Package 'stevedata'

May 9, 2026

Type Package

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

Depends R (>= 3.5.0)

Version 1.8.0

Maintainer Steve Miller <steve@svmiller.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 7.3.2

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] (ORCID: <https://orcid.org/0000-0003-4072-6263>)

Repository CRAN

Date/Publication 2025-11-12 10:20:02 UTC

Contents

`african_coups` *Modeling Coups in Africa, 1960 to 1975 (1982)*

Description

A data set on modeling coups in Africa using data from the period between 1960 and 1975 (1982). These data offer a partial replication of Jackman (1978).

Usage

```
african_coups
```

Format

A data frame with the following 11 variables.

`iso3c` a three-character ISO code for state identification

`country` an English country name

`jci` Jackman's (1978) coup index from 1960 to 1975

`tmis` Johnson et al.'s (1984) total military involvement score

`agperc` an estimate of the percentage of the country's labor force in agriculture and animal husbandry

`indperc` an estimate of the percentage of the country's labor force in industry

`literacy_cnts` an estimate of countrywide literacy from around 1965

`literacy_ba` another estimate of countrywide literacy from around 1965

`leperc` an estimate of the size of the largest ethnic group, as a percentage

`partydom` the percentage of the vote received by the largest party in the country prior to independence

`turnout` the turnout (as a percentage) at the independence referendum

Details

Data exist for instructional purposes, especially about modeling interactions. Reading Jackman (1978) and Johnson et al. (1984) will provide more information about the data and hypotheses. There was a follow-up symposium on this in 1986 in *American Political Science Review* that may be an interesting read and provide even more context about what's at stake in this debate.

English country names are country names from around the time of publication. Take note of older names of "Dahomey", "Swaziland", "Upper Volta", and "Zaire." The three-character ISO codes are current, mostly for ease of doing other things with the data. However, this comes with the acknowledgment that Dahomey and Zaire used to have different ISO codes under their older names. Both codes for Dahomey (`DHY`) and Zaire (`ZAR`) were retired in 1977 and 1997, respectively.

Ideally, I'd have Morrison's (1972) *Black Africa*, but I do not. I have a copy of a 1989 update, though. That's what I consulted in constructing this data set.

Jackman (1978) is deceptively opaque on what he's doing for the ethnic group variable and arguably misleads on what his turnout variable is actually from. In the case of the ethnic group variable, it's the difference between saying the largest ethnic group in Rwanda is 98% of the population versus 80% of the population. In short, it's the difference of saying whether there are any Tutsi at all in the country. Basically, I'm uncertain with what he's doing with what Morrison et al. (1989) define as "ethnic clusters".

Related: the agricultural variable is a midway point between columns B and columns C in Table 3.11 of Morrison et al. (1989). I do not think this is too far removed from what Jackman was looking at in an older version of the same data, but there will be slight differences. It's the difference of "these variables came from 1966" versus "this is an interpolation of 1960 to 1970". The latter is what I offer here. I can only do so much.

To continue this theme of the opacity in trying to reconstruct the data, Jackman (1978, p. 1265) says his social mobilization index incorporates the percentage of the labor force that is not employed in agriculture. The summary statistics he provides in fn. 4 on p. 1265 are consistent with this, at least (for the most part) in this reconstruction of the data. However, other statistical results and other language from Jackman are consistent with him using the percentage of the labor force that is employed in industry. This is not a trivial distinction either. Using the percentage of the country's labor force in industry would, in a literal sense, not strictly be "the simple sum of the percentage of the labor force in nonagricultural occupations". It would exclude those working in service industries. The data provide the opportunity to use either the industrial percentage variable or to manually create a non-agricultural labor force percentage variable as the difference between 100 and the agricultural labor force percentage variable. It makes the most sense to do the latter. The industry percentage variable comes from Table 3.14 in Morrison et al. (1989) and is likewise a midway point between 1960 and 1970.

Mercifully, the coup variables come from a replication by Johnson et al. (1984). Based on Morrison et al.'s (1989) updated data, it's not clear how Jackman could've derived some of these estimates using the formula he said he used. For example, Benin should have a score of 39 based on the information in Table 2.10 (p. 373 in Morrison et al. (1989)). Cameroon should have a 1 and not a 2, per Table 5.10 (p. 399). My comments here work under the assumption that Morrison et al. are adding information and not revising information in the second edition of the *Black Africa* handbook.

To be more precise, both the Jackman coup index and total military involvement variables are directly copied from Table 2 in Johnson et al. (1984) on p. 627. Missingness in the Jackman coup index variable communicates the country was not included in his original study, but was included in the Johnson et al. replication.

The literacy variables have suffices communicating where I obtained them. The Cross-National Time Series Database has a variable effectively communicating this information that I was using first in trying to recreate these data. These data come from 1965 in that data set. Jackman and Johnson et al. are assuredly using Morrison's almanac. That information is in Table 4.11 of Morrison et al. (1989), though it's possible the estimates contained therein are slightly different than what was reported in the first edition. I cannot know for sure.

Ethiopia is conspicuously missing in the party dominance variable. That's by design, and apparently its omission warranted ample discussion both by Jackman (1978) and Johnson et al. (1984). Johnson et al. (1984, fns. 4,5) argue it's a curious choice that can situationally influence the results that Jackman reports, but there are also lots of other choices made by Jackman (1978) that can influence these results.

I am 99.9% sure the turnout variable is Table 5.9 in Morrison et al. (1989). Jackman (1978) says this is from *before* independence but I'm confident he meant it was the turnout at the independence referendum.

References

Jackman, Robert W. 1978. "The Predictability of Coups d'etat: A Model with African Data." *American Political Science Review* 72(4): 1262-75.

Jackman, Robert W., Rosemary H. T. O'Kane, Thomas H. Johnson, Pat McGowan, and Robert O. Slater. 1986. "Explaining African Coups d'Etat." *American Political Science Review* 80(1): 225-49.

Johnson, Thomas H., Robert O. Slater, and Pat McGowan. 1984. "Explaining African Military Coups d'Etat, 1960-1982." *American Political Science Review* 78(3): 622-40.

Morrison, Donald George, Robert Cameron Mitchell, and John Naber Paden. 1989. *Black Africa: A Comparative Handbook* (2nd ed.). New York, NY: The Free Press.

`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)

`AJR5` *The Colonial Origins of Comparative Development (Table 5)*

Description

A data set to reproduce Table 5 in Acemoglu et al. (2001).

Usage

```
AJR5
```

Format

A data frame with 163 observations on the following variables.

`shortnam` a three-character code, ostensibly an ISO code

`catho80` the percentage of the country that is estimated to be Catholic

`muslim80` the percentage of the country that is estimated to be Muslim

`lat_abst` the latitude of the country (absolute value)

`no_cpm80` the percentage of the country that is esimated to be another religion

`f_brit` a dummy variable indicating whether the observation is a former British colony

`f_french` a dummy variable indicating whether the observation is a former French colony

`avexpr` average protection against expropriation risk, 1985-1995

`sjlofr` a dummy variable for whether the legal origin of the country's commercial code is French

`logpgp95` log-transformed GDP per capita (PPP) in 1995

`logem4` log-transformed European settler mortality

`baseco` a dummy variable indicating whether the observation is in the 'base sample'

Details

Acemoglu et al. (2001) are fairly transparent about what their data are and where you can read more about the sources they're using. La Porta et al. (1999) will feature prominently in some of these variables.

References

Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation". *American Economic Review* 91(5): 1369–1401.

La Porta, Rafael, Florencio Lopez-de-Silanes, Andrei Shleifer and Robert W. Vishny. 1999.*Journal of Law, Economics, and Organization* 15(1): 222-79.

`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` *Major Party (Democrat, Republican) Thermometer Index Data (1978-2012)*

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 *Abortion Attitudes (ANES, 2012)*

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 abortion if pregnancy would hurt woman

fatal oppose/"NFNO"/favor abortion if pregnancy would cause woman to die

incest oppose/"NFNO"/favor abortion if pregnancy was caused by incest

rape oppose/"NFNO"/favor abortion if pregnancy was caused by rape

bd oppose/"NFNO"/favor abortion if fetus would be born with serious birth defect

fin oppose/"NFNO"/favor abortion if having child would impose financial hardship

sex oppose/"NFNO"/favor abortion if the child will not be the sex the woman wants

choice oppose/"NFNO"/favor abortion if woman chooses to have one

`pid` respondent's partisanship (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

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

Details

"NFNO" = "Neither Favor Nor Oppose". All abortion prompts are on a 0-2 scale where 0 is oppose, 1 is "NFNO", and 2 is favor. The respondent's party identification is on a similar scale where 0 = "Democrat", 1 = "Independent", and 2 = "Republican". The additive scale of abortion scores has a minimum of 0 and a maximum of 16.

Source

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

---

anes_vote84 *Simple Data for a Simple Model of Individual Voter Turnout (ANES, 1984)*

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 *Arctic Sea Ice Extent Data, 1901-2015*

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 *Simple Mean Tariff Rate for Argentina*

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` *Aviation Safety Network Statistics, 1942-2019*

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.

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.

chile88 *Voting Intentions in the 1988 Chilean Plebiscite*

Description

A data set on voting intentions in the 1988 Chilean plebiscite, which ultimately ended the military junta rule of Augusto Pinochet.

Usage

```
chile88
```

Format

A data frame with 2700 observations on the following 8 variables.

`region` a character vector for the region of Chile in which the respondent lives

`pop` the population size of the respondent's community

`sex` a numeric vector that equals 1 if the respondent is a woman

`age` a numeric vector for the age of the respondent

`educ` a character vector indicating whether the respondent has a primary (P), secondary (S), or post-secondary (PS) education

`income` a numeric vector for respondent's monthly income (in pesos)

`sq` a numeric vector for the scale of support for the status quo in Chile

`vote` a character vector for the vote intention of the respondent (see details)

Details

Data were manually downloaded from John Fox's website. You will see his version of these data as `Chile` in the **carData** package. I changed a few things that are ultimately cosmetic. It's basically this data set.

The vote variable communicates vote intentions, whether to vote "Yes" (Y) to continue the Pinochet regime, to vote "No" (N) to end the Pinochet regime, to abstain (A) from a vote, or whether the respondent is undecided (U). 168 respondents did not answer the question.

Fox (2008, 336) does not say much about the status quo variable, and on what scale it is. It can only be easily inferred that higher values = more support for the status quo.

You may find it in your interest to relabel the "region" variable. In these data, the regions are Central ("C"), Metropolitan Santiago area ("M"), North ("N"), South ("S"), and the city of Santiago ("SA").

More information about the underlying source of the data would be more than welcome. Any information about these data, beyond the kind of R documentation files about its pedagogical use, is hard to find. This is a roundabout way of saying be cautious about any "real-world" use of these data beyond learning statistical methods. That is ultimately its intended use.

References

Fox, John. 2008. *Applied Regression Analysis and Generalized Linear Models* (2nd ed.). Los Angeles, CA: Sage.

`china_peace` *Drivers of China's Peace Engagement in Conflict-affected Countries*

Description

A data set on the correlates of Chinese peace engagement in conflict-affected countries.

Usage

```
china_peace
```

Format

A data frame with 884 observations on the following 13 variables.

`country` a character vector for an English country name of the conflict-affected country

`year` a numeric vector for the year

`iso3c` a three-character ISO code of the conflict-affected country

`region` a character vector for the region of the conflict-affected country

`chn_oda` a numeric vector for the amount of Chinese ODA in the conflict-affected country

`n_chnproj` a numeric vector for the number of Chinese ODA projects in the conflict-affected country

`other_oda` a numeric vector for the amount of ODA from other sources for the conflict-affected country

`gdppc` a numeric vector for GDP per capita of the conflict-affected country

`tnrr` a numeric vector for total natural resource rents in the conflict-affected country

`chn_unvs` a numeric vector for the similarity in voting with China in the UN General Assembly

`fdi_stock` a numeric vector for Chinese FDI stock in the conflict-affected country

`bdp100k` a numeric vector for battle-deaths in the conflict-affected country

`logotherodadiff1y` a numeric vector for the logged and lagged difference in Chinese ODA and ODA from other sources

Details

The `logotherodadiff1y` column comes as is from the replication data set. It seems to make use of ODA information prior to the start of the panel that is not available in their replication data. You could recreate it by taking a log-transformation of the `other_oda` column and subtracting the first-order lag from the second-order lag.

I will defer to the user to do their own log transformations of these data for the cause of replication. Read the supporting article for more information.

References

Jung, Yeonju and Karina Shyrokykh. 2024. "Needs or Interests: Drivers of China's Peace Engagement in Conflict-affected Countries." *International Peacekeeping* 31(5): 553-74.

`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. 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 CO2 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 CO2 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 4264 observations on the following 6 variables.

`country` a character vector for the country

`iso2c` a two-character vector of the country's ISO code

`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)

`pop` a numeric vector for the population size, in units of humans

Details

Coffee consumption data come from the International Coffee Organization, of which I feel I should be a member. Population data come from the UN Population Division. Observations for Yugoslavia in 1990 and 1991 are imputed manually. The observation for 1990 comes from a UN Population Division report. The 1991 observation comes from the Yugoslavian census.

The membership variable is agnostic to when a state became a member.

Belgium and Luxembourg are one reporting unit from 1990 to 1998. They are disaggregated here for those years. Each year is weighted by the relative proportion of each state's population. In practice, that means Belgium is getting about 95% of the value for those years with Belgium getting the remaining 5% or so.

Observations for the People's Republic of China are broken into components of China (Mainland), Hong Kong, and Macao. The consumption data for the People's Republic of China are simply the sum of the `value` variable for those three observations in a given year. The population variable is not; it codes the entire of Chinese population (including Hong Kong and Macao). Use with that in mind.

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 the from 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.

country_isocodes *ISO 3166 Country Codes (Two-Character, Three-Character, Numeric)*

Description

A data set of country ISO codes, for my ease and for the ease of my students.

Usage

```
country_isocodes
```

Format

A data frame with 249 observations on the following 4 variables.

`iso2c` a two-character ISO code

`iso3c` a three-character ISO code

`iso3n` a three-digit numeric ISO code

`name` an English country name

Details

This is a simple, abbreviated port and rename of the `ISO3_166_1` data in the **ISOcodes** package.

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.

`DAPO` *Determinants of Arab Public Opinion*

Description

A reduced form of data set for reproducing an analysis on the determinants of Arab public opinion in seven countries toward 13 different countries.

Usage

```
DAPO
```

Format

A data frame with 91 observations on the following variables.

`subjname` a three-character ISO code for the Arab (subject) country

`objname` an ALL-CAPS English name for the target/object country

`affect` an affect rating by the subject country to the object country

`capsub` the composite index of national capabilities (capability ratio) of the subject country

`capobj` the composite index of national capabilities (capability ratio) of the object country

`securtie` a dummy variable indicating at least an informal security tie between the subject and object

`export` the volume of exports from the subject to the object

`import` the volume of imports to the subject from the object

`subgdp` the gross domestic product (GDP) of the subject

`islam` a dummy variable that equals 1 if the object is a predominantly Muslim country

`west` a dummy variable that equals 1 if the object ia a Western country

Details

Exact coding issues/peculiarities are best addressed by reading the reference article. To maximally reproduce the article's analyses, the user will need to create some variables. The information is here, but you'll need to create a variable for dyadic trade (and as a percentage of the subject's GDP), GDP-adjusted imports, a means to filter out Israel from the analysis, and some of the information reported in Table 1. However, I think this is a learning experience for students.

References

Furia, Peter A. and Russell E. Lucas. 2006. "Determinants of Arab Public Opinion" *International Studies Quarterly* 50: 585-605.

`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

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*.

DCE12 *Domestic Conflict Events, 2012*

Description

A data set on domestic conflict events in 2012 as recorded by the Cross-National Time Series Database. Data exist for teaching about count models.

Usage

```
DCE12
```

Format

A data frame with 198 observations on the following 19 variables.

`iso2c` a two-character ISO code

`country` a character name for the country corresponding with the ISO code

`assassinations` the count of assassinations in 2012

`strikes` the count of general strikes in 2012

`guerwar` the count of guerilla warfare events in 2012

`govtcrises` the count of government crises in 2012

`purges` the count of purges in 2012

`riots` the count of riots in 2012

`revolutions` the count of revolutions in 2012

`agd` the count of anti-government demonstrations in 2012

`wci` the weighted conflict index in 2012

`area` the land area in square kilometers

`adultpop` the adult (15+) population (in 1000s)

`youthpop` the youth (15-29) population (in 1000s)

`gdppc` GDP per capita (in constant 2015 USD)

`urbanshare` urban population over total population (as percentage)

`tpop` total population (in 1000s)

`polyarchy` electoral democracy index, an estimate of democracy

`perctser` percentage of tertiary school-aged population enrolled in tertiary school

Details

Conflict events data come from the Cross-National Time Series Database. I've used these data before for published papers, but the relative opacity of a data set for yearly purchase comes with a bit of a caveat emptor for the important question of real-world inference.

Data on the democracy estimate and tertiary school enrollment rate come from the Varieties of Democracy project. Democracy estimate for Palestine comes as a simple average of the two Palestinian territories collected by the Varieties of Democracy project. These are West Bank and Gaza. The tertiary school enrollment variable, which originally comes from a data project by Barro and Lee (2013), is "filled" to the referent year from the most recent year available in the data. That would be 2010. It's fine for this purpose.

Population estimates come from the UN Population Division. GDP per capita comes from the World Bank. The estimate of land area (in square kilometers) comes from the CNTS. Country name comes from CNTS as well.

In all but the case of the data from CNTS, and the "filled" case of the tertiary school enrollment variable, the referent year for the data is 2011. Not that anyone is going to care too much for a simple data set like this, but this would be the ol' endogeneity concern.

`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 recent update I feel like including. Extensions shouldn't be too difficult with existing packages.

Usage

```
DJIA
```

Format

A data frame with 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 Dec. 31 1991 come from Pinnacle Systems. Observations from Jan. 1, 1992 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

`ddfat` 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

EBJ *The Economic Benefits of Justice*

Description

A data set on the apparent economic benefits of post-conflict justice

Usage

```
EBJ
```

Format

A data frame with 95 observations on the following 12 variables.

`testnewid_lag` an apparent identifier variable, of some description

`ccode` a Correlates of War(?) state code for the location of a conflict

`id` an apparent identifier variable, of some description

`pcj` a dummy variable for whether there was some kind of post-conflict justice institution created after a conflict

`fdi` a variable on net FDI inflows over a 10-year period after a conflict (in millions USD)

`econ_size` GDP, as an estimate of economic size

`econ_devel` GDP per capita, as an estimate of economic development

`econ_growth` GDP per capita change, as an estimate of economic growth

`kaopen` KAOPEN index score, as an estimate of capital openness

`xr` exchange rate fluctuations, as an indicator of exchange rate instability

`lf` labor force size

`lifeexp` average life expectancy for women, in years

Details

Data are taken Appell and Loyle's (2012) replication data set. Users should read their article in *Journal of Peace Research* for more information about the topic, the stake, and how the data were collected. This is just a simple, reduced form of the data they make available that is minimally sufficient for reproducing the first model of their Table I.

References

Appell, Benjamin J. and Cyanne E. Loyle. 2012. "The Economic Benefits of Justice: Post-conflict Justice and Foreign Direct Investment" 49(5): 685–99.

eight_schools *The Effect of Special Preparation on SAT-V Scores in Eight Randomized Experiments*

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 *State-Level Education and Voter Turnout in 2016*

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.

`epl_odds` *Odds for 2024-25 English Premier League Clubs*

Description

A data set on the odds for relegation and winning the table for English Premier League clubs for the 2024-25 season. Data are useful for illustrating what exactly odds are.

Usage

```
epl_odds
```

Format

A data frame with 20 observations on the following 7 variables.

`club` a character communicating the name of the club in the Premier League

`bet365_r` a numeric vector for the odds against relegation, by way of Bet365

`betfair_r` a numeric vector for the odds against relegation, by way of Betfair

`unibet_r` a numeric vector for the odds against relegation, by way of Unibet

`bet365_w` a numeric vector for the odds against winning the table, by way of Bet365

`betfair_w` a numeric vector for the odds against winning the table, by way of Betfair

`unibet_w` a numeric vector for the odds against winning the table, by way of Unibet

Details

Data come `oddschecker.com` as of Oct. 20, 2024, assuming these are preseason odds. Raw data are available on the project's website for your consideration. Don't bet on sports, unless you've been visited by Biff Tannen from the future.

Fractions have been converted into decimals for ease of maintaining the data. Raw odds (in fraction form) for those clubs most likely to be relegated are available in the raw data files on the project's Github.

Odds are typically(?) communicated as "odds against" in the sports betting world. It's why the highest odds for relegation and lowest odds for winning coincide with the biggest, most successful clubs. Context clues help, and are useful for understanding what these odds are saying.

It's possible that the language "win outright" is doing some heavy-lifting in how Bet365 lets you place wagers on winning the table.

I'm also aware of the reason these odds do not sum to 1, and in fact exceed it. If anything, I think "overrounding" is its own pedagogical tool for why odds can be wonky things to learn in relation to its use in the statistical modeling context.

`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

`eduyrs` 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

`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 scale, from 0 to 10

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

`cntry` 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

`eduyrs` 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 scale, from 0 to 10

Details

See accompanying blog post at https://svmiller.com/blog/2020/03/what-explains-british-attitudes-toward-i

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

`cntry` 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.

`eduyrs` 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 *Eurostat Country Codes*

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 *EU Member States (Current as of 2019)*

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.

`eu_ua_fta24` *A Roll Call Vote on Extending Temporary Trade Liberalization Measures Applicable to Ukrainian products under the EU/Euratom/Ukraine Association Agreement*

Description

A data set on an April 2024 roll call vote to extend an emergency free trade agreement with Ukraine.

Usage

```
eu_ua_fta24
```

Format

A data frame with 705 observations on the following 9 variables.

`member_id` a numeric identifier for a member of the European Parliament

`first_name` a first name of the member of the European Parliament

`last_name` a last name of the member of the European Parliament

`position` a character vector indicating the member's position on the measure ("For", "Against", "Did Not Vote", or "Abstain")

`iso2c` a two-character ISO code for the country the member represents

`country` an English country name for the country the member represents

`group_code` an acronym/code for the coalition of the member

`group_label` a character vector for the full name of the coalition of the member

`group_short_label` a short label for the coalition of the member

Details

Information about the exact measures are available from the European Union. Search: `A9-0077/2024`. Data in question are the raw data made available by `HowTheyVote.eu`

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: https://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 * 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.

gas_demand *Gasoline Demand in the OECD, 1960-1978*

Description

A data set on gasoline demand in the OECD countries from 1960 to 1978

Usage

```
gas_demand
```

Format

A data frame with 342 observations on the following 6 variables.

`country` a character vector for an English country name

`year` a numeric vector for the year of observation

`gas` gasoline consumption per car, log-transformed

`income` real per capita income, log-transformed

`price` real gasoline price, log-transformed

`cars` the stock of cars per capita, log-transformed

Details

The data are a simple port from the **AER** package. Users should read Baltagi and Griffin (1983) for more information. The data are purely for illustration about panel models.

Generally, per capita income should not be negative when log-transformed, especially for this set of countries. While it is clear that such negative values arise from the logarithmic transformation of values less than 1, but more than 0, it is not clear why per capita income would be on that particular scale.

The same curiosities emerge for the stock of cars per capita and real gasoline price, though proportional values between 0 and 1 are seemingly plausible (absent my willingness to look further into these details).

No matter, the data are sufficient for replication of Baltagi and Griffin (1983) without any further effort from the user. That's always nice.

References

Baltagi, Badi H. and James M. Griffin. 1983. "Gasoline Demand in the OECD: An Application of Pooling and Testing Procedures" *European Economic Review* 22: 117-137.

gatt_members *The 128 Countries That Had Signed GATT by 1994*

Description

A data set on GATT membership.

Usage

```
gatt_members
```

Format

A data frame with 128 observations on the following 3 variables.

`country` a character vector for a country name

`iso3c` a three-character ISO code

`date` the date the country joined GATT

Details

Data come from the World Trade Organization. Three-character ISO codes should be used with some caution as they mostly match what these states are now, if not what they were when they signed GATT (see: Benin, Democratic Republic of Congo). The conspicuous exception here is Yugoslavia, which has Yugoslavia's three-character ISO code of `YUG`

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.

The aforementioned details come, copied and pasted, from `GunPolicy.org`. As of my most recent check of these data (April 2024), this agency appeared to close due to lack of funding. This is unfortunate, but it is worth noting for matters of reproducibility and the use of these data in applied research questions.

`GHR04` *Comparative Public Health: The Political Economy of Human Misery and Well-Being*

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

`pubhlthexppgdp` a numeric vector for public health expenditures as a percentage of GDP

`totexphlth` 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 descriptive statistics correspond with what the authors report in their analyses for all variables, except the Polity variable. I have no way of knowing how they got the median they report. It should be 6, not 7.

The only real confusion on my end is why I ended up with one more observation than they report in Tables 1 and 3, and two more observations than they report in Table 2. 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.

`abhlth` 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 *Attitudes Toward National Spending in the General Social Survey (2018)*

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)

`occrecode` 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)

`educcat` 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 the GSS Data Explorer at the National Opinion Research Center (NORC) at the University of Chicago. 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, additive index 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. The "Quick UDS" estimate has the added quirk that converting the quantity to a probability (by way of `pnorm()`) communicates a probability that the observation in question is a 1 (i.e. a democracy). Try it out with some of the highest and lowest observations to see this in practice.

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 Jaggers. 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

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

Mitchell68 *Inequality and Insurgency: A Statistical Study of South Vietnam (Mitchell, 1968)*

Description

A data set on the correlates of government control in 26 provinces in South Vietnam, to replicate a study by Mitchell (1968).

Usage

```
Mitchell68
```

Format

A data frame with 26 observations on the following 9 variables.

`id` a numeric vector (a simple identifier)

`province` a character vector for the name of the province

`gc` a numeric vector for government control in the province (as a percent)

`ool` a numeric vector for owner-operated land (as a percent)

`cvlhs` a numeric vector for the coefficient of variation of the distribution of land-holdings, by size

`vl` a numeric vector for Vietnamese land, subject to transfer (as a percent of all land)

`fl` a numeric vector for French land, subject to transfer (as a percent of all land)

`m` a numeric vector for area of mobility

`pd` a numeric vector for population density (per square kilometer)

Details

The data are gathered from Table 1 in the document. You should also read the article for more information as to what's happening and for what purpose. Mitchell (1968) is quite clear about where else he's getting these data. Much of what follows can be discerned in the first few pages of Mitchell (1968), which jumps right into a conversation about research design after a brief introduction.

Province names are taken "as is" from Mitchell (1968). Since South Vietnam no longer exists, and these observations are about 60 years old, some of these province names may no longer exist. You may have to search for some old provincial maps of the former Republic of Vietnam in order to understand where some of these provinces are/were (especially if you're interested in the regional variation noted by Paige (1970)).

Los Angeles Times maps inform the government control variable, and there are assumptions that Mitchell makes about the nature of control by the government (South Vietnam), the Vietcong, or the areas that are contested. The "control" here ultimately refers to South Vietnam.

The observations for government control variable are from 1965. Mitchell's footnote in his Table 1 says all other variables (except for population density) correspond with information from 1960. The population density estimate comes from 1964.

The coefficient of variation variable is defined as the standard deviation of land-holding size divided over the mean. If every landholding is of equal size, the observation is 0. Larger values suggest more variability in size of land-holdings with the implication being larger land-holdings are conspicuous in the province. It's a crude, but interesting, measure of inequality with that in mind.

The owner-operated land variable is another crude, but interesting/novel measure. An obvious percentage, 100 implies complete land ownership. 0 implies universal tenancy where peasants work on land they do not own. Some familiarity with the peculiarities of South Vietnamese society at the time is strongly suggested.

The "French land" and "Vietnamese land" variables refer to a specific agrarian reform measure ("Ordinance 57"). The Vietnam version includes both expropriated and redistributed land. The French version includes just expropriated land, per Mitchell. The logic is the Vietnamese version suggests higher values = lower inequality since the measure (partly) includes redistributed land. The French land, being just expropriation, has a single owner (the South Vietnamese government). That suggests higher inequality for higher values. This logic is interesting but questionable, and we'll just have to roll with the premise for the nature of the intended use of these data (i.e. replication). Paige's (1970) objection is more about regional variation in South Vietnam and its varied patterns of land use, and not about the particulars of these two measure (per se).

The mobility measure is a percentage, referring to the percentage of the province that is composed of plains and hills without dense forest.

The data are faithfully (to my level best!) scraped from Table 1 of his article. However, the results that come from a linear model do not perfectly reproduce his results (Equation 2, p. 432). I don't know why this is the case, nor is it that important. It is worth noting that this kind of "step-wise" procedure he employs for selecting a linear model is 100% not how you should do it, and that 33rd footnote he includes on p. 432 would be an automatic rejection at any quantitatively-oriented journal today.

It may interest the user to see re-analyses of Mitchell (1968) from around this time. I include those in the references for your consideration. Briefly, Paige's (1970) objection is that Mitchell (1968) includes radically different land-holding types into assorted measures of inequality and that Mitchell is selecting on 1965 (a watershed moment of insurgency during the war). Paranzino's (1972) critique is primarily statistical, though incorporates some of the issues raised by Paige (1970). Importantly, he correctly notes what the results of the linear model should be (p. 567).

References

Mitchell, Edward J. 1968. "Inequality and Insurgency: A Statistical Study of South Vietnam." *World Politics* 20(3): 421–38.

Paige, Jeffery M. 1970. "Inequality and Insurgency in Vietnam: A Re-Analysis." *World Politics* 23(1): 24–37.

Paranzino, Dennis. 1972. "Inequality and Insurgency in Vietnam: A Further Re-Analysis." *World Politics* 24(2): 565–78.

mmb_war *Mutual Military Build-Ups and War*

Description

A simple data set on mutual military build-ups and war, useful for teaching about a long-standing empirical debate in international relations by way of basic tests (like a chi-square test).

Usage

```
mmb_war
```

Format

A data frame with 2324 observations on the following 9 variables.

`ccode1` a Correlates of War state code

`ccode2` another Correlates of War state code

`tssr_id` a rivalry identifier

`micnum` the start year of a confrontation between the two states

`year` the start year of a confrontation between the two states

`dyfatmin` the minimum estimated dyadic fatalities in the confrontation

`dyfatmax` the maximum estimated dyadic fatalities in the confrontation

`sumevents` the total number of events in the confrontation

`mmb` a dummy variable that equals 1 if the confrontation came after the start of a mutual military build-up

Details

The unit of analysis for these data are non-directed dyadic confrontations for strategic rivals. Be mindful that confrontations start with the first event of any kind. See Gibler and Miller (2024a, 2024b) for more about events and confrontations. See Thompson et al. (2021) for more information about strategic rivalries.

Mutual military build-ups (MMBs, for short) are a slightly more evasive label I'm using for the more familiar "arms race." They are operationalized largely from Gibler et al. (2005). Briefly: mutual military build-ups are any episode in 1) a rivalry relationship where 2) each dyadic partner is increasing their military expenditure *or* personnel, 3) eight percent or more from the previous year, 4) for at least three years where 5) historical evidence largely corroborates a directionality in the mobilization of the kind we would broadly conceptualize (a la Richardson, 1939). In other words, the mutual mobilization isn't coincidental or a function of other priorities.

The data I recreate here follow Gibler et al. (2005), but use newer capabilities and rivalry data. I further employ some case exclusion rules that would not otherwise be evident in a reading of Gibler et al. (2005). First, I take some care to exclude cases where it is pretty clear that what Gibler et al. (2005) call an arms race is more accurately just the mobilization of the war itself. For example,

their arms race #26 between China and Japan occurs between 1940 and 1944, though the ongoing war between both comfortably covers it. Related, I employ an admittedly ad hoc termination date to end when we might comfortably note a war is ongoing (see: the various World War 1 arms races). Further, I often extend a year to an arms race if one side started mobilizing first and the other side only started mobilizing the next year and/or one side continued mobilizing for a year after the other stopped. This is why, for example, I have an extra year in the Spain-Morocco build-up in the early 1970s (Spain mobilized through 1975). There were some cases where I disagreed that something could be considered an arms race/mutual military build-up by this metric. For example, the build-up observed between Somalia and Ethiopia in the 1970s (their arms race #44) is an interesting case where it's clear Ethiopia is mobilizing. However, the data suggest only one year of mobilization for Somalia (1974). I remove those cases from my recreation.

References

Gibler, Douglas M. 2005. "Taking Arms against a Sea of Troubles: Conventional Arms Races during Periods of Rivalry" *Journal of Peace Research* 42(2): 131-47.

Gibler, Douglas M., and Steven V. Miller. 2024a. "The Militarized Interstate Confrontation Dataset, 1816-2014." *Journal of Conflict Resolution* 68(2–3): 562–86.

Gibler, Douglas M., and Steven V. Miller. 2024b. "The Militarized Interstate Events (MIE) Dataset, 1816–2014." *Conflict Management and Peace Science* 41(4): 463–81.

Richardson, Lewis F. 1939. *Generalized Foreign Politics*. Cambridge University Press.

Thompson, William R., Kentaro Sakuwa, and Prashant Hosur Suhas. 2021. *Analyzing Strategic Rivalries in World Politics: Types of Rivalry, Regional Variation, and Escalation/De-escalation*. Springer.

`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 Carloss 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.

`plantype` the plan coverage of the respondent, as a factor

`age` the age of the respondent

`blackhisp` 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.

`plantype` 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://jrnold.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

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 knows'' or missing info) are coded as 99. Non-drinkers are coded as \c 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 *Medical-Care Expenditure: A Cross-National Survey (Newhouse, 1977)*

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 *Ozone Depleting Gas Index Data, 1992-2022*

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

`ch3cl` chloromethane concentration in parts per trillion

`ch3br` bromomethane concentration in parts per trillion

`ccl4` carbon tetrachloride concentration in parts per trillion

`ch3ccl3` 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 coverage ratio

`polity` the familiar Polity measure of democracy, from -10 to 10

`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.

Parvin73 *Economic Determinants of Political Unrest (Parvin, 1973)*

Description

A data set on the economic determinants of political unrest, for replicating a publication from 1973.

Usage

```
Parvin73
```

Format

A data frame with 26 observations on the following 9 variables.

`country` a character vector for a country name

`levviol` a numeric vector for the level of violence

`pci` a numeric vector for per capita income

`incdist` a numeric vector for income distribution

`d_pci` a numeric vector for per capita income growth

`sem` a numeric vector for socioeconomic mobility

`comint` a numeric vector for communication intensity

`concfac` a numeric vector for concentration factor

`pop` a numeric vector for population size

Details

The bulk of these data come from Russett's (1964) *World Handbook of Political and Social Indicators*. The data themselves are transcribed from the appendix of the article, which allows a replication of the results that Parvin (1973) reports. You should read that article for more information as to what's happening and for what purpose.

I did not catch Parvin (1973) mentioning this in the article, but there must be some kind of additive constant in the level of violence variable because the logarithmic transformations he reports would be undefined for the observations (like Denmark) where the level of violence is zero. The easiest way to approximate whatever Parvin (1973) did is to add .001 to the level of violence variable before taking its logarithmic transformation. That would allow a near perfect replication of Table 1.

It should go without saying that the population reported for Belgium, in the appendix, is likely a transcription error. Belgium's population is reported here as 9184, not "91.84.00".

The United Arab Republic was the short-lived union of Egypt and Syria, if you were curious what that is in the data.

References

Parvin, Manoucher. 1973. "Economic Determinants of Political Unrest: An Econometric Approach". *Journal of Conflict Resolution* 17(2): 271–96.

postcol_growth *Post-Colonial Growth in the African Continent*

Description

A simple data set on post-colonial growth trajectories in the African continent, for intended use to instruct students about *t*-tests around the application of colonial legacies.

Usage

```
postcol_growth
```

Format

A data frame with 53 observations on the following 11 variables.

`ccode` a Correlates of War state code

`cw_name` a Correlates of War state name

`styear` the start year for latest system entry for the state

`IndFrom` a Correlates of War state code, if applicable, identifying the state from which the state identified in the `ccode` gained independence

`colmast` a character vector largely corresponding with the information in `IndFrom` with only slight changes

`mrgdppcind` an estimate of GDP per capita for the year identified in the `styear` column, itself largely corresponding with independence from the state identified in `IndFrom` and `colmast`

`mrgdppc5` an estimate of GDP per capita 5 years from the year identified in the `styear` column

`mrgdppc10` an estimate of GDP per capita 10 years from the year identified in the `styear` column

`mrgdppc15` an estimate of GDP per capita 15 years from the year identified in the `styear` column

`mrgdppc20` an estimate of GDP per capita 20 years from the year identified in the `styear` column

`mrgdppc25` an estimate of GDP per capita 25 years from the year identified in the `styear` column

Details

Data are generated with assistance from **isard**, another R package I maintain.

Data are sliced to record only latest system entry into the CoW data, which concerns states (like Morocco, Tunisia, and Ethiopia) that were temporarily occupied and eliminated.

I take some liberties classifying former colonial masters in the `colmast` column. Namely, I elect to not record Ethiopia's independence from Italy as suggesting Italy was a colonial master of Ethiopia. I do not code South Africa's independence from the United Kingdom as noteworthy for the sake of this analysis (given the topic of interest to me for creating these data). Senegal nominally gains independence from Mali when it leaves the Mali Federation, but I attributes its independence to being ultimately from France. Morocco and Tunisia were protectorates of France though the ICOW measure of colonial history says it gains independence from the Ottoman Empire. I do not consider

Namibia (South Africa) or Eritrea (Ethiopia) to be colonial under the states from which they gained independence.

The estimates of GDP per capita are real GDP per capita in prices constant across countries and over time (in 2011 international dollars, PPP). These data are sourced from the Maddison project database but are the product of simulations by Farris et al. (2022). You can read a bit more about these in the sources in the reference section, or in the documentation for the `cw_gdppop` data frame in the **isard** package.

Colonial history data come by way of ICOW (v. 1.1).

References

Bolt, Jutta, Robert Inklaar, Herman de Jong, and Luiten Janvan Zanden. 2018. "Rebasing 'Maddison': New Income Comparisons and the Shape of Long-Run Economic Development." *Maddison Project Working paper 10*.

Fariss, Christopher, J., Therese Anders, Jonathan N. Markowitz, and Miriam Barnum. 2022. "New Estimates of Over 500 Years of Historic GDP and Population Data." *Journal of Conflict Resolution* 66(3): 553–91.

Hensel, Paul R. 2018. "ICOW Colonial History Data Set, version 1.1." Available at https://www.paulhensel.org/icowcol.html.

Other Points of Departure:

The intended use of these data is to instruct students about *t*-tests with an application to the development trajectories of former colonies in the African continent. This particular topic is definitely fraught with caveats to consider, and such a simple data set intended to teach students rudimentary methods around a question that might understand just cannot cover all these issues. Please consider the following scholarship on this topic.

Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation". *American Economic Review* 91(5): 1369–1401.

Bolt, Jutta and Dirk Bezemer. 2009. "Understanding Long-Run African Growth: Colonial Institutions or Colonial Education?" *The Journal of Development Studies* 45(1): 24–54.

Gallup, John Luke, Jeffrey D. Sachs, and Andrew D. Mellinger. 1999. "Geography and Economic Development." *International Regional Science Review* 22(2): 179–232.

Glaeser, Edward L., Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer. 2004. "Do Institutions Cause Growth?" *Journal of Economic Growth* 9: 271–303.

La Porta, Rafael, Florencio Lopez-de-Silanes, Andrei Shleifer and Robert W. Vishny. 1999. *Journal of Law, Economics, and Organization* 15(1): 222-79.

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 (among other things). The data offer a means to assess Garrett's (1998) argument about left-wing governments' ability to stem the tide of globalization's effect on decreased government spending. Data also draw inspiration from Rodrik (1998) and Garrett (2001).

Usage

```
PPGE
```

Format

A data frame with the following 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.

`ud` union density, as a percentage (i.e. union members/working employees)

`urbanperc` the percentage of the population living in urban areas)

`gdppc` GDP per capita, in constant 2015 USD

`tpop` total population size, in units of individual humans

`depratio` dependency ratio (see details)

Details

The data are an unbalanced panel with assorted quirks during its construction. Data missingness affecting Switzerland means it would only appear in the panel 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. Union density coverage is spotty for states like Greece and Portugal. The data also have some obvious COVID weirdness for 2020. Use that to inform whatever case or variable selection you would like to do. It may make sense to employ a temporal domain of something like 1980 to 2005, or whatever. I don't know. There's also the issue of what to do about the recession.

The dependency ratio is defined as the population aged 0-14, or 65 and above, divided over the "working-age" population of 15-64 (x 100).

Briefly: the government spending/GDP data come from the International Monetary Fund. The trade/GDP data come from the World Bank's API, as do the population, GDP per capita, and urbanization data (see their details). The more conventional union density data come from OECD/ICWSS. 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%), 2) A moderate share of the population (about 5 to 15%). 3) A large share of the population (about 16 % to 25%). 4) A very large share of the population (about 26% or more).

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 Gjerløw, 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 Rydén, Johannes von Römer, Brigitte Seim, Rachel Sigman, Svend-Erik Skaaning, Jeffrey Staton, Aksel Sundström, 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

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.

Garrett, Geoffrey. 2001. "Globalization and Government Spending around the World." *Studies in Comparative International Development* 35(4): 3-29.

Rodrik, Dani. 1998. "Why Do More Open Economies Have Bigger Government?" *Journal of Political Economy* 106: 997-1032.

`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` *U.S. Presidents and Their Terms in Office*

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` *Penn World Table (10.0) Macroeconomic Data for Select Countries, 1950-2019*

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)

`rnna` 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

Anscombe, 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-cor

Source

National Bureau of Economic Research (NBER)

---

rok_unga *The Correlates of Dyadic Voting Similiarities in the UN General Assembly for South Korea*

Description

A data set on dyadic voting similarity for South Korea in relation to other states, from 1991 to 2022.

Usage

```
rok_unga
```

Format

A data frame with the following variables.

`ccode1` a numeric vector, and constant, identifies the Correlates of War state code for South Korea (732)

`ccode2` a numeric vector for the Correlates of War state code for the other state in the dyad

`iso3c` a three-character ISO code corresponding with the Correlates of War state code for `ccode2`

`year` a numeric vector for a year

`agree` the percentage of the time South Korea and the other state in the dyad agreed on a vote in a given year

`v_agree` the percentage of the time South Korea and the other state in the dyad agreed on a vote in a given year, as calculated by Voeten et al. in their data

`kappa` weighted Cohen's kappa for dyadic foreign policy similarity as derived from the UN voting data

`ip1` the ideal point estimate for South Korea for a given year, as derived from UN voting data

`ip2` the ideal point estimate for `ccode2`, as derived from UN voting data

`ipd` the absolute distance between `ip1` and `ip2`

`gdppc1` estimated GDP per capita in 2015 USD for South Korea in the referent year

`gdppc2` estimated GDP per capita in 2015 USD for `ccode2` in a given year

`v2x_polyarchy1` the Varieties of Democracy estimate for the "polyarchy" for South Korea in the referent year

`v2x_polyarchy2` the Varieties of Democracy estimate for the "polyarchy" for `ccode2` in a given year

`xm_euds1` Xavier Marquez' estimate for the extended Unified Democracy Score for South Korea in the referent year

`xm_euds2` Xavier Marquez' estimate for the extended Unified Democracy Score for `ccode2` in a given year

`capdist` the distance between Seoul and the capital of `ccode2` in the year

Details

Voeten et al's codebook cautions that their agreement variable is there for comparison and should not be used for a serious analysis of dyadic foreign policy similarity. The `agree` variable I calculate is based on *all* votes, whereas (I think) Voeten et al. exclude amendments and votes on paragraphs.

Cohen's (weighted) kappa is suggested by Haege (2011) for use measuring dyadic foreign policy similarity. This measure is likewise calculated by me for *all* votes. I forget how Haege (2011) does this for his calculations and if he is excluding votes on amendments or paragraphs. Its interpretation differs from how one might use the ideal point distance variable. This is a chance-corrected correlation. Higher values indicate more similarity whereas higher values in the ideal point distance variable communicate more dissimilarity.

GDP per capita include some imputations by way of a semiparametric Bayesian Gaussian copulas. This prominently concerns Venezuela. Data are otherwise derived from the World Bank's open data.

Xavier Marquez' "extended Unified Democracy Scores" approximate a normal distribution with a standard deviation of 1. Invoking `pnorm()` on a particular estimate provides a kind of probabilistic assessment of whether the observation in question is a democracy. In both the Varieties of Democracy estimate and the Marquez estimate, higher values = "more democracy". See also: the `Lipset59` documentation in this same package.

Capital-to-capital distance is calculated using the Vicenty method ("as the crow flies"), and is done by way of a **peacesciencer** call and its `add_capital_distance()` function. There are unusual cases where a capital moved (i.e. Burundi, Kazakhstan, Myanmar, Nigeria). In those cases, the capital on Jan. 1 of the given year is treated as the capital.

Russett64 *Inequality and Instability: The Relation of Land Tenure to Politics (Russett, 1964)*

Description

A data set on inequality and political instability, to replicate an analysis from Russett (1964).

Usage

```
Russett64
```

Format

A data frame with 47 observations on the following 10 variables.

`country` a character vector for the country

`demcat` a character vector for the type of political system, either "Stable Democracies", "Unstable Democracies", or "Dictatorships"

`gini` a numeric vector for the GINI coefficient

`perc_farmsh` a numeric vector for the percent of farms with half the land

`perc_farmsr` a numeric vector for the percent of farms that are rented

`pi` a numeric vector for personnel instability

`eiw` a numeric vector for Eckstein's internal war measure

`deaths` a numeric vector for deaths from civil group violence per one million

`gnppc` a numeric vector for gross national product per capita

`perc_lfa` a numeric vector for the percent of the labor force in agriculture

Details

The data are scraped from Table 1 and Table 3 of his article, to the best of my ability. I was initially hoping this could be a problem for ChatGPT and its OCR functionality. However, ChatGPT made an absolute mess of Table 1. The bulk of this is hand-coded. The data, as of right now, can effectively reproduce what Russett (1964) reports in his analyses, but it is not identical.

You should read the article to see the assorted citations and data caveats that Russett (1964) reports. Those are ultimately suppressed/ignored here.

References

Russett, Bruce M. "Inequality and Instability: The Relation of Land Tenure to Politics." *World Politics* 16(3): 442–54

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 "debtrestructuring"

`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 *Sulfur Dioxide Emissions, 1980-2020*

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")

states_war *State Performance in Inter-State Wars*

Description

A data set on state performance in inter-state wars. This data is useful for evaluating Valentino et al.'s (2010) "Bear Any Burden" analysis using more current data.

Usage

```
states_war
```

Format

A data frame with the following variables.

`micnum` a numeric for the confrontation code

`ccode` a numeric for the Correlates of War state code

`stdate` a character vector communicating participant start date. See details for more.

`enddate` a character vector communicating participant start date. See details for more.

`mindur` a numeric vector communicating minimum duration in confrontation. See details for more.

`maxdur` a numeric vector communicating minimum duration in confrontation. See details for more.

`sidea` a numeric vector communicating whether participant was on side that initiated confrontation

`orig` a numeric vector communicating whether participant was in confrontation on day one

`hiact` a numeric vector communicating highest action during confrontation

`fatalmin` a numeric vector for minimum estimated fatalities for participant

`fatalmax` a numeric vector for maximum estimated fatalities for participant

`oppfatalmin` a numeric vector for minimum estimated fatalities by participant against opponents

`oppfatalmax` a numeric vector for maximum estimated fatalities by participant against opponents

`milex` an estimate of military expenditures (in thousands)

`milper` an estimate of the size of military personnel (in thousands) for the state

`cinc` The Composite Index of National Capability ("CINC") score

`tpop` an estimate of the total population size of the state (in thousands)

`v2x_polyarchy` the Varieties of Democracy "polyarchy" estimate

`polity2` the the `polity2` score from the Polity project

`xm_qudsest` an extension of the Unified Democracy Scores (UDS) estimates, made possibly by the **QuickUDS** package from Xavier Marquez.

`wbgdp2011est` a numeric vector for the estimated natural log of GDP in 2011 USD (log-transformed)

`wbpopest` a numeric vector for the estimated population size (log-transformed)

`wbgdppc2011est` a numeric vector for the estimated GDP per capita (log-transformed)

Details

Start date and end date are in "MM/D(D)/YYYY" format. You can extract this information into multiple columns with a `separate` function from the **tidyr** package. This is mostly for convenience. Be mindful of two things: First, dates are dates of first and last action, and not necessarily the escalation to war, per se. Second, dates can be "missing". These are -9s, and are commonplace when archival research can't pinpoint an exact day something happened.

Observations select at the *confrontation*-level where maximum fatalities are greater than 1,000 and at the *participant*-level where (1) the participant engaged in at least an attack during this confrontation, (2) there are no instances where a participant dropped in/out on the same side of a multilateral confrontation or switched sides, and (3) the confrontation doesn't have an instance where a participant incurred fatalities while themselves not initiating a use of force. For illustration's sake, the Taiwan Straits Crises saw several appearances by the United States, but only one instance (for six

days in Feb. 1953) where the U.S. engaged in an attack. World War II is a classic case of participants switching sides (France did so three times), but it also happened in the War of Latvian Independence as well (MIC#2604). The War of Attrition also saw the Russians reappear twice. Cases like these aren't included, mostly for convenience sake. In total, 41 cases with 1,000 maximum fatalities or more at the confrontation-level are excluded because of this. Of these 41 cases, World War II and the Vietnam War are the most conspicuous by their absence. Data come from version 1.01 of the Militarized Interstate Confrontation data.

Opponent fatalities are strictly dyadic and are derived from the Militarized Interstate Events data.

Capabilities, GDP, and democracy data come from **peacesciencer** for a forthcoming v. 1.2.0 release. See package for more information, though references are also included below. Variables are mostly lagged to the year prior to the participant observation year. However, there are several cases in the data that are born into war (see: India, Pakistan, North and South Korea, North and South Vietnam). In cases of missing data, information from the observation year is used.

The `tpop` and `wbpopest` columns are measuring the same thing but are derived from two different data sets with two different data-generating procedures. Use whichever one you like, but be mindful of what you're doing and for what purpose you're doing it.

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.

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.

Gibler, Douglas M., and Steven V. Miller. Forthcoming. "The Militarized Interstate Events (MIE) Dataset, 1816–2014." *Conflict Management and Peace Science.*

Gibler, Douglas M., and Steven V. Miller. 2023. "The Militarized Interstate Confrontation Dataset, 1816-2014." Journal of Conflict Resolution 68(2–3): 562–86

Marshall, Monty G., Ted Robert Gurr, and Keith Jaggers. 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

Miller Steven V. 2022. "peacesciencer: An R Package for Quantitative Peace Science Research." *Conflict Management and Peace Science*, 39(6), 755–779.

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.

Singer, J. David, Stuart Bremer, and John Stuckey. (1972). "Capability Distribution, Uncertainty, and Major Power War, 1820-1965." in Bruce Russett (ed) *Peace, War, and Numbers*, Beverly Hills: Sage, 19-48.

Singer, J. David. 1987. "Reconstructing the Correlates of War Dataset on Material Capabilities of States, 1816-1985" *International Interactions*, 14: 115-32.

Valentino, Benjamin A., Paul K. Huth, and Sarah E. Croco. 2010. "Bear Any Burden? How Democracies Minimize the Costs of War." *Journal of Politics* 72(2): 528-544

`stevesteps` *Steve's Steps in a Day, 2017-2025*

Description

These are data from my Fitbit for steps in a day.

Usage

```
stevesteps
```

Format

A data frame with 2875 observations on the following 2 variables.

`date` a date

`value` estimated steps in a day

`steves_clothes` *Steve's (Professional) Clothes, as of March 20, 2022*

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

`ftobama1` a thermometer rating for Barack Obama

Details

The survey period was April 10-18, 2020 and was done entirely online. Thermometer ratings are on a 0 to 100 scale, where higher values indicate more "warmth".

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

https://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 Recemt 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 *U.S. Inbound/Outbound Migration Data, 1990-2017*

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 *State Abbreviations, Names, and Regions/Divisions*

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` *U.S. Trade and GDP, 1790-2018*

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.

USFAHR *U.S. Foreign Aid and Human Rights in Assorted Years*

Description

A data set on economic aid allocation by the United States for assorted years. These are useful for illustrative cross-sectional relationships between human rights and U.S. aid allocation at what amounts to midway points for various presidential administrations.

Usage

```
USFAHR
```

Format

A data frame with 1654 observations on the following 18 variables.

`country` an English country name

`ccode` a Correlates of War state code

`region` a region in which the country resides, per Greenbook

`year` a year

`nomoblig` economic aid obligations in nominal U.S. dollars

`constoblig` economic aid obligations in constant 2019 U.S. dollars

`clphy` a physical violence index, bound between 0 and 1

`civlib` a civil liberties index, bound between 0 and 1

`fpsusa` foreign policy similarity with the United States

`fpsrus` foreign policy similarity with the Soviet Union/Russia

`mindistusa` minimum distance of the country from the United States

`mindistrus` minimum distance of the country from the USSR/Russia

`gdp` an estimate of GDP in constant 2011 U.S. dollars

`pop` an estimate of population size

`usaimp` a value of how much the U.S. imports from the country (in thousands USD)

`usaexp` a value of how much the U.S. exports to the country (in thousands USD)

`milex` an estimate of military expenditures (in thousands USD)

`cinc` a composite index of national capabilities

Details

Matching is done on Correlates of War state codes. Thus, the exact "population" is an amalgam of U.S. aid and Correlates of War state system membership. Regions are offered, as is, from USAID Data Services.

Data on aid are "obligations" and not "disbursements", and thus may better reflect donor intent. These come from US Overseas Loans & Grants ("Greenbook") and were prepared by USAID Data Services on July 14, 2021.

Greenbook only offers information about dollar amounts of aid, contingent on receiving aid. Observations were added, based on Correlates of War state system membership, about countries that could've received aid but did not. Countries that never received aid at all had to have regions assigned to them ex post. I don't think the regions imputed for these observations are problematic. This concerns Andorra, Czechoslovakia, Dominica, German Democratic Republic, German Federal Republic, Liechtenstein, Luxembourg, Monaco, Nauru, Republic of Vietnam, San Marino, St. Lucia, St. Kitts and Nevis, Switzerland, Tuvalu, Yemen Arab Republic, Yemen People's Republic, and Zanzibar.

Higher values of the physical violence index and civil liberties index communicate better human rights records. Data are lagged a year.

Foreign policy similarity is Cohen's (1960) kappa based on valued United Nations General Assembly voting. Data come from Haege (2011) by way of **peacesciencer**'s `add_fpsim()` function. Please read **peacesciencer** documentation for more information about these measures, along with what you should cite for any serious use of these data. Higher values for these measures = more foreign policy similarity.

Minimum distance is calculated using the Vincenty method ("as the crow flies"). Measurement is in kilometers and data come from **peacesciencer** and its `add_minimum_distance()` function. Check package documentation for appropriate citation for any serious use.

Estimates of gross domestic product ("GDP") and population come by way of **peacesciencer** and its `add_sdp_gdp()` function. Check package documentation for appropriate citations for any serious use. GDP is in actual dollars.

Trade data come from Correlates of War trade data by way of **peacesciencer** and its `add_cow_trade()` function. Check package documentation for appropriate citations for any serious use.

Military expenditure and capabilities data come from Correlates of War by way of **peacesciencer** and its `add_cow_trade()` function. Check package documentation for appropriate citations for any serious use.

`voteincome` *Sample Turnout and Demographic Data from the 2000 Current Population Survey*

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 *A Simple Panel drawn from World Bank Open Data*

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 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).

The human capital index is on a 0 to 1 scale.

---

wb_groups *World Bank Country Groups*

Description

A data set on World Bank country groups/classifications, for ease of selecting three-character ISO codes of interest.

Usage

```
wb_groups
```

Format

A data frame with 2085 observations on the following 4 variables.

`wbgc` a three-character code for the World Bank group

`wbgn` a more informative name for the World Bank group

`iso3c` a three-character ISO code

`name` a name for the country that corresponds with the three-character ISO code

Details

Data are for the current 2025 fiscal year. The World Bank's Data Help Desk will offer more information about specific criteria for things like income.

Weede84 *Military Conflict or War Involvement, 1960-1980 (Weede, 1984)*

Description

A data set on military conflict/war involvement and democracy.

Usage

```
Weede84
```

Format

A data frame with 101 observations on the following 3 variables.

`ccode` a Correlates of War state code

`country` a slightly more informative identifier for the Correlates of War state code

`dem` a binary indicator for whether Weede (1984, Table 3) identifies the state as a democracy

`butter` a count(?) variable of conflicts with over 100 casualties

`kende` a count(?) variable of wars, according to Kende (1982)

`ssiw` a count of involvement in Singer and Small (1972) inter-state wars

`ssiw_id` an identifier of years in inter-state war, where applicable

`ssew` a count of involvement in Singer and Small (1972) extra-state wars

`ssew_id` an identifier of years in extra-state war, where applicable

Details

Data come from the appendix. The `_id` inputs were the parentheses in the table. Weede (1984) identifies the democracies in Table 3 (p. 658) of his article.

Butterworth's temporal domain is 1960-1974. Kende and Singer and Small cover 1960-1980.

I'll admit I have never seen the Kende (1982) data before, and I like to think I'm well-versed in this stuff.

References

Weede, Erich. 1984. "Democracy and War Involvement." *Journal of Conflict Resolution* 28(4): 649–664.

`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

https://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 *Attitudes on the Justifiability of Abortion in the United States (World Values Survey, 1982-2011)*

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` *Region Categories for the United States in the World Values Survey*

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 "Northwest" 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