Package 'CASIdata'

February 12, 2026

Title Datasets from Computer Age Statistical Inference

Version 0.2.1

Description Provides the datasets from Efron & Hastie (2016, ISBN: 9781108107952), ``Computer Age Statistical Inference: Algorithms, Evidence, and Data Science", in an accessible R format for those who want to use them for study or to try to reproduce analyses from the book.

License GPL (>= 3)

Encoding UTF-8

Depends R (>= 3.5.0)

Suggests knitr, rmarkdown, vcdExtra, car

VignetteBuilder knitr

RoxygenNote 7.3.3

Language en-US

URL https://github.com/friendly/CASIdata, https://friendly.github.io/CASIdata/

BugReports https://github.com/friendly/CASIdata/issues

NeedsCompilation no

Author Michael Friendly [aut, cre]

Maintainer Michael Friendly <friendly@yorku.ca>

Repository CRAN

Date/Publication 2026-02-12 08:00:20 UTC

Contents

`als` *ALS Data*

Description

Data on amyotrophic lateral sclerosis (Lou Gehrig's disease) from Section 17.2. There are 1822 observations on individuals with ALS. The goal is to predict the rate of progression dFRS of a functional rating score, using 369 predictors based on measurements (and derivatives of these) obtained from patient visits.

Format

A data frame with 1822 rows and 371 variables. The key variables are `testset` (logical indicator for training/test split) and `dFRS` (response: rate of progression of the ALS functional rating score). The 369 predictor variables include:

- Demographics: `Age`, `Sex.Male`, `Sex.Female`, and race indicators (`Race...Caucasian`, `Race...Asian`, etc.)
- Family history of neurological diseases in relatives (e.g., `Father`, `Mother`, `Brother`, `Sister`)
- Neurological disease indicators (e.g., `Neurological.Disease.ALS`, `Neurological.Disease.PARKINSON.S.DISEASE`)
- Site of onset (`Site.of.Onset.Onset..Bulbar`, `Site.of.Onset.Onset..Limb`)
- Symptoms (`Symptom.Atrophy`, `Symptom.Cramps`, `Symptom.Fasciculations`, `Symptom.Speech`, etc.)
- Study arm indicators (`Study.Arm.ACTIVE`, `Study.Arm.PLACEBO`)
- Clinical measurements with summary statistics (first, last, min, max, mean, sd, slope): ALSFRS scores, blood pressure, forced/slow vital capacity (`fvc.liters`, `svc.liters`), respiratory rate, weight, height
- ALSFRS subscale items: `climbing.stairs`, `cutting`, `dressing`, `handwriting`, `salivation`, `speech`, `swallowing`, `turning`, `walking`

Details

These data were kindly provided by Lester Mackey and Lilly Fang, who won the DREAM challenge prediction prize in 2012 (Kuffner et al., 2015). It includes some additional variables created by them. Their winning entry used Bayesian trees, not too different from random forests.

Source

https://hastie.su.domains/CASI_files/DATA/ALS.txt

References

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press, Section 17.2.

Examples

```
data(als)
str(als)
```

---

baseball *Baseball Batting Averages*

Description

Batting averages for 18 Major League players in the 1970 season, from Table 7.1. This dataset illustrates empirical Bayes estimation, where early-season performance is used to predict full-season batting averages.

Format

A data frame with 18 rows and 3 variables:

Player Player ID number

MLE Batting average based on the first 90 at-bats of the season

TRUTH Batting average for the remainder of the 1970 season

Source

https://hastie.su.domains/CASI_files/DATA/baseball.txt

References

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press, Table 7.1.

Examples

```
data(baseball)
str(baseball)
```

bivnorm *Bivariate Normal Data*

Description

40 points generated from a bivariate normal distribution, with some entries missing. From Figure 9.3.

Format

A data frame with 40 rows and 2 variables:

X1 First variable

X2 Second variable

Source

https://hastie.su.domains/CASI_files/DATA/bivnorm.csv

References

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press, Figure 9.3.

Examples

```
data(bivnorm)
str(bivnorm)
```

butterfly *Butterfly Species Data*

Description

Number of butterfly species seen a given number of times each in two years of trapping. From Table 6.2. This is a frequency data frame.

Format

A data frame with 24 rows and 2 variables:

k Number of times a species was trapped

count Number of species seen exactly k times (e.g., 118 species trapped just once, 74 trapped twice each)

Source

https://hastie.su.domains/CASI_files/DATA/butterfly.txt

References

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press, Table 6.2.

Examples

```
data(butterfly)
str(butterfly)
```

`cellinfusion` *Cell Infusion Data*

Description

Human cell colonies infused with mouse nuclei in 5 different ratios over 1 to 5 days. From Table 8.2.

Format

A data frame with 25 rows and 4 variables:

thrived Number of cells that thrived

N Colony size (number of cells)

ratio Ratio of mouse nuclei to human cells (1-5)

time Day of observation (1-5)

Source

https://hastie.su.domains/CASI_files/DATA/cellinfusion.txt

References

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press, Table 8.2.

Examples

```
data(cellinfusion)
str(cellinfusion)
```

cholesterol *Cholesterol Data*

Description

Cholestyramine, a proposed cholesterol lowering drug, was administered to 164 men for an average of seven years each. From Figure 20.1.

Format

A data frame with 164 rows and 2 variables:

compliance Fraction of intended dose actually taken (standardized)

cholesterol.decrease Decrease in cholesterol level over the course of the experiment

Source

https://hastie.su.domains/CASI_files/DATA/cholesterol.txt

References

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press, Figure 20.1.

Examples

```
data(cholesterol)
str(cholesterol)
```

diabetes *Diabetes Data*

Description

Data from 442 diabetes patients used in Section 7.3. The response is a quantitative measure of disease progression one year after baseline. There are ten baseline predictors: age, sex, body-mass index, average blood pressure, and six blood serum measurements.

Format

A data frame with 442 rows and 12 variables:

X Row index

age Age of patient

sex Sex of patient

bmi Body mass index

map Average blood pressure (mean arterial pressure)

tc Total cholesterol (serum measurement)

ldl Low-density lipoproteins (serum measurement)

hdl High-density lipoproteins (serum measurement)

tch Total cholesterol / HDL (serum measurement)

ltg Log of triglycerides (serum measurement)

glu Blood sugar level (serum measurement)

prog Response: quantitative measure of disease progression

Details

First used in the LARS paper (Efron et al., 2004).

Note: In Table 7.2, the centered predictor variables were standardized to unit L2 norm. In Table 20.1 they were standardized to unit variance.

Source

https://hastie.su.domains/CASI_files/DATA/diabetes.csv

References

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least Angle Regression. *Annals of Statistics*, 32(2), 407-499.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press, Section 7.3.

Examples

```
data(diabetes)
str(diabetes)
```

---

doseresponse *Dose Response Data*

Description

Data from 11 groups of mice (10 each) exposed to drug Xilathon at different doses. From Figure 8.2.

Format

A data frame with 11 rows and 2 variables:

Dose Log dose level (each step is a doubling)

Proportion Proportion of mice that died at that dose

Source

https://hastie.su.domains/CASI_files/DATA/doseresponse.txt

References

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press, Figure 8.2.

Examples

```
data(doseresponse)
str(doseresponse)
```

DTI *DTI Brain Imaging Data*

Description

Diffusion Tensor Imaging (DTI) data comparing 6 dyslexic children with 6 normal controls, from Figures 15.9 and 15.10. Z scores were computed at 15,443 three-dimensional brain coordinates (voxels).

Format

A data frame with 15443 rows and 4 variables:

x Voxel coordinate: back to front

y Voxel coordinate: left to right

z Voxel coordinate: bottom to top

Zscore Z score comparing dyslexic vs normal controls at this voxel

Source

https://hastie.su.domains/CASI_files/DATA/DTI.csv

References

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press, Figures 15.9 and 15.10.

Examples

```
data(DTI)
str(DTI)
```

galaxy *Galaxy Data*

Description

Counts of galaxies binned by redshift and magnitude, from Table 8.5. The data have been reshaped into long format with variables for magnitude, redshift category, and frequency count.

Format

A data frame with 270 rows and 3 variables:

mag Magnitude category (1-18)

red Redshift category (1-15)

freq Number of galaxies in this bin

Source

https://hastie.su.domains/CASI_files/DATA/galaxy.txt

References

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press, Table 8.5.

Examples

```
data(galaxy)
str(galaxy)

library(car)

## Fit a main effects Poisson GLM
# This treats `mag` and `red` as numeric
galaxy.mod0 <- glm(freq ~ mag + red,
                   data = galaxy, family = poisson)
Anova(galaxy.mod0)

## Fit response surface model
galaxy.mod1 <- glm(freq ~ poly(mag,2) +
                     poly(red, 2) +
                     mag : red,
                   data = galaxy, family = poisson)
Anova(galaxy.mod1)
summary(galaxy.mod1)
```

haplotype *Human Ancestry Haplotype Data*

Description

Genotype data for 197 US individuals from 4 racial groups (African American, European, Japanese, and African) at 100 SNP locations. From Section 13.5.

Format

A data frame with 197 rows and 102 variables. The first column `X` is a row index, `race` is the racial/ethnic group identifier, and the remaining 100 columns (`Snp1` through `Snp100`) contain genotype values (0, 1, or 2) at each SNP location, with `NA` for missing values.

Source

https://hastie.su.domains/CASI_files/DATA/haplotype.csv

References

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press, Section 13.5.

Examples

```
data(haplotype)
str(haplotype)
```

insurance *Insurance Life Table Data*

Description

Insurance company life table from Table 9.1. At each age, gives the number of policy holders and the number of deaths.

Format

A data frame with rows for each age group and 3 variables:

age Age of policy holders

n Number of policy holders at this age

y Number of deaths at this age

Source

https://hastie.su.domains/CASI_files/DATA/insurance.txt

References

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press, Table 9.1.

Examples

```
data(insurance)
str(insurance)
```

`leukemia_small` *Leukemia Gene Expression Data (Small)*

Description

Gene expression measurements on 72 leukemia patients: 47 ALL (acute lymphoblastic leukemia) and 25 AML (acute myeloid leukemia). From the landmark Golub et al. (1999) Science paper. This smaller subset contains 3571 genes and is used in Section 19.1.

Format

A data frame with 3571 rows (genes) and 72 columns (patients). Column names indicate the class label (ALL or AML) for each patient.

Details

A larger dataset with 7128 genes is also available from the CASI website.

Source

https://hastie.su.domains/CASI_files/DATA/leukemia_small.csv

References

Golub, T.R., et al. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286, 531-537.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press, Section 19.1.

Examples

```
data(leukemia_small)
str(leukemia_small)
```

ncog *NCOG Head and Neck Cancer Data*

Description

Head and neck cancer survival data from the Northern California Oncology Group (NCOG), from Section 9.2. Patients were randomized to one of two treatment arms.

Format

A data frame with survival time information and variables:

t Time in months until death or censoring

d Death indicator: 1 = death observed, 0 = censored

arm Treatment arm: "A" = Chemotherapy, "B" = Chemotherapy + Radiation

day Day of event/censoring

month Month of event/censoring

year Year of event/censoring

Source

https://hastie.su.domains/CASI_files/DATA/ncog.txt

References

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press, Section 9.2.

Examples

```
data(ncog)
str(ncog)
```

nodes *Lymph Nodes Cancer Data*

Description

Data on lymph nodes removed from 844 cancer patients, from Figure 6.3.

Format

A data frame with 844 rows and 2 variables:

n Number of lymph nodes removed

x Number of nodes found to be positive (malignant)

Source

https://hastie.su.domains/CASI_files/DATA/nodes.txt

References

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press, Figure 6.3.

Examples

```
data(nodes)
str(nodes)
```

---

pediatric *Pediatric Cancer Survival Data*

Description

Survival data on 1620 children with cancer, from Section 9.4 and Table 9.6.

Format

A data frame with 1620 rows and 7 variables:

sex Sex: 1 = male, 2 = female

race Race: 1 = white, 2 = nonwhite

age Age in years

entry Calendar date of entry in days since July 1, 2001

far Home distance from treatment center in miles

t Survival time in days

d Death indicator: 1 = death observed, 0 = censored

Source

https://hastie.su.domains/CASI_files/DATA/pediatric.txt

References

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press, Section 9.4, Table 9.6.

Examples

```
data(pediatric)
str(pediatric)
```

police *Police Racial Bias Data*

Description

Z scores for 2749 New York City police officers, from Figure 15.7. A large value suggests racial bias in policing behavior.

Format

A data frame with 2749 rows and 1 variable:

z Z score measuring potential racial bias

Source

https://hastie.su.domains/CASI_files/DATA/police.txt

References

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press, Figure 15.7.

Examples

```
data(police)
str(police)
```

prostz *Prostate Cancer Z-values*

Description

Vector of 6033 z-values comparing gene expression between prostate cancer patients and controls, as pictured in Figure 3.4. These were computed as described on page 272.

Format

A data frame with 6033 rows and 1 variable:

z Z-value for each gene comparing cancer vs control expression

Source

https://hastie.su.domains/CASI_files/DATA/prostz.txt

References

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press, Section 3.3, Figure 3.4.

Examples

```
data(prostz)
str(prostz)
```

`student_score` *Student Score Data*

Description

Test scores for 22 students on 5 different exams, from Tables 3.1 and 10.1.

Format

A data frame with 22 rows and 5 variables:

mech Mechanics exam score

vecs Vectors exam score

alg Algebra exam score

analy Analysis exam score

stat Statistics exam score

Source

https://hastie.su.domains/CASI_files/DATA/student_score.txt

References

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press, Tables 3.1 and 10.1.

Examples

```
data(student_score)
str(student_score)
```

`supernova` *Type Ia Supernova Data*

Description

Measurements from 39 Type Ia supernovas, from Figure 12.1 and Table 12.1. These supernovas were close enough to Earth to observe their actual magnitudes. The goal is to predict magnitude from spectral energy measurements.

Format

A data frame with 39 rows and 11 variables:

Magnitude Actual observed magnitude of the supernova

E1 Spectral energy in frequency band 1

E2 Spectral energy in frequency band 2

E3 Spectral energy in frequency band 3

E4 Spectral energy in frequency band 4

E5 Spectral energy in frequency band 5

E6 Spectral energy in frequency band 6

E7 Spectral energy in frequency band 7

E8 Spectral energy in frequency band 8

E9 Spectral energy in frequency band 9

E10 Spectral energy in frequency band 10

Source

https://hastie.su.domains/CASI_files/DATA/supernova.txt

References

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press, Figure 12.1, Table 12.1.

Examples

```
data(supernova)
str(supernova)
```

`vasoconstriction` *Vasoconstriction Data*

Description

Data on vasoconstriction (lung constriction) response, from Table 13.2.

Format

A data frame with 39 rows and 2 variables:

volume Volume measurement

constriction Logical: TRUE if constriction occurred, FALSE otherwise

Source

https://hastie.su.domains/CASI_files/DATA/vasoconstriction.txt

References

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press, Table 13.2.

Examples

```
data(vasoconstriction)
str(vasoconstriction)
```

Index