User’s Guide: Delta-GLM function for the R language/environment

(Version 1.7.2, revised 07-06-2006)

This function for the R language/environment calculates a relative index of abundance based on two generalized linear models (GLMs). The first GLM estimates the probability of a positive observation using a binomial likelihood and a logit link function. The second GLM estimates the mean response for positive observations, assuming that the error distribution is either gamma, lognormal, or inverse gaussian. The final index is the product of the back-transformed year effects from the two GLMs. See Stefánsson (1996) for an example that uses gamma errors. This R code was inspired by Erik Williams’ “glmdelta.get.”

An optional jackknife routine estimates CVs for the index. To stabilize the jackknife routine, the data are filtered prior to the analysis. If any factor level has less than 2 positive observations, all records within that factor level are deleted from the data set prior to calculating the index. The user can change the threshold number of positive observations (see below).

General information about the R language/environment is found at:

http://www.r-project.org/

Download and install R:

1. Download the appropriate (platform-specific) installation file and save it to your local hard drive:

   http://cran.stat.ucla.edu/
   (see the R website for other download sites)

2. Follow installation instructions.

Load the function into the R workspace:

3. Start R

4. From the R prompt, use the function `dget()` to load the file ‘delta_glm_1-7-2.get’ into the workspace (where you can give it any name you like). The example below loads the file ‘delta_glm_1-7-2.get’ from the directory c:/temp, and gives it the name ‘dglm’:

   ```
   dglm <- dget(“c:/temp/delta_glm_1-7-2.get”)
   ```

Load your data set into the R workspace:

5. This example uses data saved as tab-delimited, ASCII text, with the first row containing header information. See the R help documentation to import other data types. Type `?read.table` at the R prompt for more information.

6. Data must be in the following format:

   Column 1: response variable (must be a continuous variable)
   Column 2: [factor that will get jackknifed error estimates, e.g. year]
   Column 3: [2nd explanatory variable] *(variables 2-p can be factors or continuous)*
   : :
   : :
   Column *p*: [*p*th explanatory variable]

   Example format for tab-delimited data with header row:

   ```
   Density        year      month       station      depth
   0              2000      1           90000        36.2
   0              1952      3           91111        42.6
   2.25           1975      12          92222        38.3
   ```

7. Use the function `read.table()` to load the data set into the R workspace. You can give the data any name you’d like:

   ```
   mydata <- read.table(“c:/temp/mydata.txt”, header = T)
   ```

   Note that it is necessary to specify that the first row contains headers (variable names). If the data are separated by another character (e.g. comma-delimited), add the argument `sep = “,”` within the parentheses.

8. Check that the data were read correctly by typing `mydata` at the prompt. To view the first 10 rows of your data, type `mydata[1:10,]`. The `dim()` function (e.g. `dim(mydata)`) will output the dimensions of the data matrix.

Run the delta-GLM function:

9. Pass your data to the delta-GLM function (you can give the output any name you’d like). By default, the function assumes that all explanatory variables are factors. To include continuous variables, use the ‘types’ argument (see below) to specify whether each variable is a factor (type ‘F’) or continuous (type ‘C’). The first variable in the data set (the response) must be of type ‘C’.

   The command below passes the example data (`mydata`) to the delta-GLM function. The ‘types’ argument identifies the response and the 4th explanatory variable as continuous variables (`types=c(‘C’, ‘F’, ‘F’, ‘F’, ‘C’)`), the ‘dist’ argument assumes a gamma distribution for positive observations (`dist=“gamma”`), and `write=T` writes the summary output to the working directory as a text file.

   ```
   myoutput <- dglm(mydata, dist=“gamma”, write=T,
                         types=c(‘C’,‘F’,‘F’,‘F’,‘C’))
   ```

   To change the working directory in R, select “Change dir…” from the “File” menu.

List of primary arguments and their defaults for the delta-GLM function:

Argument	Description	Default
data.set*	data in the form of a 'data.frame' object	[no default]
dist	specifies the error structure for positive observations (other options: "lognormal", "inverse.gaussian")	"gamma"
types	variable types (continuous, C, or factor, F)	all factors (F)
J	Estimate errors using jackknife? (TRUE or FALSE)	FALSE
minpos	minimum # of positive observations per factor level	2
write	writes summary output to working directory as text file	FALSE

* required

Description of output:

Once the function starts running, use the scroll bar or the "Page Down" key to view onscreen output.

Screen Output (this is the same as the text file generated by the 'write' argument):

1. Total number of records removed by the data filter.
2. Total number of positive records removed by filter.
3. Counts off each iteration of the jackknife routine (if argument J=T)
4. Contents of either 'deltagamma.summary', 'deltalognormal.summary', or 'deltainvgau.summary' (see below)

The function saves two 'list' objects in the R workspace, which in turn contain other objects. All screen output is saved as either 'deltagamma.summary', 'deltalognormal.summary,' or 'deltainvgau.summary', depending on the choice of error distribution. Detailed information about model fits can be extracted from the glm objects contained within the lists named 'deltagamma.results', 'deltalognormal.results', or 'deltainvgau.results'. See R-help for more information.

LIST *(object name depends on the choice of distribution)*	**CONTAINS THE FOLLOWING OBJECTS**	**DESCRIPTION OF OBJECT**
deltagamma.summary (or) deltalognormal.summary (or) deltainvgau.summary	1. error.distribution 2. binomial.formula 3. positive.formula 4. deltaGLM.index 5. effects 6. data.filter 7. levels.deleted.by.filter 8. AIC	1. distribution of positive data 2. model formula for binomial GLM 3. model formula for "positive" GLM 4. final index and results from jackknife routine (if activated) 5. effects for explanatory variables 2-p 6. current setting of 'minpos' argument 7. levels deleted by the filter 8. AIC scores for both models, and MLE of dispersion parameter
deltagamma.results (or) deltalognormal.results (or) deltainvgau.results	1. positive.glm 2. binomial.glm 3. positive.index 4. binomial.index	1. 'glm' object (GLM for positive data) 2. 'glm' object (GLM for binary data) 3. back-transformed mean responses for positive data 4. probabilities of positive observations

The not-so-fine print:

1. The linear predictor in each GLM is automatically defined as a "main effects only" model (no interactions). All variables present in the data set are automatically included in the model formulas.

2. Both GLMs use the same set of explanatory variables, as defined by the input data set. The individual glm objects can be extracted from the object "`delta_____.results`". See R help for more information regarding the use of the functions `update()` and `anova()`, which make it easy to evaluate alternative models (e.g. interactions, etc.).

3. The jackknife routine only produces error estimates for the explanatory variable in the second column of the data set. This variable must be of type 'factor' ('F'). If you want jackknifed error estimates for other factors, simply make the variable of interest the second column of the data set.

4. Some other (less used) arguments in the function are:
   a. `lnorm.init`: (default = FALSE)
   Sometimes the 'glm' function in R has trouble with convergence of gamma and inverse gaussian models when using the default starting values. Setting `lnorm.init = TRUE` borrows the regression coefficients from a lognormal model to initialize either the gamma or inverse gaussian models, which might help.
   b. `ig.scale:` (default = 1)
   Initial value for the scale parameter in the inverse gaussian model. If the program doesn't converge, try another starting value.
   c. `jack.noise:` (default = 1)
   Each iteration of the jackknife routine is initialized with regression coefficients from the GLMs fit to the complete data set. This argument adds some uniform 'noise' to these values so the program doesn't converge immediately (underestimate the variability). See the R help file for the `jitter()` function for more information.
   d. `weight.vector:` (default = vector of 1's (equal weights))
   This argument allows the user to provide weights for each observation in the GLM for the positive data. The user must define the vector of weights separately, prior to running the delta-GLM function. By default, all observations are weighted equally.

References

Stefánsson, G. 1996. Analysis of groundfish survey abundance data: combining the GLM and delta approaches. ICES Journal of Marine Science, 53: 577-588.

Searle, S. R. et al. 1980. Population marginal means in the linear model: an alternative to least squares means. The American Statistician 34(4): 216-221.

McCullagh, P., & Nelder J. A. 1989. Generalized Linear Models, 2nd ed., London: Chapman & Hall.