

Generalized linear models

3 Oct 2023

```
## it's nice to include packages at the top
## (and NOT automatically install them)
## try not to carry over packages you don't use
library(ggplot2); theme_set(theme_bw())
## diagnostics
library(performance)
library(DHARMA)
## downstream model evaluation
library(broom)
library(dotwhisker)
library(emmeans)
library(effects)
library(marginaleffects)
library(parameters)
## library(ggeffects)
```

Basics

- assume $\mathbf{y}_i \sim \text{Dist}(g^{-1}((\mathbf{X}\beta)_i))$
- g = **link function**
- $\eta = \mathbf{X}\beta$ = **linear predictor**
- **link scale, data or response scale**
- GLMs inverse-transform η , they don't transform y
- allows:
 - separate control of heteroscedasticity and nonlinearity
 - almost as convenient/efficient as LMs
 - equivalent to MLE in many cases
- in practice almost all GLMs are logistic (binary data) or Poisson

- lots of inference, diagnostics, etc. inherited from LM framework

Exponential family

- $f(x|\theta) = h(x)g(\theta) \exp(\eta(\theta)T(x))$
- e.g. Poisson: $f(x|\theta) = \theta^x \exp(-\theta)/x! = (1/x!) \exp(-\theta) \exp(x \log(\theta))$
- $h(x) = 1/x!; g(\theta) = \exp(-\theta); \eta(\theta) = \log(\theta); T(x)$
- η is the **canonical link** function for the family (nice mathematical properties)
- binomial, Poisson, Gamma (inverse Gaussian, von Mises distribution ...)

Mean-variance relations

- can show that all we need for computation is the link function and the **variance function** $V = f(\mu)$ (may also depend multiplicatively on a **scale** or **dispersion parameter**, e.g. $V = \mu$ for Poisson, $V = \sigma^2$)

Link functions

- canonical doesn't always work best (e.g. Gamma/inverse link)
- probit vs logit; not much difference
- cloglog; *log-hazard* scale
- inverse link: linear changes in the *rate* of events

Log-hazards and log-hazard offsets

- if hazard is h , probability is $1 - \exp(-h)$
- $C(\mu) = \log(-\log(1 - \mu))$
- $C^{-1}(\eta) = 1 - \exp(-\exp(\eta))$
- $C^{-1}(\eta + \log(\Delta t)) = 1 - \exp(-\exp(\eta) \cdot \Delta t)$
- $\rightarrow 1 - (1 - \mu_0)^{\Delta t}$

Computation

- iteratively reweighted least squares
- needs starting values, but almost always robust to them

in R

- “family” functions contain all of the components needed for GLM fitting, prediction, etc.
- some of the components are weird (e.g. `$aic`)
- canonical link is used by default

```
names(binomial)
```

NULL

Offsets

- allow for differential search effort, ratios, etc.
- typically add $\log(e)$
- e.g. $y \sim \text{Poisson}(\mathbf{X}\beta + \log(A))$ is equivalent to modeling the response y/A , but without messing up the mean-variance relationship

Offset/link tricks

- fit an exponential curve with constant variance: `family = gaussian(link = "log")`
- Ricker function $y = ax \exp(-bx)$: `log-link, y ~ x + offset(log(x))`
- Michaelis-Menten $y = ax/(b + x) \rightarrow 1/y = (b/a) \cdot (1/x) + 1/a$: `inverse-link, y ~ I(1/x)`

Model interpretation, visualization, testing

Parameter interpretation

- log scale: easy
- logit scale: $\approx \log$ for low baseline, $\approx \log(1 - x)$ for high baseline, slope $\beta/4$ for intermediate values
- cloglog: **log-hazard** scale

Inference

- Wald tests (no finite-size corrections!)
- approximate Wald CIs (compute then back-transform)
- profile CIs

Overdispersion

- too much variance
- SSQ of Pearson residuals $\sim \chi^2(n - p)$
 - quasi-likelihood (also handles **underdispersion**)
 - compounded models (negative binomial, beta-binomial)
 - observation-level random effects

Extended distributions

- VGAM, glmmTMB packages

Complete separation

- there is some linear combination of predictors that separates 0 from 1 responses (or 0 from non-zero responses in the case of count models)
- infinite MLE
- Hauck-Donner effect screws up Wald tests
- likelihood ratio tests still OK (sort of)
- Firth logistic regression (brglm2 package), Bayesian priors (arm::bayesglm)

Zero-inflation/hurdle models

- finite mixture models

Most common GLM problems

- binomial/Poisson models with non-integer data
- failing to specify family (default Gaussian: \rightarrow linear model); using glm() for linear models (unnecessary)
- predictions on effect scale
- using (k, N) rather than $(k, N - k)$ with family=binomial
- back-transforming SEs rather than CIs
- neglecting overdispersion
- Poisson for *underdispersed* responses
- equating negative binomial with binomial rather than Poisson
- worrying about overdispersion unnecessarily (binary/Gamma)
- ignoring random effects

References