Parametric Bootstrap Inference for Transformation Models

Master Thesis in Biostatistics (STA495)

by

Muriel Lynnea Buri

S08-911-968

supervised by

Prof. Dr. Torsten Hothorn

University of Zurich

Master Program in Biostatistics

Zurich, May 2016

University of Zurich[UZH]

Abstract

The purpose of this master thesis is to use the parametric bootstrap resampling method for doing statistical model inference on transformation models. Based on previous research completed by Hothorn *et al.* in 2015 (Hothorn, T., Möst, L., and Bühlmann, P. (2015). Most Likely Transformations. arXiv:1508.06749. Technical report, v2. URL `http://arxiv.org/abs/1508.06749`), this project utilizes the implementation of maximum likelihood-based estimation for transformation models. The framework of conditional transformation models as well as the bootstrap resampling method is profoundly explained within this thesis.

To practically illustrate the use of these approaches, the in R publicly available data set from the German Breast Cancer Study Group-2 (GBSG2) trials was used. A conditional transformation model estimates the conditional distribution of the response variable Y defined from the GBSG2 data set. Consequently, the parametric bootstrap resampling method can be applied to draw B new response variables $Y_1^\star, \ldots, Y_B^\star$ from the conditional distribution function. This procedure resulted in B new conditional transformation models, which were subsequently used for the parametric bootstrap inference. We used log-likelihood ratio statistics as a likelihood based measurement for comparing the bootstrap generated model to the original transformation model. The statistical inference of the bootstrap generated transformation models was carried out in two ways: first, on the model parameters and the distribution thereof; and second, on the data specific prediction functions, *e.g.* the density function, the empirical cumulative distribution function, the survivor function, etc.

Furthermore, this research has shown that the degrees of freedom of the Chi-squared distributed log-likelihood ratio statistics are not defined as they are expected to be. Regarding the not as expected log-likelihood ratio statistics distribution, this thesis does not definitively provide a solution, however, simulations have been included to prove the presumption that a correction of the degrees of freedom in instances of multiply occurring model coefficients is essential. In conclusion, the results of this thesis advance the understanding of graphical model inference of the model parameters of a conditional transformation model as well as the inference of the conditional transformation model itself.

Keywords Unconditional transformation model, conditional transformation model, distribution regression, parametric bootstrap, graphical model inference, likelihood ratio test statistic, degrees of freedom

Acknowledgements

I would like to express my gratitude to Professor Torsten Hothorn for all the help and guidance he has provided me during the work on this thesis. I am very excited to be able to continue working with him for my PhD.
To my friends and former fellow students of the Master Program in Biostatistics at the University of Zurich, - thank you for making the study of biostatistics so enjoyable. Thanks for the great discussions and for your help throughout the semesters. I would also like to thank the professors and lecturers from the Master Program in Biostatistics for their dedicated and passionate work.
I thank my parents, sister and brother for their faith in me and for allowing me to be as ambitious as I want. Thanks, Mum and Dad, for supporting me no matter what path I choose to take.
My extended family and my friends were always eager to know how the project was coming. Thank you for your support and encouragement through this process. And - thank you, Afua, - for your great review and valuable suggestions.

Muriel Lynnea Buri
Zurich, May 2016

"So many people along the way, whatever it is you aspire to do, will tell you it can't be done. But all it takes is imagination. You dream. You plan. You reach. There will be obstacles. There will be doubters. There will be mistakes. But with hard work, with belief, with confidence and trust in yourself and those around you, there are no limits." Michael Phelps

Contents

List of Figures

List of Tables

List of R-Code

1. Introduction

In the context of frequently used regression models the estimation of the conditional mean of the response variable Y is usually in focus. Whenever a distribution is too challenging to analyse, one might tend to simplify the distribution by only concentrating on the mean as the sole comprehensible real number value that describes the centre of the distribution. One often forgets that the mean as a characteristic of a distribution hides other important characteristics such as: variance, skewness and kurtosis. In contrast, the model class of transformation models, which is widely utilized in this thesis, is advantageous in that the higher moments of the conditional distribution are allowed to depend on the explanatory variables. In the framework of transformation models, the whole conditional distribution of the response variable Y is estimated with the help of a strictly monotone increasing transformation function $h(y)$. The paper *Most Likely Transformation*, authored by Hothorn *et al.* in 2015, has revolutionized the state of research of transformation models. For the first time, a full likelihood procedure was introduced for estimating the transformation function along with the model parameters. The combination of such transformation models with the parametric bootstrap resampling method is the main theme of this thesis. By applying the parametric bootstrap to the concept of transformation models, it is possible to obtain additional response variables. Based on these additional responses, additional transformation models are estimated and subsequently used for (graphical) model inference, the so-called *Parametric Bootstrap Inference for Transformation Models*.

1.1. Outline

In a nutshell, the thesis is structured as follows: The second chapter *Theory and Methods* (cf. Chapter 2) gives an overview, as the name suggests, of the theories and methods used in this thesis. This includes an introduction to the concept of transformation models (cf. Section 2.2) as well as an introduction to the bootstrap resampling methods (cf. Section 2.4). The implementation, application and evaluation of the combination of these two concepts are the focus of the third chapter *Modelling and Analysis* (cf. Chapter 3). Afterwards, a graphical parameter and model inference is explained in Sections 3.2 and 3.3, respectively. The fourth chapter - *Discussion* (cf. Chapter 4) - provides the conclusion, summarises the limitations of this body of work, in addition to avenues for future research.

This thesis forms part of the Master Program in Biostatistics at the University of Zurich. For the sake of brevity, long and detailed equations and proofs have been intentionally excluded. A reader that seeks details will find further information by following the in-text references.

1.2. Notation

The notation used for this thesis is inspired by Efron (1979) and Hothorn *et al.* (2015). Vectorial parameters $\boldsymbol{\vartheta}$ are printed in boldface to make it easier to distinguish them from scalar parameters ϑ. A hat on a letter indicates an estimate, such as $\hat{\vartheta}$ (respectively $\hat{\boldsymbol{\vartheta}}$). As in Held and Bové (2013), independent *uni*variate random variables Y_i from a certain distribution contribute to a random sample $Y_{1:N} = (Y_1, \ldots, Y_N)$, whereas n independent *multi*variate random variables $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ik})^T \in \mathbb{R}^k$ are denoted as $\boldsymbol{Y}_{1:n} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N) \in \mathbb{R}^{k \times n}$. A superscript "$\star$" indicates a bootstrap random variable, *e.g.* $Y_i^\star$ indicates a bootstrap random variable from data set $\boldsymbol{Y}$.

$f_Y(y)$ describes the density (or probability mass) function, $F_Y(y)$ the empirical cumulative distribution function, $S_Y(y)$ the survivor function and $\lambda_Y(y)$ the hazard function of $\boldsymbol{Y}_i$. The notation $y_i \overset{\text{iid}}{\sim} F$ for $i = 1, 2, \ldots, n$ indicates an independent and identically distributed sample of size n drawn from the distribution F. The conditional distribution function of Y given $X = x$ is denoted as $F_{Y|X}(y|x)$ or $F(Y \leq y | \boldsymbol{X} = \boldsymbol{x})$.

1.3. Software

All analyses were performed in the **R** system of statistical software (R version 3.3.0 (2016-05-03)), which is freely available at `http://www.r-project.org/`. The following Base packages `grid`, `stats`, `graphics`, `grDevices`, `utils`, `datasets`, `methods`, `base` and other packages `xtable`, `beanplot`, `Hmisc`, `ggplot2`, `Formula`, `lattice`, `SDMTools`, `colorspace`, `MASS`, `survival`, `sltm`, `mlt`, `basefun`, `variables`, `knitr` were used for the analyses and for the compilation of this report. The computing environment on the author's personal computer had the following specifications: OS X Yosemite, Version 10.10.5 (Operating system), 1.7 GHz Intel Core i7 (Processor) and 8 GB 1600 MHz DDR3 (Memory).

2. Theory and Methods

This chapter provides an introduction to the theory and an overview of the methods used in this thesis. Overall, it is divided into five subsections. Section 2.1 puts the transformation models in context of other (well known) regression models; whereas the concept of transformation models for unconditional cases is introduced in Section 2.2. In Section 2.3, conditional transformation models along with their application in **R** are discussed. Section 2.4 highlights an overview of the bootstrap resampling method, followed by a brief introduction to the data set utilized (cf. Section 2.5).

2.1. From Normal Linear Regression Models to Transformation Models

Beginning with the normal linear regression model (NLRM) as shown in the transformation model (cf. Figure 2.1) below, each type of the regression model will be elaborated upon in a clockwise manner.

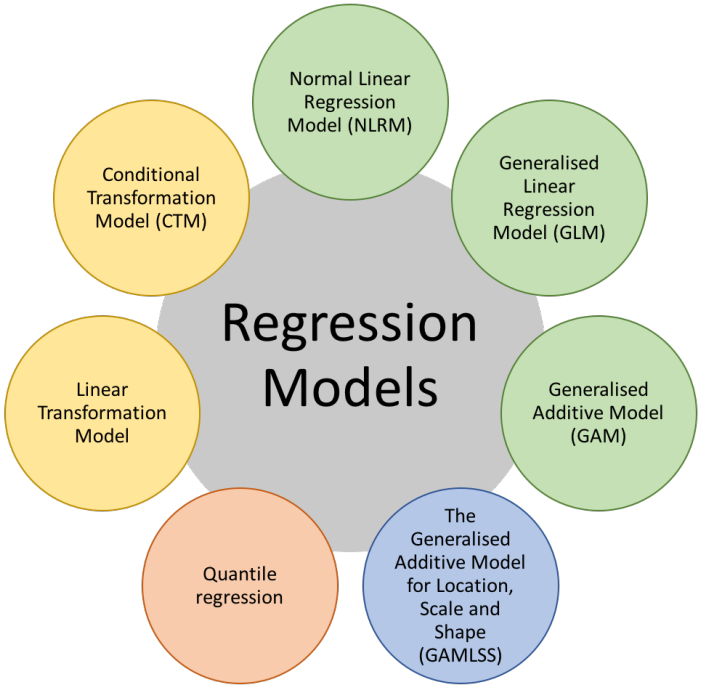

Figure 2.1.: From Normal Linear Regression to Conditional Transformation Models

It is known that most regression models are used to estimate the conditional mean of a response variable Y. In the setup of the normal linear regression model (NLRM) let Y denote a continuous normal distributed

response variable and k denote independent covariates $\boldsymbol{X}_{1:k}$. The latter is used as model inputs to estimate the conditional mean of the response variable Y. $\boldsymbol{x}_{ik}$ whereby, the i-th observation ($i = 1, 2, \ldots, n$) of the k-th covariate, y_i is the i-th observation of the response variable, respectively. The functional form of the NLRM for the i-th response is known as:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_k x_{ik} + \epsilon_i,\ i = 1, \ldots n \\ &= x_{i.}\boldsymbol{\beta} + \epsilon_i = \boldsymbol{X}_i\boldsymbol{\beta} + \epsilon_i \ \Rightarrow \boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{X} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2) \end{aligned} \tag{2.1}$$

Equation 2.1 also depicts the vector form where $\boldsymbol{X} \in \mathbb{R}^{n \times k}$ is the so-called design matrix, $\beta \in \mathbb{R}^k$ the parameter vector and ϵ_i the error term with variance σ^2. The errors $\epsilon_1, \epsilon_2, \ldots, \epsilon_N$ are independent and identically distributed (i.i.d) with $\mathbb{E}(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$, *i.e.* $\epsilon_i \sim N(0, \sigma^2)$. The property of constant variance σ^2 across all the error terms ϵ_i is referred to as *homoscedasticity* (Fahrmeir *et al.*, 2007). This implies that the error terms are independent across the covariates.

In the case of normally distributed errors, we get as a result of the NLRM, an estimator for the conditional expected value of Y given the covariates $\boldsymbol{X}_{1:k}$: $\hat{\mathbb{E}}(y|\boldsymbol{X}) = \hat{\mathbb{E}}(y|x_1, x_2, \ldots, x_k)$. $\hat{\mathbb{E}}(y|\boldsymbol{X})$ is also known as the conditional mean of Y given the covariates $\boldsymbol{X}$. The normal linear regression model is parametric in the sense that we assume $\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{X}$ to be normally distributed. The model has $(k+1)$ parameters: $\beta_1, \beta_2, \ldots, \beta_k, \sigma^2$ and the regression coefficients are perceived to be dependent on the variable x_i to which they belong to. The conditional mean $\mathbb{E}(\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{X})$ increases by β_i when $\boldsymbol{x}_i$ increases by one unit. Therefore, β_i depends on both, the scale of x_i and the possible transformation of $\boldsymbol{x}_i$. Nevertheless, the explanatory variables in the NLRM only have an impact on the mean of the response variable Y, but not on the variance σ^2. Since the covariates of the model only influence the conditional mean of the response variable Y but not the higher moments of the distribution function, it can be inferred to also applicable to both the generalised linear model (GLM) and the generalised additive model (GAM). The higher moments of the distribution function are assumed to be fixed. Due to these similar characteristics, the three models - NLRM, GLM, GAM - are depicted in Figure 2.1 with the same colour. As the name suggests, the GLMs can be interpreted as a generalization of the NLRM and also incorporates more general types of distributions for the response variable Y, *i.e.* distributions from the exponential family (Fahrmeir *et al.*, 2007, p. 301). GAMs can be considered as a concept that incorporates nonlinear forms of the predictors. The linear form $\sum_{i=1}^{n} \beta_i X_i$ gets replaced by a sum of smooth functions $\sum_{i=1}^{n} \beta_i f(X_i)$. GAMs were originally developed by Hastie and Tibshirani (1986).

The model class of generalised additive models for location, scale and shape (GAMLSS) was introduced by Stasinopoulos and Rigby in 2007. This was one of the first attempts to illustrate how the explanatory variables influence higher moments of the distribution function. GAMLSS are statistical (regression) models where the location, scale, skewness and kurtosis parameters for the distribution of the response variable Y can be modelled explicitly as a function of the explanatory variables, *i.e.* covariates.

All of the above mentioned regression models assumed a parametric distribution for the response variable and were considered to "require the definition of a parametric distribution for the response variable" (Möst, 2014). Applied to Figure 2.1 and hence figuratively explained, all the regression models aligned on the right side of the circle assume a parametric distribution for the response variable. In contrast, the approach of the quantile regression (Koenker, 2005) which according to Möst (2014) is a popular approach that does not make any assumptions about the parametric distribution function of the response variable. The quantile regression therefore models the conditional quantile functions of Y given the explanatory variables $\boldsymbol{X}$. As a consequence of the fact that we fit separate models for a grid of probabilities τ to estimate the whole

conditional quantile function, the logical monotonicity of the conditional quantiles is not considered explicitly, and therefore quantile crossing is a familiar problem associated with quantile regression (Möst, 2014; Hothorn *et al.*, 2014). The disadvantage of quantile crossing then, is its ability to lead to an invalid distribution for the response, consequently, there is an obvious transition to the conditional transformation models (CTMs) bearing in mind that the conditional quantile function is the inverse of the conditional distribution function and vice versa. The CTMs blend the favourable properties of the GAMLSS and the quantile regression: The conditional distribution function of the response variable is modelled directly and therefore the mean and all higher moments are influenced by the explanatory variables $\boldsymbol{X}$. Another potential problem of quantile crossing is in the framework of CTMs losing its usefulness, as all conditional quantiles are estimated simultaneously with the conditional distribution function. A more detailed introduction to the framework of transformation models is elaborated upon in the next section.

2.2. The Concept of Transformation Models

In general, transformation models are useful mainly because the whole conditional distribution function of Y is modelled directly and influenced by the explanatory variables $\boldsymbol{X}$. Indeed, the class of transformation models is rich, has been thoroughly researched and has a close connection with the conditional distribution function. Nonetheless, a brief introduction to the general class of linear transformation models followed by an in depth discussion of the linear transformation model is given in Section 2.3.1. Later, the general conditional transformation model will be discussed in Section 2.3.3.

Möst (2014) points out that the origin of transformation models is given by the parametric response transformation suggested by Box and Cox (1964). The authors presented a family of transformations for a non-negative response variable Y depending on a parameter λ. The Box-Cox-Transformation is scaled to be continuous at $\lambda = 0$:

$$h_Y(y|\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(y), & \lambda = 0 \end{cases} \quad \Rightarrow h_Y(y|\lambda) = \tilde{x}^T\beta + \epsilon$$

After the transformation $h_Y(y|\lambda)$ a normal, homoscedastic, linear model is valid:

$$h_Y(y|\lambda) - \tilde{x}^T\beta = \epsilon \sim N(0, \sigma^2)$$
$$\mathbb{P}(Y \leq y|x) = F(y|x) = \Phi\left(\frac{h_Y(y|\lambda) - \tilde{x}^T\beta}{\sigma}\right)$$

The linear transformation models are an extension of the parametric Box-Cox transformation models. In the framework of simple linear transformation models the response transformation $h(y) = -\boldsymbol{x}^T\boldsymbol{\beta} + \epsilon$ is not specified. However, the strictly increasing transformation $h(y)$ is dependent on linear covariate effects and the distribution function F of the random error term ϵ is completely specified:

$$h(y) + \boldsymbol{x}^T\boldsymbol{\beta} = \epsilon \sim F.$$

The conditional distribution function for the linear transformation model is therefore defined as follows:

$$\begin{aligned} \mathbb{P}(Y \leq y|X = x) &= \mathbb{P}(h(Y) \leq h(y)) \\ &= F\big(h(y) + \boldsymbol{x}^T\beta\big). \end{aligned}$$

The response distribution $F\big(h(y) + \boldsymbol{x}^T\beta\big)$ includes a linear shift due to the explanatory variables, hence only the conditional mean of the transformed response is influenced (Möst, 2014). The model complexity of the linear transformation model is restricted as the model is linear in $\boldsymbol{x}$ and does not allow for interaction terms between the response and the explanatory variables. The class of linear transformation models includes the proportional hazards model and the proportional odds model as special cases, consequently the transformation function $h(y)$ is sometimes also called the baseline function.

2.2.1. The Likelihood Function of the Transformation Function h

The content of this section is spurred by the Sections 2 and 3 of the paper *Most Likely Transformations*, authored by Hothorn *et al.* in 2015. As a result, the definitions, notations and corollaries are adopted from this publication and appropriately cited. In order to elaborate on the technical derivations of this paper and to gain a better understanding, a flowchart has been created in the appendix (cf. Appendix A.1, Figure A.1). The latter represents the main steps required to estimate a transformation model in a full likelihood framework.

Hothorn *et al.* (2015) posit that many authors have studied different approaches to estimate the transformation functions, however, a full likelihood estimation procedure is still lacking. Hothorn *et al.* (2015) therefore sought to address this issue by introducing a strictly monotone transformation of some absolute continuous random variable. Whereby, "the likelihood function of the transformed variable can then be characterised by this transformation function. The parameters of appropriate parameterisations of the transformation function, and thus the parameters of the conditional distribution function we are interested in, can then be estimated by maximum likelihood under simple linear constraints allowing classical asymptotic likelihood inference [...]" (Hothorn *et al.*, 2015, Chapter 1. Introduction).

Let $(\Omega, \mathfrak{A}, \mathbb{P})$ denote a probability space, for which Ω is the sample space, the set of all possible outcomes. $\mathfrak{A}$ is the set of events and $\mathbb{P}$ the assignment of probabilities to the events, respectively. The function $\mathbb{P}$ can be understood as a function from the events to probabilities. Let $(\Xi, \mathfrak{C})$ describe a measureable space with at least ordered sample space Ξ. The motivation for setting up the transformation model is our interest in inferring about the distribution $\mathbb{P}_Y$ of a random variable Y, *i.e.* the probability space $(\Xi, \mathfrak{C}, \mathbb{P}_Y)$ defined by the $\mathfrak{A} - \mathfrak{C}$ measureable function $Y : \Omega \to \Xi$. For the sake of notational simplicity, we here only present the results for the unconditional and ordered cases are presented. The distribution $\mathbb{P}_Y = f_Y \odot \mu$ is dominated by some measure μ and characterised by its density function f_Y, distribution function $F_Y(y)$, quantile function $F_Y^{-1}(p)$, hazard function $\lambda_Y(y)$, or cumulative hazard function $\Lambda_Y(y)$. As in Hothorn *et al.* (2015), we assume strict monotonicity of F_Y, *i.e.* $F_Y(y_1) < F_Y(y_2)\, \forall y_1 < y_2 \in \Xi$, with the aim of obtaining an estimate $\hat{F}_{Y,N}$ of the distribution function F_Y from a random sample $Y_1, \ldots, Y_N \overset{\text{iid}}{\sim} \mathbb{P}_Y$. The path to achieving this goal is not straightforward, and requires further investigation. Hereafter, we show that it is always possible to write the potentially complex distribution function F_Y as a composition of a much simpler *a priori* specified distribution function F_Z and a strictly monotone transformation function h. The estimation of F_Y is then reduced to obtaining an estimate $\hat{h}$. Since these definitions are technically and conceptually attractive, it is further elaborated upon in the subsequent paragraph.

Let $(\mathbb{R}, \mathfrak{B})$ denote the Euclidian space with Borel σ-algebra and $Z : \Omega \to \mathbb{R}$ a $\mathfrak{A} - \mathfrak{B}$ measureable function such that the measure $\mathbb{P}_Z = f_Z \odot \mu_L$ is absolute continuous (μ_L denotes the Lebesgue measure) in the probability space $(\mathbb{R}, \mathfrak{B}, \mathbb{P}_Z)$. The corresponding distribution and quantile function are F_Z and F_Z^{-1}, respectively. In addition, we assume $f_Z(z) : \mathbb{R} \to (0, \infty)$ and the existence of the first two derivatives of

$f_Z(z)$ with respect to z and $F_Z(z) : (-\infty, \infty) \to [0,1]$. By definition, all parameters for F_Z have to be known and possible choices for F_Z include:

- the standard normal distribution: $F_Z(z) = \Phi(z)$
- the standard logistic (SL) distribution: $F_Z(z) = F_{\text{SL}}(z) = (1 + \exp(-z))^{-1}$
- the minimum extreme value (MEV) distribution: $F_Z(z) = F_{\text{MEV}}(z) = 1 - \exp(-\exp(z))$

Our final goal is to obtain $\hat{F}_{Y,N}$ of the distribution function F_Y. But first, we show that there always exists a unique and strictly monotone transformation g, such that the unknown and potentially complex distribution $\mathbb{P}_Y$ can be generated from the simple and known distribution $\mathbb{P}_Z$ via $\mathbb{P}_Y = \mathbb{P}_{g \circ Z}$. Due to the existence and uniqueness of g, it is defined as corollary (seen in Hothorn *et al.* (2015)):

Corollary 1. *For all random variables Y and Z, there exists a unique strictly monotone increasing transformation g such that $\mathbb{P}_Y = \mathbb{P}_{g \circ Z}$.*

Proof. Let $g = F_Y^{-1} \circ F_Z$ and $Z \sim \mathbb{P}_Z$. Then $U := F_Z(Z) \sim \text{U}[0,1]$ and $Y = F_Y^{-1}(U) \sim \mathbb{P}_Y$ by the probability integral transform. Let $h : \Xi \to \mathbb{R}$ such that $F_Y(y) = F_Z(h(y))$. From

$$F_Y(y) = (F_Z \circ F_Z^{-1} \circ F_Y)(y) = F_Z(F_Z^{-1}(F_Y(y))) = F_Z\big(F_Z^{-1}\big(F_Z(h(y))\big)\big) \iff h = F_Z^{-1} \circ F_Y$$

the uniqueness of h and therefore g is given. Corollary 1 also covers the discrete case.

The quantile function F_Z^{-1} and the distribution function F_Y exist by assumption and are both strictly monotone and right-continuous. Therefore, h and g are both strictly monotone and right-continuous. □

The following corollaries are also taken over from Hothorn *et al.* (2015).

Corollary 2. *For $\mu = \mu_L$, we have $g = h^{-1}$ and $h'(y) = f_Z((F_Z^{-1} \circ F_Y)(y))^{-1} f_Y(y)$.*

This result for absolute continuous random variables Y can be found in many textbooks (for example in Lindsey, 1996).

Corollary 3. *For the counting measure $\mu = \mu_C$, $h = F_Z^{-1} \circ F_Y$ is a right-continuous step-function because F_Y is a right-continuous step-function with steps at $y \in \Xi$.*

Example The classical textbook example for transformations of random variables is $Y = Z^2 \sim \chi_1^2$ from $Z \sim \text{N}(0,1)$, *i.e.* using the non-monotone transformation z^2:

$$f_Z(z) = \sqrt{2\pi} \exp\left(-\frac{z^2}{2}\right)$$
$$f_Y(y) = \sqrt{2\pi} \exp\left(-\frac{y}{2}\right) = f_Y(z^2)$$

Alternatively, we can write $Z = h(Y)$ and $Y = g(Z)$ with $h = \Phi^{-1} \circ F_{\chi_1^2}$ and $g = h^{-1} = F_{\chi_1^2}^{-1} \circ \Phi$. The functions g and h are unique and strictly monotone transformations switching between the standard normal and the χ_1^2 distribution. The χ_1^2 distribution can be generated from the standard normal by the transformation $g = F_{\chi_1^2}^{-1} \circ \Phi$ and the back-transformation is $h = \Phi^{-1} \circ F_{\chi_1^2}$.

The next steps are:

- characterisation of the distribution F_Y by the corresponding transformation function h,

- setting-up the corresponding likelihood $\mathcal{L}(h)$ of such a transformation function h and
- estimating the transformation function based on this likelihood.

To demonstrate this idea, let $\mathcal{H} = \{h : \Xi \to \mathbb{R} | \mathfrak{C} - \mathfrak{B} \text{ measureable}, h(y_1) < h(y_2) \, \forall y_1 < y_2 \in \Xi\}$ denote the space of all strictly monotone transformation functions. Once the transformation function h is established, F_Y can be evaluated as $F_Y(y|h) = F_Z(h(y)) \, \forall y \in \Xi$. This indicates that g does not necessarily follow, consequently it is essential to study the transformation h. Further, due to the different types of response variables Y, we have different definitions for the density function:

- for absolute continuous variables Y ($\mu = \mu_L$):

$$\frac{\partial F_Y(y|h)}{\partial y} = \frac{\partial F_Z(h(y))}{\partial y} \iff f_Y(y|h) = f_Z(h(y))h'(y)$$

- for discrete responses Y ($\mu = \mu_C$) with finite sample space $\Xi = \{y_1, \dots, y_K\}$:

$$f_Y(y_k|h) = \begin{cases} F_Z(h(y_k)) & k = 1 \\ F_Z(h(y_k)) - F_Z(h(y_{k-1})) & k = 2, \dots, K-1 \\ 1 - F_Z(h(y_{k-1})) & k = K \end{cases}$$

- for countably infinite sample spaces $\Xi = \{y_1, y_2, y_3, \dots\}$

$$f_Y(y_k|h) = \begin{cases} F_Z(h(y_k)) & k = 1 \\ F_Z(h(y_k)) - F_Z(h(y_{k-1})) & k > 1. \end{cases}$$

With the conventions $F_Z(h(y_0)) := F_Z(h(-\infty)) := 0$ and $F_Z(h(y_K)) := F_Z(h(\infty)) := 1$ only the more compact notation $f_Y(y_k|h) = F_Z(h(y_k)) - F_Z(h(y_{k-1}))$ will be used.

As Lindsey (1996) defined and Hothorn *et al.* (2015) reiterated, for a given transformation function h, the likelihood contribution of a datum $C = (\underline{y}, \bar{y}] \in \mathfrak{C}$ is determined in terms of the distribution function:

$$\mathcal{L}(h|Y \in C) := \int_C f_Y(y|h) d\mu(y) = F_Z(h(\bar{y})) - F_Z(h(\underline{y})).$$

The aforementioned definition particularly applies to most practically interesting scenarios, oftentimes allowing for discrete and (conceptually) continuous, as well as censored or truncated observations of C.

Hothorn *et al.* (2015) has summarised the likelihood contribution of an "exact continuous" or left, right or interval-censored continuous or discrete observation $(\underline{y}, \bar{y}]$ as follows:

$$\mathcal{L}(h|Y \in (\underline{y}, \bar{y}]) = \begin{cases} f_Z(h(y))h'(y) & y = (\underline{y} + \bar{y})/2 \in \Xi \quad \text{"exact continuous"} \\ 1 - F_Z(h(\underline{y})) & y \in (\underline{y}, \infty) \cap \Xi \quad \text{"right-censored"} \\ F_Z(h(\bar{y})) & y \in (-\infty, \bar{y}] \cap \Xi \quad \text{"left-censored"} \\ F_Z(h(\bar{y})) - F_Z(h(\underline{y})) & y \in (\underline{y}, \bar{y}] \cap \Xi \quad \text{"interval-censored"}, \end{cases}$$

under the assumption of random censoring. Klein and Moeschberger (2003) (p. 69) attribute accidental deaths or the migration of human populations as typical examples, whereby the random censoring times may

be thought to be independent of the main event time of interest. Klein and Moeschberger (2003) further highlight the fact that the likelihood is more complex under dependent censoring. This body of work unfortunately does not elaborate on this idea.

In the case of truncated observations in the interval $(y_l, y_r] \subset \Xi$, Hothorn *et al.* (2015) define the above likelihood contribution differently in terms of the distribution function conditional on the truncation

$$F_Y(y|Y \in (y_l, y_r]) = F_Z(h(y)|Y \in (y_l, y_r]) = \frac{F_Z(h(y))}{F_Z(h(y_r)) - F_Z(h(y_l))} \quad \forall y \in (y_l, y_r]$$

and thus the likelihood contribution changes to (Klein and Moeschberger, 2003)

$$\frac{\mathcal{L}(h|Y \in (\underline{y}, \bar{y}])}{F_Z(h(y_r)) - F_Z(h(y_l))} = \frac{\mathcal{L}(h|Y \in (\underline{y}, \bar{y}])}{\mathcal{L}(h|Y \in (y_l, y_r])} \quad \text{when } y_l < \underline{y} < \bar{y} \leq y_r.$$

Lindsey (1999) emphasizes the importance of the fact, that the likelihood is always defined in terms of a distribution function. Therefore, it makes sense to directly model the distribution function of interest after it. Hothorn *et al.* (2015) state that the ability to uniquely characterise this distribution function by the transformation function h, gives rise to the following definition of the most likely transformation estimator $\hat{h}_N$.

Definition 1 (Most likely transformation)**.**

Let $C_1, \ldots, C_N$ denote an independent sample of possibly censored or truncated observations from $\mathbb{P}_Y$. The estimator

$$\hat{h}_N := \arg\max_{\tilde{h} \in \mathcal{H}} \sum_{i=1}^{N} \log(\mathcal{L}(\tilde{h}|Y \in C_i))$$

is called the most likely transformation (MLT).

Example For absolute continuous Y the likelihood and log-likelihood for h are approximated by the density and log-density evaluated at $y = (\underline{y} + \bar{y})/2$, respectively:

$$\begin{aligned} \mathcal{L}(h|Y \in (\underline{y}, \bar{y}]) &\approx f_Z(h(y))h'(y) \\ \log(\mathcal{L}(h|Y \in (\underline{y}, \bar{y}])) &\approx \log(f_Z(h(y))) + \log(h'(y)). \end{aligned}$$

Strict monotonicity of the transformation function h is required, otherwise the likelihood is not defined. The term $\log(h'(y))$ is not a penalty term but the likelihood favours transformation functions with large positive derivative at the observations. If we assume $Y \sim \mathrm{N}(\alpha, \sigma^2)$ and for the choice $F_Z \sim \mathrm{N}(0, 1)$ with $F_Z = \Phi$ and $f_Z = \phi$, then h can be restricted to linear functions $h(y) = (y - \alpha)\sigma^{-1}$. The likelihood reduces to

$$\begin{aligned} \mathcal{L}(h|Y \in (\underline{y}, \bar{y}]) \approx f_Z(h(y))h'(y) &= \underbrace{\phi((y - \alpha)\sigma^{-1})}_{f_Z(h(y))} \underbrace{\sigma^{-1}}_{h'(y)} \\ &= \phi_{\alpha,\sigma^2}(y) \\ &= f_Y(y|\alpha, \sigma^2). \end{aligned}$$

Along with this example Hothorn *et al.* (2015) have emphasized that it is only within this simple location-scale family, that the most likely transformation is characterised by the parameters of the normal distribution of Y. Consequently, for other choices of F_Z, the most likely transformation is non-linear. Nevertheless, the distribution function $F_Y = F_Z(h(y))$ is invariant with respect to F_Z because we can always write h as $F_Z^{-1} \circ$

F_Y. In other words, with $F_Z \neq \Phi$ normal responses Y can still be modelled, but only with a non-linear transformation function h.

Henceforth, we do not assume any specific form of the transformation function but parameterise h in terms of a basis function. Consequently, this parameterisation, a corresponding family of distributions, a maximum likelihood estimator and a large class of models for unconditional and conditional distributions will be introduced in the subsequent paragraphs below.

2.2.2. Maximum Likelihood Transformation Models

A basis function $\boldsymbol{a} : \Xi \to \mathbb{R}^P$ parametrises the transformation function $h(y) = \boldsymbol{a}(y)^\top \boldsymbol{\vartheta}, \boldsymbol{\vartheta} \in \mathbb{R}^P$ in such a way that $h(y)$ is a linear function of the basis-transformed argument y and the parameter vector $\boldsymbol{\vartheta}$. The choice of the basis function $\boldsymbol{a}$ is in close connection with the Bernstein polynomials, which are introduced and discussed in Section 2.2.3 of this thesis. The exact likelihood $\mathcal{L}$ only requires evaluation of h, however, the approximation for "exact" observations of absolute continuous random variables makes the evaluation of the first derivative of $h(y)$ with respect to y necessary. The derivative with respect to y is given by $h'(y) = \boldsymbol{a}'(y)^\top \boldsymbol{\vartheta}$ and we assume that $\boldsymbol{a}'$ is available. In the style of Hothorn *et al.* (2015), we subsequently use the notation $h = \boldsymbol{a}^\top \boldsymbol{\vartheta}$ and $h' = \boldsymbol{a}'^\top \boldsymbol{\vartheta}$ for the transformation function and its first derivative omitting the argument y. We assume that h and h' are bounded away from $-\infty$ and ∞.

For a specific choice of F_Z and $\boldsymbol{a}$, the transformation family of distributions consists of all distributions $\mathbb{P}_Y$ whose distribution function F_Y is given as the composition $F_Z \circ \boldsymbol{a}^\top \boldsymbol{\vartheta}$; Hothorn *et al.* (2015) refer to this as a *Transformation family*.

Definition 2 (Transformation family)**.**

The distribution family

$$\mathbb{P}_{Y,\Theta} = \{F_Z \circ \boldsymbol{a}^\top \boldsymbol{\vartheta} | \boldsymbol{\vartheta} \in \Theta\}$$

with parameter space $\Theta = \{\boldsymbol{\vartheta} \in \mathbb{R}^P | \boldsymbol{a}^\top \boldsymbol{\vartheta} \in \mathcal{H}\}$ *is called transformation family of distributions* $\mathbb{P}_{Y,\boldsymbol{\vartheta}}$ *with transformation functions* $\boldsymbol{a}^\top \boldsymbol{\vartheta} \in \mathcal{H}$*,* μ*-densities* $f_Y(y|\boldsymbol{\vartheta})$*,* $y \in \Xi$*, and error distribution function* F_Z*.*

Hothorn *et al.* (2015) also hypothesize that the classical definition of a transformation family relies on the idea of invariant distributions, *i.e.* only the parameters of a distribution are changed by a transformation function but not the distribution itself. Throughout this thesis, the transformation functions that do change the shape of the distribution are explicitly allowed. The transformation function $\boldsymbol{a}^\top \boldsymbol{\vartheta}$ is, at least in principle, flexible enough to generate any distribution function $F_Y = F_Z \circ \boldsymbol{a}^\top \boldsymbol{\vartheta}$ from the distribution function F_Z. As a result, the term "error distribution" function for F_Z as seen in Fraser (1968) is introduced. To estimate $\hat{F}_{Y,N}$ of the unknown distribution function F_Y has been our original goal. By redefining F_Y to $F_Z(h(y))$ with a known F_Z, the problem reduces to estimating the unknown transformation function h with the parameter vector $\boldsymbol{\vartheta}$. But thanks to the known likelihood function $\mathcal{L}(\boldsymbol{a}^\top \boldsymbol{\vartheta} | Y \in (\underline{y}, \bar{y}])$, it reduces further and it remains a maximisation of the likelihood function such that the estimator of $\boldsymbol{\vartheta}$ can be defined as the maximum likelihood estimator.

Definition 3 (Maximum likelihood estimator)**.**

$$\hat{\boldsymbol{\vartheta}}_N := \arg\max_{\boldsymbol{\vartheta} \in \Theta} \sum_{i=1}^{N} \log(\mathcal{L}(\boldsymbol{a}^\top \boldsymbol{\vartheta} | Y \in C_i))$$

As a result of defining the maximum likelihood estimator $\hat{\boldsymbol{\vartheta}}_N$, the plug-in estimators of the most likely transformation function along with the corresponding estimator of our target distribution F_Y should be defined:

- Plug-in estimators of most likely transformation function: $\hat{h}_N := \boldsymbol{a}^\top \hat{\boldsymbol{\vartheta}}_N$
- Estimator of our target distribution F_Y: $\hat{F}_{Y,N} := F_Z \circ \hat{h}_N$

Since the original aim of characterising the distribution F_Y by the corresponding transformation function h is still intact, an estimate $\hat{F}_{Y,N}$ of the distribution function F_Y will be elucidated from the random sample $Y_1, \ldots, Y_N \overset{\text{iid}}{\sim} \mathbb{P}_Y$. Thanks to the above definitions, the estimation of the target distribution $\hat{F}_{Y,N}$ is now embedded in the maximum likelihood framework. Hence, only the regularity conditions (cf. Definition 4.1, p. 80, Held and Bové, 2013) remain to be shown in order to benefit from the well-established asymptotic theory. In such instances, the asymptotic analysis benefits from the standard results extracted from the asymptotic behaviour of maximum likelihood estimators. Therefore, it is possible to derive the score function and Fisher information function depending on the different characteristics of the response variable Y. Building upon this idea, the standard likelihood inference on the model parameters $\boldsymbol{\vartheta}$ can be performed.

Further, Hothorn *et al.* (2015) discuss three additional theorems which are omitted here. These theorems point out the conditions on the densities of the error function f_Y and on the basis function $\boldsymbol{a}$ to ensure consistency and asymptotic normality of the sequence of maximum likelihood estimators $\hat{\boldsymbol{\vartheta}}_N$. Additionally, an estimator of the asymptotic covariance matrix of $\hat{\boldsymbol{\vartheta}}_N$ is given in Hothorn *et al.* (2015).

For now, we complete this theoretical introduction by formally defining the class of transformation models according to Hothorn *et al.* (2015).

Definition 4 (Transformation model)**.**

The triple $(F_Z, \boldsymbol{a}, \boldsymbol{\vartheta})$ *is called transformation model.*

The transformation model $(F_Z, \boldsymbol{a}, \boldsymbol{\vartheta})$ fully defines the distribution of Y via $F_Y = F_Z \circ \boldsymbol{a}^\top \boldsymbol{\vartheta}$ and thus the corresponding likelihood $\mathcal{L}(\boldsymbol{a}^\top \boldsymbol{\vartheta} | Y \in (\underline{y}, \bar{y}])$. Our definition of transformation models as $(F_Z, \boldsymbol{a}, \boldsymbol{\vartheta})$ is strongly tied to the idea of structural inference. Fraser (1968) described a measurement model $\mathbb{P}_Y$ for Y by an error distribution $\mathbb{P}_Z$ and a structural equation $Y = g \circ Z$ where g is a linear function.

Hothorn *et al.* (2015) define such a transformation family or model as "parametric" when F_Z and the basis function $\boldsymbol{a}$ correspond to a distribution F_Y and its parameters are directly linked to the model coefficients $\boldsymbol{\vartheta}$. A semi-parametric transformation model only partially specifies parameters of F_Y through $\boldsymbol{\vartheta}$, and a non-parametric model is characterised by the invariance of $\hat{F}_{Y,N}$ with respect to F_Z (Hothorn *et al.*, 2015).

A flowchart that summarises these concepts that allow for the estimation of a transformation model in a full likelihood framework can be found in the appendix (cf. Appendix A.1, Figure A.1). As a side note, there also exists a fully Bayesian treatment of transformation models despite being excluded from this thesis.

2.2.3. The Bernstein Polynomials

In the context of estimating a transformation model in a full likelihood framework the Bernstein polynomials (for an overview see Farouki (2012)) are important regarding the choice of the basis function $\boldsymbol{a} : \Xi \to \mathbb{R}^P$ for the parametrisation of the transformation function $h(y) = \boldsymbol{a}(y)^\top \boldsymbol{\vartheta}$. In case of order M ($P = M + 1$) the

Bernstein polynomial is defined on the interval $[\underline{y}, \bar{y}]$ as:

$$h(y) = \boldsymbol{a}_{\text{Bs},M}(y)^\top \boldsymbol{\vartheta} = \sum_{m=0}^{M} \vartheta_m f_{\text{Be}(m+1,M-m+1)}(\tilde{y})/(M+1)$$
$$h'(y) = \boldsymbol{a}'_{\text{Bs},M}(y)^\top \boldsymbol{\vartheta} = \sum_{m=0}^{M-1} (\vartheta_{m+1} - \vartheta_m) f_{\text{Be}(m+1,M-m)}(\tilde{y}) \times M/((M+1)(\bar{y} - \underline{y})),$$

where $\tilde{y}$ coupled with $f_{\text{Be}(m,M)}$ is $\tilde{y} = (y - \underline{y})/(\bar{y} - \underline{y}) \in [0,1]$ and the density of the Beta distribution, respectively. An important assumption regarding the Bernstein polynomial is its monotonicity due to the linear constraints on the parameters $\vartheta_m \leq \vartheta_{m+1}$ for all $m = 0, \ldots, M$. This monotonicity is especially important in the context of transformation models as the transformation function needs to be strictly monotone increasing. Obviously, it is convenient to choose a Bernstein polynomial as the basis function $\boldsymbol{a}_{\text{Bs},M}$ to parametrise the transformation function $h(y) = \boldsymbol{a}_{\text{Bs},M}(y)^\top \boldsymbol{\vartheta}$ so that one can ensure a strict monotone increasing transformation function h.

The question that arises therefore, is to what degree is the Bernstein polynomial optimal? In the vignette of the `mlt` package (Hothorn, 2016b, p. 10), it is stated that neither extremes, - a too small nor a too high degree - should be chosen. On the one hand, $\boldsymbol{a}_{\text{Bs},1}$ would only allow linear transformation functions of the distribution function F_Z to occur, consequently, F_Y is restricted to the distribution family of F_Z, but on the other hand, a model with basis function $\boldsymbol{a}_{\text{Bs},N-1}$ has one parameter for each observation, meaning the model is overfitted. In applications, it seems best to test the effects of the degree of Bernstein polynomial depending on the Akaike information criterion (AIC) of the model. The degree of Bernstein polynomial which leads to the smallest AIC is the one to be chosen for the model. However, there is the difficulty for some transformation models to define the right degree of freedom (cf. Section 3.1.2 for details), consequently for some models the AIC is doubted being correctly defined.

2.3. The Conditional Transformation Model (CTM)

In this section, the concept of conditional transformation models will be introduced. This will be achieved by highlighting the special cases of this model class, and giving an example of how the normal linear regression model (NLRM) is estimated within the framework of conditional linear transformation models with a linear shift.

The class of conditional transformation models includes transformation models with transformation functions that depend on the explanatory variables $\boldsymbol{X} \in \mathcal{X}$. Those transformation functions are usually of the form $h(\cdot|\boldsymbol{x}) : \Xi \to \mathbb{R}$. The corresponding distribution function $F_{Y|\boldsymbol{X}=\boldsymbol{x}}$ can be written as $F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y) = F_Z(h(y|\boldsymbol{x}))$. Like in the unconditional case introduced in Corollary 1 (cf. Section 2.2.1), there also exists a strictly monotone transformation function for the conditional case $h(\cdot|\boldsymbol{x}) = F_Z^{-1} \circ F_{Y|\boldsymbol{X}=\boldsymbol{x}}$ such that $F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y) = F_Z(h(y|\boldsymbol{x}))$.

2.3.1. The Linear Transformation Model

A linear transformation model with a linear shift is the simplest form of a regression model in the class of conditional transformation models. The conditional distribution function is:

$$\mathbb{P}(Y \leq y|\boldsymbol{X} = \boldsymbol{x}) = F_Z(h(y|\boldsymbol{x})) = F_Z(h_Y(y) - \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}) = F_Z(\boldsymbol{c}(y,\boldsymbol{x})^\top \boldsymbol{\vartheta}) = F_Z(\boldsymbol{a}(y)^\top \boldsymbol{\vartheta}_1 - \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}).$$

In other words, the conditional transformation function is parametrised as $h(y|\boldsymbol{x}) = \boldsymbol{c}(y,\boldsymbol{x})^\top \boldsymbol{\vartheta}$ and no longer $h(y) = \boldsymbol{a}^\top \boldsymbol{\vartheta}$. As a result, the basis function $\boldsymbol{b} : \mathcal{X} \to \mathbb{R}^Q$ for the explanatory variables is introduced. As suggested in Hothorn *et al.* (2015), the joint basis for both y and $\boldsymbol{x}$ is called $\boldsymbol{c} : \Xi \times \mathcal{X} \to \mathbb{R}^{d(P,Q)}$, and its dimension $d(P,Q)$ depends on the way the two basis functions $\boldsymbol{a}$ and $\boldsymbol{b}$ are combined (for example $\boldsymbol{c} = (\boldsymbol{a}^\top, \boldsymbol{b}^\top)^\top \in \mathbb{R}^{P+Q}$ or $\boldsymbol{c} = (\boldsymbol{a}^\top \otimes \boldsymbol{b}^\top)^\top \in \mathbb{R}^{PQ}$).

The simple transformation function $h(y|\boldsymbol{x}) = h_Y(y) + h_{\boldsymbol{x}}(\boldsymbol{x})$ where the explanatory variables only contribute a shift $h_{\boldsymbol{x}}(\boldsymbol{x})$ to the conditional transformation function is an important special case. Hothorn *et al.* (2015) state that this shift is often assumed to be linear in $\boldsymbol{x}$, so the function $m(\boldsymbol{x}) = \boldsymbol{b}(\boldsymbol{x})^\top \boldsymbol{\beta} = \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}$ will be used to denote linear shifts. $\boldsymbol{b}(\boldsymbol{x}) = \tilde{\boldsymbol{x}}$ is to be understood as one row of the design matrix without intercept. The conditional transformation function $\boldsymbol{c}(y,\boldsymbol{x})^\top \boldsymbol{\vartheta} = \boldsymbol{a}(y)^\top \boldsymbol{\vartheta}_1 + \boldsymbol{b}(\boldsymbol{x})^\top \boldsymbol{\vartheta}_2$ is split into the two terms $h_Y(y) = \boldsymbol{a}(y)^\top \boldsymbol{\vartheta}_1$ and $h_{\boldsymbol{x}}(\boldsymbol{x}) = \boldsymbol{b}(\boldsymbol{x})^\top \boldsymbol{\vartheta}_2 = m(\boldsymbol{x}) = -\tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}$, whereas the conditional distribution function is

$$\begin{aligned}
\mathbb{P}(Y \leq y | \boldsymbol{X} = \boldsymbol{x}) = F_Z(h(y|\boldsymbol{x})) &= F_Z(h_Y(y) + h_{\boldsymbol{x}}(\boldsymbol{x})) = F_Z(\boldsymbol{c}(y,\boldsymbol{x})^\top \boldsymbol{\vartheta}) \\
&= F_Z(\boldsymbol{a}(y)^\top \boldsymbol{\vartheta}_1 + \boldsymbol{b}(\boldsymbol{x})^\top \boldsymbol{\vartheta}_2) \\
&= F_Z(\boldsymbol{a}(y)^\top \boldsymbol{\vartheta}_1 - \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta})
\end{aligned}$$

The three theorems, which were previously mentioned in Section 2.2.2 and are discussed in Hothorn *et al.* (2015) are also applicable here. Consequently, the performance of standard likelihood inference on the model parameters $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2)^\top$ of the conditional transformation model is also possible.
For a better understanding of the concept, we examine the normal linear regression model (NLRM) from the perspective of linear transformation models.

The linear regression model reviewed from the linear transformation model perspective

We rewrite the classical normal linear model $Y = \boldsymbol{X}\boldsymbol{\beta} + \epsilon, \quad Y \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2)$ in the context of conditional transformation models: $Y \sim \mathrm{N}(\alpha + m(\boldsymbol{x}), \sigma^2)$ with conditional distribution function

$$F_{Y|\boldsymbol{X}=\boldsymbol{x}}(y) = \Phi\left(\frac{y - \alpha - m(\boldsymbol{x})}{\sigma}\right) = \Phi(h(y|\boldsymbol{x}))$$

and transformation function

$$\begin{aligned}
h(y|\boldsymbol{x}) = h_Y(y) + h_{\boldsymbol{x}}(\boldsymbol{x}) &= \underbrace{y/\sigma - \alpha/\sigma}_{h_Y(y)} - \underbrace{m(\boldsymbol{x})/\sigma}_{h_{\boldsymbol{x}}(\boldsymbol{x})} \\
&= \boldsymbol{c}(y,\boldsymbol{x})^\top \boldsymbol{\vartheta} = \underbrace{\boldsymbol{a}(y)^\top \boldsymbol{\vartheta}_1}_{h_Y(y)} + \underbrace{\boldsymbol{b}(\boldsymbol{x})^\top \boldsymbol{\vartheta}_2}_{h_{\boldsymbol{x}}(\boldsymbol{x})} \\
&= \underbrace{(y,1)}_{\boldsymbol{a}(y)^\top} \cdot \underbrace{(\sigma^{-1}, -\sigma^{-1}\alpha)^\top}_{\boldsymbol{\vartheta}_1} + \underbrace{(\tilde{\boldsymbol{x}})}_{\boldsymbol{b}(\boldsymbol{x})^\top} \cdot \underbrace{(-\sigma^{-1}\boldsymbol{\beta}^\top)^\top}_{\boldsymbol{\vartheta}_2},
\end{aligned}$$

where $\boldsymbol{a}(y)$, $\boldsymbol{b}(\boldsymbol{x})$ and $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2)^\top$ are the basis functions and parameters, respectively. Nonetheless, there is the constraint $\sigma > 0$. In a more compact notation, one can write: $(\Phi, (y, 1, \tilde{\boldsymbol{x}}^\top)^\top, \boldsymbol{\vartheta})$.

Hothorn *et al.* (2015) note that this model is parametric even though the parameters are the inverse standard deviation and the inverse negative coefficient of variation as opposed to the mean and variance of the original normal distribution.

2.3.2. Specifying a Linear Transformation Model in R

This following section contains an applied example in contrast to the before introduced technical steps on how to fit a transformation model and will be based on the vignette document of the `mlt` package (Hothorn, 2016b). The example is in the context of survival analysis and highlights all the important functions that are needed to fit a linear transformation model in the software environment **R**.

Data from the German Breast Cancer Study Group-2 (GBSG2) trial is used to explain this concept in more detail in Section 2.5.1. The focus in this section therefore, is to demonstrate the **R** implementation of the model. In so doing, an estimation of the recurrence-free survival time (positive absolutely continuous response variable) of the GBSG2 trial conditional on all covariates given in the data set will be realised.

The GBSG2 data set will be loaded from the `TH.data` package (Hothorn, 2015). The definition of a formula for the covariables of the model is necessary because the `as.basis()` method of the `basefun` package (Hothorn, 2016a) needs a formula (or factor) as its argument. The `as.basis()` function returns a function itself for the evaluation of the basis functions with corresponding `model.matrix` and has two arguments: (1) `remove_intercept` removes the intercept after appropriate contrasts were computed and (2) `negative` multiplies the model matrix with -1.

```
# Load data
data("GBSG2", package = "TH.data")
# Define covariables
xvar <- names(GBSG2)
xvar <- xvar[!(xvar %in% c("time", "cens"))]
mlt_covariates <- as.formula(paste("~", xvar, collapse = "+"))
basis_x <- as.basis(mlt_covariates, data = GBSG2,
                    remove_intercept = TRUE)
# Define survival object
GBSG2$y <- with(GBSG2, Surv(time, cens))
y_var <- numeric_var("y", support = c(0, max(GBSG2$time) + 0.1))
basis_y <- Bernstein_basis(y_var, order = 10, ui = "increasing")
# Model specification
ctm_coxph_mod <- ctm(basis_y, shifting = basis_x, todistr = "MinExtrVal")
# Model estimation
mlt_coxph_mod <- mlt(ctm_coxph_mod, data = GBSG2, scale = TRUE, checkGrad = FALSE)
```

R-Code 2.1: Explaining in a step-by-step manner how the conditional transformation model $(F_Z, (\boldsymbol{a}_{\text{Bs},10}(y)^\top, \boldsymbol{b}(\boldsymbol{x})^\top)^\top, \hat{\boldsymbol{\vartheta}}_N)$ is estimated by using the framework specific functions of the `basefun` (Hothorn, 2016a), `variables` (Hothorn, 2016c) and `mlt` (Hothorn, 2016b) packages in **R**

`numeric_var()` from `variables` package (Hothorn, 2016c) saves a formal description of a discrete numeric variables with integer-valued `support` argument which is later passed to the `Bernstein_basis()` function from `basefun` package (Hothorn, 2016a). A Bernstein polynomial (cf. Section 2.2.3, for a more detailed overview see Farouki (2012)) is used as a parametrisation of the continuous response. The associated `Bernstein_basis()` function implemented in the `basefun` package (Hothorn, 2016a) returns such a function for the evaluation of the basis functions with corresponding `model.matrix` and `predict` methods. As the name suggests, the argument `order` defines the order of the polynomial (here: 10) and the argument `ui` is a character describing the possible constraints (here: "`increasing`") on the Bernstein polynomial. As

explained in Section 2.2.3 of this thesis, neither a too small nor a too high order for the Bernstein polynomial should be chosen. After having defined the basis functions for the response (`basis_y`) as well as for the explanatory variables (`basis_x`), we are ready to specify the conditional transformation model by using the `ctm()` function from `mlt` package (Hothorn, 2016b). The conditional transformation model $(F_Z, (\boldsymbol{a}_{\text{Bs},10}(y)^\top, \boldsymbol{b}(\boldsymbol{x})^\top)^\top, \hat{\boldsymbol{\vartheta}}_N)$ is now fully defined by the parametrisation $h(y|\boldsymbol{x})$ and F_Z. The latter is specified using the `todistr` argument. The transformation function $h(y|\boldsymbol{x})$ therefore depends on the settings for the arguments `interacting` and `shifting`. The shift term is positive by default. The `response` argument (first argument of `ctm()` function) requires the Bernstein polynomial of the response variable as input (here: `basis_y`). `basis_x` is the right hand side of the model formula and defines the basis function for the shift term in the classical formula language. Note that the actual observations are not referenced during the specification of the model. As a result, the model estimation follows by applying the `mlt()` function from `mlt` package (Hothorn, 2016b) to the `ctm_coxph_mod` object of class `ctm`. The resulting object `mlt_coxph_mod` is from the `mlt` class. It contains the following objects itself specified by the `mlt_coxph_mod` environment:

- `bounds`
- `call`
- `coef`
- `convergence`
- `cpar`
- `data`
- `df`
- `feval`
- `fn.reduction`
- `gradient`
- `hessian`
- `iter`
- `loglik`
- `message`
- `model`
- `offset`
- `optimfct`
- `par`
- `parm`
- `quiet`
- `response`
- `scale`
- `score`
- `theta`
- `todistr`
- `trace`
- `value`
- `weights.`

Moreover, the following methods are available for objects of class `mlt`:

- `bounds`
- `coef`
- `coef<-`
- `confband`
- `Gradient`
- `Hessian`
- `logLik`
- `mkgrid`
- `plot`
- `predict`
- `print`
- `simulate`
- `summary`
- `variable.names`
- `vcov`
- `weights.`

The result of applying the function `coef()` to the `mlt_coxph_mod` object returns the original model parameter vector $\hat{\vartheta}_N$ (cf. **R**-Code 2.2).

```
coef(mlt_coxph_mod)

##        Bs1(y)        Bs2(y)        Bs3(y)        Bs4(y)        Bs5(y)
## -7.6072969832 -1.0350803074 -1.0350803074 -1.0350803074 -0.9545167564
##        Bs6(y)        Bs7(y)        Bs8(y)        Bs9(y)       Bs10(y)
## -0.3439678514 -0.3439678514 -0.3439678514 -0.1845162887  0.2937487631
##       Bs11(y)      horThyes           age  menostatPost         tsize
##  0.2937487631 -0.3490523654 -0.0099262360  0.2676696591  0.0077713879
##      tgrade.L      tgrade.Q        pnodes       progrec        estrec
##  0.5600910540 -0.2018493613  0.0487467451 -0.0022101686  0.0001833764
```

R-Code 2.2: Model parameter vector $\hat{\vartheta}_N$ of original transformation model $(F_Z, (\boldsymbol{a}_{\text{Bs},10}(y)^\top, \boldsymbol{b}(\boldsymbol{x})^\top)^\top, \hat{\boldsymbol{\vartheta}}_N)$ presented as original **R**-Code output

The output of **R**-Code 2.2 can be distinguished between model parameters that correspond to the Bernstein polynomial (beginning with `Bs`) and model parameters that correspond to the model covariables (`horTHyes`, `age`, ..., `estrec`). We have set the order of the Bernstein polynomial to $M = 10$ hence we get $M + 1 = 11$ coefficients that correspond to the Bernstein polynomial including the intercept of the Bernstein polynomial. Note that some Bernstein polynomial coefficients are equal to others (`Bs2(y)=Bs3(y)=Bs4(y)`, `Bs6(y)=Bs7(y)=Bs8(y)`, `Bs10(y)=Bs11(y)`. We will refer back to that characteristic and analyse the consequences of it in Section 3.1.2.

2.3.3. Conditional Transformation Models with Multiple Basis Functions

The conditional transformation model with multiple basis functions can be interpreted as an extension of the linear transformation model. Since the transformation function $h(y|\boldsymbol{x})$ depends simultaneously on y and $\boldsymbol{X}$, the model complexity of conditional transformation model is higher.

Hothorn *et al.* (2014) define the transformation function $h(y|\boldsymbol{x})$ as an additive decomposition of J partial transformation functions. Models of this class $(\cdot, \boldsymbol{c}, \boldsymbol{\vartheta})$ are called conditional transformation models (CTMs) and can be written in the following way:

$$\begin{aligned}\mathbb{P}(Y \leq y|\boldsymbol{X} = \boldsymbol{X}) &= \mathbb{P}\big(h(Y|\boldsymbol{X}) \leq h(y|\boldsymbol{x})\big) \\ &= F\big(h(y|\boldsymbol{x})\big) = F\big(\sum_{j=1}^{J} h_j(y|\boldsymbol{x})\big)\end{aligned}$$

The functions $h_j(y|x) : \mathbb{R} \to \mathbb{R}$ have to be monotonically increasing in y. The additive decomposition of the partial transformation functions $h_j(y|\boldsymbol{x})$ can be understood as a parametrisation of multiple basis functions $\boldsymbol{a}_j(y), \boldsymbol{b}_j(\boldsymbol{x}), j = 1, \ldots J$ via the joint basis

$$\boldsymbol{c} = (\boldsymbol{a}_1^\top \otimes \boldsymbol{b}_1^\top, \ldots, \boldsymbol{a}_J^\top \otimes \boldsymbol{b}_J^\top)^\top$$

Hothorn *et al.* (2014) proposed a boosting algorithm for the estimation of transformation functions h for exact continuous responses Y. As mentioned in Hothorn *et al.* (2015), in the likelihood framework conditional transformation models can be fitted under arbitrary schemes of censoring and truncation and classical likelihood

inference for the model parameters $\boldsymbol{\vartheta}$ becomes feasible. In contrast to the boosting algorithm, in the likelihood framework the number of model terms J and their complexity is limited because the likelihood does not contain any penalty terms inducing smoothness in the x-direction (Hothorn *et al.*, 2015). A more detailed overview of the class of conditional transformation models can be found in Möst (2014).

The ability to display complex relationships between the explanatory variables and the response is a great benefit of the CTMs. In addition, we have the advantage of CTMs in that, all parameters of the distribution function F_Y respective to F_Z may depend on the explanatory variables $\boldsymbol{X} \in \mathcal{X}$. The disadvantage of CTMs however, is the challenging model interpretation because of the high flexibility of the models. Möst (2014) also points out that the lack of orthogonality of the model components in CTMs constricts insights into model structure due to the fact that the model components are not separable.

2.4. The Bootstrap Resampling Method

The bootstrap resampling method was first mentioned by Efron (1979). He posits that this method is an appropriate "technique for making certain kinds of statistical inferences" (Efron and Tibshirani, 1993). Within the last decade, due to the increased efficiency of computing power and reduced cost, the method has become popular. Within this study, the differences between the parametric and the non-parametric bootstrap methods are elucidated and the assumptions about independent and identically distributed (iid) observations are valid for both cases. If these assumptions are not fulfilled, the bootstrap is misleading.

Throughout this thesis, the focus will be set on the parametric bootstrap approach (cf. Section 2.4.2), nevertheless, for the sake of completeness, we also introduce the non-parametric bootstrap procedure (cf. Section 2.4.1).

2.4.1. The Non-Parametric Bootstrap

The basic idea of the non-parametric bootstrap approach is to create additional data from the given observations. We normally compute an estimator $\hat{\vartheta}_N = g(Y_1, \dots Y_N)$ from the realisations $Y_1, \dots Y_N \overset{\text{iid}}{\sim} F$ (cf. Table 2.1). $\hat{\vartheta}_N$ is a known function g of the data $Y_1, \dots Y_N$ and with the help of the central limit theorem it is possible to estimate the asymptotic variance. Consequently, we obtain the asymptotic distribution of $\hat{\vartheta}_N$. In the framework of non-parametric bootstrap resampling technique with replacement, we draw many new data sets $Y_i^\star = (Y_1^\star, \dots, Y_N^\star) \overset{\text{iid}}{\sim} \hat{F}_N$ from the empirical distribution. On each of the original observed values $y_1, \dots, y_N$ we assign a probability of $1/N$ by the empirical distribution function. In other words, we draw a random sample of size n with replacement from the given observations. Based on these newly obtained observations $Y_1^\star, \dots, Y_N^\star$ the estimator $\hat{\vartheta}_N^\star$ can be computed. This process is repeated several times in order to obtain the approximate distribution of the simulated estimators $\hat{\vartheta}_N^\star$. Table 2.1 shows the outline of the non-parametric bootstrap and serves as a means to better understand the concept.

	Real world	**non-parametric** bootstrap world
distribution function	F	$\hat{F}_N$
data	$Y_1, \ldots, Y_N \overset{\text{iid}}{\sim} F$	$Y_1^\star, \ldots, Y_N^\star \overset{\text{iid}}{\sim} \hat{F}_N$
parameter	$\vartheta = g(F)$	$\hat{\vartheta}_N = g(\hat{F}_N)$
estimator	$\hat{\vartheta}_N = g(Y_1, \ldots, Y_N)$	$\hat{\vartheta}_N^\star = g(Y_1^\star, \ldots, Y_N^\star)$

Table 2.1.: Comparison of the real world (observed) and the world of the **non-**parametric bootstrap (Source: Own representation based on Geyer (2015))

The resulting distribution $\hat{F}_N$ is a step function (cf. Figure 2.2) since only the weight $1/N$ can be drawn from the observations. That is to say, staying at the realisations generates no new numbers.

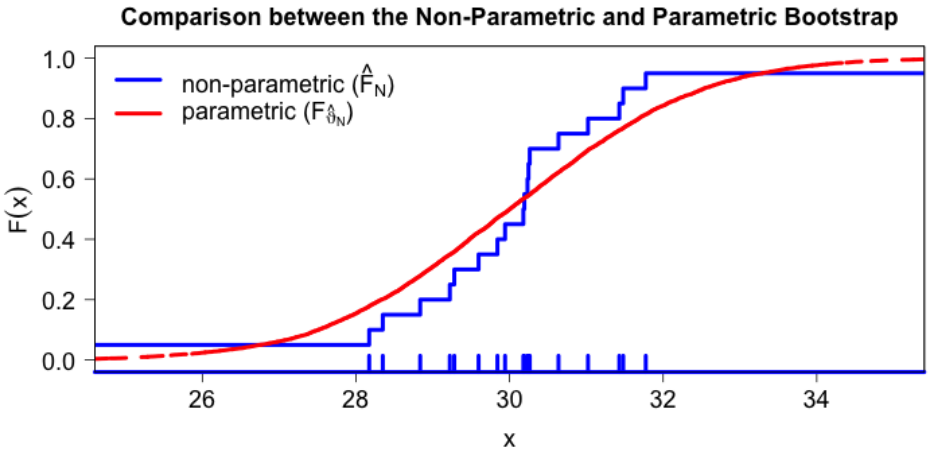

Figure 2.2.: Comparison between the non-parametric and parametric bootstrap: The resulting distribution $\hat{F}_N$ of the non-parametric bootstrap is a step function (blue), whereas the resulting distribution $F_{\hat{\vartheta}_N}$ of the parametric bootstrap is a smooth function (red).

2.4.2. The Parametric Bootstrap

The theory of the parametric bootstrap is similar to that of the non-parametric bootstrap. In contrast to the non-parametric bootstrap, the samples are now drawn from the estimated parametric distribution $F_{\hat{\vartheta}_N}$ instead of the empirical distribution $\hat{F}_N$. Efron and Tibshirani (1993) explain the parametric bootstrap approach as follows: Instead of estimating F by the empirical distribution function $\hat{F}_N$, $F_{\hat{\vartheta}_N}$ is estimated from a *parametric* model of the data. The ideal bootstrap estimate $\hat{\vartheta}_N^\star$ is then approximated by bootstrap sampling. Instead of sampling with replacement from the available data as in the non-parametric case, the bootstrap samples of size n are drawn from the estimated parametric distribution $F_{\hat{\vartheta}_N}$ of the population: $X_i^\star = (Y_1^\star, \ldots, Y_N^\star) \overset{\text{iid}}{\sim} F_{\hat{\vartheta}_N}$.

	Real world	**parametric** bootstrap world
distribution function	F	$F_{\hat{\vartheta}_N}$
data	$Y_1, \ldots, Y_N \overset{\text{iid}}{\sim} F$	$Y_1^\star, \ldots, Y_N^\star \overset{\text{iid}}{\sim} F_{\hat{\vartheta}_N}$
parameter	$\vartheta = g(F)$	$\hat{\vartheta}_N = g(F_{\hat{\vartheta}_N})$
estimator	$\hat{\vartheta}_N = g(Y_1, \ldots, Y_N)$	$\hat{\vartheta}_N^\star = g(Y_1^\star, \ldots, Y_N^\star)$

Table 2.2.: Comparison of the real world (observed) and the world of the parametric bootstrap (Source: Own representation based on Geyer (2015))

Table 2.2 shows the outline of the parametric bootstrap and serves as a support for a better understanding of the concept. The resulting distribution $F_{\hat{\vartheta}_N}$ is a smooth function (cf. Figure 2.2) as the bootstrap samples are drawn from the whole distribution function $F_{\hat{\vartheta}_N}$ in contrast to the empirical distribution function $\hat{F}_N$ in the non-parametric bootstrap (cf. Section 2.4.1, Figure 2.2).

For the sake of completeness, caution must be exercised whenever the underlying parametric model is wrong, as the application of the parametric bootstrap resampling will then also lead to wrong results. The bootstrap method should therefore not be used in such instances. Chernick and LaBudde (2014) suggest the need to compare both the non-parametric and the parametric bootstrap in order to review the parametric assumptions. In their opinion, the parametric bootstrap is especially essential when the parametric distribution is difficult to derive or as Good (2001) argues, the parametric bootstrap provides more accurate answers than textbook formulas.

2.5. The Data Set

The focus of this thesis is on different model approaches as well as the subsequent inference of these models. In other words, the data set used to estimate such models is not of primary interest as we do not want to analyse something in relation to the data more over we want to make inference about the fitted models. Nevertheless, for the future model interpretation and understanding it helps to have a background knowledge of the used data set.

2.5.1. German Breast Cancer Study Group-2 (GBSG2) Trials

The **G**erman **B**reast **C**ancer **S**tudy **G**roup-2 (GBSG2) Trial data set contains 686 (female) patients. Only patients not older than 65 years, who have tested positive for regional lymph nodes but lack distant metastases were included in the study. The data set was collected between July 1984 and December 1989 and is publicly available in **R**, where it can be downloaded through the `TH.data` package (Hothorn, 2015) and the command `data("GBSG2", package = "TH.data")` . The following continuous and factor variables are included in the data set:

- hormonal therapy (factor, 2 levels: yes, no)
- age of the patients in years (numerical)

- menopausal status (factor, 2 levels: premenopausal, postmenopausal)
- tumour size in mm (numerical)
- tumour grade (ordered factor, 3 levels: I $<$ II $<$ III)
- number of positive nodes (numerical)
- progesterone receptor in fmol (numerical)
- estrogen receptor in fmol (numerical)
- time in days describing the recurrence free survival (RFS) time (numerical)
- censoring indicator (factor, 2 levels: 0 censored, 1 event)

The survival or recurrence-free survival (RFS) time is the primary outcome variable. Out of 686 women, 246 received hormonal therapy whereas the control group of 440 women did not receive hormonal therapy. As stated in Sauerbrei *et al.* (1999), after a median follow-up time of nearly 5 years, 299 events for RFS and 171 deaths were observed. The statistical analysis is performed by fitting a Cox proportional hazards model with explicitly specified log cumulative baseline hazard function.

3. Modelling and Analysis

This chapter combines the two previously introduced concepts - the transformation model and the parametric bootstrap resampling method - with the aim being to implement, apply and evaluate the simulation-based inference for the transformation model. There are a plethora of reasons for doing so. First, the developed concept is applicable to all sizes of data sets due to the absence of any assumptions about the asymptotic behaviour of the estimators. As there are no asymptotic assumptions to be made, no data sets are considered to be too small for this approach. Second, the setting of a type I error rate α, - the probability of rejecting the null hypothesis given that it is true, - is not needed because the concept concentrates on graphical interpretation of the inference plots.

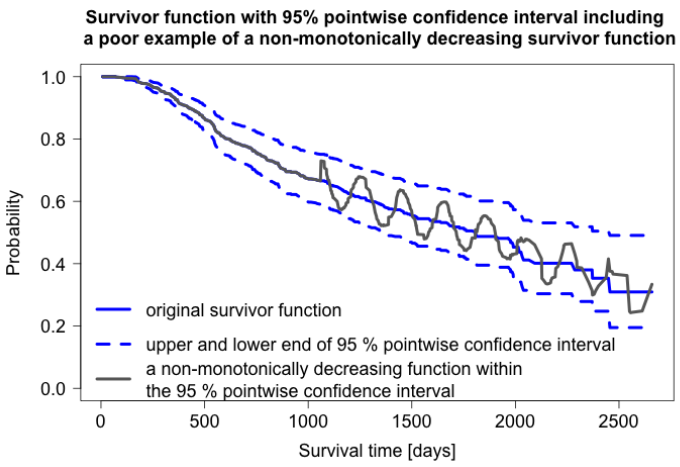

Figure 3.1.: Survivor function with 95 % pointwise confidence interval including a poor example of a non-monotonically decreasing survivor function

Third, the model-based parametric bootstrap analogue of the pointwise confidence interval only contains valid functions, *e.g.* all the model-based survivor functions in the inference plot are monotonically decreasing. This feature stands in contrast to the conventional pointwise confidence intervals that may also include not strictly monotone decreasing functions. The survivor function coloured in blue in Figure 3.1 demonstrates a poor example where the function is theoretically included in the 95 % pointwise confidence interval but is not a valid survivor function in itself since it is not monotonically decreasing. Lastly, the resulting simulation-based inference plots (shown and explained into more detail later) are easily accessible and have a straightforward interpretation.

The sections of this chapter are structured as follows: The implementation in **R** will serve as the basis for the subsequent inferences. We shall begin by explaining the effect of combining the parametric resampling method and the framework of transformation models (cf. Section 3.1). Following this, the subsequent statistical inference will be performed on the model parameters (cf. Section 3.2) as well as the data specific prediction functions, *e.g.* the density function, the empirical cumulative distribution function, the survivor function, etc. (cf. Section 3.3).

3.1. Parametric Bootstrap Resampling Method Applied to Conditional Transformation Models

3.1.1. Implementation

A conditional transformation model $(F_Z, \boldsymbol{c}(y,\boldsymbol{x}), \boldsymbol{\vartheta})$ fully defines the distribution of the response variable Y conditional on on the covariables $\boldsymbol{X}$ via $F_Y = F_Z\big(\boldsymbol{c}(y,\boldsymbol{x})^\top \boldsymbol{\vartheta}\big)$. The corresponding likelihood $\mathcal{L}(\boldsymbol{c}(y,\boldsymbol{x})^\top \boldsymbol{\vartheta} | Y \in (\underline{y}, \bar{y}])$ can directly be interpreted from the definition of the model, as introduced in Section 2.2. These characteristics enable the approach of applying the parametric bootstrap resampling method to transformation models.

Thanks to the knowledge of the conditional distribution function $F_Y = F_Z\big(\boldsymbol{c}(y,\boldsymbol{x})^\top \boldsymbol{\vartheta}\big)$, it is possible to draw new response variables $Y_1^\star, \ldots, Y_B^\star$ which are conditional on the given explanatory variables by applying the parametric bootstrap resampling method (cf. Section 2.4.2). With the information obtained, new transformation models are estimated and the newly obtained estimators are used for the subsequent model and parameter inference.

The following enumeration systematically explains in pseudo code how B new transformation models are generated by using the parametric bootstrap resampling method.

(1) Let $(F_Z, \boldsymbol{c}(y,\boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_N)$ denote the transformation model fitted to the original data set $\boldsymbol{x}_{1:N} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ with response y. F_Z defines the corresponding distribution function; $\boldsymbol{c}(y,\boldsymbol{x})$ is the joint basis that is used to transform Y conditional on $\boldsymbol{X}$; $\hat{\boldsymbol{\vartheta}}_N$ is the maximum likelihood estimator for a specific parametrisation of the transformation function. How to fit such a transformation model in **R** was explained with the help of **R**-Code 2.1 in Section 2.3.2. Furthermore, $(F_Z, \boldsymbol{c}(y,\boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_N)$ fully defines the distribution of the original response variable via $F_{Y|\boldsymbol{X}=\boldsymbol{x}} = F_Z\big(\boldsymbol{c}(y,\boldsymbol{x})^\top \hat{\boldsymbol{\vartheta}}_N\big)$ and the corresponding likelihood $\mathcal{L}(\boldsymbol{c}(y,\boldsymbol{x})^\top \hat{\boldsymbol{\vartheta}}_N | Y \in (\underline{y}, \bar{y}])$.

(2) Generate B parametric bootstrap samples $Y_1^\star, \ldots, Y_B^\star$:
`for (b in c(1:B)){` % B = number of bootstrap samples to be generated

- Generate N random variables $U_{1:N}$ from $\mathrm{U}[0,1]$, where N is the number of rows of the original data set.
- Use $u_1, \ldots, u_N$ and the parametric bootstrap resampling method to obtain additional response variables $y_{1,b}^*, \ldots, y_{N,b}^*$
  - Following from $(F_Z, \boldsymbol{c}(y,\boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_N)$, it is known:
    $Y \sim F_Y = F_Z\big(\boldsymbol{c}(y,\boldsymbol{x})^\top \hat{\boldsymbol{\vartheta}}_N\big) \Leftrightarrow F_Y(y) = F_Z\big(\boldsymbol{c}(y,\boldsymbol{x})^\top \hat{\boldsymbol{\vartheta}}_N\big) \in [0,1]$
  - $Y_{i,b}^* \sim F_Y^{-1}(u_i | \boldsymbol{x}_i)$ while keeping $\boldsymbol{x}_i$ fix for a given u_i, $i = 1, \ldots N$
    % This approach is also known as *probability integral transformation*.

`}`

(3) Executing the `for`-loop B times, the results are B new data sets each consisting of the original explanatory variables $\boldsymbol{x}_{1:N} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ and the newly generated bootstrap samples $y_{1,b}, \ldots, y_{N,b}$, $b = 1, \ldots, B$ as the response variables.

(4) The last step of the procedure is to fit B transformation models to the B newly generated data sets. It is executed by another for-loop:

`for (b in c(1:B)){` % B = number of bootstrap samples / generated data sets

- $Y^*_{i,b} \sim F_{Y^*_b} = F_Z \circ \boldsymbol{c}(y, \boldsymbol{x})^\top \hat{\boldsymbol{\vartheta}}^\star_b, i = 1, \ldots N \Rightarrow$ Transformation model $(F_Z, \boldsymbol{c}(y, \boldsymbol{x}), \hat{\boldsymbol{\vartheta}}^\star_b)$

}

$\Rightarrow$ The goal of generating B parametric bootstrap simulation-based transformation models $(F_Z, \boldsymbol{c}(y, \boldsymbol{x}), \hat{\boldsymbol{\vartheta}}^\star_b)$ with $b = 1, \ldots, B$ is fulfilled.

The implementation and execution of this pseudo code explicitly defines the B different bootstrap model parameter vectors $\hat{\boldsymbol{\vartheta}}^\star_1, \ldots, \hat{\boldsymbol{\vartheta}}^\star_B$ in addition to the original model parameter vector $\hat{\boldsymbol{\vartheta}}_N$.

The **R**-Code 3.1 explains how to implement the parametric bootstrap resampling method to generate B new transformation models in the **R** software environment. The information of the original transformation model $(F_Z, \boldsymbol{c}(y, \boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_N)$ is saved in the variable `mlt_coxph_mod`. The second `for`-loop (see pseudo code Section **(2)** above) which we use to generate the B parametric bootstrap samples $Y^\star_1, \ldots, Y^\star_B$ is here in the **R**-Code shortened on account of the function `simulate`. The `simulate()` function (**R** Development Core Team, 2009) simulates responses from the distribution corresponding to a fitted model object (here: `mlt_coxph_mod`). The argument `nsim` equals the number of response vectors to be simulated (here: `nsim` = `n_sim` = B = 1000).

```
# Define copy of GBSG2 data set for simulation
GBSG2_sim <- GBSG2
# Simulate "n_sim" responses from the distribution corresponding to a fitted model object.
y_sim <- simulate(mlt_coxph_mod, nsim = n_sim, seed = 880906)
# Prepare list to save parametric bootstrap results
mlt_coxph_mod_summary <- vector("list", n_sim)
# Refit model to the new simulated data sets
for (i in 1:n_sim){
   # overwrite the original DEXfat_y with the simulated ones
   GBSG2_sim$y <- y_sim[[i]]
   # refit/-estimate model to the simulated responses
   mlt_coxph_mod_summary[[i]] <- mlt(ctm_coxph_mod, data = GBSG2_sim,
                                     scale = TRUE, checkGrad = FALSE)
}
# Save objects as dataset
setwd(path_saved_R_objects)
save(mlt_coxph_mod_summary,
     file = paste("mlt_coxph_mod_summary_", n_sim, ".RData", sep=""))
```

R-Code 3.1: How to apply the parametric bootstrap resampling method for the estimation of B transformation models in **R**

The `for`-loop in the **R**-Code 3.1 can be linked to the `for`-loop mentioned in the pseudo code Section **(4)** above. Here, the fitted B transformation models obtained from the B newly generated data sets are **R** internally saved in the list object named `mlt_coxph_mod_summary` and **R** externally as a `Rdata` file. The latter prevents us from being forced to run the `for`-loop again and again.

3.1.2. Likelihood Based Inference Measures

The subsequent parametric bootstrap inference is based on a likelihood approach. This is possible since all model parameter vectors $\hat{\boldsymbol{\vartheta}}_N, \hat{\boldsymbol{\vartheta}}^\star_1, \hat{\boldsymbol{\vartheta}}^\star_2, \ldots, \hat{\boldsymbol{\vartheta}}^\star_B$ are maximum likelihood estimators for a specific parametrisation

of the transformation function (cf. Section 2.2.1). The generic `logLik()` function from the `stats` package in **R** can be used to extract the log-likelihood value from a **R** object of class `mlt`. It is important to note that the log-likelihood $l(\hat{\boldsymbol{\vartheta}})$ given a parameter vector $\hat{\boldsymbol{\vartheta}}$ is bounded by $-\infty$ and 0 ($-\infty < l(\hat{\boldsymbol{\vartheta}}) < 0$). The likelihood $\mathcal{L}(\hat{\boldsymbol{\vartheta}})$ given a parameter vector $\hat{\boldsymbol{\vartheta}}$ is also bounded by both the 0 and 1 ($0 < \mathcal{L}(\hat{\boldsymbol{\vartheta}}) < 1$) and the evaluated log-likelihood function of the maximum likelihood estimated parameter vector $\hat{\boldsymbol{\vartheta}}_N$ is equal to 0. The relative log-likelihood quantifies the relative probabilities of other parameter vectors, *e.g.* $\hat{\boldsymbol{\vartheta}}_b^\star$, in comparison to the maximum likelihood estimated parameter vector $\hat{\boldsymbol{\vartheta}}_N$. In the context used here, the relative log-likelihood is defined as:

$$\sum_{i=1}^{N} \log(\mathcal{L}(\boldsymbol{a}^\top \hat{\boldsymbol{\vartheta}}_b^\star | Y_i^\star) - \sum_{i=1}^{N} \log(\mathcal{L}(\boldsymbol{a}^\top \hat{\boldsymbol{\vartheta}}_N | Y_i^\star) \text{ for } b = 1, \ldots, B$$

to calculate the relative log-likelihood for each of the B bootstrap generated transformation models. Adapted from the definition of Held and Bové (2013) (Chapter 2.1.2), the relative log-likelihood can also be written as:

Definition 5 (Relative (Log-)Likelihood)**.**

- *Relative Likelihood:* $\tilde{\mathcal{L}}(\hat{\boldsymbol{\vartheta}}) = \mathcal{L}(\hat{\boldsymbol{\vartheta}}_b^\star) / \mathcal{L}(\hat{\boldsymbol{\vartheta}}_N), b = 1, \ldots, B$
- *Relative Log-Likelihood:* $\tilde{l}(\hat{\boldsymbol{\vartheta}}) = \log(\tilde{\mathcal{L}}(\hat{\boldsymbol{\vartheta}})) = l(\hat{\boldsymbol{\vartheta}}_b^\star) - l(\hat{\boldsymbol{\vartheta}}_N), b = 1, \ldots, B$

Additionally, let $-2\big(l(\hat{\boldsymbol{\vartheta}}_b^\star) - l(\hat{\boldsymbol{\vartheta}}_N)\big) = -2\tilde{l}(\hat{\boldsymbol{\vartheta}}) \in \mathbb{R}_{+0}$ denote the log-likelihood ratio statistic (LLRS). In general, the LLRS is the test statistic of the likelihood ratio test to compare the goodness of fit between two nested models. It is shown in Chapter 5.4.4 of Held and Bové (2013) that the LLRS asymptotically follows a Chi-squared distribution with k degrees of freedom where k is equal to the difference in the number of parameters between the two nested models:

$$-2\big(l_{\text{alt}} - l_0\big) \overset{\text{a}}{\sim} \chi^2_{\text{k}}\,.$$

The alternative model (the more complex model) can be transformed into the null model (the simpler model) by imposing a set of constraints on the parameters. The more complex model will always fit the data at least as well as the null model hence the alternative model has a greater or equal log-likelihood than the null model with less parameters. However, this likelihood ratio test is not applicable to our case since the bootstrap generated models $(F_Z, \boldsymbol{c}(y, \boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_b^\star)_{b=1,\ldots,B}$ and the original model $(F_Z, \boldsymbol{c}(y, \boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_N)$ are not hierarchically nested models. In other words, all the investigated models imply the same covariates and therefore the dimensionality of the parameter vectors is the same. Nevertheless, the LLRS can be calculated as follows:

$$-2\big(l(\hat{\boldsymbol{\vartheta}}_b^\star) - l(\hat{\boldsymbol{\vartheta}}_N)\big) = -2\tilde{l}(\hat{\boldsymbol{\vartheta}}) \overset{\text{a}}{\sim} \chi^2_{\text{p}}, \tag{3.1}$$

but cannot be interpreted as in the sense of the likelihood ratio test. Here, the degree of freedom p of the Chi-squared distribution is expected to be equal to the dimension of the parameter vectors $\boldsymbol{\vartheta}_b^\star$ and $\hat{\boldsymbol{\vartheta}}_N$.

The distribution of the computed LLRS for the investigated models is visualized in the histogram of Figure 3.2. The two plots shown differ in the number of drawn bootstrap samples: $B_1 = 1000$ (left panel) and $B_2 = 2000$ (right panel). This is done to minimize a potential approximation error.

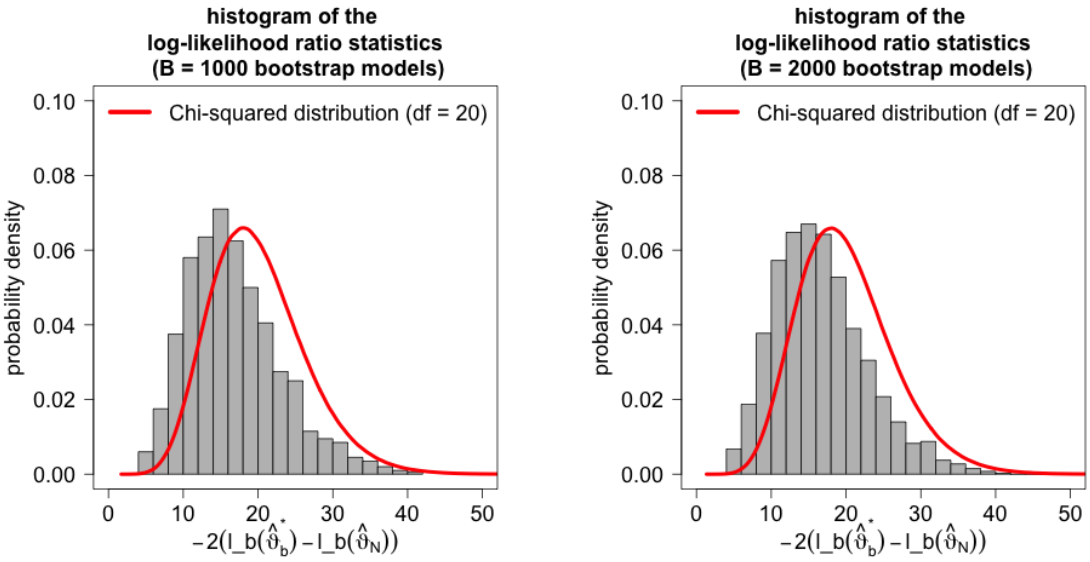

Figure 3.2.: Histogram of the $B_1 = 1000$ (left panel) and $B_2 = 2000$ (right panel) log-likelihood ratio statistics in comparison to the probability density functions of the Chi-squared distribution with degree of freedom = 20 (red line).

The LLRS is by definition always positive hence the x-axes of both histograms in Figure 3.2 only display positive numbers. The original as well as the bootstrap generated models are characterised by a parameter vector of dimension 20, *i.e.* $\hat{\boldsymbol{\vartheta}}_N, \hat{\boldsymbol{\vartheta}}_1^\star, \hat{\boldsymbol{\vartheta}}_2^\star, \ldots, \hat{\boldsymbol{\vartheta}}_B^\star \in \mathbb{R}^{20}$, consequently the LLRS should asymptotically follow a Chi-squared distribution with a degree of freedom (df) equal to 20 ($\chi^2_{\mathrm{p}} = \chi^2_{20}$, cf. Equation 3.1). However, by looking at Figure 3.2, it is obvious that the expected Chi-squared distribution with df = 20 (red line) does not fit the distribution of the LLRS (visualized as histograms). Under those circumstances, we believe that the deviation of the probability density function and the histogram can be described in connection with the multiply occurring Bernstein polynomial coefficients (cf. Section 2.2.3 and 2.3.2).

Table 3.1 represents the coefficients of the model parameter vector $\hat{\vartheta}_N$ of the original transformation model $(F_Z, \boldsymbol{c}(y, \boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_N)$. All duplicated coefficients are highlighted in **boldface**. There are 3 $(-1.0351, -0.344, 0.2937)$ coefficients that are not uniquely estimated among the Bernstein

(Detailed) Name of MLT Model Coefficient	Coefficient value
(1) Bernstein polynomial coefficient 1	-7.6073
(2) Bernstein polynomial coefficient 2	**-1.0351**
(3) Bernstein polynomial coefficient 3	**-1.0351**
(4) Bernstein polynomial coefficient 4	**-1.0351**
(5) Bernstein polynomial coefficient 5	-0.9545
(6) Bernstein polynomial coefficient 6	**-0.344**
(7) Bernstein polynomial coefficient 7	**-0.344**
(8) Bernstein polynomial coefficient 8	**-0.344**
(9) Bernstein polynomial coefficient 9	-0.1845
(10) Bernstein polynomial coefficient 10	**0.2937**
(11) Bernstein polynomial coefficient 11	**0.2937**
(12) hormonal therapy: yes	-0.3491
(13) age [years]	-0.0099
(14) menopausal status: post	0.2677
(15) tumour size [mm]	0.0078
(16) tumour grade: II	0.5601
(17) tumour grade: III	-0.2018
(18) # positive nodes	0.0487
(19) progesterone receptor [fmol]	-0.0022
(20) estrogen receptor [fmol]	2e-04

Table 3.1.: Rounded coefficients (4 digits) of the model parameter vector $\hat{\vartheta}_N$ of the original transformation model $(F_Z, \boldsymbol{c}(y, \boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_N)$ presented in table format

polynomial coefficients. Consequently, there is the assumption that the existence of such multiply occurring coefficients restricts the parameter space $\Theta = \{\boldsymbol{\vartheta} \in \mathbb{R}^P | \boldsymbol{a}^\top \boldsymbol{\vartheta} \in \mathcal{H}\}$ and a subsequent correction for the degree of freedom of the Chi-squared distributed LLRS seems essential.

The subsequent Figure 3.3 also supports this presumption regarding the essential correction for the degrees of freedom of the Chi-squared distribution. Figure 3.3 visualizes the relative frequency of the LLRS in histograms. Additionally, there are several probability density functions of Chi-squared distributions with different degrees of freedom added to the plots. And again, to minimize potential approximation errors, the amount of bootstrap samples were increased from $B_1 = 1000$ to $B_2 = 2000$ (cf. Figure 3.3, from left to right panel). However, the relative frequency displayed by the histograms in Figure 3.3 does not seem to differ much between 1000 and 2000 bootstrap samples.

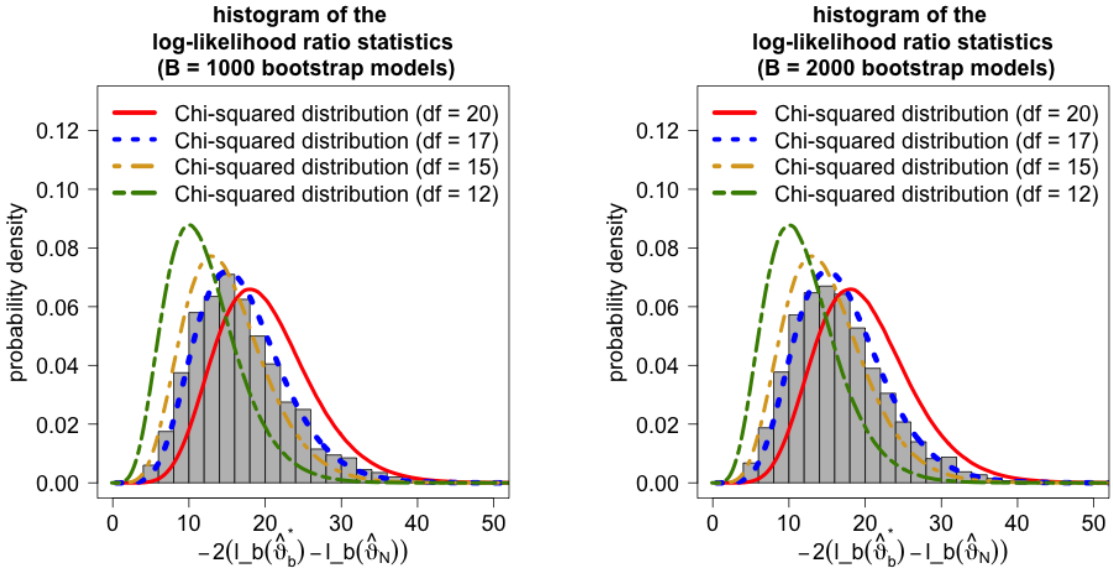

Figure 3.3.: Histogram of the $B_1 = 1000$ and $B_2 = 2000$ log-likelihood ratio statistics in comparison to several probability density functions of the Chi-squared distribution with different degrees of freedom

The following bullet points justify the possible corrections of the degrees of freedom of the Chi-squared distributed probability density functions of Figure 3.3. All the investigated corrections have a negative sign so that the degree of freedom becomes smaller after correction. This is due to the assumption that the multiply occurring Bernstein polynomial coefficients restrict the parameter space, consequently, the "corrected" degree of freedom of the Chi-squared distribution of the LLRS has to be reduced.

- Chi-squared distribution with **df=20**: The (original) model parameter vector $\hat{\vartheta}_N \in \mathbb{R}^{20}$ has dimension 20 hence the distribution of the LLRS is - according to Equation 3.1 - expected to be χ^2_{20}. The probability density function of the Chi-squared distribution with df=20 can somewhat be understood as the baseline.

- Chi-squared distribution with **df=17**: The probability density function of the Chi-squared distribution with df=17 was corrected by -3 compared to the baseline. The correction of -3 stems from the 3 multiply occurring Bernstein polynomial coefficients (-1.0351, -0.344, 0.2937, cf. Table 3.1).

- Chi-squared distribution with **df=15**: The probability density function of the Chi-squared distribution with df=15 was corrected by -5 compared to the baseline. The correction of -5 stems from the amount of repeated Bernstein polynomial coefficients (`Bs3(y)`, `Bs4(y)`, `Bs7(y)`, `Bs8(y)`, `Bs11(y)`, cf. Table 3.1).
- Chi-squared distribution with **df=12**: The probability density function of the Chi-squared distribution with df=12 was corrected by -8 compared to the baseline. The correction of -8 stems from reducing all the multiply occurring Bernstein polynomial coefficients (`Bs2(y)`, `Bs3(y)`, `Bs4(y)`, `Bs6(y)`, `Bs7(y)`, `Bs8(y)`, `Bs10(y)`, `Bs11(y)`, cf. Table 3.1).

Considering Figure 3.3, the probability density function of the Chi-squared distribution with df=17 (drawn in blue) corresponds best to the relative frequency of the LLRS visualized in the histogram. However, this does not give evidence that the suggested correction of subtracting the amount of multiply occurring Bernstein polynomial coefficients is correct.

To investigate further the possible corrections of the degrees of freedom, the original model $(F_Z, \boldsymbol{c}(y, \boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_N)$ fitted to the GBSG2 data and the bootstrap generated models $(F_Z, \boldsymbol{c}(y, \boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_b^\star)_{b=1,\dots,B}$ were again estimated for different orders of the Bernstein polynomial basis function $\boldsymbol{a}(y) = \boldsymbol{a}_{\text{Bs},i}(y)$ with $i = 1, 2, \dots, 15$. Table 3.2 summarises parameter specific information of the refitted models.

model name	order of Bernstein polynomial	AIC	total amount of parameters	amount of model parameters corresponding to Bernstein polynomial	amount of model parameters corresponding to model covariables	total amount of duplicated parameters	amount of multiply occurring model parameters
MLT model 1	1	5425	11	2	9	0	0
MLT model 2	**2**	**5299**	**12**	**3**	**9**	**0**	**0**
MLT model 3	**3**	**5242**	**13**	**4**	**9**	**1**	**1**
MLT model 4	4	5201	14	5	9	2	1
MLT model 5	5	5180	15	6	9	1	1
MLT model 6	6	5168	16	7	9	3	1
MLT model 7	**7**	**5162**	**17**	**8**	**9**	**3**	**3**
MLT model 8	**8**	**5159**	**18**	**9**	**9**	**3**	**2**
MLT model 9	9	5158	19	10	9	4	2
MLT model 10	10	5158	20	11	9	5	3
MLT model 11	11	5160	21	12	9	4	2
MLT model 12	**12**	**5161**	**22**	**13**	**9**	**5**	**4**
MLT model 13	**13**	**5163**	**23**	**14**	**9**	**7**	**5**
MLT model 14	14	5165	24	15	9	6	4
MLT model 15	15	5167	25	16	9	7	5

Table 3.2.: Overview from the results of the simulation study where the original transformation model $(F_Z, (\boldsymbol{a}(y)^\top, \boldsymbol{b}(\boldsymbol{x})^\top)^\top, \hat{\boldsymbol{\vartheta}}_N)$ was refitted to the GBSG2 data by using different Bernstein polynomials as basis function $\boldsymbol{a}_{\text{Bs},i}(y), i = 1, 2, \dots, 15$. The model names correspond to the order of the Bernstein polynomial used as basis function for the specific model. For the **boldface** printed rows, we additionally show the histogram of the LLRS as well as the probability density function of the Chi-squared distributions in the appendix (cf. Appendix A.2, Figure A.2).

The model names listed in the first column of Table 3.2 are intentionally defined to correspond with the order of the Bernstein polynomial $\boldsymbol{a}_{\text{Bs},i=1,\,2,\,\dots,\,15}(y)$ used as the basis function $\boldsymbol{a}(y)$. If the order of the polynomial is M, then the amount of model parameters that correspond to the Bernstein polynomial is $M + 1$ (cf. Table 3.2, columns 2 and 5). The amount of model parameters that correspond to the model covariables

is equal to 9 for all listed models since each model includes all covariables from the fitted GBSG2 data set (cf. Section 2.5.1). The last two columns of Table 3.2 are a bit more complex for the sake of an in depth analysis: The total amount of duplicated parameters determines how many parameters are duplicates of parameters with smaller subscripts. The amount of multiply occurring model parameters is gained by applying a unique command to the total amount of duplicated parameters. The amount of parameters that lead to the correction corresponding to the green density function in Figure 3.3 is not separately listed in Table 3.2 but equals the sum of the total amount of duplicated parameters and amount of multiply occurring model parameters.

As previously mentioned in the caption of Table 3.2, there are additional histograms of the LLRS as well as the probability density function of the Chi-squared distributions for the **boldface** printed rows of Table 3.2 to be found in the appendix (cf. Appendix A.2).

Now, to provide more evidence for the essentiality regarding the correction of the degrees of freedom of the Chi-squared distributed LLRS, let us additionally introduce another simulation study. This additional simulation study was based on randomly generated $\mathcal{N}(0,1)$ distributed response variables Y. An unconditional transformation model $(F_Z = \Phi, \boldsymbol{a}_{\text{Bs},i}, \hat{\boldsymbol{\vartheta}}_n)$ was fitted to the original data set, after which the parametric bootstrap resampling method was applied. The setup of the simulation study varied depending on the three parameters below:

`p_ord` = i = **order of Bernstein polynomial**	`n` = **# rows of data set**	`n_sim` = **# Bootstrap samples**
• 5	• 250	• 500
• 7	• 500	• 1000
• 10	• 1000	
	• 2000	

The combination of these parameters produced 24 additional simulated data sets. In summary, these models $(F_Z = \Phi, \boldsymbol{a}_{\text{Bs},i=5,\,7,\,10}, \hat{\boldsymbol{\vartheta}}_{n=250,\,500,\,1000,\,2000})$ particularly, those drawn with 500 bootstrap samples as well as 1000 were fitted.

	`Bs1(y)`	`Bs2(y)`	`Bs3(y)`	`Bs4(y)`	`Bs5(y)`	`Bs6(y)`
`n = 250`	-1.2259	-0.7501	-0.0701	0.3062	0.8068	1.2832
`n = 500`	-1.2871	-0.7545	-0.0577	0.1348	0.7311	1.2408
`n = 1'000`	-1.2872	-0.7112	-0.4598	0.4769	0.6868	1.2742
`n = 2'000`	-1.2667	-0.7952	-0.2047	0.2350	0.7768	1.3000

Table 3.3.: Parameter vectors of the unconditional transformation models $(F_Z = \Phi, \boldsymbol{a}_{\text{Bs},5}, \hat{\boldsymbol{\vartheta}}_{n=250,\,500,\,1000,\,2000})$. The row names of the table correspond to the amount of rows of the data set used to estimate the model.

Table 3.3 displays the parameter vectors of the unconditional transformation models $(F_Z = \Phi, \boldsymbol{a}_{\text{Bs},5}, \hat{\boldsymbol{\vartheta}}_{n=250,\,500,\,1000,\,2000})$ that were estimated during the simulation study. The order of the Bernstein polynomial has been set to 5, consequently the parameter vector is of dimension $5 + 1 = 6$. Since there are no multiply occurring coefficients, the degrees of freedom of the Chi-squared distributed LLRS is adequately estimated with df = $6 = 5 + 1$. This can also be determined by looking at Figure A.3 in the Appendix A.3.1. Figure A.3 additionally shows that 500 bootstrap generated samples (plots in top row of Figure A.3) are not sufficient for an asymptotic behaviour of the distribution of the LLRS.

	Bs1(y)	Bs2(y)	Bs3(y)	Bs4(y)	Bs5(y)	Bs6(y)	Bs7(y)	Bs8(y)
n = 250	-1.287	-1.0159	-0.4563	**0.0837**	**0.0837**	0.5461	0.8942	1.2696
n = 500	-1.2869	-0.8503	-0.3063	-0.1741	0.0695	0.8215	0.8717	1.28
n = 1'000	-1.2782	-0.9136	**-0.3471**	**-0.3471**	0.2385	0.4354	0.9136	1.2719
n = 2'000	-1.279	-0.8658	-0.3379	-0.1901	0.1755	0.7022	0.8435	1.2752

Table 3.4.: Parameter vectors of the unconditional transformation models $(F_Z = \Phi,\ \boldsymbol{a}_{\text{Bs},7},\ \hat{\boldsymbol{\vartheta}}_{n=250,\ 500,\ 1000,\ 2000})$. The coefficients occurring multiple times within the same model (within the same row) are highlighted in **boldface**. The row names of the table correspond to the amount of rows of the data set used to estimate the model.

Table 3.4 displays the parameter vectors of the unconditional transformation models $(F_Z = \Phi,\ \boldsymbol{a}_{\text{Bs},7},\ \hat{\boldsymbol{\vartheta}}_{n=250,\ 500,\ 1000,\ 2000})$ that were estimated during the simulation study. The order of the Bernstein polynomial has been set to 7, therefore the parameter vector is of dimension $7 + 1 = 8$, and the probability density function of the Chi-squared distributed LLRS is expected to be with degrees of freedom df = $7 + 1 = 8$. However, the two models $(F_Z = \Phi,\ \boldsymbol{a}_{\text{Bs},7},\ \hat{\boldsymbol{\vartheta}}_{n=250})$ and $(F_Z = \Phi,\ \boldsymbol{a}_{\text{Bs},7},\ \hat{\boldsymbol{\vartheta}}_{n=1000})$ have multiply occurring model parameters that correspond to the Bernstein polynomial. Hence, the probability density function of the Chi-squared distributed LLRS with degrees of freedom df = $7 + 1 = 8$, no longer adequately fits the histogram of the LLRS. This finding is also determined by looking at Figure A.4 in the Appendix A.3.2. Figure A.4 further indicates that the 500 bootstrap generated samples (plots in top row of Figure A.4) are not sufficient for an asymptotic behaviour of the distribution of the LLRS.

	Bs1(y)	Bs2(y)	Bs3(y)	Bs4(y)	Bs5(y)	Bs6(y)	Bs7(y)	Bs8(y)	Bs9(y)	Bs10(y)	Bs11(y)
n = 250	-1.2632	-1.0292	**-0.8304**	**-0.8304**	-0.1487	**0.2152**	**0.2152**	0.2231	**1.0303**	**1.0303**	1.2659
n = 500	-1.2896	-1.0458	-0.9468	**-0.4639**	**-0.4639**	**-0.4639**	**0.4849**	**0.4849**	0.6803	0.9986	1.2674
n = 1'000	-1.2789	-1.0296	-0.8171	**-0.4068**	**-0.4068**	-0.0442	**0.4457**	**0.4457**	0.7498	1.0563	1.2913
n = 2'000	-1.2864	-1.0427	**-0.6742**	**-0.6742**	-0.2004	**0.1521**	**0.1521**	0.4291	0.8946	1.031	1.273

Table 3.5.: Parameter vectors of the unconditional transformation models $(F_Z = \Phi,\ \boldsymbol{a}_{\text{Bs},10},\ \hat{\boldsymbol{\vartheta}}_{n=250,500,1000,2000})$. The coefficients occurring multiple times within the same model (within the same row) are highlighted in **boldface**. The row names of the table correspond to the amount of rows of the data set used to estimate the model.

Table 3.5 displays the parameter vectors of the unconditional transformation models $(F_Z = \Phi,\ \boldsymbol{a}_{\text{Bs},10},\ \hat{\boldsymbol{\vartheta}}_{n=250,\ 500,\ 1000,\ 2000})$ that were estimated during the simulation study. The order of the Bernstein polynomial has been set to 10, consequently the parameter vector is of dimension $10 + 1 = 11$ and the probability density function of the Chi-squared distributed LLRS is expected to be with degrees of freedom df = $10 + 1 = 11$. However, all of the models - $(F_Z = \Phi,\ \boldsymbol{a}_{\text{Bs},10},\ \hat{\boldsymbol{\vartheta}}_{n=250})$, $(F_Z = \Phi,\ \boldsymbol{a}_{\text{Bs},10},\ \hat{\boldsymbol{\vartheta}}_{n=500})$, $(F_Z = \Phi,\ \boldsymbol{a}_{\text{Bs},10},\ \hat{\boldsymbol{\vartheta}}_{n=1000})$ and $(F_Z = \Phi,\ \boldsymbol{a}_{\text{Bs},10},\ \hat{\boldsymbol{\vartheta}}_{n=2000})$ - have multiply occurring model parameters that correspond to the Bernstein polynomial; hence, the probability density function of the Chi-squared distributed LLRS with degrees of freedom df = $10 + 1 = 11$ no longer adequately fits the histogram of the LLRS. This finding is also determined by looking at Figure A.5 in the Appendix A.3.3. Figure A.5 additionally shows that 500 bootstrap generated samples (plots in top row of Figure A.5) are not sufficient for an asymptotic behaviour of the distribution of the LLRS.

3.1.2.1. Concluding remarks

The initial intention of this section was to briefly introduce the *Likelihood Based Inference Measures*, however, after delving deeper into the different measures it became clear that it is essential to elaborate on this concept. As a result, this research investigated further on the effect of restricting the degrees of freedom of the Chi-squared distributed LLRS. Now, although a definitive solution for the latter was not attained, in summary, this simulation study uncovered additional evidence to support the idea that the expected degrees of freedom of the Chi-squared distributed LLRS are violated in instances where the original model parameters imply multiply occurring coefficients. For this reason, the simulation study is considered to be expedient. Nontheless, this finding also indicates that the LLRS might not be the best likelihood based measure for the application of the parametric bootstrap inference of transformation models.

Thus, the focus of this thesis is shift back again to the parametric bootstrap inference of transformation models. Due to the aforementioned finding, the subsequent graphical inference only uses the relative log-likelihood (RLL) as a measure for defining the accuracy of the bootstrap estimated models in contrast to the original transformation model.

The focus of the subsequent statistical inference of the bootstrap generated transformation models is carried out in two ways: first, on the model parameters and the distribution thereof (cf. Section 3.2); and second, on the data specific prediction functions (cf. Section 3.3). For the latter case, the fully specified distribution function F_Y of the response variables $Y, Y_1^\star, \ldots, Y_B^\star$ is used as a basis for making inference about additional functions which can be derived from the distribution function, *e.g.* the density function, the survivor function, etc.

3.2. Parametric Bootstrap Inference for Parameters of Transformation Models

Let $(F_Z, \boldsymbol{c}(y, \boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_N)$ and $(F_Z, \boldsymbol{c}(y, \boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_b^\star)_{b=1,\ldots,B}$ be the transformation model of the original data set and the bootstrap transformation models of the B data sets generated from the parametric bootstrap, respectively. The model parameters $\hat{\boldsymbol{\vartheta}}_N, \hat{\boldsymbol{\vartheta}}_1^\star, \hat{\boldsymbol{\vartheta}}_2^\star, \ldots, \hat{\boldsymbol{\vartheta}}_B^\star$ are maximum likelihood estimators for a specific parametrisation of the transformation function h. Consequently, the standard likelihood procedures are applicable to the subsequent parameter inference. The bootstrap based model parameters $\hat{\boldsymbol{\vartheta}}_1^\star, \hat{\boldsymbol{\vartheta}}_2^\star, \ldots, \hat{\boldsymbol{\vartheta}}_B^\star$ are asymptotically multivariate normal distributed $N_p(\hat{\boldsymbol{\vartheta}}_N, \boldsymbol{I}(\hat{\boldsymbol{\vartheta}}_N)^{-1})$ with mean $\hat{\boldsymbol{\vartheta}}_N$ and covariance matrix equal to the inverse observed Fisher information. The subsequent simulation-based, *i.e.* bootstrap generated, parameter inference has an advantage in that, it is unnecessary to make assumptions about the asymptotic distribution of its parameters.

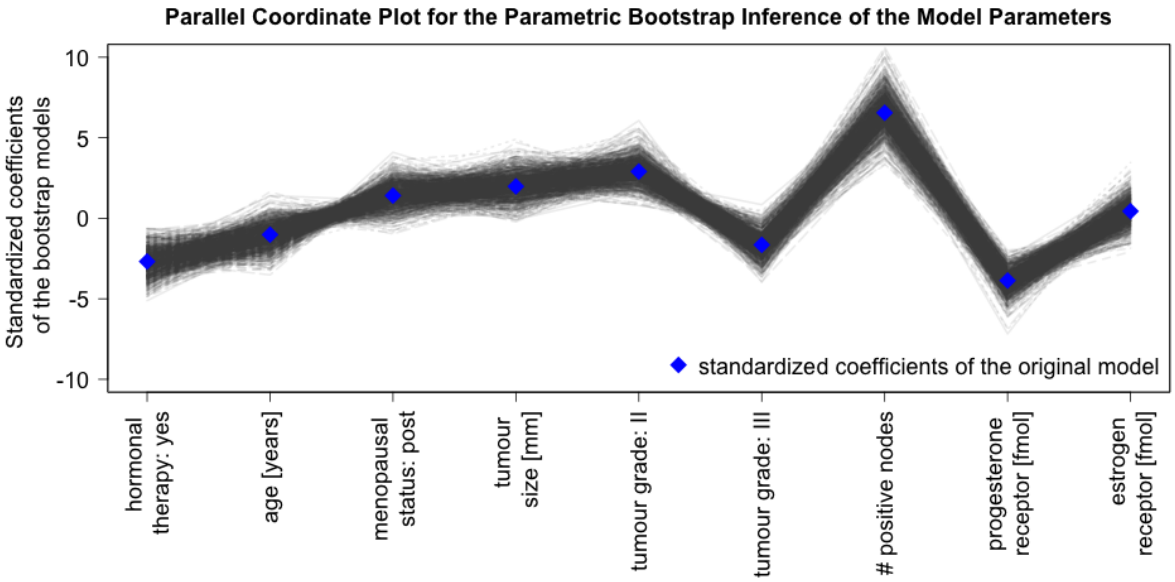

Figure 3.4.: Parallel coordinate plot of the $\hat{\vartheta}^{\star}_{b=1,\ldots,B}$ model coefficients from the $B = 1000$ bootstrap generated transformation models. The coefficients of the original model $\hat{\vartheta}_N$ serve as reference and are represented with blue dots.

Figure 3.4 shows a parallel coordinate plot of the standardized coefficients of the B bootstrap generated conditional transformation models. The blue dots refer to the coefficients $\hat{\vartheta}_N$ of the original model $(F_Z, \boldsymbol{c}(y, \boldsymbol{x}), \hat{\vartheta}_N)$. The parallel coordinate plot can be understood as showing the joint distribution of all p coefficients of the B bootstrap generated models $(\hat{\vartheta}^{\star}_{1,b}, \ldots, \hat{\vartheta}^{\star}_{p,b})_{b=1,\ldots,B}$ connected with single grey lines. Whenever the lines in the parallel coordinate plot between two coefficients (on the x-axis) are parallel, it can be interpreted as coefficients that positively correlate with one another. In contrast, the crossing lines highlight the negative correlation between the first few model coefficients in Figure 3.4. It is important to note that the asymptotically multivariate normal distribution of the parameters $(\hat{\vartheta}^{\star}_{1,b}, \ldots, \hat{\vartheta}^{\star}_{p,b})_{b=1,\ldots,B}$ around the original maximum likelihood estimator $\hat{\vartheta}_N$ is clearly visible in Figure 3.4. The asymptotic normality of the maximum likelihood estimator is one of the most important results of the likelihood theory, as thoroughly explained by Held and Bové (2013). The conventional likelihood framework is inapplicable to analyses where the underlying data set only consists of a few observations. This is because the assumptions regarding the asymptotic behaviour would not be fulfilled. On the contrary, the parametric bootstrap approach is also applicable to analyses of small data sets, as this approach does not require any assumptions regarding the asymptotic behaviour of the model parameters. In addition, there is no need to estimate the fisher information matrix as by applying the parametric bootstrap inference to model parameters, the asymptotic behaviour of the model parameters are in a sense finitely illustrated. Consequently, the plot summarizes this idea thereby requiring no previous assumptions.

Figure 3.4 serves as a nice overview of the parameter inference, however, it is difficult to interpret. By looking at Figure 3.4, it is impossible to distinguish between models that are similar to the original model and others that are not. For this reason, we use the previously introduced measure of the relative log-likelihood (RLL, cf. Section 3.1). We obtain B different values for the RLL based on the B estimated bootstrap models. The empirical cumulative distribution function (ECDF) of these B RLLs is plotted in the left panel of

Figure 3.5. The empirical cumulative distribution function included a cutoff line at probability 5 % that helped distinguishing between the extreme models ($\hat{F}_b(\text{RLL}) < 0.05$) and those with $\hat{F}_b(\text{RLL}) \geq 0.05$. The models with $\hat{F}_b(\text{RLL}) \geq 0.05$ are characterised by a RLL value close to zero, *i.e.* they are similar to the original model. The extreme models with $\hat{F}_b(\text{RLL}) < 0.05$ are plotted in the right panel of Figure 3.5.

The colour gradient for the right panel of Figure 3.5 and the left panel of Figure 3.6 is the same. The gradient is based on $\hat{F}_b(\text{RLL})$. The colouring starts just at the cutoff line at probability 5 % of the empirical cumulative distribution function: The darker the colour, the smaller the RLL and therefore, the less similar the bootstrap model is in comparison to the original transformation model. Based on this definition for the colour gradient, we can infer that the lines drawn close to the coefficients of the original model are brighter in comparison to the ones further away.

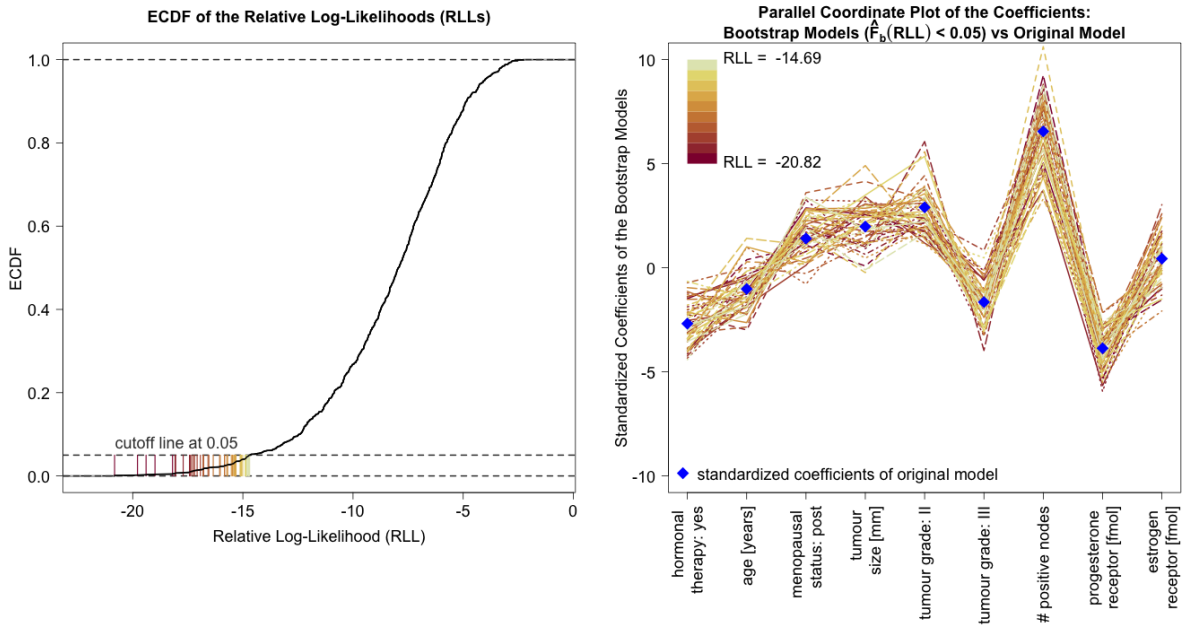

Figure 3.5.: Empirical cumulative distribution function of the relative log-likelihoods calculated by comparing the $B = 1000$ bootstrap generated transformation models $(F_Z, \boldsymbol{c}(y, \boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_b^\star)_{b=1,\ldots,B}$ versus the original transformation model $(F_Z, \boldsymbol{c}(y, \boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_N)$ of the original data set (left panel). The parallel coordinate plot of the coefficients (right panel) only shows the extreme ($\hat{F}_b(\text{RLL}) < 0.05$) bootstrap models.

Figure 3.6 shows the model parameters of all the B bootstrap generated transformation models. The bootstrap generated model with $\hat{F}_b(\text{RLL}) < 0.05$ are plotted in the left panel, whereas the bootstrap generated model with $\hat{F}_b(\text{RLL}) \geq 0.05$ are plotted in the right panel. Here, we have $B = 1000$ bootstrap samples, hence, the left plot shows 50 and the right plot shows 950 parallel coordinate lines.

Overall, it is striking that the plot in the right panel of Figure 3.6 still shows obvious deviation of the coefficients from the original model even though the extreme cases with $\hat{F}_b(\text{RLL}) < 0.05$ are removed. All the while, it is important to bear in mind that depending on the RLL, the peaks of the right panel are not as extreme as the ones visible in the left panel.

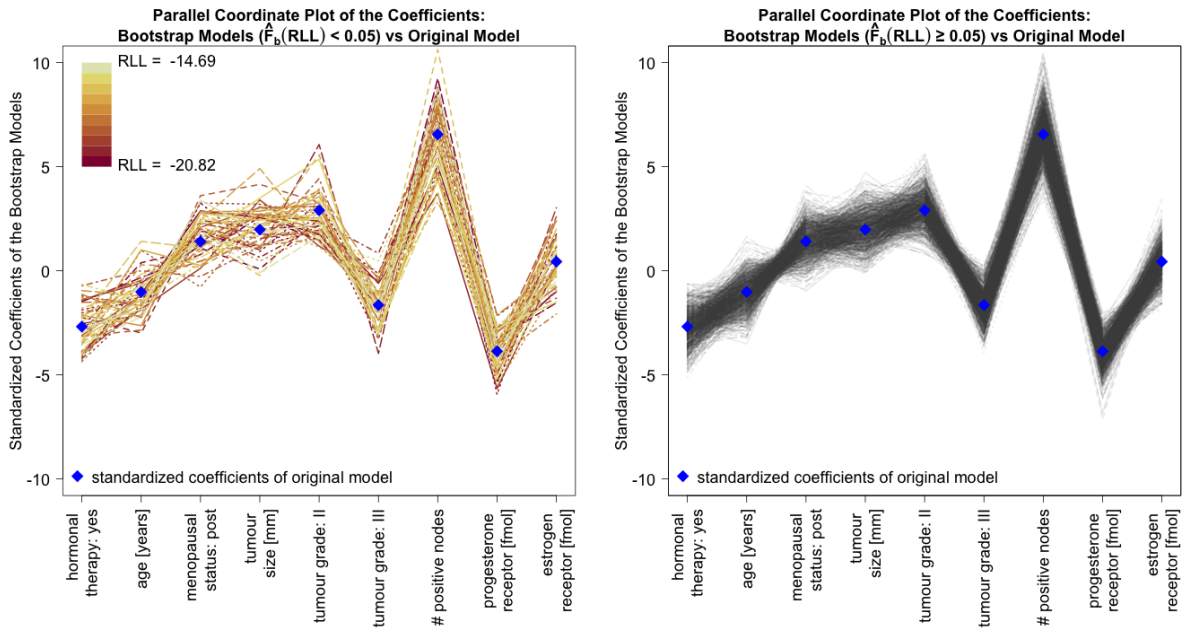

Figure 3.6.: Parallel coordinate plot of the bootstrap model coefficients versus the original model distinguished between $\hat{F}_b(\text{RLL}) < 0.05$ (left panel) and $\hat{F}_b(\text{RLL}) \geq 0.05$ (right panel)

3.3. Parametric Bootstrap Inference for Functions Obtained from the Conditional Distribution Function of the Transformation Models

Let $(F_Z, \boldsymbol{c}(y, \boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_N)$ and $(F_Z, \boldsymbol{c}(y, \boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_b^\star)_{b=1,\ldots,B}$ be the transformation model of the original data set and the bootstrap transformation models of the B data sets generated from the parametric bootstrap, respectively. Here, we shall compare the original conditional distribution function of the original responses Y to the conditional distribution functions of the B bootstrap generated $Y_1^\star, \ldots, Y_B^\star$. It can be recalled, that the underlying model is a Cox proportional hazard model in perspective of a conditional transformation model. Consequently, it can be viewed as a Cox proportional hazard model with an explicit specified log cumulative baseline hazard function. The primary outcome variable Y is the survival or recurrence-free survival (RFS) time. In other words, the analysis is based on a data set "for which the outcome variable of interest is time until an event occurs" (Kleinbaum and Klein, 1996). The following list puts the distribution function F_Y of the response variable Y obtained from a conditional transformation model $F_Z\left(\boldsymbol{c}(y, \boldsymbol{x})\boldsymbol{\vartheta}\right)$ into context with other important functions of the framework of survival analysis:

- **Distribution function**: $F_Y(y|\boldsymbol{x}) = F_Z\left(\boldsymbol{c}(y, \boldsymbol{x})\boldsymbol{\vartheta}\right)$
- **Survivor function**: $S_Y(y|\boldsymbol{x}) = 1 - F_Y(y) = 1 - F_Z\left(\boldsymbol{c}(y, \boldsymbol{x})\boldsymbol{\vartheta}\right)$
- **Density function**: $f_Y(y|\boldsymbol{x}) = \partial S_Y(y|\boldsymbol{x})/\partial y = \partial\left(1 - F_Y(y|\boldsymbol{x})\right)/\left(\partial y\right) = \partial\left(1 - F_Z\left(\boldsymbol{c}(y, \boldsymbol{x})\boldsymbol{\vartheta}\right)\right)/\left(\partial y\right)$
- **Hazard rate / function**:

$$\lambda_Y(y|\boldsymbol{x}) = f_Y(y|\boldsymbol{x})/\left(1 - F_Y(y|\boldsymbol{x})\right) = \left(\partial\left(1 - F_Y(y|\boldsymbol{x})\right)/\partial y\right)/\left(1 - F_Y(y|\boldsymbol{x})\right)$$
$$= \left(\frac{\partial\left(1 - F_Z\left(\boldsymbol{c}(y, \boldsymbol{x})\boldsymbol{\vartheta}\right)\right)}{\partial y}\right)/\left(1 - F_Z\left(\boldsymbol{c}(y, \boldsymbol{x})\boldsymbol{\vartheta}\right)\right)$$

The function in focus for the subsequent analysis is the *survivor function*, also known as *survival function*. Here, this function captures the probability of survival or recurrence-free survival beyond a specified time. In the subsequent paragraphs, a description follows of how to obtain the survivor functions when given the estimated transformation models $(F_Z, \boldsymbol{c}(y, \boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_N)$, $(F_Z, \boldsymbol{c}(y, \boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_b^\star)$ with $b = 1, \ldots, B$ and the conditional distribution functions of the response variables $Y_N, Y_1^\star, Y_2^\star, \ldots, Y_B^\star$ along with the likelihood functions of the models.

In order to predict the survivor function, a hypothetical observation, *i.e.* a hypothetical patient from the GBSG2 data set (cf. Section 2.5.1), needs to be defined. Later, this hypothetical patient is used as baseline for the estimated functions. In the appendix (cf. Appendix A.4), there is the **R**-code that explains how this hypothetical observation is obtained from the original data set in a step-by-step manner. The specification differs for numerical and factorial covariates. Regarding the numerical covariates, the hypothetical observation is set to be equal to the median. In terms of the factorial covariates, the hypothetical observation is set equal to the mode, *i.e.* the factor level which occurs most often for the specific covariable of the original data set. Although the covariates have been expressed in the most logical way, it is possible that the predicted function is based on an observation that does not exactly exist in such a way that it appears in the original data set. Nontheless, the hypothetical observation reflects an observation that would most likely be observable.

Moreover, the above introduced approach ensures that the estimated functions are not based on an outlier as this would lead to instable predictions as seen in the right panel of Figure 3.7. Figure 3.7 also shows by means of the survivor function, how the estimation of such a function depends on the baseline covariables. Figure 3.7 illustrates the estimated survivor function based on the Cox proportional hazard model (blue plotted step function) with a 95 % pointwise confidence interval (blue dashed lines). The grey lines are the bootstrap generated survivor functions based on the Cox proportional hazard model which is fitted in the framework of transformation models. All of the estimations done for the left panel are performed using the covariables of the *hypothetical patient* as it was introduced before. However, the estimations done for the right panel of Figure 3.7 utilized covariables that are differently defined. We refer to these as the observation of a *hypothetical patient* 2. The 95 % pointwise confidence interval as well as the band generated from the bootstrap estimated models is wider in the right panel than the one in the left panel. This is due to the fact, that the observations of the *hypothetical patient* 2 are less probable, which reduces the probability of estimating the function correctly. As a result, the estimation process includes higher variance.

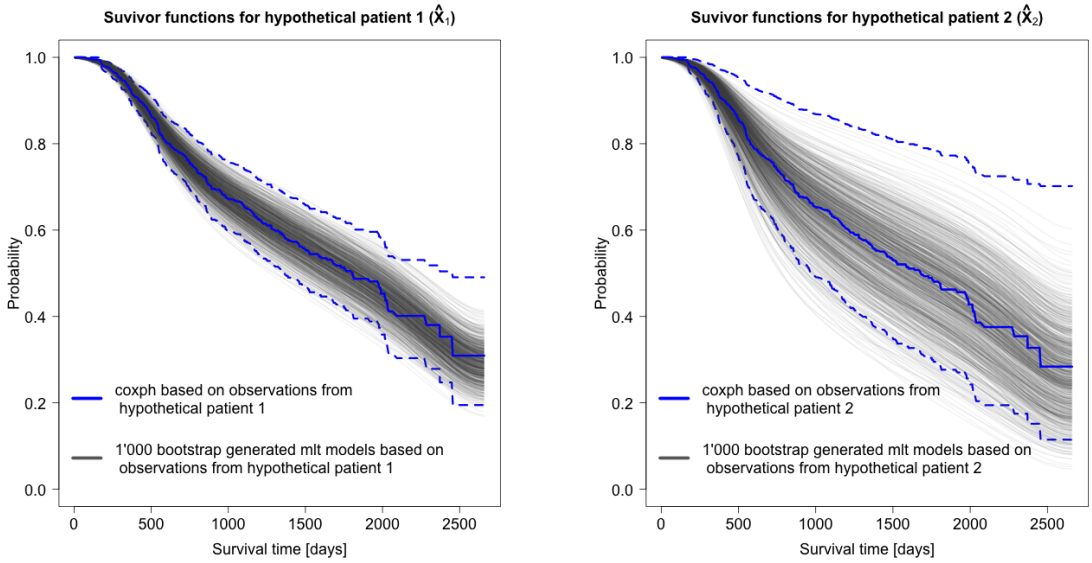

Figure 3.7.: Survivor functions (blue step functions) based on the Cox proportional hazard model including the 95 % pointwise confidence interval (dashed blue lines) and the bootstrap generated survivor functions based on the conditional transformation models $(F_Z, \boldsymbol{c}(y, \boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_b^\star)$ with $b = 1, \ldots, B$. The left and right panel are different regarding the baseline variables used for the estimation of the functions. The left panel uses the common *hypothetical patient* as its baseline variables whereas the right panel uses differently defined baseline variables referred to as *hypothetical patient* 2.

```
# Generate grid for Prediction
y_grid <- mkgrid(basis_y, n = n_sim)$y
y_grid <- y_grid[y_grid > 0] # delete the zeros
# Prediction of the £B£ survivor functions
predict_survivor_mlt_coxph_summary <- lapply(mlt_coxph_mod_summary,
                                             function(predict_survivor) {
                                                predict(predict_survivor,
                                                        newdata = hypo_obs,
                                                        q = y_grid,
                                                        type = "survivor")
                                             })
# Save objects as dataset
setwd(path_saved_R_objects)
save(predict_survivor_mlt_coxph_summary,
     file = paste("predict_survivor_mlt_coxph_summary_", n_sim, ".RData", sep=""))
# Prediction of the survivor function for original model
predict_survivor_mlt_coxph_orig <- predict(mlt_coxph_mod,
                                           newdata = hypo_obs,
                                           q = y_grid, type = "survivor")
```

R-Code 3.2: How to calculate B survivor functions based on B generated transformation models as well as the survivor function based on the original mlt model

The **R**-code 3.2 explains how to predict the survivor function based on the hypothetical observation `hypo_obs`. First, we define a grid (here: `y_grid`). Further, the `mkgrid()` function from the `variables` package (Hothorn, 2016c) generates a grid of observations from the variable description `basis_y`. The argument `n` of func-

tion `mkgrid()` defines the amount of data points for the grid. It is important to note that, the more data points used, the smoother the subsequent predicted function. Here, the grid contains of as many data points as B, *i.e.* `n = n_sim` $= B = 1000$. The prediction of the B survivor functions was obtained through the `lapply()` and `predict()` functions. The `lapply()` returns a list of the same length as the list entries of `mlt_coxph_mod_summary`, each element of which is the result of applying `predict()` to the corresponding element of `mlt_coxph_mod_summary` with `type`="`survivor`" specification. The survivor function estimated from the original model is obtained by applying the `predict()` function to the original model `mlt_coxph_mod` coupled with the specification of `type`="`survivor`".

The **R**-code 3.2 can be generalised by changing the argument `type` of the `predict()` function to the following selections: "`distribution`, "`survivor`", "`density`", "`logdensity`", "`hazard`", "`loghazard`", "`cumhazard`", "`quantile`", "`trafo`". All of the above mentioned functions can be estimated based on the mlt object in **R**. As has been noted, the focus for the subsequent analysis lies on the survivor function, however, further information on the graphical inference plots for the distribution (cf. Appendix A.5.1), the density (cf. Appendix A.5.2) and the hazard function (cf. Appendix A.5.3) can be found in the appendix.

The obtained survivor functions in **R**-Code 3.2 are used as an example for the subsequent graphical inference for functions obtained from the conditional distribution function of transformation models. Figure 3.8 shows the empirical cumulative distribution function (ECDF) of the B relative log-likelihoods (RLL) in the left panel. As introduced before (cf. Section 3.2), the cutoff line at probability 5 % helps distinguishing between the extreme models ($\hat{F}_b(\text{RLL}) < 0.05$) and those with $\hat{F}_b(\text{RLL}) \geq 0.05$. The models with $\hat{F}_b(\text{RLL}) \geq 0.05$ are characterised by a RLL value close to zero, *i.e.* they are similar to the original model. The survivor function of the extreme models with $\hat{F}_b(\text{RLL}) < 0.05$ are plotted in the right panel of Figure 3.8.

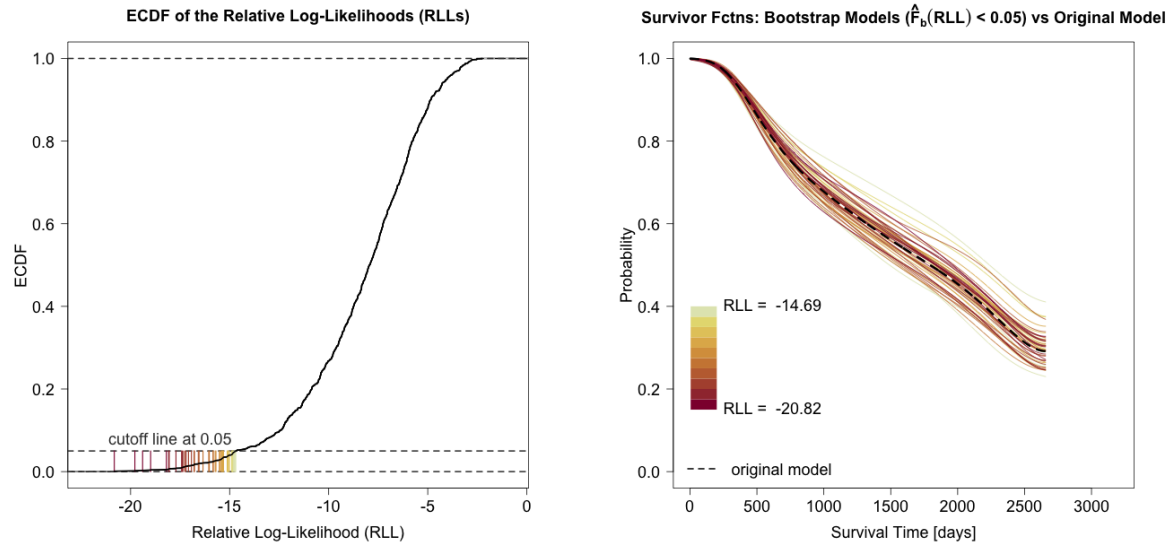

Figure 3.8.: Empirical cumulative distribution function of the relative log-likelihoods calculated by comparing the $B = 1000$ bootstrap generated transformation models $(F_Z, \boldsymbol{c}(y, \boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_b^\star)_{b=1,\ldots,B}$ versus the original transformation model $(F_Z, \boldsymbol{c}(y, \boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_N)$ of the original data set (left panel). Estimated survivor functions of the extreme ($\hat{F}_b(\text{RLL}) < 0.05$) bootstrap models (right panel).

The colour gradient for the right panel of Figure 3.8 and the left panel of Figure 3.9 is the same. The gradient is based on $\hat{F}_b(\text{RLL})$. The colouring starts just at the cutoff line at probability 5 % of the empirical cumulative distribution function: The darker the colour, the smaller the RLL; hence the less similar the bootstrap model is in comparison to the original transformation model. Based on this definition of the colour gradient, it makes sense that the lines drawn close to the survivor function of the original model (black dotted line) are brighter in comparison to the ones further away.

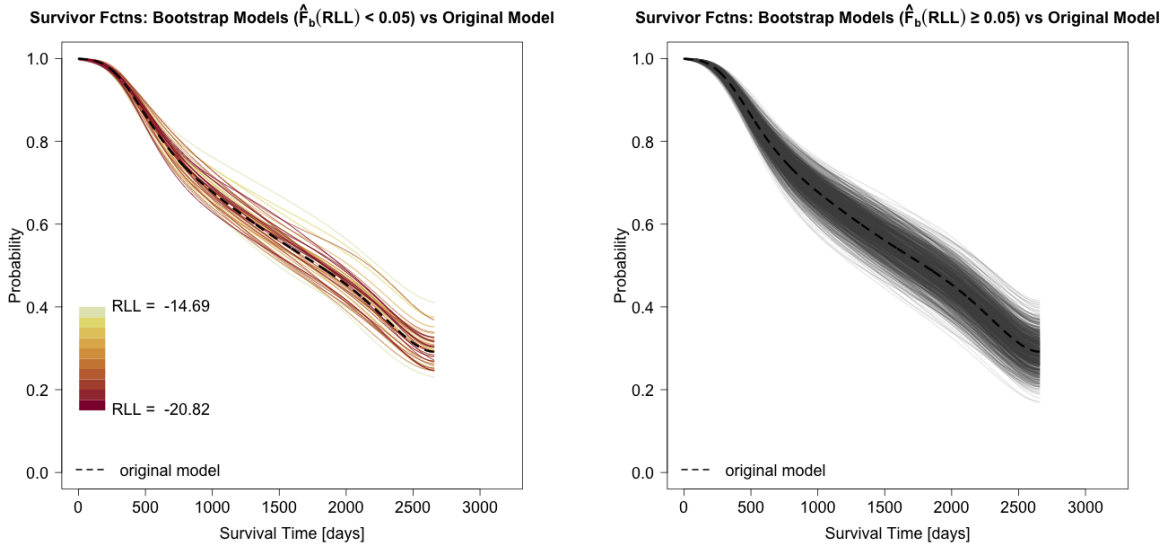

Figure 3.9.: Survivor functions of the bootstrap generated models versus the original model distinguished between $\hat{F}_b(\text{RLL}) < 0.05$ (left panel) and $\hat{F}_b(\text{RLL}) \geq 0.05$ (right panel)

Figure 3.9 shows the estimated survivor functions for all the B bootstrap generated transformation models. The survivor functions of model with $\hat{F}_b(\text{RLL}) < 0.05$ are plotted in the left panel, whereas the survivor functions of model with $\hat{F}_b(\text{RLL}) \geq 0.05$ are plotted in the right panel. Here, we have B = 1000 bootstrap samples. Hence, the left plot contains 50 and the right plot contains 950 estimated survivor functions. The estimated survivor function of the original model is added to both panels with a black dashed line so that it can be compared to the bootstrap generated models. The grey survivor functions from the bootstrap generated models (right panel) can be interpreted as a band around the survivor function. This band includes simulated survivor functions based on the estimated survivor function from the original transformation model.

The underlying model that is used for the prediction of the survivor functions in Figure 3.8 and 3.9 is known as the Cox proportional hazard model with explicitly specified log cumulative baseline hazard function that has been fitted in the framework of the transformation models: $1 - F_Z\left(h(y) - \tilde{\boldsymbol{x}}\boldsymbol{\vartheta}_2^\top\right)$. The advantage of fitting a proportional hazards model in the framework of transformation models in comparison to the conventional framework of Cox proportional hazards model is the explicitly specified log cumulative hazard baseline $h(y)$. In other words, the framework of transformation models makes up for the disadvantage of the `coxph` models which do not explicitly specify the log cumulative baseline hazard function $h(y)$. However, the coefficients obtained for the two estimated models are practically equivalent as can be seen in Table 3.6.

	coefficients of **MLT** fitted model	coefficients of **coxph** fitted model
hormonal therapy: yes	-0.349052	-0.346278
age [years]	-0.009926	-0.009459
menopausal status: post	0.267670	0.258445
tumour size [mm]	0.007771	0.007796
tumour grade: II	0.560091	0.551299
tumour grade: III	-0.201849	-0.201091
# positive nodes	0.048747	0.048789
progesterone receptor [fmol]	-0.002210	-0.002217
estrogen receptor [fmol]	0.000183	0.000197

Table 3.6.: Model parameters (corresponding to the model covariates) of original transformation model in comparison to the Cox proportional hazard model fitted with the function `coxph()` in **R**

Figure 3.10 illustrates a graphical representation of the comparison between **(1)** the survival curves generated from the `coxph()` and the `survfit()` function using the package `survival` (Therneau, 2015) and **(2)** the survival curves generated from the `mlt()` and the `predict()` function using the package `mlt` (Hothorn, 2016b).

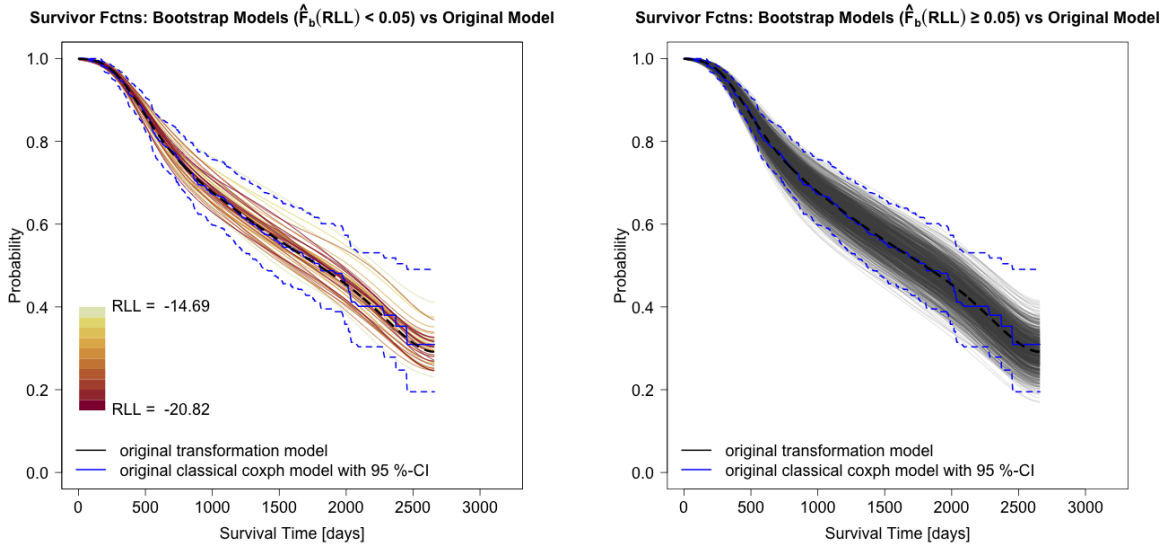

Figure 3.10.: Survivor functions of the bootstrap generated models in comparison to the original model. The extreme models ($\hat{F}_b(\text{RLL}) < 0.05$) are shown in the left panel, the models with $\hat{F}_b(\text{RLL}) \geq 0.05$ in the right panel. In addition the predicted survivor function of the Cox proportional hazard model (using `coxph()` and `survfit()` in **R**) is added to the plot with its 95 % pointwise confidence intervals.

As can be seen in Figure 3.10, both the common Cox proportional hazards model (blue dashed lines) and the Cox proportional hazards model fitted in the framework of transformation models with explicitly specified log cumulative baseline hazard function, allow for heteroscedasticity, *i.e.* the variance of the confidence band gets bigger with increasing survival time. In general, the pointwise confidence interval for the survivor function (blue dashed lines) of the Cox proportional hazards model fits the band obtained from the bootstrap sample

well. Nevertheless, a closer look reveals a higher variance for the conventional pointwise confidence interval for the survivor function of the Cox proportional hazard model, compared to the variance of the bootstrap generated "confidence band" at survival time equal to 2700 days. This is because the bootstrap generated responses $Y_1^\star, \ldots, Y_B^\star$ are not censored, therefore, their corresponding models do not include censoring, whereas the Cox proportional hazard model is based on the original data set where some of the original responses Y are censored. Censored observations occur when the information of an observation about their survival time is incomplete. In other words, the Cox proportional hazard model with the explicitly specified log cumulative baseline hazard function fitted in the framework of transformation models has "more" data than actually available in the original data set. For this reason, it is obvious that the model fit is better and the variance from the common Cox proportional hazard model is smaller.

4. Discussion

The parametric bootstrap inference for transformation models enables a graphically interpretable likelihood based model inference. An advantage of this procedure is that it eliminates the need to make any assumptions about the asymptotic behaviour which makes the procedure applicable to small data sets. Furthermore, there is no need to make assumptions about the error independency since a parametric distribution is not required for the error distribution (Davino *et al.*, 2013) in the framework of transformation models. As mentioned in the vignette of the `mlt` package (Hothorn, 2016b) for the framework of transformation models, the inspection of the parameter estimates is not essential as the models are better looked at by means of the estimated distribution, density, survivor, quantile and hazard functions. This results from the characteristic of the framework of transformation models that examines how covariates influence the entire conditional response distribution (Koenker, 2005). In addition, the parametric bootstrap generated band for all kinds of functions that are derived from the conditional distribution function often serves as a relatively easy interpretable inference. The resulting plots of the procedure are also illustrative for persons who are not very familiar with statistical inference. This is due to the fact that the descriptive colour shading is based on the log-likelihood functions of the model and reflects the probability of the estimated functions.

In comparison to the conventional pointwise confidence interval for the survivor function of the conventional Cox proportional hazard model, the band based on the parametric bootstrap generated functions only consist of functions that are correctly defined for that specific case (cf. Figure 3.1 for a poor example). More specifically, the bootstrap generated band around the survivor function only contains monotonically decreasing functions. Alternatively, the integral of each of the probability density functions that are contained in the bootstrap generated band around the probability density function of the original model is always equal to one and the probability density function itself is everywhere non-negative.

Regarding the finding about the not as expected log-likelihood ratio statistics distribution in cases of multiply occurring model coefficients, this thesis does not definitively provide a solution. However, simulations have been included to prove the presumption that a correction of the degrees of freedom in instances of multiply occurring model coefficients is essential. In conclusion, the results of this thesis advance the understanding of graphical model inference of the model parameters of a conditional transformation model as well as the inference of the conditional transformation model itself.

4.1. Limitations

Throughout this thesis, we imply that the original transformation model $(F_Z, \boldsymbol{c}(y, \boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_N)$ is a good model that fits the underlying response variable Y, - the recurrence-free survival (RFS) time, well. However, it is crucial to bear in mind that the procedure of applying the parametric bootstrap resampling method to a transformation model is limited in cases where the underlying transformation model is not adequately established. In other words, caution must be exercised whenever the underlying parametric model is wrong, as the application of

the parametric bootstrap resampling will also lead to wrong results. To quote Box (1979), “All models are wrong but some are useful.”, so to speak, the original model has to be considered useful.

Furthermore, the dependency on the original model itself could be looked at as a limitation of the procedure. Some may criticise the fact that bootstrap samples are based on a single sample data set (here: GBSG2) for a given population. A phenomenon which causes the replications to be limited to a finite number of replications (“bootstrap resampling variability”, Davino *et al.*, 2013).

4.2. Outlook

In summary, it can be said that this thesis forms a solid foundation for the application of the parametric bootstrap resampling method to the framework of transformation models. Nevertheless, the research findings presented in this thesis have created several new ideas that should be further explored. These findings are addressed in the paragraph below.

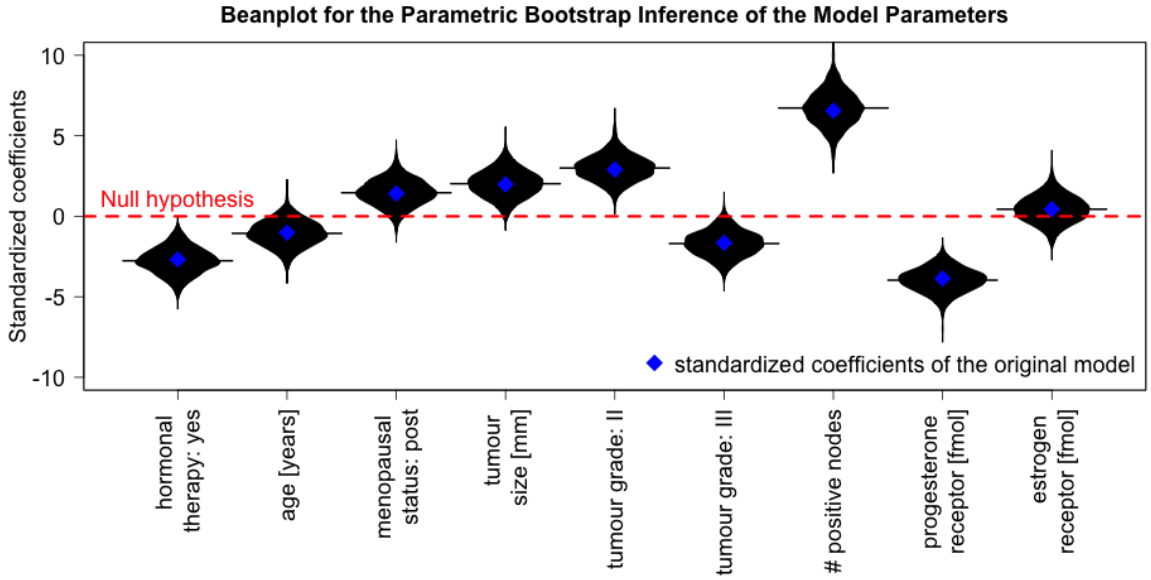

Figure 4.1.: Beanplot of the standardized model coefficients. The estimated density of the distribution as well as the mean of the bootstrap generated coefficients can directly be interpreted from the plot. The blue dots represent the coefficients of the original model.

The parametric bootstrap inference for the parameters of transformation models (cf. Section 3.2) can be looked at as a starting point for a future implementation of a (visualized) Type-1-error hypothesis test. Figure 4.1 roughly exhibits this idea. The shown beanplots represent the standardized distribution of the bootstrap generated model coefficients together with the original model coefficients visualized with a blue dot. The ticks on each beanplot mark the 50 % quantile, *i.e.* the mean. The distributions of the model coefficients shown in Figure 4.1 could be interpreted in the following way: The red dashed line represents the null hypothesis which states that the i-th coefficient $\hat{\vartheta}_i$ of the model is equal to zero. Hence, the coefficients close to the red dashes, *i.e.* distributed around $\hat{\vartheta}_i = 0$, do not affect the response variable Y of the model as much as the coefficients that are further away from the red dashes, *i.e.* distributed around $\hat{\vartheta}_i \neq 0$. This Type-1-error hypothesis test

might as well be useful for variable selection.

Now in comparing the computing time of the bootstrap resampling method to the most likely transformation models in **R** is dependent on B: the more bootstrap samples B that are drawn, the more time-consuming the computations for estimating the model gets. Especially the process of the maximization of the gradient does take a while. For future research projects, the performance of the `mlt()` function could be improved by outsourcing the maximization of the gradient into the programming languages *C++* or *Python* in order to use their speed as an advantage.

Further, future research projects can explore a specific parametric bootstrap based inference methodology for transformation models which are estimated in a survival analysis framework. This approach should then additionally be able to consider the issue of censored observations (cf. Section 3.3).

And lastly, the most obvious future research induced by this thesis, is the need to investigate the degrees of freedom of the Chi-squared distributed LLRS. We dare to hypothesize that not using the Bernstein polynomial as a basis function, reduces the doubled coefficients entries which then corrects the expected degree of freedom of the Chi-squared distributed LLRS.

References

Box, G. E. (1979). Robustness in the Strategy of Scientific Model Building. *Robustness in Statistics*, **1**, 201–236.

Box, G. E. and Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, **26**, 211–252.

Chernick, M. R. and LaBudde, R. A. (2014). *An Introduction to Bootstrap Methods with Applications to **R***. John Wiley & Sons.

Davino, C., Furno, M., and Vistocco, D. (2013). *Quantile Regression: Theory and Applications*. John Wiley & Sons, New York, U.S.A.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, **7**, 1–26.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. CRC press.

Fahrmeir, L., Kneib, T., and Lang, S. (2007). *Regression*. Springer.

Farouki, R. T. (2012). The Bernstein Polynomial Basis: A Centennial Retrospective. *Computer Aided Geometric Design*, **29**, 379–419.

Fraser, D. A. S. (1968). *The Structure of Inference*. John Wiley & Sons, New York, U.S.A.

Geyer, C. J. (2015). Statistics 5102: Bootstrap. University Lecture, University of Minnesota, School of Statistics. `http://www.stat.umn.edu/geyer/5102/slides/s8.pdf`.

Good, P. I. (2001). *Resampling Methods*. Springer.

Hastie, T. and Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science*, **1**, 297–310.

Held, L. and Bové, D. S. (2013). *Applied Statistical Inference: Likelihood and Bayes*. Springer Science & Business Media.

Hothorn, T. (2015). *TH.data: TH's Data Archive*. **R** package version 1.0-6, URL `https://CRAN.R-project.org/package=TH.data`.

Hothorn, T. (2016a). *basefun: Infrastructure for Computing with Basis Functions*. **R** package version 0.0-30, URL `https://CRAN.R-project.org/package=basefun`.

Hothorn, T. (2016b). *mlt: Most Likely Transformations*. **R** package version 0.0-30, URL `https://CRAN.R-project.org/package=mlt`.

Hothorn, T. (2016c). *variables: Variable Descriptions*. **R** package version 0.0-30, URL `https://CRAN.R-project.org/package=variables`.

Hothorn, T., Kneib, T., and Bühlmann, P. (2014). Conditional Transformation Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76**, 3–27.

Hothorn, T., Möst, L., and Bühlmann, P. (2015). Most Likely Transformations. arXiv:1508.06749. Technical report, v2. URL `http://arxiv.org/abs/1508.06749`.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer Science & Business Media.

Kleinbaum, D. G. and Klein, M. (1996). *Survival analysis*. Springer.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

Lindsey, J. (1999). Some Statistical Heresies. *The Statistician*, **48**, 1–40.

Lindsey, J. K. (1996). *Parametric Statistical Inference*. Oxford University Press.

Möst, L. (2014). *Conditional Transformation Models*. PhD thesis, Ludwig-Maximilian-Universität München. Full title: Conditional Transformation Models - Interpretable Parametrisations and Censoring.

R Development Core Team (2009). ***R**: A Language and Environment for Statistical Computing*. **R** Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org`.

Sauerbrei, W., Royston, P., Bojar, H., Schmoor, C., Schumacher, M., Group, G. B. C. S.,*et al.* (1999). Modelling the Effects of Standard Prognostic Factors in Node-Positive Breast Cancer. *British Journal of Cancer*, **79**, 1752.

Stasinopoulos, D. M. and Rigby, R. A. (2007). Generalized Additive Models for Location, Scale and Shape (GAMLSS) in **r**. *Journal of Statistical Software*, **23**, 1–46.

Therneau, T. M. (2015). *A Package for Survival Analysis in S*. **R** package version 2.38, URL `http://CRAN.R-project.org/package=survival`.

A. Appendix

A.1. Flowchart: How to Estimate a Transformation Model in a Step-by-Step Manner

The theory of transformation models was introduced in Section 2.2. This flowchart summarises at a glance main steps required to estimate a transformation model in a full likelihood framework.

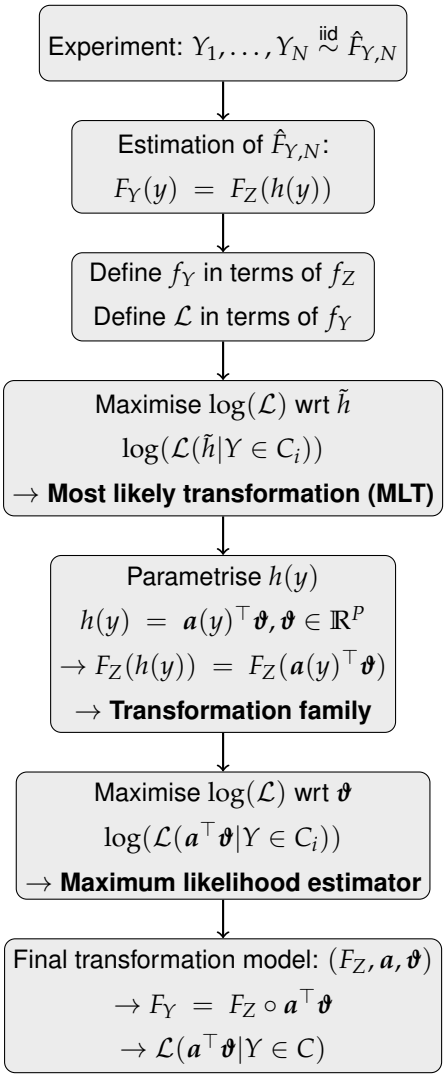

Figure A.1.: Flowchart illustrating the main steps required to estimate a transformation model in a full likelihood framework

A.2. Histogram of the Log-Likelihood Ratio Statistics Calculated from the Original Transformation Model and the Bootstrap Generated Models with Bernstein Polynomials of Different Order

We introduced the log-likelihood ratio statistic as a measure for doing likelihood based inference in Section 3.1.2. The following histograms visualize the log-likelihood ratio statistics from the original transformation model $(F_Z, \boldsymbol{c}(y, \boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_N)$ and the bootstrap generated models $(F_Z, \boldsymbol{c}(y, \boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_b^\star)_{b=1,\dots,B}$ with Bernstein polynomial of different order: $(F_Z, (\boldsymbol{a}_{\text{Bs},i=2,\,3,\,7,\,8,\,12,\,13}(y), \boldsymbol{b}(\boldsymbol{x})^\top)^\top, \hat{\boldsymbol{\vartheta}}_N)$. The coloured lines are probability density functions of the Chi-squared distribution with different degrees of freedom according to the summary of Table 3.2 in Section 3.1.2. The colours red, blue, yellow and green also correspond to the explaination given in Section 3.1.2.

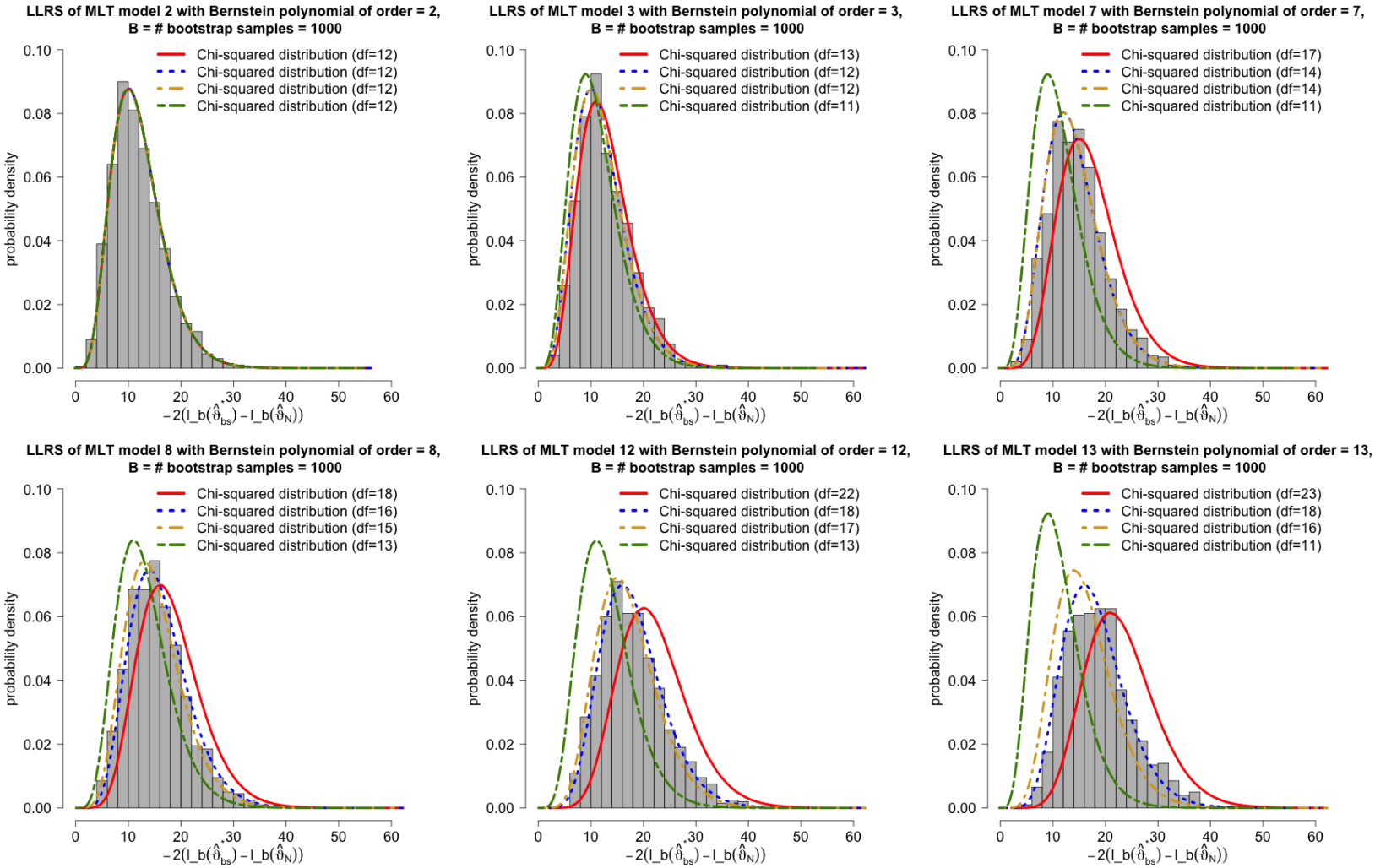

Figure A.2.: Histograms of the LLRS from the original transformation model $(F_Z, \boldsymbol{c}(y,\boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_N)$ and the bootstrap generated models $(F_Z, \boldsymbol{c}(y,\boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_b^\star)_{b=1,\ldots,B}$ with Bernstein polynomial of different order: $(F_Z, (\boldsymbol{a}_{\text{Bs},i=2,3,7,8,12,13}(y), \boldsymbol{b}(\boldsymbol{x})^\top)^\top, \hat{\boldsymbol{\vartheta}}_N)$. The probability density functions of the Chi-squared distribution are added for different degrees of freedom according to the summary of Table 3.2 in Section 3.1.2.

A.3. Simulation Study for the Distribution of the Log-Likelihood Ratio Statistics (LLRS)

A.3.1. Histograms of the LLRS Resulting from the Simulation Study with Bernstein Polynomial Order = 5

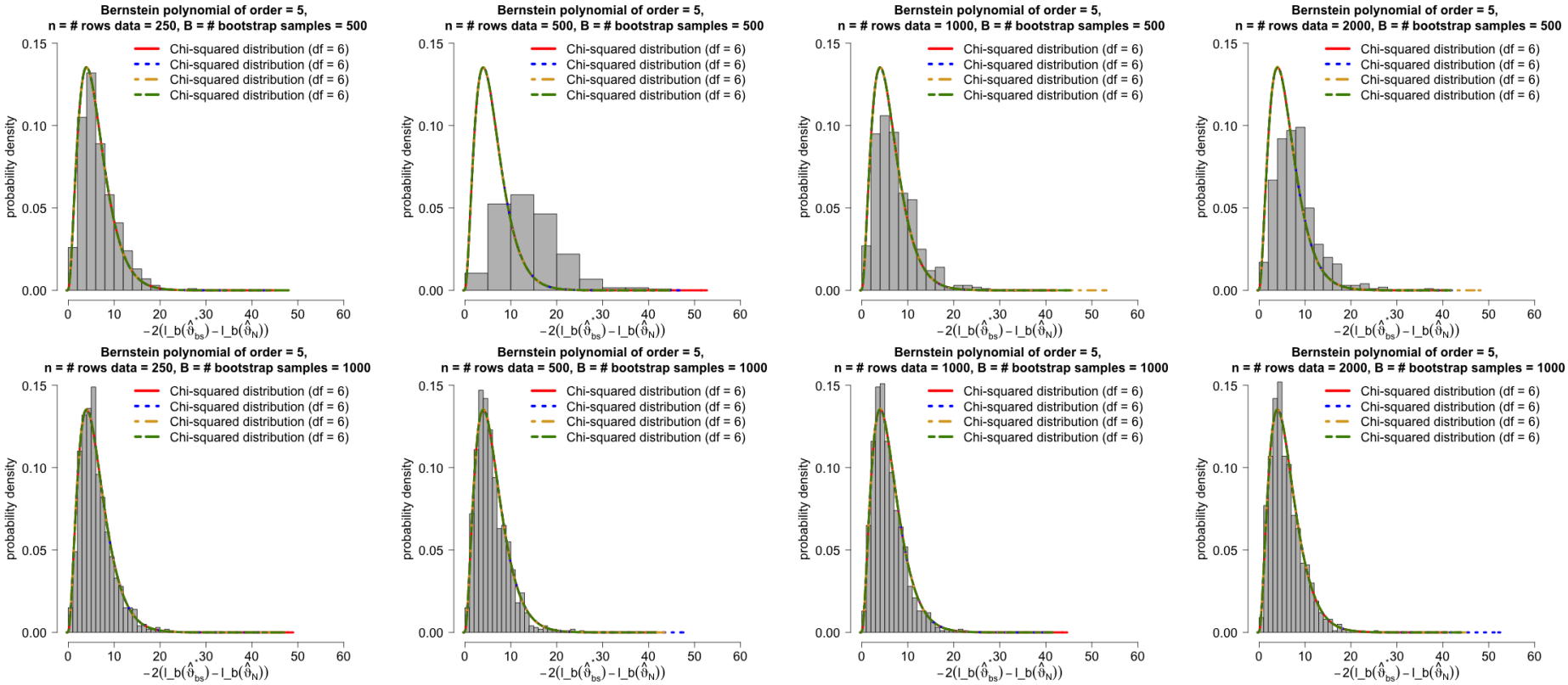

Figure A.3.: Histograms of the LLRS resulting from the simulation study based on $\mathcal{N}(0,1)$ distributed data. The model $(F_Z, \boldsymbol{a}_{\text{Bs},5}, \hat{\boldsymbol{\vartheta}}_N)$ serves as original model. The top row shows the results of the simulation study based on the $B = 500$ bootstrap samples whereas the results from the bottom row are based on the $B = 1000$ bootstrap samples. The more to the right, the bigger the underlying data set for the simluations ($n =$ # rows of data set = $250,\ 500,\ 1000,\ 2000$).

A.3.2. Histograms of the LLRS Resulting from the Simulation Study with Bernstein Polynomial Order = 7

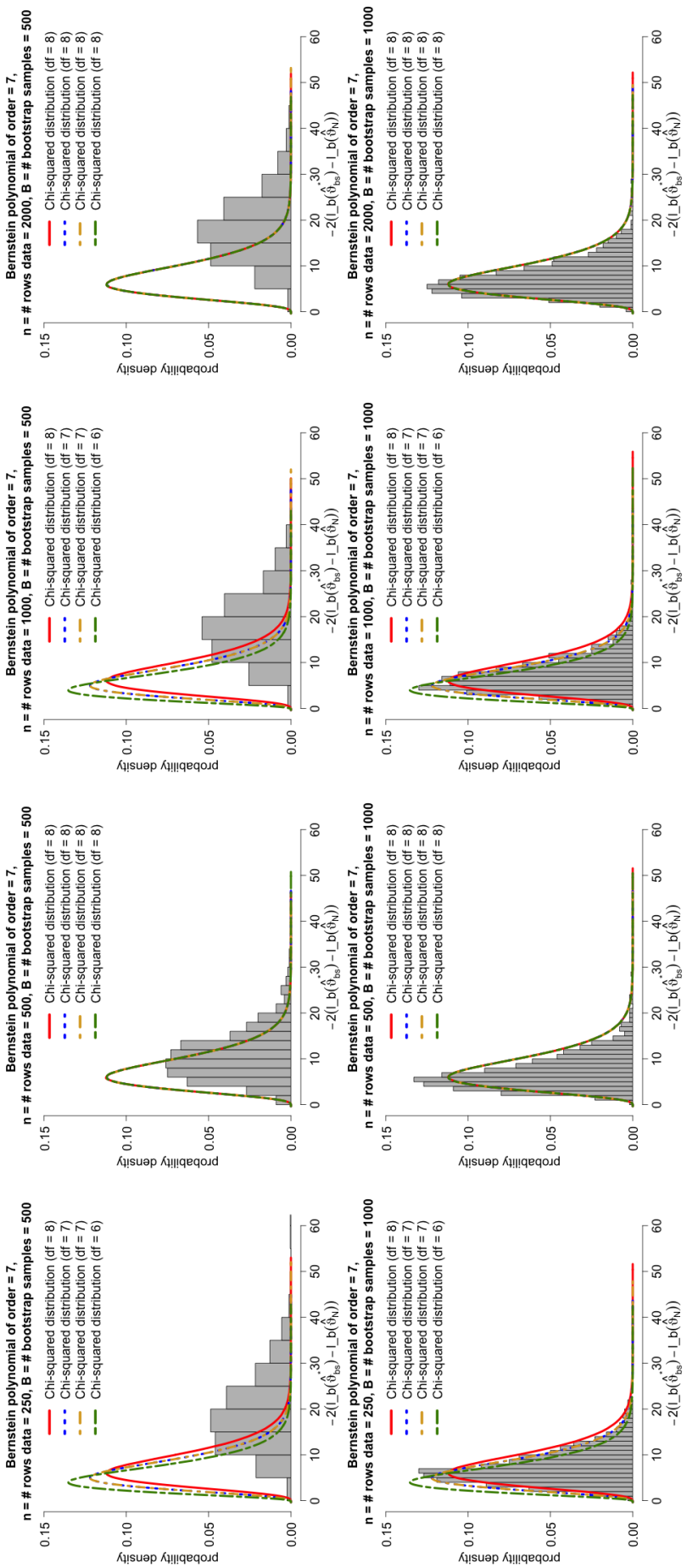

Figure A.4.: Histograms of the LLRS resulting from the simulation study based on $\mathcal{N}(0,1)$ distributed data. The model $(F_Z, \boldsymbol{a}_{\text{Bs},7}, \hat{\boldsymbol{\vartheta}}_N)$ serves as original model. The top row shows the results of the simulation study based on the $B = 500$ bootstrap samples whereas the results from the bottom row are based on the $B = 1000$ bootstrap samples. The more to the right, the bigger the underlying data set for the simluations ($n =$ # rows of data set $= 250,\ 500,\ 1000,\ 2000$).

A.3.3. Histograms of the LLRS Resulting from the Simulation Study with Bernstein Polynomial Order = 10

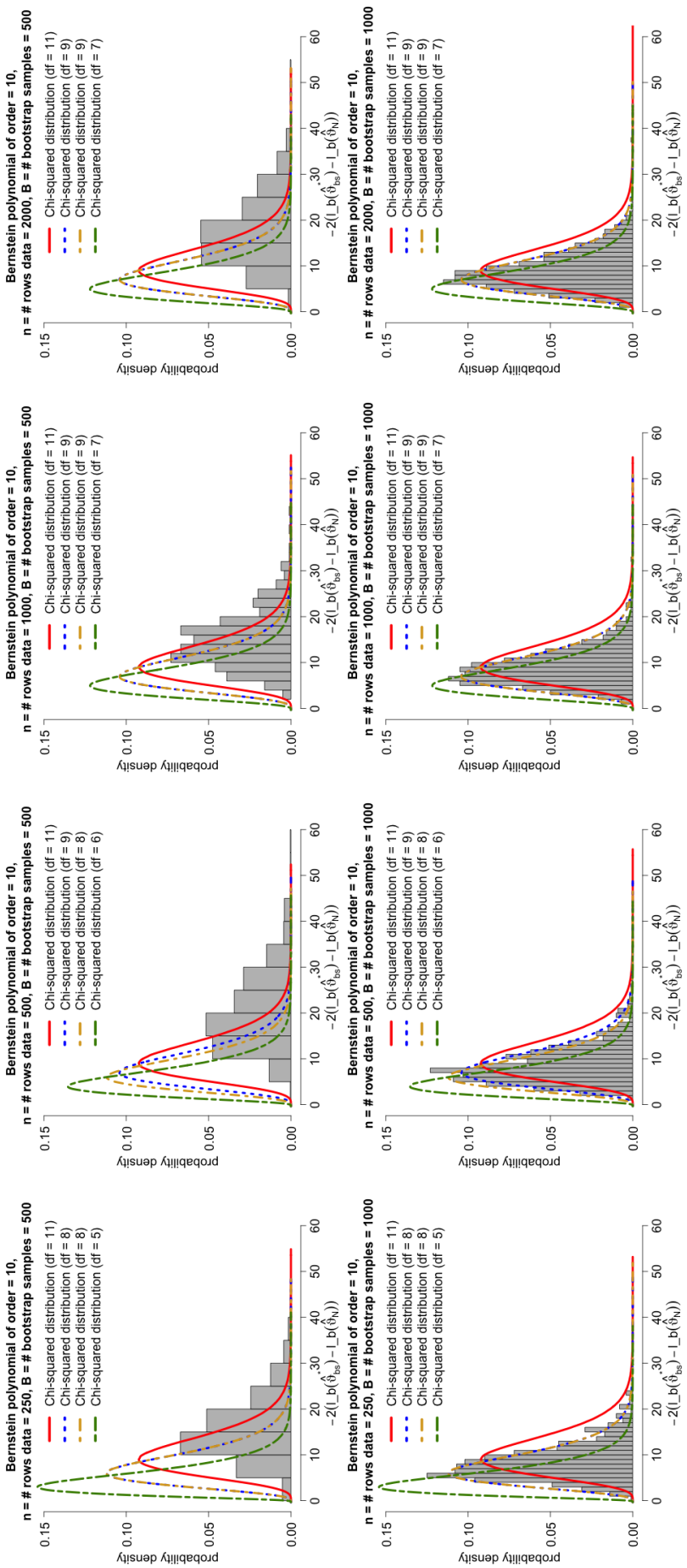

Figure A.5.: Histograms of the LLRS resulting from the simulation study based on $\mathcal{N}(0,1)$ distributed data. The model $(F_Z, \boldsymbol{a}_{\text{Bs},10}, \hat{\boldsymbol{\vartheta}}_N)$ serves as original model. The top row shows the results of the simulation study based on the $B = 500$ bootstrap samples whereas the results from the bottom row are based on the $B = 1000$ bootstrap samples. The more to the right, the bigger the underlying data set for the simluations ($n =$ # rows of data set $= 250,\ 500,\ 1000,\ 2000$).

A.4. R-Code on How to Define the *Hypothetical Observation*

In Section 3.3, the *hypothetical observation* was introduced and explained. The following **R**-Code A.1 shows how the definition of such a hypothetical observation can be implemented in **R**.

```
# ********************************************
# Define hypothetical observation for prediction
# ********************************************
GBSG2_covariates <- GBSG2[,xvar]
summary(GBSG2_covariates)
# Define sub data frames.
GBSG2_num_covariates <- GBSG2_covariates[,c("age", "tsize", "pnodes",
                                            "progrec", "estrec")]
GBSG2_cat_covariates <- GBSG2_covariates[,c("horTh", "menostat", "tgrade")]
# Calculate median of all numerical covariates
hypo_obs_num_covariates <- data.frame(t(data.frame(apply(GBSG2_num_covariates,
                                                         2, median))))
hypo_obs_num_covariates <- round(hypo_obs_num_covariates, 0)
# Calculate median of all categorical covariates
hypo_obs_cat_covariates <- rep(NA, ncol(GBSG2_cat_covariates))
names(hypo_obs_cat_covariates) <- colnames(GBSG2_cat_covariates)
for( i in colnames(GBSG2_cat_covariates)){
   hypo_obs_cat_covariates[i] <- names(sort(table(GBSG2_cat_covariates[,i]),
                                            decreasing=TRUE)[1])
}
#
hypo_obs <- cbind(hypo_obs_num_covariates,
                  t(data.frame(hypo_obs_cat_covariates)))
hypo_obs <- hypo_obs[,xvar]
#
levels(hypo_obs$horTh) <- levels(GBSG2$horTh)
levels(hypo_obs$menostat) <- levels(GBSG2$menostat)
hypo_obs$tgrade <- ordered(hypo_obs$tgrade, levels=c("I", "II", "III"))
#
summary(hypo_obs)
```

R-Code A.1: Explaining in a step-by-step manner how the *hypothetical observation*, *i.e. hypothetical patient*, is defined

A.5. Parametric Bootstrap Inference for Functions Obtained from the Conditional Distribution Function of the Transformation Models

In Section 3.3, the parametric bootstrap inference for functions obtained from the conditional distribution function of the transformation model was introduced. It has been noted (cf. **R**-Code 3.2) that depending on the argument `type` of the `predict()` function the following functions can be predicted "`distribution`", "`survivor`", "`density`", "`logdensity`", "`hazard`", "`loghazard`", "`cumhazard`", "`quantile`", "`trafo`". Here, we concentrate on the distribution function (cf. Appendix A.5.1), the density function (cf. Appendix A.5.2) and the hazard function (cf. Appendix A.5.3).

A.5.1. The Cumulative Distribution Function

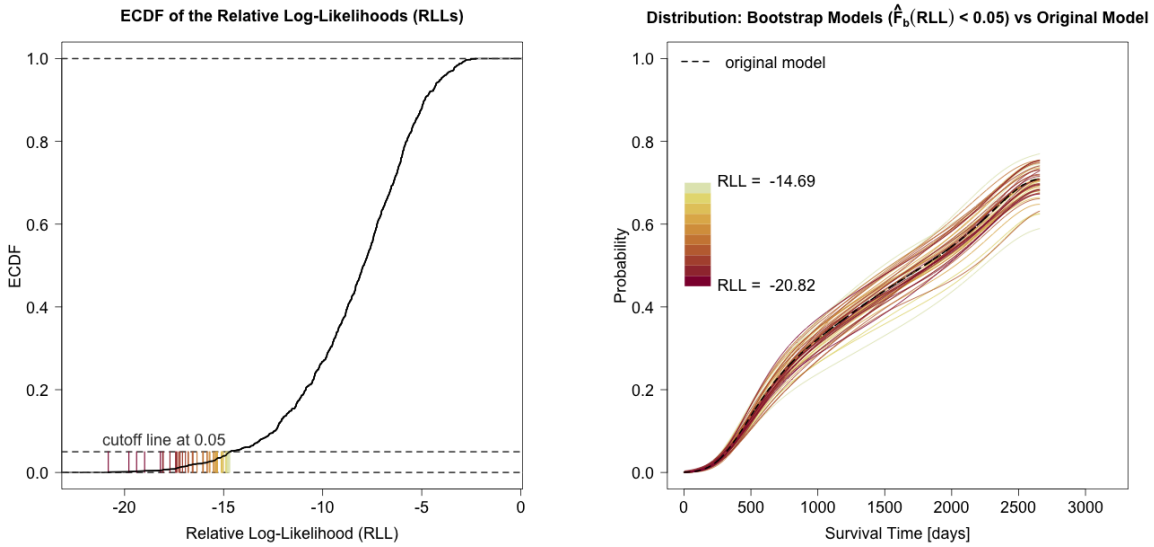

Figure A.6.: Empirical cumulative distribution function of the relative log-likelihoods calculated by comparing the $B = 1000$ bootstrap generated transformation models $(F_Z, \boldsymbol{c}(y, \boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_b^\star)_{b=1,\ldots,B}$ versus the original transformation model $(F_Z, \boldsymbol{c}(y, \boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_N)$ of the original data set (left panel). Estimated distribution functions of the extreme ($\hat{F}_b(\text{RLL}) < 0.05$) bootstrap models (right panel).

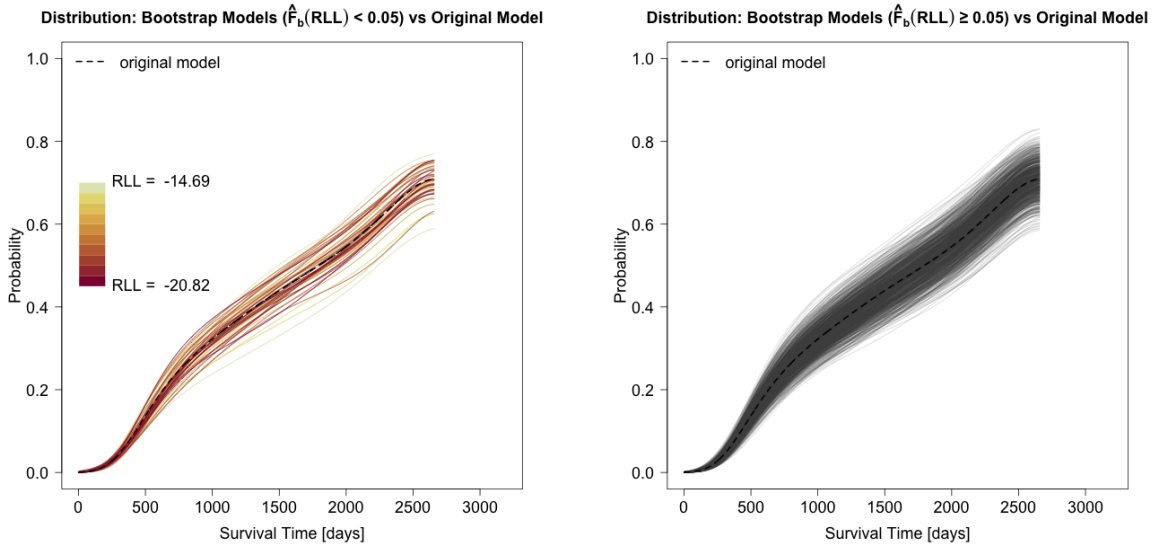

Figure A.7.: Distribution functions of the bootstrap generated models versus the original model distinguished between $\hat{F}_b(\text{RLL}) < 0.05$ (left panel) and $\hat{F}_b(\text{RLL}) \geq 0.05$ (right panel)

A.5.2. The Density Function

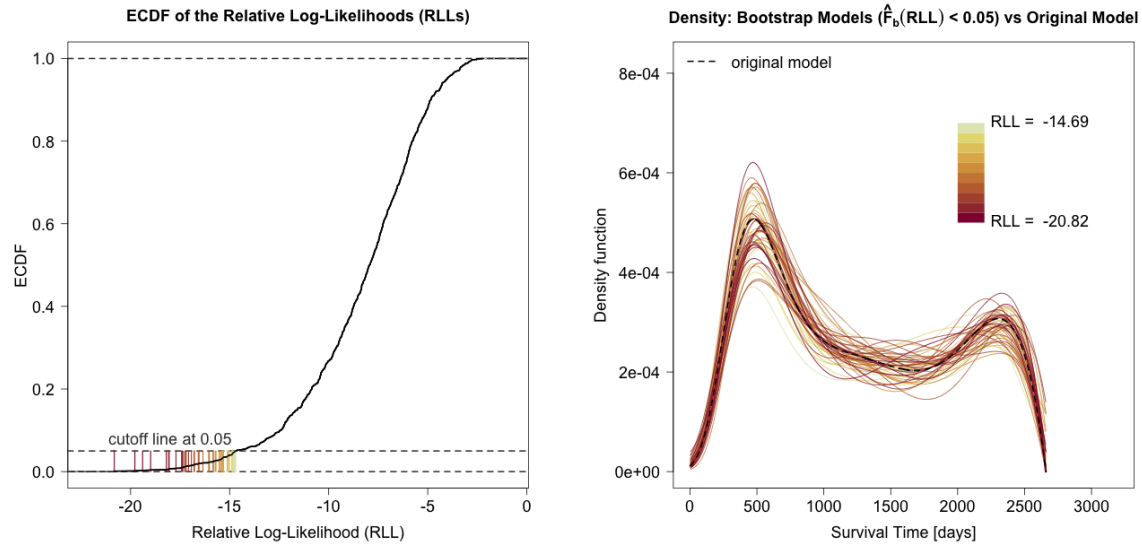

Figure A.8.: Empirical cumulative distribution function of the relative log-likelihoods calculated by comparing the $B = 1000$ bootstrap generated transformation models $(F_Z, \boldsymbol{c}(y, \boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_b^\star)_{b=1,\ldots,B}$ versus the original transformation model $(F_Z, \boldsymbol{c}(y, \boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_N)$ of the original data set (left panel). Estimated probability density functions of the extreme $(\hat{F}_b(\text{RLL}) < 0.05)$ bootstrap models (right panel).

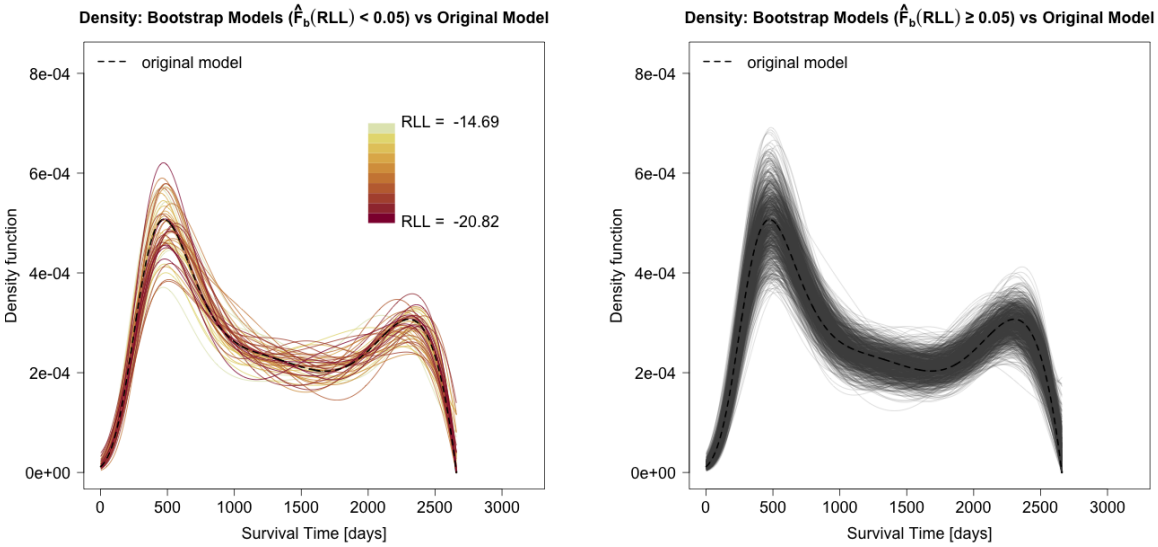

Figure A.9.: Probability density functions of the bootstrap generated models versus the original model distinguished between $\hat{F}_b(\text{RLL}) < 0.05$ (left panel) and $\hat{F}_b(\text{RLL}) \geq 0.05$ (right panel)

A.5.3. The Hazard Function

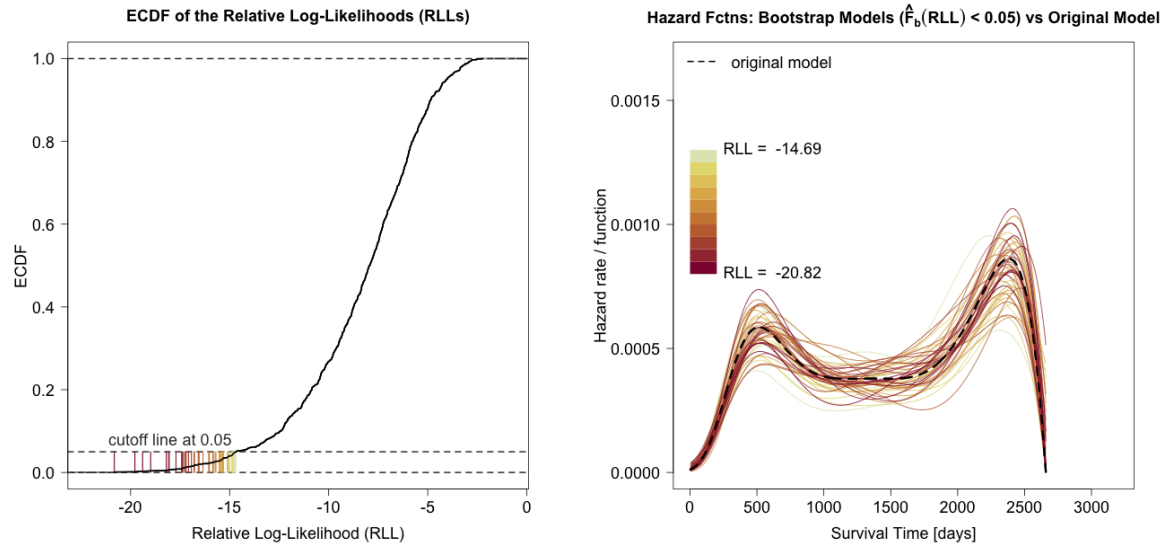

Figure A.10.: Empirical cumulative distribution function of the relative log-likelihoods calculated by comparing the $B = 1000$ bootstrap generated transformation models $(F_Z, \boldsymbol{c}(y, \boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_b^\star)_{b=1,\dots,B}$ versus the original transformation model $(F_Z, \boldsymbol{c}(y, \boldsymbol{x}), \hat{\boldsymbol{\vartheta}}_N)$ of the original data set (left panel). Estimated hazard functions of the extreme $(\hat{F}_b(\text{RLL}) < 0.05)$ bootstrap models (right panel).

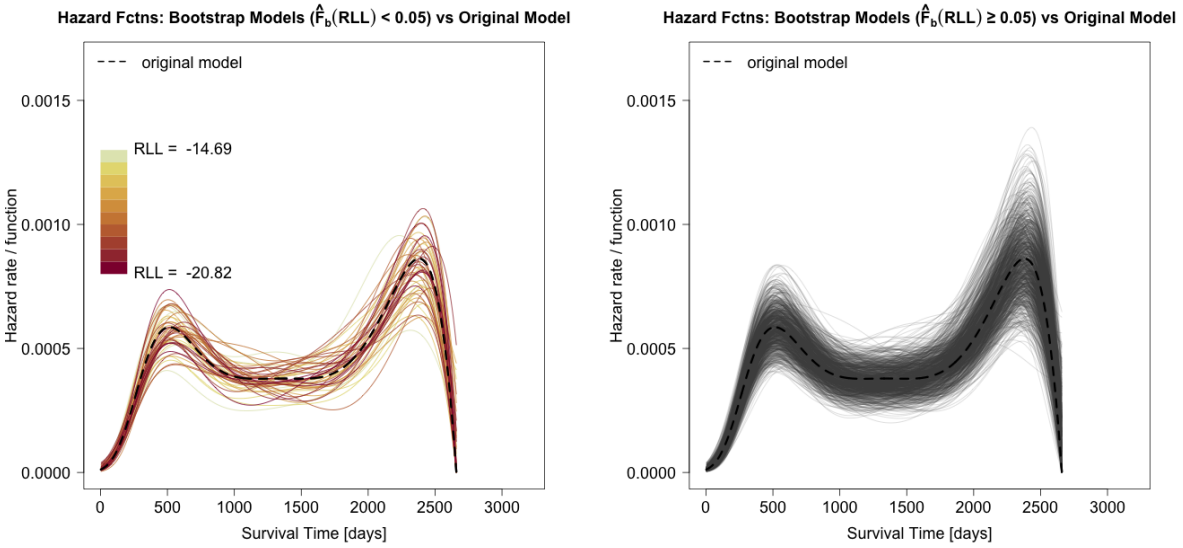

Figure A.11.: Hazard functions of the bootstrap generated models versus the original model distinguished between $\hat{F}_b(\text{RLL}) < 0.05$ (left panel) and $\hat{F}_b(\text{RLL}) \geq 0.05$ (right panel)