Package 'causaldata'

May 8, 2026

Type Package

Title Example Data Sets for Causal Inference Textbooks

Version 0.1.4

Description Example data sets to run the example problems from causal inference textbooks. Currently, contains data sets for Huntington-Klein, Nick (2021 and 2025) ``The Effect'' <https://theeffectbook.net>, first and second edition, Cunningham, Scott (2021 and 2025, ISBN-13: 978-0-300-25168-5) ``Causal Inference: The Mixtape'', and Hernán, Miguel and James Robins (2020) ``Causal Inference: What If'' <https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/>.

License MIT + file LICENSE

Depends R (>= 2.10)

Imports tibble

Encoding UTF-8

LazyData true

RoxygenNote 7.3.2

URL https://github.com/NickCH-K/causaldata

BugReports https://github.com/NickCH-K/causaldata/issues

NeedsCompilation no

Author Nick Huntington-Klein [aut, cre] (ORCID: <https://orcid.org/0000-0002-7352-3991>), Malcolm Barrett [aut] (ORCID: <https://orcid.org/0000-0003-0299-5825>)

Maintainer Nick Huntington-Klein <nhuntington-klein@seattleu.edu>

Repository CRAN

Date/Publication 2024-10-24 20:40:02 UTC

Contents

`abortion` *Data on abortion legalization and sexually transmitted infections*

Description

This data looks at the effect of abortion legalization on the incidence of gonnorhea among 15-19 year olds, as a measure of risky behavior. Treatment is whether abortion is legalized at the time that the eventual 15-19 year olds are born.

Usage

```
abortion
```

Format

A data frame with 19584 rows and 22 variables

fip State FIPS code

age Age in years

race Race - 1 = white, 2 = black

year Year

t Year but counted on a different scale

sex Sex: 1 = male, 2 = female

totpop Total population

ir Incarcerated Males per 100,000

crack Crack index

alcohol Alcohol consumption per capita

income Real income per capita

ur State unemployment rate

poverty Poverty rate

repeal In a state with an early repeal of abortion prohibition

acc AIDS mortality per 100,000 cumulative in t, t-1, t-2, t-3

wht White Indicator

male Male Indicator

lnr Logged gonnorhea cases per 100,000 in 15-19 year olds

younger From the younger group

fa State-younger interaction

pi Parental involvement law in effect

bf15 Is a black female in the 15-19 age group

Details

This data is used in the *Difference-in-Differences* chapter of *Causal Inference: The Mixtape* by Cunningham.

Source

Cunningham, Scott, and Christopher Cornwell. 2013. "The Long-Run Effect of Abortion on Sexually Transmitted Infections." American Law and Economics Review 15 (1): 381–407.

References

Cunningham. 2021. Causal Inference: The Mixtape. Yale Press. https://mixtape.scunning.com/index.html.

`adult_services` *Data from a survey of internet-mediated sex workers*

Description

This data comes from a survey of 700 internet-mediated sex workers in 2008 and 2009, asking the same sex workers standard labor market information over several time periods.

Usage

```
adult_services
```

Format

A data frame with 1787 rows and 31 variables

id Provider identifier

session Client session identifier

age Age of provider

age_cl Age of Client

appearance_cl Client Attractiveness (Scale of 1 to 10)

bmi Body Mass Index

schooling Imputed Years of Schooling

asq_cl Age of Client Squared

provider_second Second Provider Involved

asian_cl Asian Client

black_cl Black Client

hispanic_cl Hispanic Client

othrace_cl Other Ethnicity Client

reg Client was a Regular

hot Met Client in Hotel

massage_cl Gave Client a Massage

lnw Log of Hourly Wage

llength Ln(Length)

unsafe Unprotected sex with client of any kind

asian race==1. Asian

black race==2. Black

hispanic race==3. Hispanic

other race==4. Other

white race==5. White

asq Age of provider squared

cohab ms==Cohabitating (living with a partner) but unmarried

married ms==Currently married and living with your spouse

divorced ms==Divorced and not remarried

separated ms==Married but not currently living with your spouse

nevermarried ms==Single and never married

widowed ms==Widowed and not remarried

Details

This data is used in the *Panel Data* chapter of *Causal Inference: The Mixtape* by Cunningham.

Source

Cunningham, Scott, and Todd D. Kendall. 2011. "Prostitution 2.0: The Changing Face of Sex Work." Journal of Urban Economics 69: 273–87.

Cunningham, Scott, and Todd D. Kendall. 2014. "Examining the Role of Client Reviews and Reputation Within Online Prostitution." In, edited by Scott Cunningham and Manisha Shah. Vol. Handbook on the Economics of Prostitution. Oxford University Press.

Cunningham, Scott, and Todd D. Kendall. 2016. "Prostitution Labor Supply and Education." Review of Economics of the Household. Forthcoming.

References

Cunningham. 2021. Causal Inference: The Mixtape. Yale Press. https://mixtape.scunning.com/index.html.

`auto` *Automobile data from Stata*

Description

This data, which comes standard in Stata, originally came from the April 1979 issue of Consumer Reports and from the United States Government EPA statistics on fuel consumption; they were compiled and published by Chambers et al. (1983).

Usage

```
auto
```

Format

A data frame with 74 rows and 12 variables

make Make and Model

price Price

mpg Mileage (mpg)

rep78 Repair Record 1978

headroom Headroom (in.)

trunk Trunk space (cu. ft.)

weight Weight (lbs.)

length Length (in.)

turn Turn Circle (ft.)

displacement Displacement (cu. in.)

gear_ratio Gear Ratio

foreign Car type; 0 = Domestic, 1 = Foreign

Details

This data is used in the *Probability and Regression Review* chapter of *Causal Inference: The Mixtape*.

Source

Chambers, J. M., W. S. Cleveland, B. Kleiner, and P. A. Tukey. 1983. Graphical Methods for Data Analysis. Belmont, CA: Wadsworth.

References

Cunningham. 2021. Causal Inference: The Mixtape. Yale Press. https://mixtape.scunning.com/index.html.

---

avocado *Data on avocado sales*

Description

This data set includes information on the average price and total amount of avocados sold across 169 weeks from 2015 to 2018. This data covers only sales of 'conventional' avocados that take place in California.

Usage

```
avocado
```

Format

A data frame with 169 rows and 3 variables:

Date Date of observation

AveragePrice Average avocado price

TotalVolume Total volume of avocados sold

Details

This data was used in the *Identification* chapter of The Effect by Huntington-Klein

Source

Kiggins, Justin. 2018. https://www.kaggle.com/neuromusic/avocado-prices/

References

Huntington-Klein. 2021. The Effect: An Introduction to Research Design and Causality. https://theeffectbook.net.

`black_politicians` *Data from "Black Politicians are More Intrinsically Motivated to Advance Blacks' Interests"*

Description

The `black_politicians` data contains data from Broockman (2013) on a field experiment where the author sent fictional emails purportedly sent by Black people to legislators in the United States. The experiment sought to determine whether the effect of the email being from "out-of-district" (someone who can't vote for you and so provides no extrinsic motivation to reply) would have a smaller effect on response rates for Black legislators than for non-Black ones, providing evidence of additional intrinsic motivation on the part of Black legislators to help Black people.

Usage

```
black_politicians
```

Format

A data frame with 5593 rows and 14 variables

leg_black Legislator receiving email is Black

treat_out Email is from out-of-district

responded Legislator responded to email

totalpop District population

medianhhincom District median household income

black_medianhh District median household income among Black people

white_medianhh District median household income among White people

blackpercent Percentage of district that is Black

statessquireindex State's Squire index

nonblacknonwhite Legislator receiving email is neither Black nor White

urbanpercent Percentage of district that is urban

leg_senator Legislator receiving email is a senator

leg_democrat Legislator receiving email is in the Democratic party

south Legislator receiving email is in the Southern United States

Details

This data is used in the *Matching* chapter of *The Effect*.

Source

Broockman, D.E., 2013. Black politicians are more intrinsically motivated to advance blacks' interests: A field experiment manipulating political incentives. *American Journal of Political Science*, 57(3), pp.521-536.

References

Huntington-Klein. 2021. The Effect: An Introduction to Research Design and Causality. https://theeffectbook.net.

`castle` *Data on castle-doctrine statutes and violent crime*

Description

This data looks at the impact of castle-doctrine statutes on violent crime. Data from the FBI Uniform Crime Reports Summary files are combined with information on castle-doctrine/stand-your-ground law impementation in different states.

Usage

```
castle
```

Format

A data frame with 19584 rows and 22 variables

year Year

post After-treatment

sid state id

robbery_gun_r Region-quarter fixed effects

jhcitizen_c justifiable homicide by private citizen count

jhpolice_c justifiable homicide by police count

homicide homicide count per 100,000 state population

robbery Region-quarter fixed effects

assault aggravated assault count per 100,000 state population

burglary burglary count per 100,000 state population

larceny larceny count per 100,000 state population

motor motor vehicle theft count per 100,000 state population

murder murder count per 100,000 state population

unemployrt unemployment rate

blackm_15_24 % of black male aged 15-24

whitem_15_24 % of white male aged 15-24

blackm_25_44 % of black male aged 25-44

whitem_25_44 % of white male aged 25-44

poverty poverty rate

l_homicide Logged crime rate

l_larceny Logged crime rate

l_motor Logged crime rate

l_police Logged police presence

l_income Logged income

l_prisoner Logged number of prisoners

l_lagprisoner Lagged log prisoners

l_exp_subsidy Logged subsidy spending

l_exp_pubwelfare Logged public welfare spending

lead1,lead2,lead3,lead4,lead5,lead6,lead7,lead8,lead9,lag0,lag1,lag2,lag3,lag4,lag5 Indicators of how many time periods until/since treatment

popwt Population weight

r20001,r20002,r20003,r20004,r20011,r20012,r20013,r20014,r20021,r20022,r20023,r20024,r20031,r20032,r20033,r200 Region-quarter fixed effects

trend_1,trend_10,trend_11,trend_12,trend_13,trend_14,trend_15,trend_16,trend_17,trend_18,trend_19,trend_2,tren State linear time trends

Details

This data is used in the *Difference-in-Differences* chapter of *Causal Inference: The Mixtape* by Cunningham.

Source

Cheng, Cheng, and Mark Hoekstra. 2013. "Does Strengthening Self-Defense Law Deter Crime or Escalate Violence? Evidence from Expansions to Castle Doctrine." Journal of Human Resources 48 (3): 821–54.

References

Cunningham. 2021. Causal Inference: The Mixtape. Yale Press. https://mixtape.scunning.com/index.html.

---

ccdrug *Data on Drug Arrests from the Crown Court Sentencing Survey*

Description

The ccdrug data contains data on drug arrests from the Crown Court Sentencing Survey between 2012 and 2015 in England and Wales, allowing for a look at differential sentencing rates for men and women, with a set of controls for features that should impact sentencing.

Usage

```
ccdrug
```

Format

A data frame with 16973 rows and 45 variables

custody Taken in to custody.

male Is a male

first_offense This is the first offense

age Age in ten-year bins

offense Offense type

prev_convictions Previous convictions, in bins of None, 1-3, 4-9, or 10+

drg_class Type of drug

drg_culpability Level of culpability for crime

drg_increasing_ser_other_1, drg_increasing_ser_other_3, drg_increasing_ser_other_4, drg_increasing_ser_other_5, A set of indicators that should increase or reduce the likelihood of being taken into custody. See variable labels for specific definitions.

Details

This data set is used in the *Partial Identification* chapter of *The Effect*.

Source

Pina Sanchez, J., & Harris, L., 2020. Sentencing gender? Investigating the presence of gender disparities in Crown Court sentences. *Criminal Law Review*, 2020(1), pp. 3-28.

References

Huntington-Klein. 2021. The Effect: An Introduction to Research Design and Causality. https://theeffectbook.net.

`close_college` *Data from Card (1995) to estimate the effect of college education on earnings*

Description

Data from the National Longitudinal Survey Young Men Cohort. This data is used to estimate the effect of college education on earnings, using the presence of a nearby (in-county) college as an instrument for college attendance.

Usage

```
close_college
```

Format

A data frame with 3010 rows and 8 variables

lwage Log wages

educ Years of education

exper Years of work experience

black Race: Black

south In the southern United States

married Is married

smsa In a Standard Metropolitan Statistical Area (urban)

nearc4 There is a four-year college in the county

Details

This data is used in the *Instrumental Variables* chapter of *Causal Inference: The Mixtape* by Cunningham.

Source

Card, David. 1995. "Aspects of Labour Economics: Essays in Honour of John Vanderkamp." In. University of Toronto Press.

References

Cunningham. 2021. Causal Inference: The Mixtape. Yale Press. https://mixtape.scunning.com/index.html.

`close_elections_lmb` *A close-elections regression discontinuity study from Lee, Moretti, and Butler (2004)*

Description

This data comes from a close-elections regression discontinuity study from Lee, Moretti, and Butler (2004). The design is intended to test convergence and divergence in policy. Major effects of electing someone from a particular party on policy outcomes *in a close race* indicates that the victor does what they want. Small or null effects indicate that the electee moderates their position towards their nearly-split electorate.

Usage

```
close_elections_lmb
```

Format

A data frame with 13588 rows and 9 variables

state ICPSR state code

district district code

id Election ID

score ADA voting score (higher = more liberal)

year Year of election

demvoteshare Democratic share of the vote

democrat Democratic victory

lagdemocrat Lagged Democratic victory

lagdemvoteshare Lagged democratic share of the vote

Details

This data is used in the *Regression Discontinuity* chapter of *Causal Inference: The Mixtape* by Cunningham.

Source

Lee, David S., Enrico Moretti, and Matthew J. Butler. 2004. "Do Voters Affect or Elect Policies: Evidence from the U.S. House." Quarterly Journal of Economics 119 (3): 807–59.

References

Cunningham. 2021. Causal Inference: The Mixtape. Yale Press. https://mixtape.scunning.com/index.html.

---

cps_mixtape *Observational counterpart to nsw_mixtape data*

Description

Data from the Current Population Survey on participation in the National Supported Work Demonstration (NSW) job-training program experiment. This is used as an observational comparison to the NSW experimental data from the nsw_mixtape data.

Usage

```
cps_mixtape
```

Format

A data frame with 15992 rows and 11 variables

data_id Individual ID

treat In the National Supported Work Demonstration Job Training Program

age Age in years

educ Years of education

black Race: Black

hisp Ethnicity: Hispanic

marr Married

nodegree Has no degree

re74 Real earnings 1974

re75 Real earnings 1975

re78 Real earnings 1978

Details

This data is used in the *Matching and Subclassification* chapter of *Causal Inference: The Mixtape* by Cunningham.

Source

Dehejia, Rajeev H., and Sadek Wahba. 1999. “Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs.” Journal of the American Statistical Association 94 (448): 1053–62.".

References

Cunningham. 2021. Causal Inference: The Mixtape. Yale Press. https://mixtape.scunning.com/index.html.

`credit_cards` *Data on Taiwanese Credit Card Holders*

Description

Data from the UCI Machine Learning Repository on Taiwanese credit card holders, the amount of their credit card bill, and whether their payment was late.

Usage

```
credit_cards
```

Format

A data frame with 30000 rows and 4 variables

LateSept Credit card payment is late in Sept 2005

LateApril Credit card payment is late in April 2005

BillApril Total bill in April 2005 in thousands of New Taiwan Dollars

AGE Age of card-holder

Details

This data is used in the *Matching* chapter of *The Effect* by Huntington-Klein.

Source

Lichman, Moshe. 2013. UCI Machine Learning Repository.

References

Huntington-Klein. 2021. The Effect: An Introduction to Research Design and Causality. https://theeffectbook.net.

`gapminder` *Gapminder data*

Description

The `gapminder` data contains data on life expectancy and GDP per capita by country and year.

Usage

```
gapminder
```

Format

A data frame with 1704 rows and 6 variables

country The country

continent The continent the country is in

year The year data was collected. Ranges from 1952 to 2007 in increments of 5 years

lifeExp Life expectancy at birth, in years

pop Population

gdpPercap GDP per capita (US$, inflation-adjusted)

Details

This data set is the same one found in the *gapminder* package in R as of 2020. This data set is used in the *Fixed Effects* chapter of *The Effect*.

Source

https://www.gapminder.org/data/

Jennifer Bryan (2017). gapminder: Data from Gapminder. R package version 0.3.0. https://CRAN.R-project.org/package=gapminder

References

Huntington-Klein. 2021. The Effect: An Introduction to Research Design and Causality. https://theeffectbook.net.

google_stock *Google Stock Data*

Description

The google_stock data contains data on daily stock returns for Google and the S&P 500 for May through Augut 2015, centering around the August 10, 2015 announcement that Google would reorganize under parent company Alphabet.

Usage

```
google_stock
```

Format

A data frame with 84 rows and 3 variables

Date The date

Google_Return Daily GOOG Stock Return (1 = 100 percent daily return)

SP500_Return Daily S&P 500 Index Return (1 = 100 percent daily return)

Details

This data was downloaded using the *tidyquant* package, and is used in the *Event Studies* chapter of *The Effect*.

Source

Matt Dancho and Davis Vaughan (2021). tidyquant: Tidy Quantitative Financial Analysis. R package version 1.0.3. https://CRAN.R-project.org/package=tidyquant

References

Huntington-Klein. 2021. The Effect: An Introduction to Research Design and Causality. https://theeffectbook.net.

gov_transfers *Data from "Government Transfers and Political Support"*

Description

The gov_transfers data contains data from Manacorda, Miguel, and Vigorito (2011) on government transfer program that was administered based on an income cutoff. Data is pre-limited to households that were just around the income cutoff.

Usage

```
gov_transfers
```

Format

A data frame with 1948 rows and 5 variables

Income_Centered Income measure, centered around program cutoff (negative value = eligible)

Education Household average years of education among those 16+

Age Household average age

Participation Participation in transfers

Support Measure of support for the government

Details

This data is used in the *Regression Discontinuity* chapter of *The Effect*.

Source

Manacorda, M., Miguel, E. and Vigorito, A., 2011. Government transfers and political support. *American Economic Journal: Applied Economics*, 3(3), pp.1-28.

References

Huntington-Klein. 2021. The Effect: An Introduction to Research Design and Causality. https://theeffectbook.net.

gov_transfers_density *Data from "Government Transfers and Political Support" for Density Tests*

Description

The `gov_transfers_density` data contains data from Manacorda, Miguel, and Vigorito (2011) on government transfer program that was administered based on an income cutoff. As opposed to the `gov_transfers` data set, this data set only contains income information, but has a wider range of it, for use with density discontinuity tests.

Usage

```
gov_transfers_density
```

Format

A data frame with 52549 rows and 1 variable:

Income_Centered Income measure, centered around program cutoff (negative value = eligible)

Details

This data is used in the *Regression Discontinuity* chapter of *The Effect*.

Source

Manacorda, M., Miguel, E. and Vigorito, A., 2011. Government transfers and political support. *American Economic Journal: Applied Economics*, 3(3), pp.1-28.

References

Huntington-Klein. 2021. The Effect: An Introduction to Research Design and Causality. https://theeffectbook.net.

greek_data *Data from a fictional randomized heart transplant study*

Description

`greek_data` is a fictional data set from Table 2.2 in Chapter 2 of Causal Inference. From the book: "Table 2.2 shows the data from our heart transplant randomized study. Besides data on treatment A (1 if the individual received a transplant, 0 otherwise) and outcome Y (1 if the individual died, 0 otherwise), Table 2.2 also contains data on the prognostic factor L (1 if the individual was in critical condition, 0 otherwise), which we measured before treatment was assigned."

Usage

```
greek_data
```

Format

A data frame with 20 rows and 4 variables:

name The name of a Greek god

l A prognostic factor

a The treatment, a heart transplant

y The outcome, death

Source

Hernán and Robins. Causal Inference. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

---

mortgages *Data from "How do Mortgage Subsidies Affect Home Ownership? Evidence from the Mid-Century GI Bills"*

Description

The `mortgages` data contains data from Fetter (2015) on home ownership rates by men, focusing on whether they were born at the right time to be eligible for mortgage subsidies based on their military service.

Usage

```
mortgages
```

Format

A data frame with 214144 rows and 6 variables

bpl Birth State

qob Quarter of birth

nonwhite White/nonwhite race indicator. 1 = Nonwhite

vet_wwko Veteran of either the Korean war or World War II

home_ownership Owns a home

qob_minus_kw Quarter of birth centered on eligibility for mortgage subsidy (0+ = eligible)

Details

This data is used in the *Regression Discontinuity* chapter of *The Effect*.

Source

Fetter, D.K., 2013. How do mortgage subsidies affect home ownership? Evidence from the mid-century GI bills. *American Economic Journal: Economic Policy*, 5(2), pp.111-47.

References

Huntington-Klein. 2021. The Effect: An Introduction to Research Design and Causality. https://theeffectbook.net.

`Mroz` *U.S. Women's Labor-Force Participation*

Description

The `Mroz` data frame has 753 rows and 8 columns. The observations, from the Panel Study of Income Dynamics (PSID), are married women.

Usage

```
Mroz
```

Format

A data frame with 753 rows and 8 variables

lfp Labor-force participation

k5 Number of children 5 years old or younger

k618 Number of children 6 to 17 years old

age Age in years

wc Wife attended college

hc Husband attended college

lwg Log expected wage rate. For women in the labor force, the actual wage rate; for women not in the labor force, an imputed value based on the regression of lwg on the other variables.

inc Family income exclusive of wife's income

Details

This data set is a lightly edited version of the one found in the *carData* package in R. It is used in the Describing Relationships chapter of *The Effect*.

Source

Mroz, T. A. (1987) The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica* 55, 765–799.

John Fox, Sanford Weisberg and Brad Price (2020). carData: Companion to Applied Regression Data Sets. R package version 3.0-4. https://CRAN.R-project.org/package=carData

References

Fox, J. (2016) *Applied Regression Analysis and Generalized Linear Models,* Third Edition. Sage.

Fox, J. (2000) *Multiple and Generalized Nonparametric Regression.* Sage.

Fox, J. and Weisberg, S. (2019) *An R Companion to Applied Regression.* Third Edition, Sage.

Long. J. S. (1997) *Regression Models for Categorical and Limited Dependent Variables.* Sage.

Huntington-Klein. 2021. The Effect: An Introduction to Research Design and Causality. https://theeffectbook.net.

`nhefs` *National Health and Nutrition Examination Survey Data I Epidemiologic Follow-up Study*

Description

`nhefs` is a cleaned data set of the data used in Causal Inference by Hernán and Robins. `nhefs` is dataset containing data from the National Health and Nutrition Examination Survey Data I Epidemiologic Follow-up Study (NHEFS). The NHEFS was jointly initiated by the National Center for Health Statistics and the National Institute on Aging in collaboration with other agencies of the United States Public Health Service. A detailed description of the NHEFS, together with publicly available data sets and documentation, can be found at https://wwwn.cdc.gov/nchs/nhanes/nhefs/.

Usage

```
nhefs
```

Format

A data frame with 1629 rows and 67 variables. The codebook is available as `nhefs_codebook`.

Source

https://wwwn.cdc.gov/nchs/nhanes/nhefs/

References

Hernán and Robins. Causal Inference. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

nhefs_codebook *NHEFS Codebook*

Description

nhefs_codebook is the codebook for nhefs and nhefs_complete.

Usage

```
nhefs_codebook
```

Format

A data frame with 64 rows and 2 variables.

variable The variable being described

description The variable description

Source

https://wwwn.cdc.gov/nchs/nhanes/nhefs/

References

Hernán and Robins. Causal Inference. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

nhefs_complete *Complete-Data National Health and Nutrition Examination Survey Data I Epidemiologic Follow-up Study*

Description

nhefs_complete is the same as nhefs, but only participants with complete data are included. The variables that need to be complete to be included are: qsmk, sex, race, age, school, smokeintensity, smokeyrs, exercise, active, wt71, wt82, and wt82_71.

Usage

```
nhefs_complete
```

Format

A data frame with 1556 rows and 67 variables. The codebook is available as nhefs_codebook.

Source

https://wwwn.cdc.gov/nchs/nhanes/nhefs/

References

Hernán and Robins. Causal Inference. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

`nsw_mixtape` *Data from the National Supported Work Demonstration (NSW) job-training program*

Description

Data from the National Supported Work Demonstration (NSW) job-training program experiment, where those treated were guaranteed a job for 9-18 months.

Usage

```
nsw_mixtape
```

Format

A data frame with 445 rows and 11 variables

data_id Individual ID

treat In the National Supported Work Demonstration Job Training Program

age Age in years

educ Years of education

black Race: Black

hisp Ethnicity: Hispanic

marr Married

nodegree Has no degree

re74 Real earnings 1974

re75 Real earnings 1975

re78 Real earnings 1978

Details

This data is used in the *Matching and Subclassification* chapter of *Causal Inference: The Mixtape* by Cunningham.

Source

Lalonde, Robert. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." American Economic Review 76 (4): 604–20.

Dehejia, Rajeev H., and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." Journal of the American Statistical Association 94 (448): 1053–62.".

References

Cunningham. 2021. Causal Inference: The Mixtape. Yale Press. https://mixtape.scunning.com/index.html.

`organ_donations` *Organ Donation Data*

Description

The `organ_donation` data contains data from Kessler and Roth (2014) on organ donation rates by state and quarter. The state of California enacted an active-choice phrasing for their organ donation sign-up questoin in Q32011. The only states included in the data are California and those that can serve as valid controls; see Kessler and Roth (2014).

Usage

```
organ_donations
```

Format

A data frame with 162 rows and 3 variables

State The state, where California is the Treated group

Quarter Quarter of observation, in "Q"QYYYY format

Rate Organ donation rate

Quarter_Num Quarter of observation in numerical format. 1 = Quarter 4, 2010

Details

This data is used in the *Difference-in-Differences* chapter of *The Effect*.

Source

Kessler, J.B. and Roth, A.E., 2014. Don't take 'no' for an answer: An experiment with actual organ donor registrations. National Bureau of Economic Research working paper No. 20378. https://www.nber.org/papers/w20378

References

Huntington-Klein. 2021. The Effect: An Introduction to Research Design and Causality. https://theeffectbook.net.

`restaurant_inspections` *Data on Restaurant Inspections*

Description

The `restaurant_inspections` data contains data on restaurant health inspections performed in Anchorage, Alaska.

Usage

```
restaurant_inspections
```

Format

A data frame with 27178 rows and 5 variables

business_name Name of restaurant/chain

inspection_score Health Inspection Score

Year Year of inspection

NumberofLocations Number of locations in restaurant chain

Weekend Was the inspection performed on a weekend?

Details

This data set is used in the *Regression* chapter of *The Effect*.

Source

Camus, Louis-Ashley. 2020. https://www.kaggle.com/loulouashley/inspection-score-restaurant-inspection

References

Huntington-Klein. 2021. The Effect: An Introduction to Research Design and Causality. https://theeffectbook.net.

`ri` *A simple simulated data set for calculating p-values*

Description

This simulated data allows for a quick and easy calculation of a p-value using randomization inference.

Usage

```
ri
```

Format

A data frame with 8 rows and 5 variables

name Fictional Name

d Treatment

y Outcome

y0 Outcome if untreated

y1 Outcome if treated

Details

This data is used in the *Potential Outcomes Causal Model* chapter of *Causal Inference: The Mixtape* by Cunningham.

Source

Cunningham. 2021. Causal Inference: The Mixtape. Yale Press. https://mixtape.scunning.com/index.html.

References

Cunningham. 2021. Causal Inference: The Mixtape. Yale Press. https://mixtape.scunning.com/index.html.

`scorecard` *Earnings and Loan Repayment in US Four-Year Colleges*

Description

From the College Scorecard, this data set contains by-college-by-year data on how students who attended those colleges are doing.

Usage

```
scorecard
```

Format

A data frame with 48,445 rows and 8 variables:

unitid College identifiers

inst_name Name of the college or university

state_abbr Two-letter abbreviation for the state the college is in

pred_degree_awarded_ipeds Predominant degree awarded. 1 = less-than-two-year, 2 = two-year, 3 = four-year+

year Year in which outcomes are measured

earnings_med Median earnings among students (a) who received federal financial aid, (b) who began as undergraduates at the institution ten years prior, (c) with positive yearly earnings

count_not_working Number of students who are (a) not working (not necessarily unemployed), (b) received federal financial aid, and (c) who began as undergraduates at the institution ten years prior

count_working Number of students who are (a) working, (b) who received federal financial aid, and (c) who began as undergraduates at the institution ten years prior

Details

This data is not just limited to four-year colleges and includes a very wide variety of institutions.

Note that the labor market (earnings, working) and repayment rate data do not refer to the same cohort of students, but rather are matched on the year in which outcomes are recorded. Labor market data refers to cohorts beginning college as undergraduates ten years prior, repayment rate data refers to cohorts entering repayment seven years prior.

Data was downloaded using the Urban Institute's `educationdata` package.

This data was used in the *Describing Variables* chapter of The Effect by Huntington-Klein

Source

Education Data Portal (Version 0.4.0 - Beta), Urban Institute, Center on Education Data and Policy, accessed June 28, 2019. https://educationdata.urban.org/documentation/, Scorecard.

References

Huntington-Klein. 2021. The Effect: An Introduction to Research Design and Causality. https://theeffectbook.net.

`snow` *Data from John Snow's 1855 study of the cause of cholera*

Description

A subset of the aggregated death rate data from Snow's legendary study of the source of the London Cholera outbreak.

Usage

```
snow
```

Format

A data frame with 4 rows and 4 variables

year Year

supplier Water pump supplier

treatment Status of water pump

deathrate Deaths per 10k 1851 population

Details

This data is used in the *Difference-in-Differences* chapter of *The Effect* by Huntington-Klein.

Source

Snow, John. 1855. 'On the Mode of Communication of Cholera'. John Churchill."

Coleman, Thomas. 2019. 'Causality in the time of cholera: John Snow as a prototype for causal inference.' SSRN 3262234."

References

Huntington-Klein. 2021. The Effect: An Introduction to Research Design and Causality. https://theeffectbook.net.

`social_insure` *Data from "Social Networks and the Decision to Insure"*

Description

The `social_insure` data contains data from Jai, De Janvry, and Saoudlet (2015) on a two-round social network-based experiment on getting farmers to get insurance. See the paper for more details.

Usage

```
social_insure
```

Format

A data frame with 1410 rows and 13 variables

address Natural village

village Administrative village

takeup_survey Whether farmer ended up purchasing insurance. (1 = yes)

age Household Characteristics - Age

agpop Household Characteristics - Household Size

ricearea_2010 Area of Rice Production

disaster_prob Perceived Probability of Disasters Next Year

male Household Caracteristics: Gender of Household Head (1 = male)

default "Default option" in experimental format assigned to. (1 = default is to buy, 0 = default is to not buy)

intensive Whether or not was assigned to "intensive" experimental session (1 = yes)

risk_averse Risk aversion measurement

literacy 1 = literate, 0 = illiterate

pre_takeup_rate Takeup rate prior to experiment

Details

This data is used in the *Instrumental Variables* chapter of *The Effect*.

Source

Cai, J., De Janvry, A. and Sadoulet, E., 2015. Social networks and the decision to insure. *American Economic Journal: Applied Economics*, 7(2), pp.81-108.

References

Huntington-Klein. 2021. The Effect: An Introduction to Research Design and Causality. https://theeffectbook.net.

texas *Data on prison capacity expansion in Texas*

Description

This data looks at the massive expansion in prison capacity in Texas that occurred in 1993 under Governor Ann Richards, and the effect of that expansion on the number of Black men in prison.

Usage

```
texas
```

Format

A data frame with 816 rows and 12 variables

statefip State FIPS code

year Year

bmprison Number of Black men in prison

wmprison Number of White men in prison

alcohol Alcohol consumption per capita

income Median income

ur Unemployment rate

poverty Poverty rate

black Percentage of the population that is Black

perc1519 Percentage of the population that is age 15-19

aidscapita AIDS mortality per 100,000 in t

state State name

Details

This data is used in the *Synthetic Control* chapter of *Causal Inference: The Mixtape* by Cunningham.

Source

Cunningham and Kang. 2019. "Studying the Effect of Incarceration Shocks to Drug Markets." Unpublished manuscript. http://www.scunning.com/files/mass_incarceration_and_drug_abuse.pdf

References

Cunningham. 2021. Causal Inference: The Mixtape. Yale Press. https://mixtape.scunning.com/index.html.

thornton_hiv *Data from HIV information experiment in Thornton (2008)*

Description

thornton_hiv comes from an experiment in Malawi looking at whether cash incentives could encourage people to learn the results of their HIV tests.

Usage

```
thornton_hiv
```

Format

A data frame with 4820 rows and 7 variables

villnum Village ID

got Got HIV results

distvct Distance in kilometers

tinc Total incentive

any Received any incentive

age Age

hiv2004 HIV results

Details

This data is used in the Potential Outcomes Causal Model chapter of Causal Inference: The Mixtape by Cunningham.

Source

Thornton, Rebecca L. 2008. 'The Demand for, and Impact of, Learning Hiv Status.' American Economic Review 98 (5): 1829–63.

References

Cunningham. 2021. Causal Inference: The Mixtape. Yale Press. https://mixtape.scunning.com/index.html.

`titanic` *Data from the sinking of the Titanic*

Description

`titanic` comes from the sinking of the Titanic, and can be used to look at survival by different demographic characteristics.

Usage

```
titanic
```

Format

A data frame with 4820 rows and 7 variables

class class (ticket)

age Age (Child vs. Adult)

sex Gender

survived Survived

Details

This data is used in the Matching and Subclassification chapter of Causal Inference: The Mixtape by Cunningham.

Source

British Board of Trade (1990), Report on the Loss of the 'Titanic' (S.S.). British Board of Trade Inquiry Report (reprint). Gloucester, UK: Allan Sutton Publishing.

References

Cunningham. 2021. Causal Inference: The Mixtape. Yale Press. https://mixtape.scunning.com/index.html.

`training_bias_reduction` *Simulated data from a job training program for a bias reduction method*

Description

This simulated data is used to demonstrate the bias-reduction method in matching as per Abadie and Imbens (2011).

Usage

```
training_bias_reduction
```

Format

A data frame with 8 rows and 4 variables

Unit Unit ID

Y Outcome

D Treatment

X Matching variable

Details

This data is used in the *Matching and Subclassification* chapter of *Causal Inference: The Mixtape* by Cunningham.

Source

Cunningham. 2021. Causal Inference: The Mixtape. Yale Press. https://mixtape.scunning.com/index.html.

References

Cunningham. 2021. Causal Inference: The Mixtape. Yale Press. https://mixtape.scunning.com/index.html.

`training_example` *Simulated data from a job training program*

Description

This simulated data, which is presented in the form of a full results, table, is used to demonstrate a matching procedure.

Usage

```
training_example
```

Format

A data frame with 25 rows and 9 variables

unit_treat Unit ID for treated observations

age_treat age for treated observations

earnings_treat earnings for treated observations

unit_control Unit ID for control observations

age_control age for control observations

earnings_control earnings for control observations

unit_matched Unit ID for matched controls

age_matched age for matched controls

earnings_matched earnings for matched controls

Details

This data is used in the *Matching and Subclassification* chapter of *Causal Inference: The Mixtape* by Cunningham.

Source

Cunningham. 2021. Causal Inference: The Mixtape. Yale Press. https://mixtape.scunning.com/index.html.

References

Cunningham. 2021. Causal Inference: The Mixtape. Yale Press. https://mixtape.scunning.com/index.html.

yule *Data on 19th century English Poverty from Yule (1899)*

Description

yule allows for a look at the correlation between poverty relief and poverty rates in England in the 19th century.

Usage

```
yule
```

Format

A data frame with 32 rows and 5 variables

location Location in England

paup Pauperism Growth

outrelief Poverty Relief Growth

old Annual growth in aged population

pop Annual growth in population

Details

This data is used in the Potential Outcomes Causal Model chapter of Causal Inference: The Mixtape by Cunningham.

Source

Yule, G. Udny. 1899. 'An Investigation into the Causes of Changes in Pauperism in England, Chiefly During the Last Two Interensal Decades.' Journal of Royal Statistical Society 62: 249–95.

References

Cunningham. 2021. Causal Inference: The Mixtape. Yale Press. https://mixtape.scunning.com/index.html.

Index