Package 'LogisticDx'

August 29, 2016

Type Package

Title Diagnostic Tests for Models with a Binomial Response

Version 0.2

Date 2015-07-01

Author Chris Dardis

Maintainer Chris Dardis <christopherdardis@gmail.com>

License GPL (>= 2)

Description Diagnostic tests and plots for GLMs (generalized linear models) with binomial/ binary outcomes, particularly logistic regression.

VignetteBuilder knitr

Depends R (>= 3.0.0)

Imports rms, stats, statmod, graphics, speedglm, RColorBrewer, data.table, pROC, aod

Suggests knitr

LazyLoad yes

Collate 'LogisticDx_package.R' 'OR.R' 'ageChd.R' 'bbdm.R' 'dx.R' 'genBinom.R' 'gof.R' 'icu.R' 'lbw.R' 'llbw.R' 'mes.R' 'mlbw.R' 'nhanes3.R' 'pcs.R' 'plotGlm.R' 'printGofGlm.R' 'printSSglm.R' 'printSigGlm.R' 'sig.R' 'ss.R' 'uis.R'

NeedsCompilation no

Repository CRAN

Date/Publication 2015-07-08 22:37:21

R topics documented:

`logisticDx2-package` *Diagnostic Tests for Models with a Binomial Response*

Description

Diagnostic Tests for Models with a Binomial Response

Details

Package:	LogisticDx
Type:	Package
Version:	0.2
Date:	2015-07-01
License:	GPL (>= 2)
LazyLoad:	yes

Diagnostic tests and plots for GLMs (generalized linear models) with binomial/ binary outcomes, particularly logistic regression.

The most commonly used functions are likely to be `dx` (diagnostics), `plot.glm` (diagnostic plots) and `gof` (goodness-of-fit tests).

There have been changes to many of the functions between Version 0.1 and 0.2 of this package.

The package should be regarded as 'in development' until release 1.0, meaning that there may be changes to certain function names and parameters, although I will try to keep this to a minimum.

There are references in many of the functions to the textbook:
Hosmer D, Lemeshow S (2003). *Applied logistic regression*, 2nd edition. New York: John Wiley & Sons, Inc. Wiley (paywall), which is herein referred to as **H&L 2nd ed.**

For bug reports, feature requests or suggestions for improvement, please try to submit to github. Otherwise, email me at the address below.

Author(s)

Chris Dardis <christopherdardis@gmail.com>

---

ageChd *Age and Coronary Heart Disease data*

Description

Age and Coronary Heart Disease data

Format

A `data.frame` with 100 observations (rows) and 3 variables (columns).

Details

Age and presence of coronary heart disease for 100 subjects.

Columns are:

ID Identification code. 1 to 100.

age Age (years).

chd Evidence of coronary heart disease? (`factor`):

- **0** no
- **1** yes

Source

Wiley FTP

References

H&L 2nd ed. Page 3, Table 1.1.

See Also

`sig` `OR`

bbdm *Benign Breast Disease Matched study data*

Description

Benign Breast Disease Matched study data

Format

A `data.frame` with 200 observations (rows) and 14 variables (columns).

Details

The relationship between the use of oral contraceptives and fibrocystic breast disease was examined in a hospital-based case-control study undertaken in New Haven, Connecticut, from 1977 to 1979.

This is a subset of the original dataset.

Columns are:

STR stratum $1 - 50$).

OBS observation within stratum (`factor`):

- **1** Case
- **2-4** Control

AGMT Age (years) at interview.

FNDX Final diagnosis (`factor`):

- **0** Control
- **1** Case

HIGD Highest grade in school. $5 - 20$.

DEG Degree (`factor`):

- **0** none
- **1** high_school
- **2** junior_college
- **3** college
- **4** masters
- **5** doctoral

CHK Regular medical checkups? (`factor`):

- **1** Yes
- **2** No

AGP1 Age (years) at first pregnancy.

AGMN Age (years) at menarche.

NLV Non-live 'births'. Number of stillbirths, miscarraiges etc. $0 - 7$.

LIV Number of live births. $0 - 11$.

WT Weight (lbs) at time of interview.

AGLP Age (years) at last menstrual period.

MST Marital status (`factor`):

- **1** married
- **2** divorced
- **3** separated
- **4** widowed
- **5** never_married

Source

Wiley FTP

References

Pastides H, Kelsey JL, LiVolsi VA, Holford TR, Fischer DB, Goldenberg IS 1983. Oral contraceptive use and fibrocystic breast disease with special reference to its histopathology. *Journal of the National Cancer Institute* **71**(1):5–9. Oxford (paywall)

Pastides H, Kelsey JL, Holford TR, LiVolsi VA 1985. The epidemiology of fibrocystic breast disease with special reference to its histopathology. *American Journal of Epidemiology* **121**(3):440–447. Oxford (paywall)

`dx` *Diagnostics for binomial regression*

Description

Diagnostics for binomial regression

Returns diagnostic measures for a binary regression model by covariate pattern

Usage

```
dx(x, ...)

## S3 method for class 'glm'
dx(x, ..., byCov = TRUE)
```

Arguments

`x`	A regression model with class `glm` and `x$family$family == "binomial"`.
`...`	Additional arguments which can be passed to: ?stats::model.matrix e.g. `contrasts.arg` which can be used for `factor` coding.
`byCov`	Return values by *covariate pattern*, rather than by individual observation.

Value

A `data.table`, with rows sorted by $\Delta\hat{\beta}_i$.

If `byCov==TRUE`, there is one row per covariate pattern with at least one observation.

The initial columns give the predictor variables $1 \dots p$.
Subsequent columns are labelled as follows:

y y_i — The *actual* number of observations with $y = 1$ in the model data.

P P_i — Probability of this covariate pattern.
This is given by the inverse of the link function, `x$family$linkinv`. See: ?stats::family

n n_i — Number of observations with these covariates.
If `byCov=FALSE` then this will be $= 1$ for all observations.

yhat $\hat{y}$ — The *predicted* number of observations having a response of $y = 1$, according to the model.
This is:

$$\hat{y_i} = n_i P_i$$

h h_i — Leverage, the diagonal of the **h**at matrix used to generate the model:

$$H = \sqrt{V} X (X^T V X)^{-1} X^T \sqrt{V}$$

Here $^{-1}$ is the inverse and ^{T} is the transpose of a matrix.
X is the matrix of predictors, given by `stats::model.matrix`.
V is an $N \times N$ sparse matrix. All elements are $= 0$ except for the diagonal, which is:

$$v_{ii} = n_i P_i (1 - P_i)$$

Leverage H is also the estimated covariance matrix of $\hat{\beta}$.
Leverage is measure of the influence of this covariate pattern on the model and is approximately

$$h_i \approx x_i - \bar{x} \quad \text{for} \quad 0.1 < P_i < 0.9$$

That is, leverage is approximately equal to the distance of the covariate pattern i from the mean $\bar{x}$.
For values of p which are large (> 0.9) or small (< 0.1) this relationship no longer holds.

Pr Pr_i — The Pearson residual, a measure of influence. This is:

$$Pr_i = \frac{y_i - \mu_y}{\sigma_y}$$

where μ_y and σ_y refer to the mean and standard deviation of a binomial distribution.
$\sigma_y^2 = Var_y$, is the variance.

$$E(y = 1) = \mu_y = \hat{y} = nP \quad \text{and} \quad \sigma_y = \sqrt{nP(1 - P)}$$

Thus:

$$Pr_i = \frac{y_i - n_i P_i}{\sqrt{n_i P_i (1 - P_i)}}$$

dr dr_i

The deviance residual, a measure of influence:

$$dr_i = \text{sign}(y_i - \hat{y}_i)\sqrt{d_i}$$

d_i is the contribution of observation i to the model deviance.
The sign above is:

- $y_i > \hat{y}_i \quad \rightarrow \text{sign}(i) = 1$
- $y_i = \hat{y}_i \quad \rightarrow \text{sign}(i) = 0$
- $y_i < \hat{y}_i \quad \rightarrow \text{sign}(i) = -1$

In logistic regression this is:

$$y_i = 1 \quad \rightarrow \quad dr_i = \sqrt{2\log(1 + \exp(f(x))) - f(x)}$$

$$y_i = 0 \quad \rightarrow \quad dr_i = -\sqrt{2\log(1 + \exp(f(x)))}$$

where $f(x)$ is the linear function of the predictors $1 \ldots p$:

$$f(x) = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_p x_{ip}$$

this is also:

$$dr_i = sign(y_i - \hat{y_i})\sqrt{2(y_i \log\frac{y_i}{\hat{y_i}} + (n_i - y_i)\log\frac{n_i - y_i}{n_i(1 - p_i)})}$$

To avoid the problem of division by zero:

$$y_i = 0 \quad \rightarrow \quad dr_i = -\sqrt{2n_i|\log 1 - P_i|}$$

Similarly to avoid $\log\infty$:

$$y_i = n_i \quad \rightarrow \quad dr_i = \sqrt{2n_i|\log P_i|}$$

The above equations are used when calculating dr_i by covariate group.

sPr sPr_i

The standardized Pearson residual.
The residual is standardized by the leverage h_i:

$$sPr_i = \frac{Pr_i}{\sqrt{(1 - h_i)}}$$

sdr sdr_i

The standardized deviance residual.
The residual is standardized by the leverage, as above:

$$sdr_i = \frac{dr_i}{\sqrt{(1 - h_i)}}$$

dChisq $\Delta\text{P}\chi_\text{i}^2$

The change in the Pearson chi-square statistic with observation i removed. Given by:

$$\Delta P\chi_i^2 = sPr_i^2 = \frac{Pr_i^2}{1 - h_i}$$

where sPr_i is the standardized Pearson residual, Pr_i is the Pearson residual and h_i is the leverage.
$\Delta P\chi_i^2$ should be < 4 if the observation has little influence on the model.

ΔD_i dDev The change in the deviance statistic $D = \sum_{i=1}^{n} dr_i$ with observation i excluded. It is scaled by the leverage h_i as above:

$$\Delta D_i = sdr_i^2 = \frac{dr_i^2}{1 - h_i}$$

$\Delta \hat{\beta}_i$ dBhat The change in $\hat{\beta}$ with observation i excluded.
This is scaled by the leverage as above:

$$\Delta \hat{\beta} = \frac{sPr_i^2 h_i}{1 - h_i}$$

where sPr_i is the standardized Pearson residual.
$\Delta \hat{\beta}_i$ should be < 1 if the observation has little influence on the model coefficients.

Note

By default, values for the statistics are calculated by *covariate pattern*. Different values may be obtained if calculated for each individual obervation (e.g. rows in a `data.frame`).

Generally, the values calculated by covariate pattern are preferred, particularly where the number of observations in a group is > 5.
In this case Pearsons chi-squared and the deviance statistic should follow a chi-squared distribution with $i - p$ degrees of freedom.

See Also

`plot.glm`

Examples

```
## H&L 2nd ed. Table 5.8. Page 182.
## Pattern nos. 31, 477, 468
data(uis)
uis <- within(uis, {
    NDRGFP1 <- 10 / (NDRGTX + 1)
    NDRGFP2 <- NDRGFP1 * log((NDRGTX + 1) / 10)
})
(d1 <- dx(g1 <- glm(DFREE ~ AGE + NDRGFP1 + NDRGFP2 + IVHX +
                    RACE + TREAT + SITE +
                    AGE:NDRGFP1 + RACE:SITE,
                    family=binomial, data=uis)))
d1[519:521, ]
```

genBinom *Generate data for binomial regression*

Description

Generates a `data.frame` or `data.table` with a binary outcome, and a logistic model to describe it.

Usage

```
genBinomDf(b = 2L, f = 2L, c = 1L, n = 20L, nlf = 3L, pb = 0.5,
  rc = 0.8, py = 0.5, asFactor = TRUE, model = FALSE, timelim = 5,
  speedglm = FALSE)

genBinomDt(b = 2L, f = 2L, c = 1L, n = 20L, nlf = 3L, pb = 0.5,
  rc = 0.8, py = 0.5, asFactor = TRUE, model = FALSE, timelim = 5,
  speedglm = FALSE)
```

Arguments

`b`	The number of **b**inomial variables (the number of predictors which are binary). These are limited to 0 or 1.
`f`	The number of **f**actor predictors. The number of predictors which are `factors`.
`c`	The number of **c**ontinuous predictors. the number of predictors which are continuous.
`n`	The **n**umber of observations (rows) in the `data.frame` or `data.table`.
`nlf`	The **n**umber of **l**evels in a **f**actor.
`pb`	The **p**robability for **b**inomnial predictors: the probability of binomial predictors being $= 1$. E.g. if `pb=0.3`, 30% will be 1s, 70% will be 0s
`rc`	The **r**atio for **c**ontinuous variables. The ratio of levels of continuous variables to the total number of observations `n`. E.g. if `rc=0.8` and `n=100`, it will be in the range 1 to 80.
`py`	The **r**atio for **y**, the ratio of 1s to the total number of observations for the binomial predictors. E.g. if `ry=0.5`, 50% will be 1s, 50% will be 0s.
`asFactor`	If `asFactor=TRUE` (the default), predictors given as `factors` will be converted to `factors` in the data frame before the model is fit.
`model`	If `model=TRUE`, will also return a model fitted with `stats::glm` or `speedglm::speedglm`
`timelim`	function will timeout after `timelim` secs. This is present to prevent duplication of rows.
`speedglm`	If `speedglm=TRUE`, return a model fitted with `speedglm` instead of `glm`. See: ?speedglm::speedglm

Value

If `model=TRUE`: a list with the following values:

`df or dt`	A `data.frame` (for `genBinomDf`) or `data.table` (for `genBinomDt`). Predictors are labelled $x1, x2, ..., xn$. The response is y. Rows represent to n observations
`model`	A model fit with `stats::glm` or `speedglm::speedglm`

If `model=FALSE` a `data.frame` or `data.table` as above.

Note

`genBinomDt` is faster and more efficient for large datasets.

Using `asFactor=TRUE` with `factors` which have a large number of `levels` (e.g. `nlf > 30`) on large datasets (e.g. $n > 1000$) can cause fitting to be excessively slow.

Examples

```
set.seed(1)
genBinomDf(speedglm=TRUE)
genBinomDt(b=0, c=2, n=100L, rc=0.7, model=FALSE)
```

---

gof *Goodness of fit tests for binomial regression*

Description

Goodness of fit tests for binomial regression

Usage

```
gof(x, ...)

## S3 method for class 'glm'
gof(x, ..., g = 10, plotROC = TRUE)
```

Arguments

`x`	A regression model with class `glm` and `x$family$family == "binomial"`.
`...`	Additional arguments when plotting the receiver-operating curve. See: ?pROC::roc and ?pROC::plot.roc
`g`	Number of groups (quantiles) into which to split observations for the Hosmer-Lemeshow and the modified Hosmer-Lemeshow tests.
`plotROC`	Plot a receiver operating curve?

Details

Details of the elements in the returned `list` follow below:

ct:
A contingency table, similar to the output of `dx`.
The following are given per *covariate group*:

n	number of observations
y1hat	predicted number of observations with $y = 1$
y1	actual number of observations with $y = 1$
y0hat	predicted number of observations with $y = 0$
y0	actual number of observations with $y = 0$

chiSq:
$P\chi^2$ tests of the significance of the model.
Pearsons test and the deviance D test are given.
These are calculated by indididual `I`, by covariate group `G` and also from the contingency table `CT` above. They are calculated as:

$$P\chi^2 = \sum_{i=1}^{n} Pr_i^2$$

and

$$D = \sum_{i=1}^{n} dr_i^2$$

The statistics should follow a χ^2 distribution with $n - p$ degrees of freedom.
Here, n is the number of observations (taken individually or by covariate group) and p is the number pf predictors in the model.
A **high** p value for the test suggests that the model is a poor fit.
The assumption of a χ^2 distribution is most valid when observations are considered by group.
The statistics from the contingency table should be similar to those obtained when caluclated by group.

ctHL:
The contingency table for the Hosmer-Lemeshow test.
The observations are ordered by probability, then grouped into `g` groups of approximately equal size.
The columns are:

P	the probability
y1	the actual number of observations with $y = 1$
y1hat	the predicted number of observations with $y = 1$
y0	the actual number of observations with $y = 0$
y0hat	the predicted number of observations with $y = 0$
n	the number of observations
Pbar	the mean probability, which is $\frac{nP}{\sum n}$

gof:
All of these tests rely on assessing the effect of adding an additional variable to the model.
Thus a **low** p value for any of these tests implies that the model is a **poor** fit.

Hosmer and Lemeshow tests: Hosmer and Lemeshows C statistic is based on: y_k, the number of observations where $y = 1$, n_k, the number of observations and $\bar{P}_k$, the average probability in group k:

$$\bar{P}_k = \sum_{i=1}^{i=n_k} \frac{n_i P_i}{n_k}, \quad k = 1, 2, \ldots, g$$

The statistic is:

$$C = \sum_{k=1}^{g} \frac{(y_k - n_k \bar{P}_k)^2}{n_k \bar{P}_k (1 - \bar{P}_k)}$$

This should follow a χ^2 distribution with `g - 2` degrees of freedom.

The **modified** Hosmer and Lemeshow test is assesses the change in model deviance D when `G` is added as a predictor. That is, a linear model is fit as:

$$dr_i \sim G, \quad dr_i \equiv \text{devianceresidual}$$

and the effect of adding G assessed with `anova(lm(dr ~ G))`.

Osius and Rojek's tests: These are based on a *power-divergence* statistic PD_λ ($\lambda = 1$ for Pearsons test) and the standard deviation (herein, of a binomial distribution) σ. The statistic is:

$$Z_{OR} = \frac{PD_\lambda - \mu_\lambda}{\sigma_\lambda}$$

For logistic regression, it is calculated as:

$$Z_{OR} = \frac{P\chi^2 - (n - p)}{\sqrt{2(n - \sum_{i=1}^{n} \frac{1}{n_i}) + RSS}}$$

where RSS is the residual sum-of-squares from a weighted linear regression:

$$\frac{1 - 2P_i}{\sigma_i} \sim X, \quad \text{weights} = \sigma_i$$

Here $\boldsymbol{X}$ is the matrix of model predictors.
A two-tailed test against a standard normal distribution $N\ (0, 1)$ should *not* be significant.

Stukels tests: These are based on the addition of the vectors:

$$z_1 = \text{Pgeq}0.5 = \text{sign}(P_i \geq 0.5)$$

and / or

$$z_2 = \text{Pl}0.5 = \text{sign}(P_i < 0.5)$$

to the existing model predictors.
The model fit is compared to the original using the score (e.g. `SstPgeq0.5`) and likelihood-ratio (e.g. `SllPl0.5`) tests. These models should *not* be a significantly better fit to the data.

R2:

Pseudo-R^2 comparisons of the predicted values from the fitted model vs. an intercept-only model.

sum-of-squares: The sum-of-squres (linear-regression) measure based on the squared Pearson correlation coefficient by *individual* is based on the mean probability:

$$\bar{P} = \frac{\sum n_i}{n}$$

and is given by:

$$R^2_{ssI} = 1 - \frac{\sum (y_i - P_i)^2}{\sum (y_i - \bar{P})^2}$$

The same measure, by *covariate group*, is:

$$R^2_{ssG} = 1 - \frac{\sum (y_i - n_i P_i)^2}{\sum (y_i - n_i \bar{P})^2}$$

log-likelihood: The log-likelihood based R^2 measure per *individual* is based on:

- ll_0, the log-likelihood of the intercept-only model
- ll_p, the log-likelihood of the model with p covariates

It is calculated as

$$R^2_{llI} = \frac{ll_0 - ll_p}{ll_0} = 1 - \frac{ll_p}{ll_0}$$

This measure per *covariate group* is based on ll_s, the log-likelihood for the *saturated* model, which is calculated from the model deviance D:

$$ll_s = ll_p - \frac{D}{2}$$

It is cacluated as:

$$R^2_{llG} = \frac{ll_0 - ll_p}{ll_0 - ll_s}$$

auc:

The area under the receiver-operating curve.
This may broadly be interpreted as:

auc	Discrimination
$\text{auc} = 0.5$	useless
$0.7 \leq \text{auc} < 0.8$	acceptable
$0.8 \leq \text{auc} < 0.9$	excellent
$\text{auc} \geq 0.9$	outstanding

$\text{auc} \geq 0.9$ occurs rarely as this reuqires almost complete separation/ perfect classification.

Value

A `list` of `data.tables` as follows:

`ct` Contingency table.

`chiSq` χ^2 tests of the significance of the model. The tests are:

PrI	test of the Pearsons residuals, calculated by *individual*
drI	test of the deviance residuals, calculated by *individual*
PrG	test of the Pearsons residuals, calculated by covariate *group*
drG	test of the deviance residuals, calculated by covariate *group*
PrCT	test of the Pearsons residuals, calculated from the *contingency table*
drCT	test of the deviance residuals, calculated from the *contingency table*

`ctHL` **C**ontingency **t**able for the **H**osmer-**L**emeshow test.

`gof` Goodness-of-fit tests. These are:

- HL Hosmer-Lemeshow's C statistic.
- mHL The modified Hosmer-Lemeshow test.
- OsRo Osius and Rojek's test of the link function.
- S Stukel's tests:

SstPgeq0.5	score test for addition of vector $z1$
SstPl0.5	score test for addition of vector $z2$
SstBoth	score test for addition of vector $z2$
SllPgeq0.5	log-likelihood test for addition of vector $z1$
SllPl0.5	log-likelihood test for addition of vector $z2$
SllBoth	log-likelihood test for addition of vectors $z1$ and $z2$

`R2` R-squared like tests:

ssI	sum-of-squares, by *individual*
ssG	sum-of-squares, by covariate *group*
llI	log-likelihood, by *individual*
llG	log-likelihood, by covariate *group*.

`auc` Area under the receiver-operating curve (ROC) with 95 % CIs.

Additionally, if `plotROC=TRUE`, a plot of the ROC.

Note

The returned `list` has the additional `class` of `"gof.glm"`.
The `print` method for this `class` shows *only* those results which have a p value.

Author(s)

Modified Hosmer & Lemeshow goodness of fit test: adapted from existing work by Yongmei Ni.

References

Osius G & Rojek D, 1992. Normal goodness-of-fit tests for multinomial models with large degrees of freedom. *Journal of the American Statistical Association*. **87**(420):1145-52. JASA (paywall). JSTOR (free):
http://www.jstor.org/stable/2290653

Hosmer D, Hosmer T, Le Cessie S & Lemeshow S (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*. **16**(9):965-80. Wiley (paywall). Duke University (free).

Mittlboch M, Schemper M (1996). Explained variation for logistic regression. *Statistics in Medicine*. **15**(19):1987-97. Wiley (paywall). CiteSeerX / Penn State University (free).

Examples

```
## H&L 2nd ed. Sections 5.2.2, 5.2.4, 5.2.5. Pages 147-167.
data(uis)
uis <- within(uis, {
    NDRGFP1 <- 10 / (NDRGTX + 1)
    NDRGFP2 <- NDRGFP1 * log((NDRGTX + 1) / 10)
})
g1 <- gof(glm(DFREE ~ AGE + NDRGFP1 + NDRGFP2 + IVHX +
              RACE + TREAT + SITE +
              AGE:NDRGFP1 + RACE:SITE,
              family=binomial, data=uis), plot=FALSE)
g1
unclass(g1)
attributes(g1$gof)
```

---

icu *Intensive Care Unit study data*

Description

Intensive Care Unit study data

Format

A `data.frame` with 200 observations (rows) and 14 variables (columns).

Details

A sample of 200 subjects who were part of a study on survival of patients admitted to an adult intensive care unit (ICU).
The observed variable values were modified to protect patient confidentiality.

Columns are:

ID Identification code.

STA Vital status (`factor`):

- **0** lived
- **1** died

AGE Age (years).

SEX Gender (`factor`):

- **0** male
- **1** female

RACE Race (`factor`):

- **1** white
- **2** black
- **3** other

SER Service, when admitted to ICU (`factor`):

- **0** Medical
- **1** Surgical

CAN Cancer part of present problem? (`factor`):

- **0** no
- **1** yes

CRN Chronic renal failure? (`factor`):

- **0** no
- **1** yes

INF Infection probable when admitted to ICU? (`factor`):

- **0** no
- **1** yes

CPR Cardiopulmonary resuscitataion prior to ICU admission? (`factor`):

- **0** no
- **1** yes

SYS Systolic blood pressure (mmHG) when admitted to ICU.

HRA Heart rate when admitted to ICU.

PRE Previous admission to ICU within 6 months? (`factor`):

- **0** no
- **1** yes

TYP Type of admission (`factor`):

- **0** elective
- **1** emergency

FRA Fracture present (long bone, multiple, neck, single area or hip)? (`factor`):

- **0** no
- **1** yes

PO2 pO2 from initial blood gases (`factor`):

- **0** >60

1 <=60

PH pH from initial blood gases (`factor`):

0 >=7.25

1 <7.25

PCO pCO2 from initial blood gases (`factor`):

0 >=18

1 <18

CRE Creatinine from initial blood gases (`factor`):

0 <=2

1 >2

LOC Level of consciousness when admitted to ICU (`factor`):

0 no_coma

1 deep_stupor

2 coma

Source

Wiley FTP

References

H&L 2nd ed. Page 22, Section 1.6.1.

Lemeshow S, Teres D, Avrunin JS, Pastides H 1988. Predicting the outcome of intensive care unit patients. *Journal of the American Statistical Association*. **83**(402):348–356. JSTOR (free)

Lemeshow S, Teres D, Klar J, Avrunin JS, Gehlbach SH , Rapoport John 1993. Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. *Journal of the American Medical Association*. **270**(20):2478–2486. JAMA (paywall)

Lemeshow S, Le Gall J 1994. Modeling the severity of illness of ICU patients: a systems update. *Journal of the American Medical Association*. **272**(13):1049–1055. JAMA (paywall)

`lbw` *Low Birth Weight study data*

Description

Low Birth Weight study data

Format

A `data.frame` with 189 observations (rows) and 11 variables (columns).

Details

This data was collected as part of a larger study at Bayside Medical Center, Springfield, Massachusetts. It contains information on 189 births to women that were seen in the obsetetrics clinic.

The observed variable values were modified to protect patient confidentiality.

Columns are:

ID Identification code.

LOW Low birth weight? (`factor`):

- **0** BWT > 2500g
- **1** BWT <= 2500g

AGE Age of mother.

LWT Weight of mother (lbs) at last menstrual period.

RACE Race (`factor`):

- **1** white
- **2** black
- **3** other

SMOKE Smoking status during pregnancy (`factor`):

- **0** no
- **1** yes

PTL Number of previous premature labors. 0 = none.

HT History of hypertension (`factor`):

- **0** no
- **1** yes

UI History of uterine irritability (`factor`):

- **0** no
- **1** yes

FTV Number of first trimester physician visits. 0 = none.

BWT Birth weight (grams).

Source

Wiley FTP

References

H&L 2nd ed. Page 24. Section 1.6.2.

See Also

`sig` `OR`

`llbw` *Longitudinal Low Birth Weight study data*

Description

Longitudinal Low Birth Weight study data

Format

A `data.frame` with 200 observations (rows) and 8 variables (columns).

Details

A hypothetical data set based on the reference below.
The woman age 45 was excluded as an outlier.
A hypothetical additional number (1 to 3) of births was generated for each woman, yielding an average of 2.6 births per woman.

This is a subset of the original dataset.

Columns are:

ID Identification code.

BIRTH Birth number. 1 to 4.

SMOKE Smoking status during pregnancy (`factor`):

- **0** no
- **1** yes

RACE Race (`factor`):

- **1** white
- **2** black
- **3** other

AGE Age of mother.

LWT Weight of mother (lbs) at last menstrual period.

BWT Birth weight (grams).

LBW Low birth weight? (`factor`):

- **0** BWT > 2500g
- **1** BWT <= 2500g

Source

Wiley FTP

References

H&L 2nd ed. Sections 1.6.2 and 8.3.

mes *Mammography Experience Study data*

Description

Mammography Experience Study data

Format

A `data.frame` with 412 observations (rows) and 7 variables (columns).

Details

A subset of data from a study to assess factors associated with women's knowledge of and attitude towards mammography.

The observed variable values were modified to protect patient confidentiality.

Columns are:

OBS Observation/ identification code.

ME Mammography experience (`factor`):

- **0** never
- **1** within_one_year
- **2** over_one_year_ago

SYMPT "You do not need a mammogram unless you have symptoms" (`factor`):

- **1** stongly_agree
- **2** agree
- **3** disagree
- **4** strongly_disagree

PB Perveived benefit of mammography.
This is the sum of five scaled responses, each on a four point scale.
A low value is indicative of a woman with strong agreement with the benefits of mammography.

HIST Mother or sister with a history of breast cancer? (`factor`):

- **0** no
- **1** yes

BSE Breast self-exam.
"Has anyone taught you how to examine your own breasts?" (`factor`):

- **0** no
- **1** yes

DETC "How likely is it that a mammogram could find a new case of breast cancer?" (`factor`):

- **1** not_likely
- **2** somewhat_likely
- **3** very_likely

Source

Wiley FTP

References

H&L 2nd ed. Page 265. Table 8.1.

Zapka JG, Stoddard A, Maul L, Costanza ME 1991. Interval adherence to mammography screening guidelines. *Medical Care* **29**(8):697–707.
JSTOR (free):
http://www.jstor.org/stable/3766098

Costanza ME, Stoddard AM, Gaw VP, Zapka JG 1992. The risk factors of age and family history and their relationship to screening mammography utilization. *Journal of the American Geriatrics Society* **40**(8):774–778. Wiley (paywall)

Zapka JG, Hosmer D, Costanza ME, Harris DR, Stoddard A 1992. Changes in mammography use: economic, need and service factors. *American Journal of Public Health* **82**(10):1345–1351. AJPH (free)

`mlbw` *Matched Low Birth Weight data*

Description

Matched Low Birth Weight data

Format

A `data.frame` with 112 observations (rows) and 9 variables (columns).

Details

This data was collected as part of a larger study at Bayside Medical Center, Springfield, Massachusetts. It contains information on 56 cases (of low birth weight deliveries) and an equal number of age-matched controls.

The observed variable values were modified to protect patient confidentiality.

A one-to-one matched set was created from the low birth weight data. For each woman who gave birth to a low birth weight baby, a mother of the same age was randomly selected who did not give birth to a low birth weight baby. For three mothers aged < 17, it was not possible to identify a match.

Columns are:

ID Identification code.

LOW Low birth weight? (`factor`):

0 BWT > 2500g

1 BWT <= 2500g

AGE Age of mother.

LWT Weight of mother (lbs) at last menstrual period.

RACE Race (`factor`):

1 white

2 black

3 other

SMOKE Smoking status during pregnancy (`factor`):

0 no

1 yes

PTD Pre-term delivery previously? (`factor`):

0 no

1 yes

HT History of hypertension (`factor`):

0 no

1 yes

UI History of uterine irritability (`factor`):

0 no

1 yes

Source

Wiley FTP

References

H&L 2nd ed. Page 230. Section 7.3.

See Also

`lbw`

`nhanes3` *NHANES III data*

Description

NHANES III data

Format

A `data.frame` with 17030 observations (rows) and 16 variables (columns).

Details

A subset of data from the National Health and Nutrition Examination Study (NHANES) III. Subjects age >=20 are included.
A sample of 39,695 subjects was selected, representing more than 250 million people living in the USA. Data was collected 1988-1994.

49 pseudo strata were created with 2 pseudo-PSU's in each stratum (primary sampling units).

This is a subset of the original dataset.

Columns are:

SEQN Respondent sequence number.

SDPPSU6 Pseudo-PSU (primary sampling unit).

SDPSTRA6 Pseudo stratum.

WTPFHX6 Statistical weight. Range 225.93 to 139744.9.

HSAGEIR Age (years).

HSSEX Gender (`factor`):

- **0** female
- **1** male

DMARACER Race (`factor`):

- **1** white
- **2** black
- **3** other

BMPWTLBS Body weight (lbs).

BMPHTIN Standing height (inches).

PEPMNK1R Average Systolic BP.

PEPMNK5R Average Diastolic BP.

HAR1 Has respondent smoked >100 cigarettes in life (`factor`):

- **1** yes
- **2** no

HAR3 Does respondent smoke cigarettes now? (`factor`):

- **1** yes
- **2** no

SMOKE Smoking (`factor`):

- **1** never (HAR1 = 2)
- **2** >100 cigs (HAR1 = 1 & HAR3 = 2)
- **3** current (HAR1 =1 & HAR3 = 1)

TCP Serum cholesterol (mg/100ml).a

HBP High blood pressure? (`factor`):

- **1** yes (PEPMNK1R > 140)
- **2** no (PEPMNK1R <= 140)

Note

Taken from:
ANALYTIC AND REPORTING GUIDELINES: The Third National Health and Nutrition Examination Survey, NHANES III (1988-94).

In the NHANES III, 89 survey locations were randomly divided into 2 sets or phases, the first consisting of 44 and the other, 45 locations. One set of primary sampling units (PSUs) was allocated to the first 3-year survey period (1988-91) and the other set to the second 3-year period (1991-94).
Therefore, unbiased national estimates of health and nutrition characteristics can be independently produced for each phase as well as for both phases combined. Computation of national estimates from both phases combined (i.e. total NHANES III) is the preferred option; individual phase estimates may be highly variable. In addition, individual phase estimates are not statistically independent.

It is also difficult to evaluate whether differences in individual phase estimates are real or due to methodological differences. That is, differences may be due to changes in sampling methods or data collection methodology over time. At this time, there is no valid statistical test for examining differences between phase 1 and phase 2.

NHANES III is based on a complex multistage probability sample design. Several aspects of the NHANES design must be taken into account in data analysis, including the sampling weights and the complex survey design. Appropriate sampling weights are needed to estimate prevalence, means, medians, and other statistics. Sampling weights are used to produce correct population estimates because each sample person does not have an equal probability of selection. The sampling weights incorporate the differential 3 probabilities of selection and include adjustments for noncoverage and nonresponse.

With the large oversampling of young children, older persons, black persons, and Mexican Americans in NHANES III, it is essential that the sampling weights be used in all analyses. Otherwise, misinterpretation of results is highly likely.

Other aspects of the design that must be taken into account in data analyses are the strata and PSU pairings from the sample design. These pairings should be used to estimate variances and test for statistical significance.

For weighted analyses, analysts can use special computer software packages that use an appropriate method for estimating variances for complex samples such as SUDAAN (Shah 1995) and WesVarPC (Westat 1996).

Although initial exploratory analyses may be performed on unweighted data with standard statistical packages assuming simple random sampling, final analyses should be done on weighted data using appropriate sampling weights.

Source

Wiley FTP

References

H&L 2nd ed. Page 215. Table 6.3.

National Center for Health Statistics (US) and others 1996. NHANES III reference manuals and reports. *National Center for Health Statistics*. CDC (free)

Examples

```
## use simpler column names
data("nhanes3", package="LogisticDx")
n1 <- c("ID", "pStrat", "pPSU", "sWt", "age", "sex",
        "race", "bWt", "h", "sysBP", "diasBP", "sm100",
        "smCurr", "smok", "chol", "htn")
names(nhanes3) <- n1
```

OR — *Odds ratio for binary regression models fit with* `glm`

Description

Odds **r**atio for binary regression models fit with `glm`

Usage

```
OR(x, ...)

## Default S3 method:
OR(x, ...)

## S3 method for class 'glm'
OR(x, ..., newdata = rep(1L, length(stats::coef(x))),
  ci = TRUE, alpha = 0.95, what = c("model", "all", "data"))
```

Arguments

`x`	A numeric object containing probabilities P. I.e. the range of P must be 0 to 1. The odds ratio OR is given by: $$OR_i = \frac{P_i}{1-P_i} = \frac{\frac{P_1}{1-P_1}}{\frac{P_0}{1-P_0}} = \frac{\text{odds}_1}{\text{odds}_0}$$ There is a method for regression models with `class(x)==glm` and `x$family$family == "binomial"`.
`...`	Not used.

`newdata` A `vector` of new variables to use.
There should be one value, in sequence, for each coefficient in the model.
By default, values are calculated for a change in the value of the coefficient for the predictor from 0 to 1.
For continuous predictors changes of > 1 unit may have more practical significance.

`ci` If `ci=TRUE` (the default), include a confidence interval for P_i and OR_i in the returned values.

`alpha` Used to cacluate the confidence interval, which is:

$$\text{CI} = x \pm Z_{1-\alpha}\sigma$$

where the normal distribution $Z \sim N(0, 1)$ and σ is the standard deviation.

`what` See **Value** below.

Value

A `data.table`. Columns give the model, the value of the link function and the associated probability P_i and odds ratio OR_i.

If `ci=TRUE`, will also give upper and lower bounds of the confidence intervals for these values.

Rows are determined by `what`:

`what="model"` The value of the link function is given for the full model.
If an intercept term is included, the value if given with *and* without the intercept.

`what="all"` The value of the link function is given for each *combination* of coefficients in the model.

`what="data"` The value of the link function is given for each set of predictors in the data with which the model was fit.
This option will ignore the argument `newdata`.

Note

In the model formulas, the intercept term is specified as `0` (absent) or `1` (present).
The variance of the values of the link function is:

$$\sigma^2 = \sum x_i^2\sigma^2(\hat{\beta}_i) + \sum 2x_i x_j cov(\hat{\beta}_i, \hat{\beta}_j)$$

where σ^2 is the variance and cov is the covariance.

See Also

?stats::predict.glm

Examples

```
if(require("graphics")){
    plot(x <- seq(from=0.1, to=0.9, by=0.05), y=OR(x))}
## H&L 2nd ed. Table 1.3. Page 10.
data(ageChd)
summary(g1 <- glm(chd ~ age, data=ageChd, family=binomial))
OR(g1)
attributes(OR(g1))
## Table 1.4. Page 20.
stats::vcov(g1)
## Table 2.3. Page 38.
data(lbw)
summary(g1 <- glm(LOW ~ LWT + RACE, data=lbw, family=binomial))
## Table 2.4. Page 42.
vcov(g1)
ageChd$gr54 <- ageChd$age > 54
OR(glm(chd ~ gr54, data=ageChd, family=binomial))
```

---

pcs *Prostate Cancer Study data*

Description

Prostate Cancer Study data

Format

A `data.frame` with 380 observations (rows) and 9 variables (columns).

Details

A subset of data from a study of patient with prostate cancer. Variables measured at the baseline patient exam were used to try to determine whether the tumor had penetrated the prostate capsule.

The observed variable values were modified to protect patient confidentiality.

Columns are:

ID Identification code.

CAPSULE Tumor penetration of prostatic capsule? (`factor`):

- **0** no
- **1** yes

AGE Age (years).

RACE Race (`factor`):

- **1** white
- **2** black

DPROS Digital rectal exam (`factor`):

- **1** no nodule
- **2** unilobar nodule (left)
- **3** unilobar nodule (right)
- **4** bilobar nodule

DCAPS Capsular involvement on rectal exam? (`factor`):

- **0** no
- **1** yes

PSA Prostate Specific Antigen Value (mg/ml).

VOL Tumor volume (cm3)

GLEASON Gleason score (total). Range 0 to 10.

Source

Wiley FTP

References

H&L 2nd ed. Page 25. Section 1.6.3.

`plot.glm` *Plot diagnostics for a binomial* `glm` *model*

Description

Standard diagnostic plots.

Usage

```
## S3 method for class 'glm'
plot(x, y = NULL, ..., toPdf = FALSE, file = "dxPlots.pdf",
  palette = c("Dark2", "Set2", "Accent", "Blues"), usePalette = TRUE,
  bg = NULL, col = "white", alpha = 0.4, cex = 2, pch = 21,
  cex.main = 1.5, inches = 0.25, identify = FALSE, devNew = TRUE)
```

Arguments

`x`	A regression model with class `glm` and `x$family$family == "binomial"`.
`y`	Not used. Present for compatibility with generic `plot()` function.
`...`	Additional arguments, which can be passed to the plotting functions. See: ?graphics::plot.default ?graphics::symbols ?graphics::par
`toPdf`	• If `toPdf=TRUE` the output will be directed to a `.pdf` file.

- If toPdf=FALSE a new device is opened for each plot.

file — Filename if writing to .pdf as above, e.g. "plots.pdf".

palette — Palette of colors to use as the 'fill'/ 'background' colors for the plots.
The options are taken from color_brewer.

usePalette — Use the colorscheme in palette above.

- If usePalette=TRUE (the default), this colorscheme will be passed to the argument bg below:
  - graphics::plot.default(bg=)
  - graphics::symbols(bg=)
- If usePalette=FALSE, then the color specified in bg below will be used instead.

bg — The 'fill' or background color(s) to use, if usePalette=FALSE.
This can be a vector of colors.

col — The 'edge' or 'foreground' color used to outline points in the plot.
The default, "white" is used to make overlapping points easier to see.
This is passed as an argument to

- graphics::plot.default(col=)
- graphics::symbols(fg=)

alpha — Transparency for colors above.
Should be in the range 0 (transparent) to 1 (opaque). See:
?grDevices::adjustcolor

cex — **C**haracter **ex**pansion.
A multiplier used for size of the plotting symbols/ characters. See:
?graphics::par

pch — **P**lotting **ch**aracter.
The symbol/ character to for the plot.
The default, pch=21 shows filled circles at each point. See:
?graphics::points

cex.main — **C**haracter **ex**pansion for the plot title and the labels for the axes.

inches — Width of circles for the bubble plot. See
?graphics::symbols

identify — If TRUE will give option to identify individual points on a number of the plots produced.
The number which appears next to the point corresponds to the relevant row as given by dx.
This may be useful for identifying outliers. See:
?graphics::identify

devNew — If devNew==TRUE (the default), dev.new will be called before each plot.
This is useful in interactive mode.
devNew==FALSE is used for vignette building by package:knitr.

Value

There is one point per observation.

The following show **probability** P_i on the x-axis:

$P_i \times h_i$	Probability vs. leverage.
$P_i \times \Delta P\chi_i^2$	Probability vs. the change in the standardized Pearsons chi-squared with observation i excluded.
$P_i \times \Delta D_i$	Probability vs. the change in the standardized deviance with observation i excluded.
$P_i \times \Delta\hat{\beta}_i$	Probability vs. the change in the standardized maximum likelihood estimators of the model coefficients with observation i excluded.
$P_i \times \Delta P\chi_i^2$	Bubbleplot of probability vs. the change in the standardized Pearsons chi-squared with observation i excluded. The area A_i of each circle is proportional to $\Delta\hat{\beta}_i$: $$A_i = \pi r_i^2 \quad r_i = \sqrt{\frac{\Delta\hat{\beta}_i}{P_i}}$$ For details see: ?graphics::symbols

The following show **leverage** h_i on the x-axis:

$h_i \times \Delta P\chi_i^2$	Leverage vs. the change in the standardized Pearsons chi-squared with observation i excluded.
$h_i \times \Delta D_i$	Leverage vs. the change in the standardized deviance with observation i excluded.
$h_i \times \Delta\hat{\beta}_i$	Leverage vs. the change in the standardized maximum likelihood estimators of the model coefficients with observation i excluded.

The correlation of $\Delta\chi_i^2, \Delta D_i \text{and} \hat{\beta}_i$. is shown in a `pairs` plot. See:
?graphics::pairs

The **Value** of dx is also returned, invisibly.

Note

A choice of colors can be found with e.g.
grDevices::colours()[grep("blue", grDevices::colours())]

Examples

```
## H&L 2nd ed. Table 4.9. Figures 5.5-5.8. Pages 177-180.
data(uis)
uis <- within(uis, {
    NDRGFP1 <- 10 / (NDRGTX + 1)
    NDRGFP2 <- NDRGFP1 * log((NDRGFP1 + 1) / 10)
})
```

```
summary(g1 <- glm(DFREE ~ AGE + NDRGFP1 + NDRGFP2 + IVHX +
                  RACE + TREAT + SITE +
                  AGE:NDRGFP1 + RACE:SITE,
                  family=binomial, data=uis))
plot(g1)
## H&L. Similar to Figure 5.3.
set.seed(133)
(g1 <- glm(sample(c(0, 1), size=100,
                  replace=TRUE, prob=c(0.5, 0.5))
           ~ 0 + I(0.08 * rnorm(n=100, mean=0, sd=sqrt(9))),
           family=binomial))$coef # approx. 0.8
plot(g1)
```

`sig` *Significance tests for a binary regression models fit with* `glm`

Description

Significance tests for a binary regression models fit with `glm`

Usage

```
sig(x, ...)

## S3 method for class 'glm'
sig(x, ..., test = c("var", "coef"))
```

Arguments

`x`	A regression model with class `glm` and `x$family$family == "binomial"`.
`...`	Not used.
`test`	What to test. • If `test="var"` (the default), will test significance for each *variable* in the model. This includes the intercept, if present. This means `factors` are tested for *all* `levels` simultaneously. • If `test="coef"`, will test significance for each *coefficient* in the model. This means the 'dummy variables' created from `factors` will be tested individually.

Value

A `list` of `data.tables` as follows:

`Wald`	The Wald test for each coefficient which is: $$W = \frac{\hat{\beta}}{S\hat{E}_\beta}$$ This should be normally distributed.

`LR` The **l**ikelihood **r**atio test for each coefficient:

$$LR = -2\log\frac{\text{likelihood without variable}}{\text{likelihood with variable}}$$

which is:

$$LR = -2\sum_{i=1}^{n}(y_i\log\frac{P_i}{y_i} + (1-y_i)\log\frac{1-P_i}{1-y_i})$$

When comparing a fitted model to a saturated model (i.e. $P_i = y_i$ and likelihood $= 1$), the LR is referred to as the model *deviance*, D.

`score` The score test, also known as the Rao, Cochran-Armitage trend and the Lagrange multiplier test.
This removes a variable from the model, then assesses the change. For logistic regression this is based on:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

and

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i n_i}{n}$$

The statistic is:

$$\text{ST} = \frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\sqrt{\bar{y}(1-\bar{y})\sum_{i=1}^{n}(x_i - \bar{x}^2)}}$$

If the value of the coefficient is correct, the test should follow a standard normal distribution.

Note

The result has the `class` `"sig.glm"`. The `print` method for this `class` shows only the model coefficients and p values.

See Also

?aod::wald.test
?statmod::glm.scoretest
For corrected score tests:
?mdscore::mdscore

Examples

```
data(ageChd)
## H&L 2nd ed. Table 1.3. Page 10.
summary(g1 <- glm(chd ~ age, data=ageChd, family=binomial))
sig(g1)
data(lbw)
## Table 2.2. Page 36.
summary(g2 <- glm(LOW ~ AGE + LWT + RACE + FTV,
                  data=lbw, family=binomial))
sig(g2)
## Table 2.3. Pages 38-39.
```

```
summary(g3 <- glm(LOW ~ LWT + RACE,
                  data=lbw, family=binomial))
sig(g3, test="coef")
## RACE is more significant when dropped as a factor
##
sig(g3, test="var")
```

---

ss *Sample size for a given coefficient and events per covariate for model*

Description

Sample **s**ize for a given coefficient and events per covariate for model

Usage

```
ss(x, ...)

## S3 method for class 'glm'
ss(x, ..., alpha = 0.05, beta = 0.8,
  coeff = names(stats::coef(x))[2], std = FALSE,
  alternative = c("one.sided", "two.sided"), OR = NULL, Px0 = NULL)
```

Arguments

x	A regression model with class `glm` and `x$family$family == "binomial"`.
...	Not used.
alpha	significance level α for the null-hypothesis significance test.
beta	power β for the null-hypothesis significance test.
coeff	Name of coefficient (variable) in the model to be tested.
std	Standardize the coefficient? If `std=TRUE` (the default), a continuous coefficent will be standardized, using the mean $\bar{x}$ and standard deviation σ_x: $$z_x = \frac{x_i - \bar{x}}{\sigma_x}$$
alternative	The default, `alternative="one.sided"`, checks the null hypothesis with `z = 1 - alpha`. If `alternative="two.sided"`, `z = 1 - alpha/2` is used instead.
OR	Odds ratio. The size of the change in the probability.
Px0	The probability that $x = 0$. If not supplied, this is estimated from the data.

Details

Gives the sample size necessary to demonstrate that a coefficient in the model for the given predictor is equal to its given value rather than equal to zero (or, if `OR` is supplied, the sample size needed to check for such a change in probability).

Also, the number of events per predictor.
This is the *smaller* value of the outcome $y = 0$ and outcome $y = 1$.

For a **continuous** coefficient, the calculation uses $\hat{\beta}$, the estimated coefficient from the model, δ:

$$\delta = \frac{1 + (1 + \hat{\beta}^2) \exp 1.25\hat{\beta}^2}{1 + \exp -0.25\hat{\beta}^2}$$

and P_0, the probability calculated from the intercept term β_0 from the logistic model
`glm(x$y ~ coeff, family=binomial)`
as $P_0 = \frac{\exp \beta_0}{1+\exp \beta_0}$ For a model with one predictor, the calculation is:

$$n = (1 + 2P_0\delta)\frac{z_{1-\alpha} + z_{\texttt{beta}} \exp 0.25\hat{\beta}^{2^2}}{P_0\hat{\beta}^2}$$

For a multivariable model, the value is adjusted by R^2, the correlation of `coeff` with the other predictors in the model:

$$n_m = \frac{n}{1 - R^2}$$

For a **binomial** coefficient, the calculation uses P_0, the probability given the null hypothesis and P_a, the probability given the alternative hypothesis and and the average probability $\bar{P} = \frac{P_0+P_a}{2}$
The calculation is:

$$n = \frac{(z_{1-\alpha}\sqrt{2\bar{P}(1-\bar{P})} + z_{\texttt{beta}}\sqrt{P_0(1-P_0) + P_a(1-P_a)})^2}{(P_a + P_0)^2}$$

An alternative given by Whitemore uses $\hat{P} = P(x = 0)$.
The lead term in the equation below is used to correct for large values of $\hat{P}$:

$$n = (1 + 2P_0)\frac{(z_{1-\alpha}\sqrt{\frac{1}{1-\hat{P}} + \frac{1}{\hat{P}}} + z_{\texttt{beta}}\sqrt{\frac{1}{1-\hat{P}} + \frac{1}{\hat{P}\exp\hat{\beta}}})^2}{(P_0\hat{\beta})^2}$$

As above these can be adjusted in the multivariable case:

$$n_m = \frac{n}{1 - R^2}$$

In this case, Pearsons R^2 correlation is between the fitted values from a logistic regression with `coeff` as the response and the other predictors as co-variates.
The calculation uses $\bar{P}$, the mean probability (mean of the fitted values from the model):

$$R^2 = \frac{(\sum i = 1^n (y_i - \bar{P})(P_i - \bar{P}))^2}{\sum i = 1^n (y_i - \bar{P})^2 \sum i = 1^n (P_i - \bar{P})^2}$$

Value

A list of:

ss	Sample size required to show coefficient for predictor is as given in the model rather than the alternative (by default = 0).
epc	Events per covariate; should be > 10 to make meaningful statements about the coefficients obtained.

Note

The returned `list` has the additional `class` of `"ss.glm"`.
The `print` method for this `class` does not show the attributes.

References

Whitemore AS (1981). Sample Size for Logistic Regression with Small Response Probability. *Journal of the American Statistical Association*. **76**(373):27-32. JASA (paywall)
JSTOR (free)
http://www.jstor.org/stable/2287036

Hsieh FY (1989). Sample size tables for logistic regression. *Statistics in Medicine*. **8**(7):795-802. Wiley (paywall). statpower (free).

Fleiss J (2003). *Statistical methods for rates and proportions. 3rd ed.* John Wiley, New York. Wiley (paywall). Google books (free preview).

Peduzzi P, Concato J, Kemper E, Holford T R, Feinstein A R (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of clinical epidemiology*. **49**(12):1373-79. JCE (paywall). ResearchGate (free).

Examples

```
## H&L 2nd ed. Section 8.5.
## Results here are slightly different from the text due to rounding.
data(uis)
with(uis, prop.table(table(DFREE, TREAT), 2))
(g1 <- glm(DFREE ~ TREAT, data=uis, family=binomial))
ss(g1, coeff="TREATlong")
## Pages 340 - 341.
ss(g1, coeff="TREATlong", OR=1.5, Px0=0.5)
## standardize
uis <- within(uis, {
    AGES <- (AGE - 32) / 6
    NDRGTXS <- (NDRGTX - 5) / 5
})
## Page 343.
## results slightly different due to rounding
g1 <- glm(DFREE ~ AGES, data=uis, family=binomial)
ss(g1, coeff="AGES", std=FALSE, OR=1.5)
## Table 8.37. Page 344.
summary(g1 <- glm(DFREE ~ AGES + NDRGTXS + IVHX + RACE + TREAT,
                  data=uis, family=binomial))
```

```
## Page 345.
## results slightly different due to rounding
ss(g1, coeff="AGES", std=FALSE, OR=1.5)
ss(g1, coeff="TREATlong", std=FALSE, OR=1.5)
```

`uis` *UMARU IMPACT Study data*

Description

UMARU IMPACT Study data

Format

A `data.frame` with 575 observations (rows) and 9 variables (columns).

Details

A subset of data from the University of Massachusets Aids Research Unit (UMARU) IMPACT study.
This came from two concurrent randomized trials of residential treatement for durg abuse, in order to compare planned durations of admission.
Site A randomized 444 participants to compare 3 and 6 month stays in a therapeutic community. They were trained to recognize triggers for relapse and taught skills to cope without using drugs.
Site B randomized 184 participants to receive either a 6 or 12 month stay in a highly structured communal therapeutic community.

This is a subset of the original dataset.

Columns are:

ID Identification code.

AGE Age (years).

BECK Beck Depression score on admission.

IVHX IV drug use history (`factor`):

- **1** never
- **2** previous
- **3** current

NDRUGTX Number of prior drug treatments. Range 5 to 20.

RACE Race (`factor`):

- **0** white
- **1** other

TREAT Treatment randomization. 'Short' is 3 months in site A, 6 months in site B. 'Long' is 6 months in site A, 12 months in site B. (`factor`):

0 short

1 long

SITE Assignment treatment site (`factor`):

0 A

1 B

DFREE Remained drug free for 12 months (`factor`):

0 no

1 yes

Source

Wiley FTP

References

H&L 2nd ed. Page 26. Section 1.6.4.

McCusker J, Vickers-Lahti M, Stoddard A, Hindin R, Bigelow C, Zorn M, Garfield F, Frost R, Love C, Lewis B 1995. Fischer DB, Goldenberg IS 1983. The effectiveness of alternative planned durations of residential drug abuse treatment. *American Journal of Public Health* **85**(10):1426–1429. APHA (free)

McCusker J, Bigelow C, Frost R, Garfield F, Hindin R, Vickers-Lahti M, Lewis B 1997. #' The effects of planned duration of residential drug abuse treatment on recovery and HIV risk behavior. *American Journal of Public Health* **87**(10):1637–1644. APHA (free)

McCusker J, Bigelow C, Vickers-Lahti M, Spotts D, Garfield F, Frost R 1997. Planned duration of residential drug abuse treatment: efficacy versus effectiveness. *Addiction* **92**(11):1467–1478. Wiley (paywall)

See Also

`dx` `plot.glm`

Index