The **binomTools** package: Performing model diagnostics on binomial regression models

Authors: Rune Haubo B Christensen and **Merete K Hansen**

DTU Informatics
Mathematical Statistics
Technical University of Denmark
mkh@imm.dtu.dk

August 18th 2011

Outline

Binary data

Binary data

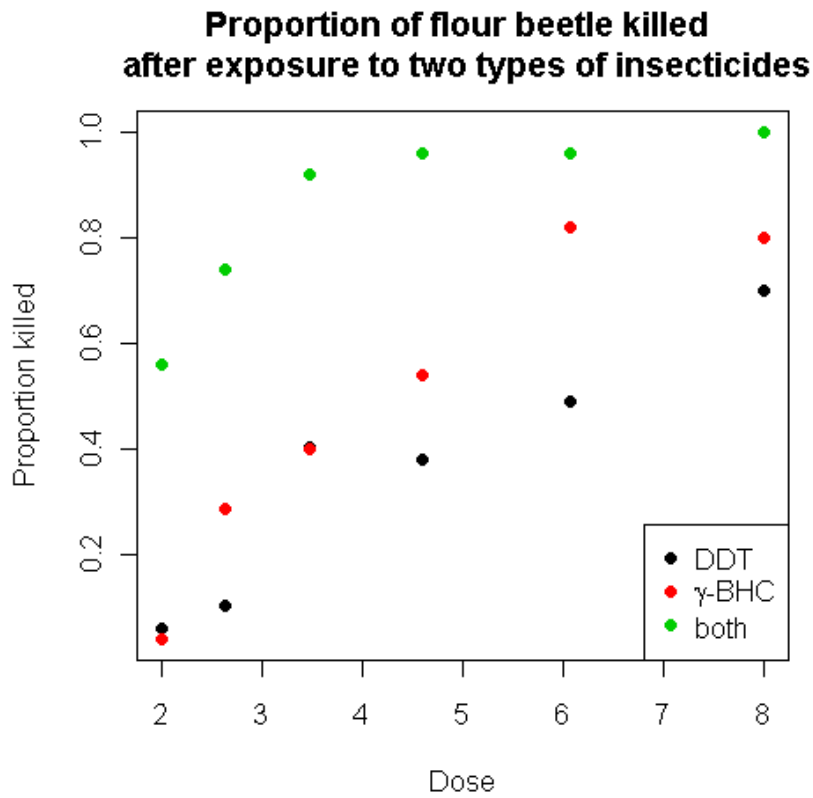
- dichotomous outcome
- yes/no, 0/1, success/failure, etc...
- e.g. $y_1 = 0, y_2 = 1, ..., y_n = 0$

Binomial data

- grouped binary data
- no. of successes / group size, e.g. $y_1 = 3/63, y_2 = 10/65, ..., y_n = 60/62$
- not possible to group binary data if all observations have distinct covariance structures

Example: Flour beetle mortality data

```
> head(flourbeetles, n=10)
type    dose    y   n
DDT     2.00    3  50
DDT     2.64    5  49
DDT     3.48   19  47
DDT     4.59   19  50
DDT     6.06   24  49
DDT     8.00   35  50
g-BHC   2.00    2  50
g-BHC   2.64   14  49
g-BHC   3.48   20  50
g-BHC   4.59   27  50
```



Fit model

Aim: model proportion of beetles dead after exposure

$$p_i = y_i/n_i$$

How:

- We fit a generalized linear model with a binomial family:

$$g(p_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki},$$

  where $g(\cdot)$ is the link function
- Logistic regression model: special case with link function $g(p_i) = \text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$
- Binomial regression model: Various link functions

Fit model in R

```
> beetles.glm <- glm(cbind(y, n-y) ~ type + log(dose),
+                    family=binomial, data=beetles)
> summary(beetles.glm)
...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.5553     0.3611 -12.613  < 2e-16 ***
typeboth      3.1305     0.2522  12.413  < 2e-16 ***
typeg-BHC     0.7128     0.1981   3.598  0.00032 ***
log(dose)     2.6958     0.2157  12.498  < 2e-16 ***
...
    Null deviance: 413.648  on 17  degrees of freedom
Residual deviance:  21.282  on 14  degrees of freedom
AIC: 92.753
...
```

Diagnostics

- Model building: iterative process of alternately model fitting and model checking
- Model inadequacy comes in several forms
  - Incorrect specification of linear predictor
  - Incorrect specification of link function
  - Discrepant observations, termed *outliers*
  - Distributional assumptions violated
- **Aim** of **binomTools:** a toolbox of diagnostic methods for binomial regression models

Existing implementations

Main functionality in R

- Various **residual types** with `residuals`, `rstandard` and `rstudent`
- Some **residual plots** with `plot(object.glm)` and `glm.diag.plots` from the **boot** package
- **Leverage** and **influence measures**, such as dfbeta, dfbetas, Cooks's distance with `influence.measures`
- **Half-normal plot** without envelopes in package **faraway** et al.
- `binom.diagnostics` in the **MLDS** package
- **car** package: A comprehensive body of **diagnostic plots** useful for examining various forms of model inadequacy
- Other implementations that (to our knowledge) only occurs sporadically

Residuals in R

- Three different methods for extraction of residuals
  - `residuals` extracts **unstandardized** deviance, Pearson, working, response and partial residuals
  - `rstandard` extracts **standardized** deviance and Pearson residuals
  - `rstudent` extracts **studentized** residuals
- Confusion terminology

It goes by many names...

A quick litterature search reveals

- Standardized Pearson residuals also called
  - studentized Pearson residuals
  - standardized residuals
  - studentized residuals
  - internally studentized residuals
- Studentized residuals
  - likelihood residuals
  - externally studentized residuals
  - deleted studentized residuals
  - jack-knife residuals

No exact definitions in the residual help files

Residuals.glm in binomTools

Method to extract residuals from a binomial regression model

```
Residuals(object, type = c("approx.deletion",
        "exact.deletion", "standard.deviance",
        "standard.pearson", "deviance", "pearson",
        "working", "response", "partial"))
```

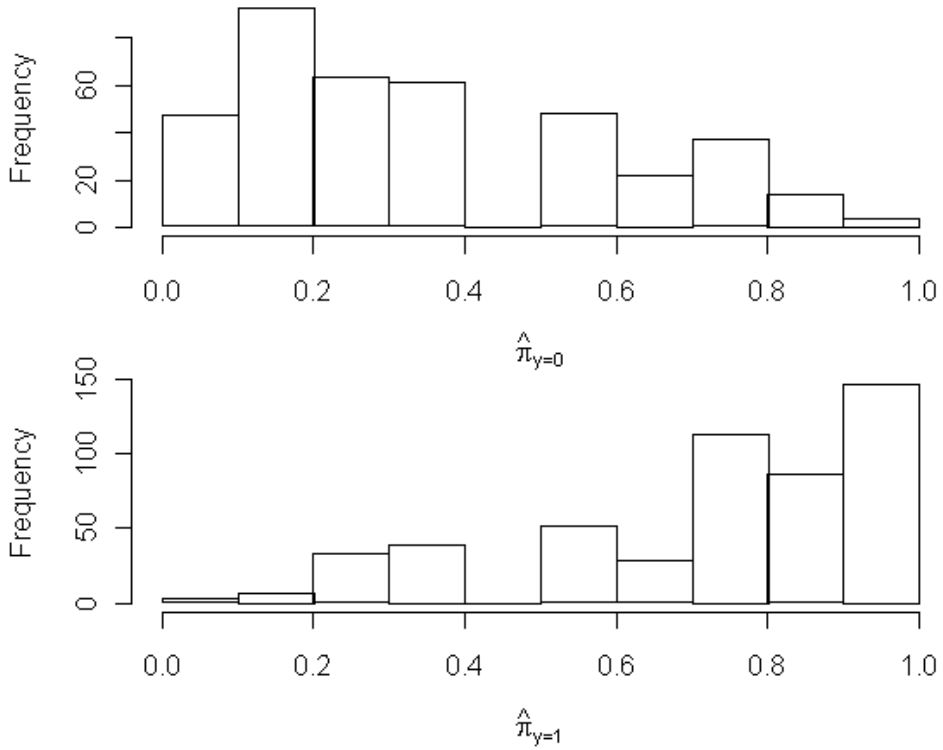
- `approx.deletion` extracted with `rstudent`
- `exact.deletion` (new function)
- `standard.deviance` extracted with `rstandard`
- `standard.pearson` extracted with `rstandard`
- remainder extracted with `residual`

Aim: Uniform syntax, enhance transparency of residual types and improve help pages with formulas

Exact deletion residuals

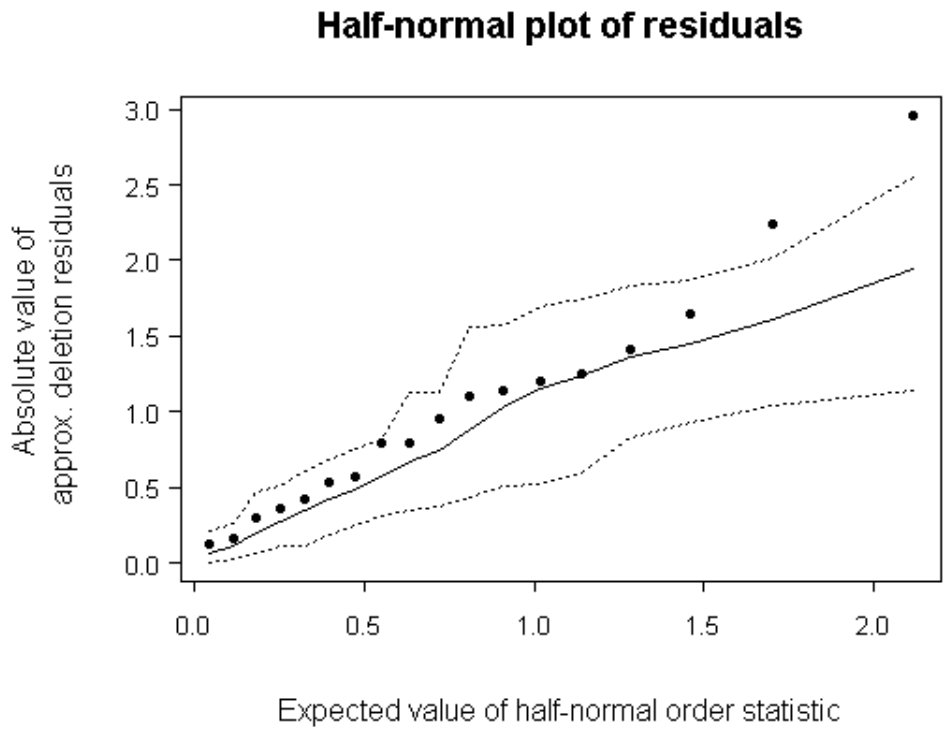
- New type of residual implemented in **binomTools**
- `approx.deletion` (`rstudent`) residuals are approximations to deletion (studentized) residuals
- `exact.deletion` are exact deletion (studentized) residuals
- Change in deviance when one observation in turn is deleted from the data
- May be computationally heavy for large data sets

Parallel histograms



- Explorative version of Hosmer-Lemeshow goodness-of-fit test (with fixed cutpoints)
- Related to confusion table
- Empicirical cumulative distribution function (ecdf) curves and empirical ROC curve also available

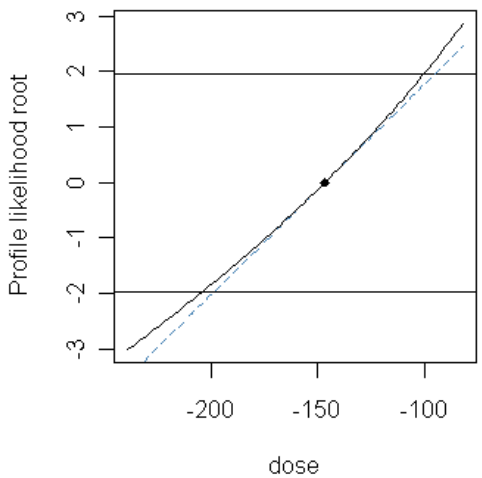
Half-normal plot

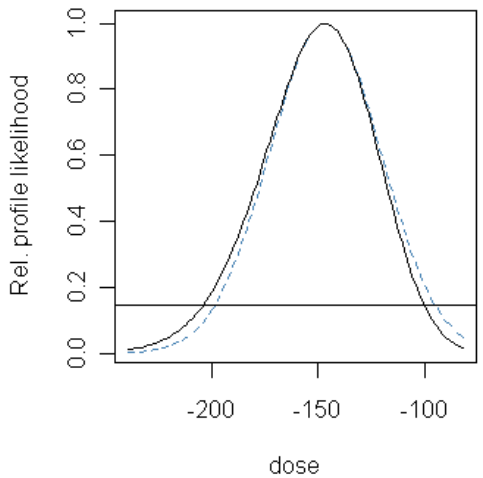


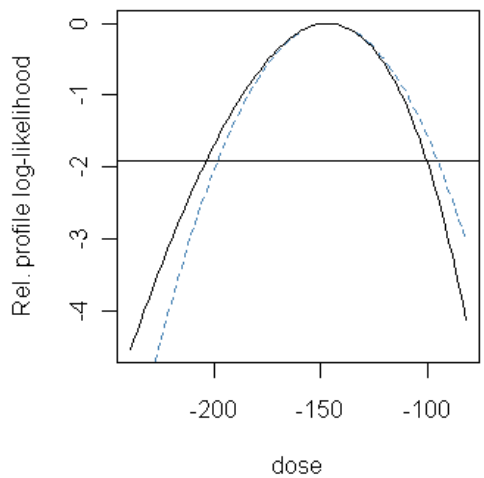
- Half-normal plot uses absolute residual values but otherwise equivalent to a normal plot
- Optional simulated envelopes to support interpretation

Profile likelihood

- Possible to assess the profile likelihood with `profile` from the **MASS** package
- Returns and plot the profile likelihood root - not the profile likelihood
- New `plot` method in **binomTools** with enhanced plot functionality (examples shown for another data set)







Miscellaneous

- Possibility to `group` binary or not completely grouped data data based on a specified covariate structure
- Goodness-of-fit tests `HLtest` and `X2GOFtest`
- Implementation of `Rsq` - a newly proposed R-square
- Empirical logit transform `empLogit` useful when at least one observation is zero or one

Future implementations in binomTools

- Enhance functionality of existing implementations
- `ungroup` data from binomial to binary form
- Empirical area under the ROC curve
- Add a generalized link function with some standard link functions as special cases. Facilitates assessment of proper specification of the link function
- Other ideas are welcome

Acknowledgments

Thank you for listening

References

References

- Atkinson A.C. (1981). Two graphical displays for outlying and influential observations in regression. *Biometrika*, **68**, 13-20.
- Collett D. (2003). *Modelling binary data*. Second edition. Chapman & Hall/CRC
- Fox J. and Weisberg S. (2011). *An R Companion to Applied Regression*. Second Edition. Sage Publications.
- Hosmer D.W. and Lemeshow S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics - Theory and Methods*, **A9**(10), 1043-1069.
- Pawitan Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press.
- Tjur T. (2009). Coefficients of determination in logistic regression models - a new proposal: The coefficient of discrimination. *The American Statistician*, **63**(4), 366-372
- Venables W.N. and Ripley B.D. (2002). *Modern Applied Statistics with S*. Fourth Edition. Springer
- Williams D.A. (1987). Generalized linear model diagnostics using the deviance and single case deletions. *Applied Statistics* **36**, 181-191