Package ‘cenGAM’

July 22, 2025

Type Package

Title Censored Regression with Smooth Terms

Version 0.5.4

Date 2025-04-11

Depends R (>= 3.2.5), mgcv (>= 1.8-19)

Imports stats

ByteCompile yes

Description

Implementation of Tobit type I and type II families for censored regression using the 'mgcv' package, based on methods detailed in Woods (2016) <doi:10.1080/01621459.2016.1180986>.

License GPL (>= 2)

Repository CRAN

Encoding UTF-8

LazyData true

NeedsCompilation no

Author Zhou Fang [aut, cre]

Maintainer Zhou Fang <`zhou.fang@bioss.ac.uk`>

Date/Publication 2025-04-11 12:40:02 UTC

Contents

`cenGAM-package` *Censored response additive modelling for mgcv*

Description

`cenGAM` provides `tobit1` and `tobit2` families for generalized additive modelling with the `gam` function in the `mgcv` package.

Under the Tobit I model, the user supplies a left and/or right threshold (optionally differing between observations, and potentially infinite to denote no censorship) and the response is assumed to be censored if it falls over this threshold. Under the Tobit II model, a more generalised model is assumed where we fit a second additive model that gives whether each observation is censored, with a possible correlation between the error in the response and the censorship.

Seen help for the individual functions for more details.

Author(s)

Zhou Fang <zhou.fang@bioss.ac.uk>

with contributions and help from Javier Palarea.

Examples

```
## see examples for tobit1 and tobit2
```

`nutrient` *Nutrient dataset*

Description

Simulated nutrient concentration data, with 1200 rows and 3 columns.

Format

This data frame contains the following columns:

day - a day of the year

location - one of four locations A B C D

y - a nutrient value, either a number or a censoring threshold

---

`tobit1` *Tobit I family for censored GAM*

Description

This function implements the Tobit I family for the mgcv package.

Usage

```
tobit1(link="identity", left.threshold=-Inf, right.threshold=Inf,
theta=NULL, initial.theta=0)
```

Arguments

`link`	The link function: one of `"log"`, `"identity"`, `"inverse"`, `"sqrt"`, or a `power` link.
`left.threshold`	Threshold value, for which response values below this will be censored. Can be a scalar or a vector of same length as the response.
`right.threshold`	Threshold value, for which response values above this will be censored. Can be a scalar or a vector of same length as the response.
`theta`	The log std error. If left NULL, it is estimated.
`initial.theta`	Optional parameter to set initial theta value if it is being estimated - change this value if the variance is very different from 1.

Details

Under the Tobit I model, given a value sigma and a conditional mean of mu, and left and right threshold values lt and rt, response values y are distributed as

lt if z < lt

rt if z > rt

z otherwise

where z is distribution Normal(mu, sigma). (Note that we have extended the model to include right as well as left censoring.

This function allows a non-linear relationship be estimated between mu and the covariates in a restricted maximum likelihood approach, via application of Wood (2016). We allow differing levels of censorship across dataset by allowing the left and right thresholds to be different between data points.

See the examples for more details of how to fit in practice.

Value

An object inheriting from class `family` for use with the mgcv package.

References

Wood, S.N., N. Pya and B. Saefken (2016), Smoothing parameter and model selection for general smooth models. Journal of the American Statistical Association. <URL: http://arxiv.org/abs/1511.03864>

See Also

gam,ziplss

Examples

```
# Generate random data
set.seed(1)
x <- matrix(2*rnorm(300), 100)
yn <- 2*x[,3] + 4*cos(x[,1]*2)
y <- yn + rnorm(100)
ycensored <- pmax(y, 0) # data left-censored at 0
ycensored <- pmin(ycensored, 4) # data right-censored at 4

par(mfrow = c(3,3))

# True model
plot(gam(y ~ s(x[,1]) + s(x[,2]) + s(x[, 3])), ylim=c(-5, 5), main = "True")

# Naive estimation
plot(gam(ycensored ~ s(x[,1]) + s(x[,2]) + s(x[, 3])), ylim=c(-5, 5), main = "Naive")

# Tobit I estimation
m <- gam(ycensored ~ s(x[,1]) + s(x[,2]) + s(x[, 3]), family = tobit1(left.threshold=0))
summary(m) #note x[,2] is not significant
m$family$getTheta(FALSE) #estimate of theta
m$family$getTheta(TRUE) #estimate of sigma = exp(theta)
plot(m, ylim = c(-5, 5), main = "Tobit I")

# More realistic dataset requires some data processing
data(nutrient)
# For this dataset, all values >1 are censored. At location D values are censored below at 0, other
# locations are censored at 0.1.
head(nutrient)
nut = data.frame(day = nutrient$day, location = factor(nutrient$location), y=as.numeric(nutrient$y))
table(nutrient$y[is.na(nut$y)])
summary(nut$y)
nut$upper = 10
nut$lower = ifelse(nut$location == "D", 0, 1)
# Recode the data to communicate which is and isn't censored
nut$y[nutrient$y %in% c("<1", "<0")] = -Inf
nut$y[nutrient$y %in% c(">10")] = Inf
# Missing values are best removed here, or can cause confusion later
nut = na.omit(nut)

# Fit including a random effect for location
m = gam(y~ s(day) + s(location, bs="re") , data = nut,
family = tobit1(left.threshold=nut$lower, right.threshold = nut$upper))
```

```
gam.vcomp(m)
anova(m)
summary(m)
```

`tobit2` *Tobit II family for censored GAM*

Description

This function implements the Tobit II family for the mgcv package.

Usage

```
tobit2(link=list("identity","identity","log","logit2" ),
censoring = FALSE, rho=NULL, eps = 1e-3)
```

Arguments

`link`	The link functions: Corresponds to `mu1`, `mu2`, `sigma` and `rho` respectively.
`censoring`	Vector of TRUE/FALSE values to denote censorship. TRUE values are censored
`rho`	Value of rho. If NULL, is estimated.
`eps`	Parameter to perturb rho in estimation if very close to -1 or 1.

Details

Under the Tobit II model, given a value sigma and a conditional mean of mu1, and a censoring parameter mu2, response values are censored if mu2 + epsilon2 < 0, and mu1 + sigma*epsilon1 otherwise.

Here epsilon1, epsilon2 are distributed Normal(0, 1) with correlated rho.

This function allows a non-linear relationship be estimated between mu1, mu2, sigma, rho and the covariates in a restricted maximum likelihood approach, via application of Wood (2016). Note that this allows for heteroskedastic errors.

Estimation of rho depends on the distributional qualities near the censorship boundary, and is hence typically very inaccurate for typical sample sizes. Hence in practice it is often better to supply a value of rho (for example 0 to imply independent censorship) instead. eps is used when estimating rho to avoid errors when rho is close to 1 or -1. Smaller values may produce more accurate results.

This method is still currently very *experimental*. It's not suggested to be used to important applications. Errors can occur if the default starting point for the function cause problems, consider changing the `start` argument to `gam`.

Value

An object inheriting from class `family` for use with the mgcv package.

References

Wood, S.N., N. Pya and B. Saefken (2016), Smoothing parameter and model selection for general smooth models. Journal of the American Statistical Association. <URL: http://arxiv.org/abs/1511.03864>

See Also

`gam`,`ziplss`

Examples

```
# Generate a small example
set.seed(1)
x <- matrix(2*rnorm(400), 200)
yn <-  x[,1]^2 + x[,2]
y <- yn + rnorm(200)
censored <- (rnorm(200) + 2*x[,2]+1) < 0 #censored according to x[,2]
ycensored <- replace(y, censored, 0)
m <- gam(c(ycensored ~ s(x[,1]) + s(x[,2]) , ~x[,1]+x[,2], ~1,~1),
family = tobit2(censoring = censored)) #estimated rho
par(mfrow = c(3,2))
plot(gam(y ~ s(x[,1]) + s(x[,2]) ), ylim=c(-5, 5), main = "True")
plot(m, ylim = c(-5, 5), main = "Tobit II estimated rho")

summary(m)
m$fitted #gives for each observation fitted mu1, mu2, sigma, rho

m2 <- gam(c(ycensored ~ s(x[,1]) + s(x[,2]) , ~x[,1]+x[,2], ~1),
family = tobit2(censoring = censored, rho=0)) #non estimated rho
plot(m2, ylim = c(-5, 5), main = "Tobit II fixed rho")

## Not run:
#Larger example
set.seed(1)
x <- matrix(2*rnorm(1500), 500)
yn <- 2*x[,3] + 4*cos(x[,1]*2)
y <- yn + 3*rnorm(500)
censored <- (rnorm(500) + 2*x[,2]) < 0 #censored according to x[,2]
ycensored <- replace(y, censored, 0)

par(mfrow = c(3,3))

# True model
plot(gam(y ~ s(x[,1]) + s(x[,2]) + s(x[, 3])), ylim=c(-5, 5), main = "True")

# Naive estimation
plot(gam(ycensored ~ s(x[,1]) + s(x[,2]) + s(x[, 3])), ylim=c(-5, 5), main = "Naive")

# Tobit II estimation
m <- gam(c(ycensored ~ s(x[,1]) + s(x[,2]) + s(x[, 3]), ~x[,1]+x[,2]+x[,3], ~1,~1),
family = tobit2(censoring = censored))
plot(m, ylim = c(-5, 5), main = "Tobit II")
```

```
#fitting with non-estimated rho
m2 <- gam(c(ycensored ~ s(x[,1]) + s(x[,2]) + s(x[, 3]), ~x[,1]+x[,2]+x[,3],~1),
family = tobit2(censoring = censored, rho=0))

## End(Not run)
```

Index