Package 'ISLR'

February 19, 2015

Type Package

Title Data for An Introduction to Statistical Learning with Applications in R

Version 1.0

Date 2013-06-10

Author Gareth James, Daniela Witten, Trevor Hastie and Rob Tibshirani

Maintainer Trevor Hastie <hastie@stanford.edu>

Suggests MASS

Description The collection of datasets used in the book ``An Introduction to Statistical Learning with Applications in R"

License GPL-2

LazyLoad yes

LazyData yes

URL http://www.StatLearning.com

Depends R (>= 2.10)

NeedsCompilation no

Repository CRAN

Date/Publication 2013-06-11 00:17:23

R topics documented:

`Auto` *Auto Data Set*

Description

Gas mileage, horsepower, and other information for 392 vehicles.

Usage

```
Auto
```

Format

A data frame with 392 observations on the following 9 variables.

`mpg` miles per gallon

`cylinders` Number of cylinders between 4 and 8

`displacement` Engine displacement (cu. inches)

`horsepower` Engine horsepower

`weight` Vehicle weight (lbs.)

`acceleration` Time to accelerate from 0 to 60 mph (sec.)

`year` Model year (modulo 100)

`origin` Origin of car (1. American, 2. European, 3. Japanese)

`name` Vehicle name

The orginal data contained 408 observations but 16 observations with missing values were removed.

Source

This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. The dataset was used in the 1983 American Statistical Association Exposition.

References

Games, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) *An Introduction to Statistical Learning with applications in R*, www.StatLearning.com, Springer-Verlag, New York

Examples

```
pairs(Auto)
attach(Auto)
hist(mpg)
```

`Caravan` *The Insurance Company (TIC) Benchmark*

Description

The data contains 5822 real customer records. Each record consists of 86 variables, containing sociodemographic data (variables 1-43) and product ownership (variables 44-86). The sociodemographic data is derived from zip codes. All customers living in areas with the same zip code have the same sociodemographic attributes. Variable 86 (`Purchase`) indicates whether the customer purchased a caravan insurance policy. Further information on the individual variables can be obtained at http://www.liacs.nl/~putten/library/cc2000/data.html

Usage

```
Caravan
```

Format

A data frame with 5822 observations on 86 variables.

Source

The data was originally supplied by Sentient Machine Research and was used in the CoIL Challenge 2000.

References

P. van der Putten and M. van Someren (eds) . CoIL Challenge 2000: The Insurance Company Case. Published by Sentient Machine Research, Amsterdam. Also a Leiden Institute of Advanced Computer Science Technical Report 2000-09. June 22, 2000. See http://www.liacs.nl/~putten/library/cc2000/
P. van der Putten and M. van Someren. A Bias-Variance Analysis of a Real World Learning Problem: The CoIL Challenge 2000. Machine Learning, October 2004, vol. 57, iss. 1-2, pp. 177-195, Kluwer Academic Publishers
Games, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) *An Introduction to Statistical Learning with applications in R*, www.StatLearning.com, Springer-Verlag, New York

Examples

```
summary(Caravan)
plot(Caravan$Purchase)
```

Carseats *Sales of Child Car Seats*

Description

A simulated data set containing sales of child car seats at 400 different stores.

Usage

```
Carseats
```

Format

A data frame with 400 observations on the following 11 variables.

`Sales` Unit sales (in thousands) at each location

`CompPrice` Price charged by competitor at each location

`Income` Community income level (in thousands of dollars)

`Advertising` Local advertising budget for company at each location (in thousands of dollars)

`Population` Population size in region (in thousands)

`Price` Price company charges for car seats at each site

`ShelveLoc` A factor with levels `Bad`, `Good` and `Medium` indicating the quality of the shelving location for the car seats at each site

`Age` Average age of the local population

`Education` Education level at each location

`Urban` A factor with levels `No` and `Yes` to indicate whether the store is in an urban or rural location

`US` A factor with levels `No` and `Yes` to indicate whether the store is in the US or not

Source

Simulated data

References

Games, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) *An Introduction to Statistical Learning with applications in R*, www.StatLearning.com, Springer-Verlag, New York

Examples

```
summary(Carseats)
lm.fit=lm(Sales~Advertising+Price,data=Carseats)
```

College *U.S. News and World Report's College Data*

Description

Statistics for a large number of US Colleges from the 1995 issue of US News and World Report.

Usage

```
College
```

Format

A data frame with 777 observations on the following 18 variables.

`Private` A factor with levels `No` and `Yes` indicating private or public university

`Apps` Number of applications received

`Accept` Number of applications accepted

`Enroll` Number of new students enrolled

`Top10perc` Pct. new students from top 10% of H.S. class

`Top25perc` Pct. new students from top 25% of H.S. class

`F.Undergrad` Number of fulltime undergraduates

`P.Undergrad` Number of parttime undergraduates

`Outstate` Out-of-state tuition

`Room.Board` Room and board costs

`Books` Estimated book costs

`Personal` Estimated personal spending

`PhD` Pct. of faculty with Ph.D.'s

`Terminal` Pct. of faculty with terminal degree

`S.F.Ratio` Student/faculty ratio

`perc.alumni` Pct. alumni who donate

`Expend` Instructional expenditure per student

`Grad.Rate` Graduation rate

Source

This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. The dataset was used in the ASA Statistical Graphics Section's 1995 Data Analysis Exposition.

References

Games, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) *An Introduction to Statistical Learning with applications in R*, www.StatLearning.com, Springer-Verlag, New York

Examples

```
summary(College)
lm(Apps~Private+Accept,data=College)
```

`Default` *Credit Card Default Data*

Description

A simulated data set containing information on ten thousand customers. The aim here is to predict which customers will default on their credit card debt.

Usage

```
Default
```

Format

A data frame with 10000 observations on the following 4 variables.

`default` A factor with levels `No` and `Yes` indicating whether the customer defaulted on their debt

`student` A factor with levels `No` and `Yes` indicating whether the customer is a student

`balance` The average balance that the customer has remaining on their credit card after making their monthly payment

`income` Income of customer

Source

Simulated data

References

Games, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) *An Introduction to Statistical Learning with applications in R*, www.StatLearning.com, Springer-Verlag, New York

Examples

```
summary(Default)
glm(default~student+balance+income,family="binomial",data=Default)
```

Hitters *Baseball Data*

Description

Major League Baseball Data from the 1986 and 1987 seasons.

Usage

```
Hitters
```

Format

A data frame with 322 observations of major league players on the following 20 variables.

`AtBat` Number of times at bat in 1986

`Hits` Number of hits in 1986

`HmRun` Number of home runs in 1986

`Runs` Number of runs in 1986

`RBI` Number of runs batted in in 1986

`Walks` Number of walks in 1986

`Years` Number of years in the major leagues

`CAtBat` Number of times at bat during his career

`CHits` Number of hits during his career

`CHmRun` Number of home runs during his career

`CRuns` Number of runs during his career

`CRBI` Number of runs batted in during his career

`CWalks` Number of walks during his career

`League` A factor with levels `A` and `N` indicating player's league at the end of 1986

`Division` A factor with levels `E` and `W` indicating player's division at the end of 1986

`PutOuts` Number of put outs in 1986

`Assists` Number of assists in 1986

`Errors` Number of errors in 1986

`Salary` 1987 annual salary on opening day in thousands of dollars

`NewLeague` A factor with levels `A` and `N` indicating player's league at the beginning of 1987

Source

This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. This is part of the data that was used in the 1988 ASA Graphics Section Poster Session. The salary data were originally from Sports Illustrated, April 20, 1987. The 1986 and career statistics were obtained from The 1987 Baseball Encyclopedia Update published by Collier Books, Macmillan Publishing Company, New York.

References

Games, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) *An Introduction to Statistical Learning with applications in R*, www.StatLearning.com, Springer-Verlag, New York

Examples

```
summary(Hitters)
lm(Salary~AtBat+Hits,data=Hitters)
```

Khan *Khan Gene Data*

Description

The data consists of a number of tissue samples corresponding to four distinct types of small round blue cell tumors. For each tissue sample, 2308 gene expression measurements are available.

Usage

```
Khan
```

Format

The format is a list containing four components: `xtrain`, `xtest`, `ytrain`, and `ytest`. `xtrain` contains the 2308 gene expression values for 63 subjects and `ytrain` records the corresponding tumor type. `ytrain` and `ytest` contain the corresponding testing sample information for a further 20 subjects.

Source

This data were originally reported in:

Khan J, Wei J, Ringner M, Saal L, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu C, Peterson C, and Meltzer P. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine, v.7, pp.673-679, 2001.

The data were also used in:

Tibshirani RJ, Hastie T, Narasimhan B, and G. Chu. Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression. Proceedings of the National Academy of Sciences of the United States of America, v.99(10), pp.6567-6572, May 14, 2002.

References

Games, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) *An Introduction to Statistical Learning with applications in R*, www.StatLearning.com, Springer-Verlag, New York

Examples

```
table(Khan$ytrain)
table(Khan$ytest)
```

NCI60 *NCI 60 Data*

Description

NCI microarray data. The data contains expression levels on 6830 genes from 64 cancer cell lines. Cancer type is also recorded.

Usage

```
NCI60
```

Format

The format is a list containing two elements: `data` and `labs`.

`data` is a 64 by 6830 matrix of the expression values while `labs` is a vector listing the cancer types for the 64 cell lines.

Source

The data come from Ross et al. (Nat Genet., 2000). More information can be obtained at http://genome-www.stanford.edu/nci60/

References

Games, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) *An Introduction to Statistical Learning with applications in R*, www.StatLearning.com, Springer-Verlag, New York

Examples

```
table(NCI60$labs)
```

OJ *Orange Juice Data*

Description

The data contains 1070 purchases where the customer either purchased Citrus Hill or Minute Maid Orange Juice. A number of characteristics of the customer and product are recorded.

Usage

```
OJ
```

Format

A data frame with 1070 observations on the following 18 variables.

`Purchase` A factor with levels `CH` and `MM` indicating whether the customer purchased Citrus Hill or Minute Maid Orange Juice

`WeekofPurchase` Week of purchase

`StoreID` Store ID

`PriceCH` Price charged for CH

`PriceMM` Price charged for MM

`DiscCH` Discount offered for CH

`DiscMM` Discount offered for MM

`SpecialCH` Indicator of special on CH

`SpecialMM` Indicator of special on MM

`LoyalCH` Customer brand loyalty for CH

`SalePriceMM` Sale price for MM

`SalePriceCH` Sale price for CH

`PriceDiff` Sale price of MM less sale price of CH

`Store7` A factor with levels `No` and `Yes` indicating whether the sale is at Store 7

`PctDiscMM` Percentage discount for MM

`PctDiscCH` Percentage discount for CH

`ListPriceDiff` List price of MM less list price of CH

`STORE` Which of 5 possible stores the sale occured at

Source

Stine, Robert A., Foster, Dean P., Waterman, Richard P. Business Analysis Using Regression (1998). Published by Springer.

References

Games, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) *An Introduction to Statistical Learning with applications in R*, www.StatLearning.com, Springer-Verlag, New York

Examples

```
summary(OJ)
plot(OJ$Purchase,OJ$PriceCH)
```

Portfolio *Portfolio Data*

Description

A simple simulated data set containing 100 returns for each of two assets, X and Y. The data is used to estimate the optimal fraction to invest in each asset to minimize investment risk of the combined portfolio. One can then use the Bootstrap to estimate the standard error of this estimate.

Usage

```
Portfolio
```

Format

A data frame with 100 observations on the following 2 variables.

`X` Returns for Asset X

`Y` Returns for Asset Y

Source

Simulated data

References

Games, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) *An Introduction to Statistical Learning with applications in R*, www.StatLearning.com, Springer-Verlag, New York

Examples

```
summary(Portfolio)
attach(Portfolio)
plot(X,Y)
```

Smarket *S&P Stock Market Data*

Description

Daily percentage returns for the S&P 500 stock index between 2001 and 2005.

Usage

```
Smarket
```

Format

A data frame with 1250 observations on the following 9 variables.

`Year` The year that the observation was recorded

`Lag1` Percentage return for previous day

`Lag2` Percentage return for 2 days previous

`Lag3` Percentage return for 3 days previous

`Lag4` Percentage return for 4 days previous

`Lag5` Percentage return for 5 days previous

`Volume` Volume of shares traded (number of daily shares traded in billions)

`Today` Percentage return for today

`Direction` A factor with levels `Down` and `Up` indicating whether the market had a positive or negative return on a given day

Source

Raw values of the S&P 500 were obtained from Yahoo Finance and then converted to percentages and lagged.

References

Games, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) *An Introduction to Statistical Learning with applications in R*, www.StatLearning.com, Springer-Verlag, New York

Examples

```
summary(Smarket)
lm(Today~Lag1+Lag2,data=Smarket)
```

Wage *Mid-Atlantic Wage Data*

Description

Wage and other data for a group of 3000 workers in the Mid-Atlantic region.

Usage

```
Wage
```

Format

A data frame with 3000 observations on the following 12 variables.

`year` Year that wage information was recorded

`age` Age of worker

`sex` Gender

`maritl` A factor with levels `1. Never Married` `2. Married` `3. Widowed` `4. Divorced` and `5. Separated` indicating marital status

`race` A factor with levels `1. White` `2. Black` `3. Asian` and `4. Other` indicating race

`education` A factor with levels `1. < HS Grad` `2. HS Grad` `3. Some College` `4. College Grad` and `5. Advanced Degree` indicating education level

`region` Region of the country (mid-atlantic only)

`jobclass` A factor with levels `1. Industrial` and `2. Information` indicating type of job

`health` A factor with levels `1. <=Good` and `2. >=Very Good` indicating health level of worker

`health_ins` A factor with levels `1. Yes` and `2. No` indicating whether worker has health insurance

`logwage` Log of workers wage

`wage` Workers raw wage

Source

Data was manually assembled by Steve Miller, of Open BI (www.openbi.com), from the March 2011 Supplement to Current Population Survey data.

http://thedataweb.rm.census.gov/TheDataWeb

References

Games, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) *An Introduction to Statistical Learning with applications in R*, www.StatLearning.com, Springer-Verlag, New York

Examples

```
summary(Wage)
lm(wage~year+age,data=Wage)
## maybe str(Wage) ; plot(Wage) ...
```

Weekly *Weekly S&P Stock Market Data*

Description

Weekly percentage returns for the S&P 500 stock index between 1990 and 2010.

Usage

```
Weekly
```

Format

A data frame with 1089 observations on the following 9 variables.

`Year` The year that the observation was recorded

`Lag1` Percentage return for previous week

`Lag2` Percentage return for 2 weeks previous

`Lag3` Percentage return for 3 weeks previous

`Lag4` Percentage return for 4 weeks previous

`Lag5` Percentage return for 5 weeks previous

`Volume` Volume of shares traded (average number of daily shares traded in billions)

`Today` Percentage return for this week

`Direction` A factor with levels `Down` and `Up` indicating whether the market had a positive or negative return on a given week

Source

Raw values of the S&P 500 were obtained from Yahoo Finance and then converted to percentages and lagged.

References

Games, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) *An Introduction to Statistical Learning with applications in R*, www.StatLearning.com, Springer-Verlag, New York

Examples

```
summary(Weekly)
lm(Today~Lag1+Lag2,data=Weekly)
```

Index