Generalized linear models

Summary

Process of modeling

1. Study design
2. Data collection
3. Selection of model class
4. Estimation
5. Model checking
6. Conclusions
7. Reporting

Type of response

- continuous $(-\infty, \infty)$
- continuous $(0, \infty)$
- continuous $(0, c)$, censored
- discrete $0, 1, 2, \ldots$
- discrete $0, 1, 2, \ldots, m$
- discrete $0, 1$
- categorical A, B, C, ordered
- categorical B, V, F, not ordered

Structure of the model

- Systematic part

$$g(\mathrm{E}(Y_i)) = \sum_{j=1}^{p} \beta_j x_{ij}$$

- Random part
  - Distribution of $Y_i - \mu_i$
  - $$\mathrm{Var}(Y_i) = \frac{\phi V(\mu_i)}{a_i}$$

GLM definition

1. The distribution of Y_i belongs to the exponential family. For the exponential family, the density function can be presented in the form

$$f_{Y_i}(y_i; \theta_i, \phi) = \exp\left(\frac{a_i(y_i\theta_i - b(\theta_i))}{\phi} + c(y_i, \phi/a_i)\right),$$

   where

   - θ_i, $i = 1, \ldots, n$ are unknown parameters (canonical parameters),
   - ϕ is the dispersion parameter (scale parameter) that can be known or unknown,
   - a_i, $i = 1, \ldots, n$ are known prior weights of each observation and
   - $b()$ and $c()$ are known functions. The first derivative $b'()$ is monotonic and differentiable.

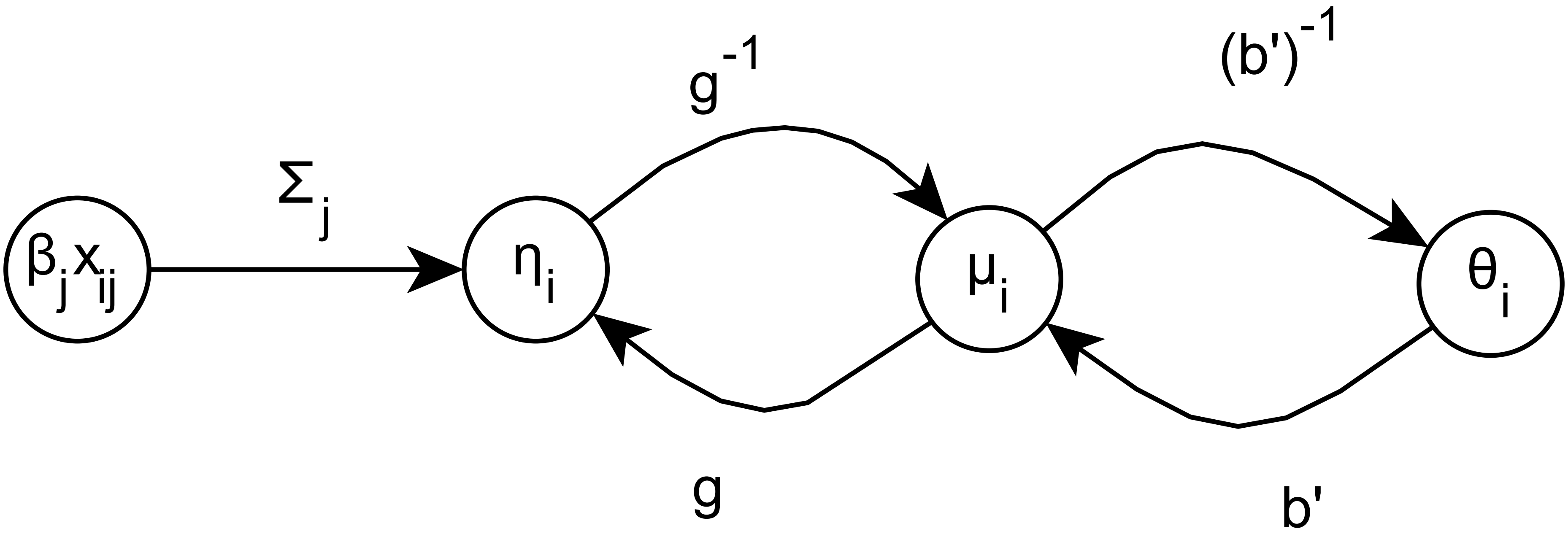
2. Random variables $Y_1, Y_2, \ldots, Y_n$ are mutually independent.
3. The expected value $\mu_i = \mathrm{E}(Y_i)$ depends on linear predictor $\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$ through monotonic and differentiable link function g

$$g(\mu_i) = \eta_i.$$

Standard distributions

	Normal	Poisson	Binomial	Gamma	Inv. Gaussian
Notation	$N(\mu, \sigma^2)$	$\text{Poisson}(\mu)$	$\text{Bin}(m, \pi)/m$	$\text{Gamma}(\lambda, \nu)$	$IG(\mu, \sigma^2)$
Range of y	$(-\infty, \infty)$	$0, 1, 2, \ldots$	$0, \frac{1}{m}, \frac{2}{m}, \ldots, 1$	$(0, \infty)$	$(0, \infty)$
ϕ	σ^2	1	$1/m$	$1/\nu$	σ^2
$b(\theta)$	$\theta^2/2$	$\exp(\theta)$	$\log(1 + e^\theta)$	$\log(-\theta)$	$-(2\theta)^{1/2}$
μ	θ	$\exp(\theta)$	$e^\theta/(1 + e^\theta)$	$\nu\lambda$	$(-2\theta)^{-1/2}$
Can. link	identity	log	logit	$1/\mu$	$1/\mu^2$
$V(\mu)$	1	μ	$\mu(1 - \mu)$	μ^2	μ^3

Parameters and functions



Likelihood is your friend

- Writing the log-likelihood for a model is a fundamental skill for a statistician.
- GLMs offer a common framework for models for different response types
  - Properties of exponential family can be applied
  - Estimation of parameters via Fisher scoring
  - Common concepts: link function, residuals, deviance, . . .

Deviance

- Deviance

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2\phi(l(\mathbf{y}; \mathbf{y}) - l(\hat{\boldsymbol{\mu}}; \mathbf{y}))$$

where $l(\mathbf{y}; \mathbf{y})$ is the log-likelihood of the saturated model (full model).

- Scaled deviance

$$D^*(\mathbf{y}; \hat{\boldsymbol{\mu}}) = \frac{D(\mathbf{y}; \hat{\boldsymbol{\mu}})}{\phi}$$

Residuals

Raw residuals (response residuals)

$$r_i = y_i - \hat{\mu}_i$$

Pearson residuals

$$r_{P,i} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)/a_i}}$$

Deviance residuals

$$r_{D,i} = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i},$$

where

$$d_i = 2a_i\left(y_i\left(\theta_i(y_i) - \theta_i(\hat{\mu}_i)\right) - b\left(\theta_i(y_i)\right) + b\left(\theta_i(\hat{\mu}_i)\right)\right).$$

Anscombe residuals where y_i's and μ_i's are transformed so that the residuals become approximately normally distributed.

Likelihood ratio test

- Hypothesis

$$\begin{aligned} H_0 : & \quad \beta_j = 0 \\ H_1 : & \quad \beta_j \neq 0 \end{aligned}$$

- Test statistics

$$-2 \log \Lambda = 2(l(\hat{\boldsymbol{\beta}}; \mathbf{y}) - l(\hat{\boldsymbol{\beta}}_0; \mathbf{y}))$$

where $\hat{\boldsymbol{\beta}} = \hat{\beta}_1, \ldots, \hat{\beta}_p$ and $\hat{\boldsymbol{\beta}}_0 = \hat{\beta}_1, \ldots, \beta_j = 0, \ldots, \hat{\beta}_p$ are the maximum likelihood estimates under the two models.

- The statistic $-2 \log \Lambda$ follows asymptotically χ^2_1 distribution.

Binary response

- Binary and binomial response
- Link functions: logit, probit, cloglog
- Overdispersion?
- Odds and log-odds
- Non-existence of MLE
- Latent variable interpretation
- Example: switching measurements

Count response

- Log-link
- Offset term
- Overdispersion?
- Example: follow-up for cardiovascular diseases

Nominal and ordinal response

- Multinomial distribution
- Logit-link
- Proportional odds model
- Latent variable interpretation for ordinal response

Positive response

- Zeros or ‘near-zeros’ in the data?
- Gamma distribution
- Inverse Gaussian distribution
- Link functions: identity, inverse, inverse squared and log
- Compound Poisson model

Time-to-event response

- Censoring and truncation
- Prospective and retrospective studies
- Survival function and hazard function
- Proportional hazards model

Extensions

- Quasi-likelihood
- Generalized linear mixed models (GLMM)
- Generalized additive models (GAM)
- Neural networks