

Package ‘causalreg’

October 1, 2026

Title Causal Generalized Linear Models

Version 0.3.0

Description An implementation of methods for causal discovery in a structural causal model where the conditional distribution of the target node is described by a generalized linear model conditional on its causal parents.

License GPL-3

Encoding UTF-8

URL <https://github.com/franciscorichter/causalreg>

BugReports <https://github.com/franciscorichter/causalreg/issues>

RoxygenNote 7.3.3

Imports mgcv, parallel, Rcpp

LinkingTo Rcpp, RcppArmadillo

Suggests knitr, rmarkdown, testthat (>= 3.0.0)

VignetteBuilder knitr

Config/testthat/edition 3

NeedsCompilation yes

Author Veronica Vinciotti [aut, cre],
Ernst C. Wit [aut],
Francisco Richter [aut]

Maintainer Veronica Vinciotti <veronica.vinciotti@unitn.it>

Repository CRAN

Date/Publication 2026-10-01 19:50:02 UTC

Contents

boot_pval_cpp	2
cgam	2
cglm	5
eval_submodels_cpp	7

2	<i>cgam</i>
fast_fit_and_stat	8
fast_glm_bic	8
fast_glm_fit	9
glm_loglik_cpp	9
pearson_stat_cpp	10
Index	11

boot_pval_cpp	<i>Bootstrap p-value for Pearson risk = 1 test (fully in C++)</i>
---------------	---

Description

Bootstrap p-value for Pearson risk = 1 test (fully in C++)

Usage

boot_pval_cpp(X, y, family, B = 100L)

Arguments

- X Design matrix
- y Response vector
- family "poisson" or "binomial"
- B Number of bootstrap replicates

Value

Two-sided bootstrap p-value

cgam	<i>Causal generalized additive model</i>
------	--

Description

This function does a search for a causal submodel within the generalized additive model provided.

Usage

```

cgam(
  formula,
  family,
  data,
  alpha = 0.05,
  pval = c("bootstrap", "chi-square"),
  B = 100,
  search = c("all", "stepwise"),
  ncores = 1L,
  fast_gam = TRUE,
  direction = c("forward", "backward"),
  ...
)

```

Arguments

formula	A formula object.
family	The response family as a character string, "poisson" or "binomial". A family object such as poisson() is not accepted.
data	A data frame containing the variables in the model.
alpha	Significance level for statistical test.
pval	If pval="bootstrap", a bootstrap test is conducted to test whether Pearson risk is 1. When family="poisson" a chi-squared test can be conducted by setting pval="chi-square".
B	Number of bootstrap samples when pval="bootstrap". Default is 100.
search	"all" (the default) fits every non-empty submodel of the candidate terms; "stepwise" runs a stepwise search in the direction set by direction.
ncores	Number of cores for parallel computation. Default is 1 (sequential). When ncores > 1, model evaluations are distributed across cores using forking (parallel::mclapply) on Unix/macOS and a PSOCK cluster (parallel::parLapply) on Windows, so parallelization is available on all platforms. The backend can be overridden with options(causalreg.parallel =) ("auto", "fork", "psock", or "sequential").
fast_gam	Logical; only relevant when pval="bootstrap". If TRUE, the smoothing parameters selected on the original data (once per candidate model) are held fixed across that model's bootstrap resamples, so each resample fit skips mgcv's REML/GCV smoothing-parameter selection. This typically gives a 3-4x speedup (composable with ncores) and the same model selection, but it is an approximation: individual bootstrap p-values can differ from the fully re-selected version, more so at small B. Default is TRUE; set fast_gam=FALSE to re-select the smoothing parameters on every resample (the exact bootstrap used in earlier versions, slower). Has no effect on cglm or on the chi-square test.
direction	Direction of the stepwise search, "forward" (default) or "backward". Ignored when 'search = "all"'. Forward starts from the intercept model and adds the variable giving the largest p-value; backward starts from the full model and

removes the variable whose removal gives the largest p-value, stopping once the model is no longer rejected. Both then prune by BIC. For a binomial response whose candidate set contains a categorical variable, the backward search is used and a message is emitted, because the Pearson risk of a binary regression on categorical covariates alone is mathematically equal to 1.

... Further arguments to be passed to the gam function.

Value

A list containing the selected causal submodel and search diagnostics.

References

Polinelli, A., V. Vinciotti and E.C. Wit. (2026). "Causal generalized linear models via Pearson risk invariance." *Journal of Causal Inference*, 14(1), 20240043. doi:[10.1515/jci20240043](https://doi.org/10.1515/jci20240043)

Examples

```
#####
#causal Poisson gam#####
n<-1000
set.seed(123)
X1<-rnorm(n,0,1)
Y<-rpois(n,exp(sin(X1)))
X2<-log(Y+1)+rnorm(n,0,0.5)
data<-data.frame(X1, X2, Y)
cm_all<-cgam(Y ~ s(X1)+s(X2),"poisson",data,pval="chi-square",search="all")
cm_all$model.opt
cm_step<-cgam(Y ~ s(X1)+s(X2),"poisson",data,pval="chi-square",search="stepwise")
cm_step$model.opt

#bigger simulation with 7 covariates
set.seed(123)
n<-1000
X1<-rnorm(n=n,sd=sqrt(0.04))
X2<-X1+rnorm(n=n,sd=sqrt(0.04))
X3<-X1+X2+rnorm(n=n,sd=sqrt(0.04))
m<-sin(X2*5)+X3^3
Z<-m+rnorm(n=n,sd=sqrt(0.04))
X4<-X2+rnorm(n=n,sd=sqrt(0.04))
X5<-Z+rnorm(n=n,sd=sqrt(0.04))
X6<-Z+rnorm(n=n,sd=sqrt(0.04))
X7<-X6+rnorm(n=n,sd=sqrt(0.04))
Y<-qpois(pnorm(Z, mean = m, sd = sqrt(0.04)), lambda=exp(m))
dat<-data.frame(X1, X2, X3, X4, X5, X6, X7,Y)
fml<- Y~s(X1)+s(X2)+s(X3)+s(X4)+s(X5)+s(X6)+s(X7)
mod.all <-cgam(fml,"poisson",dat,pval="chi-square",search="all")
mod.all$model.opt
mod.step <-cgam(fml,"poisson",dat,pval="chi-square",search="stepwise")
mod.step$model.opt
#####
#causal logistic gam#####
```

```

n<-1000
set.seed(123)
X1<-rnorm(n,0,1)
Y<-rbinom(n,1,exp(X1)/(1+exp(X1)))
flip<-rbinom(n,1,0.1)
X2<-(1-flip)*Y+rnorm(n,0,0.3)
data<-data.frame(X1, X2, Y)
set.seed(1)
cm_all<-cgam(Y ~ s(X1)+s(X2),"binomial",data,pval="bootstrap",search="all")
cm_all$model.opt
set.seed(1)
cm_step<-cgam(Y ~ s(X1)+s(X2),"binomial",data,pval="bootstrap",search="stepwise")
cm_step$model.opt

```

cglm

Causal generalized linear model

Description

This function does a search for a causal submodel within the generalized linear model provided.

Usage

```

cglm(
  formula,
  family,
  data,
  alpha = 0.05,
  pval = c("bootstrap", "chi-square"),
  B = 100,
  search = c("all", "stepwise"),
  ncores = 1L,
  use_cpp = TRUE,
  direction = c("forward", "backward"),
  ...
)

```

Arguments

formula	A formula object.
family	The response family as a character string, "poisson" or "binomial". A family object such as poisson() is not accepted.
data	A data frame containing the variables in the model.
alpha	Significance level for statistical test
pval	If pval="bootstrap", a bootstrap test is conducted to test whether Pearson risk is 1. When family="poisson" a chi-squared test can be conducted by setting pval="chi-square".

B	Number of bootstrap samples when <code>pval="bootstrap"</code> . Default is 100.
search	"all" (the default) fits every non-empty submodel of the candidate terms; "stepwise" runs a stepwise search in the direction set by <code>direction</code> .
ncores	Number of cores for parallel computation. Default is 1 (sequential). When <code>ncores > 1</code> , model evaluations are distributed across cores using forking (<code>parallel::mclapply</code>) on Unix/macOS and a PSOCK cluster (<code>parallel::parLapply</code>) on Windows, so parallelization is available on all platforms. The backend can be overridden with <code>options(causalreg.parallel =)</code> ("auto", "fork", "psock", or "sequential").
use_cpp	Logical; if TRUE (default), use fast C++ implementations for supported families.
direction	Direction of the stepwise search, "forward" (default) or "backward". Ignored when <code>search = "all"</code> . Forward starts from the intercept model and adds the variable giving the largest p-value; backward starts from the full model and removes the variable whose removal gives the largest p-value, stopping once the model is no longer rejected. Both then prune by BIC. For a binomial response whose candidate set contains a categorical variable, the backward search is used and a message is emitted, because the Pearson risk of a binary regression on categorical covariates alone is mathematically equal to 1.
...	Further arguments to be passed to the <code>glm</code> function.

Value

A list containing the selected causal submodel and search diagnostics.

References

Polinelli, A., V. Vinciotti and E.C. Wit. (2026). "Causal generalized linear models via Pearson risk invariance." *Journal of Causal Inference*, 14(1), 20240043. doi:[10.1515/jci20240043](https://doi.org/10.1515/jci20240043)

Examples

```
#####
#causal Poisson glm#####
n<-1000
set.seed(123)
X1<-rnorm(n,0,1)
Y<-rpois(n,exp(X1))
X2<-log(Y+1)+rnorm(n,0,0.3)
data<-data.frame(X1, X2, Y)
cm_all<-cglm(Y ~ X1+X2,"poisson",data,pval="chi-square",search="all")
cm_all$model.opt
cm_step<-cglm(Y ~ X1+X2,"poisson",data,pval="chi-square",search="stepwise")
cm_step$model.opt

#####
#causal logistic glm#####
n<-2000
set.seed(123)
X1<-rnorm(n,0,1)
```

```

Y<-rbinom(n,1,exp(X1)/(1+exp(X1)))
flip<-rbinom(n,1,0.1)
X2<-(1-flip)*Y+rnorm(n,0,0.3)
data<-data.frame(X1, X2, Y)
set.seed(1)
cm_all<-cglm(Y ~ X1+X2,"binomial",data,pval="bootstrap",search="all")
cm_all$model.opt
set.seed(1)
cm_step<-cglm(Y ~ X1+X2,"binomial",data,pval="bootstrap",search="stepwise")
cm_step$model.opt
#bigger simulation with 5 covariates
set.seed(12)
n<-3000
X1<-rnorm(n,0,1)
X2<-rnorm(n,X1,0.5)
X3<-rnorm(n,0,1)
X4<-rnorm(n,X2,.5)
Y<-rbinom(n,1,exp(.8*X2-.9*X3)/(1+exp(.8*X2-.9*X3)))
flip<-rbinom(n,1,0.1)
X5<-(1-flip)*Y+flip*(1-Y)+rnorm(n,0,.3)
dat<-data.frame(X1, X2, X3, X4, X5,Y)
set.seed(1)
mod.all <-cglm(Y~X1+X2+X3+X4+X5,"binomial",dat,pval="bootstrap",search="all")
mod.all$model.opt
set.seed(1)
mod.step <-cglm(Y~X1+X2+X3+X4+X5,"binomial",dat,pval="bootstrap",search="stepwise")
mod.step$model.opt

```

eval_submodels_cpp	<i>Evaluate multiple submodels in C++ (for exhaustive search)</i>
--------------------	---

Description

Given a full design matrix and a list of column index vectors, fit each submodel and return Pearson risk, p-value, BIC.

Usage

```
eval_submodels_cpp(X_full, y, family, col_indices, pval_method, B = 100L)
```

Arguments

X_full	Full design matrix (with intercept)
y	Response vector
family	"poisson" or "binomial"
col_indices	List of integer vectors, each specifying columns of X_full
pval_method	"chi-square" or "bootstrap"
B	Number of bootstrap replicates

Value

List of vectors: pearson_risk, pval, bic

fast_fit_and_stat	<i>Fit GLM and compute Pearson risk, BIC, and optionally bootstrap p-value</i>
-------------------	--

Description

Fit GLM and compute Pearson risk, BIC, and optionally bootstrap p-value

Usage

fast_fit_and_stat(X, y, family)

Arguments

X	Design matrix
y	Response vector
family	"poisson" or "binomial"

Value

List with pearson_stat, pearson_risk, bic

fast_glm_bic	<i>Compute BIC for a GLM fit</i>
--------------	----------------------------------

Description

Compute BIC for a GLM fit

Usage

fast_glm_bic(y, mu, family, p, n)

Arguments

y	Response vector
mu	Fitted values
family	"poisson" or "binomial"
p	Number of parameters
n	Number of observations

Value

BIC value

fast_glm_fit	<i>Fast GLM fit via IRLS</i>
--------------	------------------------------

Description

Fast GLM fit via IRLS

Usage

```
fast_glm_fit(X, y, family, max_iter = 25L, tol = 1e-08)
```

Arguments

X	Design matrix (n x p), including intercept column if needed
y	Response vector (n)
family	"poisson" or "binomial"
max_iter	Maximum IRLS iterations (default 25)
tol	Convergence tolerance (default 1e-8)

Value

List with fitted_values (mu) and converged flag

glm_loglik_cpp	<i>Compute log-likelihood for Poisson or Binomial GLM</i>
----------------	---

Description

Compute log-likelihood for Poisson or Binomial GLM

Usage

```
glm_loglik_cpp(y, mu, family)
```

Arguments

y	Response vector
mu	Fitted values
family	"poisson" or "binomial"

Value

Log-likelihood value

pearson_stat_cpp	<i>Compute Pearson chi-square statistic from y and mu</i>
------------------	---

Description

Compute Pearson chi-square statistic from y and mu

Usage

```
pearson_stat_cpp(y, mu, family)
```

Arguments

y	Response vector
mu	Fitted values
family	"poisson" or "binomial"

Value

Pearson chi-square statistic (sum of squared Pearson residuals)

Index

boot_pval_cpp, [2](#)

cgam, [2](#)

cglm, [5](#)

eval_submodels_cpp, [7](#)

fast_fit_and_stat, [8](#)

fast_glm_bic, [8](#)

fast_glm_fit, [9](#)

glm_loglik_cpp, [9](#)

pearson_stat_cpp, [10](#)