

Models for count data

Bert van der Veen

Department of Mathematical Sciences, NTNU

Outline

- ▶ Models for count data
- ▶ Residual diagnostics in GLMs
- ▶ Other useful models

Questions about yesterday?



The binomial GLM

Data: r the number of successes in N trials

Parameters: probability p (now: π_i)

Goal: estimate π_i for each observation

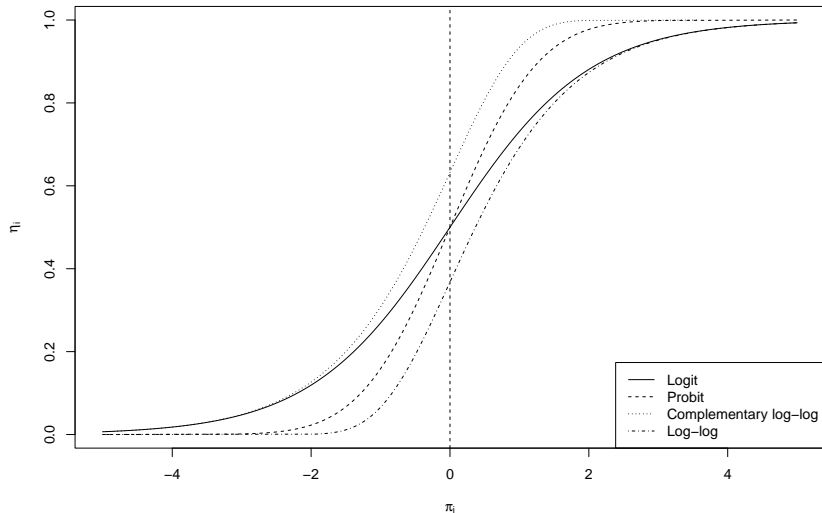
Binomial GLM use

- ▶ When a linear regression is not appropriate :)
- ▶ For binary data or counts of successes/failures

In ecology

- ▶ Predicting species' distributions
- ▶ Number of germinated plant seeds
- ▶ Prevalence of disease in a population
- ▶ Probability of observing a behavior
- ▶ Proportion of orchids 🙄

Binomial link functions (2)



Log-linear regression

Log-linear regression is a class of models that uses the log-link function:

$$\begin{aligned}
 \log\{\mathbb{E}(y_i|x_i)\} &= \eta_i = \alpha + x_i\beta \\
 \mathbb{E}(y_i|x_i) &= \lambda_i = \exp(\alpha + x_i\beta)
 \end{aligned}
 \tag{1}$$

Log-linear regression is commonly used to analyse count data

Typical count cases

- ▶ Caught fish
- ▶ Plants at a site
- ▶ Pidgeons in a city
- ▶ **Bigfoot reports**
- ▶ Wrongful convictions
- ▶ Stars in the night sky

The Poisson GLM

Data: k_i the count

Parameters: mean λ

Goal: estimate λ_i for each observation

The Poisson distribution

$$\mathcal{L}(y_i; \Theta) = \exp\{y_i \log(\lambda_i) - \lambda_i - \log(y_i!)\} \quad (2)$$



The Poisson parameter λ is the mean of the counting process

The Poisson distribution: rates

Alternatively we can write:

$$\mathcal{L}(y_i; \Theta) = \exp\{y_i \log(rt) - rt - \log(y_i!)\} \quad (3)$$

so, $\lambda = rt$

- ▶ r is the rate at which counts occur, per time period t
- ▶ we could also record counts within a certain period, instead of the total

Catching fish

- ▶ We go fishing for an hour and catch $\lambda = 5$ fish
- ▶ On average we caught $r = \frac{5}{60}$ fish per $t = 1$

Is Poisson regression really a GLM?

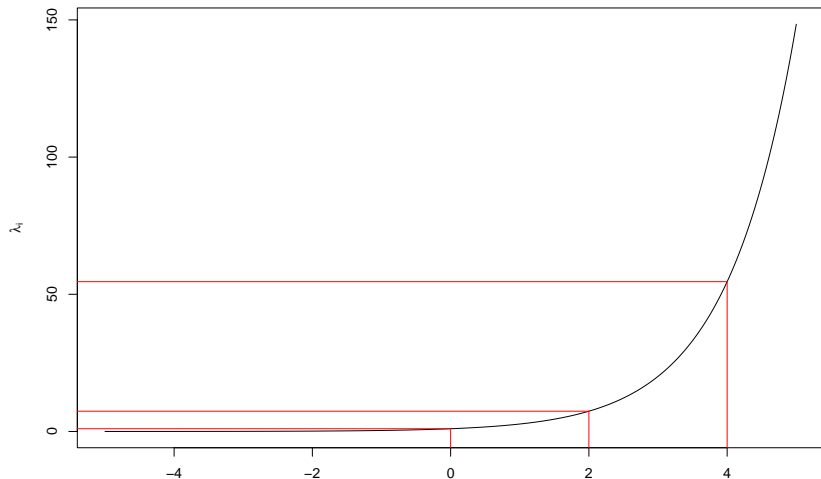
$$\mathcal{L}(y_i; \Theta) = \exp \left\{ \frac{y_i \log(\lambda) + \log(\lambda)}{1} + \log(y_i!) \right\} \quad (4)$$

All GLMs can be formulated as:

$$\mathcal{L}(y_i; \Theta) = \exp \left\{ \frac{y_i \eta_i - b(\eta_i)}{a(\phi)} + c(y_i, \phi) \right\} \quad (5)$$

Log-link function

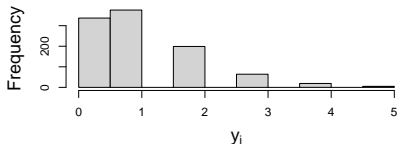
So log is the canonical link. This looks like:



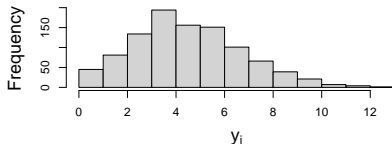
So, you caught twice as many fish!

The Poisson distribution visually

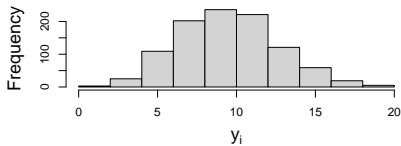
$\lambda = 1$



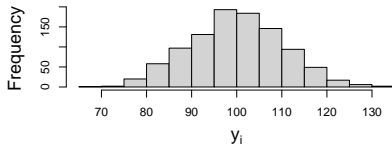
$\lambda = 5$



$\lambda = 10$



$\lambda = 100$



Poisson assumptions

- ▶ An event can occur $0 \dots \infty$ times
- ▶ Events are independent
- ▶ The rate of events is constant
- ▶ Events cannot occur simultaneously
- ▶ Variance equals the mean

Example: horseshoe crabs

Counts of male crabs (“satellites”) near female crabs



Figure 1: nwf.org



Figure 2: uwm.edu

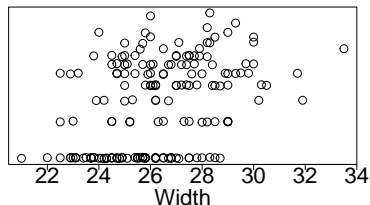
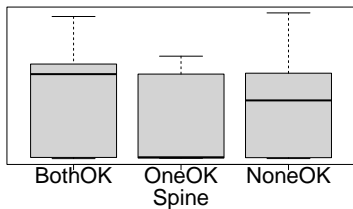
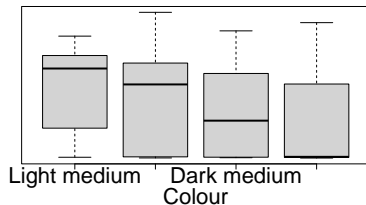
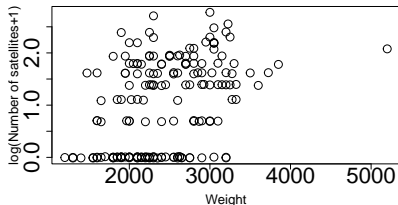
Horseshoe crabs: the data

Data originally from Brunswick (1996) via Agresti (2007) via Dunn and Smyth (2018)

- ▶ 173 observations
- ▶ 4 traits: colour, spine condition, width (cm), weight (g)

Col	Spine	Width	Sat	Wt
M	NoneOK	28.3	8	3050
DM	NoneOK	22.5	0	1550
LM	BothOK	26.0	9	2300
DM	NoneOK	24.8	0	2100
DM	NoneOK	26.0	4	2600
M	NoneOK	23.8	0	2100
LM	BothOK	26.5	0	2350

Horseshoe crabs: the data



What can we tell about the number of satellites?

Horseshoe crabs: fit the model

```

model <- glm(Sat ~ Spine + Colour + Width + Weight,
              family = "poisson", data = hcrabs)
  
```

Horseshoe crabs: interpreting parameters

##	Estimate	Std. Error	z value	Pr(> z)
## (Intercept)	-0.3600	0.97000	-0.37	0.7100
## SpineOneOK	-0.1500	0.21000	-0.70	0.4800
## SpineNoneOK	0.0870	0.12000	0.73	0.4700
## ColourMedium	-0.2600	0.17000	-1.60	0.1200
## ColourDark medium	-0.5100	0.20000	-2.60	0.0086
## ColourDark	-0.5300	0.23000	-2.30	0.0190
## Width	0.0170	0.04900	0.34	0.7300
## Weight	0.0005	0.00017	3.00	0.0028

(Intercept) = Light-medium coloured females with both spines in good condition, Width and Weight = 0

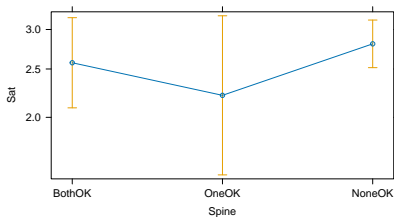
Horseshoe crabs: interpreting parameters, centered

##		Estimate	Std. Error	z value	Pr(> z)
##	(Intercept)	1.3000	0.15000	8.70	4.4e-18
##	SpineOneOK	-0.1500	0.21000	-0.70	4.8e-01
##	SpineNoneOK	0.0870	0.12000	0.73	4.7e-01
##	ColourMedium	-0.2600	0.17000	-1.60	1.2e-01
##	ColourDark medium	-0.5100	0.20000	-2.60	8.6e-03
##	ColourDark	-0.5300	0.23000	-2.30	1.9e-02
##	Width	0.0170	0.04900	0.34	7.3e-01
##	Weight	0.0005	0.00017	3.00	2.8e-03

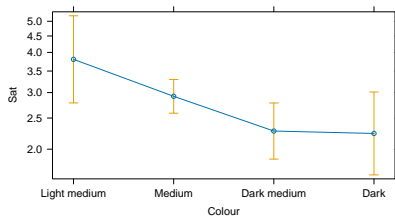
- ▶ (Intercept) = Light-medium coloured females with both spines in good condition, Weight \approx 2437g, Width \approx 26.3cm
- ▶ Width: $\exp(0.0170) = 1.017$, so double the number of satellites by $\log(2)/0.01709 = 40.56\text{cm}$ above average
- ▶ Weight: $\exp(0.0005) = 1.0005$, so double the number of satellites by $\log(2)/0.0005 = 1386.3\text{g}$ above average

Horseshoe crabs: visual interpretation

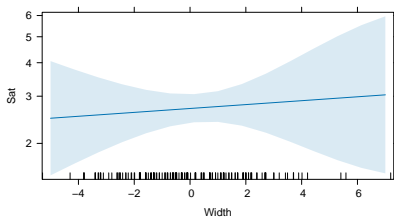
Spine effect plot



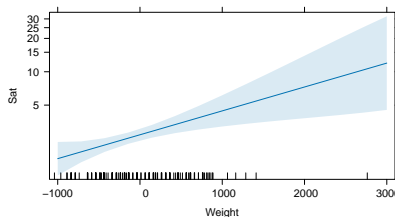
Colour effect plot



Width effect plot



Weight effect plot



Some other options

- ▶ Negative binomial (two types, with dispersion)
- ▶ Conway-Maxwell Poisson (with dispersion)
- ▶ Generalized Poisson (with dispersion)
- ▶ Skellam distribution (difference of counts)
- ▶ Binomial distribution (counts with a maximum)
- ▶ Truncated distributions
- ▶ Quasi-likelihood models

Do we have a good model?

More on this after the break

- ▶ Overdispersion or underdispersion
- ▶ Zero-inflation

Overdispersion

Our assumption: $\lambda = \text{var}(\mathbf{y})$

Reality: $\lambda \geq \text{var}(\mathbf{y})$

- ▶ Mean = variance
- ▶ If there is more variation, this assumption fails
- ▶ Consequences: CIs underestimate, biased parameter estimates, inflation in model selection

For our example: many females have few satellites, but some females have very many.

Underdispersion

Our assumption: $\lambda = \text{var}(\mathbf{y})$

Reality: $\lambda \leq \text{var}(\mathbf{y})$

Considerably less common than overdispersion.

Detecting overdispersion

- ▶ Residual diagnostics
- ▶ $D(\mathbf{y}; \hat{\boldsymbol{\mu}})/(n - k)$: should be close to 1
- ▶ `performance::check_overdispersion` (relies on asymptotics)
- ▶ Simulation (later today)

Dealing with dispersion: options

- ▶ Correct for it (calculate dispersion)
- ▶ Fit a different model
 - ▶ Negative binomial (overdispersion, MASS package)
 - ▶ Conway-Maxwell Poisson (over- and underdispersion.)
 - ▶ Generalized Poisson(over- and underdispersion)
 - ▶ Quasi-likelihood models
 - ▶ Mixed models (not covered here)

Quasi-likelihood models

Introduced by **Wedderburn (1974)**

- ▶ No “real” likelihood is specified for the data
- ▶ Means no AIC, but deviance exists
- ▶ Largely defined by its variance function

For Poisson responses: does not correct the parameter estimates

Negative-binomial

$$\mathcal{L}(y_i; \Theta) = \frac{\Gamma(y_i + \phi)}{\Gamma(\phi)y_i!} \left(\frac{\phi}{\mu_i + \phi} \right)^\phi \left(\frac{\mu_i}{\mu_i + \phi} \right)^{y_i} \quad (7)$$

- ▶ $\text{var}(\mathbf{y}) = \boldsymbol{\mu} + \frac{\boldsymbol{\mu}^2}{\phi}$
- ▶ For large ϕ Poisson!
- ▶ Requires more data/information due to extra parameter

Horseshoe crabs: Negative-binomial

```
modelnb <- MASS::glm.nb(Sat ~ Spine + Colour + Width + Weight,
                        data = hcrabs)
```

and compare the models:

```
AIC(model, modelnb)
```

##	df	AIC
## model	8	920.8833
## modelnb	9	763.3204

Horseshoe crabs: comparing estimates

##	Poisson estimate	NB estimate	Poisson SE	NB SE
## (Intercept)	1.3000	1.4000	0.15000	0.34000
## SpineOneOK	-0.1500	-0.2400	0.21000	0.40000
## SpineNoneOK	0.0870	0.0430	0.12000	0.25000
## ColourMedium	-0.2600	-0.3200	0.17000	0.37000
## ColourDark medium	-0.5100	-0.6000	0.20000	0.42000
## ColourDark	-0.5300	-0.5800	0.23000	0.47000
## Width	0.0170	-0.0024	0.04900	0.10000
## Weight	0.0005	0.0007	0.00017	0.00036

- ▶ SEs have doubled
- ▶ Coefficients have changed, but largely the same conclusions
- ▶ Except perhaps for the effect of “Width”

Summary

- ▶ Counts are analysed with log-linear models
- ▶ Due to the Poisson assumption, dispersion issues can arise
- ▶ This biases parameter and uncertainty estimates
- ▶ Negative-binomial models are useful for overdispersion problems
- ▶ Conway-Maxwell & Generalized Poisson can be used as well, and for underdispersion
 - ▶ `glmmTMB` or `VGAM` packages
- ▶ Zero-inflation (later)