GLM extensions

Ben Bolker

06:54 28 June 2015

Basics/reminders

Distributions (`family`)

- how to pick distribution?
- use knowledge, not statistical testing – but cf. Firth (1988), Dick (2004)
- 99% of GLMs are Gaussian, binomial (usu. Bernoulli), Poisson (or overdispersed equivalents)
- log-Normal usually more practical for continuous data than Gamma

Link functions

- linearizing transformation:
  e.g. exp ↔ log, logistic ↔ logit
- *canonical* link usually OK (binomial=logistic, Poisson=log)
- differences (e.g. probit vs logit) mostly have to do with interpretation or culture
- log generally more practical than inverse link for Gamma

Parameterization

- simple but not easy
- default *treatment* contrasts; *sum-to-zero* contrasts
- interpreting interactions
- R formula (Wilkinson-Rogers) notation
- "what if I want to know the value for each group?":
  fit with `-1` or use `lsmeans`, `effects`, `rockchalk` packages
- centering and scaling (Schielzeth 2010)
- linear models apply **on linear predictor scale**
- rules of thumb for interpreting effects:
  - log changes ≈ proportional
  - logit ≈ proportional at ends, $r/4$ near 50%

Top GLM mistakes

- ignoring overdispersion

- applying discrete models (Poisson, binomial) to non-discrete data: **don't divide!**
- equating negative binomial with binomial rather than Poisson
- confusion in interpreting effects
- worrying about marginal rather than conditional distributions of data
- back-transforming standard errors
- using (k, N) rather than $(k, N-k)$ in binomial models
- getting confused by predictions on the linear predictor scale
- using GLMs where linear models will do (i.e. `glm` instead of `lm`) (*mostly harmless*)
- forgetting to use `type="response"` using `predict.glm()`
- ignoring blocking factors (failing to use mixed models where necessary)

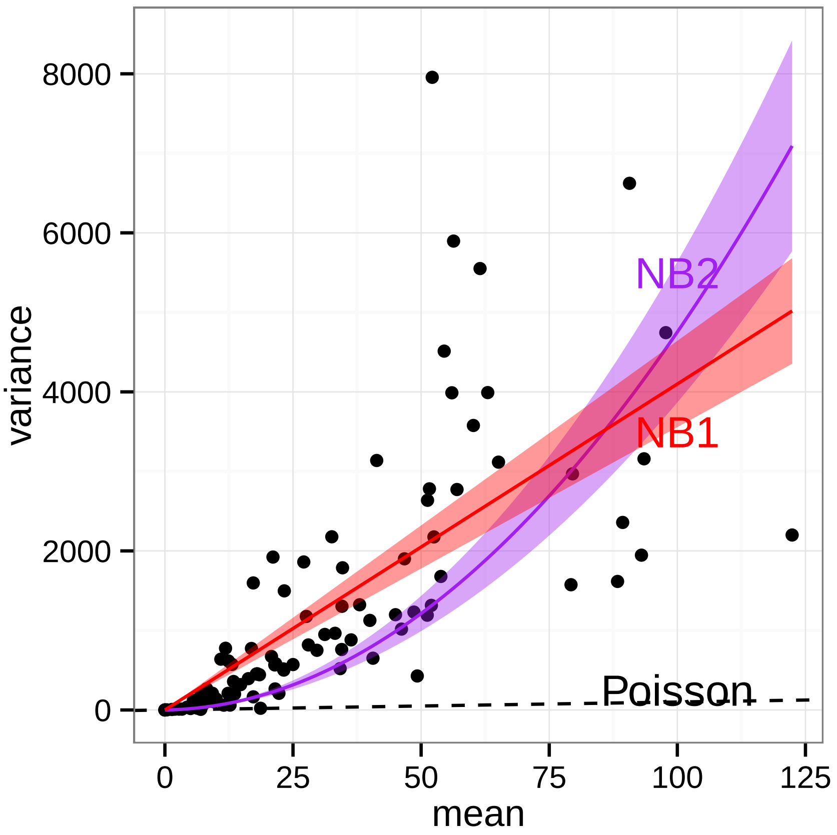
Relaxing distributional assumptions

Overdispersion

- *scale parameter* is fixed to 1 for Poisson (variance=mean), binomial (variance=$Np(1-p)$
- often untrue!
- checking: compute `deviance(fit)/df.residual(glm1)`, or use `aods3:gof()`
- across-the-board variance inflation, not outliers/bad model
- more important to do **something** about overdispersion than exactly what you do

Methods

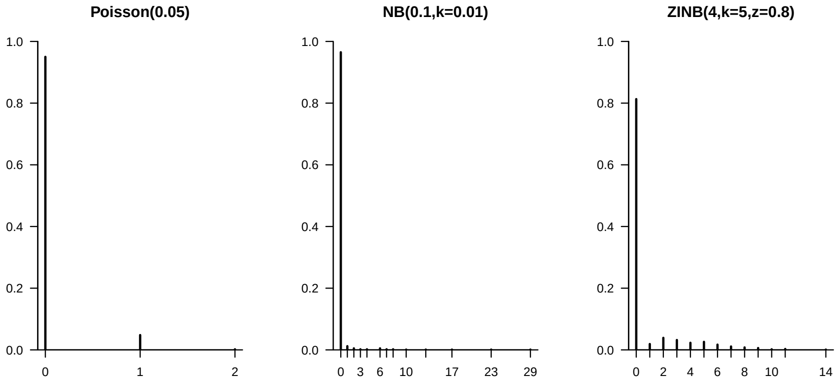
- quasi-likelihood (`glm(...,family=quasibinomial)`); point estimates **don't change**, just adjusts std errors/CIs/p-values
- extended/conjugate models (neg binomial [`MASS::glm.nb`], betabinomial); `gamlss`, `bbmle`
- different mean-variance relationships (Hardin and Hilbe 2007)
  - NB2: $V = \mu + \mu^2/k$ $(k > 0)$
  - NB1: $V = \tau\mu$ $(k > 1)$
- observation-level random effects in mixed models (lognormal-Poisson; logit-normal-binomial) (`lme4`)
- what about *underdispersion*?
  - less common
  - quasi-likelihood OK


8000
6000
4000
2000
0
variance
0
25
50
75
100
125
mean
NB2
NB1
Poisson

  - ordinal models
  - more exotic (COM-Poisson)

Zero-alteration

- zero-inflation: **too many** zeros, not just lots of zeros (Warton 2005)



- zero-alteration: maybe too few zeros?
- zero-*truncation*: no zeros
  - truncated Poisson/binomial
  - or just assume $X \sim \text{Poisson}(\lambda) + 1$ (easier)
- *zero-inflation* (mixture) and *hurdle* models: `pscl` package
- don't throw out zeros - but remember that a data set with mostly zeros is not very informative!
- zero-inflated *continuous* data?
  - two-stage (binomial + positive distribution)
  - censoring model (Tobit)
  - Tweedie models

Beta regression

- for proportion data where we *don't* know the denominator
- `betareg`, `bbmle`, `glmmADMB` packages
- exact 0 and 1 values are problematic

Extend location model

Polynomials

- adding quadratic term can make a big difference

- link-function polynomials are more reasonable (e.g. quadratic + log-link=Gaussian)
- probably not worth taking too seriously/going beyond cubic

Additive models

- smooth piecewise cubic functions (partitioned at *knots*) (Wood 2006)
- simple: `splines::ns(.,df=k)`
- `mgcv`: penalized spline models
- harder to interpret
- slightly more 'expensive' than simpler nonlinear models
- harder to constrain (linear extrapolation)
- extensions: 2-D, monotonic, convex ...

Link tricks

- use alternate links, or transform X variable
  - log-link, `y~x`: exponential
  - log-link, `y~log(x)`: power-law \ $(Y = \exp(a + b\log(X)) \to Y = cX^d$
- binomial model with log-link gives exponential (or saturating-exponential) model (Strong et al. 1999; Tiwari et al. 2006)
  - `y~x-1, family=binomial(link="log")` $\to Y = \exp(bx)$
- inverse link with `y~1/x` gives *hyperbolic* (Michaelis-Menten/Holling type II) type models
- power-logistic
- defining upper limit for logistic
- use `glm(...,family=gaussian(link="log"))` to fit exponential models with *constant* variance (cf. `lm(log(y)~x)`)

Nonlinear models

- complete flexibility
- formula interface of `bbmle` simplifies things
- need to pick starting values; worry more about parameterization

Misc tricks

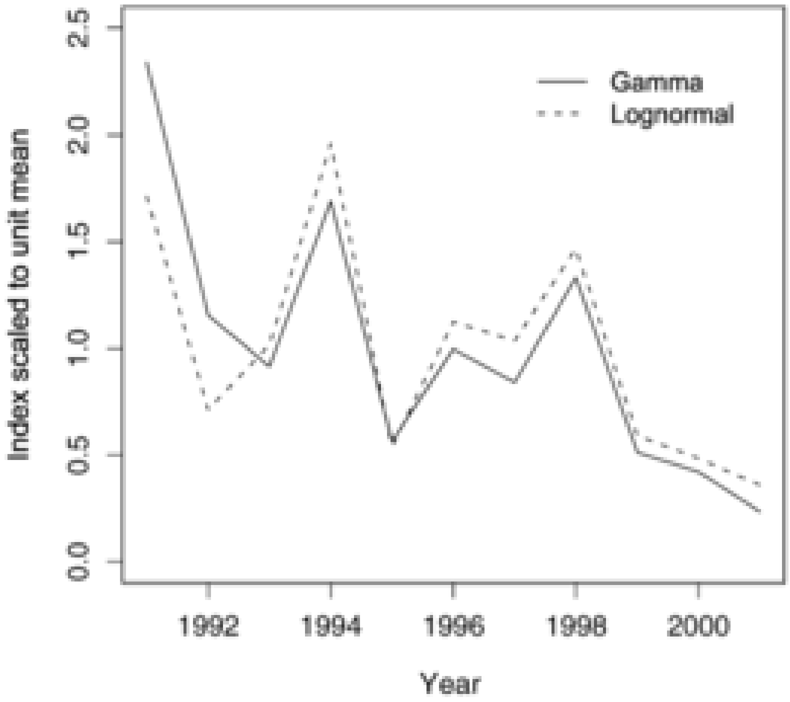
Offset tricks

- constants added to the linear predictor

- Poisson model: add log(area) (or whatever) to model a ratio of counts by exposure
- $Y = \exp(b_0 + b_1 x + \mathbf{\log A}) \rightarrow Y/A = \exp(b_0 + b_1 x)$
- survival/mortality model: use log(exposure) on **log-hazard** scale (`binomial(link="cloglog")`), or power-logistic ("Mayfield ratios")
- for convenience in resetting null model. e.g. log-link, `y~log(x)+offset(log(x))` tests difference from isometric model

Complete separation

- all-zero/all-one categories in binomial, Poisson models
- GLM estimate *should* be infinite (but is just large); maybe `glm.fit` warning
- *bias-reduced* estimate (Firth); `brglm`, `logistf` packages
- standard errors and p-values from `summary()` (*Wald* approximation) are crazy
- Bayesian priors: `arm::bayesglm`



References

Dick, E.J. 2004. "Beyond 'Lognormal Versus Gamma': Discrimination Among Error Distributions for Generalized Linear Models." *Fisheries Research* 70 (2–3): 351–66. doi:10.1016/j.fishres.2004.08.013.

Firth, David. 1988. "Multiplicative Errors: Log-Normal or Gamma?" *Journal of the Royal Statistical Society. Series B (Methodological)* 50 (2): 266–68. `http://www.jstor.org/stable/2345764`.

Hardin, James William, and Joseph Hilbe. 2007. *Generalized Linear Models and Extensions*. Stata Press.

Schielzeth, Holger. 2010. "Simple Means to Improve the Interpretability of Regression Coefficients." *Methods in Ecology and Evolution* 1: 103–13. doi:10.1111/j.2041-210X.2010.00012.x.

Strong, D. R., A. V. Whipple, A. L. Child, and B. Dennis. 1999. "Model Selection for a Subterranean Trophic Cascade: Root-Feeding Caterpillars and Entomopathogenic Nematodes." *Ecology* 80: 2750–61.

Tiwari, Manjula, Karen A. Bjorndal, Alan B. Bolten, and Benjamin M. Bolker. 2006. "Evaluation of Density-Dependent Processes and Green Turtle *Chelonia Mydas* Production at Tortuguero, Costa Rica." *Marine Ecological Progress Series* 326: 283–93.

Warton, David I. 2005. "Many Zeros Does Not Mean Zero Inflation: Comparing the Goodness-of-Fit of Parametric Models to Multivariate Abundance Data." *Environmetrics* 16 (3): 275–89. doi:10.1002/env.702.

Wood, Simon N. 2006. *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC.