“Pseudo-binomial data analysis”

Jonàs Oliva Palau

Ulf Olsson

Binomial models are often presented to programs as

$$y/n = \text{(some linear model)}$$

y: number of successes

n: number of trials

We assume a binomial distribution and a logit link.

However, for some types of “percentage data” we do not have a proper n.

Examples

Percentage of a tree stump covered by…

Percentage of a leaf area covered by…

Percentage of a plot of land covered by…

These are not binomial. Or…?

Nelder and McCullagh (1989),

(based on data from Wedderburn, 1974)

Data of this type can be analyzed as “Pseudo-binomial”.

That is, the observed proportions *p* are modeled as binomial with mean

$$E(p)=\mu$$

And variance

$$Var(p)=\phi\mu(1-\mu)$$

i.e. the variance is changed by an over-dispersion parameter ϕ.

Alternative if this does not work:

Use the variance function

$$Var(p)=\mu^2(1-\mu)^2$$

Example data: Oliva (2014)

9 locations

Two-factor experiments at each location

Protection	2 levels: Rotstop or None.
Ha_species	3 levels: P S or None

Three replicates at each site (but with exceptions…)

Response: Percentage of the area of tree stumps covered by fungus.

The percentage was assessed from digital pictures (“pixel count”)

SAS implementation

An over-dispersion parameter is included in Proc GLIMMIX by adding the statement

RANDOM _residual_;

To the program. Example program:

```
PROC GLIMMIX data=work.Pg_field ;
CLASS Location Protection Ha_species;
MODEL Ha_total = protection|Ha_species@2
   / link=logit dist=bin ddfm=KR;
RANDOM _residual_;
RANDOM Location;
run;
```

The alternative variance function can be used by adding the line

```
_variance_ = _mu_**2 * (1-_mu_)**2;
```

Transformations?

As another alternative, we can use a Normal theory Mixed models on transformed response.

Square root: $$y = \sqrt{p}$$

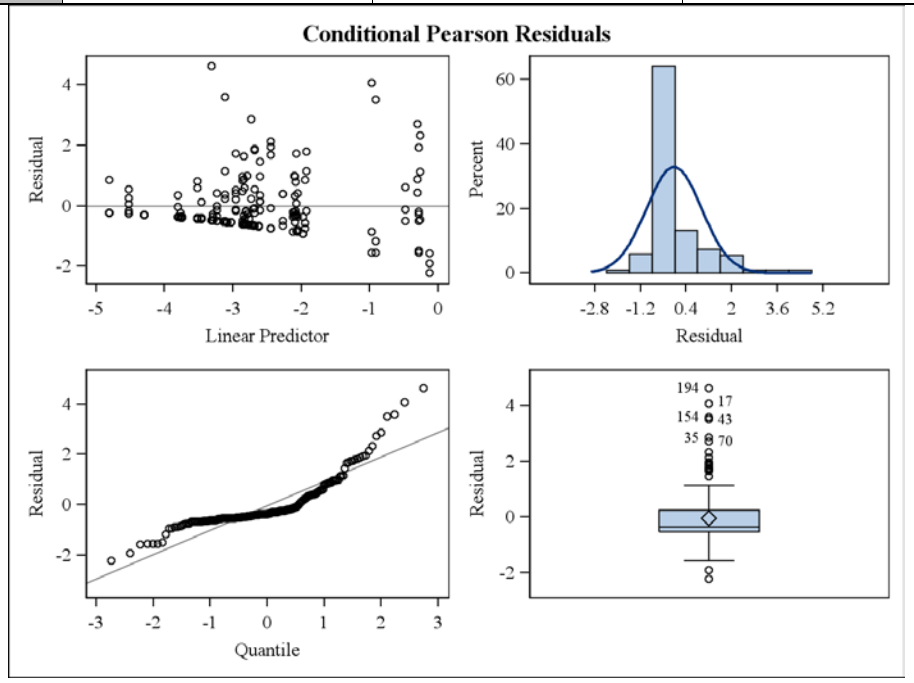
Arcsine: A classic suggestion for binomial data (Snedecor and Cochran, 1960):

$$y = arcsin\left(\sqrt{p}\right)$$

Problem: With transformed data, confidence intervals for p may fall above 1 or below 0.

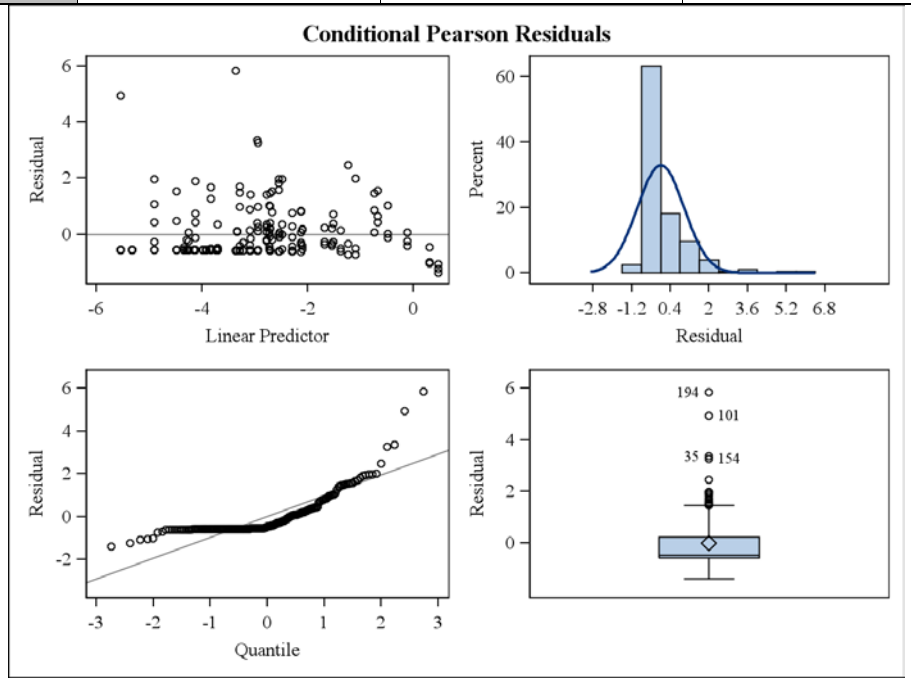
Model 1: “Pseudo binomial”

Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Protection	1	188.3	33.34	<.0001
Ha_species	2	191.4	8.59	0.0003
Protectio*Ha_species	2	187.5	7.69	0.0006



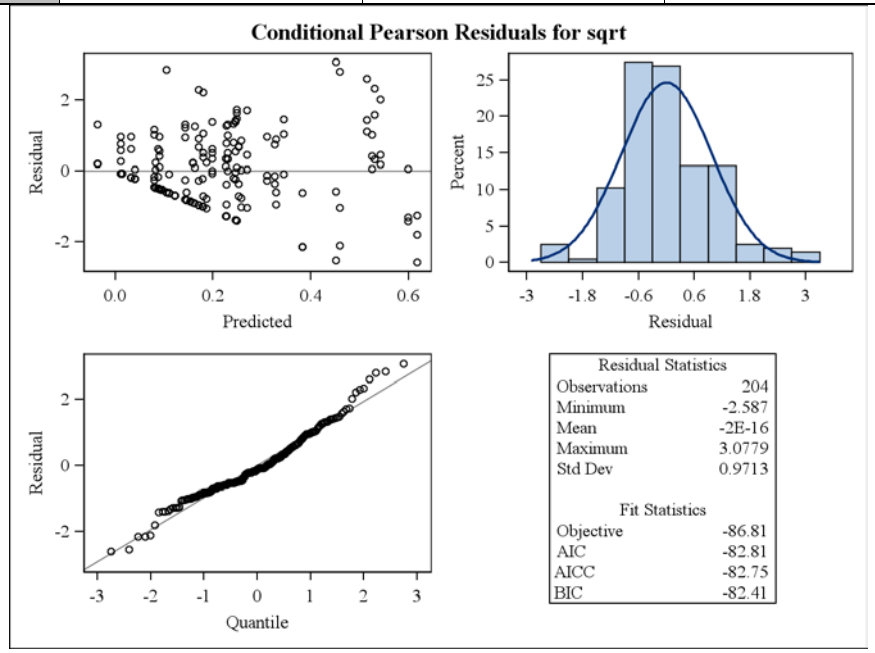
Model 2: Pseudo binomial, other variance function

Type III Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Protection	1	191.4	39.02	<.0001
Ha_species	2	193.4	10.50	<.0001
Protectio*Ha_species	2	190.7	3.19	0.0436



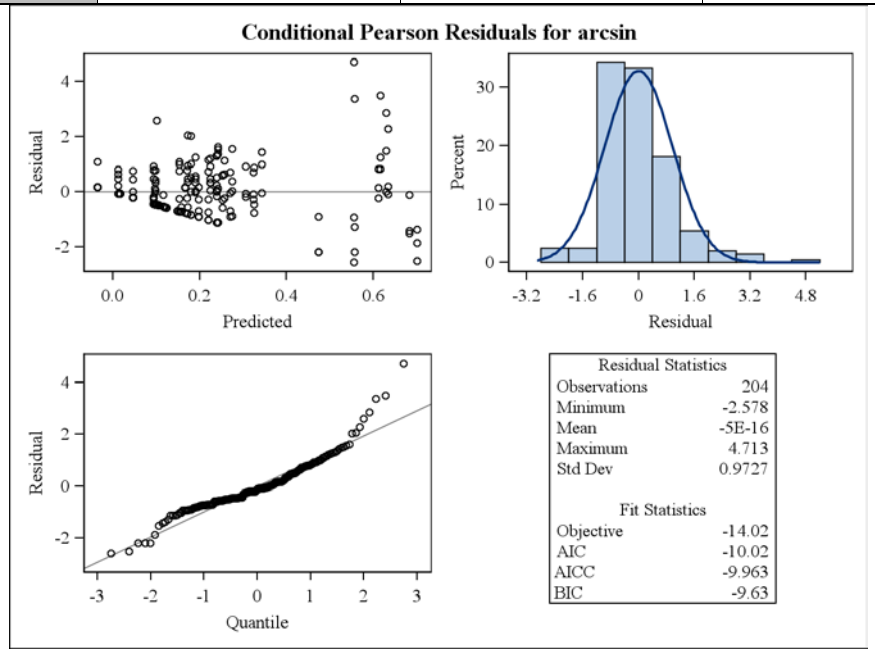
Model 3: Proc Mixed , square root transformed

Type 3 Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Protection	1	190	56.77	<.0001
Ha_species	2	190	16.26	<.0001
Protectio*Ha_species	2	190	11.31	<.0001



Model 4: Proc Mixed, arcsine-transformed

Type 3 Tests of Fixed Effects				
Effect	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Protection	1	190	52.50	<.0001
Ha_species	2	190	18.09	<.0001
Protectio*Ha_species	2	190	13.70	<.0001



Conclusion

Observations in the form of Percentages are common in biological research

The distribution of percentages (other than binomial ones) is unknown

However, the “pseudo binomial” approach is an interesting option for analyzing such data

References

Littell, R., Milliken, G., Stroup, W. Wolfinger, R. and Schabenberger O. (2006): SAS for mixed models, second ed. Cary, N. C., SAS Institute Inc.

McCullagh, P. and Nelder, J.A. (1989): Generalized linear models. London, Chapman and Hall.

Olsson, Ulf (2002): Generalized linear models: an applied approach. Lund, Studentlitteratur.

SAS Institute Inc. (2011): SAS/Stat 9.3 user's guide. Cary, N.C., SAS Institute Inc.

http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_glimmix_sect016.htm

Snedecor, G, W. and Cochran, W. G. (1960): Statistical methods. Ames, Iowa State University Press.