Package 'GJRM.data'

July 21, 2025

Version 0.1-1

Author Giampiero Marra [aut, cre]

Maintainer Giampiero Marra <giampiero.marra@ucl.ac.uk>

Title Data Sets for Copula Additive Distributional Regression Using R

Description Data sets used in the book Marra and Radice (2025, ISBN:9781032973111) ``Copula Additive Distributional Regression Using R'', for illustrating the fitting of various joint (and univariate) regression models, with several types of covariate effects, in the presence of equations' errors association.

Depends R (>= 3.6.0)

Suggests GJRM

License GPL (>= 2)

NeedsCompilation no

Repository CRAN

Date/Publication 2025-06-29 10:30:02 UTC

Contents

areds *AREDS: Age-related Eye Disease Study*

Description

Real dataset of bivariate interval and right censored data with 628 subjects and three covariates. The dataset is a reshaped version of the AREDS data from the `CopulaCenR` package. The dataset was selected from the Age-related Eye Disease Study (AREDS Group, 1999). The two events are the progression times (in years) to late-AMD in the left and right eyes.

Usage

```
data(areds)
```

Format

`war` is a 628 row data frame with the following columns:

t11, t12 left and right bounds of the intervals for the left eye. If `t12 = NA` then the observation is right-censored.

t21, t22 left and right bounds of the intervals for the right eye. If `t22 = NA` then the observation is right-censored.

SevScore1, SevScore2 baseline AMD severity scores for left and right eyes, respectively. Possible values are: 4, 5, 6, 7, 8.

age age at baseline.

rs2284665 a genetic variant covariate highly associated with late-AMD progression. Possible values are: 0, 1, 2.

cens1, cens2 type of censoring for left and right eyes.

cens joint censoring indicator for left and right eyes.

Source

Data are from:

AREDS Group (1999), The Age-Related Eye Disease Study (AREDS): design implications. AREDS report no. 1. *Control Clinical Trials*, 20, 573-600.

bpc *Blood pressure data in children*

Description

Blood pressure data in 11 year old children. The dataset is a subsample from Solomon-Moore et al. (2020).

Usage

```
data(bpc)
```

Format

bpc is a 1052 row data frame with the following columns:

sbp Systolic Blood Pressure (mmHg).

dbp Diastolic Blood Pressure (mmHg).

gender 1 = Male, 2 = Female.

bmi Body Mass Index.

mvpa Average minutes of moderate to vigorous physical activity per day.

sed Average sedentary minutes per day.

Source

Data are from Solomon-Moore E, Salway R, Emm-Collison L, Thompson JL, Sebire SJ, Lawlor DA, Jago R (PI), 2020.

cd4 *ACDIS data*

Description

Fictitious data designed to closely replicate the characteristics and patterns observed in the Africa Centre Demographic Information System (ACDIS).

Usage

```
data(cd4)
```

Format

cd4 is a 2645 row data frame with the following columns:

cd4.count CD4 count measurements.

hiv Binary variable indicating whether an individual is HIV positive (hiv = 1) or not (hiv = 0).

age Age in years.

location Three levels: PER, RUR, URB.

marital Six levels: Married, Polygamous, Divorced/Separated/Widowed, Engaged, Never Married, Under Legal Age.

water If present or not.

education Four levels: None, Primary, Junior Secondary, Upper Secondary.

distance1 Km to nearest primary school.

distance2 Km to nearest secondary school.

Source

The data have been produced as described in:

Tanser F. at al., (2007), Cohort Profile: Africa Centre Demographic Information System (ACDIS) and population-based HIV survey. *International Journal of Epidemiology*, 37(5), 956-962.

---

dataDE *Simulated data with two endogenous variables*

Description

Simulated data with two endogenous variables and binary outcome.

Usage

```
data(dataDE)
```

Format

dataDE is a 2000 row data frame with the following columns:

y1 First endogenous variable.

y2 Second endogenous variable.

y3 Binary outcome.

x1, x2 Covariates.

x3 Covariate influencing only y1.

x4 Covariate influencing only y2.

Examples

```
# Data have been simulated as shown below

n  <- 2000
x1 <- round(runif(n))
x2 <- runif(n)
x3 <- runif(n)
x4 <- rnorm(n)
u  <- rnorm(n)

y1 <- ifelse(-1.55 +              x1 - x2 + x3 + u + rnorm(n) > 0, 1, 0)
y2 <- ifelse(-0.25 -          0.5*x1 + x2 + x4 + u + rnorm(n) > 0, 1, 0)
y3 <- ifelse(-0.75 + 0.5*y1 - y2 + x1 + x2      + u + rnorm(n) > 0, 1, 0)

dataDE <- data.frame(y1, y2, y3, x1, x2, x3, x4)
```

`dataDSS` *Simulated data with double sample selection*

Description

Simulated data with double sample selection and binary outcome.

Usage

```
data(dataDSS)
```

Format

`dataDSS` is a 10000 row data frame with the following columns:

y1 First selection.

y2 Second selection.

y3 Binary outcome.

x1, x2 Covariates.

x3 Covariate influencing only `y1`.

x4 Covariate influencing only `y2`.

y3.o Original outcome, without missingness.

Examples

```
# Data have been simulated as shown below

n  <- 10000
x1 <- round(runif(n))
x2 <- runif(n)
x3 <- runif(n)
x4 <- rnorm(n)
u  <- rnorm(n)

y1 <- ifelse(-1.55 +     x1 - x2 + x3 + u + rnorm(n) > 0, 1, 0)
y2 <- ifelse(-0.25 - 0.5*x1 + x2 + x4 + u + rnorm(n) > 0, 1, 0)
y3 <- y3.o <- ifelse( -0.75 + x1 + x2 + u + rnorm(n) > 0, 1, 0)

y2 <- y2*y1
y3 <- y3*y2
y3 <- ifelse(y2 == 0, NA, y3)

dataDSS <- data.frame(y1, y2, y3, x1, x2, x3, x4, y3.o)
```

---

happy *World Happiness Report Data*

Description

Data from the 2019 World Happiness Report, an annual publication of the United Nations Sustainable Development Solutions Network.

Usage

```
data(happy)
```

Format

happy is a 155 row data frame with the following columns:

country Country.

gdp Gross domestic product per capita.

support Indicator of social support (or having someone to count on in times of trouble) calculated at national level.

hle Indicator of healthy life expectancies at birth.

freedom Freedom to make life choices is the national average of responses to the question: Are you satisfied or dissatisfied with your freedom to choose what you do with your life?

generosity Generosity is the residual of regressing national average of response to the question: Have you donated money to a charity in the past month? on GDP per capita.

corruption Corruption Perception: The measure is the national average of the survey responses to two questions in the: Is corruption widespread throughout the government or not? and Is corruption widespread within businesses or not? The overall perception is just the average of the two 0-or-1 responses.

score Subjective well-being. 1 low, 2 medium low, 3 medium, 4 high.

`hie` *Hiring Incentive Experiment - HIE*

Description

Full description available at the web link below.

Usage

```
data(hie)
```

Format

`hie` is a 7734 row data frame with the following columns:

agree Equal to 1 if the individual is in the HIE group and agreed to participate, and 0 if the individual is assigned to the control group or refuses to participate.

bonus Random allocation variable equal to 1 if the individual/employer was assigned to the hiring incentive experiment group and 0 to the control group. This is the IV.

benefit Weekly benefit amount + dependents' allowance.

unemp.dur Weeks of benefits.

status Equal to 1 if unemp.dur < 26 and 0 otherwise.

age Age of claimant.

gender 1 = male and 0 = female.

ethnicity 1 = black and 0 otherwise.

prearn Claimant's pre-claim earnings.

Source

https://www.upjohn.org/data-tools/employment-research-data-center/illinois-unemployment-incentive-experiments

`hiv` *HIV Zambian data*

Description

HIV Zambian data by region, together with polygons describing the regions' shapes.

Usage

```
data(hiv)
data(hiv.polys)
```

Format

`hiv` is a 6416 row data frame with the following columns:

consent binary variable indicating consent to test for HIV.

status binary variable indicating whether an individual is HIV positive (status = 1) or not (status = 0).

age age in years.

education years of education.

wealth wealth index.

region code identifying region, and matching `names(hiv.polys)`. It can take nine possible values: 1 central, 2 copperbelt, 3 eastern, 4 luapula, 5 lusaka, 6 northwestern, 7 northern, 8 southern, 9 western.

marital never married, currently married, formerly married.

std had a sexually transmitted disease.

highhiv had high risk sex.

partner number of partners.

condom used condom during last intercourse.

aidscare equal to 1 if would care for an HIV-infected relative.

knowsdiedofaids equal to 1 if know someone who died of HIV.

evertestedHIV equal to 1 if previously tested for HIV.

smoke smoker or not.

ethnicity bemba, lunda (luapula), lala, ushi, lamba, tonga, luvale, lunda (northwestern), mbunda, kaonde, lozi, chewa, nsenga, ngoni, mambwe, namwanga, tumbuka, other.

language English, Bemba, Lozi, Nyanja, Tonga, other.

interviewerID interviewer identifier.

agehadsex age the individual had sex.

religion four categories.

sw survey weights.

`hiv.polys` contains the polygons defining the areas in the format described below.

Details

The data frame `hiv` relates to the regions whose boundaries are coded in `hiv.polys`. `hiv.polys[[i]]` is a 2 column matrix, containing the vertices of the polygons defining the boundary of the ith region. `names(hiv.polys)` matches `hiv$region` (order unimportant).

Source

The data have been produced as described in:

McGovern M.E., Barnighausen T., Marra G. and Radice R. (2015), On the Assumption of Joint Normality in Selection Models: A Copula Approach Applied to Estimating HIV Prevalence. *Epidemiology*, 26(2), 229-237.

References

Marra G., Radice R., Barnighausen T., Wood S.N. and McGovern M.E. (2017), A Simultaneous Equation Approach to Estimating HIV Prevalence with Non-Ignorable Missing Responses. *Journal of the American Statistical Association*, 112(518), 484-496.

`hospital` *U.S. hospital data from the state of Virginia*

Description

Data on 978 randomly selected patients admitted between January and September 2014 to an over-500-bed medical center (Lewis Gale Medical Center) in the state of Virginia.

Usage

```
data(hospital)
```

Format

`hospital` is a 978 row data frame with the following columns:

los Patient length of hospital stay (in days).

died In-hospital mortality. 1 dead, 0 alive.

age Age of the patient.

gender Either male or female

bmi Body mass index.

severity Subjective assessment of severity level of patient. Value between 1 and 4, with 1 representing the lowest severity level.

risk Subjective assessment of risk of dying. Value between 1 and 4, with 1 representing the lowest level.

sp02 Oxygen saturation level.

sbp Systolic blood pressure.

dbp Diastolic blood pressure.

pulse Pulse rate.

respiratory Respiratory rate.

avpu AVPU score (A: alert, V: responding to voice, P: responding to painful stimuli, U: unresponsive).

temp Temperature.

Source

Azadeh-Fard N, Ghaffarzadegan N, Camelio JA (2016), Can a Patient's In-Hospital Length of Stay and Mortality Be Explained by Early-Risk Assessments?, PLoS ONE 11(9): e0162976.

`infants` *Infant statistic data from North Carolina*

Description

Individual-level infant mortality data on 20000 randomly selected births of female babies in the U.S. state of North Carolina, in 2008, together with polygons describing the county shapes.

Usage

```
data(infants)
data(NC.polys)
```

Format

`infants` is a 20000 row data frame with the following columns:

county Number code identifying North Carolina county in which birth occurred, and matching `names(NC.polys)`. It can take 100 possible values.

age Age of mother.

wksgest Completed weeks of gestation.

marital Equal to 1 if married, and 0 otherwise.

grams Infant's birth weight.

lbw Equal to 1 if infant's birth weight < 2500 grams, and 0 otherwise.

ethnicity Four categories of ethnicity: White, Hispanic, Black, Other.

educ Education of mother: Primary, Secondary, Tertiary.

smoke Equal to 1 if smoker, and 0 otherwise.

firstbirth Equal to 1 if it was the mother's first birth, and 0 otherwise.

ptb Equal to 1 if completed weeks of gestation < 37.

`NC.polys` contains the polygons defining the areas in the format described below.

Details

The data frame `infants` relates to the counties whose boundaries are coded in `NC.polys`. `NC.polys[[i]]` is a 2 column matrix, containing the vertices of the polygons defining the boundary of the ith county. `names(NC.polys)` matches `infants$county` (order unimportant).

Source

The data were compiled by the North Carolina State Center for Health Statistics (https://schs.dph.ncdhhs.gov/).

`meps` *MEPS: Medical Expenditure Panel Survey (year 2012)*

Description

Subsample of the 2012 MEPS data, collected and published by the U.S. Agency for Healthcare Research and Quality.

Usage

```
data(meps)
```

Format

`meps` is a 10638 row data frame with the following columns:

general General health: 1 excellent, 2 very good, 3 good, 4 fair, 5 poor.

mental Mental health (as above).

bmi Body mass index.

income Income.

age Age.

gender Male 1, Female 0.

ethnicity 1 white, 2 black, 3 native american, 4 others.

education Education in years.

region 1 Northeast, 2 Midwest, 3 South, 4 West.

hypertension Equal to 1 if hypertension present and 0 otherwise.

hyperlipidemia Equal to 1 if hyperlipidemia present and 0 otherwise.

dvisit Number of doctor (physicians) visits.

ndvisit Number of non doctor visits (non-physician providers).

dvexpend Expenditure on doctor visits.

ndvexpend Expenditure on non doctor visits.

Source

https://meps.ahrq.gov

`war` *Civil war data*

Description

Civil war data from Fearon and Laitin (2003).

Usage

```
data(war)
```

Format

`war` is a 6326 row data frame with the following columns:

onset equal to 1 for all country-years in which a civil war started.

instab equal to 1 if unstable government.

oil equal to 1 for oil exporter country.

cwar equal to 1 if the country had a distinct civil war ongoing in the previous year.

gdp GDP per capita (measured as thousands of 1985 U.S. dollars) lagged one year.

ncontig equal to 1 for non-contiguous state.

nwstate equal to 1 for new state.

lpop log(population size).

lmnt log(mountainous).

ethfrac measure of ethnic fractionalization (calculated as the probability that two randomly drawn individuals from a country are not from the same ethnicity).

relfrac measure of religious fractionalisation.

poldem measure of political democracy (ranges from -10 to 10) lagged one year.

Source

Data are from:

Fearon J.D., Laitin D.D. (2003), Ethnicity, Insurgency, and Civil War. *The American Political Science Review*, 97, 75-90.

Index