Generalized Linear Models and Extensions

Clarice Garcia Borges Demétrio

ESALQ/USP

Piracicaba, SP, Brasil

March 2013

email: Clarice.demetrio@usp.br

Course Outline

Session 1 - Generalized linear models

- Introduction
- Motivating examples
- History
- Generalized linear models
- Definition of generalized linear models
- Model fitting
- Inferential aspects

Session 2: Normal models

- Summary
- Examples
- Residual analysis and diagnostics
- Box-Cox transformation
- Transform or link

Session 3: Binary and binomial data

- Summary – Binomial models

- Analysis of dose-response models
- Examples
- Residuals for glm's

Session 4: Poisson and multinomial data

- Summary – Poisson models
- Example
- Dilution assays
- 2-way contingence tables
- Simple 2-way table
- Binomial logit and Poisson log-linear models
- Multinomial response data

Session 5: Overdispersion

- Overdispersion in glm's: causes and consequences; examples
- Overdispersion models:
  - mean-variance models
  - two-stage models
- Estimation methods
- Examples
- Extended overdispersion models

Introduction

- Agricultural Science - diferent types of data: continuous and discrete.
- Model selection - important part of the research: search for a simple model which explains well the data (Parsimony).
- All models envolve:
  - a systematic component - regression, analysis of variance, analysis of covariance;
  - a random component - distributions;
  - a link between systematic and random components.

Motivating examples

Melon organogenesis

	Eldorado				AF-522			
Replicates	0.0	0.1	0.5	1.0	0.0	0.1	0.5	1.0
1	0	0	7	8	0	0	4	7
2	0	2	8	8	0	2	7	8
3	0	0	8	8	0	0	7	8
4	0	1	5	8	0	1	8	8
5	0	0	7	5	0	1	8	7

Considerations

- Response variable: Y – number of explants (cuts of cotyledon) regenerated out of $m = 8$ explants.
- Distribution: Binomial.
- Systematic component: factorial 2×4 (2 varieties, 4 concentrations of BAP(mg/l)), completely randomized tissue culture experiment.
- Aim: to see how organogenesis is affected by variety and concentration of BAP.

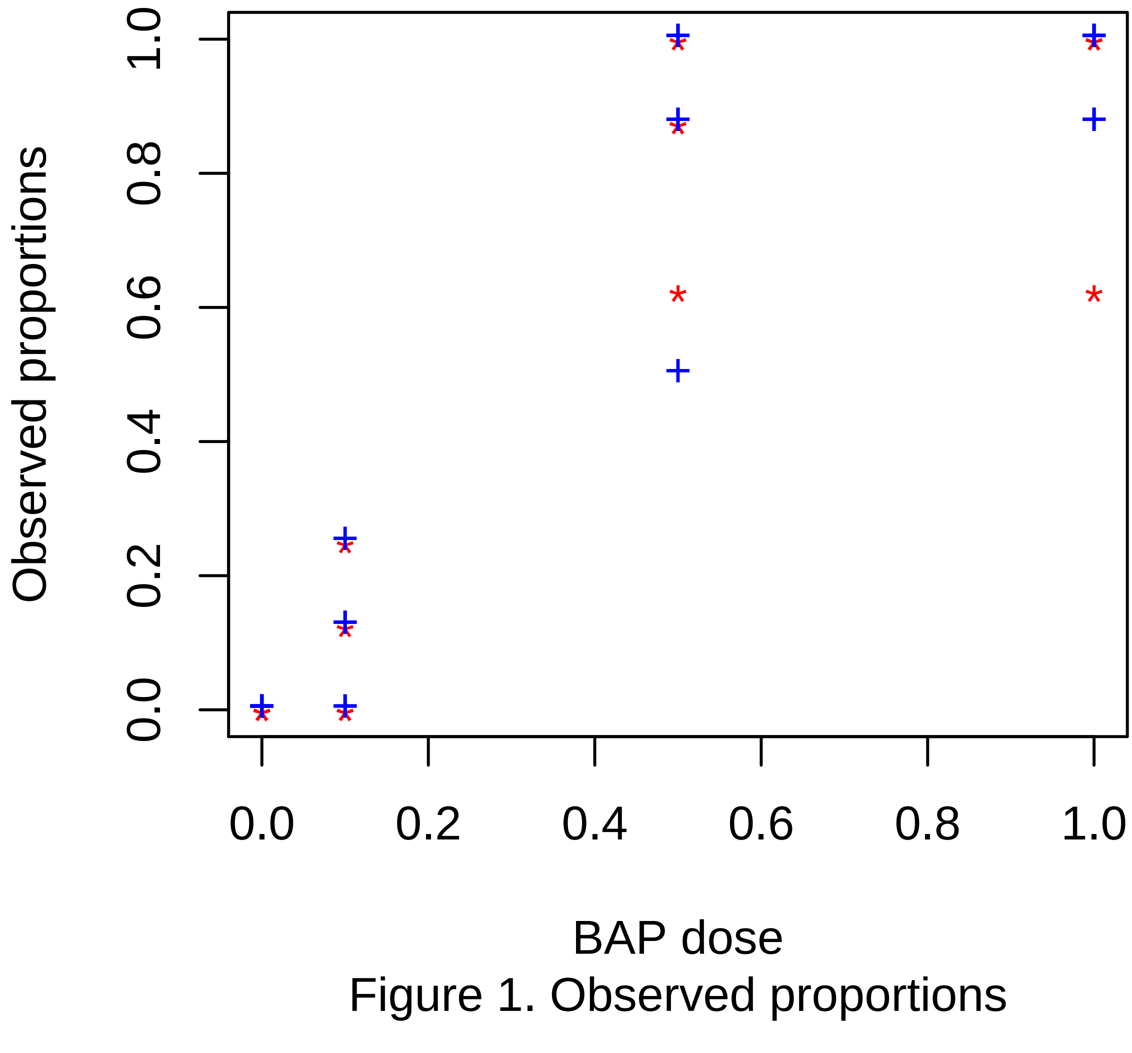

Figure 1: Melon organogenesis. Scatterplot

Carnation meristem culture

0,0			0,1			0,3			0,5			1,0			2,0		
b	c	v	b	c	v	b	c	v	b	c	v	b	c	v	b	c	v
1	2.5	0	3	5.5	1	5	4.8	1	9	2.8	0	10	2.0	1	12	1.7	1
2	2.5	0	2	4.3	1	5	3.0	1	10	2.3	1	8	2.3	1	15	2.5	1
1	3.0	0	6	3.3	0	4	2.7	0	8	2.7	1	12	2.0	1	15	2.3	1
2	2.5	1	3	4.3	0	4	3.1	1	11	3.2	0	13	1.0	1	12	1.5	1
1	4.0	0	4	5.4	0	5	2.9	0	8	2.9	1	14	2.8	1	13	1.7	1
1	4.0	0	3	3.8	1	6	3.3	1	8	1.5	1	14	2.0	1	16	2.0	1
2	3.0	0	3	4.3	1	6	2.1	1	8	2.5	0	14	2.7	1	17	1.7	1
1	3.0	0	4	6.0	1	5	3.7	1	8	2.8	0	9	1.8	1	15	2.0	1
1	5.0	0	3	5.0	1	4	3.8	1	8	1.8	1	13	1.8	1	17	2.0	1
1	4.0	0	2	5.0	0	5	3.8	1	11	2.0	0	9	2.1	1	14	2.3	1
1	2.0	1	3	4.5	0	6	3.3	0	9	2.7	1	15	1.3	1	16	2.5	1
1	4.0	0	3	4.0	1	6	2.6	1	12	1.8	1	15	1.2	1	21	1.3	1
2	3.0	0	4	3.3	0	5	2.3	0	12	2.3	1	16	1.2	1	18	1.3	1
2	3.5	1	3	4.3	1	4	3.6	1	10	1.5	1	9	1.0	1	16	1.8	1
1	3.0	0	3	4.5	1	3	4.8	1	10	1.5	1	13	1.7	1	18	1.0	1
2	3.0	0	2	3.8	0	4	2.0	0	7	1.0	1	14	1.7	1	20	1.3	1
2	5.5	0	3	4.7	1	6	1.7	0	8	3.0	1	16	1.3	1	22	1.5	1
1	3.0	0	4	2.2	0	5	2.5	0	12	2.0	1	13	1.8	1	20	1.3	1
1	2.5	0	2	3.8	1	5	2.0	0	9	3.0	1						
1	2.0	0	3	5.0	0	5	2.0	0									

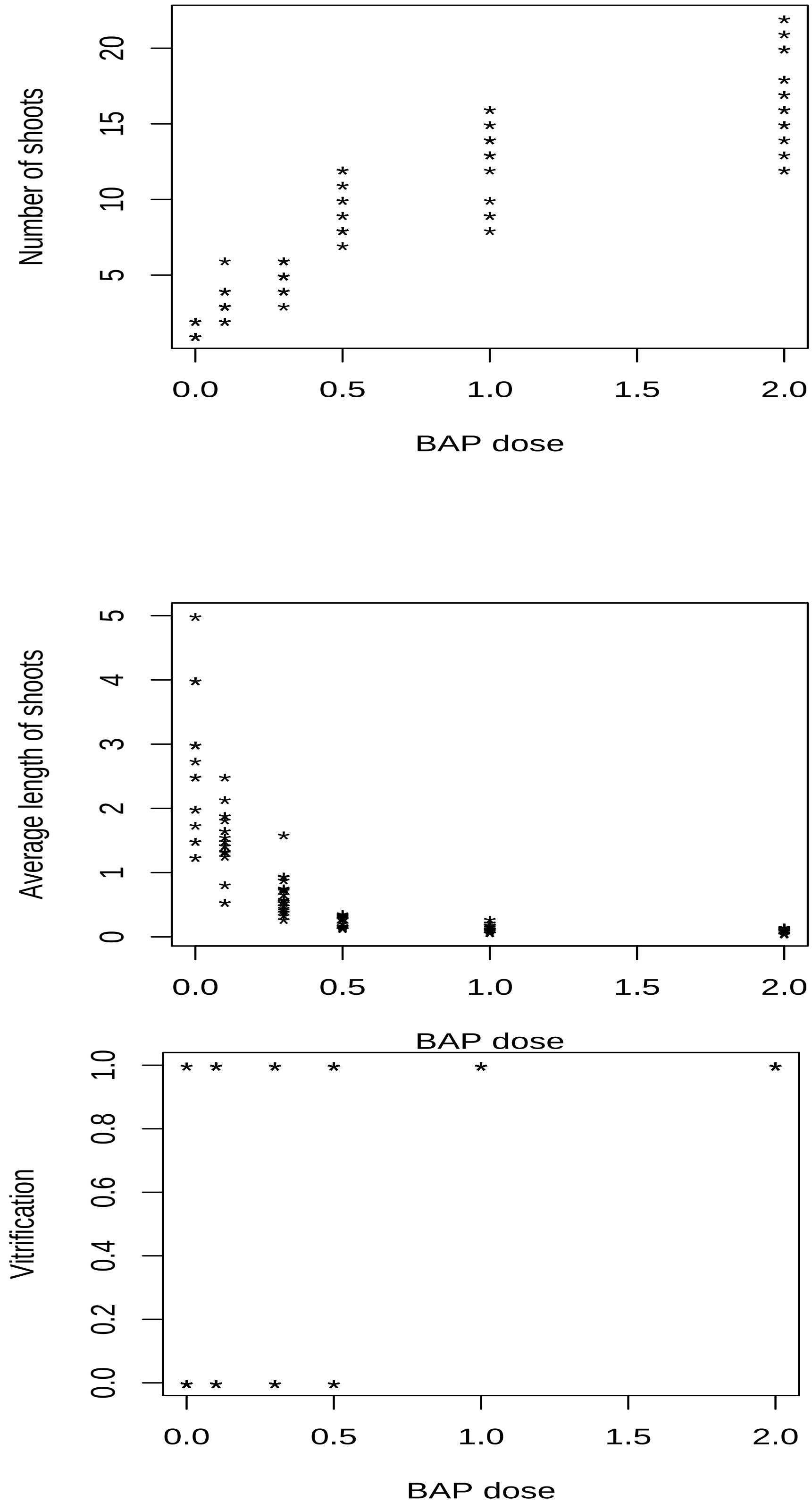

Figure 2: Carnation meristem culture. Scatterplots

Rotenon toxicity

Dose (d_i)	m_i	y_i
0.0	49	0
2.6	50	6
3.8	48	16
5.1	46	24
7.7	49	42
10.2	50	44

- Response variable: Y_i – number of dead insects out of m_i insects (Martin, 1942).
- Distribution: Binomial.
- Systematic component: regression model, completely randomized experiment.
- Aim: Lethal doses.

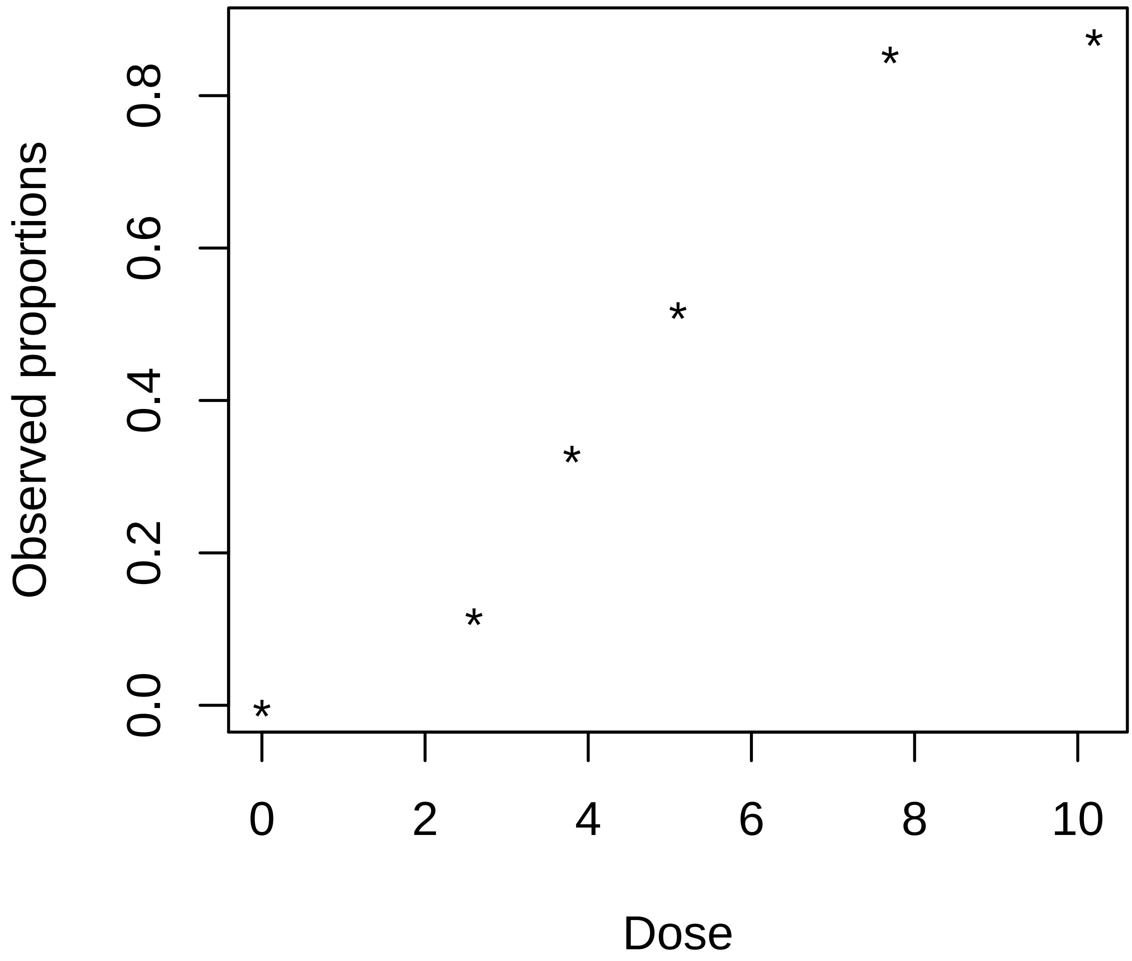

Figure 3: Rotenon - Scatterplot

Germination of Orobanche seed

O. aegyptiaca 75		*O. aegyptiaca 73*	
Bean	Cucumber	Bean	Cucumber
10/39	5/6	8/16	3/12
23/62	53/74	10/30	22/41
23/81	55/72	8/28	15/30
26/51	32/51	23/45	32/51
17/39	46/79	0/4	3/7
	10/13		

Considerations

- Response variable: Y_i – number of germinated seeds out of m_i seeds (Crowder, 1978).
- Distribution: Binomial.
- Systematic component: factorial 2×2 (2 species, 2 extracts), completely randomized experiment.
- Aim: to see how germination is affected by species and extracts.
- Problem: overdispersion.

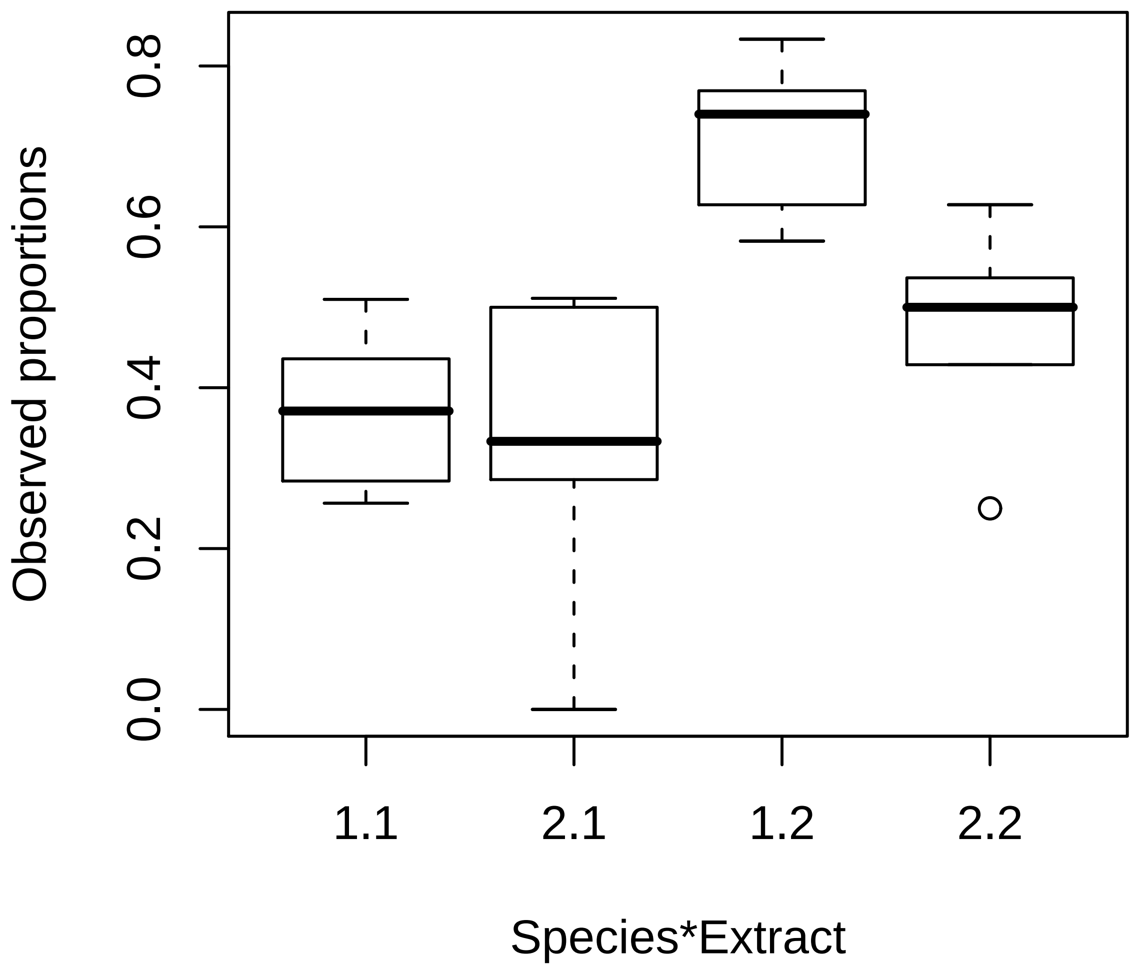

Figure 4: Orobanche - Boxplot

Apple tissue culture

- 4x2 factorial micropropagation experiment of the apple variety Trajan – a 'columnar' variety.
- Shoot tips of length 1.0-1.5 cm were placed in jars on a standard culture medium.
- 4 concentrations of cytokinin BAP added

  High concentrations of BAP often inhibit root formation during micropropagation of apples, but maybe not for 'columnar' varieties.

- Two growth cabinets, one with 8 hour photoperiod, the other with 16 hour.

  Jars placed at random in one of the two cabinets

- Response variable: number of roots after 4 weeks culture at 22°C.

	Photoperiod							
	8				16			
BAP (μM)	2.2	4.4	8.8	17.6	2.2	4.4	8.8	17.6
No. of roots								
0	0	0	0	2	**15**	**16**	**12**	**19**
1	3	0	0	0	0	2	3	2
2	2	3	1	0	2	1	2	2
3	3	0	2	2	2	1	1	4
4	6	1	4	2	1	2	2	3
5	3	0	4	5	2	1	2	1
6	2	3	4	5	1	2	3	4
7	2	7	4	4	0	0	1	3
8	3	3	7	8	1	1	0	0
9	1	5	5	3	3	0	2	2
10	2	3	4	4	1	3	0	0
11	1	4	1	4	1	0	1	0
12	0	0	2	0	1	1	1	0
>12	13,17	13	14,14	14				
No. of shoots	30	30	40	40	30	30	30	40
Mean	5.8	7.8	7.5	7.2	3.3	2.7	3.1	2.5
Variance	14.1	7.6	8.5	8.8	16.6	14.8	13.5	8.5
Overdispersion index	1.42	-0.03	0.13	0.22	**4.06**	**4.40**	**3.31**	**2.47**

Considerations about the data

- Many zeros for 16 hour photoperiod
- Overdispersion for 16 hour photoperiod
  Is this caused by excess zeros?
- Not much overdispersion for the 8 hour photoperiod.
  mean$\approx$variance for concentrations 1, 2 and 4 of BAP.
- For the 8 hour photoperiod the lowest concentration has smallest mean and largest variance
- For the 16 hour photoperiod the conclusion is not so clear cut.

History

The developments leading to the general overview of statistical modelling, known as generalized linear models, extend over more than a century. This history can be traced very briefly as follows (McCullagh & Nelder, 1989, Lindsey, 1997):

- multiple linear regression – a normal distribution with the identity link, $\mu_i = \boldsymbol{\beta}'\mathbf{x}_i$ (Legendre, Gauss, early XIX-th century);
- analysis of variance (ANOVA) designed experiments – a normal distribution with the identity link, $\mu_i = \boldsymbol{\beta}'\mathbf{x}_i$ (Fisher, 1920 to 1935);
- likelihood function – a general approach to inference about any statistical model (Fisher, 1922);
- dilution assays – a binomial distribution with the complementary log-log link, $\log[-\log(1 - \mu_i/m_i)] = \boldsymbol{\beta}'\mathbf{x}_i$ (Fisher, 1922);
- exponential family – a class of distributions

with suficient statistics for the parameters (Fisher, 1934);

- probit analysis – a binomial distribution with the probit link, $\Phi^{-1}(\mu_i/m_i) = \boldsymbol{\beta}'\mathbf{x}_i$ (Bliss, 1935);
- logit for proportions – a binomial distribution with the logit link, $\log \frac{\mu_i}{m_i-\mu_i} = \boldsymbol{\beta}'\mathbf{x}_i$ (Berkson, 1944, Dyke & Patterson, 1952);
- item analysis – a Bernoulli distribution with the logit link, $\log \frac{\mu_i}{1-\mu_i} = \boldsymbol{\beta}'\mathbf{x}_i$ (Rasch, 1960);
- log linear models for counts – a Poisson distribution with the log link, $\log \mu_i = \boldsymbol{\beta}'\mathbf{x}_i$ (Birch, 1963);
- regression models for survival data - – an exponential distribution with the reciprocal or the log link, $\frac{1}{\mu_i} = \boldsymbol{\beta}'\mathbf{x}_i$ or $\log \mu_i = \boldsymbol{\beta}'\mathbf{x}_i$ (Feigl & Zelen, 1965, Zippin & Armitage, 1966, Gasser, 1967);
- inverse polynomials – a gamma distribution with the reciprocal link, $\frac{1}{\mu_i} = \boldsymbol{\beta}'\mathbf{x}_i$ (Nelder, 1966).

Generalized Linear Models (glms)

Unifying framework for much statistical modelling.

First introduced by Nelder & Wedderburn (1972) as an extension to the standard normal theory linear model.

- single response variable Y
- explanatory variables $x_1, x_2, \ldots, x_p$, $(x_1 \equiv 1)$
- random sample: n observations $(y_i, \mathbf{x}_i)$, where $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{pi})^T$

For more details see, for example:

- McCullagh & Nelder (1989) – theory, applications
- Dobson (2002) – a simple introduction.
- Aitkin *et al* (2009) – practical application of glms using R

Definition of glm

Three components of a generalized linear model are:

- independent random variables Y_i, $i = 1, \ldots, n$, from a linear exponential family distribution with means μ_i and constant scale parameter ϕ,

  $$f(y) = \exp\left\{ \frac{y\theta - b(\theta)}{\phi} + c(y, \phi) \right\}$$

  where $\mu = \mathbf{E}[Y] = b'(\theta)$ and $\mathrm{Var}(Y) = \phi b''(\theta)$.

- a linear predictor vector $\boldsymbol{\eta}$ given by

  $$\boldsymbol{\eta} = X\boldsymbol{\beta}$$

  where $\boldsymbol{\beta}$ is a vector of p unknown parameters and $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^T$ is the $n \times p$ design matrix;

- a link function $g(\cdot)$ relating the mean to the linear predictor, i.e.

  $$g(\mu_i) = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$$

Table 1: Identifiers for exponencial family distributions

Distribution	$a(\phi)$	θ	$b(\theta)$	$c(y;\phi)$	$\mu(\theta)$	$V(\mu)$
$\mathsf{N}(\mu,\sigma^2)$	σ^2	μ	$\frac{\theta^2}{2}$	$-\frac{1}{2}\left[\frac{y^2}{\sigma^2}+\log\,(2\pi\sigma^2)\right]$	θ	1
$\mathsf{P}(\mu)$	1	$\log\,\mu$	e^{θ}	$-\log\,y!$	e^{θ}	μ
$\mathsf{B}(m,\pi)$	1	$\log\left(\frac{\pi}{1-\pi}\right)$	$m\,\log\,(1+e^{\theta})$	$\log\binom{m}{my}$	$\frac{e^{\theta}}{1+e^{\theta}}$	$\frac{1}{m}\mu(m-\mu)$
$\mathsf{NB}(k)$	1	$\log\left(\frac{\mu}{\mu+k}\right)$	$-k\,\log\,(1-e^{\theta})$	$\log\left[\frac{\Gamma(k+y)}{\Gamma(k)\,y!}\right]$	$k\,\frac{e^{\theta}}{1-e^{\theta}}$	$\mu\left(\frac{\mu}{k}+1\right)$
$\mathsf{G}(\mu,\nu)$	ν^{-1}	$-\frac{1}{\mu}$	$-\log\,(-\theta)$	$\nu\,\log\,(\nu y)-\log\,y-\log\,\Gamma(\nu)$	$-\frac{1}{\theta}$	μ^2
$\mathsf{IG}(\mu,\sigma^2)$	σ^2	$-\frac{1}{2\mu^2}$	$-(-2\theta)^{\frac{1}{2}}$	$-\frac{1}{2}\left[\log\,(2\pi\sigma^2y^3)+\frac{1}{\sigma^2 y}\right]$	$(-2\theta)^{-\frac{1}{2}}$	μ^3

Normal Models

Continuous response variable – Y
Normal distribution, constant variance

$$\boxed{\begin{gathered} Y_i \sim N(\mu_i, \sigma^2), \quad i = 1, \ldots, n \\ \mu_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} = \boldsymbol{\beta}^T \mathbf{x}_i \end{gathered}}$$

- Regression models
  continuous explanatory variables
  – fitting, testing, model checking
- Analysis of variance
  categorical explanatory variables
  – *ANOVA* - balanced designs
  – *regression* - general unbalanced designs
- Analysis of covariance
  mixture of continuous and categorical explanatory variables

Binomial regression models

Y_i counts of successes out of samples of size m_i, $i = 1, \ldots, n$.
Writing

$$\mathsf{E}[Y_i] = \mu_i = m_i \pi_i,$$

a glm models the expected proportions π_i in terms of explanatory variables $\mathbf{x}_i$

$$g(\pi_i) = \boldsymbol{\beta}' \mathbf{x}_i,$$

For $Y_i \sim \mathsf{Bin}(m_i, \pi_i)$ the variance function is

$$\mathsf{Var}(Y_i) = m_i \pi_i (1 - \pi_i).$$

the canonical link function is the logit

$$g(\mu_i) = \log\left(\frac{\mu_i}{m_i - \mu_i}\right) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \eta_i$$

Other common choices are

- probit $g(\mu_i) = \Phi^{-1}(\mu_i / m_i) = \Phi^{-1}(\pi_i)$
- complementary log-log (CLL) link

$$g(\mu_i) = \log\{-\log(1 - \pi_i)\}.$$

Poisson regression models

If Y_i, $i = 1, \ldots, n$, are counts with means μ_i, the standard Poisson model assumes that $Y_i \sim \text{Pois}(\mu_i)$ with variance function

$$\text{Var}(Y_i) = \mu_i.$$

The canonical link function is the log

$$g(\mu_i) = \log(\mu_i) = \eta_i,$$

For different observation periods/areas/volumes:

$$Y_i \sim \text{Pois}(t_i \lambda_i)$$

Taking a log-linear model for the rates,

$$\log(\lambda_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$

results in the following log-linear model for the Poisson means

$$\log(\mu_i) = \log(t_i \lambda_i) = \log(t_i) + \mathbf{x}_i^T \boldsymbol{\beta},$$

where the $\log(t_i)$ is included as a fixed term, or *offset*, in the model.

Estimation and model fitting

- Maximum likelihood estimation.
- Estimation algorithm (Nelder & Wedderburn, 1972) – Iteratively weighted least squares (IWLS)

$$X^T W X \boldsymbol{\beta} = X^T W \mathbf{z}$$

where

$X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^T$ is a design matrix $n \times p$,

$W = \text{diag}\{W_i\}$ – depends of the prior weights, variance function (distribution) and link function

$W_i = \frac{1}{V(\mu_i)} \left(\frac{d\mu_i}{d\eta_i} \right)^2$

$\boldsymbol{\beta}$ – parameter vector $p \times 1$

$\mathbf{z}$ – a vector $n \times 1$ (adjusted response variable) – depends on y and link function

$z_i = \eta_i + (y_i - \mu_i) \frac{d\eta_i}{d\mu_i}$

Inferential aspects

Measures of discrepancy:

Deviance

$$\mathsf{S} = \frac{\mathsf{D}}{\phi} = -2[\log \mathsf{L}(\hat{\boldsymbol{\mu}}, \mathbf{y}) - \log \mathsf{L}(\mathbf{y}, \mathbf{y})]$$

where $\mathsf{L}(\hat{\boldsymbol{\mu}}, \mathbf{y})$ e $\mathsf{L}(\mathbf{y}, \mathbf{y})$ are the likelihood function values for the current and saturated models

Generalized Pearson X^2

$$X^2 = \sum \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

- In general, comparisons involve nested models and deviance differences (Analysis of deviance).
- Many interesting comparisons involve non-nested models
- Use of Akaike Information Criterion (AIC) or Bayes Information Criterion (BIC) for model selection

$\mathsf{AIC} = -2 \log \mathsf{L} + 2$ (number of fitted parameters)
$\mathsf{BIC} = -2 \log \mathsf{L} + \log n$ (number of fitted parameters)

Table 2: Deviance Table – An example.

Model	DF	Deviance	Deviance Diff.	DF Diff.	Meaning
Null	$rab-1$	D_1			
			$D_1 - D_A$	$a-1$	A ignoring B
A	$a(rb-1)$	D_A			
			$D_A - D_{A+B}$	$b-1$	B including A
A+B	$a(rb-1)-(b-1)$	D_{A+B}			
			$D_{A+B} - D_{A*B}$	$(a-1)(b-1)$	Interaccion AB included A and B
A+B+A.B	$ab(r-1)$	D_{A*B}			
			D_{A*B}	$ab(r-1)$	Residual
Saturated	0	0			

References

[1] Aitkin, M.A., Francis, B.F., Hinde, J.P. and Darnell, R. (2009) *Statistical Modelling in R*. Oxford University Press.

[2] Collet, D. (1994). *Modelling binary data*. Chapman and Hall, London.

[3] Dobson, A.J. (2002). *An Introduction to Generalized Linear Models*. Chapman and Hall, London.

[4] Gbur, E.E.; Stroup, W.W.; McCarter, K.S.; Durham, S.; Young, L.J.; Christman, M.; West, M.; Kramer, M. *Analysis of Generalized Linear Mixed Models in the Agricultural and Natural Resources Sciences*. American Society of Agronomy, Soil Science Society of America and Crop Science Society of America, Madison.

[5] Hardin, J.W.; Hilbe, J.M. (2007) *Generalized linear models and extensions*. Stata Press.

[6] Madsen, H.; Thyregod, P. (2011) *An Introduction to General and Generalized Linear Models*. Chapman and Hall, London.

[7] McCullagh, P. e Nelder, J.A. (1983, 1989). *Generalized linear models*. Chapman and Hall, London.

[8] Myers, R.H.; Montgomery, D.C.; Vining, G.G. (2002) *Generalized linear models with Applications in Engineering and the Sciences* John Wiley, New York.