1

Distributional Regression Models

This chapter serves as a general introduction to the types of regression models discussed in this book and has the following four major aims.

- Provide the context for the GAMLSS way of distributional regression modeling.
- Give a brief history of the development of GAMLSS and its relation to other regression approaches.
- Introduce a first example illustrating some advantages and potential of GAMLSS analyses.
- Compare GAMLSS with, and delineate from, competing distributional regression approaches.

We start with setting the scene in Section 1.1, where we describe how our presentation fits into the data analysis circle for working with research data. Afterwards, we introduce basic terminology for and ingredients of statistical regression models (Sections 1.2 and 1.3) and briefly summarize the historical developments that led to the introduction of GAMLSS (Section 1.4). We then introduce the structure of GAMLSS in a more general and formal way (Section 1.5) and close the chapter with a discussion of alternative distributional regression approaches (Section 1.6).

1.1 The Data Analysis Circle

The statistical analysis of research data can be seen as a circular process consisting of the following phases (see Figure 1.1):

(i) Planning the data collection: In the first phase, the process of collecting the data has to be decided upon. The exact strategy strongly depends on the research question of interest as well as the purpose of the analysis, and may involve aspects such as sample size calculation, defining the relevant study population, choosing an experimental design, defining the sampling process, investigating sources for already available data, etc. In addition, this step may feature developing a fixed data analysis plan (for confirmatory analysis, e.g. in clinical studies), pre-registration of the trial, and other aspects that determine steps taken later on in the analysis of the data.

https://doi.org/10.1017/9781009410076.003 Published online by Cambridge University Press

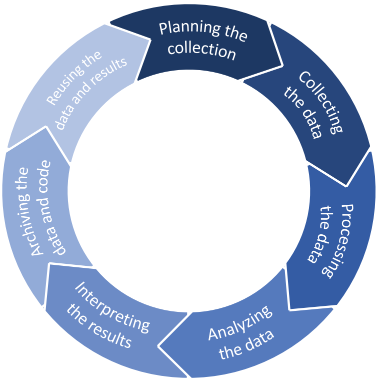

Figure 1.1 The data analysis circle.

(ii) Collecting the data: The second phase implements data collection, which may range from generating data in experimental settings, to collecting observational data or scraping data from existing sources.

(iii) Processing the data: To turn the raw data collected in the second phase into a dataset to be analyzed, it is of utmost importance to pre-process the data. This includes diverse aspects such as checking for inconsistencies, treating missing values, implementing transformations, calculating indices, graphical visualizations, etc.

(iv) Analyzing the data: The fourth phase implements the actual analysis. Throughout this book, this analysis will be based on statistical modeling, but of course very different approaches may be taken depending on the purpose of the analysis. Graphical tools are of particular relevance here to check the validity of model decisions and assumptions.

(v) Interpreting the results: After having analyzed the data, interpreting the results with respect to the original research question is crucial. This includes deriving required quantities from the raw model results, communication with subject matter scientists, and also publishing the results in appropriate formats, communicating the results to the general public, or deriving policy advice. Again, graphical displays feature prominently in this task.

(vi) Archiving the data and code: To ensure reproducibility and to make data that have been collected with considerable effort accessible to other researchers, it has become common practice in many areas to archive research data in repositories (possibly including their publication and attaching a persistent identifier to the data) and to publish the code utilized for their analysis. Of course there may be limitations, for example owing to data confidentiality, proprietary rights, etc., but publishing at least a basic set of information including analysis code is nowadays considered good scientific practice.

https://doi.org/10.1017/9781009410076.003 Published online by Cambridge University Press

(vii) Reusing the data and results: To close the data analysis circle, data collected for a specific purpose can often be reused for further analyses in different settings or for different research questions. Similarly, the results achieved in one analysis often form the basis for future research. In both cases, we considerably benefit from archived data and code as discussed in the previous step.

The focus in our book is on phases (iv) and (v) of the data analysis circle, that is, we are concerned with statistical modeling as a specific instance of data analysis and the thorough interpretation of the results obtained from the statistical model. However, the other aspects are of course also very relevant when performing data analyses, even though we do not discuss them in detail in this book. In particular, for a data analysis to be useful, the following properties are required from the data.

- The data should accurately represent the *population* under study. The population consists of the subjects we would like to investigate and the sample (that is, the data) should be representative of it. For example, this can be ensured by taking a random sample from the population.
- The data should be collected with *integrity* so there is no intentional bias.
- *Extreme values* in the data should be genuine values and not human or machine errors.
- *Missing values* should be properly treated rather than ignored, see, for example, van Buuren (2018).
- *Data contamination* should be small, if it exists at all. By data contamination we mean that part of the data is unintentionally corrupt. This phenomenon is usually treated by "robust" statistical analysis; for example, Aeberhard et al. (2021) propose robust methods for GAMLSS.

The decisions made in any of the steps of the data analysis circle depend of course very much on the purpose of the analysis. In particular, data are analyzed with very different aims, which often can be classified as follows.

- Confirmatory, when a specific hypothesis about an effect of interest is studied. This often goes along with the goal of establishing a causal link between observables, which requires either specific choices concerning the data collection process (e.g. an experimental set-up with random treatment assignment) or additional assumptions on observational data (e.g. using instrumental variables or graphical models to establish causal effects);
- Exploratory, when the goal is to derive additional (or new) knowledge about an empirical phenomenon of interest. In this case, the main focus is on identifying relevant associations rather than establishing causal relations;
- Prediction-oriented, when the analyst is interested in determining a model that not only describes the observed data well, but is also able to predict new observations. In this case, one may set interpretability aside in favour of better prediction, as is commonly done in machine learning methodology such as deep neural networks.

https://doi.org/10.1017/9781009410076.003 Published online by Cambridge University Press

Most of our applications fall into the category of exploratory data analysis, but we will also address applications with an emphasis on prediction. In data mining, the models are typically exploratory and are used to find interesting patterns in the data. Medical statistics and econometrics often use confirmatory models, where the estimated coefficients of the model and their interpretation play an important role. It is here where likelihood-based and Bayesian ways of fitting the model are prominent. In machine learning (including classical boosting) and artificial intelligence, the focus is often on finding models that sacrifice interpretability for the sake of optimized predictive ability. Statistical boosting approaches (Chapter 7) represent a mixture between classical statistical modeling and machine learning, allowing the fitting of an interpretable model with competitive prediction accuracy.

1.2 Statistical Models

This book follows the general scientific principle that "all models are wrong but some are useful" (Box, 1979) in the sense that no statistical model will usually be complex enough to fully describe reality. However, statistical models can provide a reasonable approximation to reality, subject to a certain level of abstraction. Statistical models consist of a *structural component* that relies on a mathematical description of how certain input variables (the covariates in a regression model) determine properties of the distribution of a response variable. In contrast to deterministic models, statistical models are equipped with a *stochastic component* that represents the deviation between the real data generating process and the approximation by the model, as well as truly random aspects of the data generation such as measurement errors or uncertainty stemming from random sampling from a population. In the linear model, the structural component (the regression predictor of the model) and the stochastic component (the error term) are nicely separated, whereas this is no longer the case for GAMLSS.

The process of statistical modeling includes the steps of making reasonable assumptions concerning the data generating process, fitting the model to the observed data, checking the validity of the model, and interpreting the results. As emphasized previously, all statistical models are derived through simplifying assumptions. If the assumptions are correct, it is more likely that the conclusions from the model will also be useful.

General requirements for choosing a good model include the model's ability to

- answer the right scientific question,
- highlight important features of the data while ignoring the less relevant,
- provide a good trade-off between fidelity to the data and complexity such that we neither overfit the data (implying restrictions on the generalizability beyond the observed data for predictions) nor underfit the data (implying a potential bias in the conclusions).

In essence, a good statistical model should be able to "allow the data to tell its story."

https://doi.org/10.1017/9781009410076.003 Published online by Cambridge University Press

Exploratory statistical modeling often relies on the idea of trying different models to the data and choosing the most appropriate, while accepting the principle that there could be more than one appropriate and useful model, or occasionally none, that adequately fits the data of interest.

1.3 Regression Models

Regression models are one specific instance of statistical models that consist of a *response* variable y (also denoted as the outcome, dependent variable, or the target variable), a number of *explanatory* variables $\boldsymbol{x}$ (also called predictors, covariates, independent variables, terms), and assumptions as to how the explanatory variables affect the response.

The set of possible values the response variable can assume (i.e. its *support*) is crucial in developing an appropriate model. The most important differentiation is between continuous and discrete responses, but additional differentiations are possible, for example relating to nonnegative responses, responses with skewed distributions, responses with bounded support, mixed discrete–continuous responses, nominal and ordinal discrete responses, etc. We will also consider models for multivariate responses. As we will see in Section 2.6, the support of the response can be the first criterion when choosing an appropriate distribution and we discuss a number of distribution classes in Chapter 2.

For the explanatory variables, we distinguish

- continuous covariates assuming values on the real line or an interval subset of the real line. For example, age and height both assume values on the positive real line. There are occasions when a continuous explanatory variable needs to be transformed. Skew distributed values, unusually large or small values, or the scaling of the explanatory variable are some of the reasons to transform continuous variables;

- spatial covariates representing either continuous coordinate information or discrete spatial information in terms of the assignment to a set of pre-specified regions;

- factors, namely categorical variables which can be *unordered*, for example, eye color where the *levels* of the factor do not have a specific order; or *ordered*, for example, disease level where the levels “severe”, “moderate”, “mild”, “none” do have a specific order. Factors are also used as grouping or clustering variables, for example the identification variable for individuals in longitudinal data.

It is important to remember that not all of the available covariates may be needed to explain the behavior of the response variable. A central part of *statistical modeling* is to determine which of the covariates are indeed important, and in what form(s) they should be included.

https://doi.org/10.1017/9781009410076.003 Published online by Cambridge University Press

1.4 From Linear Models to GAMLSS

In this section, we provide a brief history of the development of regression methodology from the linear model to GAMLSS. Our goal is not to provide a complete literature review or a complete enumeration of all regression approaches, but to motivate the important generalizations of and differences to the most well-known regression set-ups.

1.4.1 The Linear Model

Historically, the most popular regression model is the *linear model*, where the response variable y_i is related to a set of covariates $x_{i1}, \ldots, x_{ip}$ as

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad \text{for } i = 1, \ldots, n, \tag{1.1}$$

where ϵ_i are the *errors* or *disturbances* that quantify the deviations between the structural part of the model $\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ and the observed responses y_i. The very basic assumption about the error terms ϵ_i is that they are i.i.d. with zero mean and constant variance. The additive composition of the error terms and the structural component yields a separable model where additional assumptions on the errors entail easy interpretation and estimation of the model.

Note that in this book we use lower-case letters to denote the responses and covariates, irrespective of whether they are random variables or their realizations. Only in cases where it cannot be easily deduced from the context, will we make explicit notational distinction between random variables and realizations.

An extra assumption for the error terms, which is particularly helpful for uncertainty assessment and hypothesis testing, is that they are i.i.d. realizations from a normal distribution with zero mean and constant variance, that is, $\epsilon_i \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma^2)$. The assumption of zero means implies that the structural part of the model determines the (conditional) expectation of the responses such that

$$\mathbb{E}(y_i | x_{i1}, \ldots, x_{ip}) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip},$$

which also yields the famous ceteris paribus (everything else fixed) interpretation of the regression coefficients: When comparing two observations that differ in one unit in covariate x_j but have the same values for all other covariates, we expect a difference of β_j in the response.

More generally, normally distributed error terms imply

$$y_i | x_{i1}, \ldots, x_{ip} \overset{\text{ind}}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \sigma^2),$$

that is, the responses themselves follow a normal distribution such that inferences for the regression coefficients can be drawn based on the implied normal likelihood. This includes not only point estimates based on maximum likelihood theory, but also the assessment of uncertainties via standard errors, confidence intervals, or statistical tests.

https://doi.org/10.1017/9781009410076.003 Published online by Cambridge University Press

In matrix notation, the linear model is expressed as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.2}$$

where $\boldsymbol{y}$ and $\boldsymbol{\epsilon}$ are n-dimensional vectors of the response variable and error terms, respectively, $\boldsymbol{X}$ is an $(n \times (p+1))$-dimensional design matrix containing the explanatory variables as columns (including a column of ones relating to the intercept β_0), and $\boldsymbol{\beta}$ is the $(p+1)$-dimensional vector of regression coefficients, which shall be estimated from the data. For the vector of error terms, we then obtain the n-dimensional multivariate normal distribution:

$$\boldsymbol{\epsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n),$$

which in particular implies $\mathbb{E}(\boldsymbol{\epsilon}) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n$. Similarly, for the responses, we find

$$\boldsymbol{y} | \boldsymbol{X} \sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n) \tag{1.3}$$

and therefore $\mathbb{E}(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$ and $\mathrm{Cov}(\boldsymbol{y}|\boldsymbol{X}) = \sigma^2 \boldsymbol{I}_n$. The corresponding likelihood for σ^2 and $\boldsymbol{\beta}$ is given by

$$L(\boldsymbol{\beta}, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right).$$

Ignoring σ^2 for the moment, maximizing the likelihood is equivalent to minimizing

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{\epsilon}^\top \boldsymbol{\epsilon} = \sum_{i=1}^{n} \epsilon_i^2,$$

i.e. the least squares criterion that is often also used as the foundation for determining regression coefficients in the linear model without relying on the normal distribution for the error terms. The solution to minimizing the least squares criterion is the ordinary least squares estimator

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}, \tag{1.4}$$

while the maximum likelihood estimator for the variance of the error terms is given by

$$\hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{n}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^\top \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right).$$

Since $\hat{\sigma}^2_{\mathrm{ML}}$ is biased, a commonly used alternative is the unbiased estimator, which can also be derived as a restricted maximum likelihood (REML) estimator:

$$\hat{\sigma}^2_{\mathrm{REML}} = \frac{1}{n-p-1}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^\top \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right).$$

We illustrate a simple linear regression model in action with data from the Fourth Dutch Growth Study Fredriks et al. (2000a,b), a cross-sectional study measuring growth and development in the Dutch population between the ages 0 and 23 years. Figure 1.2(a) shows $n = 3,512$ observations of the body mass index (BMI) and age of boys between 10 and 20 years of age. The response variable is BMI, and there is only

https://doi.org/10.1017/9781009410076.003 Published online by Cambridge University Press

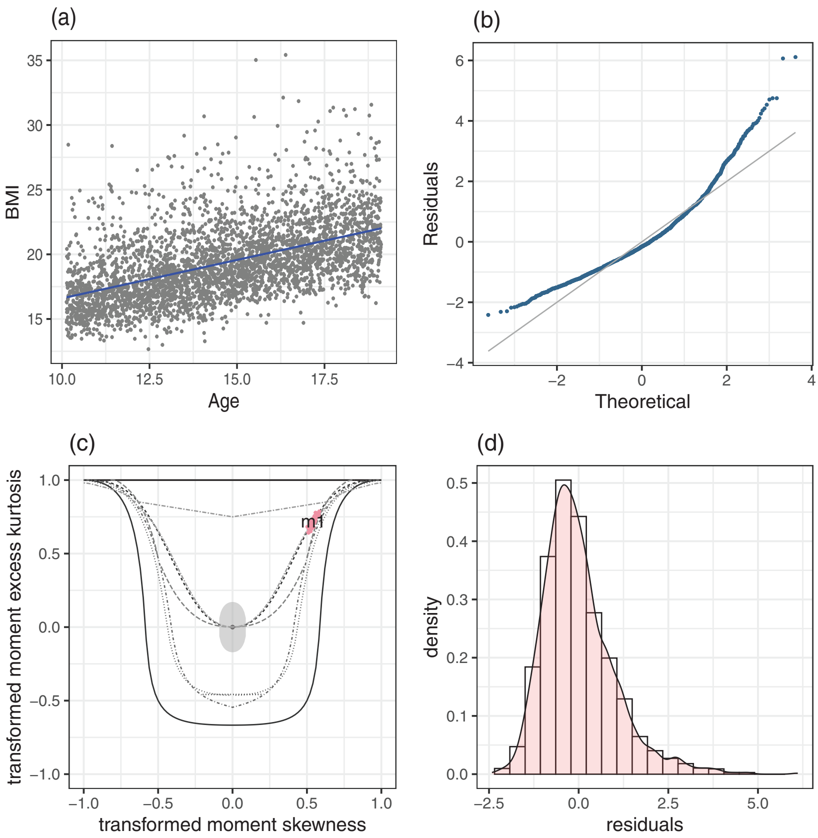

Figure 1.2 Dutch boys' BMI for boys aged between 10 and 20 years: (a) the data and the fitted least squares line; (b) QQ-plot of the residuals from the linear model `m1`; (c) bucket plot for checking the moment transformed skewness and kurtosis of the residuals of model `m1`; (d) histogram and density estimate of the residuals.

one explanatory variable, `age`. The fitted least squares line shown in Figure 1.2(a) is $\hat{y} = 10.87 + 0.577\,\texttt{age}$ and it captures the trend in growth well. We refer to this linear model as `m1`.

Unfortunately, other features in the data are not captured well by the linear model. In Figure 1.2, panels (b) and (c) show residual diagnostics from the linear model. Panel (b) shows a QQ-plot of the normalized quantile residuals (defined in Section 4.7.1) of the linear model, which checks the normality assumption. Most points in the QQ-plot are far from the diagonal line, indicating strong deviations from normality. Panel (c) shows a *bucket* plot, which is a diagnostic graphical tool checking the skewness and kurtosis assumption. (See Section 4.7.3 and also De Bastiani et al. (2022).) The point marked as `m1` in Figure 1.2(c) is the transformed moment skewness (on the x axis) against the transformed moment kurtosis (on the y axis) of the residuals of model `m1`. The cloud of red points around `m1` are 99 bootstrap points of

https://doi.org/10.1017/9781009410076.003 Published online by Cambridge University Press

transformed moment skewness and kurtosis points, obtained by bootstrapping the original residuals. This shows the variability of the measures of skewness and kurtosis of `m1`. The important feature is the shaded area in the middle of the figure around the point $(0, 0)$, which represents the normal distribution. This shaded area is a 95% confidence region based on the Jarque–Bera test for simultaneously testing whether skewness and kurtosis exist in the data. The point `m1` is far from the 95% confidence region of the the Jarque–Bera test, indicating that the residuals of the model `m1` show considerable skewness and kurtosis. The model `m1` has not taken the skewness and kurtosis observed in the data, into account.

Even without diagnostic tools, one can spot more variation above the fitted line in Figure 1.2(a) than below it. To highlight this, Figure 1.2(d) shows a histogram and density estimate of the residuals of the of `m1` model, which highlights considerable skewness. This is unlikely to be modeled adequately by the assumption of normality inherited by the linear model. The fact is, that while we fitted a reasonable model for the location parameter of the data (the mean in this case), the basic assumptions of the linear model are inevitably broken and any inference on the parameters or other features of the model would be affected by this.

Figure 1.2(a) is based on $n = 3,512$ observations for children and teenagers from 10 to 20 years old. The situation becomes more complicated if we consider the original dataset with $n = 7,294$ observations with age range of 0.30 to 22.7 years, shown in Figure 1.3(a). Obviously, the least squares fit of the linear model will fail miserably on the complete data. The curve shown in Figure 1.3(a) was fitted using P-splines (Eilers and Marx, 1996, 2021), one of the techniques we will discuss extensively in Chapter 3. In addition, there are other features in the data which indicate that the assumptions of normally distributed error terms with constant variance, are not appropriate here. There is evidence in Figure 1.3(a) of *heteroscedasticity* (the variance varies with age); *skewness*; and possibly *kurtosis* (since there exist a number of observations further away from the central line, in both directions, suggesting heavier tails than the normal distribution). These features suggest that skewness and kurtosis may also vary with age. The GAMLSS model introduced in Section 1.5 accommodates these features.

Generally, GAMLSS enables us to

- consider a much wider range of response distributions than the normal distribution,
- deal with heterogeneity, possibly covariate-dependent, in various distributional features of the response distribution (not only the mean), and
- relax the assumption of a linear predictor such that various kinds of complex regression relations can be accommodated.

In fact, equation (1.3) provides the simplest example of a GAMLSS, in which normally distributed responses are assumed with purely linear effects on only the mean while the variance is the same for all observations. In Sections 1.4.2 to 1.4.4 we discuss various extensions that preceded the development of GAMLSS. It is convenient

https://doi.org/10.1017/9781009410076.003 Published online by Cambridge University Press

at this point to introduce a different notation for Equation (1.3) that emphasizes that the elements of the response vector are independent but not identically distributed. More precisely, we rewrite model (1.3) as

$$\boldsymbol{y} \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{\mu}, \sigma^2), \qquad \boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}, \tag{1.5}$$

where the notation indicates that each element y_i of $\boldsymbol{y}$ is independently distributed as $y_i \sim \mathcal{N}(\mu_i, \sigma^2)$ for $i = 1, \ldots, n$. The mean parameter $\boldsymbol{\mu}$ is a linear function of the explanatory variables constituting the columns of $\boldsymbol{X}$.

1.4.2 Generalized Linear Models

A big step in the development of distributional regression models was the introduction of Generalized Linear Models (GLMs) by Nelder and Wedderburn (1972). The GLM was popularized by McCullagh and Nelder (1989) and Dobson and Barnett (2018) and also by the introduction of the first interactive statistical package **GLIM**; see, for example, Francis et al. (1993) and Aitkin (2018).

In a GLM, the normal response distribution in (1.5) is replaced by the exponential family of distributions such that

$$\boldsymbol{y} \overset{\text{ind}}{\sim} \mathcal{E}(\boldsymbol{\mu}, \phi),$$

where $\mathcal{E}$ denotes the exponential family, $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{y})$ is the expectation of the response and $\phi > 0$ is a scale parameter. The exponential family includes many important distributions such as the normal, Bernoulli, Poisson, gamma, inverse Gaussian and Tweedie distributions, therefore providing a unifying framework for regression analyses in variety of settings. Most importantly, the framework includes the linear model as a special case, but also allows the analysis of binary responses (based on the Bernoulli distribution), count responses (based on the Poisson distribution), nonnegative continuous responses (based on the gamma and inverse Gaussian distributions), and nonnegative continuous responses supplemented with a positive probability of observing zero (based on the Tweedie distribution).

Regression effects are now assumed for the expectation $\boldsymbol{\mu}$, with a further generalization of the linear model, allowing a monotonic *link function* $g(\cdot)$ that relates $\boldsymbol{\mu}$ to the linear predictor $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$:

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}.$$

This opens up the relationship between $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$ to a variety of shapes not possible under the linear model. For example, if $g(\cdot)$ is the logarithmic function, this implies a multiplicative relationship between the covariates and the mean response since

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

and therefore

$$\mu_i = \exp(\beta_0) \cdot \exp(\beta_1 x_{i1}) \cdots \exp(\beta_p x_{ip}). \tag{1.6}$$

https://doi.org/10.1017/9781009410076.003 Published online by Cambridge University Press

The inverse of the link function $h(\cdot) = g^{-1}(\cdot)$ is called the response function and maps the linear predictor to the expectation of the response, namely

$$\boldsymbol{\mu} = h(\boldsymbol{\eta}).$$

Since the domain of $\boldsymbol{\mu}$ is often restricted (e.g. to the unit interval in case of the Bernoulli distribution or to the positive half axis in case of Poisson, gamma, and inverse Gaussian distributions), the link function also serves as a convenient way of constraining the distribution parameter $\boldsymbol{\mu}$ to the appropriate range when modeled as a function of the explanatory variables.

Note that unlike in the linear model, GLMs in general entail a non-separable structure where the regression predictor (the structural part of the model) cannot easily be disentangled from the random component (the response distribution). Rather, one specific aspect of this distribution (namely the mean) is related to the structural model component.

Unifying various regression models under the umbrella of the exponential family allows us to derive general principles and implementations for statistical inference. In particular, iteratively weighted least squares (IWLS) estimation provides a convenient way of implementing Fisher scoring iterations for determining the maximum likelihood estimate for the regression coefficients $\boldsymbol{\beta}$ (see McCullagh and Nelder (1989) for details). Furthermore, theoretical properties of the exponential family result in asymptotic normality and the validity of likelihood ratio tests. An important property of GLM models (which is shared with the generalized estimating equation approach (Hardin and Hilbe, 2002)) is that it is always consistent in estimating the population mean.[1] The problem is that if the distribution is not correct it could be a very inefficient way of doing so. Another important theoretical implication of assuming a response distribution in the exponential family is that the variance of the responses is intrinsically linked to the expectation based on a variance function that is specific to the chosen member of the exponential family. More precisely, we find

$$\mathbb{V}(y_i) = V(\mu_i)\phi,$$

that is, the variance is determined by the product of a variance function $V(\cdot)$ and the scale parameter ϕ. For example, in case of the normal distribution, the variance function and scale parameter are given by $V(\mu) = 1$ and $\phi = \sigma^2$, providing one example where indeed the variance does not depend on the expectation μ but only on the scale parameter, which then coincides with the error variance. For other members of the exponential family, variance function and scale parameter are given by, for example, $V(\mu) = \mu$, $\phi = 1$ (Poisson distribution), $V(\mu) = \mu(1 - \mu)$, $\phi = 1$ (Bernoulli distribution), and $V(\mu) = \mu^2$, $\phi > 0$ (gamma distribution).

1.4.3 Generalized Additive Mixed Models

While generalized linear models enable considerable flexibility with respect to the response distribution, they keep the restrictive assumption of a purely linear re-

[1] The concept of the population of interest is introduced in Chapter 4.

https://doi.org/10.1017/9781009410076.003 Published online by Cambridge University Press

gression predictor $\boldsymbol{X\beta}$ determining the conditional expectation of the response, via $\boldsymbol{\mu} = h(\boldsymbol{X\beta})$. Various extensions have been introduced to overcome this limitation, following the advent of generalized additive models (Hastie and Tibshirani, 1986, 1990) that expanded the predictor to include nonlinear effects of continuous covariates, yielding

$$\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + s_1(x_{i1}) + \cdots + s_J(x_{iJ}), \tag{1.7}$$

where the $s_j(\cdot)$ are nonlinear smooth functions for the explanatory variables x_{ij}. In this book, we will rely on penalized splines for modeling these nonlinear effects, as discussed in more detail in Section 3.1. In the wake of generalized additive models, it became apparent that a multitude of other effects could be integrated into the regression predictor in similar ways. For example, extended model classes include

- spatial effects $s_{\text{spat}}(z_i)$ where z_i denotes information on the spatial allocation of individual i, in terms of either coordinates or administrative regions,
- varying coefficient terms $x_{i1}s(x_{i2})$, where the effect of x_{i1} (the interaction variable) smoothly varies with respect to the value of covariate x_{i2} (the effect modifier), and
- interaction surfaces $s(x_{i1}, x_{i2})$ of two continuous covariates.

Various approaches for defining such terms will be discussed in Chapter 3.

Figure 1.3(a) shows the benefit of using smoothing techniques (otherwise known as smoothers) for modeling the relationship between an explanatory term and the response. The fitted curve for all $n = 7294$ observations of the BMI dataset fits the trend in the data very well and it is hard to imagine we could have achieved the same effect using parametric curve fitting. [2] The GLM/GAM framework provides three distributions appropriate for modeling a continuous response variable such as BMI: the normal, the gamma, and the inverse Gaussian distributions. We have fitted all three distributions; the fitted values for the conditional mean of those distributions were very similar and indistinguishable from the line shown in Figure 1.3(a), which plots the fitted values for the inverse Gaussian model. The inverse Gaussian model had the lowest AIC[3] value of the three GAMs. Figure 1.3(b) shows the QQ-plot of the normalized quantile residuals from the three GLM/GAM fitted models. None of them fits the data well. Figure 1.3(c) shows the bucket plot of the three fitted distribution models. The points `N`, `G`, and `I` represent the transformed skewness and the transformed kurtosis of the normalized quantile residuals of the three GAM fitted distributions. All points are far from the 95% confidence region of the Jarque–Bera test. This provides additional evidence that none of the three GLM/GAM distributions adequately fits skewness and kurtosis in the BMI data. The inability of the exponential family to model skewness and kurtosis of the BMI data is also partially shown in Figure 1.3(d), in which the centile curves[4] at centiles 3, 10, 25, 50, 75, 90, and 97, of the fitted inverse Gaussian distribution model, are plotted.

[2] Note that, to improve the fit, we used the transformed variable $x = \texttt{age}^{1/3}$ rather than `age`.
[3] Use of the AIC as a way of choosing between models is discussed in Section 4.4.1.
[4] A centile is a quantile multiplied by 100.

https://doi.org/10.1017/9781009410076.003 Published online by Cambridge University Press

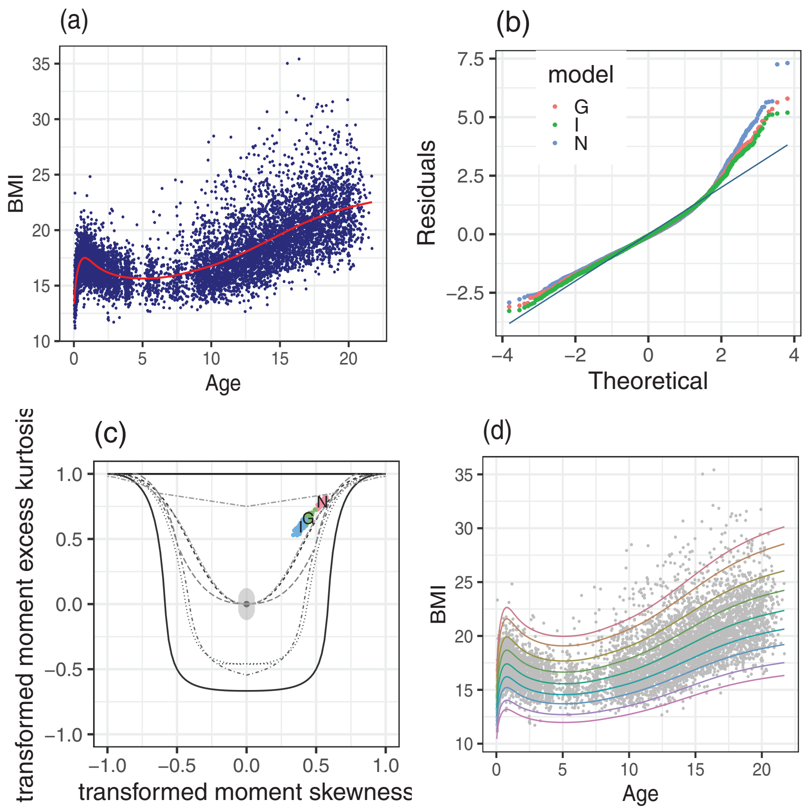

Figure 1.3 Dutch boys' BMI for boys aged between 0 and 23 years: (a) the data and the fitted smooth curve from a GAM model using an inverse Gaussian response distribution; (b) QQ-plots of the residuals from GAMs with normal distribution (`N`) in blue, gamma distribution (`G`) in red, and inverse Gaussian distribution (`I`) in blue; (c) bucket plot of the normalized quantile residuals from the three fitted GAM models: `N` `G`, and `I`; (d) fitted centile curves at centiles values 3, 10, 25, 50, 75, 90, and 97, using the inverse Gaussian GAM model.

In general, we would expect α% of the data to be below the α centile curve and $(100 - \alpha)$% above. For example, in Figure 1.3(d) we would expect 3% of the data to be below the 3 centile curve (the curve at the bottom of the plot). The observed percentage is 1.78%. Above the 97 centile curve (at the top of the plot), we would expect 3% of the data while actually there are 4.2%. The difference does not sound large, but when the centile curves are used for risk stratification, this could be crucial.

Another extension concerns the ability to adjust for potential correlation induced by unobserved heterogeneity associated with grouping structures in the data. In combination with linear predictors, random effect models were introduced by Laird and Ware (1982), and popularized by Pinheiro and Bates (2000). These are models which accommodate correlation between observations through the use of random

https://doi.org/10.1017/9781009410076.003 Published online by Cambridge University Press

intercepts and (optionally) random slopes. (See Section 3.6 for details.) A *mixed effects* model in which fixed effect and random effect coefficients coexist is written as

$$\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \boldsymbol{z}_i^\top \boldsymbol{\alpha},$$

where $\boldsymbol{x}_i$ contains explanatory variables associated with the linear fixed effect coefficients $\boldsymbol{\beta}$, while $\boldsymbol{z}_i$ contains explanatory variables associated with the random effects (coefficients) $\boldsymbol{\alpha}$. The $\boldsymbol{\alpha}$ are assumed to be normally distributed with zero mean and covariance matrix $\boldsymbol{Q}$, namely $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q})$.

The most prominent case for the application of random effects are longitudinal data with repeated observations on the same set of statistical units (subjects, individuals). In this case, within-subject correlation is accounted for by the random effects component. An alternative perspective is that (individual-specific) random effects account for individual-specific, unobserved heterogeneity between the statistical units. However, the application of random effects models is actually much broader since most smoothers $s_j(\cdot)$ in the GAM equation (1.7) (and also most of the extensions mentioned earlier) can be represented as random effects models. This connection of the smoothers to random effect models led to the further understanding and development of smoothers and also to different ways of estimating their smoothing parameters. The family of smoothers which fall into this category were called *structured additive terms* by Fahrmeir et al. (2004).

The combination of random effects with additive model structures leads to generalized additive mixed models, see Ruppert et al. (2003), Wood (2017), and Fahrmeir et al. (2021) for overviews on the state of the art for this model class.

1.4.4 Mean and Dispersion Models

As an important step towards relaxing the common focus on exclusively modeling the mean of the response variable in terms of (possibly complex) regression effects, Aitkin (1987) introduced a model with normally distributed response, in which both the mean and the variance of the model are functions of explanatory variables:

$$\begin{aligned} y_i &\sim \mathcal{N}(\mu_i, \sigma_i^2) \\ g_1(\mu_i) &= \boldsymbol{x}_{i1}^\top \boldsymbol{\beta}_1 \\ g_2(\sigma_i) &= \boldsymbol{x}_{i2}^\top \boldsymbol{\beta}_2, \end{aligned} \tag{1.8}$$

where $\boldsymbol{x}_{i1}$ and $\boldsymbol{x}_{i2}$ are design vectors containing explanatory variables associated with the mean and the standard deviation, while the link functions $g_1(\cdot)$ and $g_2(\cdot)$ are taken to be the identity and log functions, respectively. Smyth (1989) extended model (1.8) with the gamma response distribution. Both authors used maximum likelihood for the estimation of model parameters. Rigby and Stasinopoulos (1996) introduced smoothers into model (1.8). Nelder and Pregibon (1987) considered the more general case of the exponential family, namely, $y_i \sim \mathcal{E}(\mu_i, \phi_i)$, using an extended quasi-likelihood function for parameter estimation.

https://doi.org/10.1017/9781009410076.003 Published online by Cambridge University Press

Mean and dispersion models provide one simple special case of generalized additive models for location, scale, and shape (GAMLSS), in which the mean and dispersion are modeled in terms of (linear) predictors. GAMLSS considerably extends this by allowing potentially all parameters characterizing the response distribution to depend on covariate information.

1.5 Generalized Additive Models for Location, Scale and Shape

1.5.1 GAMLSS as a Distributional Regression Model

In a distributional regression model, the relationship between the response y and the covariates $\boldsymbol{x}$ is of a *stochastic* nature. The response y depends on $\boldsymbol{x}$ through the *conditional* distribution $f(y|\boldsymbol{x})$, which is the main subject of interest since it provides rich information about how the covariates $\boldsymbol{x}$ affect various aspects of the (conditional) distribution of y.

GAMLSS provides a parametric framework for *statistical inference* in distributional regression, in which we approximate the conditional distribution $f(y|\boldsymbol{x})$ by a parametric distribution $f(y|\boldsymbol{\theta}(\boldsymbol{x}))$, where $\boldsymbol{\theta}(\boldsymbol{x}) = (\theta_1(\boldsymbol{x}), \theta_2(\boldsymbol{x}), \ldots, \theta_K(\boldsymbol{x}))^\top$ is a K-dimensional vector of (unknown) model parameters which themselves depend on explanatory terms. The basic idea of statistical inference is to use $f(y|\boldsymbol{\theta}(\boldsymbol{x}))$ to say something sensible about the population distribution $f(y|\boldsymbol{x})$, see Chapter 4.

The notation $\boldsymbol{\theta}(\boldsymbol{x})$ emphasizes that any of the model parameters in $\boldsymbol{\theta}$ can be functions of any of the explanatory variables $\boldsymbol{x}$, not only the mean as in equation (1.5). This is one of the main features of the distributional regression model on which we focus in this book. The implication of such an approach is that the shape of the model distribution for y can change according to the values of explanatory variables $\boldsymbol{x}$. By modeling all the parameters of $f(y|\boldsymbol{\theta}(\boldsymbol{x}))$ as functions of the explanatory terms, we explicitly *simultaneously* model all the characteristics of the distribution including location, scale (variability), quantiles, moments, skewness, and kurtosis. Modeling only the mean allows shifts exclusively in the location of the distribution with all other distribution parameters remaining constant. Consequently, modeling all parameters allows for various types of changes in the shape of the distribution of a response variable, based on one or more explanatory variables (e.g., the age of a child).

The distinction between the basic assumptions of the standard regression model and those of a GAMLSS is shown in Figure 1.4, in which we depict simulated samples of a response variable y with a single explanatory variable x. Figure 1.4(a) demonstrates the distributional assumptions of the linear model. The mean of the normal distribution of the response y varies linearly with x; the shape of the distribution remains the same over the range of x, since the variance of y is constant. Figure 1.4(b) illustrates how the assumptions of a GAMLSS model may operate. A GAMLSS model allows a nonlinear (smooth) relationship between x and the location parameter of the distribution, but also allows all the parameters of the distribution to vary with ex-

https://doi.org/10.1017/9781009410076.003 Published online by Cambridge University Press

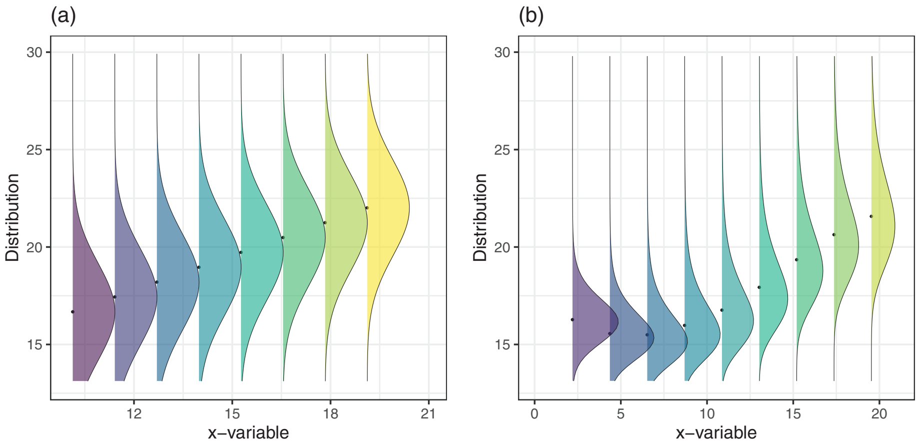

Figure 1.4 Different distributional regression models assumptions: (a) the linear regression model; (b) the GAMLSS model.

planatory terms. It allows the shape of the response distribution to change according to different values of x.

Figure 1.5(a) displays the BMI data and fitted values for the location parameter μ of a GAMLSS model fitted using the Box–Cox t (`BCTo`) distribution, which is a four-parameter distribution introduced by Rigby and Stasinopoulos (2006). The parameters are μ (location parameter, approximately the median), σ (scale parameter, approximately the coefficient of variation), and ν and τ as skewness and kurtosis parameters, respectively (see Section 2.2.2 and Rigby et al. (2019)). The model was fitted using smoothers for all of the parameters as functions of `age`.[5] The QQ-plot of the GAMLSS model, in Figure 1.5(b), shows that the `BCTo` distribution fits the data very well. It is only in the lower tail that a few points deviate from the diagonal line. Given that there are more than 7,000 observations, this behaviour is not unusual. The bucket plot in Figure 1.5(c) shows that the fitted `BCTo` distribution corrects properly for skewness and kurtosis in the data, as the value of the transformed skewness and transformed kurtosis of the residuals from the `BCTo` model fall very close to the origin $(0, 0)$ (representing the normal distribution) and within the 95% confidence region of the Jarque–Bera test. The fitted centiles from the `BCTo` distribution shown in Figure 1.5(d) provide further graphical evidence that the distribution fits well.

In the following, we introduce GAMLSS more formally and discuss the different ingredients of a GAMLSS specification.

1.5.2 Response Distributions

In a GAMLSS model, the responses are assumed to be generated from a K-parametric family of distributions with (covariate-dependent) parameters $\boldsymbol{\theta}(\boldsymbol{x}) = (\theta_1(\boldsymbol{x}), \theta_2(\boldsymbol{x}),$

[5] The transformed variable $x = \texttt{age}^{1/3}$ was fitted instead of `age`.

https://doi.org/10.1017/9781009410076.003 Published online by Cambridge University Press

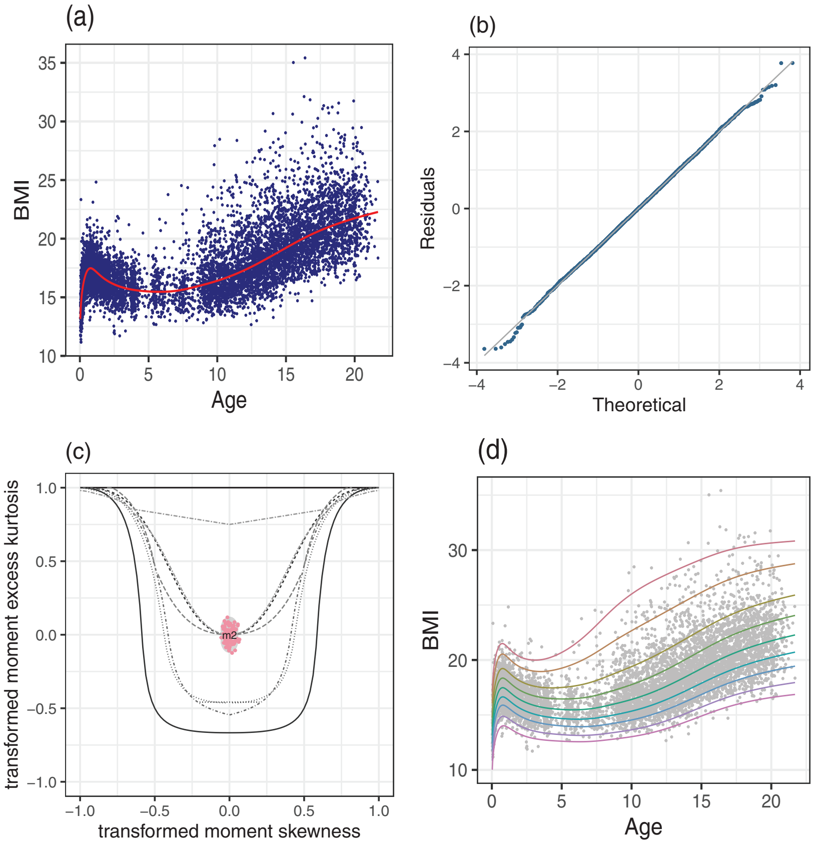

Figure 1.5 Dutch boys BMI for boys aged between 0 and 23 years: (a) the data and the fitted smooth curve from a GAMLSS model with BCT response distribution; (b) QQ-plots of the normalized quantile residuals from the GAMLSS model; (c) bucket plot of the residuals from the GAMLSS models; (d) fitted centile curves fitted at centile values 3, 10, 25, 50, 75, 90, and 97, from the GAMLSS model.

$\ldots, \theta_K(\boldsymbol{x}))^\top$. Subsequently we allow those K parameters to possibly differ for each observation i, for $i = 1, \ldots, n$, so we introduce the notation $\boldsymbol{\theta}_{[i]}(\boldsymbol{x}_i) = (\theta_{i1}(\boldsymbol{x}_i), \theta_{i2}(\boldsymbol{x}_i), \ldots, \theta_{iK}(\boldsymbol{x}_i))^\top$ for $i = 1, \ldots, n$. By suppressing the explicit dependence of θ_{ik} on $\boldsymbol{x}_i$ to keep the notation short, we have $\boldsymbol{\theta}_{[i]} = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iK})^\top$. We assume that there is one common type of distribution applying to all observations (such as normal, Poisson, etc.) but that the K parameters of this distribution are allowed to vary over the individual observations, that is,

$$y_i \overset{\text{ind}}{\sim} \mathcal{D}(\theta_{i1}, \ldots, \theta_{iK}), \qquad \text{for } i = 1, \ldots, n.$$

We denote the density and the cumulative distribution function of this distribution as $f(y_i|\boldsymbol{\theta}_{[i]})$ and $F(y_i|\boldsymbol{\theta}_{[i]})$, respectively. We note that each θ_{ik} for $k = 1, \ldots, K$ may depend on different subsets of $\boldsymbol{x}_i$.

https://doi.org/10.1017/9781009410076.003 Published online by Cambridge University Press

$\mathcal{D}$ denotes the response distribution: The GAMLSS framework allows a multitude of response types, including (but not limited to) (i) models for continuous responses, enabling us to not only deal with but to systematically study phenomena such as heteroscedasticity and skewness, (ii) models for continuous nonnegative responses, potentially featuring a discrete point mass at zero, (iii) count responses potentially featuring zero inflation and/or overdispersion, (iv) continuous fractional or bounded responses (e.g. proportions), again including the option for discrete point masses at one or both endpoints, and (v) multivariate response distributions. More details on potential choices for response distributions are provided in Chapter 2.

For many univariate continuous distributions defined on the real line $\mathbb{R}$, the first two parameters θ_1 and θ_2 are related to location and scale (or dispersion), but this is not always the case. For $K > 2$, the remaining parameter(s) are generally shape parameters, although they may also capture specialized features such as zero inflation. In special cases, θ_3 and θ_4 are true skewness and true kurtosis parameters (see Rigby et al. (2019) for the definitions of these concepts). Rigby and Stasinopoulos (2005) and subsequent publications by those authors use $K = 4$ parameters with the notation μ, σ, ν, and τ, respectively. Note, however, that neither the original definition of GAMLSS (1.10) nor its original fitting algorithms have restrictions on the number of parameters K.

1.5.3 Link Functions

For each of the distribution parameters θ_k, a monotonic link function $g_k(\cdot)$ and corresponding response function $h_k(\cdot) = g_k^{-1}(\cdot)$ relate the regression predictor η_k with the corresponding parameter, namely,

$$\eta_k = g_k(\theta_k) \quad \text{and} \quad \theta_k = h_k(\eta_k) \ .$$

The response function $h(\cdot)$ is often chosen to map the regression predictors from the real line (where the predictor η_k can take its values) to the correct support for θ_k, ensuring that parameters are appropriately constrained. For example, standard deviation and variance have to be positive while parameters representing probabilities are restricted to the unit interval $[0, 1]$. This is an important and useful feature of the link function, but note that link functions reflect the relationship between parameter and covariates. For example, in a model with an identity link for θ_k, the contribution of each explanatory variable to the distribution parameter θ_k is additive, while for a model with a log link the effect is multiplicative, as shown in equation (1.6).

While default choices for the link functions exist (e.g. the logarithmic link for positive parameters such as variances, or the logit or probit links for parameters restricted to the unit interval), it is important to emphasize that the choice for a link function also implies a modeling decision. This decision determines the exact relation between the covariates and the conditional response distribution and has consequences for both the fit of the model and the interpretation of the estimated regression effects. It therefore makes sense to consider competing specifications for the link function and

https://doi.org/10.1017/9781009410076.003 Published online by Cambridge University Press

to include the decision on a specific link function in the model building and model checking process.

An interesting avenue to circumvent the difficulties arising from the need to identify the most appropriate link function, is to consider flexible link functions estimated from the data along with the regression effects of interest. In GLMs, this has, for example, been addressed under the notion of single index models; see Ichimura (1993), where kernel density estimates are used to determine the response function, or Yu and Ruppert (2002), Muggeo and Ferrara (2008), and Yu et al. (2017) who employ penalized splines for the specification of the response function. Estimated link functions have also been combined with additive model specifications, for example in Tutz and Petry (2016) and Spiegel et al. (2019). Another way of relaxing the assumption of one given link function are composite links, where multiple transformations of linear predictors are additively combined to one composite model specification; see Thompson and Baker (1981). In this book, we will not pursue these ideas further but will rather focus on fixed, pre-specified link functions.

1.5.4 Structured Additive Predictors

The simplest case of a GAMLSS is a *fully parametric model*, where a linear predictor is specified for each of the distribution parameters, i.e.

$$\eta_{ik} = \beta_0^{\theta_k} + \beta_1^{\theta_k} x_{i1}^{\theta_k} + \cdots + \beta_{J_k}^{\theta_k} x_{iJ_k}^{\theta_k}$$

leading for n observations to the n-dimensional vector

$$\boldsymbol{\eta}_k = \boldsymbol{X}_k \boldsymbol{\beta}_k$$

of predictor evaluations. While looking rather restrictive at first glance, considerable flexibility can already be achieved in the parametric setting, by considering various types of transformations such as polynomials or interactions. (See also Section 3.8.1.) Still more flexibility is achieved when assuming that each of the regression predictors η_k is additively composed of an intercept $\beta_0^{\theta_k}$ and a sum of J_k functions $s_j^{\theta_k}(\boldsymbol{x}_i)$ (or $s_{jk}(\boldsymbol{x}_i)$ for simplicity), $j = 1, \ldots, J_k$, leading to the structured additive predictor

$$\eta_{ik} = \beta_0^{\theta_k} + s_1^{\theta_k}(\boldsymbol{x}_i) + \cdots + s_j^{\theta_k}(\boldsymbol{x}_i) + \cdots + s_{J_k}^{\theta_k}(\boldsymbol{x}_i) \tag{1.9}$$

or the more compact variant

$$\eta_{ik} = \beta_{0k} + s_{1k}(\boldsymbol{x}_i) + \cdots + s_{jk}(\boldsymbol{x}_i) + \cdots + s_{J_k k}(\boldsymbol{x}_i) \; .$$

The functions $s_j^{\theta_k}(\boldsymbol{x}_i)$ are used as a generic notation that may represent a variety of different effects, as discussed in more detail in the following, and in Chapter 3. In particular, the functions can simply represent a linear effect or more complex effects such as nonlinear effects of continuous covariates, spatial effects, or random effects. Notationally, we allow each function to depend on the complete covariate vector, although in practice each effect will usually only depend on a small subset of $\boldsymbol{x}_i$. However, to avoid notational complexity, we do not make this explicit. Furthermore,

https://doi.org/10.1017/9781009410076.003 Published online by Cambridge University Press

to make the model identifiable, appropriate centering constraints have to be applied to the different functions.

Model equation (1.9) can also include terms that do not fit into the framework of structured additive terms. For example, the local regression smoothers (`loess`) of Cleveland et al. (2017) were part of the original implementation of the GAM models in **Splus** since the early 1990s. Decision trees, neural networks, and the fitting of nonlinear terms have been implemented in the **gamlss** package since 2010. The original GAMLSS algorithms of Rigby and Stasinopoulos (2005) allow the inclusion of any statistical regression-type technique which allows prior weights in its implementation. However, while for the structured additive terms there is a strong theoretical justification (see Chapter 3), the justification for the techniques mentioned above comes from the fact that empirically they work well. Dimensionality reduction techniques such as lasso regression (Tibshirani, 1996) and principal component regression have also been implemented within **gamlss** (Stasinopoulos et al., 2022), see Section 5.4.

There is great potential to be gained by merging some of *machine learning* techniques with distributional regression. Machine learning originated in the computer science world, and as a result its language is somewhat different from that of statistical modeling. It generally encompasses algorithms and computational techniques designed to produce a prediction of an *output* (response variable) on the basis of given *inputs* (explanatory variables). In this respect its aim is similar to statistical modeling. The difference arises from the fact that while statistical modeling aims to interpret and understand the underlying structure of relationships, machine learning takes a *black box* approach. Both approaches can be helpful in different circumstances but caution and knowledge of their limitations are crucial.

1.5.5 Basis Function Representation

We use basis function expansions to represent nonlinear effects in the structured additive predictors, i.e. each function is approximated in terms of a linear combination of basis functions such that (after dropping the parameter index θ_k and the function index j for notational convenience) we obtain

$$s(\boldsymbol{x}_i) = \sum_{l=1}^{L} \gamma_l B_l(\boldsymbol{x}_i),$$

where γ_l are the basis amplitudes while $B_l(\boldsymbol{x}_i)$ represent different types of basis functions (discussed in detail in Chapter 3). In matrix notation, each of the predictors can be written for all observations as

$$\boldsymbol{\eta} = \beta_0 \mathbf{1}_n + \boldsymbol{B}_1\boldsymbol{\gamma}_1 + \cdots + \boldsymbol{B}_J\boldsymbol{\gamma}_J.$$

To enforce specific properties of the function estimates such as smoothness or shrinkage, each parameter vector $\boldsymbol{\gamma}_j$, $j = 1, \ldots, J$, is supplemented by a quadratic penalty

https://doi.org/10.1017/9781009410076.003 Published online by Cambridge University Press

term

$$\text{pen}(\boldsymbol{\gamma}_j) = \lambda_j \boldsymbol{\gamma}_j^\top \boldsymbol{K}_j \boldsymbol{\gamma}_j$$

that is, augmented to the likelihood, $\lambda_j \geq 0$ is the smoothing parameter determining the impact of the penalty and $\boldsymbol{K}_j$ is a positive semi-definite penalty matrix. In a Bayesian framework, the penalty is replaced by the equivalent prior distribution

$$f(\boldsymbol{\gamma}_j | \tau_j^2) \propto \left(\tau_j^2\right)^{-\frac{\text{rank}(\boldsymbol{K}_j)}{2}} \exp\left(-\frac{1}{2\tau_j^2} \boldsymbol{\gamma}_j^\top \boldsymbol{K}_j \boldsymbol{\gamma}_j\right) \mathbb{1}(\boldsymbol{A}_j \boldsymbol{\gamma}_j = \boldsymbol{0}),$$

where the prior variance τ_j^2 is related inversely to the smoothing parameter, the penalty matrix $\boldsymbol{K}_j$ plays the role of a prior precision matrix, and $\boldsymbol{A}_j$ is an appropriate constraint matrix that ensures identifiability of the model. In more general cases, the penalty or prior distribution may involve multiple smoothing parameters and/or it may be notationally convenient to absorb the smoothing parameter into the penalty matrix. We then write $\boldsymbol{K}_j(\boldsymbol{\lambda}_j)$ to emphasize that, indeed, the penalty term $\text{pen}(\boldsymbol{\gamma}_j) = \boldsymbol{\gamma}_j^\top \boldsymbol{K}_j(\boldsymbol{\lambda}_j)\boldsymbol{\gamma}_j$ depends on a (possibly vector-valued) hyperparameter λ.

1.5.6 Compact Summary

The GAMLSS model is expressed in matrix notation as

$$\boldsymbol{y} \overset{\text{ind}}{\sim} \mathcal{D}(\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K) \tag{1.10}$$

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k \tag{1.11}$$

$$\boldsymbol{\eta}_k = \beta_{0k} \mathbf{1}_n + \boldsymbol{B}_{1k} \gamma_{1k} + \cdots + \boldsymbol{B}_{J_k k} \gamma_{J_k k} \tag{1.12}$$

$$\boldsymbol{\gamma}_{jk} \sim \mathcal{N}(\boldsymbol{0}, \tau_{jk}^2 \boldsymbol{K}_{jk}^-), \tag{1.13}$$

where $\mathcal{D}(\theta_1, \ldots, \theta_K)$ is a K-parametric distribution and the vectors $\boldsymbol{\theta}_k$ and $\boldsymbol{\eta}_k$ are of length n, i.e. $\boldsymbol{\theta}_k = (\theta_{1k}, \theta_{2k}, \ldots, \theta_{nk})^\top$ for $k = 1, \ldots, K$. Notice the difference between the K-dimensional vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)^\top$, which represents the distribution parameters in general for any $\boldsymbol{x}$, the K-dimensional vector $\boldsymbol{\theta}_{[i]} = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iK})^\top$, which represents the distribution parameters for the ith observation, and the n-dimensional vector $\boldsymbol{\theta}_k = (\theta_{1k}, \theta_{2k}, \ldots, \theta_{nk})^\top$, which represents the kth distribution parameter for n observations.[6]

The assumptions of the GAMLSS models defined by equations (1.10)–(1.13) are scrutinized and discussed throughout this book. Equation (1.10) concerns the distributional assumptions of a GAMLSS model: Chapter 2 covers some aspects related to the type of distribution appropriate for the distributional assumption, and gives practical advice for choosing an appropriate response distribution. Chapter 3 covers the different terms appropriate for equation (1.12). Link functions appropriate for equation (1.11) are not particularly targeted in this book and are usually chosen by default to map the predictors $\boldsymbol{\eta}_k$ onto the appropriate support of the parameter $\boldsymbol{\theta}_k$.

[6] We use the following terminology: *distribution parameters* for the $\boldsymbol{\theta}$'s; *coefficients* for the β's and γ's; and *hyperparameters* or *smoothing parameters* for the τ^2's or equivalently the λ's.

https://doi.org/10.1017/9781009410076.003 Published online by Cambridge University Press

Chapter 4 introduces general ideas underlying statistical modeling and inference, and some general tools for working with GAMLSS, in particular with respect to model choice and interpretation. Different methods of estimating the parameters of equation (1.12) and the hyperparameters in equation (1.13) are discussed in Chapters 5, 6 and 7.

1.6 Other Distributional Regression Approaches

We outline here some alternative approaches to distributional regression, that is, other approaches of overcoming the focus on mean-based regression analyses. This treatment is by no means exhaustive and focuses on quantile regression and conditional transformation model as specific model classes. More extensive reviews are provided in Kneib (2013) and Kneib et al. (2023)

1.6.1 Quantile Regression

In GAMLSS, the whole distribution of the response is estimated simultaneously, making all its characteristics available to the researcher based on one convenient and coherent model assumption. The downside of this approach is that we are strongly relying on the assumption that we are able to specify a single distribution that fits all the data well. Quantile regression, in contrast, does not aim at inferring all aspects of the conditional distribution of a response variable given covariates, but rather focuses on local features of this conditional distribution, namely conditional quantiles for given quantile levels. As an advantage, it does not require the assumption of a specific response distribution,[7] alleviating the risk of distribution model misspecification.

Under suitable assumptions on the data generating process, quantile regression provides us with consistent and asymptotically unbiased estimates of the underlying population quantiles (given the model is approximately correct). Since a set of quantiles also provides an (indirect) characterization of the conditional response distribution, including the possibility to study features such as variability and skewness, quantile regression is also a distributional regression approach. There is an important general point to be made here. While the distribution-free approach sounds appealing, it makes it more difficult to check the adequacy of the fitted model. It seems that the more assumptions we make, the easier it is to check the adequacy of those assumptions. For example, if we do assume a distribution we can easily define the residuals of the model and through those residuals check the adequacy of the distribution. If we do not specify a distribution, the residuals are more difficult to obtain and therefore it is more difficult to check the model adequacy. Put another way, within statistical modeling there is no free lunch.

In the following, we briefly sketch the basic approach to quantile regression for models with linear predictors, while providing some references for more general model variants at the end of this section.

[7] It does assume that the conditional cdf $F(y|x)$. exists, but does not specify the exact form for it.

https://doi.org/10.1017/9781009410076.003 Published online by Cambridge University Press

The classical linear model specifies the conditional mean of the response variable y given covariate $\boldsymbol{x}$ as

$$\mathbb{E}(y|\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}$$

and estimates for the regression are typically obtained by minimizing the least squares criterion

$$S_2(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^2$$

with respect to the regression coefficients $\boldsymbol{\beta}$. For i.i.d. samples, it is well known that minimizing the sum of absolute deviations from a central tendency measure yields the median, such that it seems natural to define regression medians as the minimizers of the absolute error criterion

$$S_1(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left|y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right|.$$

More generally, considering the asymmetrically weighted absolute error criterion

$$S_q(\boldsymbol{\beta}) = (1-q) \sum_{i:y_i<\boldsymbol{x}_i^\top \boldsymbol{\beta}_q} \left|y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_q\right| + q \sum_{i:y_i\geq\boldsymbol{x}_i^\top \boldsymbol{\beta}_q} \left|y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_q\right| \tag{1.14}$$

for $0 < q < 1$ yields regression quantiles, with the special case $q = 0.5$ reducing to the regression median.

An alternative perspective that emphasizes the model structure underlying quantile regression starts from the regression specification

$$y = \boldsymbol{x}^\top \boldsymbol{\beta}_q + \varepsilon_q,$$

where, instead of assuming $\mathbb{E}(\varepsilon_q) = 0$ as in mean-based regression, we assume that $Q_q(\varepsilon_q) = 0$, that is, the q-quantile of the error term ε_q is assumed to be zero. This implies that

$$Q_q(y) = \boldsymbol{x}^\top \boldsymbol{\beta}_q,$$

that is, the regression predictor determines the q-quantile of the response distribution.

Comparing quantile regression to GAMLSS, the two main advantages of quantile regression are the absence of a global distributional assumption; and the robustness with respect to outliers which is inherent to the definition of quantiles. A typical example in which quantile regression can work better than GAMLSS is when the conditional distribution is bimodal while the assumed GAMLSS distribution is unimodal.

While individual quantiles, estimated using quantile regression, are consistent and asymptotically unbiased, a set of estimated quantiles based on the same data may not be. The locality of the model assumed for quantile regression implies that in fact no globally consistent model can be defined except for the trivial case of $\boldsymbol{\beta}_q \equiv \boldsymbol{\beta}$ independent of the quantile level q. Indeed, if the separate quantile regressions are

https://doi.org/10.1017/9781009410076.003 Published online by Cambridge University Press

not exactly parallel to each other, the fitted quantiles will inevitably cross at some point. In many cases, this will only happen well outside the range of the observed covariates, but for a dense set of quantile levels and small samples, quantile crossing may also be an issue inside the range of observed covariates.

This is a rather unpleasant feature of quantile regression. There are several attempts in the literature to rectify the “crossing” problem. The joint estimation of multiple quantile regressions called *quantile sheets* was introduced by Schnabel and Eilers (2013) as a valuable alternative, but unfortunately the methodology only applies to models with a single explanatory variable; see also Sottile and Frumento (2021) for a recent attempt. The non-crossing is usually achieved by adding more constraints to the estimating function of quantile regression (see equation (1.14)). It is hard though to see how those added constraints will not affect the consistency and unbiasedness of the resulting estimates.

We advocate a “dual” approach, particularly when the focus of the analysis is on the quantiles of the conditional distribution, as for example in centile estimation for growth curves. This consists of fitting a GAMLSS distribution model to the data and then using different quantile regression curves to check it, or vice versa. Note that quantile regression is necessarily restricted to continuous response distributions, while GAMLSS accommodates continuous, discrete, and mixed discrete–continuous distributions.

The estimation of quantile regression models usually relies on linear programming such that flexible extensions, for example, models including random effects or penalized splines, are more difficult to derive. However, one can use the fact that the quantile estimating function of equation (1.14) is identical to an asymmetric Laplace distribution probability function, and use GAMLSS to fit it.[8] An alternative estimation scheme for quantile regression is statistical boosting (see Chapter 7).

Because of its local nature, quantile regression does not have residuals in the conventional sense, neither is there a general measure for goodness of fit. The only residuals are binary residuals, which indicate whether an observation is above or below the fitted quantile curve. While a GAMLSS model is more difficult to find, when it is found, it provides far more information about the data generating mechanism and its properties, and checking of its assumptions is easier.

In summary, the effect of a covariate on any part of the conditional response distribution can be easily assessed with quantile regression without needing the specify a parametric distribution. We illustrate the method on the BMI measurements for Dutch boys. Figure 1.6 shows quantile curves for the quantile levels $q = 0.03$, 0.1, 0.25, 0.5, 0.75, 0.9, and 0.97. Panel (a) shows data on boys aged between 10 and 20 years. A linear term for age has been fitted. Panel (b) shows the complete data, with a nonparametric smoothing term for age. (As previously, the transformed variable $x = \texttt{age}^{1/3}$ was used instead of `age` in the fitting process.) The corresponding

[8] The asymmetric Laplace distribution is a special case of the GAMLSS distribution `SEP3`, with fixed parameters $\sigma = \tau = 1$ and $\nu = [(1 - q)/q]^{0.5}$ where q is the quantile value.

https://doi.org/10.1017/9781009410076.003 Published online by Cambridge University Press

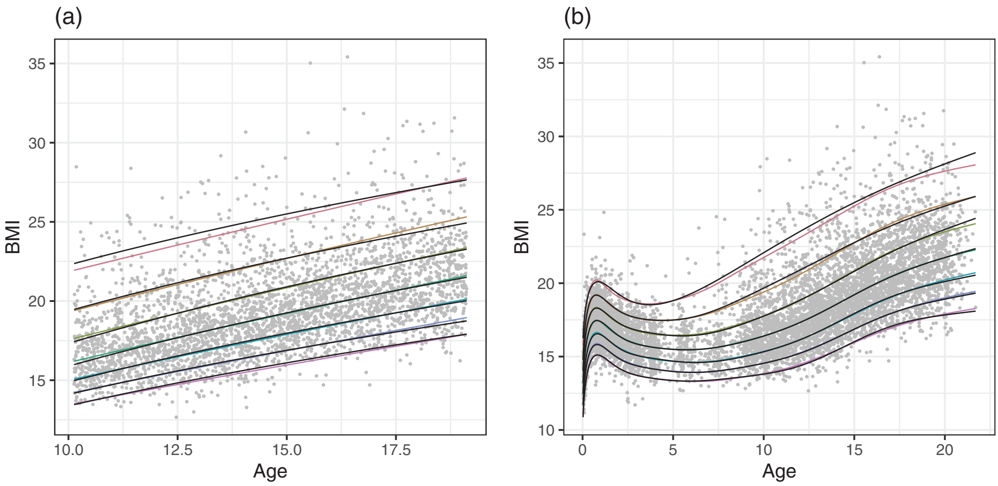

Figure 1.6 Quantile regression performed on the Dutch boys' BMI: (a) boys aged between 10 and 20 years, (b) boys aged between 0 and 23 years. The fitted quantile curves are evaluated at quantiles 0.03, 0.10, 0.25, 0.50, 0.75, 0.90, and 0.97 and are compared in panel (b) to the corresponding centiles of the GAMLSS model fitted using the BCT distribution as in Figure 1.5(d).

centiles of the GAMLSS model fitted using the BCT distribution as in Figure 1.5(d) are also shown, for comparison. As the curves are approximately parallel across the age range, we can conclude that the effect of age on BMI is similar in all regions of the distribution of BMI. The quantile regression was implemented using `gamlss()` with the asymmetric Laplace response distribution, that is, `SEP3` with $\sigma = \tau = 1$ and $\nu = [(1-q)/q]^{0.5}$. This produced similar results to the `qgam()` function in the **qgam** package (Fasiolo et al., 2021).

An extensive treatment of quantile regression methodology is provided in the classical textbook of Koenker (2005). A more recent overview of current developments is available in the *Handbook of Quantile Regression* (Koenker et al., 2020), which also discusses advances on additive quantile regression approaches (see also Fenske et al., 2011; Waldmann et al., 2013; Fasiolo et al., 2021) and ways of circumventing crossing quantiles (see also Chernozhukov et al., 2009; Bondell et al., 2010; Rodrigues and Fan, 2017).

Although quantile regression was originally developed as a distribution-free approach which does not lend itself well to a Bayesian treatment, Bayesian quantile regression has been suggested utilizing the asymmetric Laplace distribution as working model (Yu and Moyeed, 2001). Since the asymmetric Laplace distribution enjoys a representation as a location–scale mixture of normals (Kozumi and Kobayashi, 2011; Yue and Rue, 2011), efficient Bayesian inference can be implemented; this allows complex predictor structures as in GAMLSS (see for example Waldmann et al., 2013).

An alternative to quantile regression is expectile regression, where instead of consider-

https://doi.org/10.1017/9781009410076.003 Published online by Cambridge University Press

ing an asymmetrically weighted absolute error criterion, an asymmetrically weighted squared error criterion is employed. Expectile regression then includes the ordinary least squared (OLS) based mean regression as a special case, but still allows for studying the complete response distribution by varying the asymmetry of the estimation criterion. Expectiles were originally suggested in Newey and Powell (1987) and have regained more interest in recent years due to their ability to accommodate flexible predictor structures (see, for example, Schnabel and Eilers, 2009; Sobotka and Kneib, 2012).

1.6.2 Conditional Transformation Models

Conditional transformation models (CTMs) are another approach to the issue of allowing the conditional distribution to be fully responsive to covariate values. Instead of directly specifying the response distribution of interest, CTMs aim at identifying the required transformation to map the conditional distribution of the responses to a simple reference distribution. This is similar in spirit to earlier attempts such as the Box–Cox transformation, which aims to make the response distribution more normal-like.

In a very general approach, CTMs can be specified as

$$h(Y|\boldsymbol{x}) \overset{\mathcal{D}}{=} Z \sim \mathcal{N}(0,1),$$

where $h(\cdot|\boldsymbol{x})$ is a covariate-dependent transformation function that is strictly increasing in y and which is chosen such that the conditional distribution of the response is matched to a standard normal. Indeed, for continuous distributions, one can show that a unique transformation of this type always exists, if we are flexible enough with respect to $h(\cdot|\boldsymbol{x})$. Note that here we are explicitly denoting random variables as capital letters to ease the understanding of the model specification.

Due to the monotonicity assumed for the transformation function, the model can be inverted to

$$Y \overset{\mathcal{D}}{=} h^{-1}(Z), \quad Z \sim \mathcal{N}(0,1).$$

Another perspective on the model is obtained when looking at the conditional cumulative distribution function (cdf) of the response variable, which is given by

$$F_{Y|\boldsymbol{x}}(y) = \mathbb{P}(Y \leq y|\boldsymbol{x}) = \Phi(h(y|\boldsymbol{x})).$$

Thus the CTM allows us to relate the cdf of Y to the cdf of a standard normal evaluated at a transformed argument. From the cdf, we can directly determine the density

$$f_{Y|\boldsymbol{x}}(y) = \phi(h(y|\boldsymbol{x})) \left| \frac{\partial}{\partial y} h(y|\boldsymbol{x}) \right|$$

which then also gives rise to likelihood-based inference.

The main difficulties with turning CTMs into practice are the choice of a suitable parametrization of the transformation function and the interpretation of the resulting

https://doi.org/10.1017/9781009410076.003 Published online by Cambridge University Press

models. For the former, ensuring monotonicity of the transformation function is the main obstacle, where solutions based on Bernstein polynomials are particularly attractive since monotonicity constraints can then be enforced via linear constraints; see Hothorn et al. (2018) for details. While the original formulation of CTMs is targeted towards univariate, continuous responses, discrete and multivariate versions have also been suggested, see Siegfried and Hothorn (2020) and Klein et al. (2022). In addition to a likelihood-based treatment of CTMs, Bayesian variants (Carlan et al., 2023) and boosting approaches (Hothorn et al., 2014; Hothorn, 2020) are also conceivable.

https://doi.org/10.1017/9781009410076.003 Published online by Cambridge University Press