Package 'ForCausality'

October 25, 2025

Type Package

Title A Curated Collection of 'Causal Inference' Datasets and Tools

Version 0.1.0

Maintainer Tomás Valderrama <tomasvm2004@gmail.com>

Description Provides a comprehensive set of datasets and tools for 'causal inference' research.
The package includes data from clinical trials, cancer studies, epidemiological surveys, environmental exposures, and health-related observational studies.
Designed to facilitate causal analysis, risk assessment, and advanced statistical modeling,
it leverages datasets from packages such as 'causalOT', 'survival', 'causalPAF', 'evident', 'melt', and 'sanon'.
The package is inspired by the foundational work of Pearl (2009) <doi:10.1017/CBO9780511803161> on causal inference frameworks.

License GPL-3

URL https://github.com/Toby-codigos/ForCausality,
https://toby-codigos.github.io/ForCausality/

BugReports https://github.com/Toby-codigos/ForCausality/issues

Encoding UTF-8

LazyData true

Suggests ggplot2, dplyr, testthat (>= 3.0.0), knitr, rmarkdown

RoxygenNote 7.3.3

Config/testthat/edition 3

VignetteBuilder knitr

NeedsCompilation no

Author Tomás Valderrama [aut, cre]

Depends R (>= 3.5.0)

Repository CRAN

Date/Publication 2025-10-25 12:40:22 UTC

Contents

`Benzene_df` *Benzene Exposure and Chromosome Damage Data*

Description

This dataset, Benzene_df, is a data frame containing indicators of chromosome damage related to benzene exposure, alcohol consumption, and smoking habits. The dataset consists of 78 observations and 5 variables, including age, exposure, and lifestyle factors. Some observations may contain missing values.

Usage

```
data(Benzene_df)
```

Format

A data frame with 78 observations and 5 variables:

age Age of the subject (integer)

exposure Benzene exposure indicator (integer)

alcohol Alcohol consumption indicator (integer)

smoking Smoking indicator (numeric)

totalplus Chromosome damage measure (numeric)

Details

The dataset name has been kept as 'Benzene_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the ForCausality package and assists users in identifying its specific characteristics. The suffix 'df' indicates that the dataset is a data frame. The original content has not been modified in any way.

Source

Data taken from the evident package version 1.0.4

`Cloth_df` *Clothianidin Concentration in Maize Plants*

Description

This dataset, Cloth_df, is a data frame containing measurements of clothianidin concentration in maize plants under different treatments. The dataset consists of 102 observations and 3 variables, including block identifiers, treatment types, and measured concentrations. Some observations may contain missing values.

Usage

```
data(Cloth_df)
```

Format

A data frame with 102 observations and 3 variables:

blk Block identifier (factor)

trt Treatment type (factor)

clo Clothianidin concentration (numeric)

Details

The dataset name has been kept as 'Cloth_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the ForCausality package and assists users in identifying its specific characteristics. The suffix 'df' indicates that the dataset is a data frame. The original content has not been modified in any way.

Source

Data taken from the melt package version 1.11.4

Colon_df *Chemotherapy Data for Stage B/C Colon Cancer*

Description

This dataset, Colon_df, contains data from a clinical trial of chemotherapy for patients with Stage B/C colon cancer. The dataset includes 1,858 observations and 16 variables, providing information on patient demographics, treatment assignment, disease characteristics, and outcomes. Some observations contain missing values.

Usage

```
data(Colon_df)
```

Format

A data frame with 1,858 observations and 16 variables:

id Patient identifier (numeric)

study Study number (numeric)

rx Treatment group (factor)

sex Sex of the patient (numeric)

age Age of the patient in years (numeric)

obstruct Obstruction present (numeric indicator)

perfor Perforation present (numeric indicator)

adhere Adherence to adjacent structures (numeric indicator)

nodes Number of lymph nodes with cancer (numeric)

status Patient status (numeric indicator)

differ Tumor differentiation (numeric)

extent Extent of local spread (numeric)

surg Surgical procedure performed (numeric indicator)

node4 At least 4 nodes positive (numeric indicator)

time Follow-up time in days (numeric)

etype Type of event (numeric indicator)

Details

The dataset name has been kept as 'Colon_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the ForCausality package and assists users in identifying its specific characteristics. The suffix 'df' indicates that the dataset is a data frame. The original content has not been modified in any way.

Source

Data taken from the survival package version 3.8-3

`ForCausality` *ForCausality: A Curated Collection of Causal Inference Datasets and Tools*

Description

Provides a comprehensive set of datasets and tools for causal inference research. The package includes data from clinical trials, cancer studies, epidemiological surveys, environmental exposures, and health-related observational studies.

Details

ForCausality: A Curated Collection of Causal Inference Datasets and Tools

A Curated Collection of Causal Inference Datasets and Tools

Author(s)

Maintainer: Tomás Valderrama <tomasvm2004@gmail.com>

See Also

Useful links:

- https://github.com/Toby-codigos/ForCausality

`Gbsg_df` *Breast Cancer Prognostic Data (German Breast Cancer Study Group)*

Description

This dataset, Gbsg_df, provides prognostic factors for breast cancer patients from the German Breast Cancer Study Group (GBSG). The dataset includes 686 observations and 11 variables, containing information on patient demographics, tumor characteristics, hormone receptor status, and outcomes. Some observations contain missing values.

Usage

```
data(Gbsg_df)
```

Format

A data frame with 686 observations and 11 variables:

pid Patient identifier (integer)

age Age at diagnosis (integer)

meno Menopausal status (integer indicator)

size Tumor size in millimeters (integer)

grade Tumor grade (integer)

nodes Number of positive lymph nodes (integer)

pgr Progesterone receptor level (integer)

er Estrogen receptor level (integer)

hormon Hormonal therapy received (integer indicator)

rfstime Relapse-free survival time in days (integer)

status Patient status (integer indicator)

Details

The dataset name has been kept as 'Gbsg_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the ForCausality package and assists users in identifying its specific characteristics. The suffix 'df' indicates that the dataset is a data frame. The original content has not been modified in any way.

Source

Data taken from the survival package version 3.8-3

Lead_df *Lead Exposure Data*

Description

This dataset, Lead_df, is a data frame comparing control and exposed groups under different hygiene and exposure levels. The dataset consists of 33 observations and 6 variables, including measures of exposure, hygiene, and calculated differences between groups. Some observations may contain missing values.

Usage

```
data(Lead_df)
```

Format

A data frame with 33 observations and 6 variables:

control Control group count (integer)

exposed Exposed group count (integer)

level Exposure level (factor with 3 levels: "high", "low", "medium")

hyg Hygiene level (factor with 3 levels: "good", "mod", "poor")

both Combined exposure and hygiene category (factor with 4 levels, e.g. "high.ok", "high.poor", ...)

dif Difference between control and exposed (integer)

Details

The dataset name has been kept as 'Lead_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the ForCausality package and assists users in identifying its specific characteristics. The suffix 'df' indicates that the dataset is a data frame. The original content has not been modified in any way.

Source

Data taken from the evident package version 1.0.4

Mouse_df *Mouse Cancer Trial Data*

Description

This dataset, Mouse_df, provides data from mouse cancer trials used in studies by Royston and Altman. The dataset includes 181 observations and 4 variables, covering information on treatment assignment, survival time, outcome, and mouse identifiers. Some observations contain missing values.

Usage

```
data(Mouse_df)
```

Format

A data frame with 181 observations and 4 variables:

trt Treatment group (factor)

days Survival time in days (numeric)

outcome Trial outcome (factor)

id Mouse identifier (integer)

Details

The dataset name has been kept as 'Mouse_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the ForCausality package and assists users in identifying its specific characteristics. The suffix 'df' indicates that the dataset is a data frame. The original content has not been modified in any way.

Source

Data taken from the survival package version 3.8-3

`Pain_df` *Chronic Pain Clinical Trial Data*

Description

This dataset, Pain_df, is a data frame containing clinical trial data for chronic pain treatments. The trial compared active treatment versus placebo across different clinical centers and diagnoses. The dataset consists of 193 observations and 4 variables. Some observations may contain missing values.

Usage

```
data(Pain_df)
```

Format

A data frame with 193 observations and 4 variables:

treat Treatment group (factor: active vs placebo)

response Response outcome (factor)

center Clinical trial center (factor)

diagnosis Diagnosis category (factor)

Details

The dataset name has been kept as 'Pain_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the ForCausality package and assists users in identifying its specific characteristics. The suffix 'df' indicates that the dataset is a data frame. The original content has not been modified in any way.

Source

Data taken from the sanon package version 1.6

`Periodontal_df` *Periodontal Disease Data*

Description

This dataset, Periodontal_df, is a data frame containing information on smoking habits, demographics, and periodontal health indicators. The dataset consists of 882 observations and 12 variables, including smoking frequency, socioeconomic indicators, and periodontal measures. Some observations may contain missing values.

Usage

```
data(Periodontal_df)
```

Format

A data frame with 882 observations and 12 variables:

SEQN Sequence identifier (numeric)

female Sex indicator (numeric)

age Age in years (numeric)

black Race indicator for Black participants (numeric)

educf Education level (ordered factor with 5 levels)

income Income measure (numeric)

cigsperday Cigarettes smoked per day (numeric)

either Count of sites with periodontal disease (integer)

neither Count of sites without periodontal disease (integer)

pcteither Percentage of sites with periodontal disease (numeric)

z Standardized measure (numeric)

mset Additional periodontal health indicator (numeric)

Details

The dataset name has been kept as 'Periodontal_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the ForCausality package and assists users in identifying its specific characteristics. The suffix 'df' indicates that the dataset is a data frame. The original content has not been modified in any way.

Source

Data taken from the evident package version 1.0.4

Pph_df *External Control Trial Data for Post-partum Hemorrhage*

Description

This dataset, Pph_df, provides data from an external control trial of treatments for post-partum hemorrhage. The dataset includes 802 observations and 17 variables, containing information on blood loss, treatment assignment, demographic characteristics, and educational background. Some observations contain missing values.

Usage

```
data(Pph_df)
```

Format

A data frame with 802 observations and 17 variables:

cum_blood_20m Cumulative blood loss at 20 minutes (numeric)

tx Treatment indicator (numeric)

age Age of the participant (numeric)

no_educ Indicator for no formal education (numeric)

... Additional variables related to treatment and outcomes (numeric)

Details

The dataset name has been kept as 'Pph_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the ForCausality package and assists users in identifying its specific characteristics. The suffix 'df' indicates that the dataset is a data frame. The original content has not been modified in any way.

Source

Data taken from the causalOT package version 1.0.2

Resp_df *Respiratory Disorder Clinical Trial Data*

Description

This dataset, Resp_df, is a data frame containing repeated measurements from a clinical trial on respiratory disorders under two treatment conditions. The dataset records demographic information (center, sex, age), baseline measures, and follow-up measurements across four visits. It consists of 111 observations and 9 variables. Some observations may contain missing values.

Usage

```
data(Resp_df)
```

Format

A data frame with 111 observations and 9 variables:

center Clinical trial center (factor)

treatment Treatment group (character)

sex Sex of the participant (character)

age Age of the participant (integer)

baseline Baseline measurement (integer)

visit1 Measurement at visit 1 (integer)

visit2 Measurement at visit 2 (integer)

visit3 Measurement at visit 3 (integer)

visit4 Measurement at visit 4 (integer)

Details

The dataset name has been kept as 'Resp_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the ForCausality package and assists users in identifying its specific characteristics. The suffix 'df' indicates that the dataset is a data frame. The original content has not been modified in any way.

Source

Data taken from the sanon package version 1.6

Rotterdam_df *Breast Cancer Prognostic Data (Rotterdam Study)*

Description

This dataset, Rotterdam_df, provides prognostic factors for breast cancer patients used in the studies of Royston and Altman. The dataset includes 2,982 observations and 15 variables, covering patient demographics, tumor characteristics, treatments, and outcomes. Some observations contain missing values.

Usage

```
data(Rotterdam_df)
```

Format

A data frame with 2,982 observations and 15 variables:

pid Patient identifier (integer)

year Year of surgery (integer)

age Age at diagnosis (integer)

meno Menopausal status (integer indicator)

size Tumor size category (factor)

grade Tumor grade (integer)

nodes Number of positive lymph nodes (integer)

pgr Progesterone receptor level (integer)

er Estrogen receptor level (integer)

hormon Hormonal therapy received (integer indicator)

chemo Chemotherapy received (integer indicator)

rtime Relapse-free survival time in days (numeric)

recur Recurrence indicator (integer)

dtime Time to death in days (numeric)

death Death indicator (integer)

Details

The dataset name has been kept as 'Rotterdam_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the ForCausality package and assists users in identifying its specific characteristics. The suffix 'df' indicates that the dataset is a data frame. The original content has not been modified in any way.

Source

Data taken from the survival package version 3.8-3

Sebor_df *Seborrheic Dermatitis Clinical Trial Data*

Description

This dataset, Sebor_df, is a data frame containing clinical trial data on seborrheic dermatitis, comparing test and placebo treatments. It records participant center, treatment assignment, dermatitis scores across three assessments, and severity indicators at the same points. The dataset consists of 167 observations and 8 variables. Some observations may contain missing values.

Usage

```
data(Sebor_df)
```

Format

A data frame with 167 observations and 8 variables:

center Clinical trial center (factor)

treat Treatment group: test or placebo (character)

score1 Dermatitis score at assessment 1 (integer)

score2 Dermatitis score at assessment 2 (integer)

score3 Dermatitis score at assessment 3 (integer)

severity1 Severity indicator at assessment 1 (integer)

severity2 Severity indicator at assessment 2 (integer)

severity3 Severity indicator at assessment 3 (integer)

Details

The dataset name has been kept as 'Sebor_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the ForCausality package and assists users in identifying its specific characteristics. The suffix 'df' indicates that the dataset is a data frame. The original content has not been modified in any way.

Source

Data taken from the sanon package version 1.6

Skin_df *Skin Condition Clinical Trial Data*

Description

This dataset, Skin_df, is a data frame containing clinical trial data on skin conditions, comparing responses under placebo and test treatments. It includes participant center, treatment assignment, disease stage, and responses across three assessments. The dataset consists of 172 observations and 6 variables. Some observations may contain missing values.

Usage

```
data(Skin_df)
```

Format

A data frame with 172 observations and 6 variables:

center Clinical trial center (factor)

treat Treatment group: placebo or test (factor)

stage Disease stage (integer)

res1 Response at assessment 1 (integer)

res2 Response at assessment 2 (integer)

res3 Response at assessment 3 (integer)

Details

The dataset name has been kept as 'Skin_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the ForCausality package and assists users in identifying its specific characteristics. The suffix 'df' indicates that the dataset is a data frame. The original content has not been modified in any way.

Source

Data taken from the sanon package version 1.6

`SmokeH_df` *Smoking and Homocysteine Data*

Description

This dataset, SmokeH_df, is a data frame containing information on smoking, homocysteine levels, demographics, and socioeconomic indicators. The dataset consists of 2,475 observations and 15 variables, including biomarkers, smoking-related measures, age, education, and poverty ratio. Some observations contain missing values.

Usage

```
data(SmokeH_df)
```

Format

A data frame with 2,475 observations and 15 variables:

SEQN Participant identifier (integer)

homocysteine Homocysteine level (numeric)

z Z score indicator (integer)

female Sex indicator (integer, 1 = female, 0 = male)

age Age in years (integer)

education Education level (integer code)

povertyr Poverty ratio (numeric)

bmi Body mass index (numeric)

cotinine Cotinine level (numeric)

st Smoking type indicator (integer)

stf Smoking type (character string)

age3 Age category (integer code)

ed3 Education category (integer code)

bmi3 BMI category (integer code)

pov2 Poverty category (logical)

Details

The dataset name has been kept as 'SmokeH_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the ForCausality package and assists users in identifying its specific characteristics. The suffix 'df' indicates that the dataset is a data frame. The original content has not been modified in any way.

Source

Data taken from the evident package version 1.0.4

Stroke_df *Ischemic Stroke Case-Control Data*

Description

This dataset, Stroke_df, contains fictional case-control data for ischemic stroke, including exposures, risk factors, and confounders. The dataset includes 16,623 observations and 21 variables, covering demographic details, lifestyle factors, biomarkers, and comorbidities. Some observations contain missing values.

Usage

```
data(Stroke_df)
```

Format

A data frame with 16,623 observations and 21 variables:

regionnn7 Geographic region (factor)

case Case indicator for ischemic stroke (numeric)

esex Sex of the participant (integer)

eage Age of the participant (integer)

htnadmbp Hypertension or blood pressure measure (numeric)

nevfcur Smoking status (factor)

global_stress2 Perceived stress indicator (factor)

whrs2tert Waist-to-hip ratio tertiles (factor)

phys Physical activity indicator (factor)

alcohfreqwk Weekly alcohol consumption frequency (factor)

dmhba1c2 Diabetes / HbA1c category (factor)

cardiacrfcat Cardiac risk factor category (factor)

ahei3tert Alternative Healthy Eating Index tertiles (factor)

apob_apoatert ApoB/ApoA ratio tertiles (factor)

subeduc Sub-education level (factor)

moteduc Mother's education level (factor)

fatduc Father's education level (factor)

subhtn Sub-hypertension indicator (factor)

whr Waist-to-hip ratio (numeric)

apob_apoa ApoB/ApoA continuous ratio (numeric)

weights Sample weights (numeric)

Details

The dataset name has been kept as 'Stroke_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the ForCausality package and assists users in identifying its specific characteristics. The suffix 'df' indicates that the dataset is a data frame. The original content has not been modified in any way.

Source

Data taken from the causalPAF package version 1.2.5

Thiam_df *Thiamethoxam Application and Crop Yield Data*

Description

This dataset, Thiam_df, is a data frame containing information on thiamethoxam applications and crop yield measurements in squash plants. The dataset consists of 165 observations and 11 variables, including treatment types, plant variety, replication, fruit counts, yield measures, and defoliation indicators. Some observations may contain missing values.

Usage

```
data(Thiam_df)
```

Format

A data frame with 165 observations and 11 variables:

trt Treatment type (factor)

var Plant variety (factor)

rep Replication block (factor)

fruit Number of fruits (numeric)

avg_mass Average fruit mass (numeric)

mass Total fruit mass (numeric)

yield Crop yield (numeric)

visit Pollinator visit count (numeric)

foliage Foliage measure (numeric)

scb Squash vine borer damage (numeric)

defoliation Defoliation percentage (numeric)

Details

The dataset name has been kept as 'Thiam_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the ForCausality package and assists users in identifying its specific characteristics. The suffix 'df' indicates that the dataset is a data frame. The original content has not been modified in any way.

Source

Data taken from the melt package version 1.11.4

Udca_df *Ursodeoxycholic Acid Trial Data*

Description

This dataset, Udca_df, contains data from a clinical trial of ursodeoxycholic acid (UDCA). The dataset includes 1,360 observations and 8 variables, covering treatment assignment, disease stage, bilirubin levels, risk scores, follow-up time, and outcomes. Some observations contain missing values.

Usage

```
data(Udca_df)
```

Format

A data frame with 1,360 observations and 8 variables:

id Patient identifier (integer)

trt Treatment group (integer)

stage Disease stage (integer)

bili Bilirubin level (numeric)

riskscore Calculated risk score (numeric)

futime Follow-up time in days (numeric)

status Patient status indicator (numeric)

endpoint Endpoint description (character)

Details

The dataset name has been kept as 'Udca_df' to avoid confusion with other datasets in the R ecosystem. This naming convention helps distinguish this dataset as part of the ForCausality package and assists users in identifying its specific characteristics. The suffix 'df' indicates that the dataset is a data frame. The original content has not been modified in any way.

Source

Data taken from the survival package version 3.8-3

Index