Package 'hglm.data'

July 23, 2014

Type Package

Title Data for The hglm Package

Version 1.0-0

Date 2014-07-23

Author Xia Shen, Moudud Alam, Lars Ronnegard

Maintainer Xia Shen <xia.shen@slu.se>

Description This data-
only package was created for distributing data used in the examples of the hglm package.

BugReports https://r-forge.r-project.org/tracker/?group_id=558

License GPL (>= 2)

LazyLoad yes

Depends R (>= 2.10), utils, Matrix, MASS

NeedsCompilation no

Repository CRAN

Date/Publication 2014-07-23 15:24:34

R topics documented:

`hglm.data-package` *Data for The hglm Package*

Description

This data-only package was created for distributing data used in the examples of the hglm package.

Details

Package:	hglm.data
Type:	Package
Version:	1.0-0
Date:	2014-07-23
Discussion:	https://r-forge.r-project.org/forum/?group_id=558
BugReports:	https://r-forge.r-project.org/tracker/?group_id=558
License:	GPL (>= 2)
LazyLoad:	yes
Depends:	R (>= 2.10)

Author(s)

Xia Shen

Maintainer: Xia Shen <xia.shen@slu.se>

References

Lars Ronnegard, Xia Shen and Moudud Alam (2010). **hglm: A Package for Fitting Hierarchical Generalized Linear Models**. *The R Journal*, **2**(2), 20-28.

Youngjo Lee, John A Nelder and Yudi Pawitan (2006) *Generalized Linear Models with Random Effect: a unified analysis via h-likelihood.* Chapman and Hall/CRC.

Xia Shen, Moudud Alam, Freddy Fikse and Lars Ronnegard (2013). **A novel generalized ridge regression method for quantitative genetics**. *Genetics*.

Moudud Alam, Lars Ronnegard, Xia Shen (2014). **Spatial modeling in hglm**. *Submitted*.

cancer *Scottish lip cancer dataset from Clayton and Kaldor (1987)*

Description

The Scottish lip cancer dataset.

Format

The 'cancer' dataset contains 4 objects as follows.

O Observed frequency.

E Offset.

Paff Fixed effects.

nbr Spatial correlation matrix D for CAR model.

Source

Clayton D, Kaldor J 1987. Empirical Bayes Estimation of Age-standardized Relative Risk for use in Disease Mapping. *Biometrics* 43, 671–681

References

Clayton D, Kaldor J 1987. Empirical Bayes Estimation of Age-standardized Relative Risk for use in Disease Mapping. *Biometrics* 43, 671–681

ohio *Ohio elementary school dataset*

Description

Data set on 1,965 Ohio elementary school buildings for 2001-2002.

Format

The 'ohio' dataset contains 6 objects as follows.

ohioSchools Original data `ohioschool.dat` from http://www.spatial-econometrics.com/ (J. LeSage and R. Pace 2009). The data set contains information on, for instance, school building ID, Zip code of the location of the school, proportion of passing on five subjects, number of teacher, number of student, etc. The variables are:
col 1: zip code
col 2: lattitude (zip centroid)
col 3: longitude (zip centroid)
col 4: buidling irn

col 5: district irn
col 6: # of teachers (FTE 2001-02)
col 7: teacher attendance rate
col 8: avg years of teaching experience
col 9: avg teacher salary
col 10: Per Pupil Spending on Instruction
col 11: Per Pupil Spending on Building Operations
col 12: Per Pupil Spending on Administration
col 13: Per Pupil Spending on Pupil Support
col 14: Per Pupil Spending on Staff Support
col 15: Total Expenditures Per Pupil
col 16: Per Pupil Spending on Instruction % of Total Spending Per Pupil
col 17: Per Pupil Spending on Building Operations % of Total Spending Per Pupil
col 18: Per Pupil Spending on Administration % of Total Spending Per Pupil
col 19: Per Pupil Spending on Pupil Support % of Total Spending Per Pupil
col 20: Per Pupil Spending on Staff Support % of Total Spending Per Pupil
col 21: irn number
col 22: avg of all 4th grade proficiency scores
col 23: median of 4th grade prof scores
col 24: building enrollment
col 25: short-term students < 6 months
col 26: 4th Grade (or 9th grade) Citizenship % Passed 2001-2002
col 27: 4th Grade (or 9th grade) math % Passed 2001-2002
col 28: 4th Grade (or 9th grade) reading % Passed 2001-2002
col 29: 4th Grade (or 9th grade) writing % Passed 2001-2002
col 30: 4th Grade (or 9th grade) science % Passed 2001-2002
col 31: pincome per capita income in the zip code area
col 32: nonwhite percent of population that is non-white
col 33: poverty percent of population in poverty
col 34: samehouse % percent of population living in same house 5 years ago
col 35: public % of population attending public schools
col 36: highschool graduates, educ attainment for 25 years plus
col 37: associate degrees, educ attainment for 25 years plus
col 38: college, educ attainment for 25 years plus
col 39: graduate, educ attainment for 25 years plus
col 40: professional, educ attainment for 25 years plus

ohioGrades The derived dataset for analyzing the percentage passed based on Zip codes. The variables are:
y: the percentage passed (4th or 9th grade) in each school
TchExp: average Teacher's experience
Subjects: for five study subjects of Citizenship, Maths, Reading, Writing and Science
Stu.Tch: student by teacher ratio
School: school index
Zip: Zip code

ohioMedian The derived dataset for analyzing the median of 4th grade scores based on school districts. The variables are:
MedianScore: the median of 4th grade prof scores
district: school districts

ohioShape A `SpatialPolygonsDataFrame` object (see package **sp**) containing the map information of ohio school districts.

ohioZipDistMat The spatial distance matrix based on Zip codes. The codes generated this matrix are:

```
Zsp <- model.matrix(~ factor(Zip) - 1, data = ohioGrades)
uzipC <- matrix(0, nrow = ncol(Zsp), ncol = 2)
Zip <- as.numeric(substr(colnames(Zsp), start = 12, stop = 16))
for (i in 1: ncol(Zsp)) {
Cord <- as.matrix(ohioSchools[(ohioSchools$V1 == Zip[i]), 2:3])
uzipC[i,] <- Cord[1,]
}
Dst <- as.matrix(dist(uzipC))
for(i in 1:nrow(Dst)) {
x <- Dst[i,]
x <- ifelse(x == 0, 0, 1/x)
Dst[i,] <- ifelse(x > 4, 4, x)
}
ohioZipDistMat <- Dst/4
```

ohioDistrictDistMat The spatial distance matrix based on school districts. The codes generated this matrix are:

```
ccNb <- poly2nb(ccShape)
W <- matrix(0, 616, 616)
for (i in 1:nrow(W)) {
tmp <- as.numeric(ccNb[[i]])
for (k in tmp) W[i,k] <- 1
}
W[353,] <- W[,353] <- 0
districtShape <- as.numeric(substr(as.character(ohioShape@data$UNSDIDFP), 3, 7))
dimnames(W) <- list(districtShape, districtShape)
districtSchool <- floor(ohioSchools[,5]/10)
districtSchool <- factor(districtSchool[districtSchool %in% districtShape])
levelsShape <- levels(factor(districtShape))
levelsSchool <- levels(districtSchool)
levels(districtSchool) <- c(levelsSchool, levelsShape[!(levelsShape %in% levelsSchool)])
ohioDistrictDistMat <- W[levels(districtSchool),levels(districtSchool)]
```

Source

J. LeSage and R. Pace (2009). *Introduction to Spatial Econometrics*. Chapman \& Hall/CRC, Boca Raton.

References

J. LeSage and R. Pace (2009). *Introduction to Spatial Econometrics*. Chapman \& Hall/CRC, Boca Raton.

M. Alam, L. Ronnegard, X. Shen (2014). **Fitting spatial models in hglm**. *Submitted*.

pump *Pump reliability data set from Gaver and O'Muircheartaigh (1987)*

Description

The 'pump' data set presents the failures of pumps in several systems of the water reactor neuclear plant Farley 1.

Format

The 'pump' data set contains 4 columns and 10 rows. A short description of the data columns are given below.

System The system number.

S Number of pumps failures.

t Time (in thousand hours) of operation.

Gr Pump groups; two levels: 1 = operated continuously, 0 = operated intermittently.

Source

Gaver, D P. and O'Muircheartaigh, I. G. 1987. Robust Empirical Bayes Analyses of Event Rates, *Technometrics* 29(1),1–15

References

Lee, Y. and Nelder, J. A. 1996. Hierarchical generalized linear models, *Journal of the Royal Statistical Association (B, Theory and Methods)* 58(4), 619–678.

QTLMAS *Simulated Data Set for the QTLMAS 2009 Workshop*

Description

The data was simulated for the QTLMAS 2009 workshop in Wageningen, The Netherlands. The data was made available at http://www.qtlmas2009.wur.nl/UK/Dataset/ and consists of markers, trait values and pedigree information. The original data set consisted of several traits and markers from several chromosomes, whereas the current data set included in this package consists of one trait ("P265"), pedigree information and data from 90 markers on chromosome number 1. There are 2025 individuals in the pedigree where 1000 individuals have trait values.

Format

A matrix containing 1000 rows and 2116 columns. The first column contains the trait values. Columns 2 to 2026 contains matrix Z, i.e. the pedigree information (as the Colesky factorization of the additive relationship matrix). Columns 2027 to 2116 contains matrix Z.marker, i.e. the marker information for the 90 markers on chromosome 1.

Source

QTLMAS 2009 Workshop http://www.qtlmas2009.wur.nl/UK/Dataset/

References

Coster, A., Bastiaansen J., Calus M., Maliepaard C. and Bink M. 2009. QTLMAS 2009: Simulated dataset. (submitted)

`salamander` *Salamander mating data set from McCullagh and Nelder (1989)*

Description

'salamader' data set presents the outcome of an experiment which was conduceted at the University of Chicago in 1986 to study the extent at which mountain dusky salamanders from different populations would interbred. More detailed description of the data is given in its original source, McCullagh and Nelder (1989).

Format

'salamander' data set contains 6 columns and 360 rows. A brief description of the data columns is given below.

Season The seasons, Spring and Summer of 1986, when the experiment was carried out.

Experiment Experiment number; 1,2,3.

TypeM Type of the male salamander; Rough Butt=R and White Side=W

TypeF Type of the female salamander; Rough Butt=R and White Side=W

Cross Cross between female and male type e.g. Cross=WR mean a White Side female was crossed with a Rough Butt male.

Male Identification number of the male salamander.

Female Identification number of the female salamander.

Mate Whether a mating was observed, Yes=1 and No=0.

Source

McCullagh P. and Nelder, J. A. 1989. *Generalized Linear Models*, Section 14.5, Chapman and Hall/CRC.

seeds *Seeds genrmination data set from Crowder (1978)*

Description

The data set was initially presented in Corder (1978) to demonstrate the problem of over dispersion with binomial response and its solution via beta-binomial ANOVA. Latter, the data set is used by may others including Breslow and Clayton (1993) and Lee and Nelder (1996) to demonstrate the usefulness of the Generalized Linear Mixed (and hierarchical) model. The seeds data set was originally obtained from a 2 by 2 factorial layout. The experiment was conducted on two types of seeds, *O. aegyptiaca 75* and *O. aegyptica 73*, and two root extracts, bean and cucumber with an equal dilution, 1/125. Experimental units (plates) were prepared with the specific roots extracts and a batch of certain seeds was brushed into the plates. The outcome is the count of germinated seed out of the total number of seeds applied in each plate.

Format

The seeds data set contans 5 columns and 21 rows. A short description of the data columns are given below.

plate Plate number.

seed Seed type; 2 levels: O75 (*O. aegyptiaca 75*) and O73 (*O. aegyptica 73*).

extract Type of roots extract; 2 levels: Bean and Cucumber.

r Response; number of seeds germinated in each plate.

n Total number of seeds applied in each plate.

Source

Crowder, M. J. 1978. Beta-binomial Anova for proportions, *Journal of the Royal Statistical Society (C, Applied Statistics)* 27(1), 34–37.

References

Breslow, N. E. and Clayton, D. G. 1993. Approximate inference in generalized linear mixed models, *Journal of the Amrecian Statistical Association* 88, 9–25.
Lee, Y. and Nelder, J. A. 1996. Hierarchical generalized linear models, *Journal of the Royal Statistical Association (B, Theory and Methods)* 58(4), 619–678.

`semiconductor` *Semiconductor data set from GenStat.*

Description

The semiconductor data set is obtained from a 2^(6-2) factorial design conducted in a semiconductor plant. The design variables, Lamination (3 factors; Temperature, Time and Pressure) and Firing (3 factors; Temperature, Cycle Time and Dew Point), are each taken at two levels. The goal of the original data analysis was to model the curvature or camber (taken in 1e-4 in./in.) as a function of the desing variables. The data set is taken from GenStat 11.1. It is also used in Lee et al. (2006) where Mayers et al. (2002) is reffered to as the the original source of the data.

Format

This data set contains 64 rows and the following columns

Device Subtrate device

x1 Lamination Temperature; two levels +1 and -1.

x2 Lamination Time; two levels: +1 and -1.

x3 Lamination Presure; two levels: +1 and -1.

x4 Firing Temperature; two levels: +1 and -1.

x5 Firing Cycle Time; two levels: +1 and -1.

x6 Firing Dew Point: two levels: +1 and -1.

y Camber measure; in 1e-4 in./in.

Source

GenStat(R) Release 11.1. VSN International Limited.

References

Lee, Y. and Nelder J. A., and Pawitan, Y. 2006. *Generalized Linear Models with Random Effectes*, Chapman and Hall/CRC.
Mayers, P. H., Montgomery, D. C. and Vining G. G. 2002. *Generalized Linear Models with Application in Engineering and Science*, John Wiley and Sons.

Index