Conditional Transformation Models - Interpretable Parametrisations and Censoring

Lisa Möst

München 2014

Conditional Transformation Models - Interpretable Parametrisations and Censoring

Lisa Möst

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität München

vorgelegt von
Lisa Möst
aus München

München, den 2. Dezember 2014

Erster Berichterstatter: Prof. Dr. Torsten Hothorn
Zweiter Berichterstatter: Prof. Dr. Thomas Kneib
Tag der Disputation: 11. Februar 2015

Zusammenfassung

Die meisten wohlbekannten Regressionsmodelle konzentrieren sich auf die Schätzung des bedingten Erwartungswerts gegeben einer Reihe von erklärenden Variablen. Höhere Momente der Verteilungsfunktion werden üblicherweise als konstant angenommen. Damit verbunden sind typischerweise starke Annahmen wie Homoskedastizität oder eine symmetrische Verteilungsfunktion. In der flexiblen Modellklasse der konditionalen Transformationsmodelle (CTMs) hingegen wird die gesamte bedingte Verteilungsfunktion direkt modelliert. Dadurch dürfen auch höhere Momente der bedingten Verteilungsfunktion (wie Varianz, Wölbung und Schiefe) von den erklärenden Variablen abhängen. CTMs enthalten wiederum lineare Transformationsmodelle (z.B. proportional hazards und proportional odds Modelle) als Spezialfall, die ein nützliches Instrument zur Datenanalyse in zahlreichen Anwendungsgebieten darstellen. Um einen umfassenden Literaturüberblick zu geben, wird die Entwicklung linearer Transformationsmodelle innerhalb der letzten 20 Jahre in der vorliegenden Dissertation zusammengefasst und bewertet. Häufig verwendete Regressionsmodelle werden aus der Transformationsmodell-Perspektive betrachtet, wodurch die gemeinsame Modellbasis dieser Regressionsmodelle verdeutlicht wird.

Der methodische Schwerpunkt dieser Arbeit liegt in der Einführung von *conditionally linear transformation models* (CLTMs), die einen wichtigen Spezialfall von CTMs darstellen und in der Erweiterung von CTMs auf zensierte Zielgrößen. Der Einfluss der erklärenden Variablen auf die ersten beiden Momente der Verteilungsfunktion ist in den vorgeschlagenen Parametrisierungen von CLTMs interpretierbar und detailliertere Einblicke in die Modellstruktur werden ermöglicht.

Für einige niedrig-parametrisierte CLTMs wird ein likelihood-basierter Schätzansatz vorgestellt, der sich leicht auf beliebige Zensierungsarten erweitern lässt. Die damit verbundene Maximum-Likelihood Theorie macht diesen Ansatz besonders attraktiv. Alternativ können CLTMs durch regularisierte Optimierung unter Verwendung eines komponentenweisen Boosting-Algorithmus geschätzt werden. Dieser Schätzansatz ist nicht auf niedrig-parametrisierte CLTMs beschränkt und kann für deren gesamte Bandbreite verwendet werden. Insbesondere für Anwendungen in der Überlebenszeitanalyse wird die Zielfunktion durch die Berücksichtigung von *inverse probability of censoring weights* auf rechtszensierte Zielgrößen erweitert.

Die Überlegenheit von C(L)TMs im Vergleich zu weniger flexiblen Standardregressionsmodellen wurde in zwei Simulationsstudien untersucht. Außerdem wurden zwei besonders wichtige Anwendungen aus dem Bereich der Biostatistik für diese Dissertation ausgewählt. Der Einfluss von Ultraschallmessungen auf das zukünftige Geburtsgewicht von Neugeborenen aus der Perinatalen Datenbank Erlangen (Deutschland) wurde mit

Hilfe niedrig-parametrisierter likelihood-basierter CLTMs analysiert. Dabei durften die Geburtsgewichte einer beliebigen Verteilung mit fötusspezifischem Erwartungswert und fötusspezifischer Varianz folgen. Zusätzlich wurden flexiblere C(L)TMs verwendet, um fötusspezifische Prädiktionsintervalle für das zukünftige Geburtsgewicht zu schätzen. In der Überlebenszeitanalyse ist die Schätzung von patientenspezifischen Überlebensfunktionen in Abhängigkeit von Patienteneigenschaften von speziellem Interesse. Dabei ist die Verwendung von CTMs besonders empfehlenswert, da die bedingte Überlebensfunktion direkt geschätzt wird. Zur näheren Illustration wurde das Überleben von Patienten, die an chronischer myeloischer Leukämie leiden unter Verwendung von CTMs analysiert.

Abstract

Most well-known regression models focus on the estimation of the conditional mean given a set of explanatory variables. Higher moments of the distribution function are usually assumed as constant. This typically implies strict assumptions such as homoscedasticity or symmetry. In contrast, in the flexible model class of conditional transformation models (CTMs), the whole conditional distribution function is modelled directly. Thereby, higher moments of the conditional distribution (*i.e.* variance, kurtosis, and skewness) are allowed to depend on explanatory variables. CTMs include linear transformation models (*e.g.*, proportional hazards and proportional odds models) as a special case, which display a powerful tool for data analysis in various fields. To provide a broad literature overview, the development of linear transformation models over the past twenty years is summarised in this thesis. Frequently used regression models are reviewed from the perspective of transformation models to clarify their common model basis.

The methodological emphasis of this thesis is the introduction of conditionally linear transformation models (CLTMs), which constitute an important special case of CTMs, and the extension of CTMs to censored response variables. In the suggested parametrisations of CLTMs, the influence of the explanatory variables on the first two moments of the distribution function is interpretable, and closer insights into model structure can be gained.

For some low-parametrised CLTMs, a likelihood-based estimation approach is presented that can be easily extended to any type of censoring. This approach is especially appealing because the associated maximum likelihood theory comes for free. Alternatively, CLTMs can be estimated based on regularised optimisation using component-wise boosting. This estimation approach is not restricted to low-parametrised CLTMs and can be used for the whole cascade of CLTMs. For applications especially in survival analysis, the target function is extended to right-censored responses by including inverse probability of censoring weights.

The superiority of C(L)TMs in comparison to less flexible standard regression models was shown in two simulation studies. Moreover, two applications of C(L)TMs in biostatistics have been selected for this thesis. The influence of ultrasound measurements on the future birth weight for newborns from the Perinatal Database Erlangen, Germany, has been analysed using low-parametrised likelihood-based CLTMs. Thereby, the birth weights were allowed to follow some arbitrary distribution with fetus-specific means and variances. Additionally, more flexible C(L)TMs have been used to estimate fetus-specific prediction intervals for the future birth weight. In survival analysis, the estimation of patient-specific survivor functions that are conditional on a set of patient characteristics is of special interest. The consideration of CTMs is advisable because the conditional survivor function

can be estimated directly. As an example, CTMs have been used to analyse the survival of patients suffering from chronic myelogenous leukaemia.

Danksagung

Ganz herzlich bedanken möchte ich mich bei ...

... Torsten Hothorn für die sehr engagierte Betreuung dieser Arbeit und die vielen fachlichen Diskussionen. Insbesondere meine zahlreichen Forschungsaufenthalte in Zürich trugen zur Entstehung dieser Arbeit bei.

... Thomas Kneib für die Bereitschaft, diese Arbeit zu begutachten.

... Matthias Schmid für die gute Zusammenarbeit am Geburtsgewichte-Paper.

... Florian Faschingbauer und Markus Pfirrmann für das Bereitstellen der Perinatalen Datenbank Erlangen und der CML–Daten.

... Karen A. Brune für die sprachliche Überarbeitung der Kapitel 4 und 5.

... der Deutschen Forschungsgesellschaft für finanzielle Unterstützung.

... meinen jetzigen und ehemaligen Kollegen in München und Zürich. Ganz besonders bedanke ich mich bei Stephi, Margret, Jona, Stephi, Tina und Giuseppe, die diese Arbeit in Teilen Korrektur gelesen haben.

... Stephi und Margret für viel Unterstützung und Geduld bei der Cover-Gestaltung.

... meiner Familie und meinen Freunden für eure großartige Unterstützung!

Contents

1. Introduction

Most well-known regression models focus only on modelling the conditional mean $\mathbb{E}(Y|\mathbf{X} = \boldsymbol{x})$ of the response variable $Y \in \mathbb{R}$ given a set of explanatory variables $\mathbf{X} = \boldsymbol{x}$. If we look at linear regression models, generalised linear models (GLMs, *e.g.*, McCullagh (1984)), or more generally at generalised additive models (GAMs, *e.g.*, Hastie and Tibshirani (1986)), we are free to choose the parametric response distribution from the exponential family, but we are restricted to model influences of the explanatory variables on the conditional mean. Higher moments of the distribution function are usually assumed as fixed. For example, *in the GLM and GAM models, the variance, skewness and kurtosis are not modelled explicitly in terms of the explanatory variables but implicitly through their dependence on* [the conditional mean] μ (p. 507, Rigby and Stasinopoulos, 2005).

Early attempts to consider higher moments of the distribution function depending on the explanatory variables were made by extending the linear regression model to the linear heteroscedastic regression model (Carroll and Ruppert, 1982). Thereby, the mean and the variance may be influenced by the explanatory variables. However, the variance is assumed to be a parametric function of the mean responses. This link between the mean responses and the variance is especially meaningful for heteroscedastic models, as the variance increases with increasing fitted values (Carroll and Ruppert, 1982).

The idea of letting higher moments of the distribution function depend on the explanatory variables is entirely considered in generalised additive models for location, scale and shape (GAMLSS) (Rigby and Stasinopoulos, 2005). This general semiparametric model class for univariate response variables displays a flexible extension of GAMs. Not only the mean, but also the remaining parameters of the response distribution can be modelled in terms of the explanatory variables via parametric or nonparametric smooth additive functions. Thereby, the response distribution can be selected from a broad family of parametric distributions for continuous or discrete responses. Highly skew or kurtotic distributions are included as well. Hence, the distributions from the exponential family in GLMs and GAMs are replaced by a more general family of distributions. In GAMLSS, estimation is based on (penalised) maximum likelihood approaches (Rigby and Stasinopoulos, 2005). Mayr et al. (2012) present a flexible approach for estimating GAMLSS using a boosting algorithm that is able to deal with high-dimensional data, where the algorithm presented in Rigby and Stasinopoulos (2005) has its limits. This approach also benefits from the boosting characteristics of intrinsic variable selection and model choice.

Nevertheless, all models mentioned so far require the definition of a parametric distribution for the response variable. The definition of a parametric distribution might be a strong assumption, and causes problems if the chosen distribution does not fit the given data properly. A popular approach that makes no assumptions about the parametric distribution of the response variable is quantile regression (Koenker, 2005). Each conditional quantile is modelled separately in terms of the explanatory variables via linear functions (linear quantile regression) or via additive nonparametric and smooth functions (additive quantile regression). A boosting approach for estimating structured additive quantile regression models is presented in Fenske et al. (2011). Due to the modelling of the conditional quantiles in separate regression models, the logical monotonicity of the conditional quantiles is not considered explicitly, and quantile crossing is a familiar problem associated with quantile regression. A nonparametric estimator for conditional quantiles that avoids the problem of quantile crossing is suggested in Dette and Volgushev (2008). Alternatively, Schnabel and Eilers (2013) introduce quantile sheets as a new approach to estimate smooth non-crossing quantile curves. By estimating quantile sheets, all possible quantile curves are estimated simultaneously, and crossing quantile curves are omitted by a sheet that is monotonically increasing with probability τ.

Closely related to estimating the conditional quantile function is the estimation of the conditional distribution function because of the conditional quantile function being the inverse conditional distribution function and vice verse. Some nonparametric kernel-based estimators for the conditional distribution function have been presented in the past. For example, Hall et al. (1999) proposed a weighted Nadaraya-Watson estimator for the conditional distribution function. Alternatively, Li and Racine (2008) proposed a new nonparametric conditional cumulative distribution function kernel estimator that is able to deal with continuous and categorical explanatory variables. Moreover, Li and Racine (2008) suggest to estimate the conditional quantile function by simply inverting the estimated conditional distribution function, and Cai (2002) proposes an estimator for the conditional quantile function based on Hall et al. (1999)'s weighted Nadaraya-Watson estimator.

The direct estimation of the conditional distribution function is an important topic because it is closely related to the estimation of the conditional quantile function and to the estimation of conditional density functions (Hall and Müller, 2003). But more importantly, it combines the useful characteristics of GAMLSS and quantile regression discussed above. First, the influence of the explanatory variables is not restricted to the conditional mean if the whole conditional distribution function is modelled directly. Instead, higher moments of the response distribution such as variance, kurtosis and skewness may be influenced by the explanatory variables as well. This is especially important if typical assumptions such as homoscedasticity and symmetry may be violated (Hothorn et al., 2014). Second, all conditional quantiles are estimated simultaneously when estimating the whole conditional distribution and problems such as quantile crossing cannot occur. In accordance with quantile regression models, the explicit assumption of a parametric response distribution can be avoided. Therefore, Hothorn et al. (2014) introduced recently the semiparametric model class of conditional transformation models (CTMs) that hold the previously mentioned characteristics. In CTMs, the whole conditional distribution function is estimated directly under rather weak assumptions. As the new methods presented in this thesis are related to CTMs or are extensions of CTMs, a short summary of the main characteristics of the model class is given below.

1.1. A short introduction to conditional transformation models

Conditional transformation models (CTMs) (Hothorn et al., 2014) model the conditional distribution function of a response $Y_{\boldsymbol{x}} = (Y|\mathbf{X} = \boldsymbol{x})$ depending on explanatory variables $\boldsymbol{x}$:

$$\mathbb{P}(Y \leq y|\mathbf{X} = \boldsymbol{x}) = F_{Y|\mathbf{X}=\boldsymbol{x}}(y) = F(h(y|\boldsymbol{x})). \tag{1.1}$$

The conditional distribution function is modelled in terms of the monotone transformation function $h : \mathbb{R} \to \mathbb{R}$, which depends on the explanatory variables $\boldsymbol{x}$. Moreover, $y \in \mathbb{R}$ denotes some arbitrary response value from the conditional response distribution, and F denotes an absolute continuous distribution function $F : \mathbb{R} \to [0, 1]$ with corresponding quantile function $Q = F^{-1}$. The transformation function h transforms the response values conditionally on $\boldsymbol{x}$, so that the transformed responses follow the distribution function F. CTMs can be understood as the inverse of a quantile regression model, as we do not model the conditional quantile function but we model the conditional distribution function of the responses directly. Thereby, we are able to estimate all quantiles simultaneously in a joint model and do not need to fit separate models for all quantiles like in quantile regression. When CTMs are estimated, the monotone transformation function h is estimated, whereas the continuous distribution function F is chosen a priori. A frequently used choice is the standard normal distribution function $F = \Phi$ with corresponding quantile function $Q = \Phi^{-1}$. Hence, model characteristics have to be defined in terms of characteristics of the transformation function h.

Furthermore, Hothorn et al. (2014) suggest an additive decomposition of the conditional transformation function h into J partial transformation functions $h_j(\cdot|\boldsymbol{x}) : \mathbb{R} \to \mathbb{R}$:

$$h(y|\boldsymbol{x}) = \sum_{j=1}^{J} h_j(y|\boldsymbol{x}). \tag{1.2}$$

Each partial transformation function h_j depends on (potentially all of) the explanatory variables $\boldsymbol{x}$, whereby higher moments of the conditional distribution function are influenced by the explanatory variables. The conditional transformation function h has to be monotonically increasing in y, but this does not necessarily imply that every partial transformation function h_j has to be monotonically increasing in y.

CTMs are estimated based on regularised optimisation using component-wise boosting resulting in consistent estimated conditional distribution functions (Hothorn et al., 2014). A more thorough introduction to the estimation of CTMs is given in Chapter 3.

Using CTMs, the conditional distribution function is estimated under rather weak assumptions. However, two main assumptions are imposed in CTMs. Most importantly, it is assumed that there is a monotone transformation from the unknown distribution of the responses to the distribution function F, and that this transformation can be displayed in terms of the bijective function h. The distribution function F is fixed and has to be chosen a priori. Second, the additive decomposition of the conditional transformation function into partial transformation functions in Equation (1.2) requires additivity on the scale of the quantile function. Nevertheless, the semiparametric model class of CTMs is by far more flexible compared to alternative standard regression models.

1.2. Scope of this work

In this thesis, we investigate several extensions and modifications of classical CTMs, and use CTMs in two applications where their characteristics are especially beneficial. The methodological emphasis is the introduction of conditionally linear transformation models (CLTMs) as an important special case of CTMs, and the extension of CTMs to censored response variables. Furthermore, the introduction of likelihood-based CTMs is an important methodological development because it creates a broad model basis accompanied by a unified estimation procedure. Additionally, the approach offers the advantageous characteristics of maximum likelihood theory. The fundamental benefits of estimating the conditional distribution function directly have already been discussed at the beginning of this chapter.

Transformation models have been a powerful tool for data analysis in the past. Most prominently, linear transformation models (Cheng et al., 1995) have been applied in survival analysis because the proportional hazards, the proportional odds, and the accelerated failure time model are important special cases of this model class. Linear transformation models are also included as a special case in CTMs. To give a broad overview, we review the development of linear transformation models over the past twenty years in Section 2.1. Furthermore, we put several commonly used regression models into the light of (linear) transformation models to clarify their importance and their wide utilisability. Thereby, one further aim is to clarify that all mentioned regression models have a common model basis, although they might seem different at first sight.

CTMs present a very complex and general model class that often benefits from its high flexibility. Nevertheless, due to this high flexibility model interpretation can be challenging. For example, a direct interpretation of the relationship between the explanatory variables and certain moments of the distribution function of the response is difficult to obtain. Therefore, we introduce conditionally linear transformation models (CLTMs) as an important special case of CTMs with reduced flexibility in this thesis. In CLTMs, the influence of the explanatory variables is restricted to the conditional mean and the conditional variance of the transformed response. In consequence, the explanatory variables' influence on the first two moments of the distribution function is interpretable. Additionally, the interpretable parametrisations in CLTMs allow closer insights into model structure. Concerning their complexity, CLTMs can be placed in between CTMs, which are more flexible, and less flexible linear transformation models. We introduce a cascade of interpretable, low-parametrised CLTMs in Chapter 2. Furthermore, the model class of CLTMs is introduced more generally for the estimation of prediction intervals in Chapter 4. The considered CLTMs include not only low-parametrised but also more complex models.

Estimation of C(L)TMs can be based on two fundamentally different approaches, which are introduced in Chapter 3. Hothorn et al. (2014) presented a component-wise boosting algorithm for the estimation of CTMs. This approach is adapted to the estimation of CLTMs in Chapter 4, and to right-censored responses in Chapter 5. On the other hand, low-parametrised CLTMs can be estimated likelihood-based, and this approach can be easily extended to any kind of censoring. Hence, the likelihood-based approach is not restricted to the analysis of right-censored observations, but also left-censored, doubly-censored or interval-censored response variables can be considered. For uncensored responses, the performance of likelihood-based C(L)TMs is investigated in terms of a simulation study in Chapter 6. Additionally, likelihood-based CLTMs are used for the analysis of the conditional distribution function of birth weight depending on ultrasound measurements for newborns in the Perinatal Database Erlangen in Chapter 7.

The usage of standard regression models is often associated with rather strict assumptions. For example, the linear regression model implies symmetry and homoscedasticity assumptions, and these assumptions transfer directly to functionals of the conditional distribution function. One frequently used functional of the conditional distribution function is the determination of prediction intervals. Using the example of linear regression models again, the resulting prediction intervals are symmetric around the conditional mean, and the interval length is constant because it is independent of the individual explanatory variables. Hence, the prediction intervals perform poorly in the presence of heteroscedasticity and skewness (Hothorn et al., 2014). In consequence, the direct estimation of the conditional distribution function under rather weak assumptions, *i.e.* characteristics such as heteroscedasticity and skewness can be identified, is very useful for the determination of prediction intervals. Hence, CTMs have the potential to determine prediction intervals more carefully and might outperform standard prediction intervals if higher moments of the distribution function depend on the explanatory variables. To confirm the benefits of estimating prediction intervals using CTMs, we analyse the future birth weight of newborns from the Perinatal Database Erlangen, Germany, in Chapter 4. A cascade of CLTMs of different model complexity is used to predict the future birth weight depending on a set of ultrasound measurements, and associated fetus-specific prediction intervals are presented. The quality of the prediction intervals is compared to prediction intervals resulting from linear regression models and from quantile regression.

Closely connected to the estimation of the conditional distribution function is the estimation of the conditional survivor function in survival analysis. More precisely, an estimate of the conditional distribution function provides directly an estimate of the conditional survivor function. Hence, extending CTMs to deal with censored observations (especially right-censored observations) is a very interesting and important topic. Compared to standard regression models in survival analysis, CTMs have several important advantages. First, most regression models used in survival analysis focus only on the estimation of hazard functions and on summary statistics, and the conditional survivor function is seldom estimated directly. Nevertheless, the estimation of patient-specific survivor functions is of special interest in personalised medicine because individual patient risk profiles are especially informative, and allow a better prognosis of the course of the disease (Mackillop and Quirt, 1997; Crowther and Lambert, 2014). Second, the proportional hazards model (which is the most frequently used regression model in survival analysis) implies the strict assumption of proportional hazards. Of course, the proportional hazards assumption can be checked and relaxed (*e.g.*, using residuals (Schoenfeld, 1982)), but this is usually rather circumstantial and model diagnosis can be only performed after model estimation. In contrast, CTMs estimate the conditional survivor function directly and flexibly, and proportional as well as non-proportional hazard models are included as special cases in the broad model class. In Chapter 5, we investigate CTMs for survivor function estimation. Therefore, CTMs are extended to right-censored observations by including inverse probability of censoring weights (*e.g.*, Van der Laan and Robins, 2003) into the target function. The resulting censored integrated log score is minimised in terms of the component-wise boosting algorithm presented in Hothorn et al. (2014).

1.3. Contributions

The work in this thesis has been mainly influenced by common research and vital discussions with my supervisor Torsten Hothorn, and it is partly based on collaborations with colleagues and researchers from related special fields. Parts of this thesis are already published or submitted as journal articles, and the remaining parts are based on yet unpublished manuscripts. The outline given below lists the titles of the manuscripts, gives a short summary of the content, and highlights the contributions from all authors.

- **Chapter 2**, **Chapter 3**, **Chapter 6** and **Chapter 7**:
  The content of these chapters is based on the working paper

  Möst, L. and T. Hothorn (2014b). Likelihood-based conditional transformation models. *Working paper.*

  In this yet unpublished manuscript, we introduce a unified likelihood-based estimation approach for low-parametrised conditional transformation models. Furthermore, it contains a broad overview of literature on linear transformation models.

Lisa Möst and Torsten Hothorn developed the idea of a likelihood-based estimation approach for conditional transformation models. Lisa Möst looked through the literature on linear transformation models currently available, and reviewed commonly used regression models from the perspective of conditional transformation models. She furthermore performed all analyses, planned and performed all simulations, and wrote the manuscript. Torsten Hothorn contributed to the conception and presentation of the article.

- **Chapter 4**:
  The content of Chapter 4 is already published in

  Möst, L., M. Schmid, F. Faschingbauer, and T. Hothorn (2014). Predicting birth weight with conditionally linear transformation models. *Statistical Methods in Medical Research. To appear.* DOI: 10.1177/0962280214532745.

  This manuscript suggests the usage of conditionally linear transformation models for predicting the future birth weight of newborns from the Perinatal Database Erlangen, Germany, based on prenatal ultrasound measurements. Conditionally linear transformation models are especially useful for the determination of prediction intervals.
  Lisa Möst and Torsten Hothorn introduced the model class of conditionally linear transformation models (CLTMs) and developed the idea of using CLTMs for the determination of prediction intervals. Lisa Möst conducted all analyses and preliminary simulations, and drafted the manuscript. Torsten Hothorn extended the R add-on package **ctmDevel** (Hothorn, 2013) to deal with CLTMs, and contributed to the conception, the presentation and the revision of the article. Matthias Schmid contributed to the Introduction, especially to the parts concerning medical expert knowledge, and reviewed the literature on birth weight prediction and obstetric management. Furthermore, he wrote the part of the Discussion concerning reference growth charts, and he contributed to the conception and the revision of the article. Florian Faschingbauer supplied us with the Perinatal Database Erlangen and with literature on birth weight prediction and obstetric management.

- **Chapter 5:**
  Chapter 5 is already published in

  Möst, L. and T. Hothorn (2015). Conditional transformation models for survivor function estimation. *International Journal of Biostatistics. To appear.* DOI: 10.1515/ijb-2014-0006.

  We suggest the direct estimation of patient-specific survival probabilities over time using conditional transformation models in this manuscript. The proposed methodology is able to deal with proportional as well as non-proportional hazard settings in survival analysis.
  Lisa Möst and Torsten Hothorn developed the concept of using conditional transformation models for survivor function estimation. Therefore, they extended the

> component-wise boosting algorithm for conditional transformation models to right-censored responses by including inverse probability of censoring weights. Lisa Möst planned and conducted the simulation study, analysed the data set of patients suffering from chronic myelogenous leukaemia, and wrote the manuscript. Torsten Hothorn provided the necessary software (R add-on package **ctmDevel** (Hothorn, 2013)), and contributed to the conception, the presentation and the revision of the manuscript.

The respective manuscripts will be cited again at the beginning of each chapter. Afterwards, I avoid the repeated citation although there are textual matches.

1.4. Software

All analyses were carried out in the R system of statistical computing (R Core Team, 2014). Estimation of CTMs and CLTMs using a component-wise boosting algorithm (Chapter 4 and Chapter 5) was performed using the R add-on package **ctmDevel** (Hothorn, 2013). The likelihood-based estimation of C(L)TMs (Chapter 6 and Chapter 7) was performed using the `constrOptim`-function from the **stats**-package, which is an extension of the `optim`-function that is able to consider linear constraints. A more thorough summary of the used R add-on packages for the various analyses and links to tutorial R examples can be found in each chapter.

2. Conditional transformation models

The content of this chapter is based on Möst and Hothorn (2014).

In the past, linear transformation models have been a powerful tool for data analysis in various fields. Most commonly, they have been used in the context of survival analysis, as the proportional hazards model and the proportional odds model are prominent members of this model class. Over the years, numerous ways to estimate linear transformation models have been proposed including estimating equations, marginal likelihoods, or nonparametric likelihoods, to name just a few. The handling of censored observations and the associated complications in estimation have been an important topic as well. Several extensions of the ordinary linear transformation model, *e.g.*, to non-linear or random effects, have been discussed extensively.

In order to give a general overview, we review the history of linear transformation models over the past twenty years with special focus on model structure and associated estimation strategies in this chapter. To display the large variety of transformation models, we examine several important regression models (*e.g.*, proportional hazards (PH) and proportional odds (PO) models, accelerated failure time (AFT) models, cumulative regression models for ordinal responses, Box-Cox transformation models) from the perspective of (linear) transformation models. Our aim is to clarify the common model basis of the considered regression models, even though they might seem rather different.

Conditional transformation models (CTMs) (Hothorn et al., 2014) are a more complex model class, which allows the estimation of the whole conditional distribution function of a response variable given a set of explanatory variables. The model class of CTMs includes linear transformation models as a special case. For reasons of interpretability and manageability, we introduce low-parametrised conditionally linear transformation models (CLTMs), which represent an important special case of CTMs that is highly relevant for various applications. We relate CTMs, CLTMs, and linear transformation models to each other in this chapter, and discuss the associated model characteristics and limitations.

2.1. A review of transformation models

Transformation models are a powerful tool for data analysis in various fields, and thus have been frequently used in the past. The history of transformation models started with the parametric response transformations suggested by Box and Cox (1964). Due to the well-explored nature of ordinary linear regression models, the authors proposed to transform the response y such that a normal, homoscedastic, linear model is valid after transformation. For this response transformation, a family of transformations depending on the parameter λ was established:

$$h_Y(y|\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(y), & \lambda = 0. \end{cases} \tag{2.1}$$

A normal distribution with mean μ and variance σ^2 is assumed for the response after the transformation. Due to the finite number of parameters, Box-Cox transformation models can be estimated using a full likelihood approach.

Later, the introduction of linear transformation models (*e.g.*, Cheng et al., 1995; Chen et al., 2002) displayed an extension of the parametric Box-Cox transformation models. Instead of specifying the transformation function $h_Y(y|\lambda)$ up to a finite-dimensional parameter λ, the response transformation $h_Y(y)$ is left unspecified, *i.e.*

$$h_Y(y) = -\boldsymbol{x}^\top \boldsymbol{\beta} + \epsilon, \tag{2.2}$$

where ϵ is a random error with completely specified distribution function F. In this semi-parametric approach for linear transformation models, the unknown, strictly increasing response transformation $h_Y(y)$ is related to linear covariate effects. The corresponding conditional distribution function is

$$\mathbb{P}(Y \leq y|\mathbf{X} = \boldsymbol{x}) = F(h_Y(y) + \boldsymbol{x}^\top \boldsymbol{\beta}). \tag{2.3}$$

As a consequence, the explanatory variables induce linear shifts of the response distribution, and thereby only the conditional mean of the transformed response is influenced. Estimating linear transformation models involves estimating the parameter vector $\boldsymbol{\beta}$ and the monotone transformation function $h_Y(\cdot) : \mathbb{R} \to \mathbb{R}$. Model complexity is considerably restricted in linear transformation models due to the avoidance of interaction terms between the response y and the explanatory variables $\boldsymbol{x}$. As linear transformation models include the proportional hazards (PH) and the proportional odds (PO) model as important special cases (Doksum and Gasko, 1990, Section 2.1.2), the response transformation function $h_Y(y)$ is sometimes also termed *baseline function*.

Recently, Hothorn et al. (2014) introduced conditional transformation models (CTMs), which display an extension of linear transformation models and include the model class as a special case. The model class of CTMs allows the estimation of the whole conditional distribution function of a response variable Y given a set of explanatory variables $\mathbf{X} = \boldsymbol{x}$:

$$\mathbb{P}(Y \leq y|\mathbf{X} = \boldsymbol{x}) = F(h(y|\boldsymbol{x})). \tag{2.4}$$

Thereby, $h(\cdot|\boldsymbol{x}) : \mathbb{R} \to \mathbb{R}$ denotes the conditional transformation function, which is allowed to be arbitrarily flexible, but has to be monotonically increasing in y. Model complexity is considerably extended, as the transformation function h depends on y and $\boldsymbol{x}$ simultaneously. For a more detailed introduction to CTMs, see Section 1.1.

The literature on transformation models in the past twenty years is mainly dominated by references on linear transformation models, extensions of linear transformation models (*e.g.*, to non-linear or random effects) and the associated estimation techniques. Furthermore, the handling of censored observations has been an important topic. To clarify the versatile application area of linear transformation models, and to summarise the various estimation strategies, we give a broad literature overview of linear transformation models hereafter.

2.1.1. Estimation strategies for linear transformation models

Concerning linear transformation models, various estimation strategies have been proposed for the vector of regression coefficients $\boldsymbol{\beta}$ and the monotonically increasing response transformation function $h_Y(y)$ (Equation (2.2)) comprising estimating equations, partial likelihoods, marginal likelihoods, nonparametric likelihoods, and Bayesian approaches. The most important strategies are summarised below, and the respective advantages and disadvantages are outlined shortly.

Estimating equations. The early references that deal with linear transformation models for possibly censored responses suggest an estimation based on estimating equations. Most approaches have in common that they suggest an estimating equation, which allows the estimation of the parameter vector $\boldsymbol{\beta}$ *without* estimation of the response transformation function $h_Y(y)$ (Equation (2.2)). Thereby, the infinite-dimensional parameter associated with the estimation of $h_Y(y)$ is treated as a nuisance parameter, and the explicit estimation of the response transformation function is avoided. Such approaches are henceforth referred to as baseline-free. This might be of advantage if only the linear effects of the explanatory variables on the response transformation are of interest. But the estimated parameters $\hat{\boldsymbol{\beta}}$ can only be interpreted easily in PH and PO models, whereas interpretation is not straightforward with a general error distribution F (Fine et al., 1998). Additionally, *e.g.*, in the case of survival times, the prediction of the individual survival probabilities over time given the patient's prognostic information requires the estimation of $\boldsymbol{\beta}$ *and* h_Y. Therefore, most authors provide an additional estimating equation for h_Y, where h_Y is assumed to be a non-decreasing step function with jumps at the observed failure times. The most important estimating equation approaches are shortly summarised below.
A class of estimating functions for possibly right-censored linear transformation models is introduced by Cheng et al. (1995). The authors derive simple estimating functions for the parameter vector $\boldsymbol{\beta}$ based on the dichotomous variables $\{I(Y_i \geq Y_j) : i \neq j,\, i, j = 1, \dots, n\}$, I denotes the indicator function. The usage of pairwise dichotomous comparisons of response values supersedes the estimation of h_Y due to the monotone transformation being rank-invariant. Nevertheless, the authors supply an estimating function for h_Y later, and present procedures to predict the survival probabilities of future patients accompanied with

pointwise and simultaneous confidence intervals (Cheng et al., 1997).
The estimates of Cheng et al. (1995) are asymptotically biased when the support of the censoring variable is shorter than the support of the response. Therefore, Fine et al. (1998) propose simple modifications of the estimating functions to obtain consistent estimates. The root-finding estimating equation of Cheng et al. (1995) is furthermore replaced by a least-squares criterion. This least-squares criterion is adapted in Fan and Fine (2013) such that the linear transformation model might include parametric covariate transformations. Cai et al. (2000) extend linear transformation models for independent event times to correlated failure time observations, *i.e.* to clustered failure time data. The authors present estimating equations to estimate $\boldsymbol{\beta}$ and h_Y simultaneously, and predicting procedures for survival probabilities. Furthermore, Cai et al. (2002) propose a linear transformation model with random effects for clustered and possibly censored failure time data, which includes the frailty Cox model as a special case. The distribution of the random effects is completely specified up to a non-negative scale parameter γ. Inference for $\boldsymbol{\beta}$, γ and h_Y is obtained using estimating equations similar to Cheng et al. (1995) and Fine et al. (1998) based on intra- and intercluster comparisons of failure times. The approach is extended to doubly censored observations in Shen (2012b).
A procedure to estimate $\boldsymbol{\beta}$ in the presence of interval-censored data is developed in Zhang et al. (2005). The proposed estimating equations are similar to Cheng et al. (1995) because inference about the regression parameters is based on rank information and the estimation of h_Y is avoided. A method for survival probability prediction and model checking techniques for the error distribution F are considered later in Zhang (2009). The approach can be extended to interval-censored and doubly truncated data (Shen, 2013). In econometrics, Lee (2008) considers a general class of semiparametric transformation models with random effects and completely known error distribution F for panel data. An estimating equation for $\boldsymbol{\beta}$ is proposed by modifying the estimating equations of Cheng et al. (1995), where weights account for dependent right-censoring that is common for panel data.
All estimation procedures mentioned so far rely on the (strong) assumption that the covariates and the censoring variable are independent because the marginal Kaplan-Meier estimator is used to estimate the censoring distribution. The assumption can only be relaxed for a finite number of possible values of the covariate vector, when conditional Kaplan-Meier estimators can be applied. Because this assumption is too restrictive in many applications, and because Cox's partial likelihood estimator does not assume independence of the covariates and the censoring variable, Chen et al. (2002) propose a new unified estimation procedure for the analysis of censored data using linear transformation models that avoids modelling the censoring distribution. The proposed martingale-based estimating equations for $\boldsymbol{\beta}$ and h_Y are easily implemented numerically. The inference procedure for linear transformation models is reliable, and in case of the PH model the estimator is equivalent to Cox's partial likelihood estimator. Both martingale equations are solved alternatingly using an iterative algorithm.
The estimating equations proposed by Chen et al. (2002) have been extended in various directions afterwards. Lu (2005) considers multivariate event times including study subjects that may experience several types of events, study subjects that experience recurrences of

the same type of event, or clustered event times. Furthermore, the estimating equations are adapted to the case-cohort design in Lu and Tsiatis (2006) by introducing weighted estimating equations. Liu and Ying (2007) combine a normal transformation model and a linear mixed effects model via latent random variables to analyse longitudinal data subject to informative right censoring. In Lu and Zhang (2010), the estimating equations of Chen et al. (2002) are generalised to incorporate non-linear covariate effects, what results in partially linear transformation models. Parametric covariate effects as well as h_Y are estimated using global estimating equations, and nonparametric covariate effects are estimated using kernel-weighted local estimating equations. Variable selection in linear transformation models is considered by Zhang et al. (2010) who present an approach for sparse and consistent estimation. A profiled score is derived from the estimating equations of Chen et al. (2002). A loss function is constructed from this profiled score, and the loss function is finally minimised using a LASSO shrinkage penalty for $\boldsymbol{\beta}$. Motivated by the PO model, Scheike (2006) proposes martingale-based estimating functions, where the linear transformation model is extended to baseline functions that depend on the covariates. The estimating equations of Cheng et al. (1995) and Chen et al. (2002) can also be used to analyse left-truncated right-censored or doubly censored observations (Shen, 2012a). Doubly censored data are also analysed in Cai and Cheng (2004), where the error distribution F is only specified up to a finite dimensional parameter using a Box-Cox-type family of distributions.

The estimated parameter vector $\hat{\boldsymbol{\beta}}$ resulting from the estimating equations proposed by Cheng et al. (1995), Chen et al. (2002), and Zhang et al. (2005) is shown to be consistent and asymptotically normal distributed. An explicit formula for the variance-covariance matrix of the limiting distribution is given, which can be estimated consistently using the plug-in method. Nevertheless, *e.g.*, the estimating equations of Chen et al. (2002) are only efficient for the PH model and loose efficiency for alternative error distributions. Therefore, *empirical likelihood* inference procedures for censored survival data under the linear transformation model are proposed by Lu and Liang (2006), Zhao (2010), Yu et al. (2011), and Zhang and Zhao (2013). The limiting distribution of the empirical likelihood ratio test statistic is more appropriate than the normal approximation in various situations, and problems of under-coverage of associated confidence regions are solved.

Partial likelihood approaches. The popularity of the Cox model (Cox, 1972) is mainly due to the easy interpretation of the regression coefficients in terms of hazard ratios, and to simplified estimation based on the partial likelihood. The estimation of $\boldsymbol{\beta}$ is usually of primary interest. The baseline hazard function is considered a nuisance parameter of high dimensionality, and thus its estimation is omitted (see also Section 2.1.2). Therefore, Cox (1975) proposed the partial likelihood to reduce the dimensionality in situations with many nuisance parameters, where maximum likelihood as a general technique usually fails. If the baseline hazard function is also of interest, an estimate can, *e.g.*, be derived using the Breslow estimator (Breslow, 1972). The partial likelihood can be extended to time-dependent regression coefficients, what can be useful to detect violations of the PH assumption. For example, estimation is based on local partial likelihood techniques in Cai and Sun (2003), where local linear fitting techniques known from scatterplot smoothing are connected with the partial likelihood. Empirical pointwise confidence intervals and simultaneous confidence bands for the time-dependent coefficients based on local partial likelihood smoothing are derived by Sun et al. (2009). The sum of a weighted negative partial log-likelihood and an adaptive LASSO penalty is minimised by Liu and Zeng (2013) to perform variable selection in semiparametric transformation models for right-censored data. The target function includes Cox's partial log-likelihood as a special case. A baseline-free approach for longitudinal data is proposed by Wu et al. (2010). The authors suggest time-varying transformation models for modelling the conditional cumulative distribution function of a response variable. A two-step smoothing method is developed to estimate the time-varying parameters and the approach can be extended to censored observations.

Marginal likelihood approaches. Considering the marginal likelihood yields a further baseline-free approach. The response transformation h_Y is eliminated by using only the rank order information of the responses instead of the exact response values. In Gu et al. (2005) a class of semiparametric transformation models under interval censoring is considered, which is more general than the linear transformation model including, *e.g.*, frailty models, heteroscedastic hazard regression models, and heteroscedastic rank regression models. The baseline-free estimation based on the marginal log-likelihood results in the maximum marginal likelihood estimator (MMLE). A three-stage MCMC stochastic approximation algorithm is presented for model estimation because the marginal log-likelihood usually has no closed analytic expression, and involves integrals of high dimension. The asymptotic properties of the MMLE are determined by Gu et al. (2014). Concerning the PO model, Pettitt (1984) minimises a marginal likelihood based on ranks. The regression coefficients in a semiparametric PO model are estimated using only the rank order information among patients by Lam and Leung (2001). For this purpose, a Monte Carlo method is used to approximate the marginal likelihood function of the rank invariant transformation of the survival times. As an extension, a semiparametric random effects PO model to analyse multivariate survival data with various types of dependence structures is proposed by Lam et al. (2002). The approach covers cluster data and repeated measurements. Multivariate normal random effects are assumed. Estimation of the regression and variance parameters is achieved by maximising a marginal rank likelihood. As an extension of Gu et al. (2005), Li et al. (2012) include a varying-coefficient component into the linear transformation model. The coefficient functions are approximated using B-splines, and estimation is based on maximising the marginal rank log-likelihood.

Parametric likelihood approaches. Parametric approaches for the estimation of linear transformation models usually imply specifying h_Y up to a finite-dimensional parameter. Such transformation functions are, *e.g.*, given in Box and Cox (1964) (Equation (2.1)), Bickel and Doksum (1981), and MacKinnon and Magee (1990). The great advantage of parametric approaches is that model estimation can be carried out using the full log-

likelihood function due to the limited number of parameters. Hence, no penalisation approaches, approximations, tuning parameter selection (*e.g.*, smoothing parameters, kernel bandwidths), etc. are needed. Although there are many suggestions for parametric transformations of the response variable, every parametric family has characteristics that are unsuitable in certain applications (MacKinnon and Magee, 1990). This is usually due to the associated restricted flexibility of the response transformation h_Y.
A parametric approach for estimating linear transformation models for time-to-event data subject to arbitrary censoring is proposed by Zhang and Davidian (2008). Thereby, baseline survival densities are approximated by a truncated series expansion. This results in likelihood-based inference associated with an adaptive choice of the degree of truncation. An alternative parametric regression model for right-censored data is the transform both sides (TBS) model, which is an extension of the Box-Cox power family (Polpo et al., 2014). Model estimation can be either carried out by maximising the right-censored log-likelihood function, or using Bayesian methods (see also Lin et al., 2012).
Estimation of parametric accelerated failure time (AFT) models has been, *e.g.*, introduced by Kalbfleisch and Prentice (1980), and Cox and Oakes (1984). Moreover, estimation of the parametric exponential PH model (Prentice, 1973), the Weibull model (Prentice, 1973), and the Pareto PH model (Davis and Feldstein, 1979) is also considered by Kay (1977), and Klein and Moeschberger (2003). All models can be estimated based on a full likelihood approach due to the finite number of parameters. For the connection of parametric AFT and parametric PH models to linear transformation models, see Section 2.1.2.

Nonparametric likelihood approaches. If no parametric assumptions are made about the response transformation h_Y, nonparametric estimation procedures have to be considered. A nonparametric estimation has the advantage that the form of h_Y is not restricted, and arbitrarily flexible forms of h_Y can be displayed. This flexibility usually comes at the price of a high-dimensional parameter vector that needs to be estimated. Nevertheless, some assumptions have to be imposed on the form of h_Y to guarantee a feasible estimation. These assumptions turn the infinite-dimensional parameter associated with the flexible response transformation h_Y into a finite number of regression parameters, and are often associated with the selection of tuning parameters. Due to the resulting finite number of parameters, $\boldsymbol{\beta}$ and h_Y can be estimated simultaneously.
A popular approach is to convert the nonparametric into a parametric estimation task by considering the nonparametric maximum likelihood estimator (NPMLE). Thereby, the transformation function h_Y is approximated by a non-decreasing step function with jumps at the observed failure times. Afterwards, a full likelihood approach is used to estimate the covariate effects $\boldsymbol{\beta}$ and the jump sizes h_{Yi}. Nevertheless, the number of parameters increases with the number of distinct event times. This could cause problems during parameter estimation because direct maximisation of the likelihood might become impossible. The NPMLE has often been considered for estimating linear transformation models: Slud and Vonta (2004) study the large-sample consistency of NPMLEs for an unknown baseline continuous cumulative-hazard type function; Chen and Tong (2010) extend linear trans-

formation models by smooth varying-coefficient terms and propose an iterative algorithm for likelihood maximisation; a semiparametric transformation frailty model for nonproportional hazards is estimated using NPMLE in Choi and Huang (2012); and the NPMLE approach is used in linear transformation models for multivariate interval-censored data (Chen et al., 2013), and current-status data (Zhang et al., 2013). Zeng and Lin (2007b) propose a very general class of transformation models for counting processes including linear transformation models, models with crossing hazards, and time-varying covariates. Afterwards, this class is further extended to dependent failure time data. Simple and stable numerical techniques to obtain the NPMLE and a very general asymptotic theory are developed. NPMLE approaches for the PO model are considered in Bennett (1983a), Murphy et al. (1997), Hunter and Lange (2002), and Chen et al. (2012). An efficient algorithm for computing NPMLEs in linear transformation models is presented in Yin and Zeng (2006). The estimation of a high-dimensional parameter is avoided because the algorithm obtains estimates by solving a finite number of equations. This is due to a dramatic reduction of the parameter space via a reparametrisation of the baseline function.

Furthermore, some mixed model approaches can be found in the literature, where estimation is based on NPML. A semiparametric PO model with random effects is presented by Zeng et al. (2005). The parameter vector $\boldsymbol{\beta}$, the variance parameters for the multivariate normal random effects, and the baseline odds function are obtained via NPML methods. This approach is further generalised to semiparametric transformation models with random effects by Zeng et al. (2008). Kosorok et al. (2004) consider a semiparametric frailty Cox model, where the frailties follow a known one-parameter family of distributions. The regression parameters, the frailty parameter, and the baseline hazard are estimated jointly. A similar approach is suggested by Huber-Carol and Vonta (2004), who consider frailty Cox models to account for possible heterogeneity among the population for arbitrarily censored and truncated data. The usual gamma frailty model is generalised to multivariate failure times by Zeng et al. (2009), where marginal linear transformation models are formulated for each type of event. The NPMLEs are obtained using the EM-algorithm and treating the random effects as missing data.

In addition, some further nonparametric estimation approaches are suggested: Zhao et al. (2007) estimate the nonparametric transformation of the survival times in PH models using local linear approximations and locally weighted least squares. The response transformation h_Y is estimated nonparametrically by considering the baseline hazard function as a log-linear spline by Cai and Betensky (2003). In this approach, the spline coefficients are treated as random effects to ensure a smooth function estimate, and estimation is based on the integrated penalised log-likelihood. Zeng and Lin (2007a) propose an approximate nonparametric maximum likelihood method for the AFT model with possibly time-dependent covariates.

Bayesian approaches. Alternatively, linear transformation models can be estimated using Bayesian approaches. Thereby, $\boldsymbol{\beta}$, h_Y, and F are assumed to be unknown, and are associated with appropriate prior distributions. Estimation is based on a MCMC algorithm.

As all unknown model components are estimated simultaneously, the prediction of survival probabilities is straightforward.
For example, Mallick and Walker (2003) propose a Bayesian semiparametric transformation model that includes the PH, the PO, and the AFT model (Walker and Mallick, 1999) as special cases, and can be extended to multivariate survival data using frailty models. A multivariate normal prior is specified for $\boldsymbol{\beta}$, a mixture of incomplete beta functions prior is assumed for h_Y, and a Pòlya tree prior is assumed for F. A Bayesian semiparametric PO model is introduced in Hanson and Yang (2007), where a mixture of finite Pòlya trees prior is assumed for the baseline survival function. In Banerjee et al. (2007) the Bayesian analysis of a class of generalised odds-rate hazards (GORH) models is considered, which includes non-proportional hazards, PO, PH, and AFT models as special cases. The general class of gamma frailty transformation models for multivariate survival data presented in de Castro et al. (2014) can be understood as a multivariate extension of Banerjee et al. (2007).

Further extensions. The error distribution F in Equation (2.2) was assumed to be completely specified so far. Although we assume F to be known throughout this thesis, we would also like to mention approaches where the error distribution is assumed to be unknown, or only specified up to a finite dimensional parameter ρ. In Zucker and Yang (2006), the authors consider a family of survival models where the error distribution belongs to the Box-Cox family of transformations depending on the one-dimensional parameter ρ. Estimation is based on a pseudo-likelihood estimator or a martingale residual estimator (Yang and Prentice, 1999). Linton et al. (2008) consider the case where the transformation function h_Y is parametric, but the error distribution F and the covariate effects are nonparametric. The model is only identifiable up to a couple of normalisations under smoothness constraints for h_Y, F, and the covariate effects, and monotonicity constraints for h_Y and F. A nonparametric transformation model where the response transformation h_Y and the error distribution F are unspecified is considered by Song et al. (2007). Estimation of the parameter vector is based on the smoothed partial rank estimator. Khan and Tamer (2007) propose the partial rank estimator for general forms of censoring.

In Horowitz (2009), the author deals with linear transformation models in econometrics. Models of the form of Equation (2.2) are used in applied econometrics for the analysis of duration data or for the estimation of hedonic price functions. If h_Y and F are known (up to finite-dimensional parameters), maximum likelihood methods are used for estimation. If h_Y is parametric (depending on some parameter α) and F is nonparametric, F can be estimated based on the empirical distribution function of the residuals, and $\boldsymbol{\beta}$ and α can be estimated using the generalised method of moments. The most flexible case of linear transformation models with nonparametric h_Y and F is similar to semiparametric single index models. Therefore, methods for single index models are suggested for the estimation of $\boldsymbol{\beta}$, and estimators for F and h_Y are additionally provided.

Semiparametric inference methods for AFT models, where the error distribution F is left unspecified (for more information on AFT models see Section 2.1.2) are, *e.g.*, considered in Prentice (1978), and Buckley and James (1979), where linear rank statistics or modified least squares normal equations are suggested. Tsiatis (1990), Ritov (1990), and Lai and Ying (1992) suggest rank-based procedures that are essentially derived from the partial likelihood principle (Wei, 1992), and Jin et al. (2003) develop a broad class of rank-based monotone estimating equations for the semiparametric AFT model.

2.1.2. Well-known regression models reviewed from the transformation model perspective

The broad model class that is provided by transformation models could already be perceived from our literature review. Transformation models have been applied in various fields, and linear transformation models can be usefully extended by including random, non-linear, or response-varying effects. Most commonly, the PH, the PO, and the AFT model are mentioned as special cases of linear transformation models in the literature, but there are a lot more regression models that can be reviewed from the perspective of (linear) transformation models (Equation (2.3)). To give a more complete list, and to clarify the common model basis of a broad range of regression models, we integrate a selection of commonly used regression models into the transformation model context (see Figure 2.1, Figure 2.2, and Figure 2.3 for a graphical overview).

Continuous responses

Proportional hazards (PH) model. The PH model or *Cox model* (Cox, 1972, 1975) is the regression model most commonly used in survival analysis. Nevertheless, the model class is not restricted to survival times, and might be useful for a non-negative (censored) response variable Y in general. The Cox model can be expressed as a linear transformation model (Doksum and Gasko, 1990)

$$\mathbb{P}(Y \leq y|\mathbf{X} = \boldsymbol{x}) = \mathcal{M}(h_Y(y) + \boldsymbol{x}^\top \boldsymbol{\beta}),$$

where $\mathcal{M}$ denotes the minimum-extreme value distribution function, and the response transformation equals the logarithm of the cumulative baseline hazard function, $h_Y(y) = \log(\Lambda_0(y))$, $\Lambda_0(y) = \int_0^y \lambda_0(u)\,du$. The PH assumption is due to the fact that no interaction terms between the covariates $\boldsymbol{x}$ and the response y are intended. Consequently, the explanatory variables influence only the conditional mean of the transformed response variable. This assumption can be easily relaxed by including interaction terms between $\boldsymbol{x}$ and y into the model equation, what results in a non-proportional hazards model. Probably the most familiar non-proportional hazards model is the Cox model with time-varying effects (*e.g.*, Zucker and Karr, 1990).

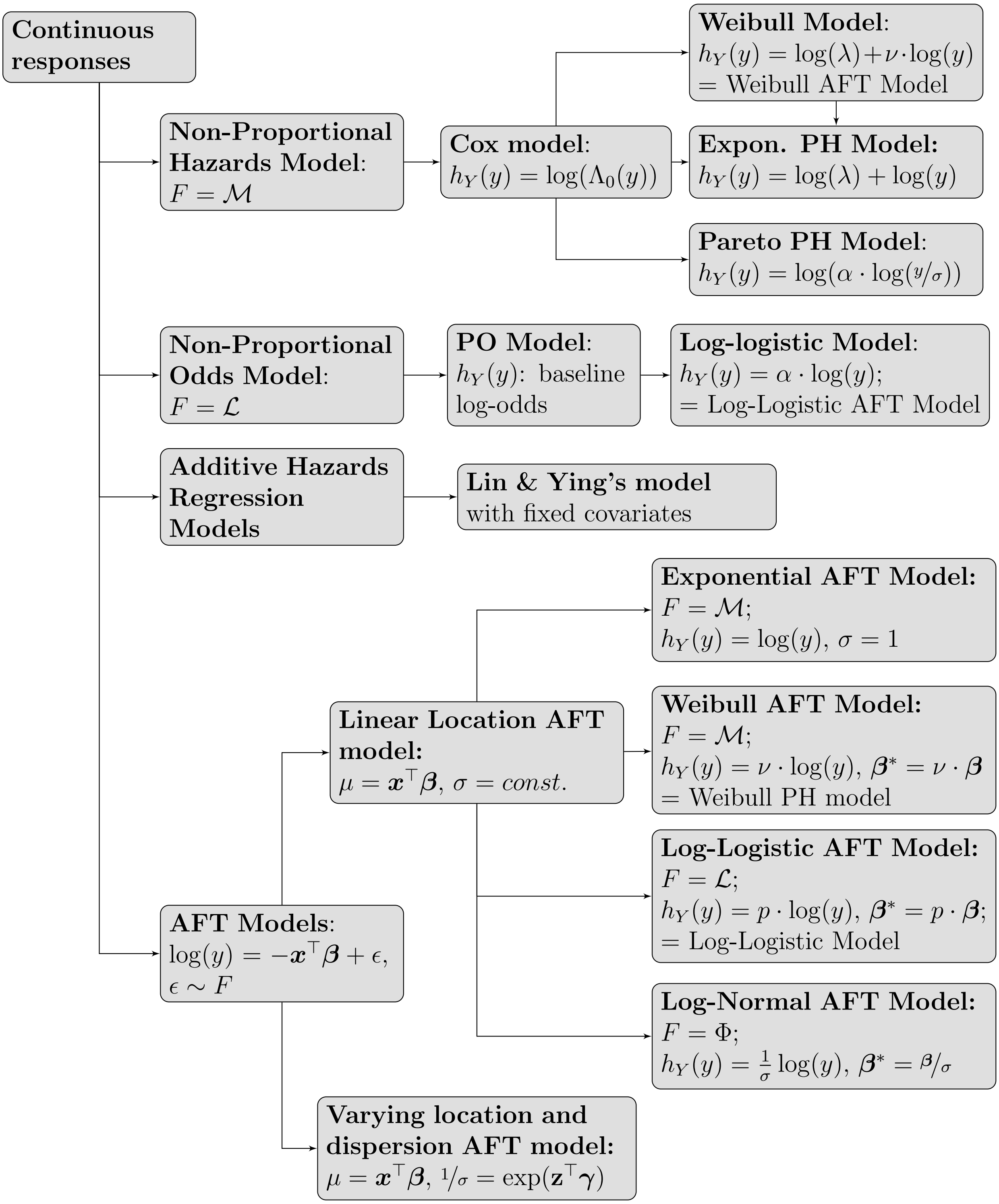

Continuous responses
Non-Proportional Hazards Model: $F = \mathcal{M}$
Cox model: $h_Y(y) = \log(\Lambda_0(y))$
Weibull Model: $h_Y(y) = \log(\lambda) + \nu \cdot \log(y)$ = Weibull AFT Model
Expon. PH Model: $h_Y(y) = \log(\lambda) + \log(y)$
Pareto PH Model: $h_Y(y) = \log(\alpha \cdot \log(y/\sigma))$
Non-Proportional Odds Model: $F = \mathcal{L}$
PO Model: $h_Y(y)$: baseline log-odds
Log-logistic Model: $h_Y(y) = \alpha \cdot \log(y)$; = Log-Logistic AFT Model
Additive Hazards Regression Models
Lin & Ying's model with fixed covariates
Exponential AFT Model: $F = \mathcal{M}$; $h_Y(y) = \log(y)$, $\sigma = 1$
Linear Location AFT model: $\mu = \boldsymbol{x}^\top\boldsymbol{\beta}$, $\sigma = const.$
Weibull AFT Model: $F = \mathcal{M}$; $h_Y(y) = \nu \cdot \log(y)$, $\boldsymbol{\beta}^* = \nu \cdot \boldsymbol{\beta}$ = Weibull PH model
Log-Logistic AFT Model: $F = \mathcal{L}$; $h_Y(y) = p \cdot \log(y)$, $\boldsymbol{\beta}^* = p \cdot \boldsymbol{\beta}$; = Log-Logistic Model
AFT Models: $\log(y) = -\boldsymbol{x}^\top\boldsymbol{\beta} + \epsilon$, $\epsilon \sim F$
Log-Normal AFT Model: $F = \Phi$; $h_Y(y) = \frac{1}{\sigma}\log(y)$, $\boldsymbol{\beta}^* = \beta/\sigma$
Varying location and dispersion AFT model: $\mu = \boldsymbol{x}^\top\boldsymbol{\beta}$, $1/\sigma = \exp(\mathbf{z}^\top\boldsymbol{\gamma})$

Figure 2.1.: Overview of specific transformation models with continuous response (Part I).

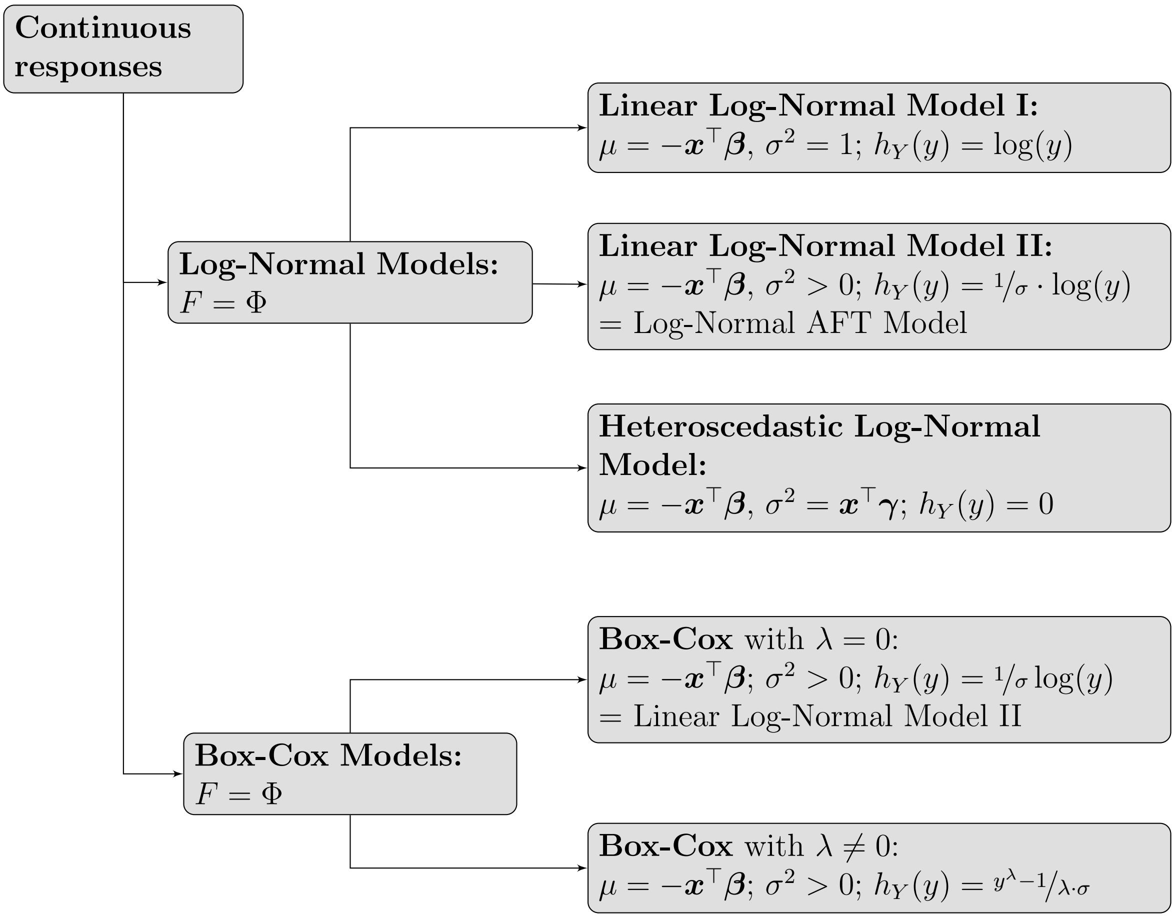

Continuous responses
Log-Normal Models: $F = \Phi$
Linear Log-Normal Model I: $\mu = -\boldsymbol{x}^\top\boldsymbol{\beta}$, $\sigma^2 = 1$; $h_Y(y) = \log(y)$
Linear Log-Normal Model II: $\mu = -\boldsymbol{x}^\top\boldsymbol{\beta}$, $\sigma^2 > 0$; $h_Y(y) = 1/\sigma \cdot \log(y)$ = Log-Normal AFT Model
Heteroscedastic Log-Normal Model: $\mu = -\boldsymbol{x}^\top\boldsymbol{\beta}$, $\sigma^2 = \boldsymbol{x}^\top\boldsymbol{\gamma}$; $h_Y(y) = 0$
Box-Cox Models: $F = \Phi$
Box-Cox with $\lambda = 0$: $\mu = -\boldsymbol{x}^\top\boldsymbol{\beta}$; $\sigma^2 > 0$; $h_Y(y) = 1/\sigma \log(y)$ = Linear Log-Normal Model II
Box-Cox with $\lambda \neq 0$: $\mu = -\boldsymbol{x}^\top\boldsymbol{\beta}$; $\sigma^2 > 0$; $h_Y(y) = y^\lambda - 1/\lambda \cdot \sigma$

Figure 2.2.: Overview of specific transformation models with continuous response (Part II).

The PH model includes some parametric regression models as special cases. The exponential PH model with constant baseline hazard function over time results from parametrising the response transformation via $h_Y(y) = \log(\lambda) + \log(y)$, where λ denotes the parameter of the exponential distribution. A monotone non-constant baseline hazard function is assumed in the Weibull model, where the response transformation is parametrised by $h_Y(y) = \log(\lambda) + \nu \cdot \log(y)$. The Weibull scale and shape parameter are denoted by λ and ν, respectively. Parametrising the response transformation by $h_Y(y) = \log(\alpha \cdot \log(y/\sigma))$ results in the PH model with responses from the Pareto (type I) distribution, where σ and α are the minimum and the index parameter of the Pareto distribution.

Proportional odds (PO) model. The PO model is commonly applied in survival analysis if the hazard ratio is not constant over time (as assumed in the PH model), but the hazards converge over time. This effect can be observed, *e.g.*, if there is an effective cure, or the treatment effect vanishes over time (Murphy et al., 1997). Bennett (1983a) introduces the PO model for a continuous survival time T, but the PO model is not restricted to survival times and might be a useful approach for a non-negative (censored) response variable Y in general. The PO model can be formulated in terms of a linear transformation model (Doksum and Gasko, 1990)

$$\mathbb{P}(Y \leq y|\mathbf{X} = \boldsymbol{x}) = \mathcal{L}(h_Y(y) + \boldsymbol{x}^\top \boldsymbol{\beta}),$$

where $\mathcal{L}$ denotes the standard logistic distribution function, and $h_Y(y)$ equals the monotone baseline log-odds. Similar to the PH model, the PO assumption results from ignoring all interactions between the response and the explanatory variables. Therefore, the explanatory variables influence only the conditional mean of the transformed response. The PO assumption can be relaxed by including interactions between $\boldsymbol{x}$ and y, what results in a non-proportional odds model.

Log-logistic regression model. The log-logistic regression model (Bennett, 1983b) is a special case of the PO model and is adequate for non-monotone hazard functions, where the hazard ratio converges to unity over time. The corresponding conditional distribution function of the response is defined via

$$\mathbb{P}(Y \leq y|\mathbf{X} = \boldsymbol{x}) = \mathcal{L}(\alpha \cdot \log(y) + \boldsymbol{x}^\top \boldsymbol{\beta}).$$

Hence, the unrestricted monotone response transformation function $h_Y(y)$ from the PO model is restricted to the parametric function $h_Y(y) = \alpha \cdot \log(y)$. In analogy to the PO model, the covariates influence the location of the distribution linearly. Because there are no interaction terms between the response and the explanatory variables, the cumulative odds ratios are constant over time. As the log-logistic model may be a poor choice for skewed or heavily tailed hazard functions, Singh et al. (1988) propose a generalisation of the log-logistic model by introducing scale parameters.

Additive hazards regression models. In the PH model it is assumed that the covariate effects act multiplicatively on some unknown baseline hazard function. An alternative to this semiparametric multiplicative hazards model is the consideration of additive hazards regression models (AHRMs) (an introduction is given, *e.g.*, by Klein and Moeschberger, 2003), where the conditional hazard function is modelled as a linear combination of the (time-dependent) covariates, *i.e.*

$$\lambda(t|\mathbf{X}(t) = \boldsymbol{x}(t)) = \beta_0(t) + \boldsymbol{x}(t)^\top \boldsymbol{\beta}(t), \tag{2.5}$$

where $\boldsymbol{\beta}(t)$ denotes varying coefficients that need to be estimated from the data. In contrast to PH models, no changes in the relative risk but changes in the additive risk over time are estimated. The best known AHRMs are Aalen's nonparametric additive hazards model and Lin and Ying's additive hazards model. We consider only Lin and Ying's model because necessary model simplifications of Aalen's model result in Lin and Ying's model anyway.

Lin and Ying (1994) propose an additive hazards regression model where the varying coefficients in Equation (2.5) are replaced by constants. Similar to the PH and the PO model, we define Lin and Ying's model for an arbitrary non-negative response variable Y, and consider only response-independent covariates, what implies the conditional hazard function

$$\lambda(y|\mathbf{X}=\boldsymbol{x}) = \beta_0(y) + \boldsymbol{x}^\top\boldsymbol{\beta}.$$

The corresponding conditional distribution function is

$$\mathbb{P}(Y \leq y|\mathbf{X}=\boldsymbol{x}) = 1 - \exp\left(-\int_0^y \lambda(u|\boldsymbol{x})\,du\right) = \mathcal{E}(B_0(y) + (\boldsymbol{x}\cdot y)^\top\boldsymbol{\beta}),$$

where $\mathcal{E}$ denotes the exponential distribution function, and $B_0(y) = \int_0^y \beta_0(u)\,du$. Because interactions between $\boldsymbol{x}$ and y are considered, additive hazards models do no longer belong to the class of linear transformation models but to the model class of CTMs instead. Due to the linear interaction terms, the explanatory variables are able to influence the conditional mean and the conditional variance of the response. Estimation of the coefficients $\boldsymbol{\beta}$ is easy if only response-independent covariates are considered, as explicit formulas exist for the estimators and their variances (*e.g.*, Klein and Moeschberger, 2003).

Accelerated failure time (AFT) models. AFT models are a class of parametric regression models that is most commonly used as an alternative to PH models. The most important difference in AFT and PH models is that the covariate effects act differently. The covariates affect the hazard function directly in PH models, whereas the covariates have a direct effect on the survivor function in AFT models by accelerating or decelerating the time scale. Nevertheless, we define AFT models for a non-negative (censored) response variable Y because AFT models are not restricted to survival times in general. The AFT model is equivalent to a location-shift model of the log-transformed response variable, *i.e.*

$$\log(y) = -\boldsymbol{x}^\top\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim F,$$

where F is the common distribution function of the i.i.d. error terms $\boldsymbol{\epsilon}$. Hence, the response transformation $h_Y(y) = \log(y)$ is completely specified in AFT models.

Two versions of the AFT model are usually distinguished:

1. **Standard Linear Location AFT Model**
   The location parameter is modelled in terms of the covariates, *i.e.* $\mu = \boldsymbol{x}^\top\boldsymbol{\beta}$, whereas the scale parameter is independent of the covariates, *i.e.* $\sigma =$ constant. For the corresponding conditional distribution function results:

   $$\mathbb{P}(Y \leq y|\mu,\sigma) = F({}^1\!/\!_\sigma \cdot (\log(y) + \boldsymbol{x}^\top\boldsymbol{\beta})) = F\left({}^1\!/\!_\sigma \cdot \log(y) + \boldsymbol{x}^\top\boldsymbol{\beta}^*\right),$$

   with scaled regression coefficients $\boldsymbol{\beta}^* = {}^{\boldsymbol{\beta}}\!/\!_\sigma$.

2. **Varying Location and Dispersion AFT Model**
   The location and the scale parameter depend linearly on the covariates (*e.g.*, Anderson, 1991): $\mu = \boldsymbol{x}^\top \boldsymbol{\beta}$; $1/\sigma = \exp(\boldsymbol{z}^\top \boldsymbol{\gamma})$. As σ has to be non-negative, the choice of a log-linear relationship for the dispersion parameter is one (common) possibility. Here, $\boldsymbol{x}$ and $\boldsymbol{z}$ may contain different sets of covariates. The conditional distribution function is
   $$\mathbb{P}(Y \leq y|\mu, \sigma) = F(\exp(\boldsymbol{z}^\top \boldsymbol{\gamma}) \cdot (\log(y) + \boldsymbol{x}^\top \boldsymbol{\beta})).$$

The linear location AFT model belongs to the model class of linear transformation models, whereas the varying location and dispersion AFT model can only be formulated in terms of CTMs due to the non-linear interaction between the explanatory variables $\boldsymbol{z}$ and the response y.
There is a variety of distribution functions F that can be used in AFT models. Nevertheless, the log-logistic model is the only parametric model with both an AFT and a PO representation. Moreover, the Weibull model is the only model that can be parametrised as a PH model and as an AFT model (Klein and Moeschberger, 2003), where the coefficients $\boldsymbol{\beta}$ are proportional with proportionality constant ν (the scale parameter of the Weibull distribution). The exponential model is a special case of the Weibull model ($\nu = 1$) and can also be parametrised in terms of a PH and an AFT model with equal parameters $\boldsymbol{\beta}$. In the following, the most important special cases of AFT models are parametrised in terms of linear transformation models.

1. **Exponential AFT model**
   The conditional distribution function of the exponential AFT model is
   $$\mathbb{P}(Y \leq y|\mathbf{X} = \boldsymbol{x}) = \mathcal{M}(\log(y) + \boldsymbol{x}^\top \boldsymbol{\beta}),$$
   where $\log(\lambda) = \boldsymbol{x}^\top \boldsymbol{\beta}$, and λ is the inverse mean of the exponential distribution.

2. **Weibull AFT model**
   The conditional distribution function of the Weibull AFT model is defined via
   $$\mathbb{P}(Y \leq y|\mathbf{X} = \boldsymbol{x}) = \mathcal{M}(1/\sigma \cdot \log(y) + \boldsymbol{x}^\top \boldsymbol{\beta}^*),$$
   where $\boldsymbol{\beta}^* = \beta/\sigma$, $\log(\lambda) = \boldsymbol{x}^\top\beta/\sigma$, and $\nu = 1/\sigma$. λ and ν denote the scale and the shape parameter of the Weibull distribution.

3. **Log-normal AFT model**
   The log-normal AFT model can be described by the conditional distribution function
   $$\mathbb{P}(Y \leq y|\mathbf{X} = \boldsymbol{x}) = \Phi\left(1/\sigma \cdot \log(y) + \boldsymbol{x}^\top \boldsymbol{\beta}^*\right),$$
   with scaled regression coefficients $\boldsymbol{\beta}^* = \beta/\sigma$. The mean μ of the log normal distribution depends linearly on the explanatory variables, $\mu = -\boldsymbol{x}^\top \boldsymbol{\beta}$, and σ denotes the standard deviation of the log-normal distribution.

4. **Log-logistic AFT model**
   The conditional distribution function of the log-logistic AFT-model is

   $$\mathbb{P}(Y \leq y | \mathbf{X} = \boldsymbol{x}) = \mathcal{L}({}^1\!/\!_\sigma \cdot \log(y) + \boldsymbol{x}^\top \boldsymbol{\beta}^*),$$

   with scaled regression coefficients $\boldsymbol{\beta}^* = {}^{\boldsymbol{\beta}}\!/\!_\sigma$, $\log(\lambda) = {}^{\boldsymbol{x}^\top \boldsymbol{\beta}}\!/\!_\sigma$, and $p = {}^1\!/\!_\sigma$. Thereby, λ denotes the location and p denotes the scale parameter of the log-logistic distribution.

Log-normal distributed responses. For example, Manning and Mullahy (2001) review log-models including the log-normal regression model as a special case. If the random variable Y is log-normal distributed with parameters μ and σ^2, $\log(Y)$ is normal distributed with parameters μ and σ^2, $\log(Y) \sim N(\mu, \sigma^2)$. The conditional distribution function of log-normal distributed responses can be formulated in terms of conditional transformation models:

1. **Linear log-normal model**
   The mean of the log-normal responses is influenced linearly by the explanatory variables, *i.e.* $\mu = -\boldsymbol{x}^\top \boldsymbol{\beta}$, and the variance is set equal to 1, *i.e.* $\sigma^2 = 1$. This implies the conditional distribution function

   $$\mathbb{P}(Y \leq y | \mathbf{X} = \boldsymbol{x}) = \Phi(\log(y) - \mu) = \Phi(\log(y) + \boldsymbol{x}^\top \boldsymbol{\beta}),$$

   where Φ is the standard normal distribution function, and $h_Y(y) = \log(y)$ is specified.

2. **Linear log-normal model with arbitrary but fixed variance**
   The mean of the log-normal responses is influenced linearly by the explanatory variables, *i.e.* $\mu = -\boldsymbol{x}^\top \boldsymbol{\beta}$, and an arbitrary non-negative but fixed variance $\sigma^2 > 0$ is assumed. The corresponding conditional distribution function is

   $$\mathbb{P}(Y \leq y | \mathbf{X} = \boldsymbol{x}) = \Phi({}^1\!/\!_\sigma \cdot \log(y) + \boldsymbol{x}^\top \boldsymbol{\beta}^*),$$

   with scaled regression coefficients $\boldsymbol{\beta}^* = {}^{\boldsymbol{\beta}}\!/\!_\sigma$. Hence, the response transformation function is $h_Y(y) = {}^1\!/\!_\sigma \cdot \log(y)$.

3. **Heteroscedastic linear log-normal model**
   Again, the mean of the log-normal responses is influenced linearly by the explanatory variables, *i.e.* $\mu = -\boldsymbol{x}^\top \boldsymbol{\beta}$. Additionally, the explanatory variables have a linear influence on the standard deviation σ, *i.e.* $\sigma = \boldsymbol{x}^\top \boldsymbol{\gamma}$. This implies a conditional distribution function of the form

   $$\mathbb{P}(Y \leq y | \mathbf{X} = \boldsymbol{x}) = \Phi(\log(y) \cdot (\boldsymbol{x}^\top \boldsymbol{\gamma})^{-1} + {}^{\boldsymbol{x}^\top \boldsymbol{\beta}}\!/\!_{\boldsymbol{x}^\top \boldsymbol{\gamma}}).$$

   To guarantee a positive standard deviation σ, $\boldsymbol{\gamma}$ is subject to constrained optimisation.

All log-normal transformation models can be extended to flexible covariate effects. Models (1.) and (2.) obviously belong to the class of linear transformation models, whereas model (3.) belongs to the model class of CTMs, due to the non-linear interaction between y and $\boldsymbol{x}$. The proposed models for log-normal responses are easily transferable to normal distributed responses. The only difference is that the response transformation is superfluous, and $\log(y)$ is replaced by y. In this case, model (3.) becomes the heteroscedastic linear regression model.

Box-Cox transformation models. The parametric Box-Cox response transformations presented in Equation (2.1) result in normal distributed transformed responses, *i.e.* $h_Y(Y|\lambda) \sim N(\mu, \sigma^2)$. Accordingly, Box-Cox transformation models can be formulated in terms of linear transformation models, where the mean μ depends linearly on the explanatory variables, *i.e.* $\mu = -\boldsymbol{x}^\top\boldsymbol{\beta}$, and the variance $\sigma^2 > 0$ is independent of the explanatory variables:

$$\mathbb{P}(Y \leq y|\mathbf{X} = \boldsymbol{x}) = \begin{cases} \Phi(\frac{y^\lambda - 1}{\lambda \cdot \sigma} + \boldsymbol{x}^\top\boldsymbol{\beta}^*), & \lambda \neq 0, \\ \\ \Phi(\frac{1}{\sigma} \cdot \log(y) + \boldsymbol{x}^\top\boldsymbol{\beta}^*), & \lambda = 0, \end{cases}$$

with scaled regression coefficients $\boldsymbol{\beta}^* = {}^{\boldsymbol{\beta}}\!/\!_{\sigma}$. Hence, in Box-Cox models the response transformation is either $h_Y(y) = {}^{(y^\lambda - 1)}\!/\!_{\lambda \cdot \sigma}$ if $\lambda \neq 0$, or $h_Y(y) = {}^{1}\!/\!_{\sigma} \cdot \log(y)$ if $\lambda = 0$.

Ordinal responses

Count data with unbounded support. Let Y be a count variable with unbounded support, $Y \in \{0, 1, 2, \ldots\}$, following an arbitrary count data distribution (*e.g.*, Poisson distribution, negative binomial distribution, or geometric distribution). The conditional distribution function can be formulated in terms of linear transformation models

$$\mathbb{P}(Y \leq y|\mathbf{X} = \boldsymbol{x}) = F(h_Y(y) + \boldsymbol{x}^\top\boldsymbol{\beta}),$$

where a non-decreasing step function as well as a non-decreasing continuous function can be chosen for $h_Y(y)$, which is only evaluated at possible response values $y \in \{0, 1, 2, \ldots\}$. The link function F is not further specified, but it is advisable to choose a link function that supports certain characteristics of the count data distribution such as a positive support or possible right-skewness. The corresponding probability density function is

$$\begin{aligned} \mathbb{P}(Y = y|\mathbf{X} = \boldsymbol{x}) &= F(h_Y(y) + \boldsymbol{x}^\top\boldsymbol{\beta}) - F(h_Y(y-1) + \boldsymbol{x}^\top\boldsymbol{\beta}), \quad y > 0, \\ \mathbb{P}(Y = 0|\mathbf{X} = \boldsymbol{x}) &= F(h_Y(0) + \boldsymbol{x}^\top\boldsymbol{\beta}), \quad y = 0. \end{aligned}$$

The unbounded count data regression models do no longer belong to the model class of linear transformation models but to CTMs instead if interactions between y and $\boldsymbol{x}$ are considered. Furthermore, the model equation can be extended easily to non-linear covariate effects.

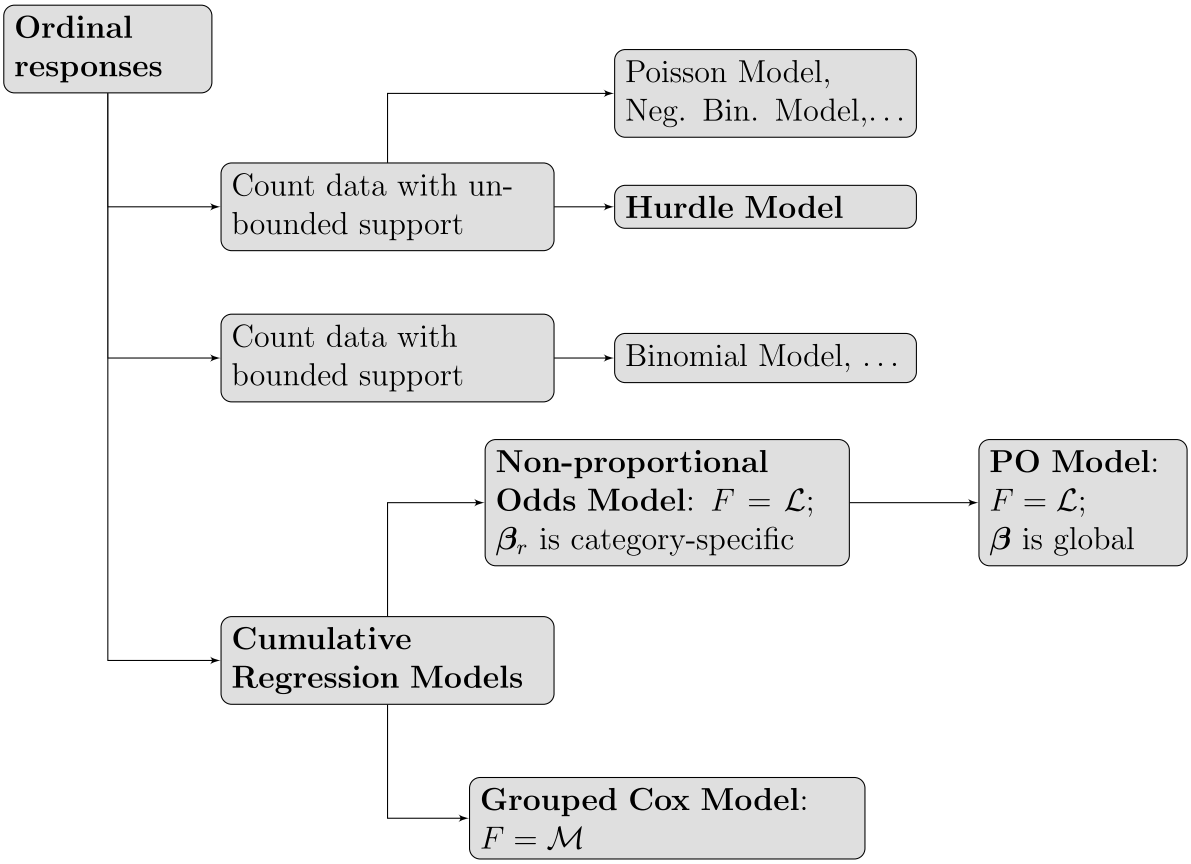

Ordinal responses
Poisson Model, Neg. Bin. Model,…
Count data with unbounded support
Hurdle Model
Count data with bounded support
Binomial Model, …
Non-proportional Odds Model: $F = \mathcal{L}$; $\boldsymbol{\beta}_r$ is category-specific
PO Model: $F = \mathcal{L}$; $\boldsymbol{\beta}$ is global
Cumulative Regression Models
Grouped Cox Model: $F = \mathcal{M}$

Figure 2.3.: Overview of specific transformation models with ordinal response.

Count data with bounded support. A typical example for a count variable with bounded support $Y \in \{0, 1, \ldots, n\}$ is a binomial distributed random variable. In terms of linear transformation models, the corresponding conditional distribution function can be formulated as for count data with unbounded support. Nevertheless, the additional constraint

$$\mathbb{P}(Y \leq n | \mathbf{X} = \boldsymbol{x}) = F(h_Y(n) + \boldsymbol{x}^\top \boldsymbol{\beta}) \overset{!}{=} 1,$$

has to be taken into account.

Hurdle models. In the social sciences or in econometrics, count data often suffer from overdispersion or zero inflation (Zeileis et al., 2008). Therefore, count data extensions like the *hurdle model* were introduced (Mullahy, 1986) that are able to deal with overdispersion and an excessive number of zeros. Even though hurdle models with hurdle at zero are most relevant in practice, the model class can be generalised to arbitrary hurdles k. A possible parametrisation of the conditional distribution function of an overdispersed count variable $Y \in \{0, 1, 2, \ldots\}$ with an excessive number of zeros ($k = 0$) in terms of conditional transformation models is

$$\begin{aligned}
&\mathbb{P}(Y \leq y|\mathbf{X} = \boldsymbol{x}) = F(h(y|\boldsymbol{x})), \text{ with} \\
&h(y|\boldsymbol{x}) = h_0(I(y \leq k)|\boldsymbol{x}) + h_1(y|\boldsymbol{x}) \cdot I(y \leq k) + h_2(y|\boldsymbol{x}) \cdot I(y > k).
\end{aligned}$$

To guarantee monotonicity of the distribution function, the constraints

$$h_0(0|\boldsymbol{x}) + h_2(y|\boldsymbol{x}) - (h_0(1|\boldsymbol{x}) + h_1(y|\boldsymbol{x})) \geq 0, \ \forall y > k,$$

have to be imposed. Additionally, both transformation functions h_1 and h_2 have to be monotonically increasing in y, *i.e.* $h_1^{\prime}(y|\boldsymbol{x}) \geq 0$ and $h_2^{\prime}(y|\boldsymbol{x}) \geq 0$. The transformation function h_0 represents the 'baseline' covariate effects below and above the hurdle k. As there are interactions between y and $\boldsymbol{x}$, the hurdle model belongs to the model class of CTMs.

Cumulative regression models for ordinal responses. McCullagh (1980) introduces cumulative regression models as a general class of regression models for ordinal data. In terms of linear transformation models, the conditional distribution function of an ordinal response with q ordered categories $Y \in \{1, \ldots, q\}$ is defined via

$$\mathbb{P}(Y \leq r|\mathbf{X} = \boldsymbol{x}) = F(\alpha_r + \boldsymbol{x}^\top \boldsymbol{\beta}), \tag{2.6}$$

where α_r denotes the category-specific intercept for category $r \in \{1, \ldots, q\}$. As the response y is ordinal, the category-specific intercepts follow the constraints $\alpha_2 - \alpha_1 \geq 0, \ldots, \alpha_q - \alpha_{q-1} \geq 0$, *i.e.* the thresholds are monotonically increasing. The corresponding probability density function is

$$\begin{aligned}
\mathbb{P}(Y = 1|\mathbf{X} = \boldsymbol{x}) &= F(\alpha_1 + \boldsymbol{x}^\top \boldsymbol{\beta}), \\
\mathbb{P}(Y = r|\mathbf{X} = \boldsymbol{x}) &= F(\alpha_r + \boldsymbol{x}^\top \boldsymbol{\beta}) - F(\alpha_{r-1} + \boldsymbol{x}^\top \boldsymbol{\beta}), \ r = 2, \ldots, q-1, \\
\mathbb{P}(Y = q|\mathbf{X} = \boldsymbol{x}) &= 1 - F(\alpha_{q-1} + \boldsymbol{x}^\top \boldsymbol{\beta}).
\end{aligned}$$

A typical choice for the link function F is the distribution function of the logistic distribution, $F = \mathcal{L}$, what results in the well-known PO model. The PO model has the property that the cumulative odds ratio for observations with explanatory variables $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$

$$\frac{\mathbb{P}(Y \leq r|\mathbf{X} = \boldsymbol{x})/\mathbb{P}(Y > r|\mathbf{X} = \boldsymbol{x})}{\mathbb{P}(Y \leq r|\mathbf{X} = \tilde{\boldsymbol{x}})/\mathbb{P}(Y > r|\mathbf{X} = \tilde{\boldsymbol{x}})} = \exp((\boldsymbol{x} - \tilde{\boldsymbol{x}})^\top \boldsymbol{\beta})$$

is independent of the category r and thus, cumulative odds ratios are proportional across all categories. Of course, the model equation in (2.6) can be further extended, *e.g.*, to include non-linear covariate effects. One important generalisation is the introduction of category-specific regression coefficients $\boldsymbol{\beta}_r$, what results in a non-proportional odds model

(Peterson and Harrell, 1990).
Another important cumulative regression model results from choosing the minimum-extreme value distribution, *i.e.* $F = \mathcal{M}$. The respective regression model is equivalent to the *grouped Cox model*, which is the discrete variant of the continuous Cox model known from survival analysis. Similar to the above extensions of the PO model, the grouped Cox model can be extended to non-proportional hazards and to more complex covariate effects as well.

2.2. Conditional transformation models

Previously, we reviewed a broad range of frequently used regression models for continuous and ordinal responses from the perspective of conditional transformation models. Our aim was to show that all considered regression models share a common model basis, the model basis of CTMs. In the following, we define the model class of CTMs for continuous and ordinal responses. Afterwards, we present the very useful simplification to the model class of conditionally linear transformation models (CLTMs), which is a special case of CTMs. Several low-parametrised and interpretable CLTMs are presented in more detail, and the associated model characteristics are described.

2.2.1. CTMs for continuous and ordinal responses

Most common regression models model only the conditional mean $\mathbb{E}(Y|\mathbf{X} = \boldsymbol{x})$ of the response Y as a function of the explanatory variables $\mathbf{X} = \boldsymbol{x}$. This assumption can be relaxed by considering CTMs (Hothorn et al., 2014). The whole conditional distribution function of Y is modelled in terms of the explanatory variables, and hence not only the conditional mean but also higher moments of the distribution function may depend on $\boldsymbol{x}$. In this chapter, we give only a brief definition of CTMs for continuous and discrete responses. For a more detailed introduction to CTMs for continuous responses including model characteristics and a precise definition of the conditional transformation function h and the distribution function F, we refer to Section 1.1 and Hothorn et al. (2014).
In CTMs, the conditional distribution function for a continuous response $Y \in \mathbb{R}$ (Equation (1.1)) and the corresponding conditional density are defined via

$$\begin{aligned} \mathbb{P}(Y \leq y|\mathbf{X} = \boldsymbol{x}) &= F_{Y|\mathbf{X}=\boldsymbol{x}}(y|\boldsymbol{x}) = F(h(y|\boldsymbol{x})), \\ f_{Y|\mathbf{X}=\boldsymbol{x}}(y|\boldsymbol{x}) &= f(h(y|\boldsymbol{x})) \cdot h^{'}(y|\boldsymbol{x}), \end{aligned} \tag{2.7}$$

where f denotes the density corresponding to the distribution function F, and $h^{'}(y|\boldsymbol{x})$ denotes the first derivative of the conditional transformation function $h(y|\boldsymbol{x})$ with respect to y. To consider important characteristics of a conditional distribution function, the conditional transformation function $h(y|\boldsymbol{x})$ is assumed to be monotonically increasing in y and smooth. The monotonicity of h transfers directly to the conditional distribution function $\mathbb{P}(Y \leq y|\mathbf{X} = \boldsymbol{x})$, which is accordingly monotonically increasing itself. Smoothness is typically expected in the direction of the response y as well as in direction of the explanatory variables $\boldsymbol{x}$. Continuous distribution functions are smooth in the response variable, what implies smoothness in the direction of y. Furthermore, we expect similar conditional distribution functions for similar values of the vector of explanatory variables $\boldsymbol{x}$.

If the discrete response variable $Y \in \{c_1, \dots, c_q\}$ is ordinal, *i.e.* the categories $c_1, \dots, c_q$ are ordered, the corresponding conditional distribution function is

$$\mathbb{P}(Y \leq c_k|\mathbf{X} = \boldsymbol{x}) = F(h(c_k|\boldsymbol{x})) = F(h_k(\boldsymbol{x})),\ k \in \{1, \dots, q\}\,, \tag{2.8}$$

where $h_1(\cdot), \dots, h_q(\cdot)$ denote separate transformation functions for each category. The category-specific transformation functions have to be monotonically increasing, *i.e.* $h_1(\cdot) \leq h_2(\cdot) \leq \ldots \leq h_q(\cdot)$, to guarantee a monotonically increasing conditional distribution function. Assumptions of smoothness may be defined problem-specific, *e.g.*, smoothing over the response categories might be meaningful if neighbouring categories are not supposed to differ considerably. The corresponding conditional probability density function is

$$\begin{aligned} \mathbb{P}(Y = c_1|\mathbf{X} = \boldsymbol{x}) &= F(h_1(\boldsymbol{x})), \\ \mathbb{P}(Y = c_k|\mathbf{X} = \boldsymbol{x}) &= F(h_k(\boldsymbol{x})) - F(h_{k-1}(\boldsymbol{x})),\ \text{for } k = 2, \dots, q. \end{aligned} \tag{2.9}$$

Of course, the conditional probabilities for the categories $c_1, \dots, c_q$ have to sum up to one. The definition of the conditional probability for the highest category c_q by $\mathbb{P}(Y = c_q|\mathbf{X} = \boldsymbol{x}) = 1 - F(h_{q-1}(\boldsymbol{x}))$ is one possibility to take this constraint into account.

In CTMs, the distribution function F plays the role of a link function that maps the values of the conditional transformation function on the interval $[0, 1]$. Hence, F is fixed and chosen a priori, and we concentrate only on characteristics of the conditional transformation function $h(y|\boldsymbol{x})$ in the following.
As we are interested in a better interpretable model class, we introduce conditionally linear transformation models (CLTMs), which are a special case of CTMs. Therefore, the conditional transformation function in CTMs needs to be restricted. The necessary restrictions of the conditional transformation function $h(y|\boldsymbol{x})$ that result in the model class of CLTMs are presented below. Additionally, several parsimonious parametrisations of the conditional transformation function in CLTMs are discussed.

2.2.2. Conditionally linear transformation models

Model class. In its most general form, the conditional transformation function $h(y|\boldsymbol{x})$ of a conditional transformation model can be displayed via

$$h(y|\boldsymbol{x}) = K_Y(y) \otimes K_{\mathbf{X}}(\boldsymbol{x}),$$

where K_Y and $K_{\mathbf{X}}$ denote some kernel functions for the response and the explanatory variables, and $\otimes$ denotes the Kronecker product. Thereby, no further assumptions on the kind of the relationship between the response variable and the explanatory variables are imposed.

In CTMs, the conditional transformation function is decomposed additively into J partial transformation functions (see Section 1.1, Hothorn et al., 2014)

$$h(y|\boldsymbol{x}) = \sum_{j=1}^{J} h_j(y|\boldsymbol{x}), \tag{2.10}$$

whereby additivity on the scale of the transformation function is assumed. Each partial transformation function $h_j(y|\boldsymbol{x})$, $j = 1, \ldots, J$, is conditional on the explanatory variables, and may have an arbitrarily complex form in direction of the explanatory variables as well as in direction of the response. The only restriction is that the conditional transformation function $h(y|\boldsymbol{x})$ has to be monotonically increasing in y. Thereby, complex relationships between the explanatory variables and the response can be displayed, and all parameters of the distribution function (*i.e.* mean, variance, skewness, and kurtosis) may depend on the explanatory variables in CTMs.

A lack of orthogonality of the model components in CTMs constricts insights into model structure because the model components are not separable. Hence, the high flexibility of CTMs often ends up in models with challenging model interpretation. For example, the effects of the explanatory variables on certain moments of the conditional distribution function are usually not interpretable. Therefore, we define less complex, low-parametrised, and interpretable transformation models belonging to the class of conditionally linear transformation models (CLTMs), which are a special case of CTMs. In CLTMs, the conditional part of the transformation function is restricted to linear functions of the response y:

$$h(y|\boldsymbol{x}) = \underbrace{h_Y(y)}_{\text{uncond. part}} + \underbrace{\beta_0(\boldsymbol{x}) + y \cdot \beta_1(\boldsymbol{x})}_{\text{conditional part}}. \tag{2.11}$$

Hence, starting from the conditional transformation function for CTMs (Equation (2.10)), we define three partial transformation functions: $h_1(y|\boldsymbol{x}) = h_Y(y)$ describes the marginal effects of the response, $h_2(y|\boldsymbol{x}) = \beta_0(\boldsymbol{x})$ describes the marginal effects of the explanatory variables, and $h_3(y|\boldsymbol{x}) = y \cdot \beta_1(\boldsymbol{x})$ describes interactions between y and $\boldsymbol{x}$, which have to be linear in y. Due to these restrictions, the explanatory variables may only influence the conditional mean and the conditional variance of the transformed response (what can be clarified by calculating means and variances in Equation (2.11)). This influence of the explanatory variables is modelled in terms of the coefficient functions $\beta_0(\boldsymbol{x})$ causing shifts of the distribution function, and $\beta_1(\boldsymbol{x})$ causing shifts and scalings of the distribution function. Higher moments of the response distribution can be modelled in terms of the response transformation $h_Y(y)$. But in contrast to CTMs, kurtosis and skewness are not affected by the explanatory variables because $h_Y(y)$ is independent of $\boldsymbol{x}$.

Similar to linear transformation models (Equation (2.2)), we assume that the unconditional part of the transformation function $h_Y(y)$ has to be monotonically increasing in y. Additionally, in accordance with CTMs, the whole conditional transformation function $h(y|\boldsymbol{x})$ has to be monotonically increasing in y.

These restrictions lead to interpretable transformation models with a clearly specified model structure, where the effects of the explanatory variables are separable. Hence, the effects of the explanatory variables on the first two moments of the distribution function are interpretable in CLTMs, whereas such interpretations are not possible in CTMs. Nevertheless, interpretability comes at the price of reduced flexibility and the assumptions imposed in CLTMs may be inadequate. Concerning their flexibility, CLTMs can be placed in between linear transformation models (Equation (2.2)), which are less flexible because no interaction terms between $\boldsymbol{x}$ and y are intended, and CTMs (Equation (2.10)), which are more flexible because all moments of the distribution function may be influenced by $\boldsymbol{x}$.

The conditional distribution function of birth weight depending on a set of ultrasound measurements for newborns in the Perinatal Database Erlangen is analysed using CLTMs. Therefore, the model class of CLTMs is also introduced in Chapter 4, and model characteristics are discussed in detail in terms of an application.

Interpretable parametrisations of CLTMs. As stated before, we are only interested in interpretable, parsimonious parametrisations of CLTMs. Therefore, we present a cascade of low-parametrised CLTMs that are highly relevant in many applications. Of course, all presented CLTMs are special cases of Equation (2.11). Depending on their complexity, the selected CLTMs are parametrised in terms of regression coefficients or in terms of smooth functions, and the corresponding model characteristics are summarised. For estimation purposes, we additionally formulate the proposed CLTMs in terms of basis functions. As the transformation function $h(y|\boldsymbol{x})$ has to be monotonically increasing in y, the corresponding linear constraints that have to be considered during estimation are presented. All CLTMs are applicable to continuous and ordinal responses, and the number of covariates is set to P, *i.e.* $\boldsymbol{x} = (x_1, \ldots, x_P)^\top$. The selected CLTMs are ordered with increasing model complexity.

- **CLTM A**:

$$\begin{aligned} h(y|\boldsymbol{x}) &= h_Y(y) + \beta_0(\boldsymbol{x}) = \\ &= h_Y(y) + \boldsymbol{x}^\top \boldsymbol{\beta}_0 = \alpha_0 + \alpha_1 \cdot y + \boldsymbol{x}^\top \boldsymbol{\beta}_0 \\ &= \mathbf{b}_{\text{lin}}(y)^\top \cdot \boldsymbol{\alpha} + \mathbf{b}_{\text{lin}}(\boldsymbol{x})^\top \cdot \boldsymbol{\beta}_0 \\ &= \begin{pmatrix} 1 & y \end{pmatrix} \cdot \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} + \begin{pmatrix} x_1 & \ldots & x_P \end{pmatrix} \cdot \begin{pmatrix} \beta_{01} \\ \vdots \\ \beta_{0P} \end{pmatrix}, \end{aligned} \tag{2.12}$$

where $\mathbf{b}_{\text{lin}}$ denotes linear basis functions, and the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_0$ contain the corresponding basis coefficients. In this least flexible CLTM, the explanatory variables induce only linear shifts of the response distribution by $\beta_0(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}_0$. Thereby, the influence of the explanatory variables is restricted to the conditional mean of the response y. As the linear interaction term $y \cdot \beta_1(\boldsymbol{x})$ was cancelled, the conditional variance of the response y is not affected by the explanatory variables, and the constant variance is only influenced by the parameter α_1. Moreover, the response transformation $h_Y(y) = \alpha_0 + \alpha_1 \cdot y$ is linear in y. Therefore, higher moments of the response distribution remain unaffected, *i.e.* skewness and kurtosis are not affected by the response transformation.
The conditional transformation function has to be monotonically increasing in y, which implies the linear constraint $h'(y|\boldsymbol{x}) = \alpha_1 > 0$.

- **CLTM B:**

$$\begin{aligned} h(y|\boldsymbol{x}) &= h_Y(y) + \beta_0(\boldsymbol{x}) = \alpha_0 + \alpha_1 \cdot y + \sum_{p=1}^{P} \beta_{0p}(x_p) \\ &= \mathbf{b}_{\text{lin}}(y)^\top \cdot \boldsymbol{\alpha} + \mathbf{b}(x_1)^\top \cdot \boldsymbol{\beta}_{01} + \ldots + \mathbf{b}(x_P)^\top \cdot \boldsymbol{\beta}_{0P} \\ &= \begin{pmatrix} 1 & y \end{pmatrix} \cdot \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} + \begin{pmatrix} b_1(x_1) & \ldots & b_m(x_1) \end{pmatrix} \cdot \begin{pmatrix} \beta_{011} \\ \vdots \\ \beta_{01m} \end{pmatrix} + \ldots + \\ &\quad \begin{pmatrix} b_1(x_P) & \ldots & b_m(x_P) \end{pmatrix} \cdot \begin{pmatrix} \beta_{0P1} \\ \vdots \\ \beta_{0Pm} \end{pmatrix}, \end{aligned} \tag{2.13}$$

where α_0 and α_1 denote regression coefficients, and $\beta_{01}(\cdot), \ldots, \beta_{0P}(\cdot)$ denote smooth functions. In the parametrisation using basis functions, $\mathbf{b}(\cdot)$ denotes a set of m smooth basis functions, and $\boldsymbol{\alpha}$, $\boldsymbol{\beta}_{01}, \ldots, \boldsymbol{\beta}_{0P}$ denote the corresponding vectors of basis coefficients. In contrast to model CLTM A, the covariates have a flexible and smooth effect on the conditional mean of the response by $\beta_0(\boldsymbol{x}) = \sum_p \beta_{0p}(x_p)$ in CLTM B. Similar to CLTM A, the conditional variance of the response is not affected by the explanatory variables because the linear interaction $y \cdot \beta_1(\boldsymbol{x})$ was cancelled. Hence, the fixed variance is only influenced by the regression coefficient α_1. As the response transformation $h_Y(y) = \alpha_0 + \alpha_1 \cdot y$ is linear in y, higher moments of the response distribution remain unaffected.
Moreover, the necessary monotonicity constraint remains $h'(y|\boldsymbol{x}) = \alpha_1 > 0$.

- **CLTM C**:

$$\begin{aligned}
h(y|\boldsymbol{x}) &= h_Y(y) + \beta_0(\boldsymbol{x}) = h_Y(y) + \boldsymbol{x}^\top \boldsymbol{\beta}_0 \\
&= \mathbf{b}(y)^\top \cdot \boldsymbol{\alpha} + \mathbf{b}_{\mathrm{lin}}(\boldsymbol{x})^\top \cdot \boldsymbol{\beta}_0 \\
&= \begin{pmatrix} b_1(y) & \dots & b_m(y) \end{pmatrix} \cdot \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix} + \begin{pmatrix} x_1 & \dots & x_P \end{pmatrix} \cdot \begin{pmatrix} \beta_{01} \\ \vdots \\ \beta_{0P} \end{pmatrix}, \qquad (2.14)
\end{aligned}$$

where $\beta_{01}, \dots, \beta_{0P}$ denote regression coefficients and $h_Y(\cdot)$ denotes a monotonically increasing, smooth function. Model CLTM C belongs to the model class of linear transformation models (Equation (2.2)). The explanatory variables induce linear shifts of the response distribution by $\beta_0(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}_0$. This results in linear influences of the covariates on the conditional mean of the transformed response $h_Y(y)$. The conditional variance of the transformed response is not influenced by the explanatory variables because the linear interaction $y \cdot \beta_1(\boldsymbol{x})$ was cancelled. Variance, skewness and kurtosis of the response distribution are modelled independently of the explanatory variables in terms of the response transformation $h_Y(y)$.
The corresponding necessary monotonicity constraints are $h^{'}(y|\boldsymbol{x}) = h_Y^{'}(y) > 0$.

- **CLTM D**:

$$\begin{aligned}
h(y|\boldsymbol{x}) &= h_Y(y) + \beta_0(\boldsymbol{x}) + y \cdot \beta_1(\boldsymbol{x}) \\
&= \alpha_0 + \alpha_1 \cdot y + \boldsymbol{x}^\top \boldsymbol{\beta}_0 + (\boldsymbol{x} \cdot y)^\top \boldsymbol{\beta}_1 \\
&= \mathbf{b}_{\mathrm{lin}}(y)^\top \cdot \boldsymbol{\alpha} + \mathbf{b}_{\mathrm{lin}}(\boldsymbol{x})^\top \cdot \boldsymbol{\beta}_0 + (\mathbf{b}_{\mathrm{lin}}(y)^\top \otimes \mathbf{b}_{\mathrm{lin}}(\boldsymbol{x})^\top) \cdot \boldsymbol{\beta}_1 \\
&= \begin{pmatrix} 1 & y \end{pmatrix} \cdot \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} + \begin{pmatrix} x_1 & \dots & x_P \end{pmatrix} \cdot \begin{pmatrix} \beta_{01} \\ \vdots \\ \beta_{0P} \end{pmatrix} + \\
&\quad \begin{pmatrix} y \cdot x_1 & \dots & y \cdot x_P \end{pmatrix} \cdot \begin{pmatrix} \beta_{11} \\ \vdots \\ \beta_{1P} \end{pmatrix}, \qquad (2.15)
\end{aligned}$$

where $\alpha_0, \alpha_1, \beta_{01}, \dots, \beta_{0P}$ and $\beta_{11}, \dots, \beta_{1P}$ denote regression coefficients, and $\otimes$ denotes the Kronecker product. The explanatory variables influence the conditional mean and the conditional variance of the response y because linear interactions between $\boldsymbol{x}$ and y are considered in addition. Thereby, $\beta_0(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}_0$ induces linear shifts of the conditional distribution function, and hence, influences the conditional mean.

Linear shifts and scalings of the distribution function are induced by $\beta_1(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}_1$, and thus, it influences the conditional mean and the conditional variance of the response. The response transformation $h_Y(y) = \alpha_0 + \alpha_1 \cdot y$ is linear in y, whereby kurtosis and skewness remain unaffected.
Due to the monotonicity of $h(y|\boldsymbol{x})$ and $h_Y(y)$, the linear constraints $h^{\prime}(y|\boldsymbol{x}) = \alpha_1 + \boldsymbol{x}^\top \boldsymbol{\beta}_1 > 0$, and $\alpha_1 > 0$ have to be considered.

- **CLTM E**:

$$\begin{aligned} h(y|\boldsymbol{x}) &= h_Y(y) + \beta_0(\boldsymbol{x}) + y \cdot \beta_1(\boldsymbol{x}) \\ &= h_Y(y) + \boldsymbol{x}^\top \boldsymbol{\beta}_0 + (\boldsymbol{x} \cdot y)^\top \boldsymbol{\beta}_1 \\ &= \mathbf{b}(y)^\top \cdot \boldsymbol{\alpha} + \mathbf{b}_{\text{lin}}(\boldsymbol{x})^\top \cdot \boldsymbol{\beta}_0 + (\mathbf{b}_{\text{lin}}(y)^\top \otimes \mathbf{b}_{\text{lin}}(\boldsymbol{x})^\top) \cdot \boldsymbol{\beta}_1 \\ &= \begin{pmatrix} b_1(y) & \dots & b_m(y) \end{pmatrix} \cdot \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix} + \begin{pmatrix} x_1 & \dots & x_P \end{pmatrix} \cdot \begin{pmatrix} \beta_{01} \\ \vdots \\ \beta_{0P} \end{pmatrix} + \\ &\quad \begin{pmatrix} y \cdot x_1 & \dots & y \cdot x_P \end{pmatrix} \cdot \begin{pmatrix} \beta_{11} \\ \vdots \\ \beta_{1P} \end{pmatrix}, \end{aligned} \tag{2.16}$$

where $\beta_{01}, \dots, \beta_{0P}$ and $\beta_{11}, \dots, \beta_{1P}$ denote regression coefficients. The flexibility of CLTM D is further increased by allowing for a monotonically increasing and smooth response transformation function $h_Y(y)$ in CLTM E. Besides their influence on the conditional mean, the explanatory variables influence the conditional variance due to the linear interaction term between $\boldsymbol{x}$ and y. Similar to CLTM D, $\beta_0(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}_0$ induces linear shifts of the distribution function, and $\beta_1(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}_1$ induces linear shifts and scalings of the distribution function. Skewness and kurtosis of the response distribution are modelled in terms of the response transformation $h_Y(y)$, but these higher moments are not affected by the explanatory variables.
The linear constraints $h^{\prime}(y|\boldsymbol{x}) = h^{\prime}_Y(y) + \boldsymbol{x}^\top \boldsymbol{\beta}_1 > 0$ have to be considered to guarantee a monotonically increasing conditional transformation function, and the linear constraints $h^{\prime}_Y(y) > 0$ have to be considered to guarantee a monotonically increasing response transformation function.

The presented cascade of low-parametrised CLTMs CLTM A – CLTM E provides a useful tool for data analysis. The considered CLTMs differ concerning their model complexity, and hence, comparing the corresponding model performances can reveal important characteristics of the conditional response distribution. In CLTM A – CLTM C, the explanatory variables influence only the conditional mean, whereas the explanatory variables are additionally able to influence the conditional variance in CLTM D and CLTM E. Skewness and kurtosis are not influenced by the explanatory variables in all CLTMs. Nevertheless, the monotonically increasing response transformation $h_Y(y)$ is able to affect the mean, variance, kurtosis, and skewness of the response distribution in CLTM C and CLTM E. In contrast, $h_Y(y)$ is only able to affect the mean and the variance of the response distribution in CLTM A, CLTM B, and CLTM D because $h_Y(y) = \alpha_0 + \alpha_1 \cdot y$ is restricted to a linear function. In general, the consideration of low-parametrised CLTMs is advantageous in terms of model estimation due to the limited number of regression and basis coefficients. Likelihood-based estimation strategies for low-parametrised CLTMs are presented in Chapter 3.

2.3. Summary

In order to give a general overview, we reviewed the development of linear transformation models (Section 2.1), which are a powerful tool for data analysis. This overview clarified that the model class of linear transformation models is often too restrictive, and hence, it often needs to be adapted, *e.g.*, by including non-linear, response-varying, or random effects. To show the huge variety of CTMs (which are a generalisation of linear transformation models), and to clarify their applicability to continuous as well as ordinal responses, we reviewed various commonly used regression models from the perspective of CTMs (Hothorn et al., 2014) in Section 2.1.2. Thereby, a further aim was to clarify the common model basis, *i.e.* the model basis of CTMs, that the considered regression models share, even though they might seem rather different.

The model class of CTMs for continuous and ordinal responses was presented in Section 2.2.1. For reasons of interpretability, we introduced the model class of CLTMs, which constitutes an important special case of CTMs. In contrast to CTMs, the effect of the explanatory variables on the conditional mean and the conditional variance of the transformed response is interpretable, and model components are separable in CLTMs. Thereby, a closer insight into the model structure can be gained. Although model complexity is reduced in CLTMs, the model class still comprises a wide range of relevant transformation and regression models (Section 2.1.2). This statement was further underlined by the introduction of a cascade of low-parametrised CLTMs with differing model complexity.

To sum up, we expect promising performances of CLTMs in various applications. The usage of more complex CLTMs enables the practitioner to check important model assumptions made in less flexible standard regression models. For example, the proportional hazards assumption in the Cox model, or the homoscedasticity assumption in linear regression models can be relaxed easily by considering CLTMs (see Section 2.1.2).

3. Estimation of conditional transformation models

The content of this chapter is based on Möst and Hothorn (2014).

Conditional transformation models can be estimated based on two fundamentally different approaches. Hothorn et al. (2014) presented a component-wise boosting algorithm for the estimation of CTMs. By considering the respective model restrictions, this estimation approach can also be applied to CLTMs. Alternatively, we suggest likelihood-based estimation for C(L)TMs. Both approaches are introduced in this chapter.

3.1. Estimation based on component-wise boosting

Using the component-wise boosting algorithm presented in Hothorn et al. (2014), CTMs are estimated by regularised optimisation of a proper scoring rule for distributional and probabilistic forecasts. This results in consistent estimated conditional distribution functions. Such proper scoring rules are, *e.g.*, the continuous ranked probability score (CRPS), which is the integrated version of the famous Brier score, or the integrated absolute loss (Gneiting and Raftery, 2007; Hothorn et al., 2014). A third scoring rule for distributional forecasts is the mean integrated logarithmic score (log score), which is explained more thoroughly because its minimisation can be performed using standard software for additive binomial regression models.

For each observation i, we observe a response value Y_i and a corresponding vector of explanatory variables $\boldsymbol{x}_i$, $i = 1, \ldots, N$. Furthermore, we define a (*e.g.*, equidistant) grid of response values $\{y_\iota | \iota = 1, \ldots, n\}$ covering their range. Our aim is to estimate the conditional distribution function $\mathbb{P}(Y \leq y_\iota | \mathbf{X} = \boldsymbol{x}_i) = F(h(y_\iota | \boldsymbol{x}_i))$ by estimating the conditional transformation function h. This estimation problem can be reformulated as estimating the probability $F(h(y_\iota|\boldsymbol{x}))$ of the binary event $I(Y \leq y_\iota)$, I denotes the indicator function, and can be solved by minimising the log score

$$\begin{aligned} LS &= -\frac{1}{N \cdot n} \sum_{i=1}^{N} \sum_{\iota=1}^{n} I(Y_i \leq y_\iota) \log(F(h(y_\iota|\boldsymbol{x}_i))) + \\ &\qquad I(Y_i > y_\iota) \log(1 - F(h(y_\iota|\boldsymbol{x}_i))). \end{aligned} \tag{3.1}$$

The log score measures the mismatch between the individual empirical distribution functions of subjects $i = 1, \dots, N$, and the corresponding probabilities of the conditional distribution function $F(h(y_\iota|\boldsymbol{x}_i))$ resulting from the CTM in terms of the negative binomial log-likelihood. The log score is evaluated on the grid of response values $\{y_\iota|\iota = 1, \dots, n\}$. The log score is minimised with respect to h, what results in a consistent estimate of the conditional transformation function $\hat{h}$ (Hothorn et al., 2014).

In Hothorn et al. (2014), a component-wise boosting algorithm is presented for the efficient estimation of CTMs (see Appendix B). This boosting algorithm indirectly controls for the functional form and complexity of the estimated conditional transformation function $\hat{h}$. A thorough and general introduction to component-wise boosting can be found, *e.g.*, in Bühlmann and Hothorn (2007) and Schmid and Hothorn (2008). The characteristics and the functionality of the boosting algorithm as well as the specification of the conditional transformation function for C(L)TMs are discussed more thoroughly in Chapter 4 and Chapter 5. The component-wise boosting algorithm is adapted to the estimation of CLTMs in Chapter 4. Furthermore, we extend the log score (Equation (3.1)) to right-censored responses by including inverse probability of censoring weights in Chapter 5. The component-wise boosting algorithm based on this censored log score is especially useful for the estimation of conditional survivor functions.

The main advantages of component-wise boosting algorithms are the properties of intrinsic variable selection and model choice. Due to the component-wise estimation, boosting algorithms are applicable to high-dimensional data, and very complex conditional transformation functions h that depend on many explanatory variables can be considered. Nevertheless, there is no large sample theory for boosting, and hence, p-values are not available. Additionally, confidence intervals cannot be obtained based on large sample theory but they can be obtained using bootstrap approaches instead, what is usually time-consuming.

3.2. Likelihood-based estimation

Our literature review in Section 2.1 clarified that multiple ways to estimate linear transformation models have been suggested in the past. These various estimation strategies had to be adapted if the model equation of linear transformation models was extended (*e.g.*, by the inclusion of non-linear covariate effects), or if different censoring patterns were considered. Hence, every model extension and every censoring pattern led to necessary modifications of the estimation algorithm. Therefore, most authors presented their own estimation algorithms that accounted for the assumed model structure and the respective censoring pattern. A unified estimation approach for linear transformation models is lacking.

Additionally, the simultaneous estimation of the regression coefficients $\boldsymbol{\beta}$ and the response transformation h_Y (Equation 2.2) often caused problems. Therefore, various baseline-free approaches, where the estimation of h_Y is omitted, have been suggested (for a summary of baseline-free approaches see Section 2.1). Nevertheless, the estimation of h_Y is essential if the prediction of the conditional distribution function of the response is of interest. Therefore, we present a new unified likelihood-based estimation approach for C(L)TMs that is able to consider conditional transformation functions of arbitrary flexibility. In this estimation approach, covariate effects $\boldsymbol{\beta}$ and the response transformation h_Y are estimated simultaneously. Furthermore, the approach can be easily adapted to any censoring pattern by considering the respective censored log-likelihood function.

3.2.1. Log-likelihoods for C(L)TMs

Censoring has been an important topic in linear transformation models because well-known models for the analysis of survival times (*e.g.*, the PH, the PO, and the AFT model) are included in the model class. However, it is important to note that we do not focus on survival times in this chapter. CTMs are a useful model class for continuous or ordinal response variables, and the response variable can be either uncensored or (arbitrarily) censored. For example, the PH model is also not restricted to survival times. Instead, it can be very useful for analysing the conditional distribution function of any non-negative (un)censored response variable. Hence, the log-likelihoods introduced below are applicable to any (continuous or ordinal) response variable. Uniformly for all response variables, we considered uncensored observations, and right-, left-, doubly-, and interval-censored responses.

The respective censoring scheme has to be carefully considered when constructing the corresponding log-likelihood (Klein and Moeschberger, 2003). The following cases have to be distinguished regarding the information content that is provided:

- Exact observations provide information on the probability of the response occurring exactly at the observed response value, *i.e.* the provided information equals the density $f(y|\boldsymbol{x})$.

- For right-censored observations, we only know that the response value is larger than the observed right-censored response value, *i.e.* the provided information equals the conditional survivor function $1 - F(y|\boldsymbol{x})$.

- For left-censored observations, we only know that the response value is smaller than the observed left-censored response value, *i.e.* the provided information equals the conditional distribution function $F(y|\boldsymbol{x})$.

- For interval-censored observations, we only know that the response value lies in an interval with lower bound L and upper bound R, *i.e.* the provided information equals $F(R|\boldsymbol{x}) - F(L|\boldsymbol{x})$.

First, we determine the log-likelihood functions for a continuous (censored) response variable $Y \in \mathbb{R}$ using the conditional distribution function and the conditional density defined for CTMs in Equation (2.7). Without loss of generality, multiple observations are ignored.

- **Log-likelihood for an exact observation**:

$$l(h; y, \boldsymbol{x}) = \log[f(h(y|\boldsymbol{x}))] + \log[h'(y|\boldsymbol{x})], \tag{3.2}$$

  where $h'(\cdot)$ denotes the first derivative of the conditional transformation function with respect to y.

- **Log-likelihood for a right-censored observation**:

$$l(h; y, \boldsymbol{x}, \delta) = \delta \cdot \{\log[f(h(y|\boldsymbol{x}))] + \log[h'(y|\boldsymbol{x})]\} + (1-\delta) \cdot \log[1 - F(h(y|\boldsymbol{x}))], \tag{3.3}$$

  where $\delta = I(Y \leq C_R)$ denotes the right-censoring indicator, and C_R denotes the right-censoring (random) variable. The right-censoring indicator δ is equal to one for an exact observation, and is equal to zero for a right-censored observation.

- **Log-likelihood for a left-censored observation**:

$$l(h; y, \boldsymbol{x}, \tilde{\delta}) = \tilde{\delta} \cdot \{\log[f(h(y|\boldsymbol{x}))] + \log[h'(y|\boldsymbol{x})]\} + (1-\tilde{\delta}) \cdot \log[F(h(y|\boldsymbol{x}))], \tag{3.4}$$

  where $\tilde{\delta} = I(Y \geq C_L)$ denotes the left-censoring indicator, and C_L denotes the left-censoring (random) variable. The left-censoring indicator $\tilde{\delta}$ is equal to one for an exact observation, and is equal to zero for a left-censored observation.

- **Log-likelihood for a doubly censored observation**:

$$\begin{aligned} l(h; y, \boldsymbol{x}, \boldsymbol{\delta}) &= \delta_E \cdot \{\log[f(h(y|\boldsymbol{x}))] + \log[h'(y|\boldsymbol{x})]\} \\ &\quad + (1-\delta_R) \cdot \log[1 - F(h(y|\boldsymbol{x}))] + (1-\delta_L) \cdot \log[F(h(y|\boldsymbol{x}))], \end{aligned} \tag{3.5}$$

  where $\delta_R = I(Y \leq C_R)$ denotes the right-censoring indicator, $\delta_L = I(Y \geq C_L)$ denotes the left-censoring indicator, and $\delta_E = 1 - I(Y < C_L) - I(Y > C_R)$ denotes the indicator for an exact observation; $\boldsymbol{\delta} = (\delta_E, \delta_R, \delta_L)$. Thereby, C_L denotes the left-censoring (random) variable and C_R denotes the right-censoring (random) variable. Hence, a doubly censored observation might be either left- or right-censored.

- **Log-likelihood for an interval-censored observation:**

$$l(h; R, L, \boldsymbol{x}) = \log[F(h(R|\boldsymbol{x})) - F(h(L|\boldsymbol{x}))], \tag{3.6}$$

  whereby the individual experiences the event of interest somewhere in the interval $(L, R]$, $L \in \mathbb{R}$, $R \in \mathbb{R}$, $L < R$.

If the response variable $Y \in \{c_1, \ldots, c_q\}$ is ordinal, the log-likelihood functions have to be adapted accordingly by using the conditional distribution function and the conditional probability function for ordinal responses defined in Equation (2.8) and Equation (2.9), respectively. Without loss of generality, multiple observations are ignored.

- **Log-likelihood for an exact observation:**

$$
\begin{aligned}
l(h; y, \boldsymbol{x}) &= \log \Bigg[\sum_{k=2}^{q} I(y = c_k)(F(h_k(\boldsymbol{x})) - F(h_{k-1}(\boldsymbol{x}))) + \\
&\qquad I(y = c_1) \cdot F(h_1(\boldsymbol{x}))\Bigg] \qquad (3.7)
\end{aligned}
$$

- **Log-likelihood for a right-censored observation:**

$$
\begin{aligned}
l(h; y, \boldsymbol{x}, \delta) &= \delta \cdot \log \Bigg[\sum_{k=2}^{q} I(y = c_k)(F(h_k(\boldsymbol{x})) - F(h_{k-1}(\boldsymbol{x}))) + \\
&\qquad I(y = c_1) \cdot F(h_1(\boldsymbol{x}))\Bigg] + \\
&\quad + (1 - \delta) \cdot \log \left\{\sum_{k=1}^{q} I(y = c_k) \cdot [1 - F(h_k(\boldsymbol{x}))]\right\}, \qquad (3.8)
\end{aligned}
$$

where $\delta = I(Y \leq C_R)$ denotes the right-censoring indicator, and C_R denotes the right-censoring (random) variable.

- **Log-likelihood for a left-censored observation:**

$$
\begin{aligned}
l(h; y, \boldsymbol{x}, \tilde{\delta}) &= \tilde{\delta} \cdot \log \Bigg[\sum_{k=2}^{q} I(y = c_k)(F(h_k(\boldsymbol{x})) - F(h_{k-1}(\boldsymbol{x}))) + \\
&\qquad I(y = c_1) \cdot F(h_1(\boldsymbol{x}))\Bigg] + \\
&\quad + (1 - \tilde{\delta}) \cdot \log \left\{\sum_{k=1}^{q} I(y = c_k) \cdot F(h_k(\boldsymbol{x}))\right\}, \qquad (3.9)
\end{aligned}
$$

where $\tilde{\delta} = I(Y \geq C_L)$ denotes the left-censoring indicator and C_L denotes the left-censoring (random) variable.

- **Log-likelihood for a doubly censored observation:**

$$\begin{aligned} l(h; y, \boldsymbol{x}, \boldsymbol{\delta}) &= \delta_E \cdot \log \left[\sum_{k=2}^{q} I(y = c_k)(F(h_k(\boldsymbol{x})) - F(h_{k-1}(\boldsymbol{x}))) + \right. \\ &\qquad \left. I(y = c_1) \cdot F(h_1(\boldsymbol{x})) \right] + \\ &\quad + (1 - \delta_R) \cdot \log \left\{ \sum_{k=1}^{q} I(y = c_k) \cdot [1 - F(h_k(\boldsymbol{x}))] \right\} \\ &\quad + (1 - \delta_L) \cdot \log \left\{ \sum_{k=1}^{q} I(y = c_k) \cdot F(h_k(\boldsymbol{x})) \right\}, \end{aligned} \tag{3.10}$$

  where $\delta_R = I(Y \leq C_R)$ is the right-censoring indicator, $\delta_L = I(Y \geq C_L)$ is the left-censoring indicator, and $\delta_E = 1 - I(Y < C_L) - I(Y > C_R)$ is the indicator for an exact observation; $\boldsymbol{\delta} = (\delta_E, \delta_R, \delta_L)$. Thereby, C_L denotes the left-censoring (random) variable and C_R denotes the right-censoring (random) variable.

- **Log-likelihood for an interval-censored observation:**

$$l(h; L, R, \boldsymbol{x}) = \log \left[\sum_{k=1}^{q} I(R = c_k) \cdot F(h_k(\boldsymbol{x})) - \sum_{k=1}^{q} I(L = c_k) \cdot F(h_k(\boldsymbol{x})) \right]. \tag{3.11}$$

  The individual experiences the event of interest somewhere in the interval $(L, R],\ L, R \in \{c_1, \ldots, c_q\},\ L < R$.

Of course, mixed censoring patterns are conceivable. If a data set consists of a mixture of exact, and left-, right- and interval-censored observations, the corresponding log-likelihood function is the sum of the different log-likelihood contributions of each censoring pattern. As an example, we presented the log-likelihood for doubly censored observations, which is a mixture of exact, left- and right-censored observations.

3.2.2. Estimation strategies

In this section, a maximum likelihood approach is presented as a unified estimation approach for C(L)TMs with various censoring patterns. So far, we presented log-likelihood functions for C(L)TMs for continuous as well as ordinal response variables, and for uncensored as well as right-, left-, doubly-, or interval-censored observations in Section 3.2.1. All log-likelihood functions depend on the conditional transformation function $h(y|\boldsymbol{x})$, which can be arbitrarily flexible in principle. Estimating C(L)TMs means estimating the conditional transformation function $h(y|\boldsymbol{x})$, *i.e.* the respective log-likelihood function has to be maximised with respect to h.

As we are interested in low-parametrised and interpretable CLTMs, we consider the conditional transformation functions $h(y|\boldsymbol{x})$ defined for the models CLTM A – CLTM E in Section 2.2.2 (Equation (2.12) – Equation (2.16)). The estimation of $h(y|\boldsymbol{x})$ is achieved by estimating the corresponding basis coefficients. Hence, models CLTM A – CLTM E are estimated by inserting the respective conditional transformation function $h(y|\boldsymbol{x})$ into one of the log-likelihood functions presented in Section 3.2.1. Afterwards, the resulting log-likelihood function is maximised with respect to the basis coefficients $\boldsymbol{\alpha}$, $\boldsymbol{\beta}_0$, and $\boldsymbol{\beta}_1$. By this, models CLTM A – CLTM E can be applied to continuous as well as ordinal response variables, and the response variable might be uncensored, or (arbitrarily) censored.
Certain important linear constraints for CLTMs have to be considered during estimation (see Section 2.2.2). Per definition, the conditional transformation function $h(y|\boldsymbol{x})$ and the response transformation $h_Y(y)$ have to be monotonically increasing in y. This results in the linear constraints $h^{'}(y|\boldsymbol{x}) > 0$ and $h^{'}_Y(y) > 0$ that have to be considered. Therefore, only optimisation algorithms that are able to handle linear constraints are appropriate for maximising the log-likelihood.

The main challenge when estimating CLTMs is the estimation of smooth functions, *i.e.* the estimation of the response transformation $h_Y(y)$, or flexible covariate effects (*e.g.*, $\beta_0(\boldsymbol{x})$ in Equation (2.11)). This is due to the estimation of smooth functions being associated with the estimation of an infinite-dimensional parameter vector. To face this problem, we present several estimation strategies for determining the conditional transformation function $\hat{h}(y|\boldsymbol{x})$, which turn the infinite-dimensional parameter estimation task into a low-parametrised estimation task. First, we suggest a parsimonious parametrisation of the response transformation $h_Y(y)$, or smooth covariate effects (*e.g.*, $\beta_0(\boldsymbol{x})$) in terms of fractional polynomials. Alternatively, the smooth functions in the conditional transformation function $h(y|\boldsymbol{x})$ can be parametrised using P-splines. As a third and fourth option, we suggest estimating CLTMs using an empirical Bayes or a full Bayesian approach. We also focus on the associated advantages and disadvantages.

Parsimonious parametrisation using fractional polynomials. Following Royston and Altman (1994), we first propose a parsimonious, parametric approach for the smooth functions in the conditional transformation function of CLTMs by considering fractional polynomials. This implies the advantage that the number of parameters, which have to be estimated, is still low. Hence, estimation can be carried out using a full maximum likelihood approach without further regularisation. Nevertheless, this parsimonious representation implies limited functional forms of the smooth functions and thus, the approach can be insufficient in some applications.

Fractional polynomials (FPs) extend the class of ordinary polynomials to a richer family of curves by including non-positive and fractional powers (Royston and Altman, 1994; Sauerbrei and Royston, 1999). Ordinary polynomial regression has the drawback that low order polynomials offer only a limited family of shapes and high order polynomials may fit poorly at extreme covariate values. Hence, FPs are useful in univariate and multivariate non-linear regression analysis due to their considerable flexibility (Royston et al., 1999; Royston and Sauerbrei, 2007), and conventional polynomials are included as special cases (Royston and Altman, 1994). In Royston and Altman (1994), the authors introduce the class of FPs

$$\phi_m(x;\boldsymbol{\xi},\mathbf{p}) = \xi_0 + \sum_{j=1}^{m}\xi_j \cdot x^{(p_j)}, \qquad x^{(p_j)} = \begin{cases} x^{p_j}, & p_j \neq 0 \\ \log(x), & p_j = 0 \end{cases},$$

where m denotes the degree of the FP, $\mathbf{p} = (p_1, \ldots, p_m)^\top$ is the vector of powers, $\boldsymbol{\xi} = (\xi_0, \ldots, \xi_m)^\top$ contains the real-valued coefficients, and $x^{(p_j)}$ denotes Box-Tidwell transformations. As m and $\mathbf{p}$ have to be chosen a priori, the estimation of a smooth function in terms of FPs means estimating the parameter vector $\boldsymbol{\xi}$. FPs are a family of curves whose powers are restricted to a predefined set of integer and non-integer values. Usually, the set $\mathcal{P} = \{-2,\ -1,\ -0.5,\ 0,\ 0.5,\ 1,\ 2, 3\}$ of powers is used (Royston and Altman, 1994).

The choice of the degree m of the FPs is important because ,*e.g.*, the class of FPs of degree 2 (FP(2)) is richer than the class of FPs of degree 1 (FP(1)). Higher degrees are seldom necessary in practice, and Sauerbrei and Royston (1999) advise against it because this may result in overcomplex and uninterpretable models. We choose $FP(1)$ for our purpose, which includes reciprocal, logarithmic, square-root, linear, and square transformations. The associated set of available functional forms is quite rich, and including more powers usually offers only slight model improvement (Royston et al., 1999).

In general, the appropriate degree and the appropriate power transformation can be selected using the deviance or likelihood-ratio tests (Royston and Altman, 1994). To avoid the suggested selection procedures for finding the most appropriate FP, we model the smooth function in terms of an additive combination of all $FP(1)$-terms. For example, the conditional transformation function $h(y|\boldsymbol{x})$ of model CLTM C (Equation (2.14)) can be written as

$$\begin{aligned} h(y|\boldsymbol{x}) \;&=\; \alpha_0 + \alpha_1 \cdot y^{-2} + \alpha_2 \cdot y^{-1} + \alpha_3 \cdot y^{-\frac{1}{2}} + \alpha_4 \cdot \log(y) + \alpha_5 \cdot y^{\frac{1}{2}} + \alpha_6 \cdot y + \\ &\quad\; \alpha_7 \cdot y^2 + \alpha_8 \cdot y^3 + \boldsymbol{x}^\top \boldsymbol{\beta}_0, \end{aligned} \tag{3.12}$$

whereby the smooth response transformation $h_Y(y)$ is formulated in terms of FPs. Coefficients α_j, $j = 1, \ldots, 8$, close to or equal to zero indicate that the corresponding FP is not able to display the functional form of $h_Y(y)$. CLTM C is low-parametrised because the estimation of $h(y|\boldsymbol{x})$ is associated with the estimation of the nine parameters $\alpha_0, \ldots, \alpha_8$, and the vector of regression coefficients $\boldsymbol{\beta}_0$. For CLTM C, the linear monotonicity constraints

$$\begin{aligned} h^{'}(y|\boldsymbol{x}) = h^{'}_Y(y) \;&=\; -2 \cdot \alpha_1 \cdot y^{-3} - \alpha_2 \cdot y^{-2} - \frac{1}{2} \cdot \alpha_3 \cdot y^{-\frac{3}{2}} + \alpha_4 \cdot \frac{1}{y} + \\ &\quad\; \frac{1}{2} \cdot \alpha_5 \cdot y^{-\frac{1}{2}} + \alpha_6 + 2 \cdot \alpha_7 \cdot y + 3 \cdot \alpha_8 \cdot y^2 \overset{!}{>} 0 \end{aligned} \tag{3.13}$$

have to be considered during estimation. There is only a unique solution for $\hat{h}_Y(y)$ if all FPs are uncorrelated. Hence, if a parsimonious and interpretable estimated function is of interest, the FPs in Equation (3.12) have to be uncorrelated first. If only an appropriate estimated functional form (possibly with a circumstantial representation) is of interest, uncorrelating the FPs is optional. Nevertheless, the FPs in Equation (3.12) are usually highly correlated. Hence, their uncorrelation is important if standard errors or confidence intervals for the basis coefficients are of interest.

Parametrisation using penalised B-spline basis functions. An alternative nonparametric approach for the smooth, non-linear functions in CLTMs is a P-spline approach, where B-spline basis functions are used for parametrisation. The main advantage of nonparametric regression approaches over parametric regression approaches (*e.g.*, ordinary polynomial regression, fractional polynomials) is that a richer pool of functional shapes can be displayed. In principle, arbitrary shapes of smooth functions can be estimated in nonparametric regression. Nevertheless, a penalisation term has to be added to the log-likelihood function (Section 3.2.1) to guarantee a smooth function estimate. Hence, higher flexibility comes at the price of a penalised maximum likelihood approach instead of a full maximum likelihood approach. Penalised maximum likelihood approaches are associated with the selection of smoothing parameters, which has to be conducted carefully. This can be time-consuming especially if more than one smoothing parameter needs to be determined. Moreover, compared to fractional polynomials, the number of parameters that have to be estimated in P-spline approaches is considerably higher.

Due to their favourable numerical properties owing to their local definition, the smooth functions in CLTMs are formulated using B-spline basis functions. For example, the B-spline representation of the conditional transformation function $h(y|\boldsymbol{x})$ of CLTM C (Equation (2.14)) is

$$h(y|\boldsymbol{x}) = \sum_{d=1}^{D} \alpha_d \cdot B_d^l(y) + \boldsymbol{x}^\top \boldsymbol{\beta}_0 = B\boldsymbol{\alpha} + \boldsymbol{x}^\top \boldsymbol{\beta}_0, \tag{3.14}$$

where $B_d^l(y)$, $d = 1, \dots, D$, denote B-spline basis functions with degree l, B denotes the B-spline design matrix, and $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_D)^\top$ denotes the vector of B-spline basis coefficients. The number D of basis functions is determined by the number of (equidistant) knots and the degree l of the basis functions. Per default, we use B-spline basis functions of degree 3 and 20 equidistant interior knots, which results in $D = 24$ basis functions. Hence, formulating $h_Y(y)$ in terms of fractional polynomials is associated with the estimation of nine regression coefficients, whereas the more flexible formulation in terms of B-spline basis functions requires the estimation of 24 basis coefficients. Additionally, the response transformation function $h_Y(y)$ is assumed to be monotonically increasing. This restriction can be considered, *e.g.*, by using T-splines (Beliakov, 2000). T-splines are a simple transformation of B-splines that guarantees a function estimate that is monotonically increasing if all basis coefficients are estimated to be positive except the first coefficient, *i.e.* $\alpha_2, \dots, \alpha_{24} > 0$.

All log-likelihoods in Section 3.2.1 require the first derivative of the transformation function $h'(y|\boldsymbol{x})$. If fractional polynomials are used to parametrise smooth functions, the first derivative can be derived analytically, but also in case of a B-spline representation the determination is straightforward. For example, $h'(y|\boldsymbol{x})$ for model CLTM C is (Fahrmeir et al., 2013):

$$h'(y|\boldsymbol{x}) = h'_Y(y) = \frac{\partial}{\partial y} \sum_d \alpha_d \cdot B_d^l(y) = l \cdot \sum_{d=2}^{D} \frac{\alpha_d - \alpha_{d-1}}{\kappa_{d+l} - \kappa_d} B_d^{l-1}(y) = B'\boldsymbol{\alpha}, \tag{3.15}$$

where κ denotes the knots. In short, B' denotes the B-spline design matrix for the first derivative $h'_Y(y)$. The first derivative of the B-spline can be obtained in terms of differences of adjacent basis coefficients, and B-spline basis functions with degree lowered by one. Hence, the B-spline approach has the appealing property that the estimation of the basis coefficients $\hat{\boldsymbol{\alpha}}$ results in an estimate for the function itself *and* for its first derivative.

A smooth function estimate is guaranteed by the introduction of a roughness penalty that prevents overfitting. A common and appropriate measure for the variability of a function is the integral over the squared second derivative because the second derivative measures the curvature of the function. For B-splines, the second derivative of an arbitrary function g can be easily approximated by the second order differences of adjacent basis coefficients (Eilers and Marx, 1996):

$$\lambda \int (g''(y))^2 \, dy \approx \lambda \sum_{d=3}^{D} (\Delta^2 \alpha_d)^2 = \lambda \boldsymbol{\alpha}^\top K_2 \boldsymbol{\alpha}, \tag{3.16}$$

where $\Delta^2\alpha_d = \alpha_d - 2\alpha_{d-1} + \alpha_{d-2}$ denotes second order differences, K_2 denotes the penalty matrix based on second differences, and $\lambda > 0$ denotes the smoothing parameter. The smoothness penalty term (Equation (3.16)) has to be added to the log-likelihood l. This results in the penalised log-likelihood

$$l_p = l - \frac{\lambda}{2} \boldsymbol{\alpha}^\top K_2 \boldsymbol{\alpha}.$$

For example, the uncensored penalised log likelihood function for model CLTM C (using Equation (3.14), Equation (3.15) and Equation (3.2)) is

$$l_p(\boldsymbol{\alpha}, \boldsymbol{\beta}_0) = \log(f(B\boldsymbol{\alpha} + \boldsymbol{x}^\top \boldsymbol{\beta}_0)) + \log(B'\boldsymbol{\alpha}) - \frac{\lambda_2}{2} \boldsymbol{\alpha}^\top K_2 \boldsymbol{\alpha} - \frac{\lambda_3}{2} \boldsymbol{\alpha}^\top K_3 \boldsymbol{\alpha},$$

and has to be optimised with respect to the vector of basis coefficients $\boldsymbol{\alpha}$ and the vector of regression coefficients $\boldsymbol{\beta}_0$. As the uncensored log-likelihood contains both h and h', we include the penalty matrices for second and third differences, K_2 and K_3, to guarantee a smooth function estimate for h and h' (Simpkin and Newell, 2013). In analogy to the full maximum likelihood approach, the linear constraints $h'(y|\boldsymbol{x}) = B'\boldsymbol{\alpha} > 0$ have to be considered during optimisation to guarantee a monotonically increasing conditional transformation function. These linear constraints are also implicitly required by the log-likelihood function, where $B'\boldsymbol{\alpha}$ is log-transformed. If we consider a T-spline approach for $h_Y(y)$, which needs to be monotonically increasing, the basis coefficients $\alpha_2, \ldots, \alpha_{24}$ have to be non-negative in addition.

The smoothing parameter λ controls the compromise between the accurateness of the estimated function to the data and smoothness of the resulting function estimate. Hence, λ has to be chosen optimally and there are many ways to do so. One common option is to use model choice criteria, *e.g.*, to determine λ by using (generalised) cross validation (for an introduction to possible model choice criteria, we refer to Fahrmeir et al., 2013). If the considered CLTM involves more than one smooth function, additional penalty terms and smoothing parameters have to be included into the penalised log-likelihood function.

P-splines reviewed from the perspective of mixed models. An optimal selection of the smoothing parameter λ is important, but the choice of λ via cross validation can be circumstantial (especially for more than one smooth function). Furthermore, penalty terms cannot be easily integrated into maximum likelihood inference theory. Therefore, the consideration of P-splines from the perspective of mixed models is appealing. Bayesian approaches can be used to estimate the spline coefficients *and* the smoothing parameter directly from the data. In contrast to the penalised maximum likelihood approach presented above, no roughness penalty term is considered to achieve a smooth function estimate. Instead, an appropriate prior distribution is imposed on the basis coefficients $\boldsymbol{\alpha}$, which are considered as random effects. In the empirical Bayes approach, the variance parameters (which are the inverse smoothing parameters) are estimated by (approximate) restricted maximum likelihood (REML) methods. For a thorough introduction to the estimation of P-splines using mixed model approaches, we refer to Fahrmeir et al. (2004) and Fahrmeir et al. (2013).

Full Bayesian approach. The mixed model based smoothing of splines belongs to empirical Bayes approaches. The spline coefficients are considered as random effects but the variance parameters are estimated likelihood-based, and hence, in a frequentist way. Full Bayesian approaches are an alternative in this framework, where additional hyperpriors are defined for the variance parameters to provide prior distributions for all unknown parameters. An introduction to Bayesian P-splines using a full Bayesian approach is, *e.g.*, given in Lang and Brezger (2004). Both Bayesian approaches have appealing properties (Fahrmeir et al., 2004): In the empirical Bayes approach estimates are obtained by maximising objective functions. Hence, questions about the convergence of MCMC or the sensitivity to the choice of hyperparameters do not arise. On the other hand, in the full Bayesian approach characteristics and functionals of posteriors can be computed without relying on large sample normality approximations. As MCMC techniques require only local computations, the approach is also applicable to massive data sets, where the empirical Bayes approach has its limits.

As a proof of concept, we present an empirical evaluation of likelihood-based C(L)TMs in Chapter 6, and analyse the birth weight of newborns from the Perinatal Database Erlangen using likelihood-based CLTMs in Chapter 7. For the estimation of the conditional transformation functions, we consider only the full maximum likelihood approach based on fractional polynomials and the penalised maximum likelihood approach based on P-splines. Although both Bayesian approaches have appealing properties, there remains one important problem to be solved. The conditional transformation function $h(y|\boldsymbol{x})$ does not only imply certain smoothness conditions but has to be monotonically increasing in addition. Hence, an appropriate prior distribution for the spline coefficients needs to consider both smoothness and monotonicity.

3.3. Concluding remarks

As seen in the review of the literature on linear transformation models in Section 2.1, various estimation strategies have been suggested over the years. These estimation strategies include estimating equations, partial and marginal log-likelihoods, and Bayesian approaches. Remarkably, a considerable number of the proposed approaches are baseline-free approaches, *i.e.* only the vector of regression coefficients $\boldsymbol{\beta}$ is estimated, whereas the response transformation $h_Y(y)$ is considered as a nuisance parameter of high dimensionality. Nevertheless, this procedure becomes a dead end if conditional distribution function values are of interest. For example, the estimation of the conditional survival probabilities over time depending on individual patient characteristics is often of special interest in survival analysis. Therefore, in our opinion, the simultaneous estimation of the regression coefficients $\boldsymbol{\beta}$ and the response transformation function $h_Y(y)$ is desirable. Additionally, many algorithms have been proposed for estimating linear transformation models, and a unified estimation procedure is lacking. Most of the cited authors suggested and implemented new algorithms for their extensions of the linear transformation model to specific censoring patterns or to specific structures of the model equation.

Therefore, we present two estimation approaches for C(L)TMs. As the whole conditional distribution function is modelled, the estimation of C(L)TMs is no baseline-free approach in general, and the response transformation and covariate effects are estimated simultaneously. First, Hothorn et al. (2014) suggested a component-wise boosting approach for the estimation of CTMs. This estimation approach is adapted to the estimation of CLTMs in Chapter 4, and to right-censored response variables in Chapter 5. The extension to alternative censoring patterns is not straightforward, and needs further investigation. Due to the component-wise estimation in the boosting algorithm, a complex conditional transformation function can be considered that might depend on a large number of explanatory variables.

Second, we presented a unified maximum likelihood approach for the estimation of

C(L)TMs, *i.e.* the same estimation approach can be used for all models irrespective of the corresponding model complexity. The direct usage of the (censored) log-likelihood for estimating transformation models has been considered very rarely in the past. The main obstacle for directly optimising the (censored) log-likelihood was the requirement of a simultaneous estimation of the response transformation and the covariate effects. For example, two related approaches can be found in Ma et al. (2014) and in Crowther and Lambert (2014). A flexible parametric approach for survival data analysis is proposed in Crowther and Lambert (2014), where the baseline hazard function and time-dependent effects are modelled using restricted cubic splines. Model estimation is based on the full censored log-likelihood. Ma et al. (2014) suggest a penalised maximum likelihood approach for the simultaneous estimation of the baseline hazard and the regression coefficients in PH models, whereby a penalty function is used to smooth the baseline hazard estimate. Additionally, the likelihood-based approach can be easily adapted to any kind of censoring (right-, left-, doubly-, or interval-censoring) by simply considering the corresponding censored log-likelihood. Standard optimisation algorithms that are able to consider linear constraints (*e.g.*, the `constrOptim`-function of the R base-package **stats**) can be used to maximise the (censored) log-likelihood. So far, the maximum likelihood estimation approach is restricted to low-parametrised CLTMs. Its extension to more flexible C(L)TMs would be worthwhile. Therefore, questions concerning algorithmic feasibility and identifiability of model components have to be answered first.

Concerning model estimation, we presented four options for the estimation of CLTMs and focused on the associated advantages and disadvantages. First, we suggested a low-parametrised approach, where smooth functions are represented in terms of fractional polynomials. This approach has the important advantage that no tuning parameters, *e.g.*, smoothing parameters, need to be determined, but this approach can only display a limited range of functional shapes. Hence, the parametrisation of smooth functions using B-splines constitutes a more flexible alternative. Nevertheless, the penalised estimation of B-splines requires the determination of smoothing parameters. This task can be especially challenging here because the estimation of CLTMs requires the determination of a two-dimensional smoothing parameter, which is usually a computationally complex task. To complete our list, we also suggested the estimation of CLTMs using empirical Bayes or full Bayesian approaches. One important advantage of Bayesian approaches is that smoothing parameters are directly estimated in a data-driven way. Nevertheless, some important questions remain to be solved, *e.g.*, how prior distributions can include monotonicity requirements.

However, we did not aim at providing a full theory for likelihood-based conditional transformation models in all its refinement. We were rather interested in putting transformation models in perspective by showing their wide applicability. Furthermore, we structured the tangled mass of estimation approaches that has been suggested for (linear) transformation models in the past in Section 2.1. It is important to note that the estimation of C(L)TMs for (un)censored response variables can be considerably simplified and unified by our likelihood-based estimation procedure. As a proof of concept, uncensored responses are analysed using low-parametrised, likelihood-based CLTMs in Chapter 6 and Chapter 7.

4. Predicting birth weight by boosting conditionally linear transformation models

The content of this chapter is already published in Möst et al. (2014).

Low and high birth weight are important risk factors for neonatal morbidity and mortality. Gynaecologists must therefore accurately predict birth weight before delivery. Most prediction formulas for birth weight are based on prenatal ultrasound measurements carried out within one week prior to birth. Although successfully used in clinical practice, these formulas focus on point predictions of birth weight but do not systematically quantify uncertainty of the predictions, *i.e.* they result in estimates of the conditional mean of birth weight but do not deliver prediction intervals.

To overcome this problem, we introduce the model class of conditionally linear transformation models (CLTMs) (see also Chapter 2) more generally in this chapter to predict future birth weight. Instead of focusing only on the conditional mean, CLTMs model the whole conditional distribution function of birth weight given prenatal ultrasound parameters. Consequently, the CLTM approach delivers both point predictions of birth weight and fetus-specific prediction intervals. Prediction intervals constitute an easy-to-interpret measure of prediction accuracy and allow identification of fetuses subject to high prediction uncertainty.

Using the Perinatal Database Erlangen, we analyse variants of CLTMs and compare them to standard linear regression estimation techniques used in the past and to quantile regression approaches. Special focus is on the quality of the associated prediction intervals, whereby the conditional coverage of the prediction intervals and the average interval length are used as quality criteria.

4.1. Introduction

Birth weight (BW) is among the most important risk indicators for neonatal morbidity and mortality (McCormick, 1985; Sappenfield et al., 1987). As shown in numerous studies, high birth weight is associated with serious maternal trauma after vaginal and surgical delivery and shoulder dystocia with fetal brachial plexus paralysis and/or clavicular fracture (Boulet et al., 2003; Ecker et al., 1997), and low birth weight increases the risk of neurological and developmental deficits during childhood (Bernstein et al., 2000; McIntire et al., 1999). The accurate estimation of birth weight is challenging for gynaecologists who need to plan the mode of delivery and organise obstetric management.

Fetal ultrasound examinations have become routine during the last 40 years (Scioscia et al., 2008) and result in readily available two-dimensional measurements highly correlated with birth weight. Most prediction formulas for birth weight incorporate biometric parameters, such as biparietal diameter (BPD), fronto-occipital diameter (FOD), head circumference (HC), abdominal transverse diameter (ATD), anterior-posterior abdominal diameter (APD), abdominal circumference (AC) and femur length (FL). Here we focus on the *statistical* aspects of prediction formulas for birth weight. Our analysis is based on prenatal ultrasound measurements recorded within seven days before delivery of $N = 8,712$ babies at the Perinatal Centre of the University Clinic Erlangen, Germany, in 2003–2011.

Statistically, the development of a prediction formula for birth weight is a regression modelling task that involves the accurate estimation of ultrasound predictor effects on birth weight:

1. Many traditional prediction formulas for birth weight have been derived by applying linear regression models with Gaussian errors (Scioscia et al., 2008; Siemer et al., 2008; Siggelkow et al., 2011; Hoopmann et al., 2010). Only little attention has been given to the frequent departure of the distribution of birth weight from the normal distribution, which could make relying on a Gaussian model suboptimal. For example, if a high percentage of the newborns are very small, the distribution of birth weight would not be normal but rather right skewed. A suitable approach to model birth weight should take this skewness into account.

2. A thorough investigation of the accuracy of the prediction formulas is essential for clinical practice because, as stated by, *e.g.*, Scioscia et al. (2008), many prediction formulas show the same tendency to under- and over-estimate birth weight at the extremes, regardless of the ultrasound parameters relied upon. To assess the performance of new prediction formulas, measures such as the relative percentage error (defined as $(\mathrm{BW}-\widehat{\mathrm{EW}})/\mathrm{BW}$) and the absolute percentage error (defined as $|\mathrm{BW}-\widehat{\mathrm{EW}}|/\mathrm{BW}$) have been commonly used, where $\widehat{\mathrm{EW}}$ denotes estimated fetal weight (*e.g.*, Scioscia et al., 2008; Dammer et al., 2013; Faschingbauer et al., 2012). As the traditional formulas for predicting birth weight estimate only the conditional mean, the aforementioned performance measures focus on the quality of the point estimates for the actual birth weight, and an appropriate measure of prediction *uncertainty* is missing. An easy-to-interpret measure of prediction accuracy accompanied with some measure of uncertainty is interval estimates that cover the true weights of newborns with a high probability. Although it is possible to construct prediction intervals around the point estimates obtained from the Gaussian modelling approach mentioned above, these intervals are subject to potential bias. First, intervals obtained from Gaussian

> models are always symmetric around the conditional mean. Consequently, these intervals might be suboptimal because the distribution of birth weight (and possibly also the distribution of the residuals in linear regression) is skewed. Second, Gaussian prediction intervals all have the same length owing to a constant residual variance term, regardless of the ultrasound measurements. This assumption is often inappropriate as the prediction accuracy may depend on the actual birth weight (via the ultrasound measurements), *e.g.*, larger fetuses might have wider prediction intervals than smaller fetuses.

To address these issues, we propose conditionally linear transformation models (CLTMs) as a novel approach to predict birth weight. Instead of considering the conditional mean only (as traditional Gaussian regression does), CLTMs model the whole conditional distribution function of birth weight given prenatal ultrasound parameters. Consequently, each quantile of the birth weight distribution can be predicted by a single CLTM. This implies that the CLTM approach not only results in point predictions of birth weight (*i.e.* , in predictions of the median) but additionally result in fetus-specific prediction intervals (whose boundaries are given, *e.g.*, by the predicted 10% and 90% quantiles). The interval estimates obtained from CLTMs represent an easy-to-interpret measure of prediction accuracy and allow identification of fetuses subject to high prediction uncertainty. Moreover, interval lengths obtained from the CLTM approach depend on individual ultrasound measurements of each fetus. This strategy results in “personalised” prediction intervals for each fetus and clearly provides more information than classical point predictions alone.

In Section 4.2, we review common prediction formulas for birth weight and associated traditional methods of estimation. We also introduce the Perinatal Database Erlangen and discuss prediction intervals for birth weight. A thorough introduction to conditionally linear transformation models, including some comments on interpretability and estimation, is given in Section 4.3. We present the results of the analysis of the Perinatal Database Erlangen in Section 4.4 and discuss the results in Section 4.5.

4.2. Prediction of birth weight

4.2.1. Review of common prediction formulas for birth weight

Since the 1970s, gynaecologists have developed numerous formulas to predict birth weights based on prenatal ultrasound measurements. Summaries of these formulas are, *e.g.*, given in Dudley (2005), Scioscia et al. (2008) and Siemer et al. (2008). A well-established prediction formula commonly used in clinical practice is that proposed by Hadlock et al. (1985):

$$\log_{10}(\widehat{\mathrm{EW}}) = 1.304 + 0.05281 \times \mathrm{AC} + 0.1938 \times \mathrm{FL} - 0.004 \times \mathrm{AC} \times \mathrm{FL},$$

where biometric parameters are measured in centimetres and estimated fetal weight ($\widehat{\text{EW}}$) is measured in grams. In addition to classical prediction formulas based on 2-D ultrasound measurements, other formulas incorporate clinical parameters (Sabbagha et al., 1989) or 3-D ultrasound measurements (Schild et al., 2008), or focus on high-risk deliveries (*e.g.*, Hart et al., 2010; Faschingbauer et al., 2012; Dammer et al., 2013). Choi et al. (2012) suggest a model with spatio-temporally varying coefficients for low birth weights. Because 3-D ultrasound measurements do not seem to improve many predictions and are poorly suited for every-day clinical practice (Scioscia et al., 2008), we focused on routinely measured 2-D biometric parameters in our study. The traditional prediction formulas for birth weights that we are aware of were derived using linear regression approaches with Gaussian errors.

4.2.2. Perinatal Database Erlangen

Our analysis is based on data of $N = 8,712$ singleton pregnancies with a complete ultrasound examination within seven days before delivery. Biometric parameters included *biparietal diameter* (BPD), *fronto-occipital diameter* (FOD), *head circumference* (HC), *abdominal transverse diameter* (ATD), *anterior-posterior abdominal diameter* (APD), *abdominal circumference* (AC) and *femur length* (FL). Additionally, the mother's body mass index (BMI) was measured. In cases in which fetus growth was followed serially, we used measurements only from the last examination before delivery. All ultrasound measurements were made by experienced examiners who underwent extensive training at University Clinic Erlangen. Birth weight was measured by the nursing staff at Erlangen University Hospital within one hour after delivery. Children with chromosomal or structural malformations and intrauterine deaths were excluded from analysis.

4.2.3. Prediction intervals

Since we are interested in some measure that quantifies the uncertainty of predictions for birth weights, we considered fetus-specific *prediction intervals* (Mayr et al., 2012). These intervals result in a range of predicted values that cover the birth weight with high probability $1 - \alpha$, where α is a pre-specified error level.

A common way to define the boundaries of a prediction interval is to use the $\alpha/2$ quantile and the $(1 - \alpha/2)$ quantile of the conditional distribution of birth weight given ultrasound measurements:

$$\widehat{\text{PI}}_{1-\alpha}(\boldsymbol{x}) = \left[\hat{q}_{\alpha/2}(\boldsymbol{x}), \hat{q}_{1-\alpha/2}(\boldsymbol{x})\right]. \tag{4.1}$$

Here, $\boldsymbol{x}$ denotes the ultrasound measurements of a new fetus, and $\hat{q}_{\alpha/2}$ and $\hat{q}_{1-\alpha/2}$ the $\alpha/2$ and the $(1 - \alpha/2)$ quantile, respectively, of the corresponding conditional distribution of birth weight. Since the estimated prediction intervals depend on the ultrasound measurements, the interval lengths and interval borders are fetus-specific. In other words, depending on the ultrasound measurements, accurate or inaccurate predictions can be made, which results in narrow or wide prediction intervals, respectively (Mayr et al., 2012). Nevertheless, the underlying assumptions of the regression model used (*e.g.*, normally distributed responses and homoscedasticity for linear regression models) in Equation 4.1 influence the form of the resulting prediction intervals. For example, the resulting prediction intervals may differ in symmetry assumptions and methods for boundary estimation. Common methods for the calculation of prediction intervals are, *e.g.*, linear regression or quantile regression approaches.

4.2.4. Existing approaches for calculation of prediction intervals

If linear regression models are used for birth weight prediction, the conditional mean of birth weight is modelled as a linear function of the (possibly transformed) prenatal ultrasound measurements. After estimation of the model parameters, symmetric prediction intervals are constructed around the point predictions based on the assumptions of homoscedasticity and normality (*e.g.*, Montgomery et al., 2012). Hence, the resulting symmetric prediction intervals are inadequate if the birth weight's distribution is skewed and if the residual variance depends on ultrasound measurements.

The use of linear or additive quantile regression approaches to determine prediction intervals for birth weight conveniently solves these problems. With quantile regression (Koenker et al., 1994; Koenker, 2005), one directly estimates the boundaries of the prediction intervals by using separate regression models for the quantiles $q_{\alpha/2}$ and $q_{1-\alpha/2}$ (Equation 4.1, Meinshausen, 2006). The influence of the ultrasound parameters on the respective quantiles is assumed to be additive. Although this approach avoids any distributional assumptions, a non-trivial problem associated with quantile regression is quantile crossing (Dette and Volgushev, 2008). The logical monotonicity requirements of the probability p ($p = q^{-1}$) are not fulfilled, and neighbouring quantile curves may cross because they are estimated independently.

To avoid quantile crossing (and also the aforementioned problems associated with linear regression), we propose CLTMs to estimate intervals for the prediction of birth weight. In contrast to quantile regression approaches, CLTMs model all conditional quantiles simultaneously by estimating the whole conditional distribution function, and the relevant quantiles are extracted afterwards. Thereby, inconsistencies between neighbouring quantiles are avoided.

4.3. Conditionally linear transformation models

4.3.1. Model class

CLTMs are a special case of CTMs that model the conditional distribution function of a response $Y_{\boldsymbol{x}} = (Y|\mathbf{X} = \boldsymbol{x})$ depending on explanatory variables $\boldsymbol{x}$. The advantages of modelling the conditional distribution function directly are summarised in Chapter 1, and a short introduction to CTMs is given in Section 1.1. We used the CTM approach to model the conditional distribution function of birth weight depending on prenatal ultrasound measurements:

$$\mathbb{P}(\mathrm{BW} \leq \upsilon|\mathbf{X} = \boldsymbol{x}) = F_{\mathrm{BW}|\mathbf{X}=\boldsymbol{x}}(\upsilon) = F(h(\upsilon|\boldsymbol{x})), \tag{4.2}$$

where $\upsilon \in \mathbb{R}$ denotes some arbitrary birth weight. The transformation function h transforms the birth weights conditionally on $\boldsymbol{x}$, so that the distribution of the transformed birth weights follows the distribution function F. As we modelled the whole conditional distribution function, higher moments (*e.g.*, the variance) may also depend on ultrasound measurements. In addition, further moments of the prediction distribution of birth weight can be modelled flexibly, *e.g.*, kurtosis and skewness.

Nevertheless, the CTMs presented in Hothorn et al. (2014) define a very complex and general class of transformation models, and therefore model interpretations can be challenging. Moreover, a lack of orthogonality of the model components constricts insights into model structure. As a consequence, direct interpretations of the relationship between the explanatory variables and certain moments of the distribution function of the response are difficult to obtain because these effects usually cannot be separated. As we are interested in a more easily interpretable version of CTMs in this application, we reduced the model complexity by imposing restrictions on CTMs and introducing CLTMs. The model class of CLTMs can be described by the following linear transformation conditional on $\boldsymbol{x}$:

$$\begin{aligned} h(Y_{\boldsymbol{x}}|\boldsymbol{x}) &= Z \sim F, \text{ with} \\ h(Y_{\boldsymbol{x}}|\boldsymbol{x}) &= h_0(Y_{\boldsymbol{x}}) \cdot \beta(\boldsymbol{x}) + \alpha(\boldsymbol{x}). \end{aligned} \tag{4.3}$$

Here, h denotes a monotone transformation function that depends on explanatory variables. The random variable Z is a transformation of the responses $Y_{\boldsymbol{x}}$ depending on explanatory variables $\boldsymbol{x}$ and follows the known distribution function F. In CLTMs, we modelled only linear functions of the transformed responses to reduce model complexity (Equation (4.3)). Hence, we considered a flexible and possibly unknown response transformation $h_0(Y_{\boldsymbol{x}})$ that depends only on the response values $Y_{\boldsymbol{x}}$. The response transformation itself was transformed by the explanatory variables via a linear function, where the coefficients $\alpha(\boldsymbol{x})$ and $\beta(\boldsymbol{x})$ depend on the explanatory variables. The coefficients $\alpha(\boldsymbol{x})$ induce shifts of the response transformation $h_0(Y_{\boldsymbol{x}})$, and the coefficients $\beta(\boldsymbol{x})$ induce shifts and scalings of the response transformation $h_0(Y_{\boldsymbol{x}})$ depending on the respective explanatory variables.

Owing to the restriction of the transformation function h to linear functions of the response transformation $h_0(Y_{\boldsymbol{x}})$, the influence of the explanatory variables $\boldsymbol{x}$ on the conditional mean and conditional variance of the response transformation can be displayed. This follows directly from calculating the conditional mean and conditional variance in Equation (4.3) and solving the equation for both $\mathbb{E}(h_0(Y_{\boldsymbol{x}})|\boldsymbol{x})$ and $\mathbb{V}(h_0(Y_{\boldsymbol{x}})|\boldsymbol{x})$:

$$\begin{aligned} \mathbb{E}(h_0(Y_{\boldsymbol{x}})|\boldsymbol{x}) &= \frac{\mathbb{E}(Z) - \alpha(\boldsymbol{x})}{\beta(\boldsymbol{x})} \\ \mathbb{V}(h_0(Y_{\boldsymbol{x}})|\boldsymbol{x}) &= \frac{\mathbb{V}(Z)}{\beta(\boldsymbol{x})^2}. \end{aligned} \tag{4.4}$$

If we assume that the transformed responses Z follow a standard normal distribution $Z \sim \mathcal{N}(0,1)$, we get $\mathbb{E}(Z) = 0$ and $\mathbb{V}(Z) = 1$, and Equation (4.4) simplifies accordingly. The coefficients $\alpha(\boldsymbol{x})$ influence only the conditional mean of the response transformation, whereas the coefficients $\beta(\boldsymbol{x})$ influence its conditional mean and its conditional variance. The influence of the explanatory variables on the conditional mean and conditional variance of the response transformation can be formulated in CLTMs, whereas such a formulation cannot be given in CTMs. This difference can also be seen by looking at the conditional quantile functions implied by CTMs and CLTMs:

$$\begin{aligned} Q_{\text{CTM}}(p|\boldsymbol{x}) &= h^{-1}(F^{-1}(p)|\boldsymbol{x}) \\ Q_{\text{CLTM}}(p|\boldsymbol{x}) &= h_0^{-1}\left(\frac{F^{-1}(p) - \alpha(\boldsymbol{x})}{\beta(\boldsymbol{x})}\right). \end{aligned}$$

For CTMs, the effect of the explanatory variables on the conditional quantile may vary with p, whereas in CLTMs, the conditional quantile is a nonlinear transformation of a linear function of $F^{-1}(p)$, where the coefficients of the latter do not depend on p.

Furthermore, we assumed additivity on the scale of the transformation function; therefore, we decomposed the monotone transformation function h into $J+1$ partial transformation functions, given the explanatory variables (Hothorn et al., 2014, and Equation (4.3)):

$$\begin{aligned} Z = h(Y_{\boldsymbol{x}}|\boldsymbol{x}) &= \sum_{j=0}^{J} h_j(Y_{\boldsymbol{x}}|\boldsymbol{x}) = \sum_{j=0}^{J} \left(h_0(Y_{\boldsymbol{x}}) \cdot \beta_j(\boldsymbol{x}) + \alpha_j(\boldsymbol{x})\right) \\ &= h_0(Y_{\boldsymbol{x}}) \cdot \sum_{j=0}^{J} \beta_j(\boldsymbol{x}) + \sum_{j=0}^{J} \alpha_j(\boldsymbol{x}). \end{aligned} \tag{4.5}$$

Despite this decomposition, the random variable Z still remains a linear function of the response transformation $h_0(Y_{\boldsymbol{x}})$.

Prominent members of the family of linear transformation models, most importantly the proportional hazards and the proportional odds model (Section 2.1.2), can be connected by restricting the above-mentioned CLTMs to the case where only shifts of the response transformation that depend on explanatory variables are allowed:

$$h(Y_{\boldsymbol{x}}|\boldsymbol{x}) = h_0(Y_{\boldsymbol{x}}) + \sum_{j=0}^{J} \alpha_j(\boldsymbol{x}) = h_0(Y_{\boldsymbol{x}}) + \alpha(\boldsymbol{x}). \tag{4.6}$$

In this model, the explanatory variables can only influence the mean $-\alpha(\boldsymbol{x})$ of the transformed response $h_0(Y_{\boldsymbol{x}})$. The transformation functions of the proportional hazards model and the proportional odds model result if we choose a CLTM (Equation (4.5)) with $\beta(\boldsymbol{x}) \equiv 1$ and an appropriate response transformation $h_0(Y_{\boldsymbol{x}})$, which is treated as a nuisance parameter in classical formulations of the proportional hazards model and proportional odds model. For linear shift functions $\alpha(\boldsymbol{x})$, a unified estimation framework has been proposed by Cheng et al. (1995).

We assumed that the response transformation $h_0(Y_{\boldsymbol{x}})$ is unknown. In the first step, we decomposed the response transformation into one part consisting only of linear functions and a more complex part representing deviations from linearity:

$$h_0(Y_{\boldsymbol{x}}) = \underbrace{\alpha_0 + \beta_0 \cdot Y_{\boldsymbol{x}}}_{\text{linear part}} + \underbrace{\tilde{h}_0(Y_{\boldsymbol{x}})}_{\text{deviations from linearity}}. \tag{4.7}$$

The decomposition in Equation (4.7) is reasonable because the model component $\tilde{h}_0(Y_{\boldsymbol{x}})$ can be used to decide whether the response variable follows a normal distribution or not, if we additionally set the link function to $F = \Phi$. If the model component $\tilde{h}_0(Y_{\boldsymbol{x}})$ is missing, we only observe a linear transformation of the conditional response, and hence we cannot leave the class of normal distributions because the normal distribution is invariant towards linear transformations. Consequently, by estimating the more complex deviations from linearity $\tilde{h}_0(Y_{\boldsymbol{x}})$, we are able to leave the class of normal distributions and model other classes of distribution functions as well.

Combining Equation (4.7) with the definition of CLTMs in Equation (4.3) leads to

$$h(Y_{\boldsymbol{x}}|\boldsymbol{x}) = (Y_{\boldsymbol{x}} + \tilde{h}_0(Y_{\boldsymbol{x}})) \cdot \beta(\boldsymbol{x}) + \alpha(\boldsymbol{x}) = Y_{\boldsymbol{x}} \cdot \beta_{\text{lin}}(\boldsymbol{x}) + \tilde{h}_0(Y_{\boldsymbol{x}}) \cdot \beta_c(\boldsymbol{x}) + \alpha(\boldsymbol{x}),$$

where $\beta_{\text{lin}}(\boldsymbol{x})$ denotes the part of $\beta(\boldsymbol{x})$ influencing the linear part of the response transformation $h_0(Y_{\boldsymbol{x}})$, and $\beta_c(\boldsymbol{x})$ denotes the part of $\beta(\boldsymbol{x})$ influencing the more complex deviations from linearity $\tilde{h}_0(Y_{\boldsymbol{x}})$.

We furthermore assumed that the more complex deviations $\tilde{h}_0(Y_{\boldsymbol{x}})$ do not depend on any explanatory variables; therefore, we set $\beta_c(\boldsymbol{x}) \equiv 1$. This is a strong assumption, but as we are interested in an interpretable model class, this is a necessary restriction of model complexity. The transformation function h with an unknown and decomposed response transformation at the start results in

$$h(Y_{\boldsymbol{x}}|\boldsymbol{x}) = Y_{\boldsymbol{x}} \cdot \beta_{\text{lin}}(\boldsymbol{x}) + \tilde{h}_0(Y_{\boldsymbol{x}}) + \alpha(\boldsymbol{x}).$$

Then, we included the decomposition of the monotone transformation function h into $J+1$ partial transformation functions (Equation (4.5)):

$$h(Y_{\boldsymbol{x}}|\boldsymbol{x}) = \tilde{h}_0(Y_{\boldsymbol{x}}) + Y_{\boldsymbol{x}} \cdot \sum_{j=0}^{J} \beta_{j,\text{lin}}(\boldsymbol{x}) + \sum_{j=0}^{J} \alpha_j(\boldsymbol{x}). \tag{4.8}$$

We furthermore set $\alpha_0(\boldsymbol{x}) \equiv \alpha_0$ and $\beta_{0,\text{lin}}(\boldsymbol{x}) \equiv \beta_0$, which we already implicitly did in Equation (4.7). By introducing the scalars α_0 and β_0, the transformation function h can be decomposed into an unconditional part (not depending on any explanatory variables) and a conditional part (depending on explanatory variables), which facilitates model interpretations. The resulting structure of the monotone transformation function is still consistent with the model class of CLTMs:

$$h(Y_{\boldsymbol{x}}|\boldsymbol{x}) = \underbrace{\alpha_0 + \beta_0 \cdot Y_{\boldsymbol{x}} + \tilde{h}_0(Y_{\boldsymbol{x}})}_{\text{unconditional part}} \; + \; \underbrace{Y_{\boldsymbol{x}} \cdot \sum_{j=1}^{J} \beta_j(\boldsymbol{x}) + \sum_{j=1}^{J} \alpha_j(\boldsymbol{x})}_{\text{conditional part}}. \tag{4.9}$$

Hence, in this model, only the linear part of the response transformation ($= Y_{\boldsymbol{x}}$) may depend on explanatory variables, whereas the function representing deviations from linearity $\tilde{h}_0(Y_{\boldsymbol{x}})$ is flexible and depends only on the response values $Y_{\boldsymbol{x}}$. Thereby, the explanatory variables solely influence the mean and variance of the transformed responses. We denote the coefficients $\beta_{j,\text{lin}}(\boldsymbol{x}), j = 1, \dots, J$ (Equation (4.8)) simply by $\beta_j(\boldsymbol{x})$ as we no longer need to distinguish the linear and the more complex part of the coefficient vector. In this model, we can estimate further characteristics of the conditional distribution function of the response (*e.g.*, skewness and kurtosis) in terms of $\tilde{h}_0(Y_{\boldsymbol{x}})$. Hence, we can only model constant kurtosis and skewness in contrast to quantile regression. A possible influence of the explanatory variables on higher moments can only be estimated in the more complex model class of CTMs. Besides, the definition of CLTMs that was used in Chapter 2 in Equation (2.11) is equivalent to Equation (4.9). To avoid confusion, please note, that regression coefficients, coefficient functions and response transformations were named differently. For example, the conditional transformation function $h_0(Y_{\boldsymbol{x}})$ is equivalent to the unconditional transformation function $h_Y(y)$ defined in Chapter 2.

By further differentiating between linear and flexible explanatory variable effects, we get:

Linear CLTM

$$h(Y_{\boldsymbol{x}}|\boldsymbol{x}) = \alpha_0 + \beta_0 \cdot Y_{\boldsymbol{x}} + \tilde{h}_0(Y_{\boldsymbol{x}}) + Y_{\boldsymbol{x}} \cdot \sum_{j=1}^{J} \beta_j \cdot \boldsymbol{x}_j + \sum_{j=1}^{J} \alpha_j \cdot \boldsymbol{x}_j,$$

where α_j and $\beta_j, j = 1, \dots, J$, are regression coefficients, and therefore the explanatory variables have a linear influence on the response transformation.

Additive CLTM

$$h(Y_{\boldsymbol{x}}|\boldsymbol{x}) = \alpha_0 + \beta_0 \cdot Y_{\boldsymbol{x}} + \tilde{h}_0(Y_{\boldsymbol{x}}) + Y_{\boldsymbol{x}} \cdot \sum_{j=1}^{J} \beta_j(\boldsymbol{x}) + \sum_{j=1}^{J} \alpha_j(\boldsymbol{x}),$$

where $\alpha_j(\boldsymbol{x})$ and $\beta_j(\boldsymbol{x}), j = 1, \dots, J$, denote smooth functions. Hence, the explanatory variables have a flexible influence on the response transformation.

Introduction of specific CLTMs for the analysis of the Perinatal Database Erlangen

For the analysis, we chose six variants of CLTMs with unknown response transformation CLTM 0 (linear) and CLTM 0 – CLTM 4, in which the models are ordered with increasing model complexity (Table 4.1). For comparison, we used the common conditional transformation model CTM as a reference model representing the most complex modelling approach. In general, we defined more complex CLTMs for the analysis of the Perinatal Database Erlangen here, but some of the considered models coinside with the CLTMs defined in Section 2.2.2. Such equivalences will be pointed out explicitly. To avoid confusion, please note again that regression coefficients, coefficient functions and response transformations were named differently in Chapter 2 and Chapter 4.

CLTM 0 (linear): Linear Transformation Model

$$h(Y_{\boldsymbol{x}}|\boldsymbol{x}) = Y_{\boldsymbol{x}} + \tilde{h}_0(Y_{\boldsymbol{x}}) + \sum_{j=1}^{J} \alpha_j \cdot \boldsymbol{x}_j \overset{\text{Equation (4.7)}}{=} h_0(Y_{\boldsymbol{x}}) + \sum_{j=1}^{J} \alpha_j \cdot \boldsymbol{x}_j.$$

CLTM 0 (linear) is denoted *Linear Transformation Model* because it belongs to the class of well-known linear transformation models (Equation (4.6)). Hence, it is equivalent to model CLTM C defined in Equation (2.14). The transformation function h is decomposed into a flexible function $h_0(Y_{\boldsymbol{x}})$ depending only on the response values $Y_{\boldsymbol{x}}$ and a part depending only on the explanatory variables. The coefficients α_j induce linear shifts of the response transformation depending on the explanatory variables $\boldsymbol{x}_j, j = 1, \dots, J$. The flexible response transformation $h_0(Y_{\boldsymbol{x}})$ is restricted to monotone functions. The transformation function results from a linear CLTM if we set $\alpha_0 = 0$, $\beta_0 = 1$ and $\beta_j = 0, j = 1, \dots, J$.

In the conditional distribution function of birth weight, these definitions result in fetus-specific means that depend linearly on the ultrasound measurements. Beyond that, the birth weights might follow some arbitrary distribution function because higher moments are modelled flexibly. The corresponding class of distribution functions is the same for all fetuses because the deviations from the normal distribution are not influenced by any ultrasound measurements.

CLTM 0: Linear Transformation Model with flexible explanatory variable effects

$$h(Y_{\boldsymbol{x}}|\boldsymbol{x}) = Y_{\boldsymbol{x}} + \tilde{h}_0(Y_{\boldsymbol{x}}) + \sum_{j=1}^{J} \alpha_j(\boldsymbol{x}) \overset{\text{Equation (4.7)}}{=} h_0(Y_{\boldsymbol{x}}) + \sum_{j=1}^{J} \alpha_j(\boldsymbol{x}).$$

CLTM 0 also represents a linear transformation model, but the influence of the explanatory variables is modelled in terms of smooth functions $\alpha_j(\boldsymbol{x}), j = 1, \ldots, J$. This results in flexible shifts of the response transformation depending on the explanatory variables. The flexible response transformation $h_0(Y_{\boldsymbol{x}})$ is again restricted to monotone functions. This transformation function results from an additive CLTM if we set $\alpha_0 = 0$, $\beta_0 = 1$ and $\beta_j = 0, j = 1, \ldots, J$.

Based on CLTM 0, fetus-specific means result that depend flexibly on the ultrasound measurements. Moreover, the birth weights may follow some arbitrary distribution, but the corresponding class of distribution functions is again the same for all fetuses. Thus, model CLTM 0 describes a very general but easy interpretable set of distributions. The explanatory variables have an additive influence only on the conditional mean and the response distribution belongs to the rich set of distributions that can be generated form the normal distribution via a monotone transformation.

CLTM 1: CLTM with linear explanatory variable effects and linear unconditional response transformation

$$h(Y_{\boldsymbol{x}}|\boldsymbol{x}) = \alpha_0 + \beta_0 \cdot Y_{\boldsymbol{x}} + Y_{\boldsymbol{x}} \cdot \sum_{j=1}^{J} \beta_j \cdot \boldsymbol{x}_j + \sum_{j=1}^{J} \alpha_j \cdot \boldsymbol{x}_j.$$

This is a linear CLTM in which $\tilde{h}_0(Y_{\boldsymbol{x}})$ is cancelled, and, therefore, the unconditional part of the response transformation is linear in $Y_{\boldsymbol{x}}$. Hence, conditional on the explanatory variables $\boldsymbol{x}$, the whole conditional transformation function $h(Y_{\boldsymbol{x}}|\boldsymbol{x})$ is linear in $Y_{\boldsymbol{x}}$. As we cancelled the deviations from linearity $\tilde{h}_0(Y_{\boldsymbol{x}})$, we assumed that the response has a normal distribution function if we additionally set the link function to $F = \Phi$ in Equation (4.2). This is due to the underlying assumption that the coefficients α_j and $\beta_j, j = 0, \ldots, J$ influence only the mean and variance of the response. These definitions result in normal distribution functions for all fetuses with fetus-specific means and variances that depend on the ultrasound measurements. Hence, model CLTM 1 is equivalent to model CLTM D defined in Equation (2.15).

CLTM 2: CLTM with linear explanatory variable effects and unconditional response transformation with monotone constraints

$$h(Y_{\boldsymbol{x}}|\boldsymbol{x}) = \underbrace{\alpha_0 + \beta_0 \cdot Y_{\boldsymbol{x}} + \tilde{h}_0(Y_{\boldsymbol{x}})}_{\text{uncond. trans. function}} + Y_{\boldsymbol{x}} \cdot \sum_{j=1}^{J} \beta_j \cdot \boldsymbol{x}_j + \sum_{j=1}^{J} \alpha_j \cdot \boldsymbol{x}_j.$$

CLTM 2 is also a linear CLTM but is more complex than CLTM 1 as the unconditional response transformation is a flexible monotone function. Hence, model CLTM 2 is equivalent to model CLTM E defined in Equation (2.16). We suggest that the distribution function of the response possibly does not belong to the class of normal distributions if we additionally set the link to $F = \Phi$. This is due to the term describing deviations from linearity $\tilde{h}_0(Y_{\boldsymbol{x}})$, which is able to affect higher moments of the distribution function of the response.

Hence, the birth weights follow some arbitrary distribution function because higher moments are modelled flexibly. Nevertheless, the corresponding class of distribution functions is again identical for all fetuses as the deviations from linearity are not influenced by any ultrasound measurements. Moreover, fetus-specific means and variances result, and the influence of the ultrasound measurements is modelled linearly.

CLTM 3: CLTM with flexible explanatory variable effects and linear unconditional response transformation

$$h(Y_{\boldsymbol{x}}|\boldsymbol{x}) = \alpha_0 + \beta_0 \cdot Y_{\boldsymbol{x}} + Y_{\boldsymbol{x}} \cdot \sum_{j=1}^{J} \beta_j(\boldsymbol{x}) + \sum_{j=1}^{J} \alpha_j(\boldsymbol{x}).$$

This model is an additive CLTM with $\tilde{h}_0(Y_{\boldsymbol{x}}) = 0$. Again, the unconditional response transformation is a linear function (compare CLTM 1), and we therefore implicitly assumed that the response follows a normal distribution. Therefore, these definitions result in normal distribution functions for all fetuses with fetus-specific means and variances. The influence of the ultrasound measurements was modelled flexibly.

CLTM 4: CLTM with flexible explanatory variable effects and unconditional response transformation with monotone constraints

$$h(Y_{\boldsymbol{x}}|\boldsymbol{x}) = \underbrace{\alpha_0 + \beta_0 \cdot Y_{\boldsymbol{x}} + \tilde{h}_0(Y_{\boldsymbol{x}})}_{\text{uncond. trans. function}} + Y_{\boldsymbol{x}} \cdot \sum_{j=1}^{J} \beta_j(\boldsymbol{x}) + \sum_{j=1}^{J} \alpha_j(\boldsymbol{x}).$$

Also this model is an additive CLTM and is the most complex CLTM considered. Comparable to CLTM 3, the influence of the explanatory variables on the linear response transformation is modelled flexibly. Additionally, the unconditional response transformation is a flexible monotone function (compare CLTM 2), in which we implicitly assumed that the response may not follow a normal distribution.

Hence, we assumed fetus-specific means and variances, whereby the influence of the ultrasound measurements was modelled flexibly. Again, birth weights for all fetuses follow some arbitrary distribution because higher moments are modelled flexibly, but the corresponding class of distribution functions is the same for all fetuses.

CTM: Conditional transformation model

$$h(Y_{\boldsymbol{x}}|\boldsymbol{x}) = \sum_{j=1}^{J} h_j(Y_{\boldsymbol{x}}|\boldsymbol{x}). \tag{4.10}$$

We define the common CTM (Hothorn et al., 2014) as our reference model because it represents a more general and more complex model class than the considered CLTMs. The transformation function $h(Y_{\boldsymbol{x}}|\boldsymbol{x})$ is decomposed additively into J partial transformation functions without any further restrictions. Thereby, we assume additivity on the scale of the transformation function, which is fundamentally different to additive mean or quantile regression, where additivity is assumed on the scale of the conditional mean or quantile function. Simulation results presented in Hothorn et al. (2014) show a better performance of CTMs compared to the parametric generalised additive models for location, scale and shape (GAMLSS) and to nonparametric kernel estimators. Since CTMs are an alternative to quantile regression models, the authors also compared the two approaches and assessed that both model classes are equally flexible. Nevertheless, CTMs have the advantages of being based on differentiable and convex proper scoring rules as risk functions that allow relatively easy optimisation algorithms to be applied, the simultaneous estimation of all quantiles in a joint model, and the dependency on only one hyperparameter (the number of boosting iterations), compared to additive quantile regresssion. Based on this CTM, we defined the model class of CLTMs and finally the special cases of CLTMs presented above.

4.3.2. Model estimation

First, we will briefly describe the model estimation in CTMs (Equation (4.10)) and then present the necessary adaptations for CLTMs. In Hothorn et al. (2014), a parametrisation of the partial transformation functions h_j, $j = 1, \dots, J$ in CTMs via basis functions is presented, and illustrates the high flexibility of the partial transformation functions in both the response variable and the explanatory variables. For example, the j-th partial transformation function is parametrised as follows:

$$h_j(Y_{\boldsymbol{x}}|\boldsymbol{x}) = \left(\mathbf{b}_j(\boldsymbol{x})^\top \otimes \mathbf{b}_0(Y_{\boldsymbol{x}})^\top\right) \boldsymbol{\gamma}_j, \tag{4.11}$$

where $\mathbf{b}_0$ is a basis along the grid of response values $Y_{\boldsymbol{x}}$, $\mathbf{b}_j$ is a basis along a grid of explanatory variables $\boldsymbol{x}$, and $\boldsymbol{\gamma}_j$ denotes the corresponding vector of basis coefficients. The two sets of basis functions are connected via a Kronecker product, thereby establishing an interaction

Model	Linear expl. variable effects		Flexible expl. variable effects		Linear uncond. transf. function	Flexible uncond. transf. function	Higher moments depend on expl. variables
	$\alpha(x)$	$\beta(x)$	$\alpha(x)$	$\beta(x)$			
CLTM 0 (linear)	×					×	
CLTM 0			×			×	
CLTM 1	×	×			×		
CLTM 2	×	×				×	
CLTM 3			×	×	×		
CLTM 4			×	×		×	
CTM			×	×			×

Table 4.1.: Overview: Relevant CLTMs (conditionally linear transformation models) and CTM (conditional transformation model).

surface between the basis for the response and the basis for the explanatory variables. The basis $\mathbf{b}_0$ defines the functional form of the response transformation (*i.e.* a linear or flexible response transformation), and the functional form of $\mathbf{b}_j$ defines how this response transformation is influenced by the explanatory variables (*i.e.* the response transformation varies linearly or flexibly with varying explanatory variables) (Hothorn et al., 2014). For example, if one chooses linear basis functions for $\mathbf{b}_0$, one gets a linear response transformation, and if one chooses B-spline basis functions for $\mathbf{b}_0$, one gets a flexible response transformation. Hence, the user is free to choose a very complex and general model framework (*e.g.*, by choosing a B-spline basis for $\mathbf{b}_0$ and $\mathbf{b}_j$) in CTMs, which often ends up in a lack of interpretability (see Subsection 4.3.1). In CTMs, one aims at obtaining an estimate for each partial transformation function h_j that is smooth in both the response and the explanatory variables, which is achieved by imposing an appropriate penalty on the Kronecker product of basis functions in Equation (4.11). For further details on parametrisation and penalty specification, see Hothorn et al. (2014).

In CTMs, model estimation is based on the minimisation of the mean integrated logarithmic score (log score)

$$\begin{aligned} LS \;=\; & -\frac{1}{N \cdot n} \sum_{i=1}^{N} \sum_{\iota=1}^{n} I(\mathrm{BW}_i \leq \upsilon_\iota) \log(\Phi(h(\upsilon_\iota | \boldsymbol{x}_i))) + \\ & \qquad I(\mathrm{BW}_i > \upsilon_\iota) \log(1 - \Phi(h(\upsilon_\iota | \boldsymbol{x}_i))), \end{aligned} \tag{4.12}$$

which is a proper scoring rule (Gneiting and Raftery, 2007; Hothorn et al., 2014). The log score is evaluated on a grid of birth weights $\upsilon_1, \ldots, \upsilon_n$ covering their range. In CTMs, the mean integrated log score is minimised in terms of a component-wise boosting algorithm (see Section 3.1 and Chapter B). As CLTMs are a special case of CTMs, we used the same approach for model estimation. All we had to adapt is the parametrisation of the partial transformation functions in Equation (4.11), which is straightforward. A derivation of the integrated log score for survival times is given in Section 5.2.1.

The choice of the functional form of $\mathbf{b}_0(Y_{\boldsymbol{x}})$ and $\mathbf{b}_j(\boldsymbol{x}), j = 1, \ldots, J$ (either linear or flexible basis functions) depends on the definition of the conditional transformation function $h(Y_{\boldsymbol{x}}|\boldsymbol{x})$. As an example, we present the parametrisation of transformation model CLTM 0 given in the previous subsection. CLTM 0 can be decomposed into the unconditional transformation function $h_0(Y_{\boldsymbol{x}})$ that depends only on the response values, and the part $\alpha(\boldsymbol{x}) = \sum_j \alpha_j(\boldsymbol{x})$ that depends only on the explanatory variables. Both parts of the transformation function are parametrised separately as special cases of Equation (4.11). First, the unconditional transformation function is parametrised via

$$h_0(Y_{\boldsymbol{x}}) = \left(\mathbf{1}_N^\top \otimes \mathbf{b}_0(Y_{\boldsymbol{x}})^\top\right) \boldsymbol{\gamma},$$

where $\mathbf{1}_N$ denotes the one-vector those length is equal to the number of observations N. As the unconditional transformation function does not depend on any explanatory variables, the basis functions for the explanatory variables $\mathbf{b}_j(\boldsymbol{x})$ are replaced by $\mathbf{1}_N$ to maintain correct dimensions. The basis functions for the response variables $\mathbf{b}_0(Y_{\boldsymbol{x}})$ are monotonic B-splines (*i.e.* T-splines, Section 3.2.2) as $h_0(Y_{\boldsymbol{x}})$ is assumed to be a flexible monotone function in the response values. Second, the function depending on the explanatory variables is parametrised by the set of basis functions

$$\alpha_j(\boldsymbol{x}) = \left(\mathbf{b}_j(\boldsymbol{x})^\top \otimes \mathbf{1}_n^\top\right) \boldsymbol{\gamma}_j,\ j = 1, \dots, J,$$

where $\mathbf{1}_n$ denotes the one-vector with length n, the number of unique υ values (a hyper parameter to the algorithm). As the functions $\alpha_j(\boldsymbol{x}), j = 1, \dots, J$, do not depend on the response variable, the corresponding basis functions $\mathbf{b}_0(Y_{\boldsymbol{x}})$ are replaced by the one-vector to maintain correct dimensions. The basis functions $\mathbf{b}_j(\boldsymbol{x}), j = 1, \dots, J$, are B-spline basis functions because the explanatory variables have a flexible influence on the mean of the transformed response in CLTM 0. The parametrisation of the other special cases of CLTMs result accordingly.

4.3.3. Computational details

All analyses were carried out in the R system for statistical computing (version 2.15.3, R Core Team, 2014). Model estimation in CLTMs and CTMs was carried out using the R add-on package **ctm** (Hothorn, 2012). To compare our proposed transformation models and established methods, we estimated a linear regression model, linear quantile regression model and additive quantile regression model. To estimate the linear regression model, we used the `lm` function in the **stats** package and fitted the linear quantile regression model using the `rq` function of the **quantreg** package (Koenker, 2012). We used component-wise boosting for the estimation of the additive quantile regression model (Fenske et al., 2011) in the **mboost** package (Hothorn et al., 2013). A tutorial R example `ex_fetus_CLTM.Rnw` including the code for estimating the proposed regression and transformation models, the calculation of prediction intervals for the birth weights, and the generation of Figure 1 is publicly available in the **ctm** package from the R-forge repository (`https://r-forge.r-project.org/projects/ctm`).

4.3.4. Evaluation of fetus-specific prediction formulas for birth weight

As we are interested in reliable prediction intervals for birth weights (see Section 4.1), we calculated fetus-specific prediction intervals based on Equation (4.1) with a coverage probability of 80%. A further goal was to identify the C(L)TM that described the Perinatal Database Erlangen best among the proposed C(L)TMs in Section 4.3.1. We considered certain aspects of model misspecification.

For the construction of prediction intervals, we considered the conditional median and the conditional $^{\alpha}/_{2}$ quantile and $1 - {^{\alpha}/_{2}}$ quantile representing the point prediction for the birth weight and the boundaries of the fetus-specific prediction intervals in Equation (4.1). Therefore, we used the well-known relationship between the conditional distribution function and the conditional quantile function to extract the relevant quantiles:

$$q_\tau(\boldsymbol{x}) = F^{-1}_{\mathrm{BW}|\mathbf{X}=\boldsymbol{x}}(\tau),$$

where $\tau = \{{}^{\alpha}\!/\!{}_{2}, 0.5, 1 - {}^{\alpha}\!/\!{}_{2}\}$ denotes the quantiles of interest, and $F_{\mathrm{BW}|\mathbf{X}=\boldsymbol{x}}$ is defined in Equation (4.2) (Mayr et al., 2012).

In the analysis of the Perinatal Database Erlangen, we used ten regression or transformation models to estimate the median birth weight and the associated interval borders. The transformation models used encompass a standard CTM and the six CLTMs [CLTM 0 (linear) and CLTM 0 – CLTM 4] presented in Section 4.3.1. For comparison, we also considered a linear regression model (LM), which served as a standard procedure in the past, a linear quantile regression model, and an additive quantile regression model (LQR and AQR).

A common strategy to check the adequacy of prediction intervals is to check their coverage probability. When we defined prediction intervals in Subsection 4.2.3, we stated that a correctly specified prediction interval $\mathrm{PI}_{1-\alpha}(\boldsymbol{x})$ for a new set of ultrasound parameters $\boldsymbol{x}$ covers a new observation BW with high probability $1 - \alpha$. The correct measure to evaluate prediction intervals adequately is the conditional coverage (Mayr et al., 2012). Therefore, we checked whether for any particular combination of ultrasound measurements $\boldsymbol{x}$ about $(1 - \alpha) \cdot 100\%$ of the corresponding observations $(\mathrm{BW}_1, \boldsymbol{x}), \ldots, (\mathrm{BW}_M, \boldsymbol{x})$ were covered by the prediction interval $\mathrm{PI}(\boldsymbol{x})$:

$$\hat{\pi}|\boldsymbol{x} = \hat{\mathbb{E}}(\mathrm{BW} \in \mathrm{PI}(\boldsymbol{x})|\mathbf{X} = \boldsymbol{x}) = \frac{1}{M}\sum_{i=1}^{M} I\left\{\mathrm{BW}_i \in \mathrm{PI}(\boldsymbol{x})\right\}, \tag{4.13}$$

where I denotes the indicator function. The conditional coverage reflects what we really expect from a prediction interval because the prediction interval for a specific combination of ultrasound parameters should cover the birth weights of 80% of the fetuses with exactly the same ultrasound measurements (Mayr et al., 2012).

In practice, the evaluation of the conditional coverage of prediction intervals is impossible because we usually only have one observation for each combination of ultrasound parameters $\boldsymbol{x}$ and more are needed with exactly the same combination of ultrasound measurements (Equation (4.13)). Especially in a regression setting with continuous explanatory variables, multiple response values for each combination of explanatory variables are unlikely to occur. Therefore, we calculated the conditional coverage of our prediction intervals using binned observations:

1. We used the ultrasound parameters AC and FL to divide the fetuses in the database into categories because these two parameters are essential for the prediction of birth weights (*e.g.*, see Sabbagha et al., 1989; Faschingbauer et al., 2012; Hoopmann et al., 2010; Scioscia et al., 2008). AC and FL were divided

quantile-based into categories, resulting in five AC categories measured in mm ($\mathbf{1}: (175, 316]$; $\mathbf{2}: (316, 331]$; $\mathbf{3}: (331, 343]$; $\mathbf{4}: (343, 357]$; $\mathbf{5}: (357, 428]$) and five FL categories measured in mm ($\mathbf{1}: (31.1, 69.6]$; $\mathbf{2}: (69.6, 71.7]$; $\mathbf{3}: (71.7, 73.4]$; $\mathbf{4}: (73.4, 75.4]$; $\mathbf{5}: (75.4, 86.6]$).

2. When we combined the five AC and five FL categories, we got 25 categories of fetuses, which results in good sample sizes of at least 102 observations for all groups. The distribution of the birth weights in the respective categories are displayed in Figure A.3.

3. To assess the conditional coverage, we generated a training data set by randomly choosing 90% of the fetuses in each of the 25 categories and generated a validation data set by choosing the remaining fetuses. We then estimated CLTM 0 (linear) - CLTM 4, CTM, LM, LQR, and AQR for the training data, and predicted the birth weights for the validation data set for each of the models. We assessed the conditional coverage (Equation (4.13)) for each of the regression and transformation models in each of the 25 categories.

In addition to the conditional coverage of the prediction intervals, we also checked their average interval lengths.

To identify the C(L)TM that described the Perinatal Database Erlangen best, we compared the performance among all CLTMs to the performance of the CTM and the LM. We fitted the models on a training data set and evaluated their predictive ability on an evaluation data set. Twenty-five training and evaluation data sets were generated by choosing randomly 50% of the original observations in each AC–FL category. The predictive ability was measured in terms of the log score given in Equation 4.12, which was used to evaluate the conditional distribution function for the whole evaluation data set and for each AC-FL category separately. As the complexities of the C(L)TMs differed, this procedure could also be used to reveal model misspecifications. We were able to detect missing covariate effects on the variance (*e.g.*, CLTM 0 against all other C(L)TMs), missing flexibility of the covariate effects on the mean or the variance (*e.g.*, CLTM 2 against CLTM 4), and missing flexibility of the response transformation (*e.g.*, CLTM 1 against CLTM 2). If even higher moments of the conditional distribution function were affected by the explanatory variables, could be checked by comparing all CLTMs to the CTM, and by comparing all CLTMs to the LM if the assumption of a normal distribution with constant variance works for the database. The out-of-sample log score cannot be calculated for the quantile regression models because quantile crossing makes the inversion of the quantile function into a distribution function impossible.

4.4. Results

4.4.1. Estimated transformation and regression models

All ultrasound parameters were included as main effects in the model equations of the regression and transformation models. One exception was the interaction between AC and FL, which has been important in many earlier prediction formulas for birth weight (*e.g.*, in Hadlock et al., 1985). Therefore, we additionally included this interaction in models CLTM 0 (linear), CLTM 1, CLTM 2, LM, LQR and AQR; we did not include this interaction in models CLTM 0, CLTM 3, CLTM 4 and CTM because the model estimation became too complex.

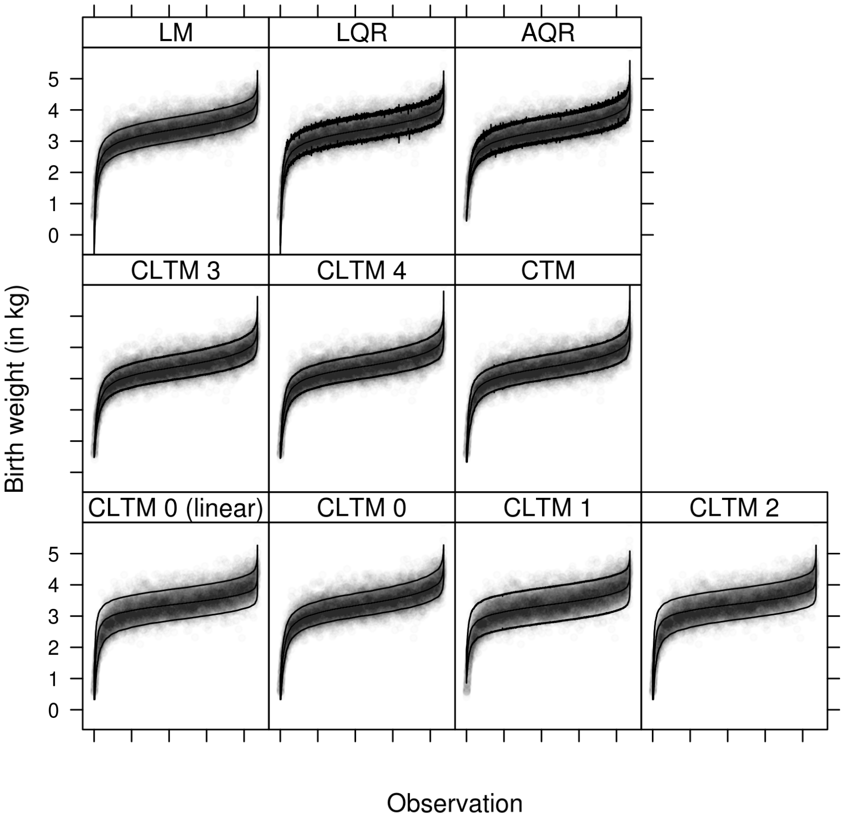


Figure 4.1.: Birth weight prediction. Observed birth weights of 8, 712 newborns (dots) ordered with respect to the predicted conditional mean (LM only) or median birth weight (central black line). The lower and upper black lines display estimated 10% and 90% quantiles of birth weights, respectively. The areas in between represent fetus-specific 80% prediction intervals. Each subplot shows the results for one of the regression or conditional transformation models: LM, linear model; LQR, linear quantile regression; AQR, additive quantile regression; CLTMs (CLTM 0 – CLTM 4), conditionally linear transformation models; and CTM, conditional transformation model.

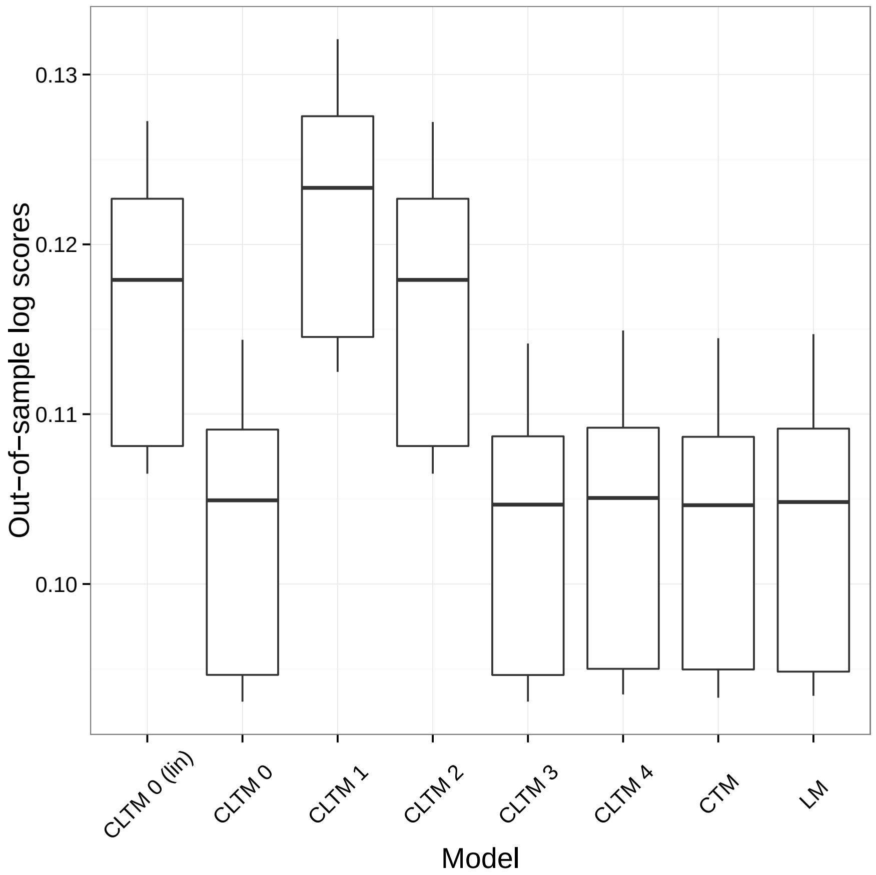


Figure 4.2.: Out-of-sample log scores for CLTM 0 – CLTM 4, LM, and CTM based on 25 randomly chosen evaluation data sets consisting of 4, 355 observations.

The estimates of the birth weights based on the prenatal ultrasound parameters are displayed in Figure 4.1. In model LM, symmetric intervals around the estimated conditional mean with equal interval lengths for all fetuses resulted, and possible heteroscedasticity, kurtosis, and skewness was ignored. Despite these restrictive assumptions, model LM provided satisfying and narrow intervals. We concluded that deviations from normality were small and no strong heteroscedasticity occurred. Nevertheless, we pursued further model improvements.

The quantile regression approaches (LQR and AQR) also provided satisfying results associated with narrow intervals. The wiggly estimates for the interval borders were due to the separate estimation of the quantiles. In contrast, smooth interval borders resulted for C(L)TMs because all quantiles were estimated simultaneously.

In CLTM 0 (linear), the influence of the ultrasound parameters on the conditional mean was modelled linearly, comparable to model LM. Owing to the unconditional transformation function, also a possible skewness and kurtosis of the distribution of the birth weights can be modelled. This led to wider intervals for CLTM 0 (linear) compared to LM, especially for extreme birth weights. In model CLTM 0, the influence of the ultrasound measurements on the conditional mean was modelled flexibly, and thus, the corresponding fetus-specific intervals were narrower than with CLTM 0 (linear).

In general, a flexible inclusion of the ultrasound parameters seems advisable because the intervals with models CLTM 0, CLTM 3 and CLTM 4 were narrower than with CLTM 1 and CLTM 2. Besides, in CLTM 1 – CLTM 4, the ultrasound parameters may influence the conditional mean and conditional variance. Hence, these models accounted for possible heteroscedasticity induced by the ultrasound measurements.

An additional slight improvement was gained by estimating the unconditional transformation function in terms of a flexible monotone function and thus accounting for possible kurtosis and skewness. This can be observed by direct comparison of CLTM 1 and CLTM 2 and of CLTM 3 and CLTM 4. Nevertheless, deviations from normality seemed to be small because the associated improvements were minor.

We were also interested in identifying the C(L)TM that described the Perinatal Database Erlangen best. We calculated the out-of-sample log scores based on 25 evaluation data sets for the proposed C(L)TMs and the LM to evaluate the estimated conditional distribution functions for new observations for the whole evaluation data set (Figure 4.2) and for each AC–FL category separately (Figure A.1 and Figure A.2). The results were in accordance with those in Figure 4.1: the out-of-sample log scores of CLTM 0, CLTM 3, CLTM 4, CTM and LM were similar, whereas those of CLTM 0 (linear), CLTM 1 and CLTM 2 were clearly higher. Hence, the inclusion of flexible covariate effects clearly improves the estimated conditional distribution functions. On the other hand, consideration of heteroscedasticity, deviations from the normality assumption, and higher moments depending on explanatory variables were of minor importance, which was also supported by the good performance of the LM.

To further illustrate important characteristics of CLTMs, we more closely examined CLTM 4, which is the most flexible among all considered CLTMs. The influence of the ultrasound measurements on the conditional mean and conditional variance was modelled flexibly, and the unconditional response transformation was modelled as a flexible monotone function. We assumed that the response values most likely do not follow a normal distribution, as the following results indicated.

Low birth weights did not exactly follow a normal distribution, *i.e.* the resulting estimated birth weight transformation function showed deviations from a linear function for low birth weights (see Equation (4.7)), whereas medium and high birth weights followed a normal distribution (Figure 4.3). Therefore, low birth weights needed to be transformed.

This conclusion can be observed clearly in normal quantile-quantile plots for original and transformed birth weights resulting from model CLTM 4 (Figure 4.4). Low original birth weights deviated from the normal distribution, but low transformed birth weights approximately followed a normal distribution. A scatterplot showing the relationship between

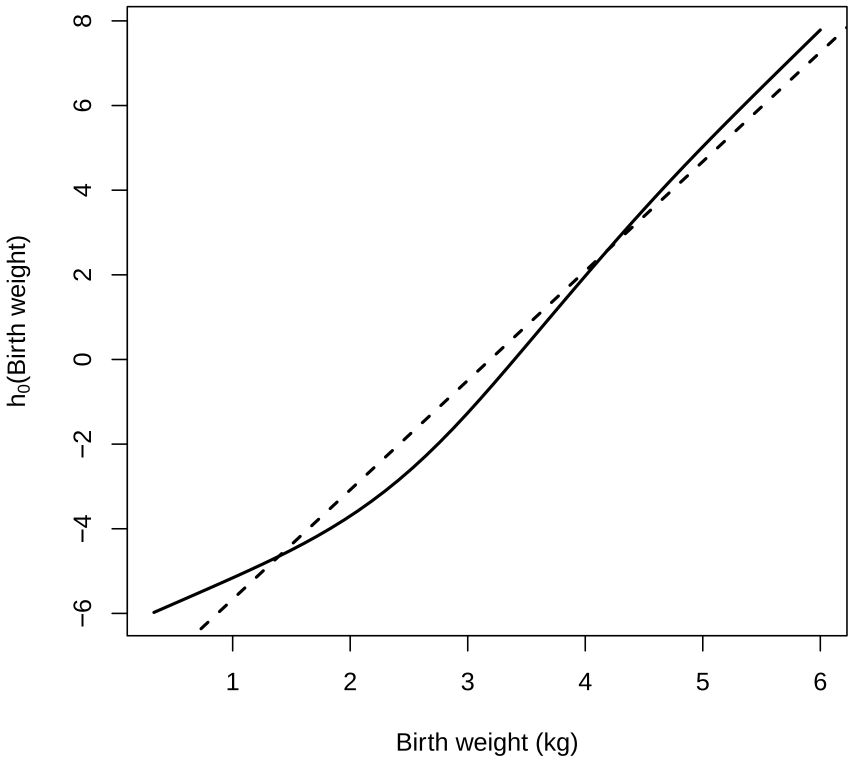
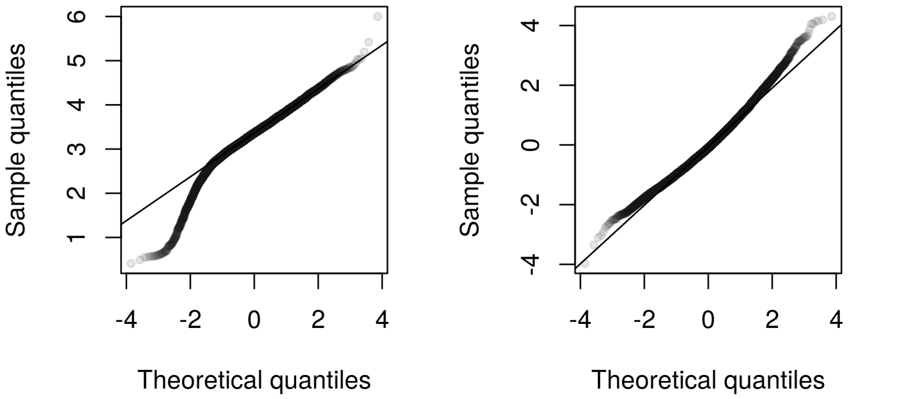


Figure 4.3.: Estimated monotone birth weight transformation function resulting from model CLTM 4. The dashed line symbolises the linear relationship between the birth weights and their monotone transformation.

original and transformed birth weights (Figure 4.5) revealed similar results. Medium and high birth weights scattered unsystematically around some linear function, whereas low birth weights deviated, which indicated that a non-linear transformation took place. Moreover, a kernel density plot (Figure 4.5) shows that the estimated density of the transformed birth weights is in good accordance with the corresponding density of the normal distribution.

These results together indicated that those regression models that allow deviations from the normal distribution assumption are more reliable when original data do not entirely follow a normal distribution.

We stressed that the main advantage of CLTMs over CTMs is the improved interpretability of the estimated effects of ultrasound measurements on moments of the distribution function of birth weights. The estimated effects of ultrasound parameters for model CLTM 4 (Figure 4.6) can be interpreted according to Equation (4.4). For almost all ultrasound



a: Normal Q-Q plot of birth weight

b: Normal Q-Q plot of transformed birth weight

Figure 4.4.: Normal Q-Q Plot of the original and the transformed birth weights resulting from model CLTM 4.

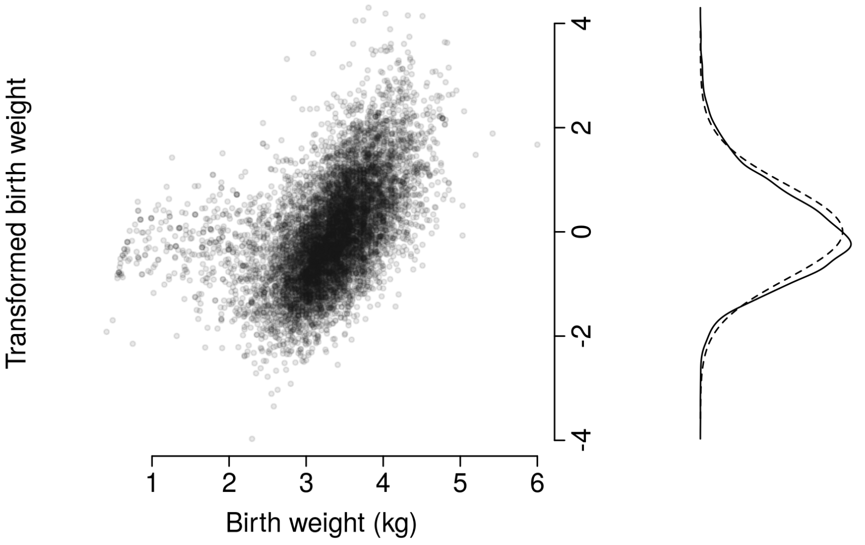


Figure 4.5.: Left: Scatterplot of the original birth weights vs. the transformed birth weights resulting from model CLTM 4. Right: Kernel density estimation of the transformed birth weights (solid line) and the corresponding normal density (dashed line).

parameters, estimated non-linear functions α and β resulted, which suggested that the ultrasound parameters influence both the conditional mean and conditional variance.

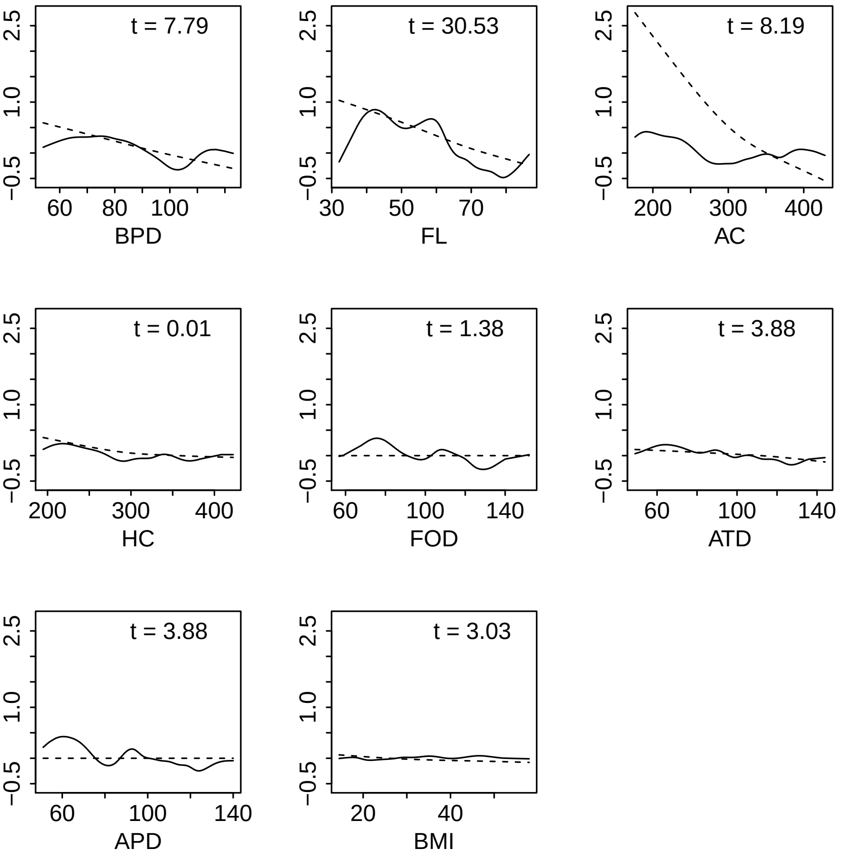


Figure 4.6.: Estimated effects of ultrasound parameters on the conditional mean and conditional variance of transformed birth weight. Solid lines represent estimated functions $\hat{\alpha}$(ultrasound parameter), and dashed lines represent estimated functions $\hat{\beta}$(ultrasound parameter). The corresponding values of t-statistics belong to the coefficients of the ordinary linear model LM. BPD, biparietal diameter; FL, femur length; AC, abdominal circumference; HC, head circumference; FOD, fronto-occipital diameter; ATD, abdominal transverse diameter; APD, anterior-posterior abdominal diameter; BMI, mother's body mass index.

4.4.2. Assessing the accuracy of the prediction intervals

We assessed the accuracy and adequacy of the (fetus-specific) prediction intervals by calculating the conditional coverage and average interval length as quality criteria.

The conditional coverage of the prediction intervals for the birth weights (Figures 4.7 and 4.8; Tables A.1 and A.2) is a measure to check the adequacy and correctness of estimated prediction intervals. We were interested in how often the postulated coverage probability of 80% was violated in the 25 AC–FL categories (defined in Subsection 4.3.4) for the ten regression models. Moreover, the accuracy of the prediction intervals can be measured by the average interval lengths given in Table 4.2.

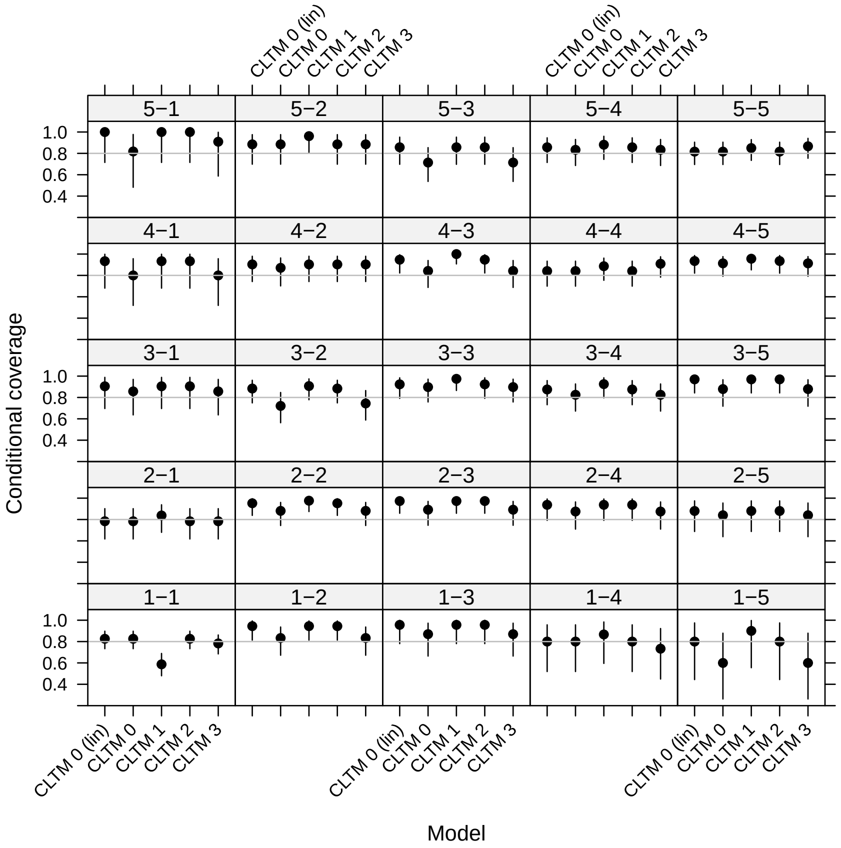


Figure 4.7.: Conditional coverage of the prediction intervals for fetuses of the 25 AC–FL categories. Points refer to the point estimates of the conditional coverage, and error bars display corresponding Clopper-Pearson confidence intervals. Grey reference lines symbolise the postulated 80% confidence level. Model estimation was carried out with CLTM 0 (linear), CLTM 0, CLTM 1, CLTM 2, and CLTM 3.

Table 4.2.: Average prediction interval length. Estimation is based on models CLTM 0 (linear), CLTM 0 – CLTM 4, CTM, LM, LQR, and AQR.

Model	Average interval length
CLTM 0 (linear)	1.042
CLTM 0	0.785
CLTM 1	1.132
CLTM 2	1.042
CLTM 3	0.790
CLTM 4	0.776
CTM	0.807
LM	0.777
LQR	0.764
AQR	0.755

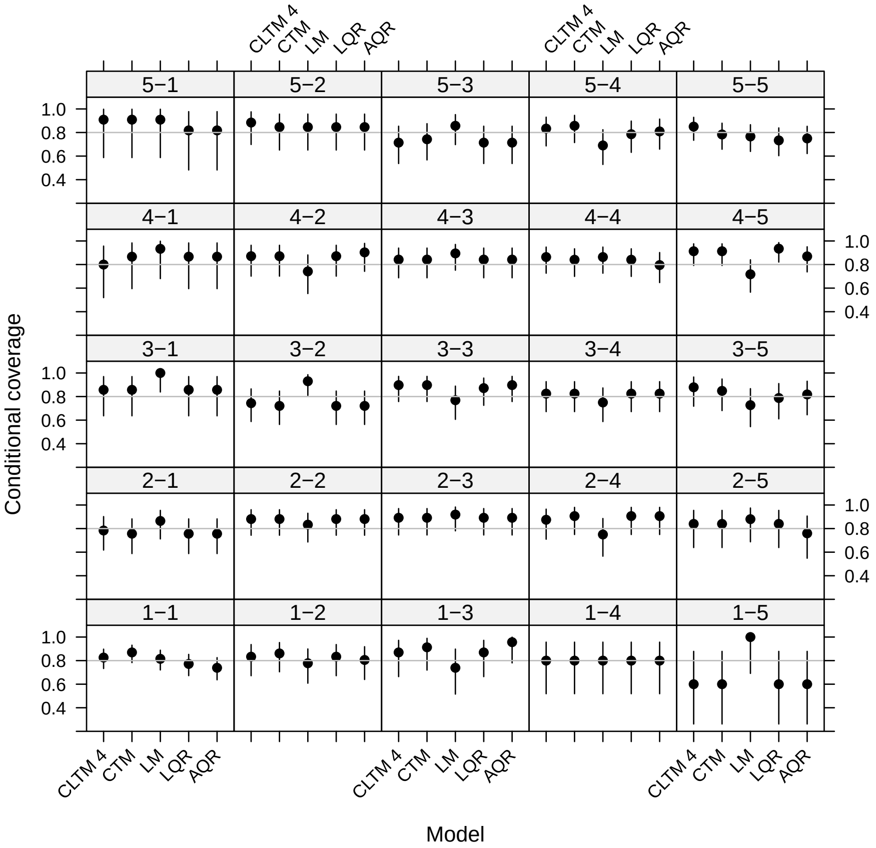


Figure 4.8.: Conditional coverage of the prediction intervals for fetuses of the 25 AC–FL categories. Points refer to the point estimates of the conditional coverage, and error bars display corresponding Clopper-Pearson confidence intervals. Grey reference lines symbolise the postulated 80% confidence level. Model estimation was carried out with CLTM 4, CTM, LM, LQR, and AQR.

The conditional coverage of all ten models was satisfying. The postulated coverage probability of 80% was not significantly violated by any of the suggested models in any of the categories. The only exception was model CLTM 1 in category 1-1 (Figure 4.7). The length of the corresponding error bars was mainly determined by the number of fetuses used for estimation. Hence, the length of the error bars was especially high in the categories 5–1, 4–1, 1–5 and 1–4.

The smallest associated average interval lengths were found for CLTM 3, CLTM 0, LM, CLTM 4, LQR and AQR (Table 4.2). Hence, regarding the accuracy of prediction intervals, our new model class of CLTMs can compete with linear regression models and quantile regression approaches.

4.5. Discussion

Although the accurate prediction of birth weight is one of the most important issues in gynaecology, traditional prediction formulas focus on point predictions and an easy-to-interpret, correct measure of quantifying prediction uncertainty is lacking. We therefore aimed at finding a new model-based strategy to predict birth weights based on prenatal ultrasound parameters, accompanied by some measure of prediction uncertainty. We introduced conditionally linear transformation models (CLTMs) – a new model class that not only results in point estimates for the median birth weight but also provides a measure of uncertainty in terms of prediction intervals.

Birth weights at the extremes have been especially over- or under-estimated by prediction formulas presented earlier (Scioscia et al., 2008). This could be due to the use of linear regression models for estimation, which are not able to deal with possible heteroscedasticity, kurtosis or skewness of the response distribution, and are accordingly inadequate in such situations. The standard approach around this problem is the use of quantile regression approaches as no distributional assumptions are made, but one often has to deal with the problem of quantile crossing instead (Dette and Volgushev, 2008).

In our novel approach of estimating CLTMs, we modelled the conditional distribution function of birth weight based on ultrasound measurements. Hence, all quantiles were estimated simultaneously, and problems such as quantile crossing were avoided. Koenker (2005) already suggested the direct estimation of the conditional distribution function via transformation models as an alternative to quantile regression models. The flexibility of the influence of the ultrasound parameters on the quantiles in CLTMs is similar to the flexible influence in quantile regression, as the ultrasound measurement effects may vary for different values of the conditional distribution function in CLTMs. The borders of the fetus-specific prediction intervals arised directly from the corresponding quantile function. In contrast to linear regression models, the fetus-specific prediction intervals showed individual interval lengths based on the ultrasound measurements and are therefore a useful measure for individual prediction accuracy. Moreover, the variance may depend on explanatory variables, and CLTMs account for possible heteroscedasticity. In addition, CLTMs can deal with skewed distributions as higher moments of the distribution of the response (*e.g.*, kurtosis and skewness) can be modelled flexibly in terms of the unconditional monotone transformation function. Hence, using CLTMs instead of linear regression models is advantageous in numerous situations, and especially in our application of predicting birth weights.

From a conceptual point of view, arose weight estimation is fundamentally different from the construction of reference growth charts of child height and weight (Cole, 1988). Growth curves are usually designed as screening tools for disease after birth (and also as reference standards for group health and economic status, *e.g.*, Wei et al. (2006)), whereas prediction of birth weight is designed to estimate the risk of neonatal mortality and morbidity *before* delivery. Consequently, although similar statistical methodology may be used for both tasks, the CLTM approach proposed here specifically addresses the problem of birth weight prediction but not the construction of reference growth curves.

Our results suggested that the best-performing CLTM variant is able to compete with quantile regression and linear regression approaches in terms of conditional coverage and average length of the prediction intervals. Although the differences to alternative methods were small, the estimation of C(L)TMs is advisable because of the aforementioned advantages of accounting for possible heteroscedasticity, kurtosis and skewness. The distribution of the birth weights showed deviations from a normal distribution (Figure 4.4), but the deviations were kept within certain limits. Therefore, the linear regression model would not be the worst choice in this application, and we would expect larger differences in favour of C(L)TMs for response variables showing more extreme deviations from normality. Consequently, our results show that prediction intervals for birth weights can be derived from a relatively easy and stable model, since the medium and high birth weights follow a normal distribution and only small birth weights show deviations from normality (Figure 4.3 and Figure 4.4). This conclusion is also underlined by the good performance of model CLTM 0 (Figure 4.2). It would have been very hard to derive such insights into the conditional distribution of birth weights from alternative models, for example additive quantile regression models. In general, the remarkably good performance of CTMs compared to alternative modelling strategies has already been investigated in simulation studies and numerous applications (Hothorn et al., 2013, 2014).

Interpretability in CLTMs is different than in linear and quantile regression models. In linear and quantile regression models, the influence of explanatory variables can be interpreted as direct effects on the conditional mean or conditional quantile, respectively. In CLTMs, in contrast, the explanatory variables influence the mean and variance of the transformed response non-linearly (compare Equation (4.4)). Nevertheless, the effects of the explanatory variables are interpretable in CLTMs, which is a main advantage over the more complex model class of CTMs. Moreover, we were primarily interested in predicting birth weights accurately, and this is accompanied by correct and precise prediction intervals.

5. Boosting conditional transformation models for survivor function estimation

The content of this chapter is based on Möst and Hothorn (2015).

Besides for the determination of prediction intervals, the characteristics of conditional transformation models are advantageous for the estimation of conditional survivor functions. In survival analysis, the estimation of patient-specific survivor functions that are conditional on a set of patient characteristics is of special interest. For example, Crowther and Lambert (2014) state that the understanding of individual patient risk profiles is of special importance in personalised medicine. In general, knowledge of the conditional survival probabilities of a patient at all relevant time points allows better assessment of the patient's risk than summary statistics, such as median survival time. Nevertheless, standard methods for analysing survival data seldom estimate the survivor function directly. Therefore, we propose the application of conditional transformation models (CTMs) for the estimation of the conditional distribution function of survival time given a set of patient characteristics. We use the inverse probability of censoring weighting approach to account for right-censored observations. Our proposed modelling approach allows the prediction of patient-specific survivor functions. In addition, CTMs constitute a flexible model class that is able to deal with proportional as well as non-proportional hazards. The well-known Cox model is included in the class of conditional transformation models as a special case. We investigate the performance of CTMs in survival data analysis in a simulation that includes proportional and non-proportional hazards settings and different scenarios of explanatory variables. Furthermore, we re-analyse the survival times of patients suffering from chronic myelogenous leukaemia and study the impact of the proportional hazards assumption on previously published results.

5.1. Introduction

The estimation of a patient's individual survival probabilities over time is a key aspect of survival analysis. Technically, we are interested in estimating the conditional survivor function, *i.e.* the probability of surviving up to a specific time point t, conditional on a set of patient-specific explanatory variables. However, common regression models for censored data seldom focus on the direct estimation of the conditional survivor function. Instead, the models concentrate either on the estimation of hazard functions or on summary statistics. In the omnipresent Cox proportional hazards model (Cox, 1972), the conditional hazard function is estimated by cleverly treating the baseline hazard function as a nuisance parameter. Only in a second step can the corresponding conditional survivor functions be derived from this model, for example by using the Breslow estimator (*e.g.*, Andersen et al., 1983). Hence, if one is interested in the conditional survival probabilities, methods for the direct estimation of the conditional survivor function are required.

Moreover, assumptions associated with common modelling strategies for survival data are restrictive. For example, the Cox model is based on the assumption of proportional hazards, the proportional odds model assumes constant odds ratios over time, and in the parametric accelerated failure time model, log-transformed responses imply survival times that are, *e.g.*, log-normal distributed or log-logistic distributed. Although remedies are available, such as stratified Cox models or time-varying effects (Sargent, 1997; Xu and O'Quigley, 2000; Scheike and Martinussen, 2004; Tian et al., 2005), and although model diagnostics (*e.g.*, based on Schoenfeld residuals or formal misspecification tests; Schoenfeld, 1982; Ng'Andu, 1997) and particularly tests for the proportional hazards assumption (*e.g.*, based on cumulative sums of martingale-based residuals or weighted residuals; Lin et al., 1993; Grambsch and Therneau, 1994) help to detect unrealistic assumptions, models making less strong assumptions would be widely welcomed.

We suggest estimating the conditional distribution function of the survival time T given a set of patient characteristics $\boldsymbol{x}$ directly in terms of conditional transformation models (CTMs). CTMs have been presented recently in Hothorn et al. (2014) and allow the direct and semiparametric estimation of the conditional distribution function $\mathbb{P}(T \leq t | \boldsymbol{X} = \boldsymbol{x})$ under rather weak assumptions. The general model class includes both the proportional odds model and the proportional hazards model as special cases. Nevertheless, the strict assumptions of proportional hazards or proportional odds are relaxed in CTMs. This is achieved by including interaction terms between the survival time and the explanatory variables (see also Section 2.1.2). For example, the CTM framework allows for varying explanatory variable effects on the hazard function and hence is able to estimate non-proportional hazards as well. However, this advantage comes at the price of a more complex model, which is not easily communicated by simple parameter estimates or even p-values. Graphic approaches are needed to interpret the model, but we can always fall back on the classical approach when the more flexible model suggests that it is safe to assume proportional hazards. P-values or confidence intervals cannot be obtained based on large sample theory but can be simulated using bootstrap approaches instead.

Transformation models play an important role in survival analysis. The one-to-one correspondences between the proportional hazards and proportional odds model to linear transformation models has already been established in Doksum and Gasko (1990) and Cheng et al. (1995). Cheng et al. (1997) extended the model class to the prediction of survival probabilities. Chen et al. (2002) introduced a unified estimation procedure for the analysis of censored data using linear transformation models, and Zeng and Lin (2006) proposed a class of semiparametric transformation models to characterise the effects of possibly time-varying covariates on the intensity functions of counting processes. For the estimation of the crude failure probabilities of a competing risk, conditional on explanatory variables, Fine (2001) proposed a semiparametric transformation model. These approaches are based on generalised estimating equations. Our approach uses component-wise gradient boosting methodology for model fitting. This approach has the advantage that it incorporates variable selection and shrinkage of coefficient estimates into the model fitting process. These regularisation techniques for regression models are necessary for the estimation of survival probabilities because patient characteristics are often highly correlated. Hence, prediction accuracy for the survival probabilities can usually be improved if only a subset of the available patient characteristics is incorporated into the prediction formula. Owing to the component-wise fitting procedure, the algorithm can deal with high-dimensional data. Van der Vaart and van der Laan (2006) and Lee et al. (2011) also considered variable selection in high-dimensional survival data. Lu and Li (2008) previously derived a component-wise boosting algorithm for the analysis of survival data in terms of non-linear transformation models.

Fully nonparametric estimation of the conditional survivor function has also been considered in the past. Making no assumptions about the form of the survivor function can be advantageous over parametric or semiparametric approaches as the underlying assumptions may be violated. Furthermore, nonparametric approaches can be used to check whether one of the more restrictive parametric or semiparametric submodels provides a good fit to the data. The well-known product limit estimator introduced by Kaplan and Meier (1958) enables nonparametric estimation of the unconditional survivor function. Dabrowska (1987), Dabrowska (1989), González Manteiga and Cadarso-Suarez (1994) and Iglesias Pérez and González Manteiga (1999) present generalisations of the product limit estimator by introducing kernel-based weights to estimate the *conditional* survivor function nonparametrically. In the light of counting process theory, McKeague and Utikal (1990) propose a general counting process regression model for estimating conditional survivor functions, and Li and Doss (1995) propose a class of estimators for the conditional survivor function based on a fully nonparametric model. The usage of local linear estimators for the conditional survivor function is suggested in Spierdijk (2008).

In contrast to kernel-based methods, tree-based approaches, and especially random forests can be used to estimate conditional distribution functions precisely without relying on strict model assumptions. For right-censored data, Hothorn et al. (2004) introduced a forest aggregation scheme that produces estimates of the conditional survivor function. The same scheme was used later by Meinshausen (2006) for uncensored observations; an alternative forest variant (random survival forests) was introduced by Ishwaran et al. (2008). Conditional inference forests (Strobl et al., 2007), based on an aggregation of conditional inference trees (Hothorn et al., 2006), use the aggregation scheme introduced by Hothorn et al. (2004) and have been shown to perform akin to other forest variants for right-censored data (Mogensen et al., 2012), and were used as a completely nonparametric competitor for conditional transformation models in our study here.

Another useful alternative to the Cox model or to linear transformation models is censored quantile regression (*e.g.*, Powell, 1986; Chernozhukov and Hong, 2002; Honoré et al., 2002; Portnoy, 2003; Peng and Huang, 2008; Wang and Wang, 2009; Wey et al., 2014). With this approach, the conditional quantiles of the survival times are modelled in terms of regression models. In contrast to our proposed CTM approach, not all conditional quantiles of the survival times are modelled simultaneously but are instead modelled separately. Hence, quantile crossing (Dette and Volgushev, 2008) is a potential problem of this procedure.

In order to illustrate CTMs for survival data, we checked the validity of the proportional hazards assumption in a re-analysis of a randomised clinical trial comparing busulfan, hydroxyurea and interferon-α treatment of chronic myelogenous leukaemia. This trial has been analysed earlier using a Cox model (Clayton and Cuzick, 1985; Aalen, 1988; McGilchrist and Aisbett, 1991; Vaida and Xu, 2000). As the proportional hazards assumption is questionable for the different treatment groups, we re-analysed the data set using the CTM approach and allowed for non-proportional effects of the patient characteristics over time.

5.2. Conditional transformation models for survival data

In the following, T denotes a positive random variable describing the time from a well-defined starting point to an event of interest, *e.g.*, death or recurrence of a disease. We consider N patients with survival times T_i, $i = 1, \dots, N$, and a vector of patient characteristics $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})$. As we do not assume that all patients experience the event of interest by the end of the study period and as some patients quit the study early, the event times sometimes are not actual event times but rather right censored. The *observed* right-censored event times $\tilde{T}_i$ are defined by $\tilde{T}_i = \min(T_i, C_i)$, $i = 1, \dots, N$, where C_i denotes the time under observation or censoring time. Furthermore, the event indicator $\delta_i = I(T_i \leq C_i)$ is 1 for observed event times and 0 for right-censored event times. A common assumption is that the survival time T and the vector of explanatory variables $\boldsymbol{X}$ are independent of the censoring time C.

The conditional survivor function S is defined as the conditional probability of being event-free up to some time point t in terms of the conditional distribution function of the survival times given the explanatory variables $\boldsymbol{x}$:

$$S(t|\boldsymbol{X} = \boldsymbol{x}) = \mathbb{P}(T > t|\boldsymbol{X} = \boldsymbol{x}) = 1 - \mathbb{P}(T \leq t|\boldsymbol{X} = \boldsymbol{x}). \tag{5.1}$$

When using CTMs, we aim at estimating the conditional distribution function of the survival times via

$$\mathbb{P}(T \leq t|\boldsymbol{X} = \boldsymbol{x}) = F(h(t|\boldsymbol{x})), \tag{5.2}$$

and the conditional survivor function can be calculated by the relationship given in Equation (5.1). Thereby, the conditional distribution function is modelled in terms of the monotone transformation function $h : \mathbb{R} \rightarrow \mathbb{R}$, which depends on the patient characteristics $\boldsymbol{x}$. A short introduction to CTMs can be found in Section 1.1.

To embed the well-known class of linear transformation models (Cheng et al., 1995) into CTMs exemplarily, we reconsider the formulation of the proportional hazards (PH) model in terms of a linear transformation model given in Doksum and Gasko (1990) (see also Section 2.1.2). The conditional distribution function of the survival times resulting from the Cox model can be written as

$$\mathbb{P}(T \leq t|\boldsymbol{X} = \boldsymbol{x}) = \mathcal{M}(h_T(t) + \boldsymbol{x}^\top \boldsymbol{\beta}), \tag{5.3}$$

where $\mathcal{M}$ denotes the distribution function of the minimum-extreme value distribution, and the transformation of the survival times $h_T(t)$ equals the logarithm of the cumulative baseline hazard. In linear transformation models, the conditional transformation function h is decomposed into a part depending only on the survival times $h_T(t)$ and a part depending only on the explanatory variables $\boldsymbol{x}^\top \boldsymbol{\beta}$. This strict decomposition results in the PH assumption.

In CTMs, the PH assumption is relaxed by allowing for interactions between the survival times and the explanatory variables in terms of the conditional transformation function $h(t|\boldsymbol{x})$. Furthermore, we assume additivity on the scale of the transformation function and decompose h into J partial transformation functions, in which each $h_j : \mathbb{R} \rightarrow \mathbb{R}$ is conditional on $\boldsymbol{x}$ (Equation (1.2)):

$$\mathbb{P}(T \leq t|\boldsymbol{X} = \boldsymbol{x}) = F(h(t|\boldsymbol{x})) = F\left(\sum_{j=1}^{J} h_j(t|\boldsymbol{x})\right). \tag{5.4}$$

In analogy to the representation of the Cox model in Equation (5.3), we choose $F = \mathcal{M}$ for the link function. In this way, we operate on the same scale of distribution functions in the CTM and the Cox model, and hence estimations from the two approaches are comparable. The CTM given in Equation (5.4) can be understood as a generalisation of the PH model to more flexible non-proportional hazard functions if F is the minimum-extreme value distribution function.

As all interaction terms between the survival time and the explanatory variables are avoided in the Cox model (Equation (5.3)), the effects of the explanatory variables are estimated to be constant and are not allowed to vary over time. This assumption is relaxed in the more flexible model class of CTMs. Interaction terms between the survival time and the explanatory variables are established in terms of the partial transformation functions h_j that depend on the survival time *and* on the explanatory variables simultaneously (Equation (5.4)). Hence, the effects of the explanatory variables are allowed to vary over time, which usually results in non-proportional hazards. We not only estimate one single param-

eter for each explanatory variable as is done in the Cox model. Instead, separate partial transformation functions are defined for each explanatory variable, whereby a smooth parameter function over time is estimated for each group of a categorical explanatory variable. For continuous explanatory variables, a smooth parameter surface is estimated that depends on the survival time and on the continuous explanatory variable.
In comparison to alternative nonparametric approaches, the estimation of the conditional survivor function is not a black box procedure in CTMs. Although the model assumptions are weak in CTMs, a certain model structure is imposed by introducing additive partial transformation functions. The resulting effects of the explanatory variable over time can be interpreted and can be illustrated graphically. Hence, concerning model complexity, semiparametric CTMs can be placed in between the less flexible semiparametric linear transformation models (*e.g.*, the Cox model) and more flexible nonparametric approaches. If one is interested in better interpretable versions of CTMs, the model class of conditionally linear transformation models (CLTMs) introduced in Chapter 2 and Chapter 4 can be considered. In CLTMs, the conditional transformation function h is restricted to transformation functions that are linear in the response transformation. Due to this restriction, the explanatory variables are only allowed to influence the conditional mean and the conditional variance of the response transformation, whereas higher moments remain unaffected. The effects of the explanatory variables on the conditional mean and the conditional variance are non-linear but can be interpreted in CLTMs. Further restrictions of the transformation function are conceivable. For example, if all interaction terms between the survival time and the explanatory variables are omitted and the effects of the explanatory variables have to be linear, the conditional transformation function of the Cox model (Equation (5.3)) results as a special case. The Cox model can even be further restricted by choosing special forms of the monotone response transformation $h_T(t)$. For example, the specification of $h_T(t) = \log(\lambda) + \nu \cdot \log(t)$ results in the Weibull model (see Section 2.1.2).

5.2.1. Estimating conditional transformation models for survival data

Hothorn et al. (2014) explain thoroughly how CTMs are estimated by the minimisation of the continuous ranked probability score (CRPS) (see Gneiting and Raftery, 2007) using a component-wise boosting algorithm. The CRPS was chosen because it constitutes a proper scoring rule for distributional and probabilistic forecasts (Hothorn et al., 2014). When we estimated CTMs for survival data, we also used a component-wise boosting algorithm to minimise an appropriate integrated loss function. First, we formulated the integrated loss function for uncensored observations, and then we extended the loss function to right-censored observations.

Integrated loss function for uncensored observations. In an uncensored survival data setup, we observed the survival or event times T_i, $i = 1, \dots, N$, for N patients under consideration. Furthermore, we considered a grid of time points $\{t_\iota : \iota = 1, \dots, n\}$ ranging from the study's starting point $t_1 = 0$ to the study's end point t_n. Typical choices for the grid points $\{t_\iota : \iota = 1, \dots, n\}$ are equally spaced grid points or a grid composed of all distinct survival and event times. Hence, we were able to observe the binary survival status $I(T_i \leq t_\iota)$ for each patient at each grid point; the status is 1 if the patient experienced the event by t_ι and is otherwise 0.

We aimed at estimating the conditional distribution function of the event times $\mathbb{P}(T \leq t_\iota | \boldsymbol{X} = \boldsymbol{x}) = F(h(t_\iota|\boldsymbol{x}))$ (see Equation (5.2)) in terms of the conditional transformation function h, where t_ι denotes some arbitrary time point in the study period. This estimation problem can be reformulated as estimating the probability $F(h(t_\iota|\boldsymbol{x}))$ for the binary event $T \leq t_\iota$ and is solved by minimising an appropriate loss function. We chose the logarithmic score (or negative binomial log-likelihood) for measuring the loss between the binary event status $T_i \leq t_\iota$ and the corresponding probability $F(h(t_\iota|\boldsymbol{x}_i))$ for N patients at a specific time point t_ι:

$$\begin{aligned} \mathrm{LS}(t_\iota) &= -\frac{1}{N} \sum_{i=1}^{N} \{ I(T_i \leq t_\iota) \log(F(h(t_\iota|\boldsymbol{x}_i))) \\ &\qquad + I(T_i > t_\iota) \log(1 - F(h(t_\iota|\boldsymbol{x}_i))) \} \,. \end{aligned} \tag{5.5}$$

Alternatively, the Brier score or the absolute error loss can be chosen as an appropriate loss function (Hothorn et al., 2014; Gneiting and Raftery, 2007; Schemper and Henderson, 2000).

Based on the logarithmic score for one specific time point t_ι (see Equation (5.5)), we defined the integrated logarithmic score over all time points, which allows estimation of the whole conditional distribution function $\mathbb{P}(T \leq t|\boldsymbol{X} = \boldsymbol{x})$ in one step:

$$\begin{aligned} \mathrm{ILS} &= -\frac{1}{N} \sum_{i=1}^{N} \int_0^{t_n} \{ I(T_i \leq t) \log(F(h(t|\boldsymbol{x}_i))) \\ &\qquad + I(T_i > t) \log(1 - F(h(t|\boldsymbol{x}_i))) \} \; dW(t), \end{aligned} \tag{5.6}$$

where $W(t)$ denotes a weight function for the time points. By choosing the same weight ${}^1\!/_n$ for all time points t_ι, $\iota = 1, \dots, n$, we get the empirical version of Equation (5.6):

$$\begin{aligned} \widehat{\mathrm{ILS}} &= -\frac{1}{N \cdot n} \sum_{i=1}^{N} \sum_{\iota=1}^{n} \{ I(T_i \leq t_\iota) \log(F(h(t_\iota|\boldsymbol{x}_i))) \\ &\qquad + I(T_i > t_\iota) \log(1 - F(h(t_\iota|\boldsymbol{x}_i))) \} \,, \end{aligned} \tag{5.7}$$

which is used as the empirical loss function in the boosting algorithm. Of course, other weight functions $W(t)$ for the time points are conceivable. The integrated logarithmic score for uncensored observations was also used for the estimation of C(L)TMs in Chapter 4 (Equation (4.12)).

When the conditional distribution function is estimated, the ultimate goal is to estimate the conditional transformation function h such that the empirical risk in Equation (5.7) is minimised. The minimisation of the empirical risk is equivalent to the minimisation of the loss between the true survival status at time point t_ι, $I(T_i \leq t_\iota)$ and the corresponding estimated survival probability $F(\hat{h}(t_\iota|\boldsymbol{x}_i))$ for all time points and all patients. In other words, the survivor function for a specific patient $\hat{S}(t_\iota|\boldsymbol{x}_i) = 1 - F(\hat{h}(t_\iota|\boldsymbol{x}_i))$, $\iota = 1, \ldots, n$, is estimated such that the survival probabilities fit the patient's true survival status best.

Integrated loss function for right-censored observations. In survival analysis, we often face right-censored survival times. We do not observe the true survival time T_i for the right-censored patients, and only the observed survival times $\tilde{T}_i = \min(T_i, C_i)$, $i = 1, \ldots, N$, are available. Van der Laan and Robins (2003) suggested the inverse probability of censoring weighting (IPCW) approach, which is one way to account for right-censored observations in model estimation, and was often used in the past (*e.g.*, see Gerds and Schumacher, 2006; Hothorn et al., 2006). For example, Robins and Finkelstein (2000) present an IPCW version of the Kaplan-Meier estimator and the log-rank test to account for noncompliance and dependent censoring. Van der Laan and Robins (2003) give an IPCW example for right-censored data with time-independent explanatory variables and censoring at random, and suggest that the full data loss function (*i.e.* the integrated logarithmic score in our case) be weighted by the inverse probability of censoring weights

$$\omega_{i\iota} = \frac{\Delta(t_\iota)}{\hat{K}(\min(T_i, t_\iota))}, \tag{5.8}$$

where $\Delta(t_\iota) = I(C_i > \min(T_i, t_\iota))$. $\hat{K}$ denotes the marginal Kaplan-Meier estimator of the censoring distribution, $\hat{K}(t) = \hat{\mathbb{P}}(T > t)$, based on $(\tilde{T}_i, 1 - \delta_i)$, $i = 1, \ldots, N$, hence on the observed survival times and the reverse censoring indicator, which is 1 for right-censored observations and 0 otherwise. Furthermore, the censoring time C_i is set to ∞ for uncensored observations.

To calculate the IPCWs for the integrated logarithmic score in Equation (5.7) based on Equation (5.8), we have to distinguish four different situations:

1. Uncensored observations ($\delta_i = 1$) that experience the event up to t_ι ($\tilde{T}_i \leq t_\iota$):

$$\omega_{i\iota} = \frac{I(\tilde{T}_i \leq t_\iota, \delta_i = 1) \cdot \overbrace{I(C_i > T_i)}^{=\Delta(t_\iota)=1}}{\hat{K}(T_i)} = \frac{1}{\hat{K}(T_i)} = \frac{1}{\hat{K}(\tilde{T}_i)}.$$

2. Uncensored observations ($\delta_i = 1$) that do not experience the event up to t_ι ($\tilde{T}_i > t_\iota$):

$$\omega_{i\iota} = \frac{I(\tilde{T}_i > t_\iota, \delta_i = 1) \cdot \overbrace{I(C_i > t_\iota)}^{=\Delta(t_\iota)=1}}{\hat{K}(t_\iota)} = \frac{1}{\hat{K}(t_\iota)}.$$

3. Right-censored observations ($\delta_i = 0$) that experience the censoring up to t_ι ($\tilde{T}_i \leq t_\iota$):

$$\omega_{i\iota} = \frac{I(\tilde{T}_i \leq t_\iota, \delta_i = 0) \cdot \overbrace{I(C_i > T_i)}^{=\Delta(t_\iota)=0}}{\hat{K}(\text{NA})} = 0.$$

4. Right-censored observations ($\delta_i = 0$) that do not experience the censoring up to t_ι ($\tilde{T}_i > t_\iota$):

$$\omega_{i\iota} = \frac{I(\tilde{T}_i > t_\iota, \delta_i = 0) \cdot \overbrace{I(C_i > t_\iota)}^{=\Delta(t_\iota)=1}}{\hat{K}(t_\iota)} = \frac{1}{\hat{K}(t_\iota)}.$$

The resulting weighting scheme corresponds exactly to the weighting scheme given in Graf et al. (1999), which results in a consistent estimator (see Gerds and Schumacher, 2006). In short, the observations are weighted by the inverse probability of not being censored up to the event time (situation 1) or up to the specific time point under consideration (situations 2 and 4). The current survival status is unknown in situation 3; consequently these observations get zero weights. Thus, censored observations contribute to the model estimation process up to their censoring time point, and those observations that have already been censored are accounted for in the inverse probability of censoring weights.
We extended the empirical logarithmic score for uncensored observations given in Equation (5.7) to right-censored observations by including the weighting scheme presented above. Hence, the empirical version of the integrated censored logarithmic score results in

$$\begin{aligned} \widehat{\text{ILS}^C} \quad = \quad & -\frac{1}{N \cdot n} \sum_{i=1}^{N} \sum_{\iota=1}^{n} \Bigg\{ I(\tilde{T}_i \leq t_\iota, \delta_i = 1) \log(F(h(t_\iota|\boldsymbol{x}_i))) \cdot \frac{1}{\hat{K}(\tilde{T}_i)} \\ & + I(\tilde{T}_i > t_\iota) \log(1 - F(h(t_\iota|\boldsymbol{x}_i))) \cdot \frac{1}{\hat{K}(t_\iota)} \Bigg\}, \end{aligned} \tag{5.9}$$

which is used as empirical loss function in the boosting algorithm.

5.2.2. Boosting conditional transformation models for survival data

In CTMs, the conditional distribution function of uncensored responses is estimated using component-wise boosting with penalisation (Section 3.1; for a detailed description, see Hothorn et al., 2014). This algorithm has to be slightly modified for the estimation of right-censored survival data. Thereby, the empirical risk given in Equation (5.9) is minimised with respect to the transformation function h. Furthermore, the parametrisation of the partial transformation functions h_j, $j = 1, \dots, J$, (Equation (5.4)) has to be slightly adapted for survival data. In component-wise boosting algorithms, regularisation is achieved by the application of penalised base-learners. The overall model complexity is regulated by the number of boosting iterations M. For a thorough introduction to component-wise boosting with smooth base-learners, see Bühlmann and Hothorn (2007) and Schmid and Hothorn (2008).

Parametrisation of the partial transformation functions. Considering the parametrisation of the partial transformation functions in Hothorn et al. (2014), we defined for the j-th partial transformation function:

$$h_j(t_\iota|\boldsymbol{x}) = \left(\boldsymbol{b}_j(\boldsymbol{x})^\top \otimes \boldsymbol{b}_T(t_\iota)^\top\right)\boldsymbol{\gamma}_j, \;\; j = 1, \dots, J, \tag{5.10}$$

where $\boldsymbol{b}_T : \mathbb{R} \to \mathbb{R}^{K_T}$ denotes the basis along the grid of time points t_ι, $\iota = 1, \dots, n$, and $\boldsymbol{b}_j : \chi \to \mathbb{R}^{K_j}$ is a basis for (a subset of) the explanatory variables $\boldsymbol{x}$. Both sets of basis functions are connected via a Kronecker product, whereby an interaction surface between the survival times and the explanatory variables is established. The vector $\boldsymbol{\gamma}_j \in \mathbb{R}^{K_j K_T}$ contains the basis coefficients for the established interaction surface. The basis $\boldsymbol{b}_T$ defines the functional form of the transformation of the survival times, and the functional form of $\boldsymbol{b}_j$ defines how the survival time transformation is influenced by the explanatory variables (Hothorn et al., 2014). Hence, one usually chooses B-spline basis functions for $\boldsymbol{b}_T$, and depending on the desired flexibility or the measurement level of the explanatory variables, one chooses linear basis functions or B-spline basis functions for $\boldsymbol{b}_j$. In more detail, linear basis functions are chosen for $\boldsymbol{b}_j$ if $\boldsymbol{x}$ is univariate and categorical or if $\boldsymbol{x}$ is univariate and continuous, and a linear influence is assumed. B-spline basis functions are chosen for $\boldsymbol{b}_j$ if $\boldsymbol{x}$ is univariate and continuous, and the influence might be more flexible. Additionally, $\boldsymbol{b}_j$ might depend on more than one explanatory variable, and appropriate multivariate basis functions have to be considered. The partial transformation functions h_j are typically expected to be smooth in the first argument t and in the conditioning variable $\boldsymbol{x}$ because continuous distribution functions have to be smooth in the response variable. Moreover, we expect similar distribution functions for similar values of the explanatory variables. Therefore, appropriate penalty matrices $P_T \in \mathbb{R}^{K_T \times K_T}$ and $P_j \in \mathbb{R}^{K_j \times K_j}$ are imposed on the basis functions defined in Equation (5.10). The penalty matrix for the Kronecker product of the basis functions is defined via $P_{Tj} = (\lambda_j P_j \otimes \mathbf{1}_{K_T} + \lambda_T \mathbf{1}_{K_j} \otimes P_T)$, where $\lambda_T \geq 0$ and $\lambda_j \geq 0$ denote smoothing parameters and $\mathbf{1}$ denotes the identity matrix.
As an example, we give the partial transformation function for the explanatory variable sex influencing the survival time transformation:

$$h_{\text{sex}}(t_\iota|\text{sex}) = \left(\boldsymbol{b}^{\text{lin}}_{\text{sex}}(\text{sex})^\top \otimes \boldsymbol{b}_T(t_\iota)^\top\right)\boldsymbol{\gamma}_{\text{sex}}.$$

As the explanatory variable sex is binary, we chose linear basis functions for $\boldsymbol{b}^{\text{lin}}_{\text{sex}}(\text{sex})$, and furthermore, we chose B-spline basis functions for $\boldsymbol{b}_T$. No penalty term P_{sex} is specified for the linear basis $\boldsymbol{b}^{\text{lin}}_{\text{sex}}$ and a smoothness penalty term based on second-order differences P_T is defined for the B-spline basis $\boldsymbol{b}_T$. The resulting interaction surface for the explanatory variable sex and the survival time can also be understood as the separate estimation of a smooth survival time transformation for males and females. Hence, the difference in the survival probabilities of males and females is allowed to vary flexibly over time and is therefore able to display non-proportional hazards for the explanatory variable sex. For further details on parametrisation and penalty specification, see Hothorn et al. (2014).

Component-wise boosting algorithm for conditional transformation models for survival data. The component-wise boosting algorithm for right-censored survival data is only a slight modification of the algorithm presented in Chapter B (Hothorn et al., 2014):

(Init) Initialise the parameters $\boldsymbol{\gamma}_j^{[0]} \equiv 0$ for $j = 1, \ldots, J$, the step-size $\nu \in (0, 1)$ and the smoothing parameters λ_j, $j = 1, \ldots, J$. Define the grid $t_1 < \tilde{T}_{(1)} < \ldots < \tilde{T}_{(N)} \leq t_n$. Calculate the inverse probability of censoring weights $\omega_{i\iota}$ for each grid point ι and each observation i.
Set $m = 0$.

(Gradient) Compute the negative gradient of the censored log score:

$$\begin{aligned} U_{i\iota} &:= -\frac{\partial}{\partial h}\rho((\tilde{T}_i \leq t_\iota, \boldsymbol{x}_i), h)\bigg|_{h=\hat{h}_{i\iota}^{[m]}} \\ &:= \left\{ I(\tilde{T}_i \leq t_\iota, \delta_i = 1)\, \frac{F^{\prime}(h(t_\iota|\boldsymbol{x}_i))}{F(h(t_\iota|\boldsymbol{x}_i))} \cdot \frac{1}{\hat{K}(\tilde{T}_i)} \right. \\ &\qquad \left. - I(\tilde{T}_i > t_\iota)\, \frac{F^{\prime}(h(t_\iota|\boldsymbol{x}_i))}{1 - F(h(t_\iota|\boldsymbol{x}_i))} \cdot \frac{1}{\hat{K}(t_\iota)} \right\}\Bigg|_{h=\hat{h}_{i\iota}^{[m]}}, \end{aligned}$$

where $F^{\prime}(\cdot)$ denotes the density of the link function F, $\hat{K}(\cdot)$ denotes the marginal Kaplan-Meier estimator of the censoring distribution and

$$\hat{h}_{i\iota}^{[m]} = \sum_{j=1}^{J} \hat{h}_j^{[m]}(t_\iota|\boldsymbol{x}_i) = \sum_{j=1}^{J} \left(\boldsymbol{b}_j(\boldsymbol{x}_i)^\top \otimes \boldsymbol{b}_T(t_\iota)^\top\right) \boldsymbol{\gamma}_j^{[m]}.$$

Fit the base-learners for $j = 1, \ldots, J$:

$$\hat{\boldsymbol{\beta}}_j = \underset{\boldsymbol{\beta} \in \mathbb{R}^{K_j \cdot K_T}}{arg\,min} \sum_{i=1}^{N} \sum_{\iota=1}^{n} \omega_{i\iota} \left\{U_{i\iota} - \left(\boldsymbol{b}_j(\boldsymbol{x}_i)^\top \otimes \boldsymbol{b}_T(t_\iota)^\top\right) \boldsymbol{\beta}\right\}^2 + \boldsymbol{\beta}^\top P_{Tj} \boldsymbol{\beta}$$

with penalty matrix P_{Tj}.
Select the best base-learner

$$j^* = \underset{j=1,\ldots,J}{arg\,min} \sum_{i=1}^{N} \sum_{\iota=1}^{n} \omega_{i\iota} \left\{U_{i\iota} - \left(\boldsymbol{b}_j(\boldsymbol{x}_i)^\top \otimes \boldsymbol{b}_T(t_\iota)^\top\right) \hat{\boldsymbol{\beta}}_j\right\}^2.$$

(Update) the parameters $\boldsymbol{\gamma}_{j^*}^{[m+1]} = \boldsymbol{\gamma}_{j^*}^{[m]} + \nu \cdot \hat{\boldsymbol{\beta}}_{j^*}$ and keep all other parameters fixed, *i.e.* $\boldsymbol{\gamma}_{j}^{[m+1]} = \boldsymbol{\gamma}_{j}^{[m]}$, $j \neq j^*$.

Iterate (Gradient) and (Update).

(Stop) if $m = M$. Output the final model

$$\begin{aligned}\hat{\mathbb{P}}(T \leq t | \mathbf{X} = \boldsymbol{x}) &= F(\hat{h}^{[M]}(t|\boldsymbol{x})) = F\left(\sum_{j=1}^{J} \hat{h}_j^{[M]}(t|\boldsymbol{x})\right) \\ &= F\left(\sum_{j=1}^{J} \left(\boldsymbol{b}_j(\boldsymbol{x})^\top \otimes \boldsymbol{b}_T(t)^\top\right) \boldsymbol{\gamma}_j^{[M]}\right)\end{aligned}$$

as a function of arbitrary $t \in \mathbb{R}^+$ and arbitrary explanatory variables $\boldsymbol{x}$.

5.2.3. Software

All analyses were carried out in the R system of statistical computing (R Core Team, 2014). CTMs were estimated using the R add-on package **ctmDevel** (Hothorn, 2013). To compare the proposed CTMs for survival data with established models, we estimated Cox models using the R add-on package **survival** (Therneau, 2013), calculated Kaplan-Meier estimators using the R add-on package **prodlim** (Gerds, 2013) and estimated conditional random forests using the R add-on package **party** (Hothorn et al., 2006). R code for reproducing the results of Section 5.3 (in `ctmDevel/inst/empeval`) and Section 5.4 (in `ctmDevel/inst/applications`) is publicly available in the **ctm** package from the R-forge repository (`https://r-forge.r-project.org/projects/ctm`).

5.3. Simulation

5.3.1. Simulation study setup

In the following simulations, we investigated the performance of CTMs in comparison to alternative semiparametric (ordinary Cox model and stratified Cox model) or nonparametric (Kaplan-Meier estimator; conditional random forests) modelling strategies in four different simulation settings with Weibull distributed survival times. We considered different scenarios of explanatory variables and proportional as well as non-proportional hazard settings. Since the handling of censored observations is an important issue, we considered different amounts of right-censored survival times. The censoring times were drawn independently from uniform distributions such that 5%, 10%, 25% and 50% right-censored observations resulted in each simulation setting.

The true hazard function and the corresponding true survivor function for Weibull distributed survival times are

$$\lambda(t) = \frac{c}{b^c} t^{c-1} \quad \text{and} \quad S(t) = \exp(-b^{-c} t^c), \tag{5.11}$$

where b and c denote the scale and shape parameter of the Weibull distribution, respectively. The choice of parameters b and c determines whether proportional hazards or non-proportional hazards result. The PH assumption is fulfilled if the explanatory variables influence only the scale parameter b and the shape parameter c is fixed. If the explanatory variables additionally influence the shape parameter c, the PH assumption is violated, which, *e.g.*, results in crossing survivor functions.

Simulation 1 In the first simulation setting, we considered the simple data setting of two treatment groups $G1$ and $G2$, which differed with respect to their survival probabilities. The survival times were Weibull distributed with $b_1 = 1$ and $c_1 = 3$ for treatment group $G1$ and $b_2 = 1.5$ and $c_2 = 3$ for treatment group $G2$. Moreover, we included a non-informative continuous covariate x. Since the shape parameters were identical, the corresponding survivor functions followed the PH assumption (Figure 5.1). This could also be recognised by rewriting the conditional Weibull distribution in terms of the Cox linear transformation model (Equation (5.3)). The conditional Weibull distribution resulted from Equation (5.11) by inserting the scale parameter $b = \beta_G + \beta_x \cdot x$, where $\beta_G = 1$ for $G1$ and $\beta_G = 1.5$ for $G2$, and the shape parameter $c = \gamma_G + \gamma_x \cdot x$ with $\gamma_G = 3$ for both treatment groups. Since x was non-influential, $\beta_x = \gamma_x = 0$:

$$\begin{aligned} 1 - S(t|G,x) &= 1 - \exp(-(\beta_G + \beta_x \cdot x)^{-\gamma_G - \gamma_x \cdot x} \cdot t^{\gamma_G + \gamma_x \cdot x}) \\ &\underset{\gamma_x = \beta_x = 0,\, \gamma_G = 3}{=} 1 - \exp(-\exp(-3 \cdot \log(\beta_G) + 3 \cdot \log(t))) \\ &= \mathcal{M}(h_T(t) + \tilde{\beta}_G), \end{aligned}$$

where $h_T(t) = 3 \cdot \log(t)$ and $\tilde{\beta}_G = -3 \cdot \log(\beta_G)$. This setting could be perfectly fitted using a Cox model as there was no interaction term between the treatment group G and the survival time t (*i.e.* the PH assumption was fulfilled), and G had a linear influence. We sampled $N_G = 200$ survival times T from the respective Weibull distribution for each treatment group and identical $N_G = 200$ independent x-values were chosen on an equidistant grid on $[0, 1]$ for the treatment groups.

Simulation 2 In analogy to Simulation 1, the survival probabilities differed for treatment groups $G1$ and $G2$, and the continuous explanatory variable x was non-informative. The parameters of the Weibull distributed survival times were $b_1 = 1.5$ and $c_1 = 3$ for treatment group $G1$ and $b_2 = 1$ and $c_2 = 1$ for treatment group $G2$. Since the scale and the shape parameters were treatment specific, the PH assumption was violated (Figure 5.1).

Again, this could be clarified by writing the conditional Weibull distribution in terms of Equation (5.3):

$$\begin{aligned} 1 - S(t|G,x) &= 1 - \exp(-(\beta_G + \beta_x \cdot x)^{-\gamma_G - \gamma_x \cdot x} \cdot t^{\gamma_G + \gamma_x \cdot x}) \\ &\underset{\beta_x = \gamma_x = 0}{=} 1 - \exp(-\exp(-\gamma_G \cdot \log(\beta_G) + \gamma_G \cdot \log(t))), \end{aligned}$$

where $\beta_G = 1.5$ for $G1$ and $\beta_G = 1$ for $G2$, and $\gamma_G = 3$ for $G1$ and $\gamma_G = 1$ for $G2$. Since there is an interaction term between G and t, the PH assumption was violated. We sampled $N_G = 200$ survival times for each treatment group from the respective Weibull distributions. The independent and identical $N_G = 200$ x-values were chosen on an equidistant grid on $[0, 1]$ for the treatment groups.

Simulation 3 The survival times differed with respect to the treatment group G *and* with respect to the continuous explanatory variable x in simulation setting 3. The survival times were Weibull distributed with scale parameters $b_1 = \exp(^1\!/_4 + x)$ for treatment group $G1$ and $b_2 = \exp(1 + x)$ for treatment group $G2$. The x-values were chosen equidistantly on $[0, 1]$. The shape parameters $c_1 = c_2 = 3$ were identical, which resulted in the PH assumption. Again, the connection to the Cox model could be established in terms of Equation (5.3). We inserted $b = \exp(\beta_G + \beta_x \cdot x)$ for the scale parameter, where $\beta_G = {}^1\!/_4$ for $G1$ and $\beta_G = 1$ for $G2$, $\beta_x = 1$, and $c = 3$:

$$\begin{aligned} 1 - S(t|G,x) &= 1 - \exp(-\exp(-3 \cdot (\beta_G + x) + 3 \cdot \log(t))) \\ &= \mathcal{M}(h_T(t) + \tilde{\boldsymbol{x}}^\top \tilde{\boldsymbol{\beta}}), \end{aligned}$$

where $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_G \;\; \tilde{\beta}_x)^\top$ and $\tilde{\boldsymbol{x}} = (G \;\; x)^\top$. More precisely, the parameters of the linear transformation model were $\tilde{\beta}_G = -^3\!/_4$ for $G1$ and $\tilde{\beta}_G = -3$ for $G2$, $\tilde{\beta}_x = -3$ and $h_T(t) = 3 \cdot \log(t)$. Hence, the simulation setting could be perfectly analysed using a Cox model as there were no interactions between the explanatory variables and the survival time, and G and x had a linear influence. First, we chose $N_G = 300$ x-values by defining an equidistant grid on $[0, 1]$ for the treatment groups. Afterwards, we sampled 300 survival times from the Weibull distributions with parameters b_1 and c for treatment group $G1$ and 300 survival times from the Weibull distributions with parameters b_2 and c for treatment group $G2$.

Simulation 4 In analogy to Simulation 3, the survival probabilities were influenced by G and x. But this time, we chose a non-proportional hazards setting by keeping the scale parameter $b = \exp(^1\!/_2)$ fixed and letting the shape parameter depend on the explanatory variables: $c_1 = 2 + x^2$ for treatment group $G1$ and $c_2 = 2.5 + x^2$ for treatment group $G2$. More general, the shape parameter c is $c = \gamma_G + \gamma_x(x)$ with $\gamma_G = 2$ for $G1$ and $\gamma_G = 2.5$ for $G2$ and a non-linear function in x, $\gamma_x(x) = x^2$. Hence, the shape parameters differed only slightly for the treatment groups and were mainly influenced non-linearly by x. Again, the

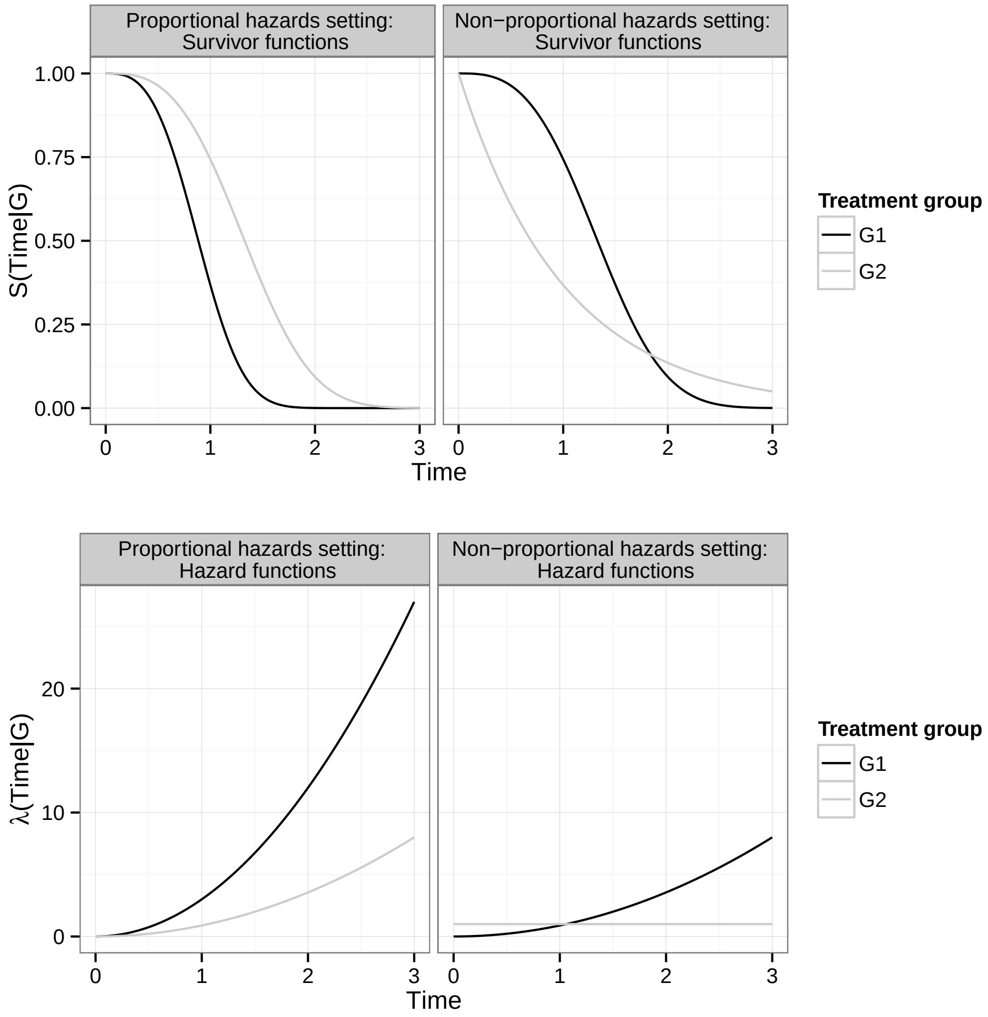


Figure 5.1.: Simulation: True survivor and hazard functions for treatment groups $G1$ and $G2$ based on Weibull distributed survival times. Proportional hazards setting (Simulation 1): $b_1 = 1$, $c_1 = 3$ (for $G1$) and $b_2 = 1.5$, $c_2 = 3$ (for $G2$); Non-proportional hazards setting (Simulation 2): $b_1 = 1.5$, $c_1 = 3$ and $b_2 = 1$, $c_2 = 1$.

conditional Weibull distribution of the survival times could be displayed as a conditional transformation model:

$$1 - S(t|G, x) = 1 - \exp\left(-\exp\left(-\frac{1}{2} \cdot \gamma_G - \frac{1}{2}x^2 + \gamma_G \cdot \log(t) + x^2 \cdot \log(t)\right)\right).$$

As there were interactions between the explanatory variables and the survival time, the PH assumption was violated. We first chose $N_G = 300$ x-values by defining an equidistant grid on $[0, 2]$. In this simulation setting, the x-values varied on $[0, 2]$ instead of $[0, 1]$, which resulted in a wider range of shape values c. Afterwards, 300 survival times were sampled from the Weibull distributions with parameters b and c_1 for treatment group $G1$ and 300 survival times were sampled from the Weibull distributions with parameters b and c_2 for treatment group $G2$.

5.3.2. Model estimation

We estimated the conditional survival curves for the treatment groups $G1$ and $G2$ and the continuous covariate x, $S(t|G, x)$, using a CTM, an ordinary Cox model, and conditional random forests in all four simulations. In the Cox model, the hazard function was modelled via $\lambda(t|G, x) = \lambda_0(t) \exp(\beta_G \cdot G + \beta_x \cdot x)$, where $\lambda_0(t)$ denotes the baseline hazard. In the CTM, a partial transformation function for each explanatory variable was defined, $h(t|G, x) = h_G(t|G) + h_x(t|x)$, in which the influence of the explanatory variables was allowed to vary over time.

Separate Kaplan-Meier estimators can only be obtained for categorical explanatory variables. As x was non-influential in simulations 1 and 2, treatment-specific Kaplan-Meier estimates were additionally provided.

A non-proportional hazards setting was considered in simulations 2 and 4. Therefore, we additionally estimated a stratified Cox model with treatment-specific baseline hazard functions: $\lambda(t|G, x) = \lambda_G(t) \cdot \exp(\beta_x \cdot x)$.

The flexibility of CTMs can be restricted to the flexibility of a Cox model by considering a conditionally linear transformation model (CLTM) (Section 2.2.2 and Section 4.3.1; Möst et al. (2014)). We avoided all interactions between the explanatory variables and the survival time and assumed linear influences for G and x in the corresponding conditional transformation function: $h(t|G, x) = h_G(1|G) + h_x(1|x) + h_T(t|1) = \beta_G \cdot G + \beta_x \cdot x + h_T(t)$. Hence, the CLTM and the Cox model could be considered as semiparametric alternatives, in which both models assumed proportional hazards.

Simulation 1 The conditional survivor functions were estimated using a CTM, a CLTM, an ordinary Cox model, the Kaplan-Meier estimator, and conditional random forests. Thereby, the treatment-specific Kaplan-Meier estimator could be understood as a nonparametric alternative to conditional random forests, in which the Kaplan-Meier estimator was expected to perform better as the non-informative explanatory variable x was ignored. The CLTM and the Cox model were semiparametric alternatives that were expected to perform comparably well, as both approaches profited from the PH assumption. The CTM was expected to perform slightly worse, as it was more flexible, but the additional flexibility was not necessary.

Simulation 2 In this non-proportional hazards setting, the conditional survivor functions were additionally estimated using a stratified Cox model. As x was non-informative, treatment-specific Kaplan-Meier estimators were obtained as both a nonparametric and a predominant alternative to conditional random forests. The CLTM and the ordinary

Cox model were expected to perform comparably poorly, as the underlying PH assumption was violated. In contrast, the stratified Cox model and the CTM were expected to perform comparably well, as treatment-specific baseline hazards and a time-varying treatment effect were allowed, respectively.

Simulation 3 We estimated the conditional survivor functions $S(t|G, x)$ using a CTM, a CLTM, a Cox model, and conditional random forests. As the chosen simulation setting could be perfectly fitted using a Cox model, the CLTM and the Cox model were expected to perform best. Conditional random forests and the CTM were expected to perform slightly worse due to the additional superfluous flexibility. In general, the identification of the linear influence of x is difficult for conditional random forests, as the linear function has to be approximated by a step function.

Simulation 4 The conditional survivor functions were estimated using a CTM, a CLTM, an ordinary Cox model and a stratified Cox model, and conditional random forests. As the PH assumption was violated the Cox model, and the CLTM should perform comparably poorly. The stratified Cox model was expected to perform slightly better, but nevertheless, it only assumed treatment-specific baseline hazards and still ignored the interaction between the survival time and x. Conditional random forests were able to account for non-proportional hazards in G and x by searching for adequate split points in the explanatory variables. Therefore, this method should outperform both Cox models and the CLTM. However, conditional random forests had difficulties in identifying the non-linear influence of x, which had to be approximated by a step function. The CTM was expected to outperform all alternative models including conditional random forests, as non-linear interactions between both explanatory variables and the survival time could be considered.

5.3.3. Model evaluation

We aimed at evaluating the goodness of the CTM, the CLTM, the Cox model (both ordinary and stratified), the Kaplan-Meier estimator, and conditional random forests for estimating the survivor functions of treatment groups $G1$ and $G2$ in all four simulation settings. Therefore, we used the out-of-sample uncensored log score (Equation (5.7)) and the mean absolute deviation (MAD) between the true and the estimated survivor functions as quality criteria.

For the evaluation, we drew $1,000$ new observations for each treatment group. In simulations 1 and 2, we simply drew $1,000$ new observations from the Weibull distributions with parameters b_1 and c_1, and b_2 and c_2, respectively. In simulations 3 and 4, we defined $1,000$ x-values by determining an equidistant grid on $[0, 1]$ and $[0, 2]$, respectively. Afterwards, we drew a new Weibull distributed survival time depending on the shape and scale parameter induced by each x-value.

Based on these new observations, we calculated separate uncensored log scores for the two treatment groups. As an example, we describe the calculation of the uncensored log score for treatment group G_1: We compared the true survivor status $I(T_l \leq t_\iota)$ for each new survival time and corresponding x-value, (T_l, x_l), $l = 1, \ldots, 1,000$, for treatment group G_1 along a grid of time points t_ι with the corresponding estimated survival probabilities $\pi(t_\iota|G_1, x_l)$. Thereby, the estimated conditional survival probabilities $\pi(t_\iota|G_1, x_l)$ resulted from the CTM, the CLTM, the Cox model (ordinary or stratified), or conditional random forests. The survival probabilities $\pi(t_\iota|G_1)$ were only treatment specific for the Kaplan-Meier estimator in simulations 1 and 2. The grid of time points t_ι consisted of all new survival times T_l, $l = 1, \ldots, 1,000$. The uncensored log score for treatment group G_2 was calculated analogously.

In addition, we calculated the MAD of the estimated survival curves and the true Weibull distribution functions for each treatment group separately. Thereby, we also considered the grid of $1,000$ x-values x_l, $l = 1, \ldots, 1,000$, and the grid consisting of the $1,000$ new survival times for each treatment group, t_ι, $\iota = 1, \ldots, 1,000$:

$$\text{MAD}(G_k) = \frac{1}{1,000 \cdot 1,000} \sum_{l=1}^{1,000} \sum_{\iota=1}^{1,000} |p(t_\iota|G_k, x_l) - \pi(t_\iota|G_k, x_l)|, \tag{5.12}$$

where p denotes the true survival probabilities and π denotes the estimated survival probabilities. Furthermore, $k \in \{1, 2\}$ denotes the index for the two treatment groups and $l = 1, \ldots, 1,000$ is the index for the new observations for each treatment group. In the simulation settings 1 and 2, the true survival probabilities $p(t_\iota|G_k, x_l)$ reduced to $p(t_\iota|G_k)$ as x was non-informative. For reasons of interpretability, the MAD values and the uncensored log scores were multiplied by 100.

This procedure was repeated for $B = 100$ simulated data sets. We calculated mean values of the resulting 100 MADs or uncensored log scores for the different treatment groups and the different estimation techniques.

Simulation 1 All estimation approaches yielded similar results. The calculated mean MADs (Table 5.1; Figure 5.2) were small for all model approaches and indicated that the estimated survivor functions were in good accordance with the true Weibull survivor functions. Only for 50% censored observations did the MADs grow larger throughout. The Cox model, the CLTM and the Kaplan-Meier estimator performed slightly better because the Cox model and the CLTM profited from the PH assumption and the Kaplan-Meier estimator ignored the non-informative explanatory variable x. Nevertheless, the uncensored log score was the more interesting quality criterion, as it evaluates how well the estimation techniques are able to predict the survivor status of *new* observations. Again, all four estimation approaches yielded similar results (Table 5.2; Figure 5.3). All uncensored log scores grew larger with an increasing amount of censored observations.

Table 5.1.: Simulation 1: Mean absolute deviations between true and estimated survival curves for each treatment group. The reported values are mean values over $B = 100$ simulations.

Treatment group	Model	Censoring			
		5%	10%	25%	50%
G1	CTM	2.87	2.96	3.45	6.18
	CLTM	2.43	2.51	3.02	5.97
	Cox	2.41	2.54	3.23	6.64
	Kaplan-Meier	2.41	2.51	3.21	6.63
	Cforest	2.59	2.67	3.34	6.74
G2	CTM	2.71	2.80	3.58	6.98
	CLTM	2.37	2.48	3.30	6.84
	Cox	2.30	2.42	3.29	6.88
	Kaplan-Meier	2.28	2.38	3.24	6.71
	Cforest	2.50	2.57	3.33	6.64

Table 5.2.: Simulation 1: Out-of-sample uncensored log score based on $1,000$ new observations for each treatment group. The reported values are mean values over $B = 100$ simulations.

Treatment group	Model	Censoring			
		5%	10%	25%	50%
G1	CTM	50.52	50.58	50.82	52.27
	CLTM	50.39	50.45	50.68	52.02
	Cox	50.40	50.47	50.78	52.41
	Kaplan-Meier	50.42	50.50	50.89	52.66
	Cforest	50.46	50.54	50.93	52.75
G2	CTM	50.54	50.63	50.97	52.79
	CLTM	50.47	50.58	50.89	52.70
	Cox	50.39	50.48	50.79	52.52
	Kaplan-Meier	50.45	50.55	50.92	52.75
	Cforest	50.49	50.60	50.95	52.74

Simulation 2 The MADs of the CTM, the stratified Cox model, the Kaplan-Meier estimator, and conditional random forests were similar throughout, whereas the ordinary Cox model and the CLTM clearly yielded higher MADs (Table 5.3; Figure 5.4). The only exception was the MADs for 50% censored observations, where all models had higher MADs. Moreover, the MADs for conditional random forests were most variable and the Kaplan-Meier estimator performed better, as it profited from ignoring x. The uncensored log scores gave similar results (Table 5.4; Figure 5.5). Again, the log scores for the CTM, the stratified Cox model, the Kaplan-Meier estimator, and conditional random forests were similar, whereas the ordinary Cox model and the CLTM clearly yielded higher values. One exception was the throughout larger uncensored log scores for 50% censored observations.

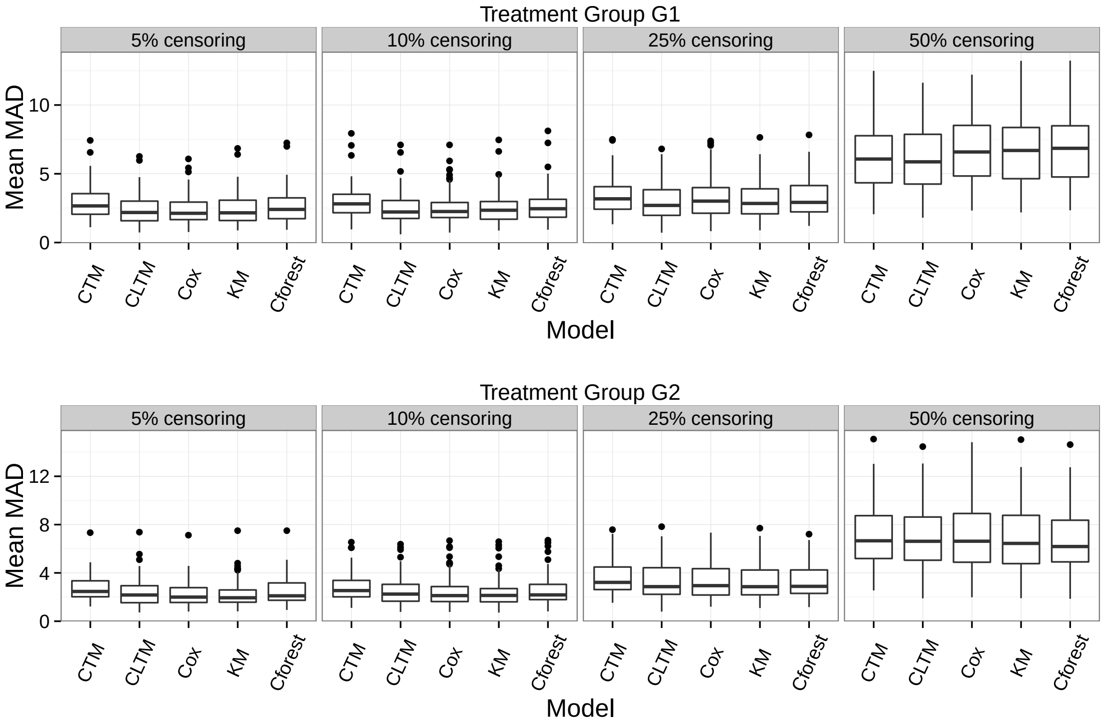


Figure 5.2.: Simulation 1: Boxplot of the treatment-specific mean MAD values based on $B = 100$ simulations for the conditional transformation model (CTM), the conditionally linear transformation model (CLTM), the Cox model (Cox), the Kaplan-Meier estimator (KM), and conditional random forests (Cforest). 5%, 10%, 25%, and 50% of right-censored observations were observed.

Table 5.3.: Simulation 2: Mean absolute deviations between true and estimated survival curves for each treatment group. The reported values are mean values over $B = 100$ simulations.

Treatment group	Model	Censoring			
		5%	10%	25%	50%
G1	CTM	2.60	2.74	3.31	6.34
	CLTM	8.14	8.08	8.27	10.26
	Cox	9.02	8.56	7.45	7.31
	Kaplan-Meier	2.58	2.74	3.43	6.84
	Cforest	2.39	2.50	3.19	6.64
	Stratified Cox	4.16	4.07	4.19	5.85
G2	CTM	3.15	3.41	5.03	10.99
	CLTM	7.60	8.20	9.80	14.30
	Cox	9.58	10.25	12.00	16.51
	Kaplan-Meier	2.55	2.88	4.83	11.15
	Cforest	2.34	2.64	4.63	10.96
	Stratified Cox	4.85	5.51	8.15	14.58

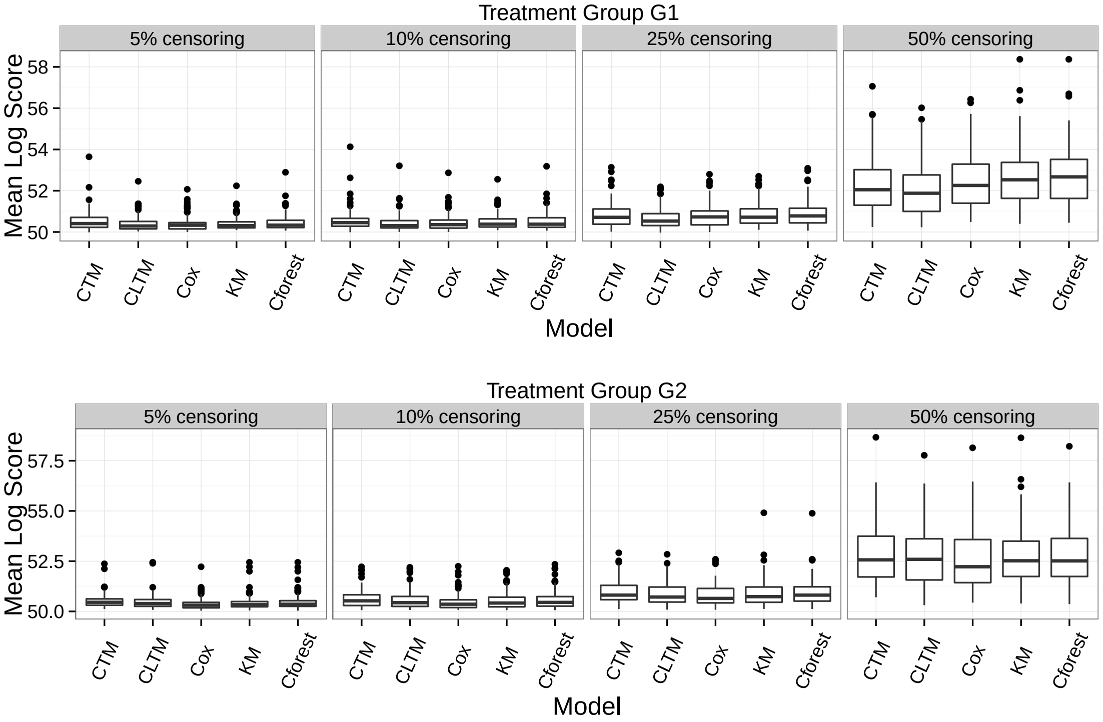


Figure 5.3.: Simulation 1: Boxplot of the out-of-sample mean uncensored log scores based on $1,000$ new observations for each treatment group and $B = 100$ simulations for the conditional transformation model (CTM), the conditionally linear transformation model (CLTM), the Cox model (Cox), the Kaplan-Meier estimator (KM), and conditional random forests (Cforest). 5%, 10%, 25%, and 50% of right-censored observations were observed.

Table 5.4.: Simulation 2: Out-of-sample uncensored log score based on $1,000$ new observations for each treatment group. The reported values are mean values over $B = 100$ simulations.

Treatment group	Model	Censoring			
		5%	10%	25%	50%
G1	CTM	50.43	50.51	50.79	52.30
	CLTM	53.12	53.14	53.41	55.06
	Cox	53.07	52.87	52.46	52.87
	Kaplan-Meier	50.46	50.54	50.91	52.66
	Cforest	50.42	50.50	50.89	52.67
	Stratified Cox	50.89	50.90	51.06	52.03
G2	CTM	50.75	50.92	51.65	55.46
	CLTM	53.47	53.85	55.13	60.09
	Cox	54.63	55.21	57.04	63.12
	Kaplan-Meier	50.52	50.70	51.53	55.45
	Cforest	50.48	50.66	51.47	55.37
	Stratified Cox	51.63	52.08	53.84	59.87

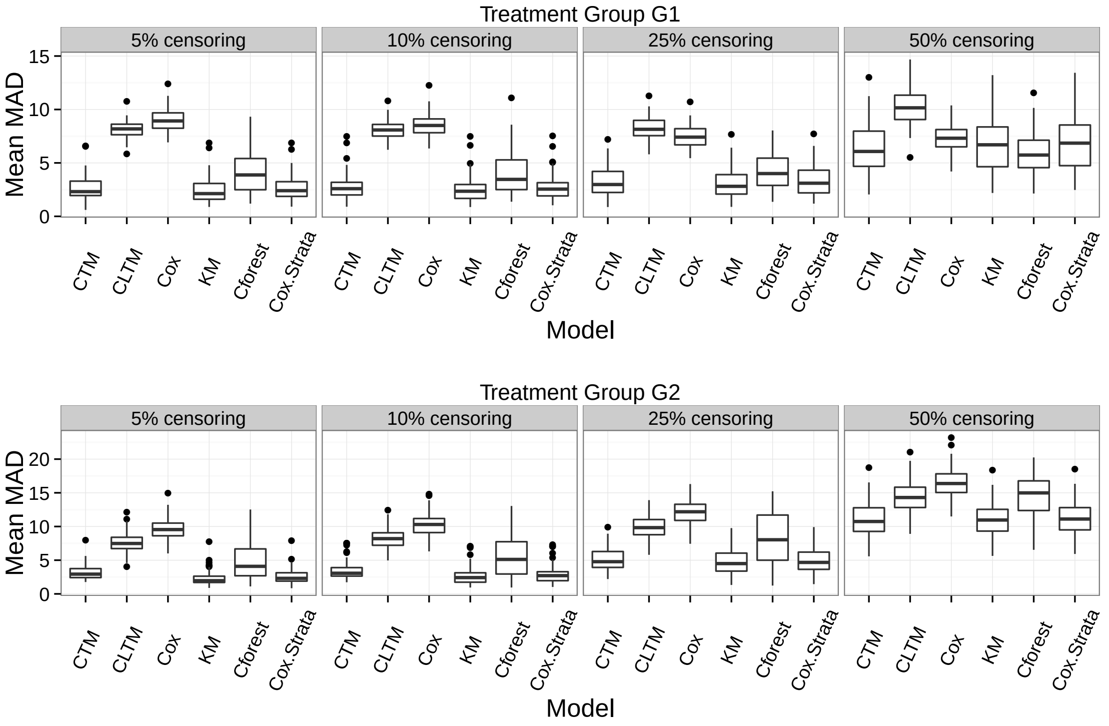


Figure 5.4.: Simulation 2: Boxplot of the treatment-specific mean MAD values based on $B = 100$ simulations for the conditional transformation model (CTM), the conditionally linear transformation model (CLTM), the Cox model (Cox), the Kaplan-Meier estimator (KM), conditional random forests (Cforest), and the stratified Cox model (Cox.Strata). 5%, 10%, 25%, and 50% of right-censored observations were observed.

Simulation 3 The Cox model and the CLTM approach performed almost equally well in the proportional hazards setting. The mean MADs (Table 5.5; Figure 5.6) and the out-of-sample uncensored log scores (Table 5.6; Figure 5.7) were similar for the Cox model and the CLTM, whereas the CTM was associated with slightly higher MADs and uncensored log scores. Conditional random forests performed worst because conditional random forests and the CTM were not able to profit from the PH assumption. Additionally, conditional random forests had difficulties in identifying the linear influence of x.

Simulation 4 The CTM performed better than all alternative modelling approaches, as it showed lower MADs for all amounts of censoring than the CLTM, the Cox model, the stratified Cox model, and conditional random forests (Table 5.7; Figure 5.8). Additionally, the CTM approach was associated with the smallest mean uncensored log scores (Table 5.8; Figure 5.9) because the CTM approach is the only approach that was able to account for the non-linear influence of x on the shape parameter of the Weibull distribution adequately. The Cox model and the CLTM performed worse owing to the PH assumption. As the non-proportionality of hazards was mainly induced by x, the stratified Cox model performed

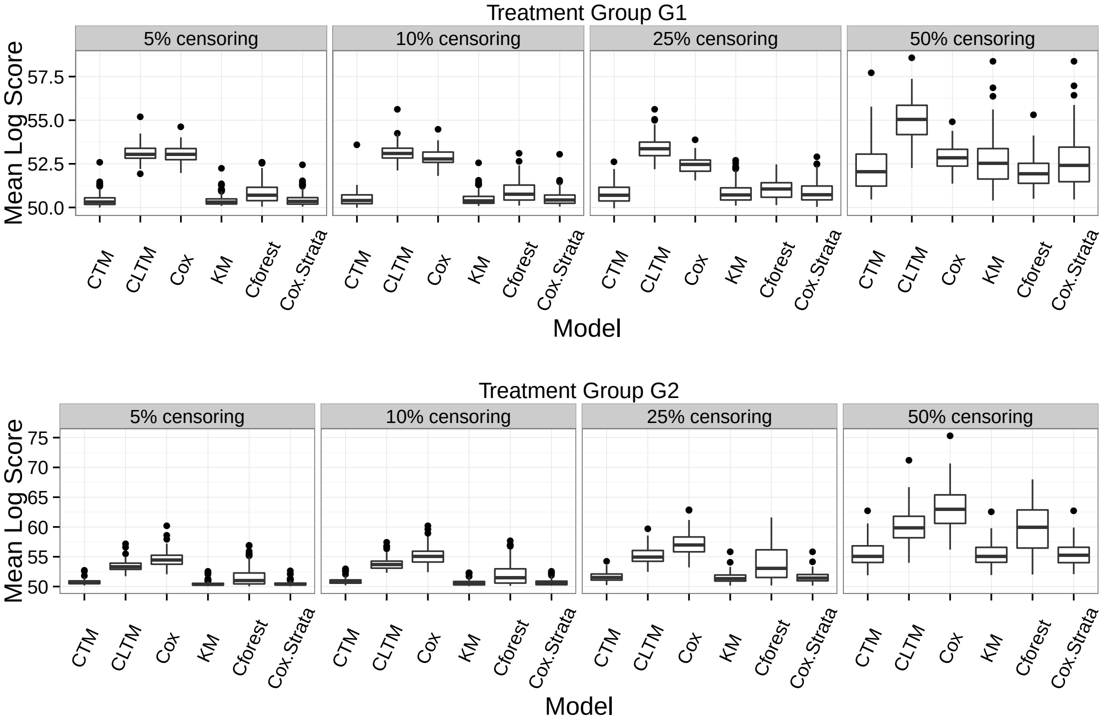


Figure 5.5.: Simulation 2: Boxplot of the out-of-sample mean uncensored log scores based on $1,000$ new observations for each treatment group and $B = 100$ simulations for the conditional transformation model (CTM), the conditionally linear transformation model (CLTM), the Cox model (Cox), the Kaplan-Meier estimator (KM), conditional random forests (Cforest), and the stratified Cox model (Cox.Strata). 5%, 10%, 25%, and 50% of right-censored observations were observed.

Table 5.5.: Simulation 3: Mean absolute deviations between true and estimated survival curves for each treatment group. The reported values are mean values over $B = 100$ simulations.

Treatment group	Model	Censoring			
		5%	10%	25%	50%
G1	CTM	1.88	1.90	2.46	4.42
	CLTM	1.67	1.69	2.28	4.37
	Cox	1.60	1.67	2.44	5.17
	Cforest	4.04	4.42	5.52	8.51
G2	CTM	1.97	2.06	2.71	5.07
	CLTM	1.74	1.78	2.43	4.81
	Cox	1.67	1.75	2.50	5.10
	Cforest	3.93	4.01	4.45	6.43

only slightly better than the ordinary Cox model in terms of the mean uncensored log score. Conditional random forests performed better than both Cox models and the CLTM as the approach can account for non-proportionality in G and x, but performed worse than the

Table 5.6.: Simulation 3: Out-of-sample uncensored log score based on 1,000 new observations for each treatment group. The reported values are mean values over $B = 100$ simulations.

Treatment group	Model	Censoring			
		5%	10%	25%	50%
G1	CTM	38.96	38.99	39.21	40.08
	CLTM	38.90	38.92	39.13	39.98
	Cox	38.93	38.97	39.25	40.44
	Cforest	40.03	40.31	41.20	44.22
G2	CTM	40.99	41.05	41.34	42.60
	CLTM	40.95	40.99	41.30	42.54
	Cox	40.86	40.91	41.20	42.40
	Cforest	41.74	41.82	42.17	44.00

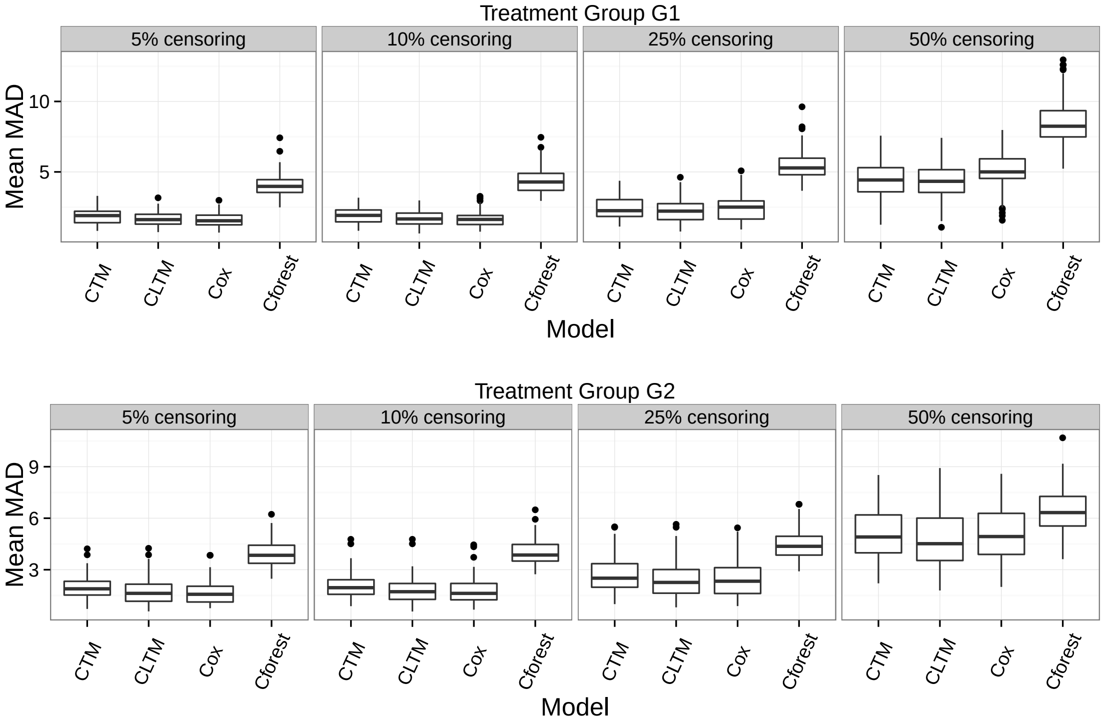


Figure 5.6.: Simulation 3: Boxplot of the treatment-specific mean MAD values based on $B = 100$ simulations for the conditional transformation model (CTM), the conditionally linear transformation model (CLTM), the Cox model (Cox), and conditional random forests (Cforest). 5%, 10%, 25%, and 50% of right-censored observations were observed.

CTM owing to the non-linear influence of x on the shape parameter. Again, conditional random forests had difficulties in identifying the non-linear influence of x.

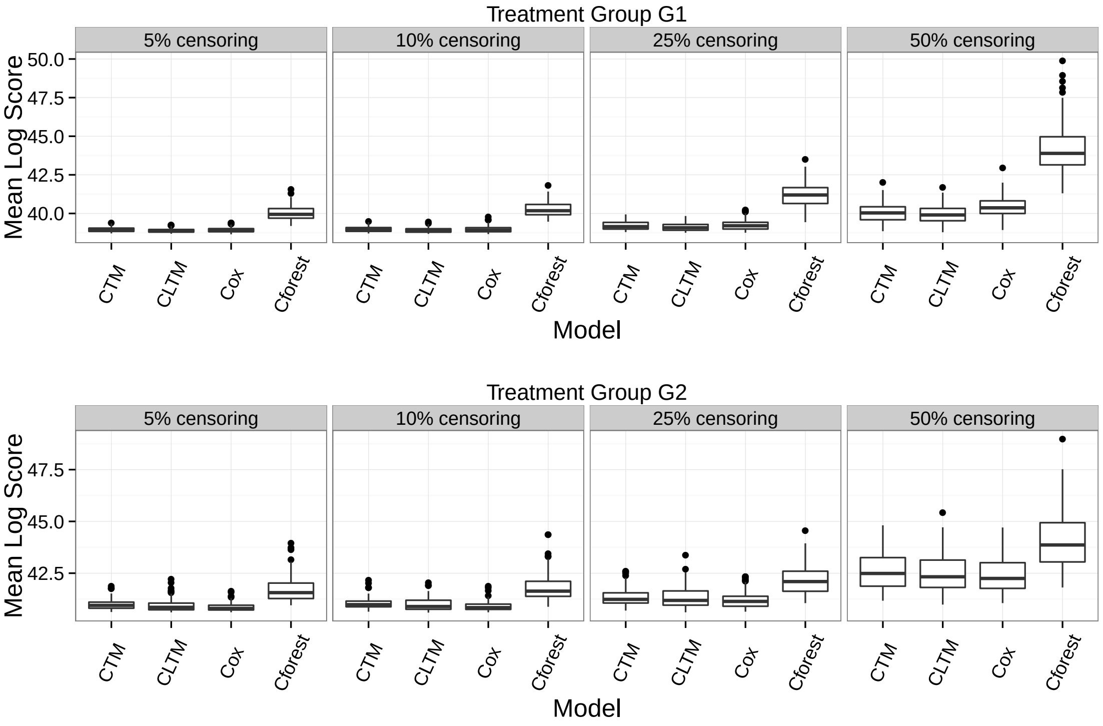


Figure 5.7.: Simulation 3: Boxplot of the out-of-sample mean uncensored log scores based on 1,000 new observations for each treatment group and $B = 100$ simulations for the conditional transformation model (CTM), the conditionally linear transformation model (CLTM), the Cox model (Cox), and conditional random forests (Cforest). 5%, 10%, 25%, and 50% of right-censored observations were observed.

Table 5.7.: Simulation 4: Mean absolute deviations between true and estimated survival curves for each treatment group. The reported values are mean values over $B = 100$ simulations.

Treatment group	Model	Censoring			
		5%	10%	25%	50%
G1	CTM	2.48	2.55	3.36	6.67
	CLTM	5.09	5.12	5.60	7.79
	Cox	5.75	5.82	6.43	8.71
	Cforest	4.34	4.42	4.92	7.14
	Stratified Cox	5.67	5.72	6.24	8.44
G2	CTM	2.38	2.46	3.26	5.78
	CLTM	4.47	4.51	4.96	6.83
	Cox	5.27	5.30	5.67	7.29
	Cforest	3.82	3.85	4.29	6.26
	Stratified Cox	5.32	5.39	5.90	7.66

Table 5.8.: Simulation 4: Out-of-sample uncensored log score based on 1,000 new observations for each treatment group. The reported values are mean values over $B = 100$ simulations.

Treatment group	Model	Censoring			
		5%	10%	25%	50%
G1	CTM	49.23	49.28	49.64	51.29
	CLTM	50.59	50.64	51.03	52.66
	Cox	50.95	51.04	51.57	53.45
	Cforest	50.71	50.78	51.20	52.97
	Stratified Cox	50.01	50.10	50.47	52.02
G2	CTM	49.23	49.28	49.55	50.66
	CLTM	50.21	50.25	50.46	51.46
	Cox	50.44	50.46	50.67	51.68
	Cforest	50.56	50.62	50.96	52.19
	Stratified Cox	49.78	49.81	50.01	51.04

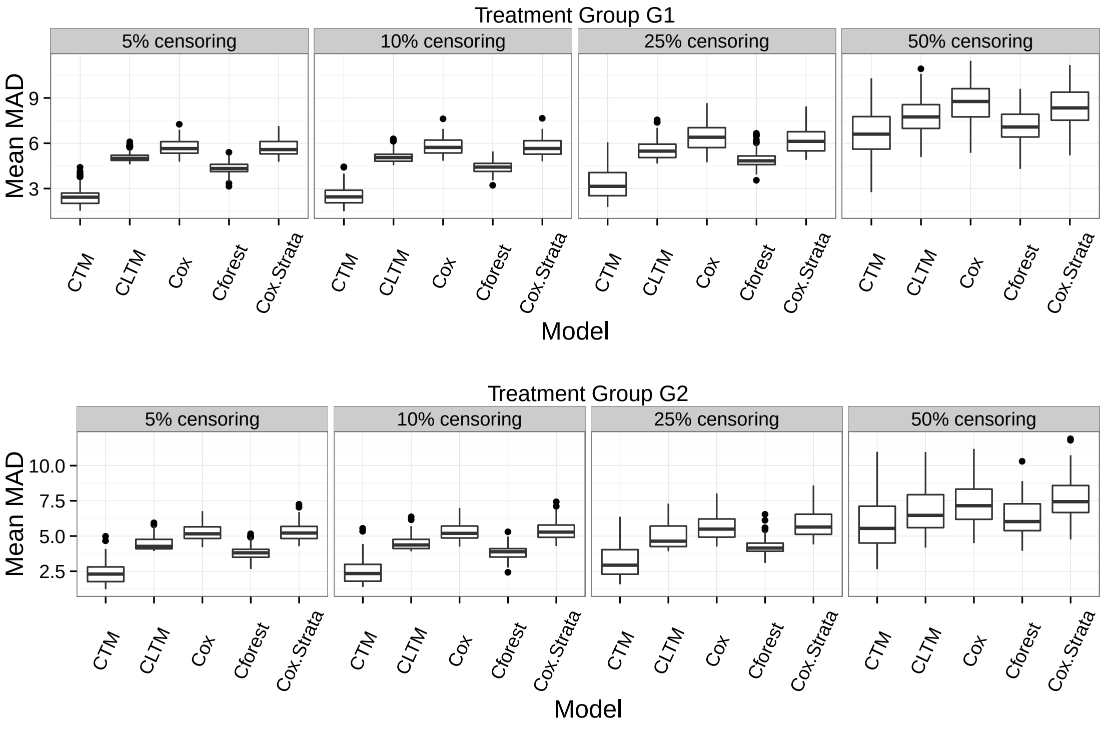


Figure 5.8.: Simulation 4: Boxplot of the treatment-specific mean MAD values based on $B = 100$ simulations for the conditional transformation model (CTM), the conditionally linear transformation model (CLTM), the Cox model (Cox), conditional random forests (Cforest), and the stratified Cox model (Cox.Strata). 5%, 10%, 25%, and 50% of right-censored observations were observed.

5.4. Chronic myelogenous leukaemia data

Curative bone marrow transplantation is feasible for only a minority of patients with chronic myelogenous leukaemia. Therefore, drug-based chemotherapy remains a treatment of cen-

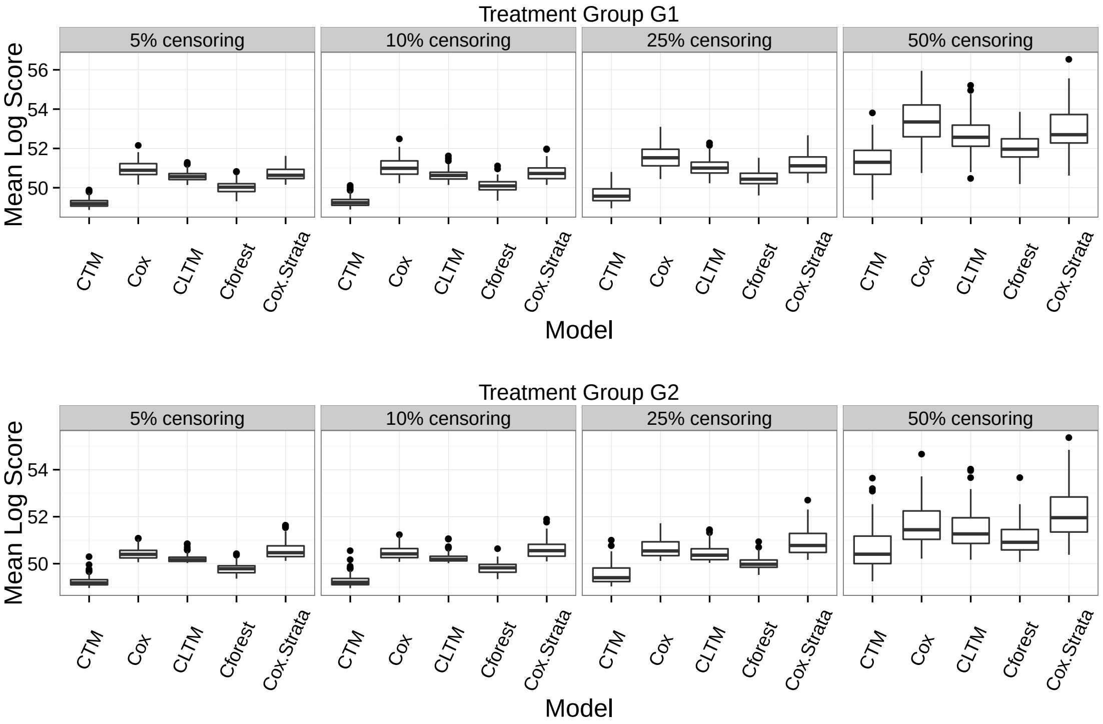


Figure 5.9.: Simulation 4: Boxplot of the out-of-sample mean uncensored log scores based on $1,000$ new observations for each treatment group and $B = 100$ simulations for the conditional transformation model (CTM), the conditionally linear transformation model (CLTM), the Cox model (Cox), conditional random forests (Cforest), and the stratified Cox model (Cox.Strata). 5%, 10%, 25%, and 50% of right-censored observations were observed.

tral interest. The standard chemotherapy has long been with the cytostatic drugs busulfan (BUS) or hydroxyurea (HU). In a multicentre, randomised study, Hehlmann et al. (1994) have shown that treatment with the drug interferon-α (IFN-α) significantly prolongs survival compared to treatment with BUS, and survival times after treatment with IFN-α or HU were not significantly different. Within the scope of the study, 516 eligible patients were recruited in 57 study centres from 1983 to 1991. For 507 of the 516 patients, complete data on sex, age and a prognostic score distinguishing between low, intermediate and high risk groups (Hasford et al., 1998) are available. Of the 507 patients, 132 random patients were treated with IFN-α, 182 were treated with BUS and 193 were treated with HU. Ninety patients were right-censored mainly due to bone marrow transplantation during the first chronic phase, and 417 patients died during the study period (Herberich and Hothorn, 2012).

5.4.1. Model estimation

Herberich and Hothorn (2012) analysed the treatment effects using a frailty Cox model (McGilchrist and Aisbett, 1991) with Gaussian frailties for the 57 study centres. Furthermore, age, sex, treatment and risk group were included as linear predictors. In our re-analysis of the CML data set, the main goals were to check the validity of the PH assumption in Cox models, which have been used for analyses in the past (*e.g.*, Herberich and Hothorn, 2012). Moreover, we were interested in possible interactions between the explanatory variables treatment, risk group, sex and age. More precisely, we were interested in whether the superiority of the IFN-α treatment found in former studies (*e.g.*, Hehlmann et al., 1994) is present in all risk groups. Additionally, treatment effectiveness might differ between men and women and patients of different age in the different risk groups. Therefore, we fitted five models to the CML data, in which the PH assumption and the considered interaction terms differed. The random effect for the study centres was excluded in all models for the purpose of model comparison, as we found its variance to be negligibly small.

First, we estimated an ordinary Cox model with linear influences for the explanatory variables treatment, risk group, sex and age:

$$\lambda(t_\iota|\boldsymbol{x}) = \lambda_0(t_\iota) \cdot \exp(\beta_{\text{tr}} \cdot x_{\text{tr}} + \beta_{\text{risk}} \cdot x_{\text{risk}} + \beta_{\text{sex}} \cdot x_{\text{sex}} + \beta_{\text{age}} \cdot x_{\text{age}}),$$

where $\lambda_0(\cdot)$ denotes the baseline hazard function and proportional hazards were assumed. To account for possible interactions between the categorical explanatory variables treatment, risk group and sex, we estimated an additional Cox model that included all two-time interactions between treatment, risk group and sex, and their three-time interaction. Again, all influences were assumed to be linear:

$$\begin{aligned}\lambda(t_\iota|\boldsymbol{x}) \quad = \quad & \lambda_0(t_\iota) \cdot \exp(\beta_{\text{tr}} \cdot x_{\text{tr}} + \beta_{\text{risk}} \cdot x_{\text{risk}} + \beta_{\text{sex}} \cdot x_{\text{sex}} + \beta_{\text{age}} \cdot x_{\text{age}} + \\ & \beta_{\text{tr:risk}} \cdot x_{\text{tr:risk}} + \beta_{\text{tr:sex}} \cdot x_{\text{tr:sex}} + \beta_{\text{risk:sex}} \cdot x_{\text{risk:sex}} + \\ & \beta_{\text{tr:risk:sex}} \cdot x_{\text{tr:risk:sex}}).\end{aligned}$$

As the Cox model assumed proportional hazards for all patient characteristics, we alternatively used a CTM for data analysis. In the CTM, the PH assumption was relaxed by allowing for flexible influences of each explanatory variable over time:

$$h(t_\iota|\boldsymbol{x}) = h_{\text{tr}}(t_\iota|\text{tr}) + h_{\text{risk}}(t_\iota|\text{risk}) + h_{\text{sex}}(t_\iota|\text{sex}) + h_{\text{age}}(t_\iota|\text{age}).$$

We defined separate partial transformation functions for treatment, risk group, sex and age that were specified in terms of basis functions:

$$\begin{aligned}
h_{\text{tr}}(t_\iota|\text{tr}) &= \left(b_{\text{tr}}^{\text{lin}}(\text{tr})^\top \otimes b_T(t_\iota)^\top\right) \boldsymbol{\gamma}_{\text{tr}}, \\
h_{\text{risk}}(t_\iota|\text{risk}) &= \left(b_{\text{risk}}^{\text{lin}}(\text{risk})^\top \otimes b_T(t_\iota)^\top\right) \boldsymbol{\gamma}_{\text{risk}}, \\
h_{\text{sex}}(t_\iota|\text{sex}) &= \left(b_{\text{sex}}^{\text{lin}}(\text{sex})^\top \otimes b_T(t_\iota)^\top\right) \boldsymbol{\gamma}_{\text{sex}} \text{ and} \\
h_{\text{age}}(t_\iota|\text{age}) &= \left(b_{\text{age}}(\text{age})^\top \otimes b_T(t_\iota)^\top\right) \boldsymbol{\gamma}_{\text{age}}.
\end{aligned}$$

In other words, we fitted a separate function over time for each treatment, for each risk group and for each sex. For the age effect, we estimated a bivariate interaction surface depending on age and the survival time.
In analogy to the Cox model, the CTM was extended to include interaction terms. Nevertheless, we always consider interactions between the survival time and the explanatory variables in CTMs. Therefore, the three-time interaction term between treatment, risk group and sex cannot currently be considered in CTMs. Furthermore, the number of two-time interactions should be restricted, which is why we chose to consider only the most interesting interaction between treatment and risk group:

$$h(t_\iota|\boldsymbol{x}) = h_{\text{tr:risk}}(t_\iota|\text{tr:risk}) + h_{\text{sex}}(t_\iota|\text{sex}) + h_{\text{age}}(t_\iota|\text{age}).$$

In contrast to the previous CTM, where three separate functions over time were estimated for each treatment and for each risk group, respectively, we estimated nine separate functions over time for all treatment–risk group combinations. By including the treatment–risk group interaction, we investigated whether different treatments should be considered depending on the specific risk group.

As a further comparative method, we analysed the CML data using conditional random forests. This nonparametric method is also able to relax the PH assumption and is able to consider interactions between the explanatory variables. More precisely, the method grew a survival tree by searching for significant split points in the explanatory variables treatment, risk group, sex and age. The estimated survival probabilities for the patients were obtained afterwards based on conditional Kaplan-Meier estimators for the observations in the final leaves. The bootstrap aggregation of conditional survival trees, which is performed when using conditional random forests as well, results in stable predictions of survival probabilities (Hothorn et al., 2004).

5.4.2. Model evaluation

In the previous section, we described the estimation of a Cox model, a Cox model with interactions, a CTM, a CTM with interactions, and conditional random forests. The evaluation of the five different models served two main goals. First, we were interested in the validity of the PH assumption. The PH assumption could be checked by comparing the performance of the Cox models to the performance of CTMs and conditional random forests, as the PH assumption was relaxed in the latter two approaches. Moreover, we were interested in possible interactions between the explanatory variables treatment, risk group, sex and age. Thereby, all possible interactions could be considered in conditional random forests. All interactions between the *categorical* explanatory variables treatment, risk group and sex were considered in the Cox model with interactions. Owing to the higher flexibility, we only considered the treatment – risk group interaction in the CTM with interactions. Hence, the importance of interactions could be investigated by comparing the models with interactions to their counterparts without interactions, and by comparing the models with interactions among each other.

The model performance was quantified by calculating the out-of-sample censored log score given in Equation (5.9). For the evaluation we used the following procedure:

1. Generate $B = 100$ bootstrap samples by randomly sampling $n = 507$ observations with replacement from the patients in the CML data set. The resulting data sets are *estimation data sets.*

2. The corresponding *evaluation data sets* consist of all observations that have not been selected for the estimation data set.

3. For each bootstrap sample $b = 1, \dots, 100$:

   a) Estimate the five different models based on the estimation data set.

   b) Predict the survival probabilities for the patients in the evaluation data set over a grid of time points.

   c) Calculate the out-of-sample censored log score (Equation (5.9)) based on the predicted survival probabilities, the grid points over time and the inverse probability of censoring weights. Separate out-of-sample censored log scores result for the five model approaches.

4. Compare the $B = 100$ out-of-sample censored log scores for the five model approaches.

Two important characteristics of the above procedure have to be noted. The inverse probability of censoring weights (IPCWs) and a grid of time points were needed when calculating the censored log score. Thereby, the grid of time points was fixed for all bootstrap data sets and consisted of all event and censoring times of the CML patients. The IPCWs were calculated beforehand for *all* patients in the CML data set over the grid of time points defined above. When we selected the patients for the estimation and the evaluation data set, we also selected their respective IPCWs.

The boxplots of the out-of-sample censored log scores for the five different models revealed that the PH assumption is problematic for the CML data when only the main effects treatment, risk group, sex and age are considered (Figure 5.10). The CTM, the CTM with interactions, and conditional random forests (*i.e.* the models that relax the PH assumption)

showed lower out-of-sample censored log scores than the Cox model. Nevertheless, the non-proportional hazards of the main effects seem to be induced by disregarding interaction terms, as the Cox model with interaction terms performed as well as the CTM, which ignored all interaction terms but assumed non-proportional hazards. Nevertheless, the out-of-sample censored log scores for the CTM, the CTM with interactions and conditional random forests were similar. Hence, the inclusion of the treatment–risk group interaction in the CTM did not lead to model improvement. All interactions between explanatory variables could be considered in conditional random forests, but the model's predictive performance was not superior. Hence, the inclusion of interaction terms is unimportant for models that allow for non-proportional hazards.
Through comparisons of models including CTMs, we found that the PH assumption is not violated for a Cox model with interactions. Hence, the best model to analyse the CML data is a Cox model that includes interaction terms between the categorical main effects.

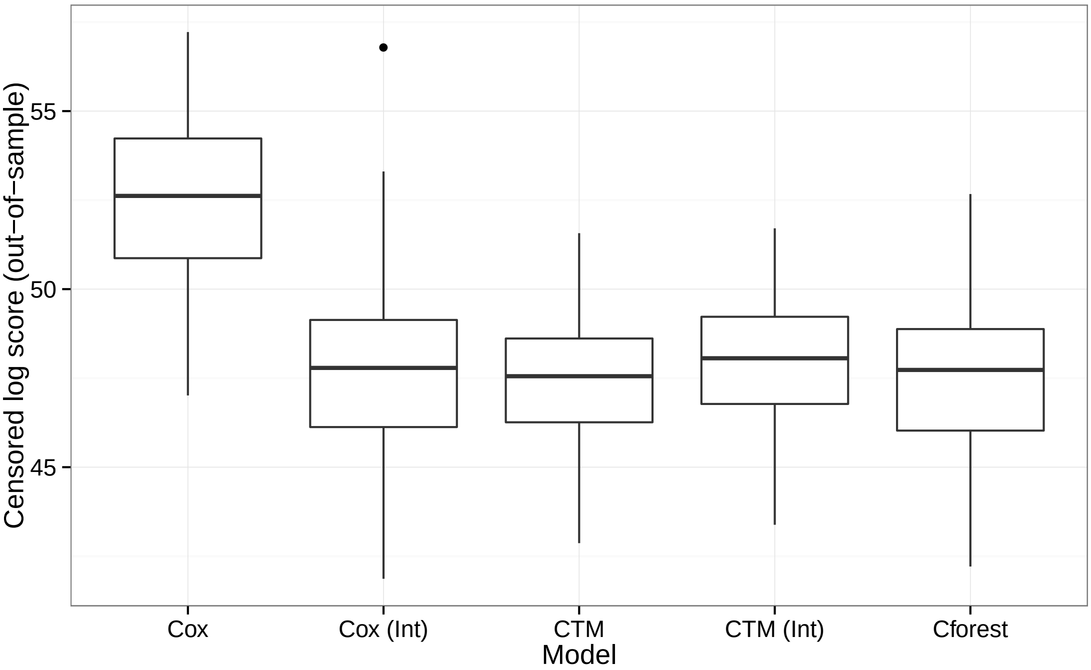


Figure 5.10.: Out-of-sample censored log scores for the Cox model (Cox), the Cox model with interactions (Cox (Int)) the CTM (CTM), the CTM with interactions (CTM (Int)), and conditional random forests (Cforest) for 100 bootstrap evaluation data sets.

5.5. Discussion

The direct estimation of the survivor function in survival data analysis is of special interest, as the reliable prediction of patient-specific survivor functions allows a better prognosis of the course of disease (Mackillop and Quirt, 1997). We propose the use of conditional transformation models (CTMs) to directly estimate the conditional survivor function of the survival times given a set of patient characteristics.

The well-known Cox model is the regression model most commonly used in survival analysis (Cox, 1972). One important restriction of the Cox model is the proportional hazards assumption. Of course, several strategies deal with or identify non-proportional hazards for some of the explanatory variables. For example, if non-proportional hazards for a categorical variable are identified, the estimation of a stratified Cox model with separate baseline hazard functions for the subgroups is frequently used. Speculation about the validity of the proportional hazards assumption in the Cox model becomes superfluous when the CTM approach is used, because the proportional hazards assumption is relaxed and can be checked easily by graphic comparisons.

In our simulation, we investigated the performance of the CTM in cases of proportional hazards and non-proportional hazards and compared the performance to that of the CLTM, the (ordinary or stratified) Cox model, the Kaplan-Meier estimator, and conditional random forests. We measured the performance in terms of the correspondence of true and estimated survival probabilities for new observations. In the simulation settings with informative binary treatment group and non-informative continuous explanatory variable, the CTM was able to keep up with the alternative methods in the case of proportional hazards. In the case of non-proportional hazards, the CTM clearly outperformed the ordinary Cox model and the CLTM and delivered results equally as good as those of the stratified Cox model, the Kaplan-Meier estimator, and conditional random forests. In the simulation settings with informative binary treatment group and informative continuous explanatory variable, the CTM performed almost as well as the ordinary Cox model and the CLTM in the proportional hazards setting. In the non-proportional hazards setting, the CTM outperformed all alternative models, as it is the only method that was able to consider non-proportionality induced non-linearly by a continuous explanatory variable. One further advantage of the CTM was that owing to the imposed smoothness penalty, smooth estimated survival curves resulted, which is more realistic than the step functions resulting from the Cox model, the Kaplan-Meier estimator, and conditional random forests. Moreover, the results of the simulation study showed that the CTM can handle up to 50% of right-censored observations without heavier losses in the quality of the resulting estimates compared to the alternative approaches.

Furthermore, we used the CTM approach to analyse survival times of patients suffering from chronic myelogenous leukaemia to check the PH assumption that has been implied when using Cox models in the past. Furthermore, we were interested in the importance of interactions between the considered explanatory variables. Therefore, the out-of-sample performances of a Cox model, a Cox model with interactions, a CTM, a CTM with interactions, and conditional random forests were compared. Our analysis revealed that the violated PH assumption for the main effects treatment, risk group, sex, and age was mainly induced by ignoring important interactions between the main effects. Furthermore, we would like to stress that models were checked *without* an extensive analysis of residuals.

The handling of right-censored observations is a main topic in survival analysis. In CTMs, the IPCW approach has been used to account for right-censored observations. The integrated Brier score or log score for right-censored observations are well-established scoring rules for model assessment and comparison, but, to the best of our knowledge, they have not yet been used as risk functions for model estimation. In the IPCW approach, the observations are reweighted by the inverse probability of remaining uncensored up to a specific time point. In CTMs, this probability is calculated in terms of the marginal Kaplan-Meier estimator of the censoring distribution. Hence, the weights are calculated based on observed data and, more importantly, it is assumed that the censoring mechanism does not depend on any explanatory variables. Especially the dependency of the censoring distribution on (some of) the explanatory variables would be a worthwhile extension and needs further investigation (Gerds and Schumacher, 2006). Nevertheless, Hothorn et al. (2014) showed the consistency of the conditional transformation function h in CTMs, which transfers to CTMs for survival data as we only adapted the weighting scheme to account for right censoring. Mackenzie (2012) previously estimated survival curves with dependent left-truncated data using Cox's model and inverse probability weighting. Thus, it would be interesting whether and how the suggested approach extends to left-truncated or interval-censored data.

Basically, three main assumptions are made when estimating CTMs for survival data. First, by assuming that the transformation function h exists, we assume that there is a monotone transformation from the unknown survival time distribution to the link function F. Second, h is decomposed additively into partial transformation functions, whereby additivity on the scale of the transformation function is assumed. Third, the event times and the right-censoring times are assumed to be independent, which is a strong but common assumption in survival data analysis. The data analyst should be aware of these model assumptions as they might be violated.

6. Empirical evaluation of likelihood-based CTMs

The content of this chapter is based on Möst and Hothorn (2014).

As a proof of concept, we estimated likelihood-based C(L)TMs (Section 3.2) for uncensored responses, and compared their performance to the performance of a standard regression model. More precisely, we compared some of the CLTMs proposed in Section 2.2.2 to a more flexible CTM, and to the Cox model in three simulation settings. We considered uncensored survival times $T_1, \dots, T_N \geq 0$ throughout, and the conditional survival time distribution was influenced by a continuous explanatory variable x in each simulation setting. In Simulation 1, we considered Weibull distributed survival times that followed the PH assumption. The transformation function in Simulation 1 was extended to a non-proportional hazards setting in Simulation 2. As the survival time distribution function was unknown in Simulation 2, we furthermore considered a non-proportional hazards setting with Weibull distributed survival times in Simulation 3. In order to test two of the proposed estimation strategies in Section 3.2.2, CLTMs were parametrised using fractional polynomials and using T-splines.

6.1. Simulation study setup

Simulation 1. In the first simulation setting, we considered Weibull distributed survival times satisfying the PH assumption. The distribution function for Weibull distributed survival times is

$$F(t) = 1 - \exp(-b^{-c} \cdot t^{c}), \tag{6.1}$$

where b and c denote the scale and the shape parameter, respectively. The choice of parameters b and c determines whether proportional or non-proportional hazards result. The PH assumption is fulfilled if the explanatory variables influence only the scale parameter b and the shape parameter c is fixed. If the shape parameter c is influenced by the explanatory variables, the PH assumption is violated.

Hence, we let the scale parameter b depend on the continuous explanatory variable x, $b = \exp(0.5 \cdot x)$, and chose a fixed shape parameter $c = 3$. The corresponding conditional

Weibull distribution function (Equation (6.1)) can be rewritten in terms of a conditional transformation model:

$$\begin{aligned}\mathbb{P}(T \leq t|X = x) &= 1 - \exp(-\exp(0.5 \cdot x)^{-3} \cdot t^3) \\ &= 1 - \exp(-\exp(-1.5 \cdot x + 3 \cdot \log(t))) \\ &= \mathcal{M}(h_T(t) + \beta_0 \cdot x) = \mathcal{M}(h(t|x)), \end{aligned} \tag{6.2}$$

with survival time transformation $h_T(t) = 3 \cdot \log(t)$, regression coefficient $\beta_0 = -1.5$, and $\mathcal{M}$ denotes the distribution function of the minimum-extreme value distribution. Hence, the resulting CLTM is a classical linear transformation model CLTM C (Equation (2.14)), and the PH assumption is fulfilled because there is no interaction term between x and t in Equation (6.2). We generated $B = 100$ data sets with $N = 500$ observations using the following procedure: First, we defined an equidistant grid $x = (x_1, \ldots, x_{500})$ on the interval $[1, 2]$. Afterwards, we sampled randomly Weibull distributed survival times with scale parameters $b_i = \exp(0.5 \cdot x_i)$, $i = 1, \ldots, 500$, and fixed shape parameter $c = 3$. The procedure was repeated 100 times, *i.e.* all data sets had equal x-values but different survival times $T_1, \ldots, T_{500}$.

Simulation 2. To generate a non-proportional hazards setting, we extended the conditional transformation function of Simulation 1 (Equation (6.2)) by a linear interaction term between t and x:

$$h(t|x) = 3 \cdot \log(t) - 1.5 \cdot x + 2 \cdot x \cdot t, \tag{6.3}$$

i.e. the survival time transformation $h_T(t)$ and β_0 remained unchanged, and the coefficient of the linear interaction term was set to $\beta_1 = 2$. The PH assumption was no longer valid due to the linear interaction term between t and x, *i.e.* the influence of the explanatory variable x varied over time. Moreover, the associated survival times were no longer Weibull distributed but followed some unknown distribution function. Therefore, the survival times had to be simulated using root-finding techniques. The conditional distribution function of the survival times is

$$\mathbb{P}(T \leq t|X = x) = \mathcal{M}(3 \cdot \log(t) - 1.5 \cdot x + 2 \cdot x \cdot t) \sim U, \tag{6.4}$$

where U is uniformly distributed on $[0, 1]$. $B = 100$ data sets with $N = 1,000$ observations were generated by first defining the explanatory variable values $x_1, \ldots, x_{1,000}$ via an equidistant grid on the interval $[1, 2]$. The survival times $T_1, \ldots, T_{1,000}$ were simulated by first sampling uniformly distributed random variables $U_1, \ldots, U_{1,000} \sim U[0, 1]$, and solving the equation (which is based on Equation (6.4)) for T_i, $i = 1, \ldots, 1,000$, afterwards:

$$\mathcal{M}^{-1}(U_i) + 1.5 \cdot x_i - 3 \cdot \log(T_i) - 2 \cdot x_i \cdot T_i \overset{!}{=} 0,$$

where $\mathcal{M}^{-1}$ denotes the quantile function of the minimum-extreme value distribution.

Simulation 3. As the survival time distribution was unknown in Simulation 2, we simulated another non-proportional hazards setting with Weibull distributed survival times. Hence, we chose a fixed scale parameter $b = \exp(2)$, and the shape parameter depended linearly on x, $c = 2 \cdot x$. As the continuous explanatory variable x was chosen equidistantly on $[1, 3]$, the shape parameter c varied on $[2, 6]$. The conditional Weibull distribution function (based on Equation (6.1)) can be written in terms of a conditional transformation model:

$$\begin{aligned}\mathbb{P}(T \leq t|X = x) &= 1 - \exp(-\exp(2)^{-2\cdot x} \cdot t^{2\cdot x}) \\ &= 1 - \exp(-\exp(-4 \cdot x + 2 \cdot x \cdot \log(t))) \\ &= \mathcal{M}(\beta_0 \cdot x + \beta_1 \cdot x \cdot \log(t)),\end{aligned} \tag{6.5}$$

with regression coefficients $\beta_0 = -4$ and $\beta_1 = 2$. Due to the interaction term between x and t, the PH assumption was violated. Moreover, there was no survival time transformation $h_T(t)$ in this simulation setting, and the interaction term between x and t was not linear but more complex because of the log-transformed survival time. Hence, the model presented in Equation (6.5) belongs to the more flexible model class of CTMs. We generated $B = 100$ data sets with $N = 2,000$ observations by first defining an equidistant grid $x = (x_1, \ldots, x_{2,000})$ on the interval $[1, 3]$. The corresponding Weibull distributed survival times were sampled randomly afterwards with fixed scale parameter $b = \exp(2)$ and shape parameters $c_i = 2 \cdot x_i$, $i = 1, \ldots, 2,000$.

6.2. Model estimation

For model analysis, we used four CLTMs with different model complexities and parametrisations, a more flexible CTM, and the Cox model. The corresponding six estimation procedures are described in more detail below.

6.2.1. CLTM: Parametrisation using fractional polynomials

The survival time transformation function $h_T(t)$ was parametrised using fractional polynomials of degree 1 (Section 3.2.2). In model CLTM C.1 (which is a special case of model CLTM C (Equation (2.14))), the linear interaction term between x and t was ignored, whereas the linear interaction term was included in model CLTM E.1 (which is a special case of model CLTM E (Equation (2.16))):

- **CLTM C.1**:

$$\begin{aligned}h(t|x) &= h_T(t) + \beta_0 \cdot x = \\ &\alpha_0 + \alpha_1 \cdot t^{-2} + \alpha_2 \cdot t^{-1} + \alpha_3 \cdot t^{-0.5} + \alpha_4 \cdot \log(t) + \alpha_5 \cdot \sqrt{t} + \alpha_6 \cdot t + \\ &\alpha_7 \cdot t^2 + \alpha_8 \cdot t^3 + \beta_0 \cdot x.\end{aligned} \tag{6.6}$$

- **CLTM E.1**:

$$\begin{aligned} h(t|x) &= h_T(t) + \beta_0 \cdot x + \beta_1 \cdot x \cdot t = \\ & \alpha_0 + \alpha_1 \cdot t^{-2} + \alpha_2 \cdot t^{-1} + \alpha_3 \cdot t^{-0.5} + \alpha_4 \cdot \log(t) + \alpha_5 \cdot \sqrt{t} + \alpha_6 \cdot t + \\ & \alpha_7 \cdot t^2 + \alpha_8 \cdot t^3 + \beta_0 \cdot x + \beta_1 \cdot x \cdot t. \end{aligned} \tag{6.7}$$

Estimation of CLTM C.1 involved the estimation of the parameters $\boldsymbol{\alpha} = (\alpha_0 \dots \alpha_8)^\top$ and β_0, and the estimation of CLTM E.1 additionally involved the estimation of β_1. The conditional transformation functions (Equation (6.6) and Equation (6.7)) yielded the uncensored log-likelihoods (Equation (3.2)):

- **CLTM C.1**: $l(\boldsymbol{\alpha}, \beta_0) = \sum_{i=1}^{N} \log(f(h_T(t_i) + \beta_0 \cdot x_i)) + \log(h'_T(t_i))$

- **CLTM E.1**: $l(\boldsymbol{\alpha}, \beta_0, \beta_1) = \sum_{i=1}^{N} \log(f(h_T(t_i) + \beta_0 \cdot x_i + \beta_1 \cdot x_i \cdot t_i)) + \log(h'_T(t_i) + \beta_1 \cdot x_i)$.

Following the definitions for linear transformation models (Equation (2.2)), the survival time transformation function $h_T(t)$ has to be monotonically increasing, *i.e.* $h'_T(t) > 0$. Furthermore, the first derivative of the conditional transformation function, $h'(t|x)$, needs to be strictly positive because it is log-transformed in the log-likelihood function. This led to the following linear constraints for models CLTM C.1 and CLTM E.1 that had to be considered during estimation:

CLTM C.1:

$$\begin{pmatrix} 0 & -2t_1^{-3} & -t_1^{-2} & -\frac{1}{2}t_1^{-\frac{3}{2}} & \frac{1}{t_1} & \frac{1}{2}t_1^{-\frac{1}{2}} & t_1 & 2t_1 & 3t_1^2 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & -2t_N^{-3} & -t_N^{-2} & -\frac{1}{2}t_N^{-\frac{3}{2}} & \frac{1}{t_N} & \frac{1}{2}t_N^{-\frac{1}{2}} & t_N & 2t_N & 3t_N^2 & 0 \end{pmatrix} \cdot \begin{pmatrix} \alpha_0 \\ \vdots \\ \alpha_8 \\ \beta_0 \end{pmatrix} > \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

CLTM E.1:

$$\begin{pmatrix} 0 & -2t_1^{-3} & -t_1^{-2} & -\frac{1}{2}t_1^{-\frac{3}{2}} & \frac{1}{t_1} & \frac{1}{2}t_1^{-\frac{1}{2}} & t_1 & 2t_1 & 3t_1^2 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & -2t_N^{-3} & -t_N^{-2} & -\frac{1}{2}t_N^{-\frac{3}{2}} & \frac{1}{t_N} & \frac{1}{2}t_N^{-\frac{1}{2}} & t_N & 2t_N & 3t_N^2 & 0 & 0 \\ 0 & -2t_1^{-3} & -t_1^{-2} & -\frac{1}{2}t_1^{-\frac{3}{2}} & \frac{1}{t_1} & \frac{1}{2}t_1^{-\frac{1}{2}} & t_1 & 2t_1 & 3t_1^2 & 0 & x_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & -2t_N^{-3} & -t_N^{-2} & -\frac{1}{2}t_N^{-\frac{3}{2}} & \frac{1}{t_N} & \frac{1}{2}t_N^{-\frac{1}{2}} & t_N & 2t_N & 3t_N^2 & 0 & x_N \end{pmatrix} \cdot \begin{pmatrix} \alpha_0 \\ \vdots \\ \alpha_8 \\ \beta_0 \\ \beta_1 \end{pmatrix} > \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

As the parametrisation of models CLTM C.1 (10 parameters) and CLTM E.1 (11 parameters) was parsimonious, estimation could be based on the maximisation of the full log-likelihood function using an optimisation algorithm that is able to consider linear constraints. We used the `constrOptim`-function from the R base-package **stats** (R Core Team, 2014) for optimisation.

6.2.2. CLTM: Parametrisation using T-splines

Alternatively, the monotonically increasing survival time transformation function $h_T(t)$ can be parametrised using T-splines (Beliakov, 2000, Section 3.2.2). In short, we parametrised

$$h_T(t) = B_t \cdot \boldsymbol{\alpha},$$

whereby $B_t \in \mathbb{R}^{N \times k}$ denotes the design matrix of the T-spline basis functions, and $\boldsymbol{\alpha} \in \mathbb{R}^k$ denotes the vector of the corresponding basis coefficients. The estimation of $h_T(t)$ involved estimating $k = 24$ parameters, $\alpha_1, \dots, \alpha_{24}$. Furthermore, T-splines inherit the B-spline characteristic that the first derivative can be displayed using the same vector of basis coefficients but adapted basis functions (Section 3.2.2):

$$h'_T(t) = B'_t \cdot \boldsymbol{\alpha}.$$

To guarantee smooth function estimates for $h_T(t)$ and $h'_T(t)$, we included the penalty matrices K_2 and K_3 based on second and third differences (Section 3.2.2) into the uncensored log-likelihood function (Simpkin and Newell, 2013, Equation (3.2)):

- **CLTM C.2**: $l_p(\boldsymbol{\alpha}, \beta_0) = \sum_{i=1}^{N} \log(f(B_t(t_i) \cdot \boldsymbol{\alpha} + \beta_0 \cdot x_i)) + \log(B'_t(t_i) \cdot \boldsymbol{\alpha})$ $-\frac{\lambda_2}{2} \cdot \boldsymbol{\alpha}^\top K_2 \boldsymbol{\alpha} - \frac{\lambda_3}{2} \cdot \boldsymbol{\alpha}^\top K_3 \boldsymbol{\alpha},$
- **CLTM E.2**: $l_p(\boldsymbol{\alpha}, \beta_0, \beta_1) = \sum_{i=1}^{N} \log(f(B_t(t_i) \cdot \boldsymbol{\alpha} + \beta_0 \cdot x_i + \beta_1 \cdot x_i \cdot t_i))$ $+ \log(B'_t(t_i) \cdot \boldsymbol{\alpha} + \beta_1 \cdot x_i) - \frac{\lambda_2}{2} \cdot \boldsymbol{\alpha}^\top K_2 \boldsymbol{\alpha} - \frac{\lambda_3}{2} \cdot \boldsymbol{\alpha}^\top K_3 \boldsymbol{\alpha},$

where the linear interaction term was ignored in CLTM C.2, and the linear interaction term was included in CLTM E.2.

To guarantee a monotonically increasing function estimate $\hat{h}_T(t)$, the T-spline coefficients $\alpha_2, \dots, \alpha_{24}$ have to be positive (α_1 remains unrestricted), and $h'(t|x)$ needs to be positive

in addition. This resulted in the linear constraints:
CLTM C.2:

$$\begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 & 0 \\ & & & & \vdots \\ & & B'_t & & \vdots \\ & & & & 0 \end{pmatrix} \cdot \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_{24} \\ \beta_0 \end{pmatrix} > \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

CLTM E.2:

$$\begin{pmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & 1 & 0 & 0 \\ & & & & \vdots & x_1 \\ & & B'_t & & \vdots & \vdots \\ & & & & 0 & x_N \end{pmatrix} \cdot \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_{24} \\ \beta_0 \\ \beta_1 \end{pmatrix} > \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

We used the `constrOptim`-function from the R base-package **stats** (R Core Team, 2014) for optimisation under linear constraints.

6.2.3. CTM

For reasons of comparison, we also fitted a CTM (Equation (1.2)), where the conditional transformation function is more flexible compared to CLTMs. In general, the conditional transformation function is able to display more complex functional relationships between t and x. Therefore, the conditional transformation function is estimated in terms of a bivariate smooth surface depending on t and x:

$$h(t|x) = (\boldsymbol{b}_x(x)^\top \otimes \boldsymbol{b}_T(t)^\top) \cdot \boldsymbol{\alpha} = B_{t,x} \cdot \boldsymbol{\alpha}.$$

Thereby, $\boldsymbol{b}_x(x)$ denotes a set of B-spline basis functions for the continuous explanatory variable x, and $\boldsymbol{b}_T(t)$ denotes a set of T-spline basis functions for t. Both sets of basis functions are connected via the Kronecker product, whereby an interaction surface is established (for additional information on the estimation of CTMs, we refer to Hothorn et al., 2014). In short, $B_{t,x}$ denotes the design matrix of the interaction surface. The interaction surface is monotonically increasing in the direction of t (due to the T-spline basis functions) and unrestricted in the direction of x. Both sets of basis functions depended on 24 basis coefficients, what resulted in $24^2 = 576$ basis coefficients that had to be estimated. Furthermore, the first derivative $h'(t|x)$ could be established via

$$h'(t|x) = (\boldsymbol{b}_x(x)^\top \otimes \boldsymbol{b}'_T(t)^\top) \cdot \boldsymbol{\alpha} = B'_{t,x} \cdot \boldsymbol{\alpha},$$

where $\boldsymbol{b}'_T(t)$ denote the first derivatives of the T-spline basis functions. To guarantee smoothness of $h(t|x)$ and $h'(t|x)$, we defined the penalty matrices K_2 and K_3 (Hothorn et al., 2014):

$$K_2 = (K_x \otimes I_{K_{T2}} + I_{K_x} \otimes K_{T2}) \text{ and } K_3 = (K_x \otimes I_{K_{T3}} + I_{K_x} \otimes K_{T3}),$$

where K_x is the penalty matrix based on second differences for $\boldsymbol{b}_x$, K_{T2} is the penalty matrix based on second differences and K_{T3} is the penalty matrix based on third differences for $\boldsymbol{b}_T$, respectively, and I denotes the identity matrix. The penalty matrices K_2 and K_3 were associated with smoothing parameters $\lambda_2 \geq 0$ and $\lambda_3 \geq 0$. All in all, the corresponding penalised uncensored log-likelihood function (Equation (3.2)) is

$$l_p(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \log(f(B_{t,x}(t_i, x_i) \cdot \boldsymbol{\alpha})) + \log(B'_{t,x}(t_i, x_i) \cdot \boldsymbol{\alpha}) - \frac{\lambda_2}{2} \cdot \boldsymbol{\alpha}^\top K_2 \boldsymbol{\alpha} - \frac{\lambda_3}{2} \cdot \boldsymbol{\alpha}^\top K_3 \boldsymbol{\alpha}.$$

In analogy to the previous models, there were certain linear constraints that had to be considered during estimation. First, the bivariate surface $B_{t,x}$ had to be monotonically increasing in the direction of t. As the basis functions $\boldsymbol{b}_T$ are T-spline basis functions, all basis coefficients $\alpha_1, \dots, \alpha_{576}$ needed to be positive except the first, the 25th, the 49th coefficient, etc. As the surface is unrestricted in the direction of x, every first coefficient in x-direction was excluded from the constraints. Moreover, the first derivative $B'_{t,x}$ had to be positive, and all necessary linear constraints could be summarised to:

$$\begin{pmatrix} 0 & 1 & 0 & \dots & & & & & & \\ \vdots & 0 & \ddots & 0 & \dots & & & & & \\ & \vdots & 0 & 1 & 0 & \dots & & & & \\ & \vdots & \vdots & & 0 & 1 & 0 & \dots & & \\ & & & & & 0 & \ddots & 0 & \dots & \dots \\ & & & & & & & 1 & 0 & \dots \\ & & & & & B'_{t,x} & & & & \end{pmatrix} \cdot \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_{576} \end{pmatrix} > \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Again, constrained optimisation was performed using the `constrOptim`-function from the R base-package **stats** (R Core Team, 2014).

6.2.4. Cox model

The Cox model is a semiparametric alternative to CLTMs, which relies on the PH assumption. The conditional hazard function was specified via

$$\lambda(t|x) = \lambda_0(t) \exp(\beta_0 \cdot x),$$

where $\lambda_0(t)$ denotes the baseline hazard function and the explanatory variable x had a linear influence. Cox models were estimated using the `coxph`-function from the R add-on package **survival** (Therneau, 2013).

Simulation 1. In the first simulation setting, we estimated CLTMs CLTM C.1, CLTM E.1, CLTM C.2 and CLTM E.2, a CTM and the Cox model.

Thereby, we expected models CLTM C.1 and CLTM C.2 to perform slightly better than models CLTM E.1 and CLTM E.2, respectively, as models CLTM E.1 and CLTM E.2 include the superfluous liner interaction term between t and x. Nevertheless, we expected models CLTM E.1 and CLTM E.2 to perform acceptably well, and the parameter β_1 of the superfluous interaction term was expected to be estimated close to zero. The comparison of CLTM C.1 to CLTM E.1, and CLTM C.2 to CLTM E.2 is important because we usually do not know the true structure of the transformation function. Therefore, we expected to get acceptable model estimates even if a too complex transformation function was used, and that the true structure of the transformation function was identified by the model. Moreover, we were interested in comparing CLTM C.1 to CLTM C.2, and CLTM E.1 to CLTM E.2, respectively, to compare the parametrisation of $h_T(t)$ using fractional polynomials to a T-spline parametrisation.

The CTM is more complex than the considered CLTMs, but the additional flexibility was superfluous in Simulation 1. Therefore, we expected the CTM to perform slightly worse than all CLTMs.

The Cox model was expected to perform best (and comparably well as CLTM C.1 and CLTM C.2, which offer the same model structure) because the PH assumption was fulfilled in Simulation 1 and hence, it perfectly fitted the given data setting.

Simulation 2. The non-proportional hazards setting in Simulation 2 was estimated using the proposed CLTMs, a CTM, and the Cox model.

As the linear interaction term between t and x was necessary in this simulation setting, we expected CLTM E.1 and CLTM E.2 to perform better than CLTM C.1 and CLTM C.2, respectively. Again, the comparison of a parametrisation of $h_T(t)$ using fractional polynomials to a parametrisation using T-splines was of interest because the parametrisation using fractional polynomials is more parsimonious.

The CTM was expected to perform slightly worse than CLTM E.1 and CLTM E.2 because CLTM E.1 and CLTM E.2 assumed the true structure of the conditional transformation function, and hence, the additional flexibility of the CTM was superfluous. Nevertheless, the CTM was expected to perform better than CLTM C.1 and CLTM C.2 because the necessary linear interaction term was ignored in the CLTMs, whereby proportional hazards were assumed.

The Cox model was assumed to perform as worse as models CLTM C.1 and CLTM C.2 because the PH assumption was violated, and proportional hazards were assumed misleadingly in these models.

Simulation 3. The Weibull distributed survival times with non-proportional hazards were estimated using the proposed CLTMs, a CTM, and the Cox model.

CLTM E.1 and CLTM E.2 were expected to perform ordinarily because the assumed structure of the conditional transformation function did not fit exactly the true structure of the conditional transformation function. Both models assumed a survival time transformation $h_T(t)$ that was not necessary, and they assumed a linear interaction term between t and x but the interaction was more complex. Nevertheless, both models were expected to perform quite well because the nonlinearity of the interaction term was expected to be partly balanced by the estimation of $h_T(t)$.
CLTM C.1, CLTM C.2 and the Cox model were expected to perform comparably, and worse than models CLTM E.1 and CLTM E.2. In all three models the interaction term between t and x was completely ignored, and a superfluous survival time transformation function $h_T(t)$ was included. Thereby, proportional hazards were assumed misleadingly.

The CTM is the only model that is able to estimate the interaction term between x and $\log(t)$ due to its higher flexibility and thus, the CTM was expected to perform best. The comparison of all considered models was expected to indicate that the assumed model structure was too restrictive in all CLTMs, and that a more complex model approach was needed.

6.2.5. Determination of smoothing parameters

The estimation of CLTM C.2, CLTM E.2 and the CTM (*i.e.* the models based on a T-spline parametrisation) involved the determination of the smoothing parameters λ_2 and λ_3. This was a challenging task because the simultaneous determination of a two-dimensional smoothing parameter is associated with a computationally extensive grid search. Nevertheless, Simpkin and Newell (2013) found no eminent difference in the performance of a sequential and a simultaneous selection of smoothing parameters for derivative estimation. Therefore, the authors recommend a sequential selection of the smoothing parameters due to reasons of computational efficiency.

Hence, our smoothing parameter selection process followed a sequential determination of the smoothing parameters, *i.e.* we first searched for the optimal λ_2 while keeping λ_3 fixed, and searched for λ_3 while keeping λ_2 at its optimal value afterwards:

1. Generate $B = 25$ training data sets and one evaluation data set with $N = 5,000$ observations using the respective data generating process in Section 6.1. Due to model

complexity, the training and evaluation data sets consisted of $N = 1,000$ observations for model CTM.

2. Determination of λ_2 (while keeping λ_3 fixed, *e.g.*, $\lambda_3 = 50$):

   a) Define a grid of possible values for λ_2:
   $\mathcal{G}_{\lambda_2} = \{0.001;\ 0.01;\ 0.1;\ 1;\ 10;\ 50;\ 100;\ 500;\ 1,000;\ 5,000\}$.

   b) For each λ_2 in $\mathcal{G}_{\lambda_2}$:

      i. Estimate the respective model for each training data set, what results in estimated models $\widehat{M}_1, \ldots, \widehat{M}_{25}$.

      ii. Predict the survival probabilities for the observations in the evaluation data set using the estimated models $\widehat{M}_1, \ldots, \widehat{M}_{25}$ over a grid of time points consisting of all survival times in the evaluation data set.

      iii. Evaluate the predicted survival probabilities using the uncensored log score (Equation (3.1)):

$$\begin{aligned} LS &= -\frac{1}{N \cdot N} \sum_{i=1}^{N} \sum_{\iota=1}^{N} I(T_i \leq t_\iota) \log(\mathcal{M}(\hat{h}(t_\iota|x_i))) + \\ &\qquad I(T_i > t_\iota) \log(1 - \mathcal{M}(\hat{h}(t_\iota|x_i))), \end{aligned} \tag{6.8}$$

      where $T_1, \ldots, T_{5,000}$ denote the survival times from the evaluation data set, $t_1, \ldots, t_{5,000}$ is the grid of time points consisting of all unique survival times from the evaluation data set, and $\mathcal{M}(\hat{h}(t_\iota|x_i))$ denotes the estimated conditional survival probability for observation i at time point t_ι. This results in 25 out-of-sample uncensored log scores $LS_1, \ldots, LS_{25}$.

      iv. Calculate the mean out-of-sample uncensored log score $\overline{LS} = \frac{1}{25}\sum_{i=1}^{25} LS_i$.

   c) The optimal value for λ_2 is the value in $\mathcal{G}_{\lambda_2}$ with the smallest corresponding mean out-of-sample uncensored log score.

3. Determination of λ_3 (while keeping λ_2 at its optimal value):

   a) Define $\mathcal{G}_{\lambda_3} = \{1;\ 10;\ 100;\ 500;\ 1,000;\ 2,500;\ 5,000;\ 7,500;\ 10,000\}$.

   b) The optimal value of λ_3 is determined in analogy to λ_2 using the same training and evaluation data sets.

6.3. Model evaluation

6.3.1. Estimated regression coefficients and response transformations

The true response transformation function $h_T(t)$ and the true regression coefficients β_0 and β_1 are known in all three simulation settings. Furthermore, the model components are separable in the Cox model and in all CLTMs, whereas the model components are inseparable in the CTM. A graphical approach was used to compare the estimated regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$, and the estimated survival time transformations $\hat{h}_T(t)$ and the first derivatives $\hat{h}'_T(t)$ to their true counterparts for models CLTM C.1, CLTM C.2, CLTM E.1, CLTM E.2, and for the Cox model. Thereby, our aim was to evaluate how well the various models were able to estimate the specific model components.

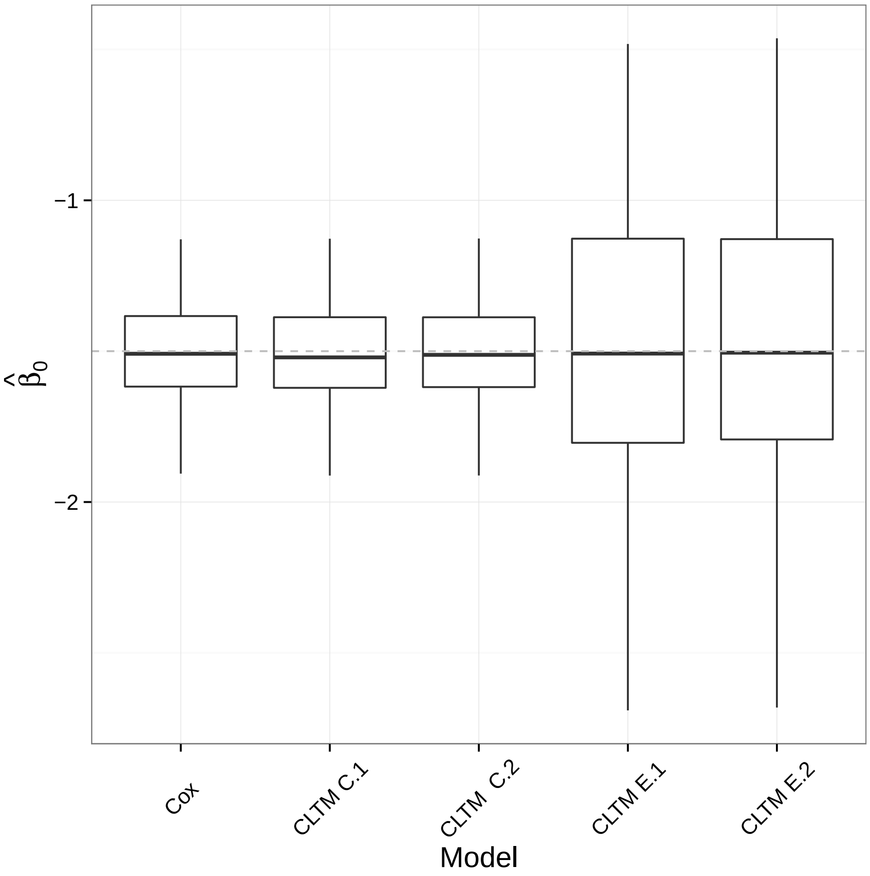


Figure 6.1.: Simulation 1: Boxplots of the estimated regression coefficients $\hat{\beta}_0$ for models CLTM C.1, CLTM C.2, CLTM E.1, CLTM E.2, and the Cox model for $B = 100$ simulated data sets. The dashed reference line indicates the true value $\beta_0 = -1.5$.

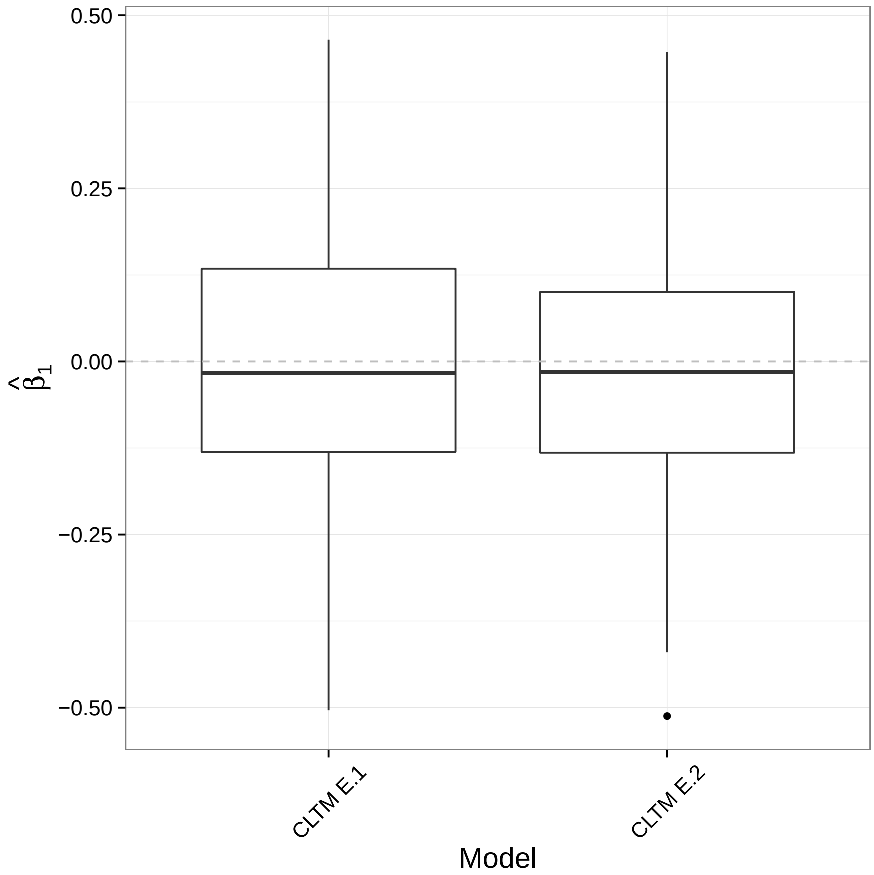


Figure 6.2.: Simulation 1: Boxplots of the estimated regression coefficients $\hat{\beta}_1$ for models CLTM E.1 and CLTM E.2 for $B = 100$ simulated data sets. The dashed reference line indicates the true value $\beta_1 = 0$.

Simulation 1. The regression coefficient β_0 was estimated in all considered models. The estimated coefficients $\hat{\beta}_0$ for the $B = 100$ simulated data sets were very similar for the Cox model and for CLTM C.1 and CLTM C.2 (Figure 6.1). This was not surprising because the same model complexity was assumed in all three models, but the corresponding estimation procedures differed. The Cox model was estimated based on a partial likelihood approach (Cox, 1975), whereas CLTM C.1 and CLTM C.2 were estimated based on a full likelihood approach. The estimates were close to the true value $\beta_0 = -1.5$. The estimated coefficients $\hat{\beta}_0$ resulting from CLTM E.1 and CLTM E.2 varied symmetrically around -1.5, but the variance was higher than in models CLTM C.1 and CLTM C.2 due to the higher complexity of the conditional transformation function $h(t|x)$.

Additionally, the superfluous regression coefficient β_1 was estimated in CLTM E.1 and CLTM E.2. Nevertheless, both models were able to identify the true structure of the conditional transformation function, and the estimated coefficients $\hat{\beta}_1$ varied symmetrically around and were close to zero (Figure 6.2).

The estimated survival time transformation functions $\hat{h}_T(t)$ based on the $B = 100$ simulated

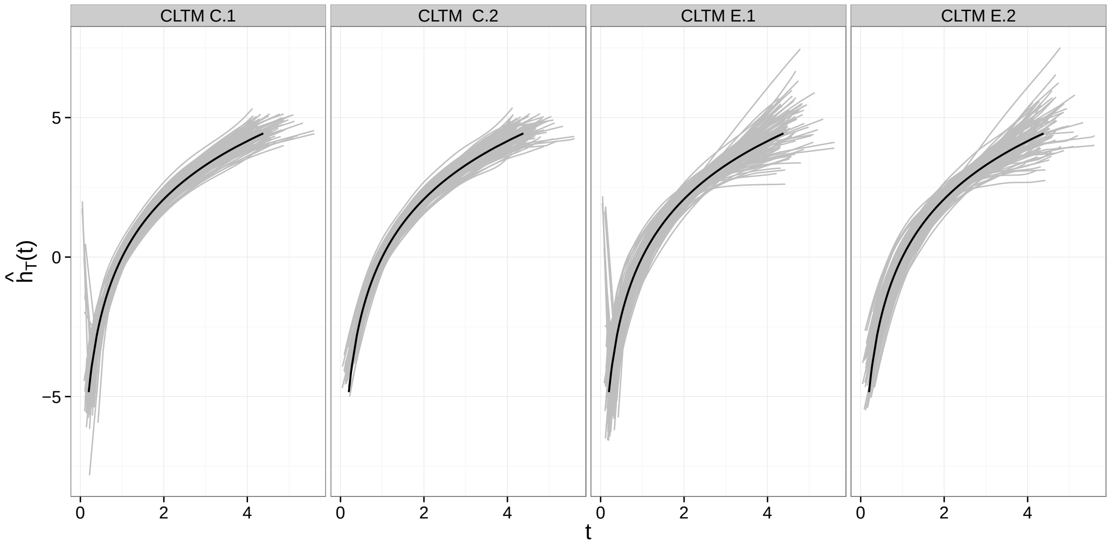


Figure 6.3.: Simulation 1: Estimated survival time transformations $\hat{h}_T(t)$ for $B = 100$ simulated data sets for models CLTM C.1, CLTM C.2, CLTM E.1, and CLTM E.2. The true survival time transformation $h_T(t) = 3 \cdot \log(t)$ is displayed in black.

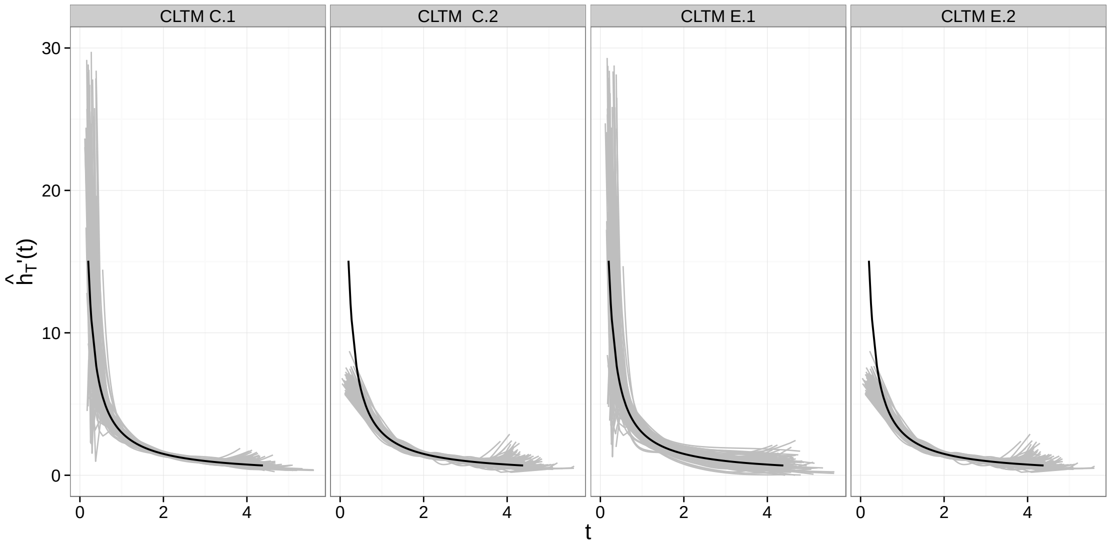


Figure 6.4.: Simulation 1: Estimated first derivatives of the survival time transformations $\hat{h}_T^{\prime}(t)$ for $B = 100$ simulated data sets for models CLTM C.1, CLTM C.2, CLTM E.1, and CLTM E.2. The true first derivative of the survival time transformation $h_T^{\prime}(t) = {}^3\!/_t$ is displayed in black.

data sets resulting for CLTM C.1, CLTM C.2, CLTM E.1 and CLTM E.2 are compared to the true survival time transformation function $h_T(t) = 3 \cdot \log(t)$ in Figure 6.3, and the corresponding estimated first derivatives $\hat{h}_T^{\prime}(t)$ are compared to the true first derivative $h_T^{\prime}(t) = {}^3\!/_t$ in Figure 6.4. For all four models, the estimated survival time transformation functions were very close to the true transformation function. Nevertheless, there occurred some inconsistencies for CLTM C.1 and CLTM E.1 for small survival times. This was most probably due to the fractional polynomials t^{-2} and t^{-1} reaching high values for small survival times. This caused instabilities for values at the lower boundary of the estimated function, which furthermore might indicate lacking flexibility of the fractional polynomials. Hence, the B-spline representation of the survival time transformation is more stable. Identical results could be observed for the first derivative of the survival time transformation.

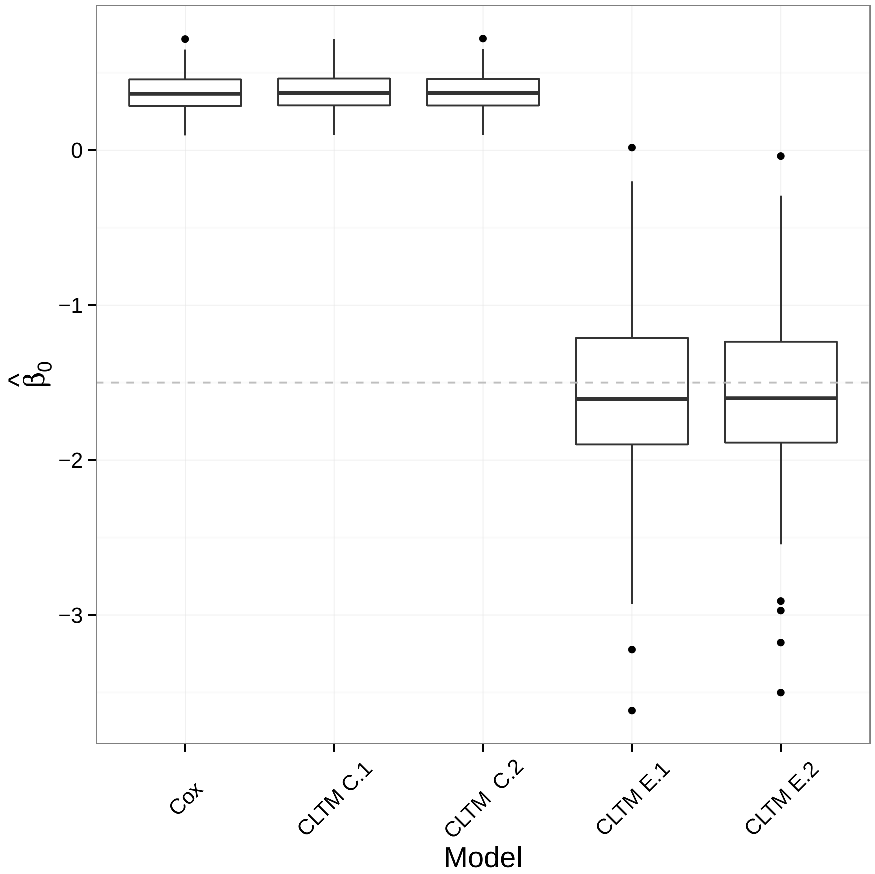


Figure 6.5.: Simulation 2: Boxplots of the estimated regression coefficients $\hat{\beta}_0$ for models CLTM C.1, CLTM C.2, CLTM E.1, CLTM E.2, and the Cox model for $B = 100$ simulated data sets. The dashed reference line indicates the true value $\beta_0 = -1.5$.

Simulation 2. Due to the same underlying model complexity, the estimated regression coefficients $\hat{\beta}_0$ for the $B = 100$ simulated data sets were almost identical for CLTM C.1, CLTM C.2, and the Cox model (Figure 6.5). Nevertheless, the estimates were considerably biased because a PH setting was assumed by the three models, and the linear interaction term between t and x was ignored. In contrast, the estimated coefficients $\hat{\beta}_0$ resulting from CLTM E.1 and CLTM E.2 were only very slightly negatively biased and close to the true

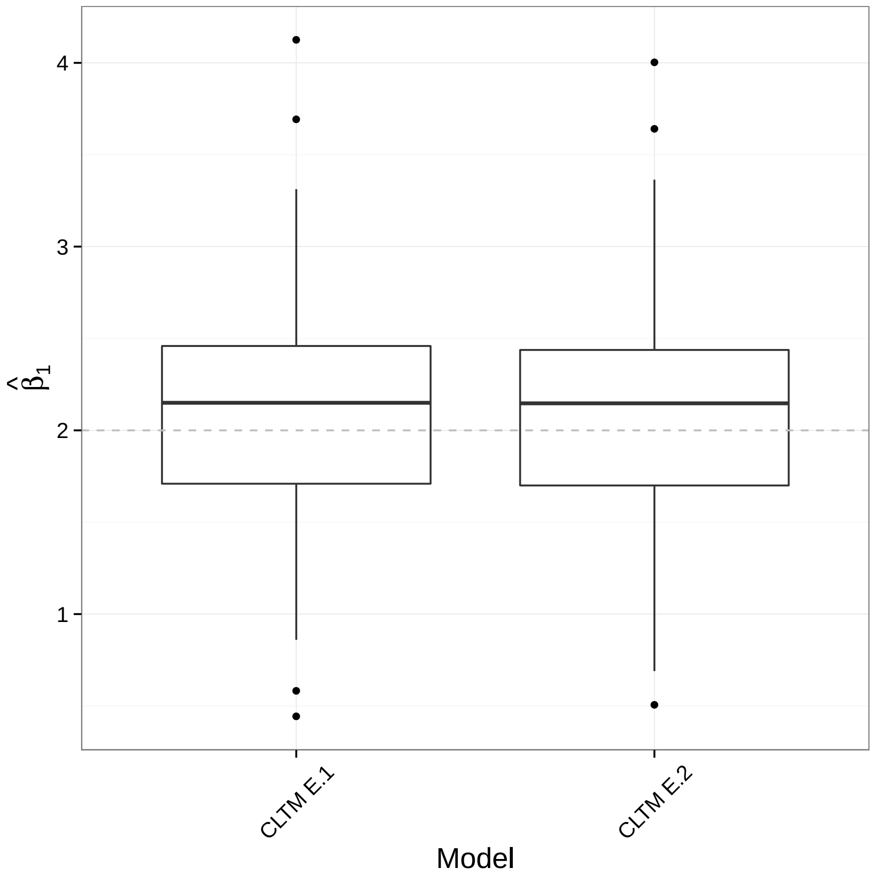


Figure 6.6.: Simulation 2: Boxplots of the estimated regression coefficients $\hat{\beta}_1$ for models CLTM E.1 and CLTM E.2 for $B = 100$ simulated data sets. The dashed reference line indicates the true value $\beta_1 = 2$.

value $\beta_0 = -1.5$ because the linear interaction term was considered. Again, the higher model complexity of CLTM E.1 and CLTM E.2 was associated with more variable estimates. Moreover, the estimated regression coefficients $\hat{\beta}_1$ resulting from CLTM E.1 and CLTM E.2 were relatively close to the true value $\beta_1 = 2$, and only slightly positively biased (Figure 6.6).

Considering the estimated survival time transformation functions $\hat{h}_T(t)$ and their first derivatives $\hat{h}'_T(t)$, CLTM C.1 and CLTM C.2 were not able to estimate the true shape of the survival time transformation properly (Figure 6.7 and Figure 6.8), which was due to the wrong structure of the underlying conditional transformation function. In contrast, CLTM E.1 and CLTM E.2 were able to display the true shape of the survival time transformation, whereby the inconsistencies for the parametrisation using fractional polynomials remained.

Simulation 3. The Cox model and all CLTMs assumed a wrong structure of the conditional transformation function. None of the considered models (except for the more flexible

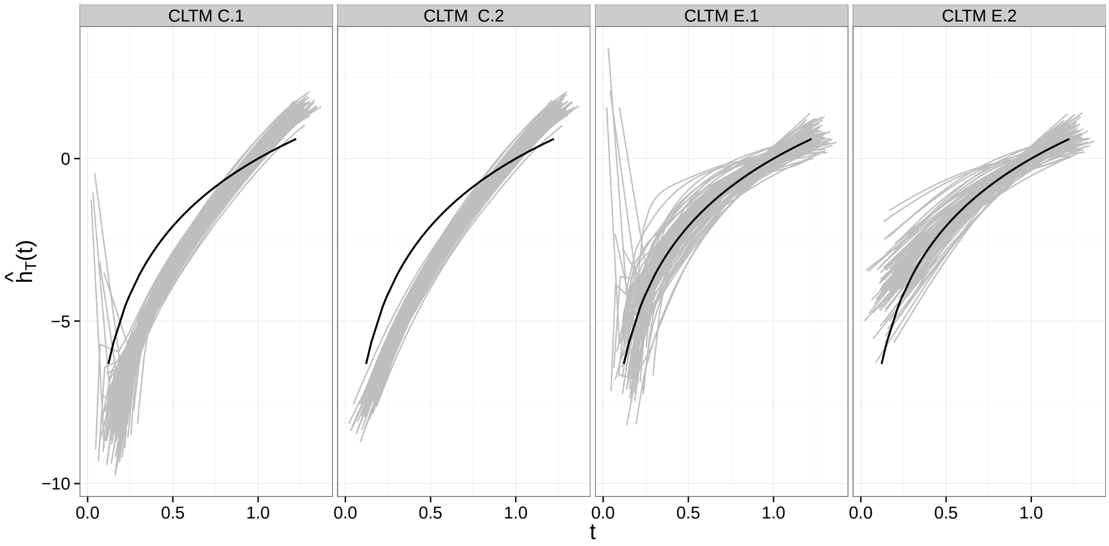


Figure 6.7.: Simulation 2: Estimated survival time transformations $\hat{h}_T(t)$ for $B = 100$ simulated data sets for models CLTM C.1, CLTM C.2, CLTM E.1, and CLTM E.2. The true survival time transformation $h_T(t) = 3 \cdot \log(t)$ is displayed in black.

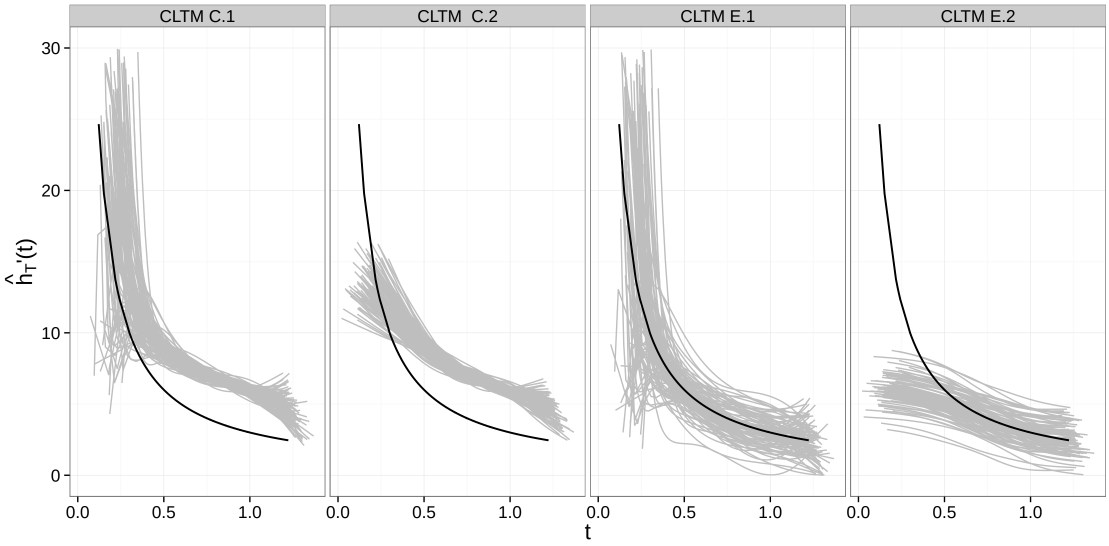


Figure 6.8.: Simulation 2: Estimated first derivatives of the survival time transformations $\hat{h}'_T(t)$ for $B = 100$ simulated data sets for models CLTM C.1, CLTM C.2, CLTM E.1, and CLTM E.2. The true first derivative of the survival time transformation $h'_T(t) = {}^3/_t$ is displayed in black.

CTM) is able to estimate a non-linear interaction between t and x. Therefore, all estimated coefficients $\hat{\beta}_0$ were considerably biased (Figure 6.9). Nevertheless, at least the consideration of a linear interaction seemed to be of advantage because the estimates resulting from CLTM E.1 and CLTM E.2 were less heavily biased. Nevertheless, the estimated regression

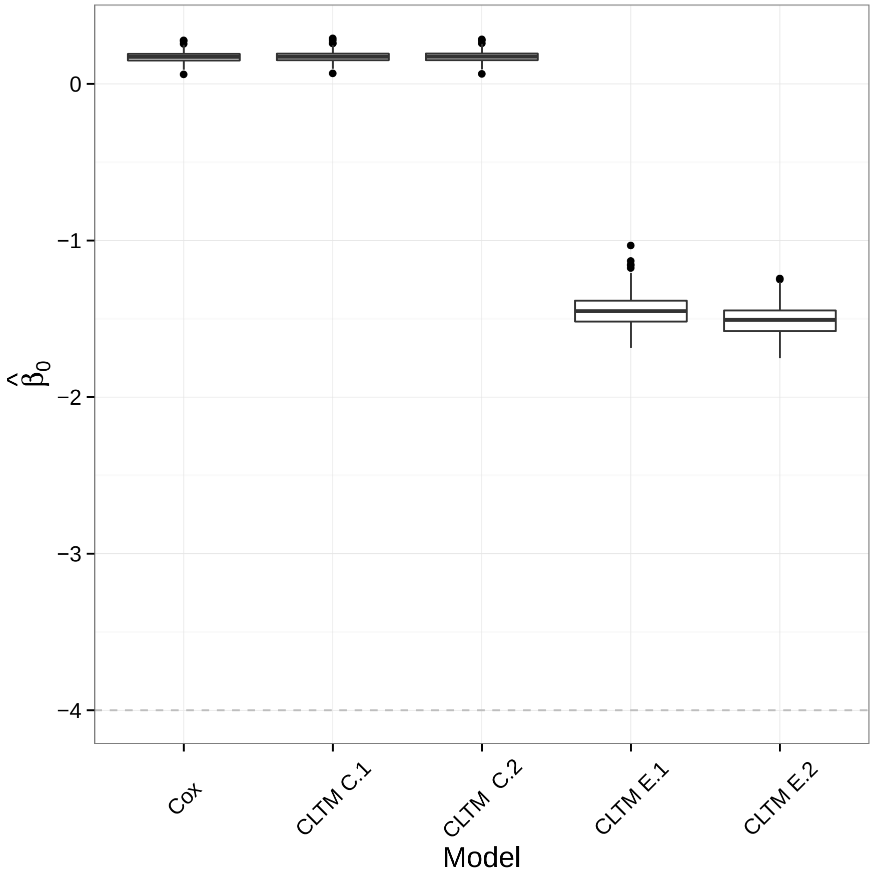


Figure 6.9.: Simulation 3: Boxplots of the estimated regression coefficients $\hat{\beta}_0$ for models CLTM C.1, CLTM C.2, CLTM E.1, CLTM E.2, and the Cox model for $B = 100$ simulated data sets. The dashed reference line indicates the true value $\beta_0 = -4$.

coefficients $\hat{\beta}_1$ for the linear interaction in CLTM E.1 and CLTM E.2 differed considerably from the regression coefficient for the true non-linear interaction, $\beta_1 = 2$ (Figure 6.10). Due to the wrong structure of the conditional transformation function, the estimated survival time transformations and their first derivatives varied considerably from their true counterparts (Figure 6.11 and Figure 6.12). Actually, there was no unconditional survival time transformation in Simulation 3. Nevertheless, a survival time transformation function was estimated in all CLTMs because the models tried to capture the nonlinearity of the interaction between t and x. The estimated functions for CLTM E.1 and CLTM E.2 showed a better fit compared to CLTM C.1 and CLTM C.2, as $\hat{h}_T(t)$ and $\hat{h}'_T(t)$ were close to zero for higher survival times.

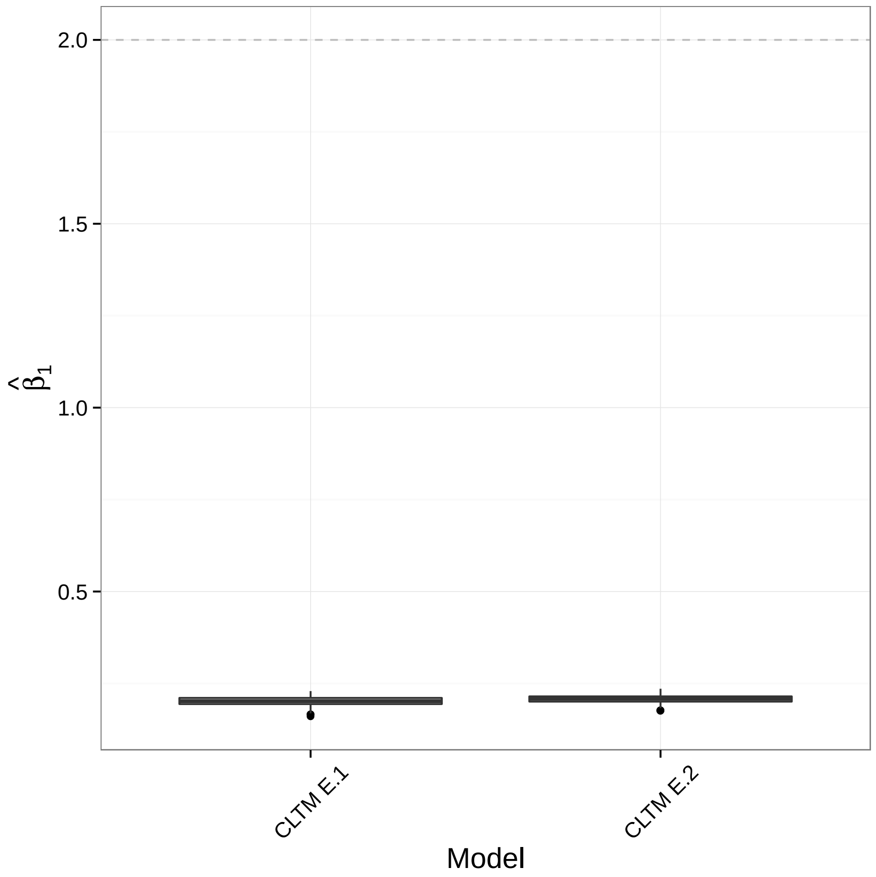


Figure 6.10.: Simulation 3: Boxplots of the estimated regression coefficients $\hat{\beta}_1$ for models CLTM E.1 and CLTM E.2 for $B = 100$ simulated data sets. The dashed reference line indicates the true value $\beta_1 = 2$.

6.3.2. MADs and out-of-sample uncensored log scores

So far, we evaluated only the CLTMs and the Cox model graphically and we focused on the accuracy of the estimated model components. Now, we additionally consider two measures to quantify the model performance of all six models that focus on the estimated conditional distribution function of the survival times.

First, we calculated the mean absolute deviation (MAD) between the true and the estimated conditional distribution functions. A separate MAD-value was calculated for each of the $B = 100$ data sets and for each model CLTM C.1 – CLTM E.2, the CTM, and the Cox model. As an example, we consider one of the data sets and one specific model strategy: For each x-value x_i, $i = 1, \dots, N$, we predicted the distribution function values over a grid of time points. This grid of time points consisted of all observed survival times t_ι, $\iota =$

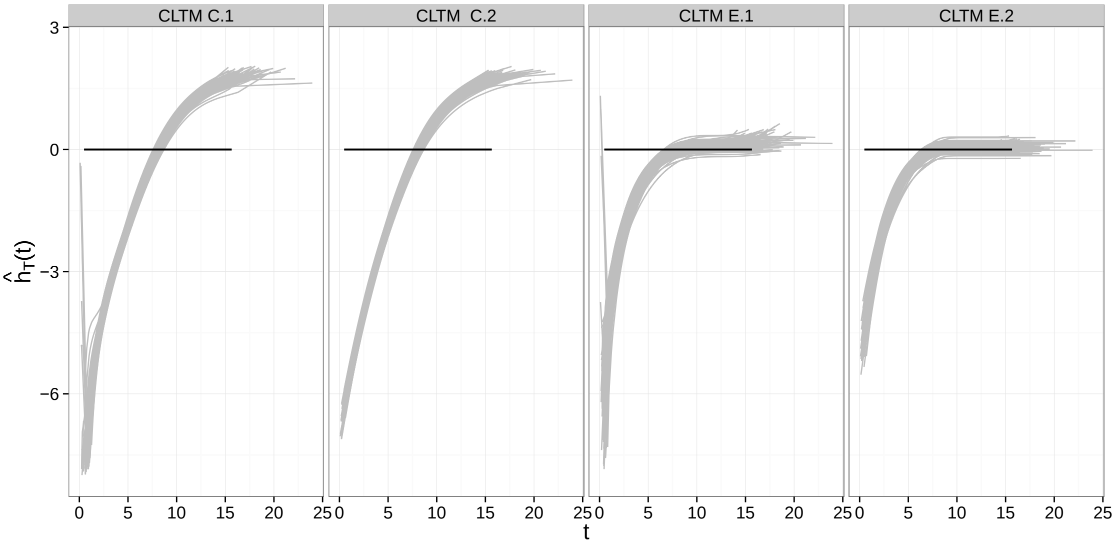


Figure 6.11.: Simulation 3: Estimated survival time transformations $\hat{h}_T(t)$ for $B = 100$ simulated data sets for models CLTM C.1, CLTM C.2, CLTM E.1, and CLTM E.2. The true survival time transformation $h_T(t) = 0$ is displayed in black.

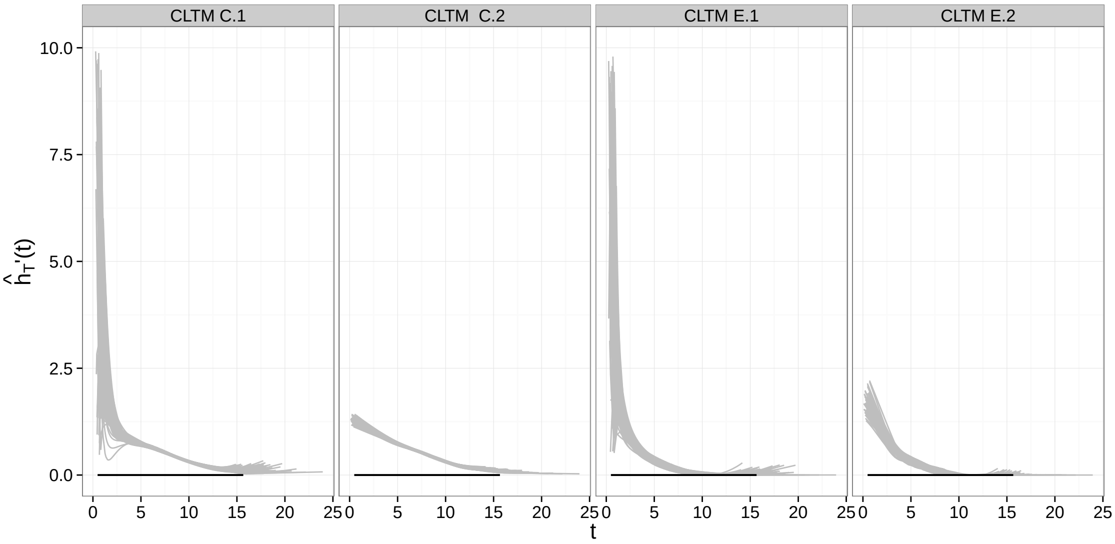


Figure 6.12.: Simulation 3: Estimated first derivatives of the survival time transformations $\hat{h}^{\shortmid}_T(t)$ for $B = 100$ simulated data sets for models CLTM C.1, CLTM C.2, CLTM E.1, and CLTM E.2. The true first derivative of the survival time transformation $h^{\shortmid}_T(t) = 0$ is displayed in black.

$1, \ldots, N$, in the data set. The predicted probabilities were compared to the corresponding probabilities from the true distribution function by calculating mean absolute deviations:

$$MAD = \frac{1}{N \cdot N} \sum_{i=1}^{N} \sum_{\iota=1}^{N} |p(t_\iota | x_i) - \pi(t_\iota | x_i)|,$$

where $p(t_\iota|x_i)$ denotes the probability for the i-th individual at time point t_ι from the true distribution function, and $\pi(t_\iota|x_i) = \mathcal{M}(\hat{h}(t_\iota|x_i))$ denotes the corresponding estimated probability. Thereby, the grid of x-values $\{x_1, \ldots, x_N\}$ is the grid we defined for data generation, *i.e.* is an equidistant grid on $[1, 2]$ in Simulation 1 and Simulation 2, and an equidistant grid on $[1, 3]$ in Simulation 3. A small MAD value is desirable and indicates a good model fit.

To evaluate the predictive ability of the considered models, we calculated the uncensored log score for a set of new observations. $N = 2,000$ new observations were generated using the respective data generating process (Section 6.1), *i.e.* we first defined an equidistant grid of $N = 2,000$ x-values, $\{x_1, \ldots, x_{2,000}\}$, on $[1, 2]$ (Simulation 1 and Simulation 2), or on $[1, 3]$ (Simulation 3), and sampled the corresponding survival times afterwards. For each new survival time and corresponding x-value, $(T_l, x_l), l = 1, \ldots, 2,000$, the true survivor status $I(T_l \leq t_\iota)$ was compared to the corresponding estimated value of the conditional distribution function, $\mathcal{M}(\hat{h}(t_\iota|x_l))$, resulting from one of the considered models. The comparison took place along a grid of time points $\{t_\iota|\iota = 1, \ldots, 2,000\}$ consisting of all new survival times $T_1, \ldots, T_{2,000}$ in terms of the uncensored log score (Equation (6.8)). Again, one out-of-sample uncensored log score resulted for each of the $B = 100$ data sets, and for each considered model. A small out-of-sample uncensored log score is desirable because it indicates a good predictive ability of the considered model.

Simulation 1. The mean MAD values for the $B = 100$ data sets of Simulation 1 reflected exactly our previous expectations of model performance (Figure 6.13). CLTM C.1 and CLTM C.2 performed best because they assumed the true structure of the transformation function. Although the Cox model assumed the same model structure, it showed very slightly higher MAD values. This might be due to the estimation of $h_T(t)$ as a step function, whereas a smooth function estimate results for CLTM C.1 and CLTM C.2. Slightly higher MAD values resulted for CLTM E.1 and CLTM E.2 due to the higher complexity of the conditional transformation function. Nevertheless, both models performed remarkably well because they were able to identify the true structure of the conditional transformation function. The highest MAD values resulted for the CTM because the higher model flexibility was superfluous in this simulation setting. Nevertheless, the resulting MAD values were still low and indicated a good model performance.

The predictive ability of the models was measured in terms of the out-of-sample uncensored log score for $N = 2,000$ new observations. The resulting out-of-sample uncensored log scores for the different models confirmed the conclusions that were drawn for the MAD values (Figure 6.14).

Simulation 2. The model performance measured in terms of MADs reflected exactly our previous expectations (Figure 6.15). CLTM E.1 and CLTM E.2 showed the lowest MAD values because the corresponding conditional transformation function included the linear

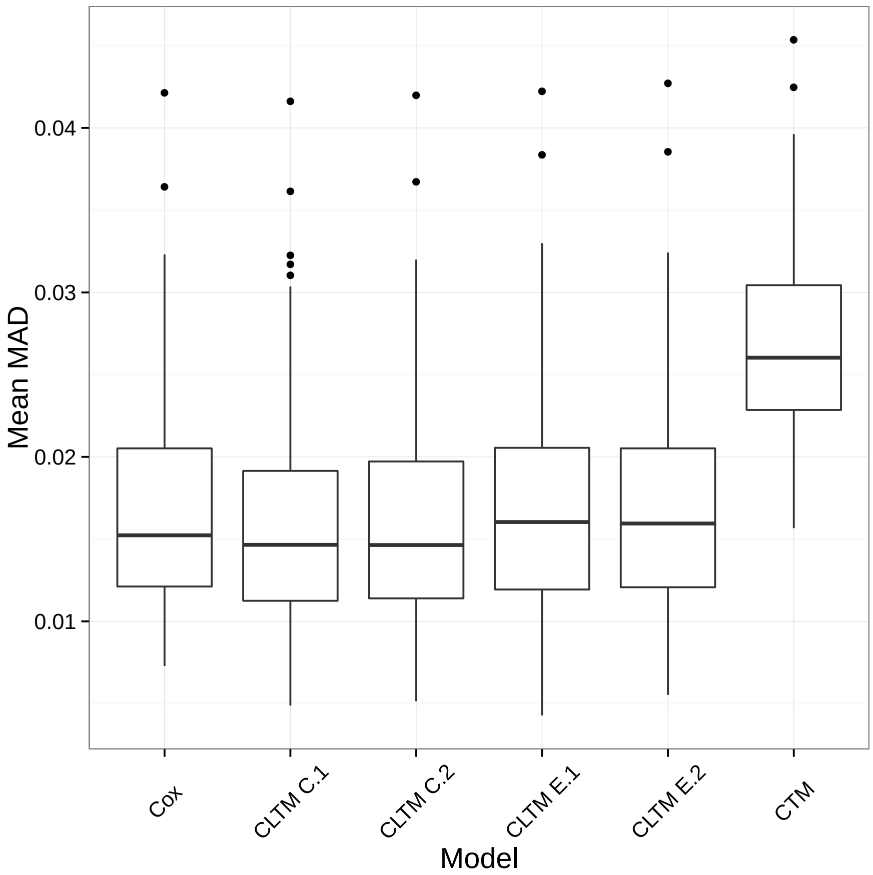


Figure 6.13.: Simulation 1: Boxplots of the mean MAD values based on $B = 100$ simulated data sets for the Cox model (Cox), conditionally linear transformation models CLTM C.1, CLTM C.2, CLTM E.1 and CLTM E.2, and for the conditional transformation model (CTM).

interaction between t and x and thus, the true structure of the conditional transformation function was assumed. The Cox model, CLTM C.1 and CLTM C.2 performed clearly worse due to the underlying PH assumption that was violated in Simulation 2. The more flexible CTM is also able to model non-proportional hazards. Hence, it performed better than CLTM C.1 and CLTM C.2, but worse than CLTM E.1 and CLTM E.2 due to the superfluous additional model complexity.

Comparing the out-of-sample uncensored log scores for the different models resulted in the same conclusions that were presented for the MAD values (Figure 6.16).

Simulation 3. The MADs for the Weibull distributed survival times with non-proportional hazards in Simulation 3 were in very good accordance with our previous expectations (Figure 6.17). As we investigated a non-proportional hazards setting, the highest MADs resulted for the Cox model, CLTM C.1 and CLTM C.2. CLTM E.1 and CLTM E.2 at least included

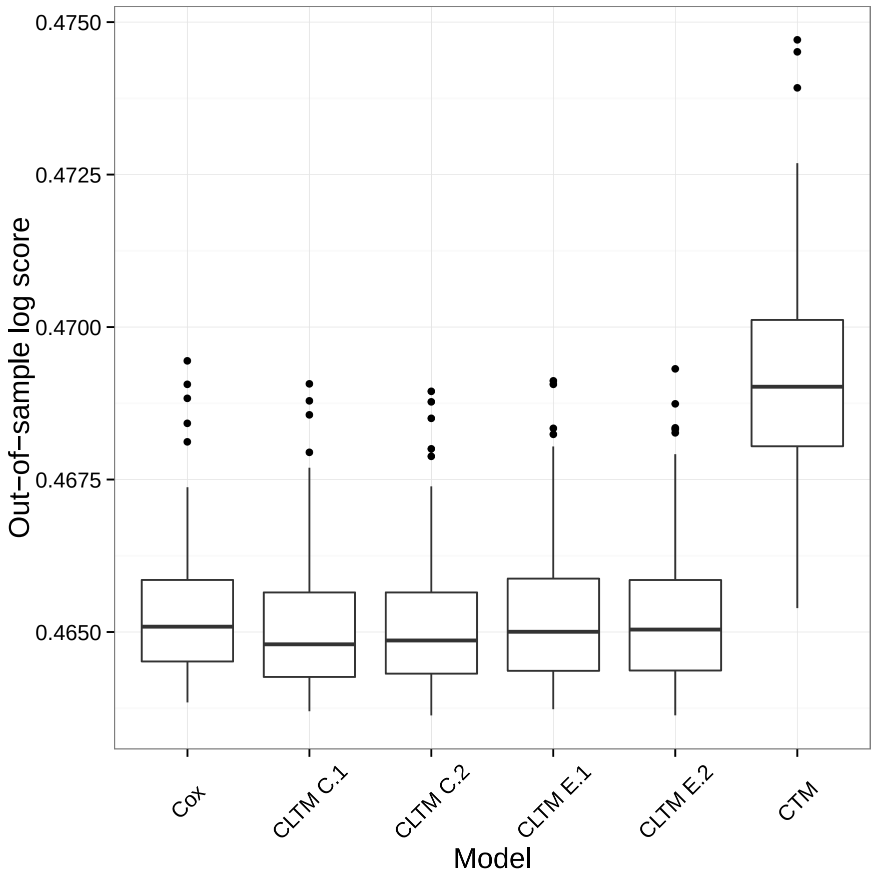


Figure 6.14.: Simulation 1: Boxplots of the out-of-sample uncensored log scores for $N = 2,000$ new observations and $B = 100$ simulated data sets for the Cox model (Cox), conditionally linear transformation models CLTM C.1, CLTM C.2, CLTM E.1 and CLTM E.2, and for the conditional transformation model (CTM).

a linear interaction between t and x and hence, the according MADs were clearly smaller. Nevertheless, the CTM is the only model that is able to account for a non-linear interaction between t and x and thus, the corresponding MADs were the smallest.

The same results could be derived from the out-of-sample uncensored log scores for the different models (Figure 6.18). Nevertheless, the CLTMs based on fractional polynomials, *i.e.* CLTM C.1 and CLTM E.1, showed higher out-of-sample uncensored log scores and a clearly higher variability compared to the CLTMs based on T-splines. This is most probably due to the fractional polynomials t^{-2} and t^{-1} that reach high values for small survival times. These inconsistencies seem to be worse for wrongly defined conditional transformation functions because the variability is especially high for CLTM C.1. Additionally, the superiority of the CTM is recognisable more clearly in terms of out-of-sample uncensored log scores.

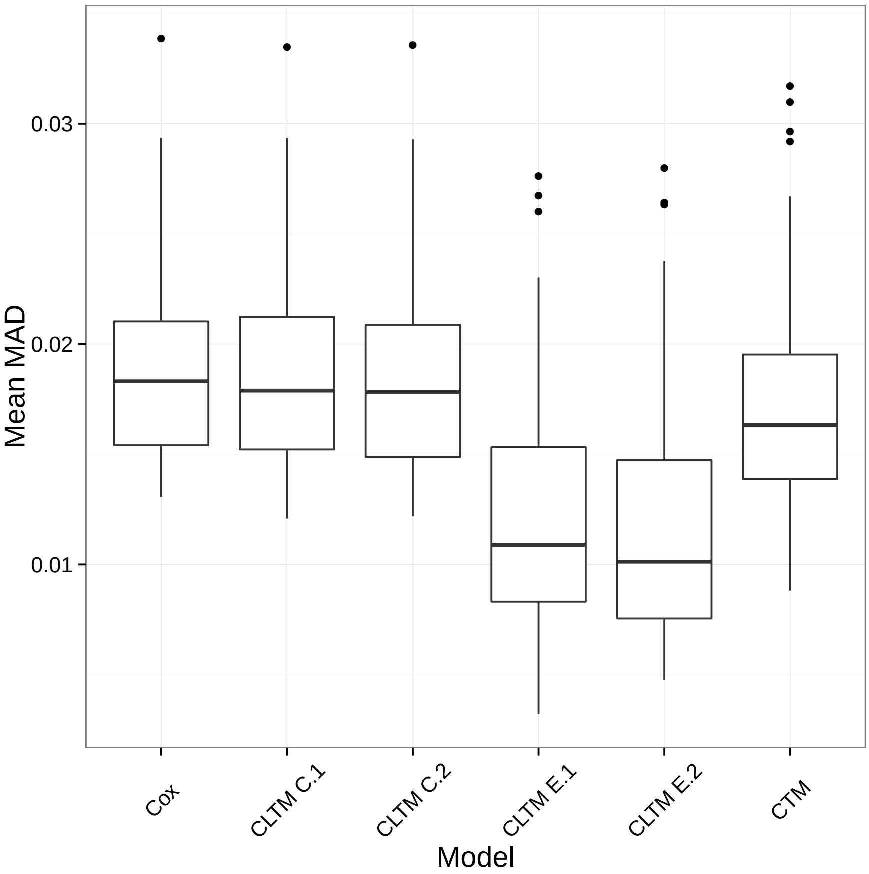


Figure 6.15.: Simulation 2: Boxplots of the mean MAD values based on $B = 100$ simulated data sets for the Cox model (Cox), conditionally linear transformation models CLTM C.1, CLTM C.2, CLTM E.1 and CLTM E.2, and for the conditional transformation model (CTM).

6.4. Summary

The likelihood-based estimation of low-parametrised C(L)TMs has been suggested in Section 3.2. As a proof of concept, we compared the performance of several likelihood-based C(L)TMs with differing model complexity to the performance of the Cox model in a simulation study. We considered three different simulation settings for uncensored survival times with proportional and non-proportional hazards.

Our simulation results indicate that CLTMs are a flexible model class that is able to deal with proportional as well as non-proportional hazards. In the proportional hazards setting, the C(L)TMs assuming proportional hazards and the Cox model showed almost identical results. However, in the non-proportional hazards setting, the C(L)TMs clearly outperformed the Cox model because the PH assumption can be relaxed easily. Moreover, we were able to show that the true structure of the conditional transformation function could be identified by comparing C(L)TMs with different model complexities. The regression

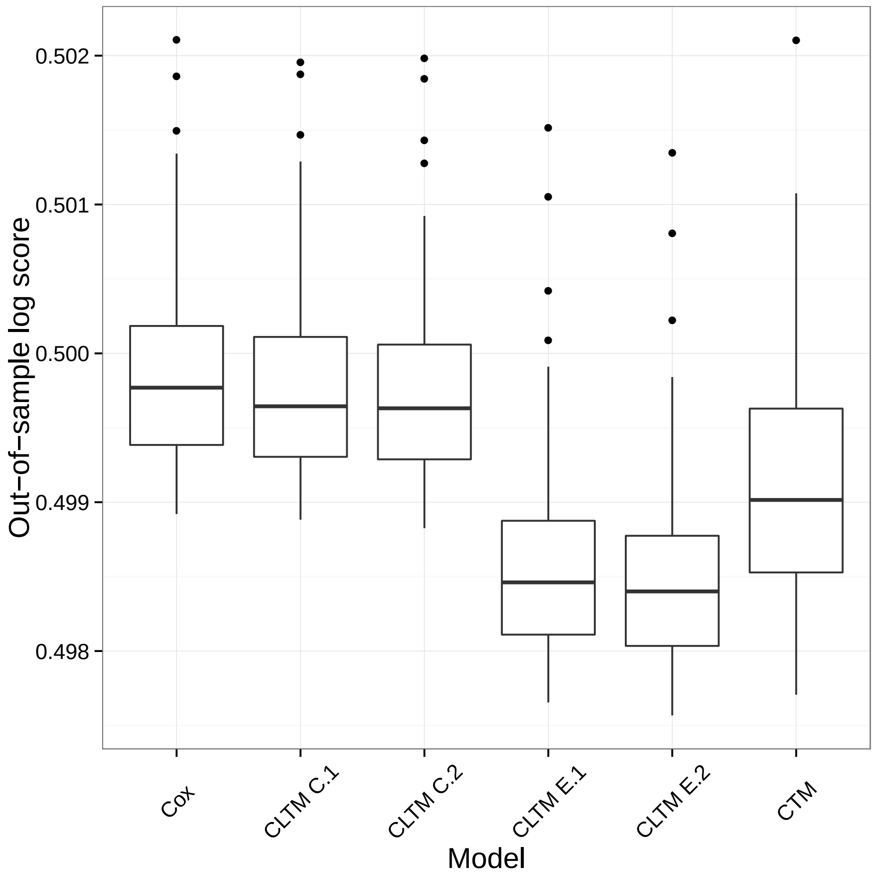


Figure 6.16.: Simulation 2: Boxplots of the out-of-sample uncensored log scores for $N = 2,000$ new observations and $B = 100$ simulated data sets for the Cox model (Cox), conditionally linear transformation models CLTM C.1, CLTM C.2, CLTM E.1 and CLTM E.2, and for the conditional transformation model (CTM).

coefficients and the survival time transformation were estimated accurately in CLTMs if the true structure of the conditional transformation function was assumed.

To test two of the proposed estimation strategies presented in Section 3.2.2, we parametrised the survival time transformation $h_T(t)$ in CLTMs using fractional polynomials and T-splines. Both parametrisations performed satisfyingly. Nevertheless, the T-spline parametrisation turned out to be more stable at the lower boundary of the survival time transformation function.

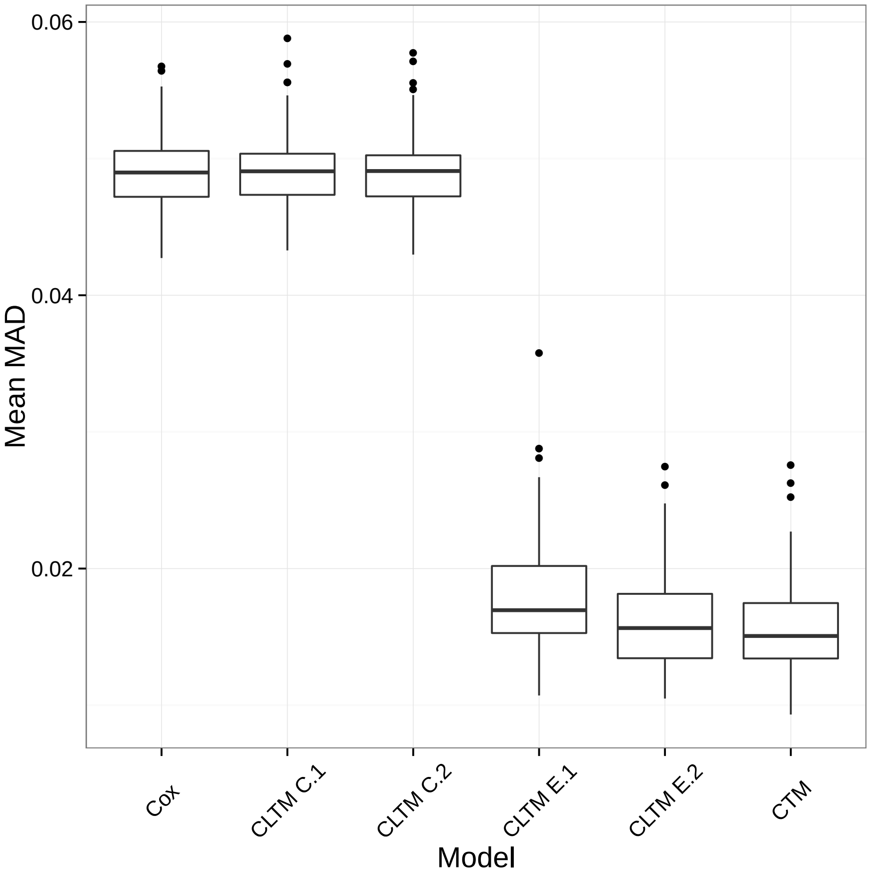


Figure 6.17.: Simulation 3: Boxplots of the mean MAD values based on $B = 100$ simulated data sets for the Cox model (Cox), conditionally linear transformation models CLTM C.1, CLTM C.2, CLTM E.1 and CLTM E.2, and for the conditional transformation model (CTM).

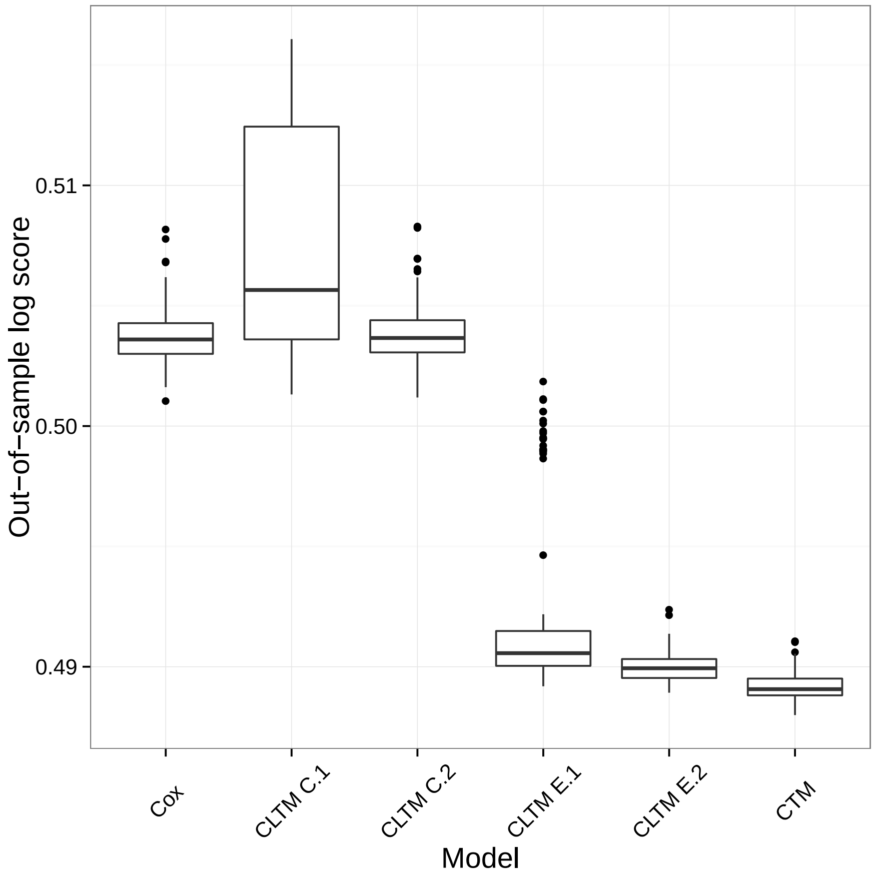


Figure 6.18.: Simulation 3: Boxplots of the out-of-sample uncensored log scores for $N = 2,000$ new observations and $B = 100$ simulated data sets for the Cox model (Cox), conditionally linear transformation models CLTM C.1, CLTM C.2, CLTM E.1 and CLTM E.2, and for the conditional transformation model (CTM).

7. Predicting birth weight with likelihood-based conditionally linear transformation models

The content of this chapter is based on Möst and Hothorn (2014).

The conditional distribution function of birth weight (BW) depending on the ultrasound parameters biparietal diameter (BPD), fronto-occipital diameter (FOD), head circumference (HC), abdominal transverse diameter (ATD), anterior-posterior abdominal diameter (APD), abdominal circumference (AC), and femur length (FL), as well as the maternal body mass index (BMI) has already been analysed in Chapter 4. Thereby, C(L)TMs of different model complexity were considered for the analysis, and estimation was based on a component-wise boosting algorithm. Accordingly, a thorough introduction to birth weight prediction, references on literature covering birth weight prediction formulas used in the past, and a description of the Perinatal Database Erlangen can be found in Chapter 4. In this chapter, we re-analysed the Perinatal Database Erlangen using likelihood-based, low-parametrised CLTMs. Therefore, we considered models CLTM A – CLTM E proposed in Section 2.2.2. Parameter estimation was based on a full maximum likelihood approach. Thereby, the link function was set to the standard normal distribution function, $F = \Phi$, because birth weights are usually assumed to be normal distributed (see Chapter 4). As all birth weights were observed, model estimation was based on the maximisation of the uncensored log-likelihood for continuous responses given in Equation (3.2).

7.1. Specific CLTMs for analysing the Perinatal Database Erlangen

When considering models CLTM A – CLTM E (Section 2.2.2) for the analysis of the Perinatal Database Erlangen, we estimated the distribution function of birth weight conditional on the ultrasound measurements and the maternal BMI. Thereby, all ultrasound parameters and the maternal BMI were considered as main effects, and we additionally considered the linear interaction between AC and FL because this interaction turned out to be important in the past (see Chapter 4, and Möst et al., 2014). Similar to our proceeding in Chapter 4, the interaction term was only allowed to influence the conditional mean but not the conditional variance of birth weight. In model CLTM B, a smooth interaction term between AC and FL would require the estimation of a smooth bivariate surface. Hence, we ignored the interaction due to reasons of model complexity. Usually, the relationship between birth weight and ultrasound measurements is analysed using ordinary linear regression models (references can be found in Möst et al., 2014, and Chapter 4), whereby normal distributed birth weights are assumed. Therefore, we also considered an ordinary linear regression model LM for reasons of comparison. The models are presented in more detail below, and they are ordered with increasing model complexity. Important model characteristics and assumptions are summarised.

LM: Linear regression model. Birth weight depends linearly on the ultrasound measurements:

$$\begin{aligned}\mathrm{BW} &= \beta_0 + \beta_{\mathrm{BPD}} \cdot \mathrm{BPD} + \beta_{\mathrm{FL}} \cdot \mathrm{FL} + \beta_{\mathrm{AC}} \cdot \mathrm{AC} + \beta_{\mathrm{HC}} \cdot \mathrm{HC} + \beta_{\mathrm{FOD}} \cdot \mathrm{FOD} + \\ &\quad \beta_{\mathrm{ATD}} \cdot \mathrm{ATD} + \beta_{\mathrm{APD}} \cdot \mathrm{APD} + \beta_{\mathrm{BMI}} \cdot \mathrm{BMI} + \beta_{\mathrm{AC:FL}} \cdot \mathrm{AC} \cdot \mathrm{FL} + \epsilon,\end{aligned}$$

where ϵ denotes a normal distributed error term with mean zero and variance σ^2. In model LM, we assumed normal distributed birth weights, a constant variance term σ^2 that is independent of the ultrasound measurements, and linear influences of the ultrasound measurements on the conditional mean of birth weight.

CLTM A. Model CLTM A (Equation (2.12)) is the transformation model analogon to model LM because the assumed model complexity is identical. The conditional transformation function is

$$h(\mathrm{BW}|\boldsymbol{x}) = \alpha_0 + \alpha_1 \cdot \mathrm{BW} + \beta_{0,\mathrm{BPD}} \cdot \mathrm{BPD} + \ldots + \beta_{0,\mathrm{BMI}} \cdot \mathrm{BMI} + \beta_{0,\mathrm{AC:FL}} \cdot \mathrm{AC} \cdot \mathrm{FL}.$$

As we chose $F = \Phi$ for the link function, the expectation of $h(\mathrm{BW}|\boldsymbol{x})$ is zero, and the effects of the ultrasound measurements on the conditional mean $\mathbb{E}(\mathrm{BW}|\boldsymbol{x})$ can be rewritten as

$$\mathbb{E}(\mathrm{BW}|\boldsymbol{x}) = {}^{(-\alpha_0 - \beta_{0,\mathrm{BPD}} \cdot \mathrm{BPD} - \ldots - \beta_{0,\mathrm{BMI}} \cdot \mathrm{BMI} - \beta_{0,\mathrm{AC:FL}} \cdot \mathrm{AC} \cdot \mathrm{FL})}\!/_{\alpha_1}.$$

Hence, *e.g.*, the effect of BPD on the conditional mean, ${}^{-\beta_{0,\mathrm{BPD}}}\!/_{\alpha_1}$, can be directly compared to β_{BPD} from model LM, and both estimates were expected to be identical. Consequently, the underlying model assumptions were the same as in model LM: We assumed normal distributed birth weights because the response transformation function is linear, *i.e.* $h_{\mathrm{BW}}(\mathrm{BW}) = \alpha_0 + \alpha_1 \cdot \mathrm{BW}$, and we are not able to leave the class of normal distribution functions by linear transformations. Moreover, we assumed a constant variance σ^2 for all birth weights, which can be calculated via $\sigma^2 = \mathbb{V}(\mathrm{BW}|\boldsymbol{x}) = {}^{1}\!/_{\alpha_1^2}$. Third, the influence of the ultrasound measurements on the conditional mean of birth weight was assumed to be linear. Model CLTM A was estimated by directly maximising the uncensored log-likelihood (Equation (3.2))

$$l(\boldsymbol{\alpha}, \boldsymbol{\beta}_0) = \sum_{i=1}^{N} \log(\phi(h(\mathrm{BW}_i | \boldsymbol{x}_i))) + \log(\alpha_1),$$

where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)$, $\boldsymbol{\beta}_0 = (\beta_{0,\mathrm{BPD}}, \ldots, \beta_{0,\mathrm{BMI}}, \beta_{0,\mathrm{AC:FL}})$, and ϕ denotes the density of the standard normal distribution. The linear constraint $\alpha_1 > 0$ had to be considered during estimation to guarantee a monotonically increasing birth weight transformation.

CLTM B. In model CLTM B (Equation (2.13)), the ultrasound measurements were allowed to have a more flexible but smooth influence on the conditional mean of birth weight. Furthermore, we assumed normal distributed birth weights, and a constant variance that is independent of the ultrasound measurements. This resulted in the conditional transformation function

$$h(\mathrm{BW}|\boldsymbol{x}) = \alpha_0 + \alpha_1 \cdot \mathrm{BW} + \beta_{0,\mathrm{BPD}}(\mathrm{BPD}) + \ldots + \beta_{0,\mathrm{BMI}}(\mathrm{BMI}).$$

To get a parsimonious model formulation, the smooth parameter functions were parametrised using fractional polynomials of degree 1 (see Section 3.2.2). For example, the function $\beta_{0,\mathrm{BPD}}(\mathrm{BPD})$ was parametrised via

$$\begin{aligned} \beta_{0,\mathrm{BPD}}(\mathrm{BPD}) \quad = \quad & \beta_{01,\mathrm{BPD}} \cdot \mathrm{BPD}^{-2} + \beta_{02,\mathrm{BPD}} \cdot \mathrm{BPD}^{-1} + \beta_{03,\mathrm{BPD}} \cdot \mathrm{BPD}^{-0.5} + \\ & \beta_{04,\mathrm{BPD}} \cdot \log(\mathrm{BPD}) + \beta_{05,\mathrm{BPD}} \cdot \sqrt{\mathrm{BPD}} + \beta_{06,\mathrm{BPD}} \cdot \mathrm{BPD} + \\ & \beta_{07,\mathrm{BPD}} \cdot \mathrm{BPD}^2 + \beta_{08,\mathrm{BPD}} \cdot \mathrm{BPD}^3, \end{aligned}$$

where the coefficients of the fractional polynomial are summarised to $\boldsymbol{\beta}_{0,\mathrm{BPD}} = (\beta_{01,\mathrm{BPD}}, \ldots, \beta_{08,\mathrm{BPD}})$. The regression coefficients in model CLTM B were estimated by maximising the uncensored log-likelihood

$$l(\boldsymbol{\alpha}, \boldsymbol{\beta}_{0,\mathrm{BPD}}, \ldots, \boldsymbol{\beta}_{0,\mathrm{BMI}}) = \sum_{i=1}^{N} \log(\phi(h(\mathrm{BW}_i | \boldsymbol{x}_i))) + \log(\alpha_1),$$

under the linear constraint $\alpha_1 > 0$.

CLTM C. In linear transformation model CLTM C (Equation (2.14)), the response transformation function $h_{\mathrm{BW}}(\mathrm{BW})$ is assumed to be smooth and monotonically increasing. Due to this birth weight transformation, we are able to leave the class of normal distribution functions, and the birth weights are allowed to follow some arbitrary distribution. This distribution function is the same for all babies, but the respective means are fetus-specific because the conditional mean is influenced linearly by the ultrasound measurements. The corresponding conditional transformation function is

$$h(\mathrm{BW}|\boldsymbol{x}) = h_{\mathrm{BW}}(\mathrm{BW}) + \beta_{0,\mathrm{BPD}} \cdot \mathrm{BPD} + \ldots + \beta_{0,\mathrm{BMI}} \cdot \mathrm{BMI} + \beta_{0,\mathrm{AC:FL}} \cdot \mathrm{AC} \cdot \mathrm{FL}.$$

To guarantee a parsimonious model formulation, the birth weight transformation function was parametrised using fractional polynomials similar to the smooth covariate functions in model CLTM B:

$$\begin{aligned} h_{\mathrm{BW}}(\mathrm{BW}) &= \alpha_0 + \alpha_1 \cdot \mathrm{BW}^{-2} + \alpha_2 \cdot \mathrm{BW}^{-1} + \alpha_3 \cdot \mathrm{BW}^{-0.5} + \alpha_4 \cdot \log(\mathrm{BW}) + \\ & \quad \alpha_5 \cdot \sqrt{\mathrm{BW}} + \alpha_6 \cdot \mathrm{BW} + \alpha_7 \cdot \mathrm{BW}^2 + \alpha_8 \cdot \mathrm{BW}^3. \end{aligned}$$

Model estimation was based on the maximisation of the log-likelihood

$$l(\boldsymbol{\alpha}, \boldsymbol{\beta}_0) = \sum_{i=1}^{N} \log(\phi(h(\mathrm{BW}_i|\boldsymbol{x}_i))) + \log(h'_{\mathrm{BW}}(\mathrm{BW}_i))$$

under the linear constraints $h'_{\mathrm{BW}}(\mathrm{BW}_i) > 0 \ \forall i$. Thereby, the coefficients are summarised to $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_8)$ and $\boldsymbol{\beta}_0 = (\beta_{0,\mathrm{BPD}}, \ldots, \beta_{0,\mathrm{BMI}}, \beta_{0,\mathrm{AC:FL}})$.

CLTM D. The ultrasound measurements are allowed to influence the conditional mean and the conditional variance of birth weight in model CLTM D (Equation (2.15)). The birth weight transformation function is linear, whereby normal distributed birth weights are assumed. The corresponding conditional transformation function is

$$\begin{aligned} h(\mathrm{BW}|\boldsymbol{x}) &= \alpha_0 + \beta_{0,\mathrm{BPD}} \cdot \mathrm{BPD} + \ldots + \beta_{0,\mathrm{BMI}} \cdot \mathrm{BMI} + \beta_{0,\mathrm{AC:FL}} \cdot \mathrm{AC} \cdot \mathrm{FL} + \\ & \quad \mathrm{BW} \cdot (\alpha_1 + \beta_{1,\mathrm{BPD}} \cdot \mathrm{BPD} + \ldots + \beta_{1,\mathrm{BMI}} \cdot \mathrm{BMI}). \end{aligned}$$

As the conditional transformation function has mean zero and variance one, the conditional mean and the conditional variance of birth weight can also be written as

$$\begin{aligned} \mathbb{E}(\mathrm{BW}|\boldsymbol{x}) &= \frac{-\alpha_0 - \beta_{0,\mathrm{BPD}} \cdot \mathrm{BPD} - \ldots - \beta_{0,\mathrm{BMI}} \cdot \mathrm{BMI} - \beta_{0,\mathrm{AC:FL}} \cdot \mathrm{AC}}{\alpha_1 + \beta_{1,\mathrm{BPD}} \cdot \mathrm{BPD} + \ldots + \beta_{1,\mathrm{BMI}} \cdot \mathrm{BMI}} \\ \mathbb{V}(\mathrm{BW}|\boldsymbol{x}) &= \frac{1}{(\alpha_1 + \beta_{1,\mathrm{BPD}} \cdot \mathrm{BPD} + \ldots + \beta_{1,\mathrm{BMI}} \cdot \mathrm{BMI})^2}. \end{aligned}$$

For model estimation, the log-likelihood

$$l(\boldsymbol{\alpha}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1) = \sum_{i=1}^{N} \log(\phi(h(\mathrm{BW}_i|\boldsymbol{x}_i))) + \log(\alpha_1 + \beta_{1,\mathrm{BPD}} \cdot \mathrm{BPD}_i + \ldots + \beta_{1,\mathrm{BMI}} \cdot \mathrm{BMI}_i)$$

had to be maximised under the linear constraints $\alpha_1 > 0$, and $\alpha_1 + \beta_{1,\text{BPD}} \cdot \text{BPD}_i + \ldots + \beta_{1,\text{BMI}} \cdot \text{BMI}_i > 0 \; \forall i$. The regression coefficients were summarised to $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)$, $\boldsymbol{\beta}_0 = (\beta_{0,\text{BPD}}, \ldots, \beta_{0,\text{BMI}}, \beta_{0,\text{AC:FL}})$, and $\boldsymbol{\beta}_1 = (\beta_{1,\text{BPD}}, \ldots, \beta_{1,\text{BMI}})$.

CLTM E. In analogy to model CLTM D, the ultrasound measurements are allowed to influence the conditional mean and the conditional variance of the transformed birth weights. Moreover, in model CLTM E (Equation (2.16)), the unconditional transformation function is a smooth and monotonically increasing function. Thereby, the birth weights are assumed to follow an arbitrary distribution function. The class of distribution functions is the same for all babies, but the means and variances are fetus-specific. The conditional transformation function is

$$\begin{aligned} h(\text{BW}|\boldsymbol{x}) \;&=\; h_{\text{BW}}(\text{BW}) + \beta_{0,\text{BPD}} \cdot \text{BPD} + \ldots + \beta_{0,\text{BMI}} \cdot \text{BMI} + \beta_{0,\text{AC:FL}} \cdot \text{AC} \cdot \text{FL} + \\ & \quad\; \text{BW} \cdot (\beta_{1,\text{BPD}} \cdot \text{BPD} + \ldots + \beta_{1,\text{BMI}} \cdot \text{BMI}). \end{aligned}$$

Similar to CLTM C, the smooth birth weight transformation function $h_{\text{BW}}(\text{BW})$ is parametrised in terms of fractional polynomials. The regression coefficients were estimated by maximising the uncensored log-likelihood

$$l(\boldsymbol{\alpha}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1) = \sum_{i=1}^{N} \log(\phi(h(\text{BW}_i|\boldsymbol{x}_i))) + \log(h'_{\text{BW}}(\text{BW}_i) + \beta_{1,\text{BPD}} \cdot \text{BPD}_i + \ldots + \beta_{1,\text{BMI}} \cdot \text{BMI}_i)$$

under the linear constraints $h'_{\text{BW}}(\text{BW}_i) > 0 \; \forall i$, and $h'_{\text{BW}}(\text{BW}_i) + \beta_{1,\text{BPD}} \cdot \text{BPD}_i + \ldots + \beta_{1,\text{BMI}} \cdot \text{BMI}_i > 0 \; \forall i$. The regression coefficients were summarised to $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_8)$, $\boldsymbol{\beta}_0 = (\beta_{0,\text{BPD}}, \ldots, \beta_{0,\text{BMI}}, \beta_{0,\text{AC:FL}})$, and $\boldsymbol{\beta}_1 = (\beta_{1,\text{BPD}}, \ldots, \beta_{1,\text{BMI}})$.

7.2. Model evaluation

To identify the CLTM that describes the Perinatal Database Erlangen best, we compared the performance of models CLTM A – CLTM E, and the LM. Therefore, we estimated all proposed models on a training data set, and evaluated their predictive ability on an evaluation data set. We generated $B = 50$ training and evaluation data sets using the following procedure (for a more detailed version see Section 4.3.4):

1. The ultrasound parameters AC and FL turned out to be essential for the prediction of birth weight in the past. Therefore, we divided the fetuses in the database into 25 AC-FL categories.

2. The training data sets were generated by choosing randomly 50% of the original observations in each AC-FL category.

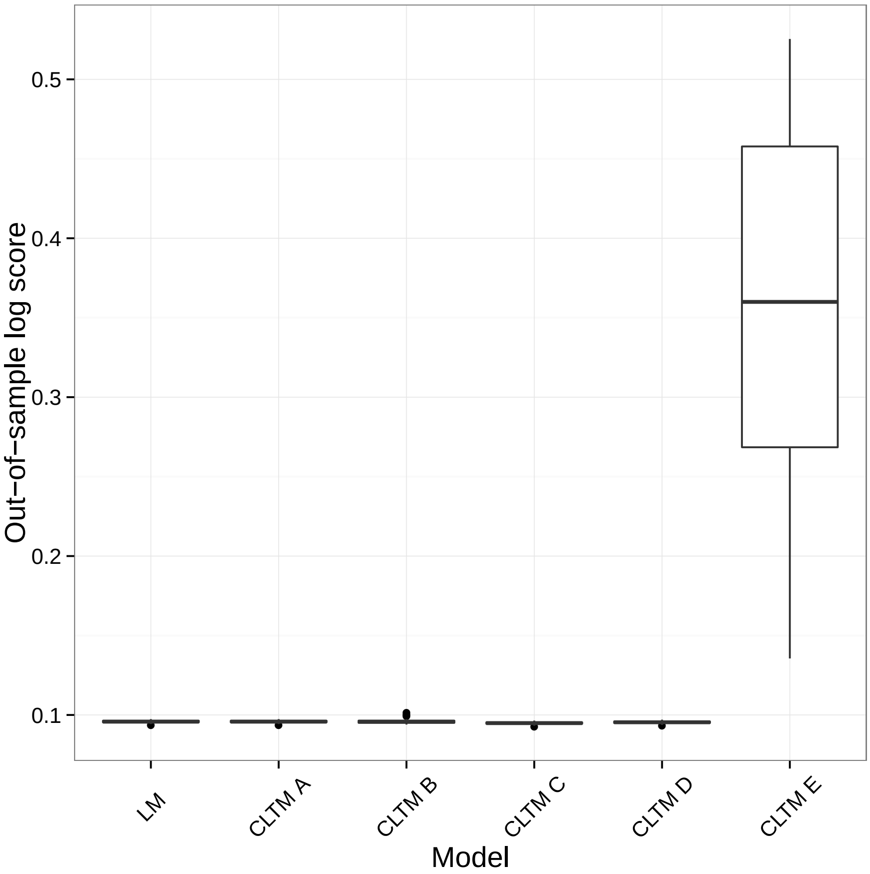


Figure 7.1.: Out-of-sample log scores for conditionally linear transformation models CLTM A – CLTM E, and the linear model (LM) based on 50 randomly chosen evaluation data sets consisting of $4,355$ observations.

3. The remaining 50% of the original observations formed the corresponding evaluation data set.

The predictive ability of the considered models was measured in terms of the uncensored log score (see Equation (4.12)):

$$LS = -\frac{1}{N \cdot n} \sum_{i=1}^{N} \sum_{\iota=1}^{n} I(\mathrm{BW}_i \leq \upsilon_\iota) \log(\Phi(\hat{h}(\upsilon_\iota | \boldsymbol{x}_i))) + I(\mathrm{BW}_i > \upsilon_\iota) \log(1 - \Phi(\hat{h}(\upsilon_\iota | \boldsymbol{x}_i))),$$

where $\{\upsilon_1, \ldots, \upsilon_n\}$ denotes a grid of birth weights, which was chosen to be an equidistant grid with length $n = 50$ from the lowest to the highest observed birth weight. The index i denotes the observations in the evaluation data set, which consisted of half of the original observations, *i.e.* $N = 4,355$. The estimated conditional transformation functions $\hat{h}$ resulted from estimating models CLTM A – CLTM E on the training data set. This procedure is useful for detecting model misspecifications because the complexities of the

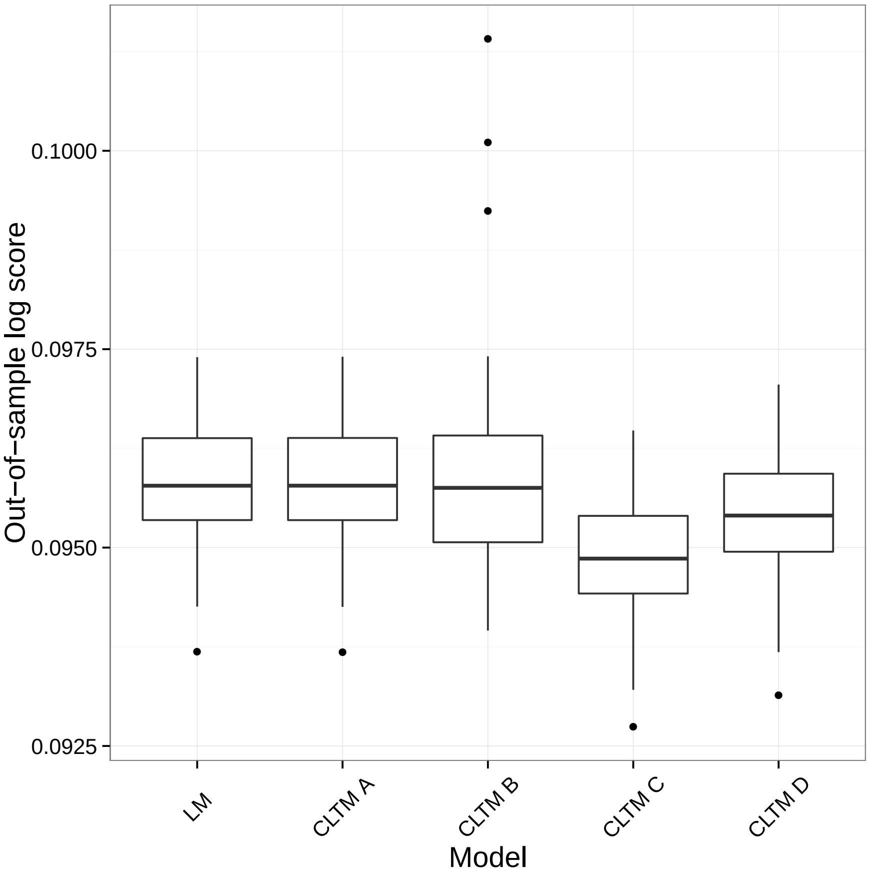


Figure 7.2.: Out-of-sample log scores for conditionally linear transformation models CLTM A – CLTM D, and the linear model (LM) based on 50 randomly chosen evaluation data sets consisting of $4,355$ observations.

considered CLTMs differed. For example, we were able to check the assumption of normal distributed birth weights by comparing model CLTM A to CLTM C, and model CLTM D to CLTM E, respectively; missing effects of the ultrasound measurements on the variance could be detected by comparing model CLTM D to model CLTM A, or model CLTM E to model CLTM C.

Model CLTM E clearly showed the highest out-of-sample log scores and thus performed worst (Figure 7.1). Hence, the consideration of arbitrarily distributed birth weights with fetus-specific means and variances does not fit the given data satisfyingly. As we had expected, models LM and CLTM A showed identical results because both models impose identical assumptions, and both models were estimated using the full log-likelihood (Figure 7.2). Moreover, models CLTM A and CLTM B performed comparably well, *i.e.* the consideration of flexible influences of the ultrasound measurements on the conditional mean of birth weight did not lead to a model improvement and hence, the consideration of linear effects is adequate. When assuming normal distributed birth weights (models LM, CLTM A, CLTM B, and CLTM D), the consideration of fetus-specific variances led to a small model improvement because model CLTM D is associated with the smallest out-of-sample log scores. Nevertheless, model CLTM C showed the smallest out-of-sample log scores throughout. Therefore, the model assuming arbitrarily distributed birth weights with fetus-specific means that are influenced linearly by the ultrasound measurements performed best, and we advise to use model CLTM C for analysing the Perinatal Database Erlangen.

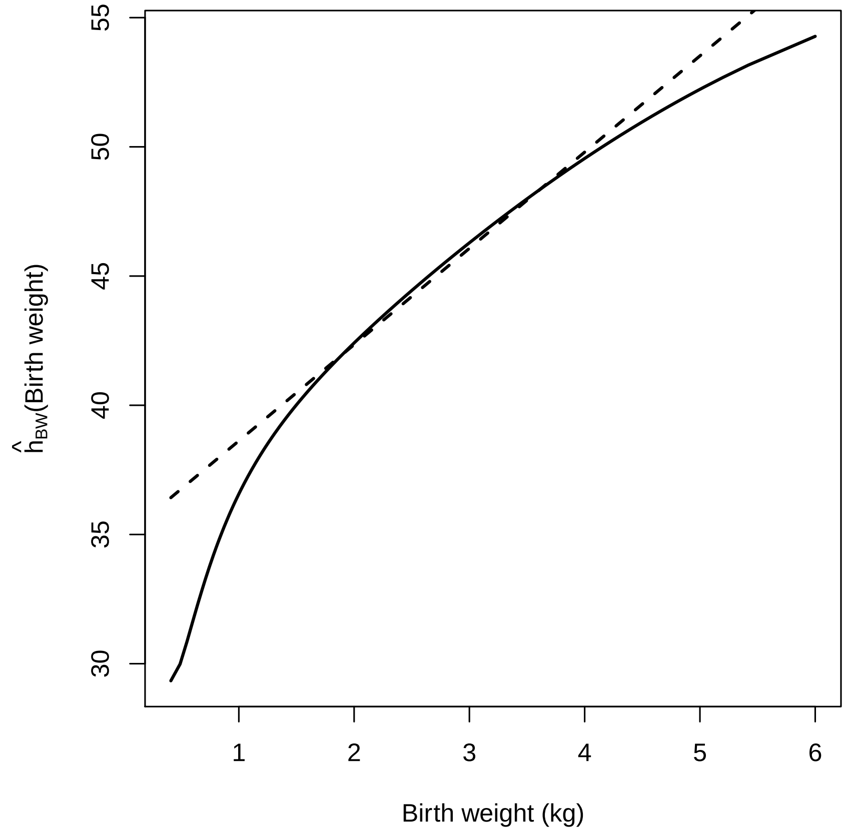


Figure 7.3.: Estimated birth weight transformation function $\hat{h}_{\text{BW}}(\text{BW})$ resulting from model CLTM C. The dashed line symbolises the linear relationship between the birth weights and their monotone transformation.

As model CLTM C performed best, we present the results from the corresponding data analysis. First, the estimated birth weight transformation function $\hat{h}_{\text{BW}}(\text{BW})$ indicated that the birth weights do not follow a normal distribution function (Figure 7.3). If the birth weights were normal distributed, $\hat{h}_{\text{BW}}(\text{BW})$ would have been estimated to be a linear function. Obviously, there are deviations from the normal distribution for birth weights at the extremes, especially for low birth weights. This was also supported by the normal quantile-quantile plots of the original and the transformed birth weights (Figure 7.4). The original birth weights deviated from the normal distribution for low birth weights, *i.e.* ,

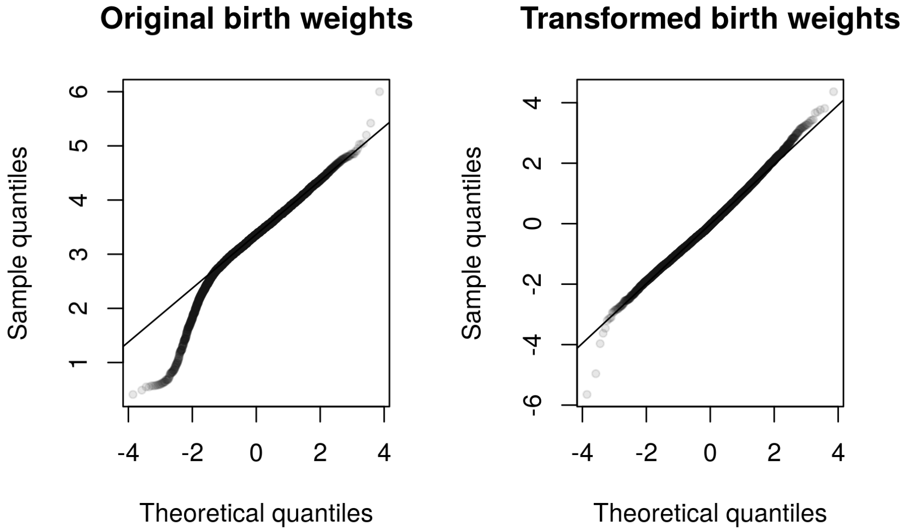


Figure 7.4.: Normal Q-Q plot of original and transformed birth weights resulting from model CLTM C.

the distribution of the original birth weights is not symmetric but rather right-skewed. These deviations from the normal distribution were perfectly captured by the birth weight transformation function $\hat{h}_{\text{BW}}(\text{BW})$ because the transformed birth weights approximately followed a normal distribution. Moreover, we were able to analyse the linear effect of the ultrasound measurements on the conditional mean of the transformed birth weights, $\mathbb{E}(h_{\text{BW}}(\text{BW})|\boldsymbol{x}) = -\beta_{0,\text{BPD}} \cdot \text{BPD} - \ldots - \beta_{0,\text{BMI}} \cdot \text{BMI} - \beta_{0,\text{AC:FL}} \cdot \text{AC} \cdot \text{FL}$, in model CLTM C. The estimated influence of all ultrasound measurements and the maternal body mass index was positive, except for the influence of the linear interaction between AC and FL, which was estimated to be negative. To visualise the importance of the ultrasound measurements for explaining the transformed birth weight, we plotted the influence of each ultrasound parameter on the conditional mean of the transformed birth weight (Figure 7.5). The remaining ultrasound parameters were fixed at their mean values. BPD, FL and AC showed the highest influence; ATD, APD and FOD showed an intermediate influence; and HC and BMI were associated with the lowest influence on the conditional mean of the transformed birth weights.

7.3. Summary

To further illustrate the likelihood-based estimation of CLTMs, we re-analysed the Perinatal Database Erlangen using a cascade of low-parametrised CLTMs. The results presented in this chapter are only comparable to a limited extent to the results presented in Chapter 4. The conditional distribution function of birth weights could be analysed in terms of

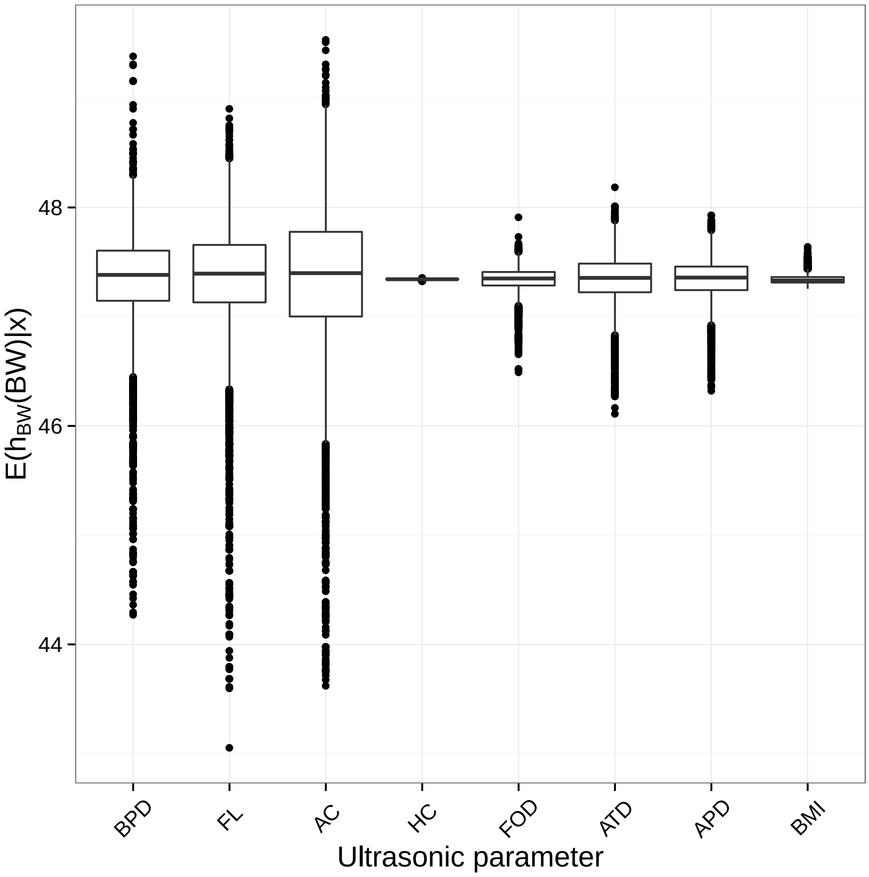


Figure 7.5.: Boxplots of the estimated influence of each ultrasound parameter biparietal diameter (BPD), femur length (FL), abdominal circumference (AC), head circumference (HC), fronto-occipital diameter (FOD), abdominal transverse diameter (ATD), anterior-posterior abdominal diameter (APD), and the mother's body mass index (BMI) on the conditional mean of the transformed birth weights, $\mathbb{E}(h_{\text{BW}}(\text{BW})|\boldsymbol{x})$, resulting from model CLTM C. The remaining ultrasound parameters were fixed at their mean values.

more complex C(L)TMs in Chapter 4 due to the estimation based on a component-wise boosting algorithm, whereas we were only able to consider low-parametrised CLTMs, here. Nevertheless, some of the results could be confirmed. Again, we found that the assumption of normal distributed birth weights is not correct, and that the deviations from the normal distribution should be considered by a suitable model. Furthermore, we found that the assumption of a linear influence of the ultrasound measurements on the conditional mean of birth weight is adequate, and that the conditional mean of the transformed birth weights is most strongly influenced by the ultrasound measurements AC, FL, and BPD. The Perinatal Database Erlangen could be described best by a likelihood-based CLTM assuming arbitrarily distributed birth weights with fetus-specific means.

8. Summary and outlook

In this concluding chapter, we give a general summary of the main developments of this thesis and their importance in concrete applications. We additionally give relevant starting points for future research. A more thorough discussion of the characteristics, advantages and limitations of the applied methodology can be found at the end of each chapter.

Conditional transformation models (CTMs) (Hothorn et al., 2014) are a model class that allows the direct estimation of the conditional distribution function. This flexible approach is beneficial in many applications because all moments of the distribution function (*i.e.* mean, variance, kurtosis, and skewness) might be influenced by the explanatory variables. In this thesis, we extended classical CTMs in two important directions. First, we used CTMs for the determination of prediction intervals, and second, we used CTMs for the estimation of conditional survivor functions. This implied the extension of CTMs to censored response variables. Concerning both topics, standard regression models that are usually used for data analysis often suffer from their strict assumptions. To illustrate the importance of CTMs for determining prediction intervals and for estimating conditional survivor functions, we selected two applications from the field of biostatistics. We analysed the influence of ultrasound measurements on the future birth weights of newborns from the Perinatal Database Erlangen in Chapter 4 using a cascade of interpretable conditionally linear transformation models (CLTMs). Our method allowed the determination of individual prediction intervals for the future birth weight of each baby. Such fetus-specific prediction intervals are highly relevant in terms of individual risk analysis and in terms of organising obstetric management. Moreover, the estimation of patient-specific survival probabilities over time depending on individual patient characteristics is of special interest in survival analysis. Thus, the CTM-based estimation of conditional survivor functions was examined in Chapter 5 and was used to analyse the survival of patients suffering from chronic myelogenous leukaemia.

From a methodological perspective, the important improvement in this thesis is the introduction of likelihood-based conditional transformation models. So far, CTMs have been estimated using a component-wise boosting algorithm, which implies the important advantages of intrinsic variable selection and model choice. This estimation approach has also been used for the determination of prediction intervals in Chapter 4, and has been extended to right-censored observations in Chapter 5. Nevertheless, a broad inference theory is lacking for boosting approaches. Because there is no large sample theory, confidence intervals and p-values can only be determined based on bootstrapping. From this perspective, a likelihood-based approach is beneficial because the large sample theory of maximum likelihood approaches comes for free and the associated inference theory is open to the practitioner. We presented a likelihood-based estimation approach for low-parametrised CLTMs in Chapter 3. For uncensored responses, the good performance of likelihood-based C(L)TMs has been investigated in Chapter 6. Furthermore, likelihood-based CLTMs were used for a re-analysis of the Perinatal Database Erlangen in Chapter 7. Likelihood-based CTMs have the important advantage that they can be extended easily to any type of censoring. Until now, the maximum likelihood approach is restricted to low-parametrised CLTMs, and the boosting approach is the preferred method for estimating C(L)TMs with more complex conditional transformation functions.

To increase interpretability in CTMs, we introduced the model class of CLTMs (Chapter 2 and Chapter 4) that allows the interpretation of the explanatory variable effects on the conditional mean and the conditional variance of the transformed response. This interpretability comes at the price of a restricted conditional transformation function. Hence, one has to be aware of the restricted model complexity that might be inappropriate in some applications. CLTMs are highly relevant in practice because they combine the increased flexibility of conditional transformation models with interpretable model results. Therefore, the introduction of CLTMs displays another important methodological emphasis of this thesis.

Comparing regression and transformation models, the perspective of transformation models is seldom taken in statistics. To stress the commonalities and to emphasis the high diversity of transformation models, we reviewed several frequently used regression models from the perspective of transformation models (Chapter 2). Thereby, our main aim was to show that all mentioned regression models share the model basis of conditional transformation models. From this awareness follows that our proposed unique likelihood-based estimation approach can be used to estimate a wide range of commonly used regression models. Moreover, the extension to any type of censoring is straightforward for each of the considered regression models.

In our analysis of the Perinatal Database Erlangen (Chapter 4 and Chapter 7) and of the patients suffering from chronic myelogenous leukaemia (Chapter 5), we considered the topic of model choice. We estimated conditional transformation models of different model complexity and standard regression models, which are usually less flexible. The performance of all models was evaluated in terms of the out-of-sample (censored) log score, which is a useful measure for the quality of estimated conditional distribution functions. This procedure can help to identify violations of important model assumptions, *e.g.*, the proportional hazards assumption in the Cox model or the homoscedasticity assumption in the linear regression model. Afterwards, there are three options how to proceed: It is save to use the standard regression model because the strict model assumptions are not violated; there are model extensions for the standard regression model that account for the violated model assumptions; the flexible conditional transformation model performs best and hence, it should be used for data analysis.

There is a great potential of enhancements for likelihood-based C(L)TMs that is left for future research. So far, we tested the performance of likelihood-based CTMs for low-parametrised CLTMs and for uncensored response variables in Chapter 6 and Chapter 7. Hence, the extension of the likelihood-based estimation approach to more flexible C(L)TMs (such models were, *e.g.*, considered in Chapter 4 and Chapter 5, and in Hothorn et al. (2014)) would be worthwhile. Therefore, questions concerning algorithmic feasibility and problems of identifiability of model components need to be solved first. We stated that likelihood-based C(L)TMs can be easily adapted to any type of censoring. The performance of likelihood-based C(L)TMs for censored response variables in terms of simulation studies and suitable applications needs to be investigated in future research. One important advantage of boosting algorithms is the intrinsic variable selection property. How variable selection should be performed in likelihood-based C(L)TMs is another important question. We presented four possible estimation strategies for likelihood-based C(L)TMs including a parsimonious parametrisation using fractional polynomials, a P-spline approach, an empirical Bayes approach, and a full Bayesian approach in Chapter 3. Thereby, a thorough investigation of the feasibility of the empirical Bayes approach and the full Bayesian approach was left for future research. For the wide applicability of likelihood-based C(L)TMs, a comprehensive software interface (*i.e.* in terms of an R package) is needed.

Furthermore, some questions concerning the estimation of CTMs in survival analysis using a component-wise boosting algorithm remain to be solved (Chapter 5). For a right-censored response variable, we included inverse probability of censoring weights into the target function to account for the censoring pattern. Thereby, we assumed that the censoring distribution is independent of all explanatory variables. This assumption might be problematic (Gerds and Schumacher, 2006) and needs further investigation. Strategies for including conditional censoring distributions into the target function would be a worthwhile development. In contrast to the likelihood-based estimation approach that can be extended to any type of censoring, we only considered right-censoring for the boosting approach. Therefore, it would be interesting if and how the boosting approach extends to other censoring or truncation mechanisms. A possible starting point is presented in Shen (2003), where the product limit estimates of left-truncated or both left-truncated and right-censored observations are expressed as inverse-probability-weighted averages. Moreover, Mackenzie (2012) previously estimated survival curves with dependent left-truncated data using Cox's model and inverse probability weighting.

In this thesis, we put transformation models into perspective by clarifying their wide applicability to continuous as well as ordinal responses. The high relevance of conditional transformation models was further supported by suggesting a unified likelihood-based estimation approach. We were successful in enhancing the interpretability in CTMs by introducing the model class of CLTMs, and the extension of CTMs to censored response variables was a further important improvement.

A. Predicting birth weight by boosting CLTMs

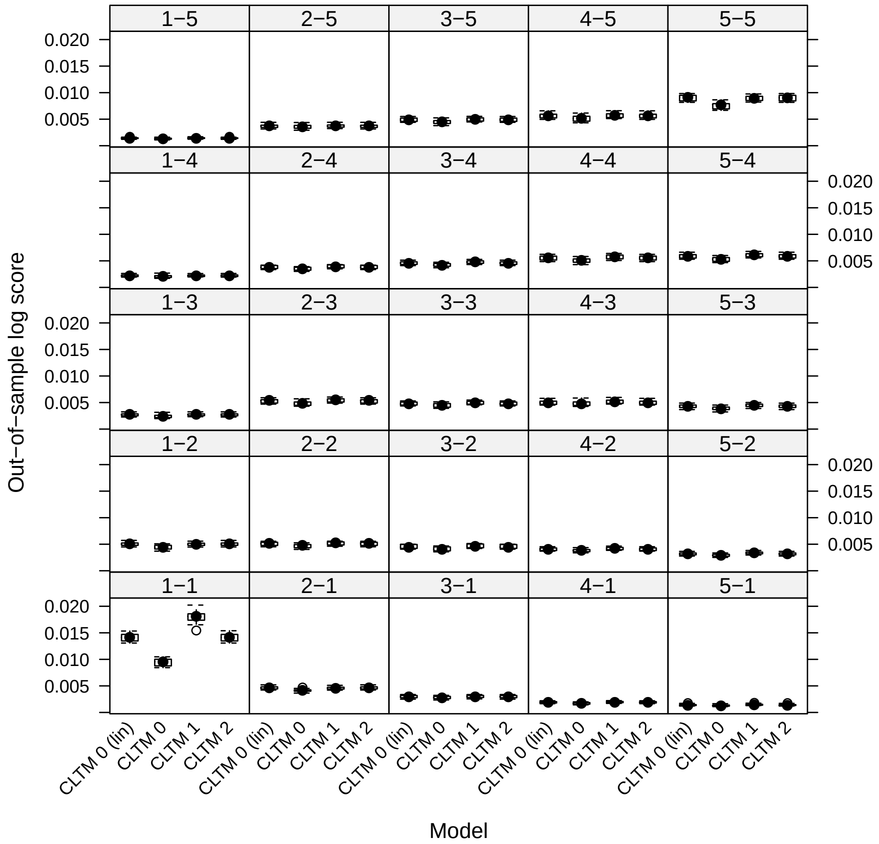


Figure A.1.: Boxplots of the out-of-sample log-scores based on 25 evaluation data sets. The log-scores were determined for the 25 categories for abdominal circumference and femur length (AC–FL) separately. Model estimation was carried out for CLTM 0 (linear), CLTM 0, CLTM 1, and CLTM 2.

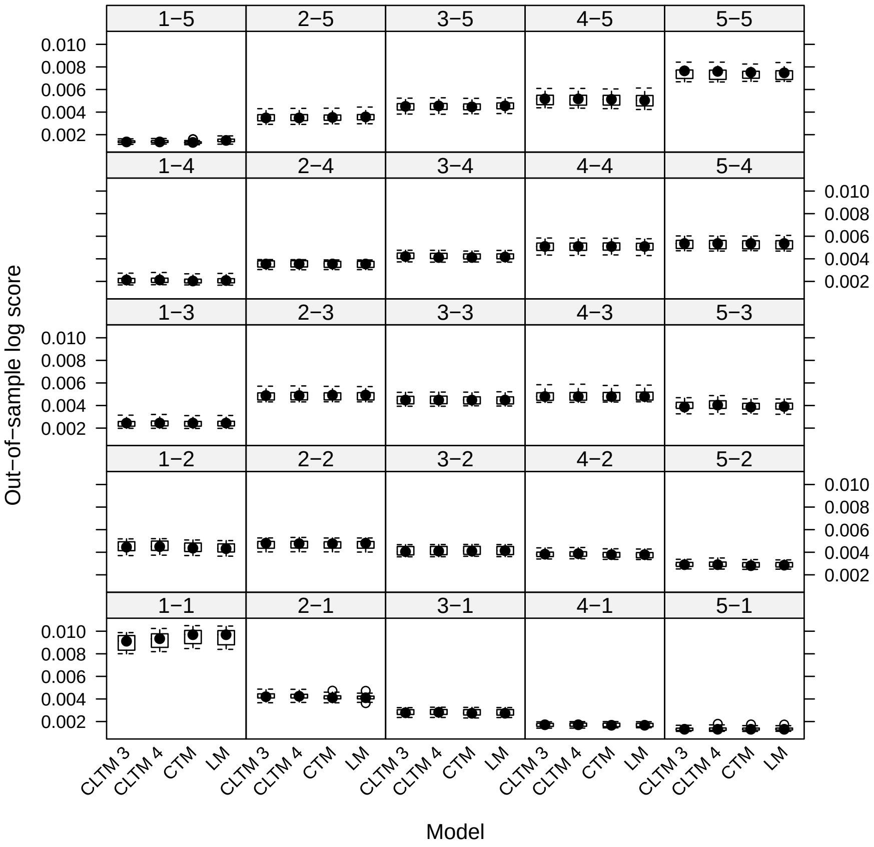


Figure A.2.: Boxplots of the out-of-sample log-scores based on 25 evaluation data sets. The log-scores were determined for the 25 categories for abdominal circumference and femur length (AC–FL) separately. Model estimation was carried out for CLTM 3, CLTM 4, CTM, and LM.

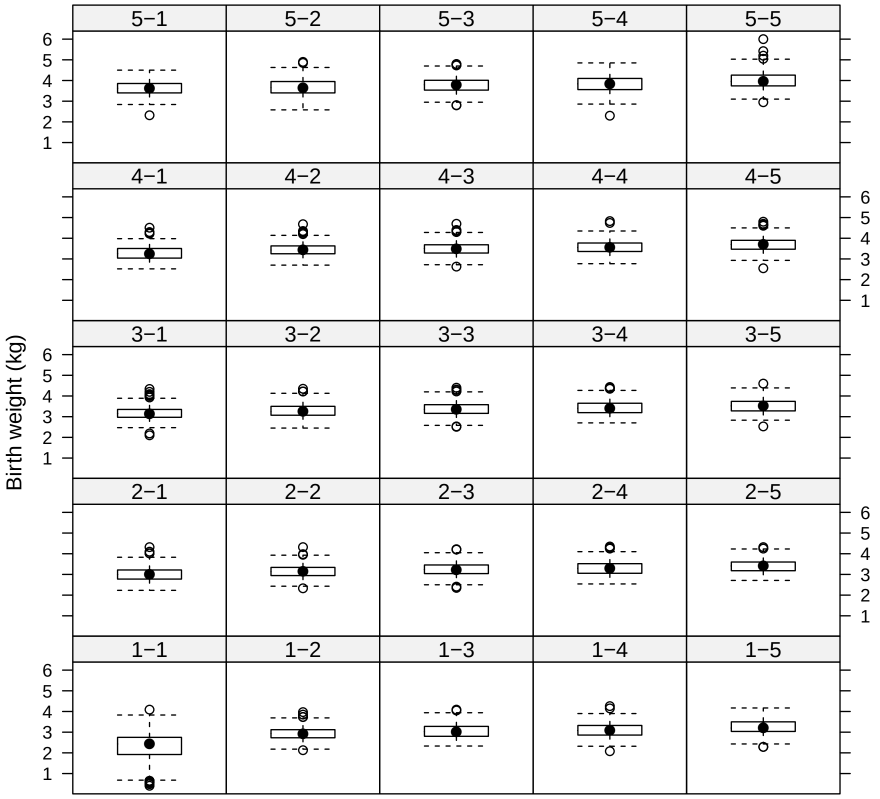


Figure A.3.: Boxplots for the birth weights in the 25 categories for abdominal circumference and femur length (AC–FL).

Table A.1.: Conditional coverage for the prediction intervals of fetuses belonging to the 25 categories defined by abdominal circumference (AC) and femur length (FL). Estimation based on the regression models CLTM 0 (lin) and CLTM 0 – CLTM 3.

AC	FL	CLTM 0 (lin)	CLTM 0	CLTM 1	CLTM 2	CLTM 3
1	1	0.826	0.826	0.587	0.826	0.783
2	1	0.784	0.784	0.838	0.784	0.784
3	1	0.905	0.857	0.905	0.905	0.857
4	1	0.933	0.800	0.933	0.933	0.800
5	1	1.000	0.818	1.000	1.000	0.909
1	2	0.944	0.833	0.944	0.944	0.833
2	2	0.952	0.881	0.976	0.952	0.881
3	2	0.884	0.721	0.907	0.884	0.744
4	2	0.903	0.871	0.903	0.903	0.903
5	2	0.885	0.885	0.962	0.885	0.885
1	3	0.957	0.870	0.957	0.957	0.870
2	3	0.973	0.892	0.973	0.973	0.892
3	3	0.923	0.897	0.974	0.923	0.897
4	3	0.947	0.842	1.000	0.947	0.842
5	3	0.857	0.714	0.857	0.857	0.714
1	4	0.800	0.800	0.867	0.800	0.733
2	4	0.938	0.875	0.938	0.938	0.875
3	4	0.875	0.825	0.925	0.875	0.825
4	4	0.841	0.841	0.886	0.841	0.909
5	4	0.857	0.833	0.881	0.857	0.833
1	5	0.800	0.600	0.900	0.800	0.600
2	5	0.880	0.840	0.880	0.880	0.840
3	5	0.970	0.879	0.970	0.970	0.879
4	5	0.935	0.913	0.957	0.935	0.913
5	5	0.817	0.817	0.850	0.817	0.867

Table A.2.: Conditional coverage for the prediction intervals of fetuses belonging to the 25 categories defined by abdominal circumference (AC) and femur length (FL). Estimation based on the regression models CLTM 4, CTM, LM, LQR, AQR.

AC	FL	CLTM 4	CTM	LM	LQR	AQR
1	1	0.826	0.870	0.815	0.772	0.739
2	1	0.784	0.757	0.865	0.757	0.757
3	1	0.857	0.857	1.000	0.857	0.857
4	1	0.800	0.867	0.933	0.867	0.867
5	1	0.909	0.909	0.909	0.818	0.818
1	2	0.833	0.861	0.778	0.833	0.806
2	2	0.881	0.881	0.833	0.881	0.881
3	2	0.744	0.721	0.930	0.721	0.721
4	2	0.871	0.871	0.742	0.871	0.903
5	2	0.885	0.846	0.846	0.846	0.846
1	3	0.870	0.913	0.739	0.870	0.957
2	3	0.892	0.892	0.919	0.892	0.892
3	3	0.897	0.897	0.769	0.872	0.897
4	3	0.842	0.842	0.895	0.842	0.842
5	3	0.714	0.743	0.857	0.714	0.714
1	4	0.800	0.800	0.800	0.800	0.800
2	4	0.875	0.906	0.750	0.906	0.906
3	4	0.825	0.825	0.750	0.825	0.825
4	4	0.864	0.841	0.864	0.841	0.795
5	4	0.833	0.857	0.690	0.786	0.810
1	5	0.600	0.600	1.000	0.600	0.600
2	5	0.840	0.840	0.880	0.840	0.760
3	5	0.879	0.848	0.727	0.788	0.818
4	5	0.913	0.913	0.717	0.935	0.870
5	5	0.850	0.783	0.767	0.733	0.750

B. Boosting algorithm for C(L)TMs

The component-wise boosting algorithm for conditional transformation models was introduced in Hothorn et al. (2014).

(Init) Initialise the parameters $\boldsymbol{\gamma}_j^{[0]} \equiv 0$ for $j = 1, \dots, J$, the step-size $\nu \in (0, 1)$ and the smoothing parameters λ_j, $j = 1, \dots, J$. Define the grid $\upsilon_1 < Y_{(1)} < \dots < Y_{(N)} \leq \upsilon_n$. Set $m = 0$.

(Gradient) Compute the negative gradient of the integrated log score:

$$\begin{aligned} U_{i\iota} \quad &:= \quad -\frac{\partial}{\partial h} \rho((Y_i \leq \upsilon_\iota, \boldsymbol{x}_i), h) \bigg|_{h=\hat{h}_{i\iota}^{[m]}} \\ &:= \quad \left\{ I(Y_i \leq \upsilon_\iota) \, \frac{F'(h(\upsilon_\iota|\boldsymbol{x}_i))}{F(h(\upsilon_\iota|\boldsymbol{x}_i))} - \; I(Y_i > \upsilon_\iota) \, \frac{F'(h(\upsilon_\iota|\boldsymbol{x}_i))}{1 - F(h(\upsilon_\iota|\boldsymbol{x}_i))} \right\} \bigg|_{h=\hat{h}_{i\iota}^{[m]}}, \end{aligned}$$

where $F'(\cdot)$ denotes the density of the link function F, and

$$\hat{h}_{i\iota}^{[m]} = \sum_{j=1}^{J} \hat{h}_j^{[m]}(\upsilon_\iota|\boldsymbol{x}_i) = \sum_{j=1}^{J} \left(\boldsymbol{b}_j(\boldsymbol{x}_i)^\top \otimes \boldsymbol{b}_0(\upsilon_\iota)^\top \right) \boldsymbol{\gamma}_j^{[m]}.$$

Fit the base-learners for $j = 1, \dots, J$:

$$\hat{\boldsymbol{\beta}}_j = \underset{\boldsymbol{\beta} \in \mathbb{R}^{K_j \cdot K_0}}{arg\,min} \sum_{i=1}^{N} \sum_{\iota=1}^{n} \omega_{i\iota} \left\{ U_{i\iota} - \left(\boldsymbol{b}_j(\boldsymbol{x}_i)^\top \otimes \boldsymbol{b}_0(\upsilon_\iota)^\top \right) \boldsymbol{\beta} \right\}^2 + \boldsymbol{\beta}^\top P_{0j} \boldsymbol{\beta}$$

with penalty matrix P_{0j}.
Select the base-learner

$$j^* = \underset{j=1,\dots,J}{arg\,min} \sum_{i=1}^{N} \sum_{\iota=1}^{n} \omega_{i\iota} \left\{ U_{i\iota} - \left(\boldsymbol{b}_j(\boldsymbol{x}_i)^\top \otimes \boldsymbol{b}_0(\upsilon_\iota)^\top \right) \hat{\boldsymbol{\beta}}_j \right\}^2.$$

(Update) the parameters $\boldsymbol{\gamma}_{j^*}^{[m+1]} = \boldsymbol{\gamma}_{j^*}^{[m]} + \nu \cdot \hat{\boldsymbol{\beta}}_{j^*}$ and keep all other parameters fixed, *i.e.* $\boldsymbol{\gamma}_j^{[m+1]} = \boldsymbol{\gamma}_j^{[m]}$, $j \neq j^*$.

Iterate (Gradient) and (Update).

(Stop) if $m = M$. Output the final model

$$\begin{aligned}
\hat{\mathbb{P}}(Y \leq \upsilon | \mathbf{X} = \boldsymbol{x}) &= F(\hat{h}^{[M]}(\upsilon|\boldsymbol{x})) = F\left(\sum_{j=1}^{J} \hat{h}_j^{[M]}(\upsilon|\boldsymbol{x})\right) \\
&= F\left(\sum_{j=1}^{J} \left(\boldsymbol{b}_j(\boldsymbol{x})^\top \otimes \boldsymbol{b}_0(\upsilon)^\top\right) \boldsymbol{\gamma}_j^{[M]}\right)
\end{aligned}$$

as a function of arbitrary $\upsilon \in \mathbb{R}$ and arbitrary explanatory variables $\boldsymbol{x}$. M denotes the previously specified maximal number of boosting iterations and displays the main tuning parameter.

References

Aalen, O. O. (1988). Heterogeneity in survival analysis. *Statistics in Medicine 7*(11), 1121–1137.

Andersen, P. K., E. Christensen, L. Fauerholdt, and P. Schlichting (1983). Measuring prognosis using the proportional hazards model. *Scandinavian Journal of Statistics 10*(1), 49–52.

Anderson, K. M. (1991). A nonproportional hazards Weibull accelerated failure time regression model. *Biometrics 47*(1), 281–288.

Banerjee, T., M.-H. Chen, D. K. Dey, and S. Kim (2007). Bayesian analysis of generalized odds-rate hazards models for survival data. *Lifetime Data Analysis 13*(2), 241–260.

Beliakov, G. (2000). Shape preserving approximation using least squares splines. *Approximation Theory and its Applications 16*(4), 80–98.

Bennett, S. (1983a). Analysis of survival data by the proportional odds model. *Statistics in Medicine 2*(2), 273–277.

Bennett, S. (1983b). Log-logistic regression models for survival data. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 32*(2), 165–171.

Bernstein, I. M., J. D. Horbar, G. J. Badger, A. Ohlsson, and A. Golan (2000). Morbidity and mortality among very-low-birth-weight neonates with intrauterine growth restriction. *American Journal of Obstetrics and Gynecology 182*(1), 198–206.

Bickel, P. J. and K. A. Doksum (1981). An analysis of transformations revisited. *Journal of the American Statistical Association 76*(374), 296–311.

Boulet, S. L., G. R. Alexander, H. M. Salihu, and M. Pass (2003). Macrosomic births in the United States: Determinants, outcomes, and proposed grades of risk. *American Journal of Obstetrics and Gynecology 188*(5), 1372–1378.

Box, G. E. P. and D. R. Cox (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological) 26*(2), 211–252.

Breslow, N. (1972). Contribution to the discussion of the paper by D. R. Cox, Regression models and life tables. *Journal of the Royal Statistical Society. Series B (Methodological) 34*(2), 216–217.

Buckley, J. and I. James (1979). Linear regression with censored data. *Biometrika 66*(3), 429–436.

Bühlmann, P. and T. Hothorn (2007). Boosting algorithms: Regularization, prediction and model fitting (with discussion). *Statistical Science 22*(4), 477–505.

Cai, T. and R. A. Betensky (2003). Hazard regression for interval-censored data with penalized spline. *Biometrics 59*(3), 570–579.

Cai, T. and S. Cheng (2004). Semiparametric regression analysis for doubly censored data. *Biometrika 91*(2), 277–290.

Cai, T., S. C. Cheng, and L. J. Wei (2002). Semiparametric mixed-effects models for clustered failure time data. *Journal of the American Statistical Association 97*(458), 514–522.

Cai, T., L. J. Wei, and M. Wilcox (2000). Semiparametric regression analysis for clustered failure time data. *Biometrika 87*(4), 867–878.

Cai, Z. (2002). Regression quantiles for time series. *Econometric Theory 18*(01), 169–192.

Cai, Z. and Y. Sun (2003). Local linear estimation for time-dependent coefficients in Cox's regression models. *Scandinavian Journal of Statistics 30*(1), 93–111.

Carroll, R. J. and D. Ruppert (1982). Robust estimation in heteroscedastic linear models. *The Annals of Statistics 10*(2), 429–441.

Chen, K., Z. Jin, and Z. Ying (2002). Semiparametric analysis of transformation models with censored data. *Biometrika 89*(3), 659–668.

Chen, K. and X. Tong (2010). Varying coefficient transformation models with censored data. *Biometrika 97*(4), 969–976.

Chen, M.-H., X. Tong, and L. Zhu (2013). A linear transformation model for multivariate interval-censored failure time data. *Canadian Journal of Statistics 41*(2), 275–290.

Chen, Y. Q., N. Hu, S.-C. Cheng, P. Musoke, and L. P. Zhao (2012). Estimating regression parameters in an extended proportional odds model. *Journal of the American Statistical Association 107*(497), 318–330.

Cheng, S. C., L. J. Wei, and Z. Ying (1995). Analysis of transformation models with censored data. *Biometrika 82*(4), 835–845.

Cheng, S. C., L. J. Wei, and Z. Ying (1997). Predicting survival probabilities with semiparametric transformation models. *Journal of the American Statistical Association 92*(437), 227–235.

Chernozhukov, V. and H. Hong (2002). Three-step censored quantile regression and extramarital affairs. *Journal of the American Statistical Association 97*(459), 872–882.

Choi, J., A. B. Lawson, B. Cai, M. Hossain, R. S. Kirby, and J. Liu (2012). A Bayesian latent model with spatio-temporally varying coefficients in low birth weight incidence data. *Statistical Methods in Medical Research 21*(5), 445–456.

Choi, S. and X. Huang (2012). A general class of semiparametric transformation frailty models for nonproportional hazards survival data. *Biometrics 68*(4), 1126–1135.

Clayton, D. and J. Cuzick (1985). Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society. Series A (General) 148*(2), 82–117.

Cole, T. J. (1988). Fitting smoothed centile curves to reference data. *Journal of the Royal Statistical Society. Series A (Statistics in Society) 151*(3), 385–418.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological) 34*(2), 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika 62*(2), 269–276.

Cox, D. R. and D. Oakes (1984). *Analysis of Survival Data.* London: Chapman&Hall.

Crowther, M. J. and P. C. Lambert (2014). A general framework for parametric survival analysis. *Statistics in Medicine*. DOI: 10.1002/sim.6300.

Dabrowska, D. M. (1987). Non-parametric regression with censored survival time data. *Scandinavian Journal of Statistics 14*(3), 181–197.

Dabrowska, D. M. (1989). Uniform consistency of the kernel conditional Kaplan-Meier estimate. *The Annals of Statistics 17*(3), 1157–1167.

Dammer, U., T. W. Goecke, F. Voigt, M. Schmid, A. Mayr, R. L. Schild, M. W. Beckmann, and F. Faschingbauer (2013). Sonographic weight estimation in fetuses with breech presentation. *Archives of Gynecology and Obstetrics 287*(5), 851–858.

Davis, H. T. and M. L. Feldstein (1979). The generalized Pareto law as a model for progressively censored survival data. *Biometrika 66*(2), 299–306.

de Castro, M., M.-H. Chen, J. G. Ibrahim, and J. P. Klein (2014). Bayesian transformation models for multivariate survival data. *Scandinavian Journal of Statistics 41*(1), 187–199.

Dette, H. and S. Volgushev (2008). Non-crossing non-parametric estimates of quantile curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(3), 609–627.

Doksum, K. A. and M. Gasko (1990). On a correspondence between models in binary regression analysis and in survival analysis. *International Statistical Review 58*(3), 243–252.

Dudley, N. J. (2005). A systematic review of the ultrasound estimation of fetal weight. *Ultrasound in Obstetrics and Gynecology 25*(1), 80–89.

Ecker, J. L., J. A. Greenberg, E. R. Norwitz, A. S. Nadel, and J. T. Repke (1997). Birth weight as a predictor of brachial plexus injury. *Obstetrics & Gynecology 89*(5, Part 1), 643–647.

Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing using B-splines and penalties. *Statistical Science 11*(2), 89–121.

Fahrmeir, L., T. Kneib, and S. Lang (2004). Penalized structured additive regression for space-time data: A Bayesian perspective. *Statistica Sinica 14*, 715–745.

Fahrmeir, L., T. Kneib, S. Lang, and B. Marx (2013). *Regression - Models, methods and applications*. Springer.

Fan, C. and J. P. Fine (2013). Linear transformation model with parametric covariate transformations. *Journal of the American Statistical Association 108*(502), 701–712.

Faschingbauer, F., B. Yazdi, T. W. Goecke, M. W. Beckmann, J. Siemer, M. Schmid, A. Mayr, and R. L. Schild (2012). A new formula for optimized weight estimation in extreme fetal macrosomia ($\geq$ 4500 g). *European Journal of Ultrasound 33*(5), 480–488.

Fenske, N., T. Kneib, and T. Hothorn (2011). Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. *Journal of the American Statistical Association 106*(494), 494–510.

Fine, J. P. (2001). Regression modeling of competing crude failure probabilities. *Biostatistics 2*(1), 85–97.

Fine, J. P., Z. Ying, and L. G. Wei (1998). On the linear transformation model for censored data. *Biometrika 85*(4), 980–986.

Gerds, T. A. (2013). *prodlim: Product Limit Estimation for event history and survival analysis.* R package version 1.3.7. Available from: http://CRAN.R-project.org/package=prodlim.

Gerds, T. A. and M. Schumacher (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal 48*(6), 1029–1040.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*(477), 359–378.

González Manteiga, W. and C. Cadarso-Suarez (1994). Asymptotic properties of a generalized Kaplan-Meier estimator with some applications. *Journal of Nonparametric Statistics 4*(1), 65–78.

Graf, E., C. Schmoor, W. Sauerbrei, and M. Schumacher (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine 18*(17–18), 2529–2545.

Grambsch, P. M. and T. M. Therneau (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika 81*(3), 515–526.

Gu, M., L. Sun, and G. Zuo (2005). A baseline-free procedure for transformation models under interval censorship. *Lifetime Data Analysis 11*(4), 473–488.

Gu, M., Y. Wu, and B. Huang (2014). Partial marginal likelihood estimation for general transformation models. *Journal of Multivariate Analysis 123*, 1–18.

Hadlock, F. P., R. B. Harrist, R. S. Sharman, R. L. Deter, and S. K. Park (1985). Estimation of fetal weight with the use of head, body, and femur measurements – A prospective study. *American Journal of Obstetrics and Gynecology 151*(3), 333–337.

Hall, P. and H.-G. Müller (2003). Order-preserving nonparametric regression, with applications to conditional distribution and quantile function estimation. *Journal of the American Statistical Association 98*(463), 598–608.

Hall, P., R. C. L. Wolff, and Q. Yao (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association 94*(445), 154–163.

Hanson, T. and M. Yang (2007). Bayesian semiparametric proportional odds models. *Biometrics 63*(1), 88–95.

Hart, N. C., A. Hilbert, B. Meurer, M. Schrauder, M. Schmid, J. Siemer, M. Voigt, and R. L. Schild (2010). Macrosomia: A new formula for optimized fetal weight estimation. *Ultrasound in Obstetrics and Gynecology 35*(1), 42–47.

Hasford, J., M. Pfirrmann, R. Hehlmann, and Others (1998). A new prognostic score for survival of patients with chronic myeloid leukemia treated with interferon alfa. *Journal of the National Cancer Institute 90*(11), 850–858.

Hastie, T. and R. Tibshirani (1986). Generalized additive models. *Statistical Science 1*(3), 297–310.

Hehlmann, R., H. Heimpel, J. Hasford, and Others (1994). Randomized comparison of interferon-α with busulfan and hydroxyurea in chronic myelogenous leukemia. *Blood 84*(12), 4064–4077.

Herberich, E. and T. Hothorn (2012). Dunnett-type inference in the frailty Cox model with covariates. *Statistics in Medicine 31*(1), 45–55.

Honoré, B., S. Khan, and J. L. Powell (2002). Quantile regression under random censoring. *Journal of Econometrics 109*(1), 67–105.

Hoopmann, M., H. Abele, N. Wagner, D. Wallwiener, and K. O. Kagan (2010). Performance of 36 different weight estimation formulae in fetuses with macrosomia. *Fetal Diagnosis and Therapy 27*(4), 204–213.

Horowitz, J. L. (2009). *Semiparametric and Nonparametric Methods in Econometrics.* New York: Springer.

Hothorn, T. (2012). *ctm: Conditional Transformation Models.* R package version 0.0-3. Available from: https://r-forge.r-project.org/projects/ctm.

Hothorn, T. (2013). *ctmDevel: Conditional Transformation Models.* R package version 0.1-0. SVN revision 56. Available from: https://r-forge.r-project.org/projects/ctm.

Hothorn, T., P. Bühlmann, S. Dudoit, A. Molinaro, and M. J. van der Laan (2006). Survival ensembles. *Biostatistics 7*(3), 355–373.

Hothorn, T., P. Bühlmann, T. Kneib, M. Schmid, and B. Hofner (2013). *mboost: Model-Based Boosting.* R package version 2.2-3. Available from: http://CRAN.R-project.org/package=mboost.

Hothorn, T., K. Hornik, and A. Zeileis (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics 15*(3), 651–674.

Hothorn, T., T. Kneib, and P. Bühlmann (2013). Conditional transformation models by example. In V. M. R. Muggeo, V. Capursi, G. Boscaino, and G. Lovison (Eds.), *Proceedings of the 28th International Workshop on Statistical Modelling*, pp. 15–26. Universitá Degli Studi Di Palermo.

Hothorn, T., T. Kneib, and P. Bühlmann (2014). Conditional transformation models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(1), 3–27.

Hothorn, T., B. Lausen, A. Benner, and M. Radespiel-Tröger (2004). Bagging survival trees. *Statistics in Medicine 23*(1), 77–91.

Huber-Carol, C. and I. Vonta (2004). Frailty models for arbitrarily censored and truncated data. *Lifetime Data Analysis 10*(4), 369–388.

Hunter, D. R. and K. Lange (2002). Computing estimates in the proportional odds model. *Annals of the Institute of Statistical Mathematics 54*(1), 155–168.

Iglesias Pérez, C. and W. González Manteiga (1999). Strong representation of a generalized product-limit estimator for truncated and censored data with some applications. *Journal of Nonparametric Statistics 10*(3), 213–244.

Ishwaran, H., U. B. Kogalur, E. H. Blackstone, and M. S. Lauer (2008). Random survival forests. *The Annals of Applied Statistics 2*(3), 841–860.

Jin, Z., D. Y. Lin, L. J. Wei, and Z. Ying (2003). Rank–based inference for the accelerated failure time model. *Biometrika 90*(2), 341–353.

Kalbfleisch, J. D. and R. L. Prentice (1980). *The Statistical Analysis of Failure Time Data.* New York: John Wiley & Sons.

Kaplan, E. L. and P. Meier (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association 53*(282), 457–481.

Kay, R. (1977). Proportional hazard regression models and the analysis of censored survival data. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 26*(3), 227–237.

Khan, S. and E. Tamer (2007). Partial rank estimation of duration models with general forms of censoring. *Journal of Econometrics 136*(1), 251–280.

Klein, J. P. and M. L. Moeschberger (2003). *Survival Analysis: Techniques for Censored and Truncated Data* (2nd ed.). New York: Springer.

Koenker, R. (2005). *Quantile Regression.* Economic Society Monographs. New York: Cambridge University Press.

Koenker, R. (2012). *quantreg: Quantile Regression.* R package version 4.94. Available from: http://CRAN.R-project.org/package=quantreg.

Koenker, R., P. Ng, and S. Portnoy (1994). Quantile smoothing splines. *Biometrika 81*(4), 673–680.

Kosorok, M. R., B. L. Lee, and J. P. Fine (2004). Robust inference for univariate proportional hazards frailty regression models. *The Annals of Statistics 32*(4), 1448–1491.

Van der Laan, M. J. and J. M. Robins (2003). *Unified Methods for Censored Longitudinal Data and Causality.* New York: Springer.

Lai, T. L. and Z. Ying (1992). Linear rank statistics in regression analysis with censored or truncated data. *Journal of Multivariate Analysis 40*(1), 13–45.

Lam, K. F., Y. W. Lee, and T. L. Leung (2002). Modeling multivariate survival data by a semiparametric random effects proportional odds model. *Biometrics 58*(2), 316–323.

Lam, K. F. and T. L. Leung (2001). Marginal likelihood estimation for proportional odds models with right censored data. *Lifetime Data Analysis 7*(1), 39–54.

Lang, S. and A. Brezger (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics 13*(1), 183–212.

Lee, K. H., S. Chakraborty, and J. Sun (2011). Bayesian variable selection in semiparametric proportional hazards model for high dimensional survival data. *The International Journal of Biostatistics 7*(1). Article 21.

Lee, S. (2008). Estimating panel data duration models with censored data. *Econometric Theory 24*(05), 1254–1276.

Li, G. and H. Doss (1995). An approach to nonparametric regression for life history data using local linear fitting. *The Annals of Statistics 23*(3), 787–823.

Li, J., M. Gu, and T. Hu (2012). General partially linear varying-coefficient transformation models for ranking data. *Journal of Applied Statistics 39*(7), 1475–1488.

Li, Q. and J. S. Racine (2008). Nonparametric estimation of conditional cdf and quantile functions with mixed categorical and continuous data. *Journal of Business & Economic Statistics 26*(4), 423–434.

Lin, D. Y., L. J. Wei, and Z. Ying (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika 80*(3), 557–572.

Lin, D. Y. and Z. Ying (1994). Semiparametric analysis of the additive risk model. *Biometrika 81*(1), 61–71.

Lin, J., D. Sinha, S. Lipsitz, and A. Polpo (2012). Semiparametric Bayesian survival analysis using models with log-linear median. *Biometrics 68*(4), 1136–1145.

Linton, O., S. Sperlich, and I. van Keilegom (2008). Estimation of a semiparametric transformation model. *The Annals of Statistics 36*(2), 686–718.

Liu, M. and Z. Ying (2007). Joint analysis of longitudinal data with informative right censoring. *Biometrics 63*(2), 363–371.

Liu, X. and D. Zeng (2013). Variable selection in semiparametric transformation models for right-censored data. *Biometrika, in press*. DOI: 10.1093/biomet/ast029.

Lu, W. (2005). Marginal regression of multivariate event times based on linear transformation models. *Lifetime Data Analysis 11*(3), 389–404.

Lu, W. and L. Li (2008). Boosting method for nonlinear transformation models with censored survival data. *Biostatistics 9*(4), 658–667.

Lu, W. and Y. Liang (2006). Empirical likelihood inference for linear transformation models. *Journal of Multivariate Analysis 97*(7), 1586–1599.

Lu, W. and A. A. Tsiatis (2006). Semiparametric transformation models for the case-cohort study. *Biometrika 93*(1), 207–214.

Lu, W. and H. H. Zhang (2010). On estimation of partially linear transformation models. *Journal of the American Statistical Association 105*(490), 683–691.

Ma, J., S. Heritier, and S. N. Lô (2014). On the maximum penalized likelihood approach for proportional hazard models with right censored survival data. *Computational Statistics & Data Analysis 74*, 142–156.

Mackenzie, T. (2012). Survival curve estimation with dependent left truncated data using Cox's model. *The International Journal of Biostatistics 8*(1). Article 29.

Mackillop, W. J. and C. F. Quirt (1997). Measuring the accuracy of prognostic judgments in oncology. *Journal of Clinical Epidemiology 50*(1), 21–29.

MacKinnon, J. G. and L. Magee (1990). Transforming the dependent variable in regression models. *International Economic Review 31*(2), 315–339.

Mallick, B. K. and S. Walker (2003). A Bayesian semiparametric transformation model incorporating frailties. *Journal of Statistical Planning and Inference 112*(1–2), 159–174.

Manning, W. G. and J. Mullahy (2001). Estimating log models: To transform or not to transform? *Journal of Health Economics 20*(4), 461–494.

Mayr, A., N. Fenske, B. Hofner, T. Kneib, and M. Schmid (2012). Generalized additive models for location, scale and shape for high dimensional data – a flexible approach based on boosting. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 61*(3), 403–427.

Mayr, A., T. Hothorn, and N. Fenske (2012). Prediction intervals for future BMI values of individual children - a non-parametric approach by quantile boosting. *BMC Medical Research Methodology 12, Part 6.* DOI: 10.1186/1471-2288-12-6.

McCormick, M. C. (1985). The contribution of low birth weight to infant mortality and childhood morbidity. *New England Journal of Medicine 312*(2), 82–90.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological) 42*(2), 109–142.

McCullagh, P. (1984). Generalized linear models. *European Journal of Operational Research 16*(3), 285–292.

McGilchrist, C. A. and C. W. Aisbett (1991). Regression with frailty in survival analysis. *Biometrics 47*(2), 461–466.

McIntire, D. D., S. L. Bloom, B. M. Casey, and K. J. Leveno (1999). Birth weight in relation to morbidity and mortality among newborn infants. *New England Journal of Medicine 340*(16), 1234–1238.

McKeague, I. W. and K. J. Utikal (1990). Inference for a nonlinear counting process regression model. *The Annals of Statistics 18*(3), 1172–1187.

Meinshausen, N. (2006). Quantile regression forests. *The Journal of Machine Learning Research 7*, 983–999.

Mogensen, U. B., H. Ishwaran, and T. A. Gerds (2012). Evaluating random forests for survival analysis using prediction error curves. *Journal of Statistical Software 50*(11), 1–23.

Montgomery, D. C., E. A. Peck, and G. G. Vining (2012). *Introduction to Linear Regression Analysis* (5th ed.). Hoboken, New Jersey: Wiley.

Möst, L. and T. Hothorn (2014). Likelihood-based conditional transformation models. *Working paper*.

Möst, L. and T. Hothorn (2015). Conditional transformation models for survivor function estimation. *International Journal of Biostatistics. To appear*. DOI: 10.1515/ijb-2014-0006.

Möst, L., M. Schmid, F. Faschingbauer, and T. Hothorn (2014). Predicting birth weight with conditionally linear transformation models. *Statistical Methods in Medical Research. To appear*. DOI: 10.1177/0962280214532745.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics 33*(3), 341–365.

Murphy, S. A., A. J. Rossini, and A. W. van der Vaart (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association 92*(439), 968–976.

Ng'Andu, N. H. (1997). An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox's model. *Statistics in Medicine 16*(6), 611–626.

Peng, L. and Y. Huang (2008). Survival analysis with quantile regression models. *Journal of the American Statistical Association 103*(482), 637–649.

Peterson, B. and F. E. Harrell (1990). Partial proportional odds models for ordinal response variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 39*(2), 205–217.

Pettitt, A. N. (1984). Proportional odds models for survival data and estimates using ranks. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 33*(2), 169–175.

Polpo, A., C. P. de Campos, D. Sinha, S. Lipsitz, and J. Lin (2014). Transform both sides model: A parametric approach. *Computational Statistics & Data Analysis 71*, 903–913.

Portnoy, S. (2003). Censored regression quantiles. *Journal of the American Statistical Association 98*(464), 1001–1012.

Powell, J. L. (1986). Censored regression quantiles. *Journal of Econometrics 32*(1), 143–155.

Prentice, R. L. (1973). Exponential survivals with censoring and explanatory variables. *Biometrika 60*(2), 279–288.

Prentice, R. L. (1978). Linear rank tests with right censored data. *Biometrika 65*(1), 167–179.

R Core Team (2014). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Rigby, R. A. and D. M. Stasinopoulos (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 54*(3), 507–554.

Ritov, Y. (1990). Estimation in a linear regression model with censored data. *The Annals of Statistics 18*(1), 303–328.

Robins, J. M. and D. M. Finkelstein (2000). Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics 56*(3), 779–788.

Royston, P. and D. G. Altman (1994). Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 43*(3), 429–467.

Royston, P., G. Ambler, and W. Sauerbrei (1999). The use of fractional polynomials to model continuous risk variables in epidemiology. *International Journal of Epidemiology 28*(5), 964–974.

Royston, P. and W. Sauerbrei (2007). Improving the robustness of fractional polynomial models by preliminary covariate tansformation: A pragmatic approach. *Computational Statistics & Data Analysis 51*(9), 4240–4253.

Sabbagha, R. E., J. Minogue, R. K. Tamura, and S. A. Hungerford (1989). Estimation of birth weight by use of ultrasonographic formulas targeted to large-, appropriate-, and small-for-gestational-age fetuses. *American Journal of Obstetrics & Gynecology 160*(4), 854–862.

Sappenfield, W. M., J. W. Buehler, N. J. Binkin, C. J. Hogue, L. T. Strauss, and J. C. Smith (1987). Differences in neonatal and postneonatal mortality by race, birth weight, and gestational age. *Public Health Reports 102*(2), 182–192.

Sargent, D. J. (1997). A flexible approach to time-varying coefficients in the Cox regression setting. *Lifetime Data Analysis 3*(1), 13–25.

Sauerbrei, W. and P. Royston (1999). Building multivariable prognostic and diagnostic models: Transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 162*(1), 71–94.

Scheike, T. H. (2006). A flexible semiparametric transformation model for survival data. *Lifetime Data Analysis 12*(4), 461–480.

Scheike, T. H. and T. Martinussen (2004). On estimation and tests of time-varying effects in the proportional hazards model. *Scandinavian Journal of Statistics 31*(1), 51–62.

Schemper, M. and R. Henderson (2000). Predictive accuracy and explained variation in Cox regression. *Biometrics 56*(1), 249–255.

Schild, R. L., M. Maringa, J. Siemer, B. Meurer, N. Hart, T. W. Goecke, M. Schmid, T. Hothorn, and M. E. Hansmann (2008). Weight estimation by three-dimensional ultrasound imaging in the small fetus. *Ultrasound in Obstetrics and Gynecology 32*(2), 168–175.

Schmid, M. and T. Hothorn (2008). Boosting additive models using component-wise P-splines. *Computational Statistics & Data Analysis 53*(2), 298–311.

Schnabel, S. K. and P. H. C. Eilers (2013). Simultaneous estimation of quantile curves using quantile sheets. *AStA Advances in Statistical Analysis 97*(1), 77–87.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika 69*(1), 239–241.

Scioscia, M., A. Vimercati, O. Ceci, M. Vicino, and L. E. Selvaggi (2008). Estimation of birth weight by two-dimensional ultrasonography: A critical appraisal of its accuracy. *Obstetrics & Gynecology 111*(1), 57–65.

Shen, P.-S. (2003). The product-limit estimate as an inverse-probability-weighted average. *Communications in Statistics - Theory and Methods 32*(6), 1119–1133.

Shen, P.-S. (2012a). Analysis of left-truncated right-censored or doubly censored data with linear transformation models. *TEST 21*(3), 584–603.

Shen, P.-S. (2012b). Semiparametric mixed-effects models for clustered doubly censored data. *Journal of Applied Statistics 39*(9), 1881–1892.

Shen, P.-S. (2013). Regression analysis of interval censored and doubly truncated data with linear transformation models. *Computational Statistics 28*(2), 581–596.

Siemer, J., N. Egger, N. Hart, B. Meurer, A. Müller, O. Dathe, T. Goecke, and R. L. Schild (2008). Fetal weight estimation by ultrasound: Comparison of 11 different formulae and examiners with differing skill levels. *European Journal of Ultrasound 29*(2), 159–164.

Siggelkow, W., M. Schmidt, C. Skala, D. Boehm, S. von Forstner, H. Koelbl, and A. Tresch (2011). A new algorithm for improving fetal weight estimation from ultrasound data at term. *Archives of Gynecology and Obstetrics 283*(3), 469–474.

Simpkin, A. and J. Newell (2013). An additive penalty P-spline approach to derivative estimation. *Computational Statistics & Data Analysis 68*, 30–43.

Singh, K. P., C. M.-S. Lee, and E. O. George (1988). On generalized log-logistic model for censored survival data. *Biometrical Journal 30*(7), 843–850.

Slud, E. V. and F. Vonta (2004). Consistency of the NPML estimator in the right-censored transformation model. *Scandinavian Journal of Statistics 31*(1), 21–41.

Song, X., S. Ma, J. Huang, and X.-H. Zhou (2007). A semiparametric approach for the nonparametric transformation survival model with multiple covariates. *Biostatistics 8*(2), 197–211.

Spierdijk, L. (2008). Nonparametric conditional hazard rate estimation: A local linear approach. *Computational Statistics & Data Analysis 52*(5), 2419–2434.

Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics 8*, 25.

Sun, Y., R. Sundaram, and Y. Zhao (2009). Empirical likelihood inference for the Cox model with time-dependent coefficients via local partial likelihood. *Scandinavian Journal of Statistics 36*(3), 444–462.

Therneau, T. M. (2013). Survival analysis. R package version 2.37-4. Available from: http://CRAN.R-project.org/package=survival.

Tian, L., D. Zucker, and L. J. Wei (2005). On the Cox model with time-varying regression coefficients. *Journal of the American Statistical Association 100*(469), 172–183.

Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics 18*(1), 354–372.

Van der Vaart, A. and M. J. van der Laan (2006). Estimating a survival distribution with current status data and high-dimensional covariates. *The International Journal of Biostatistics 2*(1). Article 9.

Vaida, F. and R. Xu (2000). Proportional hazards model with random effects. *Statistics in Medicine 19*(24), 3309–3324.

Walker, S. and B. K. Mallick (1999). A Bayesian semiparametric accelerated failure time model. *Biometrics 55*(2), 477–483.

Wang, H. J. and L. Wang (2009). Locally weighted censored quantile regression. *Journal of the American Statistical Association 104*(487), 1117–1128.

Wei, L. J. (1992). The accelerated failure time model: A useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine 11*(14-15), 1871–1879.

Wei, Y., A. Pere, R. Koenker, and X. He (2006). Quantile regression methods for reference growth charts. *Statistics in Medicine 25*(8), 1369–1382.

Wey, A., L. Wang, and K. Rudser (2014). Censored quantile regression with recursive partitioning-based weights. *Biostatistics 15*(1), 170–181.

Wu, C. O., X. Tian, and J. Yu (2010). Nonparametric estimation for time-varying transformation models with longitudinal data. *Journal of Nonparametric Statistics 22*(2), 133–147.

Xu, R. and J. O'Quigley (2000). Estimating average regression effect under nonproportional hazards. *Biostatistics 1*(4), 423–439.

Yang, S. and R. L. Prentice (1999). Semiparametric inference in the proportional odds regression model. *Journal of the American Statistical Association 94*(445), 125–136.

Yin, G. and D. Zeng (2006). Efficient algorithm for computing maximum likelihood estimates in linear transformation models. *Journal of Computational and Graphical Statistics 15*(1), 228–245.

Yu, W., Y. Sun, and M. Zheng (2011). Empirical likelihood method for linear transformation models. *Annals of the Institute of Statistical Mathematics 63*(2), 331–346.

Zeileis, A., C. Kleiber, and S. Jackman (2008). Regression models for count data in R. *Journal of Statistical Software 27*(8), 1–25.

Zeng, D., Q. Chen, and J. G. Ibrahim (2009). Gamma frailty transformation models for multivariate survival times. *Biometrika 96*(2), 277–291.

Zeng, D. and D. Y. Lin (2006). Efficient estimation of semiparametric transformation models for counting processes. *Biometrika 93*(3), 627–640.

Zeng, D. and D. Y. Lin (2007a). Efficient estimation for the accelerated failure time model. *Journal of the American Statistical Association 102*(480), 1387–1396.

Zeng, D. and D. Y. Lin (2007b). Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(4), 507–564.

Zeng, D., D. Y. Lin, and X. Lin (2008). Semiparametric transformation models with random effects for clustered failure time data. *Statistica Sinica 18*(1), 355–377.

Zeng, D., D. Y. Lin, and G. Yin (2005). Maximum likelihood estimation for the proportional odds model with random effects. *Journal of the American Statistical Association 100*(470), 470–483.

Zhang, B., X. Tong, J. Zhang, C. Wang, and J. Sun (2013). Efficient estimation for linear transformation models with current status data. *Communications in Statistics - Theory and Methods 42*(17), 3191–3203.

Zhang, H. H., W. Lu, and H. Wang (2010). On sparse estimation for semiparametric linear transformation models. *Journal of Multivariate Analysis 101*(7), 1594–1606.

Zhang, M. and M. Davidian (2008). "Smooth" semiparametric regression analysis for arbitrarily censored time-to-event data. *Biometrics 64*(2), 567–576.

Zhang, Z. (2009). Linear transformation models for interval-censored data: Prediction of survival probability and model checking. *Statistical Modelling 9*(4), 321–343.

Zhang, Z., L. Sun, X. Zhao, and J. Sun (2005). Regression analysis of interval-censored failure time data with linear transformation models. *The Canadian Journal of Statistics 33*(1), 61–70.

Zhang, Z. and Y. Zhao (2013). Empirical likelihood for linear transformation models with interval-censored failure time data. *Journal of Multivariate Analysis 116*, 398–409.

Zhao, X., X. Zhou, and X. Wu (2007). Local linear regression in proportional hazards model with censored data. *Communications in Statistics - Theory and Methods 36*(15), 2761–2776.

Zhao, Y. (2010). Semiparametric inference for transformation models via empirical likelihood. *Journal of Multivariate Analysis 101*(8), 1846–1858.

Zucker, D. M. and A. F. Karr (1990). Nonparametric survival analysis with time-dependent covariate effects: A penalized partial likelihood approach. *The Annals of Statistics 18*(1), 329–353.

Zucker, D. M. and S. Yang (2006). Inference for a family of survival models encompassing the proportional hazards and proportional odds models. *Statistics in Medicine 25*(6), 995–1014.

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12.07.11, §8, Abs. 2 Pkt. .5.)

Hiermit erkläre ich an Eidesstatt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, den 2. Dezember 2014

Lisa Möst