Package 'fishMod'

September 23, 2016

Title Fits Poisson-Sum-of-Gammas GLMs, Tweedie GLMs, and Delta
Log-Normal Models

Version 0.29

Author Scott D. Foster

Description Fits models to catch and effort data. Single-species models are 1) delta log-
normal, 2) Tweedie, or 3) Poisson-gamma (G)LMs.

Maintainer Scott Foster <scott.foster@csiro.au>

License GPL

SystemRequirements C++11

NeedsCompilation yes

Repository CRAN

Date/Publication 2016-09-23 00:44:13

R topics documented:

`fishMod-package` *How to use fishMod*

Description

fishMod is a package that contains some very specific functions for a) fitting extensions of generalised linear models (GLMs) that allow for non-negative observations with exact zeros, and b) targetted fisheries catch and effort data using a (quite complex) mixture model. The package is intended to be used in reasonable anger on real data sets but *it does* have limitations. Primarily, these arise from not nesting the models within the hierarchy of standard R functions. This means that there are currently no generic methods, such as 'plot' and 'summary' for the objects returned by the functions. Instead, the analyst must fish around (excuse the pun) for the bits and peices required for the task at hand, for example residuals and fitted values for diagnostic plots.

The important functions in the fishMod namespace can be arranged into two groups: 1) dPoisGam, pPoisGam, rPoisGam, dPoisGamDerivs, dTweedie, pTweedie, rTweedie; and 2) deltaLN, pgm, tglm, tglm.fit, simReg. These correspond to functions for 1) utility functions for Tweedie densities and Poisson-gamma representations of Tweedies; and 2) extensions of GLMs for non-negative data.

Note that there is an extended version of this package (available from the author but not from CRAN) that also contains functions for multi-species catch and effort data that attempt to account for targetting.

Author(s)

Scott D. Foster

`deltaLN` *Fitting models based on the Delta Log-Normal distribution.*

Description

Fits a compound model that assumes a Delta Log-Normal distribution. The mean of the log-normal process and the mean of the binary process are allowed to change with covariates.

Usage

```
deltaLN( ln.form, binary.form, data, residuals=TRUE)
```

Arguments

ln.form   an object of class "formula" (or one that can be coerced to that class). This is a symbolic representation of the model for the log-normal variable. Note that offset terms (if any) should be included in this part of the model.

binary.form	an object of class "formula" (or one that can be coerced to that class). This is a symbolic representation of the model for the binary variable and should not contain an outcome (e.g. ~1+var1+var2).
data	a data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the variables in the model.
residuals	boolean indicating if the quantile residuals should be calculated. Default is TRUE indicating residuals are to be calculated.

Details

The observed random variables y_i are assumed to arise from a process that has a non-zero probability that y_i is greater than zero; further, the distribution of y_i conditional on y_i>0 follows a log-normal distribution. This modelling framework models the mean of the conditional distribution and the probability of obtaining a non-zero.

The means of each component of the model are specified in ln.form and binary.form for the log-normal and the zero/non-zero model components respectively. The binary.form formula should not contain an outcome. The binary part of the model is done using a logistic link funciton.

If residuals are requested then two types are returned: Pearson residuals and randomised quantile residuals, described in general by Dunn and Smyth (1996).

Value

pgm returns an object of class "DeltaLNmod", a list with the following elements

coef	the estiamted coefficients from the fitting process. A list with an element for the binary and log-normal parts of the model as well as an element for the standard deviation of the log-normal.
logl	the maximum log likelihood (found at the estimates).
AIC	an Information Criteria.
BIC	Bayesian Information Criteria.
fitted	fitted values of the delta log-normal variable.
fitted.var	variance of the fitted delta log-normal variable.
residuals	a 2-column matrix whose first column contains the randomised quantile residuals and whose second column contains the Pearson residuals.
n	the number of observations used to fit the model.
ncovars	the number of parameters in the combined model.
nzero	the number of non-zero elements.
lnMod	the lm object obtained from fitting the log-normal (non-zero) part of the model.
binMod	the glm object obtained from fitting the zero / non-zero part of the model.

Author(s)

Scott D. Foster

References

Aitchison J. (1955) On the Distribution of a Positive Random Variable Having a Discrete Probability Mass at the Origin. *Journal of the American Statistical Association* **50** 901-908.

Dunn P. K and Smyth G. K (1996) Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics* **5**: 236-244.

Foster, S.D. and Bravington, M.V. (2013) A Poisson-Gamma Model for Analysis of Ecological Non-Negative Continuous Data. *Journal of Environmental and Ecological Statistics* **20**: 533-552

---

dPoisGam   *Density, derivatives, distribution function, and random generation for the Poisson-Gamma distribution.*

Description

Density, derivatives, distribution function, and random generation for the Poisson-Gamma distribution. The distribution is parameterised by the Poisson mean, the Gamma mean and the Gamma dispersion parameter. Derivatives are with respect to these three parameters.

Usage

```
dPoisGam( y, lambda, mu.Z, alpha, LOG=TRUE)
dPoisGamDerivs( y, lambda, mu.Z, alpha, do.checks=TRUE)
pPoisGam( q, lambda, mu.Z, alpha)
rPoisGam( n, lambda, mu.Z, alpha)
```

Arguments

y,q	vector of quantiles
n	number of random draws
lambda	scalar or vector (length matches y or equal to n) of Poisson means
mu.Z	scalar or vector (length matches y or equal to n) of Gamma means
alpha	scalar or vector (length matches y or equal to n) of Gamma dispersions
LOG	indication of return scale. If TRUE (default) then the density is returned on the log scale.
do.checks	boolean indicating if checks on arguments should be performed. If TRUE (default) then checks will be performed.

Details

The observed random variables y_i are assumed to arise from the process: $y_i = sum(z_{i1} + z_{i2} + ... + z_{in_i})$ where n_i is a Poisson variable with mean lambda and z_{ij} is a Gamma variable with mean mu.Z and variance mu.Z^2 / alpha.

The density calculation is based on the series summation method described in Dunn and Smyth (2005). However, the parameterisation used here is different and follows that described in Section

2 of Smyth (1996). The details of density calculation are subsequently different from Dunn and Smyth (2005).

The derivatives are calculated in a similar manner to the density. The derivatives are for the log-density.

Value

dPoisGam returns a numeric vector containing the (log-)densities.

dPoisGamDeriv returns a matrix with three columns, one for each of the distributional parameters.

pPoisGam returns a numeric vector containing the values of the distribution function.

rPoisGam returns a numeric vector containing the random variables.

Author(s)

Scott D. Foster

References

Smyth, G. K. (1996) Regression analysis of quantity data with exact zeros. *Proceedings of the Second Australia–Japan Workshop on Stochastic Models in Engineering, Technology and Management*. Technology Management Centre, University of Queensland, pp. 572-580.

Dunn P. K. and Smyth G. K. (2005) Series evaluation of Tweedie exponential dispersion model densities. *Statistics and Computing* **15**: 267-280.

Foster, S.D. and Bravington, M.V. (2013) A Poisson-Gamma Model for Analysis of Ecological Non-Negative Continuous Data. *Journal of Environmental and Ecological Statistics* **20**: 533-552

Examples

```
my.seq <- seq( from=0, to=20, length=200)
par( mfrow=c( 1,2))
plot( my.seq, dPoisGam( y=my.seq, lambda=8, mu.Z=1.2, alpha=0.6, LOG=FALSE), type='l',
 xlab="Variable", ylab="Density")
plot( my.seq, pPoisGam( q=my.seq, lambda=8, mu.Z=1.2, alpha=0.6), type='l',
 xlab="Variable", ylab="Distribution")
```

---

dTweedie	*Density, derivatives, distribution function, and random generation for the Tweedie distribution.*

Description

Density, derivatives, distribution function, and random generation for the Tweedie distribution. The distribution is parameterised by the mean, dispersion parameter and the power parameter so that the distribution's variance is given by disperion * mean^power.

Usage

```
dTweedie( y, mu, phi, p, LOG=TRUE)
pTweedie( q, mu, phi, p)
rTweedie( n, mu, phi, p)
```

Arguments

y,q	vector of quantiles
n	number of random draws
mu	scalar or vector (length matches y or equal to n) of means
phi	scalar or vector (length matches y or equal to n) of dispersion parameters
p	scalar or vector (length matches y or equal to n) of power parameters
LOG	indication of return scale. If TRUE (default) then the density is returned on the log scale.

Details

The density calculation is based on the series summation method described in Dunn and Smyth (2005). These functions are really just wrappers to the equivalent functions dPoisGam, pPoisGam and rPoisGam. The functions are equivalent up to parameterisation of the distribution.

The distribution function is calculated by adaptive quadrature (a call to the integrate function). However, the point discontinuity at zero is handled explicitly by assuming the convention that the distribution function evaluated at zero is equal to the density at zero.

Value

```
dTweedie returns a numeric vector containing the (log-)densities.

pTweedie returns a numeric vector containing the values of the distribution function.

rTweedie returns a numeric vector containing the random variables.
```

Author(s)

Scott D. Foster

References

Dunn P. K. and Smyth G. K. (2005) Series evaluation of Tweedie exponential dispersion model densities. *Statistics and Computing* **15**: 267-280.

Foster, S.D. and Bravington, M.V. (2013) A Poisson-Gamma Model for Analysis of Ecological Non-Negative Continuous Data. *Journal of Environmental and Ecological Statistics* **20**: 533-552

Examples

```
my.seq <- seq( from=0, to=20, length=200)
par( mfrow=c( 1,2))
plot( my.seq, dTweedie( y=my.seq, mu=5, phi=2.1, p=1.6, LOG=FALSE),
 type='l', xlab="Variable", ylab="Density")
plot( my.seq, pTweedie( q=my.seq, mu=5, phi=2.1, p=1.6),
 type='l', xlab="Variable", ylab="Distribution")
```

---

pgm	*Fitting models based on the Poisson-Gamma model.*

Description

Fits a model that assumes a Poisson-Gamma distribution. The mean of the unobserved Poisson variable and the mean of the unobserved Gamma random variables are allowed to change with covariates.

Usage

```
pgm( p.form, g.form, data, wts=NULL, alpha=NULL, inits=NULL,
  vcov=TRUE, residuals=TRUE, trace=1)
```

Arguments

p.form	an object of class "formula" (or one that can be coerced to that class). This is a symbolic representation of the model for the unobserved Poisson variable. Note that offset terms (if any) should be included in this part of the model.
g.form	an object of class "formula" (or one that can be coerced to that class). This is a symbolic representation of the model for the unobserved Gamma variables. Note that any outcome and offset terms included in this part of the model are ignored.
wts	relative contribution to the log-likelihood contributions. Must be the same length as the number of observations.
data	a data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the variables in the model.
alpha	a positive scalar whose presence indicates that a profile likelihood should be used for estimating all parameters except alpha (set to supllied value).
inits	numeric vector containing the starting values for the coefficient estimates. The ordering of the vector is:coefficients for p.form, coefficients for g.form, and log(alpha), where alpha is the dispersion parameter for the unobserved gamma variables. Defaults to zero for each coefficient.
vcov	boolean indicating if the variance-covariance matrix of the coefficient estimates should be calculated or not. Default is TRUE indicating that it will be calculated.
residuals	boolean indicating if the quantile residuals should be calculated. Default is TRUE indicating residuals are to be calculated.

trace	non-negative integer value indicating how often during estimation the updated coefficients should be calculated. Default is 1, indicating printing at every iteration. A value of 0 indicates that no printing will occur.

Details

The observed random variables y_i are assumed to arise from the process: $y_i = sum(z_{i1} + z_{i2} + ... + z_{in_i})$ where n_i is a Poisson variable with mean lambda and z_{ij} is a Gamma variable with mean mu.Z and variance mu.Z^2 / alpha.

The means of each of the unobserved random variables are modelled using a log-link and the formula given in p.form and g.form.

The distribution is equivalent, up to parameterisation and model interpretation, to the Tweedie GLM model described in Smyth (1996) and Dunn and Smyth (2005). Their models contain model the overall mean only and not the separate processes. Also the Poisson-Gamma distribution is a subset of the more general Tweedie distribution.

The residuals are quantile residuals, described in general by Dunn and Smyth (1996). Calculation of the quantile residuals ignores any specified weights.

Value

pgm returns an object of class "pgm", a list with the following elements

coef	the estiamted coefficients from the fitting process.
logl	the maximum log likelihood (found at the estimates).
score	the score of the log-likelihood (at the estimates).
vcov	the variance-covariance matrix of the estimates. If vcov is FALSE then this will be NULL.
conv	the convergence message from the call to nlminb.
message	the convergence message from nlminb.
niter	the number of iterations taken to obtain convergence.
evals	the number of times the log-likelihood and its derivative were evaluated.
call	the call used to invoke the function.
fitted	matrix with three columns containing:the total modelled mean, the Poisson mean, and the gamma mean.
residuals	the vector of quantile residuals.

Author(s)

Scott D. Foster

References

Smyth, G. K. (1996) Regression analysis of quantity data with exact zeros. *Proceedings of the Second Australia–Japan Workshop on Stochastic Models in Engineering, Technology and Management.* Technology Management Centre, University of Queensland, pp. 572-580.

Dunn P. K and Smyth G. K (1996) Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics* **5**: 236-244.

Dunn P. K. and Smyth G. K. (2005) Series evaluation of Tweedie exponential dispersion model densities. *Statistics and Computing* **15**: 267-280.

Foster, S.D. and Bravington, M.V. (2013) A Poisson-Gamma Model for Analysis of Ecological Non-Negative Continuous Data. *Journal of Environmental and Ecological Statistics* **20**: 533-552

Examples

```
my.coef <- c(0.6, 1.2, 0, -0.3, 0, -0.5, 0.85)
sim.dat <- simReg( n=250, lambda.tau=my.coef[1:3], mu.Z.tau=my.coef[4:6], alpha=my.coef[7])
fm <- pgm( p.form=y~1+x1+x2, g.form=~1+x1+x2, data=sim.dat)
tmp <- matrix( c( my.coef, fm$coef, sqrt( diag( fm$vcov))), ncol=3)
tmp[nrow( tmp),1] <- log( tmp[nrow( tmp),1])  #putting values on same scale
colnames( tmp) <- c("actual","estiated","SE")
rownames( tmp) <- names( fm$coef)
print( tmp)
```

---

simReg	Simulate Poisson-Gamma data whose component means vary with co-variates.

Description

simReg is used to simulate data from a combined model for the Poisson and Gamma components of a Poisson-Gamma distribution. This formuation allows the distribution to vary with covariates.

Usage

```
simReg( n, lambda.tau, mu.Z.tau, alpha, offset1=NULL, X=NULL)
```

Arguments

n	number of observations
lambda.tau	vector of coefficients for Poisson part of model.
mu.Z.tau	vector of coefficients for Gamma component of model. Its length must be equal to the length of lambda.tau.
alpha	scalar for Gamma dispersion.
offset1	vector of offset values (for Poisson part of the process). If NULL (default) a vector of zeros is created and used.
X	a design matrix with appropriate elements for simulation. If NULL (default) then one will be created.

Details

The observed random variables y_i are assumed to arise from the process: $y\_i=sum(z\_\{i1\}+z\_\{i2\}+...+z\_\{in\_i\})$ where n_i is a Poisson variable with mean lambda and $z\_\{ij\}$ is a Gamma variable with mean mu.Z and variance mu.Z^2 / alpha.

The Poisson mean is given by lambda=exp(X %*% lambda.tau) where X is a suitable design matrix whose first column is full of 1s and whose remain columns are random draws from a standard normal.

The Gamma mean is given by mu.Z=exp(X %*% mu.Z.tau) where X is identical to before.

Value

A data frame containing the random draws, the offset (if not NULL), and the covariates. The data f

Author(s)

Scott D. Foster

References

Foster, S.D. and Bravington, M.V. (2013) A Poisson-Gamma Model for Analysis of Ecological Non-Negative Continuous Data. *Journal of Environmental and Ecological Statistics* **20**: 533-552

Examples

```
sim.dat <- simReg( n=250, lambda.tau=c(0.6, 1.2, 0), mu.Z.tau=c(-0.3, 0, -0.5), alpha=0.85, X=NULL)
print( head( sim.dat))
```

---

tglm   *Fits a Tweedie GLM via maximum likelihood.*

Description

Fits a log-link Tweedie GLM model using maximum likelihood for all parameters, if wanted. The dispersion and power parameters (phi and p) can be set if required.

Usage

```
tglm( mean.form, data, wts=NULL, phi=NULL, p=NULL, inits=NULL, vcov=TRUE,
 residuals=TRUE, trace=1, iter.max=150)
tglm.fit( x, y, wts=NULL, offset=rep( 0, length( y)), inits=rnorm( ncol( x)),
  phi=NULL, p=NULL, vcov=TRUE, residuals=TRUE, trace=1, iter.max=150)
```

Arguments

mean.form	an object of class "formula" (or one that can be coerced to that class). This is a symbolic representation of the model for the mean of the observations. Offset terms can be included in this formula.
data	a data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the variables in the model.
wts	relative contribution to the log-likelihood contributions. Must be the same length as the number of observations.
phi	a positive scalar whose presence indicates that a profile likelihood should be used for estimating all parameters except the dispersion phi (set to the supllied value).
p	a value in the interval [1,2] whose presence indicates that a profile likelihood should be used for estimating all parameters except the power parameter p (set to the supplied value).
inits	(tglm) numeric vector containing the starting values for the coefficient estimates. The ordering of the vector is:coefficients for mean.form, phi and p. Defaults to zero for each the mean coefficients, 1 for phi, and 1.6 for p. (tglm.fit) numeric vector that must be specified (no default). It has length equal to the number of covariates plus 2.
vcov	boolean indicating if the variance-covariance matrix of the coefficient estimates should be calculated or not. Default is TRUE indicating that it will be calculated.
residuals	boolean indicating if the quantile residuals should be calculated. Default is TRUE indicating residuals are to be calculated.
trace	non-negative integer value indicating how often during estimation the updated coefficients should be calculated. Default is 1, indicating printing at every iteration. A value of 0 indicates that no printing will occur.
iter.max	the maximum number of iterations allowed.
x, y, offset	x is a design matrix, y is the vector of outcomes and offset are the offset values. None of these values are checked. The design matrix has dimension pXn where p is the number of covariates and n is the number of observations.

Details

The means of each of the unobserved random variables are modelled using a log-link and the formula given in mean.form. This function implements the model described in Smyth (1996) and Dunn and Smyth (2005) but does so by maximising the complete likelihood (ala Dunn and Smyth, 2005). The model specification is restricted to the log link.

The residuals are quantile residuals, described in general by Dunn and Smyth (1996).

tglm.fit is the workhorse function: it is not normally called directly but can be more efficient where the response vector and design matrix have already been calculated. Completely analgous to the function glm.fit.

Value

tglm returns an object of class "tglm", a list with the following elements

coef	the estiamted coefficients from the fitting process.
logl	the maximum log likelihood (found at the estimates).
score	the score of the log-likelihood (at the estimates).
vcov	the variance-covariance matrix of the estimates. If vcov is FALSE then this will be NULL.
conv	the convergence message from the call to nlminb.
message	the convergence message from nlminb.
niter	the number of iterations taken to obtain convergence.
evals	the number of times the log-likelihood and its derivative were evaluated.
call	the call used to invoke the function.
fitted	matrix with one column containing the modelled mean.
residuals	the vector of quantile residuals.

Author(s)

Scott D. Foster

References

Smyth, G. K. (1996) Regression analysis of quantity data with exact zeros. *Proceedings of the Second Australia–Japan Workshop on Stochastic Models in Engineering, Technology and Management*. Technology Management Centre, University of Queensland, pp. 572-580.

Dunn P. K and Smyth G. K (1996) Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics* **5**: 236-244.

Dunn P. K. and Smyth G. K. (2005) Series evaluation of Tweedie exponential dispersion model densities. *Statistics and Computing* **15**: 267-280.

Foster, S.D. and Bravington, M.V. (2013) A Poisson-Gamma Model for Analysis of Ecological Non-Negative Continuous Data. *Journal of Environmental and Ecological Statistics* **20**: 533-552

Examples

```
print( "Data generated using Poisson-sum-of-gammas model and fitted using a Tweedie GLM.")
print( "A great fit is not expected -- especially for the first covariate")
my.coef <- c(0.6, 1.2, 0, -0.3, 0, -0.5, 0.85)
sim.dat <- simReg( n=250, lambda.tau=my.coef[1:3], mu.Z.tau=my.coef[4:6], alpha=my.coef[7])
fm <- tglm( mean.form=y~1+x1+x2, data=sim.dat)
tmp <- matrix( c( my.coef[1:3] + my.coef[4:6], NA, (2+0.85)/(1+0.85), fm$coef), ncol=2)
colnames( tmp) <- c("actual","estimated")
rownames( tmp) <- c( names( fm$coef)[1:3], "phi", "p")
print( tmp)
```

`TigerFlathead`	*Data for Tiger Flathead in South East Survey*

Description

TigerFlathead is a a data set arising from a south-east Australia fish survey, see Bax and Williams (1999).

Usage

```
TigerFlathead
```

Format

Lon The Longitude of the trawl.

Lat The Latitude of the trawl.

GA_BATHY Bathymetry (depth) of the trawl, taken from longitude/latitude of trawls.

STRING The distance along the 100m depth contour. The origin is arbitrary, but is northwards of all trawls.

AREA_HA The swept area of the shot, giving a measure of effort.

Tiger Flathead_no The number of tiger flathead caught.

Tiger Flathead_wt The total biomass of tiger flathead caught.

References

Bax N, Williams A (eds) (1999) Habitat and fisheries production in the South East Fishery. *Final report to FRDC Project 94/040*, CSIRO Marine Research, Hobart, Tasmania, Australia

Foster, S.D. and Bravington, M.V. (2013) A Poisson-Gamma Model for Analysis of Ecological Non-Negative Continuous Data. *Journal of Environmental and Ecological Statistics* **20**: 533-552

Index