Transformation Analysis

Torsten Hothorn

October 6, 2020

Contents

List of Figures

Preface

Today is March 17, 2020, the first day of the lock down period in Switzerland due to the COVID-19 outbreak. This morning, I taught survival analysis to a completely empty class, because no student is allowed anywhere near campus. Later today I struggled uploading the videos I produced. Most of the existing video streaming facilities suffered extreme delays as it seemed I was not the only lecturer trying to communicate with students online.

The experiences I made today finally convinced me that I need to stop procrastinating writing this book. The necessary LaTeXand R infrastructure was set-up in April 2017, but there was always something more interesting to do than to sit down and go through the pain of writing. Now, my students suffer from probably naive video taping and are in need for written lecture notes for their lectures. The university rector announced today that the University of Zurich operate on skeleton staff, on-site activities have been reduced drastically. School for the kids is off, and I guess such drastic changes in our daily routines also justify to shift priorities from writing yet another research paper to drafting this book.

The book covers two courses I teach: Generalised Regression and Survival Analysis. In a nutshell, Generalised Regression introduces many regression models as special cases of transformation models. Model interpretation is put first in this class, less emphasis is given to actually estimating model parameters. Details of likelihood inference are discussed in Survival Analysis, especially different forms of the likelihood for randomly censored observations or in the presence of truncation. Because students have already been exposed to transformation models, it is, or shall I write “should be”, easy for them to understand Cox’ proportional hazards model.

This book covers more topics than discussed in the lecture room: Transformation models for count data, correlated observations, machine-learning approaches for parameter estimation and other advanced topics target students interested in going beyond the syllabus. The book almost exclusively uses the theory of transformation models to explain core concepts and more specialised issues in Generalised Regression and Survival Analysis. My main motivation to do so is the simplicity of transformation models on the one hand, and their broadness on the other hand. Computer scientists have the concept of “write once, run everywhere” when implementing software in a portable way. Transformation models enable students to follow the “think once, apply everywhere” concept.

Chapter 1

Introduction

Extreme postpartum blood loss, or postpartum haemorrhage, is a major cause of maternal mortality and, although several risk factors are known, often occurs unexpectedly. The incidence of postpartum haemorrhaging and related maternal morbidity and mortality are a serious health issue, literally a question of life and death. Before we dive into the theory of transformation models in later chapters of this book, let us motivate *why* it is worth spending time and grey matter on these models for regression analysis by looking at the distribution of postpartum blood loss recorded prospectively in the delivery ward of the University Hospital Zurich, Switzerland. I got to know these premises from first-hand experience twice, so I can assure readers that statistical analyses of this data are everything but bloodless.

Haslinger et al. (2020) report data on $N = 1309$ deliveries, and we shall concentrate on postpartum blood loss measured from these mothers. The first natural question we want to ask is "What is the distribution of postpartum blood loss?", at least in the population of mothers-to-be considering giving birth in this hospital. Postpartum blood loss, the amount of which was measured in milliliter (ml), is a positive variable. It is probably also bounded from above but, as we will see later, it is hard to put a concrete number to the maximal blood loss physiologically possible.

How was postpartum blood loss measured? Kahr et al. (2018) explain the method applied (you should not read this publication if you have a weak stomach or plan to start a family soon). Basically, a plastic bag is put under the pelvis of the parturient, and the volume of collected blood is measured. As anybody having suffered, witnessed, or just heard of the circumstances of child birth can imagine, the reported numbers are far from being exact. Thus, typical values recorded in the data base are 100 ml, 450 ml, and only rarely values like 220 ml or 1175 ml were observed. Clearly, the persons responsible for measuring the blood volume recorded rounded values. With a range from 100 ml to 5700 ml (yes, this number is right!), we only observe 39 unique values. Clearly, this measurement is not the true postpartum blood loss, but a relatively rough guess. Therefore, we shall denote this variable as “measured blood loss”.

Let us ignore these obstacles for the time being and have a look at the empirical cumulative distribution function of measured blood loss in Figure 1.1. The good news is that the median measured blood loss is 400 ml and 80% of women in labour loose less than 700 ml. However, the conceptually continuous distribution function of postpartum blood loss looks very much discrete, due to the discreteness of the measured blood loss. Switching to the quantile function also doesn’t help: The boxplot on top of the empirical cumulative distribution function in Figure 1.1 is essentially a poor-man’s approach to visualise the quantile function. The estimated quantile function itself is also discrete. The derivation of a continuous density function from this discrete estimate of the cumulative distribution function seems hopeless. More importantly, it is hard to see much in Figure 1.1 because the largest observation of 5700 ml stretches the abscissa and the bulk of the data is concentrated in the left part.

Removing the observation with maximum measured blood loss is, of course, not an option. But we could apply a standard trick from the statistician’s toolbox: We simply transform the observations in a way that differences between larger values become relatively small, for example by looking at the logarithm or square root of measured blood loss. The resulting Figure 1.2 depicts the empirical cumulative distribution function for log-measured blood loss. The result is much nicer: The boxplot looks less skewed (but still not really symmetric) and the estimated distribution func-

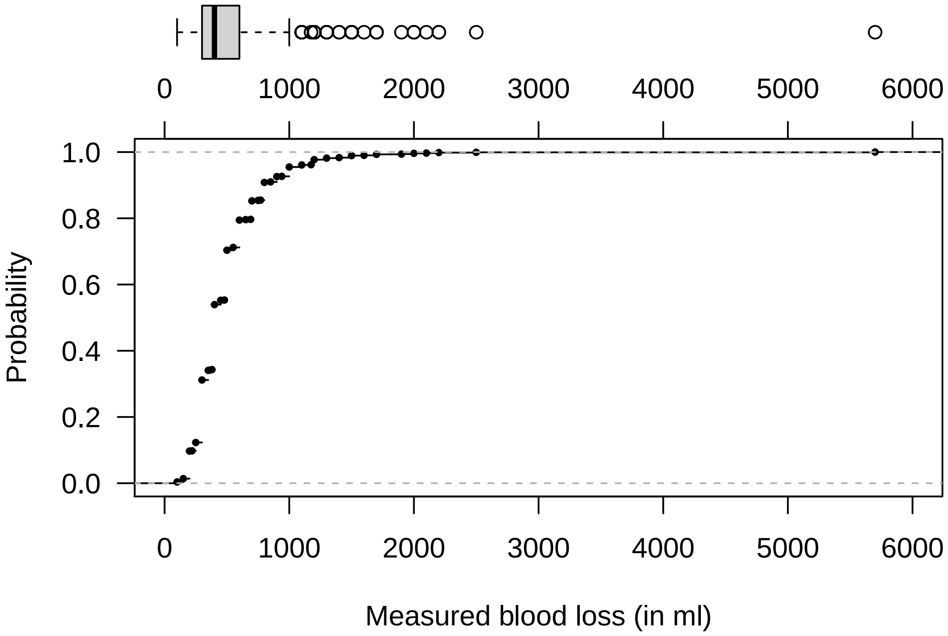


Figure 1.1: Measured blood loss. Empirical cumulative distribution function and boxplot of measured blood loss for $N = 1309$ deliveries at the University Hospital Zurich, Switzerland.

tion seems a little smoother. This is, however, an artefact because the size of the jumps are still the same as in Figure 1.1. In a nutshell, transformation models apply the same principle of transformating the observations. Instead of just applying the logarithm or square-root transformation, transformation models determine the most suitable transformation in a data-driven way.

Before we begin to understand how this might work, we need to study the details of the empirical cumulative distribution function a little closer. Unfortunately, we need some symbols to describe what is going on: $Y > 0$ is the real continuous variable representing postpartum blood loss, the vari-

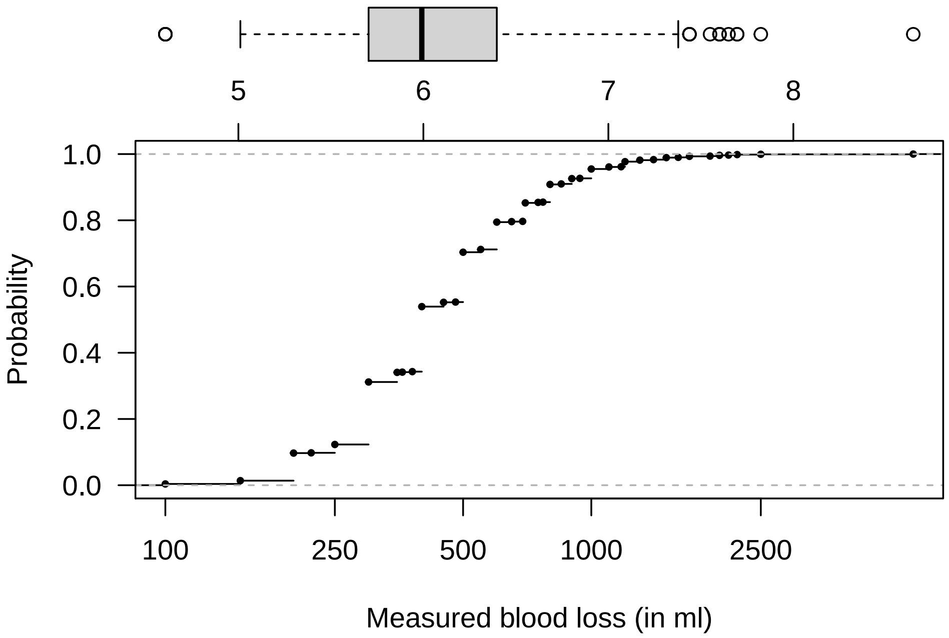


Figure 1.2: Measured blood loss. Empirical cumulative distribution function of measured blood loss on the log-scale for $N = 1309$ deliveries at the University Hospital Zurich, Switzerland. The boxplot on top presents log-quantiles.

able those cumulative distribution function $F_Y(y) = \mathbb{P}(Y \leq y)$ we want to estimate. The cumulative distribution function $F_Y(y)$ is a continuous monotonically increasing function of it's argument y with $F_Y(0) = 0$ (no way giving birth without spilling some drops) and $F_Y(\infty) = 1$ (humans have between 4000 and 6000 ml blood, even with blood infusion the quantile y for which $F_Y(y) = 1$ is much less than ∞). The empirical cumulative distribution function $\hat{F}_Y$ depicted for the untransformed measured blood loss in Figure 1.1 is computed from the measured blood loss readings

$y_1, \dots, y_{N=1309}$ as

$$\hat{F}_Y(y) = \sum_{i=1}^{N} \mathbb{1}(y_i \leq y).$$

As we will see later, $\hat{F}_Y(y)$ is the nonparametric maximum likelihood estimator of F_Y, thus optimising the nonparametric maximum likelihood. For the moment, we just note that such a thing exists. The estimated distribution functions for transformed measured blood loss presented in this chapter all maximise the same nonparametric likelihood, and thus can be compared directly.

Transforming the data, using the logarithm or any other monotonically increasing function h, does not change the probabilities derived from the empirical cumulative distribution function because

$$\hat{F}_Y(y) = \sum_{i=1}^{N} \mathbb{1}(\log(y_i) \leq \log(y)) = \sum_{i=1}^{N} \mathbb{1}(h(y_i) \leq h(y)).$$

This argument also holds for the unknown theoretical distribution function of postpartum blood loss

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(\log(Y) = \log(y)) = \mathbb{P}(h(Y) = h(y))$$

and for any monotonically increasing function h. We don't know the distribution of Y, because this is the distribution of postpartum blood loss we want to estimate. We also don't know the distribution of $\log(Y)$. Instead of asking the question "What is the distribution of $h(Y)$?" for some given transformation h, transformation models ask the question "What is the transformation h such that $h(Y)$ follows an apriori defined distribution?". For example, we could require that the transformed postpartum blood loss $h(Y)$ is normally distributed with mean zero and variance one: $h(Y) \sim \mathrm{N}(0, 1)$. Using the symbol $Z := h(Y) \sim \mathrm{N}(0, 1)$ for the transformed postpartum blood loss, we *define* that $Z \sim \mathrm{N}(0, 1)$. The transformation h can be inverted and thus we can write postpartum blood loss $Y = h^{-1}(Z)$ as a transformation of a standard normal Z. Such a transformation h does exist (arguments are layed out in later chapters) and thus, instead of estimating the distribution function F_Y, transformation models estimate the unknown transformation h. Once we have a reasonable estimate $\hat{h}$ of h, we can evaluate the corresponding distribution function by the plug-in $\Phi(\hat{h}(y))$ of $F_Y(y) = \Phi(h(y))$. Of course, in the simple situation we study, we can compute $\hat{h}$ directly from the empirical cumulative distribution function $\hat{h}(y) = \Phi^{-1}(\hat{F}_Y(y))$. Figure 1.3 depicts this model: The boxplot of $\hat{h}(y_i)$ is know almost symmetric, and the empirical cumulative distribution function of $\hat{h}(y_i)$ can be *exactly* interpolated by Φ, the cumulative distribution function of the standard normal.

What is the relevance of this academic exercise? The estimated probabilities did not change and are still discrete, the quantile function is still discrete and hard to compute, and the same applies to densities, for example. The most important point to consider is that the transformation model $F_Y(y) = \Phi(h(y))$ is much easier to extend than the unstructured distribution function F_Y, especially when studying conditional distribution functions.

We use these preliminaries to answer the question “Do women undergoing Cesarean section loose more blood?” and, if the answer is “yes”, how much more? We translate this question in the language of transformation models by proposing *two* distribution functions, one for postpartum blood loss by vaginal deliveries and one by Cesarean sections:

$$\begin{aligned}
\mathbb{P}(Y \le y \mid \text{Vaginal delivery}) &= \Phi(h(y)) \\
\mathbb{P}(Y \le y \mid \text{Cesarean section}) &= \Phi(h(y) - \beta).
\end{aligned}$$

The two *conditional* (on mode of delivery) distribution functions share the same transformation h. The shift parameter β, however, only applies to the distribution of postpartum blood loss under Cesarean sections. Clearly, if $\beta = 0$, the two distributions are identical. An alternative way of formulating this model is $h(Y) \sim \text{N}(0, 1)$ for vaginal deliveries and $h(Y) \sim \text{N}(\beta, 1)$ for Cesarean sections, or, in the language of a normal linear regression model

$$h(Y) = \beta x + Z, \quad Z \sim \text{N}(0, 1) \tag{1.1}$$

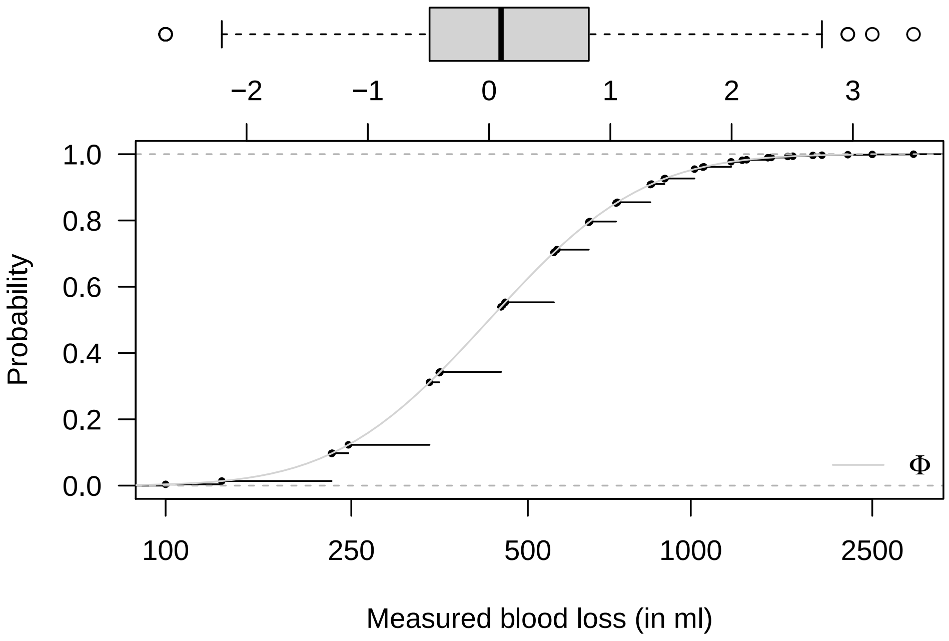


Figure 1.3: Measured blood loss. Empirical cumulative distribution function of measured blood loss on a transformed scale for $N = 1309$ deliveries at the University Hospital Zurich, Switzerland. The transformation are obtained such that the transformed values are as close as possible to a standard normal distribution. The boxplot on top presents quantiles on the transformed scale. The grey solid line depicts the cumulative distribution function Φ of the standard normal distribution.

with $x = 0$ for vaginal deliveries and $x = 1$ for Cesarean sections being a treatment contrast for mode of delivery. The unknowns in this model are the transformation h and the shift term β. It should be noted that the standard deviation of the error term Z is one by definition and that the intercept term is included in the transformation function h.

Figure 1.5 presents the two empirical cumulative distribution functions of measured blood loss – one fitted to each group according to mode of

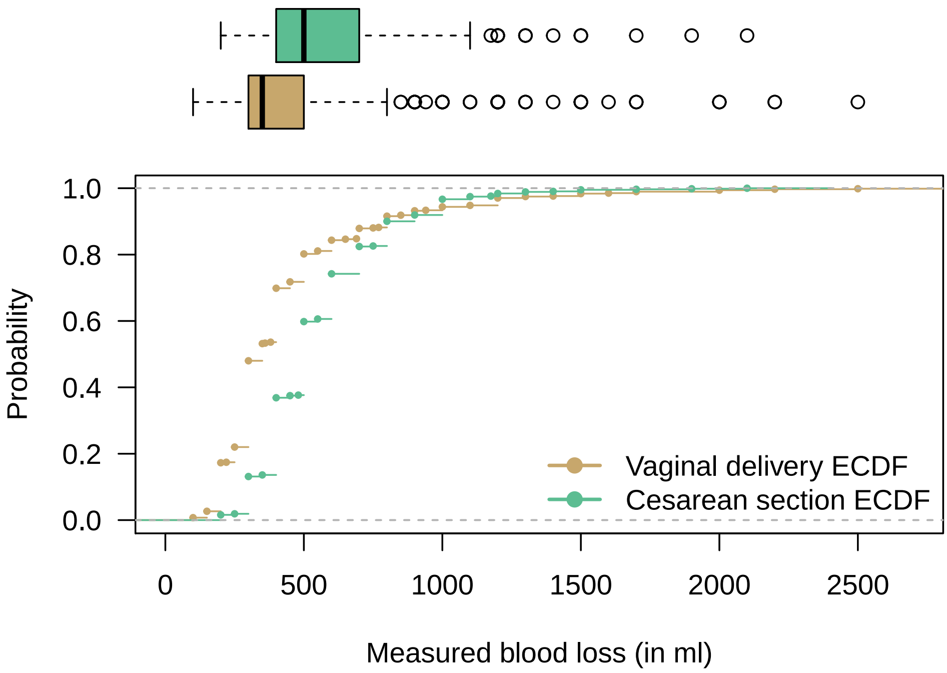


Figure 1.4: Measured blood loss. Empirical cumulative distribution functions (ECDFs) and boxplots of untransformed measured blood loss for $N_{\text{VD}} = 677$ vaginal deliveries and $N_{\text{CS}} = 632$ Cesarean section at the University Hospital Zurich, Switzerland. The two curves and boxplots were estimated separately for the two groups. One extreme observation of 5700 ml is not shown.

delivery. In Figure 1.5, these empirical cumulative distribution functions are overlayed with the two conditional distribution functions obtained from model (1.1). The estimates $\hat{h}$ and $\hat{\beta}$ were computed by nonparametric maximum likelihood estimation, *i.e.* by maximising the same log-likelihood maximsed by the empirical cumulative distribution function. We obtain $\hat{\beta} = 0.754$ with 95% confidence interval $0.639, 0.869$ for the shift term β.

There is a significant difference between the two approaches presented jointly in Figure 1.5: The two empirical cumulative distribution functions

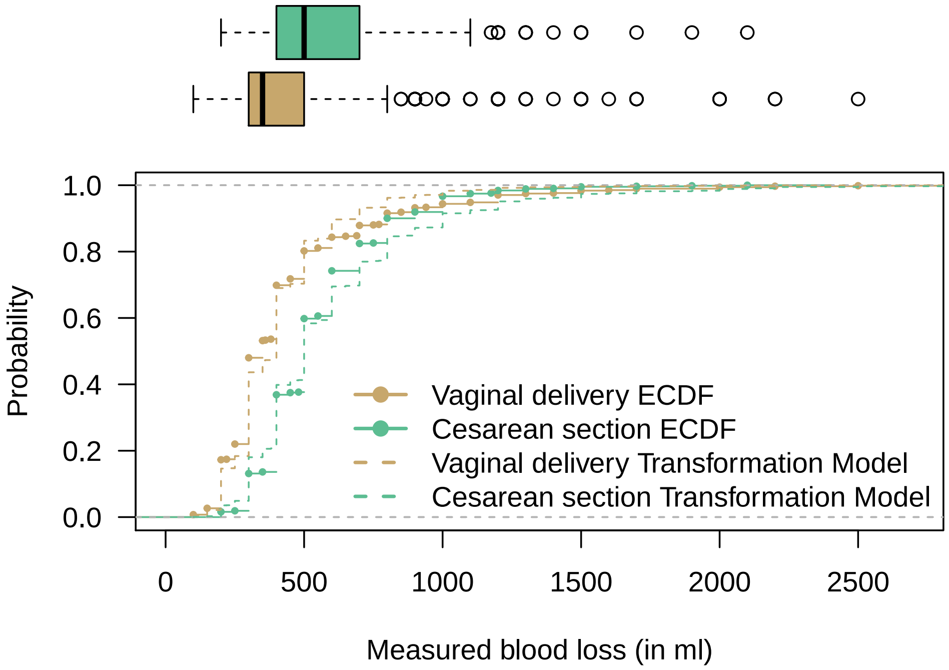


Figure 1.5: Measured blood loss. Empirical cumulative distribution functions (ECDFs) and boxplots of untransformed measured blood loss for $N_{\text{VD}} = 677$ vaginal deliveries and $N_{\text{CS}} = 632$ Cesarean section at the University Hospital Zurich, Switzerland. The two curves and boxplots were estimated separately for the two groups. The dashed curves depict the estimated distribution functions from transformation model (1.1), where the discrepancy between the two distribution functions is modelled by a shift parameter β. One extreme observation of 5700 ml is not shown.

are distribution- and model-free, whereas the two conditional distribution functions in the transformation model only differ by a shift term β. They are distribution-free, because we do not impose any assumptions on the transformation h, and thus to the distribution under vaginal deliveries, but they are not model-free because of the shift parameter β governs how the two groups differ.

Naturally, the question how to interpret the result $\hat{\beta} = 0.754$ arises. From model (1.1) we know that $\mathbb{E}(h(Y) \mid x = 0) = 0$ for vaginal deliveries and $\mathbb{E}(h(Y) \mid x = 1) = \beta$ for Cesarean sections. This allows to interpret β as the mean difference $\mathbb{E}(h(Y) \mid x = 1) - \mathbb{E}(h(Y) \mid x = 0)$ of measured blood loss between Cesarean sections and vaginal deliveries *after* the data were transformed by h. Unfortunately, h is not linear and thus it is hard to say anything about the mean difference $\mathbb{E}(Y \mid x = 1) - \mathbb{E}(Y \mid x = 0)$ for the original, untransformed measured blood loss. However, we can say at least something about the magnitude of the difference. The effect is 0.754 times the standard deviation (which is one). To get an idea about the relevance of such an effect we can investigate the sample size necessary to demonstrate an effect of 0.754 in two independent normally distributed samples with standard deviation one. For a nominal level of 5% and a power of 80%, we would need two samples of size 29 each (when applying a two-sample t-test). This would be a very small experiment, so the two distributions describe a relevant discrepancy (not “difference”!) between Cesarean sections and vaginal deliveries with respect to measured blood loss. Even when the effect was only as large as the lower bound of the confidence interval (0.639), we would only need 40 observations in each of two groups.

This way of interpreting the shift effect is perfect for causing confusion. The difficulties arise from the choice $Z \sim \mathrm{N}(0, 1)$. Later chapters will switch to alternative distributions for F_Z allowing simpler ways of interpreting parameters in transformation models.

Finally, we switch attention to the other technical problems mentioned earlier. The discreteness of the empirical cumulative distribution function, or the conditional distribution functions obtained from the transformation model (1.1) presented in Figure 1.5, makes the derivation of densities or inversion for computing quantiles difficult. Conceptually, the transformation function h for postpartum blood loss is continuous and monotonically increasing but we obtained a discrete estimate $\hat{h}$. If we chose a continuous parameterisation for the transformation h instead, we also obtain a smooth estimate $\hat{h}$.

The top panel in Figure 1.6 shows the two empirical cumulative distribution functions for measured blood loss, estimated separately for vaginal deliveries and Cesarean sections (this is identical to Figure 1.4). The additional smooth curves are the estimated distribution functions for these two groups of women:

$$\begin{aligned}\mathbb{P}(Y \leq y \mid \text{Vaginal delivery}) &= \Phi(h_{\text{VD}}(y))\\ \mathbb{P}(Y \leq y \mid \text{Cesarean section}) &= \Phi(h_{\text{CS}}(y)).\end{aligned}$$

The two transformation functions $h_{\text{VD}}(y)$ for vaginal deliveries and $h_{\text{CS}}(y)$ Cesarean sections are smooth and monotonically increasing in their argument y, and so are the corresponding estimates $\hat{h}_{\text{VD}}(y)$ and $\hat{h}_{\text{CS}}(y)$. The functions $\Phi(\hat{h}_{\text{VD}}(y))$ and $\Phi(\hat{h}_{\text{CS}}(y))$ smoothly approximate the discrete empirical cumulative distribution functions. One big advantage of this approach is that the density for measured blood loss from vaginal deliveries can be obtained simply from the derivative $\phi(\hat{h}_{\text{VD}}(y))\hat{h}'(y)$ of the estimated distribution function $\Phi(\hat{h}_{\text{VD}}(y))$ with respect to y, with $\phi(z) = \Phi'(z)$ denoting the standard normal density. The same applies to measured blood loss from Cesarean sections. The lower panel in Figure 1.6 depicts the two densities: Of course, they are far from being normal. We will later discuss the question if the smoothly estimated distribution functions and densities are distribution-free in a similar way as the empirical cumulative distribution functions.

Model (1.1), which describes the discrepancy between the two smooth distribution functions by a shift term β, can also be estimated such that we obtain a smooth estimate $\hat{h}$ of the transformation function h. Figure 1.7 presents the results. We obtain $\hat{\beta} = 0.754$ with 95% confidence interval $0.640, 0.868$ for the shift term β, these values are very close to the results obtained from a discretely parameterised transformation function (Figure 1.4). However, the model is still unrealistic, because the two distribution functions cross at about 850 ml, and this fact also explains that the simpler model fits the data less well. Figure 1.6 highlights a relatively good fit of the two smooth model-based curves for measured blood loss up to about 500 ml, but the smooth probabilities for vaginal deliveries are much higher than the corresponding probabilities obtained from the empirical cu-

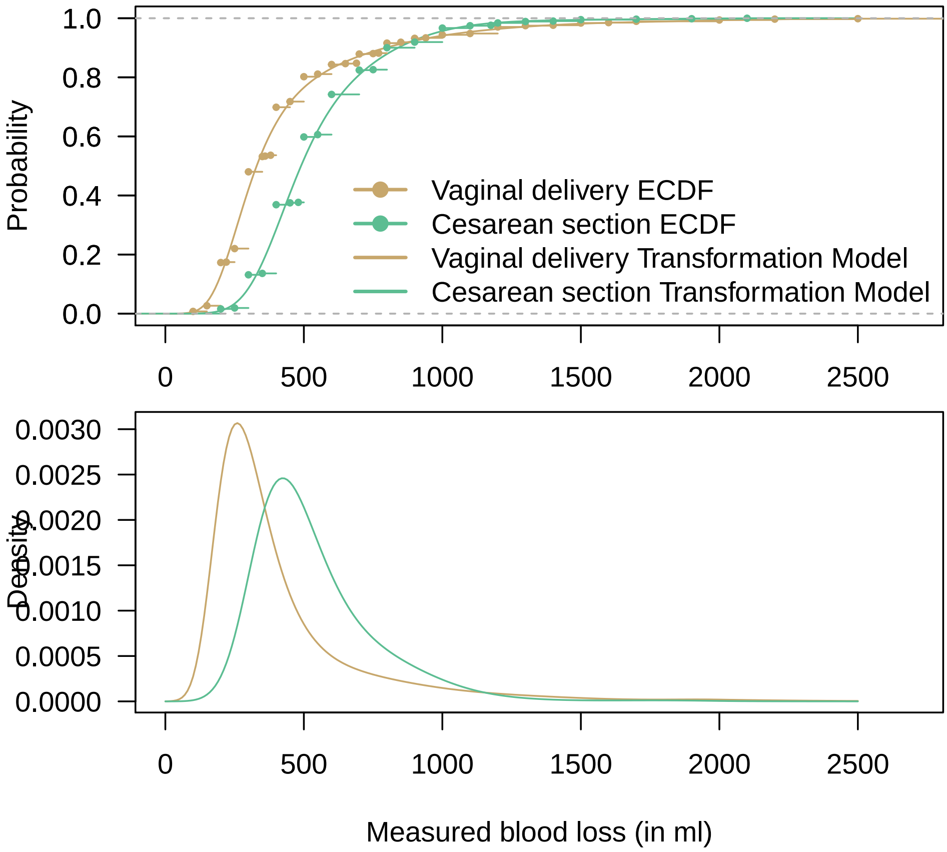


Figure 1.6: Measured blood loss. Top: Empirical cumulative distribution functions (ECDFs) of untransformed measured blood loss for $N_{\text{VD}} = 677$ vaginal deliveries and $N_{\text{CS}} = 632$ Cesarean section at the University Hospital Zurich, Switzerland. The two curves were estimated separately for the two groups. The solid lines depict the estimated distribution functions from two smoothly parameterised transformation models. Bottom: Estimated densities derived from two smoothly parameterised transformation models. One extreme observation of 5700 ml is not shown.

mulative distribution function. For Cesarean sections, smooth probabilities are smaller than their discrete counterpart. This effect is also visible from a quantitative comparison of the transformation models fitted separately

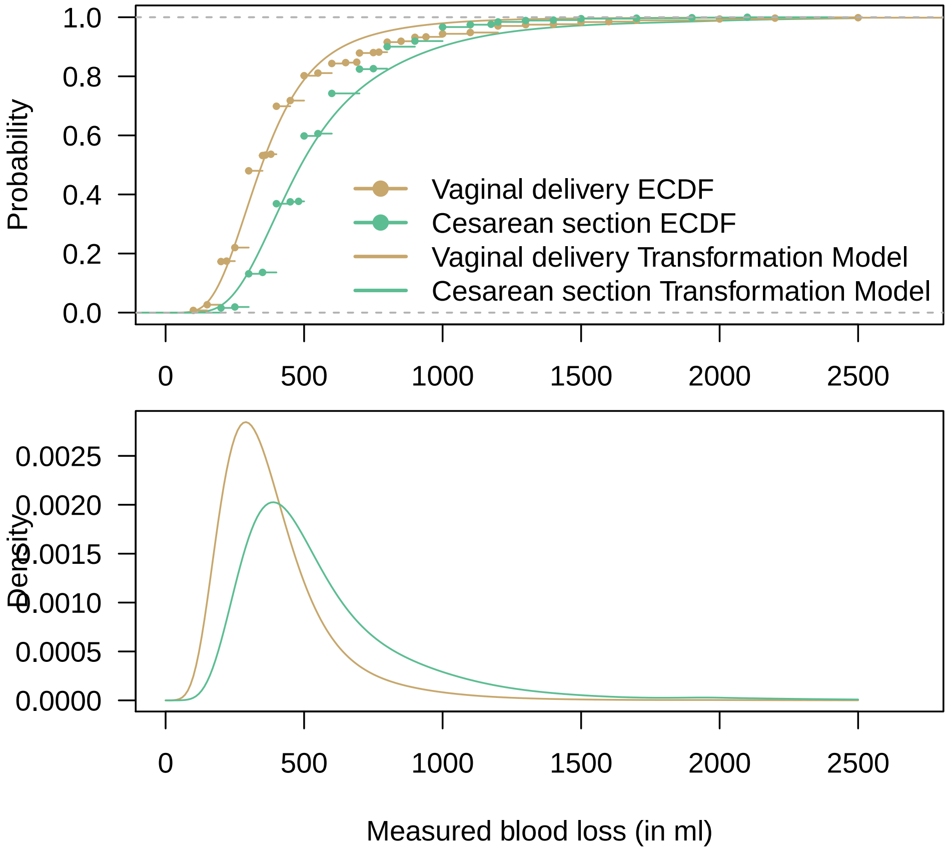


Figure 1.7: Measured blood loss. Top: Empirical cumulative distribution functions (ECDFs) of untransformed measured blood loss for $N_{\text{VD}} = 677$ vaginal deliveries and $N_{\text{CS}} = 632$ Cesarean section at the University Hospital Zurich, Switzerland. The two curves were estimated separately for the two groups. The solid depict the estimated distribution functions from a smoothly parameterised transformation model (1.1), where the discrepancy between the two distribution functions is modelled by a shift parameter β. Bottom: Estimated densities derived from the smoothly parameterised transformation model featuring a shift parameter. One extreme observation of 5700 ml is not shown.

(Figure 1.6) and jointly assuming a shift effect β on the probit scale (Figure 1.7). The parameters of the former transformation model correspond to a nonparametric log-likelihood of -4214.705 whereas the latter model gives a nonparametric log-likelihood of -4278.176. The term "nonparametric log-likelihood" for model (1.1) involving a smooth transformation function h and a shift parameter β is puzzling, but we will justify it's use later.

There are many technical issues of model estimation, model interpretation, model inference, model diagnostics or comparison, and model improvement to be discussed in this respect, and we will look at them in later chapters. For the moment, this Introduction only aims at outlining the most important aspects and potential practical benefits of transformation models. Of course, the unconditional case or the two-sample situation illustrated here are the simplest setups. In fact, many more explanatory variables are available, such as age of the mother, gestational age, if mothers expect one or more babies, various laboratory blood parameters known to be prognostic for bleeding in surgical contexts, and other variables have been recorded in the prospective study by Haslinger et al. (2020). How the distribution of postpartum blood loss depends on these explanatory variables will be inferred from more complex transformation models. We will also discuss better fitting alternatives to the normal assumption in model (1.1) in later Chapters.

With $\boldsymbol{x}$ denoting the possibly many or even unstructured explanatory variables and $\tilde{\boldsymbol{x}}$ a corresponding real-valued representation (contrasts etc.), we will later study details of transformation models for conditional distribution functions of the following forms. The simplest special case is the linear transformation model

$$\mathbb{P}(Y \le y \mid \boldsymbol{x}) \quad = \quad F_Z(h(y) - \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}).$$

In addition to the transformation function h, the model features a linear predictor $\tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}$ with unknown regression coefficients $\boldsymbol{\beta}$, acting as a shift term. Large values of the linear predictor $\tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}$ correspond to large values of Y in a sense we will make precise later. The key aspect of this model is a clear separation of terms depending on y and $\boldsymbol{x}$, there is no interaction between the two. As a consequence, the impact of changes in the explanatory variables of the distribution of Y will be relatively easy to be understood and communicated.

This above assumption is relaxed in models where the regression coefficients may depend on y. These response-varying effect models, also known as distribution regression in econometrics or time-varying effects in survival analysis,

$$\mathbb{P}(Y \le y \mid \boldsymbol{x}) \quad = \quad F_Z(h(y) - \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}(y))$$

are useful for assessing potential violations of the assumption of constant regression coefficients β in linear transformation models.

Another extension of the linear transformation model is the shift transformation model, where the linear predictor is replaced by a more complex function $\beta(\boldsymbol{x})$ of the explanatory variables

$$\mathbb{P}(Y \le y \mid \boldsymbol{x}) \quad = \quad F_Z(h(y) - \beta(\boldsymbol{x})).$$

Such models are useful for unstructured explanatory variables $\boldsymbol{x}$ or for situations where higher-order interactions among the components of $\boldsymbol{x}$ are important. The model complexity might be high in the shift term $\beta(\boldsymbol{x})$, however, the model still clearly separates between terms involving y and $\boldsymbol{x}$ and thus at least some of the interpretability of linear transformation models is preserved

Finally, one may want to drop almost all assumptions and allow for conditional transformation functions in conditional transformation models:

$$\mathbb{P}(Y \le y \mid \boldsymbol{x}) \quad = \quad F_Z(h(y \mid \boldsymbol{x})).$$

Here, we basically fall-back to the unconditional case and allow $\boldsymbol{x}$-specific transformation functions for each configuration of the explanatory variables. Of course, some constraints need to be implemented, such as a certain similarity of the transformation functions h in a neighbourhood of $\boldsymbol{x}$. These vague statements will be defined more precisely later on.

The maybe most important aspect of transformation models is the clear separation of model structure and parameter estimation (at least to a very large extend, because there are exceptions). Evaluation of the likelihood only requires evaluation of a distribution function $\mathbb{P}(Y \le y \mid \boldsymbol{x})$. Transformation models are defined on the scale of a distribution function and thus readily provide us with everything we need to set-up and optimise the likelihood. The big advantage is that we can discuss the likelihood of transformation models without the need to worry about the concrete form of the model or specific aspects of this parameterisation.

This aspect of transformation models, and the fact that many (but not all) classical models in regression or survival analysis can be understood as a transformation model, explains my fondness of this model class for teaching. Being able to discuss technical issues in a general way allows me to devote more time to model interpretation and model criticism, aspects of the professional life of statisticians I consider more important and interesting than technical aspects of model estimation. The remainder of this book was written having this goal in mind, but of course I failed to achieve it in many places.

Chapter 2

The Likelihood Function

Most books start with models and discuss ways of fitting these models to data in a second step. The first part is more fun and the second part often quite technical. At the risk of scaring readers off, I chose to reverse the order of presentation: In this chapter, the likelihood function as the central device of model fitting is introduced in a generic way.

The reason for this decision is that we need to get some common misconceptions out of our way: The likelihood is usually defined in terms of a probability density (for very good reasons!) but it is more convenient for us interested in transformation models to define the likelihood in terms of a cumulative distribution function. This must not come as a big surprise to readers of Chapter 1: Transformation models are models for distribution functions and thus we will benefit from likelihood functions defined by distribution functions in later chapters. The concept will be illustrated for all practically relevant measurement scales of a univariate response variable: binary, ordered and unordered categorical, count, and several forms of continuous response variables will be discussed.

To keep things simple and the notation as light as possible, we only consider unconditional distributions of univariate responses from independent observations in this chapter. Extensions to regression models via conditional distributions, to multivariate responses, and to correlated observations shall be discussed later on.

2.1 Probabilities and Distribution Functions

The response Y we are interested in follows a distribution $\mathbb{P}_Y$ and we write $Y \sim \mathbb{P}_Y$ understanding that a probability space $(\Xi, \mathfrak{C}, \mathbb{P}_Y)$, consisting of a sample space Ξ, a corresponding σ-algebra $\mathfrak{C}$, and our probability measure $\mathbb{P}_Y$, exists. The sample space consists of all possible *outcomes* $\upsilon \in \Xi$ of the corresponding experiment. We will discuss plenty of examples of sample spaces appropriate for modelling many different measurement scales of our response variable Y. The σ-algebra is a set of appropriate subsets $\boldsymbol{C} \subseteq \Xi$ of the sample space Ξ. The elements $\boldsymbol{C} \in \mathfrak{C}$ of the σ-algebra are called *events*, and describe all possible results of our experiment. It is important to note that an observation is, conceptually, always an event, *i.e.* a set $\boldsymbol{C} \in \mathfrak{C}$. We never directly observe outcomes, *i.e.* elements υ of the sample space Ξ, directly. For discrete sample spaces with finite cardinality $K = \mid \Xi \mid < \infty$, however, events $\boldsymbol{C} = \{\upsilon\}$ consisting of a single outcome υ might very well be observed. Of course, we also have $\mathbb{P}_Y(Y \in \Xi) = 1$ for $\boldsymbol{C} = \Xi$.

The distinction between outcomes υ and events $\boldsymbol{C}$ is important, because we can only compute probabilities for the latter. The probability measure is defined for events $\boldsymbol{C}$, more formally we write $\mathbb{P}_Y : \mathfrak{C} \to [0, 1]$. Thus, $\mathbb{P}_Y(\boldsymbol{C}) = \mathbb{P}_Y(Y \in \boldsymbol{C}) \in [0, 1]$ is the probability of observing $\boldsymbol{C}$. The distribution function $F_Y(y) = \mathbb{P}_Y(Y \leq y) = \mathbb{P}_Y(Y \in \{\upsilon \in \Xi : \upsilon \leq y\})$ is defined for outcomes $\upsilon \in \Xi$ under a specific ordering, and thus we define $F_Y : \Xi \to [0, 1]$. Here, $F_Y(y)$ is the probability of observing an event consisting of all outcomes less or equal y. When we increase y, the corresponding event (a set of outcomes) contains more elements and thus the corresponding probability increases (possibly not strictly). It follows that the distribution function is a monotonically, but not necessarily strictly, increasing function with $F_Y(y_1) \leq F_Y(y_2)$ for all $y_1 < y_2 \in \Xi$.

This theoretical setup is generic and applies to all experiments. In contrast, concrete choices of samples spaces Ξ and probability measures $\mathbb{P}_Y$ very much depend on the circumstances of the experiment we are interested in. It is very important for the data analyst to carefully consider appropriate choices and to understand how these choices impact the result of an analysis. The σ-algebra is of less interest to the practitioner.

2.2 Sample Spaces and Measurement Scales

The choice of an appropriate sample space Ξ very much depends on the measurement scale of the response Y. Most situations can be classified into binary, ordered or unordered categorical, count, and bounded or unbounded absolutely continuous responses Y.

Binary Response

The response can only take two outcomes, υ_1 or υ_2 and the sample space simply consists of these two elements: $\Xi = \{\upsilon_1, \upsilon_2\}$. We understand these two outcomes as being truely categorical and explicitly exclude dichotomisation: Binary variables like “Younger than 65 years” vs. “Older than 65 years” are modelled by a sample space appropriate for age as a numeric variable. The corresponding distribution function can be used to compute probabilities for the event “Younger than 65 years”. This seems overly complex, but when your boss eventually decides that an age-cutoff at 70 years is more appropriate than the initially chosen cut-off at 65 years, you only need to evaluate the distribution function at 70 instead of 65 years. The numerical sample space for age, and thus your theoretical foundation, doesn’t requires any change. I admit that this example is a bit academic, but we will see later that this issue has had dramatic consequences in some research areas.

Example 2.1. The two outcomes $\upsilon_1 =$ failure and $\upsilon_2 =$ success are typical. The machinery introduced above can be used to express the probability of failing as $F_Y(\upsilon_1) = \mathbb{P}_Y(Y \in \{\upsilon_1\}) = \mathbb{P}_Y(\{\text{failure}\})$. The probability of either failing or succeeding is $F_Y(\upsilon_2) = 1$ and the probability of success can be computed by $\mathbb{P}_Y(\{\text{success}\}) = F_Y(\upsilon_2) - F_Y(\upsilon_1)$. It is of course allowed to switch to the simpler and more common notation $\mathbb{P}_Y(Y = \text{success})$, however, we need to bear in mind that probabilities can only be computed for events, not outcomes. □

Ordered Categorical Response

Instead of just two categorical outcomes, the sample space may consist of $K < \infty$ of such categories: $\Xi = \{\upsilon_1, \upsilon_2, \dots, \upsilon_K\}$. These categories are a priori known and must not be changed during the course of the experiment. Again, the outcomes are truely categorical and must not have been obtained by some form of categorisation.

Example 2.2. Happiness of people is impossible to measure exactly in an objective way. Thus, a subjective self-reported response Y is used, for example with a sample space Ξ consisting of the outcomes $\upsilon_1 =$ very unhappy, $\upsilon_2 =$ not to happy, $\upsilon_3 =$ somewhat happy, and $\upsilon_4 =$ very happy. There is a natural ordering $\upsilon_1 < \upsilon_2 < \upsilon_3 < \upsilon_4$.

Evaluating the distribution function at υ_1 gives $F_Y(\upsilon_1) = \mathbb{P}_Y(Y = \upsilon_1) = \mathbb{P}_Y(\text{very unhappy})$, the probability of being very unhappy. Of course, the probability of observing any of the outcomes $F_Y(\upsilon_4) = 1$ is one. The probabilities of υ_2, υ_3, and υ_4 can be computed from the differences $F_Y(\upsilon_2) - F_Y(\upsilon_1) = \mathbb{P}_Y(\text{not to happy})$, $F_Y(\upsilon_3) - F_Y(\upsilon_2) = \mathbb{P}_Y(\text{somewhat happy})$, and $F_Y(\upsilon_4) - F_Y(\upsilon_3) = \mathbb{P}_Y(\text{very happy})$. □

Example 2.3. Subjective measurements are often coded numerically but are, in fact, ordinal categorical readings. One example is the ALS functional rating scale for assessing disease severity in patients suffering Amyotrophic Lateral Sclerosis (ALS). The scale consists of twelve items rating a patient's ability to perform common daily tasks, *i.e.* climbing stairs, writing, dressing, cutting food or handling cutlery. Each of the items is scored with a five-point scale ranging from 0 to 4, where 0 is “can't do” and 4 refers to “normal ability”. Clearly, each of these subscores are ordered categorical. The final sum score Y ranges between 0 and 48 but we must *not* interpret differences on this scale. Thus, treating the labels $0, \dots, 48$ as ordinal elements of a sample space Ξ is appropriate. □

Unordered Categorical Response

The sample space $\Xi = \{\upsilon_1, \upsilon_2, \dots, \upsilon_K\}$ consists of K categorical elements, but in contrast to the ordinal case, there is no a priori order of these elements given. We can chose any ordering, and each possible ordering of the K elements defines a corresponding cumulative distribution function. The distribution function F_Y is therefore not unique. However, the probabilities $\mathbb{P}_Y(Y = \upsilon_k) = F_Y(\upsilon_k) - F_Y(\upsilon_{k-1})$ are invariant with respect to the chosen ordering.

Example 2.4. Faculties at a university, maybe υ_1 = Medicine, υ_2 = Natural Sciences, υ_3 = Philosophy and so on, can be ordered with respect to many different aspects. Number of students, number of faculty, number of journal publications, and other “performance indicators” are often used to justify increasing or decreasing budgets. There is, however, no a priori ordering as the statement $\upsilon_1 < \upsilon_2$ “Medicine is less than the Natural Sciences” does not make sense. □

Count Response

Sample spaces for categorical data consist of a finite number of outcomes. When the response Y represents counts of something, the sample space $\Xi = \mathbb{N}$ with outcomes $\upsilon_k = k$ for $k = 0, 1, \dots, \infty$ is still discrete but contains an infinite number of elements. It is, however, still a countable set. In some situations restrictions might apply such that the maximal number of counts which can possibly observed is K. One can either chose the sample space $\Xi = \{1, \dots, K\}$ or require $F_Y(K) = 1$.

Example 2.5. Consider Y being number of wildlife-vehicle collisions counted per year on a specific road segment. The distribution function $F_Y(k)$ denotes the probability of observing at most k such collisions per year on this road segment and $F_Y(k) - F_Y(k-1)$ is the probability $\mathbb{P}_Y(Y \in \{k\}) = \mathbb{P}_Y(Y = k)$ of observing exactly k collisions. The distribution function jumps at integers k only (you can't run over half a hare), so $F_Y(y) = F_Y(\lfloor y \rfloor)$, where $\lfloor y \rfloor$ is the largest integer smaller or equal to $y \in \mathbb{R}$. □

Absolutely Continuous Response

Models for objectively measureable quantities, or quantities which seem to be measureable in an objective way because we know that there is always an observer effect, rely on response variables $Y \in \mathbb{R}$. The sample space is a subset of the real numbers, and can be unbounded ($\Xi \subseteq \mathbb{R}$), bounded ($\Xi = [\underline{\upsilon}, \bar{\upsilon}]$), or positive ($\Xi = [0, \infty]$). Of course, other intervals are also suitable choices of the sample space. In most cases, it is appropriate to assume that the distribution function $F_Y(y)$ is absolutely continuous, but sometimes exceptions apply.

Example 2.6. Pain scales are sometimes measured on a visual analog scale with sample space $\Xi = [0, 1]$, instead on an ordered categorical scale with $\upsilon_1 =$ no pain, $\dots, \upsilon_{10} =$ extreme pain. The latter scale forces study participants to pick one out of ten possible pain categories. The visual analog scale is a ten centimeter line on a sheet of paper, and participants are asked to mark the pain they are suffering, with little pain to the left and extreme pain to the right. The reader simply measures the distance between the left end of the line and the mark in centimeter and this number, divided by ten, is recorded. It makes sense to compare differences on the visual analog scale, whereas differences for ordinal outcomes must not be interpreted. □

Example 2.7. Survival analysis is a sub-field of statistics interested in the analysis of time-to-event variables. The most common example is maybe time to death Y of a patient after having been diagnosed with a severe disease. One cannot go back in time, so the sample space is $\Xi = (0, \infty)$, although the probability $F_Y(y)$ of surely being dead will be one for y much less than ∞. □

Example 2.8. The age of a person Y is also a positive real number, and thus $\Xi = (0, \infty)$. It is usually measured in years, but the clock ticks every split-second, so the sample space is continuous, as is the corresponding distribution function Y. We can represent the event "person is 44 years old" by the interval $(44, 45]$. □

Example 2.9. We discussed postpartum blood loss in the Introduction. The volume of blood lost after delivery of a baby is positive and conceptually continuous, so $\Xi = (0, \infty)$ is appropriate. The measured blood loss, that is the actual recorded measurements, is highly discrete, and we will come back to this example in later chapters. □

Example 2.10. The sample space $\Xi = [0, \infty)$ is appropriate for the amount of precipitation for a meteorological station at one day. It is important to note that $F_Y(0) > 0$, because the probability of no rain or snow at all is, luckily, larger than zero. This is an example of a distribution with a discrete part (the probability at zero), and a continuous part (the amount of precipitation when it is actually raining). □

Example 2.11. It is very hard to find a convincing example for $\Xi = \mathbb{R}$, although the majority of text books on regression start with this assumption: As soon as we write $Y \sim \mathrm{N}(\mu, \sigma^2)$ we operate under this sample space. □

2.3 The Likelihood Function

With the probability space $(\Omega, \mathfrak{A}, \mathbb{P})$ and distribution function F_Y in place, we are ready to define the likelihood function. Of course, we do not know the distribution function and, in fact, estimating F_Y from data is the primary business of ours. For the moment, we restrict our consideration to the simplest setup, namely the unconditional distribution function of a univariate response Y. We do not, however, treat different measurement scales separately.

We parameterise the unknown distribution function in terms of a vector of unknown parameters $\boldsymbol{\vartheta} \in \Theta \subseteq \mathbb{R}^P$ and write $F_Y(y \mid \boldsymbol{\vartheta}) = \mathbb{P}_Y(Y \le y \mid \boldsymbol{\vartheta})$. This is an important step, because the complex problem of estimating a function F_Y was replaced by the simpler problem of estimating parameters $\boldsymbol{\vartheta}$.

As stressed above, we observe an event $\boldsymbol{C} \in \mathfrak{C}$ as the result of a single random experiment. The likelihood function of parameters $\boldsymbol{\vartheta}$ for such a single observation $l : \Theta \times \mathfrak{C} \to [0, 1]$ is defined as

$$l(\boldsymbol{\vartheta} \mid \boldsymbol{C}) = \mathbb{P}_Y(Y \in \boldsymbol{C} \mid \boldsymbol{\vartheta}).$$

The most important special case are events $\boldsymbol{C} = (\underline{y}, \bar{y}]$ represented by intervals $(\underline{y}, \bar{y}] = \{\upsilon \in \Xi : \underline{y} < \upsilon \le \bar{y}\}$, where we can compute the likelihood by the difference of the distribution function evaluated at the bounds of this interval:

$$l(\boldsymbol{\vartheta} \mid \boldsymbol{C}) = \mathbb{P}_Y(Y \in \boldsymbol{C} \mid \boldsymbol{\vartheta}) = F_Y(\bar{y} \mid \boldsymbol{\vartheta}) - F_Y(\underline{y} \mid \boldsymbol{\vartheta}). \tag{2.1}$$

Example 2.12. For a binary response $Y \in \Xi = \{\upsilon_1 = \text{failure}, \upsilon_2 = \text{success}\}$ we possibly observe a failure $\boldsymbol{C} = \{\upsilon_1\} = (\upsilon_1, \upsilon_1]$. The likelihood is $l(\boldsymbol{\vartheta} \mid \boldsymbol{C}) = F_Y(\upsilon_1 \mid \boldsymbol{\vartheta}) - 0 = \mathbb{P}_Y(\text{failure})$. Alternatively, we might observe a success $\boldsymbol{C} = \{\upsilon_2\} = (\upsilon_1, \upsilon_2]$ and the likelihood changes to $l(\boldsymbol{\vartheta} \mid \boldsymbol{C}) = F_Y(\upsilon_2 \mid \boldsymbol{\vartheta}) - F_Y(\upsilon_1 \mid \boldsymbol{\vartheta}) = 1 - F_Y(\upsilon_1 \mid \boldsymbol{\vartheta}) = \mathbb{P}_Y(\text{success})$. □

Example 2.13. For an ordered categorical response measuring happiness (Example 2.2), a study participant is undecided between being "not to happy" and "somewhat happy". Instead of forcing her to chose either category, we observe the event $\{\text{not to happy}, \text{somewhat happy}\}$. This event can be written as the interval $\boldsymbol{C} = (\text{very unhappy}, \text{somewhat happy}]$ and the likelihood $l(\boldsymbol{\vartheta} \mid \boldsymbol{C}) = F_Y(\text{somewhat happy} \mid \boldsymbol{\vartheta}) - F_Y(\text{very unhappy} \mid \boldsymbol{\vartheta})$ is directly computed from the distribution function $F_Y(\cdot \mid \boldsymbol{\vartheta})$. □

Example 2.14. We observe $\boldsymbol{C} = \{15\} = (14, 15]$ wildlife-vehicle collisions (Example 2.5), so the likelihood if $l(\boldsymbol{\vartheta} \mid \boldsymbol{C}) = F_Y(15 \mid \boldsymbol{\vartheta}) - F_Y(14 \mid \boldsymbol{\vartheta})$. Or we maybe lost the records for December and observed 10 wildlife-vehicle collisions between January and November. The observation is then $\boldsymbol{C} =$

$\{k \in \mathbb{N} : k \geq 10\} = (9, \infty)$ and the likelihood is $l(\boldsymbol{\vartheta} \mid \boldsymbol{C}) = 1 - F_Y(9 \mid \boldsymbol{\vartheta})$. □

Example 2.15. We study the age at which a baby starts to walk. Some parents surely will have noted the exact date k (in days after birth) in their kids' diary, so we have $\boldsymbol{C} = \{k\} = (k-1, k]$ and likelihood $l(\boldsymbol{\vartheta} \mid \boldsymbol{C}) = F_Y(k) - F_Y(k-1)$. Some parents will tell us "Well, I don't remember the exact day, but it surely happened before her first birthday!". We then have $\boldsymbol{C} = (0, 365)$ and likelihood $l(\boldsymbol{\vartheta} \mid \boldsymbol{C}) = F_Y(364)$ (note the 364 days, because it was *before* the first birthday). □

Example 2.16. We study the time wanna-be-parents need for the (or one of the) female part to successfully conceive. The sample space Ξ counts the number of menstrual cycles which went by without resulting in being pregnant. Let's say a couple tried for seven months and later decides participation in our study is too much stress (and maybe even counterproductive), so they drop-out. But we still know that seven unsuccessful cycles went by, so $\boldsymbol{C} = \{8, \dots, \infty\}$ and we have $l(\boldsymbol{\vartheta} \mid \boldsymbol{C}) = 1 - F_Y(7 \mid \boldsymbol{\vartheta})$. □

Example 2.17. A woman giving birth lost 250 ml blood during the delivery. This measured blood loss is, of course, inaccurate, because the probability $\mathbb{P}_Y(Y = 250)$ of lossing *exactly* 250 ml is zero. So, this measured blood loss in fact represents an interval. We don't know the exact length of this interval, but let's assume that $\boldsymbol{C} = (225, 275]$ is a good proxy. So the likelihood is $l(\boldsymbol{\vartheta} \mid \boldsymbol{C}) = F_Y(275 \mid \boldsymbol{\vartheta}) - F_Y(225 \mid \boldsymbol{\vartheta})$. □

A common feature of all these examples is that the likelihood is evaluated for sets and thus can be computed (but not interpreted) as a probability: Given the parameters $\boldsymbol{\vartheta}$, the likelihood $l(\boldsymbol{\vartheta} \mid \boldsymbol{C})$ is equal to the probability of the datum $\boldsymbol{C} \in \mathfrak{C}$. As a function of the parameters $\boldsymbol{\vartheta}$, the likelihood l must not be interpreted as a probability. Therefore, the term "likelihood" is used instead of the term "probability", although their refer to the very same number.

So far, we only discussed the likelihood contribution of a single observation. Of course, reasonable statistical inference requires a sample of $i = 1, \dots, N$ observations $Y_i \sim \mathbb{P}_Y$. We restrict ourself to this simple situation for the time being and assume the observations are independently distributed. In light of what we just agreed upon, the notation $Y_1, \dots, Y_N \overset{\text{iid}}{\sim} \mathbb{P}_Y$ is problematic, because we cannot write $l(\boldsymbol{\vartheta} \mid Y_i)$. Each random variable Y_i takes values in Ξ and thus gives us an outcome $\upsilon \in \Xi$. The likelihood, however, is defined for events. Instead, we use the symbol $\boldsymbol{C}_1, \dots, \boldsymbol{C}_N$ to denote observations, that is events $\boldsymbol{C}_i \in \mathfrak{C}$, for which is is possible to write $l(\boldsymbol{\vartheta} \mid \boldsymbol{C}_i)$. The joint likelihood $L : \Theta \times \mathfrak{C}^N \rightarrow [0, 1]$ can now, assuming independence, be written as

$$\begin{aligned} L(\boldsymbol{\vartheta} \mid \boldsymbol{C}_1, \dots, \boldsymbol{C}_N) &= \mathbb{P}\left(\bigcap_{i=1}^N Y_i \in \boldsymbol{C}_i \mid \boldsymbol{\vartheta}\right) \\ &= \prod_{i=1}^N \mathbb{P}_Y(Y_i \in \boldsymbol{C}_i \mid \boldsymbol{\vartheta}) = \prod_{i=1}^N l(\boldsymbol{\vartheta} \mid \boldsymbol{C}_i). \end{aligned}$$

It is common to suppress the dependency of the likelihood on the observations by writing

$$L(\boldsymbol{\vartheta}) = \prod_{i=1}^N l_i(\boldsymbol{\vartheta}), \quad l_i(\boldsymbol{\vartheta}) := l(\boldsymbol{\vartheta} \mid \boldsymbol{C}_i).$$

Of course, we later want to optimise the joint likelihood with respect to it's argument $\boldsymbol{\vartheta}$ while treating the observations $\boldsymbol{C}_1, \dots, \boldsymbol{C}_N$ as fix. The resulting maximum likelihood estimator $\hat{\boldsymbol{\vartheta}}$ is the solution of the following optimisation problems

$$\hat{\boldsymbol{\vartheta}} = \arg\max_{\boldsymbol{\vartheta} \in \Theta} L(\boldsymbol{\vartheta}) = \arg\max_{\boldsymbol{\vartheta} \in \Theta} \log(L(\boldsymbol{\vartheta})) = \arg\max_{\boldsymbol{\vartheta} \in \Theta} \ell(\boldsymbol{\vartheta}).$$

The log-likelihood $\ell(\boldsymbol{\vartheta}) = \sum_{i=1}^N \ell_i(\boldsymbol{\vartheta})$ with N individual log-likelihood contributions $\ell_i(\boldsymbol{\vartheta}) = \log(l_i(\boldsymbol{\vartheta}))$ is theoretically and computationally convenient, because it is much easier to optimise sums instead of products.

Although the likelihood (2.1) applies to all measurement scales, we need to discuss aspects of three important special cases in a little more detail.

Likelihood for a Categorical Response

The sample space is $\Xi = \{\upsilon_1, \dots, \upsilon_K\}$ with $K = 2$ being the binary case. The distribution function is a step-function with possible jumps at each $\upsilon_k \in \Xi$. With the convention $\xi_0 \equiv 0 \leq \xi_1 \leq \xi_2 \leq \dots \leq \xi_{K-1} \leq \xi_K \equiv 1$ we can parameterise the distribution function by parameters $\boldsymbol{\xi} = (\xi_1, \dots, \xi_{K-1})^\top$

$$F_Y(\upsilon_k \mid \boldsymbol{\xi}) = \mathbb{P}_Y(Y \leq \upsilon_k \mid \boldsymbol{\xi}) = \xi_k, \quad k = 1, \dots, K.$$

The probabilities $\mathbb{P}_Y(Y \in \{\upsilon_k\} \mid \boldsymbol{\xi}) = \mathbb{P}_Y(Y = \upsilon_k \mid \boldsymbol{\xi}) = \xi_k - \xi_{k-1} =: \pi_k$ allow an alternative parameterisation by the probability density $F_Y(\upsilon_k \mid \boldsymbol{\pi}) = \pi_k$ for parameters $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)^\top$. The parameter space is $\boldsymbol{\pi} \in \Theta \subset \mathbb{R}^K$ under the constraint $\sum_{k=1}^K \pi_k = 1$.

We observe sets $\boldsymbol{C}_i = \{y_i\}$, *i.e.* each event consists of one single outcome $y_i = \upsilon_k$ only. For this data, we can write the corresponding likelihood as a function of parameters $\boldsymbol{\pi}$

$$L(\boldsymbol{\pi}) = \prod_{i=1}^N \sum_{k=1}^K \pi_k \mathbb{1}(y_i = \upsilon_k) = \prod_{k=1}^K \pi_k^{n_k}, \quad n_k = \sum_{i=1}^N I(y_i = \upsilon_k).$$

This is the likelihood of the multinomial $\mathrm{M}(\pi_1, \dots, \pi_K, N)$ distribution for single-valued observations *only*. It is well-known that L is maximised by $\hat{\pi}_k = n_k/N$. This looks different from (2.1) but the two are identical:

$$L(\boldsymbol{\pi}) = L(\boldsymbol{\xi}) = \prod_{k=1}^K (\xi_k - \xi_{k-1})^{n_k}.$$

Our definition (2.1) of the likelihood allows for more general observations to be handled: The observation $\boldsymbol{C}_i = \{\upsilon_1, \upsilon_2\}$, for example, means that the observer was undecided between the outcomes υ_1 and υ_2. The likelihood contribution for such an observation is $l_i(\boldsymbol{\vartheta}) = \pi_1 + \pi_2$. In the more complex notation $\boldsymbol{C}_i = \{\upsilon_1, \upsilon_2\} = (-\infty, \upsilon_2]$ we get $l_i(\boldsymbol{\vartheta}) = F_Y(\upsilon_2) - F_Y(-\infty) = F_Y(\upsilon_2) - 0 = \pi_1 + \pi_2$.

The Nonparametric Likelihood

The multinomial likelihood plays an important role for nonparametric inference. Consider any probability space $(\Xi, \mathfrak{C}, \mathbb{P}_Y)$ and a random sample

$Y_1, \dots, Y_N \overset{\text{iid}}{\sim} \mathbb{P}_Y$ from the corresponding experiment. Instead of working in the original space, we supplement a second probability space $(\tilde{\Xi}, \tilde{\mathfrak{C}}, \mathbb{P}_{\tilde{Y}})$ and call this the "empirical" probability space. The empirical sample space $\tilde{\Xi}$, consists of the unique *observed* outcomes $\upsilon_1 < \upsilon_2 < \dots < \upsilon_{K(N)}$ in the sample $Y_1 \le Y_2 \le \dots \le Y_N$. Outcomes from Ξ which have not been observed are considered impossible and thus not present in $\tilde{\Xi}$. The number of distinct outcomes $K(N) \le N$ depends on N and is discrete, regardless of the discreteness of the original sample space Ξ. If $\Xi \subseteq \mathbb{R}$ and there are no ties in the observations $Y_1 < Y_2 < \dots < Y_N$, we have $K(N) = N$.

The cumulative distribution function defining the distribution $\mathbb{P}_Y$ is also discrete with jumps exclusively at $\upsilon_k \in \tilde{\Xi}$, that is $\mathbb{P}_{\tilde{Y}}(\tilde{Y} \le \upsilon_k) = F_{\tilde{Y}}(\upsilon_k) = \xi_k$. We are thus exactly in the multinomial case discussed above, with $\hat{\pi}_k = n_k/N$ and thus

$$\hat{\xi}_k = \hat{F}_{\tilde{Y}}(\upsilon_k) = \hat{\pi}_1 + \dots + \hat{\pi}_k = N^{-1} \sum_{i=1}^{N} I(Y_i \le \upsilon_k) =: \hat{F}_N(\upsilon_k).$$

This is the empirical cumulative distribution function, the maximiser of the multinomial likelihood in the empirical probability space $(\tilde{\Xi}, \tilde{\mathfrak{C}}, \mathbb{P}_{\tilde{Y}})$. The big advantage of switching from a sample space $\Xi \subseteq \mathbb{R}$ with continuous distribution $\mathbb{P}_Y$ to a discrete sample space $\tilde{\Xi}$ with a discrete distribution $\mathbb{P}_{\tilde{Y}}$ is that $\mathbb{P}_{\tilde{Y}}(\upsilon) > 0$ but $\mathbb{P}_Y(\upsilon) \equiv 0$ for $\upsilon \in \{Y_1, \dots, Y_N\}$.

For an observation $Y_i = \upsilon_k$ we obtain the log-likelihood contribution $\ell_i(\boldsymbol{\xi}) = \log(F_{\tilde{Y}}(\upsilon_k \mid \boldsymbol{\xi}) - F_{\tilde{Y}}(\upsilon_{k-1} \mid \boldsymbol{\xi})) = \log(\xi_k - \xi_{k-1})$. This form of the likelihood reveals that we evaluate the probability of events $\boldsymbol{C}_i = (\upsilon_{k-1}, \upsilon_k)$ and for some other parameterisation we could also write $\ell_i(\boldsymbol{\vartheta}) = \log(F_Y(\upsilon_k \mid \boldsymbol{\vartheta}) - F_Y(\upsilon_{k-1} \mid \boldsymbol{\vartheta}))$. Thus, nonparametric maximum-likelihood estimation consists of two parts: (i) Defining data-driven events by $\boldsymbol{C}_i$ and (ii) parameterising the distribution function as a step function with jumps at υ_k.

The log-likelihood $\ell(\boldsymbol{\xi}) = \sum_{i=1}^{N} \ell_i(\boldsymbol{\xi})$ is called the nonparametric or empirical log-likelihood. The empirical cumulative distribution function $\hat{F}_N$ maximises this log-likelihood and is thus called the nonparametric maximum-likelihood estimator. From the classical maximum-likelihood point of view, were we parameterise distributions in terms of parameters $\boldsymbol{\vartheta}$ of fixed dimension P *before* seeing the data, the parameterisation by $\boldsymbol{\xi}$ is data-dependent and, quite distressingly, the number of parameters depends on the sample size N. We will see in later chapters that, from a practical point of view, these two seemingly orthogonal concepts are much less far apart from each other.

A Likelihood Approximation

Finally, we consider the case of an absolutely continuous random variable $Y \in \Xi \subseteq \mathbb{R}$ and an absolutely continuous distribution function F_Y, those derivative $f_Y(y) = F_Y'(y)$ exists everywhere. Technically, we have $\mathbb{P}_Y = f_Y \odot \mu_L$ with density f_Y dominated by the Lebesgue measure μ_L and $\mathfrak{C}$ the Borel σ-algebra.

It is important to note that it is impossible to observe events $\boldsymbol{C} = \{y\}$ consisting of a single real number $y \in \mathbb{R}$, simply because the sets $\boldsymbol{C} = \{y\}$ are empty sets with $\mathbb{P}_Y(Y \in \{y\}) = 0$. Instead, we always observe intervals $\boldsymbol{C} = (\underline{y}, \bar{y}]$. To keep things simple, let's say we have $(\underline{y}, \bar{y}] = (y - \epsilon, y + \epsilon]$ and $\epsilon > 0$ is neglegibly small. The likelihood is then *proportionally approximated* as

$$\begin{aligned} l(\boldsymbol{\vartheta} \mid (y-\epsilon, y+\epsilon]) &= F_Y(y+\epsilon \mid \boldsymbol{\vartheta}) - F_Y(y-\epsilon \mid \boldsymbol{\vartheta}) \\ &= \int_{(y-\epsilon, y+\epsilon]} f_Y \, d\mu_L = \int_{y-\epsilon}^{y+\epsilon} f_Y(u \mid \boldsymbol{\vartheta}) \, du \\ &\approx f_Y(y \mid \boldsymbol{\vartheta}) \times 2\epsilon \propto f_Y(y \mid \boldsymbol{\vartheta}). \end{aligned}$$

Using this approximation, it now perfectly makes sense to evaluate the likelihood for outcomes, that is, for real numbers $y \in \mathbb{R}$ by the density $f_Y(y \mid \boldsymbol{\vartheta})$. Figure 2.1 illustrates how this approximation works.

For a sample $Y_1, \dots, Y_N \overset{\text{iid}}{\sim} \mathbb{P}_Y$, very often the log-likelihood for both discrete and continuous responses is often *defined* as

$$\ell(\boldsymbol{\vartheta}) = \sum_{i=1}^{N} \ell_i(\boldsymbol{\vartheta}) = \sum_{i=1}^{N} \log(f_Y(Y_i \mid \boldsymbol{\vartheta})).$$

This notation carefully hides the fact that, suddenly, we evaluate the log-likelihood for outcomes and not events. This isn't much of a problem when

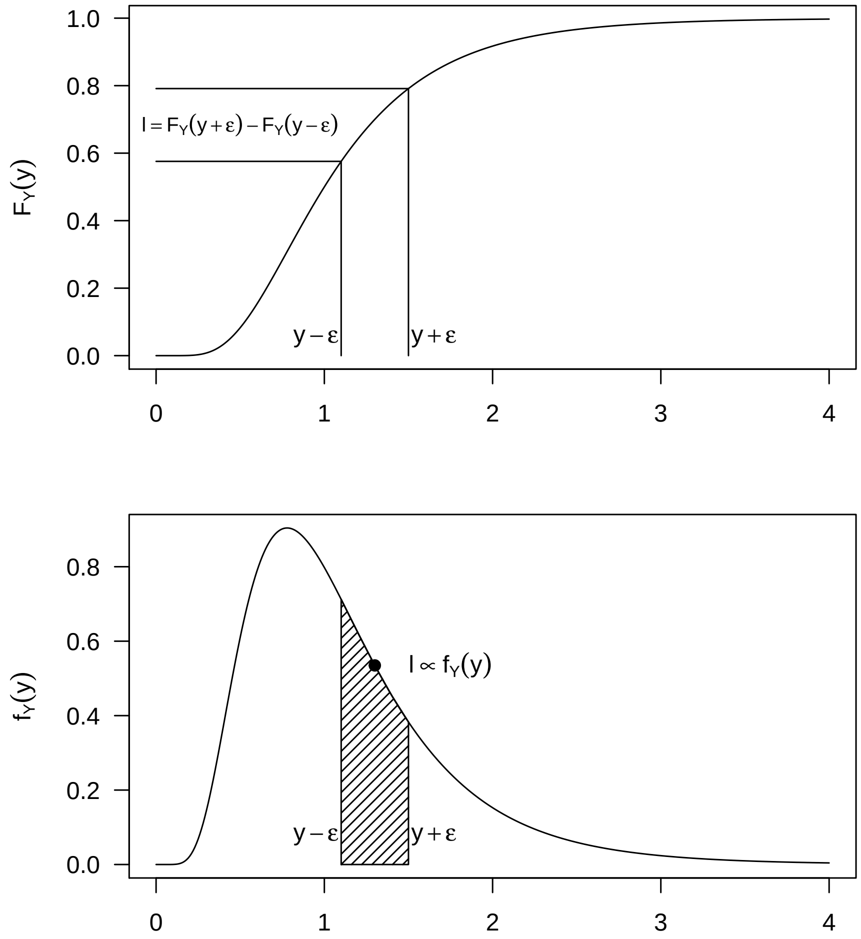


Figure 2.1: Approximation of a likelihood by a Lebesgue density. The likelihood l of the interval $(y - \epsilon, y + \epsilon]$ is given by the difference of the probabilities (top panel). The sample likelihood is the shaded area under the Lebesgue density $f_Y(y) = F'_Y(y)$. For short intervals, with ϵ small, this area can be approximated by $f_Y(y)2\epsilon$.

the sample space is discrete *and* we observe outcomes instead of events, because $\boldsymbol{C} = \{\upsilon_k\}$ is perfectly possible and this definition of the likelihood coincides with our definition (2.1). However, for the absolutely continuous case, it is an approximation to (2.1) and *not* the real thing. When using this approximation, technical difficulties arise, as $f_Y(y \mid \boldsymbol{\vartheta}) = \infty$ is perfectly possible for some distributions (the β-distribution density evaluated at $y = 0$ zero with both parameters being less than one, for example). Why is this approximation so popular? The reason is that estimators arising from optimising the approximate log-likelihood are much easier to compute (analytically or numerically) and their theoretical properties much easier to study mathematically. These are perfectly sound reasons to rely on approximations, but we must not forget to go back to the original concept in cases where this approximation is inappropriate.

Example 2.18. Consider the normal distribution $\mathbb{P}_Y = \mathrm{N}(\mu, \sigma^2)$ with mean and standard deviation parameters $\boldsymbol{\vartheta} = (\mu, \sigma)^\top, \sigma > 0$. The Lebesgue density is $f_Y(y \mid \boldsymbol{\vartheta}) = (\sqrt{2\pi}\sigma)^{-1} \exp(-(y-\mu)^2(2\sigma)^{-1})$ and the absolutely continuous distribution function is $F_Y(y \mid \boldsymbol{\vartheta}) = \Phi((y-\mu)/\sigma)$. For some observation $(y - \epsilon, y + \epsilon]$ the log-likelihood contribution is

$$\log\left(\Phi\left(\frac{y+\epsilon-\mu}{\sigma}\right) - \Phi\left(\frac{y-\epsilon-\mu}{\sigma}\right)\right).$$

This is a nightmare to evaluate numerically: Φ is not available analytically and the corresponding integral has to be approximated numerically. The log-likelihood contribution depends on two parameters, and we would have to maximise with respect to both μ and σ simultaneously. Instead, the approximation by the log-density

$$\log(f_Y(y \mid \boldsymbol{\vartheta})) \propto -\log(\sigma) - (y-\mu)^2/(2\sigma)$$

is extremely simple to evaluate. Furthermore, the maximum likelihood estimator for μ based on the observations $Y_1, \dots, Y_N \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$ can be derived *analytically* without paying attention to σ. Of course, $\hat{\mu} = N^{-1} \sum_{i=1}^N Y_i$ is the sample mean. This again is a linear function of normals, so we have the exact analytical result $\hat{\mu} \sim \text{N}(\mu, \sigma^2/N)$. This and other theoretical properties of such estimators maximising the approximate log-likelihood are relatively easy to derive, whereas such an exercise is much harder when insisting to use the log-likelihood (2.1). □

Example 2.19. The famous Boston Housing data has been used as a benchmark for "regression" models since 1978. The response is median housing price in a specific area. Generations of statisticians and machine learners have analysed this response under squared error $(y-\mu)^2$, *i.e.* using the log-density of a normal. However, observations with values $y = 50$ in the data set represent median housing values *equal to or larger than* 50, so the log-probability $\log(1 - \mathbb{P}_Y(50 \mid \boldsymbol{\vartheta}))$ of the interval $[50, \infty)$ is the correct contribution of the likelihood and $\log(f_Y(50 \mid \boldsymbol{\vartheta}))$ is not meaningful. With today's computing power, the evalution of the correct log-likelihood is not such a big deal. However, implementations of standard models (for example `lm()` or `randomForest()` in the R system for statistical computing) do not allow such intervals to be handled correctly. □

Example 2.20. For count data $Y_1, \dots, Y_N \overset{\text{iid}}{\sim} \text{Po}(\lambda)$ with $Y_i \in \Xi = \mathbb{N}$ the discrete density of the Poisson distribution is

$$f_Y(k) = \frac{\exp(-\lambda)\lambda^k}{k!}$$

and thus the log-likelihood contribution is $-\lambda k \log(\lambda) - \log(k!)$. This likelihood is commonly implemented, such as in `glm()` in the R system for statistical computing. However, the log-likelihood contribution of the event "at least 10 counts" $\{10, 11, \dots\}$ is

$$\log\left(\sum_{k=10}^{\infty} \frac{\exp(-\lambda)\lambda^k}{k!}\right) = \log(1 - F_Y(9 \mid \lambda)).$$

Such a likelihood is practically relevant, however, standard implementations, such as `glm()`, simply do not allow for such events to be observed as a consequence of understanding the log-likelihood as a log-density. □

It is fine to cut corners when appropriate, but the common practise of *defining* the log-likelihood as a log-density is clearly overdoing it. The likelihood is defined by events and thus by probabilities. The cumulative distribution function is central in computing these probabilities, and models directly formulated in terms of distribution functions are readily applicable when a likelihood shall be computed. Transformation models operate on the scale of the distribution function and are thus attractive candidates for likelihood-based inference.

Further Reading

My colleguage Leonhard Held suggested the paper by Lindsey (1996) to me, and this was really an eye-opener.

Chapter 3

Univariate Distributions

The discussion in Chapter 2 on sample spaces Ξ appropriate for different measurement scales and the form of the corresponding distribution functions F_Y might have left the reader with the impression that the structure of the distribution function directly follows from what is measured: Discrete responses Y and corresponding discrete sample spaces give rise to the distribution function being a step-function with potential jumps at the outcomes $\upsilon \in \Xi$. The distribution of continuous responses $Y \in \mathbb{R}$ would, in this line of thinking, be modelled by a continuous distribution function F_Y. I plead guilty to intentionally raising this understanding, because I believe this is an appropriate way of modelling distributions. This is, however, not always agreed upon, as the next example shows.

Example 3.1. The body mass index (BMI), defined as the ratio of body mass (in kilogram) and squared body size (in meter^2), is an important epidemiological assessment. Conceptually, this is a positive variable Y with $\Xi = (0, \infty)$. The World Health Organisation (WHO) defines four BMI categories: underweight $Y \in (0, 18.5]$, normal weight $Y \in (18.5, 25]$, overweight $Y \in (25, 30]$, and obese $Y \in (35, \infty)$. Many experiments are designed for the sample space $\Xi = \{\text{underweight}, \text{normal weight}, \text{overweight}, \text{obese}\}$ and the actual measurements are one of these categories, although somebody must have measured the original BMI ratio in the first place. Only recording one out of four categories limits our ability to evaluate probabilities for cut-off points other than the ones chosen by the WHO. In fact, it is very hard, or even impossible, to compare two studies based on different categorisations and thus different sample spaces and distribution functions. These problems disappear when modelling BMI as a continuous variable with continuous distribution function *although* the observations might be intervals (such as, for example, $(18.5, 25]$). □

Example 3.2. For a truely binary response, such as failure or success, the binomial model $Y \sim \text{B}(\pi, 1)$ with distribution function $F_Y(\text{failure}) = \pi$ and $F_Y(\text{success}) = 1$ is the only option and thus undisputed. □

Example 3.3. The distribution of happiness (Example 2.2) is modelled by the $Y \sim \text{M}(\pi_1, \dots, \pi_4, 1)$ multinomial distribution, with $\pi_1 = \mathbb{P}_Y(\text{very unhappy})$ and so on. There is basically no alternative to this choice. □

Example 3.4. Models for count data Y, such as wildlife-vehicle collisions discussed in Example 2.5, are a bit less clear. The multinomial distribution with $\pi_k = \mathbb{P}_Y(Y = k)$ for $k = 0, \dots, K < \infty$ is possible when there is a finite maximal number of counts conceptually observable. When K is large, the number of parameters is also large, and in some circumstances, K and thus Y might approach really large numbers. In these cases, adding structure to the distribution function is helpful, and chosing a low-dimensional parametric model (for example the negative binomial distribution featuring two instead of K parameters), might be attractive. □

The goal of this chapter is to study options we have for parameterising distribution functions $F_Y(\cdot \mid \boldsymbol{\vartheta})$ in terms of a relatively small set of parameters $\boldsymbol{\vartheta} \in \Theta \subseteq \mathbb{R}$. Of course, our motivation for this exercise comes from our desire to estimate these unknown parameters $\boldsymbol{\vartheta}$ later on, for example by the maximum-likelihood estimator $\hat{\boldsymbol{\vartheta}} = \arg\max_{\boldsymbol{\vartheta} \in \Theta} \ell(\boldsymbol{\vartheta})$. We still only consider the univariate unconditional case; even the lightweight notation in this simple setup will cause enough headaches for the moment.

3.1 Parameters Defining Distributions

For finite discrete sample spaces of cardinality $K =| \Xi |< \infty$, the structure of the distribution function is determined by the corresponding binomial ($K = 2$) or multinomial ($K > 2$) distribution. In these cases, we basically have no choice of picking a different distribution. The more interesting and practically relevant case is a numeric (not necessarily real) sample space $\Xi \subseteq \mathbb{R}$. We now study different possible parameterisations of F_Y.

Sometimes, the scientific question we try to answer with our experiment can be formulated in terms of a single event, such as "Age younger than 65 years?", or "Less than 5 wildlife-vehicle collisions per year and road segment?". In some experiments these cut-off points have been carved in stone a priori and we are not interested in other cut-off points. This means we can partion the sample space in two dijunct intervals $\Xi = (-\inf, \upsilon_1] \dot{\cup} (\upsilon_1, \inf)$, with $\upsilon_1 \in \mathbb{R}$ being the a priori fixed cut-off point.

In this setup, the distribution function we are interested in can be parameterised in terms of the probability $\xi_1 = \mathbb{P}_Y(Y \le \upsilon_1)$ as

$$F_Y(y \mid \xi_1) = \begin{cases} \xi_1 & y \le \upsilon_1 \\ 1 & y > \upsilon_1. \end{cases}$$

Example 3.5. Postpartum haemorrhage (introduced in the Introduction) is often defined as blood loss of more than 1000 ml in the first day after giving birth to a child. Thus, $\upsilon_1 = 1000$ is the cut-off point we are interested in and $1 - F_Y(1000)$ the probability of postpartum haemorrhage. Analysts exclusively interested in this probability will be perfectly happy with a distribution function parameterised in terms of ξ_1 above. □

Instead of one single cut-off point υ_1, we might choose a set of $K - 1$ cut-off points $\upsilon_1 < \upsilon_2 < \cdots < \upsilon_{K-1} \in \Xi$ and thus partition the sample space into K disjunct intervals

$$\Xi = \dot{\bigcup_{k=1}^{K}} (\upsilon_{k-1}, \upsilon_k]$$

with $\upsilon_0 = -\infty$ and $\upsilon_K = \infty$. A discrete density can now be defined by the probability of each of these K events $\pi_k = \mathbb{P}_Y(Y \in (\upsilon_{k-1}, \upsilon_k])$ for $k = 1, \dots, K$. The corresponding distribution function jumps at υ_k by the amount π_k. The parameters $\xi_k = \pi_1 + \dots + \pi_k$ describe the probabilities $\xi_k = \mathbb{P}_Y(Y \le \upsilon_k) \in [0, 1]$. This allows parameterising the distribution function in terms of parameters $\boldsymbol{\xi} = (\xi_1, \dots, \xi_{K-1})^\top$ as

$$F_Y(y \mid \boldsymbol{\xi}) = \begin{cases} \xi_1 & y \le \upsilon_1 \\ \xi_2 & y \le \upsilon_2 \\ \vdots & \\ \xi_{K-1} & y \le \upsilon_{K-1} \\ 1 & y > \upsilon_{K-1}. \end{cases}$$

Example 3.6. The WHO defined the cut-off points $\upsilon_1 = 18.5, \upsilon_2 = 25, \upsilon_3 = 25$ for their four BMI categories. Distribution functions defined in terms of ξ_1, ξ_2, ξ_3 as above would be appropriate *if* we commit to these cut-off points before analysing or even observing the data. □

Example 3.7. The definition of postpartum haemorrhage as blood loss of more than 1000 ml in the first day after giving birth to a child is controversial. Different cut-off points $\upsilon_1 = 500$, $\upsilon_2 = 750$, or $\upsilon_3 = 1000$ are discussed and $1 - F_Y(\upsilon_k)$ the probability of postpartum haemorrhage for the k cut-off point. We can parameterise the distribution function by corresponding parameters ξ_1, ξ_2, ξ_3 and estimate all parameters simultaneously. □

So far we understood that the cut-off points υ_k were motivated from an a priori posed scientific question. The nonparametric approach simply determines the cut-off points $\upsilon_1 < \upsilon_2 < \dots < \upsilon_{K-1}$ from a sample $Y_1, \dots, Y_N \overset{\text{iid}}{\sim} \mathbb{P}_Y$, *i.e. after* seeing the data. Each υ_k occurs at least once in $Y_1, \dots, Y_N$ and we defined $\upsilon_0 = -\infty$ and $\upsilon_K = \max_i Y_i$. The only difference to the categorical case is that K potentially depends on N.

Example 3.8. Measured blood loss values $\upsilon_1 = 100 < \upsilon_2 = 150 < \upsilon_3 = 200 < \dots < \upsilon_{39} = 5700$ were observed by Haslinger et al. (2020). We can thus parameterise the distribution function by parameters $\boldsymbol{\xi} = (\xi_1, \dots, \xi_{38})^\top$ such that $\xi_k = \mathbb{P}(Y \leq \upsilon_k), k = 1, \dots, 38$. □

Once we parameterised the distribution function $F_Y(y \mid \boldsymbol{\xi})$ in terms of the probability parameters $\xi_k = \mathbb{P}(Y \leq \upsilon_k), k = 1, \dots, K-1$, we can estimate ξ_k by the empirical cumulative distribution function $\hat{\xi}_k = N^{-1} \sum_{i=1}^N I(Y_i \leq \upsilon_k) = \hat{F}_Y(\upsilon_k)$. This choice is very attractive, both from a practical and a theoretical point of view. A plot of the empirical cumulative distribution function does not only show an estimate of the true unknown distribution F_Y but is also an efficient display of the observations: The steps υ_k denote the values occuring in the observations $Y_1, \dots, Y_N$. The frequency of υ_k in the data is given by $N\pi_k = N(\xi_k - \xi_{k-1})$. Thus, we can reconstruct the raw observations $Y_1, \dots, Y_N$ from a plot (at high enough resolution) depicting the corresponding empricical cumulative distribution function $\hat{F}_Y$. Theoretically, we also have $\hat{F}_Y(\upsilon) \overset{\text{as}}{\to} F_Y(y)$ uniformly for all y by the famous Glivenko-Cantelli theorem.

In this context of distribution functions represented by step-functions, a transformation model $F_Y(y) = F_Z(h(y))$ is just a convenient reparameterisation. For the moment, we understand that F_Z is an absolutely continuous cumulative distribution function defined on the real line (we will make this more precise later in this chapter). With $\xi_k = F_Y(\upsilon_k \mid \boldsymbol{\xi}) = F_Z(h(\upsilon_k \mid \boldsymbol{\vartheta}))$ we replace our probability parameters $\xi_k \in [0, 1]$ by parameters $\vartheta_k = F_Z^{-1}(\xi_k) \in \mathbb{R}$ and obtain

$$F_Y(y \mid \boldsymbol{\vartheta}) = F_Z(h(y \mid \boldsymbol{\vartheta})) = \begin{cases} F_Z^{-1}(\vartheta_1) & y \leq \upsilon_1 \\ F_Z^{-1}(\vartheta_2) & y \leq \upsilon_2 \\ \vdots & \\ F_Z^{-1}(\vartheta_{K-1}) & y \leq \upsilon_{K-1} \\ 1 & y > \upsilon_{K-1}. \end{cases}$$

The transformation function is therefore

$$h(y \mid \boldsymbol{\vartheta}) = \begin{cases} \vartheta_1 & y \le \upsilon_1 \\ \vartheta_2 & y \le \upsilon_2 \\ \vdots & \\ \vartheta_{K-1} & y \le \upsilon_{K-1} \\ \infty & y > \upsilon_{K-1}. \end{cases}$$

This exericise seems completely unnecessary. However, using $\boldsymbol{\vartheta}$ instead of $\boldsymbol{\xi}$ we got rid of an inconvenient constraint: The elements ξ_k are increasing probabilities between zero and one, whereas the elements ϑ_k are increasing real numbers. This reparameterisation will help a great deal when estimating parameters in more complex models. In the unconditional setup here, we of course have $\hat{\vartheta}_k = F_Z^{-1}(\hat{\xi}_k)$ and we, unfortunately, loose the direct interpretation of the parameters $\boldsymbol{\vartheta}$ in terms of probabilities.

Before turning to our final argument in this section, note that we can write our transformation function $h(y \mid \boldsymbol{\vartheta})$ as a linear function (think of computing later on!) of it's parameters $\boldsymbol{\vartheta}$. The *basis function* $\boldsymbol{a}(y)$ is a unit vector of length K. This means that this vector is zero except for one of it's elements. The argument y determines the position of the only element taking the value one: For $y \in (\upsilon_{k-1}, \upsilon_k]$, the kth element of $\boldsymbol{a}(y)$ is one. The parameters are $\boldsymbol{\vartheta} = (\vartheta_1, \dots, \boldsymbol{\vartheta}_{K-1}, \infty)^\top$ and with the convention $\infty \times 0 = 0$ we can write $h(y \mid \boldsymbol{\vartheta}) = \boldsymbol{a}(y)^\top \boldsymbol{\vartheta}$. This notation seems overly academic in the simple setup we consider here, but it is of utmost importance for our treatment of more complex transformation models in later chapters.

Now, finally, we turn back to argument presented in the beginning: If the sample space $\Xi = \mathbb{R}$ suggests a continuous true unknown distribution function F_Y, shouldn't we expect a continuously estimated distribution function $\hat{F}_Y$ from our preferred statistical method? This is not only an aesthetic question but, in contrast to an estimated step function jumping at discrete places only, a smooth $\hat{F}_Y$ also gives us the freedom to evaluate probabilties for arbitrary events $(-\infty, y]$ for all possible $y \in \mathbb{R}$. With the preparations we performed so far, setting-up a smoothly parameterised distribution function seems a straightforward exercise in a transformation model: We pick a transformation function $h(y) = \boldsymbol{a}(y)^\top \boldsymbol{\vartheta}$ being continuous in y with parameters $\boldsymbol{\vartheta} \in \mathbb{R}^P$, where the number of parameters P may or may not depend on the sample size N. We already got a taste of such smooth transformation models in the Introduction.

Example 3.9. Figure 3.1 shows three estimates of the distribution of post-partum blood loss, using three different parameterisations. First, we are only interested in the cut-off points $\upsilon_1 = 500 < \upsilon_2 = 750 < \upsilon_3 = 1000$. These cut-off points are commonly referred to in the medical literature, and it thus makes sense to report the corresponding probabilities $\hat{\xi}_1 = 0.704$, $\hat{\xi}_2 = 0.854$, and $\hat{\xi}_3 = 0.955$.

The empirical cumulative distribution function $\hat{F}_Y$ evaluated at υ_1, υ_2, and υ_3 agrees with these numbers, because the values $\upsilon_1 = 500 < \upsilon_2 = 750 < \upsilon_3 = 1000$ were actually observed in the data. The graph of the empirical cumulative distribution function jumps at all 39 uniquely observed values, and not only at the three a priori chosen values $\upsilon_1 = 500 < \upsilon_2 = 750 < \upsilon_3 = 1000$.

A smooth estimate of the continous distribution function F_Y interpolates the empirical cumulative distribution function. The dot-dashed line in Figure 3.1 is based on a certain spline featuring $P = 7$ parameters.

□

Before we can proceed with the practical aspects of discrete and smooth transformation models, we need to lay out a bit of theoretical foundation of the transformations involved.

3.2 Probabilities and Distributions

One cannot understand statistical models without understanding the concept of a distribution. We sketch the most important ideas of probability theory and the concept of a distribution in this section before introducing the concept of a transformation function.

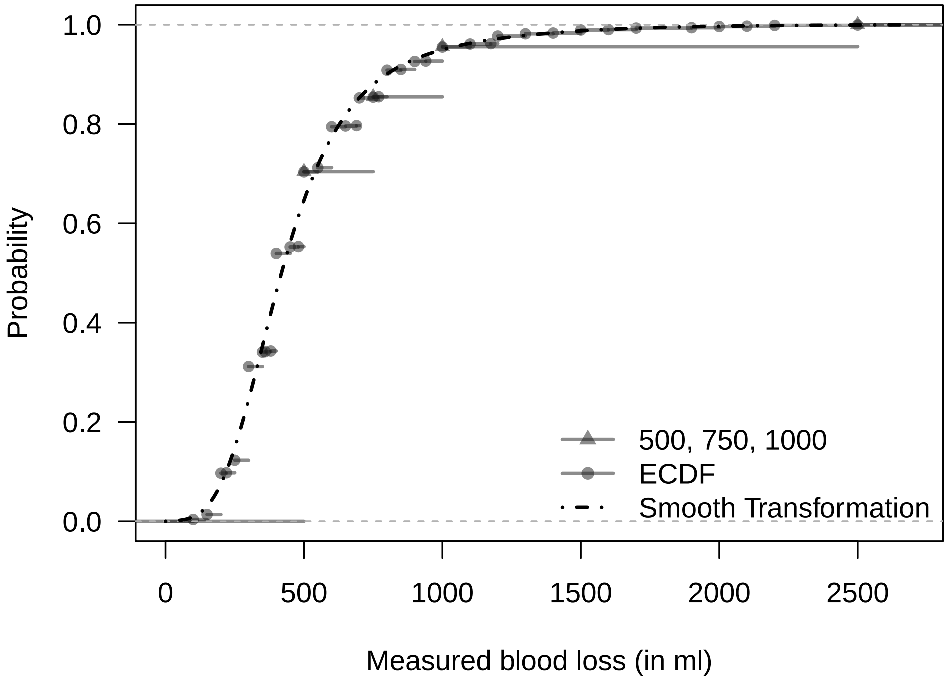


Figure 3.1: Measured blood loss. Probability of loosing less than 500, 750, and 1000 ml blood (triangles), empirical cumulative distribution function (dots), and smooth distribution function (dot-dashed line) obtained from a spline-based transformation model.

The foundation of every distribution is a probability space $(\Omega, \mathfrak{A}, \mathbb{P})$. This triple consists of the sample space Ω, the event space $\mathfrak{A}$ and a probability measure $\mathbb{P}$ and describes an experiment. The possible outcomes $\omega \in \Omega$ of the experiment are also called elementary events. Subsets of such outcomes $\boldsymbol{A} \subseteq \Omega$ are called events and the set of all events is the event space $\mathfrak{A}$, also called the σ-algebra of Ω. The probability measure $\mathbb{P}$ assigns the probability

$$\mathbb{P} : \mathfrak{A} \to [0, 1]; \quad \mathbb{P}(\boldsymbol{A}) \in [0, 1] \, \forall \boldsymbol{A} \in \mathfrak{A} \tag{3.1}$$

to all events $\boldsymbol{A} \in \mathfrak{A}$. The space $(\Xi, \mathfrak{C})$, *i.e.* the probability space without a probability measure, is called measureable space.

Dealing with arbitrary sample spaces Ω is very tedious. The idea of a random variable Y is to *transform* potentially complex outcomes ω to an element $Y(\omega) \in \Xi$ of a simpler sample space Ξ. The term "random variable" for Y is extremely misleading, in fact, it is a function whose properties are explained next.

3.3 Distributions of Random Variables

A random variable is a function $Y : \Omega \to \Xi$. The function $y = Y(\omega)$ maps each element of Ω onto an element $y \in \Xi$. Random variables are $\mathfrak{A} - \mathfrak{C}$ measureable functions. This property ensures that for all events $\boldsymbol{C} \in \mathfrak{C}$, the preimage

$$Y^{-1}(\boldsymbol{C}) = \{\omega \in \Omega \mid Y(\omega) \in \boldsymbol{C}\} \in \mathfrak{A} \tag{3.2}$$

is again an event in the event space of the probability space $(\Omega, \mathfrak{A}, \mathbb{P})$. One can therefore obtain probabilities for all events $\boldsymbol{C} \in \mathfrak{C}$ from $\mathbb{P}$ as

$$\mathbb{P}_Y(\boldsymbol{C}) = \mathbb{P}\left(Y^{-1}(\boldsymbol{C})\right). \tag{3.3}$$

The probability measure $\mathbb{P}_Y$ is also called the distribution of the random variable Y, in short $Y \sim \mathbb{P}_Y$.

The probability space $(\Omega, \mathfrak{A}, \mathbb{P})$ is often neglected and one directly defines probability measures $\mathbb{P}_Y$ in the probability space $(\Xi, \mathfrak{C}, \mathbb{P}_Y)$ (as we did in Chapter 2). This fact is reflected in the standard notation. One commonly writes $Y = y$ meaning $Y(\omega) = y$ or

$$\mathbb{P}(Y \in \boldsymbol{C}) = \mathbb{P}_Y(\boldsymbol{C}) \tag{3.4}$$

although, strictly speaking, the probability measure $\mathbb{P}$ is not defined for elements $\boldsymbol{C} \in \mathfrak{C}$ and Y is a function with $Y(\omega) \in \Xi$.

The advantage of dealing with the probability space $(\Xi, \mathfrak{C}, \mathbb{P}_Y)$ is the relative simplicity of defining an appropriate probability measure $\mathbb{P}_Y$. We assume that the elements of Ξ are ordered, something that is difficult or impossible to assume for Ω.

Example 3.10. The sample space $\Omega = \{\upsilon_1, \dots, \upsilon_K\}$ represents a population of K persons, so the outcome of our experiment is a single person. We cannot order persons, but we can measure something for each person $\upsilon_k \in \Omega$. For example, $Y(\omega_k) = y \in \mathbb{R}^+$ is the BMI of person k. We can now order the BMI values in the new sample space Ξ and thus can understand a distribution function as

$$\begin{aligned} F_Y(y) &= \mathbb{P}_Y(\{\upsilon \in \Xi : \upsilon \leq y\}) = \mathbb{P}(\{\omega \in \Omega : Y(\omega) \leq y\}) \\ &\overset{\mathbb{P}(\{\omega\}) \equiv K^{-1}}{=} K^{-1} \mid \{\omega \in \Omega : Y(\omega) \leq y\} \mid . \end{aligned}$$

The last equation assumes a discrete uniform distribution $\mathbb{P}$ on Ω, assigning probability K^{-1} to each outcome υ_k. □

Density and Distribution Functions

For an ordered sample space Ξ, we define the distribution function of the random variable Y

$$F_Y(y) = \mathbb{P}_Y(\{\upsilon \in \Xi \mid \upsilon \leq y\}).$$

It is important to note that this definition applies to finite, countably infinite, and infinite sample spaces Ξ. The probability measure is dominated by a dominating measure μ in the sense that $\mu(\boldsymbol{C}) = 0 \Rightarrow \mathbb{P}_Y(\boldsymbol{C}) = 0$ for some empty set $\boldsymbol{C} \in \mathfrak{C}$. The two most important dominating measures are the counting measure $\mu = \mu_\#$ for discrete or the Lebensgue measure $\mu = \mu_L$ for continuous distributions.

Example 3.11. The distribution of the amount of precipitation for a meteorological station at one day is neither dominated by the counting nor the Lebesgue measure, because the probability of no rain $F_Y(0) > 0$ but $\mu_L(0) = 0$. A mixed discrete-continuous distribution is required to model this response. □

The famous Radon-Nikodým theorem ensures the existence of a density function f_Y for the probability measure $\mathbb{P}_Y$. We write $\mathbb{P}_Y = f_Y \odot \mu$, so the distribution F_Y is charaterised by a probability density function $f_Y : \Xi \rightarrow \mathbb{R}^+$ and the dominating measure μ. We can evaluate probabilities by Lebesgue integration

$$\mathbb{P}_Y(\boldsymbol{C}) = \int_{\boldsymbol{C}} d\mathbb{P}_Y = \int_{\boldsymbol{C}} df_Y \odot \mu = \int_{\Xi} \mathbb{1}(\upsilon \in \boldsymbol{C}) f_Y(\upsilon)\, d\mu(\upsilon). \qquad (3.5)$$

The two most important special cases are absolutely continuous responses with Lebesgue measure $\mu = \mu_L$ where the Lebesgue integral reduces to the Riemann integral

$$\mathbb{P}_Y(\boldsymbol{C}) = \int_{\boldsymbol{C}} f_Y(\upsilon)\, d\upsilon \overset{\boldsymbol{C}=(-\infty,y]}{=} \int_{-\infty}^{\upsilon} f_Y(\upsilon)\, d\upsilon = F_Y(y)$$

which allows evaluation of the distribution function $F_Y(y)$ from a given density f_Y by integrating over the interval $(-\infty, y]$, and, consequently, the density $f_Y(y) = F_Y'(y)$ is the derivative of the distribution function. It is important to note that $f_Y(y) \geq 0$ is *not* a probability and $f_Y(y) > 1$ (and even $f_Y(y) = \infty$) is very well possible.

For discrete responses and the counting measure $\mu = \mu_\#$ we have

$$\mathbb{P}_Y(\boldsymbol{C}) = \sum_{\upsilon \in \boldsymbol{C}} f_Y(\upsilon) \overset{\boldsymbol{C}=\{\upsilon:\upsilon \leq y\}}{=} \sum_{\upsilon:\upsilon \leq y} f_Y(y) = F_Y(y)$$

and the corresponding density $f_Y(\upsilon_k) = F_Y(\upsilon_k) - F_Y(\upsilon_{k-1})$ has a direct interpretation as a probability, so $f_Y(y) > 1$ is not possible.

The density f_Y or the distribution function F_Y uniquely characterise the distribution $\mathbb{P}_Y$. Other means to characterise this probability measure are the survivor function, the odds function, the hazard and cumulative hazard function, and many more other options.

Survivor Functions

The survivor function $S_Y(y)$ evaluates the probability of observing an outcome larger then y, that is $S_Y(y) = 1 - F_Y(y) = \mathbb{P}_Y(Y > y)$. Because the cumulative distribution function F_Y is monotonically increasing, the survivor function is monotonically decreasing. The name of the function origines from responses Y measuring time-to-death. In this context $S_Y(y)$

is the probability of surviving time y, and for obvious reasons the notation $S_T(t) = \mathbb{P}_T(T > t)$ is used.

Odds Functions

The odds of a probability p is given as $p/(1-p)$. We define the odds function $O_Y : \Xi \to \mathbb{R}^+$ for the probability $\mathbb{P}_Y(Y \le y)$ by $O_Y(y) = F_Y(y)/(1-F_Y(y))$. This function is positive and monotonically increasing. The positivity constraint is removed when considering the log-odds function $\log(O_Y(y)) = \log(F_Y(y)) - \log(1 - F_Y(y)) = \log(F_Y(y)) - \log(S_Y(y))$.

Quantile Functions

The quantile function $Q_Y : [0, 1] \to \Xi$ for some probability $p \in [0, 1]$ returns the quantile $y(p) = Q_Y(p) = F_Y^{-1}(p) = \inf\{y \in \Xi \mid F_Y(y) \ge p\}$. This function is monotonically increasing and takes values in the sample space.

Hazard Functions

The hazard function $\lambda_Y : \Xi \to \mathbb{R}^+$ is defined as $\lambda_Y(y) = f_Y(y)/(1 - F_Y(y)) = f_Y(y)/S_Y(y)$. For discrete sample spaces, the hazard of y is equivalent to the conditional probability $\mathbb{P}_Y(Y = y \mid Y \ge y)$. In survival analysis with Y being a discrete time-to-death (in days after being admitted to an intensive-care-unit, maybe), this probability has a special interpretation: $\mathbb{P}_Y(Y = y \mid Y \ge y)$ is the conditional probability of dying at time y given that one has survived to that time. Unfortunately, time is usually continuous and for time-to-event $Y > 0$ the hazard is an intensity of dying between time y and $y+\epsilon$ conditional on having survived until y with the interval becoming very small ($\epsilon \to 0$). For absolute continuous distributions, the hazard function is equivalent to the derivative of $\log(1 - F_Y(y))$ with respect to y. The hazard function is positive but otherwise unconstrained. So, the log-hazard function $\log(\lambda_Y(y))$ allows a completely free and unconstrained characterisation of a distribution F_Y. If you want to "invent"

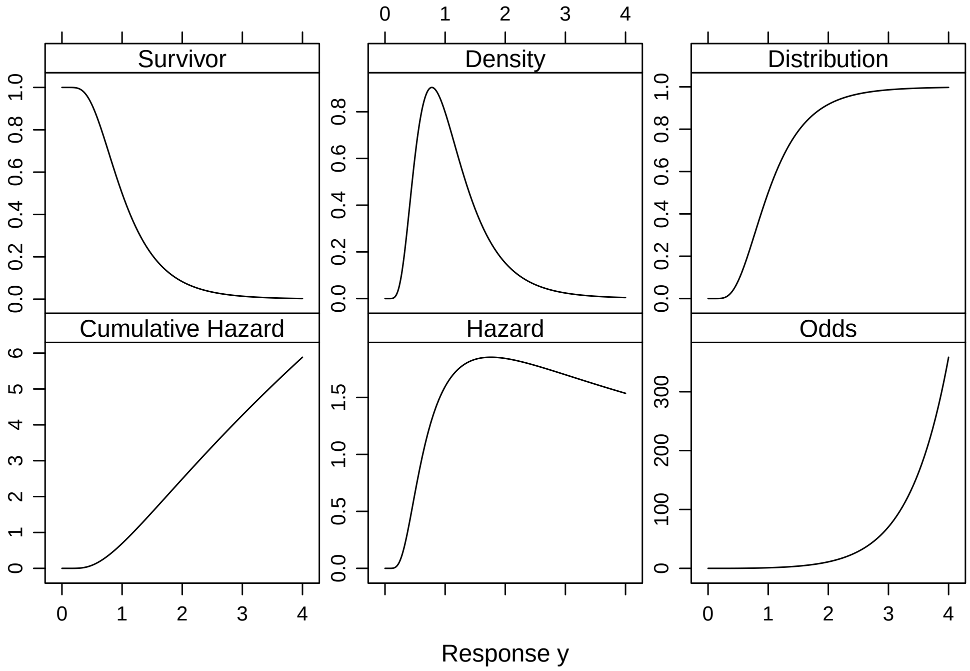


Figure 3.2: Six different characterisation of a distribution $\mathbb{P}_Y$ of a positive continuous random variable Y. The quantile function is the inverse of the distribution function.

your own probability distribution, just sketch a funny-looking log-hazard function and put your name next to it.

Example 3.12. In the context of Example 2.16 and Y denoting the number of menstrual cycles needed to conceive, the hazard is the probability to score during the ongoing yth cycle conditional on $y - 1$ unsuccessful attempts (cycles, not intercourses!). □

The reverse time hazard is $f_Y(y)/F_Y(y)$ which, in the discrete case, corresponds to $\mathbb{P}_Y(Y = y \mid Y \leq y)$ and in the continuous case to the derivative of $\log(F_Y(y))$ with respect to y.

Cumulative Hazard Functions

The cumulative hazard function $\Lambda_Y : \Xi \rightarrow \mathbb{R}^+$ integrates the hazard function

$$\Lambda_Y(y) = \int_\Xi \mathbb{1}(\upsilon \leq y)\lambda_Y(\upsilon)\, d\mu(\upsilon).$$

A simpler relationship to the survivor function exists via $\Lambda_Y(y) = -\log(1 - F_Y(y))$. For continuous responses Y, the Riemann integral

$$\Lambda_Y(y) = \int_{-\infty}^{y} \lambda_Y(\upsilon)\, d\upsilon$$

computes the cumulative hazard function, and in the discrete case we have

$$\Lambda_Y(y) = \sum_{\upsilon:\upsilon \leq y} \lambda_Y(\upsilon).$$

The different functions characterising a distribution $\mathbb{P}_Y$ are illustrated in Figure3.2 for a positive continuous response Y and in Figure 3.3 for a discrete response. In the next section, we introduce the transformation function h, which can also be understood as an additional unique characterisation of a distribution $\mathbb{P}_Y$.

3.4 Transformations

The probabilty space $(\Xi, \mathfrak{C}, \mathbb{P}_Y)$ might be easier to deal with than the probability space $(\Omega, \mathfrak{A}, \mathbb{P})$, but maybe still not convenient enough for our purposes. So, we define an *additional real* random variable h which is $\mathfrak{C} - \mathfrak{B}$ measureable (with $\mathfrak{B}$ denoting the Borel σ-algebra for $\mathbb{R}$) such that $h : \Xi \rightarrow \mathbb{R}$. The composion $Z = h \circ Y : \Omega \rightarrow \mathbb{R}$ is again a random variable and $\mathfrak{A} - \mathfrak{C}$ measureable. Instead of defining the probability measure $\mathbb{P}_Z$ by the preimage $\mathbb{P}(Z^{-1}(\boldsymbol{B}))$ for events $\boldsymbol{B} \in \mathfrak{B}$, we *define* $\mathbb{P}_Z$ a priori, for example by $\mathbb{P}_Z := \mathrm{N}(0, 1)$. We require that $\mathbb{P}_Z = f_Z \odot \mu_L$ is dominated by the Lebesgue measure with log-concave Lebesgue density $f_Z(z)$, *i.e.* $\log(f_Z(\alpha z_1 + (1-\alpha)z_2)) \geq \alpha \log(f_Z(z_1)) + (1-\alpha)\log(f_Z(z_2))$ for all $z_1, z_2 \in \mathbb{R}$ and $\alpha \in (0, 1)$ and $f_Z(z) > 0$ for all $z \in \mathbb{R}$. Furthermore, the

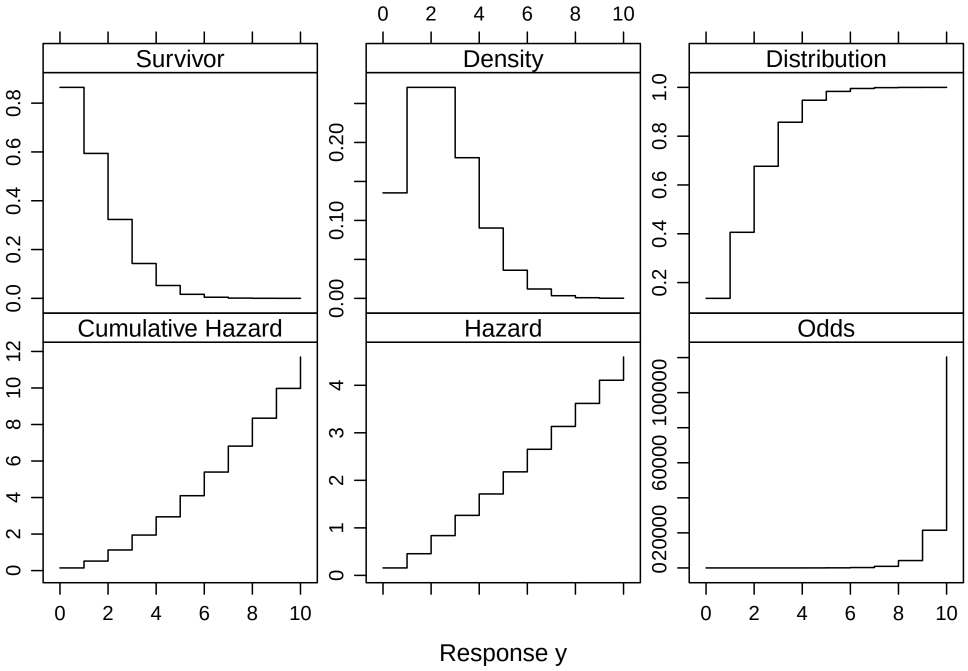


Figure 3.3: Six different characterisation of a distribution $\mathbb{P}_Y$ of a discrete random variable Y from sample space $\Xi = \mathbb{N}$. In contrast to all other functions, the density is only defined for $y \in \mathbb{N}$, therefore the choice of it's graph is not really appropriate here. On the other hand, it closely resembles a histogram.

existence of the first two derivatives of the density $f_Z(z)$ with respect to z is assumed; both derivatives shall be bounded. It is important to note that there are no free unknown parameters in the distribution of F_Z.

In the first step, we will show that there always exists a unique and strictly monotonic transformation g such that the unknown and potentially complex distribution $\mathbb{P}_Y$ that we are interested in can be generated from the simple and known distribution $\mathbb{P}_Z$ via $\mathbb{P}_Y = \mathbb{P}_{g \circ Z}$. More formally, let $g : \mathbb{R} \rightarrow \Xi$ denote a $\mathfrak{B} - \mathfrak{C}$ measureable function. The composition $g \circ Z$ is a random variable on $(\Xi, \mathfrak{C}, \mathbb{P}_{g \circ Z})$. We can now formulate the existence and uniqueness of g as a corollary to the probability integral transform.

Corollary 3.1. *For all random variables Y and Z, there exists a unique*

$$\begin{tikzcd}
(\Omega, \mathfrak{A}, \mathbb{P}) \arrow[r, bend left, "Y"] \arrow[rr, bend left=60, "Z = h \circ Y"] & (\Xi, \mathfrak{C}, \mathbb{P}_Y) \arrow[l, bend left, "Y^{-1}"] \arrow[r, bend left, "h"] & (\mathbb{R}, \mathfrak{B}, \mathbb{P}_Z) \arrow[l, bend left, "g = h^{-1}"] \arrow[ll, bend left=60, "Z^{-1} = Y^{-1} \circ h^{-1}"]
\end{tikzcd}$$

Figure 3.4: Probability spaces and random variables.

strictly monotonically increasing transformation g*, such that* $\mathbb{P}_Y = \mathbb{P}_{g \circ Z}$*.*

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

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

we get the uniqueness of h and therefore g. The quantile function F_Z^{-1} and the distribution function F_Y exist by assumption and are both strictly monotonic and right-continuous. Therefore, h is strictly monotonic and right-continuous and so is g. □

Corollary 3.2. *For* $\mu = \mu_L$*, we have* $g = h^{-1}$ *and* $h'(y) = \frac{\partial h(y)}{\partial y} = f_Z((F_Z^{-1} \circ F_Y)(y))^{-1} f_Y(y)$*.*

This result for absolutely continuous random variables Y can be found in many textbooks (*e.g.* Lindsey, 1996), Corollary 3.1 also covers the discrete case.

Corollary 3.3. *For the counting measure* $\mu = \mu_{\#}$*,* $h = F_Z^{-1} \circ F_Y$ *is a right-continuous step function because* F_Y *is a right-continuous step function with steps at* $y \in \Xi$*.*

Example 3.13. The classical textbook example for transformations of random variables is $Y = Z^2 \sim \chi^2_1$ from $Z \sim \mathrm{N}(0,1)$, *i.e.* using the non-

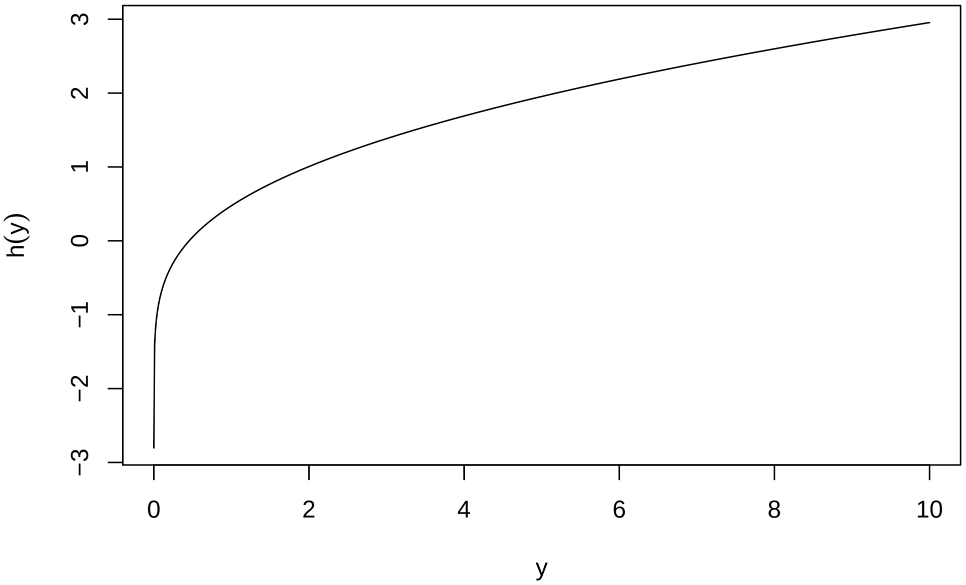


Figure 3.5: χ^2_1 distribution. Transformation function $h = \Phi^{-1} \circ F_{\chi^2_1}$ transformating $Y \sim \chi^2_1$ to a standard normal $Z = h(Y) \sim \mathrm{N}(0, 1)$.

monotone transformation z^2. Alternatively, we can write $Z = h(Y)$ and $Y = g(Z)$ with $h = \Phi^{-1} \circ F_{\chi^2_1}$ and $g = h^{-1} = F^{-1}_{\chi^2_1} \circ \Phi$. The transformations g and h are unique and strictly monotone transformations switching between the standard normal and the χ^2_1 distribution. The χ^2_{df} distribution can be generated from the standard normal by the transformation $g = F^{-1}_{\chi^2_{\text{df}}} \circ \Phi$ and the back-transformation is $h = \Phi^{-1} \circ F_{\chi^2_{\text{df}}}$.

□

Example 3.14. Let $Y \sim \mathrm{Po}(\lambda)$. With $g(z) = \inf\{y \in \Xi \mid F_Y(y \mid \lambda) \geq \Phi(z)\}$ we can transform a standard normal into a Poisson via $Y = g(Z)$. The back-transformation is, however, not possible. □

We now characterise the distribution $\mathbb{P}_Y$ by the corresponding transformation function h. For an a priori choice of F_Z, the distribution function

$F_Y(y) = F_Z(h(y))$ is uniquely defined by the transformation function h. Thus, knowing h automatically provides us with the distribution function F_Y we are interested in. The reasoning behind this setup is also computationally: The distribution function F_Y is a monotonically increasing function in $[0,1]$, whereas the transformation function takes values in $\mathbb{R}$. Estimating h thus requires to pay attention to the monotonicity constraint only.

Let $\mathcal{H} = \{h : \Xi \to \mathbb{R} \mid \mathfrak{C} - \mathfrak{B} \text{ measureable}, h(y_1) < h(y_2) \, \forall y_1 < y_2 \in \Xi\}$ denote the space of all strictly monotonic transformation functions. With the transformation function h, we can evaluate F_Y as $F_Y(y \mid h) = F_Z(h(y)) \, \forall y \in \Xi$. Therefore, we only need to study the transformation function h; the inverse transformation $g = h^{-1}$ (used to define a "group family" or "group model" by Lehmann, 1983; Bickel et al., 1993) is not necessary.

From the distribution function $F_Y(y \mid h) = F_Z(h(y))$ we can now derive the corresponding survivor function $S_Y(y \mid h) = 1 - F_Z(h(y))$. The density for absolutely continuous variables Y $(\mu = \mu_L)$ is now given by

$$f_Y(y \mid h) = f_Z(h(y))h'(y).$$

For discrete responses Y $(\mu = \mu_\#)$ with finite sample space $\Xi = \{\upsilon_1, \dots, \upsilon_K\}$, the density is

$$f_Y(\upsilon_k \mid h) = \begin{cases} F_Z(h(\upsilon_k)) & k = 1 \\ F_Z(h(\upsilon_k)) - F_Z(h(\upsilon_{k-1})) & k = 2, \dots, K-1 \\ 1 - F_Z(h(\upsilon_{k-1})) & k = K \end{cases}$$

and for countably infinite sample spaces $\Xi = \{\upsilon_1, \upsilon_2, \upsilon_3, \dots\}$, we get the density

$$f_Y(\upsilon_k \mid h) = \begin{cases} F_Z(h(\upsilon_k)) & k = 1 \\ F_Z(h(\upsilon_k)) - F_Z(h(\upsilon_{k-1})) & k > 1. \end{cases}$$

With the conventions $F_Z(h(\upsilon_0)) := F_Z(h(-\infty)) := 0$ and $F_Z(h(\upsilon_K)) := F_Z(h(\infty)) := 1$, we use the more compact notation $f_Y(\upsilon_k \mid h) = F_Z(h(\upsilon_k)) - F_Z(h(\upsilon_{k-1}))$ in the sequel.

The odds function $O_Y(y \mid h) = F_Z(h(y))/(1 - F_Z(h(y)))$, hazard function $\lambda_Y(y \mid h) = f_Y(y \mid h)/(1 - F_Z(h(y)))$ and cumulative hazard functions $\Lambda_Y(y \mid h) = -\log(1 - F_Z(h(y)))$ can also be computed directly, but these characterisation of the distribution are more directly related to the transformation function for special choices of F_Z.

Example 3.15. Choosing a minimum extreme value (also called Gompertz) distribution with cumulative distribution function $F_Z(z) = F_{\text{MinEV}}(z) = 1 - \exp(-\exp(z))$ for $z \in \mathbb{R}$ we can write the survivor function as follows

$$S_Y(y \mid h) = \exp(-\exp(h(y)) = \exp(-\Lambda_Y(y \mid h)).$$

Therefore, the transformation function $h(y) = \log(\Lambda_Y(y \mid h))$ is the log-cumulative hazard function. This also means that from a given distribution function F_Y, we can compute to the corresponding transformation function $h(y) = \text{cloglog}(F_Y(y)) = \log(-\log(1-F_Y(y))$ by the complementary log-log link function (the quantile function of the Gompertz distribution). □

Example 3.16. For the maximum extreme value (or Gumbel) distribution with cumulative distribution function $F_Z(z) = F_{\text{MaxEV}}(z) = \exp(-\exp(-z))$ for $z \in \mathbb{R}$ we have $h(y) = \text{loglog}(F_Y(y)) = -\log(-\log(F_Y(y))$, *i.e.* the transformation function is computed by the log-log link function (the quantile function of the Gumbel distribution). □

Example 3.17. The standard logistic distribution with distribution function $F_Z(z) = F_{\text{SL}}(z) = \text{expit}(z) = (1 + \exp(-\boldsymbol{x}))^{-1})$ leads to the following expression for the odds function

$$O_Y(y \mid h) = \frac{F_Z(h(y))}{1 - F_Z(h(y))} = \frac{(1 + \exp(-h(y)))^{-1}}{1 - (1 + \exp(-h(y)))^{-1}} = \exp(h(y))$$

and thus the transformation function $h(y) = \log(O_Y(y \mid h))$ is the log-odds function. Equivalently, we can also write $h(y) = \text{logit}(F_Y(y))$ using the logit link $\text{logit}(p) = \text{expit}^{-1}(p) = \log(p/(1 - p))$, *i.e.* the quantile function of the standard logistic. □

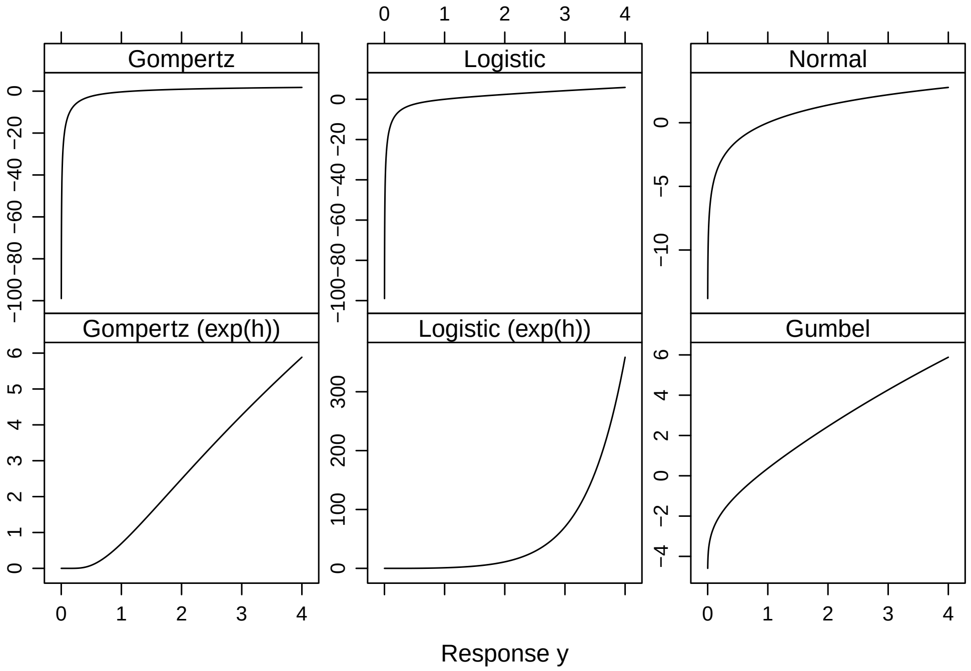


Figure 3.6: Transformation function h for four different choices $Z \sim F_Z$. For the logistic and Gompertz distributions F_Z, $\exp(h(y))$ is plotted in addition.

Example 3.18. Unfortunately, the seemingly most obvious candidate $F_Z = \Phi$ does not lead to a simple interpretation of the transformation function h. We can write $h(y) = \text{probit}(F_Y(y)) = \Phi^{-1}(F_Y(y))$ in terms of the probit link function but this doesn't give us any insights. A more fruitful exercise is to restrict h to a linear function $h(y) = \vartheta_1 + \vartheta_2 y$ of two parameters $\vartheta_1 \in \mathbb{R}$ and $\vartheta_2 \in \mathbb{R}^+$ (to ensure monotonicity). We get

$$F_Y(y \mid h) = F_Z(h(y)) = \Phi(\vartheta_1 + \vartheta_2 y) = \Phi\left(\frac{y - \mu}{\sigma}\right)$$

with parameters $\sigma^{-1} = \vartheta_2$ and $\mu = -\vartheta_1/\vartheta_2$. It should not be too surprising that a linear transformation $Y = g(Z) = h^{-1}(Z) = (Y - \vartheta_1)/\vartheta_2$ of a standard normal Z results in a normal Y. Of course, this nice property vanishes for nonlinear transformations h. □

Further Reading

Chapter 4

Most Likely Transformations

We are now putting together the generic form of the likelihood introduced in Chapter 2 and transformation functions from Chapter 3 to estimate distribution functions via their most likely transformation.

4.1 Transformation Likelihood

For a given transformation function h, the likelihood contribution of a datum $\boldsymbol{C} = (\underline{y}, \bar{y}] \in \mathfrak{C}$ is defined in terms of the distribution function. For the moment, we evaluate the likelihood $l : \mathcal{H} \times \mathfrak{C}$ for a transformation function h and not the corresponding parameters. With (2.1) we have

$$l(h \mid \boldsymbol{C}) := \int_{\boldsymbol{C}} f_Y(y \mid h) d\mu(y) = F_Z(h(\bar{y})) - F_Z(h(\underline{y})).$$

This definition of the likelihood for a transformation function h applies to finite, countably infinite and infinite sample spaces. For an observation $y = \upsilon_k$ from a discrete sample space, the event $\boldsymbol{C}$ is given by $\bar{y} = \upsilon_k$ and $\underline{y} = \upsilon_{k-1}$, such that $l(h \mid \boldsymbol{C}) = f_Y(y \mid h) = F_Z(h(\upsilon_k)) - F_Z(h(y_{k-1})$. Of course, if multiple outcomes $\boldsymbol{C} \subseteq \Xi$ were observed, the likelihood is the sum of the corresponding densities $l(h \mid \boldsymbol{C}) = \sum_{\upsilon \in \boldsymbol{C}} f_Y(\upsilon \mid h)$.

For absolutely continuous responses $Y \in \mathbb{R}$, events $\boldsymbol{C} = (\underline{y}, \bar{y}] \in \mathfrak{C}$ directly lead to the likelihood $l(h \mid \boldsymbol{C}) = F_Z(h(\bar{y})) - F_Z(h(\underline{y}))$. Because the event $\boldsymbol{C}$ is an interval, the term "interval censoring" is often used for this situation. It is often dealt with as a rather special case, but readers of Chapter 2 know that this is, in fact, the most common type of observations in practise.

Definition 4.1 (Interval-censoring)**.** *The event* $\boldsymbol{C} = (\underline{y}, \bar{y}] \in \mathfrak{C}$ *is called an interval-censored observation with likelihood function*

$$l(h \mid \boldsymbol{C}) = F_Z(h(\bar{y})) - F_Z(h(\underline{y})).$$

Right- and left-censoring are special cases of interval-censoring. A right-censored observation provides us with a lower bound $\underline{y}$ of the actual value y which, however, cannot be observed directly. This setup is most common in survival analysis.

Example 4.1. The midwife reported that a women in labor lost between 250 and 300 ml blood. The event $\boldsymbol{C} = (250, 300]$ (the half-open interval is a bit academic here) gives rise to her contribution to the likelihood of

$$l(h \mid \boldsymbol{C}) = F_Z(h(300)) - F_Z(h(250)).$$

□

Definition 4.2 (Right-censoring)**.** *The event* $Y > \underline{y}$ *with* $\boldsymbol{C} = (\underline{y}, \infty) \in \mathfrak{C}$ *is called a right-censored observation with likelihood function*

$$l(h \mid \boldsymbol{C}) = 1 - F_Z(h(\underline{y})).$$

Example 4.2. A patient suffering a potentially terminally disease was under treatment for 132 days. She feels much better and decides to spend the rest of her life on a tropical island and vanishes from our radar. We *know* that she survived for at least 132 days. We don't know, and will probabily never know, if she is still alive or if and when she actually died. So our observation is $\boldsymbol{C} = (132, \infty)$ and her contribution to the likelihood is $\mathbb{P}_Y(Y > 132) = 1 - F_Z(h(132))$. □

Definition 4.3 (Left-censoring)**.** *The event* $Y < \bar{y}$ *with* $\boldsymbol{C} = (-\infty, \bar{y}] \in \mathfrak{C}$ *is called a left-censored observation with likelihood function*

$$l(h \mid \boldsymbol{C}) = F_Z(h(\bar{y})).$$

Example 4.3. In a study the age at onset of drinking alcohol in teenagers is the response of interest. Some study participants will remember the day they were exposed to alcohol exactly (maybe because they felt so sick), but one participant reports that "I don't remember exactly, but it was at or before my 15th birthday". So, we observe $\boldsymbol{C} = (0, 15]$ and the likelihood contribution is $\mathbb{P}_Y(Y \leq 15) = F_Z(h(15))$. □

Requires random censoring.

The likelihood for censored observations assumes random censoring, essentially meaning that the bounds of the interval do not depend on Y directly. We will come back to this issue in Chapter 11.

Using intervals, we implicitly defined the terms interval-, right-, and left-censoring for a real sample space $\Xi = \mathbb{R}$. Of course, they also apply to discrete sample spaces in the very same way. The likelihood is then, as noted above, always obtained by summing-up the density evaluated at each outcome.

In some circumstances, we are able to measure the response rather precisely, so we obtain observations $\boldsymbol{C} = (y - \epsilon, y + \epsilon]$. For narrow intervals with $\epsilon > 0$ rather small, we can *approximate* the likelihood by the density $f_Y(y \mid h)$.

Definition 4.4 (Approximate Likelihood)**.** *For events* $\boldsymbol{C} = (y - \epsilon, y + \epsilon]$, *the likelihood for* $\epsilon > 0$ *small is*

$$l(h \mid \boldsymbol{C}) \approx l(h \mid y) = f_Z(h(y))h'(y).$$

We should note that the notation switched from a likelihood $l(h \mid \boldsymbol{C})$ for events $\boldsymbol{C}$ to a likelihood $l(h \mid y)$ for some real $y \in \Xi \subseteq \mathbb{R}$. The meaning is clear from the context, so we don't introduce a new symbol for this approximate likelihood function. A downside of using this approximation is $l(h \mid y) > 0$ might take very large values, even $l(h \mid y) = \infty$ is possible.

Example 4.4. We measure the daily temperature at 12:00 hours and obtain a reading of 24.68° Celsius using a very expensive and precise thermometer. The likelihood can then be approximated by $f_Z(h(24.68))h'(23.86)$. □

Sometimes observations need to be excluded from the sample, or were not even observed in the first place, on ground of the observation y. This is called truncation and leads to conditional distributions.

Example 4.5. A study compares the distribution of the response $Y =$ age of inhabitants of nursing homes for the elderly to the age of elderly living on their own. The study looks only at people at least 65 years old as an inclusion criterion. Being younger than 65 years is thus an exclusion criterion and the age of those people is not even recorded. It is said that the sample was left-truncated at 65 years. □

Example 4.6. A faculty dean want's to understand the age distribution of students when earning their first Bachelor's degree. She draws a sample of students younger than 60 years, because she is not interested in senior students. The sample was right-truncated at 60 years. □

For interval-truncated observations in $(y_l, y_r] \subset \Xi$ (left- and right-truncation are special cases with $y_r = \infty$ and $y_l = -\infty$, respectively), we are interested in the conditional distribution function

$$F_Y(y \mid Y \in (y_l, y_r]) = \mathbb{P}_Y(y \leq Y \wedge Y \in (y_l, y_r]) = \frac{F_Y(y)}{F_Y(y_r) - F_Y(y_l)}$$

for $y_l < y \leq y_r$. This translates into the transformation world as

$$\begin{aligned} F_Y(y \mid Y \in (y_l, y_r]) &= F_Z(h(y) \mid Y \in (y_l, y_r]) \\ &= \frac{F_Z(h(y))}{F_Z(h(y_r)) - F_Z(h(y_l))} \quad \forall y \in (y_l, y_r] \end{aligned}$$

and thus the likelihood contribution of an observation $\boldsymbol{C} \in \mathfrak{C}$ changes to

$$\frac{l(h \mid \boldsymbol{C})}{F_Z(h(y_r)) - F_Z(h(y_l))} = \frac{l(h \mid \boldsymbol{C})}{l(h \mid (y_l, y_r])} \quad \text{when } y_l < \underline{y} < \bar{y} \leq y_r.$$

It is important to note that the likelihood is always *defined* in terms of a distribution function, and it therefore makes sense to directly model the distribution function of interest. Transformation models directly target the distribution function. The transformation function h uniquely characterises the distribution function F_Y and we can thus formally define a maximum-likelihood estimator $\hat{h}$ for h.

Definition 4.5 (Most likely transformation)**.** *Let $\boldsymbol{C}_1, \dots, \boldsymbol{C}_N$ denote an independent sample of possibly randomly censored or truncated observations from $\mathbb{P}_Y$. The estimator*

$$\hat{h} := \operatorname*{arg\,max}_{h \in \mathcal{H}} \ell(h) = \operatorname*{arg\,max}_{h \in \mathcal{H}} \sum_{i=1}^{N} \log(l(h \mid \boldsymbol{C}_i)) = \operatorname*{arg\,max}_{h \in \mathcal{H}} \sum_{i=1}^{N} \ell_i(h)$$

is called the most likely transformation (MLT).

It is very important to note that the individual contributions ℓ_i to the log-likelihood ℓ can be a mix of log-likelihood terms for interval-, left-, or right-censored observations $\boldsymbol{C}_i$ or even contributions from "exact" observations $y \in \mathbb{R}$, where the likelihood is approximated by the density.

Log-concavity of f_Z ensures concavity of the log-likelihood (except when all observations are right-censored) and thus ensures the existence and uniqueness of $\hat{h}$.

Example 4.7. For an absolutely continuous response Y the likelihood and log-likelihood for h are approximated by the density and log-density evaluated at $y = (\underline{y} + \bar{y})/2$, respectively:

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

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

$$l(h \mid Y \in (\underline{y}, \bar{y}]) \approx f_Z(h(y))h'(y) = \phi((y - \alpha)\sigma^{-1})\sigma^{-1} = \phi_{\alpha,\sigma^2}(y) = f_Y(y \mid \alpha, \sigma^2).$$

In this simple location-scale family, the most likely transformation is characterised by the parameters of the normal distribution of Y. It is important to note that for other choices of F_Z, the most likely transformation is non-linear; however, the distribution function $F_Y = F_Z(h(y))$ is invariant with respect to F_Z because we can always write h as $F_Z^{-1} \circ F_Y$. In other words, with $F_Z \neq \Phi$, we can still model normal responses Y; however, a non-linear transformation function h is required. □

4.2 Most Likely Transformations

Maximum-likelihood estimation of a transformation function h of course requires a proper parameterisation of this function in terms of parameters $\boldsymbol{\vartheta} \in \Theta \subseteq \mathbb{R}^P$.

4.3 Transformation Analysis

We parameterise the transformation function $h(y)$ as a linear function of its basis-transformed argument y using a basis function $\boldsymbol{a} : \Xi \to \mathbb{R}^P$, such that $h(y) = \boldsymbol{a}(y)^\top \boldsymbol{\vartheta}, \boldsymbol{\vartheta} \in \mathbb{R}^P$. The choice of the basis function $\boldsymbol{a}$ is problem specific and will be discussed in depth in Section 4.5 and in later chapters. The likelihood l only requires evaluation of h, and only an approximation thereof using the Lebesgue density of "exact continuous" observations makes the evaluation of the first derivative of $h(y)$ with respect to y necessary. In this case, the derivative with respect to y is given by $h'(y) = \boldsymbol{a}'(y)^\top \boldsymbol{\vartheta}$, and we assume that $\boldsymbol{a}'$ is available.

In the following, we will write $h = \boldsymbol{a}^\top \boldsymbol{\vartheta}$ and $h' = \boldsymbol{a}'^\top \boldsymbol{\vartheta}$ for the transformation function and its first derivative, omitting the argument y, and we assume that both functions are bounded away from $-\infty$ and ∞. For a specific choice of F_Z and $\boldsymbol{a}$, the transformation family of distributions consists of all distributions $\mathbb{P}_Y$ whose distribution function F_Y is given as the composition $F_Z \circ \boldsymbol{a}^\top \boldsymbol{\vartheta}$; this family can be formally defined as follows.

Definition 4.6 (Transformation family)**.** *The distribution family*

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

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

The classical definition of a transformation family relies on the idea of invariant distributions, *i.e.* only the parameters of a distribution are changed by a transformation function but the distribution itself is not changed. The normal family characterised by affine transformations is the most well-known example (*e.g.* Fraser, 1968; Lindsey, 1996). Here, we explicitly allow and encourage transformation functions that change the shape of the distribution. The transformation function $\boldsymbol{a}^\top \boldsymbol{\vartheta}$ is, at least in principle, flexible enough to generate any distribution function $F_Y = F_Z \circ \boldsymbol{a}^\top \boldsymbol{\vartheta}$ from the distribution function F_Z. We borrow the term "error distribution" function for F_Z from Fraser (1968), because Z can be understood as an error term in some of the models discussed in later chapters. The problem of estimating the unknown transformation function h, and thus the unknown distribution function F_Y, reduces to the problem of estimating the parameter vector $\boldsymbol{\vartheta}$ through maximisation of the likelihood function. We assume that the basis function $\boldsymbol{a}$ is such that the parameters $\boldsymbol{\vartheta}$ are identifiable.

Definition 4.7 (Maximum likelihood estimator)**.**

$$\hat{\boldsymbol{\vartheta}} := \underset{\boldsymbol{\vartheta} \in \Theta}{\arg\max} \sum_{i=1}^{N} \log(l(\boldsymbol{a}^\top \boldsymbol{\vartheta} \mid \boldsymbol{C}_i)) = \sum_{i=1}^{N} \ell_i(\boldsymbol{a}^\top \boldsymbol{\vartheta})$$

Based on the maximum likelihood estimator $\hat{\boldsymbol{\vartheta}}$, we define plug-in estimators of the most likely transformation function and the corresponding estimator of our target distribution F_Y as $\hat{h} := \boldsymbol{a}^\top \hat{\boldsymbol{\vartheta}}$ and $\hat{F}_Y := F_Z \circ \hat{h}$. In other words, we can obtain an estimator for the unknown cumulative distribution function $F_Y(y)$ as $\hat{F}_Y(y) = F_Z(\hat{h}(y))$ for all $y \in \Xi$.

Because the problem of estimating an unknown distribution function is now embedded in the maximum likelihood framework, the asymptotic analysis benefits from standard results on the asymptotic behaviour of maximum likelihood estimators. We begin with deriving the score function and Fisher information for the parameters $\boldsymbol{\vartheta}$.

For an observation $\boldsymbol{C}_i = (\underline{y}, \bar{y}]$ (the constant terms $F_Z(\boldsymbol{a}(-\infty)^\top \boldsymbol{\vartheta}) = F_Z(-\infty) = 0$ and $F_Z(\boldsymbol{a}(\infty)^\top \boldsymbol{\vartheta}) = F_Z(\infty) = 1$ vanish), the score contribution $\boldsymbol{s} : \Theta \times \mathfrak{C} \to \mathbb{R}^P$ is

$$\begin{aligned}
\boldsymbol{s}(\boldsymbol{\vartheta} \mid \boldsymbol{C}_i) &= \frac{\partial \ell_i(\boldsymbol{a}^\top \boldsymbol{\vartheta} \mid (\underline{y}_i, \bar{y}_i])}{\partial \boldsymbol{\vartheta}} \\
&= \frac{\partial \log(F_Z(\boldsymbol{a}(\bar{y}_i)^\top \boldsymbol{\vartheta}) - F_Z(\boldsymbol{a}(\underline{y}_i)^\top \boldsymbol{\vartheta}))}{\partial \boldsymbol{\vartheta}} \\
&= \frac{f_Z(\boldsymbol{a}(\bar{y}_i)^\top \boldsymbol{\vartheta})\boldsymbol{a}(\bar{y}_i) - f_Z(\boldsymbol{a}(\underline{y}_i)^\top \boldsymbol{\vartheta})\boldsymbol{a}(\underline{y}_i)}{F_Z(\boldsymbol{a}(\bar{y}_i)^\top \boldsymbol{\vartheta}) - F_Z(\boldsymbol{a}(\underline{y}_i)^\top \boldsymbol{\vartheta})}. \qquad (4.1)
\end{aligned}$$

The contribution $\boldsymbol{F} : \mathbb{R}^P \times \mathfrak{C} \to \mathbb{R}^{P,P}$ to the Fisher information is

$$\begin{aligned}
\boldsymbol{F}(\boldsymbol{\vartheta} \mid \boldsymbol{C}) &= -\frac{\partial^2 \log(F_Z(\boldsymbol{a}(\bar{y}_i)^\top \boldsymbol{\vartheta}) - F_Z(\boldsymbol{a}(\underline{y}_i)^\top \boldsymbol{\vartheta}))}{\partial \boldsymbol{\vartheta} \partial \boldsymbol{\vartheta}^\top} \\
& -\frac{f'_Z(\boldsymbol{a}(\bar{y}_i)^\top \boldsymbol{\vartheta})\boldsymbol{a}(\bar{y}_i)\boldsymbol{a}(\bar{y}_i)^\top - f'_Z(\boldsymbol{a}(\underline{y}_i)^\top \boldsymbol{\vartheta})\boldsymbol{a}(\underline{y}_i)\boldsymbol{a}(\underline{y}_i)^\top}{F_Z(\boldsymbol{a}(\bar{y}_i)^\top \boldsymbol{\vartheta}) - F_Z(\boldsymbol{a}(\underline{y}_i)^\top \boldsymbol{\vartheta})} \\
& +\frac{[f_Z(\boldsymbol{a}(\bar{y}_i)^\top \boldsymbol{\vartheta})\boldsymbol{a}(\bar{y}_i) - f_Z(\boldsymbol{a}(\underline{y}_i)^\top \boldsymbol{\vartheta})\boldsymbol{a}(\underline{y}_i)]}{[F_Z(\boldsymbol{a}(\bar{y}_i)^\top \boldsymbol{\vartheta}) - F_Z(\boldsymbol{a}(\underline{y}_i)^\top \boldsymbol{\vartheta}]^2} \\
& \times[f_Z(\boldsymbol{a}(\bar{y}_i)^\top \boldsymbol{\vartheta})\boldsymbol{a}(\bar{y}_i)^\top - f_Z(\boldsymbol{a}(\underline{y}_i)^\top \boldsymbol{\vartheta})\boldsymbol{a}(\underline{y}_i)^\top]. \qquad (4.2)
\end{aligned}$$

For very small real intervals $\boldsymbol{C}_i = (y_i - \epsilon, y_i + \epsilon]$ the score contribution from an absolutely continuous distribution is approximated by the gradient of the log-density

$$\begin{aligned}
\boldsymbol{s}(\boldsymbol{\vartheta} \mid \boldsymbol{C}_i) \approx \frac{\partial \log(f_Y(y_i \mid \boldsymbol{\vartheta}))}{\partial \boldsymbol{\vartheta}} &= \frac{\partial \log(f_Z(\boldsymbol{a}(y_i)^\top \boldsymbol{\vartheta}))) + \log(\boldsymbol{a}'(y_i)^\top \boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} \\
&= \boldsymbol{a}(y_i)\frac{f'_Z(\boldsymbol{a}(y_i)^\top \boldsymbol{\vartheta})}{f_Z(\boldsymbol{a}(y_i)^\top \boldsymbol{\vartheta})} + \frac{\boldsymbol{a}'(y_i)}{\boldsymbol{a}'(y_i)^\top \boldsymbol{\vartheta}}. \qquad (4.3)
\end{aligned}$$

with corresponding contribution to the Fisher information

$$\begin{aligned}
\boldsymbol{F}(\boldsymbol{\vartheta} \mid \boldsymbol{C}_i) &\approx -\frac{\partial^2 \log(f_Y(y_i \mid \boldsymbol{\vartheta}))}{\partial \boldsymbol{\vartheta} \partial \boldsymbol{\vartheta}^\top} \\
&= -\boldsymbol{a}(y_i)\boldsymbol{a}(y_i)^\top \left\{ \frac{f''_Z(\boldsymbol{a}(y_i)^\top \boldsymbol{\vartheta})}{f_Z(\boldsymbol{a}(y_i)^\top \boldsymbol{\vartheta})} - \left[\frac{f'_Z(\boldsymbol{a}(y_i)^\top \boldsymbol{\vartheta})}{f_Z(\boldsymbol{a}(y_i)^\top \boldsymbol{\vartheta})}\right]^2 \right\} \\
& +\frac{\boldsymbol{a}'(y_i)\boldsymbol{a}'(y_i)^\top}{(\boldsymbol{a}'(y_i)^\top \boldsymbol{\vartheta})^2} \qquad (4.4)
\end{aligned}$$

(NB: the weight to $\boldsymbol{a}(y_i)\boldsymbol{a}(y_i)^\top$ is constant one for $F_Z = \Phi$).

The score function for a sample of N independent observations $\boldsymbol{C}_1, \dots, \boldsymbol{C}_N$ is the sum of the score contributions

$$\boldsymbol{s}(\boldsymbol{\vartheta}) = \boldsymbol{s}(\boldsymbol{\vartheta} \mid \boldsymbol{C}_1, \dots, \boldsymbol{C}_N) = \sum_{i=1}^{N} \boldsymbol{s}(\boldsymbol{\vartheta} \mid \boldsymbol{C}_i).$$

The Fisher information for the whole sample is

$$\boldsymbol{F}(\boldsymbol{\vartheta}) = \boldsymbol{F}(\boldsymbol{\vartheta} \mid \boldsymbol{C}_1, \dots, \boldsymbol{C}_N) = \sum_{i=1}^{N} \boldsymbol{F}(\boldsymbol{\vartheta} \mid \boldsymbol{C}_i).$$

For a sample truncated to the interval $(y_l, y_r]$, the score function is $\boldsymbol{s}(\boldsymbol{\vartheta}) - \boldsymbol{s}(\boldsymbol{\vartheta} \mid (y_l, y_r])$ and the Fisher information is $\boldsymbol{F}(\boldsymbol{\vartheta}) - \boldsymbol{F}(\boldsymbol{\vartheta} \mid (y_l, y_r])$.

We will first discuss the asymptotic properties of the maximum likelihood estimator $\hat{h}$ in the parametric setting with fixed parameters $\boldsymbol{\vartheta}$ in both the discrete and continuous case.

Parametric Inference

Conditions on the densities of the error distribution f_Z and the basis functions $\boldsymbol{a}$ ensuring consistency and asymptotic normality of the sequence of maximum likelihood estimators $\hat{\boldsymbol{\vartheta}}_N$ and an estimator of their asymptotic covariance matrix are given in the following three theorems. Due to the full parameterisation of the model, the proofs are simple standard results for likelihood asymptotics, and a more complex analysis (as required for estimation equations in the presence of a nuisance parameter h_Y, for example in Cheng et al., 1995; Chen et al., 2002) is not necessary. We will restrict ourselves to absolutely continuous or discrete random variables Y, where the likelihood is given in terms of the density $f_Y(y \mid \boldsymbol{\vartheta})$. Furthermore, we will only study the case of a correctly specified transformation $h = \boldsymbol{a}^\top \boldsymbol{\vartheta}$ and refer the reader to Hothorn et al. (2014), where consistency results for arbitrary h are given.

Theorem 4.1. *For $Y_1, \dots, Y_N \overset{iid}{\sim} \mathbb{P}_{Y,\boldsymbol{\vartheta}_0}$ and under the assumptions* (A1) *the parameter space Θ is compact and* (A2) $\mathbb{E}_{\boldsymbol{\vartheta}_0}[\sup_{\boldsymbol{\vartheta} \in \Theta} |\log(f_Y(Y \mid \boldsymbol{\vartheta}))|] < \infty$

where $\boldsymbol{\vartheta}_0$ is well-separated:

$$\sup_{\boldsymbol{\vartheta}; |\boldsymbol{\vartheta}-\boldsymbol{\vartheta}_0| \geq \epsilon} \mathbb{E}_{\boldsymbol{\vartheta}_0}[\log(f_Y(Y \mid \boldsymbol{\vartheta}))] < \mathbb{E}_{\boldsymbol{\vartheta}_0}[\log(f_Y(Y \mid \boldsymbol{\vartheta}_0))],$$

the sequence of estimators $\hat{\boldsymbol{\vartheta}}_N$ converges to $\boldsymbol{\vartheta}_0$ in probability, $\hat{\boldsymbol{\vartheta}}_N \xrightarrow{\mathbb{P}} \boldsymbol{\vartheta}_0$, as $N \to \infty$.

Proof. The log-likelihood is continuous in $\boldsymbol{\vartheta}$, and due to (A2), each log-likelihood contribution is dominated by an integrable function. Thus, the result follows from van der Vaart (1998) (Theorem 5.8 with Example 19.7; see note at bottom of page 46). □

Remark 4.1. *Assumption* (A1) *is made for convenience, and relaxations of such a condition are given in van de Geer (2000) or van der Vaart (1998). The assumptions in* (A2) *are rather weak: the first one holds if the functions* $\boldsymbol{a}$ *are not arbitrarily ill-posed, and the second one holds if the function* $\mathbb{E}_{\boldsymbol{\vartheta}_0}[\log(f_Y(Y \mid \boldsymbol{\vartheta}))]$ *is strictly convex in* $\boldsymbol{\vartheta}$ *(if the assumption would not hold, we would still have convergence to the set argmax*$_{\boldsymbol{\vartheta}}\mathbb{E}_{\boldsymbol{\vartheta}_0}[\log(f_Y(Y \mid \boldsymbol{\vartheta}))]$*).*

Theorem 4.2. *Under the assumptions of Theorem 4.1 and in addition* (A3)

$$\mathbb{E}_{\boldsymbol{\vartheta}_0}\left(\sup_{\boldsymbol{\vartheta}} \left|\left|\frac{\partial \log f_Y(Y \mid \boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}}\right|\right|^2\right) < \infty,$$

(A4) $\mathbb{E}_{\boldsymbol{\vartheta}_0}(\boldsymbol{a}(Y)\boldsymbol{a}(Y)^\top)$ *and (for the absolutely continuous case $\mu = \mu_L$ only)* $\mathbb{E}_{\boldsymbol{\vartheta}_0}(\boldsymbol{a}'(Y)\boldsymbol{a}'(Y)^\top)$ *are nonsingular, and* (A5) $0 < f_Z < \infty$, $\sup|f_Z'| < \infty$ *and* $\sup|f_Z''| < \infty$*, the sequence* $\sqrt{N}(\hat{\boldsymbol{\vartheta}}_N - \boldsymbol{\vartheta}_0)$ *is asymptotically normal with mean zero and covariance matrix*

$$\Sigma_{\boldsymbol{\vartheta}_0} = \left(\mathbb{E}_{\boldsymbol{\vartheta}_0}\left(-\frac{\partial^2 \log f_Y(Y \mid \boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta} \partial \boldsymbol{\vartheta}^\top}\right)\right)^{-1},$$

as $N \to \infty$.

Proof. Because the map $\boldsymbol{\vartheta} \mapsto \sqrt{f_Y(y \mid \boldsymbol{\vartheta})}$ is continuously differentiable in $\boldsymbol{\vartheta}$ for all y in both the discrete and absolutely continuous case and the matrix

$$\mathbb{E}_{\boldsymbol{\vartheta}_0}\left(\left[\frac{\partial \log f_Y(Y \mid \boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}}\right]\left[\frac{\partial \log f_Y(Y \mid \boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}}\right]^\top\right)$$

is continuous in $\boldsymbol{\vartheta}$ as given in (4.3) and (4.1), the transformation family $\mathbb{P}_{Y,\Theta}$ is differentiable in quadratic mean with Lemma 7.6 in van der Vaart (1998). Furthermore, assumptions (A4-5) ensure that the expected Fisher information matrix is nonsingular at $\boldsymbol{\vartheta}_0$. With the consistency and (A3), the result follows from Theorem 5.39 in van der Vaart (1998). □

Remark 4.2. *Assumption* (A4) *is valid for the densities f_Z of the normal, logistic and minimum extreme value distribution. The Fisher information (4.4) and (4.2) evaluated at the maximum likelihood estimator $\hat{\boldsymbol{\vartheta}}_N$ can be used to estimate the covariance matrix $\Sigma_{\boldsymbol{\vartheta}_0}$.*

Theorem 4.3. *Under the assumptions of Theorem 4.2 and assuming $\mathbb{E}_{\boldsymbol{\vartheta}_0}|\boldsymbol{F}(\boldsymbol{\vartheta}_0 \mid Y)| < \infty$, a consistent estimator for $\Sigma_{\boldsymbol{\vartheta}_0}$ is given by*

$$\hat{\Sigma}_{\boldsymbol{\vartheta}_0,N} = \left(N^{-1}\sum_{i=1}^{N}\boldsymbol{F}(\hat{\boldsymbol{\vartheta}}_N \mid Y_i)\right)^{-1}.$$

Proof. With the law of large numbers we have

$$N^{-1}\sum_{i=1}^{N}\boldsymbol{F}(\boldsymbol{\vartheta}_0 \mid Y_i) = N^{-1}\sum_{i=1}^{N} -\frac{\partial^2 \log f_Y(Y_i \mid \boldsymbol{\vartheta})}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^\top} \xrightarrow{\mathbb{P}} \mathbb{E}_{\boldsymbol{\vartheta}_0}\left(-\frac{\partial^2 \log f_Y(Y \mid \boldsymbol{\vartheta})}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^\top}\right) = \Sigma_{\boldsymbol{\vartheta}_0}^{-1}.$$

Because the map $\boldsymbol{\vartheta} \mapsto \boldsymbol{F}(\boldsymbol{\vartheta} \mid y)$ is continuous for all y (as can be seen from (4.4) and (4.2)), the result follows with Theorem 4.1. □

Based on Theorems 4.1-4.3, we can perform standard likelihood inference on the model parameters $\boldsymbol{\vartheta}$. In particular, we can construct confidence intervals and confidence bands for the conditional distribution function from confidence intervals and bands for the linear functions $\boldsymbol{a}^\top\boldsymbol{\vartheta}$. We complete this part by formally defining the class of transformation models.

Definition 4.8 (Transformation model)**.** *The triple $(F_Z, \boldsymbol{a}, \boldsymbol{\vartheta})$ is called transformation model.*

The transformation model $(F_Z, \boldsymbol{a}, \boldsymbol{\vartheta})$ fully defines the distribution of Y via $F_Y = F_Z \circ \boldsymbol{a}^\top\boldsymbol{\vartheta}$ and thus the corresponding likelihood $l(\boldsymbol{a}^\top\boldsymbol{\vartheta} \mid Y \in (\underline{y}, \bar{y}])$. Our definition of transformation models as $(F_Z, \boldsymbol{a}, \boldsymbol{\vartheta})$ is strongly tied to the idea of structural inference (Fraser, 1968) and group families (Lehmann, 1983) or group models (Bickel et al., 1993). Fraser (1968) described a measurement model $\mathbb{P}_Y$ for Y by an error distribution $\mathbb{P}_Z$ and a structural equation $Y = g \circ Z$, where g is a linear function, thereby extending the location-scale family $Y = \alpha + \sigma Z$ introduced by Fisher (1934) and refined by Pitman (1939). Group models consist of distributions generated by possibly non-linear g. The main difference to these classical approaches is that we parameterise h instead of $g = h^{-1}$. By extending the linear transformation functions g dealt with by Fraser (1968) to non-linear transformations, we approximate the potentially non-linear transformation functions $h = g^{-1} = F_Z^{-1} \circ F_Y$ by $\boldsymbol{a}^\top \boldsymbol{\vartheta}$, with subsequent estimation of the parameters $\boldsymbol{\vartheta}$. For given parameters $\boldsymbol{\vartheta}$, a sample from $\mathbb{P}_Y$ can be drawn by the probability integral transform, *i.e.* $Z_1, \dots, Z_N \overset{\text{iid}}{\sim} \mathbb{P}_Z$ is drawn and then $Y_i = \inf\{y \in \Xi \mid \boldsymbol{a}(y)^\top \boldsymbol{\vartheta} \geq Z_i\}$. This generalises the method published by Bender et al. (2005) from the Cox model to all conditional transformation models.

4.4 Smooth Transformation Functions

For absolutely continuous responses Y, the distribution function F_Y is a smooth continuous function. In a transformation model with $F_Y(y) = F_Z(h(y \mid \boldsymbol{\vartheta}))$, the parameterisation $h(y) = \boldsymbol{a}(y)^\top \boldsymbol{\vartheta}$ of the transformation function h should therefore also be smooth in y. Consquently, the estimated distribution function $\hat{F}_Y(y) = F_Z(h(y \mid \hat{\boldsymbol{\vartheta}}))$ is then also a smooth function of y. This property also makes the derivation of densities via $\hat{f}_Y(y) = \hat{F}_Y'(y)$ possible. The same applies to quantile functions, because the task of inverting a smooth function is relatively simple.

This demand for smooth models describing the distribution of absolutely continuous responses is controversial and, in fact, many prominent transformation models are fitted such that the resulting estimated distribution or survivor functions are discrete step-functions. Many smooth parameterisations for special transformation models have been suggested and we will discuss some options now.

In principle, any polynomial or spline basis is a suitable choice for $\boldsymbol{a}$. The simplest choice is a linear transformation $\boldsymbol{a}(y)^\top\boldsymbol{\vartheta} = (1, y)^\top\boldsymbol{\vartheta} = \vartheta_1+\vartheta_2 y$ with $\boldsymbol{\vartheta} = (\vartheta_1, \vartheta_2)^\top$. Because the transformation function must be monotone nondecreasing, the scale parameter ϑ_2 comes with a positivity constraint $\vartheta_2 > 0$. More advanced are log-linear transformation functions parameterised as $\boldsymbol{a}(y)^\top\boldsymbol{\vartheta} = (1, \log(y))^\top\boldsymbol{\vartheta} = \vartheta_1 + \vartheta_2 \log(y)$ for positive random variables $Y > 0$. Again, the scale parameter ϑ_2 must be positive.

Example 4.8. The model $\mathbb{P}(Y \le y) = \Phi(\vartheta_1 + \vartheta_2 y)$ with linear transformation function is equivalent to $Y \sim \mathrm{N}(-\vartheta_1\vartheta_2^{-1}, \vartheta_2^{-2})$, so ϑ_2 is the inverse standard deviation and ϑ_1 the negative standardised mean. □

Example 4.9. The model $\mathbb{P}(Y \le y) = \Phi(\vartheta_1 + \vartheta_2 \log(y))$ with log-linear transformation function is equivalent to a log-normal model which is explicitly defined by log-transformation of a normal response: $\log(Y) \sim \mathrm{N}(-\vartheta_1\vartheta_2^{-1}, \vartheta_2^{-2})$. □

Example 4.10. The model $\mathbb{P}(Y \le y) = 1-\exp(-\exp(\vartheta_1+\vartheta_2 \log(y)))$ with log-linear transformation function is equivalent to an exponential model (with $\vartheta_2 \equiv 1$) because the distribution function simplifies to

$$1 - \exp(-\exp(\vartheta_1)y)) = 1 - \exp(-\lambda y)$$

with rate $\lambda = \exp(\vartheta_1) > 0$.

Under the less strict constraint $\vartheta_2 > 0$, the model is a reparameterisation of the Weibull distribution function. The cumulative hazard function is $\exp(\vartheta_1 + \vartheta_2 \log(y))$ and thus the log-linear transformation function $h(y \mid \boldsymbol{\vartheta}) = \vartheta_1 + \vartheta_2 \log(y)$ has a direct interpretation as log-cumulative hazard function. □

Other popular and simple transformations are the square-root transformation or the arcsin transformation. A common theme here is that the transformation function $h(y \mid \boldsymbol{\vartheta}) = \boldsymbol{a}(y)^\top\boldsymbol{\vartheta}$ is linear in the parameters $\boldsymbol{\vartheta}$

and potential nonlinearity is injected by nonlinear transformations of y in the basis function $\boldsymbol{a}(y)$.

Example 4.11. The Box-Cox power transformation is an example where the parameters of the transformation define a nonlinear function

$$h(y \mid \lambda, \boldsymbol{\vartheta}) = \vartheta_1 + \vartheta_2 \frac{y^\lambda - 1}{\lambda}$$

for $\lambda > 0$ and $h(y \mid \lambda, \boldsymbol{\vartheta}) = \vartheta_1 + \vartheta_2 \log(y)$ for $\lambda = 0$. Technically, this transformation cannot be written in terms of a basis function $\boldsymbol{a}(y)$, because the basis function $\boldsymbol{a}(y \mid \lambda)$ would depend on an unknown paramater λ. □

In a transformation model $\mathbb{P}(Y \leq y) = F_Z(h(y))$ we want to estimate the transformation function h as accurately and as quickly as possible. A parameterisation in terms of basis functions with $h(y \mid \boldsymbol{\vartheta}) = \boldsymbol{a}(y)^\top \boldsymbol{\vartheta}$ is computationally convenient because any updates with respect to the parameters only require the linear function to be evaluated while the basis functions $\boldsymbol{a}(y)$ have to be computed only once. Two issues remain important: The resulting estimate $\boldsymbol{a}(y)^\top \hat{\boldsymbol{\vartheta}}$ must be monotone nondecreasing in y. Second, the derivative of the transformation function should be easy to evaluate in case we want to derive densities and corresponding approximations to the likelihood.

Many spline variants have been proposed, implemented, and evaluated for this task in different model domains. We shall concentrate on polynomials in Bernstein form, mostly because of computational convenience.

Polynomials in Bernstein form are always defined on an a priori given interval $[\underline{\upsilon}, \bar{\upsilon}] \in \Xi$. For bounded random variables (such as for visual analog scales, Example 2.6), this interval is defined from the experiment. For positive or unbounded variables, one or both bounds have to be specified. The basis functions of order M (with $P = M + 1$ parameters and thus $\boldsymbol{a}_{\text{Bs},M}(y) \in \mathbb{R}^P$) are

$$\boldsymbol{a}_{\text{Bs},M}(y) \quad = \quad \frac{(f_{\text{Be}(1,M+1)}(\tilde{y}), \dots, f_{\text{Be}(m,M-m+1)}(\tilde{y}), \dots, f_{\text{Be}(M+1,1)}(\tilde{y}))^\top}{M+1}$$

where $\tilde{y} = (y - \underline{\upsilon})/(\bar{\upsilon} - \underline{\upsilon}) \in [0, 1]$ is the argument y scaled to the unit interval and $f_{\text{Be}(m,M)}$ is the density of the β-distribution on $[0, 1]$ with parameters m and M.

The parameterisation of the transformation function h in Bernstein form is then

$$h(y \mid \boldsymbol{\vartheta}) \quad = \quad \boldsymbol{a}_{\text{Bs},M}(y)^\top \boldsymbol{\vartheta} = \sum_{m=0}^{M} \vartheta_{m+1} \frac{f_{\text{Be}(m+1,M-m+1)}(\tilde{y})}{M+1}.$$

This choice is computationally attractive because strict monotonicity can be formulated as a set of M linear constraints on the parameters $\vartheta_p < \vartheta_{p+1}$ for all $p = 1, \dots, P-1$. The basis contains an intercept.

The derivative of this transformation function is available in closed form

$$h'(y \mid \boldsymbol{\vartheta}) \quad = \quad \boldsymbol{a}'_{\text{Bs},M}(y)^\top \boldsymbol{\vartheta} = \sum_{m=0}^{M-1} (\vartheta_{m+2} - \vartheta_{m+1}) \frac{f_{\text{Be}(m+1,M-m)}(\tilde{y}) M}{(M+1)(\bar{\upsilon} - \underline{\upsilon})}.$$

In fact, this is a polynomial in Bernstein form of order $M-1$.

For unbounded responses, it might become necessary to evaluate the transformation function outside the interval $[\underline{\upsilon}, \bar{\upsilon}]$. Linear extrapolation can be used, with a straight line of slope $h'(\underline{\upsilon} \mid \boldsymbol{\vartheta})$ going through $(\underline{\upsilon}, h(\underline{\upsilon} \mid \boldsymbol{\vartheta}))$ when $y < \underline{\upsilon}$. It is sometimes helpful to constrain the second derivative at the boundary $h''(\underline{\upsilon} \mid \boldsymbol{\vartheta})$, which again is a polynomial in Bernstein form of order $M-2$, to zero. Extrapolating beyond $\bar{\upsilon}$ works in the very same way.

It is worth noting that the coefficients $\boldsymbol{\vartheta} \in \mathbb{R}^P$ of polynomials in Bernstein form have a direct interpretation in terms of function approximation. Assume that for some function $h : [\underline{\upsilon}, \bar{\upsilon}] \to \mathbb{R}$ we are seeking an approximating polynomial $h(y \mid \boldsymbol{\vartheta})$ of order M such that

$$|h(y) - h(y \mid \boldsymbol{\vartheta})| \leq \varepsilon \quad \forall y \in [\underline{\upsilon}, \bar{\upsilon}]$$

for arbitrary $\varepsilon > 0$. An polynomial $h(y \mid \boldsymbol{\vartheta})$ fullfilling this condition is said to uniformly approximate h. The existance of such polynomials for any function h is stated in Weierstrass' approximation theorem. Sergei Natanovich Bernstein's proof of this theorem is constructive and states that a Bernstein polynomial with parameters

$$\vartheta_p = h\left(\underline{\upsilon} + \frac{(\bar{\upsilon} - \underline{\upsilon})(p-1)}{M}\right), \quad p = 1, \dots, P = M+1$$

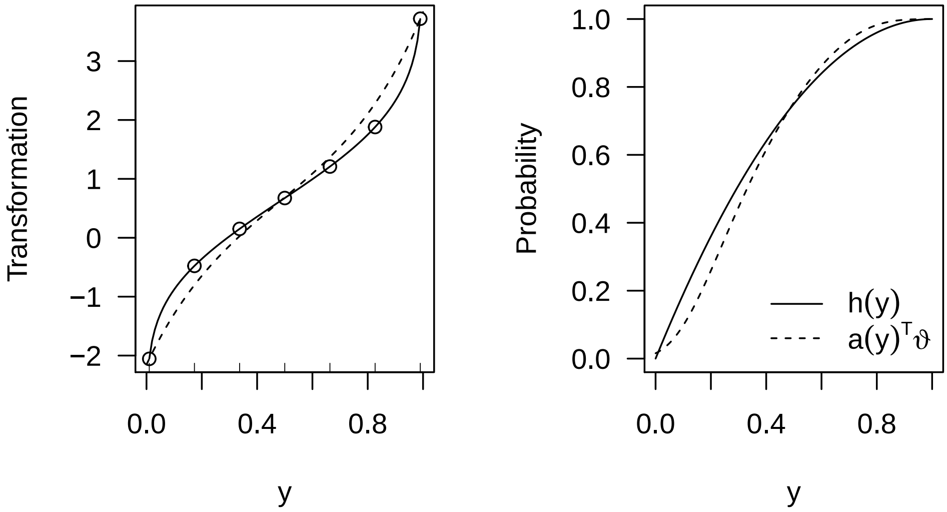


Figure 4.1: Bernstein polynomial approximation of the transformation function of a $\mathrm{B}(1,2)$ distribution. Rugs indicate knots and open dots the true transformation function evaluated at the knots.

and sufficiently high order M exists which uniformly approximates any function h. The arguments of h are the $M+1$ knots of a Bernstein polynomial placed equidistantly on the interval $[\underline{\upsilon}, \bar{\upsilon}]$. The two most extreme knots are $\underline{\upsilon}$ and $\bar{\upsilon}$ and $h(\underline{\upsilon} \mid \boldsymbol{\vartheta}) = \vartheta_1$ and $h(\bar{\upsilon} \mid \boldsymbol{\vartheta}) = \vartheta_P$. Bernstein polynomials of order $M = 1$ are thus linear functions and thus constants and linear functions h can be exactly recovered by Bernstein polynomials.

Example 4.12. Let $Y \sim \mathrm{B}(1,2)$ denote a β-distributed response on $[0,1]$ with shape parameters 1 and 2. The distribution function is given by the incomplete regularised beta function $\mathbb{P}(Y \leq y) = F_Y(y) = I_y(1,2)$. The true transformation function in the transformation model $F_Y(y) = \Phi(h(y))$ is therefore $h(y) = \Phi^{-1}(I_y(1,2))$.

For $M = 6$, say, the Bernstein polynomial $h(y \mid \boldsymbol{\vartheta})$ approximating the "true" transformation function $h(y)$ is defined by the $M+1$ coefficients

$$\vartheta_p = \Phi^{-1}(I_{(p-1)/M}(1,2)), \quad p = 1, \dots, 7$$

evaluated on the knots $(p-1)/M$. The corresponding approximations on the scale of the transformation function $h(y \mid \boldsymbol{\vartheta})$ and distribution function $\Phi(h(y \mid \boldsymbol{\vartheta}))$ are given in Figure 4.1. Increasing M to, say, 15 or 20 considerably improves the approximation. □

4.5 Applications

We are now going to fit a series of different models to the unconditional distribution of postpartum blood loss introduced in Chapter 1. Different forms of the likelihood discussed in Chapter 2 and different formulations of transformation models highlighted in this chapter will be used.

Example 4.13. Suppose we are only interested in the probabilities to loose at most 500, 750, 1000 ml of blood and thus we only need to estimate three parameters, one for each cut-off point. This is the situation discussed in Example 3.7. We choose a transformation model

$$F_Y(y \mid \boldsymbol{\vartheta}) = \Phi(h(y \mid \boldsymbol{\vartheta})) = \begin{cases} 0 & y \le 0 \\ \Phi(\vartheta_1) & y \le 500 \\ \Phi(\vartheta_2) & y \le 750 \\ \Phi(\vartheta_3) & y \le 1000 \\ 1 & y > 1000. \end{cases}$$

with three parameters $\vartheta_1 < \vartheta_2 < \vartheta_3$ and $\boldsymbol{\vartheta} = (\vartheta_1, \vartheta_2, \vartheta_3)^\top$. Although numeric values of blood loss were recorded, the corresponding likelihood is the same as the likelihood for an ordered categorical outcome. Let's say a mother lost 250 ml blood in the first 24 hours after giving birth to her child. The log-likelihood contribution is then

$$\log(F_Y(250 \mid \boldsymbol{\vartheta}) - F_Y(0 \mid \boldsymbol{\vartheta})) = \log(\Phi(\vartheta_1)).$$

A mother who lost more than 1000 ml contributes

$$\log(F_Y(\infty \mid \boldsymbol{\vartheta}) - F_Y(1000 \mid \boldsymbol{\vartheta})) = \log(1 - \Phi(\vartheta_3))$$

to the log-likelihood and someone with 550 ml, say, contributes

$$\log(F_Y(750 \mid \boldsymbol{\vartheta}) - F_Y(500 \mid \boldsymbol{\vartheta})) = \log(\Phi(\vartheta_2) - \Phi(\vartheta_1)).$$

The exact recorded values of blood loss do not matter, only their placing between the different cut-off points is relevant for defining the likelihood.

From optimising this log-likelihood $\ell(\hat{\boldsymbol{\vartheta}}) = -1182.573$ we obtain the parameter estimates $\hat{\boldsymbol{\vartheta}}$ and their covariance matrix $\hat{\Sigma}$, the inverse of the Fisher information matrix, as $\boldsymbol{F}(\hat{\boldsymbol{\vartheta}})$

$$\begin{aligned}\hat{\boldsymbol{\vartheta}} &= (0.535, 1.054, 1.695)^\top \\ \hat{\Sigma} &= \begin{pmatrix} 0.001 & 0.001 & 0.001 \\ 0.001 & 0.002 & 0.001 \\ 0.001 & 0.001 & 0.004 \end{pmatrix}\end{aligned}$$

The estimated transformation function $h(y \mid \hat{\boldsymbol{\vartheta}}) = \boldsymbol{a}_{\text{Bs},6}(y)^\top \hat{\boldsymbol{\vartheta}}$ and distribution function $\Phi(h(y \mid \hat{\boldsymbol{\vartheta}}))$ are shown in Figure 4.2. The covariance matrix $\hat{\Sigma}$ describes the uncertainty around the estimated parameters $\hat{\boldsymbol{\vartheta}}$ and the limiting multivariate normal distribution $\mathrm{N}(\hat{\boldsymbol{\vartheta}}, \hat{\Sigma})$ can be used to obtain simultaneous confidence intervals for the three parameters, plotted as grey lines in Figure 4.2.

□

In this example, we basically treated the measured blood loss as if it were an ordered categorical variable. The results would have been the same if the data were recorded as blood loss at most 500 ml, between 500 and 750 ml, between 750 and 1000 ml or more than 1000 ml.

In the next example, we actually make use of the numerical values. The core idea of the previous example, that is, a categorisation of the response, is applied in the very same way, with just much more cut-off points.

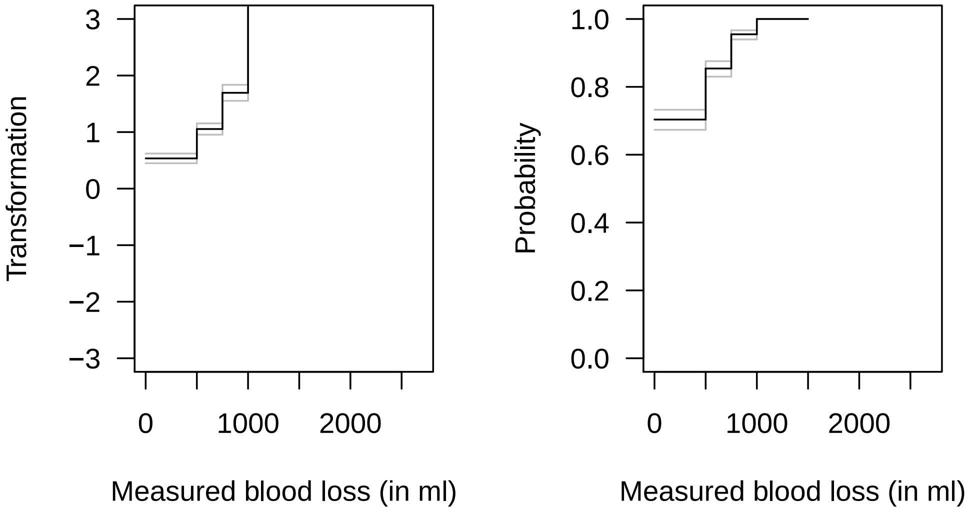


Figure 4.2: Measured blood loss. Probability of loosing less than 500, 750, and 1000 ml blood. The transformation (left) and distribution (right) functions are plotted with confidence bands.

Example 4.14. We extend the possible cut-off points in this model to all observed values of measured blood loss as in Example 3.8 and employ the corresponding nonparametric log-likelihood for parameter estimation. In a sense, we perform the (in this simple case unnecessarily complex) task of estimating the empirical cumulative distribution function on the probit scale.

From optimising the nonparametric log-likelihood to $\ell(\hat{\boldsymbol{\vartheta}}) = -3251.192$

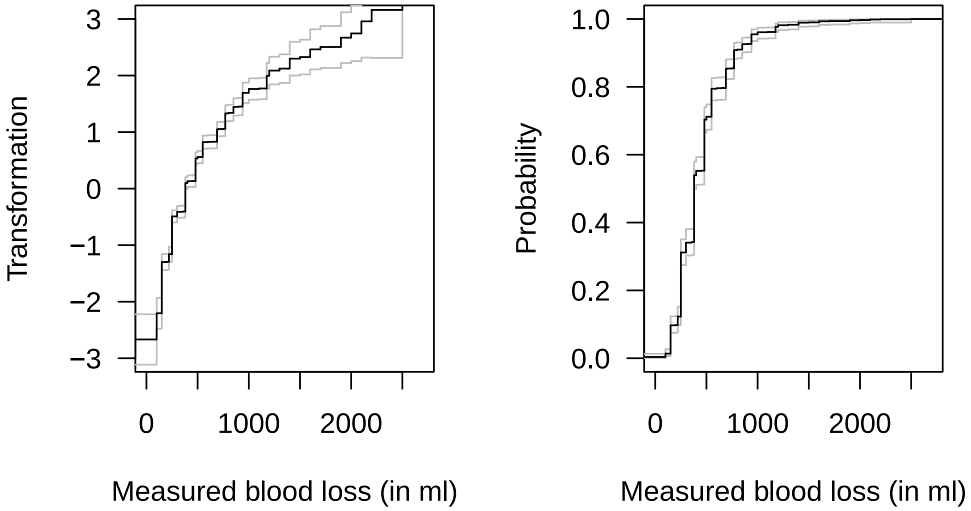


Figure 4.3: Measured blood loss. Empirical cumulative distribution function (right) estimated on the probit scale (left), with 95% confidence bands.

we obtain the 38 parameter estimates

$$\begin{aligned}
\hat{\boldsymbol{\vartheta}} &= (-2.668, -2.204, -1.299, -1.294, -1.160, -0.491, \dots)^\top \\
\hat{\Sigma} &= \begin{pmatrix}
0.022 & 0.007 & 0.001 & 0.001 & 0.001 & 0.000 & \dots \\
0.007 & 0.008 & 0.002 & 0.002 & 0.001 & 0.001 & \\
0.001 & 0.002 & 0.002 & 0.002 & 0.002 & 0.001 & \\
0.001 & 0.002 & 0.002 & 0.002 & 0.002 & 0.001 & \\
0.001 & 0.001 & 0.002 & 0.002 & 0.002 & 0.001 & \\
0.000 & 0.001 & 0.001 & 0.001 & 0.001 & 0.001 & \\
\vdots & & & & & & \ddots
\end{pmatrix}
\end{aligned}$$

The estimates $\hat{\boldsymbol{\vartheta}}$ are plotted as step-function in the left panel of Figure 4.3, along with a confidence band derived from the limiting $\mathrm{N}(\hat{\boldsymbol{\vartheta}}, \hat{\Sigma})$ distribution.

In the next step, we change the model but continue to apply the nonparametric log-likelihood. A smooth transformation model $\mathbb{P}(Y \le y) =$

$\Phi(\boldsymbol{a}_{\text{Bs},6}(y)^\top \boldsymbol{\vartheta})$ with a transformation function parameterised in Bernstein form of order $M = 6$ is fitted by maximising the nonparametric log-likelihood. This likelihood is equivalent to the likelihood for interval-censored data, where the interval limits are given by the unique values of measured blood loss $(\upsilon_{k-1}, \upsilon_k]$ (Chapter 2 and Example 3.8).

Before fitting the model, some details regarding the basis functions defining the polynomial in Bernstein form need to be discussed. Measured blood loss is positive, whoever, childbirth without at least some minor blood loss is unrealistic and, in fact, the smallest reported value was 100 ml. Thus, we use $\underline{\upsilon} = 100$ as the lower bound of the interval defining the Bernstein bases. For the upper bound we note that 1308 our of 1309 mothers suffered blood loss of at most 2500 ml and in only one case an extreme 5700 ml was lost. We are not interested in modelling extreme tail probabilities and thus use $\bar{\upsilon} = 2500$ as upper bound, so essentially the range of the response was used to define this interval. Linear extrapolation is employed when evaluating the likelihood of the largest observation. The basis functions $\boldsymbol{a}_{\text{Bs},6}(y)$ are visualised in the left panel of Figure 4.4. The transformation function $h(y \mid \boldsymbol{\vartheta}) = \boldsymbol{a}_{\text{Bs},6}(y)^\top \boldsymbol{\vartheta}$ is the sum of these basis functions scaled by the corresponding parameter ϑ_p (right panel of Figure 4.4).

From optimising th nonparametric log-likelihood to $\ell(\hat{\boldsymbol{\vartheta}}) = -4372.005$ we obtain the $M + 1 = 7$ parameter estimates

$$\begin{aligned}
\hat{\boldsymbol{\vartheta}} &= (-2.360, 1.740, 1.740, 1.740, 2.726, 2.726, 3.192)^\top \\
\hat{\Sigma} &= \begin{pmatrix}
0.008 & -0.018 & 0.032 & -0.037 & 0.026 & -0.009 & 0.002 \\
-0.018 & 0.074 & -0.160 & 0.206 & -0.154 & 0.055 & -0.011 \\
0.032 & -0.160 & 0.423 & -0.609 & 0.496 & -0.188 & 0.041 \\
-0.037 & 0.206 & -0.609 & 1.029 & -0.925 & 0.394 & -0.082 \\
0.026 & -0.154 & 0.496 & -0.925 & 1.002 & -0.499 & 0.113 \\
-0.009 & 0.055 & -0.188 & 0.394 & -0.499 & 0.399 & -0.070 \\
0.002 & -0.011 & 0.041 & -0.082 & 0.113 & -0.070 & 0.097
\end{pmatrix}
\end{aligned}$$

Confidence bands can be derived from contrasts of the limiting normal distribution $\text{N}(\hat{\boldsymbol{\vartheta}}, \hat{\Sigma})$. For a grid of blood loss values $y_1 < \dots, y_K$ the Bernstein

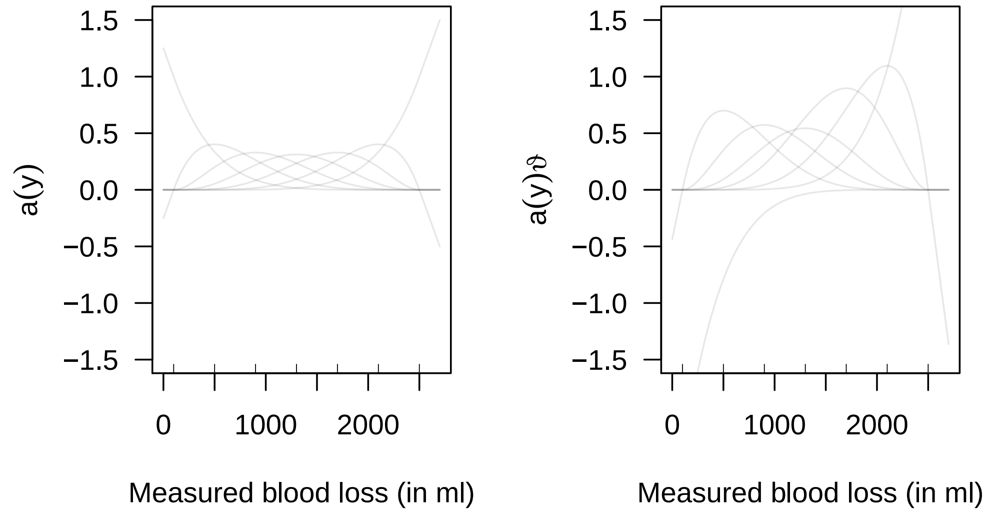


Figure 4.4: Measured blood loss. Bernstein bases (left) on the interval $[100, 2500]$ and Bernstein bases scaled by the corresponding parameter (right). Rugs indicate knots.

basis functions are evaluated and stored in an $K \times P$ matrix

$$\boldsymbol{A} = \left(\boldsymbol{a}(y_1)^\top, \dots, \boldsymbol{a}(y_k)^\top\right)^\top$$

and the limiting normal distribution of the contrast $(h(y_1 \mid \hat{\boldsymbol{\vartheta}}), \dots, h(y_K \mid \hat{\boldsymbol{\vartheta}}))^\top = \boldsymbol{A}\hat{\boldsymbol{\vartheta}}$ with covariance $\boldsymbol{A}\hat{\Sigma}\boldsymbol{A}^\top$ defines the confidence band. The uncertainty is very small, as can be infererred from the small bands around the transformation and distribution functions in Figure 4.7.

An important question concerns model complexity, that is, the order M of the polynomial in Bernstein form. The model was fitted with orders $M = 1, \dots, 25$, where $M = 1$ corresponds to a normal response. The log-likelihood as a function of M is presented in Figure 4.5. The different transformation and distribution functions obtained from increasing orders M are given in Figure 4.6, where the linear transformation function corresponds to the normal model with $M = 1$. This normal model is clearly

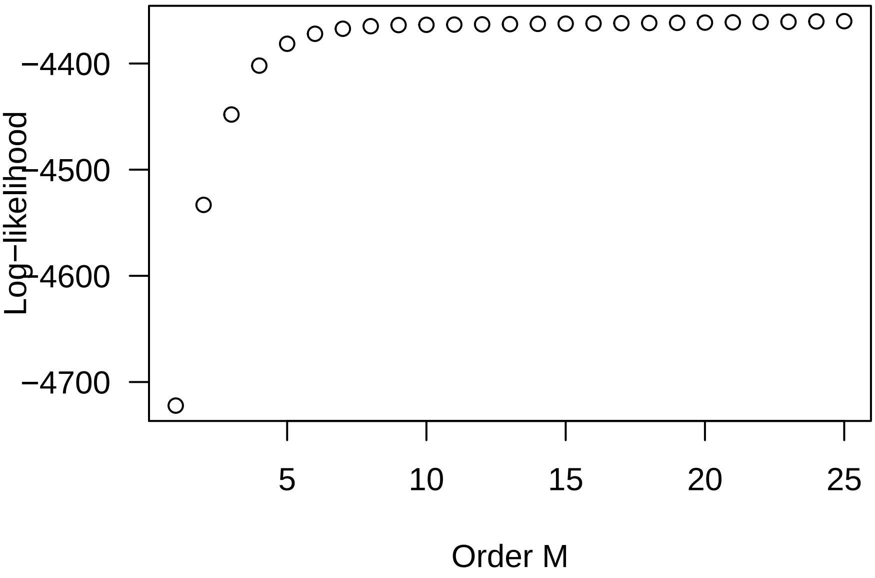


Figure 4.5: Measured blood loss. Log-likelihood as a function of the order M of the polynomial in Bernstein form used to define the smooth transformation model.

not appropriate and the log-Likelihood is much larger for higher orders. It should be noted that orders larger than $M = 6$ do not lead to overfitting, as the log-likelihood remains constant despite the increasing number of parameters. The reason for this quite pleasant behaviour is the monotonicity constraint on the transformation function h and thus the parameters $\boldsymbol{\vartheta}$: Thanks to this constraint, the transformation function cannot become overly erratic. We use $M = 6$ from now on.

One advantage of the smooth over the discrete model is that we can easily invert the estimated transformation function and corresponding confidence band to obtain an median estimate. Because for the median blood loss $y_{0.5}$ we have $h(y_{0.5} \mid \boldsymbol{\vartheta}) = 0$, we can find the root of $h(y_{0.5} \mid \hat{\boldsymbol{\vartheta}})$ numerically to obtain an estimate of the median blood loss, resulting in 418.7 with

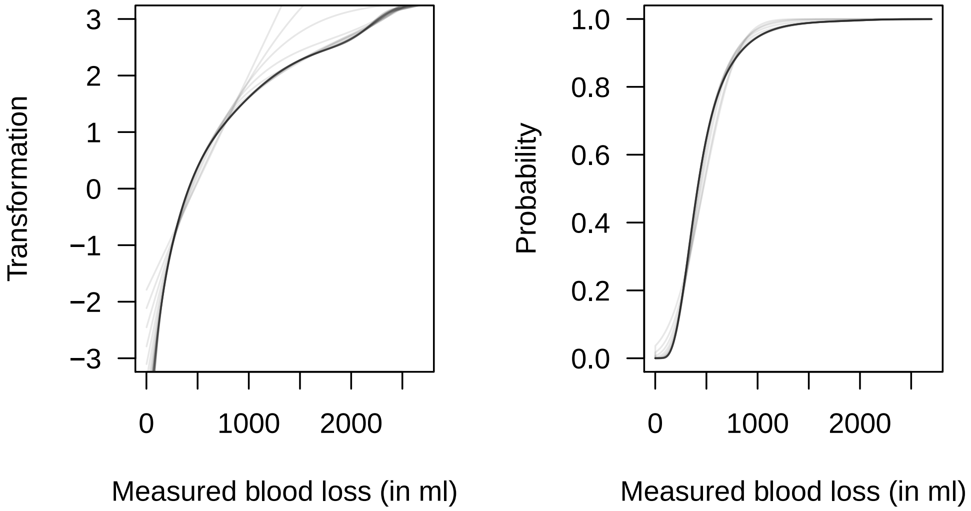


Figure 4.6: Measured blood loss. Transformation (left) and distribution functions (right) for a smooth transformation model defined by polynomials in Bernstein form of increasing orders M.

confidence interval $(402.0, 436.0)$.

□

Both the discrete and smooth transformation models were fitted by the nonparametric log-likelihood defined by all distinct values of measured blood loss. These intervals might be rather long, for example, the interval of the larges observation is $(2500, 5700]$, or 2.2 liters of blood, with the actual observation at the right limit of the interval. In the next example, we try to find more realistic intervals around the observed values and employ the corresponding interval-censored log-likelihood to fit a discrete and a smooth transformation model.

Example 4.15. For a value of, say, 250 ml the interval $(225, 275]$ might represent the rounding error in a more realistic way (see Example 2.17). For extremely large values over 1000 ml, an interval length of 100 and thus more

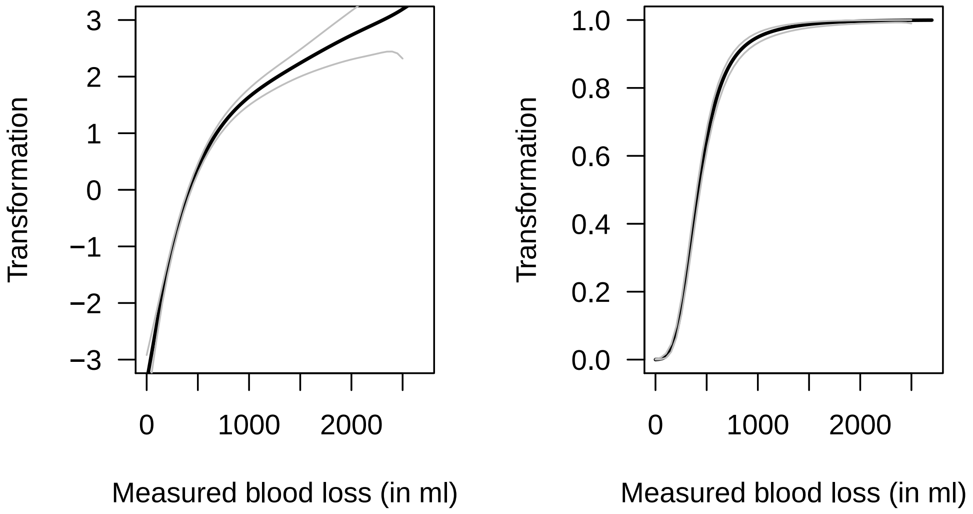


Figure 4.7: Measured blood loss. Transformation function and distribution function estimated smoothly from a transformation model $\mathbb{P}(Y \leq y) = \boldsymbol{a}_{\text{Bs},6}(y)^\top \boldsymbol{\vartheta})$ fitted by optimising the nonparametric log-likelihood. 95% confidence bands are plotted in grey.

uncertainty around the recorded value is chosen. Some values in the data "look" more precise, for example 220, 480, or 690, and we use an interval of length 10 around these values. The log-likelihood is then the log-probability of the intervals $(y_i - 25, y_i + 25]$ for $y_i < 1000$, $(y_i - 50, y_i + 50]$ for $y \geq 1000$, and $(y_i - 5, y_i + 5]$ for values y_i not divisible by 50.

A discrete nonparametric transformation model jumps at each possible interval limit and comprises 46 parameters. The nonparametric maximum likelihood estimator for interval-censored data is known as the Turnbull estimator.

We optimise the interval-censored log-likelihood $\ell(\hat{\boldsymbol{\vartheta}}) = -3211.611$ and obtain the 46 parameter estimates

$$\begin{aligned}
\hat{\boldsymbol{\vartheta}} &= (-2.668, -2.204, -2.204, -1.294, -1.160, -0.491, \dots)^\top \\
\hat{\Sigma} &= \begin{pmatrix}
0.023 & 0.007 & 0.001 & 0.001 & 0.001 & 0.001 & \dots \\
0.007 & 0.008 & 0.001 & 0.002 & 0.001 & 0.001 & \\
0.001 & 0.001 & 0.913 & 0.002 & 0.001 & 0.001 & \\
0.001 & 0.002 & 0.002 & 0.002 & 0.002 & 0.001 & \\
0.001 & 0.001 & 0.001 & 0.002 & 0.002 & 0.001 & \\
0.001 & 0.001 & 0.001 & 0.001 & 0.001 & 0.001 & \\
\vdots & & & & & & \ddots
\end{pmatrix}
\end{aligned}$$

As a smooth alternative to this discrete model, we employ the smooth model $\mathbb{P}(Y \le y) = \Phi(\boldsymbol{a}_{\text{Bs},6}(y)^\top \boldsymbol{\vartheta})$ and optimise the interval-censored log-likelihood $\ell(\hat{\boldsymbol{\vartheta}}) = -3741.217$ to obtain the $M + 1 = 7$ parameter estimates

$$\begin{aligned}
\hat{\boldsymbol{\vartheta}} &= (-2.615, 1.768, 1.768, 1.768, 2.695, 2.752, 3.068)^\top \\
\hat{\Sigma} &= \begin{pmatrix}
0.011 & -0.025 & 0.041 & -0.040 & 0.021 & -0.001 & -0.001 \\
-0.025 & 0.087 & -0.166 & 0.180 & -0.097 & 0.006 & 0.004 \\
0.041 & -0.166 & 0.381 & -0.459 & 0.268 & -0.017 & -0.011 \\
-0.040 & 0.180 & -0.459 & 0.656 & -0.423 & 0.037 & 0.025 \\
0.021 & -0.097 & 0.268 & -0.423 & 0.351 & -0.039 & -0.025 \\
-0.001 & 0.006 & -0.017 & 0.037 & -0.039 & 0.052 & 0.040 \\
-0.001 & 0.004 & -0.011 & 0.025 & -0.025 & 0.040 & 0.039
\end{pmatrix}
\end{aligned}$$

The discrete and smooth models are compared on the scale of the transformation and distribution function in Figure 4.8. The smooth model nicely interpolates the discrete transformation and distribution functions.

□

For the sake of completeness, we fit a smooth transformation model by optimising the log-density, *i.e.* by replacing the nonparametric or interval-censored log-likelihood by this approximation.

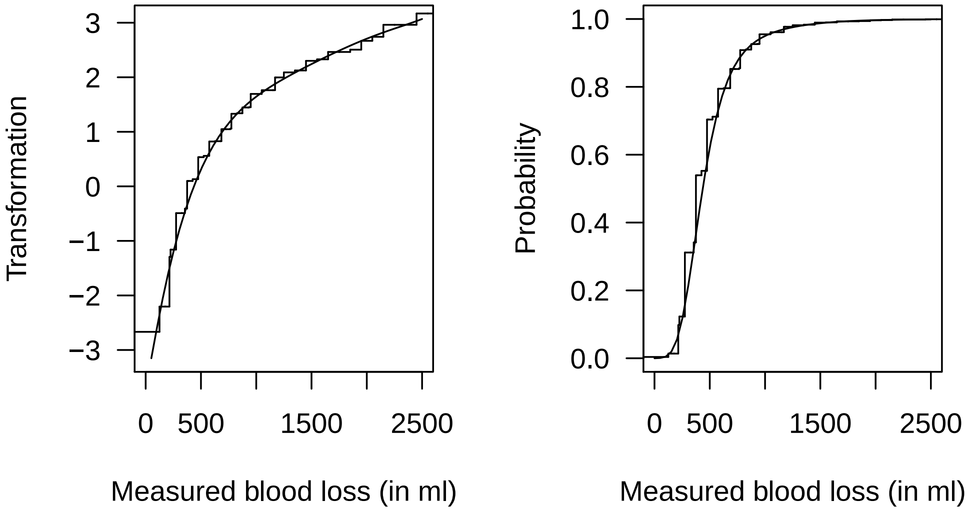


Figure 4.8: Measured blood loss. Discrete and smooth transformation models estimated by maximisation of the interval-censored log-likelihood.

Example 4.16. From optimising the log-density $\ell(\hat{\boldsymbol{\vartheta}}) = -8906.534$ we obtain the $M+1=7$ parameter estimates

$$\begin{aligned}
\hat{\boldsymbol{\vartheta}} &= (-2.612, 1.769, 1.769, 1.769, 2.482, 2.900, 2.969)^\top \\
\hat{\Sigma} &= \begin{pmatrix}
0.010 & -0.024 & 0.038 & -0.037 & 0.019 & -0.000 & -0.000 \\
-0.024 & 0.082 & -0.156 & 0.168 & -0.089 & 0.003 & 0.002 \\
0.038 & -0.156 & 0.358 & -0.428 & 0.248 & -0.008 & -0.007 \\
-0.037 & 0.168 & -0.428 & 0.609 & -0.388 & 0.021 & 0.019 \\
0.019 & -0.089 & 0.248 & -0.388 & 0.318 & -0.024 & -0.022 \\
-0.000 & 0.003 & -0.008 & 0.021 & -0.024 & 0.042 & 0.040 \\
-0.000 & 0.002 & -0.007 & 0.019 & -0.022 & 0.040 & 0.039
\end{pmatrix}
\end{aligned}$$

for the smooth transformation model with transformation function $\boldsymbol{a}_{\text{Bs},6}(y)^\top\boldsymbol{\vartheta}$.

□

One may ask what we gained from looking at the plethora of models and estimation techniques for the distribution of postpartal blood loss. The

	$h(500 \mid \hat{\boldsymbol{\vartheta}})$	$h(750 \mid \hat{\boldsymbol{\vartheta}})$	$h(1000 \mid \hat{\boldsymbol{\vartheta}})$
4.13 discrete	0.5348 ± 0.0365	1.0541 ± 0.0426	1.6946 ± 0.0604
4.14 discrete	0.5346 ± 0.0365	1.0539 ± 0.0426	1.6940 ± 0.0604
4.14 smooth	0.3756 ± 0.0314	1.1730 ± 0.0404	1.6414 ± 0.0522
4.15 smooth	0.3079 ± 0.0311	1.1554 ± 0.0403	1.6448 ± 0.0517
4.16 smooth	0.3086 ± 0.0310	1.1490 ± 0.0400	1.6255 ± 0.0511

Table 4.1: Measured blood loss. Estimated transformation functions $h(y \mid \hat{\boldsymbol{\vartheta}})$ for quantiles $y = 500, 750, 1000$ and corresponding standard errors for several discrete and smooth transformation models.

most important point is the understanding that model parameterisation and the specific form of the likelihood are not necessarily in a one-to-one relationship. At least three different forms of the likelihood (nonparametric, interval-censored, and log-density) were used to estimate the model parameters of a smoothly parameterised transformation model.

The smooth model fits closely approximate the discrete model fits, but the differences can still be substancially. For three selected quantiles $y = 500, 750, 1000$, Table 4.1 presents the estimated transformation functions $h(y \mid \hat{\boldsymbol{\vartheta}})$ for discrete and smooth models estimated by different forms of the likelihood. The values are quite different for $y = 500$ and less diverging for the larger quantiles. It is important to note that the standard deviation of the estimators is largest for the discrete models and smallest for the model fitted by optimising the log-density. This is not very surprising, because the latter likelihood understands the observations as precise measurements whereas the nonparametric and interval-censored log-likelihoods pay attention to the uncertainty reflected in the interval lengths.

The most important aspect of a smooth transformation model is that we can look at the model from many different perspectives. A plot showing different characteristics of the estimated distribution, similar to the one in Figure 3.2, is given in Figure 4.9. Many readers will probably prefer the density function, but also the hazard function λ has a direct interpretation in this application: The "risk" that bleeding will stop at y ml given that y ml were already lost is given by $\lambda(y)$. The hazard is largest at 569 ml and

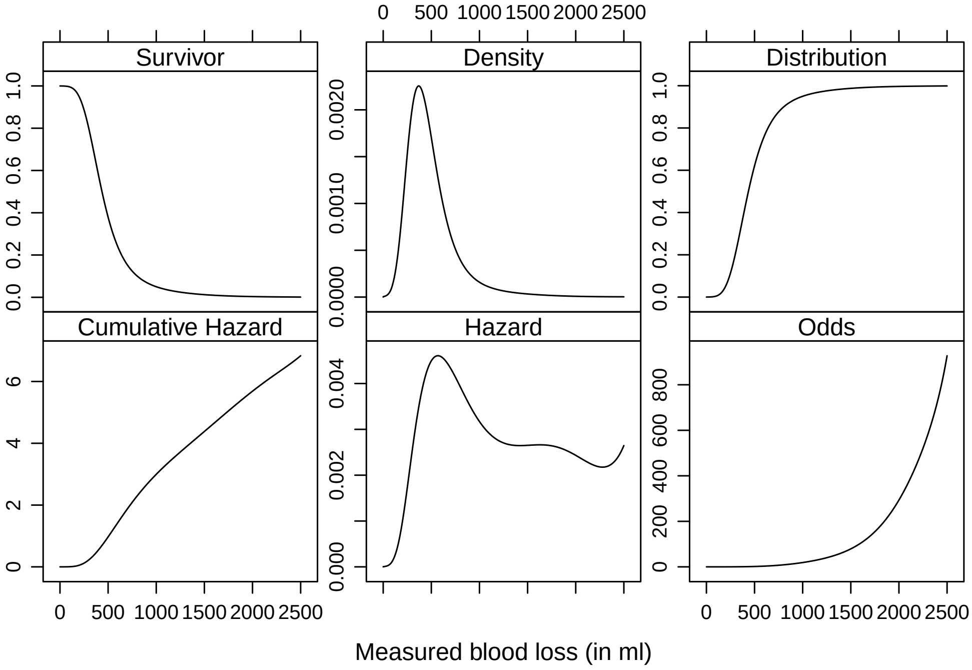


Figure 4.9: Measured blood loss. Transformation model fitted to interval-censored data displayed via six different characterisations of the estimated distribution. The quantile function is the inverse of the distribution function.

declines rapidly, however, a substantial risk remains for extreme values of postpartum blood loss.

Further Reading

(Klein and Moeschberger, 2003)

Farouki (2012)

Chapter 5

Two Sample Comparisons

After dealing with the theoretical foundations, we will now come back to the motivating example introduced in the Introduction: How does the distribution of postpartum blood loss differ between vaginal births and births by Cesarean section? The aim is to compare two conditional distributions, one for the response observed in each of two groups. This seamingly simple situation concerns the simplest regression models and is ideally suited to introduce many important concepts in transformation models and regression in general.

5.1 Nonparametric Comparisons

With $Y \in \Xi$ we refer to an response and with $X \in \{0, 1\}$ to a group indicator, so $X = 0$ denotes the reference group we want to compare the second group with $X = 1$ to. The conditional distributions we strive to compare are $F_0(y) = \mathbb{P}(Y \leq y \mid X = 0)$ and $F_1(y) = \mathbb{P}(Y \leq y \mid X = 1)$. If, in fact, the response Y is independent of group membership X, the two conditional distributions are equal to the marginal distribution function F_Y

$$\mathbb{P}(Y \leq y) = F_Y(y) = F_0(y) = F_1(y), \quad \forall y \in \Xi.$$

If we can find outcomes y for which $F_0(y) \neq F_1(y)$, the two conditional distributions differ somehow, and the changing the group membership thus changes the conditional distribution of the response.

In the language of null hypothesis significance testing (NHST), one relevant null hypothesis H_0 could be equivalence of F_1 and F_2

$$H_0 : F_0(y) = F_1(y), \quad \forall y \in \Xi$$

against the omnibus alternative H_1

$$H_1 : \exists\, y \in \Xi : F_0(y) \neq F_1(y).$$

In this general form, this framework and especially the omnibus alternative H_1 are not very useful. For large enough sample sizes, the null hypothesis can always be rejected in real experiments, because any reasonable division of observational units into two groups will have an effect on the distribution of some reasonable choice of response variable. Such a "significant" result, however, only tells the data analyst that there is some discrepancy. If this finding is relevant and in which respect the two distributions differ remains unclear. The whole point about regression models is to characterise the discrepanies between the two, or in later Chapters many more, conditional distributions. The term *model* highlights the fact that some form of simplification will be used to reduce the overwhelming complexity of the alternative H_1.

Example 5.1. Figure 1.4 shows the two empirical distribution functions $\hat{F}_0$ and $\hat{F}_1$, estimated separately for vaginal ($x = 0$) and Cesarean births ($x = 1$). The Kolmogorov-Smirnov statistic comparing the two distribution functions is

$$\sup_y \left| \hat{F}_0(y) - \hat{F}_1(y) \right| = 0.400$$

Under the conditions of the null hypothesis $H_0 : F_0 \equiv F_1$, observing a Kolmogorov-Simirnov statistic as large or larger than 0.400 is very unlikely (the probability, or "P-value", is $P < 10^{16}$). This significance statement is not very helpful in our context, because even the statistically illiterate would have guessed from Figure 1.4 that something is going on here. After rejecting the null hypothesis H_0, that is, the simplest model assuming equal distributions, the Kolmogorov-Smirnov tests leaves the data analyst in the air on the question of how a better model for our data might actually look like. □

In this context, the term *nonparametric* is commonly used to describe the Kolmogorov-Smirnov or similar procedures. It means that the test is distribution-free and model-free, because we neither assume a certain shape of the distribution function F_0 or F_1 under H_0 nor do we assume anything about the deviation between the two functions under the alternative H_1. Without a model, there is no hope of making progress regarding the identification of appropriate and more specific alternatives. Parametric procedures introduce such models, however at the price of also sacrificing the distribution-freedom as shall be discussed in the next section.

5.2 Parametric Comparisons

In the classical parametric setup, we apply the exact opposite of a distribution- and model-free procedure by considering a certain family of distributions only. On the one hand, this restricts the generality of the procedure and thus, potentially, the validity of our results. On the other hand, we can discuss results in light of a model which simplifies the complexity of the state of affairs under the alternative H_1 substantially.

Maybe the most prominent choice of distributions is a pair of homoscedastic normal distributions $Y \mid X = 0 \sim \text{N}(\mu_0, \sigma^2)$ and $Y \mid X = 1 \sim \text{N}(\mu_2, \sigma^2)$. The null hypothesis of equal distributions simplifies to equality of means $H_0 : \mu_0 = \mu_1$. Because the absolute values of μ_0 and μ_1 are less relevant, it is common to introduce a treatment effect $\beta = \mu_1 - \mu_0$ defined as the difference of means and we write $H_0 : \beta = 0$. For any alternative $H_1 : \beta \neq 0$ we can evaluate the two distribution functions in the shift model

$$\Phi_{\mu_1,\sigma^2}(y) = F_1(y) = F_0(y - \beta) = \Phi_{\mu_0,\sigma^2}(y - \beta).$$

This constitutes a considerable progress: We can express, evaluate, and communicate the discrepancy between the two distribution functions by a scalar treatment effect β with straightforward interpretation. On top of these nice properties, we were also given an optimal procedure for testing H_0 against H_1: the famous two-sample t-test is the test with highest power. There is, however, a price to be paid: The model assumes normal distributions with equal variances in both groups.

In many applications, normality is out of question: Discrete responses can not be normal and thus our model is not applicable. For continuous responses, skewed and thus non-normal distributions are common. Even if we can assume symmetry around μ_0 and μ_1, respectively, the variability often differs between both groups. Unfortunately, the many formal test procedures available for the assessment of model assumptions, such as the Shapiro-Wilk test, are useless in this context: They put the nice and simple model in their null hypothesis. The data analyse thus finds herself in the position not wanting to reject the normal model and thus the null hypothesis. The easiest way to go is to throw away all data, obtain a large P-value due to lack of power, and simply continue with the seemingly appropriate parametric procedure.

The big question arises if we can find a compromise between distribution- and model-free nonparametric procedures and parametric methods resulting in a model-based but still distribution-free approach? This questions is, of course, rhetorical and, as will become clear in the next section, transformation models are part of the answer.

5.3 Distribution-free Comparisons

In the normal model $Y \mid X = 0 \sim \mathrm{N}(\mu_0, \sigma^2)$ and $Y \mid X = 1 \sim \mathrm{N}(\mu_2, \sigma^2)$ we wrote

$$F_1(y) = F_0(y - \beta). \tag{5.1}$$

One straightforward idea would be to drop the normality assumption, allowing an arbitrary distribution function F_0 from which the distribution function F_1 would be generated under the shift model 5.1. For as long as the response $Y \in \mathbb{R}$ is defined on the real line, this approach is reasonable. For positive, bounded, or discrete variables Y, the model is not defined because $y - \beta$ is not necessarily an element of the sample space Ξ.

A workaround this problem could be to evaluate $F_0(y)$, transform the resulting probability to the normal scale $\Phi^{-1}(F_0(y))$, and only then apply the shift term β:

$$F_1(y) = \Phi(\Phi^{-1}(F_0(y)) - \beta).$$

The distribution function F_1 is well-defined for all $y \in \Xi$ and all shift effects $\beta \in \mathbb{R}$. The meaning of β, however, is harder to interpret because, unlike in (5.1), this parameter is no longer defined on the same scale as the response Y. The next two examples discuss this problem.

Example 5.2. With the distribution $Y \mid X = 0 \sim \mathrm{N}(\mu_0, \sigma^2)$ in the first group, the shift model

$$F_1(y) = \Phi\left(\frac{y - \mu_0}{\sigma} - \beta\right)$$

induces the distribution $Y \mid X = 1 \sim \mathrm{N}(\mu_0 + \sigma\beta, \sigma^2)$ in the second group. The difference of means can be expressed as

$$\mathbb{E}(Y \mid X = 1) - \mathbb{E}(y \mid X = 0) = \sigma\beta$$

and therefore the shift parameter β is the standardised difference of means. □

Example 5.3. In a distribution-free setup, we allow arbitrary distribution functions F_0 for the response in the reference group $Y \mid X = 0$. We can transform $Y \mid X = 0$ into a standard normal distribution by the transformation $Z = h(Y) = \Phi^{-1}(F_0(Y))$. In the shift transformation model $F_1(y) = \Phi(h(y) - \beta)$, the shift parameter measures the difference of means on this latent standard normal scale. □

.

There is no reason to strict ourself to a latent standard normal scale and a more general form model for the two-sample situation is

$$F_1(y) = F_Z(F_Z^{-1}(F_0(y)) - \beta) = F_Z(h(y) - \beta) \tag{5.2}$$

The null hypothesis $H_0 : F_0 \equiv F_1$ is now equivalent to $H_0 : \beta = 0$ and under this condition we get

$$F_1(y) = F_Z(F_Z^{-1}(F_0(y))) = F_0(y) = F_Z(h(y)).$$

This is, clearly, equivalent to an unconditional transformation model as introduced in Chapter 4. The alternative

$$H_1 : \beta \neq 0$$

is now understood under the conditional transformation shift model

$$F_1(y) = F_Z(h(y) - \beta).$$

In addition to β, the monotone nondecreasing transformation function $h : \Xi \rightarrow \mathbb{R}$ with $h(y) = F_Z^{-1}(F_0(y))$ is unknown and thus both h and β define the model. The model is distribution-free, because we do not assume anything about F_0 and thus h. At the same time, it is model-based because the alternative is given by the shift-transformation model (5.2). The distribution function F_Z plays an important role, because it defines the scale on which the two distribution functions F_0 and F_1 differ by a constant term β. This choice introduces a model assumption which might or might not be appropriate for a specific experiment.

In the two-sample situation, this assumption is relatively easy to check. The reformulation

$$F_Z^{-1}(F_1(y)) = F_Z^{-1}(F_Z(F_Z^{-1}(F_0(y)) - \beta)) = F_Z^{-1}(F_0(y)) - \beta \quad \forall y \in \Xi$$

highlights that the two distribution functions, after having been transformed by the quantile or link function F_Z^{-1} differ additively by a constant β.

Given two estimators $\hat{F}_0$ and $\hat{F}_1$, for example the two empirical cumulative distribution functions fitted separately to the two groups, we can plot $F_Z^{-1}(\hat{F}_0(y_i))$ and $F_Z^{-1}(\hat{F}_1(y_i))$ against the observations $y_1, \dots, y_N$. If these two nonlinear functions of y_i share the same shape but are vertically shifted, the constant difference

$$\beta \approx F_Z^{-1}(\hat{F}_0(y_i)) - F_Z^{-1}(\hat{F}_1(y_i)), \quad i = 1, \dots, N$$

corresponds to the shift term β. The connection to the Kolmogorov-Smirnov statistic

$$\sup_y \left| \hat{F}_0(y) - \hat{F}_1(y) \right|$$

is striking: Instead of taking vertical differences on the scale of the distribution functions, we take vertical differences on the transformed scale

$$\left| F_Z^{-1}(\hat{F}_0(y)) - F_Z^{-1}(\hat{F}_1(y)) \right|$$

and assume, or validate, that these are constant for all possible outcomes $y \in \Xi$.

The rather straightforward exercise to generalise the normal model underlying the t-test (Section 5.2) to the transformation shift model (5.2) substantially reduced the distributional assumptions while maintaining the simplicity of a simple scale shift treatment effect β describing the discrepancy of the two distributions under the alternative. Unlike the normal model, the transformation model is applicable to arbitrary response variables, including those with discrete and skewed distributions, does not assume symmetry or homoscedasticity, and the remaining assumption of additivity of the distribution functions on the transformed scale is relatively simple to verify.

We will now discuss properties of the shift model (5.2) for $F_Z = \Phi$ and for three alternative choices of F_Z resulting in three different forms of Lehmann alternatives, namely

$$\begin{aligned} F_1(y) &= F_0(y)^{\exp(\beta)} && \text{“reverse time proportional hazards”} \\ F_1(y) &= 1 - (1 - F_0(y))^{\exp(-\beta)} && \text{“proportional hazards”} \\ \frac{F_1(y)}{1 - F_1(y)} &= \exp(-\beta) \frac{F_0(y)}{1 - F_0(y)} && \text{“proportional odds”} \end{aligned}$$

for arbitrary distribution functions F_0 in a distribution-free fashion.

Normal Shift Alternatives

Normal shift alternatives

$$F_1(y) = \Phi(h(y) - \beta)$$

are based on the transformation function $h(y) = \Phi^{-1}(F_0(y))$, see also Example 5.3. One can interpret the shift parameter as a difference of means on a latent standard normal scale. If a certain value of β is small or large can be judged by asking the question: How large is the power (for a given sample size, or the necessary sample size for a given power as on page 12) when comparing the two distributions $\text{N}(0, 1)$ and $\text{N}(\beta, 1)$? For a total sample size of $N = 100$ observations, the power curve for different values of β is plotted, and the least we can say is that absolute values larger .5 are interesting and values larger than one indicate very interesting effect sizes.

Example 5.4. Is this model appropriate for the comparison of postpartum blood loss between vaginal and Cesarean births? In Figure 5.2, the two corresponding empirical cumulative distribution functions are plotted after probit transformation. Because the two empirical distribution functions cross, any shift model is “wrong”, but the problem only occurs for relatively large values. For measured blood loss of up to 750 ml, say, the shape of the two curves is rather close and the vertical difference more or less constant. □

It is natural to ask how the distribution functions F_1 look like under different alternatives.

Example 5.5. We assume $Y \mid X = 0$ to follow a standard normal, a χ^2 distribution and a t distribution, the latter two with five degrees of freedom. The corresponding distribution and density functions under the null ($\beta = 0$) and alternative hypothesis ($\beta \neq 0$) are plotted in Figure 5.3. In the first column, we have a standard normal $F_0 = \Phi$ and, consequently,

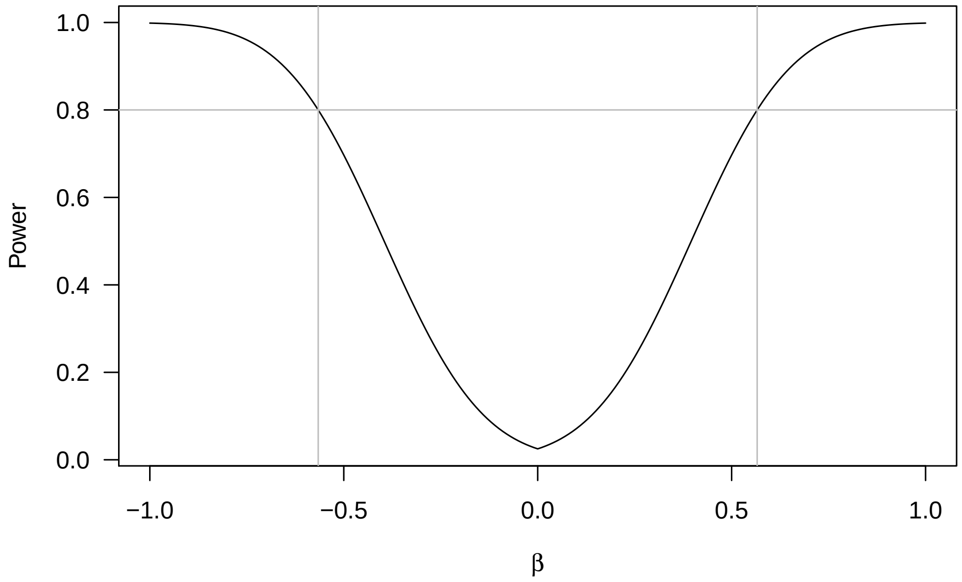


Figure 5.1: Power of a two-sample t-test with $N_0 = N_1 = 50$ observations per group and variance one as a function of the difference of means β.

also the alternative is normal with the same variance but mean β (because the variance is one). The assumptions of the t-test are met in this situation and we see the nicely shifted distributions with identical shape and variance which is, unfortunately, rather unique.

For the skewed χ^2 distribution, the whole shape of the distribution and density function changes under the alternative, in fact, $Y \mid X = 1$ is not χ^2 distributed. With increasing values of β, the location increased but also the variance. The same phenomenon is also visible for symmetric but heavy-tailed t distributions. In these two cases, a t-test is clearly inappropriate and differences of means do not describe the discrepancies between F_0 and F_1 in any meaningful way.

□

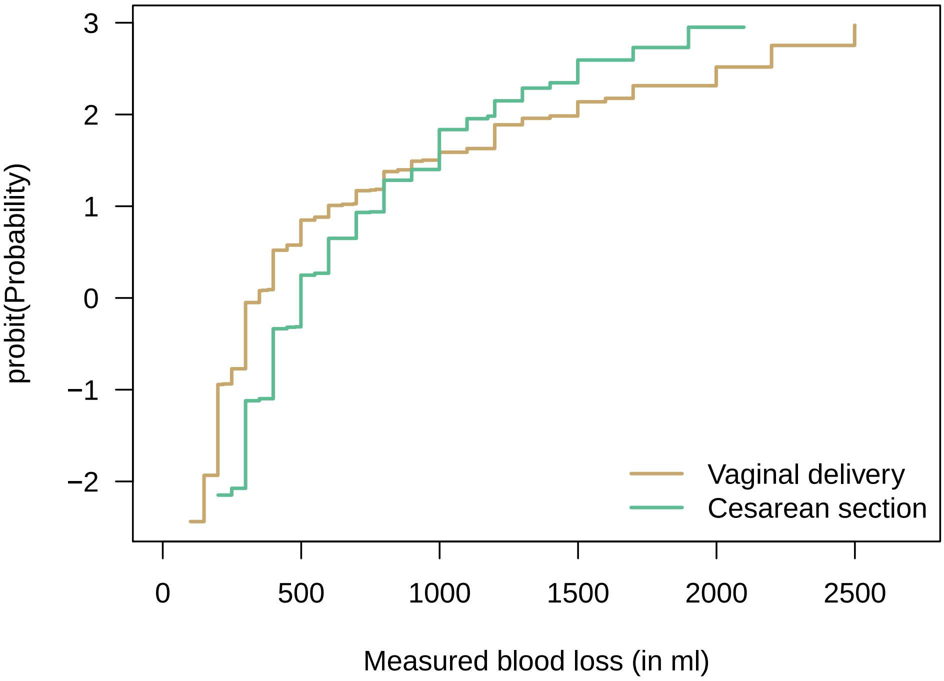


Figure 5.2: Measured blood loss. Empirical cumulative distribution functions (ECDFs) of untransformed measured blood loss on the probit scale for $N_{\text{VD}} = 677$ vaginal deliveries and $N_{\text{CS}} = 632$ Cesarean section at the University Hospital Zurich, Switzerland. The two curves were estimated separately for the two groups. One extreme observation of 5700 ml is not shown.

Reverse Time Hazard-ratio Alternatives

With the choice $F_Z(z) = \text{loglog}^{-1}(z) = \exp(-\exp(-z))$ and $F_Z^{-1}(p) = \text{loglog}(p) = -\log(-\log(p))$ the transformation $h(y) = \text{loglog}(F_0(y)) = -\log(-\log(F_0(y)))$ defines the shift transformation model

$$F_1(y) = \text{loglog}^{-1}(h(y) - \beta).$$

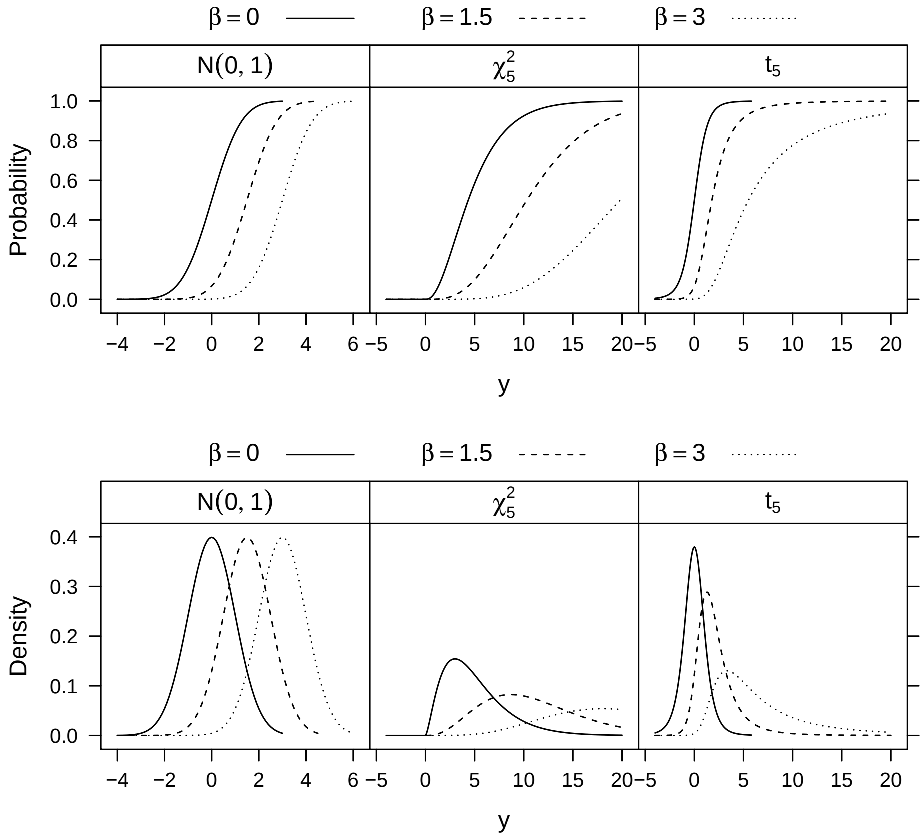


Figure 5.3: Normal Shift-models for two-sample experiments. The lines for $\beta = 0$ correspond to the standard normal, χ^2_5, and t_5 distributions for the first group.

Some rearrangements show that we can express the conditional distribution function F_1 is a relatively simple way as a function of F_0 and β:

$$\begin{aligned}
F_1(y) &= F_Z(h(y) - \beta) = F_Z(F_Z^{-1}(F_0(y)) - \beta) \\
&= \exp(-\exp(-(-\log\{-\log[F_0(y)]\} - \beta))) \\
&= \exp(-(-\log(F_0(y))\exp(\beta)) \\
&= F_0(y)^{\exp(\beta)}
\end{aligned}$$

The reverse time hazard is then

$$\frac{\partial \log(F_1(y))}{\partial y} = \beta \frac{\partial \log(F_0(y))}{\partial y}$$

and thus β can be interpreted as the ratio of the two reverse time hazards, which is, surprisingly, constant for all $y \in \Xi$. The model is also known as Lehmann alternative type I.

Another very interesting interpretation is in terms of maximum-value theory. For $\beta = \log(n), n \in \mathbb{N}$ and $Y_1, \dots, Y_n \sim F_0$ a sample of n independent observations of $Y \mid X = 0$, the distribution function $F_1(y) = F_0(y)^n$ is the distribution function of $\max\{Y_1, \dots, Y_n\}$ because

$$\mathbb{P}(\max\{Y_1, \dots, Y_n\} \le y) = \prod_{i=1}^{n} \mathbb{P}(Y_i \le y) = \prod_{i=1}^{n} F_0(y) = F_0(y)^n.$$

The Gumbel distribution with distributon function $F_Z(z) = \exp(-\exp(-z))$ and quantile function $F_Z^{-1}(p) = \text{loglog}(p)$ is, for this reason, also known as the maximum extreme value or double-exponential distribution.

Example 5.6. The empirical distribution functions for measured blood loss after loglog-Transformation are shown in Figure 5.4.

□

Example 5.7. The shape of the distribution functions F_1 for different values of β under normal, χ^2, and t-distributions for F_0 are given in Figure 5.5. In none of these cases, a simple location shift is visible, the whole shape of the distribution functions under the alternative differ.

□

Hazard-ratio Alternatives

Using $F_Z(z) = \text{cloglog}^{-1}(z) = 1-\exp(-\exp(z))$ and $\text{cloglog}(p) = \log(-\log(1-p))$ we express the transformation function as $h(y) = \text{cloglog}(F_0(y)) = \log(-\log(1 - F_0(y)))$. The interpretation of the shift parameter in the model $F_1(y) = 1 - \exp(-\exp(h(y) - \beta)$ follows from

$$\begin{aligned} F_1(y) &= F_Z(h(y) - \beta) = F_Z(F_Z^{-1}(F_0(y)) - \beta) \\ &= 1 - \exp(-\exp(\log\{-\log[1 - F_0(y)]\} - \beta)) \\ &= 1 - \exp(-(-\log(1 - F_0(y)))\exp(-\beta)) \\ &= 1 - (1 - F_0(y))^{\exp(-\beta)} \end{aligned}$$

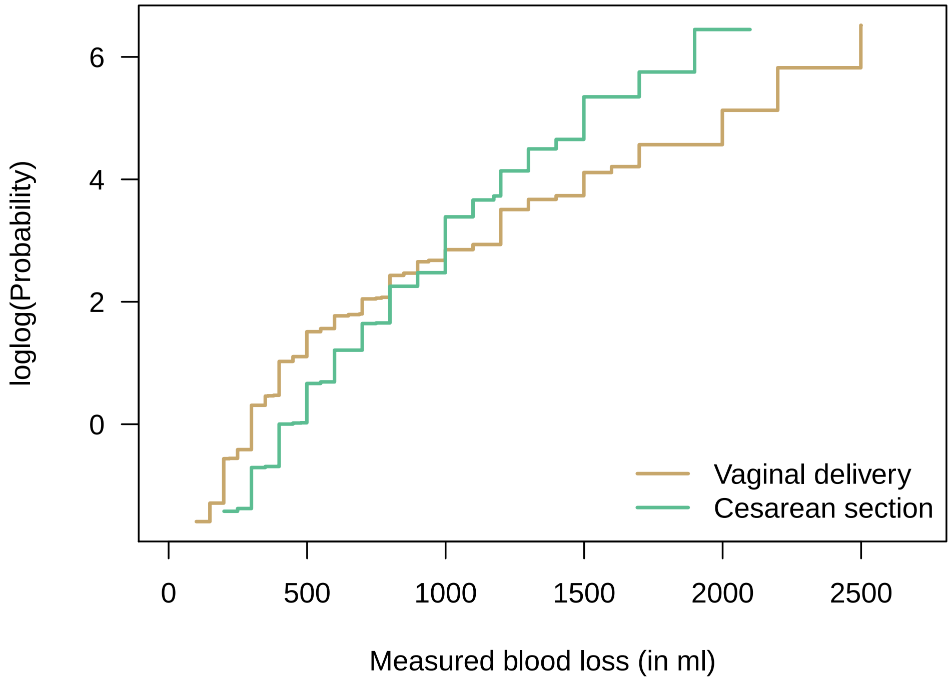


Figure 5.4: Measured blood loss. Empirical cumulative distribution functions (ECDFs) of untransformed measured blood loss on the log-log scale for $N_{\text{VD}} = 677$ vaginal deliveries and $N_{\text{CS}} = 632$ Cesarean section at the University Hospital Zurich, Switzerland. The two curves were estimated separately for the two groups. One extreme observation of 5700 ml is not shown.

The cumulative hazard functions in the two groups are $\Lambda_0(y) = -\log(1 - F_0(y))$ and $\Lambda_1(y) = -\log(1 - F_1(y))$ and thus

$$\begin{aligned} \exp(-\Lambda_1(y)) &= 1 - F_1(y) = \exp(-(-\log(1 - F_0(y))) \exp(-\beta)) \\ &= \exp(-\Lambda_0(y) \exp(-\beta)) \end{aligned}$$

and therefore $\Lambda_1(y) = \Lambda_0(y) \exp(-\beta)$ for all $y \in \Xi$. This result is significant, because $\exp(-\beta)$ is the constant ratio ${}^{\Lambda_1(y)}/_{\Lambda_0(y)}$ for all arguments y.

For $-\beta = \log(n), n \in \mathbb{N}$ and $Y_1, \dots, Y_n \sim F_0$ a sample of n independent

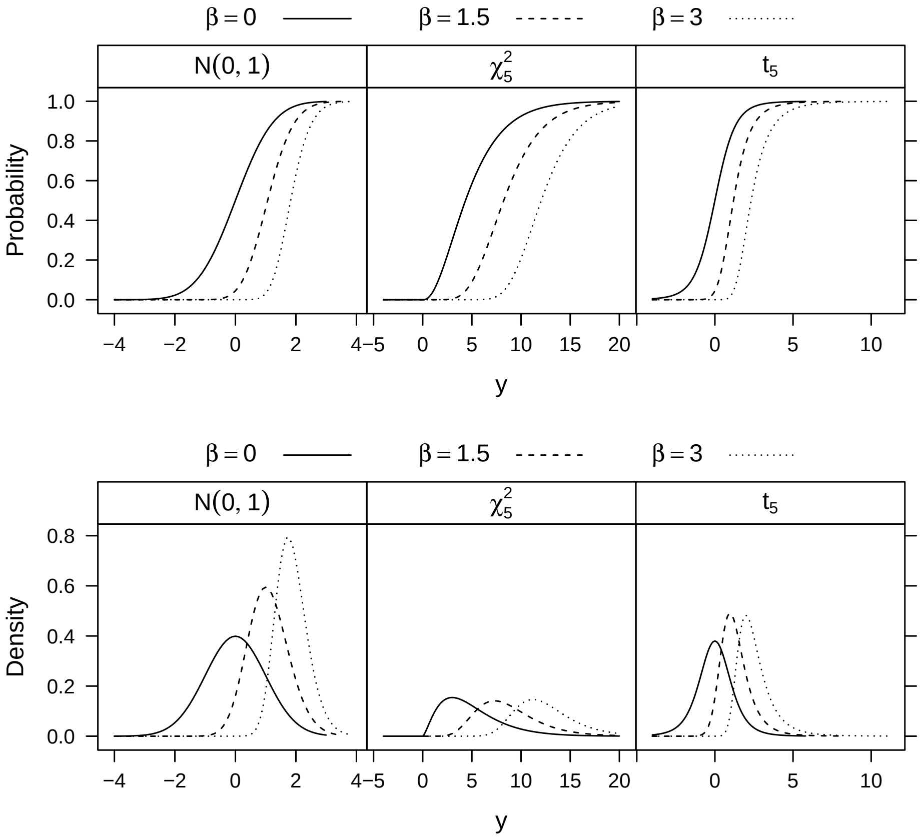


Figure 5.5: Reverse time hazard ratio models for two-sample experiments. The lines for $\beta = 0$ correspond to the standard normal, χ^2_5, and t_5 distributions for the first group.

observations of $Y \mid X = 0$, the survivor function $1 - F_1(y) = (1 - F_0(y))^n$ is the survivor function of $\min\{Y_1, \dots, Y_n\}$ because

$$\begin{aligned}
\mathbb{P}(\min\{Y_1, \dots, Y_n\} \leq y) &= \mathbb{P}\left(\bigcup_{i=1}^n Y_i \leq y\right) = 1 - \mathbb{P}\left(\bigcap_{i=1}^n Y_i > y\right) \\
&= 1 - \prod_{i=1}^n \mathbb{P}(Y_i > y) = 1 - \prod_{i=1}^n (1 - F_0(y)) \\
&= 1 - (1 - F_0(y))^n.
\end{aligned}$$

The Gompertz distribution with distribution function $F_Z(z) = 1 - \exp(-\exp(z))$ and quantile function $pZ^{-1}(p) = \text{cloglog}(p)$ is also known as the minimum

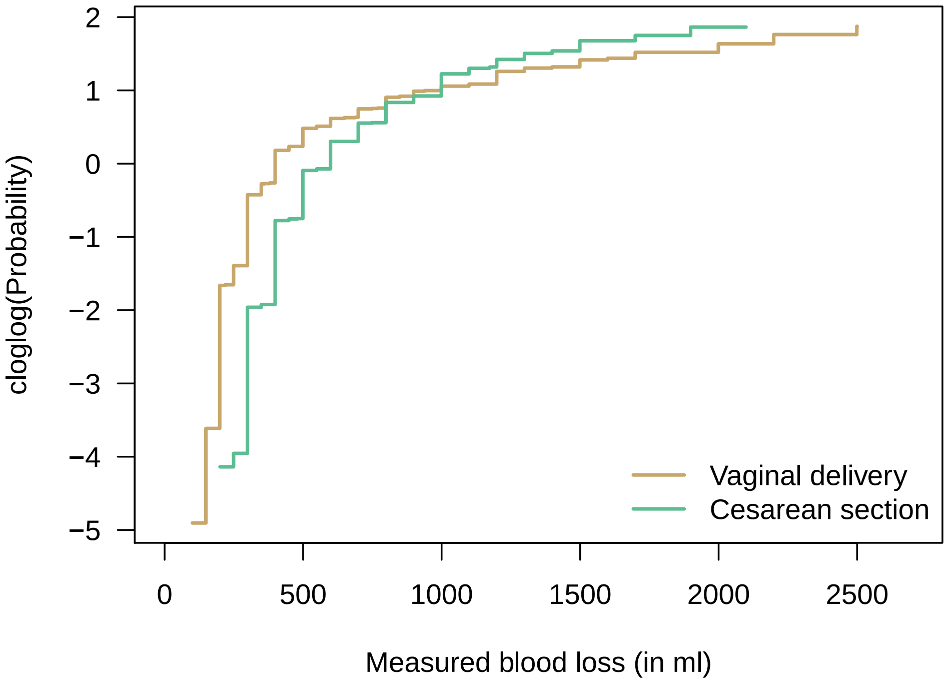


Figure 5.6: Measured blood loss. Empirical cumulative distribution functions (ECDFs) of untransformed measured blood loss on the complementary log-log scale for $N_{\text{VD}} = 677$ vaginal deliveries and $N_{\text{CS}} = 632$ Cesarean section at the University Hospital Zurich, Switzerland. The two curves were estimated separately for the two groups. One extreme observation of 5700 ml is not shown.

extreme value distribution.

Example 5.8. The empirical distribution functions for measured blood loss after cloglog-Transformation are shown in Figure 5.6.

□

Example 5.9. The shape of the distribution functions F_1 for different values of β under normal, χ^2, and t-distributions for F_0 are given in Figure 5.7. In none of these cases, a simple location shift is visible, the whole shape of

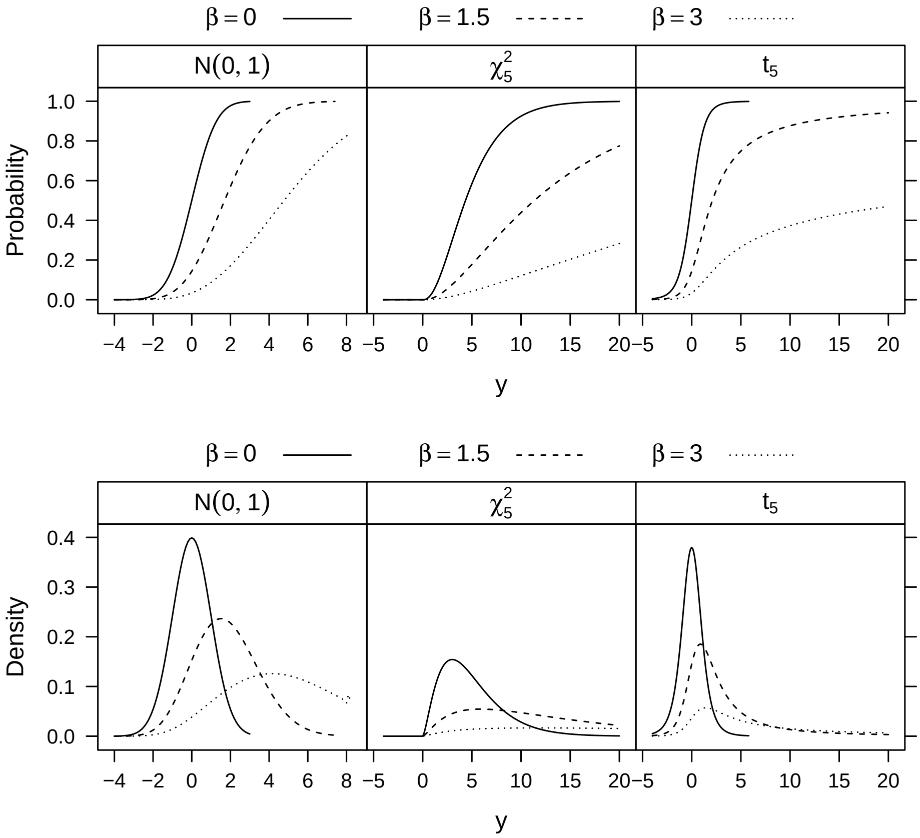


Figure 5.7: Hazard ratio models for two-sample experiments. The lines for $\beta = 0$ correspond to the standard normal, χ^2_5, and t_5 distributions for the first group.

the distribution functions under the alternative differ.

□

Odds-ratio Alternatives

Finally, we study models relying on the choice $F_Z(z) = \text{expit}(z) = (1 - \exp(-z))^{-1}$ and $F_Z^{-1}(p) = \text{logit}(p) = \log(p/1-p)$ with transformation function $h(y) = \text{logit}(F_0(y))$. We can write the distribution function F_1 as

$$\begin{aligned} F_1(y) &= \text{expit}\left(\log\left(\frac{F_1(y)}{1-F_1(y)}\right)\right) = \text{expit}\left(\log\left(\frac{F_0(y)}{1-F_0(y)}\right) - \beta\right) \\ &= \text{expit}(h(y) - \beta). \end{aligned}$$

Therefore, on the scale of the odds function we have

$$\frac{F_1(y)}{1-F_1(y)} = \exp(-\beta)\frac{F_0(y)}{1-F_0(y)}, \quad \forall y \in \Xi.$$

The shift parameter β has therefore a direct interpretation as log-odds ratio.

Example 5.10. The empirical distribution functions for measured blood loss after logit-Transformation are shown in Figure 5.8.

□

Example 5.11. The shape of the distribution functions F_1 for different values of β under normal, χ^2, and t-distributions for F_0 are given in Figure 5.9. In none of these cases, a simple location shift is visible, the whole shape of the distribution functions under the alternative differ.

□

5.4 Connection to Other Models

The line of argumentation in this chapter might have left the reader with the impression that transformation models are rather special. In fact, the opposite is true and there are many strong connections to well-established and prominent classical statistical models. The language of transformation models allows us to communicate ideas without paying special attention to the measurement scale of the response. Unlike the classical classification in methods into procedures for binary, ordered, metric, count, survival or other kinds of data, transformation models handle all measurement scales in a unified way. This section shall highlight the links to some important classical models.

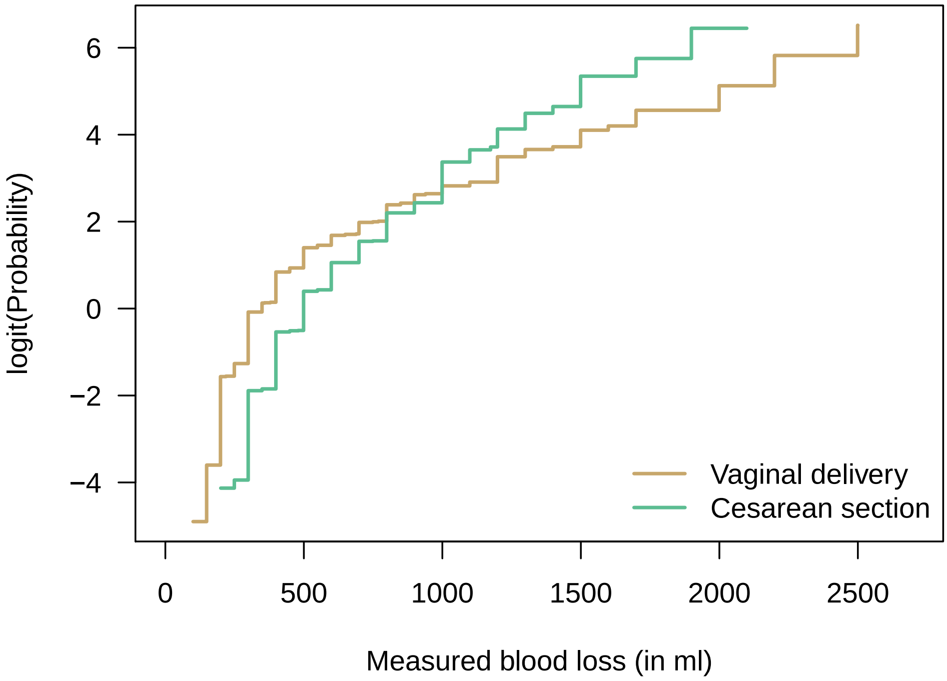


Figure 5.8: Measured blood loss. Empirical cumulative distribution functions (ECDFs) of untransformed measured blood loss on the logit scale for $N_{\text{VD}} = 677$ vaginal deliveries and $N_{\text{CS}} = 632$ Cesarean section at the University Hospital Zurich, Switzerland. The two curves were estimated separately for the two groups. One extreme observation of 5700 ml is not shown.

We first simplify the notation a bit by writing

$$\mathbb{P}(Y \le y \mid X = x) = F_Z(h(y) - \beta x)$$

instead of F_0 and F_1. The distribution function of $Y \mid X = 0$ becomes $F_0(y) = F_Z(h(y) - \beta 0) = F_Z(h(y))$ and for $Y \mid X = 1$ we now say $F_1(y) = F_Z(h(y) - \beta 1) = F_Z(h(y) - \beta)$.

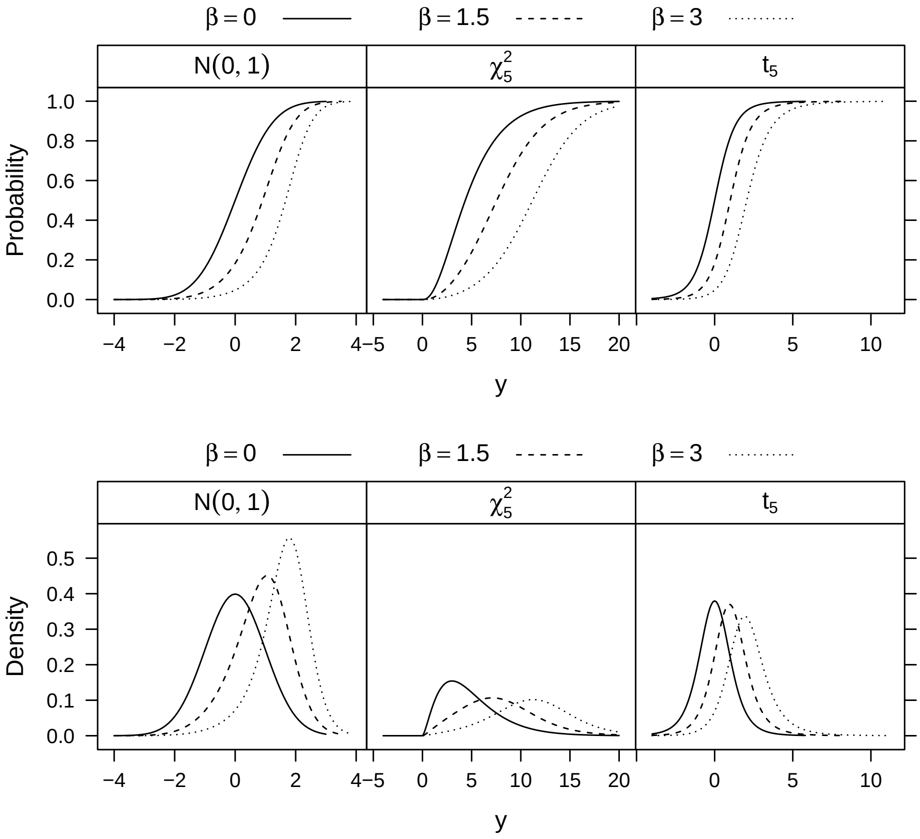


Figure 5.9: Odds ratio models for two-sample experiments. The lines for $\beta = 0$ correspond to the standard normal, χ^2_5, and t_5 distributions for the first group.

Binary Response

The well-known binary generalised linear model for binary response $Y \in \{\upsilon_1, \upsilon_2\}$, with υ_1 meaning success and υ_2 failure for example, can be written as

$$\mathbb{P}(Y \le \upsilon_1 \mid X = x) = \mathbb{P}(Y = \upsilon_1 \mid X = x) = F_Z(\vartheta - \beta x)$$

with link function F_Z and two unknown parameters ϑ and β. The first parameter ϑ is the intercept describing the probability of success $F_0(\upsilon_1) = F_Z(\vartheta)$. One should note that $\mathbb{P}(Y \le \upsilon_1 \mid X = x)$ is equal to the conditional mean $\mathbb{E}(\mathbb{1}(Y \le \upsilon_1) \mid X = x) = \mathbb{E}(\mathbb{1}(Y = \upsilon_1) \mid X = x) =: \pi(\boldsymbol{x})$ which motivates the model formulation

$$F_Z^{-1}(\pi(\boldsymbol{x})) = \vartheta - \beta x$$

with $\mathbb{1}(Y = \upsilon_1) \mid X = x \sim \text{B}(1, \pi(\boldsymbol{x}))$. This model formulation featured in most textbooks is equivalent to our definition as a transformation model. The function F_Z^{-1} is called link function and it's inverse F_Z is the response function in the language of generalised linear models. Biostatisticians and epidemiologists seldomly deviate from the choice $F_Z = \text{expit}$ for reasons explained in the next example.

Example 5.12. For a binary response $Y \in \{\upsilon_1, \upsilon_2\}$, one often formulates the discrepancy between the two conditional distributions ($X = 0$ vs. $X = 1$) by their odds ratio

$$\text{OR} = \frac{\frac{F_1(\upsilon_1)}{1-F_1(\upsilon_1)}}{\frac{F_0(\upsilon_1)}{1-F_0(\upsilon_1)}}$$

or log-odds ratio

$$\log(\text{OR}) = \log\left(\frac{F_1(\upsilon_1)}{1 - F_1(\upsilon_1)}\right) - \log\left(\frac{F_0(\upsilon_1)}{1 - F_0(\upsilon_1)}\right).$$

We can write this equation as

$$\log\left(\frac{F_1(\upsilon_1)}{1 - F_1(\upsilon_1)}\right) = \log\left(\frac{F_0(\upsilon_1)}{1 - F_0(\upsilon_1)}\right) + \log(\text{OR})$$

or even

$$F_1(\upsilon_1) = \frac{1}{1 + \exp\left(-\left(\log\left(\frac{F_0(\upsilon_1)}{1-F_0(\upsilon_1)}\right) + \log(\text{OR})\right)\right)}.$$

Noting that $\text{expit}(z) = (1 + \exp(-z))^{-1}$ and $\text{logit}(p) = \text{expit}^{-1}(p) = \log(p/(1-p))$ are the distribution and quantile functions of the standard logistic distribution, the formula reads

$$F_1(\upsilon_1) = \text{expit}(\text{logit}(F_0(\upsilon_1)) + \log(\text{OR})).$$

With $\vartheta = \text{logit}(F_0(\upsilon_1))$ and $-\beta = \log(\text{OR})$ we see the equivalence to our transformation model.

This model is an interesting special case where the parameter estimates can be computed analytically.

	$x = 0$	$x = 1$
$Y = \upsilon_1$	N_{10}	N_{11}
$Y = \upsilon_2$	N_{20}	N_{21}

$N = N_{10} + N_{11} + N_{20} + N_{21}$

$$\begin{aligned}
\mathbb{P}(Y = y_1 \mid x = 0) &= \text{expit}(\vartheta_1) \\
\mathbb{P}(Y = y_2 \mid x = 0) &= 1 - \text{expit}(\vartheta_1) \\
\mathbb{P}(Y = y_1 \mid x = 1) &= \text{expit}(\vartheta_1 - \beta) \\
\mathbb{P}(Y = y_2 \mid x = 1) &= 1 - \text{expit}(\vartheta_1 - \beta)
\end{aligned}$$

We obtain OR $= (N_{10}N_{21})/(N_{11}N_{20})$ as the cross-product ratio. From this solution it becomes clear that the odds ratio is invariant when the roles of Y and X are interchanged.

□

We now move on to an at least ordered sample space Ξ and write

$$F_1(y) = \text{expit}(\text{logit}(F_0(y)) + \log(\text{OR})) = \text{expit}(\text{logit}(F_0(y)) - \beta)$$

assuming that the two distribution functions increase with increasing y but that the odds ratio remains constant. The treatment effect $-\beta$ is then the constant log-odds ratio.

Ordered Categorical Response

Cumulative model for ordered categorical response $Y \in \{\upsilon_1 < \cdots < \upsilon_K\}$ with

$$\mathbb{P}(Y \le \upsilon_k \mid X = x) = F_Z(\vartheta_k - \beta x), \quad k = 1, \dots, K - 1$$

with response function F_Z and $\mathbb{P}(Y \le \upsilon_K \mid X = x) = 1$. With $F_Z = \text{expit}$, the parameter $-\beta$ is a contant log-odds ratio and thus the model is called proportional odds logistic regression model.

Example 5.13. For each possible cut-off point $\upsilon_1, \dots, \upsilon_{K-1}$, we set-up a 2×2 table

	$x = 0$	$x = 1$
$Y \leq \upsilon_k$	N_{k10}	N_{k11}
$Y > \upsilon_k$	N_{k20}	N_{k21}

For each of these tables, we deal with a binary logistic regression model

$$\mathbb{P}(Y \leq \upsilon_k \mid X = x) = \operatorname{expit}(\vartheta_k - \beta x).$$

The subtle detail here is that we allow the intercept terms to change with the cut-off point υ_k but we keep the log-odds ratio $-\beta$ constant. □

Although $F_Z = \operatorname{expit}$ is the default choice in many areas of research, alternative choices often lead to better interpretable parameters.

Example 5.14. In the context of Example 3.12, with Y denoting the number of menstrual cycles needed to conceive, the hazard is the probability to score during the ongoing yth cycle conditional on $y - 1$ unsuccessful attempts. Let's say we run a randomised trial with the aim to evaluate the effect of a stress-reduction programme ($X = 1$) compared to some wellness placebo ($X = 0$) on the distribution of Y. One way of measuring the efficacy of the treatment or lack thereof would be to compare the hazards by

$$\exp(-\beta) = \frac{\lambda_Y(y \mid X = 1)}{\lambda_Y(y \mid X = 0)}.$$

The ratio of the probabilities to conceive in any cycle given the so-far unsuccessful cycles compares the treated and placebo group. This parameter β has a clear and meaningful interpretation and can be estimated in a cumulative model with cloglog link function. □

Continuous Response

Binary models and models for ordered categorical responses are ubiquitous in statistics but, quite surprisingly, cumulative models for $Y \in \Xi \subseteq \mathbb{R}$

$$\mathbb{P}(Y \leq y \mid X = x) = F_Z(\boldsymbol{a}(y)^\top \boldsymbol{\vartheta} - \beta x) \tag{5.3}$$

are less well studied. There is one exception, namely the famous and ground-breaking Cox proportional hazards model. As the name suggests, the cloglog link function is used and $-\beta$ is a log-hazard ratio (the model is usally defined with a positive shift term). The terminology is rather subtle here: Cox suggested to estimate the model with a nonparametric parameterisation of the transformation function $\boldsymbol{a}(y)^\top \boldsymbol{\vartheta}$, which is equivalent to the log-cumulative hazard function in this model. The famous partial likelihood can be applied to estimate the log-hazard ratio $-\beta$.

Over three decades, it was suggested time and again to parameterise $\boldsymbol{a}(y)^\top \boldsymbol{\vartheta}$ by some form of splines, and also polynomials in Bernstein form were paid attention to. The simplest choice is a linear baseline hazard function $\boldsymbol{a}(y) = c(1, \log(y))^\top$, leading to a Weibull model. We will come back to these important models in Chapter 11.

Only recently were models assuming a constant odds ratio, *i.e.* with logit-link, scrutinised. There are very interesting connections to two-sample rank tests, but we will postpone the discussion until after the next section focusing on inference procedures for β.

5.5 Inference for the Shift Parameter

When comparing two groups under model (5.1), the main interest is in the shift parameter β. This is a dangerous statement to make, because I believe one should be interested in both distributions and not just the way how they might differ.

There are two conceptually different approaches for inference about β: One can either fit all unknown model parameters $\boldsymbol{\vartheta}$ and β simultaneously

$$(\hat{\boldsymbol{\vartheta}}^\top, \hat{\beta})^\top = \operatorname*{arg\,max}_{\boldsymbol{\vartheta} \in \mathbb{R}^P, \beta \in \mathbb{R}} \ell((\boldsymbol{\vartheta}^\top, \beta)^\top)$$

under appropriate constraints in $\boldsymbol{\vartheta}$ first and assess the variability of $\hat{\beta}$ in a second step. Alternatively, one can estimate a model under the additional constraint $\beta = \beta_0$, where one estimates the model under an hypothetical value β_0 of the shift paramater by

$$\hat{\boldsymbol{\vartheta}}(\beta_0) = \underset{\boldsymbol{\vartheta} \in \mathbb{R}^P}{\arg\max}\, \ell((\boldsymbol{\vartheta}^\top, \beta_0)^\top)$$

first, followed by an assessment of how well this model fits the data. Deviations from this model indicate that the true shiftparm β might be far off our hypothetical value β_0. For this restricted maximum likelihood estimator we write $(\hat{\boldsymbol{\vartheta}}(\beta_0)^\top, \beta_0)^\top$.

In both cases, the score function is $P + 1$ dimensional

$$\boldsymbol{s}((\boldsymbol{\vartheta}^\top, \beta)^\top) = (\boldsymbol{s}(\boldsymbol{\vartheta})^\top, \boldsymbol{s}(\beta))^\top$$

with the first P elements corresponding to the parameters $\boldsymbol{\vartheta}$ of the transformation function and the last element corresponding to the shift parameter β.

The Fisher information matrix has $(P+1)$ rows and $P+1$ columns and is partitioned into four submatrices

$$\boldsymbol{F}((\boldsymbol{\vartheta}^\top, \beta)^\top) = \left(\begin{array}{cc} \boldsymbol{F}_{11}((\boldsymbol{\vartheta}^\top, \beta)^\top) & \boldsymbol{F}_{12}((\boldsymbol{\vartheta}^\top, \beta)^\top) \\ \boldsymbol{F}_{12}((\boldsymbol{\vartheta}^\top, \beta)^\top)^\top & \boldsymbol{F}_{22}((\boldsymbol{\vartheta}^\top, \beta)^\top) \end{array} \right).$$

The $P \times P$ block $\boldsymbol{F}_{11}((\boldsymbol{\vartheta}^\top, \beta)^\top)$ is the Fisher information for the parameters $\boldsymbol{\vartheta}$ and the scalar $\boldsymbol{F}_{22}((\boldsymbol{\vartheta}^\top, \beta)^\top)$ is the Fisher information for β. The inverse Fisher information matrix is

$$\Sigma((\boldsymbol{\vartheta}^\top, \beta)^\top) = \left(\begin{array}{cc} \Sigma_{11}((\boldsymbol{\vartheta}^\top, \beta)^\top) & \Sigma_{12}((\boldsymbol{\vartheta}^\top, \beta)^\top) \\ \Sigma_{12}((\boldsymbol{\vartheta}^\top, \beta)^\top)^\top & \Sigma_{22}((\boldsymbol{\vartheta}^\top, \beta)^\top) \end{array} \right) = \boldsymbol{F}((\boldsymbol{\vartheta}^\top, \beta)^\top)^{-1}$$

with the same block structure. The bottom-right part Σ_{22} can be computed from the Schur complement $\Sigma_{22} = (\boldsymbol{F}_{22} - \boldsymbol{F}_{12}^\top \boldsymbol{F}_{11}^{-1} \boldsymbol{F}_{12})^{-1}$. We abbreviate the inverse Fisher information evaluated at the estimates by $\hat{\Sigma} = \Sigma((\hat{\boldsymbol{\vartheta}}^\top, \hat{\beta})^\top)$.

We are now exclusively study the situtation when the data were indeed generated from the model

$$\mathbb{P}(Y \le y \mid X = x) = F_Z(\boldsymbol{a}(y)^\top \boldsymbol{\vartheta} - \beta_0 x),$$

that is, the true model parameters are ϑ and $\beta = \beta_0$. For all possible parameterisations in terms of basis functions $\boldsymbol{a}$ and all forms of the corresponding likelihood (nonparametric, discrete, censored, ...), the following three important statements can be made

$$\begin{aligned}
\mathrm{W}(\beta_0) &:= \frac{\hat{\beta} - \beta_0}{\sqrt{\Sigma_{22}((\hat{\boldsymbol{\vartheta}}^\top, \hat{\beta})^\top)}} \xrightarrow{D} \mathrm{N}(0,1) \\
\mathrm{R}(\beta_0) &:= \sqrt{\Sigma_{22}(\hat{\boldsymbol{\vartheta}}(\beta_0)^\top, \beta_0)^\top} \boldsymbol{s}(\beta_0) \xrightarrow{D} \mathrm{N}(0,1) \\
\mathrm{LR}(\beta_0) &:= -2(\ell((\hat{\boldsymbol{\vartheta}}(\beta_0)^\top, \beta_0)^\top) - \ell((\hat{\boldsymbol{\vartheta}}^\top, \hat{\beta})^\top)) \xrightarrow{D} \chi^2_1.
\end{aligned}$$

$\mathrm{W}(\beta_0)$ is called the Wald statistic (named after Abraham Wald), $\mathrm{R}(\beta_0)$ the Rao score statistic (named after Calyampudi Radhakrishna Rao, the statistic is also known as Lagrange Multiplier statistic), and $\mathrm{LR}(\beta_0)$ is the likelihood ratio statistic. Maximum-likelihood-based inference relies on either of these three statistics and we will look at them separately in the next sections.

Wald Tests and Intervals

The Wald statistic can be used to formally test hypotheses about the shift parameter β:

$$\begin{aligned}
&H_0 : \beta \le \beta_0 \,\text{vs.}\, H_1 : \beta > \beta_0 \quad \text{“greater”} \\
&H_0 : \beta \ge \beta_0 \,\text{vs.}\, H_1 : \beta < \beta_0 \quad \text{“less”} \\
&H_0 : \beta = \beta_0 \,\text{vs.}\, H_1 : \beta \neq \beta_0 \quad \text{“two-sided”}.
\end{aligned}$$

In the NHST framework, we would reject H_0 favour of a “greater” alternative H_1 at level $0 < \alpha < 0.5$ when the observed Wald statistic $\mathrm{W}(\beta_0) = w$ is larger than $\Phi^{-1}(1 - \alpha)$, thus controlling the type I error

$$\sup_{\beta \le \beta_0} \mathbb{P}(\mathrm{W}(\beta_0) > \Phi^{-1}(1 - \alpha) \mid \beta) = \mathbb{P}(\mathrm{W}(\beta_0) > \Phi^{-1}(1 - \alpha) \mid \beta = \beta_0) \le \alpha$$

for large sample sizes $N \to \infty$. The P-value describes the probability to observe a Wald statistic as large or larger than w

$$\mathbb{P}(\mathrm{W}(\beta_0) > w \mid \beta = \beta_0) \to 1 - \Phi(w) \,\text{for}\, N \to \infty$$

and $w > \Phi^{-1}(1-\alpha)$ is equivalent to $1 - \Phi(w) \le \alpha$.

We one is interested in a "less" alternative, the null hypothesis would be rejected when $\text{W}(\beta_0) = w$ is smaller than $\Phi^{-1}(\alpha)$ or, equivalently, when $\Phi(w) \le \alpha$. In the two-sided case, the null is rejected at level 2α when $w < \Phi^{-1}(\alpha)$ or $w > \Phi^{-1}(1-\alpha)$. The two-sided P-value is then $2 \times \min(\Phi(\text{W}(\beta_0)), 1 - \Phi(\text{W}(\beta_0)))$. Anyhow, it seems a bit rediculus to go through the pain of reading 114 pages introducing transformation models only to end-up with a yes/no decision against or in favour of some alternative H_1. At the very least, we would like to know which subset of the parameter space for β is in line with the observed data.

A $(1-2\alpha) \times 100\%$ confidence interval is given by the set $\{\beta_0 \in \mathbb{R} \mid \Phi^{-1}(\alpha) < \text{W}(\beta_0) \le \Phi^{-1}(1-\alpha)\}$. Because for large sample sizes $N \to \infty$ we have

$$\mathbb{P}(\text{W}(\beta_0) \le \Phi^{-1}(\alpha) \mid \beta = \beta_0) = \alpha, \, 0 < \alpha < .5$$

we can solve $\text{W}(\beta_0) \le \Phi^{-1}(\alpha)$ and $\text{W}(\beta_0) \le \Phi^{-1}(1-\alpha)$ for β and obtain the confidence interval

$$\left[\hat{\beta} - \Phi^{-1}(\alpha)\sqrt{\Sigma_{22}((\hat{\boldsymbol{\vartheta}}^\top, \hat{\beta})^\top)}, \hat{\beta} + \Phi^{-1}(\alpha)\sqrt{\Sigma_{22}((\hat{\boldsymbol{\vartheta}}^\top, \hat{\beta})^\top)}\right].$$

This result is very convenient because we only need the maximum likelihood estimator of $\hat{\beta}$ and the corresponding estimated standard error $\text{SE} = \sqrt{\Sigma_{22}((\hat{\boldsymbol{\vartheta}}^\top, \hat{\beta})^\top)}$ for the computation of these Wald-intervals at arbitrary levels 2α. One-sided $(1-\alpha) \times 100\%$ intervals $(-\infty, \hat{\beta} + \Phi^{-1}(\alpha)\text{SE}]$ or $[\hat{\beta} - \Phi^{-1}(\alpha)\text{SE}, \infty)$ are simply the limits of the two-sided $(1-2\alpha) \times 100\%$ interval.

Example 5.15. $\mathbb{P}(Y \le y \mid X = x) = \exp(-\exp(-(h(y) - \beta)))$ such that $\mathbb{P}(Y \le y \mid \text{Cesarean}) = \mathbb{P}(Y \le y \mid \text{Vaginal})^{\exp(\beta)}$. The transformation function $h(y)$ is modelled by a polynomial in Bernstein form of order $M = 6$ and the nonparametric log-likelihood is optimed for parameter estimation. The fit is contrasted to the empiricial cumulative distribution functions in Figure 5.10.

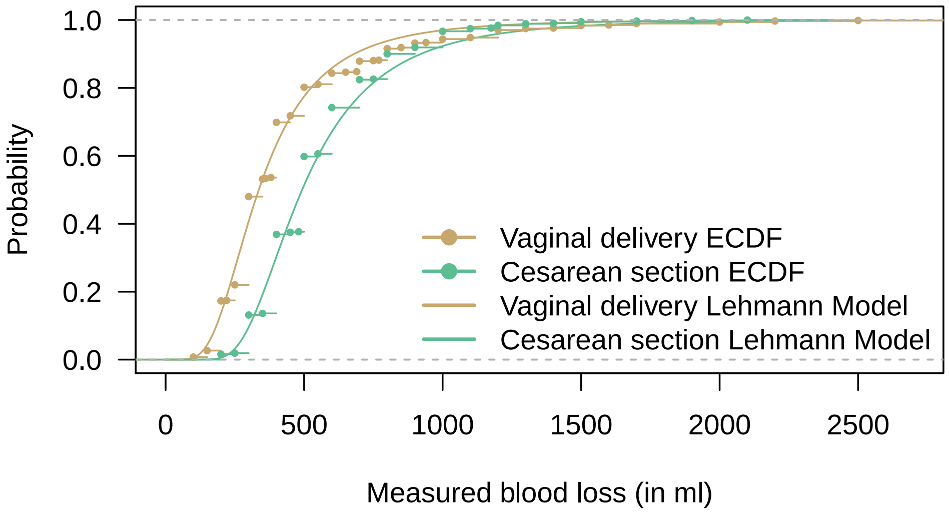


Figure 5.10:

The parameter estimates and their corresponding covariance matrix are

$$
\begin{aligned}
\hat{\boldsymbol{\vartheta}} &= (-2.794, 1.435, 2.309, 3.722, 4.637, 5.448, 6.258)^\top \\
\hat{\beta} &= 0.961 \\
\hat{\Sigma} &= \left(\begin{array}{rrrrrrr|r}
0.061 & -0.148 & 0.249 & -0.274 & 0.164 & -0.029 & -0.018 & -0.001 \\
-0.148 & 0.409 & -0.741 & 0.868 & -0.541 & 0.099 & 0.063 & 0.004 \\
0.249 & -0.741 & 1.516 & -1.905 & 1.278 & -0.237 & -0.147 & 0.003 \\
-0.274 & 0.868 & -1.905 & 2.691 & -1.972 & 0.417 & 0.263 & 0.001 \\
0.164 & -0.541 & 1.278 & -1.972 & 1.770 & -0.429 & -0.261 & 0.004 \\
-0.029 & 0.099 & -0.237 & 0.417 & -0.429 & 0.313 & 0.212 & 0.002 \\
-0.018 & 0.063 & -0.147 & 0.263 & -0.261 & 0.212 & 0.231 & 0.003 \\
\hline
-0.001 & 0.004 & 0.003 & 0.001 & 0.004 & 0.002 & 0.003 & 0.004
\end{array}\right)
\end{aligned}
$$

With a standard error for $\hat{\beta}$ of SE $= 0.061$ we obtain a Wald statistic for the hypothesis $H_0 : \beta = 0$ of $\mathrm{W}(0) = 15.715$. In light of Figure 5.10 showing clearly more blood loss from Cesarean sections, a NHST for this hypothesis is simply not interesting. A 95% Wald interval for β is $(0.841, 1.081)$ sheds more light on the matters of affair but, unfortunately, these numbers are also not directly interpretable.

From the estimated model parameters, we can obtain the median blood loss for vaginal deliveries by solving $\exp(-\exp(-\boldsymbol{a}(y)^\top \hat{\boldsymbol{\vartheta}})) = 0.5$ for y and obtain $y_{0.5} = 337$ after rounding to integers. We get $\hat{\mathbb{P}}(Y \le 337 \mid \text{Vaginal}) = 0.501$ and, due to our choice $F_Z(z) = \text{loglog}^{-1}(z)$, we also have $\hat{\mathbb{P}}(Y \le 337 \mid \text{Cesarean}) = 0.165 = 0.501^{\exp(0.961)}$. This means that 50% of mothers giving birth by vaginal delivery loose at most 337 ml whereas this is only the case in 16% of Cesarean sections. When we switch to log-probabilities, we get

$$\exp(0.961) = 2.614 = \frac{\log(\hat{\mathbb{P}}(Y \le y \mid \text{Cesarean}))}{\log(\hat{\mathbb{P}}(Y \le y \mid \text{Vaginal}))} \quad \forall y \in \mathbb{R}^+.$$

Transformating the Wald confidence interval by exp() gives $(2.319, 2.947)$ and covers the true unknown ratio $\exp(\beta)$ of the log-probabilities at any quantile y with probability 95%. Because the model assumes proportional reverse time hazard ratios, the ratio of the reverse time hazards between Cesarean and vaginal birth is constant and an interval-estimate for this constant is given by $(2.319, 2.947)$. We could also say that only one out of 2 to 3 vaginal deliveries results in blood loss as severe as a single Cesarean section.

Because Wald intervals are simple to compute it is tempting to look at intervals for increasing confidence levels. In Figure 5.11, the confidence curve (or P-value function) is shown.

□

Unfortunately, neither the Wald test nor the corresponding confidence intervals are invariant with respect to reparameterisation of β. The confidence interval for $\exp(\beta)$ presented in Example 5.15 maintains its nominal level of 95% but it might not be the best available interval. Score and likelihood intervals are invariant with respect to reparameterisations and can immediately be used to compute confidence intervals for $\exp(\beta)$, however, at the expense of more laborous computations. Before we look into these issues, we will review some ways of formulating relevant hypotheses.

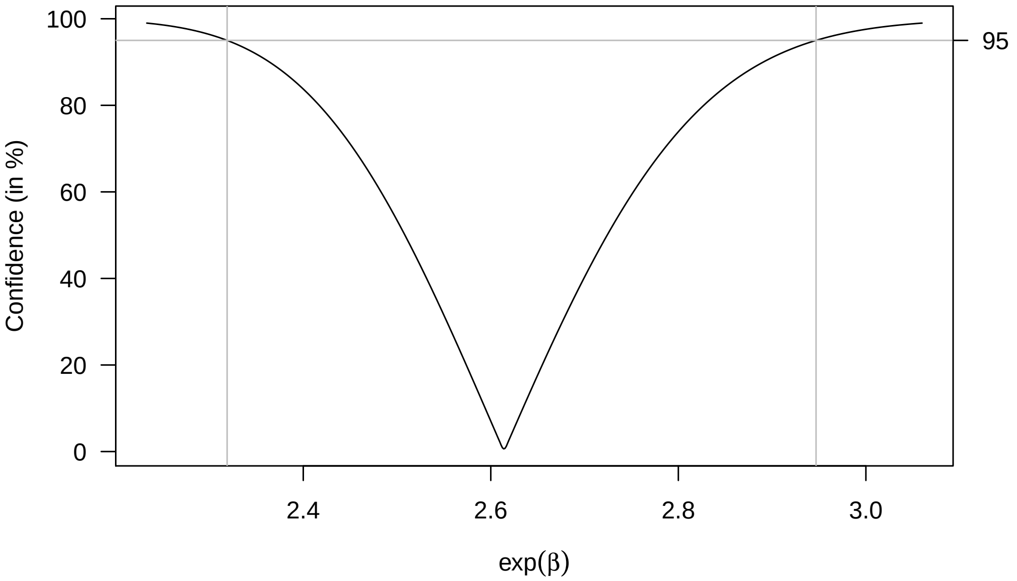


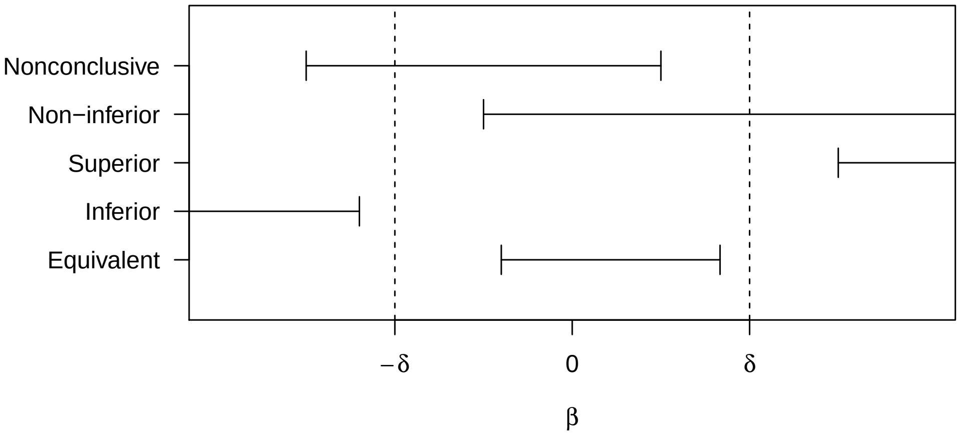
Figure 5.11: Confidence is the two-sided P-value.

Very often the default $\beta_0 = 0$ is used to define the null hypotheses. If $\beta > 0$ corresponds to a benefitial effect, for example of a new treatment compared to an established therapy, one might be interested in slightly different pairs of hypothesis and alternative. For some relevant effect $\delta > 0$ one may want to test superiority, non-inferiority, or equivalence:

$$H_0 : \beta \le \delta \,\text{vs.}\, H_1 : \beta > \delta \quad \text{“superior”}$$
$$H_0 : \beta \le -\delta \,\text{vs.}\, H_1 : \beta > -\delta \quad \text{“non-inferior”}$$
$$H_0 : \beta < -\delta \vee \beta > \delta \,\text{vs.}\, H_1 : -\delta < \beta \le \delta \quad \text{“equivalent”}$$

Superiority means that the novel treatment is better by a relevant margin, so small improvements $0 < \beta < \delta$ are not considered relevant and are, together with a negative treatment effect $\beta < 0$, part of the null hypothesis.

Non-inferority refers to the situtation where the novel treatment is allowed to be a little bit less effective than the control $(-\delta < \beta < 0)$, equally effective $(\beta = 0)$, better but not relevantly so $(0 < \beta < \delta)$ or even superior



($\beta > \delta$) under the alternative. The null hypothesis only consists of the situation where the novel treatment is inferior to the control ($\beta < -\delta$).

Equivalence alternatives formulate that the novel treatment is neither superior nor inferior, with $|\beta| < \delta$. The null hypothesis thus combines the superior and inferior situations.

It is relatively simple to assess these hypotheses by confidence intervals. If the lower bound of a one-sided confidence interval is greater than $-\delta$, non-inferiority can be concluded. Equivalence is given when the two-sided confidence interval is contained in the equivalence range $[-\delta, \delta]$. The big problem is to define the relevance margin δ but shift parameters in transformation models (5.1) are a good basis for the choice of δ.

Example 5.16. With $F_Z = \text{loglog}^{-1}$ we have

$$\exp(\beta) = \frac{\log(\mathbb{P}(Y \le y \mid X = 1)}{\log(\mathbb{P}(Y \le y \mid X = 0)} \quad \forall y \in \Xi$$

and we could define $\exp(\beta) > 1.1$, that is at least 10% increase from $\log(\mathbb{P}(Y \le y \mid X = 0)$ to $\log(\mathbb{P}(Y \le y \mid X = 1)$, as our relevance margin with $\delta = \log(1.1)$. The advantage here is that the choice of δ does not depend on the measurement scale of Y. The same line of argumentation can be applied for hazard ratios $\exp(\beta)$ (using $F_Z = \text{cloglog}^{-1}$) and odds ratios $\exp(\beta)$ (using $F_Z = \text{expit}$). □

Example 5.17. Y describes the reduction of diastolic blood pressure (in mm Hg) observed in patients suffering from major depression after four weeks of daily treatment with $0.2 - 0.4$ mg Moxonodin or $25 - 50$ mg Captopril. We assume $Y_\text{M} \sim \text{N}(\mu_\text{M}, \sigma^2)$ and $Y_\text{C} \sim \text{N}(\mu_\text{C}, \sigma^2)$ and test the equivalence hypothesis $H_0 : (\mu_\text{C} - \mu_\text{M})/\sigma \le -\delta_1$ or $(\mu_\text{C} - \mu_\text{M})/\sigma \ge \delta_2$ against the alternative $H_1 : -\delta_1 < (\mu_\text{C} - \mu_\text{M})/\sigma < \delta_2$ with $(-\delta_1, \delta_2) = (-.5, 1)$.

The parameter $\beta = (\mu_\text{C} - \mu_\text{M})/\sigma$, the standardised difference of means, is the shift parameter in a normal linear regression model parameterised in terms of a shift transformation model (5.1) with conditional distributions

$$\begin{aligned}
\mathbb{P}(Y_\text{M} \le y) &= \Phi\left(\frac{1}{\sigma} y - \frac{\mu_\text{M}}{\sigma}\right) = \Phi((1, y)\boldsymbol{\vartheta}) \\
\mathbb{P}(Y_\text{C} \le y) &= \Phi\left(\frac{1}{\sigma} y - \frac{\mu_\text{M}}{\sigma} - \frac{\mu_\text{C} - \mu_\text{M}}{\sigma}\right) = \Phi((y, 1)\boldsymbol{\vartheta} - \beta)
\end{aligned}$$

In contrast to the normal linear regression model in its classical parameterisation in terms of the mean difference $\mu_\text{C} - \mu_\text{M}$, the transformation model directly contains the standardised difference of means as a parameter and it is straightforward to compute a Wald confidence interval.

The parameter estimates and their corresponding covariance matrix are

$$\begin{aligned}
\hat{\boldsymbol{\vartheta}} &= (-0.637, 0.153)^\top \\
\hat{\beta} &= 0.463 \\
\hat{\Sigma} &= \left(\begin{array}{rr|l} 0.092 & -0.002 & 0.077 \\ -0.002 & 0.000 & 0.001 \\ \hline 0.077 & 0.001 & 0.171 \end{array}\right)
\end{aligned}$$

and the 95% Wald interval for the standardised difference of means is $(-0.348, 1.274)$. The interval is not completely contained in the equivalence range $(-0.5, 1)$ and thus one cannot reject the null in favour of the equivalence hypothesis H_1. □

Score Tests and Intervals

The definition of the Rao score statistic $\mathrm{R}(\beta_0)$ is worth a closer look. For the true unknown parameters $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}_0$ and $\beta = \beta_0$ we have the classical result

$$\frac{1}{\sqrt{N}} \boldsymbol{s}((\boldsymbol{\vartheta}_0^\top, \beta_0)^\top) \xrightarrow{D} \mathrm{N}_{P+1}\left(\boldsymbol{0}, \boldsymbol{I}((\boldsymbol{\vartheta}_0^\top, \beta_0)^\top)\right)$$

involving the expected Fisher information

$$\boldsymbol{I}((\boldsymbol{\vartheta}^\top, \beta)^\top) = \left(\begin{array}{cc} \boldsymbol{I}_{11}((\boldsymbol{\vartheta}^\top, \beta)^\top) & \boldsymbol{I}_{12}((\boldsymbol{\vartheta}^\top, \beta)^\top) \\ \boldsymbol{I}_{12}((\boldsymbol{\vartheta}^\top, \beta)^\top)^\top & \boldsymbol{I}_{22}((\boldsymbol{\vartheta}^\top, \beta)^\top) \end{array} \right).$$

Because the observed Fisher information is a consistent estimate of the expected Fisher information $N^{-1}\boldsymbol{F}((\boldsymbol{\vartheta}_0^\top, \beta_0)^\top) \xrightarrow{\mathbb{P}} \boldsymbol{I}((\boldsymbol{\vartheta}_0^\top, \beta_0)^\top)$ one can use the quadratic score statistic

$$\boldsymbol{s}((\boldsymbol{\vartheta}_0^\top, \beta_0)^\top)^\top \boldsymbol{F}((\boldsymbol{\vartheta}_0^\top, \beta_0)^\top)^{-1} \boldsymbol{s}((\boldsymbol{\vartheta}_0^\top, \beta_0)^\top) = \\ \boldsymbol{s}((\boldsymbol{\vartheta}_0^\top, \beta_0)^\top)^\top \Sigma((\boldsymbol{\vartheta}_0^\top, \beta_0)^\top) \boldsymbol{s}((\boldsymbol{\vartheta}_0^\top, \beta_0)^\top)$$

for testing the simple hypothesis

$$H_0 : \boldsymbol{\vartheta} = \boldsymbol{\vartheta}_0, \beta = \beta_0.$$

This hypothesis is not interesting at all, because we are only interested in inference about β but not in testing $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}_0$. Thus, we are concerned with the composite hypothesis

$$H_0 : \beta = \beta_0$$

treating $\boldsymbol{\vartheta}$ as a nuisance parameter. (From a modelling point of view, the term *nuisance parameter* is misleading, because we should want to estimate full distributions, not just shift effects.) We estimate $\boldsymbol{\vartheta}$ under the constraint $\beta = \beta_0$ and obtain $\hat{\boldsymbol{\vartheta}}(\beta_0)$. By definition, the score function corresponding to $\boldsymbol{\vartheta}$ evaluated at $\hat{\boldsymbol{\vartheta}}(\beta_0))$ is zero: $\boldsymbol{s}(\hat{\boldsymbol{\vartheta}}(\beta_0))) = \boldsymbol{0}$. The quadratic score statistic evaluated at $\hat{\boldsymbol{\vartheta}}(\beta_0)$ and β_0 simplifies to

$$(\boldsymbol{s}(\hat{\boldsymbol{\vartheta}}(\beta_0))^\top, \boldsymbol{s}(\beta_0))^\top \Sigma((\hat{\boldsymbol{\vartheta}}(\beta_0))^\top, \beta_0)^\top)(\boldsymbol{s}(\hat{\boldsymbol{\vartheta}}(\beta_0))^\top, \boldsymbol{s}(\beta_0)) = \\ \Sigma_{22}((\hat{\boldsymbol{\vartheta}}(\beta_0))^\top, \beta_0)^\top) \boldsymbol{s}(\beta_0)^2 \xrightarrow{D} \chi^2_1$$

and thus we define the linear score statistic $\mathrm{R}(\beta_0)$ by

$$\Sigma_{22}((\hat{\boldsymbol{\vartheta}}(\beta_0)^\top, \beta_0)^\top)^{1/2}\boldsymbol{s}(\beta_0) \xrightarrow{D} \mathrm{N}(0, 1).$$

Computing $\mathrm{R}(0)$ for the shift model (5.1) under $\beta = \beta_0 = 0$ is equivalent to computing the maximum likelihood estimator for $\boldsymbol{\vartheta}$ in an unconditional model as discussed in Chapter 4. A simultaneous estimation of $\boldsymbol{\vartheta}$ and β is not necessary. However, the statistic requires in addition to compute the variance term $\Sigma_{22}((\hat{\boldsymbol{\vartheta}}(\beta_0)^\top, \beta_0)^\top)$, so one can not get away without at least an implementation of the score function and the observed Fisher information of model (5.1). The rejection regions and P-values regarding less, greater, and two-sided alternatives are the same as those of the Wald test.

The ancillary statistic $\mathrm{R}(\beta_0)$ can be used to construct $(1 - 2\alpha) \times 100\%$ score intervals $\{\beta_0 \in \mathbb{R} \mid \Phi^{-1}(\alpha) < \mathrm{R}(\beta_0) \leq \Phi^{-1}(1-\alpha)\}$ where the quantiles are such that $\mathbb{P}_\beta(\mathrm{R}(\beta) \leq \Phi^{-1}(\alpha)) \leq \alpha$. Unfortunately, solving $\mathrm{R}(\beta_0) = \Phi^{-1}(\alpha)$ and $\mathrm{R}(\beta_0) = \Phi^{-1}(1 - \alpha)$ is not possible analytically and numerical inversion has to be applied to construct the confidence interval. This also means that the restricted maximum likelihood estimator $\hat{\boldsymbol{\vartheta}}(\beta_0)$ must be computed for several possible values of β_0, so this procedure is more time consuming that the computation of a Wald interval. Score intervals are invariant with respect to reparameterisations and, unlike Wald intervals, not necessarily symmetric around the maximum likelihood estimator $\hat{\beta}$.

Example 5.18. For the model in Example 5.15 we obtain the restricted maximum likelihood estimator $\hat{\boldsymbol{\vartheta}}(0)$ and the corresponding covariance ma-

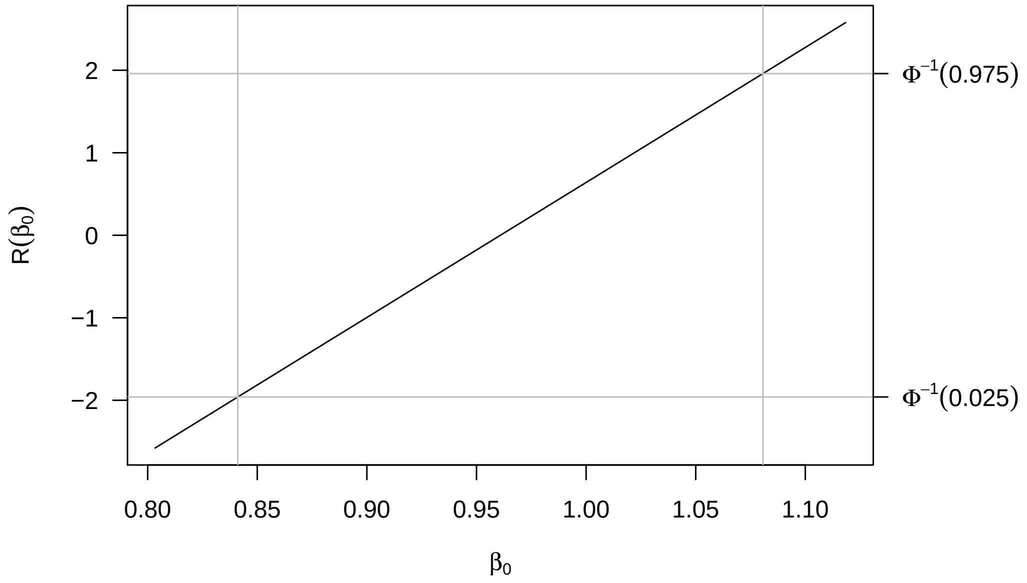


Figure 5.12: Numerical inversion computing 95% score intervals.

trix as

$$
\begin{aligned}
\hat{\boldsymbol{\vartheta}} &= (-2.682, 0.722, 1.488, 3.425, 3.987, 4.825, 5.663)^\top \\
\hat{\beta} &= 0.000 \\
\hat{\Sigma} &= \left(\begin{array}{rrrrrrr|r}
0.048 & -0.120 & 0.208 & -0.234 & 0.143 & -0.025 & -0.016 & -0.000 \\
-0.120 & 0.338 & -0.636 & 0.760 & -0.483 & 0.090 & 0.056 & 0.001 \\
0.208 & -0.636 & 1.339 & -1.721 & 1.169 & -0.223 & -0.136 & 0.005 \\
-0.234 & 0.760 & -1.721 & 2.486 & -1.849 & 0.397 & 0.245 & -0.001 \\
0.143 & -0.483 & 1.169 & -1.849 & 1.691 & -0.412 & -0.244 & 0.004 \\
-0.025 & 0.090 & -0.223 & 0.397 & -0.412 & 0.303 & 0.198 & 0.002 \\
-0.016 & 0.056 & -0.136 & 0.245 & -0.244 & 0.198 & 0.221 & 0.002 \\
\hline
-0.000 & 0.001 & 0.005 & -0.001 & 0.004 & 0.002 & 0.002 & 0.004
\end{array}\right)
\end{aligned}
$$

The score function for the shift parameter is $\boldsymbol{s}(0) = -251.137$ and, multiplied with the standard error 0.065, gives $\mathrm{R}(0) = -16.204$. The 95% score interval for β is $(0.841, 1.081)$ and for $\exp(\beta)$ we get $(2.319, 2.947)$. The numerical inversion of the score statistic $\mathrm{R}(\beta_0)$ is illustrated in Figure 5.12.

□

Likelihood Tests and Intervals

The Wald statistic is based on the unrestricted maximum likelihood estimator $(\hat{\boldsymbol{\vartheta}}, \hat{\beta})$ and needs an estimate of the variance of $\hat{\beta}$ through the inverse observed Fisher information. The score statistic requires the restricted maximum likelihood estimator $\hat{\boldsymbol{\vartheta}}(\beta)$ and also the inverse of the observed Fisher information. Computing the likelihood ratio statistic needs both the unrestricted and the restricted maximum likelihood estimators but not the observed Fisher information.

The likelihood ratio test rejects the null in favour of a two-sided alternative when $\mathrm{LR}(\beta_0) > q(1-\alpha)$, where $q(\alpha)$ is the α quantile of the χ^2_1 distribution with $\mathbb{P}(\chi^2_1 \le q(\alpha)) \le \alpha$ for all $0 < \alpha < 1$. The P-value is $1 - \mathbb{P}(\chi^2_1 \le \mathrm{LR}(\beta_0))$.

Likelihood confidence intervals at level $(1-2\alpha) \times 100\%$ are defined by $\{\beta_0 \in \mathbb{R} \mid \mathrm{LR}(\beta_0) \le q(1-2\alpha)\}$, with $q(1-2\alpha)$ denoting the $1-2\alpha$ quantile of the χ^2_1 distribution. The two roots of $\mathrm{LR}(\beta) = q(1-2\alpha)$ define the confidence interval and have to be computed by numerical inversion.

Example 5.19. In Example 5.15, the restricted maximum likelihood estimator $\hat{\boldsymbol{\vartheta}}(0)$ maximises the log-likelihood at -4363.890, the unconstraint estimator $(\hat{\boldsymbol{\vartheta}}^\top, \hat{\beta}))^\top$ results in a log-likelihood of -4241.103; the likelihood ratio statistic is minus two times the difference of these log-likelihoods, resulting in $\mathrm{LR}(0) = 245.574$. The numerical inversion method for computing a 95% likelihood confidence interval is illustrated in Figure 5.13. The confidence interval for β is $(0.841, 1.081)$ and for $\exp(\beta)$ we obtain $(2.319, 2.947)$.

□

Like score intervals, likelihood intervals are potentially asymmetric, invariant with respect to reparameterisations, and computationally more intensive than Wald tests because the restricted maximum likelihood estimator $\hat{\boldsymbol{\vartheta}}(\beta)$ needs to be computed for several candidate values of β.

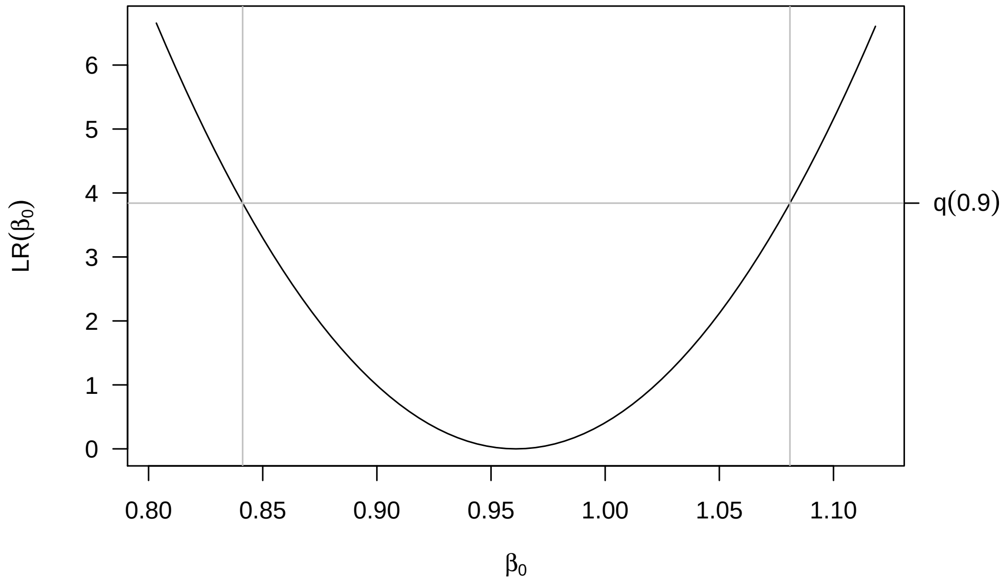


Figure 5.13: Numerical inversion 95% likelihood confidence interval.

Example 5.20. For the equivalence hypothesis considered in Example 5.17 we compare Wald, score, and likelihood intervals for the standardised differences of means of normal variables Y_M and Y_C. In addition, we consider more flexible transformation models

$$\begin{aligned} \mathbb{P}(Y_\text{M} \le y) &= \Phi\left(\boldsymbol{a}(y)^\top \boldsymbol{\vartheta}\right) \\ \mathbb{P}(Y_\text{C} \le y) &= \Phi\left(\boldsymbol{a}(y)^\top \boldsymbol{\vartheta} - \beta\right) \end{aligned}$$

such that $\boldsymbol{a}(Y_\text{M})^\top \boldsymbol{\vartheta} \sim \text{N}(0, 1)$ and $\boldsymbol{a}(Y_\text{C})^\top \boldsymbol{\vartheta} \sim \text{N}(\beta, 1)$. The transformation function $\boldsymbol{a}(y)^\top \boldsymbol{\vartheta}$ is in Bernstein form with order $M = 6$. The shift parameter $\beta = \mathbb{E}(\boldsymbol{a}(Y_\text{M})^\top \boldsymbol{\vartheta}) - \mathbb{E}(\boldsymbol{a}(Y_\text{C})^\top \boldsymbol{\vartheta})$ is a standardised difference of means on the transformed scale. Unlike the model in Example 5.17, normality of Y_M and Y_C is not assumed.

The different 95% confidence intervals are given in Table 5.1. For the normal model, the discrepancies between Wald, score, and likelihood inter-

	2.5 %	97.5 %
Normal Wald	-0.34767	1.27395
Normal Score	-0.34767	1.27395
Normal Likelihood	-0.34750	1.27411
Non-normal Wald	-0.42813	1.27220
Non-normal Score	-0.35839	1.21078
Non-normal Likelihood	-0.39027	1.23872

Table 5.1:

vals are neglegible. For the more complex non-normal model, the different procedures lead to substantially different intervals.

□

5.6 Connections to Rank Tests

To derive the score statistic $S(0)$ for model (5.2), consider a sample of N independent observations $(y_i, x_i), i = 1, \dots, N$ with $P \leq N$ distinct ordered outcomes $\upsilon_1 < \dots < \upsilon_P$ in $y_1, \dots, y_N$. The ith observation contributes

$$\ell_i(\beta, \boldsymbol{\vartheta}) = \log(F_Z(h(y_i \mid \boldsymbol{\vartheta}) - \beta x_i) - F_Z(h(y_i- \mid \boldsymbol{\vartheta}) - \beta x_i)) \qquad (5.4)$$

to the nonparametric log-likelihood function of model (5.2). The transformation function h is parameterised nonparametrically through parameters $\boldsymbol{\vartheta} \in \mathbb{R}^{P-1}$, more specifically through $\vartheta_0 = -\infty < \vartheta_1 < \dots < \vartheta_{P-1} < \vartheta_P = \infty$, as

$$h(\upsilon_p \mid \boldsymbol{\vartheta}) = \vartheta_p \text{ and } h(\upsilon_p- \mid \boldsymbol{\vartheta}) = \vartheta_{p-1},$$

i.e. one parameter is assigned to each but the largest of the distinct outcomes.

Under $H_0 : \beta = 0$, we assign upranks $R_1, \dots, R_N$ to $y_1, \dots, y_N$, with R_i being the number of observations not larger than y_i. The marginal distribution function is estimated by the nonparametric maximum likelihood estimator $\hat{\mathbb{P}}(Y \leq y_i) = {}^{R_i}/_{N}$, *i.e.* we apply the empirical cumulative distribution function. We obtain $h(y_i \mid \hat{\boldsymbol{\vartheta}}(0)) = F_Z^{-1}({}^{R_i}/_{N})$ and the score statistic for the nonparametric log-likelihood is then

$$\begin{aligned}
S(0) &= \left.\frac{\partial \log(F_Z(h(y_i \mid \hat{\boldsymbol{\vartheta}}(0)) - \beta x_i) - F_Z(h(y_i - \mid \hat{\boldsymbol{\vartheta}}(0)) - \beta x_i))}{\partial \beta}\right|_{\beta=0} \\
&= \sum_{i=1}^{N} \frac{F_Z'(h(y_i \mid \hat{\boldsymbol{\vartheta}}(0)) - 0x_i) - F_Z'(h(y_i - \mid \hat{\boldsymbol{\vartheta}}(0)) - 0x_i)}{F_Z(h(y_i \mid \hat{\boldsymbol{\vartheta}}(0)) - 0x_i) - F_Z(h(y_i - \mid \hat{\boldsymbol{\vartheta}}(0)) - 0x_i)} x_i.
\end{aligned}$$

Normal Scores Test

$F_Z(z) = \Phi$, $F_Z^{-1} = \Phi^{-1}$ and $f_Z(z) = F_Z'(z) = \phi(z)$.

Savage or Log-rank Test

$F_Z'(z) = \exp(z - \exp(z))$

Wilcoxon-Rank-Sum Test

We have $F_Z = \text{expit}$ and $F_Z^{-1} = \text{logit}$. With $F_Z'(z) = F_Z(z)(1 - F_Z(z))$, $F_Z(h(y_i \mid \hat{\boldsymbol{\vartheta}}(0)) - 0x_i) = {}^{R_i}/_N$ and $F_Z(h(y_i - \mid \hat{\boldsymbol{\vartheta}}(0)) - 0x_i) = {}^{R_i - t_i}/_N$ (t_i being the number of ties equal to y_i) this simplifies to

$$S(0) = \sum_{i=1}^{N} \frac{N - 2R_i + t_i}{N} x_i.$$

For the two-sample case ($x_i \in \{0,1\}$) with $N_1 := \sum_{i=1}^N x_i$ and $N_0 := N - N_1$, we obtain

$$\begin{aligned}
\frac{N}{2}S(0) &= \frac{N}{2}\sum_{i:x_i=1}\left(1-\frac{2}{N}\left(R_i-\frac{t_i}{2}\right)\right) \\
&= \frac{N}{2}\left(N_1-\frac{2}{N_0+N_1}\sum_{i:x_i=1}\left(R_i-\frac{t_i-1}{2}-\frac{1}{2}\right)\right) \\
&= \frac{N_1}{2}+\frac{N_1(N_0+N_1)}{2}-\sum_{i:x_i=1}\left(R_i-\frac{t_i-1}{2}\right) \\
&= \underbrace{\frac{N_0N_1}{2}+\underbrace{\frac{N_1(N_1+1)}{2}-\underbrace{\sum_{i:x_i=1}\left(R_i-\frac{t_i-1}{2}\right)}_{=W}}_{=-U}}_{=-(U-\mathbb{E}(U))}
\end{aligned}$$

where W is the Wilcoxon statistic, *i.e.* the sum of the mid-ranks $R_i - (t_i-1)/2$ in the sample with $x_i = 1$, and $U = W - N_1(N_1+1)/2$ is the corresponding Mann-Whitney U-statistic with $\mathbb{E}(U) = N_0N_1/2$ under H_0.

5.7 Assessment of Model Assumptions

We already discussed that we can check the model assumption, namely the presence of a constant shift effect β, by simply comparing the two distribution functions on the scale of the link function F_Z^{-1}. Even more convenient is the ability to model potential deviations from this model assumption.

We can always write the two conditional distribution functions F_0 and F_1 in terms of their transformation functions $F_0(y) = F_Z(h_0(y))$ and $F_1(y) = F_Z(h_1(y))$ by application of the principles discussed for unconditional distributions in Chapter 3. Doing so puts us back in the distribution- and model-free world because, with flexible enough transformation functions h_0 and h_1, we can describe all pairs of distributions in such a way.

These two separate models can be cast into one model, still enjoying the distribution- and model-free properties, by writing

$$\begin{aligned}
\mathbb{P}(Y \le y \mid X = x) &= F_Z(h_0(y)(1-x) + h_1(y)x) \\
&= F_Z(h_0(y) + (h_1(y) - h_0(y))x) \\
&= F_Z(h_0(y) + \beta(y)x).
\end{aligned}$$

This looks like a shift-transformation model (5.2), however, with a potentially non-constant shift parameter $\beta(y)$. The function $\beta(y) = (h_1(y) - h_0(y)) = F_Z^{-1}(F_1(y)) - F_Z^{-1}(F_0(y))$ describes the difference of the two conditional distribution functions on the scale of the link function F_Z^{-1}. It is of utmost importance to note that the conditional distribution function $\mathbb{P}(Y \le y \mid X = x)$ is invariate with respect to the choice of F_Z: Any choice of F_Z will result in the same conditional distribution function, but of course in different functions $h_0(y)$ and $\beta(y)$.

We apply the same parameterisation as before, but to both functions characterising the model:

$$\mathbb{P}(Y \le y \mid X = x) = F_Z\left(\boldsymbol{a}(y)^\top \boldsymbol{\vartheta}_0 + \boldsymbol{a}(y)^\top \boldsymbol{\vartheta}_1 x\right).$$

The constraints on the parameters in this model are a bit more complex, because we have to ensure that $\boldsymbol{a}(y)^\top \boldsymbol{\vartheta}_0$ and $\boldsymbol{a}(y)^\top \boldsymbol{\vartheta}_0 + \boldsymbol{a}(y)^\top \boldsymbol{\vartheta}_1$ are monotone nondescreasing, however, $\beta(y) = \boldsymbol{a}(y)^\top \boldsymbol{\vartheta}_1$ is not subject to a monotonicity constraint.

The big advantage of being able to fit such a model is that we can construct a confidence band around $\hat{\beta}(y) = \boldsymbol{a}(y)^\top \hat{\boldsymbol{\vartheta}}_1$ which allows to investigate, under appropriate error control of the family-wise error rate, for which outcomes y the two distribution functions are different.

Example 5.21. The difference of the empirical cumulative distribution functions for measured blood loss (plotted in Figure 5.4) is shown in Figure 5.14.

The confidence band excludes zero for measured blood losses of less than 745 and we see higher probabilities for these measured blood losses for vaginal deliveries. This means that more women giving birth the natural way

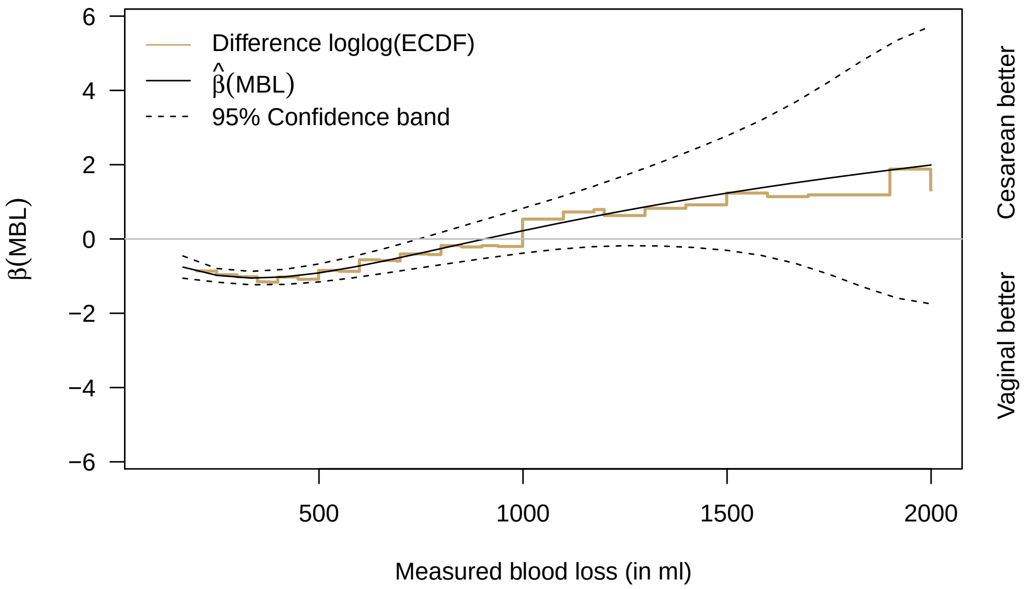


Figure 5.14: Measured blood loss. Difference of empirical cumulative distribution functions (ECDFs) of untransformed measured blood loss on the log-log scale for $N_{\text{CS}} = 632$ Cesarean section minus $N_{\text{VD}} = 677$ vaginal deliveries delivered at the University Hospital Zurich, Switzerland. The two curves were estimated separately for the two groups. The same difference was estimated by a transformation model (black line) with corresponding 95% confidence band.

suffer blood losses of at most 745. For larger cut-off points, the uncertainty is too large to make any definitive statements, but we see larger probabilities of suffering less than a certain about of blood larger than 745 under Cesarean sections. This may mean that blood loss management was better and maybe quicker in the OR compared to delivery rooms, but we are entering the area of wild speculation here. In any case, the picture we see is much more informative than the common difference of means (which is, by the way, 109.632 ml with 95% confidence interval $(76.0, 143.3)$ as obtained from Welch's test) .

□

Chapter 6

Linear Transformation Models

6.1 Simple Regression

The simplest extension of the transformation model (5.2)

$$\mathbb{P}(Y \le y \mid X = x) = F_Z(h(y) - \beta x)$$

defined for the group indicator $x \in \{0, 1\}$ is to allow a conditioning explanatory variable $x \in \mathbb{R}$. Regression models conditioning on a single explanatory variable x, and thus featuring a single regression coefficient β, are often referred to as a “simple” model. Changes in the variable x induce changes in the conditional distribution of Y unless $\beta = 0$. Such a variable is also called independent variable (and the response is the dependent variable), the terms exogenous variable for x and endogenous variable for Y are often used in economics, x is sometimes called the design variable, prognostic variable, or predictor variable. These different terminologies all reflect that idea that changes in x cause changes in Y. We are entering muddy waters here, because it is one thing to write down such a model but a completely different thing to infer such a causal effect from data. For the moment, we don’t bother with these difficulties in real life and try to understand the meaning and interpretation of such models before commenting on estimation problems.

A one unit-increase in x, that is a move from the conditional distribution of Y given x to the conditional distribution of Y given $x + 1$, causes the conditional transformation function to decrease by β:

$$h(y) - \beta(x+1) = h(y) - \beta x - \beta.$$

This interpretation also holds for the two-sample situation, where we move from $x = 0$ to $x = 1$ when switching from the first to the second group. The meaning of the shift β on other scales of course depends on F_Z: With $F_Z = \text{expit}$, for example, we can say that a one unit increase in x reduces the log-odds function by β or, equivalently, leads to an $\exp(-\beta)$-fold multiplicative change in the odds function. Regression models allowing a linear interpretation of regression coefficients β are called "linear regression models". Such models are very popular because the nature of the impact of x on Y can be understood and communicated by statistically literate professionals.

Example 6.1. For a binary response $Y \in \{y_1, y_2\}$, the conditional probability of observing y_1 given some explanatory variable $x \in \mathbb{R}$ can be written as

$$\pi_1(x) = \mathbb{P}(Y \le y_1 \mid X = x) = F_Z(\vartheta_1 - \beta x).$$

This simple binary logistic regression model with $F_Z = \text{expit}$ can also be written as a multiplicative model for the conditional odds function

$$O_{Y\mid \boldsymbol{X}=\boldsymbol{x}}(y_1 \mid x) = \frac{\pi_1(x)}{1-\pi_1(x)} = \exp(\vartheta_1 - \beta x)$$

which allows a simple comparison of the two conditional distributions $\mathbb{P}(Y \le y_1 \mid X = x)$ and $\mathbb{P}(Y \le y_1 \mid X = x+1)$ in terms of their odds ratio

$$\frac{O_{Y\mid \boldsymbol{X}=\boldsymbol{x}}(y_1 \mid x)}{O_{Y\mid \boldsymbol{X}=\boldsymbol{x}}(y_1 \mid x+1)} = \frac{\frac{\pi_1(x)}{1-\pi_1(x)}}{\frac{\pi_1(x+1)}{1-\pi_1(x+1)}} = \frac{\exp(\vartheta_1 - \beta x)}{\exp(\vartheta_1 - \beta(x+1))} = \exp(\beta)$$

or

$$\frac{O_{Y\mid \boldsymbol{X}=\boldsymbol{x}}(y_1 \mid x+1)}{O_{Y\mid \boldsymbol{X}=\boldsymbol{x}}(y_1 \mid x)} = \frac{\frac{\pi_1(x+1)}{1-\pi_1(x+1)}}{\frac{\pi_1(x)}{1-\pi_1(x)}} = \exp(-\beta).$$

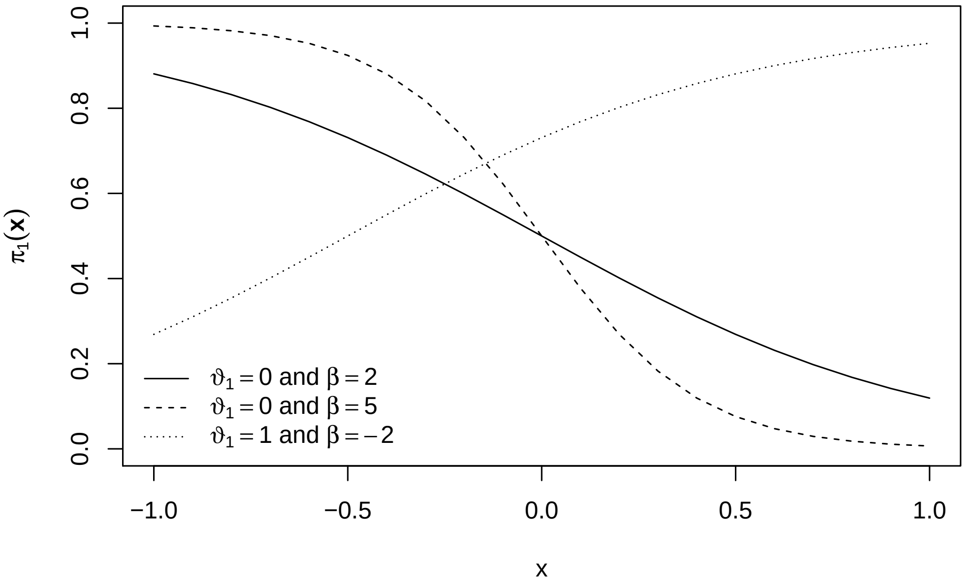


Figure 6.1: Simple binary logistic regression model: Dependency of the conditional probability $\pi_1(x) = \mathbb{P}(Y \leq y_1 \mid X = x)$ on the explanatory variable x for several choices of intercept ϑ_1 and regression coefficient β.

The conditional probability $\pi_1(x)$ is plotted as a function of x for three choices of the intercept ϑ_1 and log-odds ratio β in Figure 6.1. Plotted on the scale of the log-odds function $\log(O_{Y|\boldsymbol{X}=\boldsymbol{x}}(y_1 \mid x))$, all three functions $\vartheta_1 - \beta x$ are linear functions of x.

□

Example 6.2. For an ordered categorical response $Y \in \{y_1, y_2, \ldots, y_K\}$, the conditional distribution function given an explanatory variable $x \in \mathbb{R}$ can be formulated in terms of a cumulative model of the form

$$\pi_k(x) = \mathbb{P}(Y \leq y_k \mid X = x) = F_Z(\vartheta_k - \beta x).$$

The conditional odds function in the proportional-odds logistic regression

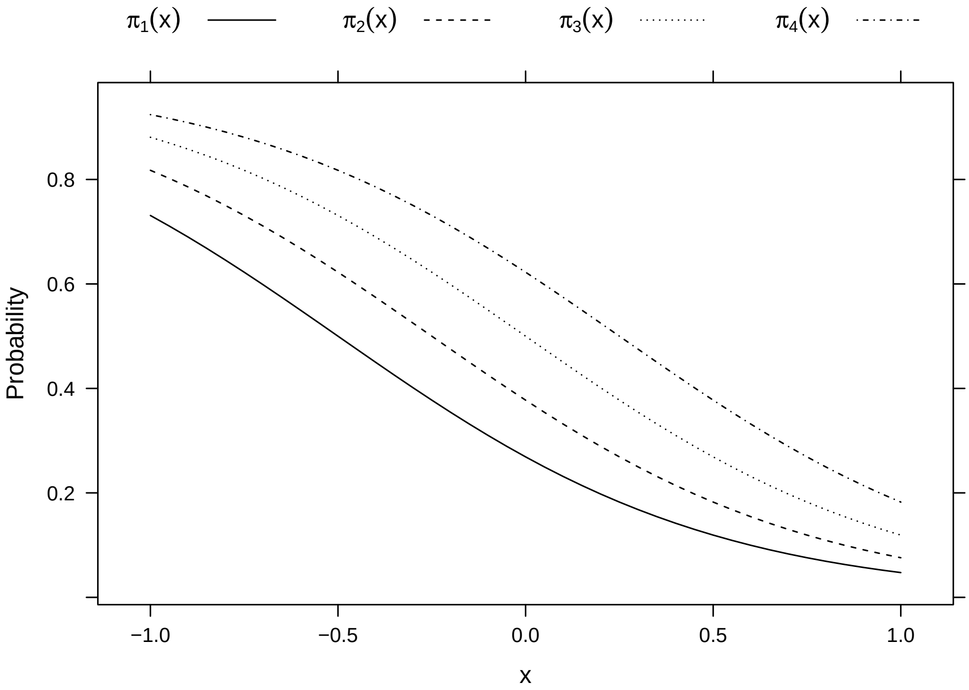


Figure 6.2: Proportional-odds logistic regression: Dependency of the conditional probabilities $\pi_k(x) = \mathbb{P}(Y \le y_k \mid X = x)$ on the explanatory variable x for intercepts $\vartheta_1 = -2, \vartheta_2 = -1, \vartheta_3 = 0, \vartheta_4 = 1$ and regression coefficient $\beta = 2$.

model defined by $F_Z = \text{expit}$ is

$$O_{Y|\boldsymbol{X}=\boldsymbol{x}}(y_k \mid x) = \frac{\pi_k(x)}{1-\pi_k(x)} = \exp(\vartheta_k - \beta x) = \exp(\vartheta_k)\exp(-\beta x)$$

and thus an increase of x by one unit leads to the common odds ratio

$$\frac{\frac{\pi_k(x)}{1-\pi_k(x)}}{\frac{\pi_k(x+1)}{1-\pi_k(x+1)}} = \frac{\exp(\vartheta_k - \beta x)}{\exp(\vartheta_k - \beta(x+1))} = \exp(\beta) \quad \forall k = 1, \dots, K-1.$$

For a log-odds ratio of $\beta = 2$, the probabilities $\pi_k(x)$ are plotted as a function of x in Figure 6.2. Presented on the log-odds scale, these four lines would be parallel, with a vertical distance β.

□

Example 6.3. For an absolute continuous response $Y \in \mathbb{R}$, a conditional normal distribution modelling the impact of an explanatory variable $x \in \mathbb{R}$ on the distribution of Y can be formulated in terms of the transformation model

$$\mathbb{P}(Y \le y \mid X = x) = \Phi(h(y) - \beta x) = \Phi(\xi y - \alpha - \beta x), \quad \xi > 0$$

which is another way of writing $Y \mid X = x \sim \mathrm{N}\left(\xi^{-1}(\alpha + \beta x), \xi^{-2}\right)$. The distribution function, for some fixed values of y, is presented as a function of x in Figure 6.3. The distribution is conceptually similar to the distributions presented in Figures 6.1 and 6.2 but, as a result of using $F_Z = \Phi \neq \text{expit}$ the log-odds function for the normal distribution will be a nonlinear function of x. Linearity in the normal model is given on the scale of the quantile function (for some $0 < \alpha < 1$)

$$Q(\alpha \mid x) = \xi^{-1}(\alpha + \beta x) + \xi^{-1}\Phi^{-1}(\alpha)$$

as shown in Figure 6.4. The vertial distance is constant for varying values of x because of the variance homogeneity in this model, that is, the residual standard deviation ξ does not depend on x. The slope of all conditional quantile functions is $\xi^{-1}\beta$. The quantile functions, and thus also the conditional mean in this symmetric distribution, increase with increasing values of x, because the shift term βx enters with negative sign in the transformation $h(y) - \beta x$.

□

Example 6.4. For an absolute continuous positive response $Y \in \mathbb{R}^+$ we model the conditional distribution given $x \in \mathbb{R}$ using a Weibull regression model, that is a transformation model with log-linear transformation function $h(y)$ and complementary log-log link:

$$\mathbb{P}(Y \le y \mid X = X) = 1 - \exp(-\exp(\xi \log(y) - \alpha - \beta x)), \quad \xi > 0$$

or, in other words, $Y \mid X = x \sim \mathrm{W}(\exp(\alpha + \beta x), \xi)$. Through the connection

$$\begin{aligned} \mathbb{P}(Y > y \mid X = x) &= \exp(-\exp(\xi \log(y) - \alpha - \beta x)) \\ &= \exp(-\Lambda_{Y \mid \boldsymbol{X} = \boldsymbol{x}}(y \mid x)) \end{aligned}$$

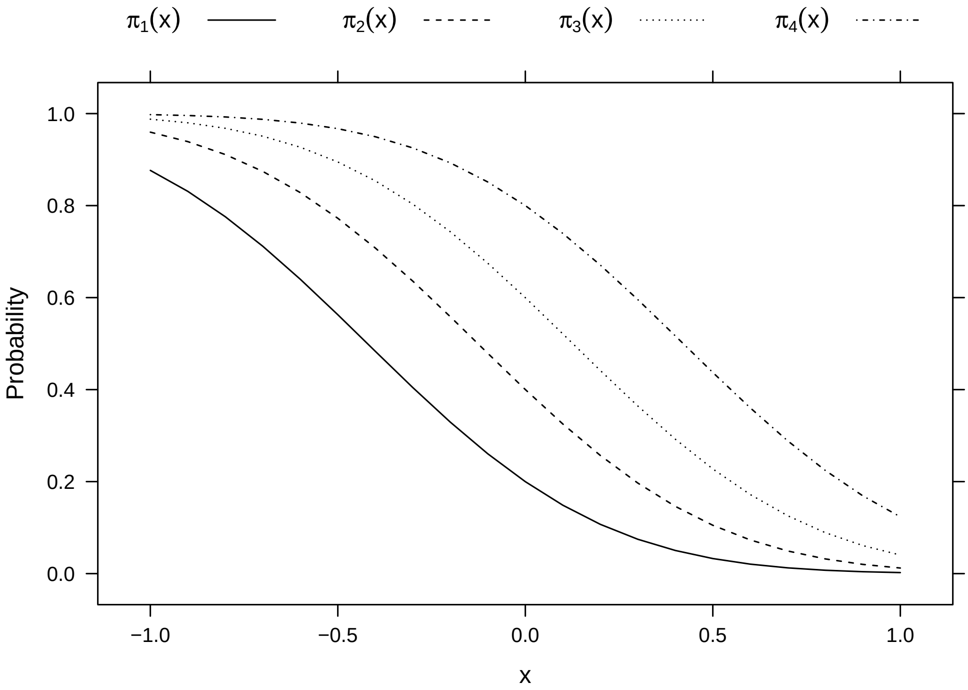


Figure 6.3: Normal linear regression model: Probabilities $\pi_1(x) = \mathbb{P}(Y \leq -0.842 \mid X = x)$, $\pi_2(x) = \mathbb{P}(Y \leq -0.253 \mid X = x)$, $\pi_3(x) = \mathbb{P}(Y \leq 0.253 \mid X = x)$, and $\pi_4(x) = \mathbb{P}(Y \leq 0.842 \mid X = x)$ as a function of x, for parameters $\xi = 1, \alpha = 0$, and $\beta = 2$.

we see that the conditional cumulative hazard function is

$$\Lambda_{Y|\boldsymbol{X}=\boldsymbol{x}}(y \mid x) = \exp(\xi \log(y) - \alpha - \beta x) = \exp(\xi \log(y) - \alpha)\exp(-\beta x).$$

The hazard function is therefore

$$\lambda_{Y|\boldsymbol{X}=\boldsymbol{x}}(y \mid x) = \Lambda'_{Y|\boldsymbol{X}=\boldsymbol{x}}(y \mid x) = \exp(\xi \log(y) - \alpha)\xi y^{-1}\exp(-\beta x).$$

We can compare the conditional distributions $\mathbb{P}(Y \leq y \mid X = x)$ and $\mathbb{P}(Y \leq y \mid X = x + 1)$ by their hazard ratio

$$\frac{\lambda_{Y|\boldsymbol{X}=\boldsymbol{x}}(y \mid x)}{\lambda_{Y|\boldsymbol{X}=\boldsymbol{x}}(y \mid x + 1)} = \exp(\beta), \quad \forall y \in \mathbb{R}^+.$$

This property is known as the proportional-hazards assumption. The distribution and quantile functions are presented for varying values of x in

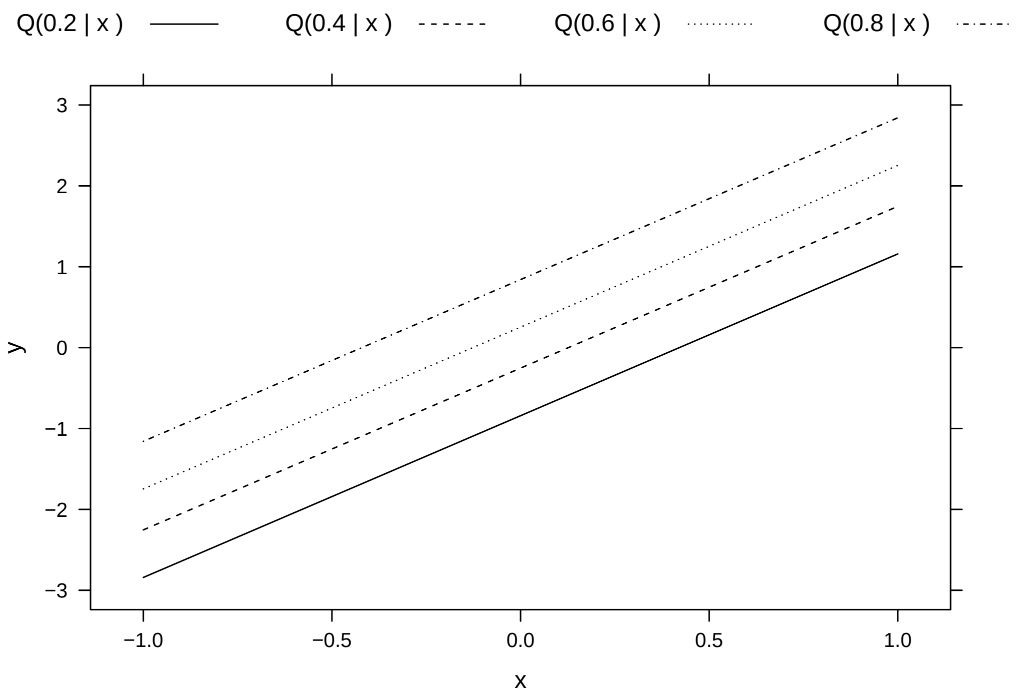


Figure 6.4: Normal linear regression model: Conditional quantile functions for $\alpha = .2, .4, .6, .8$ as a function of x.

Figures 6.5 and 6.6. It is important to note that the variability of the response increases with increasing values of x.

□

The rational behind this choice of examples was to demonstrate that seemingly unrelated models, such as a binary logistic, normal, and Weibull model, can be motivated from the unifying transformation model point of view. The most important difference is the interpretation of the regression coefficient β as log-odds ratio, log-hazard ratio, or standardised slope. All models can be used to compute conditional distribution functions and all other functions commonly used for characterising distributions, as discussed in Chapter 3. Which scale the model is best presented and communicated with is context specific: Hazard functions are often used in survival analysis (Chapter 11), quantile functions are popular in economics and financial

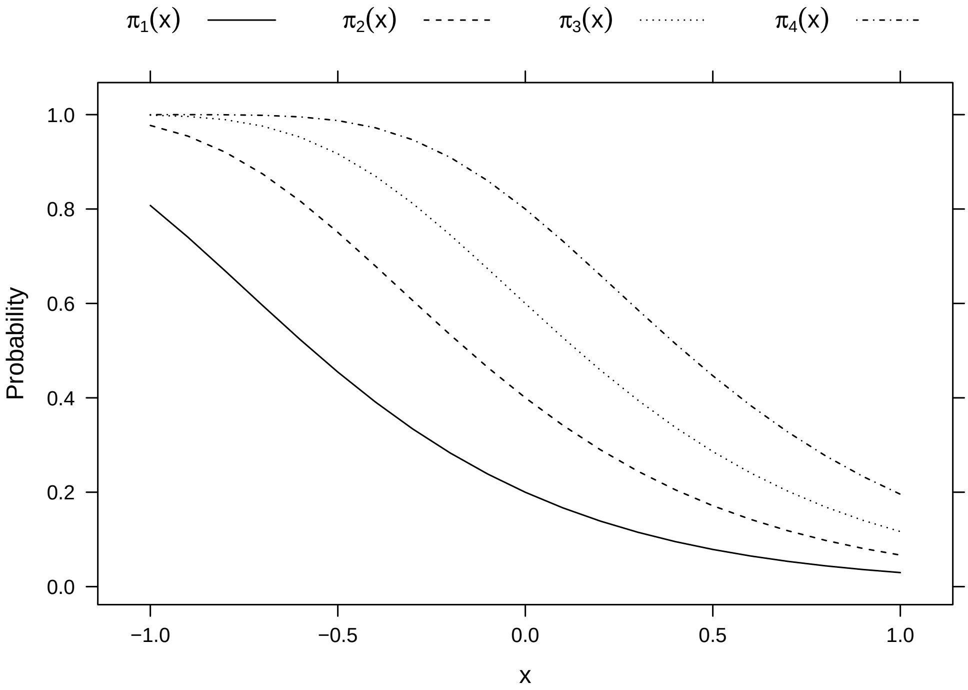


Figure 6.5: Weibull regression model: Probabilities $\pi_1(x) = \mathbb{P}(Y \le 0.223 \mid X = x)$, $\pi_2(x) = \mathbb{P}(Y \le 0.511 \mid X = x)$, $\pi_3(x) = \mathbb{P}(Y \le 0.916 \mid X = x)$, and $\pi_4(x) = \mathbb{P}(Y \le 1.609 \mid X = x)$ as a function of x, for parameters $\xi = 1$ and $\beta = 2$. .

applications, and odds functions have a direct application in some medical fields.

Example 6.5. Figure 6.7 shows a scatterplot of $N = 1309$ neonatal weights and measured maternal blood losses, overlayed with the conditional median as well as 10 and 90% quantiles estimated by the proportional-odds transformation model

$$\mathbb{P}(Y \le y \mid X = x) = \text{expit}(h(y) - \beta \times (x - 3500)/100)$$

where the explanatory variable x is the weight of the baby. For a baby born at some "normal" 3500 g, the conditional distribution of measured blood loss is given by the simple expression

$$\mathbb{P}(Y \le y \mid X = 3500) = \text{expit}(h(y) - 0).$$

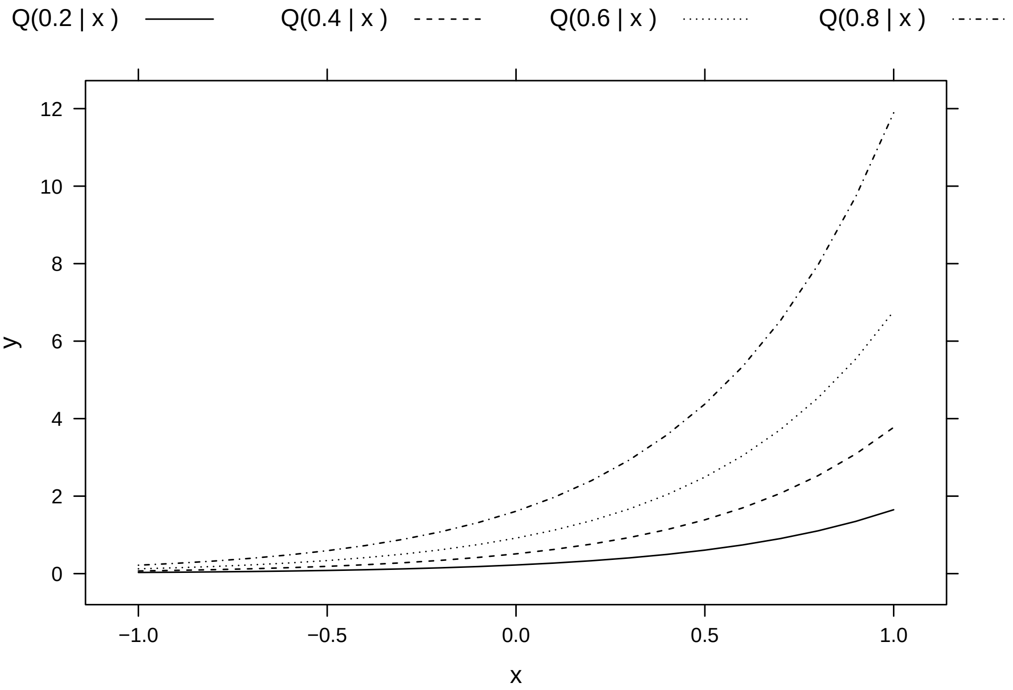


Figure 6.6: Weibull regression model: Conditional quantile functions for $\alpha = .2, .4, .6, .8$ as a function of x. .

We obtain $\hat{\beta} = 0.00656$. This number means that an additional 100 g of birth weight changes the conditional distribution to

$$\mathbb{P}(Y \leq y \mid X = 3600) = \text{expit}(h(y) - 0.00656).$$

This conditional distribution for a baby with 3600 g is stochastically larger than the conditional distribution for a baby with 3500 g, and thus the measured blood loss is stochastically larger for mothers of heavier babies. The conditional log-odds function of measured blood loss given a 3600 g baby is the conditional log-odds function given a 3500 g baby minus 0.00656. The conditional odds function given a 3600 g baby is 0.99346 times the conditional odds function given a 3500 g baby. The odds ratio of any event $Y \leq y$, say measured blood loss less than 500, or 750, or 1000 ml, comparing mothers of 3500 to 3600 g heavy babies, is thus 0.99346. Physicians often operate with cut-off points instead of the underlying continuous distributions, and are used to assess effects on the odds-ratio scale for binary responses, so this model extends the binary to the continuous case in a natural way.

Using a transformation function $\boldsymbol{a}(y)^\top \boldsymbol{\vartheta}$ in Bernstein form, the joint maximum likelihood estimates of $\boldsymbol{\vartheta}$ and the log-odds ratio β and their covariance matrix are

$$\begin{aligned}
\hat{\boldsymbol{\vartheta}} &= (-6.542, 2.087, 2.087, 2.087, 4.388, 5.099, 5.810)^\top \\
\hat{\beta} &= 0.007 \\
\hat{\Sigma} &= \left(\begin{array}{rrrrrrr|r}
0.221 & -0.412 & 0.587 & -0.560 & 0.314 & -0.040 & -0.023 & -0.000 \\
-0.412 & 0.937 & -1.480 & 1.533 & -0.905 & 0.128 & 0.078 & -0.000 \\
0.587 & -1.480 & 2.603 & -2.941 & 1.866 & -0.287 & -0.179 & -0.000 \\
-0.560 & 1.533 & -2.941 & 3.734 & -2.601 & 0.477 & 0.311 & -0.000 \\
0.314 & -0.905 & 1.866 & -2.601 & 2.172 & -0.500 & -0.334 & -0.000 \\
-0.040 & 0.128 & -0.287 & 0.477 & -0.500 & 0.358 & 0.268 & -0.000 \\
-0.023 & 0.078 & -0.179 & 0.311 & -0.334 & 0.268 & 0.272 & -0.000 \\
\hline
-0.000 & -0.000 & -0.000 & -0.000 & -0.000 & -0.000 & -0.000 & 0.000
\end{array}\right)
\end{aligned}$$

From these estimated model parameters we can obtain the conditional quantiles as a function of neonatal weight x by numerical inversion. Given $0 < \alpha < 1$ and x, solving the equation

$$\alpha = \hat{\mathbb{P}}(Y \le \hat{Q}(\alpha \mid x) \mid X = x) = \operatorname{expit}(\boldsymbol{a}(\hat{Q}(\alpha \mid x))^\top \hat{\boldsymbol{\vartheta}} - \hat{\beta} \times (x - 3500)/100)$$

gives us the estimated conditional $\alpha \times 100\%$ quantile $\hat{Q}(\alpha \mid x)$ and we plot these conditional quantiles for $\alpha = .1, .5, .9$ as a function of neonatal weight x in Figure 6.7.

The median measured blood loss for a baby born at 500 g is 393.6 ml and increases to a median measured blood loss of 427.9 ml for a 5000 g baby. The variability of this asymmetric distribution (the vertical difference between the conditional 10% quantile and the conditional median is smaller than the vertical difference between the median and the conditional 90% quantile) increases with heavier babies. The small magnitude of the effect of increasing birth weight is, however, a bit surprising. One would have

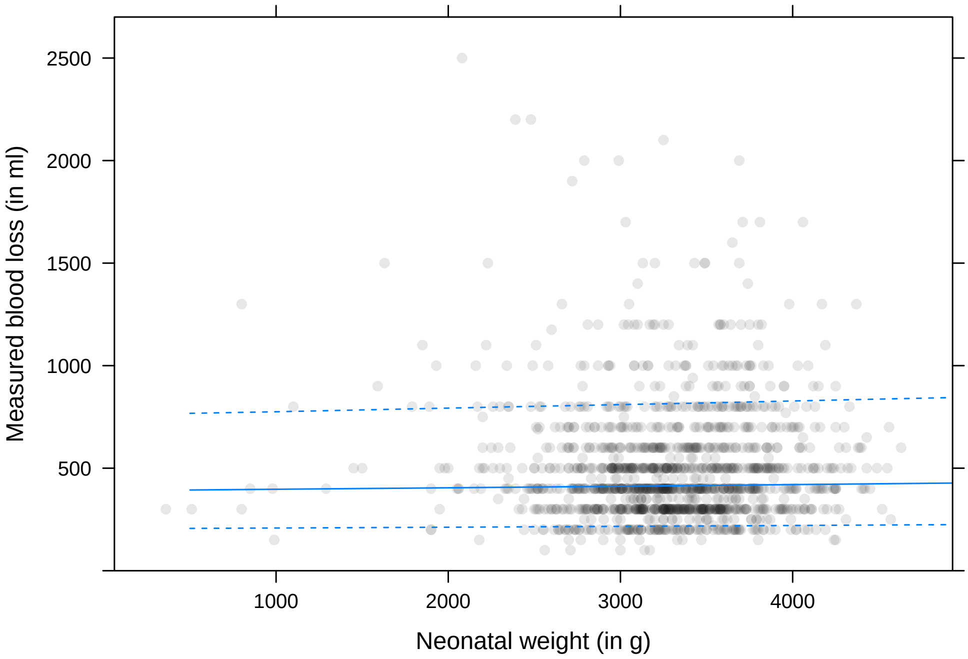


Figure 6.7: Measured blood loss. Scatterplot of neonatal weight and measured blood loss, overlayed with conditional median (solid line) and conditional 10% and 90% quantiles obtained from a simple proportional-odds model explaining measured blood loss by neonatal weight.

expected a more pronounced association of birth weight and measured blood loss.

In the more formal NHST framework, the 95% likelihood confidence interval for β is $(-0.01179, 0.02488)$, and thus does not provide any evidence against the null hypothesis $H_0 : \beta = 0$ of independence of measured blood loss and birth weight. Later examples will demonstrate that this counter-intuitive result is due to our ignorance towards other important explanatory variables in this too simple model. □

The proportional-odds model discussed in Example 6.5 is surely not the default model most data analysts would prefer analysing the data shown in Figure 6.7. Instead, a normal linear regression model assuming a conditional normal distribution would raise fewer eyebrows and thus we discuss this option in the next example.

Example 6.6. The simple normal linear regression model for measured blood loss given neonatal weight written as a model for a conditional distribution function reads

$$\mathbb{P}(Y \le y \mid X = x) = \Phi(\vartheta_2 y - \vartheta_1 - \beta \times (x - 3500)/100)$$

such that $Y \mid X = x \sim \mathrm{N}(\vartheta_2^{-1}(\vartheta_1 + \beta \times (x - 3500)/100), \vartheta_2^{-2})$. It follows that

$$\mathbb{E}(Y \mid X = x) = \vartheta_2^{-1}\vartheta_1 + \vartheta_2^{-1}\beta \times (x - 3500)/100 = \mu + \gamma \times (x - 3500)/100.$$

The mean measured blood loss when giving birth to a 3500 g baby is μ. An additional 100 g birth weight induce a γ ml increase in mean measured blood loss in this model and this simple and straightforward interpretation of this classical parameterisation explains much of the popularity of this model.

Using a linear transformation function $\boldsymbol{a}(y)^\top \boldsymbol{\vartheta} = (y, 1)\boldsymbol{\vartheta}$ the maximum likelihood estimates of $\boldsymbol{\vartheta}$ and β and their covariance matrix are

$$\begin{aligned}
\hat{\boldsymbol{\vartheta}} &= (-1.715, 0.004)^\top \\
\hat{\beta} &= -0.005 \\
\hat{\Sigma} &= \left(\begin{array}{rr|r} 0.002 & -0.000 & 0.000 \\ -0.000 & 0.000 & 0.000 \\ \hline 0.000 & 0.000 & 0.000 \end{array}\right)
\end{aligned}$$

and thus $\hat{\mu} = 485.65917$ and $\hat{\gamma} = 1.52745$ and

$$\hat{Q}(\alpha \mid x) = \hat{\mu} + \hat{\gamma} \times (x - 3500)/100) + \hat{\vartheta}_2^{-1}\Phi^{-1}(\alpha)$$

gives us the conditional $\alpha \times 100\%$ quantile $\hat{Q}(\alpha \mid x)$ and we plot these conditional quantiles for $\alpha = .1, .5, .9$ as a function of neonatal weight x in Figure 6.8. The above formula shows that these quantiles are always symmetric around the mean and median ($\Phi^{-1}(.5) = 0$), which is a highly unrealistic assumption for this data.

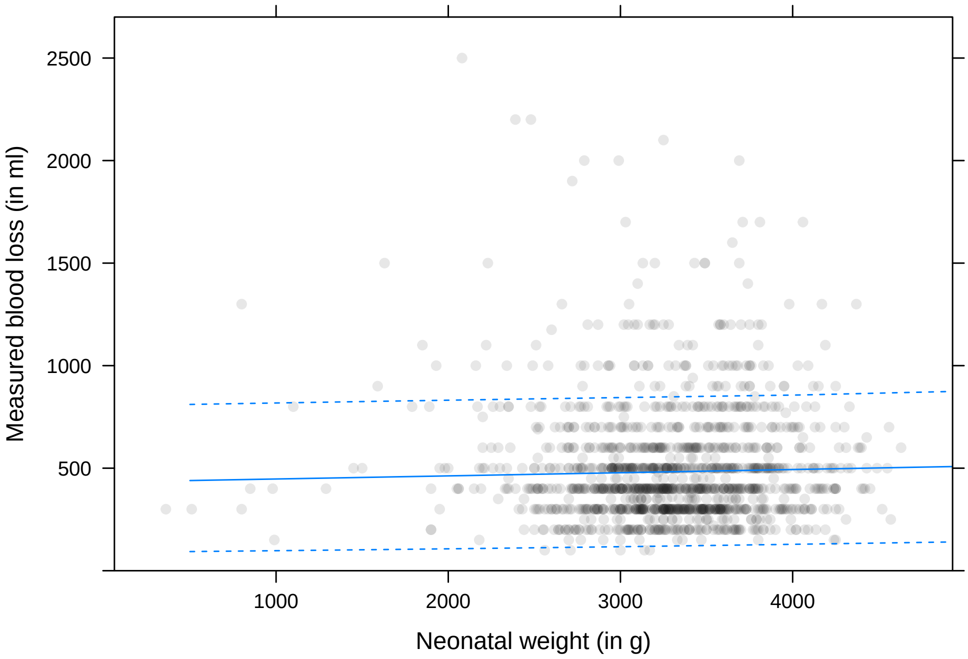


Figure 6.8: Measured blood loss. Scatterplot of neonatal weight and measured blood loss, overlayed with conditional median (solid line) and conditional 10% and 90% quantiles obtained from a simple normal linear regression model measured blood loss by neonatal weight.

It should be noted that the results reported here are not least-squares estimates but maximum likelihood estimates obtained by maximising the interval-censored log-likelihood.

□

There is no need to discuss specific inference procedures here. All technical steps for the estimation of $\boldsymbol{\vartheta}$ and β as well as the derivation of tests and confidence intervals for β introduced in Chapter 5 can be applied in the very same way to the simple transformation models of this section. However, this technically uninteresting statement has very interesting practical consequences. Many classical procedures, such as the Wilcoxon-Rank-Sum-Test or the Log-Rank test can be applied not only in the two-sample set-up but also simple regression situations where one is interested in analysing a single numeric explanatory variable.

Example 6.7. The 95% Wald $(0.98831, 1.02519)$, score $(0.98835, 1.02516)$, and likelihood $(0.98827, 1.02519)$ confidence intervals for the odds ratio β modelled in Example 6.5 all rely on asymptotic arguments. The permutation score confidence interval $(0.98755, 1.02576)$ conditions on the observed data and computes a confidence interval based on the conditional permutation distribution under $H_0 : \beta = 0$. The interval is a bit wider than the other three intervals, so relying on asymptotic approximations might not be a good idea in this case.

□

6.2 Multiple Regression

An important step ahead is to condition on multiple explanatory variables $x_1, x_2, \dots$ in a so-called "multiple linear regression model". We can plot and at least partially understand what is going on in absense of a formal model when only three variables, that is two explanatory variables x_1, x_2 and the response y, have been observed. Things quickly become unmanageble for more than two or at most three explanatory variables and we need a model to study the changes in the conditional distribution induced by multiple explanatory variables. We focus on a linear transformation model

$$\mathbb{P}(Y \le y \mid \boldsymbol{X} = \boldsymbol{x}) = F_Z(h(y) - \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}) \tag{6.1}$$

with linear predictor

$$\tilde{\boldsymbol{x}}^\top \boldsymbol{\beta} = \sum_{q=1}^{Q} x_q \beta_q.$$

The name of the model is highly misleading, because the transformation function $h(y)$ is nonlinear in many models, and the name comes from the linear combination $\tilde{\boldsymbol{x}}^\top \beta$ capturing the impact of $\boldsymbol{x}$ on the conditional distribution of Y. A linear predictor is the scalar product of the vector of regression coefficients $\boldsymbol{\beta} = (\beta_1, \dots, \beta_Q)^\top \in \mathbb{R}^Q$ and some conversion $\tilde{\boldsymbol{x}} \in \mathbb{R}^Q$ of the explanatory variables $\boldsymbol{x} \in \chi$. The vectors $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$ are typically not the same, because $\boldsymbol{x} \in \chi$ which might be of a different dimension than Q. Instead, $\tilde{\boldsymbol{x}}$ is a suitable and a priori known conversion of $\boldsymbol{x}$ which, unlike the transformation function $h(y)$, is not a parameter and will not be estimated.

The main motivation for linear regression models (6.1) and other models involving a linear predictor $\tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}$ is complexity reduction. We assume that the impact of some x_q on Y is linear, the slope is given by a single number β_q, and additive to the impact caused by other explanatory variables. If such a simple model describes the true state of affairs well enough, we can enjoy almost the same simplicity of model description and communication as in the two-sample situation or in the simple regression set-up. More often, however, conditional distributions cannot be approximated well enough by such simple models and one of the most difficult tasks in regression analysis is to detect these cases.

The regression coefficients β_q are also called partial effects, because they describe the change in the linear predictor $\tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}$ induced by a one-unit change in the qth variable when all other variables are held constant. One can write the partial effect as the derivative

$$\beta_q = \frac{\partial \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}}{\partial \tilde{x}_q}.$$

Partial effects depends on the scale of the corresponding explanatory variable and it is important to report regression coefficients on a meaningful metric. Of similar importance is the suitable choice of a reference configuration. A configuration $\boldsymbol{x}_0$ for which $\tilde{\boldsymbol{x}}_0^\top \boldsymbol{\beta} = 0$ for all $\boldsymbol{\beta} \in \mathbb{R}^Q$ is called a reference, or reference configuration. One should chose the reference in an appropriate way such that

$$\mathbb{P}(Y \le y \mid \boldsymbol{X} = \boldsymbol{x}_0) = F_Z(h(y) - \tilde{\boldsymbol{x}}_0^\top \boldsymbol{\beta}) = F_Z(h(y))$$

is a conditional distribution with reasonable interpretation, that is a conditional distribution function for a relevant population.

Example 6.8. We extend the model introduced in Example 6.5 with a variable for mode of delivery and are interested in the conditional distribution of measured blood loss given mode of delivery (vaginal or Cesarean) and neonatal weight. One could start with the linear predictor

$$\tilde{\boldsymbol{x}}^\top \boldsymbol{\beta} = \beta_1 \times \text{Cesarean delivery} + \beta_2 \times \text{Neonatal weight in g}$$

whose reference is a vaginal birth of a fetus with weight 0 gr. Clearly, the corresponding transformation function h doesn't have any reasonable interpretation, so we switch to the linear predictor

$$\tilde{\boldsymbol{x}}^\top \boldsymbol{\beta} = \beta_1 \times \text{Cesarean delivery} + \beta_2 \times (\text{Neonatal weight in g} - 3500)$$

such that the reference is a vaginal birth of a 3500 g baby. The scale of β_2 might not be very smart, because β_2 is the change in the linear predictor induced by a one gram increase in neonatal weight. Finally, we define

$$\tilde{\boldsymbol{x}}^\top \boldsymbol{\beta} = \beta_1 \times \text{Cesarean delivery} + \beta_2 \times (\text{Neonatal weight in g} - 3500)/100.$$

Here we do not work with the original neonatal weight, but with the conversion $\tilde{x}_2 = (x_2 - 3500)/100$. With the choice $F_Z = \text{expit}$ we define the model as

$$\begin{aligned}&\mathbb{P}(Y \leq y \mid \boldsymbol{X} = \boldsymbol{x}) = \\ &\quad \text{expit}(h(y) - (\beta_1 \times \text{Cesarean delivery} + \beta_2 \times (\text{Neonatal weight in g} - 3500)/100))\end{aligned}$$

where the conditional log-odds function decreases by an amount of β_1 when changing from a vaginal to a Cesarean delivery for a fixed neonatal weight. It decreases by β_2 when 100 g are added to the neonatal weight. The model is a multiplicative model for the odds function.

Using a transformation function $\boldsymbol{a}(y)^\top \boldsymbol{\vartheta}$ in Bernstein form, the joint maximum likelihood estimates of $\boldsymbol{\vartheta}$ and the log-odds ratios $\boldsymbol{\beta}$ and their covariance matrix are

$$
\begin{aligned}
\hat{\boldsymbol{\vartheta}} &= (-8.124, 1.615, 1.615, 1.615, 3.741, 4.475, 5.210)^\top \\
\hat{\boldsymbol{\beta}} &= (1.516, 0.032)^\top \\
\hat{\Sigma} &= \left(\begin{array}{rrrrrrr|rr}
0.267 & -0.479 & 0.674 & -0.627 & 0.351 & -0.044 & -0.025 & -0.008 & -0.000 \\
-0.479 & 1.111 & -1.709 & 1.737 & -0.999 & 0.150 & 0.094 & -0.015 & -0.000 \\
0.674 & -1.709 & 2.929 & -3.238 & 2.006 & -0.323 & -0.204 & 0.015 & 0.000 \\
-0.627 & 1.737 & -3.238 & 4.017 & -2.730 & 0.520 & 0.342 & -0.020 & -0.000 \\
0.351 & -0.999 & 2.006 & -2.730 & 2.229 & -0.519 & -0.344 & 0.005 & -0.000 \\
-0.044 & 0.150 & -0.323 & 0.520 & -0.519 & 0.358 & 0.263 & -0.005 & -0.000 \\
-0.025 & 0.094 & -0.204 & 0.342 & -0.344 & 0.263 & 0.270 & -0.004 & -0.000 \\
\hline
-0.008 & -0.015 & 0.015 & -0.020 & 0.005 & -0.005 & -0.004 & 0.011 & 0.000 \\
-0.000 & -0.000 & 0.000 & -0.000 & -0.000 & -0.000 & -0.000 & 0.000 & 0.000
\end{array}\right)
\end{aligned}
$$

In contrast to marginal log-odds ratios $\tilde{\beta}_1$ of mode of delivery in the two-sample model

$$\mathbb{P}(Y \le y \mid \text{Mode of delivery}) = \text{expit}(h(y) - \tilde{\beta}_1 \times \text{Cesarean delivery})$$

and neonatal weight $\tilde{\beta}_2$ the simple regression model

$$\mathbb{P}(Y \le y \mid \text{Neonatal weight}) = \text{expit}(h(y) - \tilde{\beta}_2 \times (\text{Neonatal weight in g} - 3500)/100))$$

the partial log-odds ratios β_1 and β_2 try to disentangle the effects of both variables. Extremely small or very large babies are certainly a reason to plan ahead for a Cesarean section and thus some dependency between the two variables mode of delivery and neonatal weight has to be expected. The multiple regression provides a way to estimate adjusted or partial effects of each variable, taking the effect of the other variable into account.

The partial odds ratio $\exp(-\hat{\beta}_1)$ compares the odds function of a Cesarean section to the odds function of a vaginal delivery for any fixed neonatal weight. The odds ratio $\exp(-\hat{\beta}_2)$ compares the odds function for babies born at some weight x g to the odds when a baby weighting $x - 100$ g is given birth, either for vaginal or Cesarean deliveries. The log-odds ratios, odds ratios, and corresponding 95% Wald confidence intervals are given in

	log-OR	OR	97.5 %	2.5 %
Mode of delivery: β_1	-1.516	0.220	0.179	0.270
Neonatal weight: β_2	-0.032	0.969	0.951	0.987

Table 6.1: Measured blood loss. Partial log-odds ratios of mode of delivery and neonatal weight with 95% likelihood confidence intervals.

Table 6.1. The partial odds ratio of neonatal weight is, in contrast to the marginal odds ratio reported in Example 6.5, significant at the 5% level.

The conditional quantile functions are plotted in Figure 6.9. The effect of neonatal weight and variance heterogeneity towards heavier babies is more pronounced than in the marginal model for neonatal weight. The same information is presented using conditional distribution functions in Figure 6.10.

□

Interaction terms

$$\tilde{\boldsymbol{x}}^\top \boldsymbol{\beta} = \beta_1 \tilde{x}_1 + \beta_2 \tilde{x}_2 + \beta_3 \tilde{x}_1 \tilde{x}_2$$

Example 6.9. Model with interaction term $\tilde{x}_1 =$ Cesarean delivery,

$\tilde{x}_2 = (\text{Neonatal weight in g} - 3500)/100$

$$\begin{aligned}&\mathbb{P}(Y \le y \mid \boldsymbol{X} = \boldsymbol{x}) = \\ &\quad \text{expit}(h(y) - (\beta_1 \times \tilde{x}_1 + \beta_2 \times \tilde{x}_2 + \beta_3 \times \tilde{x}_1 \times \tilde{x}_2))\end{aligned}$$

or, with $\tilde{x}_3 = \tilde{x}_1 \times \tilde{x}_2$

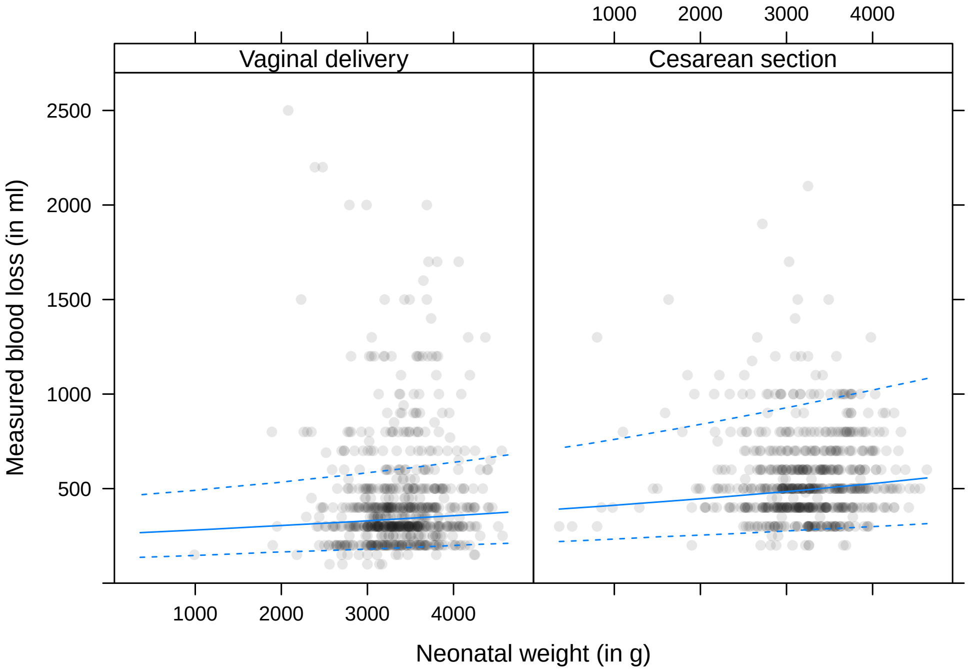


Figure 6.9: Measured blood loss. Scatterplot of neonatal weight and measured blood loss for vaginal deliveries and Cesarean sections, overlayed with conditional median (solid line) and conditional 10% and 90% quantiles obtained from a multiple proportional-odds model explaining measured blood loss by mode of delivery and neonatal weight.

$$\hat{\boldsymbol{\vartheta}} = (-8.105, 1.648, 1.648, 1.648, 3.765, 4.502, 5.239)^\top$$

$$\hat{\boldsymbol{\beta}} = (1.462, 0.021, 0.032)^\top$$

$$\hat{\Sigma} = \left(\begin{array}{rrrrrrr|rrr}
0.267 & -0.478 & 0.673 & -0.627 & 0.351 & -0.044 & -0.025 & -0.008 & -0.000 & -0.000 \\
-0.478 & 1.112 & -1.709 & 1.738 & -0.998 & 0.152 & 0.095 & -0.016 & -0.001 & 0.001 \\
0.673 & -1.709 & 2.929 & -3.239 & 2.008 & -0.325 & -0.205 & 0.015 & 0.000 & 0.000 \\
-0.627 & 1.738 & -3.239 & 4.018 & -2.731 & 0.523 & 0.344 & -0.021 & -0.001 & 0.000 \\
0.351 & -0.998 & 2.008 & -2.731 & 2.232 & -0.521 & -0.345 & 0.004 & -0.000 & 0.001 \\
-0.044 & 0.152 & -0.325 & 0.523 & -0.521 & 0.359 & 0.264 & -0.005 & -0.000 & 0.000 \\
-0.025 & 0.095 & -0.205 & 0.344 & -0.345 & 0.264 & 0.270 & -0.005 & -0.000 & 0.000 \\
\hline
-0.008 & -0.016 & 0.015 & -0.021 & 0.004 & -0.005 & -0.005 & 0.012 & 0.000 & -0.001 \\
-0.000 & -0.001 & 0.000 & -0.001 & -0.000 & -0.000 & -0.000 & 0.000 & 0.000 & -0.000 \\
-0.000 & 0.001 & 0.000 & 0.000 & 0.001 & 0.000 & 0.000 & -0.001 & -0.000 & 0.000
\end{array}\right)$$

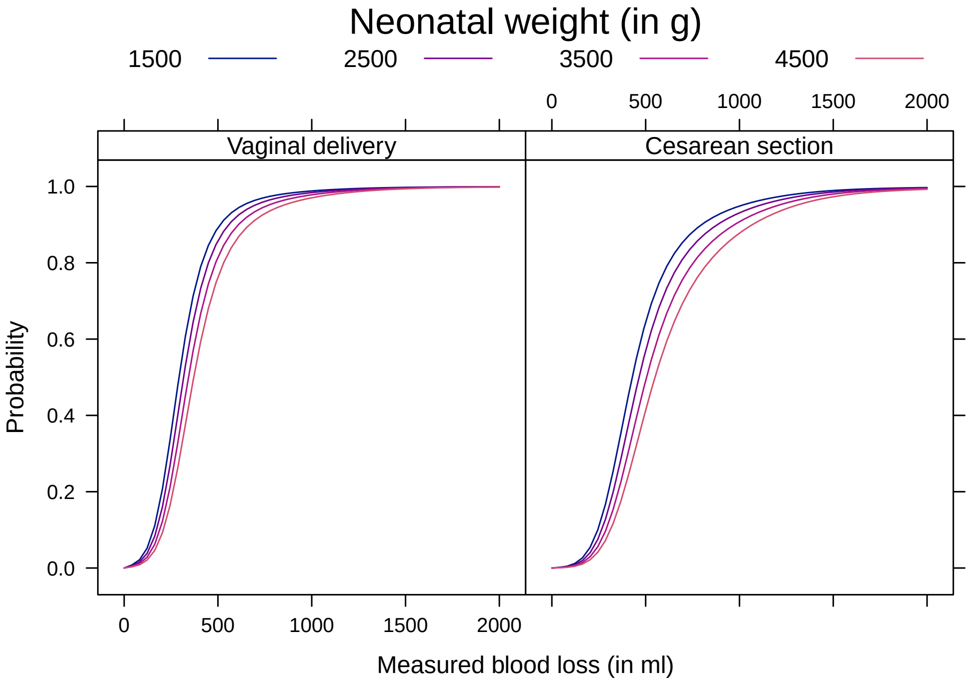
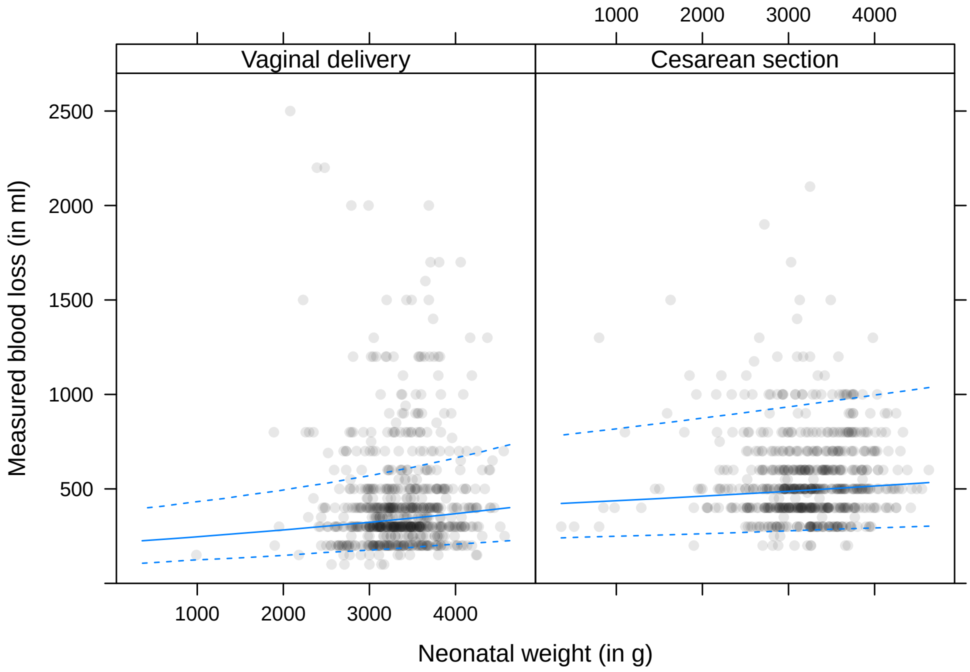


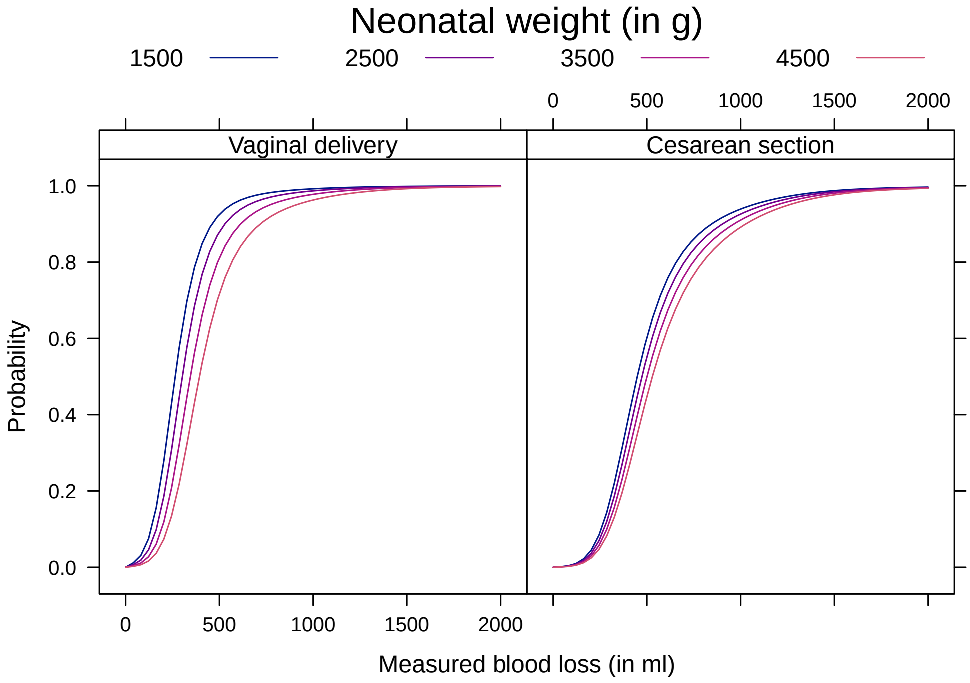
Figure 6.10: Measured blood loss. Conditional distibution functions of measured blood loss for vaginal deliveries and Cesarean sections and selected neonatal weights obtained from a multiple proportional-odds model explaining measured blood loss by mode of delivery and neonatal weight.

	log-OR	OR	97.5 %	2.5 %
Mode of delivery: β_1	-1.462	0.232	0.187	0.288
Neonatal weight: β_2	-0.021	0.979	0.957	1.002
Interaction: β_3	-0.032	0.969	0.932	1.007

Table 6.2: Measured blood loss. Partial log-odds ratios of mode of delivery and neonatal weight with 95% likelihood confidence intervals.

□


Vaginal delivery
Cesarean section
Measured blood loss (in ml)
2500
2000
1500
1000
500
Neonatal weight (in g)
1000
2000
3000
4000


Neonatal weight (in g)
1500
2500
3500
4500
Vaginal delivery
Cesarean section
Probability
1.0
0.8
0.6
0.4
0.2
0.0
Measured blood loss (in ml)
0
500
1000
1500
2000

6.3 Inference for Regression Coefficients

6.4 Model Residuals

$$\mathbb{P}(Y \le y \mid \boldsymbol{X} = \boldsymbol{x}) = F_Z(\boldsymbol{a}(y)^\top \boldsymbol{\vartheta} - \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta} - \mu), \quad \mu \equiv 0$$

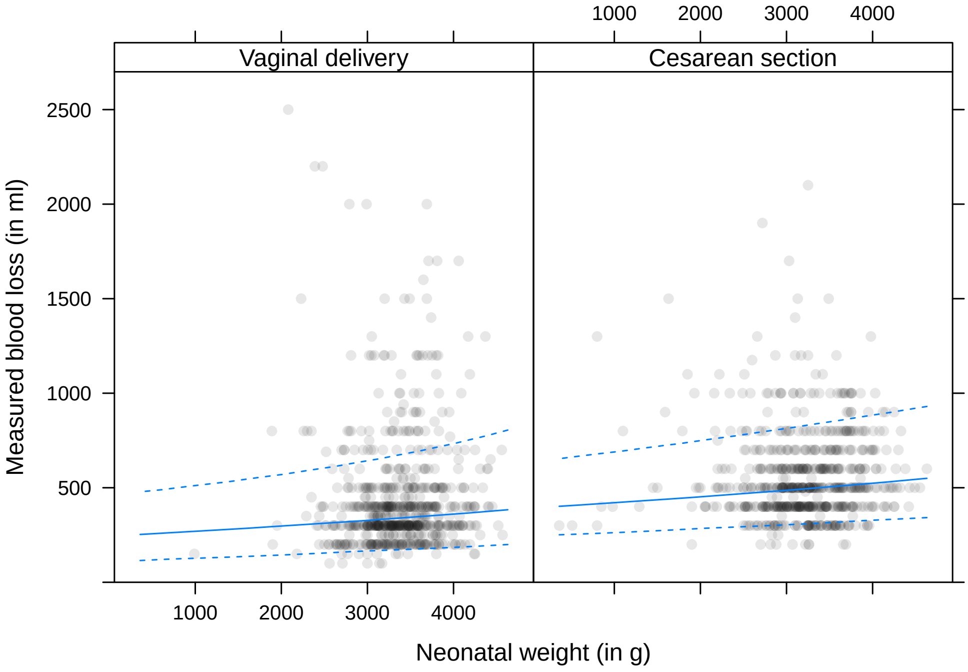
$$\begin{aligned}
\boldsymbol{s}((\hat{\boldsymbol{\vartheta}}^\top, \hat{\boldsymbol{\beta}}^\top, \mu)^\top) &= (\boldsymbol{s}(\hat{\boldsymbol{\vartheta}})^\top, \boldsymbol{s}(\hat{\boldsymbol{\beta}}))^\top, \boldsymbol{s}(\mu))^\top \\
\boldsymbol{s}(\mu = 0) &= \left.\frac{\partial \ell((\boldsymbol{\vartheta}^\top, \boldsymbol{\beta}^\top, \mu)^\top)}{\partial \mu}\right|_{\boldsymbol{\vartheta} = \hat{\boldsymbol{\vartheta}}, \boldsymbol{\beta} = \hat{\boldsymbol{\beta}}, \mu = 0}
\end{aligned}$$

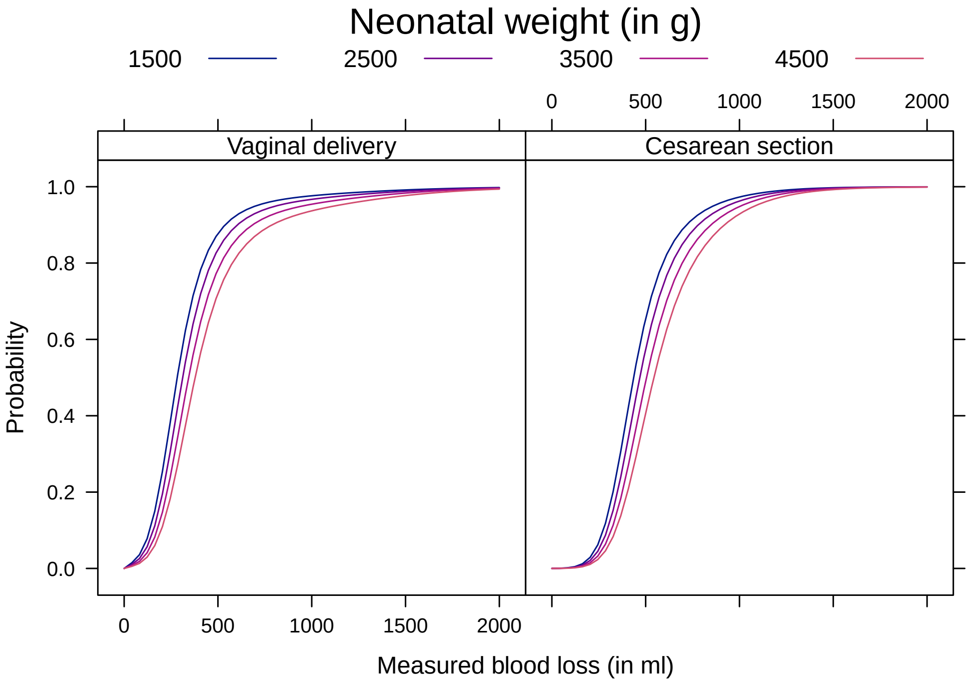
Example 6.10. □

Chapter 7

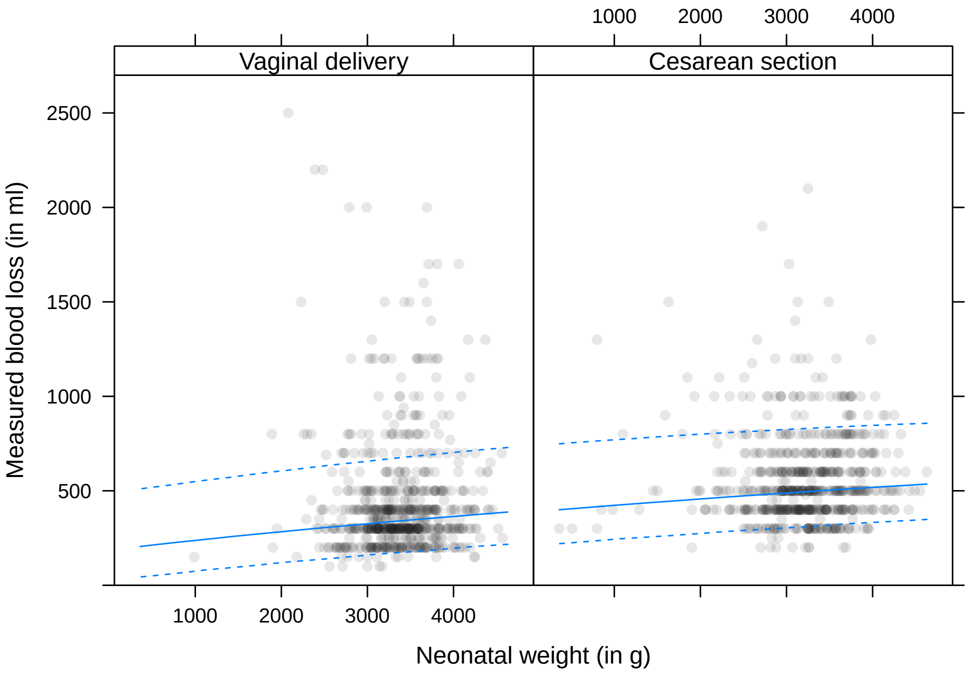
Distribution Regression

```
## Warning in model.matrix.box_bases(object = structure(list(iresponse
= structure(function (data, :  use scale = TRUE in as.basis.formula
with sumcontr = TRUE
## Warning in model.matrix.box_bases(object = structure(list(iresponse
```

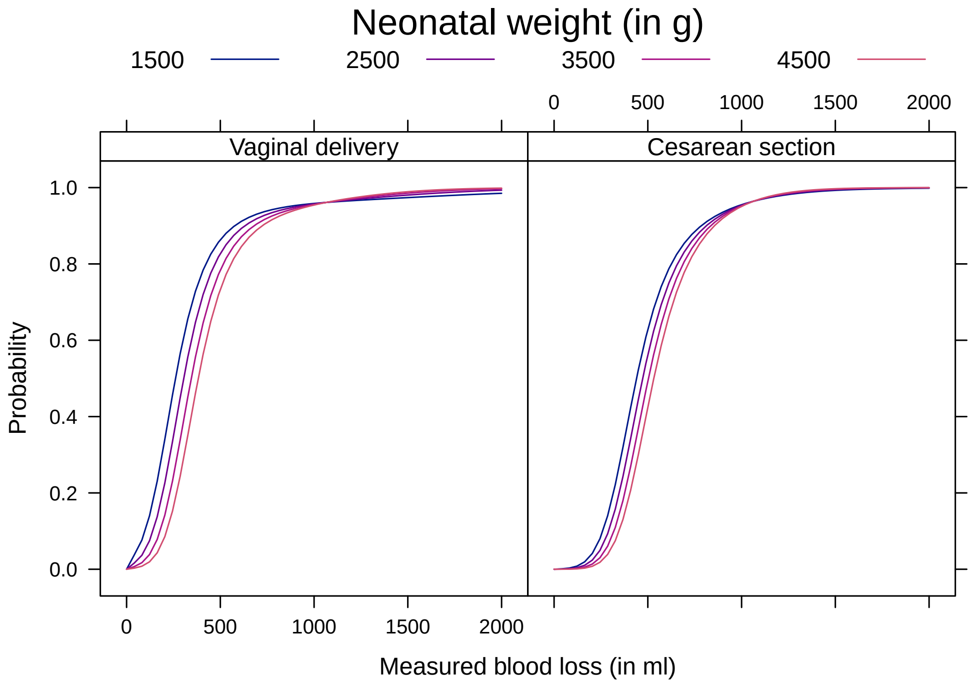




```
= structure(function (data, :  use scale = TRUE in as.basis.formula
with sumcontr = TRUE
## Warning in model.matrix.box_bases(object = structure(list(iresponse
= structure(function (data, :  use scale = TRUE in as.basis.formula
with sumcontr = TRUE
## Warning in model.matrix.box_bases(object = structure(list(iresponse
= structure(function (data, :  use scale = TRUE in as.basis.formula
with sumcontr = TRUE
## Warning in model.matrix.box_bases(object = structure(list(iresponse
= structure(function (data, :  use scale = TRUE in as.basis.formula
with sumcontr = TRUE
```


Vaginal delivery
Cesarean section
Measured blood loss (in ml)
2500
2000
1500
1000
500
1000
2000
3000
4000
Neonatal weight (in g)

```
## Warning in model.matrix.box_bases(object = object, data =
newdata, dim = dim, :  use scale = TRUE in as.basis.formula
with sumcontr = TRUE
```



Chapter 8

Conditional Transformation Models

Chapter 9

Ordinal Regression

Chapter 10

Count Regression

Chapter 11

Survival Analysis

Chapter 12

Appendix

12.1 Introduction

Models

```
### ECDF via Polr
blood$MBLc <- cut(blood$MBL, breaks = c(-Inf, sMBL), ordered = TRUE)
wght <- c(table(blood$MBLc))
y <- sort(unique(blood$MBLc))
mhECDF <- Polr(y ~ 1, weights = wght, method = "probit")

### Shift model (1.1) with discrete parameters
mVDPolr <- Polr(MBLc ~ vd, data = blood, method = "probit")

### smooth unconditional transformation model
i <- unclass(blood$MBLc)
blood$MBLsurv <- Surv(c(0, sMBL[-length(sMBL)])[i], sMBL[i], type = "interva
mVDmlt <- BoxCox(MBLsurv | vd ~ 1, data = blood, order = 15,
               bounds = c(0, Inf), support = c(0, 2000),
               extrapolate = TRUE)
llVDmlt <- logLik(mVDmlt)

### smooth shift transformation model
mVDshift <- BoxCox(MBLsurv ~ vd, data = blood, order = 15,
                bounds = c(0, Inf), support = c(0, 2000),
                extrapolate = TRUE)
llVDshift <- logLik(mVDshift)
```

```
## Error in as.mlt(mh):  object 'mh' not found
## Error in as.mlt(mh):  object 'mh' not found
## Error in eval(expr, envir, enclos):  object 'mh' not found
## Error in h(simpleError(msg, call)):  error in evaluating the
argument 'x' in selecting a method for function 'plot':  could not
find function "FZ"
## Error in h(simpleError(msg, call)):  error in evaluating the
argument 'x' in selecting a method for function 'plot':  could not
find function "FZ"
```

12.2 Most Likely Transformations

Models

```
### hot start
theta <- qnorm(ecdf(blood$MBL)(brk_o[-c(1, length(brk_o))]))
m_TB <- Polr(MBLsurv_o ~ 1, data = blood, method = "probit", theta = theta)
s <- predict(m_TB, newdata = data.frame(1), type = "survivor")
```

```
sf <- survfit(MBLsurv_i ~ 1, data = blood)
d <- max(abs(s[-length(s)] - summary(sf, time = brk_o[-1])$surv))
stopifnot(d < 1e-4)
max(d)

## [1] 1.391652e-06
```

12.3 Two Sample Comparisons

The linear model for Example 5.17 was fitted by

```
### Wellek (2010) Testing Statistical Hypotheses of
### Equivalence and Noninferiority, Table 6.3
bp <- data.frame(bp = c(10.3, 11.3,  2.0, -6.1,  6.2,  6.8,
                         3.7, -3.3, -3.6, -3.5, 13.7, 12.6,
                         3.3, 17.7,  6.7, 11.1, -5.8,  6.9,
                         5.8,  3.0,  6.0,  3.5, 18.7,  9.6),
                 gp = gl(2, 12, labels = c("Moxonidin",
                                           "Captopril")))
m1 <- Lm(bp ~ gp, data = bp)
```

and the estimated coefficients and their covariance is

```
coef(as.mlt(m1))

## (Intercept)          bp gpCaptopril
##  -0.6374578   0.1526841   0.4631430

vcov(as.mlt(m1))

##              (Intercept)            bp gpCaptopril
## (Intercept)  0.091798973 -0.0020276980 0.077182649
## bp          -0.002027698  0.0004856762 0.001473218
## gpCaptopril  0.077182649  0.0014732177 0.171135427
```

Wald, score, and likelihood intervals

```
### Wald
confint(m1)

##                 2.5 %   97.5 %
## gpCaptopril -0.347665 1.273951
```

```
### score
confint(score_test(m1))

##                    2.5 %   97.5 %
## gpCaptopril -0.3476658 1.273951

### likelihood
confint(profile(m1))

##      2.5 %     97.5 %
## -0.3475044  1.2741115
```

Proportional reverse time hazard ratio models for postpartum blood loss were estimated by

```
mi

## [1] 6

m <- Lehmann(MBLsurv ~ vd, data = blood, order = mi,
             bounds = c(0, Inf), support = c(0, 2000),
             extrapolate = TRUE)
summary(m)

##
##   Proportional Reverse Time Hazards Linear Regression Model
##
## Call:
## Lehmann(formula = MBLsurv ~ vd, data = blood, order = mi, bounds = c(0,
##     Inf), support = c(0, 2000), extrapolate = TRUE)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## vdCesarean section  0.96089    0.06114   15.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood:
##  -4241.103 (df = 8)
## Likelihood-ratio Test: Chisq = 245.574 on 1 degrees of freedom; p = < 2.2
```

The Wald statistic is shown as z-value and the last line presents the likelihood ratio statistic with a corresponding P-value. The in-sample log-likelihood is -4241.103

Wald intervals

```
confint(m)

##                         2.5 %   97.5 %
## vdCesarean section 0.8410557 1.080732
```

Score intervals

```
(st <- score_test(m))

##
##  Transformation Score Test
##
## data:  Lehmann(formula = MBLsurv ~ vd, data = blood, order = mi, bounds =
## Z = -16.204, p-value < 2.2e-16
## alternative hypothesis: true lehmann parameter for vdCesarean section is
## 95 percent confidence interval:
##  0.8411185 1.0806694
## sample estimates:
## lehmann parameter for vdCesarean section
##                                 0.960894

confint(st)
```

```
##                         2.5 %   97.5 %
## vdCesarean section 0.8411185 1.080669
```

Likelihood-ratio intervals

```
confint(profile(m))

##     2.5 %    97.5 %
## 0.8412106 1.0809235
```

12.4 Linear Transformation Models

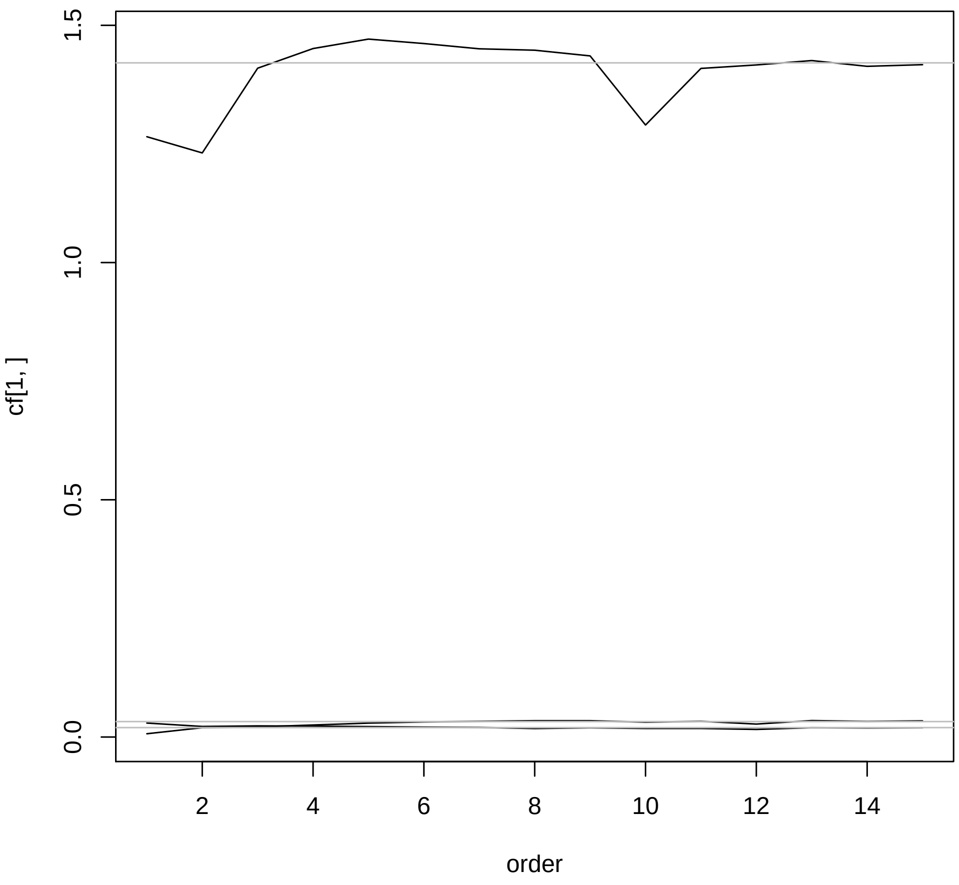
MBLsurv is nonparametric likelihood, orm is also based on this likelihood (but needs MBL as response)

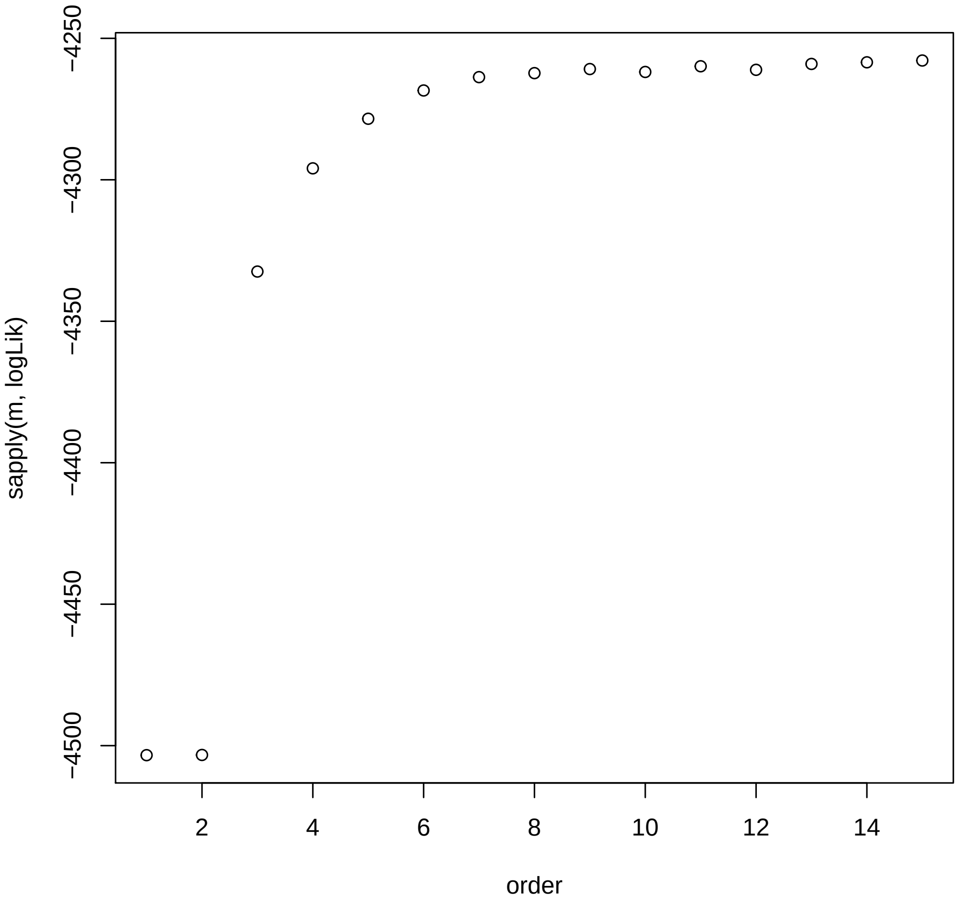
```
order <- 1:15
m <- lapply(order, function(i)
    Colr(MBLsurv ~ vdC*fw, data = blood, order = i,
             bounds = c(0, Inf), support = c(0, 2000),
             extrapolate = i > 1)
)
mo <- orm(MBL ~ vdC*fw, data = blood)
```

ToDo: add confidence intervals

-4257.838 -3147.384

```
nd <- data.frame(order = order)
cf <- sapply(m, coef)
plot(order, cf[1,], type = "l", ylim = range(cf))
lines(order, cf[2,])
lines(order, cf[3,])
abline(h = -rev(coef(mo))[1:3], col = "grey")
```




−4250
−4300
−4350
−4400
−4450
−4500
sapply(m, logLik)
2
4
6
8
10
12
14
order

Bibliography

Ralf Bender, Thomas Augustin, and Maria Blettner. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, 24(11):1713–1723, 2005. doi: 10.1002/sim.2059.

Peter J. Bickel, Chris A. J. Klaassen, Ya'acov Ritov, and Jon A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models.* The Johns Hopkins University Press, Baltimore, U.S.A. and London, U.K., 1993.

K. Chen, Z. Jin, and Z. Ying. Semiparametric analysis of transformation models with censored data. *Biometrika*, 89(3):659–668, 2002. doi: 10.1093/biomet/89.3.659.

S. C. Cheng, L. J. Wei, and Z. Ying. Analysis of transformation models with censored data. *Biometrika*, 82(4):835–845, 1995. doi: 10.1093/biomet/82.4.835.

Rida T. Farouki. The Bernstein polynomial basis: A centennial retrospective. *Computer Aided Geometric Design*, 29(6):379–419, 2012. doi: 10.1016/j.cagd.2012.03.001.

Ronald Aylmer Fisher. Two new properties of mathematical likelihood. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 53:285–306, 1934.

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

Christian Haslinger, Wolfgang Korte, Torsten Hothorn, Romana Brun, Charles Greenberg, and Roland Zimmermann. The impact of prepartum factor XIII activity on postpartum blood loss. *Journal of Thrombosis and Haemostasis*, 2020. doi: 10.1111/jth.14795.

Torsten Hothorn, Thomas Kneib, and Peter Bühlmann. Conditional transformation models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):3–27, 2014. doi: 10.1111/rssb.12017.

Maike Katja Kahr, Romana Brun, Roland Zimmermann, Denise Franke, and Christian Haslinger. Validation of a quantitative system for real-time measurement of postpartum blood loss. *Archives of Gynecology and Obstetrics*, 298:1071–1077, 2018. doi: 10.1007/s00404-018-4896-0.

John P. Klein and Melvin K. Moeschberger. *Survival Analysis. Techniques for Censored and Truncated Data.* Springer, New York, U.S.A., 2nd edition, 2003.

Erich L. Lehmann. *Theory of Point Estimation.* John Wiley & Sons, New York, U.S.A, 1983.

Jim K. Lindsey. *Parametric Statistical Inference.* Clarendon Press, Oxford, UK, 1996.

Edwin James George Pitman. The estimation of the location and scale parameters of a continuous population of any given form. *Biometrika*, 30 (3/4):391–421, 1939.

Sarah van de Geer. *Empirical Processes in M-Estimation.* Cambridge University Press, Cambridge, UK, 2000.

Aart W. van der Vaart. *Asymptotic Statistics.* Cambridge University Press, Cambridge, UK, 1998.