Package 'limai'

April 17, 2025

Title Linear modeling and AI decision making

Version 0.0.1

Author Xu Liu [aut,cre]

Maintainer Xu Liu <liu.xu@sufe.edu.cn>

Description Data sets used in the book ``Linear modeling and AI decision making''

License GPL (>=2)

Encoding UTF-8

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.2

Depends R (>= 3.5.0)

Imports Matrix, glmnet

LazyData true

R topics documented:

`breastcancer` *Breast cancer data*

Description

The dataset contains gene expression and gene copy number information from 89 subjects.

Usage

```
data(breastcancer)
```

Arguments

dna	Copy number variation (CNV) data representing genomic DNA amplification or deletion events in tumor samples.
rna	Gene expression profiles measured via RNA transcript levels (e.g., microarray or RNA-seq data).
chrom	Chromosome numbers (1-22, X, Y) corresponding to the genomic location of the measured genes.
nuc	Nucleotide positions (start/end coordinates) of genes or probes on the chromosome (e.g., hg18/hg19 reference).
gene	Unique gene identifiers (e.g., Entrez Gene IDs or probe IDs) linked to genomic features.
genenames	Official gene symbols or names (e.g., BRCA1, ERBB2) standardized by HUGO Gene Nomenclature Committee (HGNC).
genechr	Chromosomal mapping information for each gene (e.g., "chr17" for TP53).
genedesc	Brief functional descriptions of genes (e.g., "tumor protein p53" or "estrogen receptor 1").
genepos	Genomic coordinates of genes (e.g., cytoband or base-pair positions like "17q21.31").

Details

The dataset is derived from molecular bioinformatics data obtained from breast cancer tissue samples treated according to the standard of care between 1989 and 1997. It primarily consists of gene expression profiles and copy number variation data across 22 chromosomal pairs in tumor tissue samples from 89 breast cancer patients. For a detailed explanation of this dataset, please refer to Chin et al. (2006).

Source

http://icbp.lbl.gov/breastcancer/ or R package PMA

References

Chin, K., DeVries, S., et al. (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer cell*, **10**(6), 529-541.

Examples

```
library(glmnet)
data(breastcancer)
dna = breastcancer$dna[breastcancer$chrom==21,]
rna = breastcancer$rna[which(breastcancer$genechr==21),]
y = dna[1,]
x = t(rna)
set.seed(100)
fit_ridge = cv.glmnet(x,y,alpha = 0)
coef(fit_ridge, s = "lambda.min")
fit_lasso = cv.glmnet(x,y,alpha = 1)
coef(fit_lasso, s = "lambda.min")
```

CHARLS *China Health and Retirement Longitudinal Study (CHARLS) data of Hebei, Shandong, and Fujian provinces.*

Description

The dataset contains the CHARLS data collected in Hebei, Shandong, and Fujian provinces.

Usage

```
data(CHARLS)
```

Format

A list object containing the following 3 variables:

Name	Type	Description
hebei	matrix	The CHARLS data of Hebei province. A matrix with 257 rows and 50 columns including y and 49 covariates v1,...,v49.
shandong	matrix	The CHARLS data of Shandong province. A matrix with 413 rows and 50 columns including y and 49 covariates v1,...,v49.
fujian	matrix	The CHARLS data of Fujian province. A matrix with 167 rows and 50 columns including y and 49 covariates v1,...,v49.

Details

Each matrix containing y, v1-v49 variables:

Name	Description
y	Annual support income of elderly people.
v1	Gender: 1 = male, 0 = female.
v2	Age.
v3	Education level: 1 = primary, 2 = junior high, 3 = high school, 4 = other.
v4	Marital status: 1 = married, 0 = unmarried.
v5	Live alone: 1 = yes, 0 = no.
v6	Live with a spouse: 1 = yes, 0 = no.
v7	Live with children: 1 = yes, 0 = no.
v8	Live with other members (e.g., parents): 1 = yes, 0 = no.
v9	Health status: 1 = disability/chronic illness, 0 = healthy.
v10	Pension income.
v11	Whether to receive a pension: 1 = yes, 0 = no.
v12	Number of surviving children: 0 = none, 1 = one, 2 = two or more.
v13	Wage income per household.
v14	Net operating income per household.
v15	Net transfer income per household.
v16	Number of children with a college degree or above.
v17	Number of children earning over 10,000 CNY annually.
v18	Emotional comfort: 1 = contact children \geq every half month, 0 = otherwise.
v19	Number of household members.
v20	Number of deceased biological children.

v21	Number of surviving adopted children.
v22	Number of surviving sons.
v23	Financial support for parents.
v24	Financial support for other relatives.
v25	Net financial support received from other relatives.
v26	Number of types of disability.
v27	Chronic illness: 0 = no, 1 = yes, 2 = other.
v28	Whether to receive a retirement pension: 1 = yes, 0 = no.
v29	Retirement pension income.
v30	New rural pension income.
v31	All other pension income.
v32	Pension income of elderly households.
v33	Total financial assets of elderly and spouses.
v34	Wage income of main household members.
v35	Government subsidies for individual families.
v36	Government subsidies for main household members.
v37	Wage income of other family members.
v38	Government subsidies for other family members.
v39	Total government subsidies for families.
v40	Government transfer income for households.
v41	Net household income excluding private transfers.
v42	Net household income.
v43	Net household income per capita.
v44	Other net private transfer income of elderly.
v45	Family shared income received by elderly.
v46	Whether to complete junior high school education: 1 = yes, 0 = no.
v47	Whether to complete high school education: 1 = yes, 0 = no.
v48	Annual net income from other sources.
v49	Financial support provided for children.

Source

The CHARLS data from https://charls.charlsdata.com/pages/Data/2015-charls-wave4/zh-cn.html

References

Ren, P., Liu, X., Zhang, X., Zhan, P., & Qiu, T. (2024). Integrative analysis of high-dimensional quantile regression with contrasted penalization. *Journal of Applied Statistics*, 1-17.

Examples

```
library(glmnet)
data(CHARLS)
data_hebei = CHARLS$hebei
y = data_hebei$y
x = data_hebei[,-1]
x = matrix(unlist(x),nrow = nrow(x))
fit_lasso = cv.glmnet(x,y,alpha = 1)
coef(fit_lasso,s = "lambda.min")
```

game *Online Gaming Behavior Dataset*

Description

This dataset captures comprehensive metrics and demographics related to player behavior in online gaming environments. It includes variables such as player demographics, game-specific details, engagement metrics, and a target variable reflecting player retention.

Usage

```
data(game)
```

Arguments

A data frame object containing 200 player entries with the following 13 variables:

Name	Type	Description
ID	integer	Unique identifier for each player
age	integer	Age of the player
gender	character	Gender of the player
location	character	Geographic location of the player
genre	character	Genre of the game the player is engaged in
time	numeric	Average hours spent playing per session
inbuy	integer	Indicates whether the player makes in-game purchases (0 = No, 1 = Yes)
difficulty	character	Difficulty level of the game
session	integer	Number of gaming sessions per week
sesslong	integer	Average duration of each gaming session in minutes
level	integer	Current level of the player in the game
achievement	integer	Number of achievements unlocked by the player
engagement	character	Categorized engagement level reflecting player retention ('High', 'Medium', 'Low')

Details

The data provides information on various aspects of online gaming behavior for 200 players, including player identifiers, demographic details, gaming - related metrics.

Source

The dataset from Kaggle website https://www.kaggle.com/datasets/rabieelkharoua/predict-online-gaming-behavior-dataset/data

Examples

```
data(game)

# Check the relationship between in - game purchases and session length
cor(game$inbuy, game$sesslong, use = "complete.obs")

# Count the number of players in each game genre
genre_counts <- table(game$genre)
```

```
barplot(genre_counts,
        main = "Number of Players in Each Game Genre",
        xlab = "Game Genre",
        ylab = "Number of Players")
```

GTEx_apoe *GTEx data associated with APOE gene*

Description

The dataset contains APOE gene and other related genes from 111 subjects.

Usage

```
data(GTEx_apoe)
```

Details

This dataset contains gene expression profiles and genomic data derived from the Genotype-Tissue Expression (GTEx) project. It is a list with 111 rows and 119 columns, and the last column is the APOE gene, the other 118 columns is the related genes.

Source

Genotype-Tissue Expression (GTEx) project, available at: https://www.gtexportal.org/home/

Examples

```
library(glmnet)
data(GTEx_apoe)
y = GTEx_apoe$APOE
x = GTEx_apoe[,-1]
x = matrix(unlist(x), nrow = nrow(x))
fit_lasso = cv.glmnet(x,y,alpha = 1)
coef(fit_lasso,s = "lambda.min")
```

---

pollution *Air Quality Index (AQI) for Chinese Cities (2022)*

Description

A multidimensional dataset containing weekly Air Quality Index (AQI), meteorological parameters, and socioeconomic indicators for 173 Chinese cities in 2022.

Usage

```
data(pollution)
```

Arguments

A list object containing 173 city entries with the following 10 variables:

Name	Type	Description
AQI	matrix	Air Quality Index, a matrix with 173 rows (cities) and 51 columns (weekly AQI values). Higher values indicate poorer air quality.
city	character	City names vector (length 173).
temp	numeric	Annual mean air temperature in °C.
dew	numeric	Annual mean dew point temperature in °C.
windD	numeric	Wind direction in degrees (0-360).
windS	numeric	Annual mean wind speed in m/s.
pres	numeric	Annual mean atmospheric pressure in hPa.
pop	numeric	Household resident population (unit: 10,000).
green	numeric	Green Covered Area as percentage of Completed Area (0-100).
second	numeric	Secondary Industry as Percentage to GRP (0-100).

Details

The data provides AQI data for 173 Chinese cities for the 51 weeks of 2022 and economic and meteorological related annual average data.

Source

- Air Quality Index form China National Environmental Monitoring Center(https://air.cnemc.cn:18007/)
- Meteorological Data from NOAA National Centers for Environmental Information (https://www.ncei.noaa.gov/)
- Socioeconomic data from China City Statistical Yearbook (https://www.stats.gov.cn/)

References

Guan, X., Li, Y., Liu, X., & You, J. (2025). Subgroup learning in functional regression models under the RKHS framework. *arXiv preprint arXiv:2503.01515*.

Examples

```
data(pollution)

# Explore AQI distribution for Beijing
bj_aqi <- as.numeric(pollution$AQI[pollution$city == "Beijing", ])
plot(bj_aqi,
     type = "l",
     main = "Weekly AQI in Beijing (2022)",
     xlab = "Week",
     ylab = "AQI")

# Correlation analysis
cor(pollution$temp, rowMeans(pollution$AQI, na.rm = TRUE))
```

---

skcm *Skin Cutaneous Melanoma (SKCM) data*

Description

The dataset contains clinical outcome measurements and high-dimensional gene expression profiles from 361 subjects with skin cutaneous melanoma.

Usage

```
data(skcm)
```

Arguments

y	A numeric vector of length 361 representing Breslow's thickness, a clinicopathologic feature of cutaneous melanoma.
gexp	A data frame with 361 rows and 2000 columns, where each column represents expression values of a gene. Gene names are provided as column names (e.g., SLC8A1, DPYD).
adj_matrix	A 2000 x 2000 adjacency matrix representing gene-gene interaction relationships in gexp. Entries are defined as: • 1: If two genes are directly interacting in the KEGG melanoma pathway (hsa05218) • 0: Otherwise (no interaction or gene not in the pathway) We obtain melanoma-related pathway data through KEGG portal: 1. Access https://www.kegg.jp/ and search "melanoma" 2. Identify pathway ID: hsa05218 (Skin Cutaneous Melanoma) 3. Contains 73 genes with curated regulatory relationships Key construction steps see Details.

Details

The dataset includes 361 samples with outcomes (Breslow's thickness measurements) and expression levels of the top 2000 most variable genes. It is derived from The Cancer Genome Atlas (TCGA) for Skin Cutaneous Melanoma (SKCM), one of the most aggressive cancer types.

adj_matrix construction follows these key steps:

1. **Pathway Data Retrieval**: Download KGML file from KEGG via:

   ```
   download.file("https://rest.kegg.jp/get/hsa05218/kgml", "melanoma_kgml.xml")
   ```

2. **Graph Parsing**: Convert KGML to adjacency matrix:

   ```
   if (!require("BiocManager", quietly = TRUE))
   install.packages("BiocManager")
   BiocManager::install("KEGGgraph")

   library(KEGGgraph)
   melanoma_kgml <- "melanoma_kgml.xml"
   kegg_graph <- parseKGML2Graph(melanoma_kgml, genesOnly = TRUE)
   adjacent <- as(kegg_graph, "matrix")
   adj_matrix_KEGG <- adjacent + t(adjacent)
   ```

3. **Gene ID Conversion**: Map KEGG IDs to Gene Symbols:

   ```
   BiocManager::install("KEGGREST")
   BiocManager::install("org.Hs.eg.db")

   library(KEGGREST)
   library(org.Hs.eg.db)
   kegg_genes <- keggLink("hsa", "hsa05218")
   kegg_ids <- gsub("hsa:", "", kegg_genes)
   ```

```
gene_symbols <- mapIds(org.Hs.eg.db,
                    keys = kegg_ids,
                    column = "SYMBOL",
                    keytype = "ENTREZID")  # Get Symbol mapping
rownames(adj_matrix_KEGG) <- gene_symbols[match(rownames(adj_matrix_KEGG), kegg_genes)]
colnames(adj_matrix_KEGG) <- rownames(adj_matrix_KEGG)
```

4. **Matrix Embedding**: Create final adjacency matrix:

```
gene_skcm <- colnames(skcm$gexp)
p <- length(gene_skcm)
adj_matrix_skcm <- matrix(0, nrow = p, ncol = p, dimnames = list(gene_skcm, gene_skcm))
common_genes <- intersect(gene_skcm, gene_symbols)
adj_matrix_skcm[common_genes, common_genes] <- adj_matrix_KEGG[common_genes, common_genes]
```

Source

The Cancer Genome Atlas (TCGA) portal: https://tcga-data.nci.nih.gov

References

The Cancer Genome Atlas Consortium. (2015). Genomic classification of cutaneous melanoma. *Cell*, **161**(7), 1681-1696.

Tan, X., Zhang, X., Cui, Y., & Liu, X. (2024). Uncertainty quantification in high-dimensional linear models incorporating graphical structures with applications to gene set analysis. *Bioinformatics*, **40**(9), btae541.

Examples

```
data(skcm)
hist(skcm$y, main = "Distribution of Clinical Outcomes", xlab = "Outcome Value")
pca_result <- prcomp(skcm$gexp[, 1:100], scale = TRUE) # Run PCA on the top 100 genes
plot(pca_result$x[, 1:2], main = "PCA of Gene Expression Data")
```

Index