

# Package ‘dispmod’

February 9, 2012

**Version** 1.0.1

**Date** 2009-02-10

**Title** Dispersion models.

**Author** Luca Scrucca <luca@stat.unipg.it>

**Maintainer** Luca Scrucca <luca@stat.unipg.it>

**Depends** R (>= 2.6)

**Description** Functions for modelling dispersion in GLM.

**License** GPL-2

**Repository** CRAN

**Date/Publication** 2009-02-10 19:01:50

## R topics documented:

glm.binomial.disp . . . . .	2
glm.poisson.disp . . . . .	3
holford . . . . .	6
lm.disp . . . . .	7
minitab . . . . .	9
orobanche . . . . .	10
salmonellaTA98 . . . . .	10

<b>Index</b>	<b>12</b>
--------------	-----------

---

glm.binomial.disp      *Overdispersed binomial logit models.*

---

### Description

This function estimates overdispersed binomial logit models using the approach discussed by Williams (1982).

### Usage

```
glm.binomial.disp(object, maxit = 30, verbose = TRUE)
```

### Arguments

object	an object of class "glm" providing a fitted binomial logistic regression model.
maxit	integer giving the maximal number of iterations for the model fitting procedure.
verbose	logical, if TRUE information are printed during each step of the algorithm.

### Details

Extra-binomial variation in logistic linear models is discussed, among others, in Collett (1991). Williams (1982) proposed a quasi-likelihood approach for handling overdispersion in logistic regression models.

Suppose we observe the number of successes  $y_i$  in  $m_i$  trials, for  $i = 1, \dots, n$ , such that

$$y_i | p_i \sim \text{Binomial}(m_i, p_i)$$

$$p_i \sim \text{Beta}(\gamma, \delta)$$

Under this model, each of the  $n$  binomials has a different probability of success  $p_i$ , where  $p_i$  is a random draw from a beta distribution. Thus,

$$E(p_i) = \frac{\gamma}{\gamma + \delta} = \theta$$

$$\text{Var}(p_i) = \phi\theta(1 - \theta)$$

Assume  $\gamma > 1$  and  $\delta > 1$ , so that the beta density is equal to zero at both zero and one, and thus  $0 < \phi \leq 1/3$ . From this, the unconditional mean and variance can be calculated:

$$E(y_i) = m_i\theta$$

$$\text{Var}(y_i) = m_i\theta(1 - \theta)(1 + (m_i - 1)\phi)$$

so unless  $m_i = 1$  or  $\phi = 0$ , the unconditional variance of  $y_i$  is larger than binomial variance.

Identical expressions for the mean and variance of  $y_i$  can be obtained if we assume that the  $m_i$  counts on the  $i$ -th unit are dependent, with the same correlation  $\phi$ . In this case,  $-1/(m_i - 1) < \phi \leq 1$ .

The method proposed by Williams uses an iterative algorithm for estimating the dispersion parameter  $\phi$  and hence the necessary weights  $1/(1 + \phi(m_i - 1))$  (for details see Williams, 1982).

**Value**

The function returns an object of class "glm" with the usual information (see `help(glm)`) and the added components:

`dispersion`      the estimated dispersion parameter.  
`disp.weights`    the final weights used to fit the model.

**Note**

Based on a similar procedure available in Arc (Cook and Weisberg, <http://www.stat.umn.edu/arc>)

**Author(s)**

Luca Scrucca, <luca@stat.unipg.it>

**References**

Collett, D. (1991), *Modelling Binary Data*, London: Chapman and Hall.  
Williams, D. A. (1982), Extra-binomial variation in logistic linear models, *Applied Statistics*, **31**, 144–148.

**See Also**

[lm](#), [glm](#), [lm.disp](#), [glm.poisson.disp](#)

**Examples**

```
data(orobanche)
attach(orobanche)
h <- factor(host)
v <- factor(variety, levels=c("0.a75", "0.a73"))

mod <- glm(cbind(germinated, seeds-germinated) ~ h + v + h*v, family=binomial(logit))
summary(mod)

mod.disp <- glm.binomial.disp(mod)
summary(mod.disp)
mod.disp$dispersion
```

---

`glm.poisson.disp`      *Overdispersed Poisson log-linear models.*

---

**Description**

This function estimates overdispersed Poisson log-linear models using the approach discussed by Breslow N.E. (1984).

**Usage**

```
glm.poisson.disp(object, maxit = 30, verbose = TRUE)
```

**Arguments**

`object` an object of class "glm" providing a fitted binomial logistic regression model.  
`maxit` integer giving the maximal number of iterations for the model fitting procedure.  
`verbose` logical, if TRUE information are printed during each step of the algorithm.

**Details**

Breslow (1984) proposed an iterative algorithm for fitting overdispersed Poisson log-linear models. The method is similar to that proposed by Williams (1982) for handling overdispersion in logistic regression models ([glm.binomial.disp](#)).

Suppose we observe  $n$  independent responses such that

$$y_i | \lambda_i \sim \text{Poisson}(\lambda_i n_i)$$

for  $i = 1 \dots, n$ . The response variable  $y_i$  may be an event counts variable observed over a period of time (or in the space) of length  $n_i$ , whereas  $\lambda_i$  is the rate parameter. Then,

$$E(y_i | \lambda_i) = \mu_i = \lambda_i n_i = \exp(\log(n_i) + \log(\lambda_i))$$

where  $\log(n_i)$  is an offset and  $\log(\lambda_i) = \beta' x_i$  expresses the dependence of the Poisson rate parameter on a set of, say  $p$ , predictors. If the periods of time are all of the same length, we can set  $n_i = 1$  for all  $i$  so the offset is zero.

The Poisson distribution has  $E(y_i | \lambda_i) = \text{Var}(y_i | \lambda_i)$ , but it may happen that the actual variance exceeds the nominal variance under the assumed probability model. Suppose now that  $\theta_i = \lambda_i n_i$  is a random variable distributed according to

$$\theta_i \sim \text{Gamma}(\mu_i, 1/\phi)$$

where  $E(\theta_i) = \mu_i$  and  $\text{Var}(\theta_i) = \mu_i^2 \phi$ . Thus, it can be shown that the unconditional mean and variance of  $y_i$  are given by

$$E(y_i) = \mu_i$$

and

$$\text{Var}(y_i) = \mu_i + \mu_i^2 \phi = \mu_i(1 + \mu_i \phi)$$

Hence, for  $\phi > 0$  we have overdispersion. It is interesting to note that the same mean and variance arise also if we assume a negative binomial distribution for the response variable.

The method proposed by Breslow uses an iterative algorithm for estimating the dispersion parameter  $\phi$  and hence the necessary weights  $1/(1 + \mu_i \hat{\phi})$  (for details see Breslow, 1984).

**Value**

The function returns an object of class "glm" with the usual information (see `help(glm)`) and the added components:

`dispersion` the estimated dispersion parameter.  
`disp.weights` the final weights used to fit the model.

**Note**

Based on a similar procedure available in Arc (Cook and Weisberg, <http://www.stat.umn.edu/arc>)

**Author(s)**

Luca Scrucca, <luca@stat.unipg.it>

**References**

Breslow, N.E. (1984), Extra-Poisson variation in log-linear models, *Applied Statistics*, **33**, 38–44.

**See Also**

[lm](#), [glm](#), [lm.disp](#), [glm.binomial.disp](#)

**Examples**

```
##-- Salmonella TA98 data

data(salmonellaTA98)
attach(salmonellaTA98)
log.x10 <- log(x+10)
mod <- glm(y ~ log.x10 + x, family=poisson(log))
summary(mod)

mod.disp <- glm.poisson.disp(mod)
summary(mod.disp)
mod.disp$dispersion

# compute predictions on a grid of x-values...
x0 <- seq(min(x), max(x), length=50)
eta0 <- predict(mod, newdata=data.frame(log.x10=log(x0+10), x=x0), se=TRUE)
eta0.disp <- predict(mod.disp, newdata=data.frame(log.x10=log(x0+10), x=x0), se=TRUE)
# ... and plot the mean functions with variability bands
plot(x, y)
lines(x0, exp(eta0$fit))
lines(x0, exp(eta0$fit+2*eta0$se), lty=2)
lines(x0, exp(eta0$fit-2*eta0$se), lty=2)
lines(x0, exp(eta0.disp$fit), col=2)
lines(x0, exp(eta0.disp$fit+2*eta0.disp$se), lty=2, col=2)
lines(x0, exp(eta0.disp$fit-2*eta0.disp$se), lty=2, col=2)

##-- Holford's data

data(holford)
attach(holford)

mod <- glm(incid ~ offset(log(pop)) + Age + Cohort, family=poisson(log))
summary(mod)

mod.disp <- glm.poisson.disp(mod)
```

```
summary(mod.disp)
mod.disp$dispersion
```

---

holford

*Holford's data on prostatic cancer deaths*

---

### Description

Holford's data on prostatic cancer deaths and mid-period population denominators for non-whites in the US by age and calendar period. Thirteen birth cohorts from 1855-59 through to 1915-19 are represented in at least one of seven 5-year age groups (50-54 through to 80-84) and one of the seven 5-year calendar periods (1935-39 through to 1965-69) for which data are provided.

### Usage

```
data(minitab)
```

### Format

This data frame contains the following columns:

**incid** number of prostatic cancer deaths.

**pop** mid-period population counts.

**Age** age groups.

**Period** calendar periods.

**Cohort** cohorts.

### Source

Holford, T.R. (1983) The estimation of age, period and cohort effects for vital rates. *Biometrics*, **39**, 311–324.

### References

Breslow, N.E. (1984), Extra-Poisson variation in log-linear models, *Applied Statistics*, **33**, 38–44.

---

lm.disp                      *Normal dispersion models.*

---

### Description

This function estimates normal dispersion regression models.

### Usage

```
lm.disp(formula, var.formula, data = list(), maxit = 30,
        epsilon = glm.control()$epsilon, subset, na.action = na.omit,
        contrasts = NULL, offset = NULL)
```

### Arguments

formula	a symbolic description of the mean function of the model to be fit. For the details of model formula specification see <code>help(lm)</code> and <code>help(formula)</code> .
var.formula	a symbolic description of the variance function of the model to be fit. This must be a one-sided formula; if omitted the same terms used for the mean function are used. For the details of model formula specification see <code>help(lm)</code> and <code>help(formula)</code> .
data	an optional data frame containing the variables in the model. By default the variables are taken from <code>'environment(formula)'</code> , typically the environment from which the function is called.
maxit	integer giving the maximal number of iterations for the model fitting procedure.
epsilon	positive convergence tolerance epsilon; the procedure converge when <code>ldev - devoldl &lt; epsilon</code> .
subset	an optional vector specifying a subset of observations to be used in the fitting process.
na.action	a function which indicates what should happen when the data contain 'NA's. The default is set by the 'na.action' setting of 'options', and is 'na.fail' if that is unset. The default is 'na.omit'.
contrasts	an optional list. See the 'contrasts.arg' of 'model.matrix.default'.
offset	this can be used to specify an a priori known component to be included in the linear predictor during fitting. An 'offset' term can be included in the formula instead or as well, and if both are specified their sum is used.

### Details

Normal dispersion models allow to model variance heterogeneity in normal regression analysis using a log-linear model for the variance.

Suppose a response  $y$  is modelled as a function of a set of  $p$  predictors  $x$  through the linear model

$$y_i = \beta'x_i + e_i$$

where  $e_i \sim N(0, \sigma^2)$  under homogeneity. Variance heterogeneity is expressed as

$$\text{Var}(e_i) = \sigma^2 = \exp(\lambda' z_i)$$

where  $z_i$  may contain some or all the variables in  $x_i$  and other variables not included in  $x_i$ ;  $z_i$  is however assumed to contain a constant term. This model can be re-expressed also as

$$E(y|x) = \beta' x$$

$$\text{Var}(y|x) = \exp(\lambda' z)$$

and is fitted by maximum likelihood following the algorithm described in Aitkin (1987).

### Value

'lm.dispmod' returns an object of 'class' "dispmod".

The function 'summary' is used to obtain and print a summary of the results.

An object of class "lm.dispmod" is a list containing the following components:

call	the matched call.
mean	an object of class "glm" giving the fitted model for the mean function.
var	an object of class "glm" giving the fitted model for the variance function.
initial.deviance	the value of the deviance at the beginning of the iterative procedure, i.e. assuming constant variance.
deviance	the value of the deviance at the end of the iterative procedure.

### Note

Based on a similar procedure available in Arc (Cook and Weisberg, <http://www.stat.umn.edu/arc>)

### Author(s)

Luca Scrucca, <luca@stat.unipg.it>

### References

Aitkin, M. (1987), Modelling variance heterogeneity in normal regression models using GLIM, *Applied Statistics*, **36**, 332–339.

### See Also

[lm](#), [glm](#), [glm.binomial.disp](#), [glm.poisson.disp](#), [ncv.test](#) (in the car library).

**Examples**

```

data(minitab)
attach(minitab)

y <- V^(1/3)
summary(mod <- lm(y ~ H + D))

summary(mod.disp1 <- lm.disp(y ~ H + D))
summary(mod.disp2 <- lm.disp(y ~ H + D, ~ H))

# Likelihood ratio test
deviances <- c(mod.disp1$initial.deviance, mod.disp2$deviance, mod.disp1$deviance)
lrt <- c(NA, abs(diff(deviances)))
cbind(deviances, lrt, p.value=1-pchisq(lrt, 1))

# quadratic dispersion model on D (as discussed by Aitkin)
summary(mod.disp4 <- lm.disp(y ~ H + D, ~ D + I(D^2)))
r <- mod$residuals
plot(D, log(r^2))
phi.est <- mod.disp4$var$fitted.values
lines(D, log(phi.est))

```

---

minitab

*Minitab tree data*


---

**Description**

Data on 31 black cherry trees sampled from the Allegheny National Forest, Pennsylvania.

**Usage**

```
data(minitab)
```

**Format**

This data frame contains the following columns:

**D** diameter 4.5 feet of the ground, inches

**H** height of the tree, feet

**V** marketable volume of wood, cubic feet

**Source**

Ryan, T.A., Joiner, B.L. and Ryan, B.F. (1976) *Minitab Student Handbook*. N. Scituate, MA: Duxbury.

**References**

Cook, R.D. and Weisberg, S. (1982) *Residuals and Influence in Regression*, New York: Chapman and Hall, p. 66.

---

orobanche

*Germination of Orobanche*

---

### Description

Orobanche, commonly known as broomrape, is a genus of parasitic plants with chlorophyll that grow on the roots of flowering plants. Batches of seeds of two varieties of the plant were brushed onto a plate of diluted extract of bean or cucumber, and the number germinating were recorded.

### Usage

```
data(orobanche)
```

### Format

This data frame contains the following columns:

**germinated** Number germinated

**seeds** Number of seeds

**slide** Slide number

**host** Host type

**variety** Variety name

### Source

Crowder, M.J. (1978) Beta-binomial anova for proportions. *Applied Statistics*, **27**, 34–37.

### References

Collett, D. (1991) *Modelling Binary Data*, London: Chapman and Hall, Chapter 6.

---

salmonellaTA98

*Salmonella reverse mutagenicity assay*

---

### Description

Data on Ames Salmonella reverse mutagenicity assay.

### Usage

```
data(salmonellaTA98)
```

**Format**

This data frame contains the following columns:

**x** dose levels of quinoline

**y** numbers of revertant colonies of TA98 Salmonella observed on each of three replicate plates testes at each of six dose levels of quinolinediameter 4.5 feet of the ground, inches

**Source**

Margolin, B.J., Kaplan, N. and Zeiger, E. (1981) Statistical analysis of the Ames Salmonella/microsome test, *Proc. Natl. Acad. Sci. USA*, **76**, 3779–3783.

**References**

Breslow, N.E. (1984), Extra-Poisson variation in log-linear models, *Applied Statistics*, **33**, 38–44.

# Index

## \*Topic **datasets**

holford, [6](#)

minitab, [9](#)

orobanche, [10](#)

salmonellaTA98, [10](#)

## \*Topic **models**

glm.binomial.disp, [2](#)

glm.poisson.disp, [3](#)

lm.disp, [7](#)

## \*Topic **regression**

glm.binomial.disp, [2](#)

glm.poisson.disp, [3](#)

lm.disp, [7](#)

glm, [3](#), [5](#), [8](#)

glm.binomial.disp, [2](#), [4](#), [5](#), [8](#)

glm.poisson.disp, [3](#), [3](#), [8](#)

holford, [6](#)

lm, [3](#), [5](#), [8](#)

lm.disp, [3](#), [5](#), [7](#)

minitab, [9](#)

ncv.test, [8](#)

orobanche, [10](#)

salmonellaTA98, [10](#)

summary.dispmod(lm.disp), [7](#)