GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

bamlss.vis

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

Stanislaus Stadlmann

7. December 2017

Georg-August University of Göttingen

Table of contents

1. Introduction

2. BAMLSS ancestry

3. Motivation for bamlss.vis

4. Case-Study

5. bamlss.vis

Introduction

Introduction

Distributional Regression

- An emerging field in regression methods
- Each parameter of a response distribution beyond the mean can be modeled using a set of predictors

Introduction

Distributional Regression

- An emerging field in regression methods
- Each parameter of a response distribution beyond the mean can be modeled using a set of predictors
- Notable frameworks:
  1. Generalized Additive Models for Location, Scale and Shape, coined by Rigby and Stasinopoulos (2001)
  2. Bayesian Additive Models for Location, Scale and Shape, coined by Umlauf, Klein, and Zeileis (2017)
- Differences: Estimation techniques - Likelihood/Bayes

Introduction

bamlss.vis

- R package based on the Shiny framework
- Built upon R package `bamlss`
- Requires a fitted `bamlss` object
- Yields the abilities to
    1. visualize predictions for user-chosen covariate combinations
    2. visualize the influence of a certain covariate on distributional moments

BAMLSS ancestry

Additive Models (AM)

Overview

- Proposed by Friedman and Stuetzle (1981)
- Dependent variable y is related to non-parametric effects in an additive way

Additive Models (AM)

Overview

- Proposed by Friedman and Stuetzle (1981)
- Dependent variable y is related to non-parametric effects in an additive way

Model specification

$$y_i = f_1(z_{i1}) + f_2(z_{i2}) + \ldots + f_K(z_{iK}) + \epsilon_i \quad \text{(only nonparametric effects)}$$

$$y_i = \sum_{j=1}^{K} f_j(z_{ij}) + \sum_{l=1}^{Q} \beta_l x_{il} + \epsilon_i \quad \text{(with parametric effects)}$$

Additive Models (AM)

Overview

- Proposed by Friedman and Stuetzle (1981)
- Dependent variable y is related to non-parametric effects in an additive way

Model specification

$$y_i = f_1(z_{i1}) + f_2(z_{i2}) + \ldots + f_K(z_{iK}) + \epsilon_i \quad \text{(only nonparametric effects)}$$

$$y_i = \sum_{j=1}^{K} f_j(z_{ij}) + \sum_{l=1}^{Q} \beta_l x_{il} + \epsilon_i \quad \text{(with parametric effects)}$$

Why additive?

- Curse of dimensionality
- Easier to separate covariate effects

Structured Additive Regression (STAR) Models

Motivation

- AM allow for non-parametric effects, but sometimes even more flexibility is needed
- STAR (Fahrmeir et al., 2003) also support structured terms, which include:

Structured Additive Regression (STAR) Models

Motivation

- AM allow for non-parametric effects, but sometimes even more flexibility is needed
- STAR (Fahrmeir et al., 2003) also support structured terms, which include:
  1. Nonlinear effects of a single variable
  2. Spatial effects of location index s
  3. Interactions between a continuous covariate and a categorical variable
  4. Nonlinear interactions between two continuous covariates
  5. Random Effects with intercept ν_0 and slope ν_j deviations from main effects

Structured Additive Regression (STAR) Models

Motivation

- AM allow for non-parametric effects, but sometimes even more flexibility is needed
- STAR (Fahrmeir et al., 2003) also support structured terms, which include:
  1. Nonlinear effects of a single variable
  2. Spatial effects of location index s
  3. Interactions between a continuous covariate and a categorical variable
  4. Nonlinear interactions between two continuous covariates
  5. Random Effects with intercept ν_0 and slope ν_j deviations from main effects

Model specification

$$y_i = \underbrace{\kappa_i^{add}}_{\text{AM components}} + f_{struc}(\mathbf{z}_{iF}) + \epsilon_i$$

where $\mathbf{z}_F$ can be a one- or multidimensional variable.

Generalized STAR Models

Motivation

- AM and STAR assume normalty and directly model $E(y)$
- Generalized STAR models use link function $g(\cdot)$ of Generalized Linear Models
- Adds ability to model $E(y)$ of all exponential families, e.g. binomial or poisson distribution

Generalized STAR Models

Motivation

- AM and STAR assume normalty and directly model $E(y)$
- Generalized STAR models use link function $g(\cdot)$ of Generalized Linear Models
- Adds ability to model $E(y)$ of all exponential families, e.g. binomial or poisson distribution

Model specification

$$
\begin{aligned}
g(\mu_i) &= \eta_i \\
\eta_i &= f_1(\mathbf{z}_{i1}) + \ldots + f_J(\mathbf{z}_{iJ})
\end{aligned}
$$

where $f_j(\cdot)$ can be any structured effect.

Structured Additive Distributional Regression

Motivation

- Often, more than just the location (Expected Value) of a distribution is of interest
- Scale/Shape (Variance, Kurtosis) might also be dependent on covariates
- Structured Additive Distributional Regression allows modeling of all distributional parameters θ_l

Structured Additive Distributional Regression

Motivation

- Often, more than just the location (Expected Value) of a distribution is of interest
- Scale/Shape (Variance, Kurtosis) might also be dependent on covariates
- Structured Additive Distributional Regression allows modeling of all distributional parameters θ_l

Model specification

Let $y \sim D(\theta_1, \ldots, \theta_L)$. Then:

$$
\begin{aligned}
g_l(\theta_{il}) &= \eta_{il} \\
\eta_{il} &= f_{1l}(\mathbf{z}_{i1l}) + \ldots + f_{J_l l}(\mathbf{z}_{iJ_l l})
\end{aligned}
$$

where every θ_l can be modeled with effect types of different subsets of $\mathbf{Z}$.

Bayesian Models for Location, Scale and Shape

Overview

- Coined by Umlauf, Klein, and Zeileis (2017)
- Bayesian variant of Structured Additive Distributional Regression
- „Full“ Bayesian inference with
    1. Posterior distribution maximisation and
    2. Markov Chain Monte Carlo Sampling

Bayesian Models for Location, Scale and Shape

Overview

- Coined by Umlauf, Klein, and Zeileis (2017)
- Bayesian variant of Structured Additive Distributional Regression
- „Full“ Bayesian inference with
  1. Posterior distribution maximisation and
  2. Markov Chain Monte Carlo Sampling

Differences/Advantages over GAMLSS

- Valid credible intervals in comparison to CI based on asymptotics
- Structured Additive Effects
- Support of Multivariate Distributions

but

- Slower estimation

Motivation for bamlss.vis

Motivation for bamlss.vis

Problem

- Often, distribution parameters θ_l do not directly equate to $E(y)$, $Var(y)$

Motivation for bamlss.vis

Problem

- Often, distribution parameters θ_l do not directly equate to $E(y)$, $Var(y)$
- Therefore hard to know influence of covariates on moments because:
    1. Link function $h_l(\cdot)$ transforms effects
    2. Transformed effects are for parameters θ_l, which are often not directly moments

Motivation for bamlss.vis

An example

Consider the censored normal distribution $y^* \sim CN(\mu = 0, \sigma^2 = 1)$ with cut-off point $a = 0$.

The Problem

- Blue line depicts the expected value
- Parameters μ and σ^2 are not the first two moments of the CN distribution!

$\Rightarrow$ Any predicted parameters $\hat{\mu}$ and $\hat{\sigma^2}$ need transformation to $E(y^*)/V(y^*)$

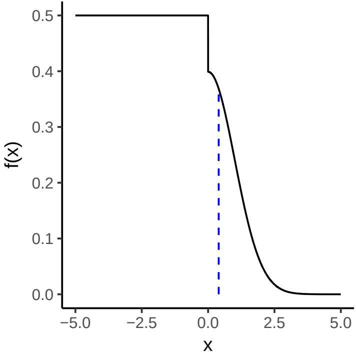

Figure 1: PDF of CN with expected value as blue line.

Motivation for bamlss.vis

Solution

- Thus: Package needed which
  1. Makes it easy to graphically display and compare predicted distributions
  2. Displays the influence of a covariate on the distributional moments
- $\Rightarrow$ `bamlss.vis` was born, solving these problems in R with a Shiny App.

Case-Study

Case-Study

$\Rightarrow$ Use real data to illustrate `bamlss.vis`' capabilities

The Data

- `Wage` dataset, by United States Census Bureau (2011)
- Depicts yearly income in 1000$ of males from the US East Coast based on:
    1. **age**
    2. **year**
    3. **race**
    4. **education**
    5. **health**

Case-Study

First look

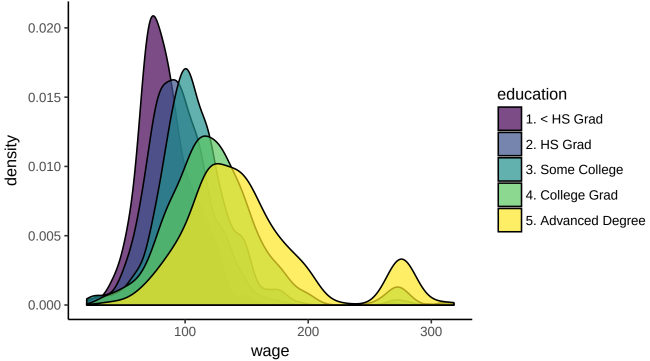

Figure 2: Gaussian kernel density estimates for wages split up by education level.

⇒ Model both μ and σ^2 depending on education level

Case-Study

The model

```
cnorm_model <- bamlss(
list(wage ~ s(age) + race + year + education + health,
     sigma ~ s(age) + race + year + education + health
        ),
data = wage_sub,
family = cnorm_bamlss()
)
```

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

bamlss.vis

Let’s start up `bamlss.vis`!

Installation

You can install `bamlss.vis` today! Run the following code:

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

Thanks!

Thanks for your attention!

Literatur

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`.

J.H. Friedman and W. Stuetzle. Projection pursuit regression. *Journal of the American statistical Association*, 76(376):817–823, 1981.

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.

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].