

# Package ‘carData’

September 30, 2018

**Version** 3.0-2

**Date** 2018-09-28

**Title** Companion to Applied Regression Data Sets

**Depends** R (>= 3.0)

**LazyLoad** yes

**LazyData** yes

**Description** Datasets to Accompany J. Fox and S. Weisberg,  
An R Companion to Applied Regression, Third Edition, Sage (forthcoming).

**License** GPL (>= 2)

**URL** <https://r-forge.r-project.org/projects/car/>,  
<https://CRAN.R-project.org/package=carData>,  
<http://socserv.socsci.mcmaster.ca/jfox/Books/Companion/index.html>

**Author** John Fox [aut, cre],  
Sanford Weisberg [aut],  
Brad Price [aut]

**Maintainer** John Fox <jfox@mcmaster.ca>

**Repository** CRAN

**Repository/R-Forge/Project** car

**Repository/R-Forge/Revision** 566

**Repository/R-Forge/DateTimeStamp** 2018-09-28 16:36:37

**Date/Publication** 2018-09-30 16:00:03 UTC

**NeedsCompilation** no

## R topics documented:

Adler . . . . .	3
AMSSurvey . . . . .	4
Angell . . . . .	5
Anscombe . . . . .	5

Arrests . . . . .	6
Baumann . . . . .	7
BEPS . . . . .	8
Bfox . . . . .	9
Blackmore . . . . .	10
Burt . . . . .	10
CanPop . . . . .	11
CES11 . . . . .	12
Chile . . . . .	13
Chirot . . . . .	14
Cowles . . . . .	14
Davis . . . . .	15
DavisThin . . . . .	16
Depredations . . . . .	17
Duncan . . . . .	17
Ericksen . . . . .	18
Florida . . . . .	19
Freedman . . . . .	20
Friendly . . . . .	20
Ginzberg . . . . .	21
Greene . . . . .	22
GSSvocab . . . . .	23
Guyer . . . . .	24
Hartnagel . . . . .	24
Highway1 . . . . .	25
KosteckiDillon . . . . .	26
Leinhardt . . . . .	27
LoBD . . . . .	28
Mandel . . . . .	30
Migration . . . . .	30
Moore . . . . .	31
MplsDemo . . . . .	32
MplsStops . . . . .	33
Mroz . . . . .	34
OBrienKaiser . . . . .	35
Ornstein . . . . .	36
Pottery . . . . .	37
Prestige . . . . .	37
Quartet . . . . .	38
Robey . . . . .	39
Rossi . . . . .	39
Sahlins . . . . .	42
Salaries . . . . .	42
SLID . . . . .	43
Soils . . . . .	44
States . . . . .	45
TitanicSurvival . . . . .	46
Transact . . . . .	47

<i>Adler</i>	3
UN . . . . .	47
UN98 . . . . .	48
USPop . . . . .	49
Vocab . . . . .	50
WeightLoss . . . . .	51
Wells . . . . .	51
Womenlf . . . . .	52
Wong . . . . .	53
Wool . . . . .	54
WVS . . . . .	55
<b>Index</b>	<b>56</b>

---

Adler	<i>Experimenter Expectations</i>
-------	----------------------------------

---

### Description

The Adler data frame has 108 rows and 3 columns.

The “experimenters” were the actual subjects of the study. They collected ratings of the apparent success of people in pictures who were pre-selected for their average appearance of success. The experimenters were told prior to collecting data that particular subjects were either high or low in their tendency to rate appearance of success, and were instructed to get good data, scientific data, or were given no such instruction. Each experimenter collected ratings from 18 randomly assigned subjects. This version of the Adler data is taken from Erickson and Nosanchuk (1977). The data described in the original source, Adler (1973), have a more complex structure.

### Usage

Adler

### Format

This data frame contains the following columns:

**instruction** a factor with levels: good, good data; none, no stress; scientific, scientific data.

**expectation** a factor with levels: high, expect high ratings; low, expect low ratings.

**rating** The average rating obtained.

### Source

Erickson, B. H., and Nosanchuk, T. A. (1977) *Understanding Data*. McGraw-Hill Ryerson.

### References

Adler, N. E. (1973) Impact of prior sets given experimenters and subjects on the experimenter expectancy effect. *Sociometry* **36**, 113–126.

---

AMSSurvey

*American Math Society Survey Data*

---

### Description

Counts of new PhDs in the mathematical sciences for 2008-09 and 2011-12 categorized by type of institution, gender, and US citizenship status.

### Usage

AMSSurvey

### Format

A data frame with 24 observations on the following 5 variables.

**type** a factor with levels I(Pu) for group I public universities, I(Pr) for group I private universities, II and III for groups II and III, IV for statistics and biostatistics programs, and Va for applied mathematics programs.

**sex** a factor with levels Female, Male of the recipient

**citizen** a factor with levels Non-US, US giving citizenship status

**count** The number of individuals of each type in 2008-09

**count11** The number of individuals of each type in 2011-12

### Details

These data are produced yearly by the American Math Society.

### Source

<http://www.ams.org/employment/surveyreports.html> Supplementary Table 4 in the 2008-09 data. See <http://www.ams.org/profession/data/annual-survey/docsgrtd> for more recent data.

### References

Fox, J. and Weisberg, S. (2011) *An R Companion to Applied Regression*, Second Edition, Sage.

Phipps, Polly, Maxwell, James W. and Rose, Colleen (2009), *2009 Annual Survey of the Mathematical Sciences*, 57, 250–259, Supplementary Table 4, <http://www.ams.org/employment/2009Survey-First-Report-Supp.pdf>

---

Angell

*Moral Integration of American Cities*

---

**Description**

The Angell data frame has 43 rows and 4 columns. The observations are 43 U. S. cities around 1950.

**Usage**

Angell

**Format**

This data frame contains the following columns:

**moral** Moral Integration: Composite of crime rate and welfare expenditures.

**hetero** Ethnic Heterogeneity: From percentages of nonwhite and foreign-born white residents.

**mobility** Geographic Mobility: From percentages of residents moving into and out of the city.

**region** A factor with levels: E Northeast; MW Midwest; S Southeast; W West.

**Source**

Angell, R. C. (1951) The moral integration of American Cities. *American Journal of Sociology* **57** (part 2), 1–140.

**References**

Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models*, Second Edition. Sage.

---

Anscombe

*U. S. State Public-School Expenditures*

---

**Description**

The Anscombe data frame has 51 rows and 4 columns. The observations are the U. S. states plus Washington, D. C. in 1970.

**Usage**

Anscombe

**Format**

This data frame contains the following columns:

**education** Per-capita education expenditures, dollars.

**income** Per-capita income, dollars.

**young** Proportion under 18, per 1000.

**urban** Proportion urban, per 1000.

**Source**

Anscombe, F. J. (1981) *Computing in Statistical Science Through APL*. Springer-Verlag.

**References**

Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models*, Second Edition. Sage.

---

Arrests

*Arrests for Marijuana Possession*

---

**Description**

Data on police treatment of individuals arrested in Toronto for simple possession of small quantities of marijuana. The data are part of a larger data set featured in a series of articles in the Toronto Star newspaper.

**Usage**

Arrests

**Format**

A data frame with 5226 observations on the following 8 variables.

**released** Whether or not the arrestee was released with a summons; a factor with levels: No; Yes.

**colour** The arrestee's race; a factor with levels: Black; White.

**year** 1997 through 2002; a numeric vector.

**age** in years; a numeric vector.

**sex** a factor with levels: Female; Male.

**employed** a factor with levels: No; Yes.

**citizen** a factor with levels: No; Yes.

**checks** Number of police data bases (of previous arrests, previous convictions, parole status, etc. – 6 in all) on which the arrestee's name appeared; a numeric vector

**Source**

Personal communication from Michael Friendly, York University.

**Examples**

```
summary(Arrests)
```

---

Baumann

*Methods of Teaching Reading Comprehension*

---

**Description**

The Baumann data frame has 66 rows and 6 columns. The data are from an experimental study conducted by Baumann and Jones, as reported by Moore and McCabe (1993) Students were randomly assigned to one of three experimental groups.

**Usage**

```
Baumann
```

**Format**

This data frame contains the following columns:

**group** Experimental group; a factor with levels: Basal, traditional method of teaching; DRTA, an innovative method; Strat, another innovative method.

**pretest.1** First pretest.

**pretest.2** Second pretest.

**post.test.1** First post-test.

**post.test.2** Second post-test.

**post.test.3** Third post-test.

**Source**

Moore, D. S. and McCabe, G. P. (1993) *Introduction to the Practice of Statistics, Second Edition*. Freeman, p. 794–795.

**References**

Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models*, Second Edition. Sage.

Fox, J. and Weisberg, S. (2011) *An R Companion to Applied Regression*, Second Edition, Sage.

---

BEPS

*British Election Panel Study*

---

### **Description**

These data are drawn from the 1997-2001 British Election Panel Study (BEPS).

### **Usage**

BEPS

### **Format**

A data frame with 1525 observations on the following 10 variables.

vote Party choice: Conservative, Labour, or Liberal Democrat

age in years

economic.cond.national Assessment of current national economic conditions, 1 to 5.

economic.cond.household Assessment of current household economic conditions, 1 to 5.

Blair Assessment of the Labour leader, 1 to 5.

Hague Assessment of the Conservative leader, 1 to 5.

Kennedy Assessment of the leader of the Liberal Democrats, 1 to 5.

Europe an 11-point scale that measures respondents' attitudes toward European integration. High scores represent 'Eurosceptic' sentiment.

political.knowledge Knowledge of parties' positions on European integration, 0 to 3.

gender female or male.

### **References**

J. Fox and R. Andersen (2006) Effect displays for multinomial and proportional-odds logit models. *Sociological Methodology* **36**, 225–255.

### **Examples**

summary(BEPS)



---

Bfox

*Canadian Women's Labour-Force Participation*

---

### Description

The Bfox data frame has 30 rows and 7 columns. Time-series data on Canadian women's labor-force participation, 1946–1975.

### Usage

Bfox

### Format

This data frame contains the following columns:

**partic** Percent of adult women in the workforce.

**tfr** Total fertility rate: expected births to a cohort of 1000 women at current age-specific fertility rates.

**menwage** Men's average weekly wages, in constant 1935 dollars and adjusted for current tax rates.

**womwage** Women's average weekly wages.

**debt** Per-capita consumer debt, in constant dollars.

**parttime** Percent of the active workforce working 34 hours per week or less.

### Warning

The value of tfr for 1973 is misrecorded as 2931; it should be 1931.

### Source

Fox, B. (1980) *Women's Domestic Labour and their Involvement in Wage Work*. Unpublished doctoral dissertation, p. 449.

### References

Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models*, Second Edition. Sage.

---

Blackmore

*Exercise Histories of Eating-Disordered and Control Subjects*

---

### Description

The Blackmore data frame has 945 rows and 4 columns. Blackmore and Davis's data on exercise histories of 138 teenaged girls hospitalized for eating disorders and 98 control subjects.

### Usage

Blackmore

### Format

This data frame contains the following columns:

**subject** a factor with subject id codes. There are several observations for each subject, but because the girls were hospitalized at different ages, the number of cases and the age at the last case vary.

**age** subject's age in years at the time of observation; all but the last observation for each subject were collected retrospectively at intervals of two years, starting at age 8.

**exercise** the amount of exercise in which the subject engaged, expressed as estimated hours per week.

**group** a factor with levels: control, Control subjects; patient, Eating-disordered patients.

### Source

Personal communication from Elizabeth Blackmore and Caroline Davis, York University.

---

Burt

*Fraudulent Data on IQs of Twins Raised Apart*

---

### Description

The Burt data frame has 27 rows and 4 columns. The "data" were simply (and notoriously) manufactured. The same data are in the dataset "twins" in the `alr3` package, but with different labels.

### Usage

Burt

**Format**

This data frame contains the following columns:

**IQbio** IQ of twin raised by biological parents

**IQfoster** IQ of twin raised by foster parents

**class** A factor with levels (note: out of order): high; low; medium.

**Source**

Burt, C. (1966) The genetic determination of differences in intelligence: A study of monozygotic twins reared together and apart. *British Journal of Psychology* **57**, 137–153.

---

CanPop	<i>Canadian Population Data</i>
--------	---------------------------------

---

**Description**

The CanPop data frame has 16 rows and 2 columns. Decennial time-series of Canadian population, 1851–2001.

**Usage**

CanPop

**Format**

This data frame contains the following columns:

**year** census year.

**population** Population, in millions

**Source**

Urquhart, M. C. and Buckley, K. A. H. (Eds.) (1965) *Historical Statistics of Canada*. Macmillan, p. 1369.

Canada (1994) *Canada Year Book*. Statistics Canada, Table 3.2.

Statistics Canada: <http://www12.statcan.ca/english/census01/products/standard/popdwell/Table-PR.cfm>.

**References**

Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models*, Second Edition. Sage.

---

CES11                                  *2011 Canadian National Election Study, With Attitude Toward Abortion*

---

### **Description**

Data are drawn from the 2011 Canadian National Election Study, including a question on banning abortion and variables related to the sampling design.

### **Usage**

```
data("CES11")
```

### **Format**

A data frame with 2231 observations on the following 9 variables.

`id` Household ID number.

`province` a factor with (alphabetical) levels AB, BC, MB, NB, NL, NS, ON, PE, QC, SK; the sample was stratified by province.

`population` population of the respondent's province, number over age 17.

`weight` weight sample to size of population, taking into account unequal sampling probabilities by province and household size.

`gender` a factor with levels Female, Male.

`abortion` attitude toward abortion, a factor with levels No, Yes; answer to the question "Should abortion be banned?"

`importance` importance of religion, a factor with (alphabetical) levels not, notvery, somewhat, very; answer to the question, "In your life, would you say that religion is very important, somewhat important, not very important, or not important at all?"

`education` a factor with (alphabetical) levels bachelors (Bachelors degree), college (community college or technical school), higher (graduate degree), HS (high-school graduate), lessHS (less than high-school graduate), somePS(some post-secondary).

`urban` place of residence, a factor with levels rural, urban.

### **Details**

This is an extract from the data set for the 2011 Canadian National Election Study distributed by the Institute for Social Research, York University.

### **References**

Fournier, P., Cutler, F., Soroka, S., and Stolle, D. (2013). Canadian Election Study 2011: Study documentation. Technical report, Canadian Opinion Research Archive, Queen's University, Kingston, Ontario.

Northrup, D. (2012). The 2011 Canadian Election Survey: Technical documentation. Technical report, Institute for Social Research, York University, Toronto, Ontario.

**Examples**

```
summary(CES11)
```

---

Chile	<i>Voting Intentions in the 1988 Chilean Plebiscite</i>
-------	---

---

**Description**

The Chile data frame has 2700 rows and 8 columns. The data are from a national survey conducted in April and May of 1988 by FLACSO/Chile. There are some missing data.

**Usage**

```
Chile
```

**Format**

This data frame contains the following columns:

**region** A factor with levels: C, Central; M, Metropolitan Santiago area; N, North; S, South; SA, city of Santiago.

**population** Population size of respondent's community.

**sex** A factor with levels: F, female; M, male.

**age** in years.

**education** A factor with levels (note: out of order): P, Primary; PS, Post-secondary; S, Secondary.

**income** Monthly income, in Pesos.

**statusquo** Scale of support for the status-quo.

**vote** a factor with levels: A, will abstain; N, will vote no (against Pinochet); U, undecided; Y, will vote yes (for Pinochet).

**Source**

Personal communication from FLACSO/Chile.

**References**

Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models*, Second Edition. Sage.

Fox, J. and Weisberg, S. (2011) *An R Companion to Applied Regression*, Second Edition, Sage.

---

 Chirot

*The 1907 Romanian Peasant Rebellion*


---

**Description**

The Chirot data frame has 32 rows and 5 columns. The observations are counties in Romania.

**Usage**

Chirot

**Format**

This data frame contains the following columns:

**intensity** Intensity of the rebellion

**commerce** Commercialization of agriculture

**tradition** Traditionalism

**midpeasant** Strength of middle peasantry

**inequality** Inequality of land tenure

**Source**

Chirot, D. and C. Ragin (1975) The market, tradition and peasant rebellion: The case of Romania. *American Sociological Review* **40**, 428–444 [Table 1].

**References**

Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models*, Second Edition. Sage.

---

 Cowles

*Cowles and Davis's Data on Volunteering*


---

**Description**

The Cowles data frame has 1421 rows and 4 columns. These data come from a study of the personality determinants of volunteering for psychological research.

**Usage**

Cowles

**Format**

This data frame contains the following columns:

**neuroticism** scale from Eysenck personality inventory

**extraversion** scale from Eysenck personality inventory

**sex** a factor with levels: female; male

**volunteer** volunteering, a factor with levels: no; yes

**Source**

Cowles, M. and C. Davis (1987) The subject matter of psychology: Volunteers. *British Journal of Social Psychology* **26**, 97–102.

---

Davis

*Self-Reports of Height and Weight*

---

**Description**

The Davis data frame has 200 rows and 5 columns. The subjects were men and women engaged in regular exercise. There are some missing data.

**Usage**

Davis

**Format**

This data frame contains the following columns:

**sex** A factor with levels: F, female; M, male.

**weight** Measured weight in kg.

**height** Measured height in cm.

**repwt** Reported weight in kg.

**reph** Reported height in cm.

**Source**

Personal communication from C. Davis, Departments of Physical Education and Psychology, York University.

**References**

Davis, C. (1990) Body image and weight preoccupation: A comparison between exercising and non-exercising women. *Appetite*, **15**, 13–21.

Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models*, Second Edition. Sage.

Fox, J. and Weisberg, S. (2011) *An R Companion to Applied Regression*, Second Edition, Sage.

---

DavisThin

*Davis's Data on Drive for Thinness*

---

### Description

The DavisThin data frame has 191 rows and 7 columns. This is part of a larger dataset for a study of eating disorders. The seven variables in the data frame comprise a "drive for thinness" scale, to be formed by summing the items.

### Usage

```
DavisThin
```

### Format

This data frame contains the following columns:

**DT1** a numeric vector

**DT2** a numeric vector

**DT3** a numeric vector

**DT4** a numeric vector

**DT5** a numeric vector

**DT6** a numeric vector

**DT7** a numeric vector

### Source

Davis, C., G. Claridge, and D. Cerullo (1997) Personality factors predisposing to weight preoccupation: A continuum approach to the association between eating disorders and personality disorders. *Journal of Psychiatric Research* **31**, 467–480. [personal communication from the authors.]

### References

Fox, J. and Weisberg, S. (2011) *An R Companion to Applied Regression*, Second Edition, Sage.



---

Depredations

*Minnesota Wolf Depredation Data*

---

**Description**

Wolf depredations of livestock on Minnesota farms, 1976-1998.

**Usage**

Depredations

**Format**

A data frame with 434 observations on the following 5 variables.

longitude longitude of the farm

latitude latitude of the farm

number number of depredations 1976-1998

early number of depredations 1991 or before

late number of depredations 1992 or later

**References**

Fox, J. and Weisberg, S. (2011) *An R Companion to Applied Regression*, Second Edition, Sage.

Harper, Elizabeth K. and Paul, William J. and Mech, L. David and Weisberg, Sanford (2008), Effectiveness of Lethal, Directed Wolf-Depredation Control in Minnesota, *Journal of Wildlife Management*, 72, 3, 778-784. <http://dx.doi.org/10.2193/2007-273>

---

Duncan

*Duncan's Occupational Prestige Data*

---

**Description**

The Duncan data frame has 45 rows and 4 columns. Data on the prestige and other characteristics of 45 U. S. occupations in 1950.

**Usage**

Duncan

**Format**

This data frame contains the following columns:

**type** Type of occupation. A factor with the following levels: prof, professional and managerial; wc, white-collar; bc, blue-collar.

**income** Percentage of occupational incumbents in the 1950 US Census who earned \$3,500 or more per year (about \$36,000 in 2017 US dollars).

**education** Percentage of occupational incumbents in 1950 who were high school graduates (which, were we cynical, we would say is roughly equivalent to a PhD in 2017)

**prestige** Percentage of respondents in a social survey who rated the occupation as “good” or better in prestige

**Source**

Duncan, O. D. (1961) A socioeconomic index for all occupations. In Reiss, A. J., Jr. (Ed.) *Occupations and Social Status*. Free Press [Table VI-1].

**References**

Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models*, Second Edition. Sage.

Fox, J. and Weisberg, S. (2011) *An R Companion to Applied Regression*, Second Edition, Sage.

---

Ericksen

*The 1980 U.S. Census Undercount*

---

**Description**

The Ericksen data frame has 66 rows and 9 columns. The observations are 16 large cities, the remaining parts of the states in which these cities are located, and the other U. S. states.

**Usage**

Ericksen

**Format**

This data frame contains the following columns:

**minority** Percentage black or Hispanic.

**crime** Rate of serious crimes per 1000 population.

**poverty** Percentage poor.

**language** Percentage having difficulty speaking or writing English.

**highschool** Percentage age 25 or older who had not finished highschool.

**housing** Percentage of housing in small, multiunit buildings.

**city** A factor with levels: city, major city; state, state or state-remainder.

**conventional** Percentage of households counted by conventional personal enumeration.

**undercount** Preliminary estimate of percentage undercount.

**Source**

Ericksen, E. P., Kadane, J. B. and Tukey, J. W. (1989) Adjusting the 1980 Census of Population and Housing. *Journal of the American Statistical Association* **84**, 927–944 [Tables 7 and 8].

**References**

Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models*, Second Edition. Sage.

Fox, J. and Weisberg, S. (2011) *An R Companion to Applied Regression*, Second Edition, Sage.

---

Florida

*Florida County Voting*

---

**Description**

The Florida data frame has 67 rows and 11 columns. Vote by county in Florida for President in the 2000 election.

**Usage**

Florida

**Format**

This data frame contains the following columns:

**GORE** Number of votes for Gore

**BUSH** Number of votes for Bush.

**BUCHANAN** Number of votes for Buchanan.

**NADER** Number of votes for Nader.

**BROWNE** Number of votes for Browne (whoever that is).

**HAGELIN** Number of votes for Hagelin (whoever that is).

**HARRIS** Number of votes for Harris (whoever that is).

**MCREYNOLDS** Number of votes for McReynolds (whoever that is).

**MOOREHEAD** Number of votes for Moorehead (whoever that is).

**PHILLIPS** Number of votes for Phillips (whoever that is).

**Total** Total number of votes.

**Source**

Adams, G. D. and Fastnow, C. F. (2000) A note on the voting irregularities in Palm Beach, FL. Formerly at '<http://madison.hss.cmu.edu/>', but no longer available there.

---

Freedman

*Crowding and Crime in U. S. Metropolitan Areas*

---

### Description

The Freedman data frame has 110 rows and 4 columns. The observations are U. S. metropolitan areas with 1968 populations of 250,000 or more. There are some missing data.

### Usage

Freedman

### Format

This data frame contains the following columns:

**population** Total 1968 population, 1000s.

**nonwhite** Percent nonwhite population, 1960.

**density** Population per square mile, 1968.

**crime** Crime rate per 100,000, 1969.

### Source

United States (1970) *Statistical Abstract of the United States*. Bureau of the Census.

### References

Fox, J. and Weisberg, S. (2011) *An R Companion to Applied Regression*, Second Edition, Sage.

Freedman, J. (1975) *Crowding and Behavior*. Viking.

---

Friendly

*Format Effects on Recall*

---

### Description

The Friendly data frame has 30 rows and 2 columns. The data are from an experiment on subjects' ability to remember words based on the presentation format.

### Usage

Friendly

**Format**

This data frame contains the following columns:

**condition** A factor with levels: Before, Recalled words presented before others; Meshed, Recalled words meshed with others; SFR, Standard free recall.

**correct** Number of words correctly recalled, out of 40 on final trial of the experiment.

**Source**

Friendly, M. and Franklin, P. (1980) Interactive presentation in multitrial free recall. *Memory and Cognition* **8** 265–270 [Personal communication from M. Friendly].

**References**

Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models*, Second Edition. Sage.

Fox, J. and Weisberg, S. (2011) *An R Companion to Applied Regression*, Second Edition, Sage.

---

Ginzberg

*Data on Depression*

---

**Description**

The Ginzberg data frame has 82 rows and 6 columns. The data are for psychiatric patients hospitalized for depression.

**Usage**

Ginzberg

**Format**

This data frame contains the following columns:

**simplicity** Measures subject's need to see the world in black and white.

**fatalism** Fatalism scale.

**depression** Beck self-report depression scale.

**adjsimp** Adjusted Simplicity: Simplicity adjusted (by regression) for other variables thought to influence depression.

**adjfatal** Adjusted Fatalism.

**adjdep** Adjusted Depression.

**Source**

Personal communication from Georges Monette, Department of Mathematics and Statistics, York University, with the permission of the original investigator.

## References

Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models*, Second Edition. Sage.

---

Greene

*Refugee Appeals*

---

## Description

The Greene data frame has 384 rows and 7 columns. These are cases filed in 1990, in which refugee claimants rejected by the Canadian Immigration and Refugee Board asked the Federal Court of Appeal for leave to appeal the negative ruling of the Board.

## Usage

Greene

## Format

This data frame contains the following columns:

**judge** Name of judge hearing case. A factor with levels: Desjardins, Heald, Hugessen, Iacobucci, MacGuigan, Mahoney, Marceau, Pratte, Stone, Urie.

**nation** Nation of origin of claimant. A factor with levels: Argentina, Bulgaria, China, Czechoslovakia, El. Salvador, Fiji, Ghana, Guatemala, India, Iran, Lebanon, Nicaragua, Nigeria, Pakistan, Poland, Somalia, Sri.Lanka.

**rater** Judgment of independent rater. A factor with levels: no, case has no merit; yes, case has some merit (leave to appeal should be granted).

**decision** Judge's decision. A factor with levels: no, leave to appeal not granted; yes, leave to appeal granted.

**language** Language of case. A factor with levels: English, French.

**location** Location of original refugee claim. A factor with levels: Montreal, other, Toronto.

**success** Logit of success rate, for all cases from the applicant's nation.

## Source

Personal communication from Ian Greene, Department of Political Science, York University.

## References

Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models*, Second Edition. Sage.

---

GSSvocab

*Data from the General Social Survey (GSS) from the National Opinion Research Center of the University of Chicago.*

---

### Description

This data set illustrates analysis of a multifactor observational study, with response given by subject's score on a vocabulary test, and factors for age group, education level, natality status, gender and year of the survey.

### Usage

```
data("GSSvocab")
```

### Format

A data frame with 28867 observations on the following 8 variables.

`year` a factor with levels 1978 1982 1984 1987 1988 1989 1990 1991 1993 1994 1996 1998 2000 2004 2006 2008 2010 2012 2014 2016. Data are included from the GSS for each of these years.

`gender` a factor with levels female male

`nativeBorn` Was the respondent born in the US? A factor with levels no and yes.

`ageGroup` a factor with levels 18-29 30-39 40-49 50-59 60+, grouped age of the respondent.

`educGroup` a factor with levels <12 yrs 12 yrs 13-15 yrs 16 yrs >16 yrs, grouped education level of the respondent. 12 years corresponds to high school graduate, 16 years to college graduate.

`vocab` Number of words out of 10 correct on a vocabulary test

`age` age of the respondent in years

`educ` years of education of the respondent

### Details

This file includes the years of the GSS for which the `vocab` and `nativeBorn` items were included.

### Source

These data were collected from the GSS data explorer <https://gssdataexplorer.norc.org>, using the data fields `year`, `id_`, `age`, `educ`, `sex`, `born` and `wordsum`. The GSS began in 1972, and has included several thousand data items, some regularly and some only once, on topics of interest to social scientists. Data have been slightly edited to change entires like `No` answer and `Not` applicable to missing value codes.

### Examples

```
data(GSSvocab)
```

---

 Guyer

*Anonymity and Cooperation*


---

### Description

The Guyer data frame has 20 rows and 3 columns. The data are from an experiment in which four-person groups played a prisoner's dilemma game for 30 trials, each person making either a cooperative or competitive choice on each trial. Choices were made either anonymously or in public; groups were composed either of females or of males. The observations are 20 groups.

### Usage

Guyer

### Format

This data frame contains the following columns:

**cooperation** Number of cooperative choices (out of 120 in all).

**condition** A factor with levels: anonymous, Anonymous choice; public, Public choice.

**sex** Sex. A factor with levels: female and male.

### Source

Fox, J. and Guyer, M. (1978) Public choice and cooperation in n-person prisoner's dilemma. *Journal of Conflict Resolution* **22**, 469–481.

### References

Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models*, Second Edition. Sage.

Fox, J. and Weisberg, S. (2011) *An R Companion to Applied Regression*, Second Edition, Sage.

---

 Hartnagel

*Canadian Crime-Rates Time Series*


---

### Description

The Hartnagel data frame has 38 rows and 7 columns. The data are an annual time-series from 1931 to 1968. There are some missing data.

### Usage

Hartnagel



**Format**

This data frame contains the following columns:

**year** 1931–1968.

**tfr** Total fertility rate per 1000 women.

**partic** Women’s labor-force participation rate per 1000.

**degrees** Women’s post-secondary degree rate per 10,000.

**fconvict** Female indictable-offense conviction rate per 100,000.

**ftheft** Female theft conviction rate per 100,000.

**mconvict** Male indictable-offense conviction rate per 100,000.

**mtheft** Male theft conviction rate per 100,000.

**Details**

The post-1948 crime rates have been adjusted to account for a difference in method of recording. Some of your results will differ in the last decimal place from those in Table 14.1 of Fox (1997) due to rounding of the data. Missing values for 1950 were interpolated.

**Source**

Personal communication from T. Hartnagel, Department of Sociology, University of Alberta.

**References**

Fox, J., and Hartnagel, T. F (1979) Changing social roles and female crime in Canada: A time series analysis. *Canadian Review of Sociology and Anthropology*, **16**, 96–104.

Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models*, Second Edition. Sage.

---

Highway1

*Highway Accidents*

---

**Description**

The data comes from an unpublished master’s paper by Carl Hoffstedt. They relate the automobile accident rate, in accidents per million vehicle miles to several potential terms. The data include 39 sections of large highways in the state of Minnesota in 1973. The goal of this analysis was to understand the impact of design variables, Acpts, Slim, Sig, and Shld that are under the control of the highway department, on accidents.

**Usage**

Highway1

**Format**

This data frame contains the following columns:

**rate** 1973 accident rate per million vehicle miles

**len** length of the Highway1 segment in miles

**adt** average daily traffic count in thousands

**trks** truck volume as a percent of the total volume

**sigs1** (number of signalized interchanges per mile times len + 1)/len, the number of signals per mile of roadway, adjusted to have no zero values.

**slim** speed limit in 1973

**shld** width in feet of outer shoulder on the roadway

**lane** total number of lanes of traffic

**acpt** number of access points per mile

**itg** number of freeway-type interchanges per mile

**lwid** lane width, in feet

**htype** An indicator of the type of roadway or the source of funding for the road, either MC, FAI, PA, or MA

**Source**

Carl Hoffstedt. This differs from the dataset Highway in the alr4 package only by addition of transformation of some of the columns.

**References**

Fox, J. and Weisberg, S. (2011) *An R Companion to Applied Regression*, Second Edition, Sage.

Weisberg, S. (2014) *Applied Linear Regression*, Fourth Edition, Wiley, Section 7.2.

---

KosteckiDillon

*Treatment of Migraine Headaches*

---

**Description**

Subset of data on migraine treatments collected by Tammy Kostecki-Dillon.

**Usage**

KosteckiDillon

**Format**

A data frame with 4152 observations on 133 subjects for the following 9 variables.

id Patient id.

time time in days relative to the onset of treatment, which occurs at time 0.

dos time in days from the start of the study, January 1 of the first year of the study.

hatype a factor with levels Aura Mixed No Aura, the type of migraine experienced by a subject.

age at onset of treatment, in years.

airq a measure of air quality.

medication a factor with levels none reduced continuing, representing subjects who discontinued their medication, who continued but at a reduced dose, or who continued at the previous dose.

headache a factor with levels no yes.

sex a factor with levels female male.

**Details**

The data consist of headache logs kept by 133 patients in a treatment program in which bio-feedback was used to attempt to reduce migraine frequency and severity. Patients entered the program at different times over a period of about 3 years. Patients were encouraged to begin their logs four weeks before the onset of treatment and to continue for one month afterwards, but only 55 patients have data preceding the onset of treatment.

**Source**

Personal communication from Georges Monette (and adapted from his description of the data).

**References**

Kostecki-Dillon, T., Monette, G., and Wong, P. (1999). Pine trees, comas, and migraines. *York University Institute for Social Research Newsletter*, 14:2.

**Examples**

```
summary(KosteckiDillon)
```

---

Leinhardt

*Data on Infant-Mortality*

---

**Description**

The Leinhardt data frame has 105 rows and 4 columns. The observations are nations of the world around 1970.

**Usage**

Leinhardt

**Format**

This data frame contains the following columns:

**income** Per-capita income in U. S. dollars.

**infant** Infant-mortality rate per 1000 live births.

**region** A factor with levels: Africa; Americas; Asia, Asia and Oceania; Europe.

**oil** Oil-exporting country. A factor with levels: no, yes.

**Details**

The infant-mortality rate for Jamaica is misprinted in Leinhardt and Wasserman; the correct value is given here. Some of the values given in Leinhardt and Wasserman do not appear in the original New York Times table and are of dubious validity.

**Source**

Leinhardt, S. and Wasserman, S. S. (1979) Exploratory data analysis: An introduction to selected methods. In Schuessler, K. (Ed.) *Sociological Methodology 1979* Jossey-Bass.

*The New York Times*, 28 September 1975, p. E-3, Table 3.

**References**

Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models*, Second Edition. Sage.

Fox, J. and Weisberg, S. (2011) *An R Companion to Applied Regression*, Second Edition, Sage.

---

LoBD

*Cancer drug data use to provide an example of the use of the skew power distributions.*

---

**Description**

A portion of an experiment to determine the limit of blank/limit of detection in a biochemical assay.

**Usage**

LoBD

**Format**

A data frame with 84 observations on the following 9 variables.

pool a factor with levels 1 2 3 4 5 6 7 8 9 10 11 12 denoting the 12 pools used in the experiment; each pool had a different level of drug.

I1L1 a numeric vector giving the measured concentration in pmol/L of drug in the assay

I1L2 a numeric vector giving the measured concentration in pmol/L of drug in the assay

I2L1 a numeric vector giving the measured concentration in pmol/L of drug in the assay

I2L2 a numeric vector giving the measured concentration in pmol/L of drug in the assay

I3L1 a numeric vector giving the measured concentration in pmol/L of drug in the assay

I3L2 a numeric vector giving the measured concentration in pmol/L of drug in the assay

I4L1 a numeric vector giving the measured concentration in pmol/L of drug in the assay

I4L2 a numeric vector giving the measured concentration in pmol/L of drug in the assay

**Details**

Important characteristics of a clinical chemistry assay are its limit of blank (LoB), and its limit of detection (LoD). The LoB, conceptually the highest reading likely to be obtained from a zero-concentration sample, is defined operationally by the upper 95% point of readings obtained from samples that do not contain the analyte. The LoD, conceptually the lowest level of analyte that can be reliably determined not to be blank, is defined operationally as true value at which there is a 95% chance of the reading being above the LoB.

These data are from a portion of a LoB/D study of an assay for a drug used to treat certain cancers. Twelve pools were used, four of them blanks of different types, and eight with successively increasing drug levels. The 8 columns of the data set refer to measurements made using different instruments I and reagent lots L.

**Source**

Used as an illustrative example for Box-Cox type transformations with negative readings in Hawkins and Weisberg (2015). For examples of its use, see [bcnPower](#).

**References**

Hawkins, D. and Weisberg, S. (2015) Combining the Box-Cox Power and Generalized Log Transformations to Accommodate Negative Responses, submitted for publication.

**Examples**

LoBD

---

Mandel	<i>Contrived Collinear Data</i>
--------	---------------------------------

---

**Description**

The Mandel data frame has 8 rows and 3 columns.

**Usage**

Mandel

**Format**

This data frame contains the following columns:

**x1** first predictor.

**x2** second predictor.

**y** response.

**Source**

Mandel, J. (1982) Use of the singular value decomposition in regression analysis. *The American Statistician* **36**, 15–24.

**References**

Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models*, Second Edition. Sage.

---

Migration	<i>Canadian Interprovincial Migration Data</i>
-----------	--

---

**Description**

The Migration data frame has 90 rows and 8 columns.

**Usage**

Migration

## Format

This data frame contains the following columns:

**source** Province of origin (source). A factor with levels: ALTA, Alberta; BC, British Columbia; MAN, Manitoba; NB, New Brunswick; NFLD, New Foundland; NS, Nova Scotia; ONT, Ontario; PEI, Prince Edward Island; QUE, Quebec; SASK, Saskatchewan.

**destination** Province of destination (1971 residence). A factor with levels: ALTA, Alberta; BC, British Columbia; MAN, Manitoba; NB, New Brunswick; NFLD, New Foundland; NS, Nova Scotia; ONT, Ontario; PEI, Prince Edward Island; QUE, Quebec; SASK, Saskatchewan.

**migrants** Number of migrants (from source to destination) in the period 1966–1971.

**distance** Distance (between principal cities of provinces): NFLD, St. John; PEI, Charlottetown; NS, Halifax; NB, Fredricton; QUE, Montreal; ONT, Toronto; MAN, Winnipeg; SASK, Regina; ALTA, Edmonton; BC, Vancouver.

**pops66** 1966 population of source province.

**pops71** 1971 population of source province.

**popd66** 1966 population of destination province.

**popd71** 1971 population of destination province.

## Details

There is one record in the data file for each migration stream. You can average the 1966 and 1971 population figures for each of the source and destination provinces.

## Source

Canada (1962) *Map*. Department of Mines and Technical Surveys.

Canada (1971) *Census of Canada*. Statistics Canada, Vol. 1, Part 2 [Table 32].

Canada (1972) *Canada Year Book*. Statistics Canada [p. 1369].

## References

Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models*, Second Edition. Sage.

---

Moore

*Status, Authoritarianism, and Conformity*

---

## Description

The Moore data frame has 45 rows and 4 columns. The data are for subjects in a social-psychological experiment, who were faced with manipulated disagreement from a partner of either of low or high status. The subjects could either conform to the partner's judgment or stick with their own judgment.

## Usage

Moore

**Format**

This data frame contains the following columns:

**partner.status** Partner's status. A factor with levels: high, low.

**conformity** Number of conforming responses in 40 critical trials.

**fcategory** F-Scale Categorized. A factor with levels (note levels out of order): high, low, medium.

**fscore** Authoritarianism: F-Scale score.

**Source**

Moore, J. C., Jr. and Krupat, E. (1971) Relationship between source status, authoritarianism and conformity in a social setting. *Sociometry* **34**, 122–134.

Personal communication from J. Moore, Department of Sociology, York University.

**References**

Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models*, Second Edition. Sage.

Fox, J. and Weisberg, S. (2011) *An R Companion to Applied Regression*, Second Edition, Sage.

---

MplsDemo

*Minneapolis Demographic Data 2015, by Neighborhood*


---

**Description**

Minneapolis Demographic Data 2015, by Neighborhood, from the 2015 American Community Survey

**Format**

A data frame with 84 observations on the following 7 variables.

neighborhood name of the neighborhood

population total population

black fraction of the population estimated to be black

white fraction of the population estimated to be white

foreignBorn fraction of the population estimated to be foreign born

hhIncome estimated median household income

poverty estimated fraction earning less than twice the poverty level

collegeGrad estimated fraction with a college degree

**Details**

The data frame [MplsStops](#) contains 2017 Minneapolis Police stop data, using the same neighborhood names as this data file.



**Source**

<http://www.mncompass.org/profiles/neighborhoods/minneapolis-saint-paul#!community-areas>

**Examples**

```
str(MplsDemo)
```

---

MplsStops

*Minneapolis Police Department 2017 Stop Data*

---

**Description**

Results of nearly all stops made by the Minneapolis Police Department for the year 2017.

**Format**

A data frame with 51857 observations on the following 14 variables.

`idNum` character vector of incident identifiers

`date` a POSIXlt date variable giving the date and time of the stop

`problem` a factor with levels suspicious for suspicious vehicle or person stops and traffic for traffic stops

`citationIssued` a factor with levels no yes indicating if a citation was issued

`personSearch` a factor with levels no yes indicating if the stopped person was searched

`vehicleSearch` a factor with levels no or yes indicating if a vehicle was searched

`preRace` a factor with levels white, black, east african, latino, native american, asian, other, unknown for the officer's assessment of race of the person stopped before speaking with the person stopped

`race` a factor with levels white, black, east african, latino, native american, asian, other, unknown, officer's determination of race after the incident

`gender` a factor with levels female, male, unknown, gender of person stopped

`lat` latitude of the location of the incident, somewhat rounded

`long` longitude of the location of the incident, somewhat rounded

`policePrecinct` Minneapolis Police Precinct number

`neighborhood` a factor with 84 levels giving the name of the Minneapolis neighborhood of the incident

`MDC` a factor with levels mdc for data collected via in-vehicle computer, and other for data submitted by officers not in a vehicle, either on foot, bicycle or horseback. Several of the variables above were recorded only in-vehicle

**Details**

A few stops have been deleted, either because the location data was missing, or a few very rare categories were also removed. The data frame `MplsDemo` contains 2015 demographic data on Minneapolis neighborhoods, using the same neighborhood names as this data file. Demographics are available for 84 of Minneapolis' 87 neighborhoods. The remaining 3 presumably have no housing.

**Source**

These are public data obtained from <http://opendata.minneapolismn.gov/datasets/police-stop-data>. A few more fields, and more data, are available at the original source

**Examples**

```
summary(MplsStops)
```

---

Mroz

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


---

**Description**

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

**Usage**

```
Mroz
```

**Format**

This data frame contains the following columns:

**lfp** labor-force participation; a factor with levels: no; yes.

**k5** number of children 5 years old or younger.

**k618** number of children 6 to 18 years old.

**age** in years.

**wc** wife's college attendance; a factor with levels: no; yes.

**hc** husband's college attendance; a factor with levels: no; yes.

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

**inc** family income exclusive of wife's income.

**Source**

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

## References

- Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models*, Second Edition. Sage.
- Fox, J. (2000) *Multiple and Generalized Nonparametric Regression*. Sage.
- Fox, J. and Weisberg, S. (2011) *An R Companion to Applied Regression*, Second Edition, Sage.
- Long, J. S. (1997) *Regression Models for Categorical and Limited Dependent Variables*. Sage.

---

 OBrienKaiser

*O'Brien and Kaiser's Repeated-Measures Data*


---

## Description

These contrived repeated-measures data are taken from O'Brien and Kaiser (1985). The data are from an imaginary study in which 16 female and male subjects, who are divided into three treatments, are measured at a pretest, posttest, and a follow-up session; during each session, they are measured at five occasions at intervals of one hour. The design, therefore, has two between-subject and two within-subject factors.

The contrasts for the treatment factor are set to  $-2, 1, 1$  and  $0, -1, 1$ . The contrasts for the gender factor are set to `contr.sum`.

## Usage

```
OBrienKaiser
```

## Format

A data frame with 16 observations on the following 17 variables.

```
treatment a factor with levels control A B
gender a factor with levels F M
pre.1 pretest, hour 1
pre.2 pretest, hour 2
pre.3 pretest, hour 3
pre.4 pretest, hour 4
pre.5 pretest, hour 5
post.1 posttest, hour 1
post.2 posttest, hour 2
post.3 posttest, hour 3
post.4 posttest, hour 4
post.5 posttest, hour 5
fup.1 follow-up, hour 1
fup.2 follow-up, hour 2
fup.3 follow-up, hour 3
fup.4 follow-up, hour 4
fup.5 follow-up, hour 5
```

**Source**

O'Brien, R. G., and Kaiser, M. K. (1985) MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin* **97**, 316–333, Table 7.

**Examples**

```
OBrienKaiser
contrasts(OBrienKaiser$treatment)
contrasts(OBrienKaiser$gender)
```

---

 Ornstein

---

*Interlocking Directorates Among Major Canadian Firms*


---

**Description**

The Ornstein data frame has 248 rows and 4 columns. The observations are the 248 largest Canadian firms with publicly available information in the mid-1970s. The names of the firms were not available.

**Usage**

```
Ornstein
```

**Format**

This data frame contains the following columns:

**assets** Assets in millions of dollars.

**sector** Industrial sector. A factor with levels: AGR, agriculture, food, light industry; BNK, banking; CON, construction; FIN, other financial; HLD, holding companies; MAN, heavy manufacturing; MER, merchandizing; MIN, mining, metals, etc.; TRN, transport; WOD, wood and paper.

**nation** Nation of control. A factor with levels: CAN, Canada; OTH, other foreign; UK, Britain; US, United States.

**interlocks** Number of interlocking director and executive positions shared with other major firms.

**Source**

Ornstein, M. (1976) The boards and executives of the largest Canadian corporations. *Canadian Journal of Sociology* **1**, 411–437.

Personal communication from M. Ornstein, Department of Sociology, York University.

**References**

Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models*, Second Edition. Sage.

Fox, J. and Weisberg, S. (2011) *An R Companion to Applied Regression*, Second Edition, Sage.

---

Pottery

*Chemical Composition of Pottery*

---

**Description**

The data give the chemical composition of ancient pottery found at four sites in Great Britain. They appear in Hand, et al. (1994), and are used to illustrate MANOVA in the SAS Manual. (Suggested by Michael Friendly.)

**Usage**

Pottery

**Format**

A data frame with 26 observations on the following 6 variables.

Site a factor with levels AshleyRails Caldicot IsleThorns Llanedyrn

Al Aluminum

Fe Iron

Mg Magnesium

Ca Calcium

Na Sodium

**Source**

Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J., and E., O. (1994) *A Handbook of Small Data Sets*. Chapman and Hall.

**Examples**

Pottery

---

Prestige

*Prestige of Canadian Occupations*

---

**Description**

The Prestige data frame has 102 rows and 6 columns. The observations are occupations.

**Usage**

Prestige

**Format**

This data frame contains the following columns:

**education** Average education of occupational incumbents, years, in 1971.

**income** Average income of incumbents, dollars, in 1971.

**women** Percentage of incumbents who are women.

**prestige** Pineo-Porter prestige score for occupation, from a social survey conducted in the mid-1960s.

**census** Canadian Census occupational code.

**type** Type of occupation. A factor with levels (note: out of order): bc, Blue Collar; prof, Professional, Managerial, and Technical; wc, White Collar.

**Source**

Canada (1971) *Census of Canada*. Vol. 3, Part 6. Statistics Canada [pp. 19-1–19-21].

Personal communication from B. Blishen, W. Carroll, and C. Moore, Departments of Sociology, York University and University of Victoria.

**References**

Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models*, Second Edition. Sage.

Fox, J. and Weisberg, S. (2011) *An R Companion to Applied Regression*, Second Edition, Sage.

---

Quartet

*Four Regression Datasets*

---

**Description**

The Quartet data frame has 11 rows and 5 columns. These are contrived data.

**Usage**

Quartet

**Format**

This data frame contains the following columns:

**x** X-values for datasets 1–3.

**y1** Y-values for dataset 1.

**y2** Y-values for dataset 2.

**y3** Y-values for dataset 3.

**x4** X-values for dataset 4.

**y4** Y-values for dataset 4.

**Source**

Anscombe, F. J. (1973) Graphs in statistical analysis. *American Statistician* **27**, 17–21.

---

 Robey

*Fertility and Contraception*


---

**Description**

The Robey data frame has 50 rows and 3 columns. The observations are developing nations around 1990.

**Usage**

Robey

**Format**

This data frame contains the following columns:

**region** A factor with levels: Africa; Asia, Asia and Pacific; Latin.Amer, Latin America and Caribbean; Near . East, Near East and North Africa.

**tfr** Total fertility rate (children per woman).

**contraceptors** Percent of contraceptors among married women of childbearing age.

**Source**

Robey, B., Shea, M. A., Rutstein, O. and Morris, L. (1992) The reproductive revolution: New survey findings. *Population Reports*. Technical Report M-11.

**References**

Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models*, Second Edition. Sage.

---

 Rossi

*Rossi et al.'s Criminal Recidivism Data*


---

**Description**

This data set is originally from Rossi et al. (1980), and is used as an example in Allison (1995). The data pertain to 432 convicts who were released from Maryland state prisons in the 1970s and who were followed up for one year after release. Half the released convicts were assigned at random to an experimental treatment in which they were given financial aid; half did not receive aid.

**Usage**

Rossi

**Format**

A data frame with 432 observations on the following 62 variables.

week week of first arrest after release or censoring; all censored observations are censored at 52 weeks.

arrest 1 if arrested, 0 if not arrested.

fin financial aid: no yes.

age in years at time of release.

race black or other.

wexp full-time work experience before incarceration: no or yes.

mar marital status at time of release: married or not married.

paro released on parole? no or yes.

prio number of convictions prior to current incarceration.

educ level of education: 2 = 6th grade or less; 3 = 7th to 9th grade; 4 = 10th to 11th grade; 5 = 12th grade; 6 = some college.

emp1 employment status in the first week after release: no or yes.

emp2 as above.

emp3 as above.

emp4 as above.

emp5 as above.

emp6 as above.

emp7 as above.

emp8 as above.

emp9 as above.

emp10 as above.

emp11 as above.

emp12 as above.

emp13 as above.

emp14 as above.

emp15 as above.

emp16 as above.

emp17 as above.

emp18 as above.

emp19 as above.

emp20 as above.

emp21 as above.

emp22 as above.

emp23 as above.



emp24 as above.  
emp25 as above.  
emp26 as above.  
emp27 as above.  
emp28 as above.  
emp29 as above.  
emp30 as above.  
emp31 as above.  
emp32 as above.  
emp33 as above.  
emp34 as above.  
emp35 as above.  
emp36 as above.  
emp37 as above.  
emp38 as above.  
emp39 as above.  
emp40 as above.  
emp41 as above.  
emp42 as above.  
emp43 as above.  
emp44 as above.  
emp45 as above.  
emp46 as above.  
emp47 as above.  
emp48 as above.  
emp49 as above.  
emp50 as above.  
emp51 as above.  
emp52 as above.

### Source

Allison, P.D. (1995). *Survival Analysis Using the SAS System: A Practical Guide*. Cary, NC: SAS Institute.

### References

Rossi, P.H., R.A. Berk, and K.J. Lenihan (1980). *Money, Work, and Crime: Some Experimental Results*. New York: Academic Press.

John Fox, Marilia Sa Carvalho (2012). The RcmdrPlugin.survival Package: Extending the R Commander Interface to Survival Analysis. *Journal of Statistical Software*, 49(7), 1-32.

**Examples**

```
summary(Rossi)
```

---

Sahlins	<i>Agricultural Production in Mazulu Village</i>
---------	--

---

**Description**

The Sahlins data frame has 20 rows and 2 columns. The observations are households in a Central African village.

**Usage**

```
Sahlins
```

**Format**

This data frame contains the following columns:

**consumers** Consumers/Gardener, ratio of consumers to productive individuals.

**acres** Acres/Gardener, amount of land cultivated per gardener.

**Source**

Sahlins, M. (1972) *Stone Age Economics*. Aldine [Table 3.1].

**References**

Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models*, Second Edition. Sage.

---

Salaries	<i>Salaries for Professors</i>
----------	--------------------------------

---

**Description**

The 2008-09 nine-month academic salary for Assistant Professors, Associate Professors and Professors in a college in the U.S. The data were collected as part of the on-going effort of the college's administration to monitor salary differences between male and female faculty members.

**Usage**

```
Salaries
```

**Format**

A data frame with 397 observations on the following 6 variables.

**rank** a factor with levels AssocProf AsstProf Prof

**discipline** a factor with levels A (“theoretical” departments) or B (“applied” departments).

**yrs.since.phd** years since PhD.

**yrs.service** years of service.

**sex** a factor with levels Female Male

**salary** nine-month salary, in dollars.

**References**

Fox J. and Weisberg, S. (2011) *An R Companion to Applied Regression*, Second Edition Sage.

---

SLID

*Survey of Labour and Income Dynamics*

---

**Description**

The SLID data frame has 7425 rows and 5 columns. The data are from the 1994 wave of the Canadian Survey of Labour and Income Dynamics, for the province of Ontario. There are missing data, particularly for wages.

**Usage**

SLID

**Format**

This data frame contains the following columns:

**wages** Composite hourly wage rate from all jobs.

**education** Number of years of schooling.

**age** in years.

**sex** A factor with levels: Female, Male.

**language** A factor with levels: English, French, Other.

**Source**

The data are taken from the public-use dataset made available by Statistics Canada, and prepared by the Institute for Social Research, York University.

**References**

Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models*, Second Edition. Sage.

Fox, J. and Weisberg, S. (2011) *An R Companion to Applied Regression*, Second Edition, Sage.

Soils

*Soil Compositions of Physical and Chemical Characteristics***Description**

Soil characteristics were measured on samples from three types of contours (Top, Slope, and Depression) and at four depths (0-10cm, 10-30cm, 30-60cm, and 60-90cm). The area was divided into 4 blocks, in a randomized block design. (Suggested by Michael Friendly.)

**Usage**

Soils

**Format**

A data frame with 48 observations on the following 14 variables. There are 3 factors and 9 response variables.

Group a factor with 12 levels, corresponding to the combinations of Contour and Depth

Contour a factor with 3 levels: Depression Slope Top

Depth a factor with 4 levels: 0-10 10-30 30-60 60-90

Gp a factor with 12 levels, giving abbreviations for the groups: D0 D1 D3 D6 S0 S1 S3 S6 T0 T1 T3 T6

Block a factor with levels 1 2 3 4

pH soil pH

N total nitrogen in %

Dens bulk density in gm/cm<sup>3</sup>

P total phosphorous in ppm

Ca calcium in me/100 gm.

Mg magnesium in me/100 gm.

K phosphorous in me/100 gm.

Na sodium in me/100 gm.

Conduc conductivity

**Details**

These data provide good examples of MANOVA and canonical discriminant analysis in a somewhat complex multivariate setting. They may be treated as a one-way design (ignoring Block), by using either Group or Gp as the factor, or a two-way randomized block design using Block, Contour and Depth (quantitative, so orthogonal polynomial contrasts are useful).

**Source**

Horton, I. F., Russell, J. S., and Moore, A. W. (1968) Multivariate-covariance and canonical analysis: A method for selecting the most effective discriminators in a multivariate situation. *Biometrics* **24**, 845–858. Originally from 'http://www.stat.lsu.edu/faculty/moser/exst7037/soils.sas' but no longer available there.

**References**

Khattree, R., and Naik, D. N. (2000) *Multivariate Data Reduction and Discrimination with SAS Software*. SAS Institute.

Friendly, M. (2006) Data ellipses, HE plots and reduced-rank displays for multivariate linear models: SAS software and examples. *Journal of Statistical Software*, 17(6), <http://www.jstatsoft.org/v17/i06>.

---

States

*Education and Related Statistics for the U.S. States*

---

**Description**

The States data frame has 51 rows and 8 columns. The observations are the U. S. states and Washington, D. C.

**Usage**

States

**Format**

This data frame contains the following columns:

**region** U. S. Census regions. A factor with levels: ENC, East North Central; ESC, East South Central; MA, Mid-Atlantic; MTN, Mountain; NE, New England; PAC, Pacific; SA, South Atlantic; WNC, West North Central; WSC, West South Central.

**pop** Population: in 1,000s.

**SATV** Average score of graduating high-school students in the state on the *verbal* component of the Scholastic Aptitude Test (a standard university admission exam).

**SATM** Average score of graduating high-school students in the state on the *math* component of the Scholastic Aptitude Test.

**percent** Percentage of graduating high-school students in the state who took the SAT exam.

**dollars** State spending on public education, in \ \$1000s per student.

**pay** Average teacher's salary in the state, in \$1000s.

**Source**

United States (1992) *Statistical Abstract of the United States*. Bureau of the Census.

## References

Moore, D. (1995) *The Basic Practice of Statistics*. Freeman, Table 2.1.

---

TitanicSurvival

*Survival of Passengers on the Titanic*

---

## Description

Information on the survival status, sex, age, and passenger class of 1309 passengers in the Titanic disaster of 1912.

## Usage

```
TitanicSurvival
```

## Format

A data frame with 1309 observations on the following 4 variables.

survived no or yes.

sex female or male

age in years (and for some children, fractions of a year); age is missing for 263 of the passengers.

passengerClass 1st, 2nd, or 3rd class.

## Details

This is part of a larger data set compiled by Thomas Cason. Many additional details are given in the sources cited below.

## Source

Data set titanic3 from <http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/DataSets>.

## References

<http://www.encyclopedia-titanica.org/>

F. E. Harrell, Jr. (2001) *Regression Modeling Strategies* New York: Springer.

## Examples

```
summary(TitanicSurvival)
```

---

Transact	<i>Transaction data</i>
----------	-------------------------

---

**Description**

Data on transaction times in branch offices of a large Australian bank.

**Usage**

Transact

**Format**

This data frame contains the following columns:

**t1** number of type 1 transactions

**t2** number of type 2 transactions

**time** total transaction time, minutes

**Source**

Cunningham, R. and Heathcote, C. (1989), Estimating a non-Gaussian regression model with multicollinearity. *Australian Journal of Statistics*, 31,12-17.

**References**

Fox, J. and Weisberg, S. (2011) *An R Companion to Applied Regression*, Second Edition, Sage.

Weisberg, S. (2014) *Applied Linear Regression*, Fourth Edition, Wiley, Section 4.6.1.

---

UN	<i>National Statistics from the United Nations, Mostly From 2009–2011</i>
----	---

---

**Description**

National health, welfare, and education statistics for 213 places, mostly UN members, but also other areas like Hong Kong that are not independent countries.

**Usage**

data(UN)

**Format**

A data frame with 213 rows on the following 7 variables.

region Region of the world: Africa, Asia, Caribbean, Europe, Latin Amer, North America, NorthAtlantic, Oceania.

group A factor with levels oecd for countries that are members of the OECD, the Organization for Economic Co-operation and Development, as of May 2012, africa for countries on the African continent, and other for all other countries. No OECD countries are located in Africa.

fertility Total fertility rate, number of children per woman.

ppgdp Per capita gross domestic product in US dollars.

lifeExpF Female life expectancy, years.

pctUrban Percent urban.

infantMortality Infant deaths by age 1 year per 1000 live births

**Note**

Similar data, from the period 2000-2003, appear in the `alr3` package under the name `UN3`. This data set was formerly named `UN11a` and replaces the older dataset named `UN`.

**Source**

All data were collected from UN tables accessed at <http://unstats.un.org/unsd/demographic/products/socind/> on April 23, 2012. OECD membership is from [www.oecd.org](http://www.oecd.org), accessed May 25, 2012.

**References**

Weisberg, S. (2014). *Applied Linear Regression*, 4th edition. Hoboken NJ: Wiley.

**Examples**

```
summary(UN)
```

---

UN98

*United Nations Social Indicators Data 1998]*

---

**Description**

Social indicators data on 207 nations distributed by the United Nations circa 1998.

**Usage**

```
data("UN98")
```



**Format**

A data frame with 207 observations on the following 13 variables.

region a factor with alphabetical levels Africa, America, Asia, Europe, Oceania.

tfr total fertility rate, number of children per woman.

contraception percentage of married women using any method of contraception.

educationMale average number of years of education for men.

educationFemale average number of years of education for women.

lifeMale expectation of life at birth for males.

lifeFemale expectation of life at birth for females.

infantMortality infant deaths per 1000 live births.

GDPperCapita gross domestic product per person in U.S. dollars.

economicActivityMale percentage of men who are economically active.

economicActivityFemale percentage of women who are economically active.

illiteracyMale percentage of males 15 years of age and older who are illiterate.

illiteracyFemale percentage of females 15 years of age and older who are illiterate.

**Details**

In a few cases where the percentages of males and females 15 and older who are illiterate were unavailable, these variables were filled in by regression imputation from the corresponding percentages 25 and older who are illiterate.

**Source**

Downloaded from <http://www.un.org/Depts/unsd/social/main.htm> in 1998.

**Examples**

```
summary(UN98)
```

---

USPop

*Population of the United States*

---

**Description**

The USPop data frame has 22 rows and 1 columns. This is a decennial time-series, from 1790 to 2000.

**Usage**

```
USPop
```

**Format**

This data frame contains the following columns:

**year** census year.

**population** Population in millions.

**Source**

U.S.-Census Bureau: <http://www.census-charts.com/Population/pop-us-1790-2000.html>, downloaded 1 May 2008.

**References**

Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models*, Second Edition. Sage.

---

Vocab

*Vocabulary and Education*

---

**Description**

The Vocab data frame has 30,351 rows and 4 columns. The observations are respondents to U.S. General Social Surveys, 1972-2016.

**Usage**

Vocab

**Format**

This data frame contains the following columns:

**year** Year of the survey.

**sex** Sex of the respondent, Female or Male.

**education** Education, in years.

**vocabulary** Vocabulary test score: number correct on a 10-word test.

**Source**

National Opinion Research Center *General Social Survey*. GSS Cumulative Datafile 1972-2016, downloaded from <http://gss.norc.org/>.

**References**

Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models*, Second Edition. Sage.

Fox, J. and Weisberg, S. (2011) *An R Companion to Applied Regression*, Second Edition, Sage.

---

WeightLoss

*Weight Loss Data*

---

**Description**

Contrived data on weight loss and self esteem over three months, for three groups of individuals: Control, Diet and Diet + Exercise. The data constitute a double-multivariate design.

**Usage**

WeightLoss

**Format**

A data frame with 34 observations on the following 7 variables.

group a factor with levels Control Diet DietEx.

w11 Weight loss at 1 month

w12 Weight loss at 2 months

w13 Weight loss at 3 months

se1 Self esteem at 1 month

se2 Self esteem at 2 months

se3 Self esteem at 3 months

**Details**

Helmert contrasts are assigned to group, comparing Control vs. (Diet DietEx) and Diet vs. DietEx.

**Source**

Originally taken from <http://www.csun.edu/~ata20315/psy524/main.htm>, but modified slightly. Courtesy of Michael Friendly.

---

Wells

*Well Switching in Bangladesh*

---

**Description**

Data on whether or not households in Bangladesh changed the wells that they were using.

**Usage**

Wells

**Format**

A data frame with 3020 observations on the following 5 variables.

switch whether or not the household switched to another well from an unsafe well: no or yes.

arsenic the level of arsenic contamination in the household's original well, in hundreds of micrograms per liter; all are above 0.5, which was the level identified as "safe".

distance in meters to the closest known safe well.

education in years of the head of the household.

association whether or not any members of the household participated in any community organizations: no or yes.

**Details**

The data are for an area of Arahazar upazila, Bangladesh. The researchers labelled each well with its level of arsenic and an indication of whether the well was "safe" or "unsafe." Those using unsafe wells were encouraged to switch. After several years, it was determined whether each household using an unsafe well had changed its well. These data are used by Gelman and Hill (2007) for a logistic-regression example.

**Source**

<http://www.stat.columbia.edu/~gelman/arm/examples/arsenic/wells.dat>.

**References**

A. Gelman and J. Hill (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

**Examples**

```
summary(Wells)
```

---

Women1f

*Canadian Women's Labour-Force Participation*

---

**Description**

The Women1f data frame has 263 rows and 4 columns. The data are from a 1977 survey of the Canadian population.

**Usage**

```
Women1f
```

**Format**

This data frame contains the following columns:

**partic** Labour-Force Participation. A factor with levels (note: out of order): fulltime, Working full-time; not.work, Not working outside the home; parttime, Working part-time.

**hincome** Husband's income, \$1000s.

**children** Presence of children in the household. A factor with levels: absent, present.

**region** A factor with levels: Atlantic, Atlantic Canada; BC, British Columbia; Ontario; Prairie, Prairie provinces; Quebec.

**Source**

*Social Change in Canada Project*. York Institute for Social Research.

**References**

Fox, J. (2008) *Applied Regression Analysis and Generalized Linear Models*, Second Edition. Sage.

Fox, J. and Weisberg, S. (2011) *An R Companion to Applied Regression*, Second Edition, Sage.

---

Wong

*Post-Coma Recovery of IQ*

---

**Description**

The Wong data frame has 331 row and 7 columns. The observations are longitudinal data on recovery of IQ after comas of varying duration for 200 subjects.

**Usage**

Wong

**Format**

This data frame contains the following columns:

**id** patient ID number.

**days** number of days post coma at which IQs were measured.

**duration** duration of the coma in days.

**sex** a factor with levels Female and Male.

**age** in years at the time of injury.

**piq** performance (i.e., mathematical) IQ.

**viq** verbal IQ.

**Details**

The data are from Wong, Monette, and Weiner (2001) and are for 200 patients who sustained traumatic brain injuries resulting in comas of varying duration. After awakening from their comas, patients were periodically administered a standard IQ test, but the average number of measurements per patient is small ( $331/200 = 1.7$ ).

**Source**

Wong, P. P., Monette, G., and Weiner, N. I. (2001) Mathematical models of cognitive recovery. *Brain Injury*, **15**, 519–530.

**References**

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

**Examples**

```
summary(Wong)
```

---

Wool

*Wool data*

---

**Description**

This is a three-factor experiment with each factor at three levels, for a total of 27 runs. Samples of worsted yarn were with different levels of the three factors were given a cyclic load until the sample failed. The goal is to understand how cycles to failure depends on the factors.

**Usage**

```
Wool
```

**Format**

This data frame contains the following columns:

**len** length of specimen (250, 300, 350 mm)

**amp** amplitude of loading cycle (8, 9, 10 min)

**load** load (40, 45, 50g)

**cycles** number of cycles until failure

**Source**

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations (with discussion). *J. Royal Statist. Soc.*, B26, 211–46.

**References**

- Fox, J. and Weisberg, S. (2011) *An R Companion to Applied Regression*, Second Edition, Sage.  
 Weisberg, S. (2014) *Applied Linear Regression*, Fourth Edition, Wiley, Section 6.3.

---

 WVS

*World Values Surveys*


---

**Description**

Data from the World Values Surveys 1995-1997 for Australia, Norway, Sweden, and the United States.

**Usage**

WVS

**Format**

A data frame with 5381 observations on the following 6 variables.

poverty “Do you think that what the government is doing for people in poverty in this country is about the right amount, too much, or too little?” (ordered): Too Little, About Right, Too Much.

religion Member of a religion: no or yes.

degree Held a university degree: no or yes.

country Australia, Norway, Sweden, or USA.

age in years.

gender male or female.

**References**

- J. Fox and R. Andersen (2006) Effect displays for multinomial and proportional-odds logit models. *Sociological Methodology* **36**, 225–255.

**Examples**

summary(WVS)

# Index

## \*Topic **datasets**

Adler, 3  
AMSSurvey, 4  
Angell, 5  
Anscombe, 5  
Arrests, 6  
Baumann, 7  
BEPS, 8  
Bfox, 9  
Blackmore, 10  
Burt, 10  
CanPop, 11  
CES11, 12  
Chile, 13  
Chiot, 14  
Cowles, 14  
Davis, 15  
DavisThin, 16  
Depredations, 17  
Duncan, 17  
Ericksen, 18  
Florida, 19  
Freedman, 20  
Friendly, 20  
Ginzberg, 21  
Greene, 22  
GSSvocab, 23  
Guyer, 24  
Hartnagel, 24  
Highway1, 25  
KosteckiDillon, 26  
Leinhardt, 27  
LoBD, 28  
Mandel, 30  
Migration, 30  
Moore, 31  
MplsDemo, 32  
MplsStops, 33  
Mroz, 34  
OBrienKaiser, 35  
Ornstein, 36  
Pottery, 37  
Prestige, 37  
Quartet, 38  
Robey, 39  
Rossi, 39  
Sahlins, 42  
Salaries, 42  
SLID, 43  
Soils, 44  
States, 45  
TitanicSurvival, 46  
Transact, 47  
UN, 47  
UN98, 48  
USPop, 49  
Vocab, 50  
WeightLoss, 51  
Wells, 51  
Womenlf, 52  
Wong, 53  
Wool, 54  
WVS, 55

Adler, 3  
AMSSurvey, 4  
Angell, 5  
Anscombe, 5  
Arrests, 6  
Baumann, 7  
bcnPower, 29  
BEPS, 8  
Bfox, 9  
Blackmore, 10  
Burt, 10  
CanPop, 11  
CES11, 12



Chile, 13  
Chirot, 14  
Cowles, 14

Davis, 15  
DavisThin, 16  
Depredations, 17  
Duncan, 17

Ericksen, 18

Florida, 19  
Freedman, 20  
Friendly, 20

Ginzberg, 21  
Greene, 22  
GSSvocab, 23  
Guyer, 24

Hartnagel, 24  
Highway1, 25

KosteckiDillon, 26

Leinhardt, 27  
LoBD, 28

Mandel, 30  
Migration, 30  
Moore, 31  
MplsDemo, 32, 34  
MplsStops, 32, 33  
Mroz, 34

OBrienKaiser, 35  
Ornstein, 36

Pottery, 37  
Prestige, 37

Quartet, 38

Robey, 39  
Rossi, 39

Sahlins, 42  
Salaries, 42  
SLID, 43  
Soils, 44  
States, 45

TitanicSurvival, 46  
Transact, 47

UN, 47  
UN98, 48  
USPop, 49

Vocab, 50

WeightLoss, 51  
Wells, 51  
Womenlf, 52  
Wong, 53  
Wool, 54  
WVS, 55