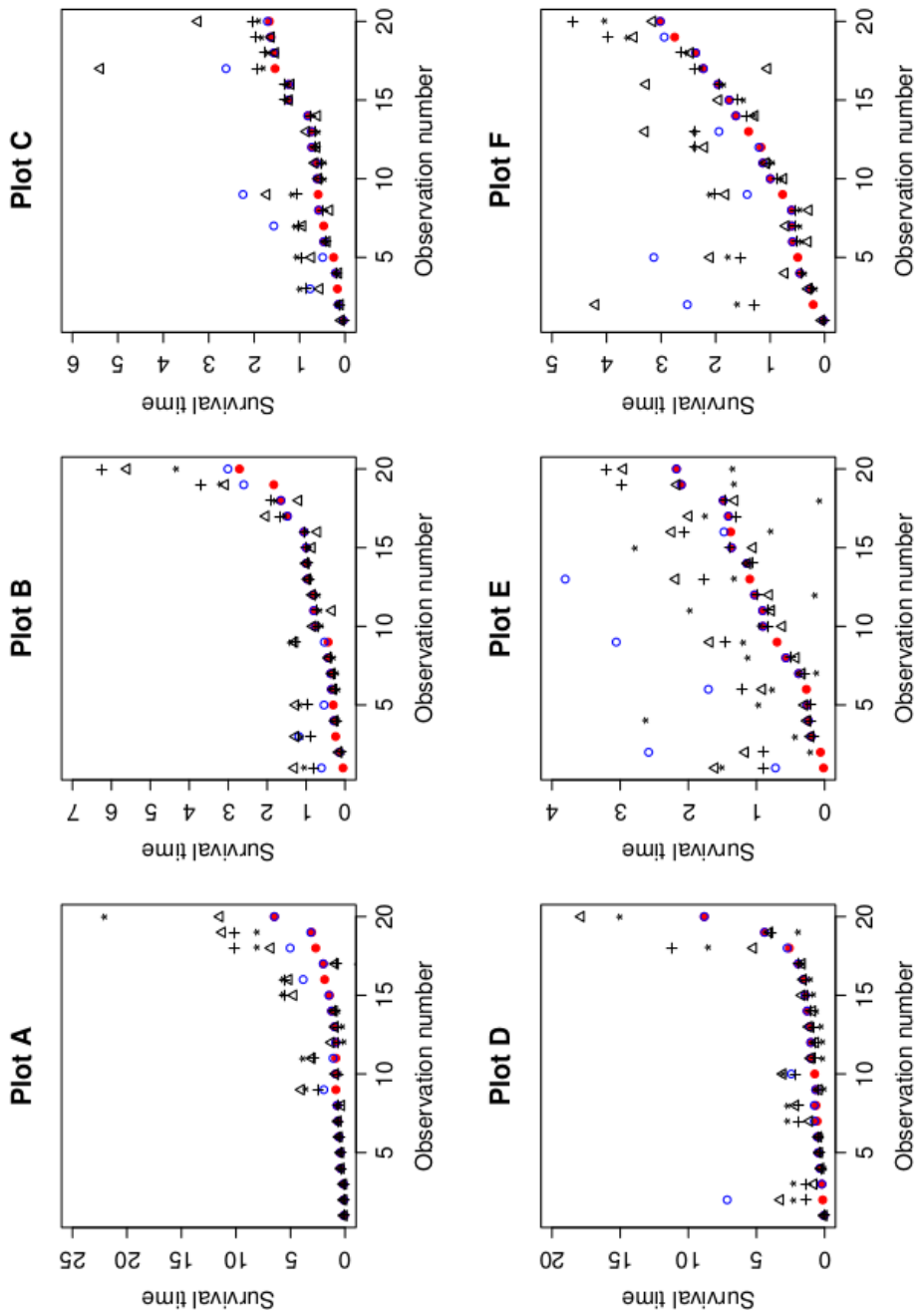

Figure 5.16: Parametric pseudo observations (△), KM pseudo observations (*), KMlog pseudo observations (+), real survival time (blue ○) and observed survival time (red ●). The data sets are the same as for figure 5.1 and figure 5.10.

5.4.2.1 The data set with 300 observations

For the large data set with 300 observations (figure 5.17), we don't see the same clear groups of pseudo observations as we did for KM and KMlog (the middle figure in 5.6 and 5.12). When we sort the parametric pseudo observations, observed survival times and real survival times in increasing order we get the figure on the right side in 5.17. From this we see that the data set consisting of parametric pseudo observations also serves as a good approximation to the real data set, but not as good as the KM pseudo observations because some of the highest pseudo observations are too high.

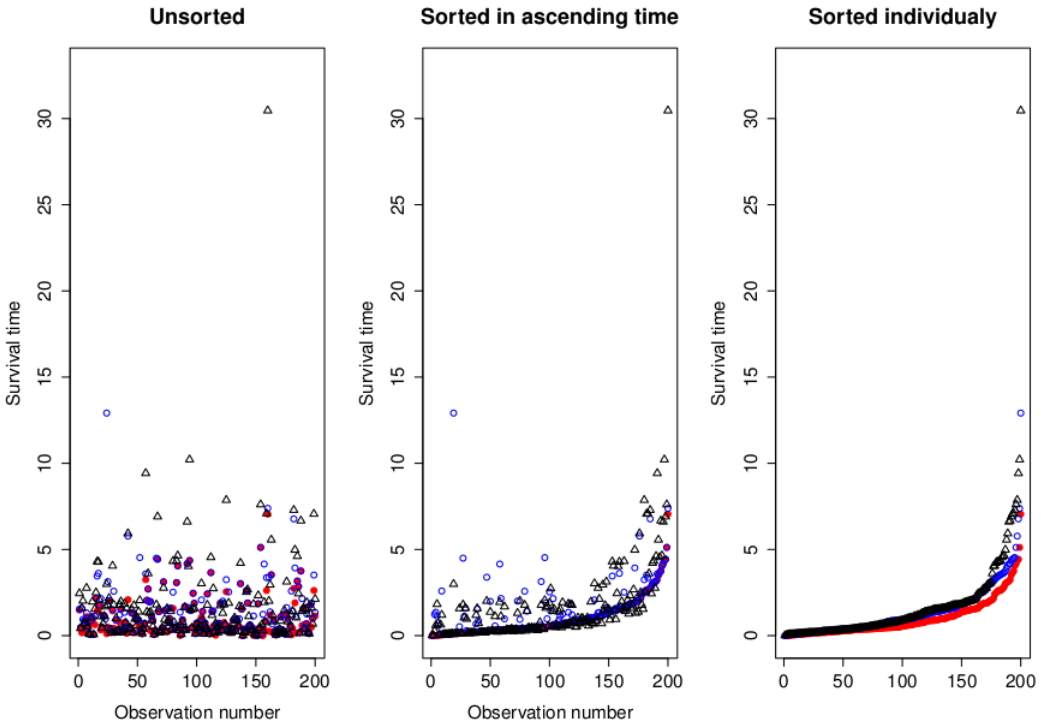

Figure 5.17: Pseudo observations (△), real survival times (blue ∘) and observed survival times (red •) for the same data set as in figure 5.6 and 5.12. Left: Unsorted. Middle: Sorted observations in increasing time. Right: sorted individually, no link between pseudo, real and observed observations.

If we plot censored observations and uncensored observations separately we get figure 5.18. Pseudo observations corresponding to uncensored observations stays in the area around the observed time, but have a tendency to be bigger for larger observed times. The same goes for uncensored observations, but they are all somewhat larger than the observed times.

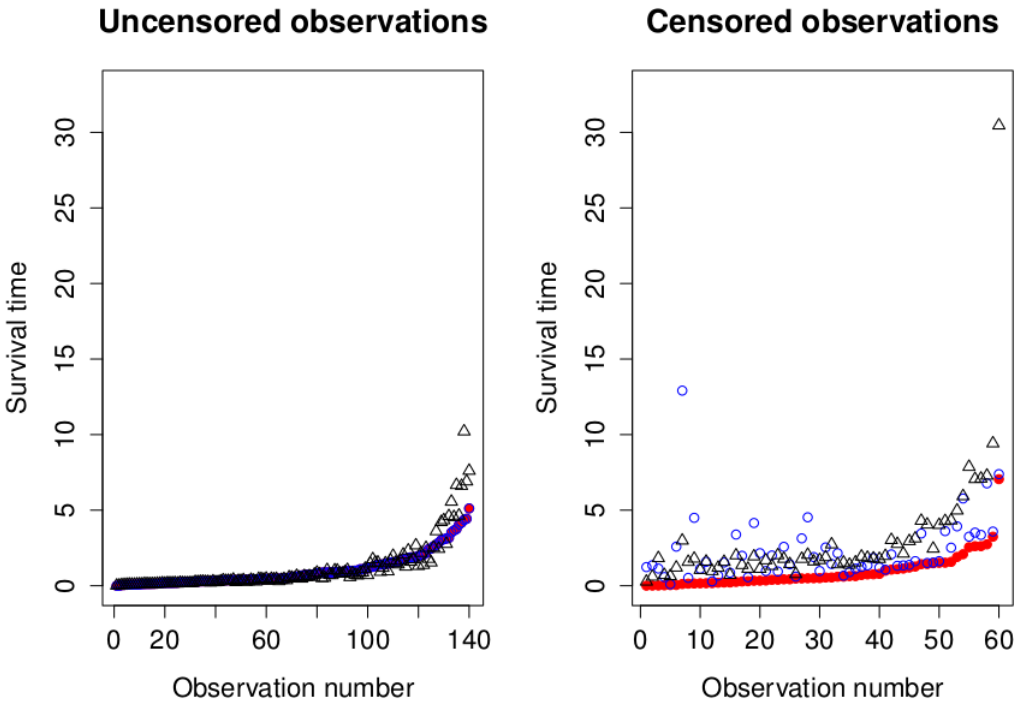

Figure 5.18: Pseudo observations (△), real survival times (blue ∘) and observed survival times (red •) for censored and uncensored observations. The data set is the same as in figure 5.17.

The trend, with too high pseudo observations for larger observed times, can be explained the same way as for KMlog. Parametric pseudo observations are also first found for $log(T)$ and then exponentiated. This makes big differences in values for small survival times for $log(T)$ small for T, and conversely for small differences for $log(T)$ for large survival times.

5.4.2.2 A more detailed study of parametric pseudo observations

Data set W1 in appendix B is simulated from a Weibull distributed AFT model as described in section 4.8. The model is

$$log(T) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \sigma\epsilon,$$

where β_0=0, β_1=0.6, β_2=0.5, β_3=0.2 and σ=2/3. x_1 is binary, x_2 is uniform[-1,1], x_3 is normal(0,1) and ϵ is Gumbel distributed. Expected value for covariates should therefore be around 0.5, 0 and 0, for x_1, x_2 and x_3 respectively, and $E[\epsilon]$=-0.5772.

In figure 5.19, parametric pseudo observations for this data set are plotted with observed times and real survival times. For each pseudo observation we need to estimate σ and β_j, and find $\bar{x}_j$ for $j = 1, 2, 3$. For the full data set, and some selected reduced data set, these values are printed in table 5.2 along with estimated pseudo observations of $log(T)$ ($\hat{\theta}_i$) and pseudo observations of T. We notice that the estimated covariates are close to the theoretical value, but also that $\bar{x}_2$ is negative. Some of the estimated $\hat{\beta}_j$ values varies quite a lot from the true values, but at least they have the right sign compared to $\bar{x}_j$. When we only have 20 observations we can't expect the estimated values to be equal to the theoretical values, and hence we can't expect that pseudo observations based on those results are very accurate either.

Table 5.2: Estimates of parameters and expected value of covariates, for the full and some reduced data sets from the W1 data set. Along with pseudo observations for $log(T)$ ($\hat{\theta}_i$), and pseudo observations for T.

Data set	Full	Without 5	Without 10	Without 15	Without 20
$\hat{\mu}$	0.0035	-0.0556	0.0171	-0.0432	0.0960
$\hat{\beta}_1$	0.8361	0.9101	0.8003	0.9751	0.4398
$\hat{\beta}_2$	-0.0814	-0.1123	-0.0251	-0.2195	0.1535
$\hat{\beta}_3$	0.3821	0.3515	0.3761	0.3729	0.4064
$\hat{\sigma}$	0.6968	0.7047	0.7081	0.7056	0.6414
$\bar{x}_1$	0.5	0.5263	0.4737	0.4737	0.4737
$\bar{x}_2$	-0.1237	-0.1338	-0.0930	-0.1094	-0.1245
$\bar{x}_3$	0.0766	0.0246	0.0311	0.0839	0.0746
$\hat{\theta}$	0.0622				
$\hat{\theta}_{-i}$		0.0439	0.0050	0.0702	-0.0515
$\hat{\theta}_i$		0.4105	1.1493	-0.0904	2.2225
Pseudo T		1.5076	3.1561	0.913	9.2307

By looking at the estimated parameters and covariates it's difficult to see which observations will result in good pseudo observations or not. The fact that estimates deviate from the theoretical values makes it even harder. Some

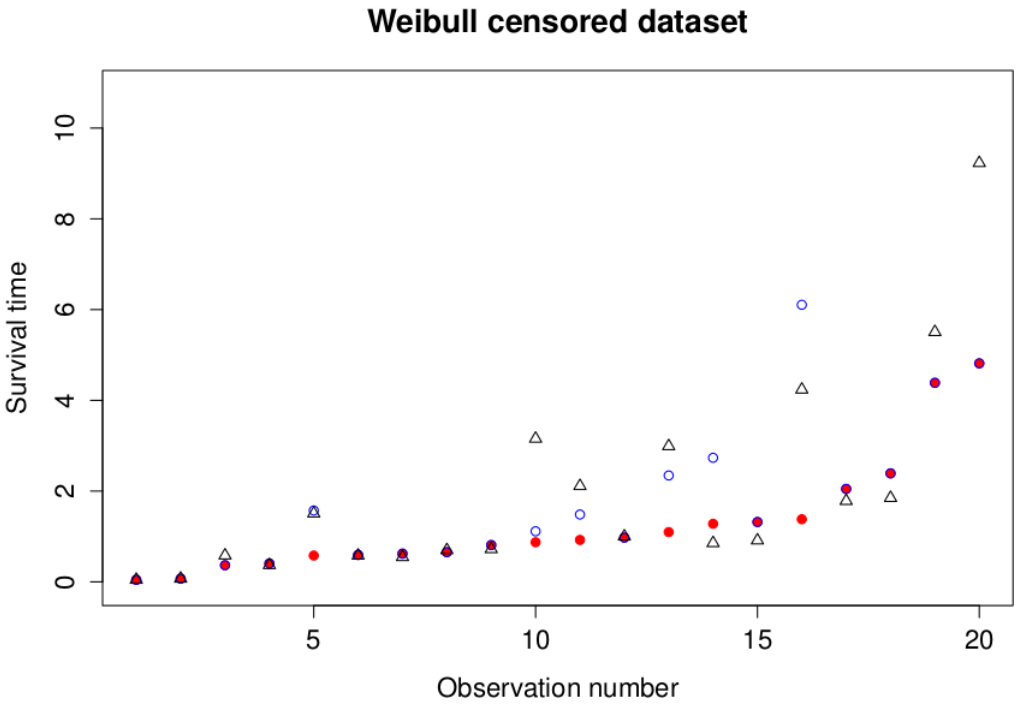

Figure 5.19: Pseudo observations (△), Observed survival times (red •) and real survival times (blue ∘) for the W1 data set.

observations, like observation 5, are very close to the real simulated survival time. Others, like observation 10 and 20 are too big, but it's difficult to see why by looking at the table.

In section 5.2.2.1 we saw that pseudo observations based on KM gives the same value for two censored observations in a row. Because we use individual estimates for covariates and parameters for each observation, this is unlikely to happen for the parametric pseudo observation.

5.4.2.3 Changing one observation

If we change censoring status or the observed time for one observation, estimators used to find pseudo observations may change for all other observations as well. Figure 5.20 shows original pseudo observations from the W1 data set, plotted against pseudo observations for two data sets where observation 16 are changed. In both cases we see that the change in pseudo observation value is largest for pseudo observations where the original value was high. This indicates that small observed times are more robust to changes in other observations.

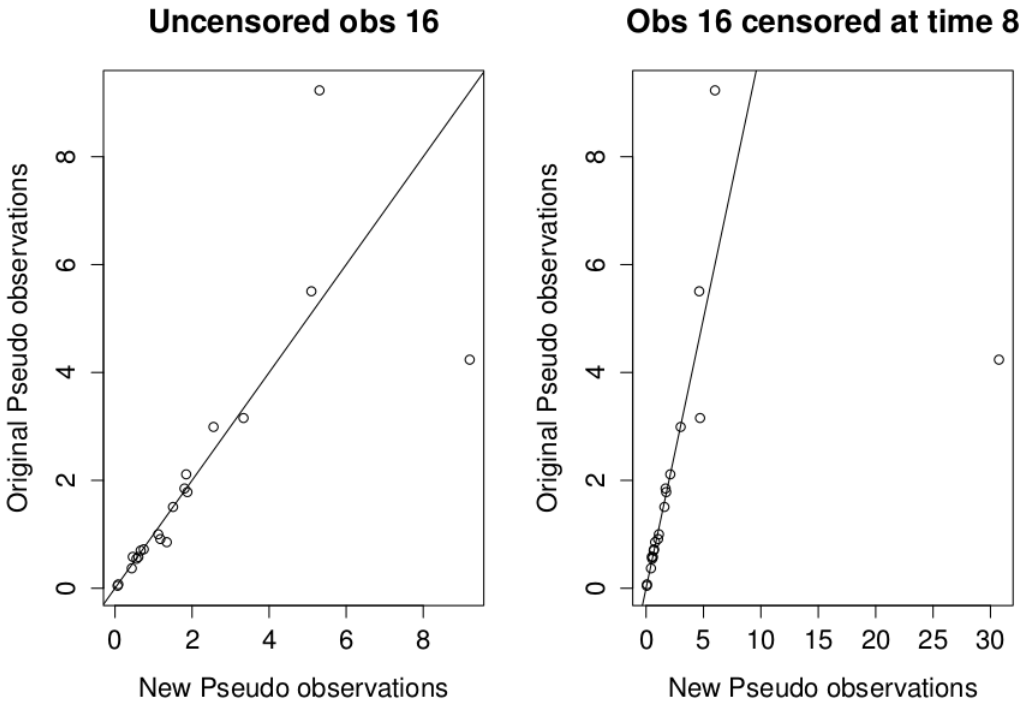

Figure 5.20: Original pseudo observations from the W1 data set plotted against pseudo observations when we changed observation 16.

5.4.2.4 Performance under different levels of censoring

How well parametric pseudo observations fit the real survival times will depend on the level of censoring. In figure 5.21 we see figures showing pseudo observations for different levels of censoring. The data set is the same as the one used for KM and KMlog. We see that the more censoring, the bigger the difference between pseudo observations and real survival times. After further studies of pseudo observations for KM, KMlog and the parametric method, we found that KM and the parametric model are less sensitive to censoring than KMlog. KM seems to be slightly better than the parametric method, but keeping in mind that KM and KMlog gives pseudo observations with little variance under high levels of censoring, the parametric method may still preform better for some data sets.

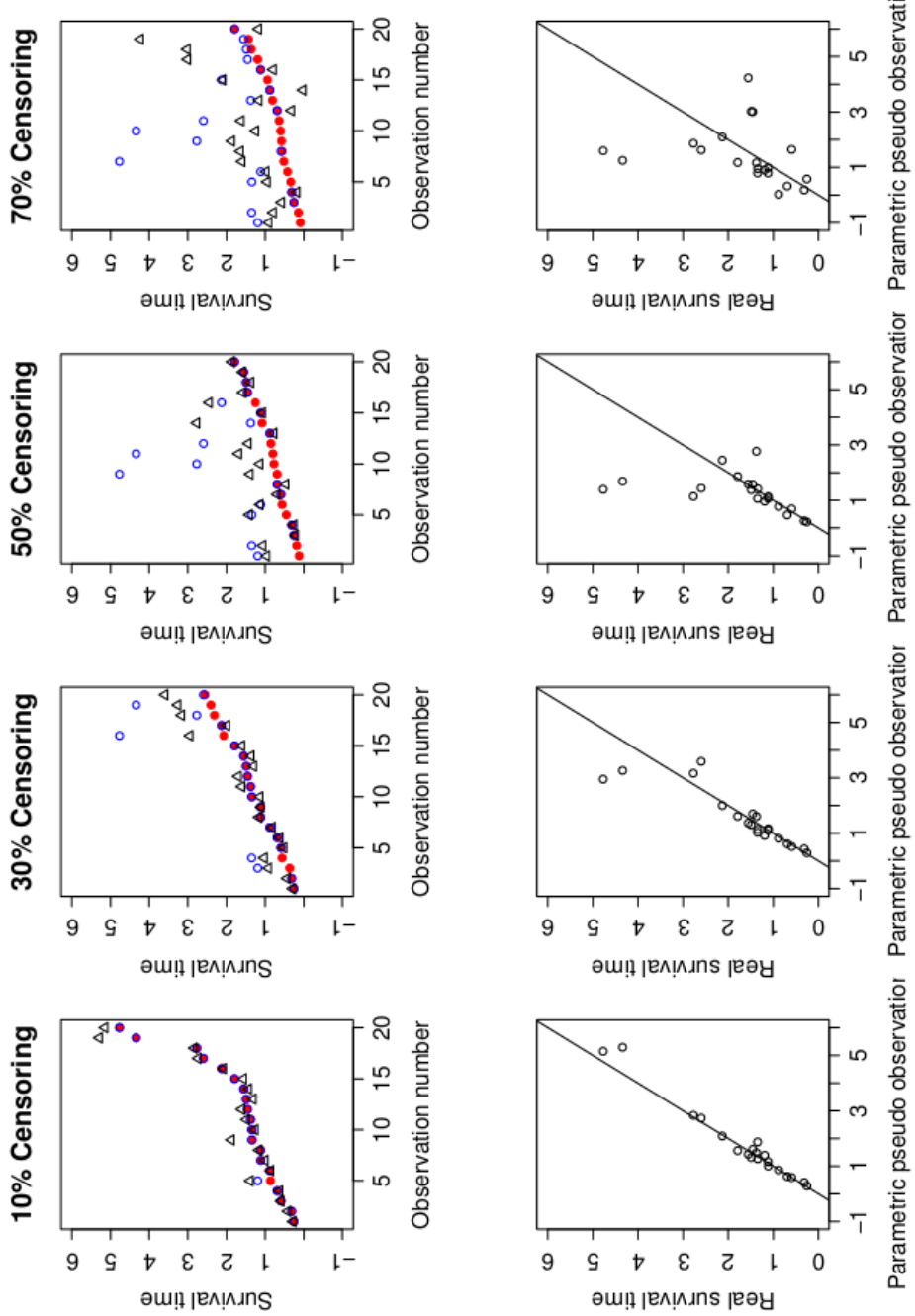

Figure 5.21: Top row: Parametric pseudo observations (△), real survival times (blue ○) and observed survival times (red ●) for different levels of censoring of the same data set as in figure 5.7 and 5.13. Bottom row: Real survival times plotted against parametric pseudo observations.

5.4.2.5 Performance for different values of σ

For parametric pseudo observations, σ is estimated for the full data set and for each pseudo observation. The value of σ will therefore influence the pseudo observations. In figure 5.22 we see parametric pseudo observations plotted against real event times for different values of σ. From this figure we see that the difference between pseudo and real times gets bigger for larger values of σ. If we compare these plots to figure 5.9 and 5.14, we see that pseudo observations based on the parametric distribution are closer to the real survival times than KM and KMlog when σ is small. When σ is around 0.6 we see that the difference between KM and parametric observations are insignificant, and as σ gets larger, KM seem to be a slightly better choice. Because σ is the constant in front of the error term in equation 4.3, a large σ will influence the error in the model. This can make it more difficult to estimate parameters, and hence pseudo observations may be further from the real event times when σ is high.

5.4.2.6 What if we guess the wrong distribution

The method for finding pseudo observations with KM and KMlog will be the same for all distributions we may assume for T. The parametric method on the other hand, use the assumed distribution and model to find pseudo observations. Therefore the code for finding parametric pseudo observations must be changed to fit the distribution and the number of covariates we assume the model to have.

In figure 5.23 we see probability density functions for some values of σ. The shape of a lognormal distribution can in many cases be similar to the shape of a Weibull distribution. Because of this similarity one can mistake a Weibull distributed data set for a lognormal one, and conversely. We will now look at this in an example.

In figure 5.24 we see parametric pseudo observations for both lognormal and Weibull distribution plotted against the real Weibull distributed survival times. From the figure we see that guessing a lognormal distribution will give pseudo observations that are close to the ones we get with Weibull. This result is not unexpected. The AFT model will be the same for both, only with different σ values. Guessing the wrong AFT model and not only the distribution, may on the other hand give very different results.

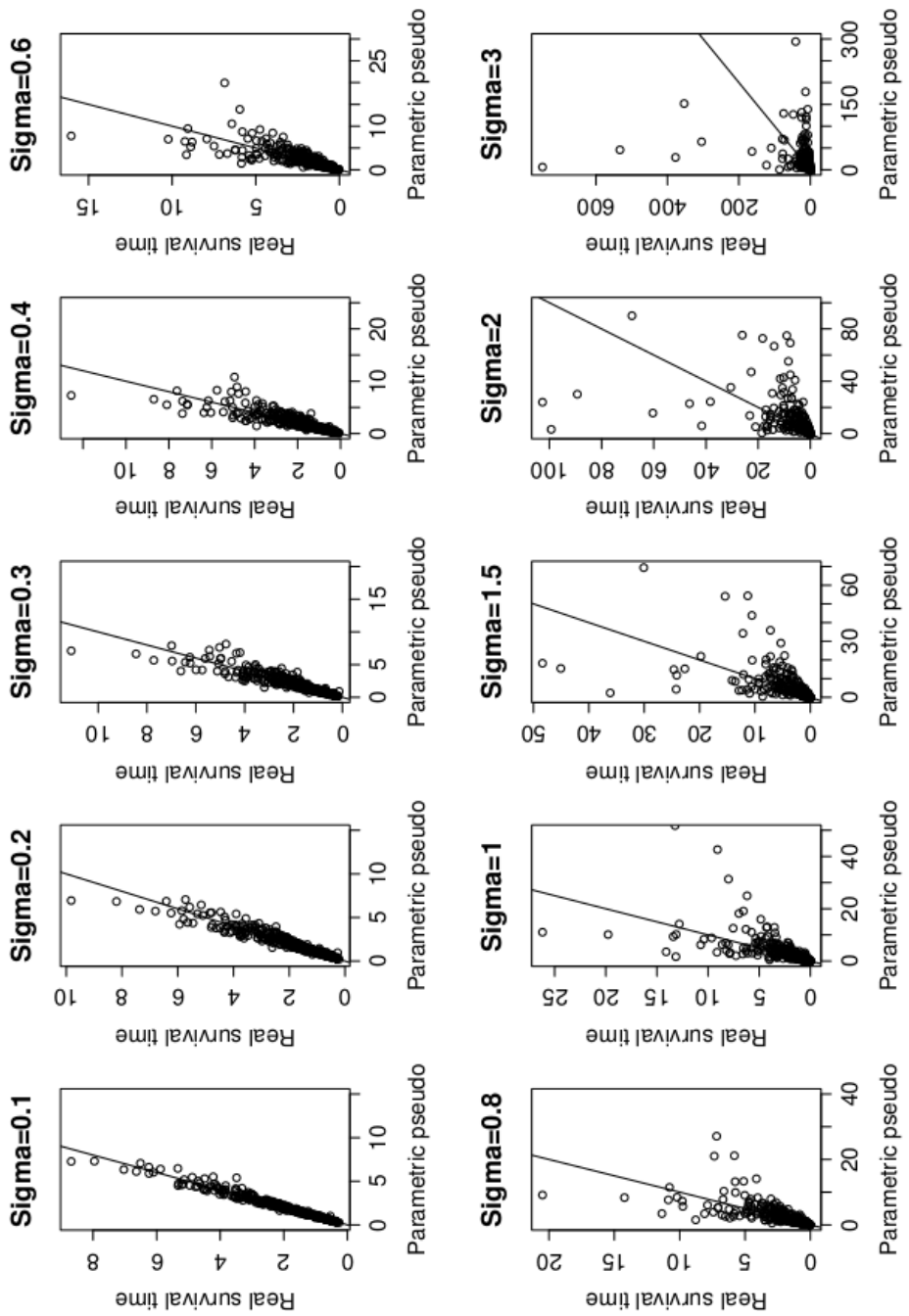

Figure 5.22: Plot of parametric pseudo observations ($\triangle$), real survival times (blue $\circ$) and observed survival times (red $\bullet$) for different values of σ. The data set is the same as in figure 5.9 and 5.14.

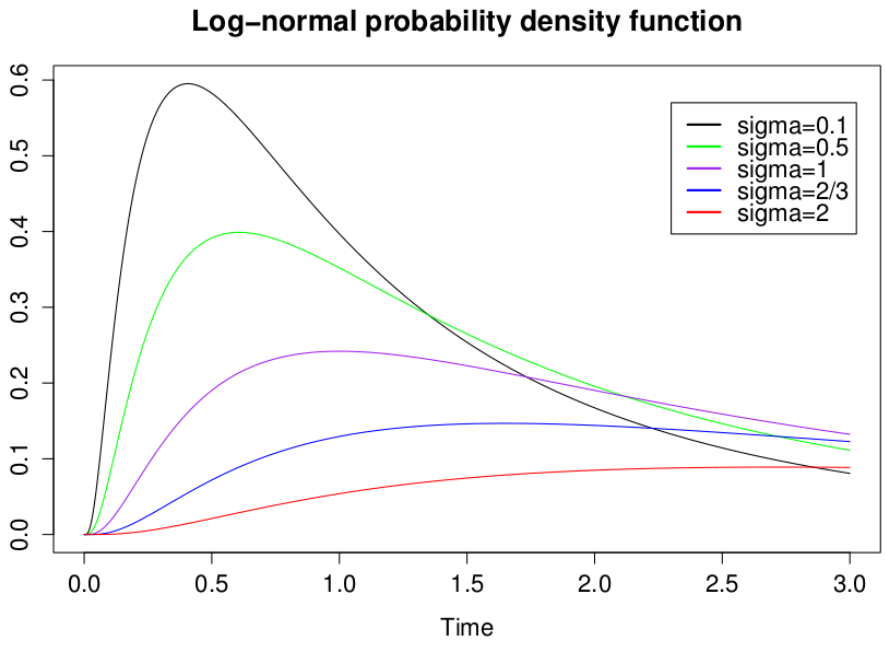

Figure 5.23: PDF for lognormal distributions with different values of σ.

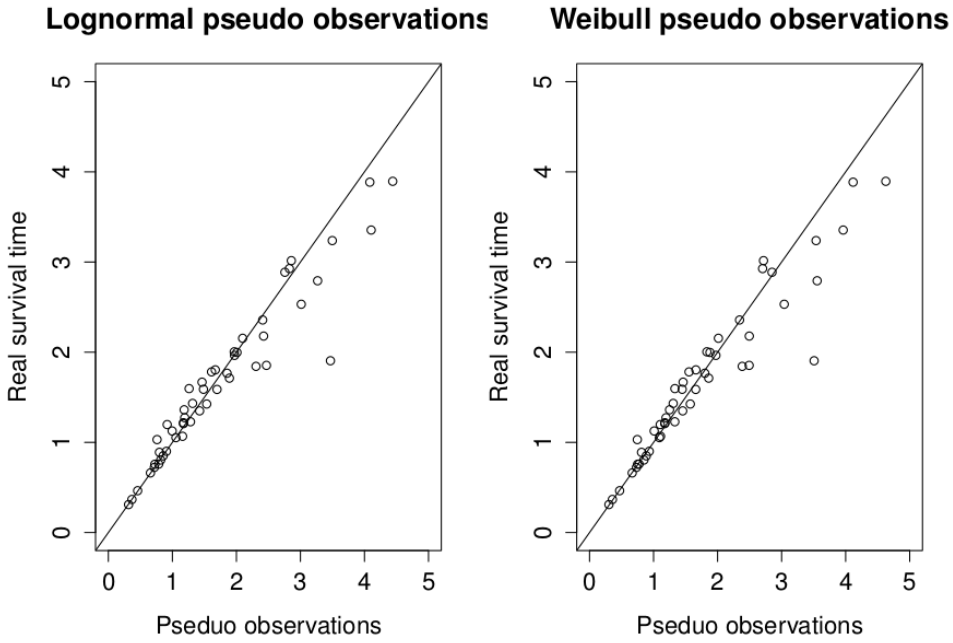

Figure 5.24: Plot of parametric pseudo observations for Lognormal and Weibull distribution against real Weibull survival times. $\boldsymbol{\beta} = (0, 0.7, 0.6, 0.2)$, $\sigma = 0.2$ and there is 30% censoring.

5.5 Estimation with pseudo observations

Once we have obtained a set of pseudo observations, we may want to use them in model regression. It is therefore of interest to study how well estimation of parameters can be done with the different methods. With R we have simulated a Weibull data set consisting of 10 000 observations from an AFT model with a single standard normal distributed covariate. For the same data set, we adjust the censoring level so that we get 0%, 25%, 50% and 75% censoring.

For all levels of censoring, we then apply the three methods to find pseudo observations. With *survreg*, we can estimate parameters from the different sets of pseudo observations. In addition we estimate parameters with *survreg* on the original data set. All estimated values are in table 5.3.

Table 5.3: Estimated parameters from the original simulated data set with 10 000 observations and three pseudo observation data sets for four levels of censoring.

0 % censoring	True	KM	KMlog	Parametric	Survreg on original
μ	0	0.00258	0.00258	0.0129	0.00258
β_1	1	0.99897	0.99897	1.0007	0.99897
σ	0.5	0.502	0.502	0.515	0.502
25 % censoring	True	KM	KMlog	Parametric	Survreg on original
μ	0	-0.0591	0.0991	0.0421	-0.00105
β_1	1	0.9719	1.2167	1.0498	0.99660
σ	0.5	0.859	0.948	0.617	0.501
50 % censoring	True	KM	KMlog	Parametric	Survreg on original
μ	0	·	1.152 *	0.4148	0.279
β_1	1	·	1.465 *	1.3316	1.171
σ	0.5	·	1.48 *	1.06	0.576
75 % censoring	True	KM	KMlog	Parametric	Survreg on original
μ	0	·	0.849	0.418	-0.00863
β_1	1	·	1.635	1.003	0.99288
σ	0.5	·	4.72	2.33	0.496

NOTES:
· For data sets where KM gives negative values we will not get any estimates from *survreg*.
* For some sets of observations *survreg* does not converge. We then set initial values to (0,1) to get the results in the table.

Based on this simulation, it seems like the parametric method is the best choice of method to obtain pseudo observations when there are censoring and we want to estimate parameters. KM will be useless in *survreg* when the

pseudo observations are negative, and KMlog gives parameter values that are too high compared to the true values. The parametric method gives values that are large compared to *survreg* on the original data set, but are still close to the real values compared to the other pseudo methods. Because we know the distribution of the data it is reasonable that the parametric method can generate parameter values similar to the true values, but as we see, not necessarily that well for high levels of censoring.

This is only one simulation, and more research should be conducted before drawing any clear conclusions.

5.6 Prostatic cancer example

In this section we will apply the three methods for finding pseudo observation on the survival times of prostatic cancer patients in a clinical trial. The data set can be found in appendix B, and are obtained from table 1.4 in Collett [7]. The trial consisted of 38 patients who received either DES treatment or placebo treatment. In table B.3 in the appendix, DES treatment is represented with 1 and placebo with 0. Individuals who died from other causes than prostatic cancer, or for other reasons where lost during the follow up period, are regarded as censored. Censoring times will be right censored, and in this data set there are 32 censored observations.

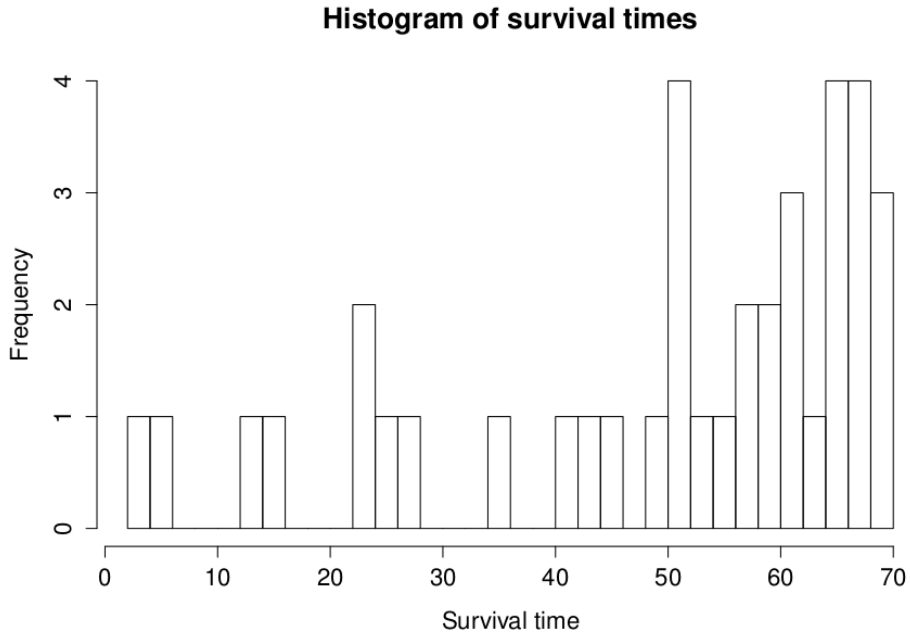

Figure 5.25: Histogram of survival times for prostatic cancer trial.

For all patients, age, Serum haemoglobin level, tumour size and Gleason

index was measured. Histogram of the observed survival times are in figure 5.25. Figures showing survival times plotted against covariates are given in figure 5.26 and 5.27. From figure 5.26 it looks like tumour size and Gleason index has the most effect on the survival times, and then age. Figure 5.27 shows survival times for both treatments, and it does not look like the treatment have any significant effect on the survival time.

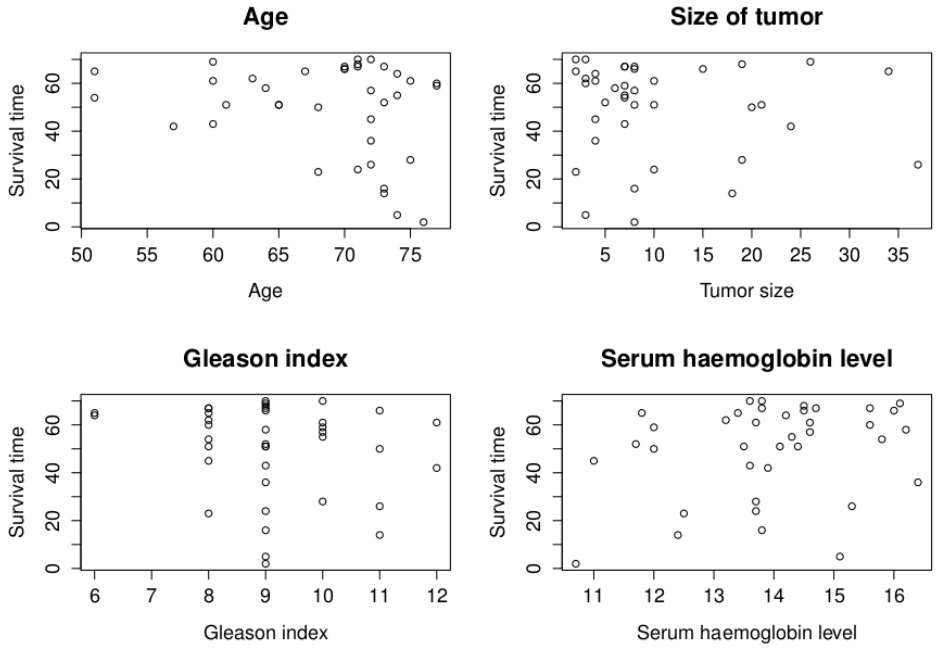

Figure 5.26: Plot of survival times versus covariates for the prostatic cancer data set.

Collett fits both a Cox-proportional hazards model (examples 3.6 and 3.10 in [7]) and an accelerated failure time model (example 6.4 in [7]) to this data set. For both models Collett found that tumour size and Gleason index where the significant covariates. Treatment was still added, giving AFT model,

$$log(T_i) = \mu + \beta_{Size}X_{Size,i} + \beta_{Index}X_{Index,i} + \beta_{Treat}X_{Treat,i} + \sigma\epsilon_i. \qquad (5.2)$$

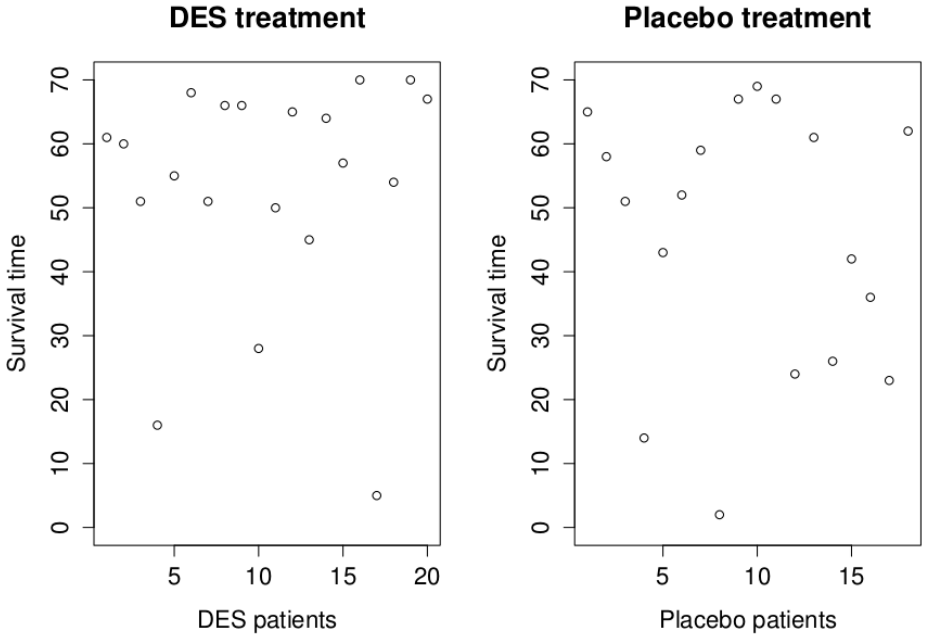

Figure 5.27: Plot of survival times for both treatments in the prostatic cancer data set.

5.6.1 Nonparametric methods

Pseudo observations for this data set found with KM and KMlog will not depend on any of the covariates or any assumed distributions. Plots of pseudo observations and the observed survival times are found in figure 5.28. Red indicates observed survival times and black indicates pseudo observations, circles are for censored observations and dots for uncensored.
Because almost all observations are censored we get a lot of pseudo observations with similar survival times for both KM and KMlog. Table B.3 in appendix B shows the value for each pseudo observations, and we see that KM and KMlog are very close for all observations.

KM and KMlog can now be treated as two new data sets. Since they are so similar we will only look at the KM pseudo observations. A histogram showing KM pseudo observations is in figure 5.29, and plots of covariates vs. KM pseudo observations can be found in figure 5.30 and 5.31.

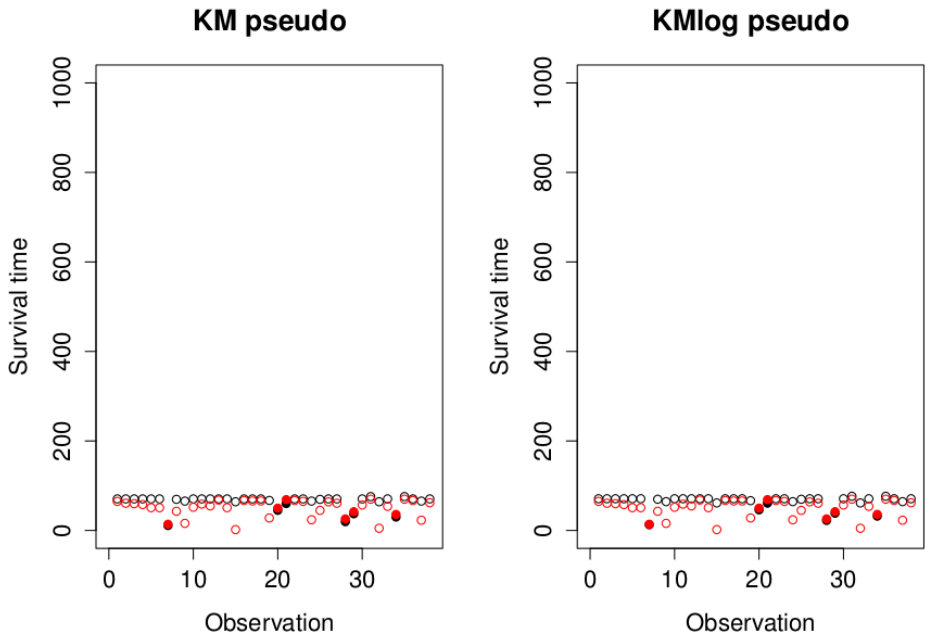

Figure 5.28: KM/KMlog pseudo observations (black) and observed survival times (red). Observed times (•) and censored times (◦) for the prostatic cancer data set

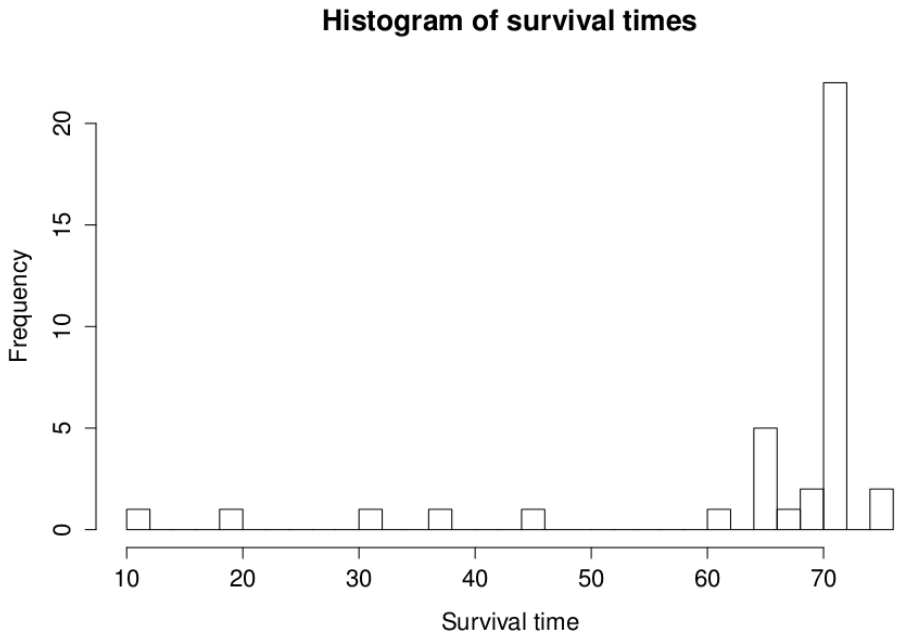

Figure 5.29: Histogram for KM pseudo observations for the prostetic cancer data set

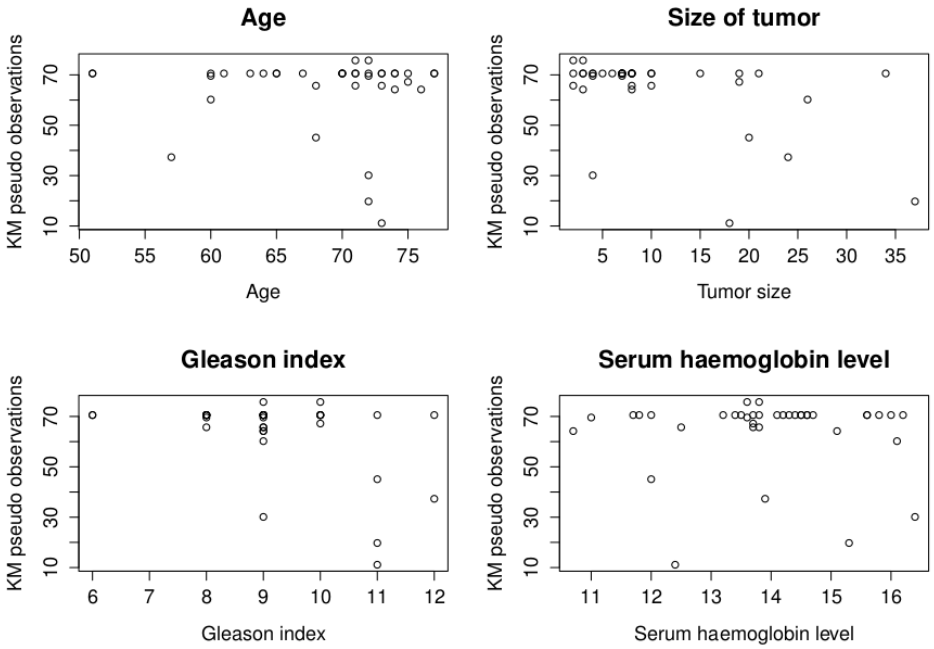

Figure 5.30: Plot of KM pseudo observations against covariates for the prostatic cancer data set.

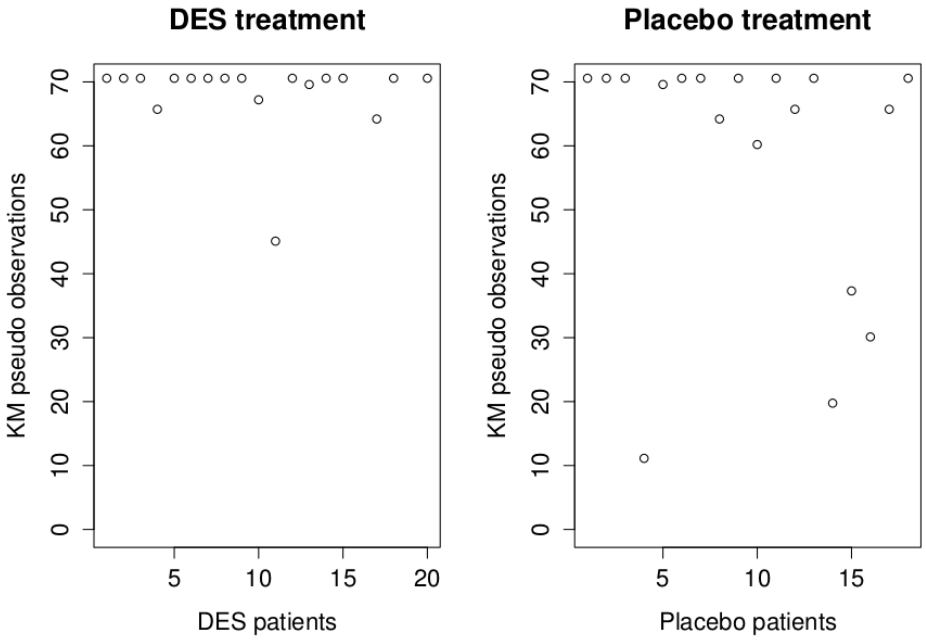

Figure 5.31: Plot of survival times for both treatments in the prostatic cancer data set.

Again we see that the survival time might depend on Gleason index and tumour size. Collett assumes that the survival times are loglogistic, but our main focus is on Weibull and lognormal distributions. Therefore we fit model 5.2 to the KM pseudo observations for Weibull and lognormal survival times. That gives us the following output in *R*:

```
Call:
survreg(formula = Surv(PCData$KMPseudo) ~ PCData$TumorSize +
    PCData$Treatment + PCData$GleasonIndex, data = PCData,
    dist = "weibull")
                       Value Std. Error       z         p
(Intercept)          4.42610    0.17803  24.862 1.91e-136
PCData$TumorSize    -0.00417    0.00271  -1.538  1.24e-01
PCData$Treatment     0.04103    0.04563   0.899  3.69e-01
PCData$GleasonIndex -0.02014    0.01886  -1.068  2.86e-01
Log(scale)          -1.97857    0.15385 -12.861  7.50e-38

Scale= 0.138

Weibull distribution
Loglik(model)= -149.1   Loglik(intercept only)= -151.4
Chisq= 4.66 on 3 degrees of freedom, p= 0.2
Number of Newton-Raphson Iterations: 21
n= 38

Call:
survreg(formula = Surv(PCData$KMPseudo) ~ PCData$TumorSize +
    PCData$Treatment + PCData$GleasonIndex, data = PCData,
    dist = "lognormal")
                       Value Std. Error      z        p
(Intercept)          4.8753    0.38330  12.72 4.63e-37
PCData$TumorSize    -0.0127    0.00662  -1.92 5.48e-02
PCData$Treatment     0.1741    0.10582   1.65 1.00e-01
PCData$GleasonIndex -0.0792    0.04300  -1.84 6.55e-02
Log(scale)          -1.1474    0.11471 -10.00 1.48e-23

Scale= 0.317

Log Normal distribution
Loglik(model)= -166.5   Loglik(intercept only)= -173.8
Chisq= 14.59 on 3 degrees of freedom, p= 0.0022
Number of Newton-Raphson Iterations: 4
n= 38
```

We see that none of the covariates are significant if we use the KM pseudo observations to estimate. This is not surprising since so many observations have similar survival times, although their covariates are different.

5.6.2 Parametric pseudo observations

To have some survival times to compare the parametric pseudo observations to we use *survreg* to fit a Weibull and lognormal model to the original data set:

```
Call:
survreg(formula = Surv(PCData$Survivaltime, PCData$Status) ~
    PCData$TumorSize + PCData$Treatment + PCData$GleasonIndex,
    data = PCData, dist = "weibull")
                    Value Std. Error      z        p
(Intercept)          7.731     1.4545  5.315 1.06e-07
PCData$TumorSize    -0.037     0.0174 -2.126 3.35e-02
PCData$Treatment     0.434     0.4633  0.937 3.49e-01
PCData$GleasonIndex -0.269     0.1162 -2.317 2.05e-02
Log(scale)          -0.990     0.3489 -2.837 4.55e-03

Scale= 0.372

Weibull distribution
Loglik(model)= -31.4   Loglik(intercept only)= -39.3
Chisq= 15.64 on 3 degrees of freedom, p= 0.0013
Number of Newton-Raphson Iterations: 7
n= 38
```

and:

```
Call:
survreg(formula = Surv(PCData$Survivaltime, PCData$Status) ~
    PCData$TumorSize + PCData$Treatment + PCData$GleasonIndex,
    data = PCData, dist = "lognormal")
                    Value Std. Error     z        p
(Intercept)          7.927     1.7234  4.60 4.24e-06
PCData$TumorSize    -0.026     0.0199 -1.30 1.92e-01
PCData$Treatment     0.775     0.4758  1.63 1.03e-01
PCData$GleasonIndex -0.329     0.1651 -1.99 4.63e-02
Log(scale)          -0.442     0.3079 -1.43 1.51e-01

Scale= 0.643

Log Normal distribution
Loglik(model)= -31.9   Loglik(intercept only)= -39.2
Chisq= 14.64 on 3 degrees of freedom, p= 0.0021
Number of Newton-Raphson Iterations: 6
n= 38
```

Plot of survival times estimated with these parameters are presented in figure 5.32. The red are the observed times and the black are the estimated survival times. Circles represents censored survival times and dots represent uncensored survival times. We see that these observations are much higher than the KM and KMlog pseudo observations in figure 5.28.

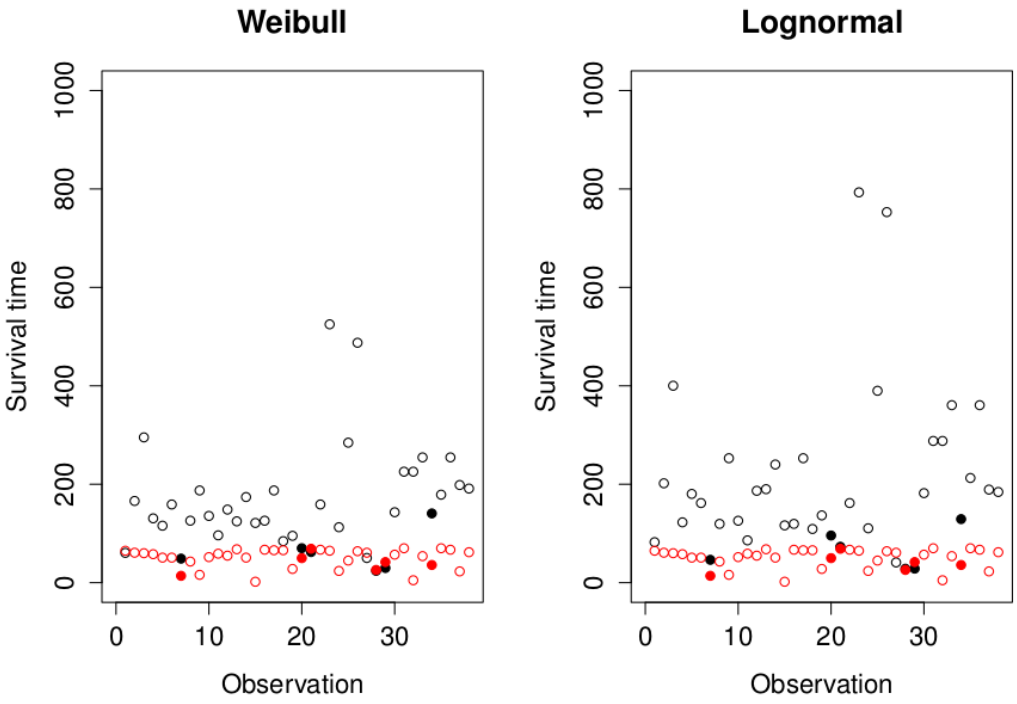

Figure 5.32: For the prostatic cancer data set: Plot of survival times estimated with *survreg* (black) and observed (red) for Weibull and lognormal distribution. Uncensored survival times (•) and censored survival times (○).

Parametric pseudo observations can take advantage of the assumed distribution and measured covariates. They may therefore be a better choice when there are large levels of censoring and we are sure that we have the right distribution. Comparing the estimated survival times in figure 5.32 to the parametric pseudo observations in figure 5.33, we see that the parametric pseudo observations are similar to the ones made just with *survreg*. To see the difference we look at table B.3 in appendix B. From the table we see that there are some very high observations for the parametric pseudo observations.

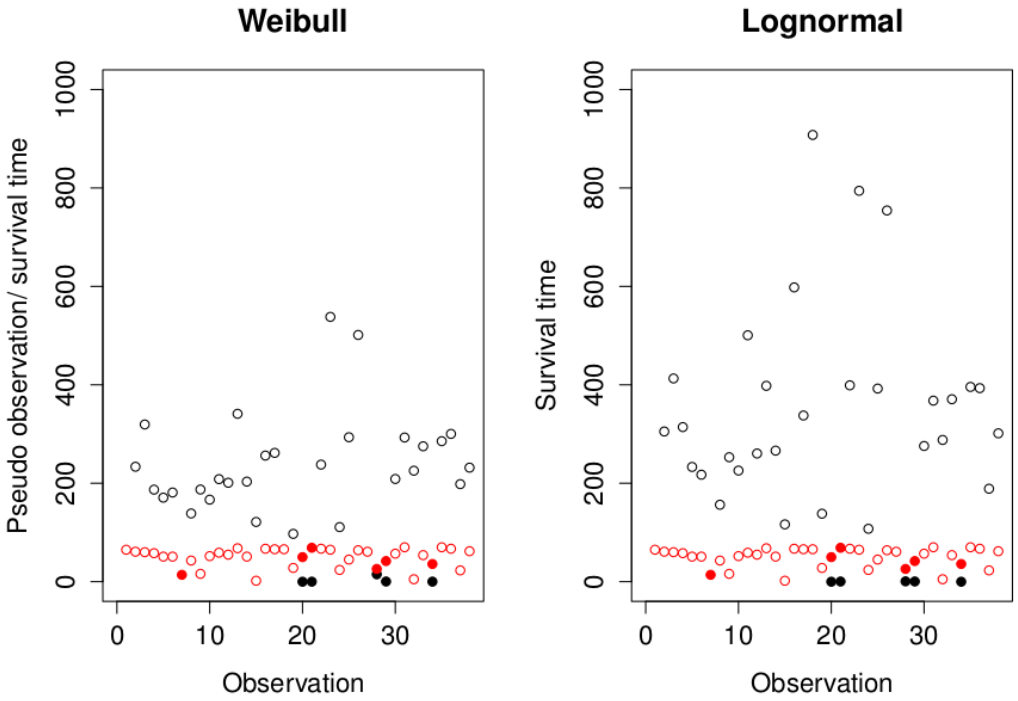

Figure 5.33: Parametric pseudo observations (black) and observed survival times (red) for Weibull and lognormal distribution. Censored survival times are (○) and uncensored survival times are (●). Obs: high pseudo observations are not included. The data set is the Prostatic cancer data set.

Estimated parameters from the parametric pseudo observations are:

```
Call:
survreg(formula = Surv(PCData$ParamPseudoweibull)
  ~ PCData$TumorSize + PCData$Treatment + PCData$GleasonIndex,
    data = PCData, dist = "weibull")
                      Value Std. Error      z        p
(Intercept)          5.2559     4.9114  1.070 2.85e-01
PCData$TumorSize    -0.0236     0.0875 -0.270 7.87e-01
PCData$Treatment    -0.1848     1.4876 -0.124 9.01e-01
PCData$GleasonIndex  0.0622     0.5179  0.120 9.04e-01
Log(scale)           1.4614     0.1510  9.680 3.67e-22

Scale= 4.31
```

and:

```
Call:
survreg(formula = Surv(PCData$ParamPseudolognormal)
  ~ PCData$TumorSize + PCData$Treatment + PCData$GleasonIndex,
    data = PCData, dist = "lognormal")
                     Value Std. Error     z        p
(Intercept)         22.754     13.124  1.73 8.30e-02
PCData$TumorSize    -0.229      0.227 -1.01 3.13e-01
PCData$Treatment    -3.622      3.623 -1.00 3.17e-01
PCData$GleasonIndex -1.645      1.472 -1.12 2.64e-01
Log(scale)           2.386      0.115 20.80 4.31e-96

Scale= 10.9

Log Normal distribution
Loglik(model)= -275   Loglik(intercept only)= -276.9
Chisq= 3.93 on 3 degrees of freedom, p= 0.27
Number of Newton-Raphson Iterations: 3
n= 38
```

None of the parameters estimated for model 5.2 are significant for either Weibull or lognormal distribution. Looking at the scale we see that both distributions lead to very wide probability density functions. This may be a result of the high level of censoring, which we have seen can lead to very high parametric pseudo observations.

A conclusion from this example will be that with very high level of censoring, neither of the three pseudo observation methods are any good for estimating parameters. The high level of censoring makes most observed survival times smaller than they should be. Not knowing enough about the actual survival times makes KM and KMlog stay close to the observed times, and we get a short spread of pseudo survival times. The parametric pseudo method on the other hand, may give too much variation. Because we assume a distribution for the data the parametric method try to fit a model with that distribution. When there are high levels of censoring the information we have may not be enough to fit this model right.

Chapter 6

Pseudo Residuals

6.1 Introduction to Pseudo Residuals

We have now seen examples of methods for creating pseudo observations that can be treated as uncensored data sets. With these new uncensored data sets we can find uncensored residuals and use them in model checking. Another option is to find pseudo residuals without calculating pseudo observations first. In this chapter we look at this for standardized and Cox-Snell residuals.

The expected value of data from a unit exponential distribution can be found using

$$\hat{\theta} = \frac{\sum_{i=1}^{n} u_i}{n}, \tag{6.1}$$

and if the data set is censored we can use

$$\hat{\theta} = \frac{\sum_{i=1}^{n} u_i}{\sum_{i=1}^{n} \delta_i}. \tag{6.2}$$

where δ_i is the event indicator introduced in section 2.2.

This can be used to estimate pseudo observations for the exponentially distributed data set. In figure 6.1 we see pseudo observations for an uncensored and a censored data set with unit exponential survival times. Censoring times are found with exponential censoring. From the figure we see that uncensored observations gives pseudo observations equal to the observed time. This can easily be found by setting equation 6.2 into equation 5.1. Pseudo observations for the censored data set behaves similar to pseudo observations created with the KM method, and can be negative.

We know that if the estimated model is close to the real model, Cox-Snell residuals will behave like a sample from a unit exponential distribution. When we have found a set of possibly censored Cox-Snell residuals, we can therefore use 6.3 to find an uncensored set of Cox-Snell residuals.

$$\hat{\theta}_i = n\frac{U}{\Delta} - (n-1)\frac{U - u_i}{\Delta - \delta_i}. \tag{6.3}$$

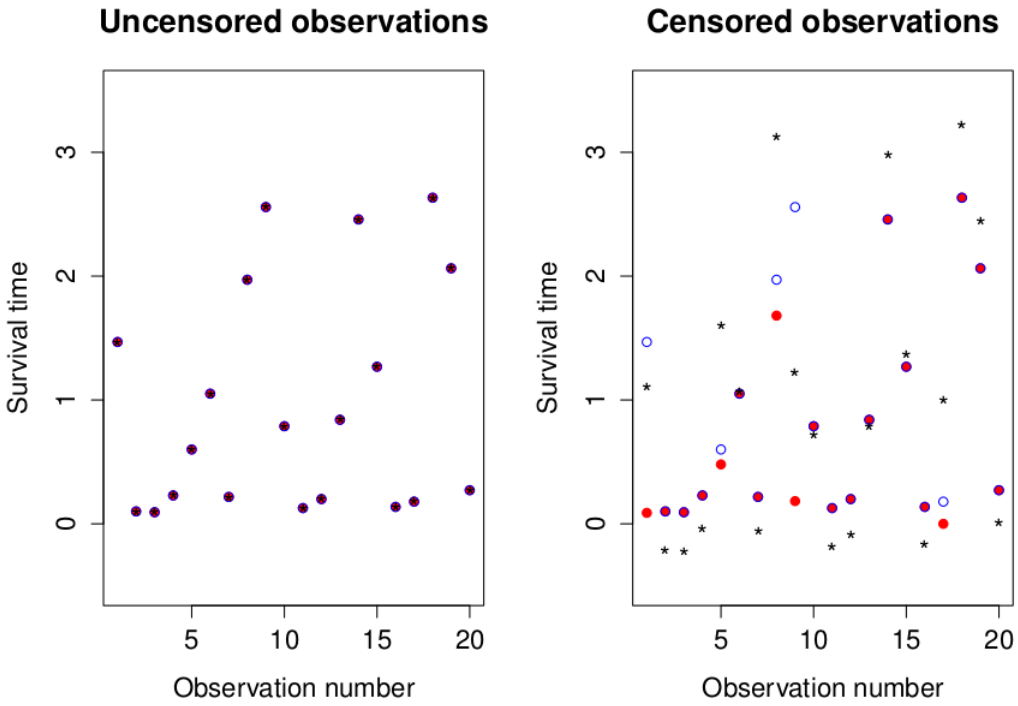

Figure 6.1: Pseudo observations for unit exponentially distributed survival times.

where u_i is Cox-Snell residual i, $U = \sum_{i=1}^{n} u_i$ and $\Delta = \sum_{i=1}^{n} \delta_i$. This is an alternative to the adjusted Cox-Snell residuals, and we will call them pseudo residuals.

6.1.1 Pseudo residuals for data set W1

For data set W1 we can find estimates for the parameters using *survreg*. That gives us the following output in R:

```
Call:
survreg(formula = Surv(Data$Time, Data$Status) ~ Data$x1 + Data$x2
    + Data$x3, data = Data, dist = "weibull")
              Value Std. Error       z      p
(Intercept)  0.00355      0.338  0.0105 0.9916
Data$x1      0.83609      0.490  1.7051 0.0882
Data$x2     -0.08139      0.526 -0.1547 0.8771
Data$x3      0.38207      0.220  1.7398 0.0819
Log(scale)  -0.36127      0.221 -1.6353 0.1020

Scale= 0.697

Weibull distribution
Loglik(model)= -18.1   Loglik(intercept only)= -22.5
```

```
Chisq= 8.77 on 3 degrees of freedom, p= 0.033
Number of Newton-Raphson Iterations: 6
n= 20
```

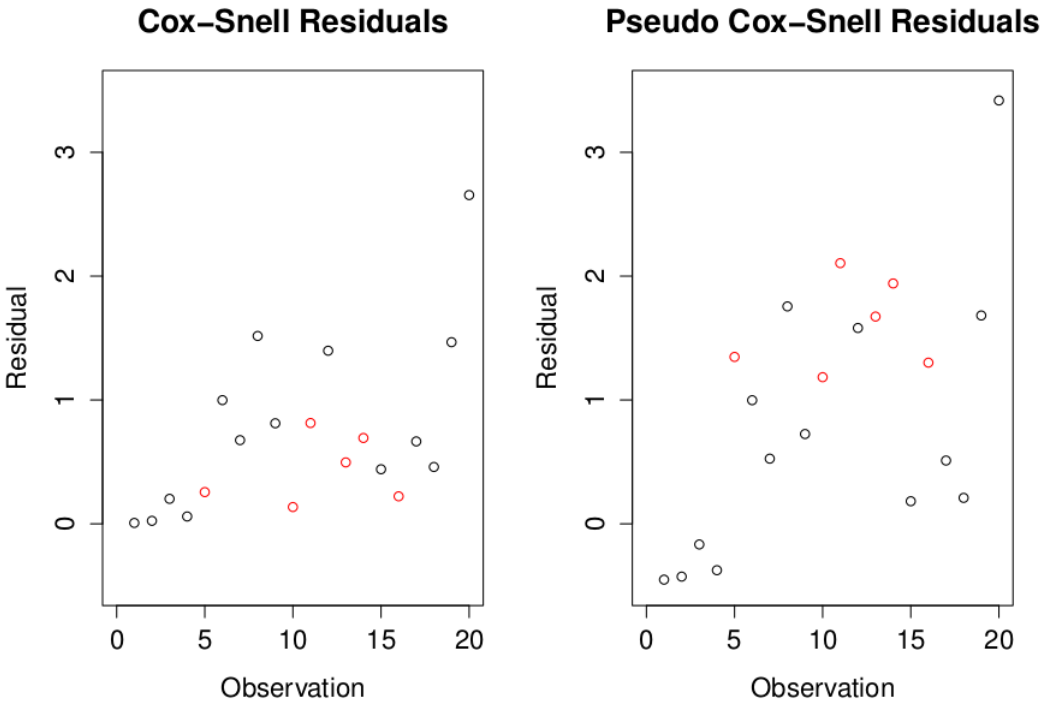

Figure 6.2: Cox-Snell residuals and Pseudo Cox-Snell residuals for W1 data set. Red ∘ indicate censored observations, and black ∘ indicate uncensored observations.

Using the estimated parameters, Cox-Snell and pseudo residuals can be found. In figure 6.2 both types of residuals are plotted. The red circles are the residuals that originate from censored observations and the black circles are from uncensored observations. The figure shows us that pseudo residuals are different from the standard Cox-Snell residuals, and that censored observations are more different than the uncensored, which is what we want. In figure 6.3 the difference between pseudo residuals and Cox-Snell residuals are plotted against the value of the Cox-Snell residuals. The black horizontal line indicates where the points would be if there were no difference, and the red horizontal line indicates where censored observations would lie if pseudo residuals were equal to 1-adjusted Cox-Snell residuals. Notice that the points lie on neither of these lines. Instead they lie on two other increasing lines, one for censored observations and one for uncensored observations. To see why we get this we rewrite equation 6.3 to

$$\hat{\theta}_i = \left(\frac{n}{\Delta} - \frac{n-1}{\Delta - \delta_i}\right)U + \left(\frac{n-1}{\Delta - \delta_i}\right)u_i.$$

Since figure 6.2 shows the difference between the Cox-Snell residual and the pseudo residual we have

$$\hat{\theta}_i - u_i = \left(\frac{n}{\Delta} - \frac{n-1}{\Delta - \delta_i}\right)U + \left(\frac{n-1}{\Delta - \delta_i} - 1\right)u_i.$$

This is on the form $f(u_i) = a_{\delta_i} + b_{\delta_i}u_i$, explaining why we get one line for $\delta_i = 0$ and another for $\delta_i = 1$, with different intercept and slope. Unlike adjusted residuals ,which adds the same value to all censorded observations, these pseudo residuals add (or subtract) a value to all Cox-Snell residuals, depending on the value of the residual, if the observation is censored or not and the total level of censoring.

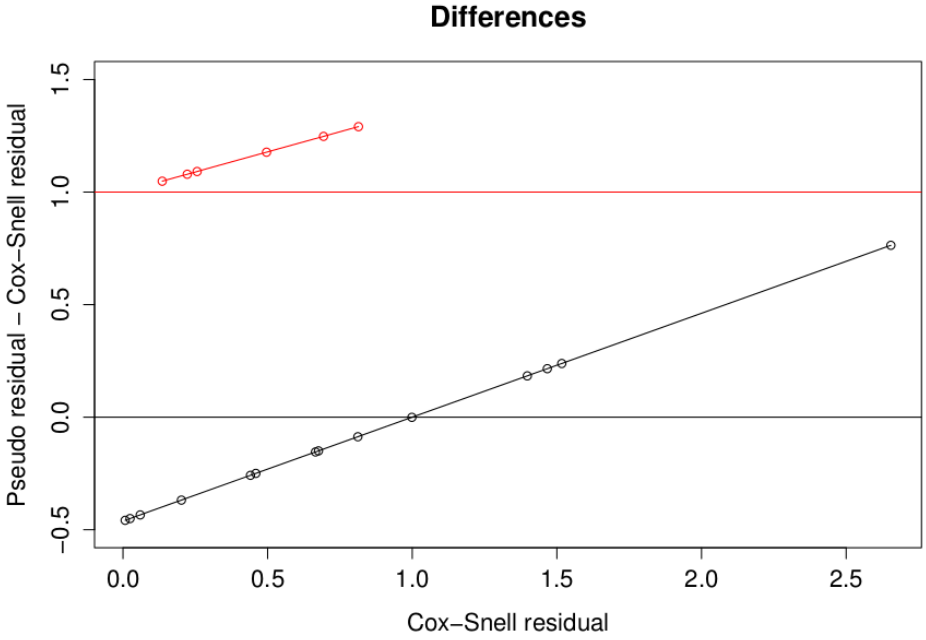

Figure 6.3: Pseudo residuals and Cox-Snell residuals for the W1 data set plotted against the value of the Cox-Snell residual.

When we use residuals in model checking and analysis, we often benefit from looking at $log(u)$ in stead of u. Unfortunately we can get negative pseudo residuals, and negative u will lead to a undefined $log(u)$. A solution may be to remove the negative pseudo residuals before continuing with the analysis. Because the negative residuals always come from uncensored observations an alternative solution may be to only replace the censored residuals with pseudo residuals and keep the uncensored ones. In [2], Andersen et al. points out that all observations must be replaced by pseudo observations. Therefore we will look at another option. That option is to treat residuals as survival times and use methods described in chapter 5 to find pseudo residuals.

6.2 KM, KMlog and parametric Cox-Snell pseudo residuals

KM and KMlog can be used on pseudo residuals the same way we use them on survival times. KM can, as residuals based on equation 6.3, give negative pseudo residuals. However, in some cases where 6.3 are negative, KM can give positive residuals. If KM also gives negative Cox-Snell residuals, we should use KMlog or the parametric method, because Cox-Snell residuals should be positive.

To make parametric pseudo residuals more robust we will use the fact that the exponential distribution is a special case of the Weibull distribution. We can therefore use a Weibull AFT model without covariates to find parametric Cox-Snell pseudo residuals. This gives us

$$ln(u_i) = \mu + \sigma\epsilon_i \tag{6.4}$$

where ϵ_i is Gumbel distributed. Treating residuals this way makes pseudo residuals more robust since it allows them to have any Weibull distribution, and not be restricted to the exponential distribution. However, if the model is appropriate we will have $\gamma = \frac{1}{\sigma}$=1, $\lambda = exp\left[\frac{-\mu}{\sigma}\right] = 1 \Rightarrow \sigma = 1$, $\mu = 0$.

The parametric and KMlog method will always give positive pseudo residuals, but can be very big compared to the Cox-Snell residuals. We will therefore recommend using KM to find pseudo residuals first, and then proceed with KMlog or the parametric method if necessary. Figure 6.4 shows pseudo residuals made with the three methods for dataset W1, plotted against Cox-Snell residuals. Lines show where points would be if the difference were 0 or 1.

6.3 Pseudo standardized residuals

Pseudo residuals for standardized residuals should be found with KM. This is first of all because standardized residuals can be negative, and KMlog and parametric pseudo observations give strictly positive pseudo residuals. Figure 6.5 shows standardized residuals and pseudo standardized residuals for dataset W1.

A conclusion for pseudo residuals is that when we know the distribution of the error term, pseudo standardized residuals will be preferred because we can use the KM method. If we don't know the distribution, Cox-Snell residuals may be easier to use because we know they should be unit exponentially distributed. For Cox-Snell residuals we recommend using KMlog. That way, the residuals are not restricted to be any distribution, and we will not have any negative pseudo residuals.

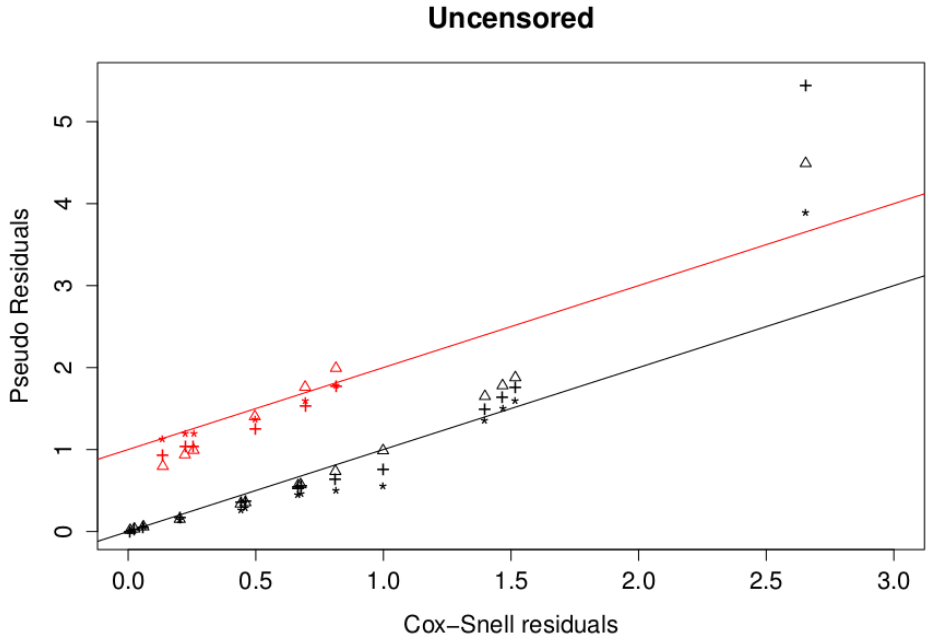

Figure 6.4: Pseudo residuals for W1 plotted against Cox-Snell residual. KM (*), KMlog (+) and Parametric (△). Censored observations are red, and uncensored observations are black. Lines show where the points would be if the difference were one (red) or zero (black).

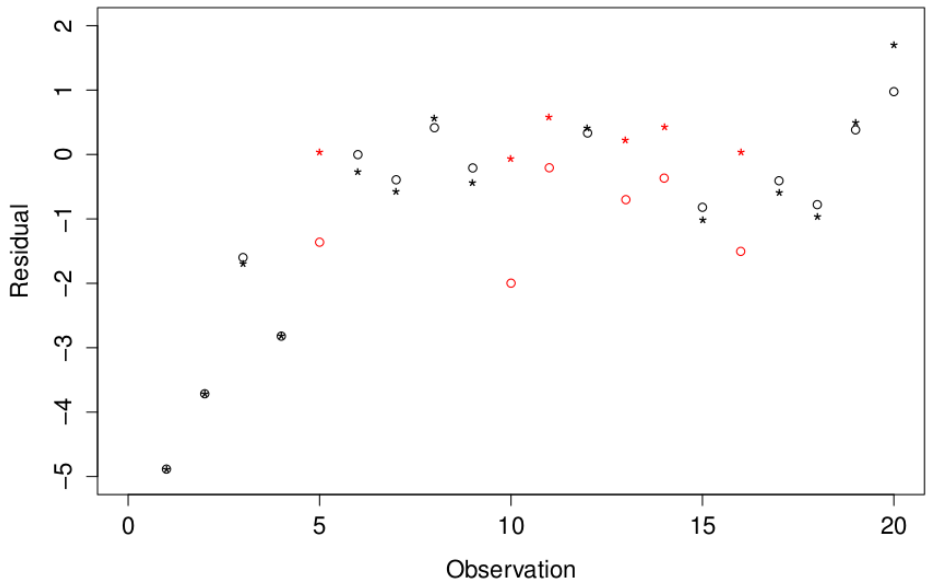

Figure 6.5: Standardized residuals (circle) and pseudo standardized residual for W1 dataset. Red observations are censored and black observations are uncensored.

Chapter 7

Model Checking and Functional Form

7.1 Analysis of residuals

After a model is fitted for our data, and parameters are estimated, residuals are common tools for analysing the fit. As mentioned, censored observations will lead to censored residuals. To compensate for this, several adjustments and smoothing methods have been applied to the residuals. Examples are the 1- and log(2)-adjusted residuals introduced in section 4.5.3. Pseudo residuals introduced in the previous chapter is an alternative to these adjustments. Yet another alternative, that we are going to look at, is to find residuals for the pseudo observations presented in chapter 5. Because pseudo observations are an uncensored version of the original data set, residuals calculated from them should not be censored.

When we have obtained residuals, several plots can be used to see if there are any reasons to suspect that the model is inappropriate. In Collett [7], a cumulative hazard or log-hazard plot of Cox-Snell residuals is suggested. This should give a straight line with unit slope and intercept in zero if the model is appropriate for the data. Dobson and Barnett [8] suggest a sequence plot where residuals are ordered after survival time and plotted in order to identify outliers and dependencies. Plots of survival times against explanatory variables can also be used to identify patterns and detect systems that indicates that the model is not correct. Yet another alternative is to plot an exponential probability plot of Cox-Snell residuals and check if they fit a unit exponential distribution.

7.1.1 Example, Nelson's superalloy data

In this example we will study Nelson's superalloy data. The data consists of 26 observations where the survival time is the number of cycles (measured in thousands) and the only explanatory variable is pseudostress measured in ksi. The data set is included in appendix B, and plot of the data can be found in figure 7.1. This data set have also been studied in Meeker and Escobar [16],

Aaserud [1] and Lindqvist et al. [15]. They studied the relationship between the logarithm of the pseudostress and the number of cycles, and assumed the following model,

$$log(T) = \mu + \beta_1 X + \beta_2 X^2 + \sigma\epsilon, \tag{7.1}$$

where ϵ is Gumbel, and hence T is Weibull.

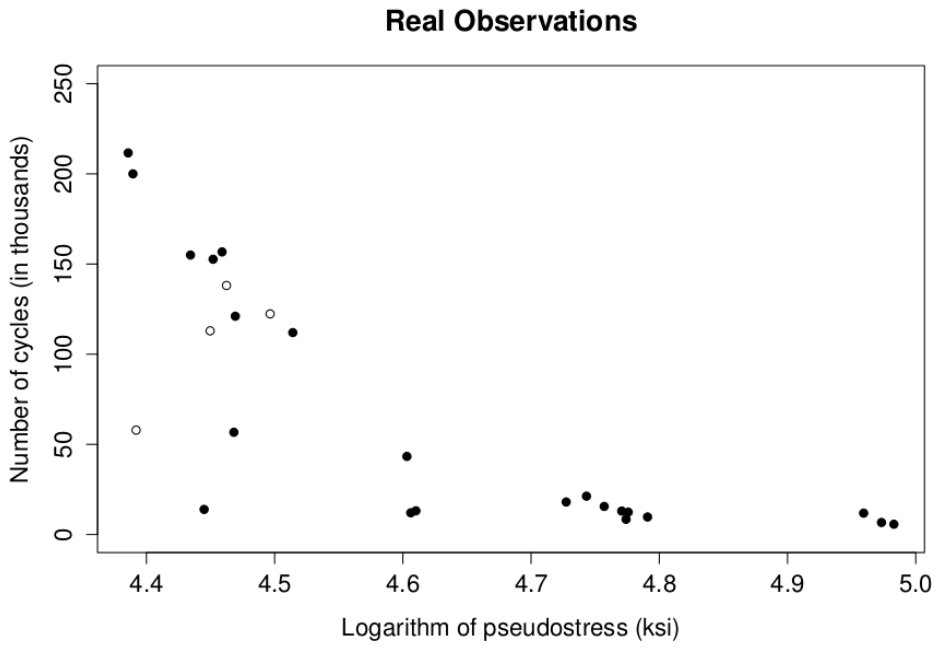

Figure 7.1: Nelsons superally data set. Number of cycles (in thousands) plotted against the logarithm of the pseudostress (measured in ksi). Censored observations are ∘, uncensored are •.

With *survreg* we can fit this model and find standard Cox-Snell residuals (figure 7.2). These residuals are censored. To compensate for this we will study two options, finding pseudo observations and their Cox-Snell residuals, or pseudo residuals from the ordinary Cox-Snell residuals.

The KM-method can be used to find pseudo observations for this data set. They are shown in figure 7.3. There are no negative pseudo observations so we do not have to use KMlog or the parametric method. From the figure, we see that the four pseudo observations corresponding to censored survival times are larger than the observed times. Among the uncensored observations we see that the biggest uncensored observations also have pseudo observation values larger than the observed time, but some of the middle size pseudo observations are smaller than the observed times.

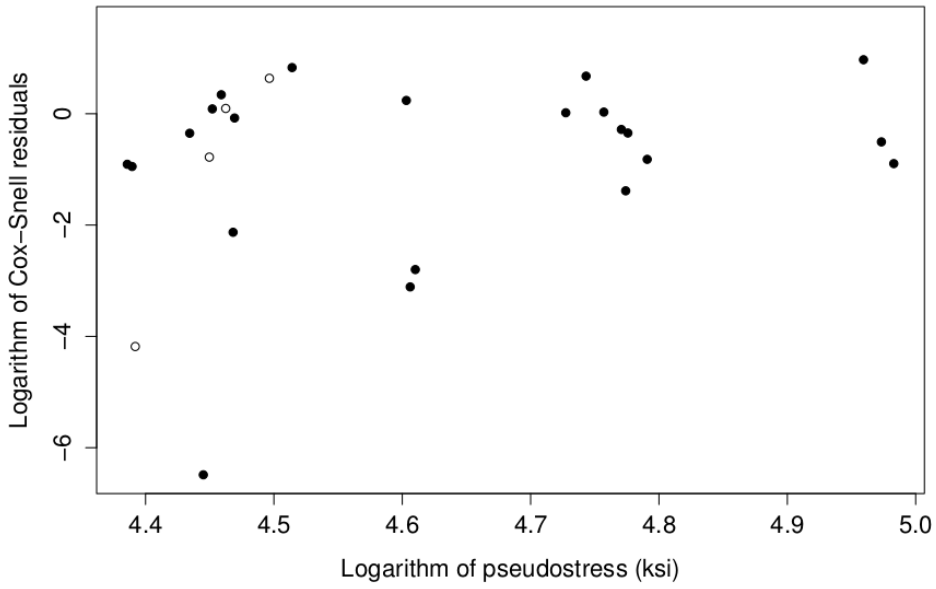

Figure 7.2: Logarithm of Cox-Snell residuals for the original superalloy data set. Censored observations are circles, uncensored are dots

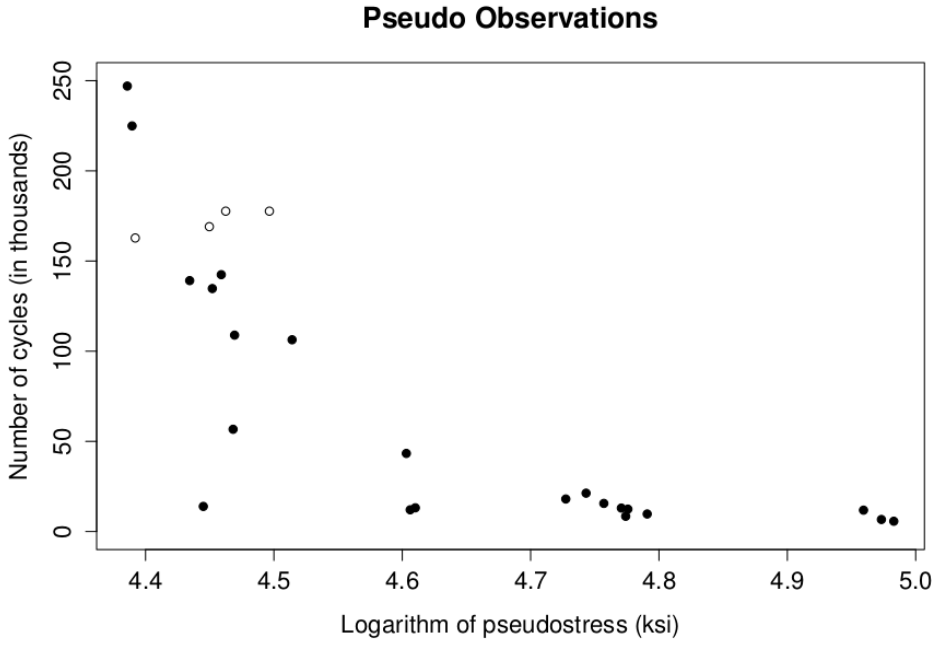

Figure 7.3: KM pseudo observations of th number of cycles in Nelson's superalloy data, plotted against log(pseudostress). Censored observations are circles, uncensored are dots

The new data set can now be used to find Cox-Snell residuals, which again can be used in model checking. Figure 7.4 shows the logarithm of the Cox-Snell residuals plotted against the covariate and an exponential probability plot. Most observations fluctuate around zero, but some values for $log(r_{c,i})$

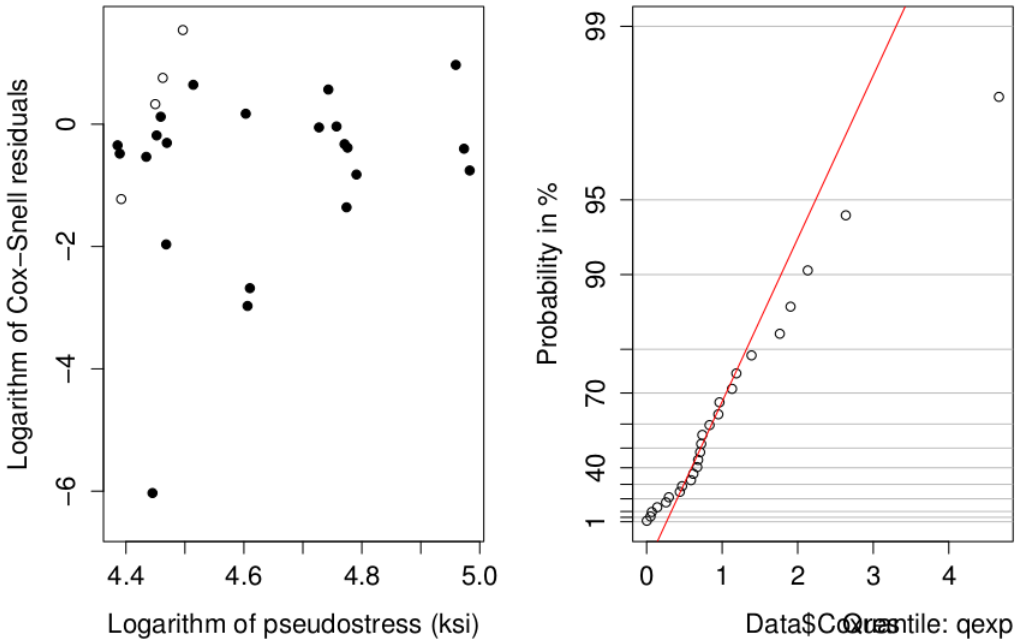

Figure 7.4: Left: Logarithm of Cox-Snell residuals for pseudo observations vs. logarithm of pseudostress. Residuals from censored observations are circles, and uncensored are dots. Right: Exponential probability plot for Cox-Snell residuals from pseudo observations.

are very negative compared to the rest. This may indicate that there is some dependency on the pseudostress, but if we remove the smallest residual, this plot is not so bad. On the other hand, the probability plot does not show a straight line, which also indicates that the model might be inappropriate.

Pseudo Cox-Snell residuals are obtained with the KM-method on the original Cox-Snell residuals. Plots similar to the ones in figure 7.4 for pseudo residuals are found in figure 7.5. From the left plot we see that pseudo Cox-Snell residuals for lower values of $log(r_{c,i})$ also are negative, and even more so than for Cox-Snell residuals from pseudo observations. It therefore looks like there is a dependency between pseudostress and the number of cycles that are not accounted for in the model. The exponential probability plot at the right looks linear, but might be slightly convex.

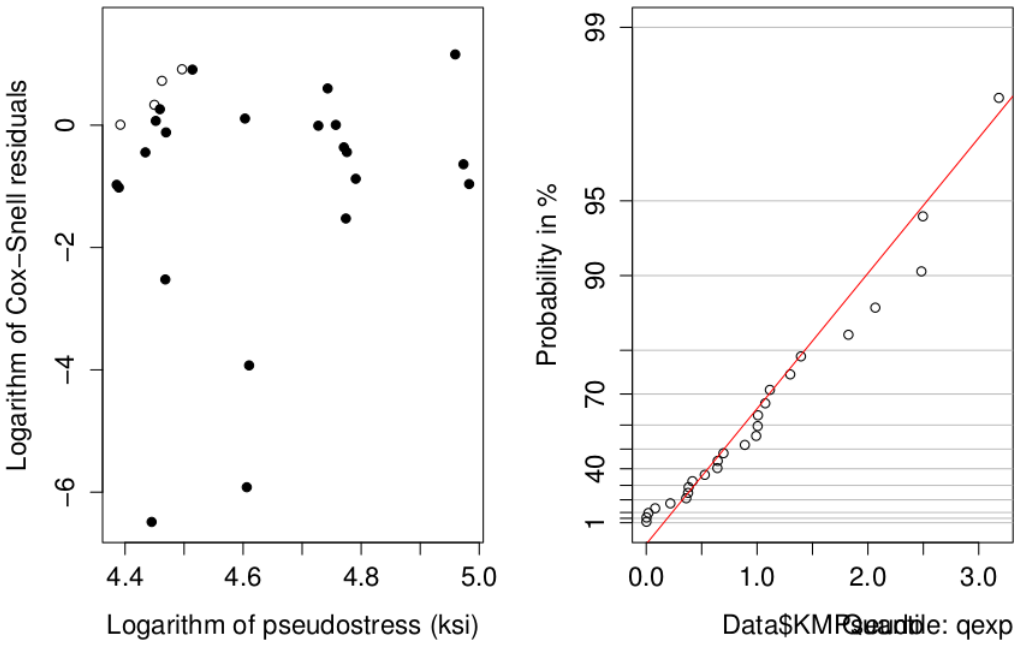

Figure 7.5: Left: Logarithm of pseudo Cox-Snell residuals vs. logarithm of pseudostress. Residuals from censored observations are circles, and uncensored are dots. Right: Exponential probability plot for pseudo Cox-Snell residuals.

7.2 Functional form for covariates

In chapter 4 we stated that AFT models can be written on the form

$$log(T) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \sigma\epsilon.$$

Alternatively, we can write

$$log(T) = f(X) + \sigma\epsilon,$$

where

$$f(X) = \mu + \beta_1 X_1 + \cdots + \beta_p X_p,$$

or

$$log(T) = \mu + f_1(X_i) + \cdots + f_p(X_p) + \sigma\epsilon,$$

where $f_j(\cdot)$, $j = 1, ..., p$ is called *covariate functions*, which tell us something about how each covariate j influences the survival.

If residual plots or other diagnostic tools make us suspect that the model we use is inappropriate, changing the functional form of one or more covariate might be a solution. Two procedures are suggested in section 5 of [15]. The first is to start out with a model without any covariates and then,

for each covariate X_j, find the best covariate function $f_j(X_j)$ for when X_j is the only covariate. The other is to change covariate functions iteratively one by one until we obtain a satisfactory model. We will now continue to follow Lindqvist et al. and see how the second approach can be done using a misspecified model and residual plots.

Assume first that all covariate functions are known except for one. The correct model can then be stated as

$$log(T) = \mu + \boldsymbol{\beta}'\mathbf{Z} + f(X) + \sigma\epsilon, \tag{7.2}$$

where X is the single component we want to find the covariate function for, and $\mathbf{Z}$ is the remaining components. To find $f(X)$, we then start by fitting a linear model

$$log(T) = \mu + \boldsymbol{\beta}'\mathbf{Z} + \gamma X + \sigma\epsilon, \tag{7.3}$$

giving us estimated model

$$log(T) = \hat{\mu} + \hat{\boldsymbol{\beta}}'\mathbf{Z} + \hat{\gamma}X + \hat{\sigma}\epsilon, \tag{7.4}$$

where $\hat{\mu}$, $\hat{\boldsymbol{\beta}}$, $\hat{\gamma}$ and $\hat{\sigma}$ are maximum likelihood estimates, possibly found using *survreg*.

Lindqvist et al. [15] (and Aaserud [1]) then refer to White [20], when they say that there, under certain conditions, exsist theoretical parameter values μ^*, $\boldsymbol{\beta^*}$, γ^* and σ^* that minimize the Kullback-Leiber distance between model 7.2 and model 7.3. For simulated data, the actual values of μ^*, $\boldsymbol{\beta^*}$, γ^* and σ^* can be found by simulating a large number of observations from model 7.2, and finding maximum likelihood estimates.

When we know μ^*, $\boldsymbol{\beta^*}$, γ^* and σ^*, we can find theoretical standardized residuals using equation 4.8,

$$R_s^* = \frac{log(T) - \mu^* - \boldsymbol{\beta^*}'\mathbf{Z} - \gamma^* X}{\sigma^*}.$$

Inserting equation 7.3 for $log(T)$ we get

$$R_s^* = \frac{\sigma}{\sigma^*}\epsilon + \frac{(\mu - \mu^*) + (\boldsymbol{\beta} - \boldsymbol{\beta}^*)'\mathbf{Z} + f(X) - \gamma^* X}{\sigma^*}, \tag{7.5}$$

and solving for $f(X)$ gives us

$$f(X) = \sigma^* R_s^* - \sigma\epsilon - (\mu - \mu^*) - (\boldsymbol{\beta} - \boldsymbol{\beta}^*)\mathbf{Z} + \gamma^* X.$$

Further, we use that $f(x) = E(f(X)|X = x)$, giving us a new expression,

$$f(x) = \sigma^* E[R_s^*|X = x] - \sigma E[\epsilon] - (\mu - \mu^*) - (\boldsymbol{\beta} - \boldsymbol{\beta}^*)' E[\mathbf{Z}|X = x] + \gamma^* x.$$

Because we assume independent covariates, we have $E[\mathbf{Z}|X = x] = E[\mathbf{Z}]$. All parts of this equation that does not depend on X will be treated as a constant in the expression for $f(x)$, resulting in

$$f(x) = \sigma^* E[R_s^*|X = x] + \gamma^* x + constant. \tag{7.6}$$

Normally one will not know μ^*, $\boldsymbol{\beta^*}$, γ^* and σ^*, but have to use estimated values. Ignoring the constant term in equation 7.6, we will have an expression for the estimated f(x):

$$\hat{f}(x) = \hat{\gamma}x + \hat{\sigma}\hat{H}(x), \tag{7.7}$$

where $\hat{H}(x)$ is an estimate for $H(x) = E(R_s^*|X = x)$. Or alternatively, using equation 4.10, $H(x) = E[\Phi_\epsilon^{-1}(1 - exp(-R_c^*))|X = x]$

Derivation of an estimator for H(x) for uncensored and adjusted residuals can be found in Lindqvist et al. [15]. We will, in this thesis, use the R function *lowess* as an estimator for $H(x)$. *lowess* is a smoother that uses weighted least square regression to smooth the points $(x_i,\hat{r}_{s,i})$. Information on *lowess* can be found in Cleveland [6], and in the R documentation.

7.2.1 Simulated Weibull example

We will now look at a simulated data set consisting of 200 Weibull distributed observations from the following model

$$log(T) = \mu + \beta_1 Z_1 + \beta_2 Z_2 + log(X) + \sigma\epsilon, \tag{7.8}$$

where Z_1 is binary with p=0.5, $Z_2 \sim N(0,1)$, $X \sim unif[0,1]$ and ϵ is Gumbel. The true parameters are set to $\mu = 0$, $\beta_1 = 0.8$, $\beta_2 = 0.3$ and $\sigma = 0.5$. We will look at four levels of censoring, 0%, 25%, 50% and 75%. For all levels of censoring we try to fit the misspecified model

$$log(T) = \mu + \beta_1 Z_1 + \beta_2 Z_2 + \gamma X + \sigma\epsilon. \tag{7.9}$$

In addition to the data set with 200 observations we also simulate a data set with 1,000,000 observations. With this set of observations we can find the theoretical estimates for model 7.9. They can be found in table 7.1. From the table we see that the theoretical values for β_1^* and β_2^* are close to the real values for all levels of censoring. This is not the case for μ^*,γ^* and σ^*, but that is expected because we have a misspecified model with unknown $f(X)$.

Table 7.1: Theoretical parameter values minimizing the Kullback-Leiber distance between model 7.8 and 7.9 for four levels of censoring.

Censoring	0%	25%	50%	75%
μ^*	-0.838	-0.940	-1.093	-1.377
β_1^*	0.799	0.812	0.831	0.845
β_2^*	0.300	0.305	0.312	0.322
γ^*	0.645	0.704	0.808	1.073
σ^*	0.603	0.617	0.639	0.681

Because we normally don't know the theoretical values, we have to use the parameters estimated from the data set with 200 observations in our model cheeking. We will now, as we did with Nelson's superalloy data, first examine the functional form with residuals based on pseudo observations, and then with pseudo residuals based on standard residuals.

Pseudo observations

When finding pseudo observations for misspecified models, we recommend using one of the non-parametric models. The parametric model uses the misspecified model to find pseudo observations, and hence residuals based on parametric pseudo observations may not be appropriate when we check the functional form of a covariate. In this case pseudo observations based on KM gives too many negative pseudo observations for the censored data sets. We will therefore use KMlog to find pseudo observations. In figure 7.6 we see pseudo observations for all levels of censoring plotted with the observed times. For 0 % censoring, the observed times and pseudo observations will be the same.

Using the obtained pseudo observations to find estimates for the parameters gives values shown in table 7.2.

Table 7.2: Parameters in model 7.9 estimated from pseudo observations

Censoring	0%	25%	50%	75%
$\hat{\mu}$	-0.475	-0.715	-0.7763	-1.092
$\hat{\beta_1}$	0.773	0.993	0.9718	1.279
$\hat{\beta_2}$	0.264	0.300	0.3005	0.328
$\hat{\gamma}$	0.486	0.550	0.6214	0.760
$\hat{\sigma}$	0.568	0.727	1.02	1.75

In this case we see that the estimated values of the parameters change quite a lot with the degree of censoring, even for β_1 and β_2.

With the estimated parameters and pseudo observations, we then find standardized residuals. In the left figure of 7.7 we see standardized residuals plotted against covariate X for 0% censoring. Because we have a Weibull distributed data set $R_s = log(R_c)$, and hence R_s should fluctuate around zero. This is not the case in figure 7.7 so we may suspect that the functional form of X is wrong. The line in the left figure is what we find with *lowess*. Using that line as an estimator for $H(x)$ in equation 7.7 we can find estimated values for $f(x)$. They are plotted against x in the right figure of 7.7. The line in this figure is the *lowess* smoother of $(x, f(x))$.

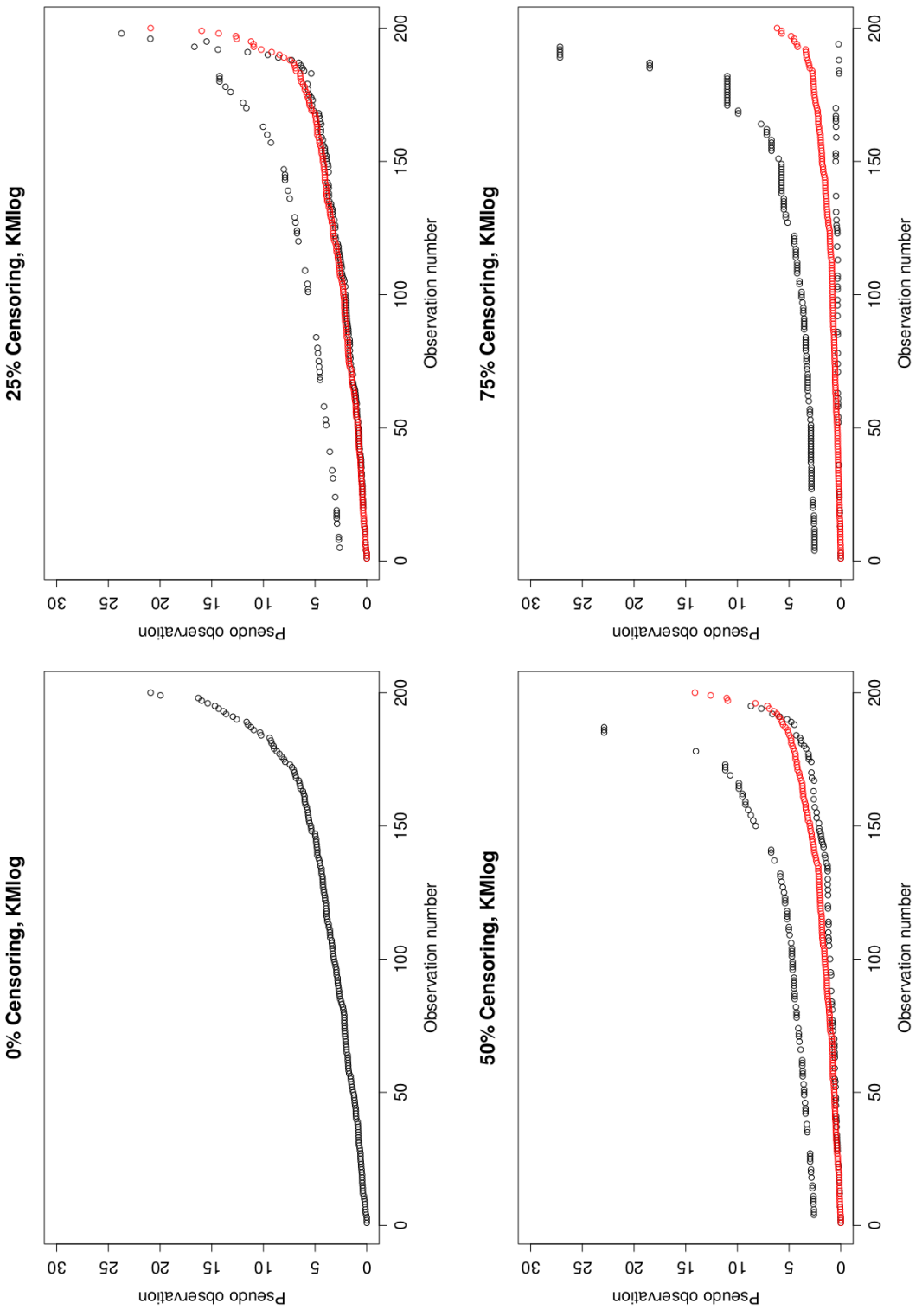

Figure 7.6: Pseudo observations (black) for four different levels of censoring plotted with the observed survival times (red). NB: Some pseudo observations are to big to be included in the plot

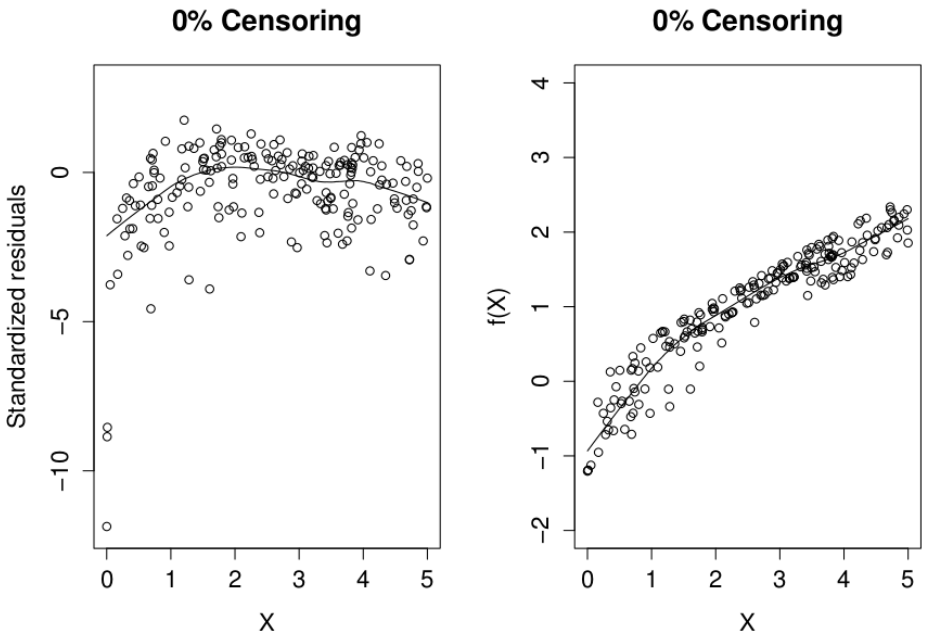

Figure 7.7: Left: Standardized residuals from pseudo observations. Right: Estimated covariate function for X

Comparing this to figure 7.8 we see that $f(x)$ may have a form like $log(x)$.

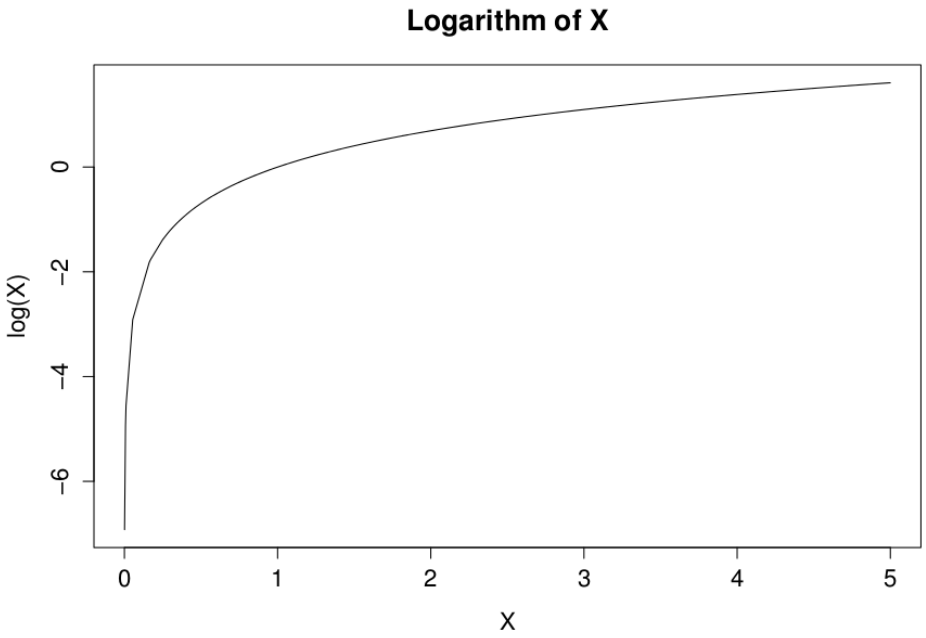

Figure 7.8: Log(x) for values of x in the simulated data set

Figures for 25%, 50% and 75% censoring, corresponding to figure 7.7, are in figure 7.9,7.10 and 7.11.

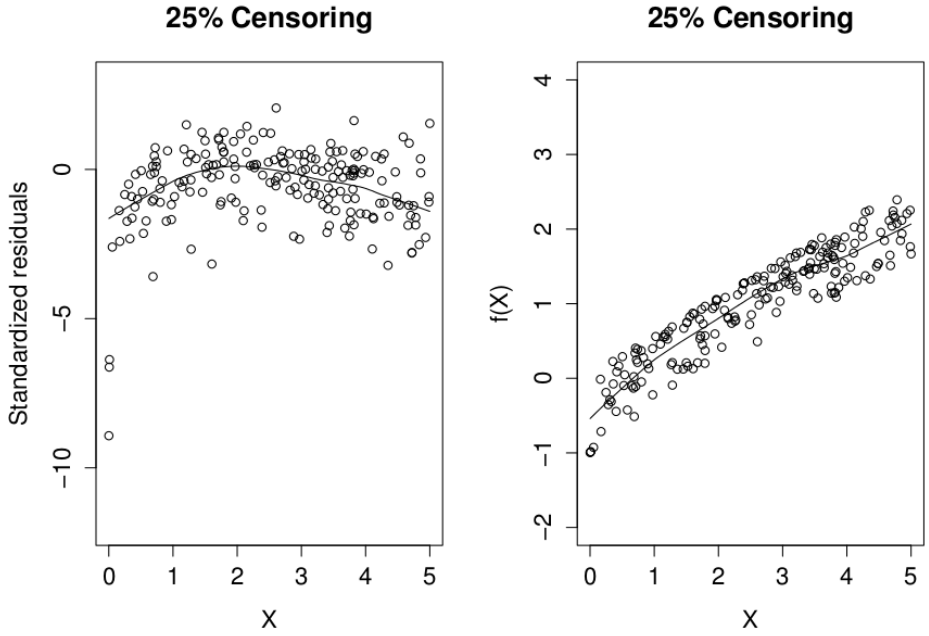

Figure 7.9: Left: Standardized residuals from pseudo observations. Right: Estimated covariate function for X

From this we see that finding the functional form with pseudo residuals are more difficult the higher the level of censoring. This is as expected, since pseudo observations and estimation of parameters from pseudo residuals are expected to be less and less accurate the higher the level of censoring.

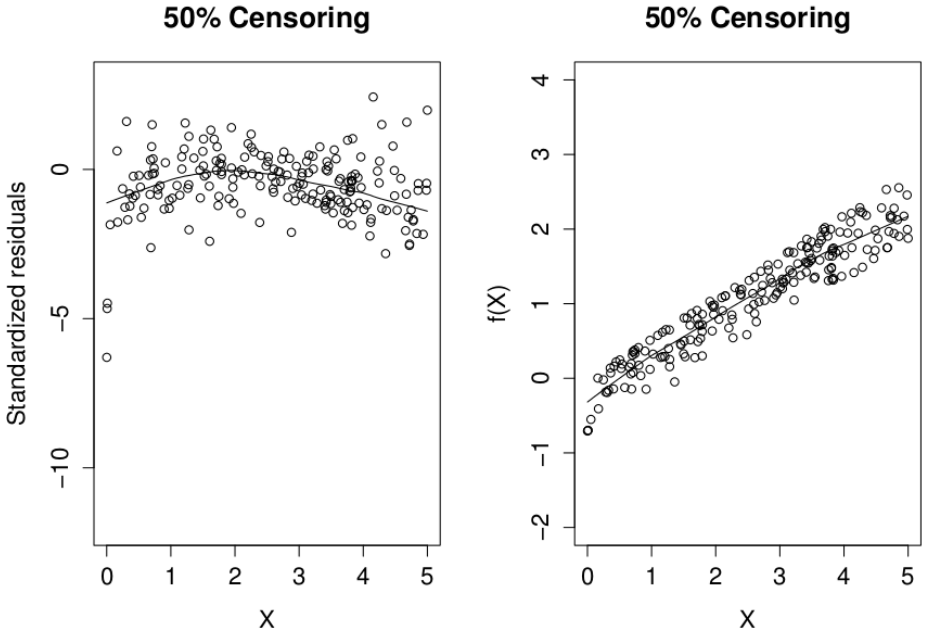

Figure 7.10: Left: Standardized residuals from pseudo observations. Right: Estimated covariate function for X

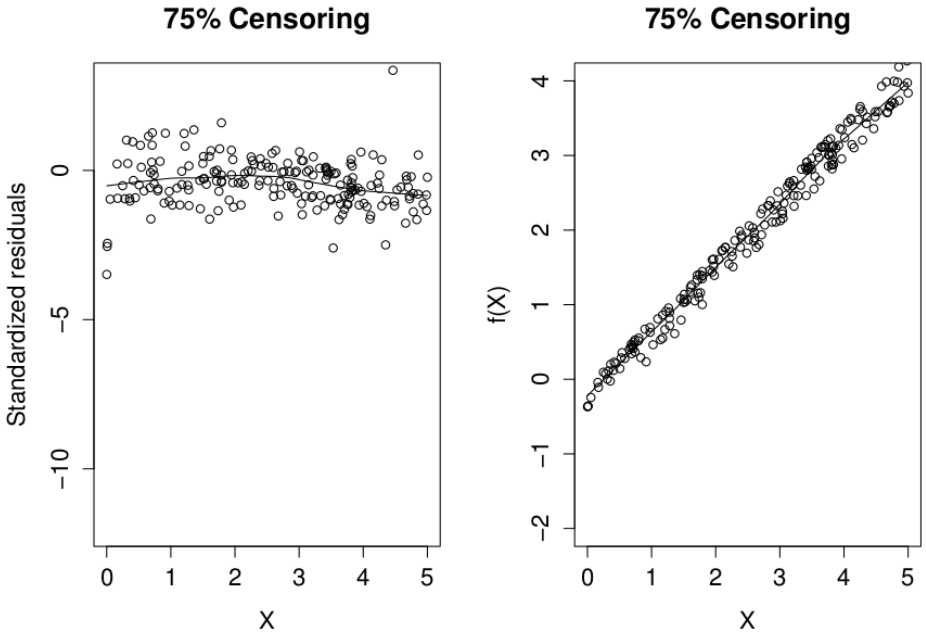

Figure 7.11: Left: Standardized residuals from pseudo observations. Right: Estimated covariate function for X

Pseudo residuals

The other option that we study is to find residuals with *survreg* from the censored data set and use any method to find pseudo residuals. Again standardized residuals are a good choice of residuals since they should fluctuate around zero when the survival times are Weibull distributed. Standardized residuals can be negative, so the KM-method will not cause any problems when we asses the model. We will therefore use the KM-method to find pseudo residuals even though other methods will be applicable. If the distribution was unknown, using Cox-Snell residuals may be a better choice, but then pseudo residuals have to be positive.

The parameters estimated for our data sets are found with *survreg* and can be found in table 7.3.

Table 7.3: Parameters in model 7.9 estimated from with *survreg*.

Censoring	0%	25%	50%	75%
$\hat{\mu}$	-0.475	-0.540	-0.604	-1.121
$\hat{\beta}_1$	0.773	0.816	0.743	0.920
$\hat{\beta}_2$	0.264	0.246	0.274	0.307
$\hat{\gamma}$	0.486	0.504	0.551	0.888
$\hat{\sigma}$	0.568	0.607	0.63	0.714

With the observed times and these parameters we find standardized residuals, and with the KM method, pseudo residuals. The left figure in 7.12 shows pseudo residuals for 0% censoring plotted against covariate X. Again we find the estimated $\hat{H}(x)$ as the line generated with *lowess*, and $f(x)$ plotted against x in the right figure.

These two figures are identical to the results we got for 0% for residuals based on pseudo observations. This comes from the fact that KMlog pseudo observations are identical to the observed times when there are 0% censoring. Standardized residuals for pseudo observations will therefore be equal to standardized residuals for the observed times. The same goes for pseudo residuals based on KM. Because KM also returns the observed times when there are no censoring, pseudo residuals found with KM will be equal to the standardized residuals. For other levels of censoring we will not get the same residuals with the two methods. Figures 7.13, 7.14 and 7.15 show the same plots for 25%, 50% and 75% censoring. From these we find the same results as we did with residuals found for pseudo observations.

Corresponding figures for standardized and 1-adjusted Cox-Snell residauls for the same data sets and level of censoring are included in appendix C.

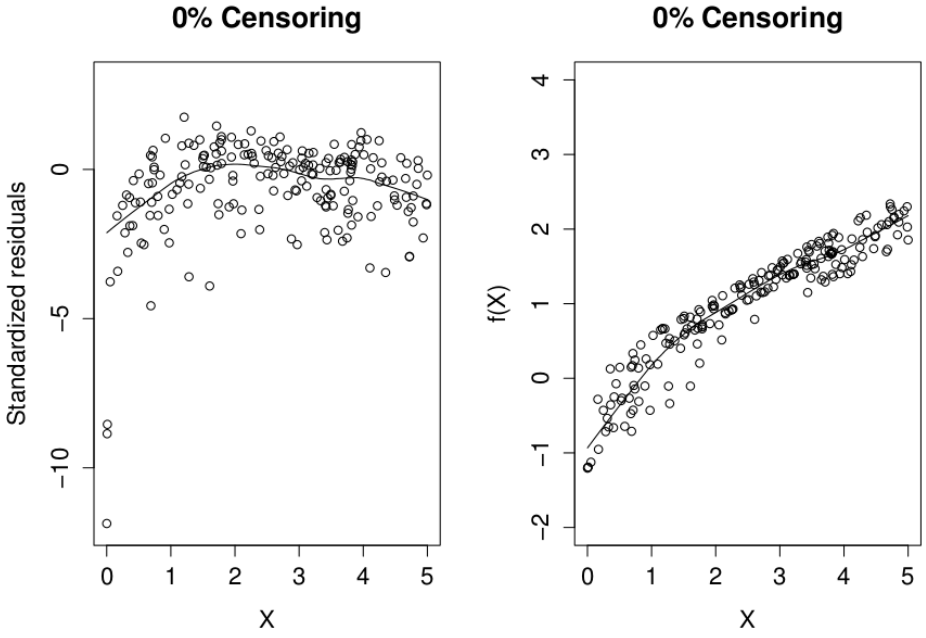

Figure 7.12: Left: Pseudo standardized residuals. Right: Estimated covariate function for X

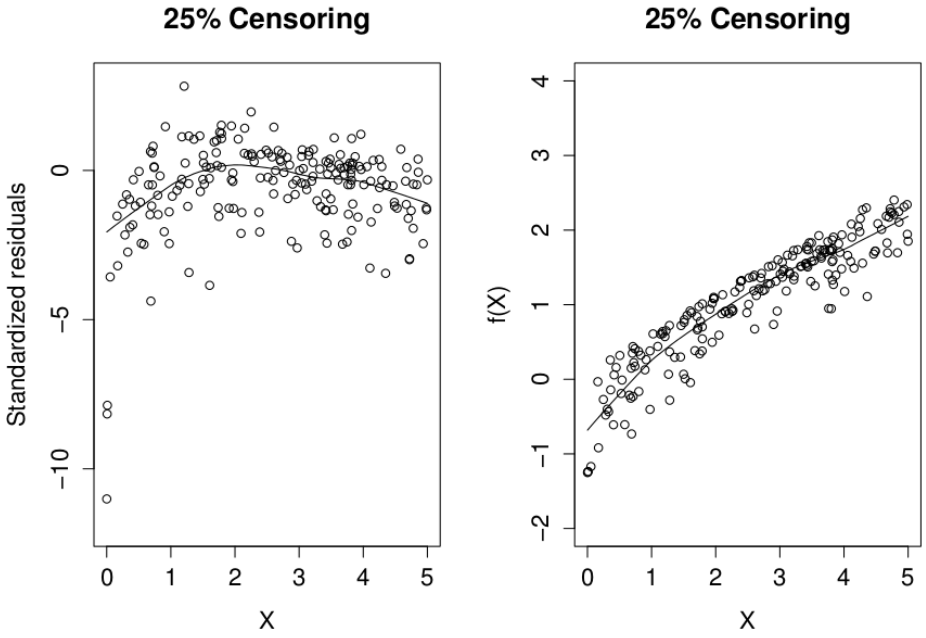

Figure 7.13: Left: Pseudo standardized residuals. Right: Estimated covariate function for X

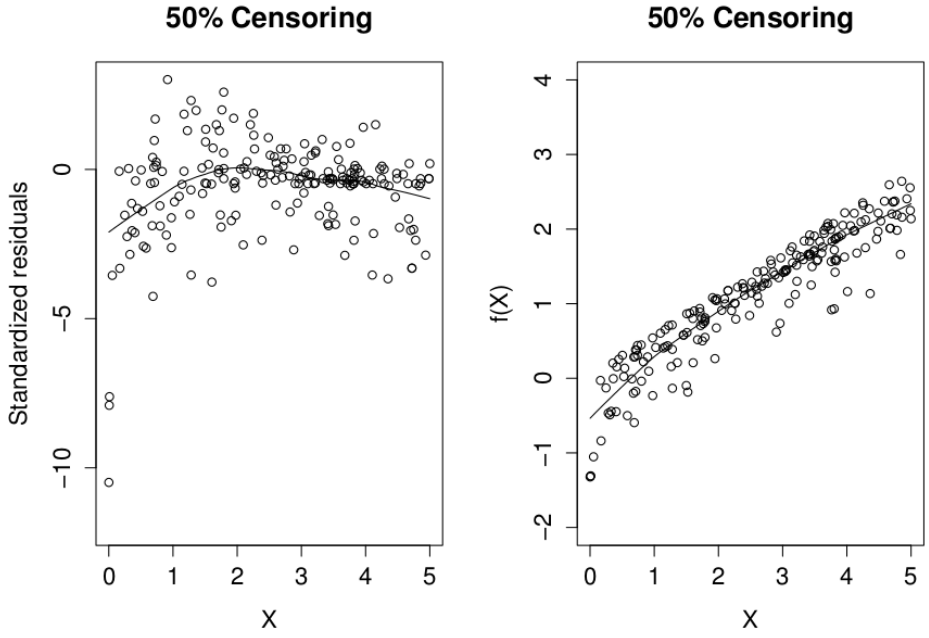

Figure 7.14: Left: Pseudo standardized residuals. Right: Estimated covariate function for X

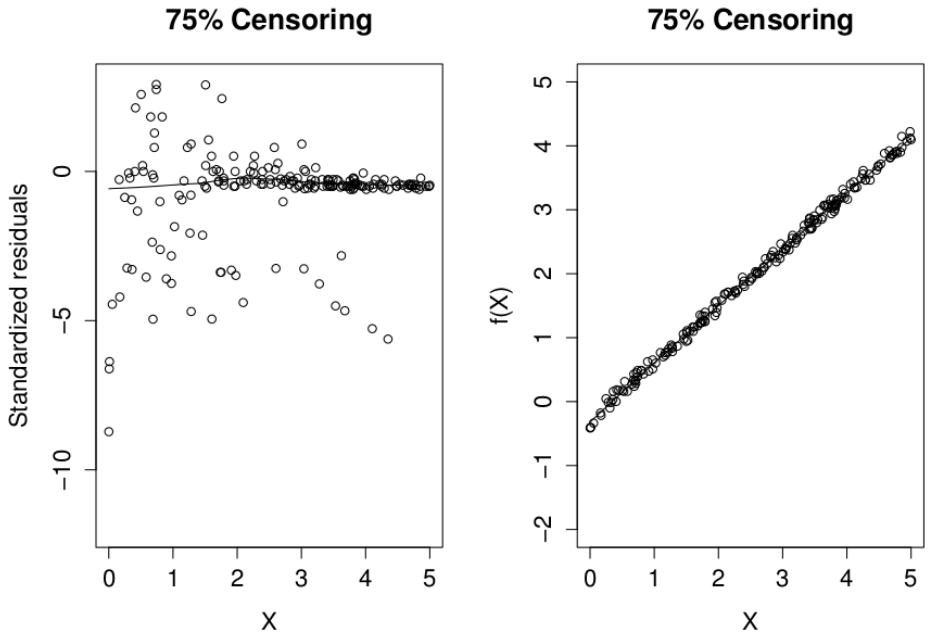

Figure 7.15: Left: Pseudo standardized residuals. Right: Estimated covariate function for X

Based on this example we see that finding the functional form for covariates with pseudo observations lead to good results for both pseudo observations and pseudo residuals. The level of censoring will affect how well we can detect that we have the wrong functional form. The more censoring, the more difficult it is to detect that the functional form is inappropriate. To say something about which method works best is difficult based on this example alone. It does however seem like the method where we find pseudo residuals gives a larger spread in the residuals which might make it easier to detect errors. Compared to the standardized residuals and the 1-adjusted Cox-Snell residuals in appendix C, we see that we get better results from the pseudo methods than the standard methods, though not so significantly compared to the 1-adjusted as for the standardized residuals.

Chapter 8

Concluding Remarks

The aim of this master's thesis has been to study pseudo observation methods for accelerated failure time models in survival analysis, and apply them in estimation and residual analysis.

Accelerated failure time models were discussed in chapter 4. On log-linear form they can be expressed as

$$Y = log(T) = \mu + \boldsymbol{\beta}'\mathbf{X} + \sigma\epsilon.$$

where T is the observed survival time, μ and σ are constants, $\boldsymbol{\beta}$ are regression parameters, $\mathbf{X}$ are the covariate vector and ϵ the error term.

Pseudo observations are known from jackknife theory as a method for resampling data sets. If $\mathbf{t} = (t_1, t_2, \cdots t_n)$ is a sample of observations of the random variable T, and $\hat{\theta} = \hat{\theta}(t)$ is an approximately unbiased estimator of $\theta = E[f(T)]$, we can a find pseudo observations of $f(T)$ by

$$\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{-i}.$$

Here $\hat{\theta}$ is found for the whole data set and $\hat{\theta}_{-i}$ for the data set without observation t_i. When this is done for i= 1,2,$\cdots$ n, we have a new data set consisting of pseudo observations.

In this report we study pseudo observations for the survival time T, and for corresponding Cox-Snell and standardized residuals. Pseudo observations for the survival time are found with three methods, KM, KMlog and the parametric method. KM and KMlog are non-parametric methods based on the Kaplan-Meier estimate of the survival function. These two methods can therefore be used when we don't know the distribution and model of our data. The parametric method is based on the specific AFT model, and will therefore not be recommended when the model is unknown. In some situations however, we can get good results even if we assume the wrong distribution, but one of the other methods should be used when in doubt.

The KM method uses $f(T) = T$. Pseudo observations obtained with this

model can be negative. For pseudo observations of the survival time this is not a good trait. Some computational methods, like *survreg*, require positive survival times. Other methods for estimating parameters should therefore be applied if one still wants to use the KM pseudo observations. Negative survival times will also be a problem when we look at standardized or Cox-Snell residuals for the pseudo observations. Both of these residuals require $log(T)$, and the logarithm of a non-positive T is undefined. A solution may be to do some type of translation before taking the logarithm or treat $log(T)$ as a missing value if T is non-positive.

KMlog and the parametric method obtain pseudo observations for $log(T)$, and these observations have to be exponentiated to get pseudo observations for T. Because of this KMlog and the parametric method always gives positive pseudo observations. Unfortunately, this also makes pseudo observations for some of the large observed survival times too high. Based on our simulations, it seems like KMlog gives higher pseudo observations at the end than the parametric method, unless the level of censoring is so high that thde KMlog pseudo observations flatten out.

The goal with pseudo observations is to create new data sets that can be treated as uncensored versions of the original data sets. When the original data set is uncensored, KM and KMlog return pseudo observations equal to the original data set, and the parametric method returns observations close to the original ones. When there is censoring, the level of censoring affects how good approximations the pseudo observations are to the real event times. The higher the level of censoring, the bigger the difference it was between the simulated real survival times and the estimated pseudo observations. It seems like censoring had most effect on KMlog, and that KM give more negative pseudo observations for higher levels of censoring. The prostatic cancer example show that for large censoring, none of the methods are very good. An advantage of the parametric approach, however, is that it seems to provide better estimated parameters than any of the other methods. Also, when there were high levels of censoring the KM and KMlog pseudo observations tend to be very even and with many equal values. That will not be a problem with the parametric method.

The effect of σ on the performance of the pseudo observation methods was studied for a Weibull distributed data set. For Weibull distributed data sets, σ is the constant in front of the error, but also the inverse of the shape parameter. Because of this, data sets with small σ values will have a smaller variance and the probability of observing data sets that are representative for the real event times are higher. KM and KMlog will therefore be more accurate for such data sets, and parameters used in the parametric model will be closer to the true parameters. Based on our simulations we found that the parametric method performed best for $\sigma \leq 0.6$, and that KM were slightly better for $\sigma > 0.6$. KMlog gave again the least good pseudo observations.

Our conclusion is that the KM method is the best method to use in situations where σ is around 0.6 and up, and the level of censoring is low. If we know the distribution of the data and σ is small the parametric method will be appropriate to use. If KM gives negative observations and we are not sure that our model is correct, KMlog is the best alternative. Some times KMlog will be preferred even though we know the distribution and model because the parametric method must be adopted to each distribution and model, while KMlog can be used without alterations.

When the model assumed for the data is appropriate, Cox-Snell residuals should behave as from an exponential distribution and standardized residuals should have approximately the same distribution as ϵ. Censored observations lead to censored residuals. Therefore one has introduced the adjusted residuals, which add a given value to residuals corresponding to censored observations. In section 4.5, the 1- and log(2)-adjusted Cox-Snell residuals was given as an example. As an alternative to these adjusted residuals, we want to use pseudo observation methods to find pseudo residuals. For Cox-Snell residuals all methods can be used, but we recommend using KMlog or the parametric method because Cox-Snell residuals should be positive. The parametric method assumes that the residuals are Weibull distributed and will therefore not be appropriate in every situation, so KMlog may be the safest choice. Standardized residuals can be both positive and negative. Therefore we recommend using pseudo residuals from equation (6,3) or KM, and among them KM seems to be the best.

The final chapter is dedicated to the use of pseudo observations and pseudo residuals in residual analysis. Here two approaches were used. The first was to find pseudo observations for the survival times, and then find residuals from them. Here, only KM and KMlog should be used. The parametric model assumes a distribution and model, and can therefore influence the outcome of the residual analysis. The other was to find residuals for the original data set and then apply pseudo observation methods to obtain pseudo residuals. Residual analysis of Nelson's superally data showed that residual plots based on the two methods are different, but the conclusion is the same. A simulated data set was then used to study the use of residuals to infer the functional form of covariates. From this we found that both of our suggested approaches preforms well, and serves as a satisfactory alternative to standard and adjusted residuals for AFT models.-.

In their paper [4], Andersen et al. state that " Regardless of the application, the pseudo observations $\hat{\theta}_i$ will always be used for all n subjects and not only for those where $f(T_i)$ is unobserved ". We have therefore replaced all observations with pseudo observations in this thesis. In figures 5.6, 5.12 and 5.17 we saw that the set of pseudo observations were close to the set of real event times. Therefore this argument makes sense, and estimation of mean and variance will be satisfactory with the new data set. Unfortunately it was not necessarily pseudo observations for the highest real event times that

became the highest pseudo observations. Estimating parameters can therefore give bad results, even though the two sets of observations are similar. An idea to further investigation of pseudo observations can therefore be to study what happens if only censored observations are replaced with pseudo observations. Perhaps that may lead to a better performance of the set of pseudo observations, although it violates the original logic of the jackknifing.

In their papers Andersen et al. uses a slightly different code than us for obtaining pseudo observations. Their method, which is the one implemented in the *Pseudo*-package in R, integrates the Kaplan-Meier estimator to the same value, t_n, for all jackknife samples. This makes the last pseudo observation potentially smaller than our method, which is to integrate to t_{n-1} for the n-th jackknife sample. This can give slightly different results, and may improve estimation and hence some of the residual plots might change. Other ways of ending the tail of the Kaplan-Meier curve may also be interesting to study. For example, when the last observation is censored, one could make the survival function be a line that goes to zero instead of being constant.

This thesis has mostly be concerned of construction of pseudo observations and pseudo residuals, and not so much on their use in practice. We have only just touched upon some applications, so there is a wealth of interesting possibilities and questions that can be investigated further.

Bibliography

[1] Aaserud, S. (2011). *Residuals and Functional Form in Accelerated Life Regression Models*, Master's thesis at NTNU

[2] Andersen, P.K., Hansen, M.G. & Klein, J.P. (2004). *Regression Analysis of Restricted Mean Survival Time Based on Pseudo-Observations*, Lifetime Data Analysis, 10, page 335–350

[3] Andersen, P.K., Klein, J.P. & Rosthøy, S. (2003). *Generalized linear models for correlated pseudo-observations, with applications to multi-state models*, Biometrika, 66, page 429-436

[4] Andersen, P.K., Hansen, M.G. & Perme, M.P. (2010). *Pseudo-observations in survival analysis*, Statistical Methods in Medical Research, 19, page 71–99

[5] Andersen, P.K., Hansen & Perme, M.P. (2008). *Checking hazard regression models using pseudo-observations*, Statistics in Medicine, 27, page 5309–5328

[6] Cleveland, W.S. (1981) *A Program for Smoothing Scatterplots by Robust Locally Weighted Regression*, The American Statistician, 35 (1), p. 54.

[7] Collett, D. (2003). *Modelling Survival Data in Medical Reaseach*, 2nd edition, Chapman and Hall/CRC

[8] Dobson, A.J. & Barnett, A.G. (2008) *A introduction to generalized linear models*, 3rd edition, Chapman and Hall/CRC

[9] Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, Chapman and Hall, London

[10] Garthwaite, P.H., Jolliffe, I.T. & Byron, J. (2002). *Statistical Inference*, 2nd edition, Oxford University Press, New York

[11] Kalbfleisch, J.D & Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd edition, John Wiley and Sons, Hoboken, New Jersey

[12] Kaplan, E.L. & Meier, P. (1958). *Nonparametric Estimation from Incomplete Observations*, Journal of the American Statistical Association, 53 (282)

[13] Klein, J.P., Gerster,M., Andersen, P.K, Tarima,S. Perme, M.P. (2008). *SAS and R functions to compute pseudo-values for censored daat regression*, Statistics in Medicine, 89, p.289-300

[14] Klein, J.P. & Moeschberger, M.L. (2003). *Survival Analysis, Techniques for Censored and Truncated Data*, 2nd edition, Springer, New York

[15] Lindqvist, B.H., Aaserud, S. & Kvaløy, J.T (2012). *Residuals and Functional Form in Accelerated Life Regression Models* preprint. Available at http://www.math.ntnu.no/preprint/statistics/2012/S13-2012.pdf.

[16] Meeker, W.Q. & Escobar, L.A. (1998). *Statisical Methods for Reliability Data*, 2nd edition, John Wiley and Sons, Hoboken, New Jersey

[17] Rausand, M. & Høyland. A. (2004). *System Reliability Theory: Models, Statisical Methods and Applications*, 2nd edition, John Wiley and Sons, Hoboken, New Jersey

[18] Therneau, T.M. (2007). *A Package for Survival Analysis in S*, Mayofoundation. Availible at http://mayoresearch.mayo.edu/mayo/research/biostat/upload/survival.pdf

[19] Dalgaard, P. (2008). *Introductory statistics with R*, Springer Verlag.

[20] White, H. (1982) *Maximum likelihood estimation of misspecied models*, Econometrica, 50 (1), page 1–25.

Appendices

Appendix A

Distributions

We will now look at the probability distributions that are used in this thesis.

Exponential distribution

The exponential distribution is one of the most used distributions in survival analysis, much because its properties makes it easy to work with. It has probability density function

$$f(t) = \lambda e^{-\lambda t} \quad for \quad t \geq 0, \lambda > 0,$$

with expectation and variance

$$E(T) = \frac{1}{\lambda},$$

$$Var(T) = \frac{1}{\lambda^2}.$$

Using equation 2.1 and 2.2 we get

$$S(t) = \int_t^\infty \lambda e^{-\lambda u} du = e^{-\lambda t},$$

$$h(t) = \frac{f(t)}{S(t)} = \lambda,$$

A consequence of the constant hazard is the "memoryless" property. It can be stated as P(T > t+Δ t | T > t)=P(T>Δ t) and says that the probability of surviving a additional period of time Δt, given survival up to time t, is independent of t.

Weibull distribution

Although the exponential function is easy to work with, it is not very realistic to have a constant hazard. An alternative is another popular distribution in survival analysis called the Weibull distribution. It is related to the exponential distribution but allows more flexibility because the hazard function does

not have to be constant. Probability density function, survival and hazard function for Weibull distributed survival times are

$$f(t) = \lambda\gamma t^{\gamma-1} exp(-\lambda t^{\gamma}) \quad for \quad t \geq 0, \lambda > 0, \gamma > 0, \tag{A.1}$$

$$S(t) = e^{-\lambda t^{\gamma}},$$

$$h(t) = \lambda\gamma t^{\gamma-1}, \tag{A.2}$$

where γ is called the shape parameter and λ the scale parameter. The hazard will still be either increasing ($\gamma > 1$) or decreasing ($\gamma < 1$) and if $\gamma = 1$ we will have an exponential distribution.

The expected value and variance for Weibull distributed lifetimes are

$$E(T) = \frac{1}{\lambda}\Gamma(\frac{1}{\gamma} + 1),$$

$$Var(T) = \frac{1}{\lambda^2}\Big(\Gamma(\frac{2}{\gamma} + 1) - \Gamma^2(\frac{1}{\gamma} + 1)\Big),$$

where Γ is the gamma function.

Gumbel distribution of smallest extreme

The cumulative distribution function for the Gumbel distribution of smallest extreme is

$$F(t) = 1 - exp\Big[-exp\Big(\frac{t-\nu}{\alpha}\Big)\Big], \quad for \quad -\infty < t < \infty$$

By setting $Y = \frac{T-\nu}{\alpha}$, we get standardized Gumbel with CDF,

$$F(y) = 1 - exp[-exp(y)], \quad for \quad -\infty < y < \infty$$

This gives us density, mean, variance and survival function

$$f(y) = exp(y)exp[-exp(y)],$$

$$E[Y] = -\phi,$$

$$Var[T] = \frac{\pi^2}{6},$$

and

$$S(y) = exp[-exp(y)]. \tag{A.3}$$

where $\phi = 0.5772$ is Euler constant.

More on extreme value distributions can be found on page 54 in Rausand & Høyland [17].

Normal distribution

The Gaussian normal distribution is the most used distribution in statistics. With mean μ and variance σ^2 we can write the probability density function as

$$f(t) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[\frac{-(t-\mu)^2}{2\sigma^2}\right], \quad for \quad -\infty < y < \infty$$

When $\mu = 0$ and $\sigma = 1$ we have what is called the standard normal distribution, with cumulative distribution function denoted by

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} exp\left[\frac{-x^2}{2}\right] dx.$$

Using the CDF of the standard normal distribution we can find a general expression for the distribution function as

$$F(t) = \Phi\left(\frac{t-\mu}{\sigma}\right).$$

Lognormal distribution

The lognormal distribution is closely related to the normal distribution. If a random variable T is lognormally distributed with parameters μ and σ then $Y = log(T)$ is normal distributed with mean μ and variance σ^2. The density and survival function for lognormal lifetimes are

$$f(t) = \frac{1}{\sqrt{2\pi}\sigma t} exp\left[\frac{-(log(t)-\mu)^2}{2\sigma^2}\right],$$

$$S(t) = \Phi\left(\frac{\mu - log(t)}{\sigma}\right). \tag{A.4}$$

Uniform distribution

For a variable T that is uniformly distributed, all observations of equal length have the same probability. That is, if T can take any value $t \, \epsilon$ [a,b], then the probability density function for T will be

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{if } t\epsilon[a,b] \\ 0, & \text{otherwise} \end{cases}.$$

The cumulative distribution function will therefore

$$F(x) = \begin{cases} 0, & t < a \\ \frac{x-a}{b-a}, & \text{if } a \leq t < b \\ 1, & t \geq b \end{cases}.$$

Appendix B

Data sets

W1

Data set W1, used for detailed studies of KM and parametric pseudo observations. Survival times and covariates are simulated from model

$$log(T) = \mu + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \sigma\epsilon$$

where $\mu = 0$, β_1=0.6, β_2=0.5, β_3=0.2 and $\sigma = 2/3$. ϵ is gumbel distributed, X_1 binary with $p_0 = p_1 = 0.5$, $X_2 \sim unif[-1, 1]$ and $X_3 \sim N(0, 1)$. Status=1 indicates uncensored observation. Table for this data set is in table B.1.

Prostatic cancer data

Data from a study of prostetic cancer patients. The measured survival time and covariates are in table B.2 and the results in table B.3. SH is serumhaemaglobin, T.S is Tumor size and G.I is gleason index.

Nelson's superalloy Data

Low-cycle fatigue life of nickel-base superalloy specimens. Survival times are measured in thousand cycles. Censored observations are indicated with status=0, and uncensored with status=1. This data set is in table B.4.

Table B.1: The W1 data set.

Obs:	x_1	x_2	x_3	ϵ	T	Status	T_{Real}	C	T_{Param}	T_{KM}	T_{KMlog}
1	0.00	-0.72	0.65	-4.29	0.05	1.00	0.05	2.88	0.05	0.05	0.05
2	0.00	-0.58	-0.29	-3.45	0.07	1.00	0.07	4.91	0.07	0.07	0.07
3	1.00	0.46	-1.81	-2.20	0.37	1.00	0.37	13.51	0.58	0.37	0.37
4	1.00	0.11	0.60	-2.51	0.41	1.00	0.41	2.61	0.37	0.41	0.41
5	0.00	0.07	1.06	0.31	0.58	0.00	1.57	0.58	1.51	2.25	1.73
6	0.00	-0.04	-1.37	-0.34	0.60	1.00	0.60	10.69	0.58	0.48	0.55
7	0.00	-0.24	-0.59	-0.36	0.62	1.00	0.62	4.11	0.55	0.51	0.58
8	0.00	-0.72	-2.03	0.51	0.65	1.00	0.65	6.24	0.70	0.54	0.61
9	0.00	-0.27	-0.23	-0.04	0.81	1.00	0.81	2.36	0.72	0.71	0.77
10	1.00	-0.71	0.94	-0.49	0.87	0.00	1.12	0.87	3.16	2.87	2.52
11	0.00	-0.48	0.06	0.93	0.92	0.00	1.49	0.92	2.11	2.87	2.52
12	0.00	-0.01	-0.69	0.17	0.97	1.00	0.97	4.76	1.00	0.39	0.73
13	0.00	-0.42	1.42	1.16	1.10	0.00	2.34	1.10	2.99	3.18	2.94
14	1.00	0.91	-0.69	0.13	1.28	0.00	2.74	1.28	0.85	3.18	2.94
15	1.00	-0.39	-0.06	-0.17	1.32	1.00	1.32	10.95	0.91	-0.09	0.74
16	1.00	-0.74	1.23	2.00	1.38	0.00	6.11	1.38	4.24	3.84	3.87
17	1.00	0.27	0.48	-0.17	2.05	1.00	2.05	10.23	1.78	0.43	1.28
18	1.00	0.85	1.68	-0.73	2.39	1.00	2.39	6.10	1.85	1.29	1.88
19	1.00	0.29	1.04	0.79	4.39	1.00	4.39	4.48	5.51	6.28	8.58
20	1.00	-0.11	0.11	1.50	4.82	1.00	4.82	23.82	9.23	7.35	10.83

Table B.2: The prostetic cancer data set.

	Treatment	Time	Status	Age	SH	T.S	G.I
1	0.00	65.00	0.00	67.00	13.40	34.00	8.00
2	1.00	61.00	0.00	60.00	14.60	4.00	10.00
3	1.00	60.00	0.00	77.00	15.60	3.00	8.00
4	0.00	58.00	0.00	64.00	16.20	6.00	9.00
5	1.00	51.00	0.00	65.00	14.10	21.00	9.00
6	0.00	51.00	0.00	61.00	13.50	8.00	8.00
7	0.00	14.00	1.00	73.00	12.40	18.00	11.00
8	0.00	43.00	0.00	60.00	13.60	7.00	9.00
9	1.00	16.00	0.00	73.00	13.80	8.00	9.00
10	0.00	52.00	0.00	73.00	11.70	5.00	9.00
11	0.00	59.00	0.00	77.00	12.00	7.00	10.00
12	1.00	55.00	0.00	74.00	14.30	7.00	10.00
13	1.00	68.00	0.00	71.00	14.50	19.00	9.00
14	1.00	51.00	0.00	65.00	14.40	10.00	9.00
15	0.00	2.00	0.00	76.00	10.70	8.00	9.00
16	0.00	67.00	0.00	70.00	14.70	7.00	9.00
17	1.00	66.00	0.00	70.00	16.00	8.00	9.00
18	1.00	66.00	0.00	70.00	14.50	15.00	11.00
19	1.00	28.00	0.00	75.00	13.70	19.00	10.00
20	1.00	50.00	1.00	68.00	12.00	20.00	11.00
21	0.00	69.00	1.00	60.00	16.10	26.00	9.00
22	0.00	67.00	0.00	71.00	15.60	8.00	8.00
23	1.00	65.00	0.00	51.00	11.80	2.00	6.00
24	0.00	24.00	0.00	71.00	13.70	10.00	9.00
25	1.00	45.00	0.00	72.00	11.00	4.00	8.00
26	1.00	64.00	0.00	74.00	14.20	4.00	6.00
27	0.00	61.00	0.00	75.00	13.70	10.00	12.00
28	0.00	26.00	1.00	72.00	15.30	37.00	11.00
29	0.00	42.00	1.00	57.00	13.90	24.00	12.00
30	1.00	57.00	0.00	72.00	14.60	8.00	10.00
31	1.00	70.00	0.00	72.00	13.80	3.00	9.00
32	1.00	5.00	0.00	74.00	15.10	3.00	9.00
33	1.00	54.00	0.00	51.00	15.80	7.00	8.00
34	0.00	36.00	1.00	72.00	16.40	4.00	9.00
35	1.00	70.00	0.00	71.00	13.60	2.00	10.00
36	1.00	67.00	0.00	73.00	13.80	7.00	8.00
37	0.00	23.00	0.00	68.00	12.50	2.00	8.00
38	0.00	62.00	0.00	63.00	13.20	3.00	8.00

Table B.3: Data and pseudo observations for the prostetic cancer data set.

Obs	T	T_{KM}	T_{KMlog}	$T_{param,wei}$	$T_{param,lognorm}$
1	65.00	70.57	70.96	2827.01	1642.96
2	61.00	70.57	70.96	233.63	305.25
3	60.00	70.57	70.96	319.52	413.04
4	58.00	70.57	70.96	187.10	314.19
5	51.00	70.57	70.96	170.83	233.25
6	51.00	70.57	70.96	181.29	217.42
7	14.00	11.13	12.86	19623.18	4915790.09
8	43.00	69.59	69.80	138.68	156.43
9	16.00	65.72	64.47	187.33	252.96
10	52.00	70.57	70.96	166.71	225.87
11	59.00	70.57	70.96	208.44	500.77
12	55.00	70.57	70.96	201.11	260.51
13	68.00	70.57	70.96	341.20	398.05
14	51.00	70.57	70.96	203.42	266.36
15	2.00	64.20	61.64	121.43	116.52
16	67.00	70.57	70.96	256.36	598.05
17	66.00	70.57	70.96	261.78	337.75
18	66.00	70.57	70.96	1088.10	907.80
19	28.00	67.20	66.69	97.24	138.22
20	50.00	45.10	46.18	0.00	0.00
21	69.00	60.21	61.13	0.14	0.40
22	67.00	70.57	70.96	237.92	399.06
23	65.00	70.57	70.96	538.11	794.36
24	24.00	65.72	64.47	110.99	107.58
25	45.00	69.59	69.80	293.55	392.51
26	64.00	70.57	70.96	501.40	754.37
27	61.00	70.57	70.96	8629.55	2934.37
28	26.00	19.76	22.56	15.11	0.73
29	42.00	37.32	38.63	0.40	0.50
30	57.00	70.57	70.96	208.77	275.82
31	70.00	75.76	76.45	293.03	367.91
32	5.00	64.20	61.64	225.58	288.07
33	54.00	70.57	70.96	275.20	370.82
34	36.00	30.12	32.11	0.20	0.00
35	70.00	75.76	76.45	285.55	395.81
36	67.00	70.57	70.96	300.36	393.34
37	23.00	65.72	64.47	198.33	188.95
38	62.00	70.57	70.96	231.79	301.55

Table B.4: Nelson's superalloy data set.

	Pseudostress	k-Cycles	Status
1	80.3	211.629	1
2	80.6	200.027	1
3	80.8	57.923	0
4	84.3	155.000	1
5	85.2	13.949	1
6	85.6	112.968	0
7	85.8	152.680	1
8	86.4	156.725	1
9	86.7	138.114	0
10	87.2	56.723	1
11	87.3	121.075	1
12	89.7	122.372	0
13	91.3	112.002	1
14	99.8	43.331	1
15	100.1	12.076	1
16	100.5	13.181	1
17	113.0	18.067	1
18	114.8	21.300	1
19	116.4	15.616	1
20	118.0	13.030	1
21	118.4	8.489	1
22	118.6	12.434	1
23	120.4	9.750	1
24	142.5	11.865	1
25	144.5	6.705	1
26	145.9	5.733	1

Appendix C

Figures

Figures used to assess the functional form of covariate X in section 7.2.1. Figures are made for standardized and 1-adjusted Cox-Snell residuals, for 0%, 25%, 50% and 75% censoring.

Standardized residuals

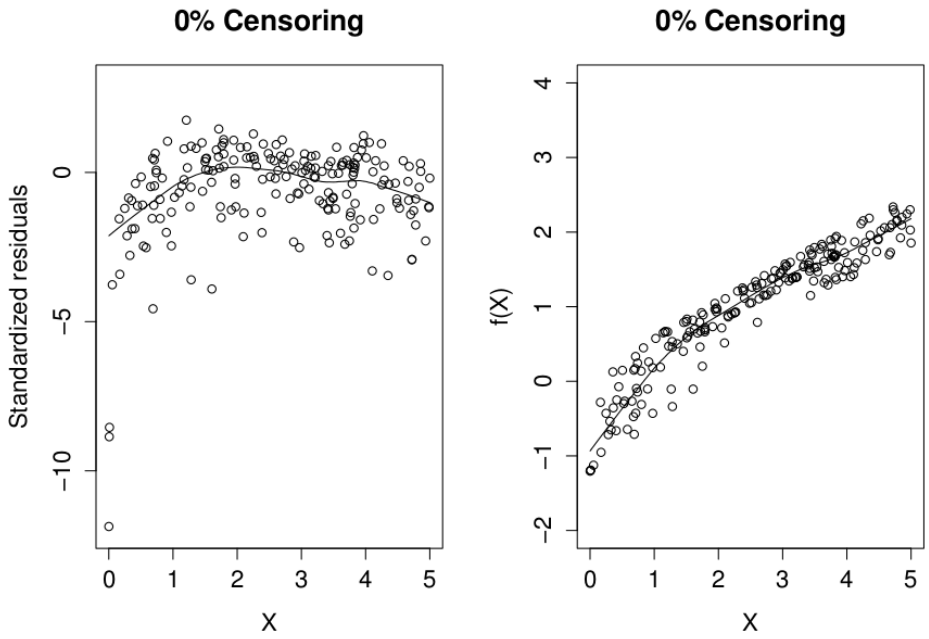

Figure C.1: Left: Standardized residuals plotted against X. Right: Estimated covariate function for X, 0% censoring

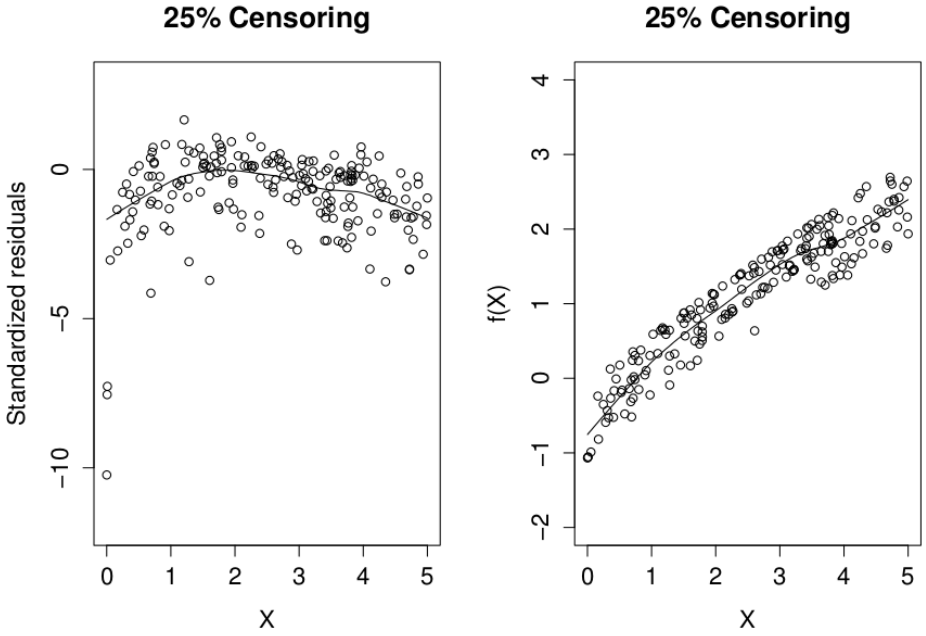

Figure C.2: Left: Standardized residuals plotted against X. Right: Estimated covariate function for X, 25% censoring

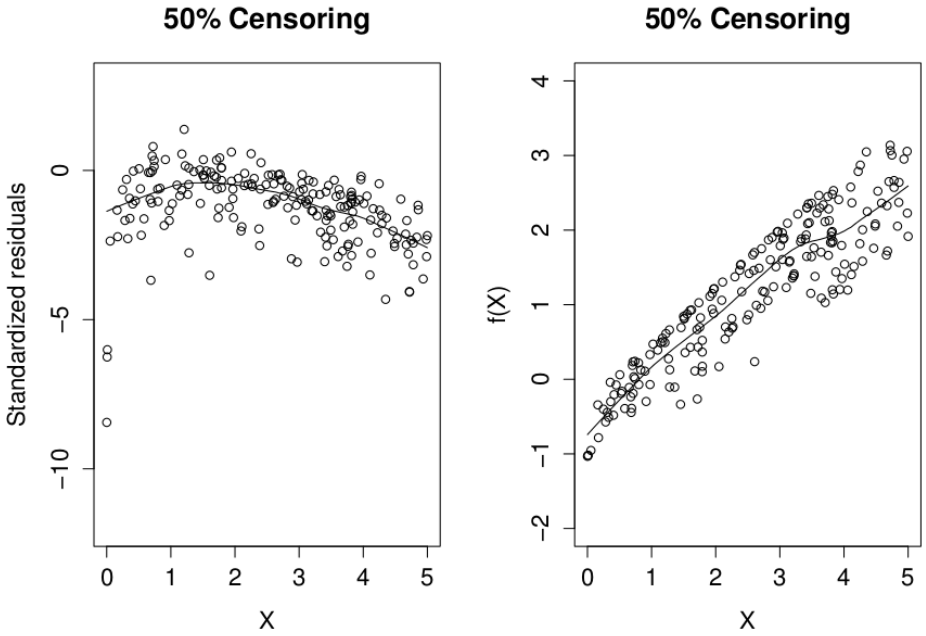

Figure C.3: Left:Standardized residuals plotted against X. Right: Estimated covariate function for X, 50% censoring

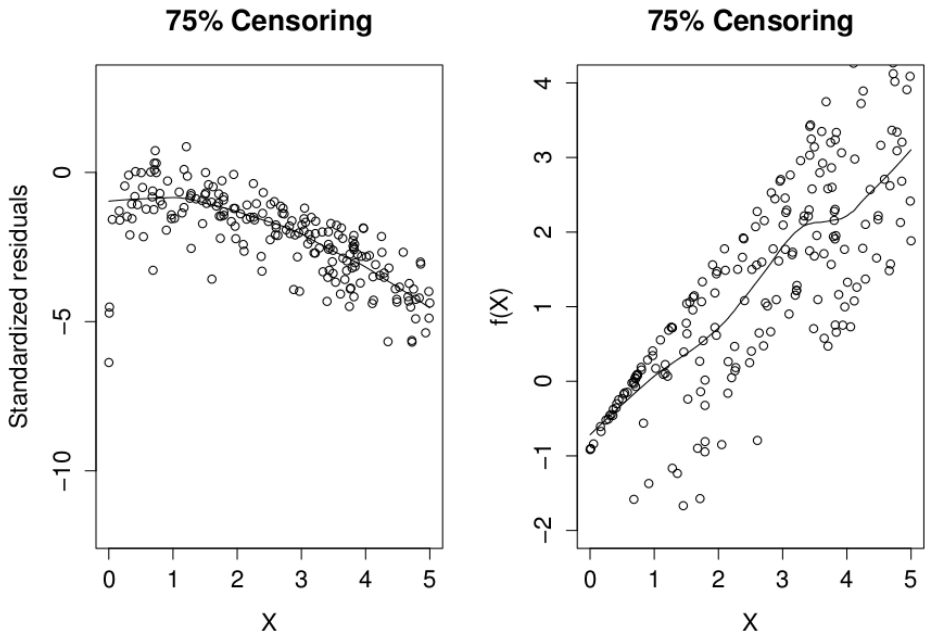

Figure C.4: Left: Standardized residuals plotted against X. Right: Estimated covariate function for X, 75% censoring

1-Adjusted Cox-Snell residuals

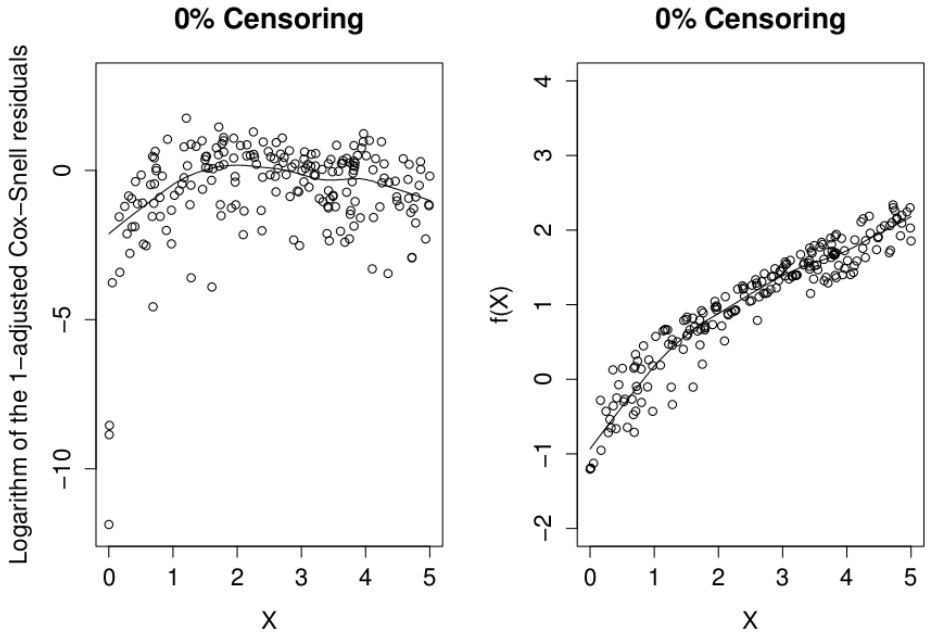

Figure C.5: Left: 1-adjusted Cox-Snell residuals. Right: Estimated covariate function for X, 0% censoring

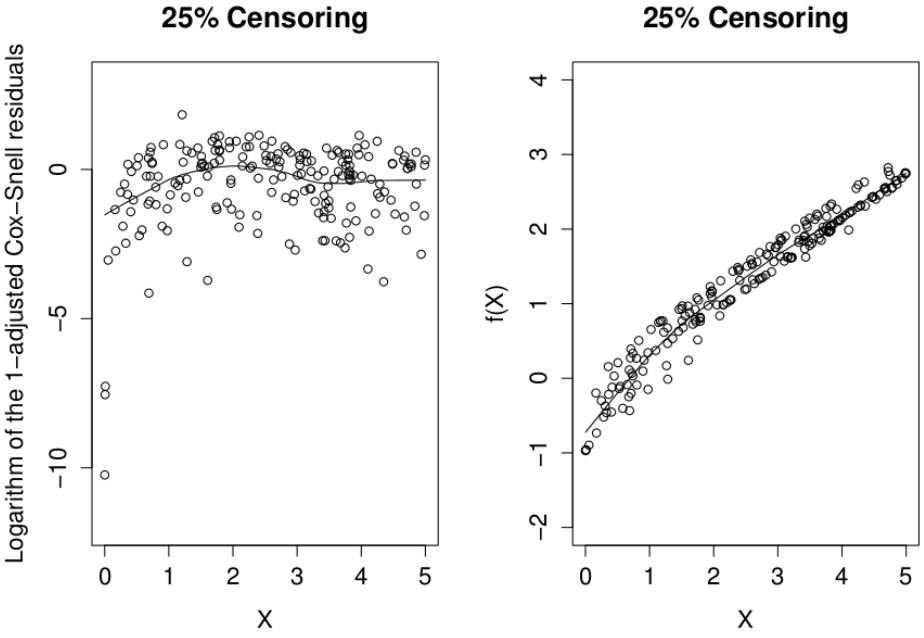

Figure C.6: Left: 1-adjusted Cox-Snell residuals. Right: Estimated covariate function for X, 25% censoring

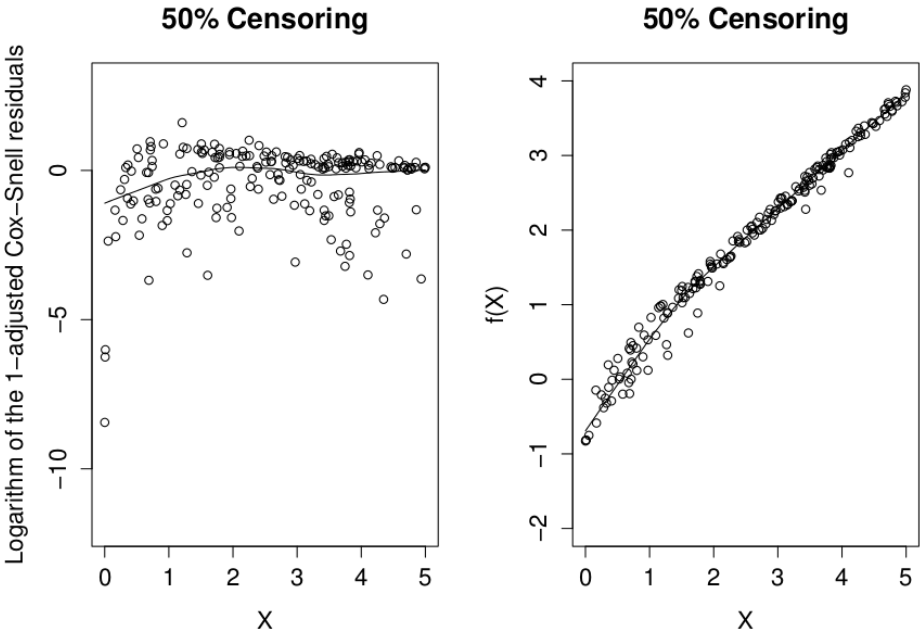

Figure C.7: Left: 1-adjusted Cox-Snell residuals. Right: Estimated covariate function for X, 50% censoring

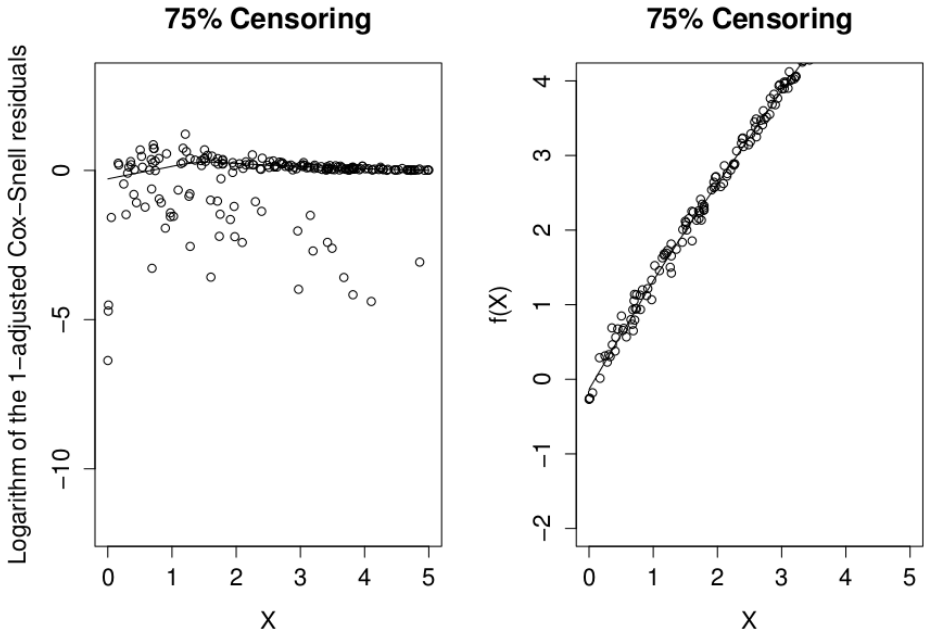

Figure C.8: Left: 1-adjusted Cox-Snell residuals. Right: Estimated covariate function for X, 75% censoring

Appendix D

R codes

.

Simulating data sets

This is an example of the code used to simulate data set. This particular code is for the AFT model discussed in 4.8.

```
# Setting the number of observations:
n=20

# Function for finding n Gumbel distributed values:
Gumbel <- function(n)
{
  u<-runif (n,0,1)
  w<-log(-log(u))
}

# Simulating covariates:
w<-Gumbel(n)   # Error term if the data is Weibull
# w<-rnorm(n)   # Error term if the data is log-normal
x1 <-round(runif(n,0,1)) # Either 0 or 1, with p=0.5
x2 <-runif(n,-1,1) # Uniformlly distributed between -1 and 1.
x3 <-rnorm(n) # Standard normal distributed

# True coefficients:
b0=0
b1=0.8
b2=0.6
b3=0.2
sigma=(2/3)
```

```
# Finding real survival times and censoring times:
realtimes<-exp(b0+b1*x1+b2*x2+b3*x3+sigma*w)
v<-runif (n,0,1)
censtimes<-(-1/0.7155)*log(v)

# Finding the observed survival time:
Time <-pmin(realtimes, censtimes)
max(Time)

# Finding the status
Status <-as.numeric(censtimes > realtimes)

# Creating the data set and order in increasing survival time:
Data<-data.frame(x1=x1,x2=x2,x3=x3,w=w,Time=Time,Status=Status,
realtimes=realtimes,censtimes=censtimes)
Data<-Data[order(Data$Time),]

# To find the percentage of censored observations:
sum<-sum(Data$Status)
censorening<-1-sum/n

# Adding a column in the data frame for observation number:
Data["Obs"]<-NA
Data$Obs<-1:n
```

Pseudo observations

KM

Code for finding pseudo observations using the method based on Kaplan-Meier.

```
# Fitting a Kaplan-Meier survival function for the observed times:
km<-survfit(Surv(Data$Time,Data$Status==1)~1)

# Finding the restricted mean of the KM-survival times:
rmean_full<-summary(km, rmean=TRUE)$table[5]

# Finding pseudo observations:
rmean<-NULL
pseudo<-NULL
n<-length(Data$Time)
for(i in 1:n){
  minus_i<-Data[-i,]
  km_minus<-survfit(Surv(minus_i$Time,minus_i$Status==1)~1)
  summary(km_minus,rmean=TRUE)
  rmean[i]<-summary(km_minus, rmean=TRUE)$table[5]
```

```
  pseudo[i]<-n*rmean_full-(n-1)*rmean[i]
}

#Adding pseudo observations to the data set:
Data["KMPseudo"]<-NA
Data$KMPseudo<-pseudo
```

KMlog

Code for finding pseudo observations based on Kaplan-Meier for log(T).

```
# Fitting a Kaplan-Meier survival function to the observed times:
km<-survfit(Surv(Data$Time,Data$Status==1)~1)

# Extracting information needed later:
kmsurv<-km$surv
kmtime<-km$time
kmlogtime<-log(kmtime)

# Function for finding restricted mean for log(T):
Rmean.log<-function(kmlogtime,kmsurv){
  k<-length(kmsurv)
  R.mean<-kmlogtime[1]
  for (i in 2:k){
    l<-(kmlogtime[i]-kmlogtime[i-1])
    R.mean<-R.mean+l*kmsurv[i-1]
  }
  return(R.mean)
}

# Using the function to find restricted mean for full model
rmean.full<-Rmean.log(kmlogtime,kmsurv)

# Finding pseudo observations:
rmean<-NULL
pseudo<-NULL
for(i in 1:n){
  minus_i<-Data[-i,]
  km.pseudo<-survfit(Surv(minus_i$Time,minus_i$Status==1)~1)
  pseudo.i.time<-km.pseudo$time
  pseudo.i.surv<-km.pseudo$surv
  pseudo.i.logtime<-log(pseudo.i.time)
  rmean[i]<-Rmean.log(pseudo.i.logtime,pseudo.i.surv)
  pseudo[i]<-n*rmean.full-(n-1)*rmean[i]
}
Pseudo<-exp(pseudo)
```

```
# Adding pseudo observations to the data set:
Data["KMlogPseudo"]<-NA
Data$KMlogPseudo<-Pseudo
```

Parameteric

The code for parameteric pseudo observations must be adjusted to the distribution and model of the data. This example is for Weibull distributed data with model

$$log(T) = \mu + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \sigma\epsilon \tag{D.1}$$

```
# Expected value for a gumbel distributed random variable
Ew =-0.5722

# Fitting a model for the full data set:
model<-survreg(Surv(Data$Time,Data$Status)~Data$x1+Data$x2+Data$x3,
data=Data, dist="weibull")
summary<-summary(model)
summary

# Extracting coefficients:
coef<-as.numeric(summary$coefficients)
b0_hat<-coef[1]
b1_hat<-coef[2]
b2_hat<-coef[3]
b3_hat<-coef[4]
sigma_hat<-summary$scale

# Finding an estimate for the expected value for the full model:
x1_bar<-sum(Data$x1)/n
x2_bar<-sum(Data$x2)/n
x3_bar=sum(Data$x3)/n
Theta_hat_full = b0_hat+b1_hat*x1_bar+b2_hat*x2_bar
+b3_hat*x3_bar+sigma_hat*Ew

# Finding the pseudo observations :
Theta_hat<-NULL
Theta_hat_min<-NULL
for(i in 1:n){
  minus_i <-Data[-i,]
  model_min<-survreg(Surv(minus_i$Time,minus_i$Status)
   ~minus_i$x1+minus_i$x2+minus_i$x3, data=minus_i, dist="weibull")
  summary_min<-summary(model__min)
  coef_min<-as.numeric(summary_min$coefficients)
  b0_hat_min<-coef_min[1]
  b1_hat_min<-coef_min[2]
  b2_hat_min<-coef_min[3]
```

```
  b3_hat_min<-coef_min[4]
  sigma_hat_min<-summary_min$scale
  x1_bar_min<-(sum(Data$x1)-Data$x1[i])/(n-1)
  x2_bar_min<-(sum(Data$x2)-Data$x2[i])/(n-1)
  x3_bar_min<-(sum(Data$x3)-Data$x3[i])/(n-1)
  Theta_hat_min[i]<-b0_hat_min+b1_hat_min*x1_bar_min
   +b2_hat_min*x2_bar_min+b3_hat_min*x3_bar_min+sigma_min*Ew
  Theta_hat[i]<-n*Theta_hat_full-(n-1)*Theta_hat_min[i]
}
pseudo<-exp(Theta_hat)

# Adding pseudo observations to the data set
Data["ParamPseudo"]<-NA
Data$ParamPseudo<-pseudo
```

For log-normal data sets, change to Ew=0 and dist='lognormal'.

Residuals

```
# Extracting coefficients
model<-survreg(Surv(Data$Time,Data$Status)~Data$x1+Data$x2+Data$x3
   ,data=Data,dist="weibull")
summary<-summary(model)
coef<-as.numeric(summary$coefficients)
b0_hat<-coef[1]
b1_hat<-coef[2]
b2_hat<-coef[3]
b3_hat<-coef[4]
sigma_hat<-summary$scale

#Standardized residuals
rs<-(log(Data$Time)-b0_hat-b1_hat*Data$x1-b2_hat*Data$x2
    -b3_hat*Data$x3)/sigma_hat

#Cox-Snell residuals
rc<-exp((log(Data$Time)-b0_hat-b1_hat*Data$x1-b2_hat*Data$x2
    -b3_hat*Data$x3)/sigma_hat)
```

Pseudo residuals

Code for pseudo residuals based on KM and KMlog is equal to the code for pseudo observations, but Time must be exchanged with the residual in question. Here r can be either standardized or Cox-Snell residuals.

Introductory pseudo residuals

```
# Calculating expected residual value for full model
```

```
Clevel<-sum(Data$Status) # The number of uncensored observations
Usum<-sum(Data$xr) # The sum of all residuals
Theta_full<-Usum/Clevel

pseudores<-NA
# Pseudo residuals:
for(i in 1:n){
  res_minus_i<-Data[-i,]
  Clevel_minus_i<-sum(res_minus_i$Status)
  Usum_minus_i<-sum(res_minus_i$Coxres)
  Theta<-(Usum_minus_i/Clevel_minus_i)
  pseudores[i]<-n*Theta_full-(n-1)*Theta
}

Data["Pseudores"]<-NA
Data$Pseudores<-pseudores
```

KM pseudo residuals

```
km<-survfit(Surv(Data$r,Data$Status==1)~1)
km.summary<-summary(km)
rmean_full<-summary(km, rmean=TRUE)$table[5]
rmean<-NULL
pseudo<-NULL
n<-length(Data$Time)
for(i in 1:n){
  res_minus_i<-Data[-i,]
  km_minus<-survfit(Surv(res_minus_i$r,res_minus_i$Status==1)~1)
  #plot(km_minus,main= 'Plot D')
  summary(km_minus,rmean=TRUE)
  rmean[i]<-summary(km_minus, rmean=TRUE)$table[5]
  pseudo[i]<-n*rmean_full-(n-1)*rmean[i]
}

Data["KMPseudores"]<-NA
Data$KMPseudores<-pseudo
```

KMlog pseudo residuals

```
km<-survfit(Surv(Data$r,Data$Status==1)~1)
kmsurv<-km$surv
kmtime<-km$time
kmlogtime<-log(kmtime)

## Restricted mean for log(T)
Rmean.log<-function(kmlogtime,kmsurv){
  k<-length(kmsurv)
```

```
  R.mean<-kmlogtime[1]
  for (i in 2:k){
    l<-(kmlogtime[i]-kmlogtime[i-1])
    R.mean<-R.mean+l*kmsurv[i-1]
  }
  return(R.mean)
}
Rmean.log(kmlogtime,kmsurv)
rmean.full<-Rmean.log(kmlogtime,kmsurv)
rmean<-NULL
pseudo<-NULL
n<-length(Data$Time)

for(i in 1:n){
  res_minus_i<-Data[-i,]
  km.pseudo<-survfit(Surv(re_minus_i$r,res_minus_i$Status==1)~1)
  pseudo.i.time<-km.pseudo$time
  pseudo.i.surv<-km.pseudo$surv
  pseudo.i.logtime<-log(pseudo.i.time)
  rmean[i]<-Rmean.log(pseudo.i.logtime,pseudo.i.surv)
  pseudo[i]<-n*rmean.full-(n-1)*rmean[i]
}

Data["KMlogPseudores"]<-NA
Data$KMlogPseudores<-pseudo
```

Weibull parametric pseudo residuals

```
model<-survreg(Surv(Data$r,Data$Status)~1, data=Data, dist="weibull")
summary<-summary(model)

#Extracting the coefficients
coef<-as.numeric(summary$coefficients)
b0_hat<-coef[1]
sigma_hat<-summary$scale

x_bar<-sum(Data$x)/n
Theta_hat_full = b0_hat+sigma_hat*Ew
Theta_hat<-NULL
Theta_hat_min<-NULL

for(i in 1:n){
  res_minus_i<-Data[-i,]
  model_theta_min<-survreg(Surv(res_minus_i$r,res_minus_i$Status)~1
  ,data=res_minus_i, dist="weibull")
  summary_theta_min<-summary(model_theta_min)
  coef_theta_min<-as.numeric(summary_theta_min$coefficients)
```

```
  b0_hat_theta_min<-coef_theta_min[1]
  sigma_hat_min<-summary_theta_min$scale
  Theta_hat_min[i]<-b0_hat_theta_min+sigma_hat_min*Ew
  Theta_hat[i]<-n*Theta_hat_full-(n-1)*Theta_hat_min[i]
}

Data["ParamPseudores"]<-NA
Data$ParamPseudores<-exp(Theta_hat)
```