

# Package ‘mpae’

May 9, 2026

**Type** Package

**Title** Metodos Predictivos de Aprendizaje Estadistico (Statistical Learning Predictive Methods)

**Version** 0.1.2

**Date** 2024-03-19

**Maintainer** Ruben Fernandez-Casal <rubenfcasal@gmail.com>

**Depends** R (>= 3.5.0), graphics

**Imports** caret, RcmdrMisc

**Suggests** car, gbm, leaps, lmtest, glmnet, mgcv, np, NeuralNetTools, pdp, vivid, plot3D, AppliedPredictiveModeling, ISLR

**Description** Functions and datasets used in the book: Fernandez-Casal, R., Costa, J. and Oviedo-de la Fuente, M. (2024) ``Metodos predictivos de aprendizaje estadistico" <[https://rubenfcasal.github.io/aprendizaje\\_estadistico/](https://rubenfcasal.github.io/aprendizaje_estadistico/)>.

**License** GPL (>= 2)

**URL** <https://github.com/rubenfcasal/mpae>

**BugReports** <https://github.com/rubenfcasal/mpae/issues>

**Encoding** UTF-8

**LazyData** true

**RoxygenNote** 7.2.3

**NeedsCompilation** no

**Author** Ruben Fernandez-Casal [aut, cre] (ORCID:  
<<https://orcid.org/0000-0002-5785-3739>>),  
Manuel Oviedo-de la Fuente [aut] (ORCID:  
<<https://orcid.org/0000-0001-7360-3249>>),  
Julian Costa-Bouzas [aut] (ORCID:  
<<https://orcid.org/0000-0001-9760-9581>>)

**Repository** CRAN

**Date/Publication** 2024-03-21 14:40:02 UTC

## Contents

mpae-package . . . . .	2
accuracy . . . . .	3
bfan . . . . .	4
bodyfat . . . . .	5
bodyfat.raw . . . . .	7
hbat . . . . .	9
pred.plot . . . . .	11
scaled.coef . . . . .	13
winequality . . . . .	14
winetaste . . . . .	15
<b>Index</b>	<b>17</b>

---

mpae-package	<i>mpae: Métodos predictivos de aprendizaje estadístico (statistical learning predictive methods)</i>
--------------	---

---

## Description

Functions and datasets used in the book Fernández-Casal, Costa and Oviedo-de la Fuente (2024) *Métodos predictivos de aprendizaje estadístico*.

## Details

For more information visit <https://rubenfcasal.github.io/mpae/>.

## References

- Fernández-Casal R., Costa J. and Oviedo-de la Fuente M. (2024). *Métodos predictivos de aprendizaje estadístico* (github).
- Fernández-Casal R., Roca-Pardiñas J., Costa J. and Oviedo-de la Fuente M. (2022). *Introducción al Análisis de Datos con R* (github).
- Fernández-Casal R., Cao R. and Costa J. (2023). *Técnicas de Simulación y Remuestreo*, segunda edición, (github).

---

accuracy	<i>Accuracy measures</i>
----------	--------------------------

---

**Description**

Computes accuracy measurements.

**Usage**

```
accuracy(pred, obs, na.rm = FALSE, tol = sqrt(.Machine$double.eps))
```

**Arguments**

pred	a numeric vector with the predicted values.
obs	a numeric vector with the observed values.
na.rm	a logical indicating whether <b>NA</b> values should be stripped before the computation proceeds.
tol	divide underflow tolerance.

**Value**

Returns a named vector with the following components:

- me mean error
- rmse root mean squared error
- mae mean absolute error
- mpe mean percent error
- mape mean absolute percent error
- r.squared pseudo R-squared

**See Also**

[pred.plot\(\)](#)

**Examples**

```
set.seed(1)
nobs <- nrow(bodyfat)
itrain <- sample(nobs, 0.8 * nobs)
train <- bodyfat[itrain, ]
test <- bodyfat[-itrain, ]
fit <- lm(bodyfat ~ abdomen + wrist, data = train)
pred <- predict(fit, newdata = test)
obs <- test$bodyfat
pred.plot(pred, obs)
accuracy(pred, obs)
```

---

bfan

*Above normal body fat data*

---

### Description

Modification of the [bodyfat](#) dataset for classification. The response bfan is a factor indicating a body fat value above the normal range. The variable bodyfat was dropped for convenience, and two new variables bmi (body mass index, in  $\text{kg}/\text{m}^2$ ) and bmi2 (alternate body mass index, in  $\text{kg}^{1.2}/\text{m}^{3.3}$ ) were computed (see examples below).

### Usage

bfan

### Format

A data frame with 246 rows and 16 columns:

**bfan** Body fat above normal range

**age** Age (years)

**weight** Weight (kg)

**height** Height (cm)

**neck** Neck circumference (cm)

**chest** Chest circumference (cm)

**abdomen** Abdomen circumference (cm)

**hip** Hip circumference (cm)

**thigh** Thigh circumference (cm)

**knee** Knee circumference (cm)

**ankle** Ankle circumference (cm)

**biceps** Biceps (extended) circumference (cm)

**forearm** Forearm circumference (cm)

**wrist** Wrist circumference (cm)

**bmi** Body mass index ( $\text{kg}/\text{m}^2$ )

**bmi2** Alternate body mass index

### Details

See [bodyfat](#) and [bodyfat.raw](#) for details.

### Source

StatLib Datasets Archive: <https://lib.stat.cmu.edu/datasets/bodyfat>.

## References

Penrose, K., Nelson, A. and Fisher, A. (1985). Generalized Body Composition Prediction Equation for Men Using Simple Measurement Techniques. *Medicine and Science in Sports and Exercise*, 17(2), 189. doi:10.1249/0000576819850400000037.

## See Also

[bodyfat](#), [bodyfat.raw](#)

## Examples

```
bfan <- bodyfat
# Body fat above normal
bfan[1] <- factor(bfan$bodyfat > 24 , # levels = c('FALSE', 'TRUE'),
                 labels = c('No', 'Yes'))
names(bfan)[1] <- "bfan"
bfan$bmi <- with(bfan, weight/(height/100)^2)
bfan$bmi2 <- with(bfan, weight^1.2/(height/100)^3.3)

fit <- glm(bfan ~ abdomen, family = binomial, data = bfan)
summary(fit)
```

---

bodyfat

*Body fat data*

---

## Description

Modification of the dataset analysed in Penrose et al. (1985). Lists estimates of the percentage of body fat determined by underwater weighing and various body measurements for 246 men.

## Usage

bodyfat

## Format

A data frame with 246 rows and 14 columns:

**bodyfat** Percent body fat (from Siri's 1956 equation)

**age** Age (years)

**weight** Weight (kg)

**height** Height (cm)

**neck** Neck circumference (cm)

**chest** Chest circumference (cm)

**abdomen** Abdomen circumference (cm)

**hip** Hip circumference (cm)

**thigh** Thigh circumference (cm)  
**knee** Knee circumference (cm)  
**ankle** Ankle circumference (cm)  
**biceps** Biceps (extended) circumference (cm)  
**forearm** Forearm circumference (cm)  
**wrist** Wrist circumference (cm)

### Details

This data set can be used to illustrate multiple regression techniques (e.g. Johnson 1996). Instead of estimating body fat percentage from body density, which is not easy to measure, it is desirable to have a simpler method that allow this to be done from body measurements.

[bodyfat.raw](#) contains the original data. According to Johnson (1996), there were data entry errors (cases 42, 48, 76, 96 and 182 of the original data) and he suggested some rules to correct them. These outliers were removed in the `bodyfat` dataset, as well as an influential observation (case 39, which has a big effect on regression estimates). Additionally, the variable `density` was dropped for convenience, and variables `height` and `weight` were transformed into metric units (centimetres and kilograms) for consistency.

See [bodyfat.raw](#) for more details.

### Source

StatLib Datasets Archive: <https://lib.stat.cmu.edu/datasets/bodyfat>.

### References

Johnson, R. W. (1996). Fitting Percentage of Body Fat to Simple Body Measurements. *Journal of Statistics Education*, 4(1). doi:10.1080/10691898.1996.11910505.

Penrose, K., Nelson, A. and Fisher, A. (1985). Generalized Body Composition Prediction Equation for Men Using Simple Measurement Techniques. *Medicine and Science in Sports and Exercise*, 17(2), 189. doi:10.1249/0000576819850400000037.

### See Also

[bodyfat.raw](#), [bfan](#)

### Examples

```
fit <- lm(bodyfat ~ abdomen, bodyfat)
summary(fit)
plot(bodyfat ~ abdomen, bodyfat)
abline(fit)
```

---

bodyfat.raw	<i>Original body fat data</i>
-------------	-------------------------------

---

### Description

Popular dataset originally analysed in Penrose et al. (1985). Lists estimates of the percentage of body fat determined by underwater weighing and various body measurements for 252 men.

### Usage

bodyfat.raw

### Format

A data frame with 252 rows and 15 columns:

**density** Density (gm/cm<sup>3</sup>; determined from underwater weighing)

**bodyfat** Percent body fat (from Siri's 1956 equation)

**age** Age (years)

**weight** Weight (lbs)

**height** Height (inches)

**neck** Neck circumference (cm)

**chest** Chest circumference (cm)

**abdomen** Abdomen 2 circumference (cm)

**hip** Hip circumference (cm)

**thigh** Thigh circumference (cm)

**knee** Knee circumference (cm)

**ankle** Ankle circumference (cm)

**biceps** Biceps (extended) circumference (cm)

**forearm** Forearm circumference (cm)

**wrist** Wrist circumference (cm)

### Details

This data set can be used to illustrate data cleaning and multiple regression techniques (e.g. Johnson 1996). Percentage of body fat for an individual can be estimated from body density, for instance by using Siri's (1956) equation:

$$bodyfat = 495/density - 450.$$

Volume, and hence body density, can be accurately measured by underwater weighing (e.g. Katch and McArdle, 1977). However, this procedure for the accurate measurement of body fat is inconvenient and costly. It is desirable to have easy methods of estimating body fat from body measurements.

"Measurement standards are apparently those listed in Benhke and Wilmore (1974), pp. 45-48 where, for instance, the abdomen 2 circumference is measured 'laterally, at the level of the iliac crests, and anteriorly, at the umbilicus'.

Johnson (1996) uses the original data in an activity to introduce students to data cleaning before performing multiple linear regression. An examination of the data reveals some unusual cases:

- Cases 48, 76, and 96 seem to have a one-digit error in the listed density values.
- Case 42 appears to have a one-digit error in the height value.
- Case 182 appears to have an error in the density value (as it is greater than 1.1, the density of the "fat free mass"; resulting in a negative estimate of body fat percentage that was truncated to zero).

Johnson (1996) suggests some rules for correcting these values (see examples below).

### Source

StatLib Datasets Archive: <https://lib.stat.cmu.edu/datasets/bodyfat>.

### References

Johnson, R. W. (1996). Fitting Percentage of Body Fat to Simple Body Measurements. *Journal of Statistics Education*, 4(1). doi:10.1080/10691898.1996.11910505.

Penrose, K., Nelson, A. and Fisher, A. (1985). Generalized Body Composition Prediction Equation for Men Using Simple Measurement Techniques. *Medicine and Science in Sports and Exercise*, 17(2), 189. doi:10.1249/0000576819850400000037.

Siri, W. E. (1956). Gross Composition of the Body, in *Advances in Biological and Medical Physics* (Vol. IV), eds. J. H. Lawrence and C. A. Tobias, Academic Press.

### See Also

[bodyfat](#), [bfan](#)

### Examples

```
bodyfat <- bodyfat.raw
# Johnson's (1996) corrections
cases <- c(48, 76, 96) # bodyfat != 495/density - 450
bodyfat$density[cases] <- 495 / (bodyfat$bodyfat[cases] + 450)
bodyfat$height[42] <- 69.5
# Other possible data entry errors
# See https://stat-ata-asu.github.io/PredictiveModelBuilding/BFdata.html
bodyfat$ankle[31] <- 23.9
bodyfat$ankle[86] <- 23.7
bodyfat$forearm[159] <- 24.9
# Outlier and influential observation
outliers <- c(182, 39)
bodyfat[outliers, ]
bodyfat <- bodyfat[-outliers, ]

# Body mass index (kg/m2)
```

```

bodyfat$bmi <- with(bodyfat, weight/(height*0.0254)^2)
# Alternate body mass index
bodyfat$bmi2 <- with(bodyfat, (weight*0.45359237)^1.2/(height*0.0254)^3.3)
# See e.g. https://en.wikipedia.org/wiki/Body_fat_percentage#From_BMI
# \text{(Adult) body fat percentage} = (1.39 \times \text{BMI})
#
# + (0.16 \times \text{age}) - (10.34 \times \text{gender}) - 9

```

hbat

*HBAT data*

## Description

A dataset containing observations of customers of the industrial distribution company HBAT. The variables can be classified into three groups: the first 6 (categorical) are shopper characteristics (data warehouse classification), variables 7 to 19 (numerical) measure shopper perceptions of HBAT and the last 5 are possible target variables (responses), the purchase outcomes.

## Usage

```
hbat
```

## Format

A data frame with 200 rows and 24 columns:

**empresa** Customer ID.

**cliente** Customer Type. Length of time a particular customer has been buying from HBAT: Menos de 1 año = less than 1 year. De 1 a 5 años = between 1 and 5 years. Más de 5 años = longer than 5 years.

**industri** Type of industry that purchases HBAT's paper products: Revista = magazine industry, Periodico = newsprint industry.

**tamaño** Employee size: Pequeña (<500) = small firm, fewer than 500 employees, Grande (>=500) = large firm, 500 or more employees.

**region** Customer location: America del norte = USA/North America, Otros = outside North America.

**distrib** Distribution System. How paper products are sold to customers: Indirecta = sold indirectly through a broker, Directa = sold directly.

**calidadp** Product Quality. Perceived level of quality of HBAT's paper products.

**web** E-Commerce Activities/Web Site. Overall image of HBAT's Web site, especially user-friendliness.

**soporte** Technical Support. Extent to which technical support is offered to help solve product/service issues.

**quejas** Complaint Resolution. Extent to which any complaints are resolved in a timely and complete manner.

**publi** Advertising. Perceptions of HBAT's advertising campaigns in all types of media.

**producto** Product Line. Depth and breadth of HBAT's product line to meet customer needs.

- imgfvent** Salesforce Image. Overall image of HBAT's salesforce.
- precio** Competitive Pricing. Extent to which HBAT offers competitive prices.
- garantia** Warranty and Claims. Extent to which HBAT stands behind its product/service warranties and claims.
- nprod** New Products. Extent to which HBAT develops and sells new products.
- facturac** Ordering and Billing. Perception that ordering and billing is handled efficiently and correctly.
- flexprec** Price Flexibility. Perceived willingness of HBAT sales reps to negotiate price on purchases of paper products.
- velocida** Delivery Speed. Amount of time it takes to deliver the paper products once an order has been confirmed.
- satisfac** Customer satisfaction with past purchases from HBAT, measured on a 10-point graphic rating scale.
- precomen** Likelihood of recommending HBAT to other firms as a supplier of paper products, measured on a 10-point graphic rating scale.
- pcompra** Likelihood of purchasing paper products from HBAT in the future, measured on a 10-point graphic rating scale.
- fidelida** Percentage of Purchases from HBAT. Percentage of the responding firm's paper needs purchased from HBAT, measured on a 100-point percentage scale.
- alianza** Perception of Future Relationship with HBAT. Extent to which the customer/respondent perceives his or her firm would engage in strategic alliance/partnership with HBAT: No = Would not consider. Si = Yes, would consider strategic alliance or partnership.

### Details

For more details, consult the reference Hair et al. (1998).

### Source

Hair et al. (1998).

### References

Hair, J. F., Anderson, R. E., Tatham, R. L., y Black, W. (1998). *Multivariate Data Analysis*. Prentice Hall.

### Examples

```
str(hbat)
as.data.frame(attr(hbat, "variable.labels"))
summary(hbat)
```

---

pred.plot	<i>Observed vs. predicted plots</i>
-----------	-------------------------------------

---

### Description

Generates plots comparing predictions with observations.

### Usage

```
pred.plot(pred, obs, ...)  
  
## Default S3 method:  
pred.plot(  
  pred,  
  obs,  
  xlab = "Predicted",  
  ylab = "Observed",  
  lm.fit = TRUE,  
  lowess = TRUE,  
  ...  
)  
  
## S3 method for class 'factor'  
pred.plot(  
  pred,  
  obs,  
  type = c("frec", "perc", "cperc"),  
  xlab = "Observed",  
  ylab = NULL,  
  legend.title = "Predicted",  
  label.bars = TRUE,  
  ...  
)
```

### Arguments

pred	a numeric vector with the predicted values.
obs	a numeric vector with the observed values.
...	additional graphical parameters or further arguments passed to other methods (e.g. to <code>RcmdrMisc::Barplot()</code> ).
xlab	a title for the x axis.
ylab	a title for the y axis.
lm.fit	logical indicating if a <code>lm</code> fit is added to the plot.
lowess	logical indicating if a <code>lowess</code> smooth is added to the plot.

type	types of the desired plots. Any combination of the following values is possible: "frec" for frequencies, "perc" for percentages or "cperc" for conditional percentages.
legend.title	a title for the legend.
label.bars	if TRUE (the default) show values of frequencies or percents in the bars.

### Details

The default method draws a scatter plot of the observed values against the predicted values.

`pred.plot.factor()` creates bar plots representing frequencies, percentages or conditional percentages of `pred` within levels of `obs`. This method is a front end to `RcmdrMisc::Barplot()`.

### Value

The default method invisibly returns the fitted linear model if `lm.fit == TRUE`.

`pred.plot.factor()` invisibly returns the horizontal coordinates of the centers of the bars.

### See Also

[accuracy\(\)](#)

### Examples

```
set.seed(1)
nobs <- nrow(hbat)
itrain <- sample(nobs, 0.8 * nobs)
train <- hbat[itrain, ]
test <- hbat[-itrain, ]

# Regression
fit <- lm(fidelida ~ velocida + calidadp, data = train)
pred <- predict(fit, newdata = test)
obs <- test$fidelida
res <- pred.plot(pred, obs)
summary(res)

# Classification
fit2 <- glm(alianza ~ velocida + calidadp, family = binomial, data = train)
obs <- test$alianza
p.est <- predict(fit2, type = "response", newdata = test)
pred <- factor(p.est > 0.5, labels = levels(obs))
pred.plot(pred, obs, type = "frec", style = "parallel")
old.par <- par(mfrow = c(1, 2))
pred.plot(pred, obs, type = c("perc", "cperc"))
par(old.par)
```

---

scaled.coef	<i>Scaled (standardized) coefficients</i>
-------------	---

---

### Description

Computes the standardized (regression) coefficients, also called beta coefficients or beta weights, to quantify the importance (the effect) of the predictors on the dependent variable in a multiple regression analysis where the variables are measured in different units.

### Usage

```
scaled.coef(object, ...)  
  
## Default S3 method:  
scaled.coef(object, scale.response = TRUE, complete = FALSE, ...)
```

### Arguments

object	an object for which the extraction of model coefficients is meaningful.
...	further arguments passed to or from other methods.
scale.response	logical indicating if the response variable should be standardized.
complete	for the default (used for lm, etc) and aov methods: logical indicating if the full coefficient vector should be returned also in case of an over-determined system where some coefficients will be set to <a href="#">NA</a> .

### Details

The beta weights are the coefficient estimates resulting from a regression analysis where the underlying data have been standardized so that the variances of dependent and explanatory variables are equal to 1. Therefore, standardized coefficients are unitless and refer to how many standard deviations a dependent variable will change, per standard deviation increase in the predictor variable. See [https://en.wikipedia.org/wiki/Standardized\\_coefficient](https://en.wikipedia.org/wiki/Standardized_coefficient) or [QuantPsyc::lm.beta](#).

Based on [QuantPsyc::lm.beta](#).

### Value

A named vector with the scaled coefficients.

### See Also

[coef\(\)](#)

### Examples

```
fit <- lm(fidelida ~ velocida + calidadp, hbat)
coef(fit)
scaled.coef(fit)
fit2 <- lm(scale(fidelida) ~ scale(velocida) + scale(calidadp), hbat)
coef(fit2)
fit3 <- lm(fidelida ~ scale(velocida) + scale(calidadp), hbat)
coef(fit3)
scaled.coef(fit, scale.response = FALSE)
```

---

winequality

*Wine quality data*

---

### Description

A subset related to the white variant of the Portuguese "Vinho Verde" wine, containing physico-chemical information (fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates and alcohol) and sensory (quality).

### Usage

winequality

### Format

A data frame with 1,250 rows and 12 columns:

**fixed.acidity** fixed acidity

**volatile.acidity** volatile acidity

**citric.acid** citric acid

**residual.sugar** residual sugar

**chlorides** chlorides

**free.sulfur.dioxide** free sulfur dioxide

**total.sulfur.dioxide** total sulfur dioxide

**density** density

**pH** pH

**sulphates** sulphates

**alcohol** alcohol

**quality** median of at least 3 evaluations of wine quality carried out by experts, who evaluated them between 0 (very bad) and 10 (very excellent)

### Details

For more details, consult <https://www.vinhoverde.pt/en/> or the reference Cortez et al. (2009).

**Source**

UCI Machine Learning Repository: <https://archive.ics.uci.edu/dataset/186/wine+quality>.

**References**

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4), 547-553.

**See Also**

[winetaste](#)

**Examples**

```
str(winequality)
```

---

winetaste

*Wine taste data*

---

**Description**

A subset related to the white variant of the Portuguese "Vinho Verde" wine, containing physico-chemical information (fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates and alcohol) and sensory (taste), which indicates the quality of the wine (it is considered good if the median of the wine quality evaluations, made by experts, who evaluated them between 0 = very bad and 10 = very excellent, is not less than 6).

**Usage**

```
winetaste
```

**Format**

A data frame with 1,250 rows and 12 columns:

**fixed.acidity** fixed acidity

**volatile.acidity** volatile acidity

**citric.acid** citric acid

**residual.sugar** residual sugar

**chlorides** chlorides

**free.sulfur.dioxide** free sulfur dioxide

**total.sulfur.dioxide** total sulfur dioxide

**density** density

**pH** pH

**sulphates** sulphates

**alcohol** alcohol

**taste** factor with levels "good" and "bad" indicating the quality of the wine

**Details**

For more details, consult <https://www.vinhoverde.pt/en/> or the reference Cortez et al. (2009).

**Source**

UCI Machine Learning Repository: <https://archive.ics.uci.edu/dataset/186/wine+quality>.

**References**

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4), 547-553.

**See Also**

[winequality](#)

**Examples**

```
winetaste <- winequality[, names(winequality)!="quality"]
winetaste$taste <- factor(winequality$quality < 6,
                        labels = c('good', 'bad')) # levels = c('FALSE', 'TRUE')
str(winetaste)
```

# Index

- \* **Monte-Carlo**
  - mpae-package, 2
- \* **bootstrap**
  - mpae-package, 2
- \* **datasets**
  - bfan, 4
  - bodyfat, 5
  - bodyfat.raw, 7
  - hbat, 9
  - winequality, 14
  - winetaste, 15
- \* **simulation**
  - mpae-package, 2

accuracy, 3  
accuracy(), 12

bfan, 4, 6, 8  
bodyfat, 4, 5, 5, 8  
bodyfat.raw, 4–6, 7

coef(), 13

hbat, 9

lm, 11  
lowess, 11

mpae (mpae-package), 2  
mpae-package, 2

NA, 3, 13

pred.plot, 11  
pred.plot(), 3

RcmdrMisc::Barplot(), 11, 12

scaled.coef, 13

winequality, 14, 16  
winetaste, 15, 15