

# Package ‘stevedata’

October 25, 2023

**Type** Package

**Title** Steve's Toy Data for Teaching About a Variety of Methodological, Social, and Political Topics

**Depends** R (>= 3.5.0)

**Version** 1.1.0

**Maintainer** Steve Miller <steven.v.miller@gmail.com>

**Description** This is a collection of various kinds of data with broad uses for teaching. My students, and academics like me who teach the same topics I teach, should find this useful if their teaching workflow is also built around the R programming language. The applications are multiple but mostly cluster on topics of statistical methodology, international relations, and political economy.

**License** GPL-2

**Encoding** UTF-8

**LazyData** true

**LazyDataCompression** xz

**RoxygenNote** 7.2.3

**URL** <http://svmiller.com/stevedata/>

**BugReports** <https://github.com/svmiller/stevedata/issues/>

**Suggests** knitr, rmarkdown, tibble, tools, testthat

**NeedsCompilation** no

**Author** Steve Miller [aut, cre] (<<https://orcid.org/0000-0003-4072-6263>>)

**Repository** CRAN

**Date/Publication** 2023-10-25 11:20:02 UTC

## R topics documented:

af_crime93	3
aluminum_premiums	4
anes_partytherms	5

anes_prochoice	6
anes_vote84	7
Arca	8
arcticseaice	9
arg_tariff	10
asn_stats	10
CFT15	11
clemson_temps	12
co2emissions	13
coffee_imports	14
coffee_price	15
commodity_prices	15
CP77	17
Datasaurus	17
Dee04	18
DJIA	19
DST	20
eight_schools	21
election_turnout	21
eq_passengercars	22
ESS10NO	23
ESS9GB	24
ESSBE5	26
eurostat_codes	27
eustates	27
fakeAPI	28
fakeHappiness	29
fakeLogit	30
fakeTSCS	30
fakeTSD	31
ghp100k	32
GHR04	33
gss_abortion	34
gss_spending	35
gss_wages	37
Guber99	38
illiteracy30	39
inglehart03	40
Lipset59	41
LOTI	42
LTPT	43
LTWT	44
min_wage	44
mm_mlda	45
mm_nhis	46
mm_randhie	47
mvprod	48
nesarc_drinkspd	49

Newhouse77 . . . . .	50
ODGI . . . . .	50
OODTPT . . . . .	51
PPGE . . . . .	52
PRDEG . . . . .	54
Presidents . . . . .	55
pwt_sample . . . . .	55
quartets . . . . .	56
recessions . . . . .	57
SBCD . . . . .	58
scb_regions . . . . .	58
SCP16 . . . . .	59
sealevels . . . . .	60
so2concentrations . . . . .	61
steves_clothes . . . . .	61
sugar_price . . . . .	62
sweden_counties . . . . .	63
thatcher_approval . . . . .	63
therms . . . . .	64
turnips . . . . .	65
TV16 . . . . .	66
ukg_eeri . . . . .	67
uniondensity . . . . .	68
usa_chn_gdp_forecasts . . . . .	69
usa_computers . . . . .	70
usa_migration . . . . .	71
usa_states . . . . .	71
usa_trade_gdp . . . . .	72
voteincome . . . . .	73
wbd_example . . . . .	73
wvs_ccodes . . . . .	74
wvs_immig . . . . .	75
wvs_justifbribe . . . . .	75
wvs_usa_abortion . . . . .	76
wvs_usa_educat . . . . .	77
wvs_usa_regions . . . . .	78
yugo_sales . . . . .	79
<b>Index</b>	<b>80</b>

af\_crime93

*Statewide Crime Data (1993)***Description**

These data are in Table 9.1 of the 3rd edition of Agresti and Finlay's *Statistical Methods for the Social Sciences*. The data are from *Statistical Abstract of the United States* and most variables were measured in 1993.

**Usage**

af\_crime93

**Format**

A data frame with 51 observations on the following 8 variables.

state a character vector for the state

violent a numeric vector for the violent crime rate (per 100,000 people in population)

murder a numeric vector for the murder rate (per 100,000 people in population)

poverty a numeric vector for the percent with income below the poverty level

single a numeric vector for the percent of families headed by a single parent

metro a numeric vector for the percent of population in metropolitan areas

white a numeric vector for the percentage of the state that is white

highschool a numeric vector for the percent of state that graduated from high school

**Details**

The data are from Statistical Abstract of the United States and most variables were measured in 1993. These data should result in regressions that would flunk a Breusch-Pagan test for heteroskedasticity.

**References**

Agresti, Alan and Barbara Finley. 1997. *Statistical Methods for the Social Sciences*. Prentice Hall. (3rd Edition)

---

aluminum\_premiums      *LME Aluminum Premiums Data*

---

**Description**

A near daily data set on the price of aluminum premiums (USD/MT) for LME in the U.S., Western Europe, East Asia, and Southeast Asia. I like these data as illustrative of some of the shortsightedness of the aluminum tariffs that Donald Trump announced in March 2018. The tariffs had no discernible effect on manufacturing employment or earnings, but they created a supply shock that made aluminum more expensive.

**Usage**

aluminum\_premiums

**Format**

A data frame with 2,812 observations on the following 3 variables.

date a date

group a factor with levels of East Asia, Southeast Asia, United States, and Western Europe

price a numeric vector for the price of the LME aluminum premium

**Details**

LME aluminum premiums (monthly contracts going out to 15 months) work alongside LME aluminum contracts to allow market participants to hedge the all-in price and physically deliver or receive premium aluminum warrants in non-queued LME premium warehouses.

---

anes_partytherms	<i>Major Party (Democrat, Republican) Thermometer Index Data (1978-2012)</i>
------------------	--

---

**Description**

A data set on thermometer ratings for the Democratic party, Republican party, "both major parties", and a major party thermometer index from the American National Election Studies (1978-2012).

**Usage**

anes\_partytherms

**Format**

A data frame with 33830 observations on the following 19 variables.

year the survey year

uid a unique identifier for each respondent, taken directly from the time-series files for potential merging

stateabb the two-character abbreviation for the state of residence for the respondent

therm\_dem the respondent's thermometer rating of the Democratic party

therm\_gop the respondent's thermometer rating of the Republican party

therm\_bmp the respondent's thermometer rating of "both major parties"

mpti the "major party thermometer index" score for the respondent. See details for more.

age the age of the respondent

educat the education-level of the respondent. 1 = 8 grades or less. 2 = high school, no diploma. 3 = high school diploma. 4 = high school "plus non-academic training". 5 = Some college, no degree (includes AA holders). 6 = BA-level degree. 7 = advanced degree, including Bachelor of Laws degrees.

urbanism 1 = central cities. 2 = suburban areas. 3 = rural/small towns/outlying areas.

- pid7 1 = Strong Democrat. 2 = Weak Democrat. 3 = Independent, lean Democrat. 4 = Independent. 5 = Independent, lean Republican. 6 = Weak Republican. 7 = Strong Republican
- incomeperc respondent's household income percentile. 1 = 0-16 percentile. 2 = 17-33. 3 = 34-67. 4 = 68-95. 5 = 96-100.
- race4 respondent's race-ethnicity summary. 1 = White, non-hispanic. 2 = Black, non-hispanic. 3 = Hispanic. 4 = Other.
- unemployed a binary numeric vector for if the respondent is temporarily unemployed.
- polint the respondent's self-reported interest in public affairs. 1 = Hardly at all. 2 = Only now and then. 3 = Some of the time. 4 = Most of the time.
- distrust\_govt the respondent's self-reported (dis)trust in the federal government's ability to do what's right. 1 = Just about always (trust the government). 2 = Most of the time. 3 = Some of the time. 4 = None of the time/never.
- govt\_crooked the respondent's assessment of how many government officials are crooked. 1 = Hardly any. 2 = Not many. 3 = Quite a few; quite a lot.
- govt\_waste the respondent's assessment of how much the government wastes in tax money. 1 = Not very much. 2 = Some. 3 = A lot.
- govt\_biginterests the respondent's assessment of whether the government is run by a few big interests. 0 = Run for the benefit of all people. 1 = Run by a few big interests.

### Details

The major party thermometer index is calculated as the thermometer rating for the Democratic party minus the thermometer rating for the Republican party. 100 is then added to that difference, which is then divided by 2. Fractional results are rounded to the next highest integer. Also note the coding of the "government distrust" measures. These are reverse-coded from their original scales.

### Source

Data come from ANES's time series file.

---

anes_prochoice	<i>Abortion Attitudes (ANES, 2012)</i>
----------------	--

---

### Description

A simple data set for in-class illustration about how to estimate and interpret interactive relationships. The data here are deliberately minimal for that end.

### Usage

anes\_prochoice

**Format**

A data frame with 5914 observations on the following 14 variables.

version version identifier from ANES

caseid time-series case identifier from ANES

health oppose/"NFNO"/favor [0:2] abortion if pregnancy would hurt woman

fatal oppose/"NFNO"/favor [0:2] abortion if pregnancy would cause woman to die

incest oppose/"NFNO"/favor [0:2] abortion if pregnancy was caused by incest

rape oppose/"NFNO"/favor [0:2] abortion if pregnancy was caused by rape

bd oppose/"NFNO"/favor [0:2] abortion if fetus would be born with serious birth defect

fin oppose/"NFNO"/favor [0:2] abortion if having child would impose financial hardship

sex oppose/"NFNO"/favor [0:2] abortion if the child will not be the sex the woman wants

choice oppose/"NFNO"/favor [0:2] abortion if woman chooses to have one

pid respondent's partisanship [0:2] (Democrat, Independent, Republican)

knowspeaker was the respondent able to correctly identify the Speaker of the House (John Boehner)

addchoice an additive scale of the abortion scores [0:16]

lchoice a continuous latent scale of pro-choice scores (from a simple graded response model)

**Details**

"NFNO" = "Neither Favor Nor Oppose"

**Source**

Data come from ANES's (2012) time series.

---

anes_vote84	<i>Simple Data for a Simple Model of Individual Voter Turnout (ANES, 1984)</i>
-------------	--

---

**Description**

This is a simple data set for estimating a simple model on voter turnout from the 1984 American National Election Studies (ANES) 1984 time-series.

**Usage**

anes\_vote84

**Format**

A data frame with 2257 observations on the following 9 variables.

`uid` a unique identifier for the respondent

`stateabb` the state where the respondent lives (as an abbreviation)

`vote` whether the respondent voted (1 = yes; 0 = no)

`age` the age of the respondent

`educ` the education-level of the respondent. See details section for more.

`female` whether the respondent is a woman (1 = female; 0 = male)

`south` does the respondent live in the south (1 = yes; 0 = no)

`polint` the political interest of the respondent in the campaigns (-1 = not much interested; 0 = somewhat interested; 1 = very much interested)

`govrace` did the respondent's state have a gubernatorial election that same November (1 = yes; 0 = no)

**Details**

The `vote` variable is deliberately coded where those with a value of 1 are respondents who said they voted and the ANES was able to confirm that with voter registration records. There are purportedly 85 responses in this raw variable where the respondent said they voted, but this could not be confirmed from registration records. Those cases are recorded as NA. The `educ` variable ranges from 1 (finished 8th grade or less than that) to 10 (respondent holds an advanced degree). The `uid` variable is a simple sequence variable ranging from 1 to 2257 and is calculated on the original 1984 time-series study (May 3, 1999 version) before other recoding was done. This should allow some reproducibility for an interested user.

**Source**

Data come from ANES's (1984) time series.

---

Arca

*NYSE Arca Steel Index data, 2017–present*

---

**Description**

Daily data on the NYSE Arca Steel Index. These data are useful for me in teaching how Trump's 2018 steel tariffs didn't do much good for the steel industry.

**Usage**

Arca



**Format**

A data frame with 966 observations on the following 6 variables.

date the date  
 close the closing price  
 open the opening price  
 high the daily high in that day's trading  
 low the daily low in that day's trading

**Details**

These data are taken from `investing.com`. See: <https://www.investing.com/indices/arca-steel-historical-data>

---

arcticseaice	<i>Arctic Sea Ice Extent Data, 1901-2015</i>
--------------	--

---

**Description**

This data set from Connelly et al. (2017) measures the Arctic sea ice extent in  $10^6$  square kilometers. It includes lower bounds and upper bounds on annual averages.

**Usage**

arcticseaice

**Format**

A data frame with 115 observations on the following 4 variables.

year the year  
 value the annual Arctic sea ice extent (in  $10^6$  sq km)  
 ub The upper bound of the value, provided by Connelly et al.  
 lb The lower bound of the value, provided by Connelly et al.

**Details**

This is for illustration of climate change for my intro students. Connelly et al. (2017) are in part a methodological paper. The data I present here are from the "rescaled (unadjusted T)" data in the second sheet from their replication files.

**References**

Connolly et al. (2017), "Re-calibration of Arctic sea ice extent datasets using Arctic surface air temperature records". *Hydrological Sciences Journal* 62(8): 1317–40.

---

arg_tariff	<i>Simple Mean Tariff Rate for Argentina</i>
------------	--

---

**Description**

Simple mean tariff rate for Argentina, starting in 1980. The goal is to keep these data current.

**Usage**

arg\_tariff

**Format**

A data frame with three variables:

country country name (Argentina)

year the year

tariffrate the simple mean tariff rate for Argentina on all products (as a percentage)

**Details**

Data come from various sources. World Bank estimates are used for 1980-1984 and 2010-2018, but see also Lora's (2012) report for the Inter-American Development Bank. The 1980-1984 estimates are actually means for 1980-1 and 1982-4 via Laird and Nogues' (1989) article in the World Bank Economic Review.

---

asn_stats	<i>Aviation Safety Network Statistics, 1942-2019</i>
-----------	--

---

**Description**

These are yearly counts on air accidents and fatalities, including measures for corporate jet accidents and hijackings. The hijackings are of particular interest to me, at least from a historical terrorism perspective.

**Usage**

asn\_stats

**Format**

A data frame with 78 observations on the following 7 variables.

`year` numeric vector for the year

`airacc` a numeric vector for the number of airliner accidents

`airfatal` a numeric vector for the number of fatalities from airliner accidents

`corpjetacc` a numeric vector for the number of corporate jet accidents

`corpjetfatal` a numeric vector for the number of fatalities from corporate jet accidents

`hijack` a numeric vector for the number of hijackings/skyjackings

`hijackfatal` a numeric vector for the number of fatalities from hijackings/skyjackings

**Details**

All fatality estimates exclude ground fatalities. All accidents are hull-loss accidents. The airliner figures are for those flights with at least 14 passengers. Check <https://aviation-safety.net/statistics/period/stats.php?cat=H2> for more.

**Source**

Aviation Safety Network, a service provided by the Flight Safety Foundation.

---

CFT15

*Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate*

---

**Description**

This is the replication data for "Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate", published in 2015 in *Journal of Causal Inference*. I use these data to teach about regression discontinuity designs.

**Usage**

CFT15

**Format**

A data frame with 1390 observations on the following 9 variables.

`state` a numeric vector for the state. This is ultimately a categorical variable.

`year` a numeric vector for the year of the election.

`vote` a numeric vector for the Democratic vote share in the *next* election (i.e. six years later).

`margin` a numeric vector for the Democratic party's margin of victory in the statewide election.

This is the running variable, in RDD parlance.

`class` a numeric vector for the class to which each Senate seat belongs.

termshouse a numeric vector for the Democratic candidate's cumulative number of terms previously served in the U.S. House.

termssenate a numeric vector for the Democratic candidate's cumulative number of terms previously served in the U.S. Senate.

population a numeric vector for the population of the Senate seat's state.

treatment a numeric vector that is 1 if margin > 0 and is 0 if margin < 0.

### Source

Cattaneo, Matias D. and Brigham R. Frandsen and Rocio Titiunik. 2015. "Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate". *Journal of Causal Inference* 3(1): 1–24.

### References

Cattaneo, Matias D. and Brigham R. Frandsen and Rocio Titiunik. 2015. "Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate". *Journal of Causal Inference* 3(1): 1–24.

Calonico, Sebastian and Matias D. Cattaneo and Max H. Farrell and Rocio Titiunik. 2017. "rdrobust: Software for regression-discontinuity designs". *The Stata Journal* 17(2):372–404.

---

clemson\_temps

*Daily Clemson Temperature Data*

---

### Description

This data set contains daily temperatures (highs and lows) for Clemson, South Carolina from Jan. 1, 1930 to the end of the most recent calendar year. The goal is to update this periodically with new data for as long as I live in this town.

### Usage

clemson\_temps

### Format

A data frame with 33,148 observations on the following 3 variables.

date the date

tmin the daily low, adjusted to Fahrenheit

tmax the daily high, adjusted to Fahrenheit

### Details

Data obtained from NOAA, via the **rnoaa** package. The station identifier is GHCND:USC00381770 for added context. The call from **rnoaa** returns these values initially as Celsius\*10. I don't know why NOAA does it this way, but there you go.

---

`co2emissions`*Carbon Dioxide Emissions Data*

---

**Description**

This is a sample data set, cobbled from various sources, about carbon dioxide emissions in the history of the planet from 800,000 BCE to the most recently concluded calendar year. I use this for a data visualization example for a lecture on climate change and international politics. Data communicate yearly averages/estimates.

**Usage**`co2emissions`**Format**

A data frame with 3,099 observations on the following 2 variables.

`year` the year (negative values = BCE)

`value` estimated carbon dioxide emissions (in ppm)

**Details**

The data come from many sources. Before 0 CE, the data come from 10 sources described by the Environmental Protection Agency ("Climate Change Indicators: Atmospheric Concentrations of Greenhouse Gases"). Observations from 0 CE to 2014 come from Meinshausen et al. (2017) doi: [10.5194/gmd1020572017](https://doi.org/10.5194/gmd1020572017). Observations from 2015 forward come from NASA ("Vital Signs").

**References**

EPICA Dome C and Vostok Station, Antarctica: approximately 796,562 BCE to 1813 CE Lüthi, D., M. Le Floch, B. Bereiter, T. Blunier, J.-M. Barnola, U. Siegenthaler, D. Raynaud, J. Jouzel, H. Fischer, K. Kawamura, and T.F. Stocker. 2008. High-resolution carbon dioxide concentration record 650,000–800,000 years before present. *Nature* 453:379–382.

Law Dome, Antarctica, 75-year smoothed: approximately 1010 CE to 1975 CE Etheridge, D.M., L.P. Steele, R.L. Langenfelds, R.J. Francey, J.-M. Barnola, and V.I. Morgan. 1998. Historical CO<sub>2</sub> records from the Law Dome DE08, DE08-2, and DSS ice cores. In: *Trends: A compendium of data on global change*. Oak Ridge, TN: U.S. Department of Energy.

Siple Station, Antarctica: approximately 1744 CE to 1953 CE Neftel, A., H. Friedli, E. Moor, H. Lötscher, H. Oeschger, U. Siegenthaler, and B. Stauffer. 1994. Historical carbon dioxide record from the Siple Station ice core. In: *Trends: A compendium of data on global change*. Oak Ridge, TN: U.S. Department of Energy.

Mauna Loa, Hawaii: 1959 CE to 2015 CE NOAA (National Oceanic and Atmospheric Administration). 2016. Annual mean carbon dioxide concentrations for Mauna Loa, Hawaii.

Barrow, Alaska: 1974 CE to 2014 CE Cape Matatula, American Samoa: 1976 CE to 2014 CE South Pole, Antarctica: 1976 CE to 2014 CE NOAA (National Oceanic and Atmospheric Administration).

2016. Monthly mean carbon dioxide concentrations for Barrow, Alaska; Cape Matatula, American Samoa; and the South Pole.

Cape Grim, Australia: 1992 CE to 2006 CE Shetland Islands, Scotland: 1993 CE to 2002 CE Steele, L.P., P.B. Krummel, and R.L. Langenfelds. 2007. Atmospheric CO<sub>2</sub> concentrations (ppmv) derived from flask air samples collected at Cape Grim, Australia, and Shetland Islands, Scotland. Commonwealth Scientific and Industrial Research Organisation.

Lampedusa Island, Italy: 1993 CE to 2000 CE Chamard, P., L. Ciattaglia, A. di Sarra, and F. Monteleone. 2001. Atmospheric carbon dioxide record from flask measurements at Lampedusa Island. In: Trends: A compendium of data on global change. Oak Ridge, TN: U.S. Department of Energy.

Meinshausen, M., Vogel, E., Nauels, A., Lorbacher, K., Meinshausen, N., Etheridge, D. M., Fraser, P. J., Montzka, S. A., Rayner, P. J., Trudinger, C. M., Krummel, P. B., Beyerle, U., Canadell, J. G., Daniel, J. S., Enting, I. G., Law, R. M., Lunder, C. R., O'Doherty, S., Prinn, R. G., Reimann, S., Rubino, M., Velders, G. J. M., Vollmer, M. K., Wang, R. H. J., and Weiss, R.: Historical greenhouse gas concentrations for climate modelling (CMIP6), *Geosci. Model Dev.*, 10, 2057-2116, 2017.

---

coffee\_imports

*Coffee Imports for Select Importing Countries*

---

### Description

A simple panel on coffee imports for importing countries.

### Usage

coffee\_imports

### Format

A data frame with 4530 observations on the following 4 variables.

country a character vector for the country

member a numeric vector indicating whether the importer is or is not a member of the International Coffee Organization

year a numeric vector for the year

value a numeric vector for the coffee imports for all select importing countries (in thousand 60-kg bags)

### Details

Data come from the International Coffee Organization, of which I feel I should be a member.

Observations for the People's Republic of China are removed because those can be obtained by adding together the values for "Macao", "Hong Kong", and "China (Mainland)".

The user may want to be mindful about when 0s in the value data are actually communicating that the entry did not exist at the time, or no longer exists. For example, there is no independent Armenia

in 1990 (and whatever imports Armenia had are lumped into the USSR value for 1990). Likewise, the 0s for the USSR in 1992 are communicating the USSR no longer exists that year and you should instead look into one of the constituent republics for the information you want. You may want to benchmark this information to some kind of state system membership data.

---

 coffee\_price

*The Primary Commodity Price for Coffee (Arabica, Robustas)*


---

### Description

This is primary commodity price data for coffee (Arabica, Robustas) from 1980 to the present. I manually update these data since FRED's coverage since 2017 has been spotty.

### Usage

coffee\_price

### Format

A data frame with the following 3 variables.

date the date (year-month)

arabica the price (monthly average) of mild Arabica, via International Coffee Organization data, in nominal US cents per pound

robustas the price (monthly average) of Robustas, via International Coffee Organization data, in nominal US cents per pound

### Details

Data come from International Monetary Fund (Primary Commodity Prices) and International Coffee Organization. The IMF adds these prices are global and the New York cash price, ex-dock

---

 commodity\_prices

*Select World Bank Commodity Price Data (Monthly)*


---

### Description

A data set on select, monthly commodity prices made available by the World Bank in its so-called "pink sheet." These data are potentially useful for applications on data gathering, inflation adjustments, indexing, cointegration, general economic riff-raff, and more.

### Usage

commodity\_prices

**Format**

A data frame with the following 11 variables.

date a date

oil\_brent crude oil, UK Brent 38° API (\$/bbl)

oil\_dubai crude oil, Dubai Fateh 32 API for years 1985-present; 1960-84 refer to Saudi Arabian Light, 34° API (\$/bbl).

coffee\_arabica coffee (ICO), International Coffee Organization indicator price, other mild Arabicas, average New York and Bremen/Hamburg markets, ex-dock (\$/kg)

coffee\_robustas coffee (ICO), International Coffee Organization indicator price, Robustas, average New York and Le Havre/Marseilles markets, ex-dock (\$/kg)

tea\_columbo tea (Colombo auctions), Sri Lankan origin, all tea, arithmetic average of weekly quotes (\$/kg).

tea\_kolkata tea (Kolkata auctions), leaf, include excise duty, arithmetic average of weekly quotes (\$/kg).

tea\_mombasa tea (Mombasa/Nairobi auctions), African origin, all tea, arithmetic average of weekly quotes (\$/kg).

sugar\_eu sugar (EU), European Union negotiated import price for raw unpackaged sugar from African, Caribbean and Pacific (ACP) under Lome Conventions, c.i.f. European ports (\$/kg)

sugar\_us sugar (United States), nearby futures contract, c.i.f. (\$/kg)

sugar\_world sugar (World), International Sugar Agreement (ISA) daily price, raw, f.o.b. and stowed at greater Caribbean ports (\$/kg).

**Details**

All data are in nominal USD. Adjust (to taste) accordingly.

Data compiled by the World Bank for its historical data on commodity prices. The oil price data come from a combination of sources, supposedly Bloomberg, Energy Intelligence Group (EIG), Organization of Petroleum Exporting Countries (OPEC), and the World Bank. Data on coffee prices come from Bloomberg, Complete Coffee Coverage, the International Coffee Organization, Thomson Reuters Datastream, and the World Bank. Data on tea prices for Colombo auctions come from the International Tea Committee, Tea Broker's Association of London, Tea Exporters Association Sri Lanka, and the World Bank. Data on tea prices for Kolkata auctions come from the International Tea Committee, Tea Board India, Tea Broker's Association of London, and the World Bank. Tea prices for Mombasa/Nairobi auctions come from African Tea Brokers Limited, International Tea Committee, Tea Broker's Association of London, and the World Bank. EU sugar price data come from International Monetary Fund, World Bank. Sugar price data for the United States come from Bloomberg and World Bank. Global sugar price data come from Bloomberg, International Sugar Organization, Thomson Reuters Datastream, and the World Bank.

This data set effectively deprecates the `sugar_price` and `coffee_price` data sets in this package. Both may be removed at a later point.



---

CP77

*Education Expenditure Data (Chatterjee and Price, 1977)*

---

**Description**

This is a simple data set provided by Chatterjee and Price (1977, p. 108) that serves as a known example of heteroscedasticity.

**Usage**

CP77

**Format**

A data frame with 50 observations on the following 6 variables.

state a character vector for the state

region a character vector for the Census region

urbanpop a numeric vector for the number of residents (per thousand) living in urban areas in 1970

incpc a numeric vector for income per capita in 1973

pop a numeric vector for residents (per thousand) under 18 years of age in 1974

edexppc a numeric vector for per capita public school expenditures in a state, projected for 1975.

**Details**

I copied these data from the robustbase package. I just didn't want to make my students install it. Note: I'm pretty sure "NB" was suppose to be "NE" and that "DY" is supposed to be "KY". I made those changes.

**References**

P. J. Rousseeuw and A. M. Leroy (1987) Robust Regression and Outlier Detection; Wiley, p.110, table 16.

---

Datasaurus

*The Datasaurus Dozen*

---

**Description**

An illustrative exercise in never trusting the summary statistics without also visualizing them.

**Usage**

Datasaurus

**Format**

A data frame with 1,846 observations on the following 3 variables.

dataset the particular data set, one of 12

x a random variable

y another random variable

**Details**

Data were created by Alberto Cairo to illustrate you should always visualize your data beyond the summary statistics. These are 12 data sets, in long form, each with a mean of x about 54.26, a mean of y about 47.83. The standard deviation for x is about 16.76 and the standard deviation of y is about 26.93. x and y will correlate weakly, about -.06.

**Author(s)**

Alberto Cairo, Justin Matejka, George Fitzmaurice

**References**

Cairo, Alberto. 2016. "Download the Datasaurus: Never trust summary statistics alone; always visualize your data". URL: <http://www.thefunctionalart.com/2016/08/download-datasaurus-never-trust-summary.html>

Matejka, Justin and George Fitzmaurice. 2017. "Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing." *ACM SIGCHI Conference on Human Factors in Computing Systems*.

---

Dee04

*Are There Civics Returns to Education?*

---

**Description**

This should be a data set for a (partial?) replication of Dee's (2004) article on the purported civics returns to education. I use these data for in-class illustration about instrumental variable analyses.

**Usage**

Dee04

**Format**

A data frame with 9227 observations on the following 8 variables.

schoolid a numeric vector that should be understood as categorical

hispanic a numeric vector for if the person is Hispanic

college a numeric vector for if the person went to college

black a numeric vector for if the person is black  
 otherrace a numeric vector for if the person is another race  
 female a numeric vector for if the person is a woman  
 register a numeric vector for if the person is registered to vote  
 distance a numeric vector for the distance to college

### Details

I should note I acquired this data set in Mexico City sitting on a two-week program at IPSA-FLACSO Mexico Summer School in 2019. The sample size here (9,227) is about two thousand short of what Dee reports in his article. It'll do, though.

### References

Dee, Thomas S. 2004. "Are there civics returns to education?" *Journal of Public Economics* 88: 1697–1720

---

DJIA

*Dow Jones Industrial Average, 1885-Present*

---

### Description

This data set contains the value of the Dow Jones Industrial Average on daily close for all available dates (to the best of my knowledge) from 1885 to the most recently concluded calendar year. Extensions shouldn't be too difficult with existing packages.

### Usage

DJIA

### Format

A data frame with 36951 observations on the following 2 variables.

date the date

value the value of the the Dow Jones Industrial Average at daily close

### Details

Observations before October 7, 1896 are from the single Dow Jones Average. Observations from October 7, 1896 to July 30, 1914 are from the first DJIA. Observations before the 1914 closure of the first DJIA in July 1914 come from MeasuringWorth. Observations from its reopening in Dec. 12, 1914 to January 28, 1985 come from Pinnacle Systems. Observations from January 29, 1985 to the most recent observation come from a quantmod call.

## References

Samuel H. Williamson, 'Daily Closing Value of the Dow Jones Average, 1885 to Present,' MeasuringWorth, 2019.

Jeffrey A. Ryan and Joshua M. Ulrich, 'quantmod: Quantitative Financial Modelling Framework,' 2018.

---

DST

*Casualties/Fatalities in the U.S. for Drunk-Driving, Suicide, and Terrorism*

---

## Description

These are fatalities (and, in the case of terrorism, casualties as well) for drunk-driving, suicide, and acts of terrorism in the U.S. spanning 1970 to 2018. Only one of these is sufficiently important to command public attention despite being the least severe public bad. Do you want to guess which one?

## Usage

DST

## Format

A data frame with 49 observations on the following 5 variables.

year the year

nkill a numeric vector for the number killed in acts of terrorism

terrtotal a numeric vector for the number killed or wounded in acts of terrorism

suicides a numeric vector for the number of suicides

ddfata a numeric vector for the number of drunk-driving fatalities

## Details

Following my own work in *Political Research Quarterly*, terror incidents with unknown fatalities or number wounded were imputed to be 1. In those cases, the GTD has reason to believe at least one person died or was wounded, but doesn't know how many. GTD is weird about 1993, so perhaps treat those observations with some care (though it does well to capture the WTC bombing that year). Suicides include only those who passed, not those who survived a suicide attempt. Drunk-driving fatalities seem to include those who were killed in a drunk-driving accident despite not being drunk themselves.

## Source

Global Terrorism Database (Sept. 2019 update), Centers for Disease Control, U.S. Department of Transportation

---

eight_schools	<i>The Effect of Special Preparation on SAT-V Scores in Eight Randomized Experiments</i>
---------------	--

---

**Description**

You've all seen these before. These are the "eight schools" that everyone gets when being introduced to Bayesian programming. Here are the full data for your consideration, which you can use instead of awkwardly searching where the data are and copy-pasting them as a list. Every damn time, Steve.

**Usage**

```
eight_schools
```

**Format**

A data frame with 8 observations on the following 6 variables.

school a letter denoting the school  
 num\_treat the number of students in the school receiving the treatment  
 num\_control the number of students in the school in the control group  
 est the estimated treatment effect  
 se the standard error of the effect estimate  
 rvar the residual variance

**Details**

Data copy-pasted from Table 1 in Rubin (1981).

**References**

Rubin, Donald B. 1981. "Estimation in Parallel Randomized Experiments." *Journal of Educational Statistics* 6(4): 377-401.

---

election_turnout	<i>State-Level Education and Voter Turnout in 2016</i>
------------------	--

---

**Description**

A simple data set on education and state-level (+ DC) turnout in the 2016 presidential election. This is inspired by what Pollock (2012) does in his book.

**Usage**

```
election_turnout
```

**Format**

A data frame with 51 observations on the following 13 variables.

year the year of the presidential election (2016)

state the state abbreviation

region the state's Census region

division the state's Census division

turnoutho voter turnout for the highest office as percent of voting-eligible population (VEP)

perhsed the percentage of the state that completed high school

percoled the percentage of the state that completed college

gdppercap an estimate of the state's GDP per capita

ss is it a "swing state?"

trumpw did Trump win the state?

trumpshare the share of the vote Trump received

sunempr the state-level unemployment rate entering Nov. 2016

sunempr12md the state-level unemployment rate (12-month difference) entering Nov. 2016. Higher values indicate the unemployment rate is increasing entering Nov. 2016 relative to what it was entering Nov. 2015.

gdp an estimate of the state's GDP

**Details**

Data were created in early 2017 for an upper-division course on quantitative methods. Educational attainment and division/region data come from the Census. Voter turnout/share data come from the Elections Project at George Mason University. GDP per capita estimates come from Bureau of Economic Analysis. Unemployment data come from the Bureau of Labor Statistics and code to generate it was derived from a forthcoming publication of mine.

---

 eq\_passengercars

*Export Quality Data for Passenger Cars, 1963-2014*


---

**Description**

Data from the International Monetary Fund for the export quality and unit/trade value of passenger cars for all available countries and years from 1963 to 2014.

**Usage**

eq\_passengercars

**Format**

A data frame with 60424 observations on the following 6 variables.

country a character vector for the country/area.

ccode a numeric vector for the Correlates of War country code.

category a factor with levels Export Quality Index, Export quality 95 percent interval - lower bound, Export quality 95 percent interval - upper bound, Unit value of exports, Unit value 95 percent interval - lower bound, Unit value 95 percent interval - upper bound, Trade value of exports

type a factor with levels 51. Transport equipment, Passenger cars. This is a constant. I just felt like making it a factor.

year a numeric vector for the year

value a numeric vector for the value of the particular category.

---

 ESS10NO

---

*Norwegian Attitudes toward European Integration (2021-2022)*


---

**Description**

This is a simple data set to illustrate the use of sampling weights from the European Social Survey.

**Usage**

ESS10NO

**Format**

A data frame with 1,411 observations on the following 24 variables.

cntry a character vector with Norway's two-character ISO code

idno a numeric identifier for the individual respondent

region a character for one of six regions recorded by the European Social Survey

inwds a date-time vector for the start of the interview

inwde a date-time vector for the end of the interview

dweight a design weight

pspwght a post-stratification weight, including the design weight

pweight a population size weight

anweight an analysis weight

prob the sampling probability

stratum the sampling stratum

psu the primary sampling unit

eu\_vote a character vector indicating how a respondent would vote if given a vote on joining the European Union  
 brnnorge a dummy variable indicating whether respondent was born in Norway or not  
 agea a numeric vector for the respondent's age in years  
 imbgeco a numeric vector for if respondent thinks immigrants are generally good or bad for Norway's economy. Higher values = good  
 imueclt a numeric vector for if respondent thinks immigrants enrich or undermine Norway's culture. Higher values = enrich more than undermine  
 imwbcnt a numeric vector for if respondent thinks immigrants make Norway a better place to live. Higher values = better place to live  
 female a numeric vector for whether the respondent is a woman  
 eduysr a numeric vector for total years of education for the respondent  
 uempl a numeric vector for whether the respondent is currently unemployed but seeking work  
 polint a dummy variable indicating political interest. 1 = very or quite interested. 0 = hardly or not at all interested.  
 hinctnta a numeric vector for household income in deciles  
 lrscale a numeric vector for the ideology of the respondent on an 11-point [0:10] scale

### Details

You'll want to convert the eu\_vote variable into something usable. Possible values include "Remain Outside", "Join EU", "Don't Know", "Not Eligible", "Blank Ballot", "Refuse to Answer", "Wouldn't Vote". Perhaps it's reasonable to make this a dummy variable comparing those who want to join versus those who want Norway to remain outside the European Union.

The data are edition 2.2 of the 10th round of European Social Survey, which was released for public consumption on 21 December 2022.

### Source

European Social Survey, Round 10

---

ESS9GB

*British Attitudes Toward Immigration (2018-19)*

---

### Description

This is a replication data originally set to accompany a blog post and presentation to students at the University of Nottingham in March 2020. However, COVID-19 led to the cancellation of the talk.

### Usage

ESS9GB



**Format**

A data frame with 1,905 observations on the following 19 variables.

`name` a character for the name of the survey

`essround` a numeric for the ESS round

`edition` a character for the particular edition of the ESS round

`idno` a numeric/unique identifier

`centry` a character vector for the country (i.e. the UK)

`region` a character vector for the region of the UK the respondent lives

`brncntr` a numeric vector for if the respondent was born in the UK

`stintrvw` a Date for the interview start date

`endintrvw` a Date for the interview end date

`imbgeco` a numeric vector for if respondent thinks immigrants are generally good or bad for UK's economy. Higher values = good

`imueclt` a numeric vector for if respondent thinks immigrants enrich or undermine UK's culture. Higher values = enrich more than undermine

`imwbcnt` a numeric vector for if respondent thinks immigrants make UK a better place to live. Higher values = better place to live

`immigsent` a numeric vector for immigration sentiment (i.e. `imbgeco` + `imueclt` + `imwbcnt`). Higher values = more pro-immigration sentiment

`agea` a numeric vector for the respondent's age in years

`female` a numeric vector for whether the respondent is a woman

`edyrs` a numeric vector for total years of education for the respondent

`uempla` a numeric vector for whether the respondent is currently unemployed but seeking work

`hinctnta` a numeric vector for household income in deciles

`lrscale` a numeric vector for the ideology of the respondent on an 11-point [0:10] scale

**Details**

See accompanying blog post at <http://svmiller.com/blog/2020/03/what-explains-british-attitudes-toward-immigrants/>

**Source**

European Social Survey, Round 9

ESSBE5

*Trust in the Police in Belgium (European Social Survey, Round 5)***Description**

This is a sample data set cobbled from the fifth round of European Social Survey data for Belgium. It offers a means to do a basic replication of some of Chapter 5 of The SAGE Handbook of Regression Analysis and Causal Inference.

**Usage**

ESSBE5

**Format**

A data frame with 1704 observations on the following 10 variables.

`essround` a numeric for the ESS round

`edition` a character for the edition number of the fifth round

`idno` a numeric id number

`centry` a character vector for the country (i.e. Belgium, or BE)

`trstplc` a numeric vector for trust in the police on an 11-point scale. Higher values indicate more trust. 0 = "no trust at all". 10 = "complete trust"

`agea` a numeric vector for the respondent's age

`female` a numeric vector for whether the respondent is a woman or not.

`eduysr` a numeric vector for years of education.

`hincfel` a numeric vector for the respondent's feeling about their household income. 1 = "living comfortably", 2 = "coping on present income", 3 = "difficult on present income", 4 = "very difficult on present income"

`plcpvcr` a numeric vector for how successful police are at preventing crimes in a country on an 11-point scale. 0 = "extremely unsuccessful". 10 = "extremely successful."

**Details**

See Chapter 5 of The SAGE Handbook of Regression Analysis and Causal Inference for more information.

**Source**

European Social Survey (Round 5)

---

eurostat_codes	<i>Eurostat Country Codes</i>
----------------	-------------------------------

---

**Description**

A data set taken from Eurostat's glossary on codes and country classifications.

**Usage**

eurostat\_codes

**Format**

A data frame on the following 3 variables.

country an English country/territorial unit name

iso2c a two-character code for the country/territorial unit

cat a category indicator for the country/territorial unit. See Details section for more.

**Details**

The ISO two-character code for Kosovo is not "XK". XK is a "user assigned" ISO 3166 code that is not used by the International Organization for Standardization, but is nevertheless in wide use by entities like the European Commission. To the best of my knowledge, Kosovo's official ISO classification is still what it was when it was a subdivision of Serbia/Yugoslavia.

A glossary on Eurostat provides the following category entries included in this data frame. "EU" is an European Union member. "EFTA" are countries outside the European Union, but still included in the free trade agreement. "UK" is the United Kingdom, because they left. "EUCC" is a category for European Union candidate countries. "PC" are potential candidates. European Union expansion led to the delineation of neighboring states to "South" and "East" as part of the European Neighbourhood Policy (ENP). "OEC" stands for "Other European Countries", but is effectively a simple indicator for Russia.

---

eustates	<i>EU Member States (Current as of 2019)</i>
----------	--

---

**Description**

European Union membership by accession date

**Usage**

eustates

**Format**

A data frame with 28 observations on the following 3 variables.

date a date indicating accession  
 country a character vector for the country  
 iso2c a character vector for iso2c

**Details**

Data come from the European Union's website.

---

 fakeAPI

---

*Hypothetical (Fake) Data on Academic Performance*


---

**Description**

This is a hypothetical universe of schools in a given territorial unit, patterned off the apipop data available in the survey package.

**Usage**

fakeAPI

**Format**

A data frame with 10000 observations on the following 8 variables.

uid a numeric vector as a unique identifier for schools  
 schooltype a character vector for school type. E = elementary school. M = middle school. H = high school  
 county a character vector for the county, named after an Ohio State All-American. "County" incidence is weighted by how many All-American honors the Ohio State player had. It's my fake data. You make your own if you have a problem with it.  
 community a character vector for the school's community, either rural, suburban, or urban.  
 api a numeric vector vector an academic performance index for the school  
 meals a numeric vector for the percentage of school students eligible for subsidized meals  
 colgrad a numeric vector for the percentage of school parents with college degrees  
 fullqual a numeric vector for the percentage of the school with teachers that are fully qualified  
 sbase a numeric vector for some base differences between schools, patterned off the school type means for api00 in the apipop data.  
 cbase a numeric vector for some base differences between counties, randomly drawn from a uniform distribution  
 e a numeric vector for random errors

**Details**

These data were generated for a blog post on my website.

**References**

Miller, Steven V. 2020. "Some Parlor Tricks with Survey-Type Analyses in R." URL: <http://svmiller.com/blog/2020/08/some-parlor-tricks-with-survey-type-analyses-in-r/>

---

fakeHappiness

*Fake Data on Happiness*

---

**Description**

This is a toy ("fake") data set I might use to illustrate the so-called curvilinear effect of age on happiness.

**Usage**

fakeHappiness

**Format**

A data frame with 1000 observations on the following 8 variables.

age a numeric vector for age.

female a numeric that equals 1 if the respondent is a woman

collegeed a numeric vector that equals 1 if the respondent says s/he has a college degree

famincr a numeric vector for the respondent's household income. Ranges from 1 to 12.

bornagain a numeric vector for whether the respondent self-identifies as a born-again Christian.

e random noise, generated from a normal distribution with a mean of 0 and a standard deviation of 3

happy an arbitrary happiness variable. See details for its construction

z\_happy the same arbitrary happiness variable, scaled to have a mean of 0 and a standard deviation of 1. This makes it seem more "latent".

**Details**

Data are randomly sampled from the TV16 data set in the same package for the age, female, college education, family income, and born-again variables. Thereafter, I created an arbitrary "happiness" variable that is equal to  $100 - .95*age + .01*(age^2) + .25*female + .05*famincr + .1*bornagain + e$ . The data are not supposed to be realistic, per se. They're supposed to be functional for this purpose.

---

`fakeLogit`*Fake Data for a Logistic Regression*

---

**Description**

This is a simple fake data set to illustrate a logistic regression.

**Usage**`fakeLogit`**Format**

A data frame with 10000 observations on the following 2 variables.

`x` a five-item functionally ordered categorical variable

`y` a binary variable that is either 0 or 1

**Details**

The data are generated such that the outcome `y` is a logistic function of the `x` variable and come from a `rbinom()` call. The estimated natural logged odds of `y` when `x` is 0 is -2.8. Each unit increase in `x` is simulated to increase the natural logged odds of `y` by 1.4. This example is very much patterned off a similar fake data set that Pollock (2012) uses to teach about logistic regression. In his case, `x` is a stand-in for hypothetical education categories and `y` is whether this fake person voted or not.

---

`fakeTSCS`*Fake Data for a Time-Series Cross-Section*

---

**Description**

This is a toy (i.e. "fake") data set created by the `fabricatr` package. There are 100 observations for 25 hypothetical countries. The outcome `y` is a linear function of a baseline for each hypothetical country, plus a yearly growth trend as well as varying growth errors for each country. `x1` is supposed to have a linear effect of .5 on `y`, all things considered. `x2` is supposed to have a linear effect of 1 on `y` for each unit change in `x2`, all things considered.

**Usage**`fakeTSCS`

**Format**

A data frame with 2500 observations on the following 8 variables.

year a numeric vector for the year

country a character vector for the country

y a numeric vector for the outcome.

x1 a continuous variable

x2 a binary variable

base a numeric vector for the baseline starting point for each country

growth\_units a numeric vector for the growth units for each country

growth\_error a numeric vector for the growth errors for each country

**Details**

x1 is generated by a normal distribution with a mean of 5 and a standard deviation of 2. x2 is drawn from a Bernoulli distribution with a probability of .5 of observing a 1.

---

 fakeTSD

---

*Fake Data for a Time-Series*


---

**Description**

This is a toy (i.e. "fake") data set created by the `fabricatr` package. There are 100 observations. The outcome  $y$  is a linear function of  $20 + (.25 * \text{year}) + (.25 * x1) + (1 * x2) + e$ . This clearly implies some autocorrelation in the data. I.e. it's a time-series.

**Usage**

```
fakeTSD
```

**Format**

A data frame with 100 observations on the following 5 variables.

year the year

y an outcome

x1 a continuous variable

x2 a binary variable

e randomly generated errors

**Details**

Errors are random-normal with a mean of 0 and a standard deviation of 1. x1 is generated by a normal distribution with a mean of 5 and a standard deviation of 2. x2 is drawn from a Bernoulli distribution with a probability of .5 of observing a 1.

---

ghp100k

*Gun Homicide Rate per 100,000 People, by Country*

---

### Description

This is the yearly rate of gun homicides per 100,000 people in the population, selecting on "Western" countries of interest.

### Usage

ghp100k

### Format

A data frame with 561 observations on the following 3 variables.

country the country

year the year

value a numeric vector for the estimated rate of gun homicide per 100,000 people

### Details

The reported, or calculated annual crude rate of completed, intentional homicide committed with a firearm, per 100,000 population, in years descending.

Where a jurisdiction's published count of 'annual homicide' includes cases of attempted (uncompleted) homicide, these figures have been disaggregated wherever possible.

In the United States, this category is confused by inaccurate and conflicting data published, suppressed or labeled as unreliable by the Centers for Disease Control and Prevention (CDC) and the Federal Bureau of Investigation (FBI). Suppression can result in zero values where in fact homicides did occur.

Incomplete classification by local agencies can also result in a significant proportion of events being categorized as 'unknown cause' or similar.

Before quoting these datasets, please follow the citation links for a description of the considerable differences between them and the reasons for data suppression.

Where a rate is calculated by GunPolicy.org, a matched population estimate is also cited.

### Source

<https://www.gunpolicy.org>



---

GHR04	<i>Comparative Public Health: The Political Economy of Human Misery and Well-Being</i>
-------	--

---

### Description

This is a data set for replicating Ghobarah et al. (2004), a reduced form of what they make available on Dataverse for replication. Variables have been renamed for legibility.

### Usage

GHR04

### Format

A data frame with 182 observations on the following 15 variables.

country a character vector denoting a country name

iso3c a three-character ISO code for the country

publ1thexppgdp a numeric vector for public health expenditures as a percentage of GDP

totexph1th a numeric vector for total expenditures on health

hale a numeric vector for health adjusted life expectancy (in years)

log\_gdppc a numeric vector for (log-transformed) GDP per capita

gini a numeric vector for income inequality

log\_educ a numeric vector for (log-transformed) educational attainment

log\_vanhanen a numeric vector for (log-transformed) racial-linguistic-religious heterogeneity

rivalry a dummy variable indicating the presence of an enduring international rivalry for the country

polity a numeric vector communicating a Polity score, as a measure of the democratic nature of the country's regime

prvhlthexpgdp a numeric vector for private spending on health as a percentage of GDP

urban\_growth a numeric vector for the pace of urbanization

cwdeaths a numeric vector for civil war deaths

contig\_cw a dummy variable communicating whether there is a civil war in a geographically contiguous territory

### Details

The three-character ISO code is the only new addition to the data. I add this because the country names they have in the data are not neat and may lead users astray if they wanted to search for a specific observation. The ISO code for Yugoslavia (Serbia and Montenegro) around this time was "SCG".

The data the authors make available come with no .do file to indicate what exactly they used. Some forensic work based on the descriptive statistics they mention led to this reduced form of their data, which almost perfectly replicates their results. The differences are typically in the hundredths, and often in the thousandths, and should be considered "good enough" for replication purposes. The only real confusion on my end is why I ended up with one more observation than they report in their analyses. This suggests one (or more?) of their variables they use has an 'NA', but I have no way of knowing what it could be.

### Source

Ghobarah, Hazem Adam, Paul Huth, and Bruce Russett. 2004. "Comparative Public Health: The Political Economy of Human Misery and Well-Being" *International Studies Quarterly* 48: 73-94

---

gss\_abortion

*Abortion Opinions in the General Social Survey*

---

### Description

This is a toy data set derived from the General Social Survey that I intend to use for several purposes. First, the battery of abortion items can serve as toy data to illustrate mixed effects modeling as equivalent to a one-parameter (Rasch) model. Second, I include some covariates to also do some basic regressions. I think abortion opinions are useful learning tools for statistical inference for college students. Third, there's a time-series component as well for understanding how abortion attitudes have changed over time.

### Usage

gss\_abortion

### Format

A data frame with 64,814 observations on the following 18 variables.

id a unique respondent identifier

year the survey year

age the respondent's age in years

race the respondent's race, as character variable

sex the respondent's gender, as character variable

hispaniccat the respondent's Hispanic ethnicity, as character variable

educ how many years the respondent spent in school

partyid the respondent's party identification, as character variable

relactiv the self-reported religious activity of the respondent on a 1:11 scale

abany a binary variable that equals 1 if the respondent thinks abortion should be legal for any reason. 0 indicates no support for abortion for any reason.

- abdefect a numeric vector that equals 1 if the respondent thinks abortion should be legal if there is a serious defect in the fetus. 0 indicates no support for abortion in this circumstance.
- abnomore a numeric vector that equals 1 if the respondent thinks abortion should be legal if a woman is pregnant but wants no more children. 0 indicates no support for abortion in this circumstance.
- abh1th a numeric vector that equals 1 if the respondent thinks abortion should be legal if a pregnant woman's health is in danger. 0 indicates no support for abortion in this circumstance.
- abpoor a numeric vector that equals 1 if the respondent thinks abortion should be legal if a pregnant woman is poor and cannot afford more children. 0 indicates no support for abortion in this circumstance.
- abrape a numeric vector that equals 1 if the respondent thinks abortion should be legal if the woman became pregnant because of a rape. 0 indicates no support for abortion in this circumstance.
- absingle a numeric vector that equals 1 if the respondent thinks abortion should be legal if a pregnant woman is single and does not want to marry the man who impregnated her. 0 indicates no support for abortion in this circumstance.
- pid partyid recoded so that 7 = NA
- hispanic a dummy variable that equals 1 if the respondent is any way Hispanic

### Details

Data include all General Social Survey observations from 1972 to 2018 for these variables. Be mindful of missing data.

---

gss_spending	<i>Attitudes Toward National Spending in the General Social Survey (2018)</i>
--------------	---

---

### Description

This is a toy data set that collects attitudes on toward national spending for various things in the General Social Survey for 2018. I use these data for in-class illustration about ordinal variables and ordinal models.

### Usage

```
gss_spending
```

### Format

A data frame with 2348 observations on the following 33 variables.

year a numeric constant for the GSS survey year (2018)

id a unique identifier for the survey respondent

age a numeric vector for the age of the respondent (min: 18, max: 89)

sex a numeric vector for the respondent's sex (1 = female, 0 = male)

educ a numeric vector for the highest year of school completed (min: 0, max: 20)

degree a numeric vector for the respondent's highest degree (0 = did not graduate high school, 1 = high school, 2 = junior college, 3 = bachelor degree, 4 = graduate degree)

race a numeric vector for the respondent's race (1 = white, 2 = black, 3 = other)

rincom16 a numeric vector for the respondent's yearly income (min: 1 (under \$1,000), max: 26 (\$170,000 or over))

partyid a numeric vector for the respondent's party identification on the familiar seven-point scale. NOTE: D to R partisanship in this variable goes from 0 to 6. 7 = supporters of other parties. You may want to recode this if you want an interval-level measure of partisanship.

polviews a numeric vector for the respondent's ideology (min: 1 (extremely liberal), max: 7 (extremely conservative))

xnorcsiz a numeric vector for the NORC size code. This is a measure of what kind of area in which the respondent took the survey (i.e. lives). 1 = city, greater than 250k residents. 2 = city, between 50k-250k residents. 3 = suburbs of a large city. 4 = suburbs of a medium-sized city. 5 = unincorporated area of a large city. 6 = unincorporated area of a medium city. 7 = city, between 10-50k residents. 8 = town, greater than 2,500 residents. 9 = smaller areas. 10 = open country.

news a numeric vector for how often the respondent reads the newspapers. 1 = everyday. 2 = a few times a week. 3 = once a week. 4 = less than once a week. 5 = never.

wrkstat a numeric vector for the respondent's work status. 1 = working full-time. 2 = working part-time. 3 = temporarily not working. 4 = unemployed/laid off. 5 = retired. 6 = in school. 7 = house-keeping work. 8 = other.

natspac a numeric vector for attitudes toward spending on the space program. See details below for this variable and all other variables beginning with nat.

natenvir a numeric vector for attitudes toward spending on improving/protecting the environment.

natheal a numeric vector for attitudes toward spending on improving/protecting the nation's health.

natcity a numeric vector for attitudes toward spending on solving the big city's problems.

natcrime a numeric vector for attitudes toward spending on halting the "rising crime rate." This question is subtly hilarious.

natdrug a numeric vector for attitudes toward spending on dealing with drug addiction.

nateduc a numeric vector for attitudes toward spending on improving the nation's education system.

natrace a numeric vector for attitudes toward spending on improving the condition of black people.

natarms a numeric vector for attitudes toward spending on the military/armaments/defense.

nataid a numeric vector for attitudes toward spending on foreign aid.

natfare a numeric vector for attitudes toward spending on welfare.

natroad a numeric vector for attitudes toward spending on highways and bridges.

natsoc a numeric vector for attitudes toward spending on social security.

natmass a numeric vector for attitudes toward spending on mass transportation.

natpark a numeric vector for attitudes toward spending on parks and recreation.  
 natchld a numeric vector for attitudes toward spending on assistance for child care.  
 natsci a numeric vector for attitudes toward spending on scientific research.  
 natenrgy a numeric vector for attitudes toward spending on alternative sources of energy.  
 sumnat a numeric vector for the sum total of responses to all the aforementioned spending variables (i.e. those that begin with nat). This creates an interval-ish measure with a nice and mostly normal distribution.  
 sumnatsoc a numeric vector for the sum of all responses toward various "social" prompts (i.e. natenvir, natheal, natdrug, nateduc, natrace, natfare, natroad, natmass, natpark, natsoc, natchld). This creates an interval-ish measure with a mostly normal (but small left skew) distribution.

### Details

For all the variables beginning with nat, note that I rescaled the original data so that -1 = respondent thinks country is spending too much on this topic, 0 = respondent thinks country is spending "about (the) right" amount, and 1 = respondent thinks country is spending too little on this topic. I do this to facilitate reading each nat prompt as increasing support for more spending (the extent to which increasing values means the respondent thinks the country spends too little on a given prompt). I think this is more intuitive.

Also, the natspac, natenvir, natheal, natcity, natcrime, natdrug, nateduc, natrace, natarms, nataid, and natfare have "alternate" prompts in later GSS waves in which a subset of respondents get a slightly different prompt. For example, one set of respondents for natcity gets a prompt of "Solving the problems of the big cities" (the legacy prompt) whereas another set of respondents gets a prompt of "Assistance to big cities" (typically noted as "version y" in the GSS). I, perhaps problematically if I were interested in publishing analyses on these data, combine both prompts into a single variable. I don't think it's a huge problem for what I want the data to do, but FYI.

### Source

General Social Survey, 2018

---

gss\_wages

*The Gender Pay Gap in the General Social Survey*

---

### Description

Wage data from the General Social Survey (1974-2018) to illustrate wage discrepancies by gender (while also considering respondent occupation, age, and education).

### Usage

gss\_wages

**Format**

A data frame with 11 variables:

year the survey year

realrinc the respondent's base income (in constant 1986 USD)

age the respondent's age in years

occ10 respondent's occupation code (2010)

occrcode recode of the occupation code into one of 11 main categories

prestg10 respondent's occupational prestige score (2010)

childs number of children (0-8)

wrkstat the work status of the respondent (full-time, part-time, temporarily not working, unemployed (laid off), retired, school, housekeeper, other)

gender respondent's gender (male or female)

educat respondent's degree level (Less Than High School, High School, Junior College, Bachelor, or Graduate)

maritalcat respondent's marital status (Married, Widowed, Divorced, Separated, Never Married)

**Details**

For further details, see <https://gssdataexplorer.norc.org>. Consult <https://census.gov> for more information about occupation codes.

---

Guber99

*School Expenditures and Test Scores for 50 States, 1994-95*

---

**Description**

A data set for a canonical case of a Simpson's paradox, useful for in-class instruction on the topic.

**Usage**

Guber99

**Format**

A data frame with 50 observations on the following 8 variables.

state a character vector for the state

expendpp a numeric vector for the current expenditure per pupil in average daily attendance in public elementary and secondary schools, 1994-95 (in thousands of dollars)

ptratio a numeric vector for the average pupil/teacher ratio in public elementary and secondary schools, Fall 1994

tsalary a numeric vector for the estimated average annual salary of teachers in public elementary and secondary schools, 1994-95 (in thousands of dollars)

perctakers a numeric vector for the percentage of all eligible students taking the SAT, 1994-95  
 verbal a numeric vector for the average verbal SAT score, 1994-95  
 math a numeric vector for the average math SAT score, 1994-95  
 total a numeric vector for the average total SAT score, 1994-95

## References

Guber, Deborah Lynne. 1999. "Getting What You Pay For: The Debate Over Equity in Public School Expenditures." *Journal of Statistics Education* 7(2).

---

 illiteracy30

*Illiteracy in the Population 10 Years Old and Over, 1930*


---

## Description

This is perhaps the canonical data set for illustrating the ecological fallacy.

## Usage

```
illiteracy30
```

## Format

A data frame with 40 observations on the following 11 variables.

state a character for the state  
 pop a numeric vector for the total population  
 pop\_il a numeric vector for the total population that is illiterate  
 nwhite a numeric vector for the total native white population  
 nwhite\_il a numeric vector for the total native white population that is illiterate  
 fpwhite a numeric vector for the total white population with "foreign or mixed parentage"  
 fpwhite\_il a numeric vector for the total white population with "foreign or mixed parentage" that is illiterate  
 fbwhite a numeric vector for the total foreign-born white population  
 fbwhite\_il a numeric vector for the total foreign-born white population that is illiterate  
 black a numeric vector for the total black population.  
 black\_il a numeric vector for the total black population that is illiterate

## Details

All population totals reflect those 10 years or older. The 1930 Census (along with Robinson (1950)) uses "negro" in lieu of black, but the variable names here eschew that older label. Note that some states are not yet states in the 1930 Census.

**Source**

U.S. Census Bureau (1933). Fifteenth Census of the United States: 1930. Population, Volume II.

**References**

Grotenhuis, Manfred Te, Rob Eisinga, and SV Subramanian. 2011. "Robinson's Ecological Correlations and the Behavior of Individuals: methodological corrections." *International Journal of Epidemiology* 40(4): 1123-25.

Robinson, WS. 1950. "Ecological Correlations and the Behavior of Individuals." *American Sociological Review* 15(3): 351-57.

---

inglehart03

*"How Solid is Mass Support for Democracy—And How Can We Measure It?"*

---

**Description**

A data set based on summary information provided in Inglehart's (2003) article in \*PS: Political Science & Politics\*. These data would be from the article itself and only indirectly from the raw World or European Values Survey.

**Usage**

inglehart03

**Format**

A data frame with 77 observations on the following 4 variables.

state\_year the state year and survey year, as provided in the article

havedem the percentage of respondents saying having a democratic political system is "very good" or "good"

strongleader the percentage of respondents saying having a strong leader unencumbered by elections or parliaments is "very good" or "good"

muslim a dummy variable that equals 1 if Inglehart codes the state as being a "predominantly Islamic society"

**Details**

Data manually entered based on Table 1 and Table 2 in Inglehart's (2003) article.

**References**

Inglehart, Ronald. 2003. "How Solid is Mass Support for Democracy—And How Can We Measure It?" \*PS: Political Science & Politics\* 36(1): 51-57.



Lipset59

*Democracy and Economic Development (Around) 1949-50***Description**

A data set on democracy and economic development for 48 countries that Lipset (1959) first described.

**Usage**

Lipset59

**Format**

A data frame with 48 observations on the following 11 variables.

country a character country for an English country name

cat a category for the country by their region and level of democracy

iso3c a three-character ISO code

wbgdp2011est an estimated gross domestic product in 2011 USD

wbpopest an estimated population size

unpop a population size (in thousands)

uninc a national income (in millions)

unincpc a national income per capita

xm\_qudsest a "Quick UDS" estimate of democracy on a latent scale (see details)

v2x\_polyarchy the Varieties of Democracy "polyarchy" estimate (see details)

polity2 the polity2 score from the Polity project (see details)

**Details**

The three variables with the prefix of un nominally come from the United Nations Statistical Division for 1949/1950, but are actually retrieved from Andic and Peacock (1961). Andic and Peacock (1961) note you should be skeptical of Soviet-style calculations of national income and thus don't include it in the data they make available.

Anything else is explicitly benchmarked to 1950 as a referent year. The GDP and population estimates come by way of Anders et al. (2020). You can manually create your own GDP per capita variable here because the GDP is demarcated in dollars and the population size is in units of 1. Take one and divide it over the other.

The democracy variables are all unique in their own way. The "Quick UDS" estimates are generated to be latent and, globally, have a mean that approximates 0 and a standard deviation that approximates 1. In the regression context, that would mean a coefficient would communicate something like a magnitude change across a standard deviation on the scale. The "polyarchy" estimate has a theoretical minimum of 0 and a theoretical maximum of 1. In the regression context, that would mean a coefficient communicates a min/max effect. The Polity project estimate comes from a usual

scale of -10 to 10 and a regression coefficient communicates something much less exotic. It's a unit change on this scale.

In all cases, higher values of democracy = more "democraticness", for lack of a better term.

## References

- Anders, Therese, Christopher J. Fariss, and Jonathan N. Markowitz. 2020. "Bread Before Guns or Butter: Introducing Surplus Domestic Product (SDP)" *International Studies Quarterly* 64(2): 392–405.
- Andic, Suphan and Alan T. Peacock. 1961. "The International Distribution of Income, 1949 and 1957." *Journal of the Royal Statistical Society*\*. Series A (General) 124(2): 206-218.
- Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Jan Teorell, David Altman, Michael Bernhard, M. Steven Fish, Adam Glynn, Allen Hicken, Anna Luhrmann, Kyle L. Marquardt, Kelly McMann, Pamela Paxton, Daniel Pemstein, Brigitte Seim, Rachel Sigman, Svend-Erik Skaaning, Jeffrey Staton, Agnes Cornell, Lisa Gastaldi, Haakon Gjerlow, Valeriya Mechkova, Johannes von Romer, Aksel Sundtrom, Eitan Tzelgov, Luca Uberti, Yi-ting Wang, Tore Wig, and Daniel Ziblatt. 2020. "V-Dem Codebook v10" Varieties of Democracy (V-Dem) Project.
- Lipset, Seymour Martin. 1959. "Some Social Requisites of Democracy: Economic Development and Political Legitimacy" *American Political Science Review*\* 53(1): 69-105.
- Marshall, Monty G., Ted Robert Gurr, and Keith Jagers. 2017. "Polity IV Project: Political Regime Characteristics and Transitions, 1800-2017." Center for Systemic Peace.
- Marquez, Xavier, "A Quick Method for Extending the Unified Democracy Scores" (March 23, 2016). doi: [10.2139/ssrn.2753830](https://doi.org/10.2139/ssrn.2753830)
- Pemstein, Daniel, Stephen Meserve, and James Melton. 2010. "Democratic Compromise: A Latent Variable Analysis of Ten Measures of Regime Type." *Political Analysis*\* 18(4): 426-449.

---

LOTI

*Land-Ocean Temperature Index, 1880-2022*

---

## Description

These data contain monthly mean temperature anomalies expressed as deviations from the corresponding 1951-1980 means. They are useful for showing how we can measure climate change.

## Usage

LOTI

## Format

A data frame with 1,716 observations on the following 2 variables.

date a date, mostly to contain information for the year and month

value the mean temperature anomaly as deviation from corresponding 1951-1980 mean

**Details**

Data are updated through most recent month, at least for last time I updated it. Data represent combined land-surface air and sea-surface water temperature anomalies. Of note: the day value in the date column has no real value. It was just a way of combining data that are aggregated by year and month.

**Source**

National Aeronautics and Space Administration's Goddard Institute for Space Studies.

---

LTPT

*Long-Term Price Trends for Computers, TVs, and Related Items*

---

**Description**

These data are a monthly time-series of changes in the consumer price index relative to a Dec. 1997 starting date for televisions, computers, and related items. I use this as in-class illustration that globalization has made consumer electronics cheaper across the board for Americans.

**Usage**

LTPT

**Format**

A data frame with 1,704 observations on the following 3 variables.

date a date

category the particular category (e.g. all items, televisions, etc.)

value the consumer price index (Dec. 1997 = 100)

**Details**

This is a web-scraping job from the U.S. Bureau of Labor Statistics. Post is titled "Long-term price trends for computers, TVs, and related items" and was published on Oct. 13, 2015.

**Source**

U.S. Bureau of Labor Statistics.

---

 LTWT

*"Let Them Watch TV"*


---

### Description

"Let Them Watch TV": These data contain price indices for various items for the general urban consumer. Categories include medical services, college tuition, college textbooks, child care, housing, food and beverages, all items (i.e. general CPI), new vehicles, apparel, and televisions. The base period in value was originally the 1982-4 average, but I converted the base period to January 2000. I use these data for in-class discussion about how liberalized trade has made consumer electronics (like TVs) fractions of their past prices. Yet, young adults face mounting costs for college, child-raising, and health care that government policy has failed to address.

### Usage

LTWT

### Format

A data frame with 2377 observations on the following 3 variables.

date a date

category a factor for the particular category

value the price index. Base: January 2000

### Details

Inspiration comes from a blog post titled "Chart of the day (century?): Price changes 1997 to 2017", which was published by the American Enterprise Institute on Feb. 2, 2018.

### Source

Bureau of Labor Statistics, via the blscrapeR package.

---

 min\_wage

*History of Federal Minimum Wage Rates Under the Fair Labor Standards Act, 1938-2009*


---

### Description

A data set on the various federal minimum wage rates.

### Usage

min\_wage

**Format**

A data frame with 23 observations on the following 5 variables.

date a date for when a new minimum wage was introduced

wage the (nominal) value of the wage

**Details**

Data come from the Department of Labor. Wages are taken from wage adjustments from the 1938 act.

**Source**

Department of Labor

---

 mm\_mlda

*Minimum Legal Drinking Age Fatalities Data*

---

**Description**

These are data you can use to replicate the regression discontinuity design analyses throughout Chapter 4 of *Mastering 'Metrics*. Original analyses come from Carpenter and Dobkin (2009, 2011).

**Usage**

mm\_mlda

**Format**

A data frame with 50 observations on the following 19 variables.

agecell a numeric

all a numeric

allfitted a numeric

internal a numeric

internalfitted a numeric

external a numeric

externalfitted a numeric

alcohol a numeric

alcoholfitted a numeric

homicide a numeric

homicidefitted a numeric

suicide a numeric

suicidefitted a numeric

mva a numeric  
 mvafitted a numeric  
 drugs a numeric  
 drugsfitted a numeric  
 externalother a numeric  
 externalotherfitted a numeric

### Details

These data are not well-documented. You guys are on your own here. Good luck.

### References

Carpenter, Christopher and Carlos Dobkin. 2009. "The Effect of Alcohol Consumption on Mortality: Regression Discontinuity Evidence from the Minimum Drinking Age". *American Economic Journal: Applied Economics* 1(1): 164–182.

Carpenter, Christopher and Carlos Dobkin. 2011. "The Minimum Legal Drinking Age and Public Health". *Journal of Economic Perspectives* 25(2): 133–156.

---

mm\_nhis

*Data from the 2009 National Health Interview Survey (NHIS)*

---

### Description

These are data from the 2009 NHIS survey. People who have read *Mastering 'Metrics* should recognize these data. They're featured prominently in that book and the authors' discussion of random assignment and experiments.

### Usage

mm\_nhis

### Format

A data frame with 18790 observations on the following 10 variables.

fml is the respondent a woman?  
 hi a numeric vector for whether respondent has at least some health insurance  
 hlth a numeric vector for a health index, broadly understood  
 nwhite is the respondent not white?  
 age the respondent's age in years  
 yedu the respondent's total years of education  
 famsize the size of the respondent's family  
 empl is the respondent employed  
 inc the respondent's household/family income  
 perweight a numeric vector for weight

**Details**

Data are already cleaned in a way that facilitates an easy replication of Table 1.1 in *Mastering 'Metrics*. Check the book's website for more information.

**Source**

National Health Interview Survey (2009).

---

 mm\_randhie

---

*Data from the RAND Health Insurance Experiment (HIE)*


---

**Description**

These are data from the RAND Health Insurance Experiment (HIE). People who have read *Mastering 'Metrics* should recognize these data. They're featured prominently in that book and the authors' discussion of random assignment and experiments.

**Usage**

mm\_randhie

**Format**

The data are a list of two data frames (or "tibbles"). The first is the baseline data.

platype the plan coverage of the respondent, as a factor

age the age of the respondent

blackhispanic whether the respondent is not white

cholest the cholesterol level of the respondent (in mg/dl)

educper the education-level of the respondent

female whether the respondent is a woman

ghindx a general health index

hosp was the respondent hospitalized last year?

income1cpi the family/household income of the respondent, adjusted for inflation

mhi a mental health index

systol the systolic blood pressure level of the respondent (in mm HG)

The second is the outcome data.

platype the plan coverage of the respondent, as a factor

ftf the number of face-to-face visits for the respondent

out\_inf the total of out-patient expenses for the respondent

totadm the number of hospital admissions for the respondent

tot\_inf the total health expenses for the respondent

**Details**

Data are already cleaned in a way that facilitates an easy replication of Table 1.3 and a partial replication of Table 1.4 in *Mastering 'Metrics*. Check the book's website for more information. I want to note that my treatment of the data leans heavily on Jeff Arnold's treatment of it. Check <https://jrnoold.github.io/masteringmetrics/> for more information. Future updates to the data may pursue a more exhaustive replication. I will only note these data are a mess and the authors of *Mastering 'Metrics* do not do a great job annotating code.

**Source**

RAND Health Insurance Experiment.

---

mvprod

*Motor Vehicle Production by Country, 1950-2019*

---

**Description**

Data, largely from Organisation Internationale des Constructeurs d'Automobiles (OICA), on motor vehicle production in various countries (and the world totals) from 1950 to 2019 at various intervals. Tallies include production of passenger cars, light commercial vehicles, minibuses, trucks, buses and coaches.

**Usage**

mvprod

**Format**

A data frame with three variables

country the country's name

year the year

value the total motor vehicles produced that year

**Details**

This is a Wikipedia web-scraping job. See: [https://en.wikipedia.org/wiki/List\\_of\\_countries\\_by\\_motor\\_vehicle\\_production](https://en.wikipedia.org/wiki/List_of_countries_by_motor_vehicle_production)

**Source**

Organisation Internationale des Constructeurs d'Automobiles (OICA)



---

nesarc\_drinkspd      *The Usual Daily Drinking Habits of Americans (NESARC, 2001-2)*

---

### Description

This toy data set is loosely modified from Wave I of the NESARC data set. Here, my main interest is the number of drinks consumed on a usual day drinking alcohol in the past 12 months, according to respondents in the nationally representative survey of 43,093 Americans.

### Usage

nesarc\_drinkspd

### Format

A data frame with 43093 observations on the following 8 variables.

`idnum` a numeric vector and sequence from 1 to the number of rows in the data

`ethrace2a` a numeric vector for the ethnicity/race. 1 = White, not Hispanic. 2 = Black, not Hispanic. 3 = AI/AN. 4 = Asian, Native Hawaiian, Pacific Islander. 5 = Hispanic or Latino.

`region` a numeric vector for the Census region. 1 = Northeast. 2 = Midwest. 3 = South. 4 = West

`age` a numeric vector for age in years

`sex` a numeric vector for sex. 1 = female. 0 = male

`marital` a numeric vector for marital status. 1 = married. 2 = living with someone as married. 3 = widowed. 4 = divorced. 5 = separated. 6 = never married

`educ` a numeric vector for education level, recoded from `s1q6a` in the original data. 1 = did not make it to/finish high school. 2 = high school graduate or equivalency. 3 = some college, but no four-year degree. 4 = four-year college degree or more.

`s2aq8b` a numeric vector for the number of drinks of any alcohol consumed on days drinking alcohol in the past 12 months. This variable is “as-is” from the original data set.

### Details

You will not want to use the `s2aq8b` variable without recoding it first. Those who cannot recall how much they typically drink (i.e. true “don’t know” or missing info) are coded as 99. Non-drinkers are coded as NA in the `s2aq8b` variable and should be recoded as 0. Any value between 1 and 98 in the variable represents the, for lack of better term, “true” number of alcoholic drinks a respondent says s/he typically consumes on a day drinking alcohol in the past 12 months, though this is evidently preposterous as a count variable. A person drinking 42 alcoholic drinks a day would not be alive to tell you they did this. The researcher may want to employ some sensible right censoring here.

### Source

National Epidemiologic Survey on Alcohol and Related Conditions (NESARC)—Wave 1 (2001–2002)

---

Newhouse77	<i>Medical-Care Expenditure: A Cross-National Survey (Newhouse, 1977)</i>
------------	---

---

**Description**

These are the data in Newhouse's (1977) simple OLS model from 1977. In his case, he's trying to explain medical care expenditures as a function of GDP per capita for these countries. It's probably the easiest OLS model I can find in print because Newhouse helpfully provides all the data in one simple table.

**Usage**

Newhouse77

**Format**

A data frame with 13 observations on the following 5 variables.

country a character vector for the country

year a numeric vector for the year

gdppc a numeric vector for the per capita GDP in USD

medsharegdp a numeric vector for the medical care share as percentage of GDP

medexppc a numeric vector for per capita medical care expenditure (in USD)

**Details**

Table 1 in Newhouse (1977) is well-annotated with background information.

**References**

Newhouse, Joseph P. 1977. "Medical-Care Expenditure: A Cross-National Survey." *Journal of Human Resources* 12(1): 115-125.

---

ODGI	<i>Ozone Depleting Gas Index Data, 1992-2022</i>
------	--

---

**Description**

The NOAA Earth System Research Laboratory has an "ozone depleting gas index" (ODGI) data set from 1992 to 2018. This dataset summarizes Table 1 and Table 2 from its website. The primary interest here (for my purposes) is the ODGI indices (including the new 2012 measure). The data set includes constituent greenhouse gases/chlorines as well in parts per trillion. The primary use here is for in-class illustration.

**Usage**

ODGI

**Format**

A data frame with 62 observations on the following 16 variables.

year the year

cat categorical variable for the Antarctic or Mid-Latitudes measurements

cfc12 CFC-12 concentration in parts per trillion

cfc11 CFC-11 concentration in parts per trillion

ch3c1 chloromethane concentration in parts per trillion

ch3br bromomethane concentration in parts per trillion

cc14 carbon tetrachloride concentration in parts per trillion

ch3cc13 methyl chloroform concentration in parts per trillion

halons aggregate concentration in parts per trillion of H-1211, H-1301 and H-2402

cfc113 trichlorotrifluoroethane concentration in parts per trillion

hcfcs aggregate concentration in parts per trillion of HCFC-22, HCFC-141b, and HCFC-142b

wmo\_minor aggregate concentration in parts per trillion of CFC-114, CFC-115, halon 2402 and halon 1201

sum the sum of all greenhouse gas concentration measurements

eesc includes consideration of lag times for transport and mixing associated with transport. New as of 2012

odgi\_old old greenhouse gas index, no longer supported as of 2012

odgi\_new new greenhouse gas index, as of 2012

**Source**

<https://gml.noaa.gov/odgi/>

---

OODTPT

*Data for "Optimal Obfuscation: Democracy and Trade Policy Transparency"*

---

**Description**

A data set for replicating an argument about the relationship between democracy and tariffs/non-tariff trade barriers.

**Usage**

OODTPT

**Format**

A data frame with 75 observations on the following 16 variables.

country a character vector for the country

isocode a character vector for the three-character ISO code of the country

tariff the mean statutory most favored nation tariff rate

corecov the core non-tariff barrier coverage ratio

qualcov the quality non-tariff barrier coverate ratio

polity the familiar [-10,10] Polity measure of democracy

iec the index of electoral competitiveness from the World Bank

lngdppc real GDP per capita in 1995 dollars

lngdp real GDP in 1995 dollars

lnexpgdp export dependence (i.e. export/GDP ratio)

reer real effective exchange rate

growth GDP per capita growth rate

dimpgdp the change in the import/GDP ratio over the past three years

lngovcons the log of country's government consumption spending as a percentage of GDP

gatt a dummy variable for GATT membership

avgtar the country's average most favored nation tariff rate

**Details**

Data downloaded Joshua Alley's Github repository on simple cross-sectional OLS models. These were originally two separate Stata files that I merged into one. Please read the Kono (2006) article for more information.

**References**

Kono, Daniel. 2006. "Optimal Obfuscation: Democracy and Trade Policy Transparency" \*American Political Science Review\* 100(3): 369-384.

**Description**

A data set on government spending in select rich countries as a function of trade/GDP, financial openness, and the state-year-level engagement in trade unions. The data offer a means to quasi-replicate Garrett's (1998) argument about left-wing governments' ability to stem the tide of globalization's effect on decreased government spending.

**Usage**

PPGE

**Format**

A data frame with the following 9 variables.

country a character vector for the country

iso3c a character vector for the three-character country ISO code

year the year

govtspendgdp total government spending over GDP

tradegdp the volume of trade over GDP

kaopen an index measuring a country's degree of capital account openness

ka\_open an alternate index measuring a country's degree of capital account openness, normalized to be between 0 and 1

v2catrauni an estimate of a country's engagement in independent trade unions, generated by way of a Bayesian item response model

v2catrauni\_ord an estimate of a country's engagement in independent trade unions, on ordinal scale. See details.

**Details**

The data are an unbalanced panel because of data missingness primarily affecting Switzerland (which would only appear in the panel in earnest starting in the mid-1990s). The Netherlands has some missing data in the mid-1970s. Spain and Portugal appear at the start of the panel, though the transition to democracy for both wouldn't start until 1974/1975. The data also have some obvious COVID weirdness for 2020. Perhaps an honest re-assessment of Garrett (1998) may want to drop Switzerland altogether, ignore 2020, and lop a few years off the Spanish and Portuguese panel. What you do with the Netherlands is up to you.

Briefly: the government spending/GDP data come from the International Monetary Fund. The trade/GDP data come from the World Bank's API. The financial openness indicators come by way of the Chinn-Ito index. The engagement in trade unions data are from the Varieties of Democracy project. The ordinal measure of the trade union estimates communicate what percentage of the population is active in independent trade unions. Values include 0) virtually no one 1) a small share of the population (less than 5 moderate share of the population (about 5 to 15 the population (about 16 (about 26

**References**

Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Jan Teorell, Nazifa Alizada, David Altman, Michael Bernhard, Agnes Cornell, M. Steven Fish, Lisa Gastaldi, Haakon Gjerl w, Adam Glynn, Sandra Grahn, Allen Hicken, Garry Hindle, Nina Ilchenko, Katrin Kinzelbach, Joshua Krusell, Kyle L. Marquardt, Kelly McMann, Valeriya Mechkova, Juraj Medzihorsky, Pamela Paxton, Daniel Pemstein, Josefine Pernes, Oskar Ryd n, Johannes von R mer, Brigitte Seim, Rachel Sigman, Svend-Erik Skaaning, Jeffrey Staton, Aksel Sundstr m, Eitan Tzelgov, Yiting Wang, Tore Wig, Steven Wilson and Daniel Ziblatt. 2022. "V-Dem [Country-Year/Country-Date] Dataset v12" Varieties of Democracy (V-Dem) Project. doi: [10.23696/vdemds22](https://doi.org/10.23696/vdemds22)

Chinn, Menzie D. and Hiro Ito. 2006. "What Matters for Financial Development? Capital Controls, Institutions, and Interactions." *Journal of Development Economics* 81(1): 163–192.

Garrett, Geoffrey. 1998. *Partisan Politics in the Global Economy*. New York, NY: Cambridge University Press.

---

PRDEG

*Property Rights, Democracy, and Economic Growth*

---

### **Description**

A data set for replicating David Leblang's (1996) analysis on property rights, democracy, and economic growth.

### **Usage**

PRDEG

### **Format**

A data frame with 147 observations on the following 10 variables.

levine a numeric vector that serves as a cross-section identifier

country a character vector for the country

decade a numeric vector for a decade

private a numeric vector for credit allocated to private enterprise

rgdp a numeric vector for the initial level of real per capita GDP

democ a numeric vector for the level of democracy

pri a numeric vector for primary school attainment

sec a numeric vector for secondary school attainment

grow a numeric vector for per capita growth rate

xcontrol a numeric vector for exchange controls

### **Details**

Data come Joshua Alley's Github repository on cross-sectional OLS regressions. Please read David Leblang's (1996) article for some more detail about the variables included in the model.

### **References**

Leblang, David. 1996. "Property Rights, Democracy, and Economic Growth." 49(1): 5-26.

---

Presidents	<i>U.S. Presidents and Their Terms in Office</i>
------------	--

---

**Description**

This should be self-evident. Here are all U.S. presidents who have completed their terms in office (i.e. excluding the current one).

**Usage**

Presidents

**Format**

A data frame with 45 observations on the following 3 variables.

president the president

start the start date of the term, as a date

end the end date of the term, as a date

**Details**

I scraped this from <https://www.presidentsusa.net/presvplist.html>. Data frame is capital-P "Presidents" to avoid a conflict with the presidents data frame from the datasets package.

---

pwt_sample	<i>Penn World Table (10.0) Macroeconomic Data for Select Countries, 1950-2019</i>
------------	---

---

**Description**

These are some macroeconomic data for 21 select (rich) countries. I've used these data before to discuss issues of grouping and skew in cross-sectional data.

**Usage**

pwt\_sample

**Format**

A data frame with 1470 observations on the following 11 variables.

country the country name

isocode The country's ISO code

year a numeric vector for the year

pop Population in millions

hc Index of human capital per person, based on years of schooling and returns to education

rgdpna Real GDP at constant 2011 national prices (in million 2017 USD)

rgdpo Output-side real GDP at chained PPPs (in million 2017 USD)

rgdpe Expenditure-side real GDP at chained PPPs (in million 2017 USD)

labsh Share of labor compensation in GDP at current national prices

avh Average annual hours worked by persons engaged.

emp Number of persons engaged (in millions)

rna Capital stock at constant 2017 national prices (in million 2017 USD)

**Source**

Taken from the pwt10 package. See: <https://www.rug.nl/ggdc/>

---

quartets

*Anscombe's (1973) Quartets*

---

**Description**

These are four x-y data sets, combined into a long format, which have the same traditional statistical properties (mean, variance, correlation, regression line, etc.). However, they look quite different.

**Usage**

quartets

**Format**

A data frame with 44 observations on the following 3 variables.

group a categorical identifier for the quartet

x a continuous variable

y a continuous variable

**Details**

Data come default in R, but I elected to change the format to be a bit more accessible.



**References**

Ansambe, Francis J. (1973). "Graphs in Statistical Analysis." *The American Statistician* 27: 17–21.

---

recessions

*United States Recessions, 1855-present*

---

**Description**

Data on U.S. recessions, past to present. Data include information on contraction, expansion, and cycle.

**Usage**

recessions

**Format**

A data frame with 35 observations on the following 8 variables.

peak the year-month of the peak, as a date

trough the year-month of the trough, as a date

peakq the peak quarter

troughq the trough quarter

p2t peak to trough (in months)

prev\_t2p previous trough to this peak (in months)

tfpt trough from previous trough (in months)

pfpp peak from previous peak (in months)

**Details**

Data come from via scraping job of <https://www.nber.org/research/data/us-business-cycle-expansions-and-con>

**Source**

National Bureau of Economic Research (NBER)

---

 SBCD

*Systemic Banking Crises Database II*


---

### Description

A data set on banking, currency, debt, and debt-restructuring crises from 1970 to 2017.

### Usage

SBCD

### Format

A data frame with 574 observations on the following 4 variables.

country the country, as it appears in the data

type the type of crisis, entered here as "banking", "currency", "debt", or "debt restructuring"

year the year of the crisis

month the month the crisis started, if known

### Details

Data are cobbled from the second and third sheets of the spreadsheet the authors provide. Country names are as entered in their spreadsheet. Liberia has an "NA" in the raw data for sovereign debt restructuring and I don't know why. I elect to keep it.

### References

Laeven, Luc and Fabian Valencia. 2020. "Systemic Banking Crises Database II". \*IMF Economic Review\* 68: 307–361.

---

 scb\_regions

*Region Codes in the Central Bureau of Statistics ("Statistiska centralbyrån") in Sweden*


---

### Description

This is a simple data set for matching region codes to the names of territorial units in Sweden, at least recorded/cataloged by the Central Bureau of Statistics in Sweden.

### Usage

scb\_regions

**Format**

A data frame with 312 observations on the following 2 variables.

region an intuitive name for a territorial unit/"region" in Sweden

region\_code an alpha-numeric code coinciding with the territorial unit/"region"

**Details**

Data were manually derived from first gathering everything the Central Bureau of Statistics had to offer. Its intended use is alongside the **pxweb** package. May it allow for more focused uses of the package without having to rely on the interactive component to do all the heavy-lifting.

---

 SCP16

---

*South Carolina County GOP/Democratic Primary Data, 2016*


---

**Description**

County-level data on vote share and various background/demographic information for the 2016 South Carolina GOP/Democratic primaries.

**Usage**

SCP16

**Format**

A data frame with 46 observations on the following 15 variables.

county the county

clinton Hillary Clinton's county-level vote share in the 2016 party primary

sanders Bernie Sanders' county-level vote share in the 2016 party primary

trump Donald Trump's county-level vote share in the 2016 party primary

cruz Ted Cruz' county-level vote share in the 2016 party primary

rubio Marco Rubio's county-level vote share in the 2016 party primary

percapinc A county-level estimate for per capita income

medhouseinc A county-level estimate for the median household income

medfaminc A county-level estimate for the median family income

illiteracy An estimate of the percent of the county lacking "basic" prose literacy skills

perblack Percentage of the county that is black

population An estimate of the county-level population

romneyshare2012 Mitt Romney's vote share at the county-level from the 2012 general election

perhsgrad Percentage of the county whose residents 25 years and older have at least a high school education

unemployment Unemployment rate for the county for January 2016

**Details**

The illiteracy estimate comes from a Department of Education report from 2003. The unemployment rate data come from the Bureau of Labor Statistics. A Github repository contains more information: <https://github.com/svmiller/sc-primary-2016>.

---

 sealevels

*Global Average Absolute Sea Level Change, 1880–2015*


---

**Description**

These data describe how sea level has changed over time, in both relative and absolute terms. Absolute sea level change refers to the height of the ocean surface regardless of whether nearby land is rising or falling.

**Usage**

```
sealevels
```

**Format**

A data frame with 136 observations on the following 5 variables.

year the year

adjlev adjusted sea level (in inches)

lb the lower bound of the estimate (in inches)

ub the upper bound of the estimate (in inches)

adjlev\_noaa NOAA's adjusted sea level (in inches)

**Source**

Environmental Protection Agency ("Climate Change Indicators: Sea Level")

**References**

CSIRO (Commonwealth Scientific and Industrial Research Organisation). 2015 update to data originally published in: Church, J.A., and N.J. White. 2011. Sea-level rise from the late 19th to the early 21st century. *Surv. Geophys.* 32:585–602.

NOAA (National Oceanic and Atmospheric Administration). 2016. Laboratory for Satellite Altimetry: Sea level rise. Accessed June 2016.

---

so2concentrations	<i>Sulfur Dioxide Emissions, 1980-2020</i>
-------------------	--

---

**Description**

This data set contains yearly observations by the Environmental Protection Agency on the concentration of sulfur dioxide in parts per billion, based on 32 sites. I use this for in-class illustration. Note that the national standard is 75 parts per billion. Data are the national trend.

**Usage**

so2concentrations

**Format**

A data frame with the following 4 variables.

year the year

value the mean concentration of sulfur dioxide in the air based on 32 trend sites, in parts per billion

ub the lower bound of the value (10th percentile)

lb the upper bound of the value (90th percentile)

**Source**

Environmental Protection Agency ("Sulfur Dioxide Trends")

---

steves_clothes	<i>Steve's (Professional) Clothes, as of March 20, 2022</i>
----------------	---

---

**Description**

I cobbled together this data set of the professional clothes (polos, long-sleeve dress shirts, pants) in my closet, largely for illustration on the origins of apparel in the U.S. for an intro lecture on trade.

**Usage**

steves\_clothes

**Format**

A data frame with 86 observations on the following 4 variables.

type Type of clothing

brand The brand of clothing (e.g. Apt. 9, Saddlebred)

color the color (and/or pattern) of the article of clothing

origin The country that produced the garment.

**Details**

If you must know, I do most of my clothes shopping at major retailers in the U.S. This is mostly Belk, J.C. Penney, and Kohl's. If that's you as well, the odds are good the distribution of my clothes will closely resemble yours. A recent move I made resulted in me donating a fair bit of my short-sleeved polo shirts. I did not buy any new shirts, though. Thus, I copied that information from a previous version of the data.

**Source**

Steve's closet. Hey, that's me!

---

sugar\_price

*IMF Primary Commodity Price Data for Sugar*

---

**Description**

This is primary commodity price data for sugar globally, in the United States, and in Europe for every month from 1980 to (roughly) the present. Prices are nominal U.S. cents per pound and are not seasonally adjusted ("NSA").

**Usage**

sugar\_price

**Format**

A data frame with 1,316 observations on the following 3 variables.

date a date

category the category (either the U.S., global, or Europe)

value the price of sugar in U.S. cents per pound (NSA, nominal)

**Details**

The price data for Europe do not appear to be updated as regularly as the global and U.S. prices. Thus, the last month in the data for Europe are June 2017. For that reason, I elected to make a data set of these data for posterity's sake.

**Source**

International Monetary Fund

---

sweden\_counties      *The Counties of Sweden*

---

**Description**

A simple data set on Sweden's counties.

**Usage**

sweden\_counties

**Format**

A data frame with 21 observations on the following 6 variables.

iso the ISO 3166-2 code for the county

nuts the Nomenclature of Territorial Units for Statistics (NUTS) code for the county

county the name of the county, in Swedish

centre the administrative centre, or centres, of the county

area the size of the county in square kilometers

pop2019 the size of the county in 2019

**Details**

This is a simple Wikipedia scrape job from 7 November 2022.

---

thatcher\_approval      *Margaret Thatcher Satisfaction Ratings, 1980-1990*

---

**Description**

A data set on satisfaction/dissatisfaction ratings during Margaret Thatcher's tenure as prime minister.

**Usage**

thatcher\_approval

**Format**

A data frame with 125 observations on the following 8 variables.

`poll_date` the effective "date" of the public opinion poll

`date` a date for the poll, to make for easier plotting

`govt_sat` the percentage of respondents saying they were satisfied with the government

`govt_dis` the percentage of respondents saying they were dissatisfied with the government

`thatcher_sat` the percentage of respondents saying they were satisfied with Margaret Thatcher

`thatcher_dis` the percentage of respondents saying they were dissatisfied with Margaret Thatcher

`opp_sat` the percentage of respondents saying they were satisfied with the leader of the opposition

`opp_dis` the percentage of respondents saying they were dissatisfied with the leader of the opposition

**Details**

Data come from Ipsos. "Leader of the opposition" was typically named in the exact poll. In the lifetime of this series, the leader of the opposition was James Callaghan until Nov. 10 1980. Thereafter, it was Michael Foot until Oct. 2 1983. Neil Kinnock replaces him for the duration of this series. Interpret "leader of the opposition" with that in mind.

The date variable is, again, for simple convenience to make for easier plotting. In the absence of a specific day provided by Ipsos, the poll benchmarks to the first of the month. In the case of a known period of days, it benchmarks to the first day provided.

---

therms

*Thermometer Ratings for Donald Trump and Barack Obama*

---

**Description**

A data set on thermometer ratings for Donald Trump and Barack Obama in 2020. I use these data for in-class illustration of central limit theorem. Basically: the sampling distribution of a population is normal, even if the underlying population is decidedly not.

**Usage**

therms

**Format**

A data frame with 3080 observations on the following 2 variables.

`fttrump1` a thermometer rating for Donald Trump [0:100]

`ftobama1` a thermometer rating for Barack Obama [0:100]

**Details**

The survey period was April 10-18, 2020 and was done entirely online.



**Source**

American National Election Studies (ANES) Exploratory Testing Survey (ETS)

---

turnips

*Turnip prices in Animal Crossing (New Horizons)*

---

**Description**

A data set on turnip prices from my experience with Animal Crossing (New Horizons)

**Usage**

turnips

**Format**

A data frame with the following 3 variables.

date a date

time a character vector referring to the particular time period of observation

price a numeric vector for the price of turnips, in bells

**Details**

Sunday prices are set for purchase and do not fluctuate. Timmy and Tommy do not accept turnips on Sunday either. Daily prices fluctuate both at opening on Nook's Cranny and at noon. This amounts to three time periods in the data. "5:00 a.m." is reserved only for Sunday purchases (i.e. when Daisy Mae arrives on the island). 8:00 a.m. is the morning price because that is when Nook's Cranny opens. 12:00 p.m. is when the price changes for the day.

Explanations for missing dates: Timmy and Tommy were renovating the shop on May 6, 2021. My wife was diagnosed with cancer and my mother in law went to the hospital on the afternoon of Dec. 27, 2021. I did not get to play the game on Jan. 9, 2022 because of errands I was running for my wife. I plain forgot to check on Feb. 7, 2022.

TV16

*The Individual Correlates of the Trump Vote in 2016***Description**

These data come from the 2016 CCES and allow interested students to model the individual correlates of the Trump vote in 2016. Code/analysis heavily indebted to a 2017 analysis I did on my blog (see references).

**Usage**

TV16

**Format**

A data frame with 64600 observations on the following 21 variables.

`uid` a numeric vector, a unique identifier for the respondent as they first appear in the CCES data.

`state` a character vector for the state in which the respondent resides

`votetrump` a numeric that equals 1 if the respondent voted says s/he voted for Trump in 2016.

`age` a numeric vector for age that is roughly calculated as 2016 - `birthyr`, as it's coded in the CCES data.

`female` a numeric that equals 1 if the respondent is a woman

`collegeed` a numeric vector that equals 1 if the respondent says s/he has a college degree

`racef` a character vector for the race of the respondent

`famincr` a numeric vector for the respondent's household income. Ranges from 1 (Less than \$10,000) to 12 (\$150,000 or more).

`ideo` a numeric vector for the respondent's ideology on a liberal-conservative discrete scale. 1 = very liberal. 5 = very conservative.

`pid7na` a numeric vector for the respondent's partisanship on the familiar 1-7 scale. 1 = Strong Democrat. 7 = Strong Republican. Other party supporters (e.g. libertarians) are coded as NA.

`bornagain` a numeric vector for whether the respondent self-identifies as a born-again Christian.

`religimp` a numeric vector for the importance of religion to the respondent. 1 = not at all important. 4 = very important.

`churchatd` a numeric vector for the extent of church attendance for the respondent. 1 = never. 6 = more than once a week.

`prayerfreq` a numeric vector for the frequency of prayer for the respondent. 1 = never. 7 = several times a day.

`angryracism` a numeric vector for how angry the respondent is that racism exists. 1 = strongly agree (i.e. is angry racism exists). 5 = strongly disagree.

`whiteadv` a numeric vector for agreement with statement that white people have advantages over others in the U.S. 1 = strongly agree. 5 = strongly disagree.

- fearraces a numeric vector for agreement with statement that the respondent fears other races. 1 = strongly disagree. 5 = strongly agree.
- racerare a numeric vector for agreement with statement that racism is rare in the U.S. 1 = strongly disagree. 5 = strongly agree.
- lrelig a numeric vector that serves as a latent estimate for religiosity from the bornagain, religimp, churchatd, and prayerfreq variables. Higher values = more religiosity.
- lcograc a numeric vector that serves as a latent estimate for cognitive racism. This is derived from the racerare and whiteadv variables.
- lemprac a numeric vector that serves as a latent estimate for empathetic racism. This is derived from the fearraces and angryracism variables.

### Details

The latent estimates for religiosity, cognitive racism, and empathetic racism come from a graded response model estimated in mirt. The concepts of "cognitive racism" and "empathetic racism" come from DeSante and Smith.

### Source

Cooperative Congressional Election Study, 2016

### References

<http://svmiller.com/blog/2017/04/age-income-racism-partisanship-trump-vote-2016/>  
<https://github.com/svmiller/2016-cces-trump-vote/blob/master/1-2016-cces-trump.R>

---

ukg\_eeri

*United Kingdom Effective Exchange Rate Index Data, 1990-2022*

---

### Description

This is a (near) daily data set on the effective exchange rate index for the United Kingdom's pound sterling from 1990 onward. The data are indexed, such that 100 equals the monthly average in January 2005. This is useful for illustrating devaluations of the pound after Black Wednesday, the financial crisis, and, more recently, the UK's separation from the European Union.

### Usage

ukg\_eeri

### Format

A data frame with 8318 observations on the following 2 variables.

date a date

value a numeric vector for the effective exchange rate index (Jan. 2005 = 100)

**Details**

Credit to the Bank of England for making these data readily available and accessible. The Bank of England's website (<https://www.bankofengland.co.uk/>) has these data with a code of XUDLBK67.

**Source**

Bank of England

---

uniondensity

*Cross-National Rates of Trade Union Density*

---

**Description**

Cross-national data on relative size of the trade unions and predictors in 20 countries. This is a data set of interest to replicating Western and Jackman (1994), who themselves were addressing a debate between Wallerstein and Stephens on which of two highly correlated predictors explains trade union density.

**Usage**

uniondensity

**Format**

A data frame with 20 observations on the following 5 variables.

`country` a character vector for the country

`union` a numeric vector for the percentage of the total number of wage and salary earners plus the unemployed who are union members, measured between 1975 and 1980, with most of the data drawn from 1979.

`left` a numeric vector tapping the extent to which parties of the left have controlled governments since 1919, due to Wilensky (1981).

`size` a numeric vector measuring the log of labor force size, defined as the number of wage and salary earners, plus the unemployed.

`concen` a numeric vector measuring the percentage of employment, shipments, or production accounted for by the four largest enterprises in a particular industry, averaged over industries (with weights proportional to the size of the industry) and the resulting measure is normalized such that the United States scores a 1.0, and is due to Pryor (1973). Some of the scores on this variable are imputed using procedures described in Stephens and Wallerstein (1991, 945).

**Details**

Data documentation are derived from Simon Jackman's `pscl` package. I just tidied up the presentation a bit.

**Source**

- Pryor, Frederic. 1973. Property and Industrial Organization in Communist and Capitalist Countries. Bloomington: Indiana University Press.
- Stephens, John and Michael Wallerstein. 1991. Industrial Concentration, Country Size and Trade Union Membership. *American Political Science Review* 85:941-953.
- Western, Bruce and Simon Jackman. 1994. Bayesian Inference for Comparative Research. *American Political Science Review* 88:412-423.
- Wilensky, Harold L. 1981. Leftism, Catholicism, Democratic Corporatism: The Role of Political Parties in Recent Welfare State Development. In *The Development of Welfare States in Europe and America*, ed. Peter Flora and Arnold J. Heidenheimer. New Brunswick: Transaction Books.

**References**

- Jackman, Simon. 2009. *Bayesian Analysis for the Social Sciences*. Wiley: Hoboken, New Jersey.

---

usa\_chn\_gdp\_forecasts *United States-China GDP and GDP Forecasts, 1960-2050*

---

**Description**

This is a toy data set to examine the time in which we should expect China to overtake the United States in total gross domestic product (GDP), given current trends. It includes an OECD long-term GDP forecast from 2014, and forecasts from the forecast and prophet packages in R.

**Usage**

usa\_chn\_gdp\_forecasts

**Format**

A data frame with 182 observations on the following 12 variables.

country a character vector (United States, China)

year a numeric vector for the year

p\_gdp y-hats (forecasted GDP) from a prophet forecast

p\_lo80 lower bound (80%) of y-hats (forecasted GDP) from a prophet forecast

p\_hi80 upper bound (80%) of y-hats (forecasted GDP) from a prophet forecast

gdp observed GDP, made available to the World Bank and OECD national accounts data. Available from 1960 to 2019.

f\_gdp forecasted GDP from 2020 to 2050, from the forecast package

f\_lo80 lower bound (80%) forecasted GDP from 2018 to 2050, from the forecast package

f\_hi80 upper bound (80%) forecasted GDP from 2018 to 2050, from the forecast package

f\_lo95 lower bound (95%) forecasted GDP from 2018 to 2050, from the forecast package

f\_hi95 upper bound (95%) forecasted GDP from 2018 to 2050, from the forecast package

oecd\_ltgdpf long-term GDP forecast from the OECD via the OECD Outlook No 95 - May 2014

**Details**

Forecasts from the forecast package and prophet package are rudimentary and bare minimum forecasts based on previous values to that point. Notice the forecast forecasts have a prefix of f\_ and the prophet forecasts have a prefix of p\_. Forecasts are not meant to be exhaustive (clearly), only illustrative for in-class discussion about the "Rise of China." Forecasts made in R on Nov. 20, 2020.

**Source**

OECD Outlook No 95 - May 2014 - Long-term baseline projections provided by Organisation for Economic Co-operation and Development (OECD)

---

usa\_computers

*Percentage of U.S. Households with Computer Access, by Year*

---

**Description**

This is a simple and regrettably incomplete time-series on the percentage of U.S. households with access to a computer, by year.

**Usage**

usa\_computers

**Format**

A data frame with 19 observations on the following 2 variables.

year the year

value the estimated percentage of households with access to a computer

**Details**

Data are spotty and regrettably this is not a perfect time-series. However, it is useful for an in-class exercise to show that the proliferation of household computers (over time) in the United States comes in part because of globalization. Use it for that purpose. The data are reasonably faithful, but don't treat it as gospel. Exact sourcing available upon request.

**Source**

Various: U.S. Census Bureau, Current Population Survey, and American Community Survey

---

usa_migration	<i>U.S. Inbound/Outbound Migration Data, 1990-2017</i>
---------------	--

---

**Description**

This data set contains counts/estimates for the number of inbound migrants in the U.S as well as outbound migrants of American origin to other countries from 1990 to 2017.

**Usage**

```
usa_migration
```

**Format**

A data frame with 3535 observations on the following 5 variables.

year a numeric vector for 1990, 1995, 2000, 2005, 2010, 2015, 2017

country a character vector/constant for the United States

category a character vector for whether the count is inbound to the U.S. from the area variable or outbound (i.e. American expats) to the area variable in a given year.

area a character vector for the area of origin (if category == "Inbound") or destination for American migrants (if category == "Outbound")

count a numeric vector for the count of inbound/outbound migrants

**Details**

"Cote d'Ivoire", "Curacao", and "Reunion" originally had UTF-8 characters, which were removed for maximal compliance with CRAN. CRAN raises a note for every non-ASCII character it sees.

**Source**

United Nations Population Division (DESA)

---

usa_states	<i>State Abbreviations, Names, and Regions/Divisions</i>
------------	--

---

**Description**

A simple data set from state.abb, state.name, state.region, and state.division (+ District of Columbia). I'd rather just have all these in one place.

**Usage**

```
usa_states
```

**Format**

A data frame with 51 observations on the following 4 variables.

stateabb the state abbreviation

statename the state's name

region the state's Census region

division the state's Census division

---

usa_tradegdp	<i>U.S. Trade and GDP, 1790-2018</i>
--------------	--------------------------------------

---

**Description**

A yearly data set on U.S. trade and GDP from 1790 to 2018. Data also include a population variable to facilitate per capita adjustments, if the user sees it useful.

**Usage**

```
usa_tradegdp
```

**Format**

A data frame with 229 observations on the following 5 variables.

year the year

gdpb U.S. GDP (nominal, in billions)

pop Population of the U.S. (in thousands)

impo The value of U.S. imports (in billions)

expo The value of U.S. exports (in billions)

**Details**

Data come from various sources (see, especially: <https://econdataus.com/tradeall.html>). Post-1989 data come from the U.S. Census Bureau. 2018 GDP comes from the IMF. 2018 population estimate comes from the U.S. Census Bureau.



---

voteincome	<i>Sample Turnout and Demographic Data from the 2000 Current Population Survey</i>
------------	--

---

**Description**

A data set on turnout and demographic data from the 2000 Current Population Survey. This is a basic port of the voteincome data from the **Zelig** package.

**Usage**

```
voteincome
```

**Format**

A data frame with 1500 observations on the following 7 variables.

state a character variable for the state, either Arkansas (AK) or South Carolina (SC)

year a numeric constant for the year (2000)

vote a dummy variable for whether the person voted (1) or did not vote

income a numeric variable for income ranging from 4 (less than \$5000) to 17 (greater than \$75000)

education a numeric variable for educational attainment ranging from 1 (less than high school education) to 4 (more than college education)

age a numeric variable for the respondent's age in years, ranging from 18 to 85

female a dummy variable for whether the respondent is a woman (1) or a man (0)

**Details**

Data come from the 2000 Current Population Survey by way of the **Zelig** package. Data should not be used for inferential applications, only for pedagogical purposes. See the appropriate CPS codebook for more information on variable coding (especially for income and education). In all likelihood, age is right-censored as well.

---

wbd_example	<i>A Simple Panel drawn from World Bank Open Data</i>
-------------	---

---

**Description**

A simple data set drawn from World Bank Open Data. I'll use it to illustrate some merge issues you might encounter in panel data.

**Usage**

```
wbd_example
```

**Format**

A data frame with 4537 observations on the following 7 variables.

country an English name for the country/territorial unit

iso2c the two-character ISO code for the country/territorial unit

iso3c the three-character ISO code for the country/territorial unit

year the year of observation

rgdppc the real GDP per capita of the country/territorial unit in that year

lifeexp the average life expectancy at birth for men and women for the country that year

hci the human capital index [0-1] for the country that year

**Details**

The idea for this data comes by way of a student encounter where we noticed this issue. Data were further generated by the wonderful **WDI** package. The underlying data come from the World Bank national accounts (GDP per capita), World Bank analyst estimates (human capital index), or the United Nations Population Division (life expectancy at birth).

---

wvs\_ccodes

*Syncing Word Values Survey Country Codes with CoW Codes*

---

**Description**

A simple data set that syncs World Values Survey country codes (s003) with corresponding country codes from the Correlates of War state system membership data.

**Usage**

wvs\_ccodes

**Format**

A data frame with 112 observations on the following 3 variables.

s003 the World Values Survey country code

country a character vector for the corresponding country name

ccode the equivalent country code from the Correlates of War state system membership data

**Details**

<http://svmiller.com/blog/2015/06/syncing-word-values-survey-country-codes-with-cow-codes/>

---

wvs\_immig

*Attitudes about Immigration in the World Values Survey*


---

**Description**

A data set on attitudes about immigration for all observations in the third to sixth wave of the World Values Survey. I use these data for in-class illustration.

**Usage**

wvs\_immig

**Format**

A data frame with 310,388 observations on the following 6 variables.

s002 the World Values Survey wave

s003 the World Values Survey country code

country the country name

s020 the survey year

uid a unique identifier for the survey respondent

e143 an attitude about immigration policy in the World Values Survey

**Details**

1 = "let anyone come". 2 = "as long as jobs are available". 3 = "strict limits". 4 = "Prohibit people from coming" for the e143 variable. See ?wvs\_ccodes for more information about naming/identifying countries.

---

wvs\_justifbribe

*Attitudes about the Justifiability of Bribe-Taking in the World Values Survey*


---

**Description**

A data set on attitudes about the justifiability of bribe-taking for all observations in the third to sixth wave of the World Values Survey. I use these data for in-class illustration about seemingly interval-level, but information-poor measurements.

**Usage**

wvs\_justifbribe

**Format**

A data frame with 348532 observations on the following 6 variables.

s002 the World Values Survey wave

s003 the World Values Survey country code

country the country name

s020 the survey year

uid a unique identifier for the survey respondent

f117 an attitude about the justifiability of bribe-taking in the World Values Survey

**Details**

1 = "never justifiable". 10 = "always justifiable". Increasing values on this 1-10 scale imply increasing permissiveness for the respondent toward this particular/blatant form of corruption.

---

wvs_usa_abortion	<i>Attitudes on the Justifiability of Abortion in the United States (World Values Survey, 1982-2011)</i>
------------------	--

---

**Description**

A data set on attitudes about the justifiability of abortion in the United States based on World Values Survey responses recorded across six waves (from 1982 to 2011). I assembled this data frame probably around 2014 and routinely use it for in-class illustration about regression, post-estimation simulation, quantities of interest, and how to think about modeling a dependent variable that is on a 1-10 scale, but has curious heaping patterns.

**Usage**

wvs\_usa\_abortion

**Format**

A data frame with 10387 observations on the following 16 variables.

wvsccode the country code for the United States (a numeric constant)

wave the survey wave

year the survey year corresponding to the survey wave

aj the justifiability of abortion on a 1-10 scale (1 = never justifiable; 10 = always justifiable)

age the age of the respondent in years

collegeed a dummy variable that equals 1 if the respondent graduated from college

female a dummy variable that equals 1 if the respondent is a woman

unemployed a dummy variable that equals 1 if the respondent is unemployed

ideology the ideological self-placement of the respondent on a 1-10 scale (1 = furthest to the left; 10 = furthest to the right)

satisfinancial the respondent's financial satisfaction with his/her life (1 = most dissatisfied; 10 = most satisfied)

postma4 the post-materialist index for the respondent (-1 = materialist; 0 = mixed, 1 = post-materialist)

cai the child autonomy index, which ranges from -2 to 2

trustmostpeople can most people be trusted (1) or "(you) never can be too careful" (0)

godimportant the importance of God to the respondent on a 1-10 scale (1 = God is not at all important; 10 = God is most important)

respectauthority would more respect for authority be a welcome change to the United States?

nationalpride a dummy that equals 1 if the respondent is very proud to be an American.

### Details

Data come from the World Values Survey. Note that the college education variable is curiously NA until the third survey wave. The child autonomy index ranges from -2 to 2 where increasing values indicate that children should learn determination and independence over obedience and religious faith. The respectauthority variable is coded where -1 means the respondent believes greater respect for authority in the United States as a future change to the country would be a bad thing. 0 means the respondent doesn't mind such a change. 1 = the respondent believes it would be a good thing.

---

wvs\_usa\_educat

*Education Categories for the United States in the World Values Survey*

---

### Description

This is a simple data set that summarizes what the education codes are in the World Values Survey for the United States.

### Usage

wvs\_usa\_educat

### Format

A data frame with 42 observations the following 6 variables.

x025 the numeric code for supposedly the highest educational level attained

x025cswvs the numeric code for supposedly the education-level attained by the respondent, with country-specific categories

n the number of observations in the World Values Survey with that unique x025cswvs code

x025cswvsmeaning the meaning behind the unique x025cswvs code

x025meaning the meaning behind the unique x025 code

educat a standardized categorical variable corresponding with that unique x025cswvs code

## Details

Observations taken from the combined seven waves of survey data made available by the World Values Survey, but isolated to just the United States. The World Values Survey unfortunately did not collect information about the education-level of the respondent in the 1981 and 1990 waves. These education categories feature in the Miller and Davis (2020) article in *Journal of Ethnicity, and Politics*, albeit before the release of the seventh wave.

## References

Miller, Steven V. and Nicholas T. Davis. Forthcoming. "The Effect of White Social Prejudice on Support for American Democracy." *Journal of Race, Ethnicity, and Politics*.

---

wvs_usa_regions	<i>Region Categories for the United States in the World Values Survey</i>
-----------------	---

---

## Description

This is a simple data set that summarizes what the region codes are in the World Values Survey for the United States.

## Usage

```
wvs_usa_regions
```

## Format

A data frame with 63 observations the following 6 variables.

x048wvs the numeric code for supposedly the region in which the interview was conducted

x048wvsmeaning the meaning behind the unique x048wvs code

stateabb the corresponding state abbreviation (if available) for the unique x048wvs code

statename the corresponding state abbreviation (if available) for the unique x048wvs code

division the corresponding division for the unique x048wvs code

region the corresponding region for the unique x048wvs code

## Details

The region codes are a mess. Some of these are informed guesses. For example, I assume "North-west" means "Pacific" and that Idaho was not included in that category. I make a similar assumption that "Rocky Mountain state" means "Mountain".

---

yugo\_sales

*Yugo Sales in the United States, 1985-1992*

---

**Description**

A data set on Yugo sales against two competing models in the United States from 1985 to 1992.

**Usage**

yugo\_sales

**Format**

A data frame with 24 observations on the following 3 variables.

year the year

car the car type, either the Hyundai Excel, Yugo, or Toyota Tercel

sales the number of units sold in the United States

**Details**

Data come from a website then known as `carsalesbase.com`. I'm aware the inclusion of the Tercel is questionable since the third generation of Tercels were quite different from the first and second generations. However, I use these data to illustrate how poorly the Yugo fared against competing models, including the first and second generation Tercels. I think the inclusion is fair for that purpose.

# Index

## \* datasets

af\_crime93, 3  
aluminum\_premiums, 4  
anes\_partytherms, 5  
anes\_prochoice, 6  
anes\_vote84, 7  
Arca, 8  
arcticseaice, 9  
arg\_tariff, 10  
asn\_stats, 10  
CFT15, 11  
clermson\_temps, 12  
co2emissions, 13  
coffee\_imports, 14  
coffee\_price, 15  
commodity\_prices, 15  
CP77, 17  
Datasaurus, 17  
Dee04, 18  
DJIA, 19  
DST, 20  
eight\_schools, 21  
election\_turnout, 21  
eq\_passengercars, 22  
ESS10N0, 23  
ESS9GB, 24  
ESSBE5, 26  
eurostat\_codes, 27  
eustates, 27  
fakeAPI, 28  
fakeHappiness, 29  
fakeLogit, 30  
fakeTSCS, 30  
fakeTSD, 31  
ghp100k, 32  
GHR04, 33  
gss\_abortion, 34  
gss\_spending, 35  
gss\_wages, 37  
Guber99, 38  
illiteracy30, 39  
inglehart03, 40  
Lipset59, 41  
LOTI, 42  
LTPT, 43  
LTWT, 44  
min\_wage, 44  
mm\_mlda, 45  
mm\_nhis, 46  
mm\_randhie, 47  
mvprod, 48  
nesarc\_drinkspd, 49  
Newhouse77, 50  
ODGI, 50  
OODTPT, 51  
PPGE, 52  
PRDEG, 54  
Presidents, 55  
pwt\_sample, 55  
quartets, 56  
recessions, 57  
SBCD, 58  
scb\_regions, 58  
SCP16, 59  
sealevels, 60  
so2concentrations, 61  
steves\_clothes, 61  
sugar\_price, 62  
sweden\_counties, 63  
thatcher\_approval, 63  
therms, 64  
turnips, 65  
TV16, 66  
ukg\_eeri, 67  
uniondensity, 68  
usa\_chn\_gdp\_forecasts, 69  
usa\_computers, 70  
usa\_migration, 71



- usa\_states, 71
  - usa\_tradegdp, 72
  - voteincome, 73
  - wbd\_example, 73
  - wvs\_ccodes, 74
  - wvs\_immig, 75
  - wvs\_justifbribe, 75
  - wvs\_usa\_abortion, 76
  - wvs\_usa\_educat, 77
  - wvs\_usa\_regions, 78
  - yugo\_sales, 79
- af\_crime93, 3
- aluminum\_premiums, 4
- anes\_partytherms, 5
- anes\_prochoice, 6
- anes\_vote84, 7
- Arca, 8
- arcticseaice, 9
- arg\_tariff, 10
- asn\_stats, 10
- CFT15, 11
- clemson\_temps, 12
- co2emissions, 13
- coffee\_imports, 14
- coffee\_price, 15
- commodity\_prices, 15
- CP77, 17
- Datasaurus, 17
- Dee04, 18
- DJIA, 19
- DST, 20
- eight\_schools, 21
- election\_turnout, 21
- eq\_passengercars, 22
- ESS10NO, 23
- ESS9GB, 24
- ESSBE5, 26
- eurostat\_codes, 27
- eustates, 27
- fakeAPI, 28
- fakeHappiness, 29
- fakeLogit, 30
- fakeTSCS, 30
- fakeTSD, 31
- ghp100k, 32
- GHR04, 33
- gss\_abortion, 34
- gss\_spending, 35
- gss\_wages, 37
- Guber99, 38
- illiteracy30, 39
- inglehart03, 40
- Lipset59, 41
- LOTI, 42
- LTPT, 43
- LTWT, 44
- min\_wage, 44
- mm\_mlda, 45
- mm\_nhis, 46
- mm\_randhie, 47
- mvprod, 48
- nesarc\_drinkspd, 49
- Newhouse77, 50
- ODGI, 50
- OODTPT, 51
- PPGE, 52
- PRDEG, 54
- Presidents, 55
- pwt\_sample, 55
- quartets, 56
- recessions, 57
- SBCD, 58
- scb\_regions, 58
- SCP16, 59
- sealevels, 60
- so2concentrations, 61
- steves\_clothes, 61
- sugar\_price, 62
- sweden\_counties, 63
- thatcher\_approval, 63
- therms, 64
- turnips, 65
- TV16, 66
- ukg\_eeri, 67

uniondensity, 68  
usa\_chn\_gdp\_forecasts, 69  
usa\_computers, 70  
usa\_migration, 71  
usa\_states, 71  
usa\_tradegdp, 72  
  
voteincome, 73  
  
wbd\_example, 73  
wvs\_ccodes, 74  
wvs\_immig, 75  
wvs\_justifbribe, 75  
wvs\_usa\_abortion, 76  
wvs\_usa\_educat, 77  
wvs\_usa\_regions, 78  
  
yugo\_sales, 79