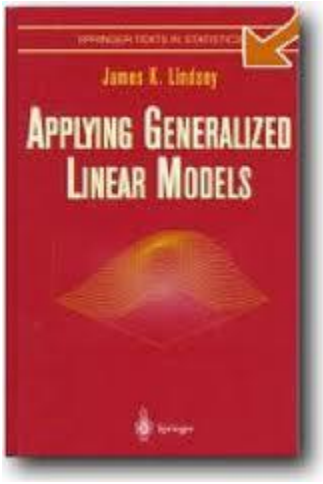
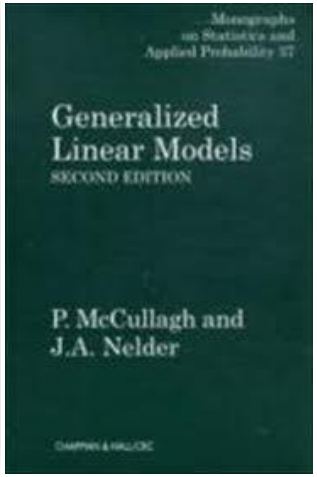
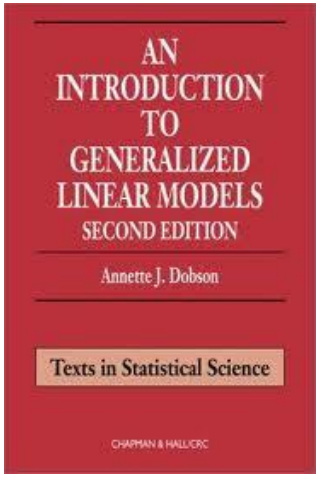
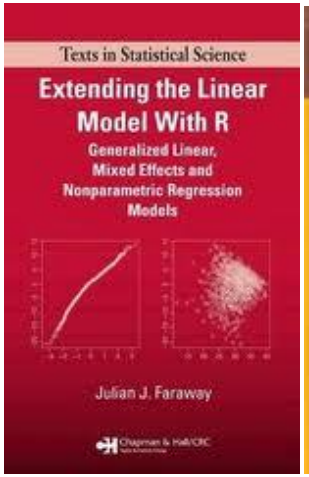

James K. Lindsey
APPLYING GENERALIZED LINEAR MODELS


Monographs on Statistics and Applied Probability 37
Generalized Linear Models
SECOND EDITION
P. McCullagh and J.A. Nelder
CHAPMAN & HALL/CRC


AN INTRODUCTION TO GENERALIZED LINEAR MODELS
SECOND EDITION
Annette J. Dobson
Texts in Statistical Science
CHAPMAN & HALL/CRC


Texts in Statistical Science
Extending the Linear Model With R
Generalized Linear, Mixed Effects and Nonparametric Regression Models
Julian J. Faraway


Springer Series in Statistics
Ludwig Fahrmeir
Gerhard Tutz
Multivariate Statistical Modelling Based on Generalized Linear Models

# Generalized Linear Modelling

Evol Ecol (2011) 25:1179–1195
DOI 10.1007/s10682-011-9469-7

ORIGINAL PAPER

## Maternal and paternal contributions to egg size and egg number variation in the blackfin pearl killifish *Austrolebias nigripinnis*

Mahmoud Moshgani · Tom J. M. Van Dooren

(La Paloma Uruguay)

# Generalized Linear Modelling



**Table 1** Maximal model fitted to each dependent variable

| Grouping factor | Description (number of parameters) |
|---|---|
| Fixed effect | Age (1) |
| | Weekday (3, Monday–Wednesday–Friday) |
| | Age × Weekday Interaction (3) |
| Random effect | |
| Females | Age (1) |
| Females | Weekday (3) |
| Random effect | |
| Males | Age (1) |
| Males | Weekday (3) |
| Random effect | |
| Female × male combination | Intercept (1) |
| Random effect | |
| Tank | Intercept (1) |
| Error term | (1, only for Reproductive effort and egg size) |

The model contains fixed effects, random effects, and for the reproductive effort and egg size traits also an error variance

Random effects can be estimated if there are repeated observations per level of a grouping factor (fish tank, females, males, or female-male combinations). The weekday variable is categorical with three levels. For the fixed effects, there is a separate parameter estimated per weekday level, for the random effects, a separate variance component is estimated per weekday

## Generalized Linear Modelling GLM

- investigate importance/relevance of explanatory variables
  These can be continuous or discrete - we can include interactions between them

- a single response variable

- response: probability distribution from exponential dispersion family

- start from a maximal model - model selection - hypothesis testing
  likelihood & frequentist approaches
-
- Checking assumptions: Investigate model fit graphically (or with a test)

**Model matrix *X***

$$\eta = X\beta$$

```
head(model.matrix(model),3)
  (Intercept) x
1           1 1
2           1 2
3           1 3
```

membership of a specific group is indicated by **dummy variables**

help(lm) example from Dobson

groups *Control* (Ctl) and *Treatment* (Trt)

```
model.matrix(glm.D9)[1,]
(Intercept)    groupTrt
          1           0
model.matrix(glm.D9)[11,]
(Intercept)    groupTrt
          1           1
```

- **Assume** we want to fit a model with - in the linear predictor - effects of one covariate *x* and one categorical variable
  the categorical variable has two levels/groups

  observations belong to group one or group two
  the covariate is observed for each observation

  in group one: $\eta_i = \beta_{01} + \beta_{11} x_i$
  in group two: $\eta_i = \beta_{02} + \beta_{12} x_i$

  by using **dummies - membership variables - indicator variables**
  we can combine this into one equation that applies to all observations

$I_1$ indicates membership of group one, then $I_1 = 1$, otherwise $I_1 = 0$
$I_2$ indicates membership of group two , then $I_2 = 1$, otherwise $I_2 = 0$

in group one: $\eta_i = \beta_{01} + \beta_{11} x_i$
in group two: $\eta_i = \beta_{02} + \beta_{12} x_i$

**combined**

$$\eta_i = \beta_{01} I_1 + \beta_{02} I_2 + \beta_{11} I_1 x_i + \beta_{12} I_2 x_i$$

parameters can be redefined without changing any prediction

$\beta'_{01} = \beta_{01}$ and $\beta'_{02} = \beta_{02} - \beta_{01}$
$\beta'_{11} = \beta_{11}$ and $\beta'_{12} = \beta_{12} - \beta_{11}$

$$\eta'_i = \beta'_{01} + \beta'_{02} I_2 + \beta'_{11} x_i + \beta'_{12} I_2 x_i$$

$$\eta'_i = \beta'_{01} + \beta'_{02} I_2 + \beta'_{11} x_i + \beta'_{12} I_2 x_i$$

the last term is called an **interaction** term, it contains a product of explanatory (in this case including dummy) variables

in this formulation, some parameters represent differences between groups.

A test whether some of these parameters are zero represents in fact a test for a difference between groups

In R:

$$\eta_i = \beta'_{01} + \beta'_{02} I_2 + \beta'_{11} x_i + \beta'_{12} I_2 x_i$$

*y~1 + groups + x + groups:x* or shorthand *y~groups*x*

$$\eta_i = \beta_{01} I_1 + \beta_{02} I_2 + \beta_{11} I_1 x_i + \beta_{12} I_2 x_i$$

*y~groups + groups:x -1* **we have to avoid that an intercept gets added automatically**


Fender
www.fender.com
Fender
Guitar Shop

**Generalized Linear Modelling**

General approach model selection / hypothesis testing

B is **nested** within A: model A – model B is equal to model A, but with one or more model "terms" deleted.

1. **Choose** family – link – model equations
2. **Fit both models**
3. **Model comparison**

Model comparison of **NESTED** models:

Usually, a more complex model fits the data better.

The deviance is generally smaller for the more complex model in the pair

If the simplest model in the pair is TRUE, what is the probability of finding the deviance reduction between both models?

Probability: what we expect if we repeat the experiment a lot of times and compare both models each time. **Frequentist**

Use "anova" techniques to compare nested models:

use F or likelihood ratio Chi-squared statistics based on model **deviances** or **scaled deviances** to test whether the more complicated model is preferred over the simplified one.

use preferably **F tests** when **dispersion parameters are estimated**

or otherwise scaled deviances and likelihood ratio tests.

1. **Choose** family – link – model equations

normally distributed error
identity link

null model without slope: coded as "*y* ~ 1"
model without intercept: "*y* ~ *x* -1"
zero model: "*y* ~ -1"

2. **Fit all models involved in comparison**

```
# with some aspects of R equation notation
model2<-glm(y~1)
summary(model2)

model3<-glm(y~x-1)
summary(model3)

model4<-glm(y~-1)
summary(model4)
```

## 3. **Model comparison**

```
anova(model2,model,test="F") # test for slope
Model 1: y ~ x
Model 2: y ~ 1
      Resid. Df Resid. Dev      Df    Deviance F        Pr(>F)
1     18        0.236
2     19        58.795          1     58.560   4474.4   < 2.2e-16 ***

anova(model3,model,test="F") # test for intercept
Model 1: y ~ x
Model 2: y ~ x - 1
      Resid. Df Resid. Dev      Df    Deviance F        Pr(>F)
1     18        0.236
2     19        229.068         1     228.833  17484    < 2.2e-16 ***

anova(model4,model,test="F") # comparison with model where overall mean is zero
Model 1: y ~ x
Model 2: y ~ -1
      Resid. Df Resid. Dev      Df    Deviance F        Pr(>F)
1     18        0.236
2     20        2116.38         2     2116.14  80844    < 2.2e-16 ***
```

**drop1()** Shorthand - for very common comparisons: fitted model and nested models with each independent term removed

```
drop1(model,test="F") # F test for slope
Single term deletions

Model:
y ~ x
       Df  Deviance AIC    F value  Pr(F)
<none>     0.236    -26.071
x      1   58.795   82.324 4474.4   < 2.2e-16 ***

drop1(model,test="Chisq") # likelihood ratio test for slope
Single term deletions

Model:
y ~ x
       Df  Deviance AIC    scaled dev.  Pr(Chi)
<none>     0.236    -26.071
x      1   58.795   82.324 110.395      < 2.2e-16 ***
```

**Parametric bootstrap**

Simulations can be used to generate probability distributions of statistics according the null hypothesis (assuming the model simulated is true)

Example simulate 500 datasets with no effect of x
refit the regression model on each simulated dataset
get student *t* statistic for the slope

```
simdata<-simulate(model2,500) # simulate simpler model
tvalsim<-numeric(500)
for(i in 1:500)
{tvalsim[i]<-summary(update(model,simdata[,i]~.))$coef[2,3]}
# refit more complicated model
hist(tvalsim) # histogram results
```

t-value of estimate in data is 72.78

```
sum(abs(tvalsim)>72.78) # probability of a more extreme value
[1] 0
```

Model comparison of **NON-NESTED** models

Easy: An **Akaike Information Criterion AIC**
**measure of model fit + complexity penalty**
models with lower AIC are preferred - These have better fit or/and fewer parameters

**glm**:
twice (- log likelihood ML fit + number of parameters in model)

AIC is in principle pure likelihood inference **weaker** than frequentist inference

```
model2$aic # no slope
[1] 82.32416
model3$aic # no intercept
[1] 109.5233
```

Model comparison of **NON-NESTED** models

More involved: use parametric bootstrap and appropriate statistics, e.g. the *J* statistic. Difficulty: choosing $H_0$ well so that inconsistent results are unlikely

An entry to this literature:

Pesaran, M. H., & Weeks, M. (2001). Non-nested hypothesis testing: an overview. *A Companion to Theoretical Econometrics*, 279-309.

https://www.repository.cam.ac.uk/bitstream/handle/1810/429/nnest.pdf?sequence=1

## Models

### Saturated model

*One parameter per data value. Fits the data perfectly.*

### Maximal model

*Linear predictor includes all potentially relevant independent variables.*

### Minimum adequate model **MAM**

*Result of model selection by means of model comparisons*
*Simplified from the maximal model without sacrificing too much fit.*
*Preferably by simplifying one parameter at a time, if hypotheses allow that.*

### Null model

*One parameter, the overall mean of the response, or two (mean & standard deviation).*

**Saturated model**

general baseline, which applies to all response variables with independent observations in categories indexed $i$ ($i$ = 1,..., $I$).

multinomial model $$P(n_1, n_2, ..., n_I) = \binom{N}{n_1 n_2 ... n_I} \prod_{i=1}^{I} \pi_i^{n_i}$$

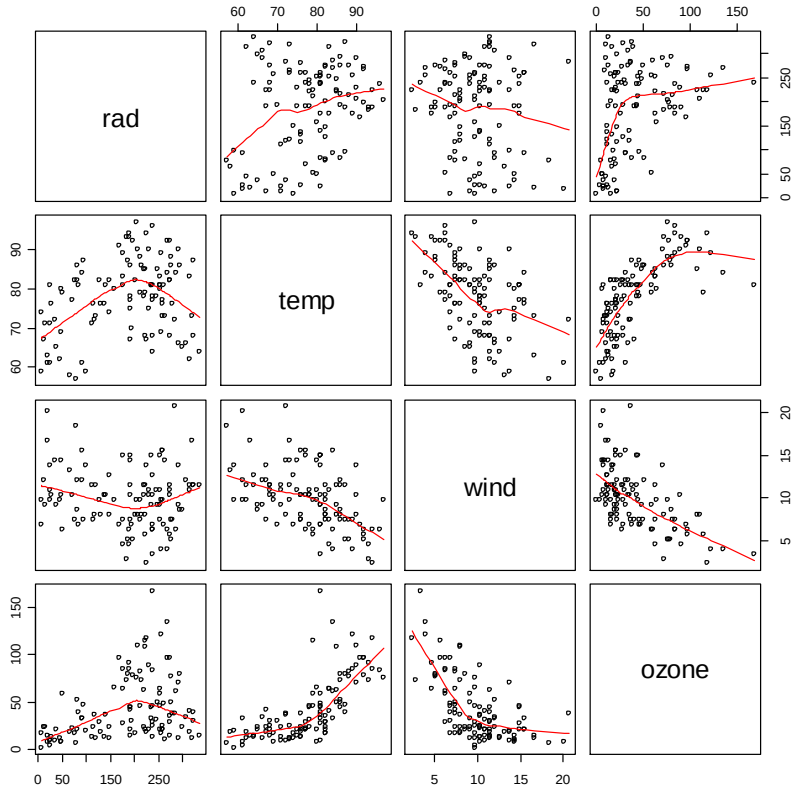
The ML-probability of getting an observation in category $i$ equals

$\hat{\pi}_i = \frac{n_i}{N}$ (for $i$ in 1, ..., $I - 1$).

We can use this to calculate likelihoods and deviances of models. The AIC becomes a simple expression, $2(I - 1)$!

# Model selection example **Interactions**: Ozone concentrations

```
ozone.pollution<-read.table("c:\\temp\\Crawley\\ozone.data.txt",header=T)
attach(ozone.pollution)
pairs(ozone.pollution,panel=panel.smooth)
```



```
model1<-glm(ozone~temp*wind*rad+I(rad^2)+I(temp^2)+I(wind^2))
drop1(model1,test="F")
Single term deletions

Model:
ozone ~ temp * wind * rad + I(rad^2) + I(temp^2) + I(wind^2)
                Df  Deviance AIC F value    Pr(F)
<none>          31742     967
I(rad^2)        1     32370     967  1.9784    0.16265
I(temp^2)       1     33624     971  5.9275    0.01668 *
I(wind^2)       1     37228     983  17.2832   6.807e-05 ***
temp:wind:rad 1     31879     965  0.4298    0.51358
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

plot(model)
```

- Remove higher-order interactions first
- Remove non-linear terms first
- Don’t remove interactions and their component terms at the same time

```
model2<-glm(ozone~temp*wind+temp*rad+wind*rad+I(rad^2)+I(temp^2)+I(wind^2))
drop1(model2,test="F")
```

| | Df | Deviance | AIC | F value | Pr(F) |
|---|---|---|---|---|---|
| <none> | | 31879 | 965 | | |
| I(rad^2) | 1 | 32440 | 965 | 1.7771 | 0.18550 |
| I(temp^2) | 1 | 33850 | 970 | 6.2447 | 0.01407 * |
| I(wind^2) | 1 | 37513 | 981 | 17.8501 | 5.246e-05 *** |
| temp:wind | 1 | 32930 | 967 | 3.3324 | 0.07088 . |
| temp:rad | 1 | 32613 | 966 | 2.3262 | 0.13034 |
| wind:rad | 1 | 32896 | 967 | 3.2224 | 0.07563 . |

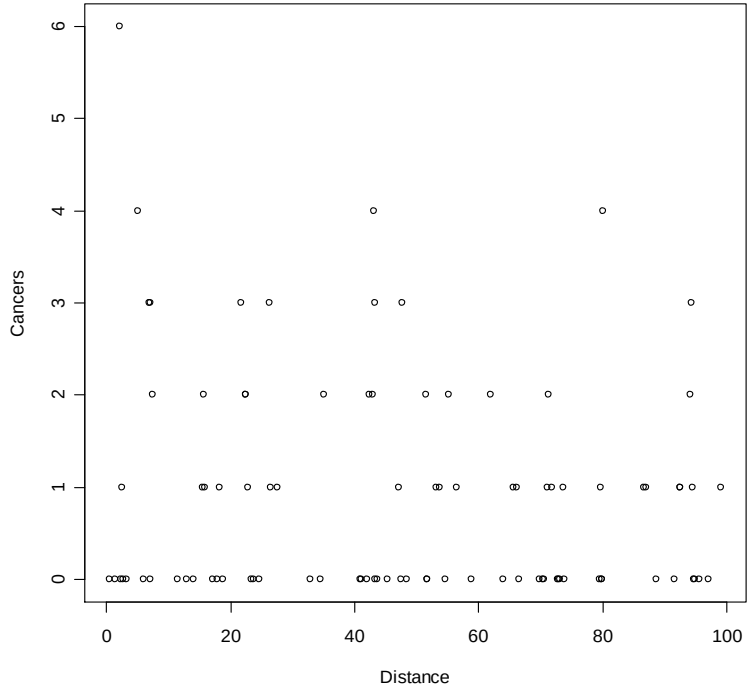
After a few rounds....

```
model6<-glm(ozone~wind+temp+rad+I(temp^2)+I(wind^2));drop1(model6,test="F")
```

| | Df | Deviance | AIC | F value | Pr(F) |
|---|---|---|---|---|---|
| <none> | | 34954 | 968 | | |
| wind | 1 | 46286 | 997 | 34.0384 | 6.054e-08 *** |
| temp | 1 | 36768 | 971 | 5.4471 | 0.021505 * |
| rad | 1 | 38547 | 976 | 10.7908 | 0.001387 ** |
| I(temp^2) | 1 | 37708 | 974 | 8.2718 | 0.004877 ** |
| I(wind^2) | 1 | 42056 | 986 | 21.3327 | 1.100e-05 *** |

## Count data

- never negative
- the variance increases with the mean
- the observations are integers



**General approach**

1. **Choose**

linear predictor $\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$

family of probability distributions: **Poisson**

*canonical* link function *g*: ***log*** function

The predicted mean of the Poisson response variable for observation *i* is $\mu_i$

$$\eta_i = h(\mu_i) = \log(\mu_i) \Leftrightarrow \mu_i = e^{\eta_i}$$

$\rightarrow \mu_i$ is always non-negative

2. **Fit** model to data using Maximum Likelihood / IWLS

```
model1<-glm(Cancers~Distance,poisson)
```

3. **Output**

```
summary(model1)
             Estimate  Std. Error       z value    Pr(>|z|)
(Intercept) 0.186865       0.188728 0.990       0.3221
Distance  -0.006138       0.003667 -1.674      0.0941 .

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 149.48  on 93  degrees of freedom
Residual deviance: 146.64  on 92  degrees of freedom

drop1(model1,test="Chisq")
Model:
Cancers ~ Distance
             Df          Deviance AIC LRT Pr(Chi)
             <none>    146.643   262.405
Distance  1      149.484   263.246   2.841       0.0919 .
```

# 4. **Check** model assumptions

For a Poisson distribution, the mean is equal to the variance.
Sometimes, the variance can be inflated relative to the mean, called **overdispersion**. It requires corrections to statistics in hypothesis testing.

```
model2<-glm(Cancers~Distance,quasipoisson);    summary(model2)
Coefficients:
             Estimate   Std. Error  t value    Pr(>|t|)
(Intercept)  0.186865   0.235341   0.794      0.429
Distance     -0.006138  0.004573   -1.342     0.183

(Dispersion parameter for quasipoisson family taken to be 1.554966)

    Null deviance: 149.48  on 93  degrees of freedom
Residual deviance: 146.64  on 92  degrees of freedom
AIC: NA

drop1(model1,test="F")
         Df      Deviance  AIC      F value   Pr(F)
         <none>  146.64    262.40
Distance 1       149.48    263.25   1.7823    0.1852
```

**Dealing with overdispersion:**

**Quasi-distributions**

Have an extra parameter in the variance function of the original distribution

Are not associated with a proper likelihood but a quasi-likelihood

**No AIC**

**No LRT (but there are statistics that can have chisq distributions)**

**Use Wald-type inference by means of *F* - tests**

We always check for overdispersion in the maximal model fitted:

**It will generally increase in simplified models**

In some cases, over/underdispersion cannot occur:

- **Bernoulli data 0/1**
- **when the largest model under consideration is equal to the saturated model.**

[There are probability distributions where over- and underdispersion are allowed and are represented by a parameter, e.g. beta-binomial, double binomial, Conway-Maxwell-Poisson]

**Log-linear analysis**:

```
data(Titanic)
margin.table(Titanic, c(2,4))
sex<-factor(c("M","M","F","F"))
survival<-factor(c("N","Y","N","Y"))
N<-c(1364,367,126,344)
drop1(glm(N~survival*sex,family=poisson),test="Chisq")
fisher.test(margin.table(Titanic, c(2,4)))
```

Variables have no clear distinction between independent and dependent
**Association** between variables is investigated

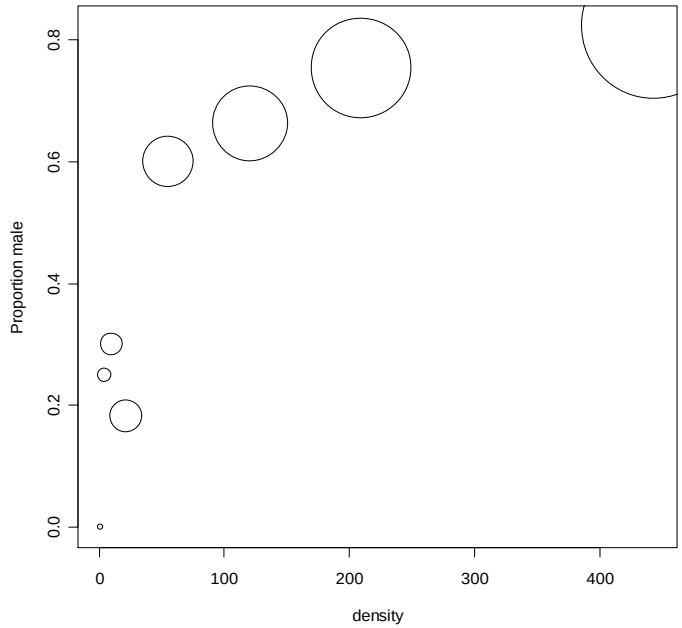
**Simpson's paradox** has to be avoided: ignoring an important covariate can flip conclusions

- The relevant test will always be on the significance of an **interaction**

- The minimal model will NEVER be the null model

## Proportion data

- proportions are between zero and one.
- Sample sizes on which proportions are calculated matter
- the variance is non-constant
- the observations are integers



**General approach**

1. **Choose**

linear predictor $\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$

family of probability distributions: **Binomial**

*canonical* link function *g*: ***logit*** function

The predicted mean of the Binomial response variable for observation *i* is $\mu_i$

$$\eta_i = h(\mu_i) = logit(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) \Leftrightarrow \mu_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$$

$\rightarrow \mu_i$ is always between zero and one.

2. **Fit:** fit model to data using Maximum Likelihood / IWLS

```
model1<-glm(y~density,quasibinomial)
```

3. **Output**

```
summary(model1)
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 0.0807368 | 0.2888702 | 0.279 | 0.7893 |
| density | 0.0035101 | 0.0009532 | 3.683 | 0.0103 * |

(Dispersion parameter for quasibinomial family taken to be 3.471613)

Null deviance: 71.159 on 7 degrees of freedom
Residual deviance: 22.091 on 6 degrees of freedom

```
drop1(model1,test="F")
```

Model:
y ~ density

| | Df | Deviance | F value | Pr(F) |
|---|---|---|---|---|
| <none> | | 22.091 | | |
| density | 1 | 71.159 | 13.327 | 0.01070 * |

References

http://www.stat.wisc.edu/courses/st849-bates/lectures/GLMH.pdf
http://www.stat.wisc.edu/courses/st849-bates/lectures/GLMDeviance.pdf

https://www.statistics.ma.tum.de/fileadmin/w00bdb/www/czado/lec2.pdf

Crawley M (2005) Statistics: An introduction using R

Faraway J (2006) Extending the Linear Model with R. CRC Press

Pawitan, Y (2001) In all Likelihood. Oxford University Press