Introduction to Generalized Linear Models

Myriam Riek

May 17, 2010

Outline

Recall the Linear Model

n independent observations of response variable Y

with $Y_i \sim \mathcal{N}(\mu_i, \sigma^2) \qquad i = 1, \ldots, n$

and $\mu_i = \mathbf{x}_i^T \beta$

Simple example: $\mathbf{x}_i^T \beta = \begin{pmatrix} 1 \\ x_i \end{pmatrix}^T \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \beta_0 + \beta_1 x_i$

Some Definitions

Linear predictor

$\eta_i = \mathbf{x}_i^T \beta$

Inverse link function

$\mu_i = g^{-1}(\eta_i)$

Link function

$\eta_i = g(\mu_i)$

Linear model: $\mu_i = \eta_i$, i.e., $g = g^{-1} =$ identity

The Generalized Linear Model

Generalization

(i) $Y_i \sim F(\mu_i)$

$F \in \mathcal{F} =$ exponential family of distributions

(ii) $g(\mu_i)$ any suitable function

preferably accounting for restrictions in μ_i and η_i

dependent on F

Exponential Family of Distributions (EFD)

Examples for F

- Normal
- Bernoulli
- Binomial
- Poisson
- Exponential
- Gamma
- . . .

Logistic Regression

n independent observations of response variable Y

Binary data $Y \in \{0,1\}$

$$Y_i \sim \text{Bern}(\mu_i) \qquad E[Y_i] = \mu_i \qquad \text{var}(Y_i) = \mu_i(1-\mu_i)$$

Proportions $Y \in (0,1)$

$$Y_i = \frac{Z_i}{m_i} \sim \mathcal{B}(m_i, \mu_i) \quad E[Y_i] = \mu_i \quad \text{var}(Y_i) = \frac{\mu_i(1-\mu_i)}{m_i}$$

Link function

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

(logit)

Inverse link function

$$\mu_i = g^{-1}(\eta_i) = \frac{e^{\eta_i}}{1+e^{\eta_i}}$$

(*logistic*)

Log-Poisson Regression

n independent observations of response variable Y

$Y_i \sim \mathcal{P}(\mu_i) \qquad \text{E}[Y_i] = \mu_i \qquad \text{var}(Y_i) = \mu_i$

Link function

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

(log)

Inverse link function

$\mu_i = g^{-1}(\eta_i) = e^{\eta_i}$

(exp)

Canonical pdf for EFD

The pdf (or pmf) of $Y_i \sim F(\mu_i) \in \mathcal{F}$ can be written in so called canonical form

$$f(y_i \big| \theta_i, \tau^2, \omega_i) = \exp(\frac{\theta_i y_i - d(\theta_i)}{\tau^2} \omega_i) h(y_i, \tau^2, \omega_i)$$

where

θ_i	canonical parameter
τ^2	dispersion parameter (fixed)
ω_i	some number ($= 1$, for binomial data $= m_i$)
$d(\theta_i)$	characteristic function for F
$h(y_i, \tau, \omega_i)$	normalizing function, characteristic for F

$$\mathrm{E}[Y_i] = \mu_i = d'(\theta_i) \qquad \mathrm{var}(Y_i) = d''(\theta_i)\frac{\tau^2}{\omega_i} = \nu(\mu_i)\frac{\tau^2}{\omega_i}$$

Examples of Canonical pdf for EFD

Normal data

$\theta_i = \mu_i, \quad d(\theta_i) = \frac{\theta_i^2}{2}, \quad \nu(\mu_i) = 1,$
$\omega_i = 1, \quad \tau^2 = \sigma^2$ (nuisance parameter)

$\rightarrow \mathrm{E}[Y_i] = \mu_i = \theta_i \qquad \mathrm{var}(Y_i) = \tau^2$

Binary data

$\theta_i = \log(\frac{\mu_i}{1-\mu_i}), \quad d(\theta_i) = \log(1 + e^{\theta_i}),$
$\nu(\mu_i) = \mu_i(1-\mu_i), \quad \omega_i = 1, \quad \tau^2 = 1$

$\rightarrow \mathrm{E}[Y_i] = \mu_i = \frac{\exp\theta_i}{1+\exp\theta_i} \qquad \mathrm{var}(Y_i) = \mu_i(1-\mu_i)$

Poisson data

$\theta_i = \log(\mu_i), \quad d(\theta_i) = e^{\theta_i}, \quad \nu(\mu_i) = \mu_i,$
$\omega_i = 1, \quad \tau^2 = 1$

$\rightarrow \mathrm{E}[Y_i] = \mu_i = e^{\theta_i} \qquad \mathrm{var}(Y_i) = \mu_i$

Canonical Link Functions

Special class of link functions with nice mathematical properties

Definition

Link η_i to the canonical parameter θ_i,

with $\theta_i = \eta_i = \mathbf{x}_i^T \beta$

Examples

Normal data	identity
Bernoulli/binomial data	logit
Poisson data	log function

In what follows we are considering the canonical link function only

Maximum Likelihood Estimators

Maximum likelihood principle used for deriving estimates of β, i.e.,

$$\left.\frac{\partial \ell(\mathbf{y},\beta)}{\partial \beta}\right|_{\beta=\hat{\beta}} = \sum_{i=1}^{n} \left.\frac{\partial \log f(y_i,\eta_i)}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta}\right|_{\beta=\hat{\beta}}$$

$$= \sum_{i=1}^{n} \left.\frac{y_i - d'(\eta_i)}{\tau^2} \omega_i \mathbf{x}_i\right|_{\beta=\hat{\beta}} = 0 \text{ (for canonical link)}$$

solved using a modification of the iterative Newton-Raphson algorithm called Fisher's method of scoring

Recall Newton-Raphson

At $(k+1)$-th iteration:

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} - \mathbf{H}^{-1}(\hat{\beta}^{(k)}) \frac{\partial \ell(\mathbf{y}, \beta)}{\partial \beta}\Big|_{\beta=\hat{\beta}^{(k)}}$$

with (for canonical link)

$$\mathbf{H}(\hat{\beta}^{(k)}) = \frac{\partial^2 \ell(\mathbf{y}, \beta)}{\partial \beta \partial \beta^T}\Big|_{\beta=\hat{\beta}^{(k)}} = -\sum_{i=1}^{n} \frac{d''(\eta_i)}{\tau^2} \omega_i \mathbf{x}_i \mathbf{x}_i^T \Big|_{\beta=\hat{\beta}^{(k)}}$$

Fisher's Method of Scoring (1)

Replaces

$$\mathbf{H}(\hat{\beta}^{(k)})$$

with

$$\mathrm{E}[\mathbf{H}(\hat{\beta}^{(k)})] = \mathrm{E}\left[\frac{\partial^2 \ell(\mathbf{y}, \beta)}{\partial\beta\partial\beta^T}\bigg|_{\beta=\hat{\beta}^{(k)}}\right] = -\mathbf{I}(\hat{\beta}^{(k)})$$

the Fisher information matrix. Thus,

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + \mathbf{I}^{-1}(\hat{\beta}^{(k)})\frac{\partial \ell(\mathbf{y}, \beta)}{\partial\beta}\bigg|_{\beta=\hat{\beta}^{(k)}}$$

For canonical link

$$\mathbf{I}(\hat{\beta}^{(k)}) = -\mathbf{H}(\hat{\beta}^{(k)})$$

i.e., Fisher scoring and Netwon-Raphson are equivalent

Fisher's Method of Scoring (2)

Since (for canonical link)

$$\mathbf{I}(\hat{\beta}^{(k)}) = \Sigma_{i=1}^{n} \frac{d''(\eta_i)}{\tau^2} \omega_i \mathbf{x}_i \mathbf{x}_i^T \Big|_{\beta=\hat{\beta}^{(k)}} \quad \text{and}$$

$$\frac{\partial \ell(\mathbf{y}, \beta)}{\partial \beta}\Big|_{\beta=\hat{\beta}^{(k)}} = \Sigma_{i=1}^{n} \frac{y_i - d'(\eta_i)}{\tau^2} \omega_i \mathbf{x}_i \Big|_{\beta=\hat{\beta}^{(k)}}$$

hence

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + (\mathbf{X}^T \mathbf{\Omega}^{(k)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{\Omega}^{(k)} \tilde{\mathbf{Y}}^{(k)}$$

where

$$\mathbf{\Omega}^{(k)} = \text{diag}(d''(\eta_1)\omega_1\big|_{\beta=\hat{\beta}^{(k)}}, \ldots, d''(\eta_n)\omega_n\big|_{\beta=\hat{\beta}^{(k)}}) \quad \text{and}$$

$$\tilde{\mathbf{Y}}^{(k)} = \left(\frac{y_1 - d'(\eta_1)}{d''(\eta_1)}\Big|_{\beta=\hat{\beta}^{(k)}}, \ldots, \frac{y_n - d'(\eta_n)}{d''(\eta_n)}\Big|_{\beta=\hat{\beta}^{(k)}}\right)^T$$

Iterative Weighted Least Squares

Equivalent to ML approach using Fisher scoring

Since

$$\begin{aligned}\hat{\beta}^{(k+1)} &= \hat{\beta}^{(k)} + (\mathbf{X}^T\mathbf{\Omega}^{(k)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Omega}^{(k)}\tilde{\mathbf{Y}}^{(k)} \\ &= (\mathbf{X}^T\mathbf{\Omega}^{(k)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Omega}^{(k)}(\mathbf{X}\hat{\beta}^{(k)} + \tilde{\mathbf{Y}}^{(k)}) \\ &= (\mathbf{X}^T\mathbf{\Omega}^{(k)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Omega}^{(k)}\mathbf{T}^{(k)}\end{aligned}$$

hence $\hat{\beta}^{(k+1)}$ are the weighted least squares solution for regressing $\mathbf{T}^{(k)}$, the linearized response, linearly on $\mathbf{X}$

Distribution of MLE

$$\hat{\beta} \sim^{app} \mathcal{N}_p(\beta, \mathbf{I}^{-1}(\hat{\beta}))$$

$$\mathbf{I}^{-1}(\hat{\beta}) = \tau^2(\mathbf{X}^T\mathbf{\Omega}(\hat{\beta})\mathbf{X})^{-1}$$

Wald Tests and CIs (1)

Assuming normality for $\hat{\beta}$ holds, the Wald test statistic W for a linear combination $\mathbf{B}\beta$ of β, $\mathbf{B}$ being a (q x p) matrix, is

$$W = (\mathbf{B}\hat{\beta} - \mathbf{B}\beta)^T \mathbf{V}^{-1} (\mathbf{B}\hat{\beta} - \mathbf{B}\beta) \sim \chi^2_q$$

with

$$\mathbf{V} = \text{cov}(\mathbf{B}\hat{\beta}) = \mathbf{B}\mathbf{I}^{-1}(\hat{\beta})\mathbf{B}^T$$

E.g.,

$$W = \frac{(\hat{\beta}_j - \beta_j)^2}{\mathbf{I}^{-1}(\hat{\beta})_{jj}} \sim \chi^2_1 \text{ or } \sqrt{W} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\mathbf{I}^{-1}(\hat{\beta})_{jj}}} \sim \mathcal{N}(0, 1)$$

Wald Tests and CIs (2)

A $(1-\alpha)100\%$-CIs for $\mathbf{B}\beta$ based on the Wald test statistic is

$$\{b : W = (\mathbf{B}\hat{\beta} - b)^T \mathbf{V}^{-1} (\mathbf{B}\hat{\beta} - b) \leq \chi^2_{q,1-\alpha}\}$$

E.g.,

$$\hat{\beta}_j \pm z_{1-\alpha/2} \sqrt{\mathbf{I}^{-1}(\hat{\beta})_{jj}}$$

Note that these CIs are symmetric around $\mathbf{B}\hat{\beta}$

Likelihood Based Tests and CIs (1)

A likelihood ratio test (LRT) compares the maximum $\mathcal{L}_{H_0}$ with the maximum $\mathcal{L}_{H_A \supset H_0}$

$$LRT-\text{statistic} = -2(\ell_{H_0} - \ell_{H_A \supset H_0}) \sim^{app} \chi^2_q \text{ (if } H_0 \text{ is true)}$$

where q is the difference in df

E.g.,

$$LRT - \text{statistic} = -2(\ell_{\hat{\beta}|\beta_j} - \ell_{\hat{\beta}}) \sim^{app} \chi^2_1$$

Likelihood Based Tests and CIs (2)

A $(1-\alpha)100\%$-CI for e.g. β_j based on the LRT statistic is

$$\{b : LRT - \text{statistic} = -2(\ell_{\hat{\beta}|\beta_j=b} - \ell_{\hat{\beta}}) \leq \chi^2_{1,1-\alpha}\}$$

Note that LRT-based CIs must not be symmetric around the MLE

Deviance - Assessing Model Fit

Measure of how good the data is represented by the fitted model (goodness of fit)

Definition

Deviance $D = -2(\ell_M - \ell_F)$ (up to a factor τ^2)

where ℓ_M = log-likelihood of fitted model M

and ℓ_F = maximized log-likelihood under the full model F (n parameters)

Distribution

$D \sim^{\text{app}} \chi^2_{n-p}$ if M is correct (approx. may be poor)

Note

- $D \not\sim^{\text{app}} \chi^2_{n-p}$, not a measure of goodness of fit for binary data
- D = residual SS for normal data and $\sim \tau^2 \chi^2_{n-p}$

Diagnostics - Assessing Adequacy of Fitted Model

Assumptions

- Independent observations
- Specified model is correct, i.e.,
  - $Y_i \sim F(\mu_i)$
  - $\mu_i = g^{-1}(\eta_i)$
  - $\eta_i = \mathbf{x}_i^T \beta$

These assumptions should to be checked $\rightarrow$ residuals

Residuals

Definition

Measure of agreement between the individual observed response and its fitted value

Example

Linear model: $y_i - \hat{\mu}_i$

- estimate for ε
- measure for each observation's contribution to the residual SS or deviance

Use

Residuals form the basis of many diagnostic techniques

Residuals for GLMs (1)

Problematic

No best way of measuring agreement between observed and fitted value

$\rightarrow$ several types of residuals

$\rightarrow$ usefulness dependent on F

Residuals for GLMs (2)

Analogons of $y_i - \hat{\mu}_i$

Raw or response residual $R_i^{(R)} = y_i - \hat{\mu}_i$

Working residual $R_i^{(W)} = R_i^{(R)} / \nu(\hat{\mu}_i) = \tilde{Y}_i$

Pearson residual $R_i^{(P)} = R_i^{(R)} / \sqrt{\frac{\nu(\hat{\mu}_i)}{\omega_i}}$

Standardized Pearson residual $R_i^{(SP)} = R_i^{(R)} / \sqrt{\frac{\hat{\text{var}}(R_i^{(R)})}{\tau^2}}$

$(\hat{\text{var}}(R_i^{(R)}) = \tau^2 \mathbf{\Omega}_{ii}^{-1} [\mathbf{I} - \mathbf{\Omega}^{1/2} \mathbf{X} (\mathbf{X}^T \mathbf{\Omega} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{\Omega}^{1/2}]_{ii} = \tau^2 \mathbf{\Omega}_{ii}^{-1} (1 - \mathbf{P}_{ii})$ evaluated at $\hat{\mu}_i)$

Residuals for GLMs (3)

Analogon of *i*-th's contribution to deviance

Deviance residual $R_i^{(D)} = \text{sgn}(R_i^{(R)})\sqrt{d_i}$

with $\Sigma_{i=1}^{n} d_i = D$

Standardized Deviance residual $R_i^{(SD)} = R_i^{(D)}/(1 - \mathbf{P}_{ii})$

Preferred by many

Diagnostic Plots (1)

Analogous approach to linear model

Tukey-Anscombe

e.g., working residuals vs $\hat{\eta}_i$ or deviance residuals vs $\hat{\eta}_i$ or $\hat{\mu}_i$
$\rightarrow$ there should be no structure (smoothing graph for ease of interpretation)

Linear predictor

plot residuals vs covariates to identify additional relevant covariates or transformation of covariates

Link function

e.g., linearized response (T) vs $\hat{\eta}$

Diagnostic Plots (2)

Distribution F

overdispersion

Outliers, influential observations

Interpretation of diagnostic plots may be very difficult, especially for binary data ($\rightarrow$ example later)

Overdispersion

Larger $\text{var}(Y_i)$ than expected given $F(\mu_i)$

Canonical pdf

$f(y_i | \eta_i, \tau^2, \omega_i) = \exp\left(\frac{\eta_i y_i - d(\eta_i)}{\tau^2}\omega_i\right) h(y_i, \tau^2, \omega_i)$

Dispersion parameter $\tau^2 = 1$

for Bernoulli, Binomial, Poisson distribution

$\Rightarrow \text{var}(Y_i) = \frac{\nu(\mu_i)}{\omega_i}$

Overdispersion

e.g., assuming $\text{var}(Y_i) = \tau^2 \frac{\nu(\mu_i)}{\omega_i}$, $\tau^2 > 1$, estimated from data

$\Rightarrow$ quasi-distribution

Quasi-Likelihood

Assuming a quasi-distribution leads to maximum quasi-likelihood estimators of β

Assumptions

- var$(Y_i) \propto \frac{\nu(\mu_i)}{\omega_i}$, i.e., same relationship as for underlying 'parent' F up to the factor τ^2
- link function g

$\Rightarrow$ same $\hat{\beta}$ and deviance as ML approach based on 'parent' F and g

$\Rightarrow \hat{\tau}^2 = \frac{1}{n-p}\Sigma_{i=1}^{n}(R_i^{(P)})^2$ (or $\frac{1}{n-p}D$)

$\Rightarrow$ cov$(\hat{\beta}) \uparrow$ and CIs $\uparrow$ compared to ML approach

$\Rightarrow$ approximate F-tests instead of LRTs

R Function glm()

glm(formula=response $\sim$ x's, family=, offset=)

For binomial data, response is given as cbind(# successes, # failures) with length n

Starting values must not be provided

$\rightarrow$ examples

Binary Data

Artificial data

$n = 100$ observations of $Y \in \{0, 1\}$

with

$$\mu_i = \frac{\exp(0.5 x_i)}{1 + \exp(0.5 x_i)}$$

Call: glm($y \sim x$, family=binomial)

$\rightarrow R$

Poisson Data

Real data

$n = 32$ observations on the number of faults in rolls of fabric

Possible model for μ_i:

$\mu_i = \mu * \text{length}_i \quad \mu = \#\text{faults/unit length}$

$\eta_i = \log \mu_i = \log \mu + \log \text{length}_i = \beta_0 + \text{offset}_i$

Call: glm(faults∼ 1, offset=log(length), family=poisson)

$\rightarrow R$

Summary and Outlook

Summary

Generalization of linear model to

- any distribution from exponential family of distributions
- any suitable link between μ_i and η_i

$\rightarrow$ we can now deal with more than just normal data

Outlook

Inclusion of random effects