



GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

distreg.vis (formerly bamLSS.vis)

Interactively visualising distributional regression models

Stanislaus Stadlmann

15 May 2017, Budapest

Georg-August University of Göttingen

Distributional Regression

What is distributional regression?

Overview

Let $y \sim D(\theta_1, \dots, \theta_K)$. With distributional regression...

- Every parameter θ_i can be modeled using a (different) set of predictors
- Predictors can take different forms, e.g. non-linear, spatial, random effects
- Parameters are connected to the predictors using link-functions that uphold the support of the parameter

Model Equations

⇒ This allows for extremely flexible model equations:

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

Why do you need distributional regression?

Motivation

⇒ Interested in modeling parameters beyond the mean.
Let's say you are interested in income distributions...

Why do you need distributional regression?

Motivation

⇒ Interested in modeling parameters beyond the mean.
Let's say you are interested in income distributions...

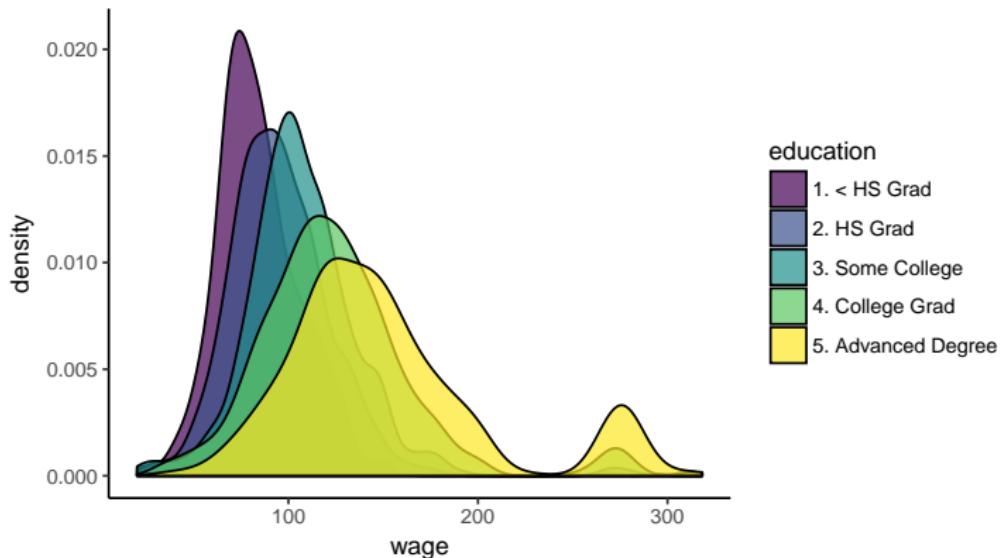


Figure 1: Kernel density estimates for wages split up by education level.

Current implementations

Available model classes

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)

Both have R implementations, called **bamlss** and **gamlss**,
respectively.

Current implementations

Available model classes

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)

Both have R implementations, called **bamlss** and **gamlss**, respectively.

Differences

- Backfitted Maximum Likelihood Estimation (GAMLSS)
- Posterior Maximisation and MCMC sampling (BAMLSS)
- Multivariate Distributions (BAMLSS)
- Structured Additive Effects (BAMLSS)

Rising Usage

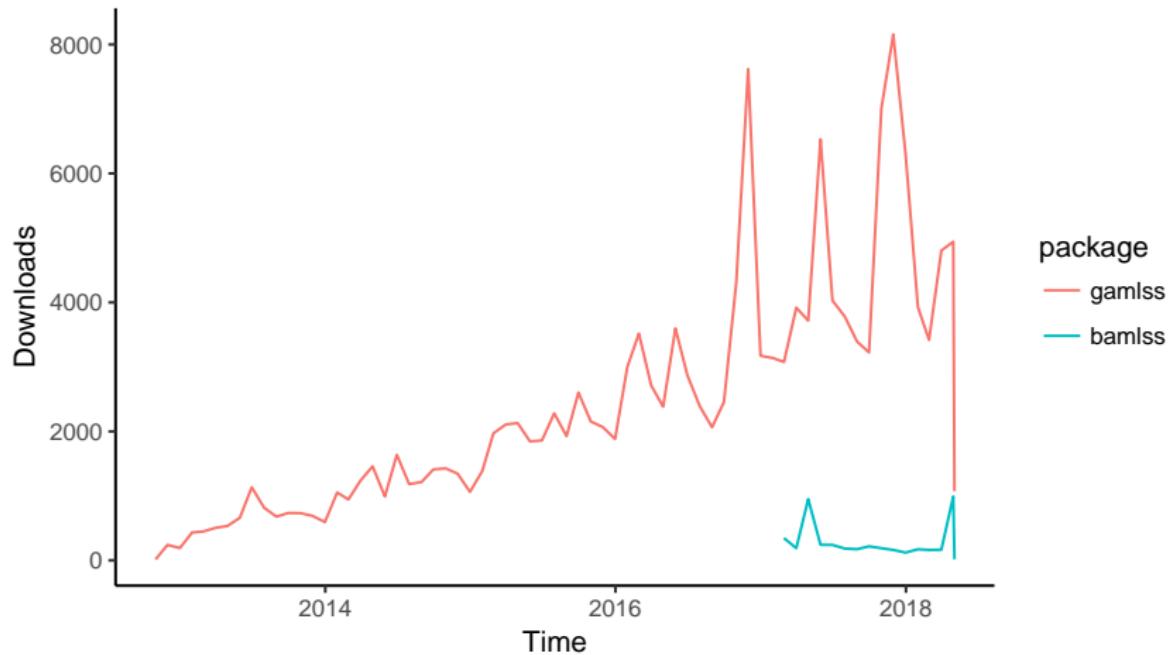


Figure 2: Monthly downloads of **gamlss** and **bamllss** R packages.

Motivation for distreg.vis

Motivation for distreg.vis

Difficulties in distributional regression

- In many cases, 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
 3. Non-Parametric effects are on their own often hard to interpret.

Motivation for distreg.vis

An example

Consider the log normal distribution $y \sim \text{LOGNO}(\mu = -1, \sigma = 0.8)$.

The Problem

- Blue line depicts the expected value
- Parameters μ and σ are not the first two moments of the log normal but of the **underlying normal distribution!**

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

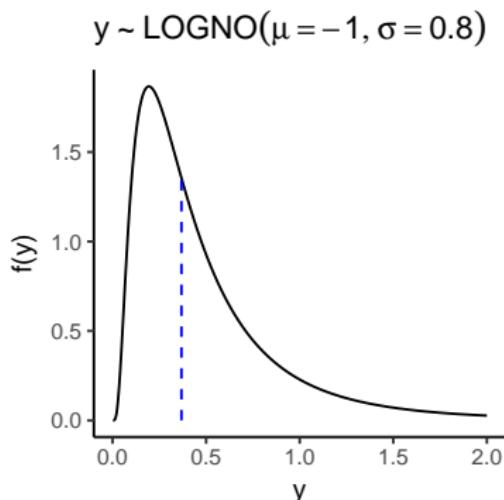


Figure 3: PDF of LOGNO with expected value as blue line.

Motivation for distreg.vis

Difficulties

```
-----  
Mu link function: identity  
Mu Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 1.326103  0.015273  86.83 <2e-16 ***  
ps(norm1)    0.020122  0.001225   16.43 <2e-16 ***  
ps(norm2)    0.021709  0.002008   10.81 <2e-16 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
-----  
Sigma link function: log  
Sigma Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 3.13103  0.21200  14.711 < 2e-16 ***  
ps(norm1)   -0.10054  0.02022  -4.973 0.00000092 ***  
ps(norm2)    0.10756  0.02577   4.174 0.00003552 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

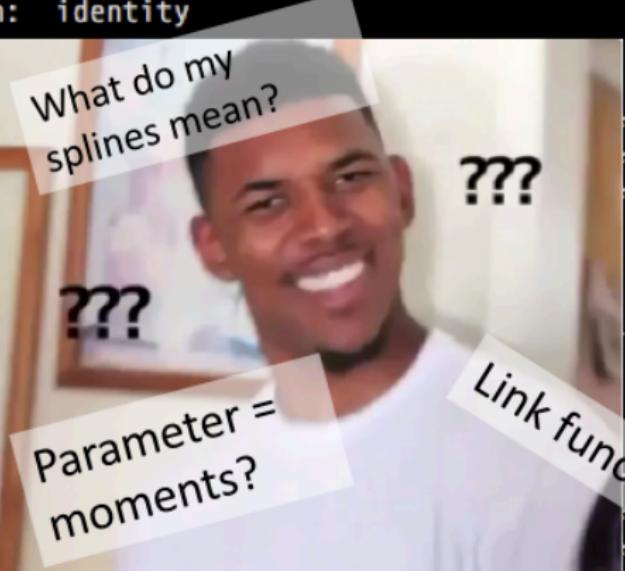
Figure 4: Output from a fitted LOGNO GAMLSS.

Motivation for distreg.vis

Difficulties

```
Mu link function: identity
Mu Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.0000000 0.0000000 -1.0000000 0.1000000
ps(norm1) 0.0000000 0.0000000 0.0000000 0.1000000
ps(norm2) 0.0000000 0.0000000 0.0000000 0.1000000
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Sigma link function: identity
Sigma Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.0000000 0.0000000 -3.0000000 0.1000000
ps(norm1) -0.0000000 0.0000000 -0.0000000 0.1000000
ps(norm2) 0.0000000 0.0000000 0.0000000 0.1000000
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



What do my splines mean?

???

Parameter = moments?

Link function?

Figure 5: Confused man.

Motivation for `distreg.vis`

Solution

Thus: Package needed which

1. Helps users view and compare their complete predicted distributions
2. Makes interpretation of covariate effects easier

⇒ `distreg.vis` was born!

distreg.vis

distreg.vis - an R Package

Distreg.vis seeks to achieve the following goals:

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

The core of the package is represented by two functions:

1. `distreg.vis::plot_dist()`
2. `distreg.vis::plot_moments()`

Both **gamlss** and **bamlss** families are supported.

The plot_dist function

The Code

```
1 # Fitting the model
2 wage_model <- gamlss(wage ~ ps(age) + race + education,
3                         ~ ps(age) + race + education,
4                         data = wage_sub, family = LOGNO())
5
6 # Obtaining predicted parameters
7 predicted_params <- preds(wage_model, newdata = ndata)
8 predicted_params
9 #> mu      sigma
10 #> 1 4.407968 0.2431475
11 #> 2 4.543124 0.2951681
12 #> 3 4.674615 0.2580149
13 #> 4 4.749154 0.3398120
14 #> 5 5.044413 0.3094951
15
16 # Make the plot
17 distreg.vis:::plot_dist(wage_model, predicted_params)
```

The `plot_dist` function

Output

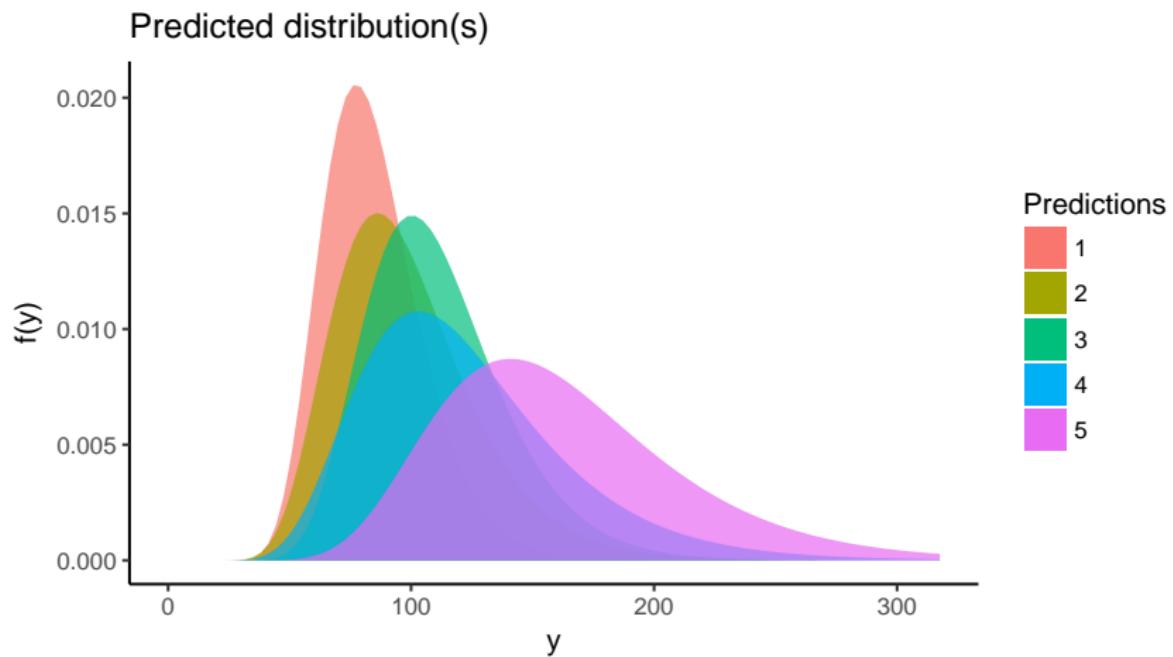


Figure 6: Predicted distributions based on `predicted_params`

The plot_moments function

Code

```
1 | plot_moments(logno_gamlss, "age", newdata)
```

Output

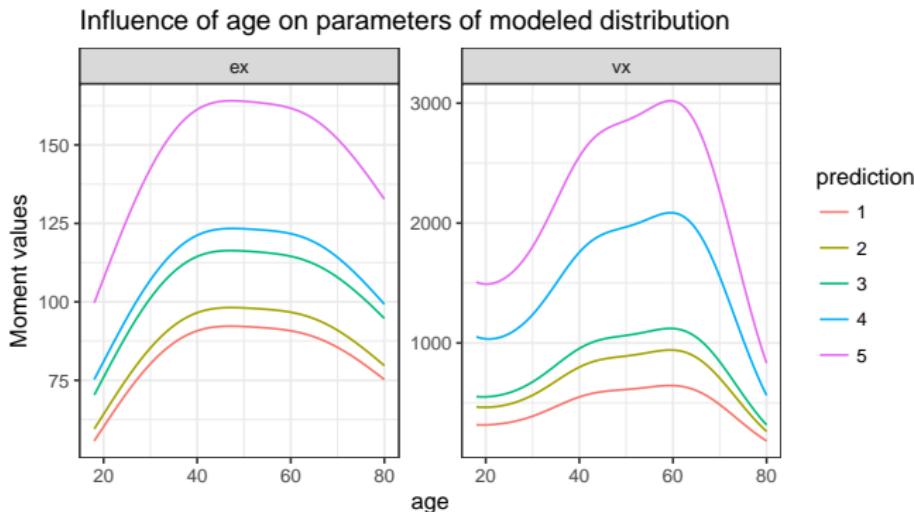


Figure 7: Predicted moments for a specific variable age



Overview

`distreg.vis` provides the right functions to help interpret distributional regression models. **But:**

- Users might have a hard time remembering the correct functions including their arguments
- Users want to quickly compare different prediction scenarios

⇒ The perfect use-case for the Shiny framework.

Shiny Application

Starting the app

```
1 | distreg.vis::vis()
```

or call via RStudio:

Shiny Application

Starting the app

```
1 | distreg.vis::vis()
```

or call via RStudio:

Shiny Application

Model Overview

Shiny Application

Model Scenarios

Shiny Application

Selecting a scenario

Shiny Application

Selecting multiple scenarios

Shiny Application

Colour palette

Shiny Application

Obtain the code

Shiny Application

Edit table tab

Shiny Application

Properties tab

Things yet to come

- ✓ Support both **gamlss** and **bamlss** objects
- Involve many people to carve out all possible bugs
- Support ALL `gamlss.dist` families
- More effects: spatial terms, tensor splines
- Possibility to include user-defined metrics which depend on parameters of a distribution (e.g. Gini coefficient)
- Support multivariate distributions

and then...

- CRAN admission
- Submit to the Journal of Statistical Software

More information



Code and Installation Guide

- To see my code and a guide, visit
github.com/Stan125/distreg.vis.
- This presentation can be found under
github.com/Stan125/erum2018.

Thanks!

Thanks for your attention!

References i

Literatur

- 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>.