

bamlss.vis: An R Package to Interactively Analyze and Visualize Bayesian Additive Models for Location, Scale and Shape (bamlss) Using the Shiny Framework

20 week Master thesis as part of the
Master of Science (M.Sc.) course “Applied Statistics”
at the University of Göttingen

Author:

Stanislaus STADLMANN,
Student ID: 21144637

Supervisors

Prof. Dr. Thomas KNEIB
Dr. Nadja KLEIN

Submitted on December 2, 2017
by Stanislaus Stadlmann,
born in Vienna, Austria



GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Contents

1	Introduction	1
2	Motivating Bayesian Additive Models for Location, Scale and Shape	2
2.1	Additive Models	2
2.2	Structured Additive Regression Models	4
2.2.1	Spatial Effects	4
2.2.2	Interaction Terms	5
2.2.3	Random Effects	7
2.3	Generalized Structured Additive Regression Models	7
2.4	Structured Additive Distributional Regression	9
2.4.1	GAMLSS	10
2.4.2	BAMLSS	11
2.5	Estimation	12
2.5.1	Generalized STAR and STAR models	13
2.5.2	GAMLSS	14
2.5.3	BAMLSS	15
3	bamlss.vis	15
3.1	Motivation	16
3.2	Case-Study	18
3.3	Application Structure & Guide	20
3.3.1	Overview Tab	21
3.3.2	Scenarios Tab	22
3.3.3	Plot Tab	24
3.3.4	Scenario Data Tab	26
3.3.5	Properties Tab	27
3.4	Additional Functions	30
3.4.1	Discrete Responses	31
3.4.2	Multivariate Responses	32
3.4.3	Multinomial Responses	34
4	Conclusion	36
A	Appendix	37
	Bibliography	42

List of Figures

1	Probability Density Function of a left-censored normal distribution with the expected value drawn as a blue line.	17
2	Gaussian kernel density estimates for wages split up by education level.	19
3	Button to start the main application of <code>bamlss.vis</code> in RStudio.	20
4	Layout of <code>bamlss.vis</code> after starting the application.	21
5	Expanded overview tab after model selection.	22
6	Scenarios tab of <code>bamlss.vis</code>	23
7	Plot tab output when specifying five different scenarios with different education levels.	24
8	Modal window with formatted and highlighted code after pressing the “Obtain Code!” button	25
9	Scenario data tab.	26
10	Influence of <code>age</code> on the first two moments of the predicted distributions for <code>cnorm_model</code>	28
11	Influence of <code>race</code> categories on the first two moments of the predicted distributions for <code>cnorm_model</code>	29
12	Expected Value and Variance for predicted distributions based on specified covariate combinations.	30
13	Plot tab with predicted probability mass functions for Poisson distributions with three specified covariate combinations in the “Scenarios” tab	32
14	Predicted distribution for a multivariate normal distributed response based on the covariate specification of the “Scenarios” tab.	33
15	“Influence graph” tab for a selected bivariate normal <code>bamlss</code> with three specified covariate combinations.	34
16	Predicted multinomial response in the “Plot tab” based on three specified covariate combinations	35
17	Influence of variable <code>norm1</code> on expected first moments of multinomial distribution	36

List of Tables

1 Introduction

Since the commercialization of the personal computer and the smartphone about two decades later the overwhelming majority of modern life in developing nations has greatly been revolutionized. To name a few advancements, the period stretching from the late 20th century until today has seen changes in the way modern human beings communicate, listen to music, work and are entertained. The common denominator of these changes is the switch from analogue to digital processes, which saw the creation of entire industries, such as Digital Image Processing. The digital revolution also started a significant growth in the number of data collection possibilities and -techniques, with the newest breakthrough, the Internet of Things (IoT), being right around the corner (O'Connor, 2016).

The exponential increase in available data points, paired with dramatic improvements in computing power, gave rise to numerous advancements in statistical sciences. Many computation-heavy models were able to be applied on a broader basis and new methods, such as Neural Nets or Generalized Additive Models could finally be realistically used (The Economist, 2015). With the increase in number of new methods and improvements in data availability, the recent past also saw a significant rise in employed statisticians. In the United States alone, the number of jobs classified as statisticians has increased by more than 120% in the years from 1997 to 2016 (Bureau of Labor Statistics, 2016).

One of the new fields that has emerged is distributional regression, where not only the mean, but each parameter of a response distribution can be modeled using a set of predictors (Klein et al., 2015). Notable frameworks called Generalized Additive Models for Location, Scale and Shape (*gamlss*) and Bayesian Additive Models for Location, Scale and Shape (*bamlss*) were invented by Rigby and Stasinopoulos (2001) in the form of a frequentist perspective and Umlauf et al. (2017) with a Bayesian approach, respectively.

Because methods have become increasingly more complex and capable over the years, it is important to make them accessible and understandable to the growing number of statistical users. In the case of distributional regression models, the interpretation of covariate effects on response moments and the expected conditional response distribution is harder than with traditional methods such as Ordinary Least Squares or Generalized Linear Models, since the moments of a distribution do not directly equate the modeled parameters, but are rather a

combination of them with a varying degree of complexity.

This thesis will introduce a framework for the visualization of distributional regression models fitted using the **bamlss** R package (Umlauf et al., 2017) as well as display an implementation as an R extension titled `bamlss.vis`. The goal of this framework is the ability to:

- See and compare the expected distribution for chosen sets of covariates and
- View the direct relationship between moments of the response distribution and a chosen explanatory variable, given a set of covariates.

Additionally, the user can obtain the code which created the graphs to potentially reproduce them later. The implementation will be done using the statistical software R (R Core Team, 2017) in the form of a Shiny application (Chang et al., 2017).

2 Motivating Bayesian Additive Models for Location, Scale and Shape

Bayesian Additive Models for Location, Scale and Shape (`bamlss`) are a form of Bayesian regression models in which every parameter of a parametric distribution with K parameters is related to a set of additive predictors. The distribution does not have to follow the exponential family, which extends the distributions available for modeling beyond the ones used in Generalized Linear Models (GLM). In similar fashion to Generalized Additive Models (GAM, Hastie and Tibshirani, 1990), the additive predictors can assume different shapes, including non-linear, fixed, random and spatial effects (Umlauf et al., 2017).

To give a sufficient depiction of this model class, this section will start with explaining Additive Models and then gradually generalize the broader frameworks to finally arrive at `bamlss`. Furthermore, a brief overview of the different estimation techniques for the covered model frameworks will be given.

2.1 Additive Models

Bamlss can be seen as a generalization of Structured Additive Regression, which are in turn a generalization of Additive Models. Additive Models, first proposed

by Friedman and Stuetzle (1981) represent a model type in which a dependent variable y is related to a set of non-parametric predictors in an additive way. Assuming conditional independence of y_1, \dots, y_n given the explanatory variables $\mathbf{z}_1, \dots, \mathbf{z}_K$, we obtain the following model equation:

$$y_i = f_1(z_{i1}) + f_2(z_{i2}) + \dots + f_k(z_{ik}) + \epsilon_i \quad (2.1)$$

where $f_j(\cdot)$ depict unspecified non-parametric functions of covariate z_j , which can include smoothing splines or local regression approaches. This makes additive models more flexible compared to standard linear regression, while still being more interpretable than non-additive models (Buja et al., 1989).

Fahrmeir et al. (2013) suggest that an Additive Model can also include parametric components. Given covariates $\mathbf{x}_1, \dots, \mathbf{x}_Q$, we can extend (2.1) to a semi-parametric regression model with the following specification:

$$y_i = \sum_{j=1}^K f_j(z_{ij}) + \underbrace{\sum_{l=1}^Q \beta_l x_{il}}_{\beta_0 + \beta_1 x_{i1} + \dots + \beta_Q x_{iQ}} + \epsilon_i \quad (2.2)$$

Eq. (2.2) combines non-parametric and parametric components. Because the model would otherwise not be identified, functions $f_j(\cdot)$ now have to be centered around zero, such that

$$\sum_{i=1}^n f_1(x_{i1}) = \dots = \sum_{i=1}^n f_K(x_{iK}) = 0$$

holds. The functions $f_j(\cdot)$ are approximated using basis functions in the following scheme:

$$f_j(z_j) = \sum_{m=1}^{d_j} \mathbf{B}_m(z_j) \gamma_{jm}$$

This allows to write the Additive Model in a matrix form, indifferent of the chosen basis:

$$\mathbf{y} = \sum_{j=1}^K \mathbf{Z}_j \boldsymbol{\gamma}_j + \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (2.3)$$

Here, the design matrices $\mathbf{Z}_1, \dots, \mathbf{Z}_K$ represent the basis functions assessed at different covariates. \mathbf{X} is constructed in equivalence to the standard linear regression model. Assumptions about the error term of a semi-parametric Additive Model

are also similar to the classic linear model, where ϵ_i are identically and independently (i.i.d) normally distributed with $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$. These properties are then also valid for the response variable, so that $y_i \stackrel{i.i.d.}{\sim} N(\hat{y}, \sigma_y^2)$ (Fahrmeir et al., 2013, chap. 9.1).

2.2 Structured Additive Regression Models

The nonparametric components in additive models open the possibility for more flexible relationships between the dependent variable and single explanatory variables, which standard linear regression methods might not capture correctly. However, sometimes the area of model application requires even more flexibility, e.g. by including spatial covariates, fixed/random effects or interaction terms. These specific types of effects extend the Additive Model to a Structured Additive Regression Model (Fahrmeir et al., 2003, STAR). This chapter will briefly describe its different components.

2.2.1 Spatial Effects

Similarly to Section 2.1, observations $(y_i, \mathbf{z}_i, \mathbf{x}_i)$ are given, where \mathbf{z}_i and \mathbf{x}_i represent vectors of covariate values for the i th observation. Additionally, a geographic location index s is known with observations s_i , which can be either discrete (e.g. region or country) as well as continuous (e.g. longitude/latitude). Extending the semi-parametric Additive Model as specified in (2.2), a geospatial effect is now added:

$$\begin{aligned} y_i &= \sum_{j=1}^K f_j(z_{ij}) + \sum_{l=1}^Q \beta_l x_{il} + f_{geo}(s_i) + \epsilon_i \\ &= \kappa^{add} + f_{geo}(s_i) + \epsilon_i \end{aligned} \tag{2.4}$$

κ^{add} includes the non-spatial effects from (2.2). The spatial effect, $f_{geo}(\cdot)$, is often viewed as a proxy for unknown covariates, such as altitude or climate data. If the geographic location index s is tracked using discrete values, $f_{geo}(\cdot)$ could represent a Markov random field. For continuous values, smoothing techniques such as Kriging (Matheron, 1963) or a multivariate tensor product spline are available. In both the discrete and the continuous case, the vector of geo-additive

components \mathbf{f}_{geo} can be written as

$$\mathbf{f}_{geo} = \mathbf{Z}_{geo}\boldsymbol{\gamma}_{geo}$$

so that it can be incorporated into the geoadditive model in matrix notation in the following way

$$\mathbf{y} = \sum_{j=1}^K \mathbf{Z}_j \boldsymbol{\gamma}_j + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_{geo}\boldsymbol{\gamma}_{geo} + \boldsymbol{\epsilon} \quad (2.5)$$

which bears similarities to the basis function approach in (2.3) (Fahrmeir et al., 2013, chap. 9.2).

2.2.2 Interaction Terms

The regression equation (2.2) of Additive Models included main nonparametric and parametric effects, but no interactions between covariates. When incorporating interaction effects, one has to differentiate between an interaction between a continuous and a categorical variable, as well as one where two continuous variables share a common effect (Fahrmeir et al., 2013, chap. 9.3).

To illustrate the first case, it is assumed that z_1 and x_1 are continuous and binary ($x_i \in (0, 1)$) covariates, respectively. Then, the interaction term $f_{z_1|x_1}(z_1) \cdot x_1$ can be included in the Additive Model from (2.2) in the following way:

$$y_i = \sum_{j=1}^K f_j(z_{ij}) + \sum_{l=1}^Q \beta_l x_{il} + \underbrace{f_{z_1|x_1}(z_{i1})x_{i1}}_{\begin{array}{ll} 0 & \text{if } x_{i1}=0 \\ f_{z_1|x_1}(z_{i1}) & \text{if } x_{i1}=1 \end{array}} + \epsilon_i$$

If $x_1 = 0$, the non-linear effects of z_1 are now

$$\begin{aligned} f_1(z_1) &\quad \text{if } x_1 = 0 \\ f_1(z_1) + f_{z_1|x_1}(z_1) + \beta_1 &\quad \text{if } x_1 = 1 \end{aligned}$$

This framework can also incorporate spatially covarying terms, where the interaction term $f_{geo|x_1}(s)$ represents an interaction between the location variable s and a categorical variable x_1 (Fahrmeir et al., 2013).

Using a Basis function approach, the vector of interaction effects

$$\mathbf{f}_{int} = (f_{z_1|x_1}(z_{11})x_{11}, \dots, f_{z_1|x_1}(z_{n1})x_{n1})$$

can also be described in matrix notation to extend (2.3) in the following way:

$$\mathbf{y} = \sum_{j=1}^K \mathbf{Z}_j \boldsymbol{\gamma}_j + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_{int} \boldsymbol{\gamma}_{int} + \boldsymbol{\epsilon}$$

Here, the design matrix \mathbf{Z}_{int} represents the Basis function values multiplied with x_1 observations (Fahrmeir et al., 2013, chap. 9.3).

The possibility of interactions between two continuous covariates is also given. In this case, the interaction between z_1 and z_2 is modeled using a two-dimensional nonparametric function $f_{z_1|z_2}(z_1, z_2)$. Common two-dimensional functions include bi-variate smooth splines and Kriging techniques. When only the two-dimensional functions without main effects ($f_1(z_1)$, $f_2(z_2)$) should be included, the model equation assumes the following form:

$$y_i = f_{z_1|z_2}(z_{i1}, z_{i2}) + f_3(z_{i3}) + \dots + f_K(z_{iK}) + \sum_{l=1}^Q \beta_l x_{il} + \epsilon_i \quad (2.6)$$

For reasons of identifiability, $f_{z_1|z_2}(z_1, z_2)$ also needs to be centered around zero. Fahrmeir et al. (2013, chap. 9.3.2) warn that for estimation of models with two-dimensional surfaces a high sample size with combinations of z_1 and z_2 is required. In cases where this requirement is not fulfilled, a simple main effects model as in (2.2) is preferred.

It is also possible to model the interaction effect of z_1 and z_2 using the two-dimensional surface $f_{z_1|z_2}(z_1, z_2)$ while still including the main effects. In this scenario, the model is specified as follows:

$$y_i = f_{z_1|z_2}(z_{i1}, z_{i2}) + f_1(z_{i1}) + f_2(z_{i2}) + \sum_{j=3}^K f_j(z_{ij}) + \sum_{l=1}^Q \beta_l x_{il} + \epsilon_i \quad (2.7)$$

The identifiability problem in this model is more complex than before. To solve it, Fahrmeir et al. (2013, chap. 9.3) state that not only all included functions have to be centered around zero, but also “all slices of the interaction $f_{z_1|z_2}(z_1, z_2)$, i.e. all one-dimensional smooths with fixed value of z_1 or z_2 ”. Using the basis function approach, the matrix representation of the model can be obtained:

$$\mathbf{y} = \sum_{j=1}^K \mathbf{Z}_j \boldsymbol{\gamma}_j + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_{z_1|z_2} \boldsymbol{\gamma}_{z_1|z_2} + \boldsymbol{\epsilon} \quad (2.8)$$

with interaction term design matrix $\mathbf{Z}_{z_1|z_2}$ (Fahrmeir et al., 2013, chap. 9.3).

2.2.3 Random Effects

When dealing with repeated measures or other longitudinal datasets it is often necessary to model cluster-specific similarities using Random Effects (Laird et al., 1982). Additive Models can also be extended with Random Effects to arrive at so called Additive Mixed Models. Assuming a longitudinal data structure with subjects $j = 1, \dots, n_i$ in clusters $i = 1, \dots, m$ and covariates \mathbf{x}_k , a parametric random coefficient model possesses the following structure:

$$y_{ij} = (\beta_0 + \nu_{0i}) + (\beta_1 + \nu_{1i})x_{ij1} + \dots + (\beta_Q + \nu_{Qi})x_{ijQ} + \epsilon_i$$

The “random” coefficients ν_{0i} (intercept) and $\nu_{1i}, \dots, \nu_{Qi}$ (slopes) represent the cluster-specific deviations from the main effects. To obtain Additive Mixed Models, the main effects are then replaced with nonparametric functions:

$$y_{ij} = f_1(x_{ij1}) + \dots + f_Q(x_{ijQ}) + \nu_{0i} + \nu_{1i}x_{ij1} + \dots + \nu_{Qi}x_{ijQ} + \epsilon_i \quad (2.9)$$

Like non-parametric main effects, Random Effects also have a matrix notation. In the case where every main effect is also modeled with cluster-specific effects, the matrix form of Additive Mixed Models is as follows:

$$\mathbf{y} = \sum_{j=1}^K \mathbf{Z}_j \boldsymbol{\gamma}_j + \mathbf{R}_0 \boldsymbol{\nu}_0 + \sum_{j=1}^K \mathbf{R}_j \boldsymbol{\nu}_j + \boldsymbol{\epsilon}$$

Here, $\boldsymbol{\nu}_0 = (\nu_{01}, \dots, \nu_{0m})'$ and $\boldsymbol{\nu}_j = (\nu_{j1}, \dots, \nu_{jm})'$ represent the Random Effects coefficients. A more in-depth look at the structure of the design matrices is given by Fahrmeir et al. (2013, chap. 9.4, p. 550)

2.3 Generalized Structured Additive Regression Models

Structured Additive Regression (STAR) models extend simple Additive Models with special model terms briefly introduced in the previous sections. These effects include:

- Nonlinear effects of z_1

- Spatial effects of location index s
- Interactions between continuous covariate z_1 and a categorical variable x_1
- Nonlinear interactions between two continuous covariates z_1, z_2
- Random Effects with intercept ν_0 and slope ν_j deviations from main effects

All of the aforementioned model terms can be included in a STAR interchangeably, including simple linear predictors $\mathbf{x}'\boldsymbol{\beta}$ (Fahrmeir et al., 2013, chap 9.5).

STAR models provide very flexible ways of modeling the influence of explanatory variables on a given response variable y_i . Note that while the components can be nonparametric, the direct modeling of y_i assumes that the response variable follows a Gaussian distribution. However, when dealing with e.g. binary or categorical responses, this assumption is violated. Then, a type of model specification is needed that directly upholds the dependent variables' support (Olsson, 2002, chap. 2). To solve this challenge, STAR models are merged with Generalized Linear Models to Generalized STAR models.

Generalized Linear Models (GLM), first coined by Nelder and Wedderburn (1972), introduce a framework where the expectation of response y is related to a linear predictor $\eta = \mathbf{x}'\boldsymbol{\beta}$ via a link function $\eta = g(E(y)) = g(\mu)$ or a response function $h = g^{-1}$ to arrive at the following model specification:

$$\begin{aligned}\mu_i &= h(\mathbf{x}'_i\boldsymbol{\beta}) \quad \text{or} \\ g(\mu_i) &= \mathbf{x}'_i\boldsymbol{\beta}\end{aligned}\tag{2.10}$$

When modeling a binomially distributed response the probability parameter π , which has a support of $\pi \in [0, 1]$, is related to predictors $\mathbf{x}'\boldsymbol{\beta}$. Using a logistic link function, we obtain a Logistic Regression Model:

$$\begin{aligned}\eta_i &= \mathbf{x}'_i\boldsymbol{\beta} \\ E(y_i) = \pi_i &= \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}\end{aligned}$$

Here, the response function ensures the correct support of π (Fahrmeir et al., 2013, chap. 5). Using a link function, the expectation of y can be the first moment of many different continuous or discrete distributions, which includes the Poisson, Binomial and Gamma distribution. However, all possible distributions need to be part of the exponential family (Rigby and Stasinopoulos, 2005).

Note in (2.10) that the effects of covariates $\mathbf{x}_1, \dots, \mathbf{x}_K$ are modeled parametrically. Generalized Additive Models (GAM), as suggested by Hastie and Tibshirani (1990), extend the class of Generalized Linear Models to allow for non-parametric effects. In particular, the linear predictor $\eta = \mathbf{x}'\boldsymbol{\beta}$ is interchanged by smooth non-parametric functions $f_j(x_j)$. Given response variable y and covariates $\mathbf{z}_1, \dots, \mathbf{z}_K$, the following model specification is obtained:

$$\begin{aligned}\eta_i &= \sum_{j=1}^K f_j(x_{ij}) \\ \mu_i &= E(y_i) = h(\eta_i)\end{aligned}\tag{2.11}$$

Now, many different response distributions as well as flexible effects for explanatory variables are supported to create a highly flexible model framework. In (2.11), only non-parametric effects are linked to η_i . However, given response y and covariates $(\mathbf{x}_i, \mathbf{z}_i)$, all specific effects of STAR models (spatial effects $f_{geo}(\cdot)$, interactions $f_{int}(\cdot)$, etc.) as well as parametric coefficients can be combined to form a Generalized Structured Additive Regression Model (Generalized STAR):

$$\begin{aligned}\eta_i &= f_1(z_{i1}) + \dots + f_K(z_{iK}) + \beta_0 + \beta_1 x_{i1} + \dots + \beta_Q x_{iQ} \\ \mu_i &= h(\eta_i)\end{aligned}\tag{2.12}$$

In semi-parametric Generalized STAR models, $f_j(\cdot)$ can have any of the structural forms described in Chapter 2.2. Modeled response variables also have to follow an exponential family distribution (Fahrmeir et al., 2013, chap. 9.5).

2.4 Structured Additive Distributional Regression

Generalized STAR models provide a framework to flexibly estimate the expected value of a previously specified distributional parameter. However, in many cases not only the first moment, but also higher-order moments are of special interest. In modeling income, for example, not only the expected income but also the shape of the overall distribution is important. A common measure for income inequality is the Gini coefficient, which can be calculated using the cumulative distribution function (cdf) (Lerman and Yitzhaki, 1984).

2.4.1 GAMLSS

First modeling approaches which go beyond the mean of a distribution were suggested by Nelder and Pregibon (1987) using parametric functions of explanatory covariates related to the dispersion parameter ϕ of an exponential family distribution. Building upon this approach, Generalized Additive Models for Location, Scale and Shape (gamlss) were introduced by Rigby and Stasinopoulos (2001). Gamlss combine the flexibility of being able to model multiple distributions with parametric or nonparametric explanatory effects and extend them for multiple response distribution parameters such that not only the location, but also the scale and shape of a distribution can be modeled simultaneously. Furthermore, gamlss relax the assumption of y following an exponential family distribution, which significantly increases the number of response modeling possibilities.

Assuming a dependent variable from a distribution with parameters $\theta_1, \dots, \theta_L$ and observations y_1, \dots, y_n , given covariates $\mathbf{z}_1, \dots, \mathbf{z}_K$ and $\mathbf{x}_1, \dots, \mathbf{x}_Q$, a gamlss can be described with the following model specification:

$$g_l(\theta_{il}) = \eta_{il} = \mathbf{x}'_{il}\boldsymbol{\beta} + \sum_{j=1}^{K_l} f_{jl}(z_{ijl}) \quad (2.13)$$

In Equation (2.13), $g_l(\cdot)$ represents a known monotonic link function, which can be different for each parameter. \mathbf{x}'_{il} depicts the subset of x variables used to model parameter θ_l in observation i , while $f_{jl}(z_{ijl})$ serves as a non-parametric effect of covariate z_j on parameter θ_l , taken from a subset of the K z variables, evaluated for the i th observation. The specific subset of covariates z with nonparametric effects on parameter θ_k has a length of K_l variables (Stasinopoulos and Rigby, 2007).

As shown above, gamlss can utilize different combinations of parametric and non-parametric effects to model each distributional parameter. Equation 2.13 displays a case in which every parameter is modeled using a non-empty subset of variables x and z . However, some parameters can also be set to a constant and not be dependent on covariates. For example, when assuming the Gaussian distribution for the dependent variable and connecting μ to parametric effects \mathbf{x}_j using the identity link function ($g(\mu) = \mu$) and the variance parameter σ^2 to a constant, we arrive at a linear model specification (Stasinopoulos and Rigby, 2007).

2.4.2 BAMLSS

As mentioned in the introduction of this thesis, not always do the modeled parameters directly equate the moments (location, scale and shape) of a distribution, but rather a combination of them. For this reason, approaches to simultaneously model the parameters of a distribution are often referred to as distributional regression, which includes `gamlss`. However, as seen in (2.13), `gamlss` in its normal form only incorporate main effect modeling. To further integrate structured additive terms, such as spatial effects, random effects and interaction terms (Brezger and Lang, 2006), distributional regression is further extended to Structured Additive Distributional Regression (Klein et al., 2015).

In 2013, Klein et al. introduced Bayesian Additive Distributional Regression, which is a model type extending `gamlss` to include structured additive predictors for modeling parameters of a specified distribution. It represents a fully Bayesian approach, in which coefficients are obtained by drawing samples from the approximate posterior effect distributions using using Markov Chain Monte Carlo (MCMC) simulations.

An implementation of Bayesian Additive Distributional Regression, called Bayesian Additive Models for Location, Scale and Shape (`bamlss`) was since created by Umlauf et al. (2017). As the authors point out, the name bears resemblance to `gamlss`, because of many similarities in its modeling approach. However, extensions of `bamlss` over `gamlss` are manifold. First, parallel to the proposed framework of Klein et al. (2013), MCMC simulations are utilized for estimation of coefficients. This is done in contrast to `gamlss`, where predictor coefficient estimates are retrieved via penalised likelihood maximization techniques. Advantages of using MCMC simulations over likelihood-based approaches include the sample-based inference, which yields more reliable confidence intervals than the intervals of `gamlss` estimates based on asymptotic properties. Second, `bamlss` offer more flexibility of specifying covariate effects with the support of structured additive predictors, like spatial effects or two-dimensional splines. Third, `bamlss` also support multivariate response distributions, which enhances `gamlss`' univariate response framework. Furthermore, the implementation of `bamlss` is designed in a way that allows for the usage of external estimation algorithms and software packages like JAGS or BayesX.

The model specification of `bamlss` is similar to the `gamlss` class. The parameters

$\theta_1, \dots, \theta_L$ of a parametric distribution \mathbf{y} with observations y_1, \dots, y_n are linked to structured additive predictors using monotonic and twice-differentiable link functions $g_l(\theta_l)$ (note that the paper uses $h_l(\theta_l)$). Based on covariates $\mathbf{x}_1, \dots, \mathbf{x}_Q$, the following model equation can be obtained:

$$g_l(\theta_l) = f_{1l}(\mathbf{x}_{1l}; \boldsymbol{\beta}_{1l}) + \dots + f_{Ql}(\mathbf{x}_{Ql}; \boldsymbol{\beta}_{Ql}) \quad (2.14)$$

Here, $f_{jl}(\cdot)$ represent unspecified functions that can attain any structured additive predictor forms, including nonparametric effects. It is also possible to describe the effects in vector form:

$$\mathbf{f}_{jl} = \begin{bmatrix} f_{jl}(\mathbf{x}_1; \boldsymbol{\beta}_{jl}) \\ \vdots \\ f_{jl}(\mathbf{x}_n; \boldsymbol{\beta}_{jl}) \end{bmatrix} = f_{jl}(\mathbf{X}_{jl}; \boldsymbol{\beta}_{jl})$$

with \mathbf{X}_{jl} ($n \times m_{jl}$) specifying the design matrix for effect $f_{jl}(\cdot)$ so that they integrate themselves into the following model equation

$$g_l(\boldsymbol{\theta}_l) = \boldsymbol{\eta}_l = \mathbf{f}_{1l} + \dots + \mathbf{f}_{Jl} \quad (2.15)$$

where \mathbf{f}_{jl} represents the j th effect of \mathbf{x}_{jl} (subvector of \mathbf{x}) on parameter θ_l . Similar to Chapters 2.1 and 2.2, effects in bamllss can also be derived through a basis function approach, such that it can be written as $\mathbf{f}_{jl} = \mathbf{X}_{jl}\boldsymbol{\beta}_{jl}$. The structure of the design matrix depends on the types of covariates and prior assumptions about $f_{jl}(\cdot)$ (Umlauf et al., 2017). As mentioned earlier in this chapter, bamllss offer very flexible ways of specifying covariate effects. Breaking through the framework of basis function approaches, bamllss also allow covariate functions $f_{jl}(\cdot)$ which are nonlinear in its parameters $\boldsymbol{\beta}_{jl}$. An example of this is the Gompertz growth curve

$$\mathbf{f}_{jl} = \beta_1 \cdot \exp(-\exp(\beta_2 + \mathbf{X}_{jl}\beta_3))$$

with nonlinear parameters $\boldsymbol{\beta}_{jl}$ (Umlauf et al., 2017).

2.5 Estimation

In this section, different estimation techniques for the regression models introduced in Section 2 will be discussed.

2.5.1 Generalized STAR and STAR models

Structured Additive Regression Models (STAR) including simple Additive Models (AM) can be estimated using backfitting, an algorithm first coined by Friedman and Stuetzle (1981). The intention of backfitting is to compute estimates for the model terms by the iterative smoothing of partial residuals. Originating from the STAR model equation

$$\mathbf{y} = \sum_{j=1}^K \mathbf{f}_j + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

one can describe \mathbf{f}_j as

$$\mathbf{f}_j \approx \mathbf{y} - \sum_{l \neq j} \mathbf{f}_l - \mathbf{X}\boldsymbol{\beta}$$

Thus, estimates $\hat{\mathbf{f}}_j, \hat{\boldsymbol{\beta}}$ can be obtained by running the following steps:

1. Initialize by setting $\hat{\mathbf{f}}_1 = \dots = \hat{\mathbf{f}}_K = \hat{\boldsymbol{\beta}} = 0$.
2. Compute estimates $\hat{\mathbf{f}}_j$ by

$$\hat{\mathbf{f}}_j = \mathbf{S}_j(\mathbf{y} - \sum_{i \neq j} \hat{\mathbf{f}}_i - \mathbf{X}\hat{\boldsymbol{\beta}})$$

3. Compute estimates $\hat{\boldsymbol{\beta}}$ through

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \sum_{j=1}^K \hat{\mathbf{f}}_j)$$

4. Repeat Steps 2-3 until the iteration deviations between estimates $\hat{\mathbf{f}}_j$ are sufficiently small.

Note that \mathbf{S}_j can be any smoothing function. In the case of penalized smooth splines, this could be the penalized least squares estimate $\mathbf{S}_j = (\mathbf{Z}'\mathbf{Z} + \lambda\mathbf{K})^{-1}\mathbf{Z}'$ with smoothing parameter λ , design matrix \mathbf{Z} and penalty matrix \mathbf{K} . The estimated smoothing parameter λ can be computed by optimizing a model choice criterion, for example GCV or AIC (Fahrmeir et al., 2013).

Generalized STAR models, which use a link function to relate structured additive predictors to expected values of exponential family distributions, are estimated based on a maximum-likelihood approach. To obtain coefficient estimates, the

penalized log-likelihood criterion

$$l_{pen}(\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_K, \boldsymbol{\beta}) = l(\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_K, \boldsymbol{\beta}) - \frac{1}{2} \sum_{j=1}^K \lambda_j \boldsymbol{\gamma}'_j \mathbf{K}_j \boldsymbol{\gamma}_j \quad (2.16)$$

is maximised in regard to $\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_K, \boldsymbol{\beta}$. The first part of the criterion, $l(\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_K, \boldsymbol{\beta})$ depicts the log-likelihood known from Generalized Linear Models, but with structural additive effects.

As suggested by Fahrmeir et al. (2013), penalized log-likelihood criterion of (2.16) can in some cases be directly maximised. For example, an R implementation for Generalized Additive Models (which can be seen as a sub-variant of STAR models), `mgcv` (Wood, 2011), uses penalized iteratively reweighted least squares to directly obtain coefficient estimates. Optimisation of the penalized log-likelihood can also be combined with backfitting, as it is done in the R package `gam`, written by Chambers and Hastie (1992).

2.5.2 GAMLSS

Generalized Additive Models for Location, Scale and Shape (`gamlss`) also rely on the penalized likelihood in (2.16) to obtain coefficient estimates. For maximisation, Rigby and Stasinopoulos (2005) present two algorithms: the “RS” algorithm, a generalization of the algorithm suggested by Rigby and Stasinopoulos (1996) for fitting mean and dispersion additive models, and “CG”, which represents a generalization of the algorithm coined by Cole and Green (1992) relying on first and second (cross) derivatives of the likelihood function with regard to the distributional parameters.

RS, abbreviated for Rigby and Stasinopoulos, uses iteratively weighted least squares (IRLS) in combination with a modified backfitting algorithm to arrive at coefficient estimates. To illustrate the algorithm system, it is broken up into inner and outer iterations, with each inner iteration depicting the fitting of one distributional parameter $\boldsymbol{\theta}_l$. Here, a working variable \mathbf{z}_l consisting of all used predictors, the first derivative of the likelihood (score function) and “iterative weights” \mathbf{w}_l , determined with a local scoring algorithm, is calculated. Then, the working variable is fit to the explanatory variables using backfitted weighted least squares and penalized weighted least squares for parametric and nonparametric coefficients, respectively. The inner iteration is repeated until the inner global

deviance has converged. This procedure is done for every θ_l , after which one outer iteration is finished. The outer iterations are further repeated, until the outer global deviance has also converged (Stasinopoulos et al., 2017, Chap. 3).

Note that while the RS algorithm uses score functions to update the working variable, the cross derivatives of the log likelihood functions with respect to the distributional parameters θ_l are not needed. The CG algorithm, however, uses cross derivatives to obtain new weights for the iteratively reweighted least squares estimation. Other contrasts between CG and RS include the iteration system of CG, where new weights w_l and the working variable z_l for iteratively reweighted least squares is not updated for fitting of every parameter, but represents the outer iteration. In the inner iterations, distributional parameters are fitted after another relying on given weights depending on the current working variable. Here, backfitting is again used to compute current estimates of regression coefficients. After the inner and outer iterations converged, model estimation is finished (Stasinopoulos et al., 2017, Chap. 3). As stated by Stasinopoulos et al. (2017), the RS algorithm is preferred in most cases since it is usually faster and converges more consistently than CG. In cases where the modeled distribution possesses highly correlated parameters, however, RS can be slower and suffer from premature convergence.

2.5.3 BAMLSS

3 bamlss.vis

The previous Sections 2.1 to 2.5 gave a description of Bayesian Additive Models for Location, Scale and Shape (bamlss) and the underlying sub-models on which they are based. This section will introduce a framework to interactively visualize covariate effects and distributional predictions of fitted bamlss models and feature its implementation as an R package. Because of the visual component, the tool will be called bamlss.vis. A small case-study based on wages of male workers in the Mid-Atlantic region will be presented to feature most of bamlss.vis' abilities.

3.1 Motivation

As discussed in previous sections, distributional regression is concerned with modeling the parameters of a known parametric distribution. After estimation of the model, the user obtains coefficients which measure the influence of an explanatory variable on η_l , which represents the transformed parameter θ_l . However, in most cases the user is not interested in specific distributional parameters but more in the moments, which often do not directly equate the parameters but are rather a combination of them.

This problem can be well illustrated using the censored normal distribution. Assume a normally distributed variable, $y \sim N(\mu, \sigma^2)$. Then, the probability density function (pdf) of a left-censored normal distribution y^* with cut-off point $a = 0$ can be obtained by

$$f(y^* = x) = \begin{cases} f(y = x) & x > 0 \\ F(y = \frac{-\mu}{\sigma}) & x \leq 0 \end{cases}$$

where $f(y)$ and $F(y)$ are the probability density functions (pdf) and the cumulative distribution function (cdf) of normally distributed variable y , respectively. It is visible in the above equation that the censored normal distribution is both discrete and continuous. While y^* shares the density with y above the cut-off point, the full remaining density in the censored normal distribution is assigned to the cut-off point a . Figure 1 shows a sample left-censored normal distribution y^* created from $y \sim N(0, 1)$ with $a = 0$ (Greene, 2012).

As visible in Figure 1, the moments of the standard normal distribution do not carry over to the censored normal distribution. In fact, while $E(y) = 0$, the expected value of y^* is $E(y^*) \approx 0.399$. To be exact, the censored normal distri-

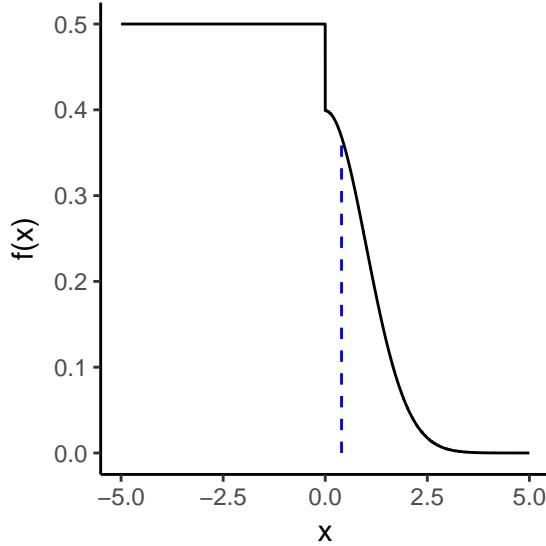


Figure 1: Probability Density Function of a left-censored normal distribution with the expected value drawn as a blue line.

butions first two moments with cut-off $a = 0$ can be calculated as follows:

$$E(y^*) = (1 - \alpha) \cdot (\mu + \sigma\beta) \quad \text{and}$$

$$Var(y^*) = \sigma^2(1 - \alpha) \cdot [(1 - \gamma) + (\frac{-\mu}{\sigma} - \beta)^2 \cdot \alpha]$$

$$\begin{aligned} \text{while: } \alpha &= \Phi(\frac{-\mu}{\sigma}) \\ \beta &= \frac{\phi(\frac{\mu}{\sigma})}{1 - \alpha} \\ \gamma &= \beta^2 - \beta \cdot (\frac{-\mu}{\sigma}) \end{aligned} \tag{3.1}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density function (pdf) and cumulative distribution function (cdf) of the standard normal distribution and μ and σ^2 are the parameters of y , respectively (Greene, 2012). Equation (3.1) shows that both the expected value and the variance of y^* are computed by a combination of the parameters of the original variable y , μ and σ^2 , and are not equal. Thus, an explanatory variable that has a positive effect on μ has both an impact on $E(y^*)$ and $Var(y^*)$. Therefore, coefficients for measuring covariate influences on those parameters are not directly translatable to the moments of the modeled distribution and might even have critically different estimates.

Furthermore, even in cases where the desired moments directly equate the modeled parameters (e.g. in gaussian or poisson-distributed responses), different link

functions for their transformation $g_l(\theta_l)$ and possibly highly complex nonparametric effects of explanatory variables can lead to coefficient estimates that are hard to interpret. In this case, a visual comparison of predicted distributions would be helpful.

To tackle both of the aforementioned interpreting problems with fitted bamllss models, this thesis will introduce a framework with two main objectives:

- Visually compare the predicted distributions (pdf or cdf) based on interactively selected covariates
- View the changes of distribution moments over the whole range of a selected variable, based on chosen explanatory covariates.

Using bamllss.vis, one can then observe the influence of a covariate on the distribution by 1. its cdf or pdf and 2. its moments.

3.2 Case-Study

While automatic testing of bamllss.vis' main functions relies on artificial data for each supported distribution in order to prove correct behavior, presenting the apps' abilities is best done with a dataset of real observations. This chapter will focus on fitting a bamllss based on "real" data for further use in bamllss.vis. The objective for a suitable dataset was that its response variable and explanatory variables are easy to understand for people without a specific scientific background. The chosen dataset, "Wage" from the ISLR R package (James et al., 2017), perfectly encompasses these requirements. "Wage", collected by the United States Census Bureau (2011), includes 3000 male individuals with records of the following variables:

- **wage**: Workers raw wage (in 1000 \$)
- **age**: Age of worker
- **year**: Year that wage information was recorded
- **race**: A factor with levels 1. White, 2. Black, 3. Asian and 4. Other
- **education**: A factor with levels 1. < HS Grad, 2. HS Grad, 3. Some College, 4. College Grad and 5. Advanced Degree

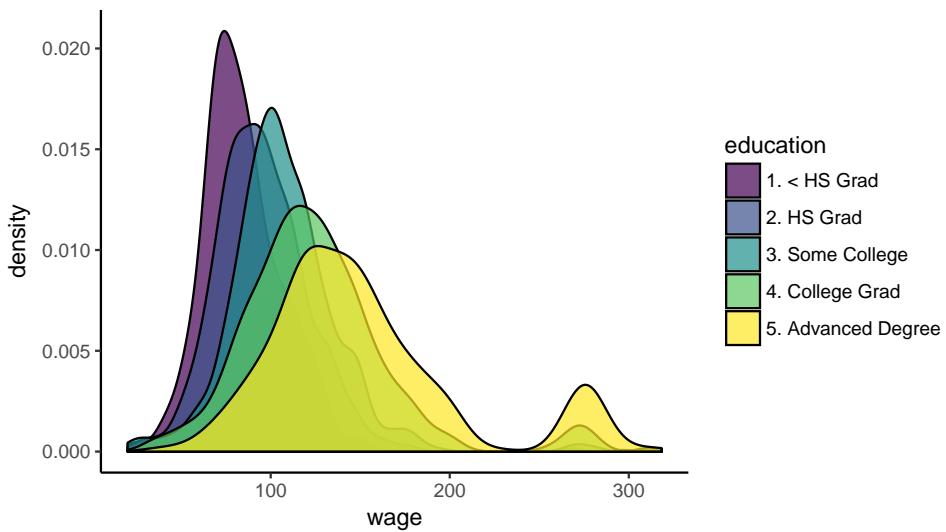


Figure 2: Gaussian kernel density estimates for wages split up by education level.

- **health:** A factor with levels 1. `<= Good` and 2. `>= Very Good` indicating health level of worker

Naturally, the variable of interest and response variable will be the male workers wage. While doing first analyses, it is clear that the wage is highly dependent on the given variables. Figure 2 shows kernel density estimates (Gaussian) for the wage distribution depending on education level.

As visible in Figure 2, the kernel density estimates are critically different for each education level. In general, we can observe the trend that a higher education level leads to a higher expected income, but also to an increased variance. Therefore, both location and shape will be modeled when fitting the bamlss. Because income cannot be smaller than zero but does otherwise not have upper limits, the censored normal distribution with cut-off $a = 0$ will be chosen as the response family. After some data preparation, model estimation can then be achieved with the bamlss R package (Umlauf et al., 2017):

```

1 model <- bamlss(
2   list(wage ~ s(age) + race + year + education + health,
3        sigma ~ s(age) + race + year + education + health),
4   data = wage_sub,
5   family = cnorm_bamlss()
6 )

```

Code-Chunk 1: R code for fitting the bamlss based on Wage dataset

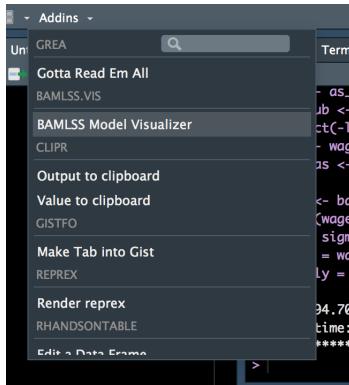


Figure 3: Button to start the main application of bamlss.vis in RStudio.

As visible in Code-Chunk 1, both μ and σ are modeled such that they relate to explanatory variables additively. Both parameters are connected to parametric effects `race`, `year`, `education` and `health`. The influence of `age` is specified with a thin-plate smooth spline.

3.3 Application Structure & Guide

As previously mentioned, bamlss.vis is implemented in the form of an R extension. For building and maintaining the package, GitHub is used. This allows users to easily install the package with the following R commands:

```
1 if (!require(devtools))
2   install.packages("devtools")
3 devtools::install_github("Stan125/bamlss.vis")
```

Furthermore, bamlss.vis is strongly based on the Shiny framework (Chang et al., 2017), which is an R package designed to create interactive visualizations with HTML code and R functions. In the words of the author, Shiny combines “the computational power of R with the interactivity of the modern web” (RStudio, Inc., 2017).

In its core, a Shiny application is built using R functions and can therefore be called similarly. In the case of bamlss.vis, there are two ways one can start the application. First, the user can run the code `bamlss.vis::vis()`. Second, bamlss.vis can also be called using the open source General User Interface RStudio. As displayed in Figure 3, one can click on the “Add-Ins” button and then select “BAMLSS Model Visualizer” if bamlss.vis is installed. This will also trigger the command `bamlss.vis::vis()`.

After executing the R code, a new browser window with the started application

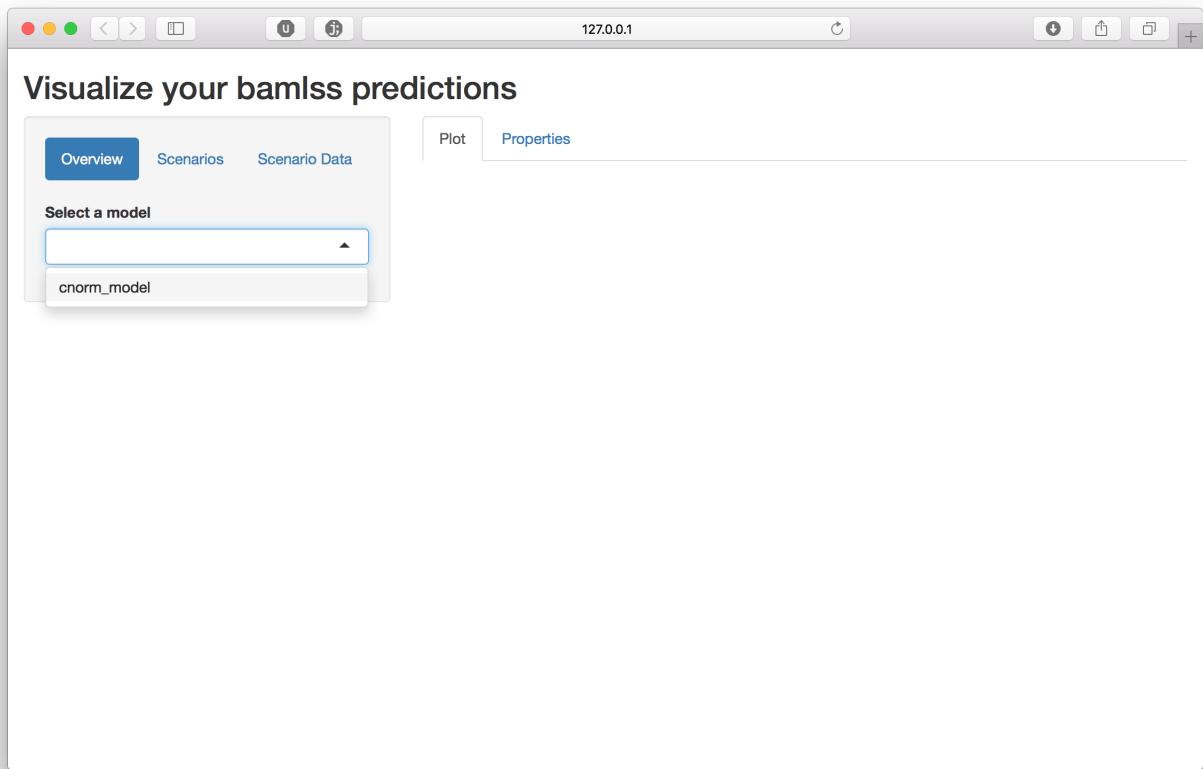


Figure 4: Layout of bamlss.vis after starting the application.

will be opened up. Figure 4 shows the layout of the application, which is then displayed in the users browser. As visible in Figure 4, the layout of bamlss.vis is divided into two segments, which have their own tabs the user can click on. In each segment, one of those tabs is always displayed. The left segment, with tabs “Overview”, “Scenarios” and “Scenario Data”, is concerned with model-related settings. The right segment, with tabs “Plot” and “Properties” is used to display graphs and properties in reaction to user inputs on the left segment.

3.3.1 Overview Tab

The overview tab is meant for displaying descriptive details about fitted bamlss. After bamlss.vis is started up, it only consists of a select list, where the user can select the model on which the further analysis is to be based. Entries in this list are created by an R function which searches the working directory of the current user for any object of the class `bamlss`. Figure 4 shows only one entry, `cnorm_model`, which represents the model fitted with the code provided in Section 3.2.

After a model was selected, the overview tab expands to show an outline of

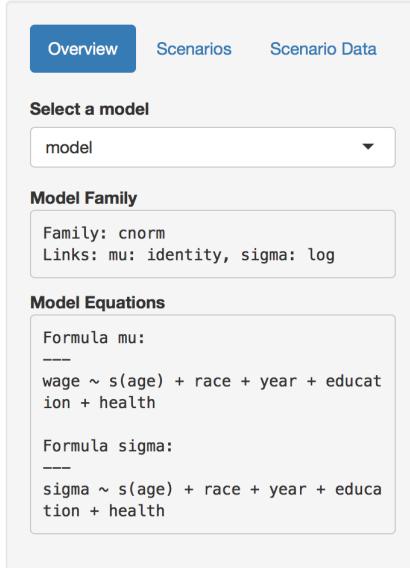


Figure 5: Expanded overview tab after model selection.

the fitted bamlss. Specifically, as shown in Figure 5, the tab displays two parts, called “Model Family” and “Model Equations”. “Model Family” shows the distributional family of the model, as well as the parameters which can be modeled including their link functions. In the case of `cnorm_model`, the family “`cnormal`” (for censored normal distribution) with parameters μ and σ and link functions “identity” and “log” can be obtained. “Model Equations” displays the way covariate effects were specified. We can confirm that for `cnormal_model`, the effect of age on both μ and σ is specified with a smooth spline. Furthermore, all other effects are included parametrically.

3.3.2 Scenarios Tab

After a bamlss was successfully selected in the first tab, one can then click on the “Scenarios” tab. In this tab, the user can specify covariate values for each explanatory variable and create one or more “Scenarios”, for which the predicted distribution is then plotted.

As displayed in Figure 6, the top of the tab consists of two buttons, “Create Scenario” and “Clear Scenarios”. Right below these buttons a box for including the intercept in predictions is portrayed. Further below, web widgets for each explanatory variable are visible. Bamlss.vis executes a check for the type of each explanatory variable and then constructs different web application elements depending on that information. Categorical covariates receive selector boxes (R function `shiny::selectInput()`) with the variables’ possible categories, while for

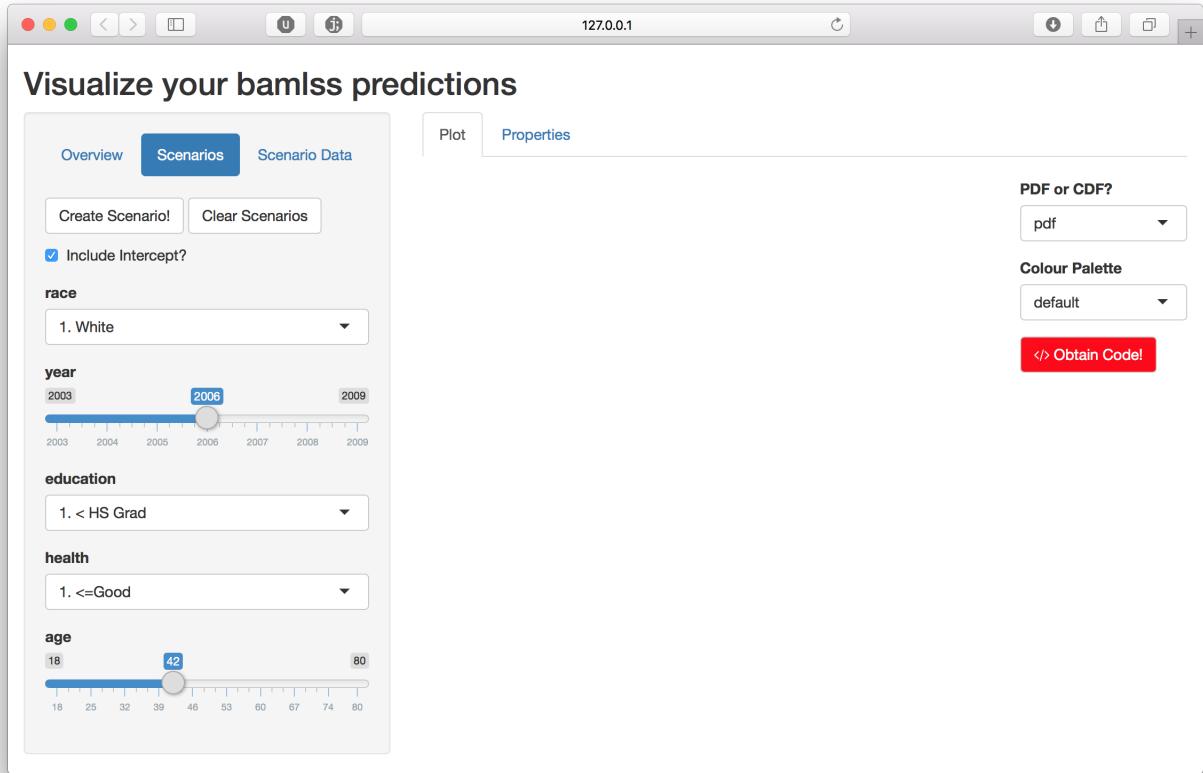


Figure 6: Scenarios tab of bamlss.vis.

numeric variables slider modules are created (`shiny::numericInput()`), ranging from the variables minimum to maximum value. The default value for numeric covariates is its arithmetic average.

To add a new scenario, one can select a value for each variable and then click on “Create Scenario”. When more than one scenario was created, predictions will be computed for each scenario, such that comparisons between scenarios are easily obtained.

Because critical differences in wage distributions depending on education were already observed in Figure 2 of Section 3.2, it will now be interesting to recreate this plot by visually comparing the modeled distributions depending on `education`, while still controlling for other variables. To achieve this, the following covariate values will be specified: `race: 1. White`, `year: 2006`, `education: 1. < HS grad`, `health: 2. >= Very Good` and `age: 42` ($= \bar{age}$). Then, the “Create Scenario” Button is pressed. This is done four more times, with each time seeing a rise in education level by one category. Thus, we can now view different wage distributions for a 42-year-old white male with very good health depending on his level of education (Figure 7).

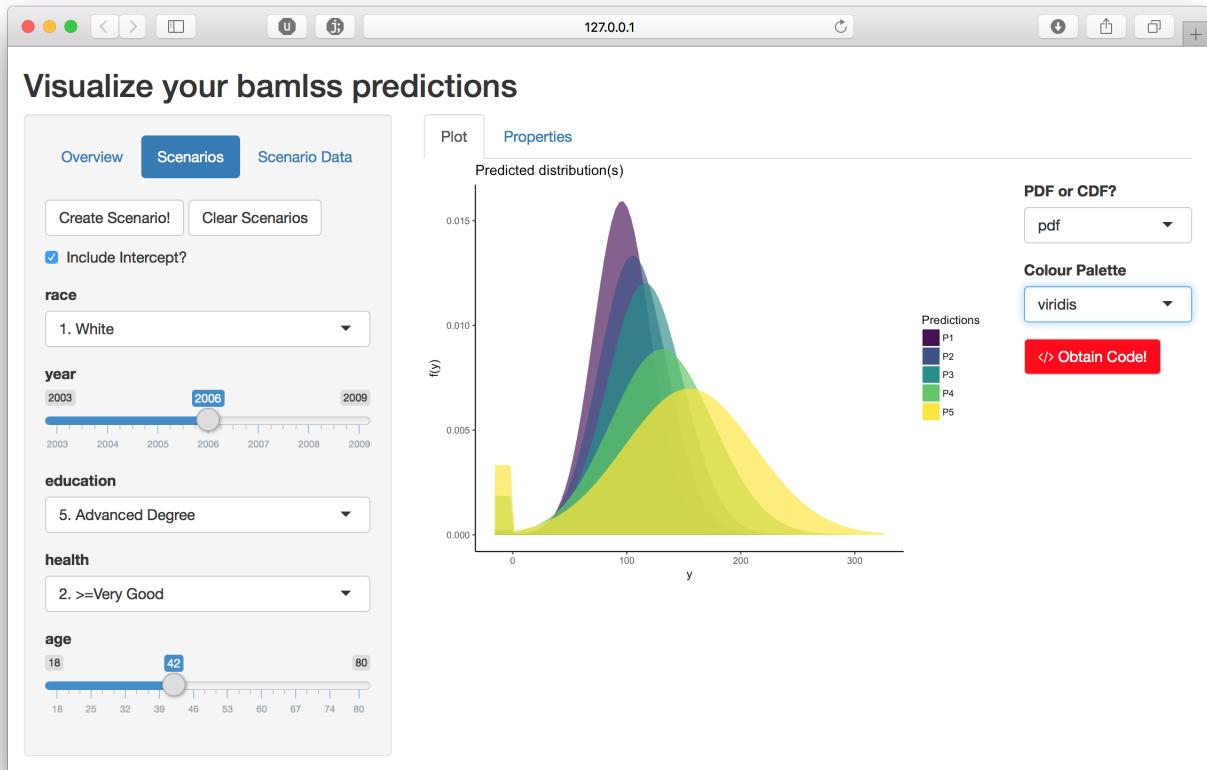


Figure 7: Plot tab output when specifying five different scenarios with different education levels.

3.3.3 Plot Tab

The “Plot” tab, which is located in the right segment of `bamlss.vis`, entirely reacts to user interaction in the Scenarios tab. Every time the “Create Scenarios” button is pressed, the tab is updated. Specifically, the data which the user inputs in the left tab is passed onto `bamlss.vis::preds()` (a customized version of `bamlss::predict.bamlss()`), which then computes predictions for the response distribution parameters by taking the arithmetic average of transformed MCMC samples from the posterior distribution. Afterwards, the predicted parameters are inserted into the probability density function for graphically visualizing the predicted distribution. This procedure is repeated for each “Scenario”.

Figure 7 shows the plot output for five different scenarios based on the `Wage` dataset described in Section 3.3.2. As visible in the aforementioned figure, the predicted distributions have both a higher expected value and a higher variance as the education level rises, similar to the kernel density estimates in Figure 2.

Also noticeable in Figure 7 to the right side of the plot are three web elements

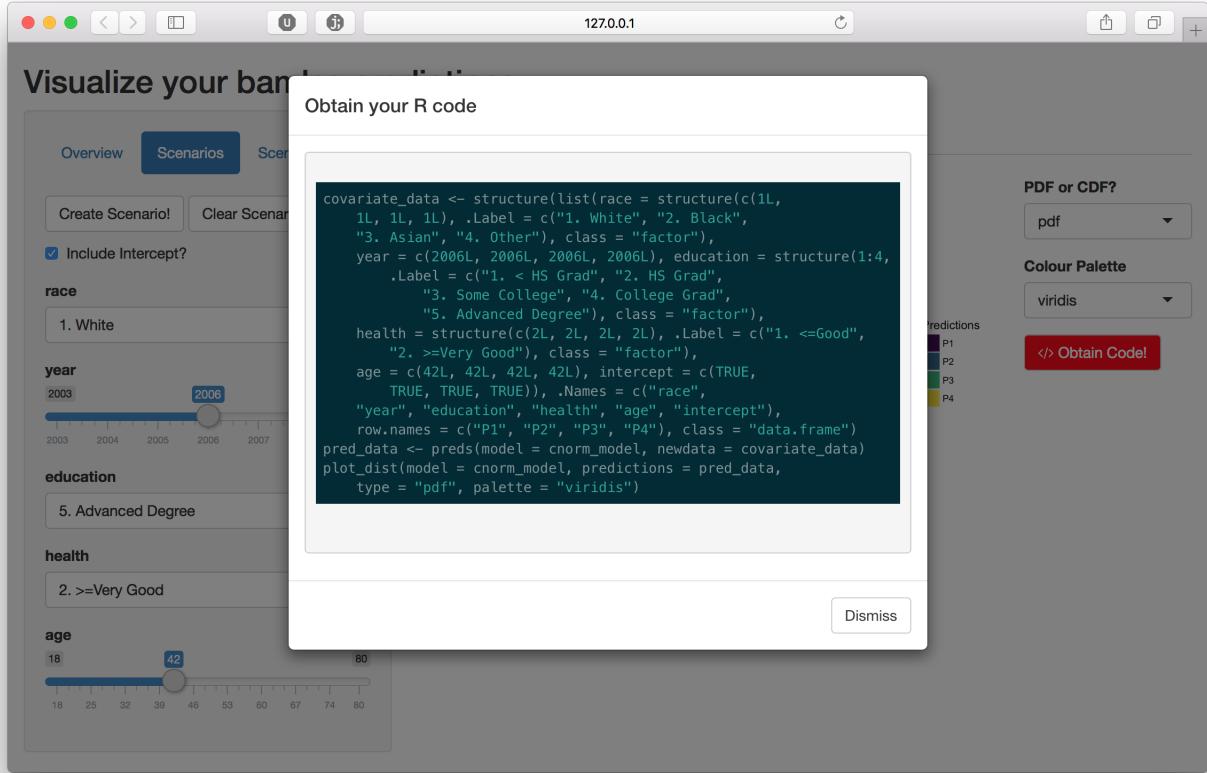


Figure 8: Modal window with formatted and highlighted code after pressing the “Obtain Code!” button

for user interaction. The first element, found below the description “PDF or CDF?”, provides the ability to switch between displaying probability density function (default) and the cumulative distribution function (Appendix: Figure A.1). The second element gives the option to select a different color palette. Its default value is “default”, which uses the built-in colour palette provided in ggplot2 (Wickham, 2009). The color palette which was selected in Figure 7 is “viridis”, which is a colorblind-friendly palette spanning over a high range of different colors (Garnier, 2017).

The third web element on the right side of the plot output, a red button with the description “Obtain Code!”, adds user reproducibility to the plot. When clicked, a modal window pops up with R commands that, if executed in the main R console with the users current working environment, directly recreate the graph in the “Plot” tab.

Figure 8 shows the modal window which arises when pressing the “Obtain Code!” button in the interface of Figure 7. Below the description “Obtain your R code!” it is possible to see a text window with a dark background. In it are three syntax-highlighted R commands which can be copied and stored for future recreation.

	race	year	education
P1	1. White	2006	1. < HS Grad
P2	1. White	2006	2. HS Grad
P3	1. White	2006	3. Some College
P4	1. White	2006	4. College Grad
P5	1. White	2006	5. Advanced Degree

Figure 9: Scenario data tab.

The first R command creates a `data.frame` object from the “Scenarios” tabs user input. The second one uses `bamlss.vis::preds()` in combination with the provided models name to compute predicted parameters for the data, which are then re-used in line three to plot the results graphically with `bamlss.vis::plot_dist()`.

Furthermore, `bamlss.vis` checks for specified plot options (type of distribution, color palette) and includes them as arguments in the last line. The formatting of code relies on the R package `formatR` (Xie, 2017). Clicking on the “Dismiss” button closes the modal window.

3.3.4 Scenario Data Tab

While the “Scenario” tab gives users the ability to quickly specify covariate values without having to type in exact values themselves, sometimes it will be important to directly compare specific covariate combinations. Furthermore, one might want to see what values were previously specified in the “Scenarios” tab. For both reasons the tab “Scenario Data” was created in `bamlss.vis`’ left segment.

Figure 9 shows the tabs layout. As visible, the only element in this tab is a table placed underneath the description “Edit scenario data here”. This table, created with the R package `rhandsontable` (Owen, 2016) represents the editable version of all data input from the “Scenarios” tab. Columns represent the specified covariates, while one row counts as one “scenario”. In Figure 9, it is possible to see that only three columns of the original six are currently visible. This cut-off was built such that overlapping with the plot is prevented. Nevertheless, the user can use the scrolling bar at the bottom to reach other columns.

To edit a categorical value, users can click on a small drop-down button in the cell which is to be edited. The table recognizes categorical variables and will then provide a menu where the desired value is to be selected. With numeric variables, users can select the cell and then input the value they wish to make predictions with. Values in logical variables can be specified by checking or unchecking a box in the cell.

Unfortunately, this also means that in numeric cases, the user is able to specify a value which is not in the original variables' range. In the case of `cnorm_model`, this could mean that one tries to make predictions for incomes in the year 2050, which is far beyond $\max(\text{year}) = 2009$. To circumvent the irresponsible usage of model predictions, `bamlss.vis` will display a warning pop-up message with the "scenario" where values were specified which are out of the original variables range (Appendix: Figure A.2) message.

3.3.5 Properties Tab

Previous chapters have described that the "Plot" tab visualizes the predicted distributions for each covariate combination specified in the "Properties" tab. However, while differences in distributions for each combination were visible (e.g. in Figure 7), it is not possible to directly infer influences of each covariate on the distributional moments.

To provide this functionality, the tab "Properties" was implemented in `bamlss.vis`, located in the right segment. When opened, the "Properties" tab reveals two sub-tabs: "Influence graph" and "Table". As visible in Figure 10, the "Influence graph" tab consists of a graph on the left side next to a small bar with additional options. The graph displays the changes in the first two moments over the whole range of a covariate that was chosen on the right side. This is repeated for each "scenario" specified on `bamlss.vis`' left segment. As visible in Figure 10, the plot is divided into a `Expected_Value` and a `Variance` segment. In this case, variable `age` was chosen and therefore represents the x-axis from the variables minimum to maximum value. The y-axis depicts values for each of the first two distributional moments. Each line displays the level which the predicted moments would assume for `age`'s possible values, depending on the chosen covariate combination in the "Scenarios" tab.

In Figure 10, five different scenarios which represent varying `education` levels

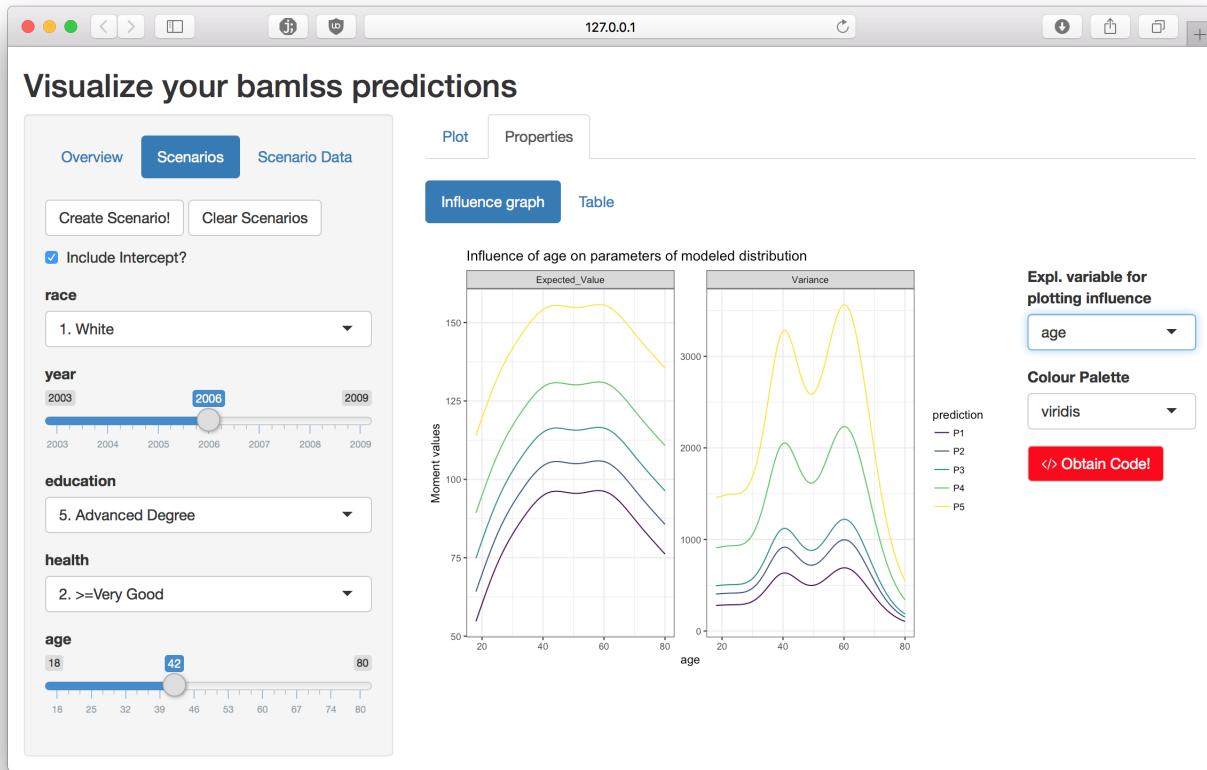


Figure 10: Influence of `age` on the first two moments of the predicted distributions for `cnorm_model`

were already specified. So in summary, this plot measures the influence of `age` on each of the first two moments of the modeled income distribution for healthy 42-year-old white males, depending on the `education` level.

For the influence of `age` on the first moment, it is possible to observe on all five lines that up until around 40 years `age` has a positive effect on the expected income, then takes a small downturn which is recovered around age 60, after which a rising age decreases the expected income. This influence structure is similar for all education levels because the effect of `age` on both parameters of the censored normal distribution was specified as a simple smooth spline with no `education` interaction.

For effects on the variance, a similar influence structure is observed, although the ups- and downs get more extreme as education level rises. This can be explained by the original model specification, where not the variance σ^2 but just the standard deviation σ is modeled in relation to explanatory covariates. Thus, its effects have to be squared because the variance of a censored normal distributed variable is dependent on σ^2 .

The effects of `age` on distributional moments can be visualized in lines because

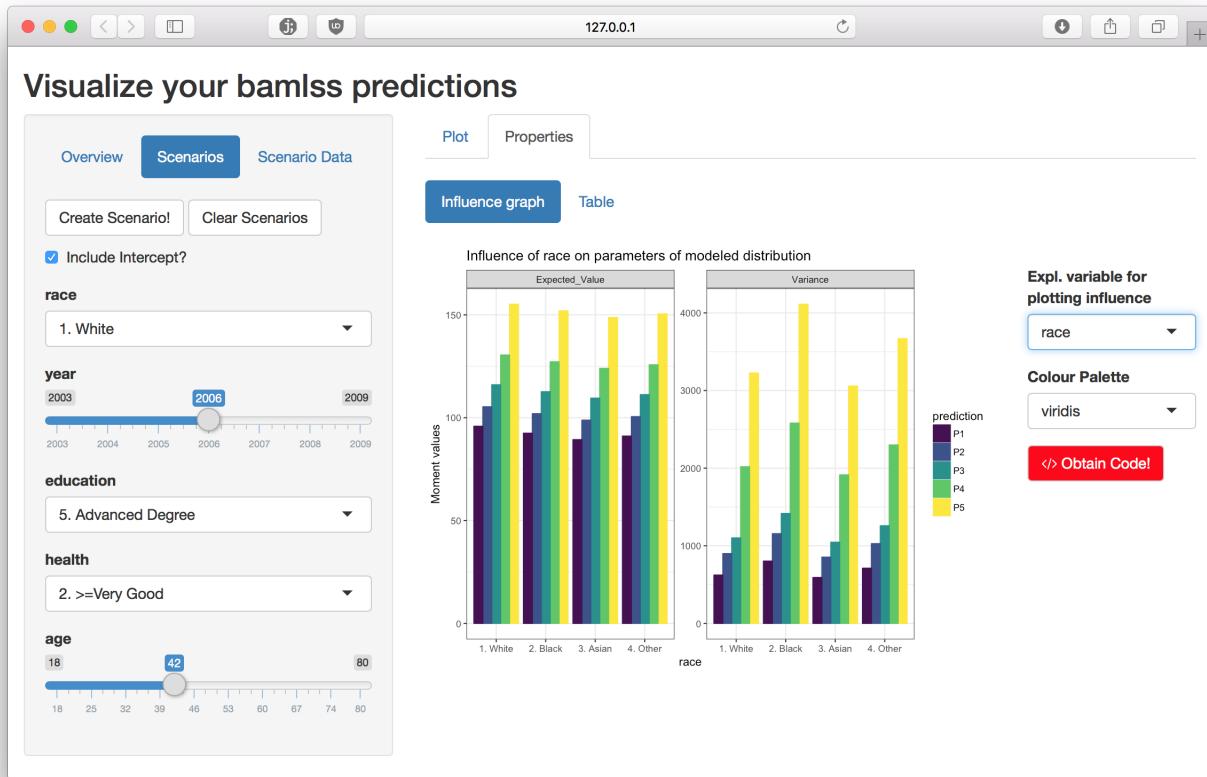


Figure 11: Influence of `race` categories on the first two moments of the predicted distributions for `cnorm_model`.

the underlying variable is measured in integers. However, `bamlss.vis` can also display differences in both moments depending on a categorical covariate. Figure 11 illustrates the influence graph when `race`, a qualitative covariate, is selected. In Figure 11, the x-axis again depicts the selected variable, `race`, with its possible outcomes `1. White` to `4. Other`. Every outcome displays five bars, one for each covariate combination specified in the “Scenarios” tab. Because the five different covariate combinations depict increasing education levels, this plot shows the effect of both `education` and `race` on the expected income level and its variance. As already observed in Figure 10, both the expected income and income variance increase with a higher education level. Furthermore, Figure 11 shows that 42-year old males are expected to earn less if they belong to the `2. Black` or `3. Asian` category than when they fall in the `1. White` category. Still, it is visible that the effect of `education` leads to more differences in the expected income than `race`.

Next to the graph in the “Influence graph” sub-tab three web elements are displayed. The first one, located under the description “Expl. variable for plotting influence”, yields the ability to select the explanatory variable for which the influ-

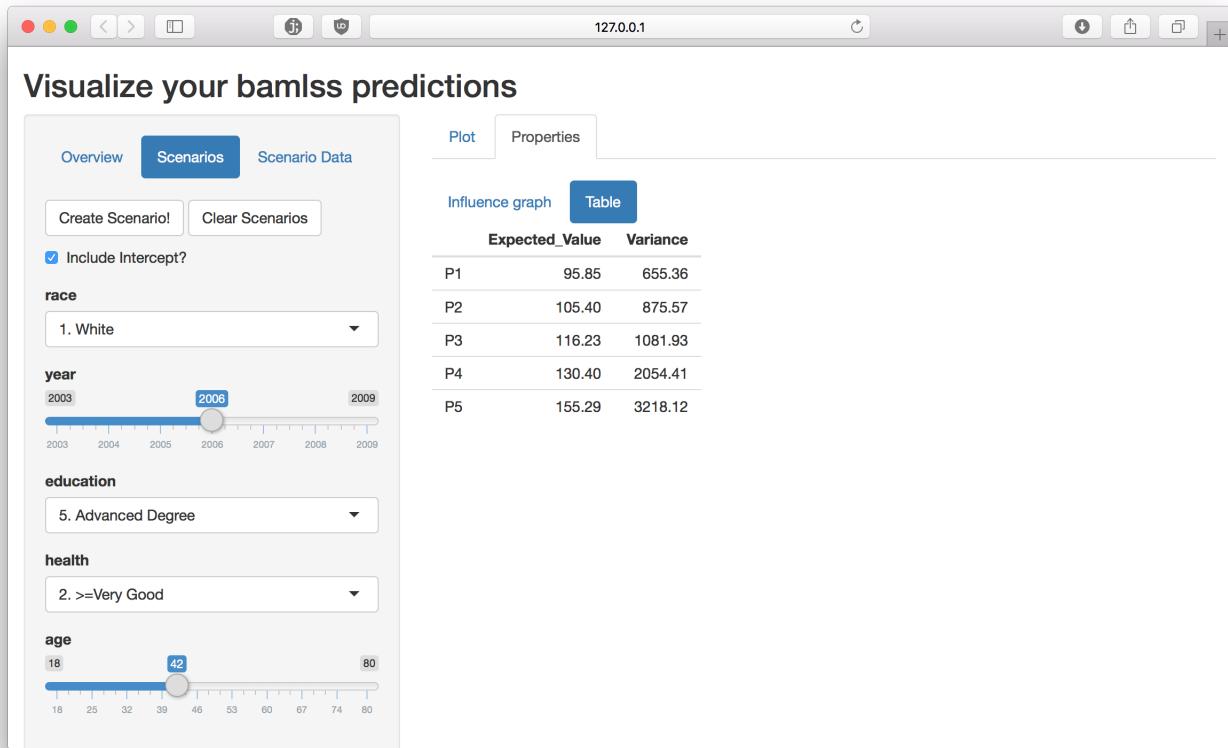


Figure 12: Expected Value and Variance for predicted distributions based on specified covariate combinations.

ence plot shall be created. The second element lets the user select a color palette, similar to Figure 7. In Figures 10 and 11, the “viridis” color palette was specified to account for colorblindness compatibility. Located below the palette selector the same red button as in Figure 7 is placed, which opens a modal window to reproduce the influence plot when it is pressed (Appendix: Figure A.3).

The other sub-tab of “Properties” called “Table” is useful if the user solely wants to see the differences in distributional moment values across specified covariate combinations in the “Scenarios tab”. Figure 12 shows the tabs layout. The only element in this tab is a table which shows two columns, **Expected_Value** and **Variance** with computed values for the first two moments. Every row depicts one covariate combination specified in the “Scenario” tab.

3.4 Additional Functions

The previous sub-chapters of Section 3.3 gave an overview of the functionality that bamlss.vis provides for continuous response distributions with a direct application on a bamlss computed from the `Wage` dataset. However, the abilities

of `bamlss.vis` go beyond the censored normal distribution and are applicable to almost all distributional families that the `bamlss` package provides, which includes discrete and bivariate distributions. This chapter will demonstrate how `bamlss.vis` handles those non-continuous response distributions.

Because most distributions included in `bamlss` are not suitable for usage with the `Wage` dataset, this section will rely on simulated data generated by `bamlss.vis ::model_fam_data()`. To sustain a correlation structure between covariates such that they can be exploited for useful modeling purposes, `model_fam_data()` simulates a three-dimensional uniform distribution with a sample space of $x_i \in [0, 1]$, $i = 1, 2, 3$. Then, inverse transform sampling with x_1 is used to obtain sample data for all supported `bamlss` distribution families. x_2 and x_3 are transformed to standard normally distributed variables. Covariates x_1 and x_2 are then used as explanatory variables for modeling the transformed response variable x_1 .

3.4.1 Discrete Responses

When a model with discrete observations is specified in the “Overview” tab, it is automatically recognized by `bamlss.vis` using the function `is.discrete()`. In comparison to the continuous case, the application structure of `bamlss.vis` does not change. Figure 13 shows the “Plot” tab of `bamlss.vis` for a `bamlss` modeling a Poisson-distributed response where three distinct covariate combinations were already specified in the “Scenarios” tab. In Figure 13, the three predicted probability mass functions for the different Poisson distributions are visible. As the Poisson distribution is a discrete family, probabilities $f(y)$ are displayed in bars on the y-axis. The x-axis represents possible outcomes of the Poisson distribution. For every outcome, three bars for three different covariate combinations specified in the “Scenarios” are displayed. For example, prediction P2 has the lowest expected probability $P(y = 0)$ compared to the other two combinations.

In theory, the Poisson distribution can attain all values $\{y \in \mathbb{Z} \mid y \geq 0\}$ up to ∞ . To appropriately visualize different expected Poisson distribution, the x axis of the graph is limited to $0, \dots, x_{lim}$ where $x_{lim} = (\max(\lambda_1, \dots, \lambda_K) \cdot 2) + 3$ for predicted parameters λ_i from covariate combinations $1, \dots, K$. Furthermore, `bamlss.vis` also has the ability to display cumulative distribution functions for discrete response distributions. To toggle this plot, the same list selector element as with continuous response distributions to the graphs right side is used.

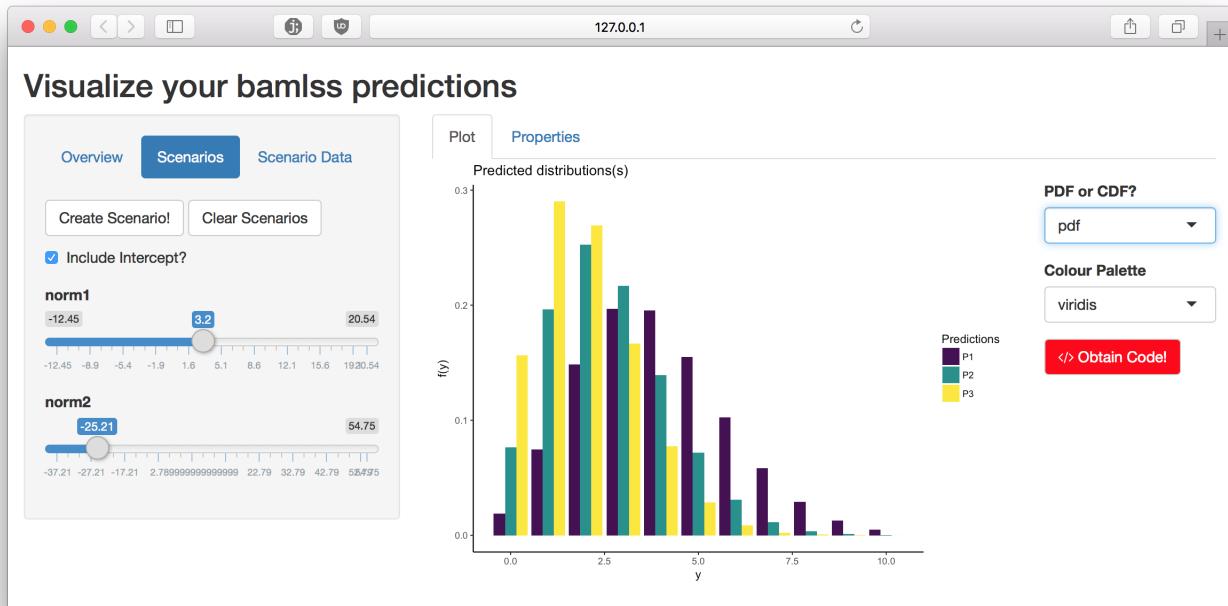


Figure 13: Plot tab with predicted probability mass functions for Poisson distributions with three specified covariate combinations in the “Scenarios” tab

The cumulative distribution function is then displayed as a step-wise graph (Appendix: Figure A.4). Moreover, specifying other color palettes and obtaining plot reproduction code by pressing the red button is also possible.

While the Poisson distribution has discrete observations, its first two moments $E(y) = Var(y) = \lambda$ assume values on a continuous scale. Therefore, the appearance of the “Properties” tab with “Influence graph” and “Table” subtabs are identical to the continuous case described in Section 3.3.5.

3.4.2 Multivariate Responses

In contrast to the `gamlss` R package without extensions, `bamlss` also supports multivariate response distributions. From the range of multivariate distributions, only the multivariate normal distribution is currently implemented in `bamlss` and therefore supported by `bamlss.vis`. Nevertheless, because of its existing 3D graphical framework, `bamlss.vis` can easily be extended for new multivariate distributions.

For graphical reasons, `bamlss.vis` can only display bivariate distributions. Figure 14 shows the “Plot” and “Scenarios” tab with a specified scenario from the left tab. As visible in Figure 14, a bivariate normal distribution is displayed. Multivariate response predictions in `bamlss.vis` are always only dependent on

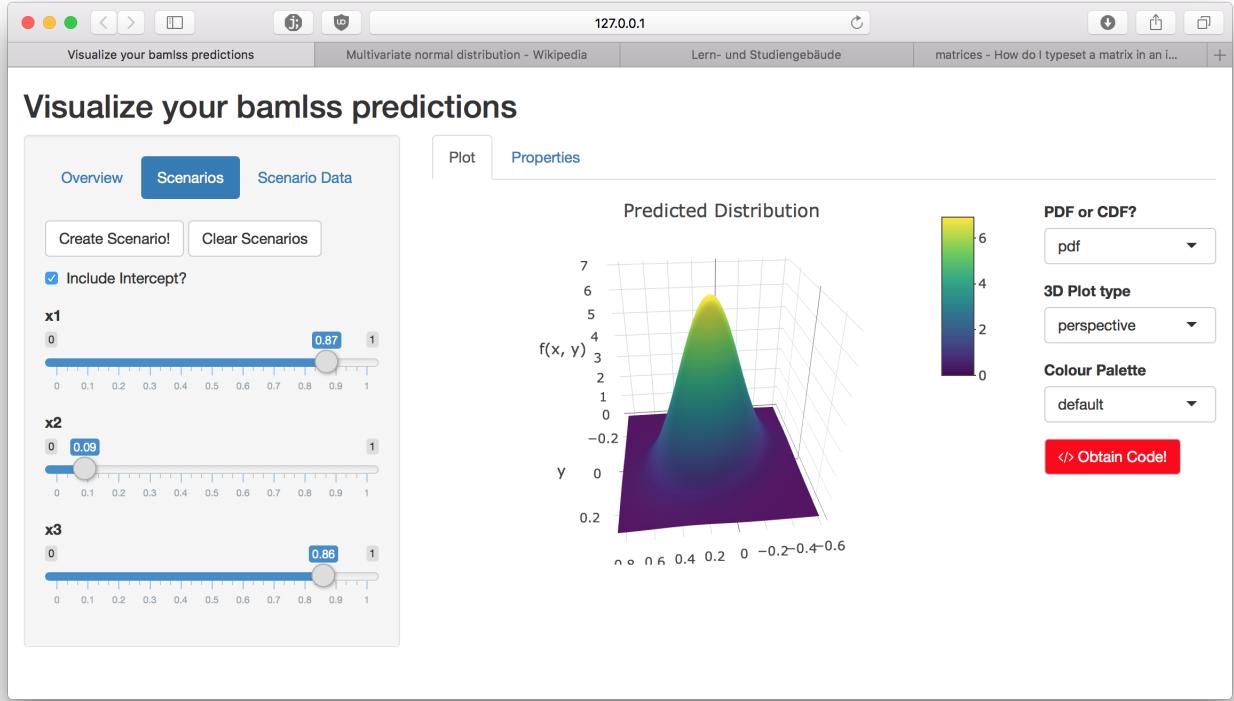


Figure 14: Predicted distribution for a multivariate normal distributed response based on the covariate specification of the “Scenarios” tab.

the last specified covariate combination, without any prediction comparisons. Furthermore, multivariate predictions are displayed using the R package `plotly` (Sievert et al., 2017), which adds a layer of interactivity to the graph. Using a graph generated with `plotly`, the user can hold and drag the distribution to view the shape from all perspectives. Moreover, hovering over the distribution surface yields current x , y and z values.

Located to the right side of the plot are web elements mostly known from previous Sections. Bamlss.vis supports the display of multivariate cumulative distribution functions, which can be selected in the first element titled “PDF or CDF?”. An example is given by Figure A.5 in the Appendix. Placed below the first element and the description “3D Plot type” is another selector box for specifying the type of 3D Plot. In addition to “perspective”, the options “contour” and “image” can be selected. Examples for both options can be found in Appendix Figures A.6 and A.7 for a contour and image plot, respectively.

Because the bivariate normal distribution consists of two univariate normal distributions with covariance $\rho\sigma_1\sigma_2$, its first two moments are $\boldsymbol{\mu} = (\mu_1 \mu_2)$ and $\boldsymbol{\Sigma} = (\begin{matrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{matrix})$. In total, this amounts to five parameters for which covariate influences can be graphically displayed. Figure 15 shows the “Influence graph”

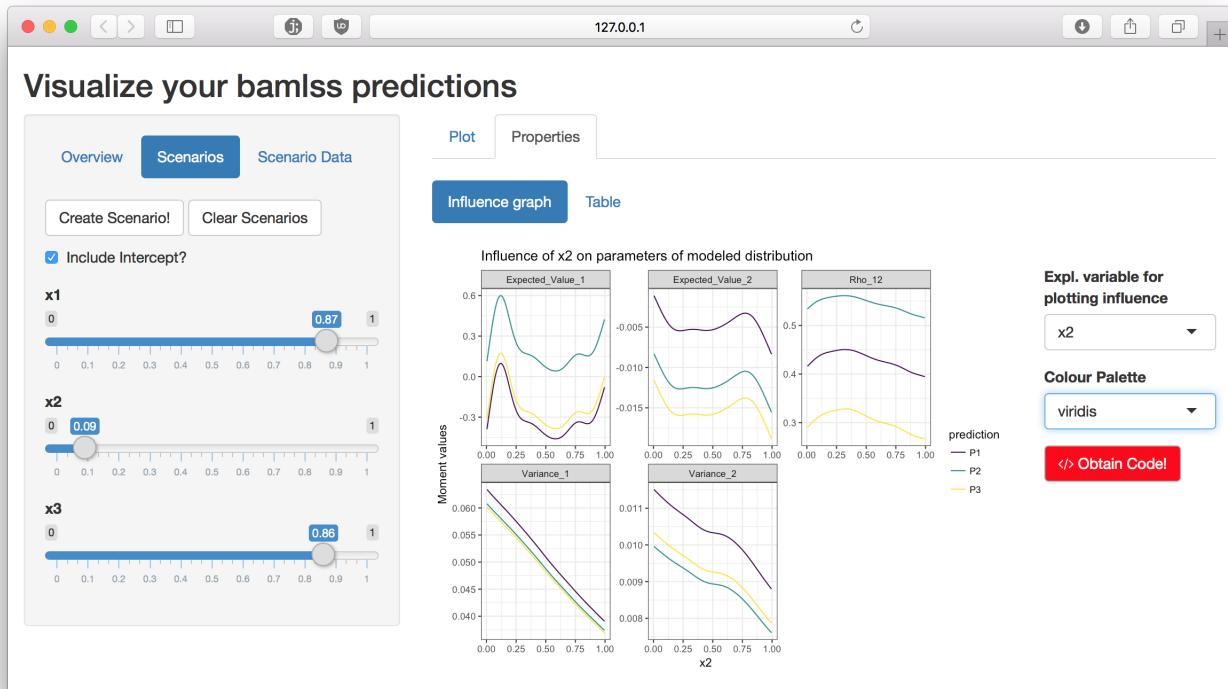


Figure 15: “Influence graph” tab for a selected bivariate normal `bamlss` with three specified covariate combinations.

sub-tab for a specified multivariate normal `bamlss`. As visible, the displayed graph is divided into five subgraphs, instead of two in the univariate case. Similar to the influence plot for univariate response distributions, the relationship between the moments and `x2` (explanatory variable selected in this case) is plotted for every previously specified covariate combination. Furthermore, the user can select different color palettes and obtain R code for plot reproduction.

The “Table” sub-tab of the “Properties” tab again differs for bi-variate normal distributions. Here, the displayed table for distributional moments depending on covariate combinations specified in the “Scenarios” tab is expanded to the five moments of the bivariate normal distribution. For an example, Figure A.8 in the Appendix is provided.

3.4.3 Multinomial Responses

For modeling unordered categorical response variables, `bamlss` supports the multinomial distribution family, a generalization of the binomial distribution. Loading a `bamlss` with a multinomial response is possible in `bamlss.vis`, although the graphical appearance of predicted distributions and the influence plot differs from usual discrete cases.

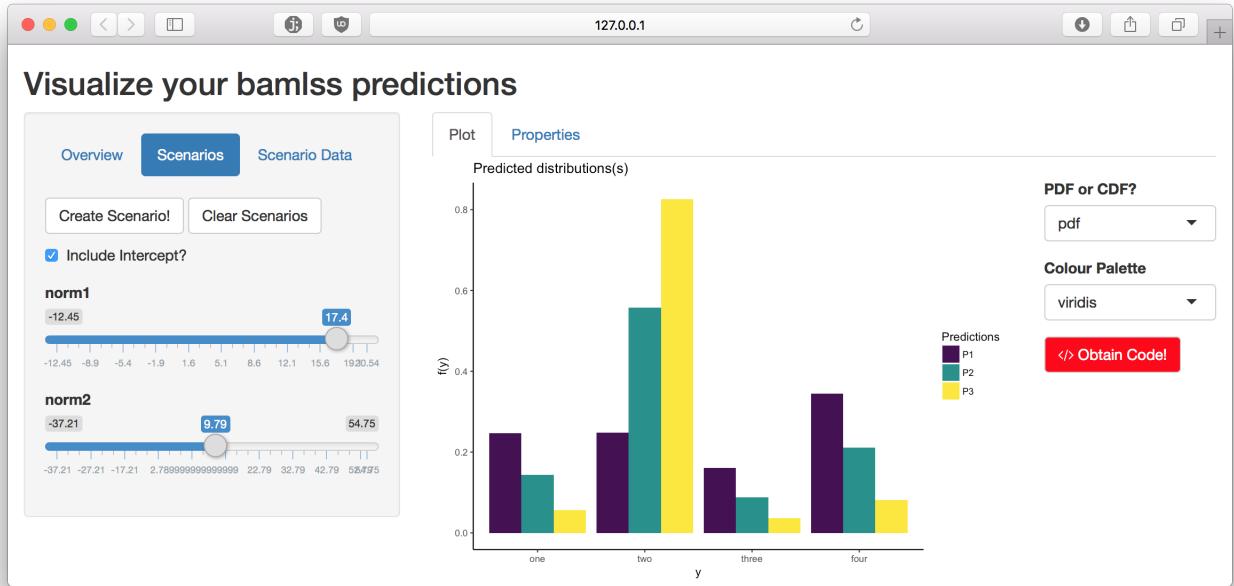


Figure 16: Predicted multinomial response in the “Plot tab” based on three specified covariate combinations

Figure 16 shows the “Plot” tab with three specified covariate combinations after selecting a sample `bamlss` with multinomial response in the “Overview” tab. The modeled sample response variable can have four different outcomes `one`, `two`, `three` and `four` which are visible on the x-axis. The predicted probabilities for each of those outcomes are displayed on the y-axis. This is realized graphically in bars for each of the specified covariate combinations, which are three in the present case. Interpreting 16, one can see that with each higher covariate combination (P_1 to P_3), the probability to fall into category `two` increases, while the probabilities for all other categories decrease.

Moreover, tweaking the plot with the web elements in the right sidebar is possible, similar to `bamlss.vis`’ behavior for discrete or continuous univariate responses. The option for displaying the cumulative distribution function is inoperative, due to the lack of a reasonable graphical representation.

Using own covariate combinations and then visualizing the probability differences for each class as in Figure 16 can already be used to obtain influence tendencies of covariates on the distributional moments, because for a multinomial distribution y with K classes and $n = 1$ draws $E(y) = \boldsymbol{\pi} = (\mu_1, \dots, \mu_K)'$ holds, which means the first moment is already graphically displayed. However, the variation of $\boldsymbol{\pi}$ over the whole range of an explanatory variable might be of interest. To solve this need, Figure 17 is provided. In Figure 17, the predicted values for π_1, \dots, π_4 are displayed over the range of the selected variable, `norm1`. This is then repeated

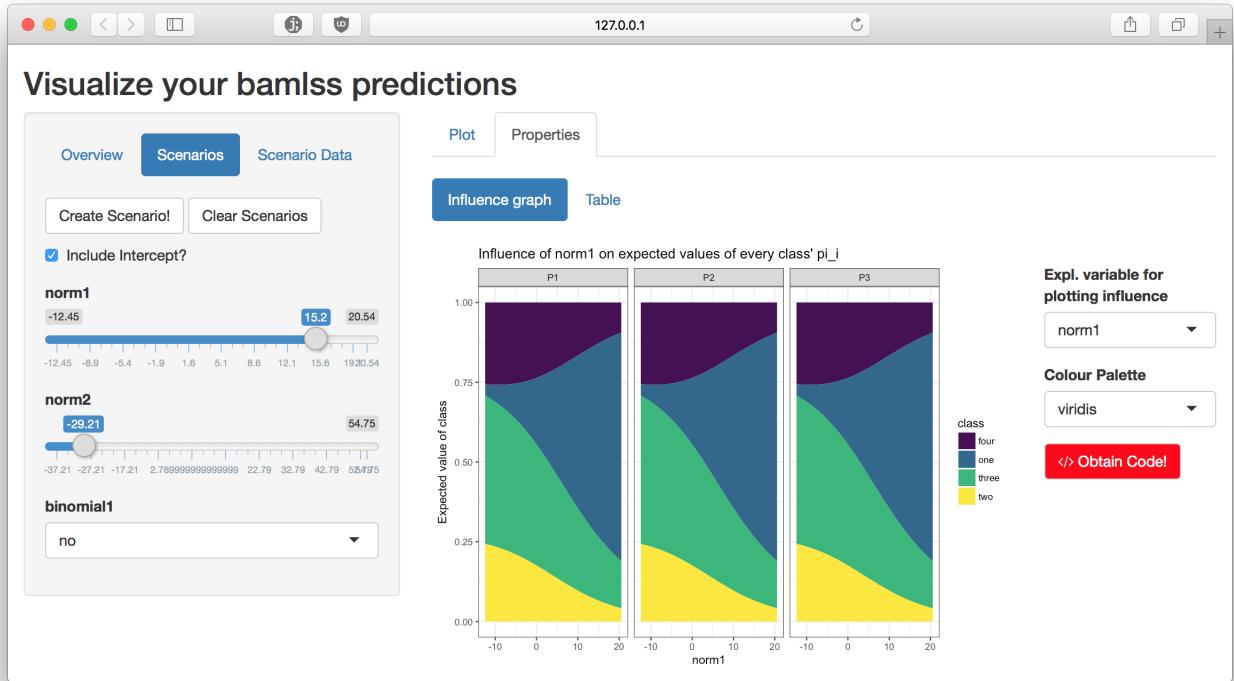


Figure 17: Influence of variable `norm1` on expected first moments of multinomial distribution

for every covariate combination which was specified in the “Scenarios” tab. As visible, the y-axis limits are always 0 and 1, due to $\mu_1 = \dots = \mu_4 \geq 0$ and $\sum_{i=1}^4 \pi_i = 1$ holding true. Furthermore, one can see that an increase of `norm1` is associated with an increase of the probability to fall into category “one” (π_1) and a decrease for all other categories, across all three covariate combinations.

Displaying the influence of explanatory covariates on probabilities π_i is also possible for categorical variables. For an example, Figure A.9 in the Appendix is provided.

4 Conclusion

A Appendix

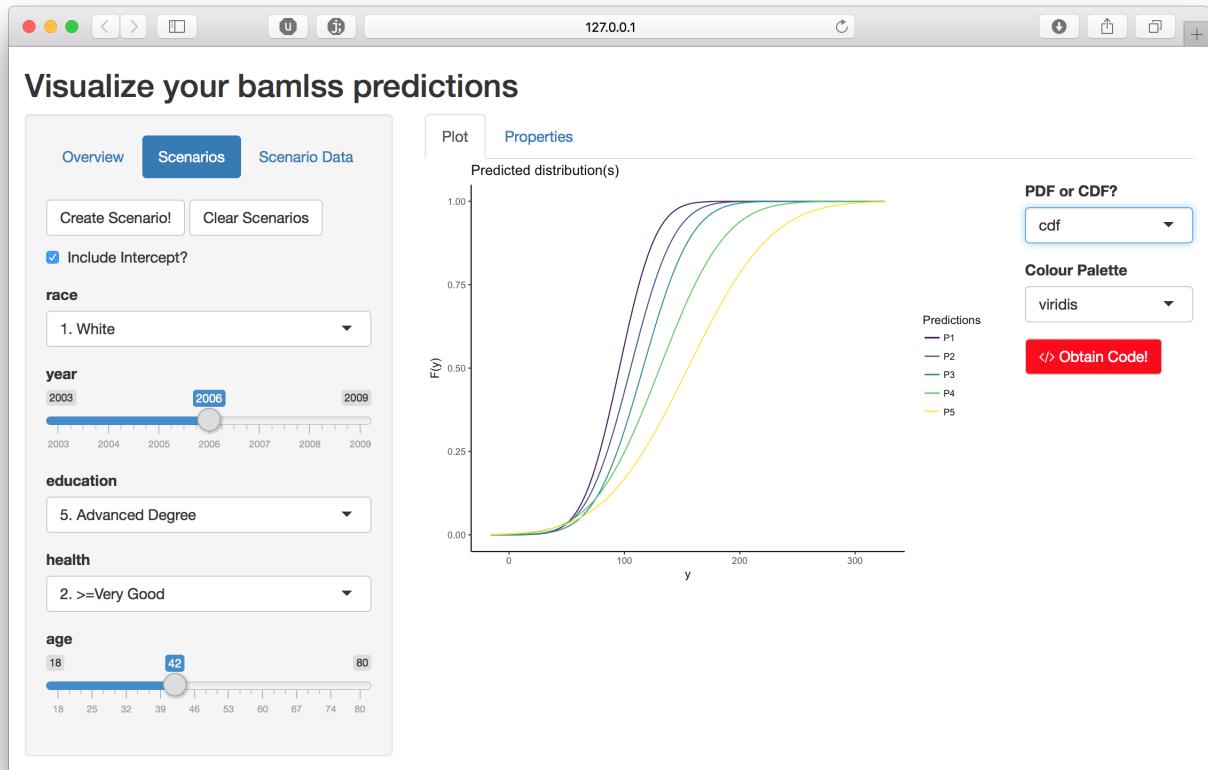


Figure A.1: Cumulative Distribution Function plot output for different education levels based on the `Wage` dataset.

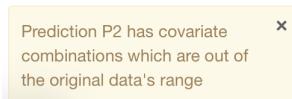


Figure A.2: Warning message when specifying covariate combinations which are out of range.

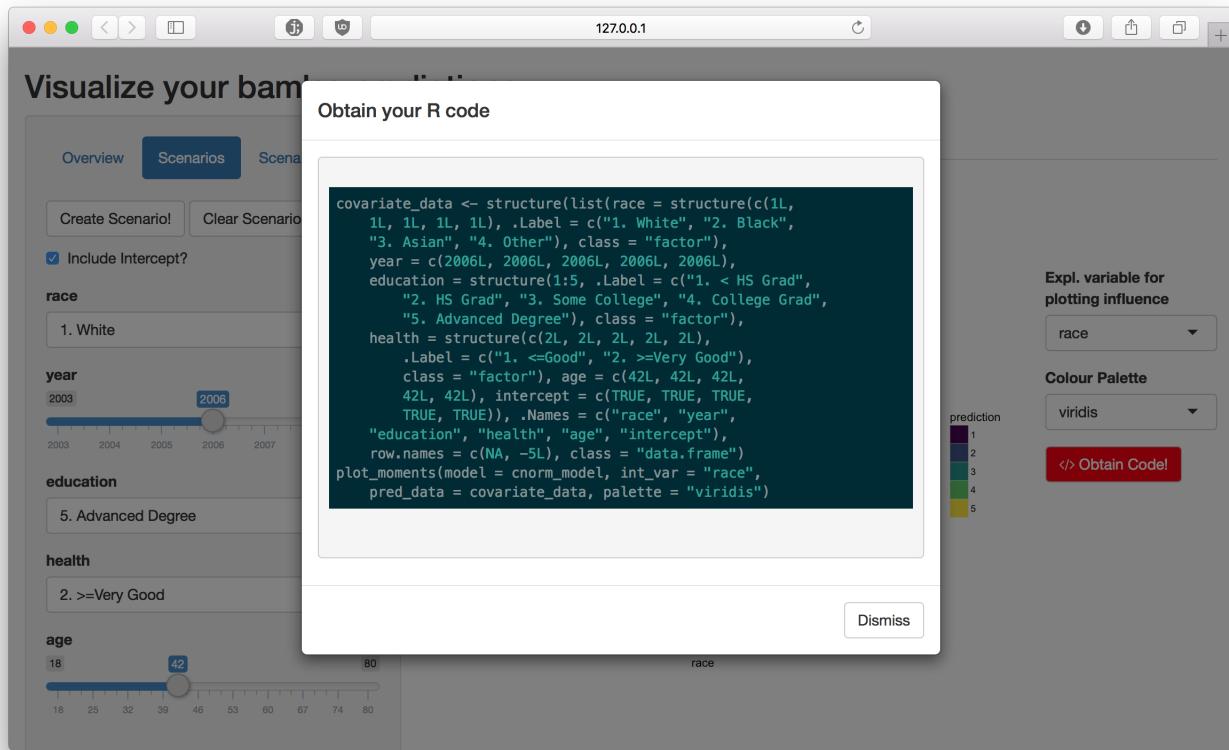


Figure A.3: Modal window to display code for reproducing the influence plot.

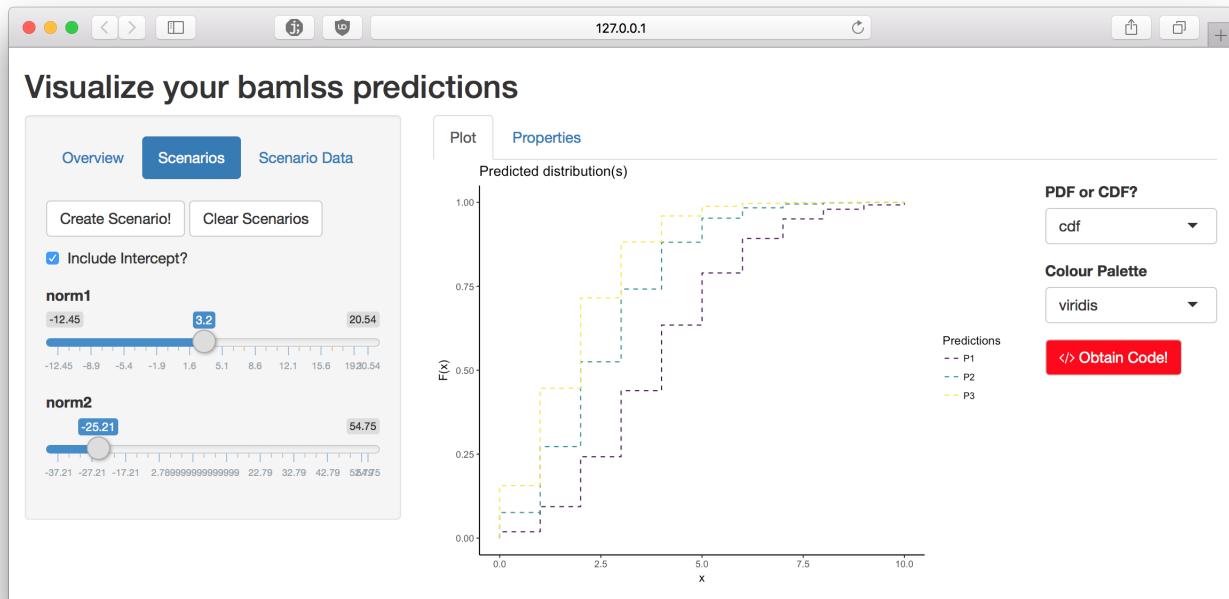


Figure A.4: Predicted Cumulative Distribution Function for three Poisson distributions based on user-input covariate values.

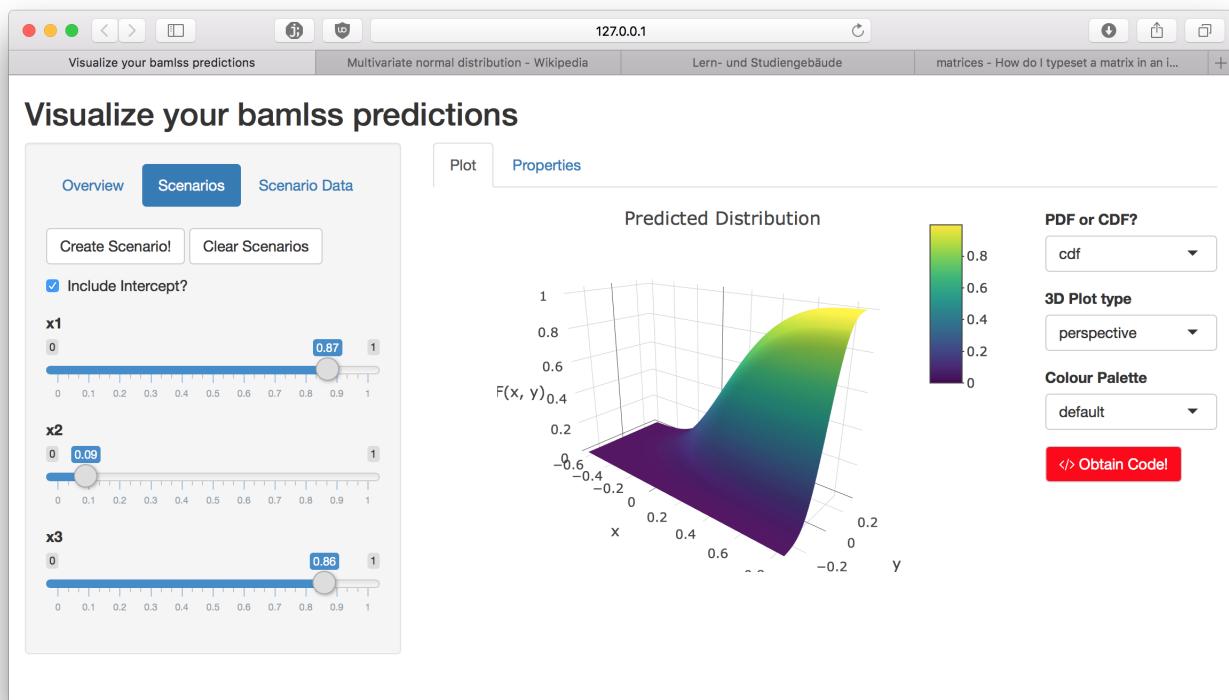


Figure A.5: Predicted distribution for a multivariate normal distributed response based on the covariate specification of the “Scenarios” tab.

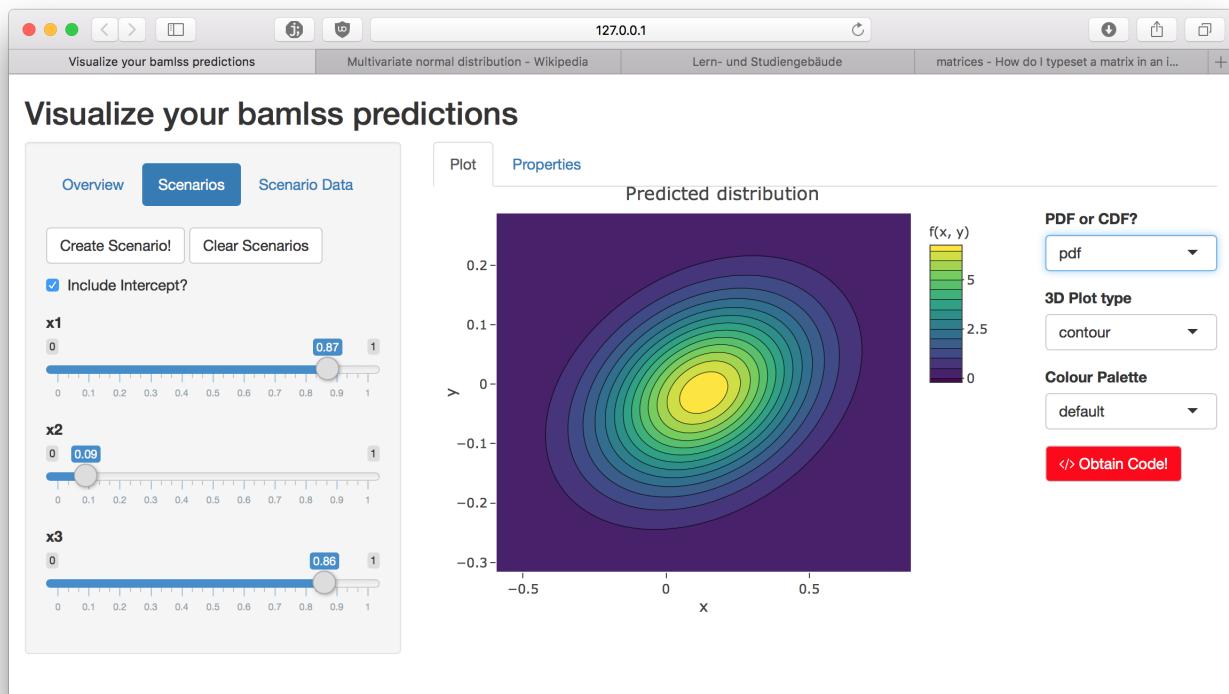


Figure A.6: Contour plot for the predicted multivariate normal distribution based on user inputs in “Scenarios” tab.

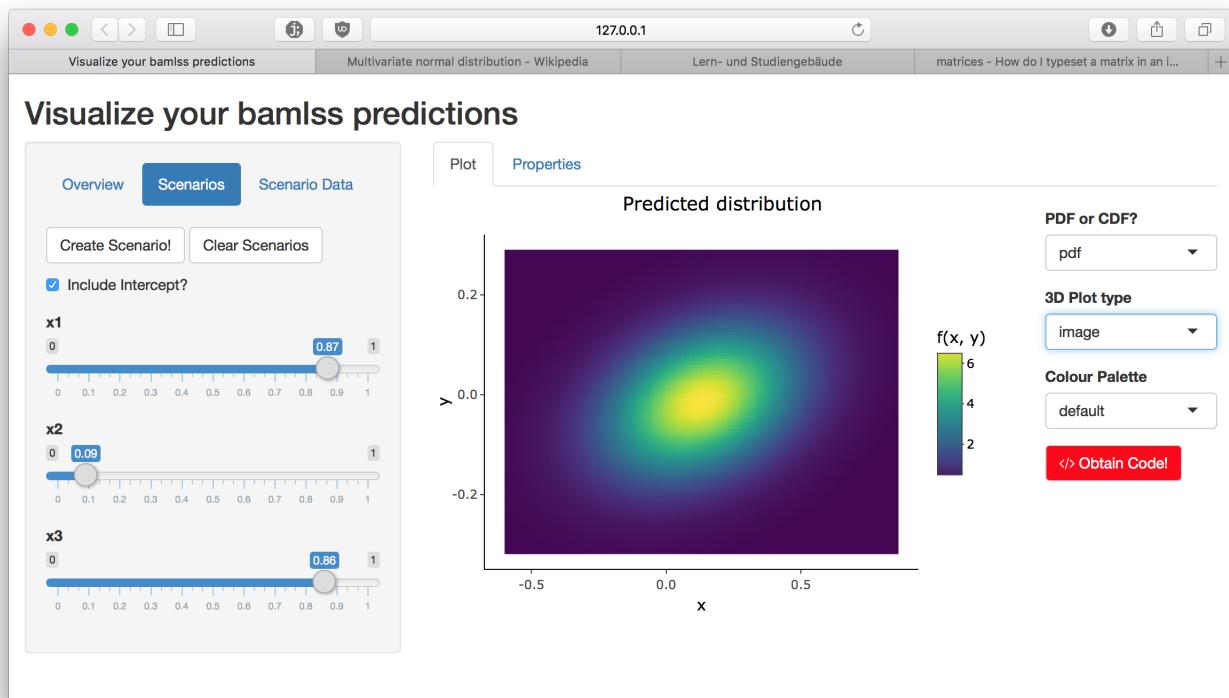


Figure A.7: Image plot for the predicted multivariate normal distribution based on user inputs in “Scenarios” tab.

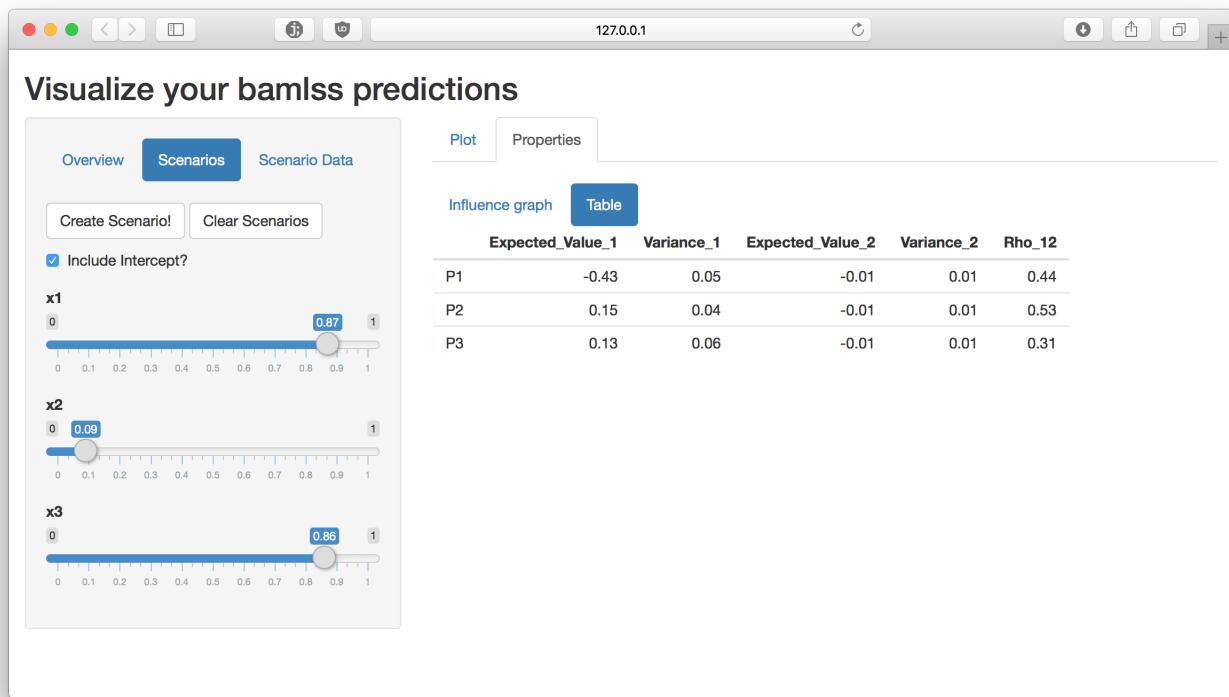


Figure A.8: Table view for expected moments of the bivariate normal distribution based on three specified covariate combinations in the “Scenarios” tab.

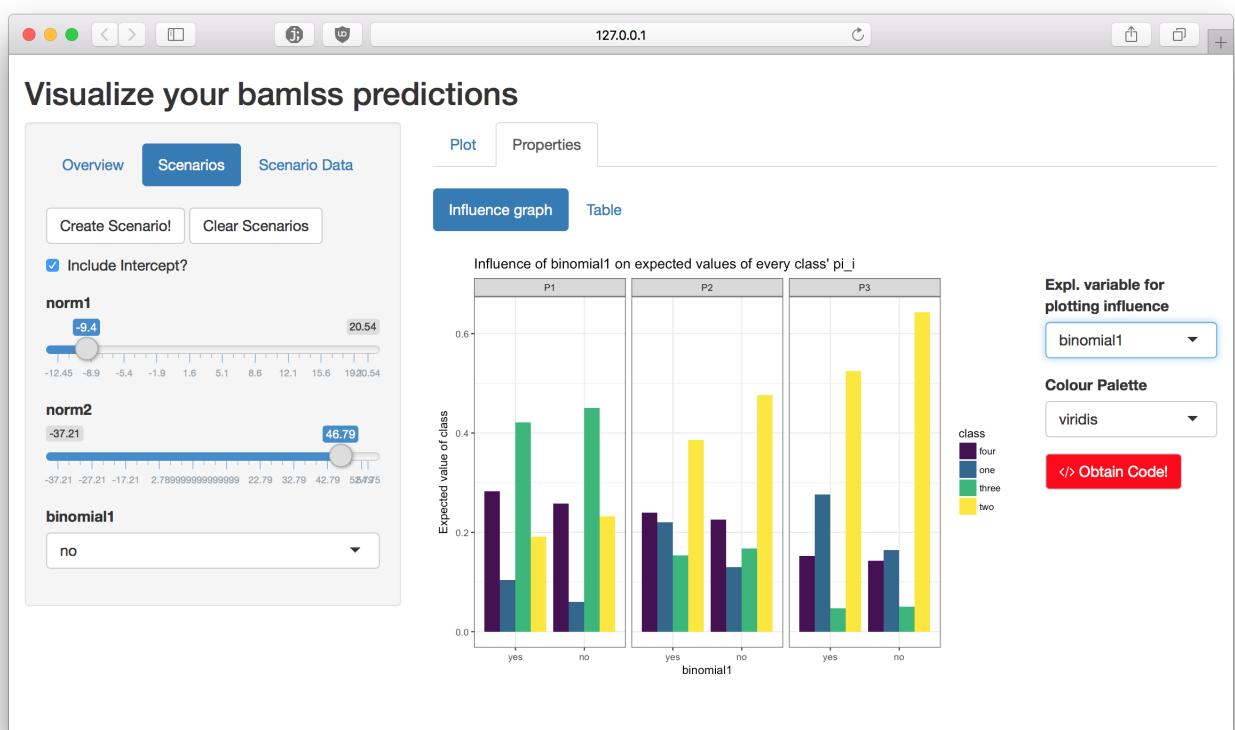


Figure A.9: Sample influence plot for categorical covariates on expected class probabilities in multinomial response `bamLSS`

Bibliography

- A. Brezger and S. Lang. Generalized structured additive regression based on bayesian p-splines. *Computational Statistics & Data Analysis*, 50:967–991, February 2006.
- A. Buja, T. Hastie, and R. Tibshirani. Linear smoothers and additive models. *The Annals of Statistics*, pages 453–510, 1989.
- Bureau of Labor Statistics. OES Data, 2016. URL <https://www.bls.gov/oes/tables.htm>. [Online; accessed 28-Nov-2017].
- J.M. Chambers and T. Hastie. *Statistical Models in S*. Wadsworth & Brooks/Cole computer science series. Wadsworth & Brooks/Cole Advanced Books & Software, 1992. ISBN 9780534167646. URL <https://books.google.de/books?id=uyfvAAAAMAAJ>.
- W. Chang, W. Cheng, J.J. Allaire, Yihui X., and J. McPherson. *shiny: Web Application Framework for R*, 2017. URL <https://CRAN.R-project.org/package=shiny>. R package version 1.0.5.
- T. J. Cole and P. J. Green. Smoothing reference centile curves: The lms method and penalized likelihood. *Statistics in Medicine*, 11(10):1305–1319, 1992. ISSN 1097-0258. doi: 10.1002/sim.4780111005. URL <http://dx.doi.org/10.1002/sim.4780111005>.
- L. Fahrmeir, T. Kneib, and S. Lang. Penalized additive regression for space-time data: a bayesian perspective, 2003. URL <http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-1687-9>.
- L. Fahrmeir, T. Kneib, S. Lang, and B. Marx. *Regression: Models, Methods and Applications*. Springer Berlin Heidelberg, 2013. ISBN 9783642343339. URL <https://books.google.de/books?id=EQxU9iJtipAC>.
- J.H. Friedman and W. Stuetzle. Projection pursuit regression. *Journal of the American statistical Association*, 76(376):817–823, 1981.
- S. Garnier. *viridis: Default Color Maps from 'matplotlib'*, 2017. URL <https://CRAN.R-project.org/package=viridis>. R package version 0.4.0.
- W.H. Greene. *Econometric Analysis*. Pearson International Edition. Pearson Education, Limited, 2012. ISBN 9780273753568. URL <https://books.google.de/books?id=-WFPYgEACAAJ>.
- T.J. Hastie and R.J. Tibshirani. *Generalized Additive Models*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1990. ISBN 9780412343902. URL <https://books.google.de/books?id=qa29r1Ze1coC>.

- G. James, D. Witten, T. Hastie, and R. Tibshirani. *ISLR: Data for an Introduction to Statistical Learning with Applications in R*, 2017. URL <https://CRAN.R-project.org/package=ISLR>. R package version 1.2.
- N. Klein, T. Kneib, and S. Lang. Bayesian structured additive distributional regression. Working Papers in Economics and Statistics 2013-23, Universität Innsbruck, Institut für Finanzwissenschaft, Innsbruck, 2013. URL <http://hdl.handle.net/10419/101101>.
- N. Klein, T. Kneib, S. Lang, and A. Sohn. Bayesian structured additive distributional regression with an application to regional income inequality in germany. *Ann. Appl. Stat.*, 9(2):1024–1052, June 2015. doi: 10.1214/15-AOAS823. URL <https://doi.org/10.1214/15-AOAS823>.
- Laird, Nan M., and James H. Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.
- R.I. Lerman and S. Yitzhaki. A note on the calculation and interpretation of the gini index. *Economics Letters*, 15(3-4):363–368, 1984.
- G. Matheron. Principles of geostatistics. *Economic geology*, 58(8):1246–1266, 1963.
- J.A. Nelder and D. Pregibon. An extended quasi-likelihood function. *Biometrika*, 74(2):221–232, 1987.
- J.A. Nelder and R.W.M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972.
- C. O’Connor. Data: the key measure of relevance in a digital revolution, 2016. URL <https://www.ibm.com/blogs/internet-of-things/data-revolution/>. [Online; accessed 28-Nov-2017].
- U. Olsson. *Generalized Linear Models: An Applied Approach*. Lightning Source, 2002. ISBN 9789144041551. URL <https://books.google.de/books?id=SP1jHQAAQAAJ>.
- J. Owen. *rhandsontable: Interface to the 'Handsontable.js' Library*, 2016. URL <https://CRAN.R-project.org/package=rhandsontable>. R package version 0.3.4.
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL <https://www.R-project.org/>.
- R.A. Rigby and D.M. Stasinopoulos. The gamlss project: a flexible approach to statistical modelling. In *New trends in statistical modelling: Proceedings of the 16th international workshop on statistical modelling*, volume 337, page 345. University of Southern Denmark, June 2001.

- R.A. Rigby and D.M. Stasinopoulos. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554, 2005.
- R.A. Rigby and M.D. Stasinopoulos. *Mean and Dispersion Additive Models*, pages 215–230. Physica-Verlag HD, Heidelberg, 1996. ISBN 978-3-642-48425-4. doi: 10.1007/978-3-642-48425-4_16. URL https://doi.org/10.1007/978-3-642-48425-4_16.
- RStudio, Inc. Shiny, 2017. URL <https://shiny.rstudio.com>. [Online; accessed 28-Nov-2017].
- C. Sievert, C. Parmer, T. Hocking, S. Chamberlain, K. Ram, M. Corvellec, and P. Despouy. *plotly: Create Interactive Web Graphics via 'plotly.js'*, 2017. URL <https://CRAN.R-project.org/package=plotly>. R package version 4.7.1.
- D.M. Stasinopoulos and R.A. Rigby. Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, 23(7):1–46, 2007.
- M.D. Stasinopoulos, R.A. Rigby, G.Z. Heller, V. Voudouris, and F. De Bastiani. *Flexible Regression and Smoothing: Using GAMLSS in R*. Chapman & Hall/CRC The R Series. CRC Press, 2017. ISBN 9781351980371. URL <https://books.google.de/books?id=1h-9DgAAQBAJ>.
- The Economist. Rise of the machines, 2015. URL <https://www.economist.com/news/briefing/21650526-artificial-intelligence-scares-peopleexcessively-so-rise-machines>. [Online; accessed 28-Nov-2017].
- N. Umlauf, N. Klein, and A. Zeileis. Bamlss: Bayesian additive models for location, scale and shape (and beyond). Working papers, Working Papers in Economics and Statistics, 2017. URL <https://EconPapers.repec.org/RePEc:inn:wpaper:2017-05>.
- United States Census Bureau. Supplement to current population survey, March 2011. URL <http://www.nber.org/cps/cpsmar11.pdf>. [Online; accessed 28-Nov-2017].
- H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. ISBN 978-0-387-98140-6. URL <http://ggplot2.org>.
- S. N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3–36, 2011.
- Y. Xie. *formatR: Format R Code Automatically*, 2017. URL <https://CRAN.R-project.org/package=formatR>. R package version 1.5.