Lecture 2: Introduction to GLM's

Claudia Czado

TU München

Overview

- Introduction to GLM's
- Goodness of fit in GLM's
- Testing in GLM's
- Estimation in GLM's

Introduction to GLM's

In generalized linear models (GLM) we also have independent response variables with covariates.

While in linear models a good scale of the response variables has to combine additivity of the covariate effects with the normality of the errors, including variance homogeneity, GLM's don't need to satisfy these scale requirements.

GLM's allow also to include nonnormal errors such as binomial, Poisson and Gamma errors.

Regression parameters are estimated using maximum likelihood.

The standard reference on GLM's is McCullagh and Nelder (1989).

Components of a GLM:

Response Y_i and independent variables $\boldsymbol{X}_i = (x_{i1}, \cdots x_{ip})$ for $i = 1, \cdots, n$.

1. **Random Component:**
   $Y_i, 1 \leq i \leq n$ independent with density from the exponential family, i.e.

$$f(y; \theta, \phi) = \exp\{\frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)\}.$$

   Here ϕ is a dispersion parameter and functions $b(), a()$ and $c(,)$ are known.

2. **Systematic Component:**
   $\eta_i(\boldsymbol{\beta}) = \boldsymbol{x_i^t \beta} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ linear predictor,
   $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)$ regression parameters

3. **Parametric Link Component:**
   The link function $g(\mu_i) = \eta_i = \boldsymbol{x_i^t \beta}$ combines linear predictor with mean μ_i of y_i. Canonical link function if $\boldsymbol{\theta} = \boldsymbol{\eta}$.

LM as GLM

$$Y_i = \boldsymbol{x_i^t \beta} + \epsilon_i = \mu_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)\ iid, \quad i = 1, .., n,$$

The density of Y_i has exponential family form since

$$\begin{aligned} f(y_i, \mu_i, \sigma) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2}(y_i - \mu_i)^2 \right\} \\ &= \exp\left\{ \frac{y_i\mu_i - \frac{\mu_i^2}{2}}{\sigma^2} - \frac{1}{2}\left[\ln(2\pi\sigma^2) + \frac{y_i^2}{\sigma^2} \right] \right\}. \end{aligned}$$

This implies for $\theta_i = \mu_i$ and $\phi = \sigma^2$

$$b(\theta_i) = \frac{\mu_i^2}{2} = \frac{\theta_i^2}{2}, a(\phi) = \sigma^2, c(y_i, \phi) = -\frac{1}{2}\left[\ln(2\pi\phi) + \frac{y_i^2}{\phi} \right]$$

Further we have the identity as link function, i.e. $g(\mu_i) = \mu_i$.

Expectation and variance in GLM's

When integration and differentiation can be exchanged, mean and variance in a GLM can be represented as

$$\begin{aligned} \mu_i &= E(Y_i) = b'(\theta_i) \\ Var(Y_i) &= a(\phi) \cdot b''(\theta_i). \end{aligned}$$

$V(\theta) := b''(\theta)$ is called the variance function of the GLM.

GLM's implemented in Splus

Distribution	Family	Link	Variance
Normal/Gaussian	`gaussian`	μ	1
Binomial	`binomial`	$\ln(\frac{\mu}{1-\mu})$	$\frac{\mu(1-\mu)}{n}$
Poisson	`poisson`	$\ln(\mu)$	μ
Gamma	`gamma`	$\frac{1}{\mu}$	μ^2
Inverse Normal / Gaussian	`inverse.gaussian`	$\frac{1}{\mu^2}$	μ^3
Quasi	`quasi`	$g(\mu)$	$V(\mu)$

For the binomial family the distribution of $\frac{Y_i}{n_i}$ is used. "Quasi" allows for user defined GLM's.

Link functions:

$$\eta_i = \boldsymbol{x}_i^t \boldsymbol{\beta} \quad \eta_i = g(\mu_i) \quad E(Y_i) = \mu_i \quad g - \text{ monotone } \uparrow$$

Normal: $\mu_i \in \mathbb{R},\ \eta_i \in \mathbb{R}$.
Often $g(\mu) = \mu$ or for $\mu > 0$

$$g_\alpha(\mu) = \begin{cases} \frac{\mu^\alpha - 1}{\alpha} & \alpha \neq 0 \\ \log(\mu) & \alpha = 0 \end{cases} \qquad g_\alpha(\mu) \to \log(\mu),\ \alpha \to 0$$

Box-Cox - transformation

Poisson: $\mu > 0,\ g : \mathbb{R}^+ \to \mathbb{R}$ monotone $\uparrow$

$$g(\mu) = \log(\mu)$$

Link functions:

Binomial: $\mu \in [0,1]$, need $g : [0,1] \to \mathbb{R}$ monotone $\uparrow$
All cdf's $F : \mathbb{R} \to [0,1]$ monotone $\uparrow \Rightarrow g(\mu) := F^{-1}(\mu)$

$a)$	$F(z) = \frac{e^z}{1+e^z}$	Logit link (symmetric, heavy-tailed)
	$\Rightarrow g(\mu) := F^{-1}(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$	Logistic regression
$b)$	$F(z) = \Phi(z)$, $\Phi(z) =$ cdf of $N(0,1)$	Probit Link (symmetric)
	$\Rightarrow g(\mu) = \Phi^{-1}(\mu)$	Probit regression
$c)$	$F(z) = 1 - \exp\{-exp\{z\}\}$	complementary Log-log distribution
	$\Rightarrow g(\mu) = \ln(\ln(1-\mu))$	(nonsymmetric)

Canonical link functions

If $\theta_i = \eta_i \; \forall i$ holds, we call the corresponding link function canonical.

Examples:

Linear model: $\theta_i = \mu_i = \eta_i \Rightarrow$ identity link canonical.

Binomial model: $\theta_i = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \eta_i \Rightarrow$ logistic link canonical

In GLM with canonical link $(\sum_{i=1}^{n} x_{i1}y_i, \ldots, \sum_{i=1}^{n} x_{ip}y_i)$ is sufficient for $(\beta_1, \ldots, \beta_p)^t$.

Goodness of fit in GLM: Deviance

Want to estimate Y_i by $\hat{\mu}_i$.
For n data points we can estimate n parameters.

Null model:

$$\hat{\mu}_i := \overline{Y} \quad \forall i, \ \overline{Y} := \frac{1}{n}\sum_{i=1}^{n} Y_i$$

one parameter $\longrightarrow$ too simple.

Saturated model:

$$\hat{\mu}_i := Y_i \quad \forall i$$

no error, n parameters used, no explanation of data possible.

Loglikelihood in GLM with $\eta_i = g(\mu_i),\ \theta_i = h(\mu_i) \quad (i = 1, \ldots, n)$

$$
\begin{aligned}
l(\boldsymbol{\beta}, \phi, \mathbf{y}) &= \sum_{i=1}^{n} \left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} - c(y_i, \phi) \right] \\
&= \sum_{i=1}^{n} \left[\frac{y_i h(\mu_i) - b(h(\mu_i))}{a(\phi)} - c(y_i, \phi) \right] \\
&= l(\boldsymbol{\mu}, \phi, \mathbf{y}) \qquad \text{“mean parametrization”}
\end{aligned}
$$

$l(\hat{\boldsymbol{\mu}}, \phi, \mathbf{y}) :=$ log likelihood maximized over $\boldsymbol{\mu}$ (ϕ known) $\hat{\mu}_i := g^{-1}(\boldsymbol{x_i}^t \hat{\boldsymbol{\beta}})$
$l(\mathbf{y}, \phi, \mathbf{y}) :=$ log likelihood attainable in saturated model i.e. $\hat{\mu}_i = Y_i \quad \forall i$

$$
\begin{aligned}
\Rightarrow \quad & -2[l(\hat{\boldsymbol{\mu}}, \phi, \mathbf{y}) - l(\mathbf{y}, \phi, \mathbf{y})] \\
= \quad & 2 \sum_{i=1}^{n} \frac{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)}{a(\phi)}, \qquad \text{where } \hat{\theta}_i := h(\hat{\mu}_i),\ \tilde{\theta}_i := h(Y_i)
\end{aligned}
$$

If $a(\phi) = \frac{\phi}{\omega} \Rightarrow$

$$
-2[l(\hat{\boldsymbol{\mu}}, \phi, \mathbf{y}) - l(\mathbf{y}, \phi, \mathbf{y})] = 2\omega \sum_{i=1}^{n} \frac{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)}{\phi} =: \frac{D(\mathbf{y}, \hat{\boldsymbol{\mu}})}{\phi} \quad \text{deviance}
$$

Ex: Deviance in Linear and Binomial Models

Linear model:

$$
\begin{aligned}
l(\boldsymbol{\beta}, \phi, \mathbf{y}) &= -\sum_{i=1}^{n} \tfrac{1}{2\sigma^2}(y_i - \mu_i)^2 - \tfrac{n}{2}\ln(2\pi\sigma^2), \quad \mu_i = \boldsymbol{x_i^t}\boldsymbol{\beta}, \quad \phi = \sigma^2 \\
&\Rightarrow -2[l(\hat{\boldsymbol{\mu}}, \phi, \mathbf{y}) - l(\mathbf{y}, \phi, \mathbf{y})] = \tfrac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \hat{\mu}_i)^2 \\
&\Rightarrow D(\mathbf{y}, \hat{\boldsymbol{\mu}}) := \sum_{i=1}^{n}(Y_i - \hat{\mu}_i)^2
\end{aligned}
$$

Binomial model:

$Y_i \sim$ binomial(n_i, p_i) independent $\quad \hat{\mu}_i := n_i\hat{p}_i \quad \hat{p}_i =$ MLE of p_i

$$
D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n}\left\{y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) + (n_i - y_i)\ln\left(\frac{n_i - y_i}{n_i - \hat{\mu}_i}\right)\right\}
$$

In binomial regression models is not $\{Y_i, i = 1, \ldots, n\}$ a GLM, but $\{\frac{Y_i}{n_i}, i = 1, \ldots, n\}$ is a GLM.

Generalized Pearson Statistic

$$\boxed{\chi^2 := \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}}$$

$$\begin{aligned} V(\hat{\mu}_i) &= \text{estimated variance function} \\ &= b''(\hat{\theta}_i)\big|_{\hat{\theta}_i = h(\hat{\mu}_i)} \end{aligned}$$

Examples:

Normal: $Y_i \sim N(\mu_i, \sigma^2)$ ind.

$$\Rightarrow \theta_i = \mu_i \qquad b(\mu_i) = \frac{\mu_i^2}{2} \Rightarrow b''(\hat{\mu}_i) = 1$$

$$\Rightarrow \chi^2 = \sum_{i=1}^{n} (Y_i - \hat{\mu}_i)^2 = D(\mathbf{y}, \hat{\boldsymbol{\mu}}).$$

Logistic Regression: $Y_i \sim \text{bin}(n_i, p_i)$ ind.

$$p_i = \frac{e^{\theta_i}}{1+e^{\theta_i}} \Rightarrow \mu_i = n_i p_i = n_i \frac{e^{\theta_i}}{1+e^{\theta_i}}$$

$$b(\theta_i) = n_i \ln(1+e^{\theta_i}) \Rightarrow b''(\theta_i) = n_i \frac{e^{\theta_i}}{(1+e^{\theta_i})^2}$$

$$\Rightarrow b''(p) = n_i p_i (1-p_i) = \mu_i (1 - \frac{\mu_i}{n})$$

$$\Rightarrow V(\hat{\mu}_i) = \hat{\mu}_i (1 - \frac{\hat{\mu}_i}{n}) = n_i \hat{p}_i (1 - \hat{p}_i)$$

$$\Rightarrow \chi^2 = \sum_{i=1}^{n} \frac{(y_i - n_i \hat{p}_i)^2}{n_i \hat{p}_i (1-\hat{p}_i)}.$$

Asymptotic distribution of Deviance and Pearson statistic

1) Normal: $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad X \in \mathbb{R}^{n \times p} \qquad \boldsymbol{\epsilon} \sim N_n(\mathbf{0}, I_n \sigma^2)$

$$\Rightarrow D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = \chi^2 = \sum_{i=1}^{n} (Y_i - \hat{\mu}_i)^2 \sim \sigma^2 \chi^2_{n-p}$$

2) For all other GLM's we have

$$\boxed{\begin{array}{lll} D(\mathbf{y}, \hat{\boldsymbol{\mu}}) & \xrightarrow{L} & \phi \chi^2_{n-p},\ n \to \infty \qquad p = \#\text{ of unknown parameters} \\ \chi^2 & \xrightarrow{L} & \phi \chi^2_{n-p},\ n \to \infty \end{array}}$$

Proof: deviance is equivalent to a likelihood ratio statistic and bχ^2 to the Wald statistic for which general asymptotic results are available (see e.g: Rao (1973))

3) For finite n one has no theoretical results whether D or χ^2 is performing better.

Nested linear models

Model			**SSE**
$M_1:$	$\mathbf{Y} = \mathbf{1_n}\beta_0 + \boldsymbol{\epsilon}$	(null model)	SSE_0
$M_2:$	$\mathbf{Y} = X_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$		$SSE(X_1)$
$M_3:$	$\mathbf{Y} = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$	(full model)	$SSE(X_1, X_2)$

$$X \in \mathbb{R}^{n\times p} \quad X_1 \in \mathbb{R}^{n\times p_1} \quad X_2 \in \mathbb{R}^{n\times p_2} \quad p_1 + p_2 = p$$

Recall: $SSE_0 = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$

$SSE(X_1) = ||\mathbf{Y} - X_1\hat{\boldsymbol{\beta}}_1^2||^2$ $\qquad \hat{\boldsymbol{\beta}}_1^2 = MLE$ in M_2

$SSE(X_1, X_2) = ||\mathbf{Y} - X_1\hat{\boldsymbol{\beta}}_1^3 - X_2\hat{\boldsymbol{\beta}}_2^3||^2$ $\qquad \hat{\boldsymbol{\beta}}_1^3, \hat{\boldsymbol{\beta}}_2^3 = MLE$ in M_3

Analysis of deviance

Let $M_1 \subset M_2 \subset \ldots \subset M_r$ a sequence of nested models with $M_1 =$ null model and $M_r =$ saturated model. That means that all covariates of M_i are contained in M_s for $s \geq i+1 \quad \forall i$.

$$
\begin{array}{llll}
\textbf{Model} & \textbf{Deviance} & & \\
M_1 \text{ (null model)} & Dev_1 & & \\
 & & > & Dev_1 - Dev_2 \\
M_2 & Dev_2 & & \\
\vdots & \vdots & & \vdots \\
M_{r-1} & Dev_{r-1} & & \\
 & & > & Dev_{r-1} - Dev_r \\
M_r \text{ (saturated model)} & Dev_r & &
\end{array}
$$

-Difference $Dev_i - Dev_{i+1}$ is considered as the variation explained by M_{i+1} minus the variation explained by $M_1, \ldots, M_i$. The variations explained by $M_{i+2}, \ldots, M_r$ are disregarded.

-Analysis of deviance depends on the order of covariates added to the models

-Since there is no exact distribution theory, it is used as a screening method to identify important covariates

Statistical hypothesis tests

Residual deviance test

$$H_0: \eta_i = g(\mu_i)\ \forall i \qquad\qquad H_1: \text{ not } H_0$$

$$\text{Reject } H_0 \Leftrightarrow Dev > \chi^2_{n-q,1-\alpha} \text{ is an asymptotic } \alpha-\text{level test}$$

Problem: Often one is interested to use this as a goodness-of-fit test, i.e. one wants to accept H_0. However the power function is unknown.

Partial deviance test

$$\begin{array}{lll} \eta = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 & \text{Model F with deviance } D_F & \boldsymbol{\beta}_1 \in \mathbb{R}^{p_1},\ \boldsymbol{\beta}_2 \in \mathbb{R}^{p_2} \\ \eta = X_1\boldsymbol{\beta}_1 & \text{Model R with deviance } D_R & p_1 + p_2 = p \end{array}$$

$$H_0: \boldsymbol{\beta}_2 = \mathbf{0} \qquad\qquad H_1: \boldsymbol{\beta}_2 \neq \mathbf{0}$$

$$\text{Reject } H_0 \Leftrightarrow D_R - D_F > \chi^2_{p-p_2=p_1,1-\alpha}$$

Residuals

Pearson residuals: $r_i^P := \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$ $\qquad i = 1, \ldots, n$

Deviance residuals: $r_i^D := sign(y_i - \hat{\mu}_i)\sqrt{d_i}$

$Dev = \sum_{i=1}^{n} d_i \qquad d_i =$ deviance contribution of i_{th} obs.

$$sign(x) = \begin{cases} 1 & x > 0 \\ -1 & x \leq 0 \end{cases}$$

- $\chi^2 = \sum_{i=1}^{n} (r_i^P)^2, \qquad Dev = \sum_{i=1}^{n} (r_i^D)^2$

- For nonnormal GLM Pearson residuals are skewed. Better to use Anscombe residuals.

Maximum Likelihood Estimation (MLE) in GLM's

Loglikelihood for obs. i:

$$l_i(y_i, \mu_i, \phi) = \frac{[y_i\theta_i - b(\theta_i)]}{a(\phi)} + c(y_i, \phi)$$

where $g(\mu_i) = \eta_i \quad \mu_i = E(Y_i) = h(\theta_i) \quad \eta_i = \boldsymbol{x_i^t}\boldsymbol{\beta} \quad \boldsymbol{\beta} \in \mathbb{R}^p$
Since

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \mu_i} \frac{d\mu_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta_j} \qquad \text{we need}$$

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij} \qquad \mathbf{x_i} = (x_{i1}, \ldots, x_{ip})^t$$

$$\frac{\partial l_i}{\partial \mu_i} = \frac{\partial l_i}{\partial \theta_i} \Big/ \frac{\partial \mu_i}{\partial \theta_i} \overset{\mu_i = b'(\theta_i)}{=} \frac{y_i - b'(\theta_i)}{a(\phi)} / b''(\theta_i) = \frac{y_i - \mu_i}{V_i}, \quad \text{since } V_i = Var(Y_i) = a(\phi) \cdot b''(\theta_i)$$

$$\Rightarrow \frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \mu_i} \frac{d\mu_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \frac{y_i - \mu_i}{V_i} \frac{d\mu_i}{d\eta_i} x_{ij}$$

For n independent observations:

$$l(\mathbf{y}, \boldsymbol{\beta}) := \sum_{i=1}^{n} l_i(y_i, \mu_i, \phi)$$

$$\Rightarrow \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i(y_i,\mu_i,\phi)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i-\mu_i}{V_i} \frac{d\mu_i}{d\eta_i} x_{ij} \qquad V_i = Var(Y_i)$$

$$\mu_i = E(Y_i)$$

Let $W_i := \frac{1}{V_i\left(\frac{d\eta_i}{d\mu_i}\right)^2} = \left(\frac{d\mu_i}{d\eta_i}\right)^2 / V_i$ since $\frac{d\eta_i}{d\mu_i} = 1 / \frac{d\mu_i}{d\eta_i}$

$$\Rightarrow \quad \boxed{s_j(\boldsymbol{\beta}) := \frac{\partial l(\mathbf{y},\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} W_i(y_i - \mu_i)\frac{d\eta_i}{d\mu_i} x_{ij} = 0 \quad j = 1,\ldots,p}$$

score equations

Newton Raphson Method

Want to solve $f(\mathbf{x}) = \begin{bmatrix} f_1(x_1, \ldots, x_n) \\ \vdots \\ f_n(x_1, \ldots, x_n) \end{bmatrix} = \mathbf{0}$. Let $\mathbf{x} = \boldsymbol{\xi}$ the solution and $\mathbf{x_0}$ a value close to $\boldsymbol{\xi}$. Then we have with first order Taylor expansion around $\mathbf{x_0}$:

$$\mathbf{0} = f(\boldsymbol{\xi}) \approx f(\mathbf{x_0}) + \underbrace{Df(\mathbf{x_0})}_{\in \mathbb{R}^{n \times n}} \underbrace{(\boldsymbol{\xi} - \mathbf{x_0})}_{\in \mathbb{R}^n} \quad \text{where}$$

$$Df(\mathbf{x_0}) := \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}_{\mathbf{x}=\mathbf{x_0}}$$

$$\Rightarrow \quad \boldsymbol{\xi} = \mathbf{x_0} - [Df(\mathbf{x_0})]^{-1} f(\mathbf{x_0})$$

Newton Raphson method is an iterative algorithm with $\mathbf{x_0}$ a starting value and

$$\mathbf{x_{i+1}} = \mathbf{x_i} - [Df(\mathbf{x_i})]^{-1} f(\mathbf{x_i})$$

There are general convergence results available.

To solve $s(\boldsymbol{\beta}) = (s_1(\boldsymbol{\beta}), \ldots, s_p(\boldsymbol{\beta}))^t = \mathbf{0}$ we need

$$H(\boldsymbol{\beta}) := \begin{bmatrix} \frac{\partial s_1}{\partial \beta_1} & \cdots & \frac{\partial s_1}{\partial \beta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial s_p}{\partial \beta_1} & \cdots & \frac{\partial s_p}{\partial \beta_p} \end{bmatrix} = \begin{bmatrix} \frac{\partial^2 l}{\partial \beta_1 \partial \beta_1} & \cdots & \frac{\partial^2 l}{\partial \beta_1 \partial \beta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 l}{\partial \beta_p \partial \beta_1} & \cdots & \frac{\partial^2 l}{\partial \beta_p \partial \beta_p} \end{bmatrix}$$

$=$ Hessian matrix $= -$ observed information matrix

$$\begin{aligned} \frac{\partial^2 l}{\partial \beta_s \partial \beta_r} &= \frac{\partial}{\partial \beta_s} \left[\sum_{i=1}^n \frac{y_i - \mu_i}{V_i} \frac{d\mu_i}{d\eta_i} x_{ir} \right] \\ &= \sum_{i=1}^n (y_i - \mu_i) \frac{\partial}{\partial \beta_s} \left[V_i^{-1} \frac{d\mu_i}{d\eta_i} x_{ir} \right] + \sum_{i=1}^n V_i^{-1} \frac{d\mu_i}{d\eta_i} x_{ir} \frac{\partial}{\partial \beta_s} (y_i - \mu_i) \end{aligned}$$

Further $\frac{\partial}{\partial \beta_s}(y_i - \mu_i) = -\frac{d\mu_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta_s} = -\frac{d\mu_i}{d\eta_i} x_{is}$. Since $\frac{\partial^2 l}{\partial \beta_s \partial \beta_r}$ depends on $\mathbf{Y}$ in general we use $E(\frac{\partial^2 l}{\partial \beta_s \partial \beta_r})$ instead. Note that for canonical link we have $\frac{\partial^2 l}{\partial \beta_s \partial \beta_r} = E(\frac{\partial^2 l}{\partial \beta_s \partial \beta_r})$.

Expected information matrix

$A(\boldsymbol{\beta}) := \left[-E(\frac{\partial^2 l}{\partial\beta_s \partial\beta_r})\right]_{s,r=1,\ldots,p}$ is called the expected information matrix

One can show that

$$-E\frac{\partial^2 l}{\partial\beta_s \partial\beta_r} = E\frac{\partial l}{\partial\beta_s}\frac{\partial l}{\partial\beta_r}$$

$$\stackrel{E\mathbf{s}(\beta)=\mathbf{0}}{\Rightarrow} \quad A(\boldsymbol{\beta}) = cov\,\mathbf{s}(\boldsymbol{\beta})$$

The both expressions are used as definition for the expected information matrix in the literature.

Fisher scoring method

$$
\begin{aligned}
E\left(\frac{\partial^2 l}{\partial\beta_s\partial\beta_r}\right) &= E\underbrace{(\cdots)}_{=0} + E\left(-\sum_{i=1}^{n}\underbrace{V_i^{-1}\left(\frac{d\mu_i}{d\eta_i}\right)^2}_{W_i} x_{is}\, x_{ir}\right) \\
&= -\sum_{i=1}^{n} W_i\, x_{is}\, x_{ir}
\end{aligned}
$$

$$
\Rightarrow A(\boldsymbol{\beta}) := \left[-E\left(\frac{\partial^2 l}{\partial\beta_s\partial\beta_r}\right)\right]_{s,r=1,\ldots,p} = +X^t W X \in \mathbb{R}^{p\times p},
$$

where $W = diag(W_1, \ldots, W_n) \in \mathbb{R}^{n\times n}$ and $X \in \mathbb{R}^{n\times p}$.

Fisher scoring method: let $\boldsymbol{\beta}^r$ the current estimation to the solution of $s(\boldsymbol{\beta}) = 0$, the new estimation value is given by

$$
\boldsymbol{\beta}^{r+1} = \boldsymbol{\beta}^r + A^{-1}(\boldsymbol{\beta}^r)s(\boldsymbol{\beta}^r)
$$

Fisher scoring as iterative weighted least squares

Since $\underbrace{A(\boldsymbol{\beta}^r)\boldsymbol{\beta}^{r+1}}_{\in\mathbb{R}^p} = A(\boldsymbol{\beta}^r)\boldsymbol{\beta}^r + s(\boldsymbol{\beta}^r)$

$$
\begin{aligned}
\Rightarrow (A(\boldsymbol{\beta}^r)\boldsymbol{\beta}^{r+1})_j \quad &= \sum_{s=1}^{p} A_{js}(\boldsymbol{\beta}^r)\beta_s^r + s_j(\boldsymbol{\beta}^r) \qquad\qquad g(\mu_i^r) = \eta_i^r = \boldsymbol{x_i^t}\boldsymbol{\beta}^r \\
&= \sum_{s=1}^{p}\sum_{i=1}^{n} W_i^r x_{ij} x_{is} \beta_s^r + \sum_{i=1}^{n} W_i^r (y_i - \mu_i^r)\tfrac{d\eta_i^r}{d\mu_i^r} x_{ij} \\
&= \sum_{i=1}^{n} W_i^r x_{ij}\ [\underbrace{\sum_{s=1}^{p} x_{is} b_s^r}_{\eta_i^r} + (y_i - \mu_i^r)\tfrac{d\eta_i^r}{d\mu_i^r}]
\end{aligned}
$$

Define the adjusted dependent variable

$$
Z_i^r := \eta_i^r + (y_i - \mu_i^r)\frac{d\eta_i^r}{d\mu_i^r} \Rightarrow
$$

$$
(A(\boldsymbol{\beta}^r)\boldsymbol{\beta}^{r+1})_j = \sum_{i=1}^{n} W_i^r x_{ij} Z_i^r
$$

On the other side we have

$$\begin{aligned}(A(\boldsymbol{\beta}^r)\boldsymbol{\beta}^{r+1})_j &= \sum_{s=1}^{p} A_{js}(\boldsymbol{\beta}^r)\beta_s^{r+1} = \sum_{s=1}^{p}\sum_{i=1}^{n} W_i^r x_{ij} x_{is} \beta_s^{r+1} \\ &= \sum_{i=1}^{n} W_i^r x_{ij} \underbrace{\sum_{s=1}^{p} x_{is} b_s^{r+1}}_{\eta_i^{r+1}}\end{aligned}$$

Therefore we have

$$\sum_{i=1}^{n} W_i^r x_{ij} Z_i^r = \sum_{i=1}^{n} W_i^r x_{ij} \eta_i^{r+1} \qquad \forall j = 1, \ldots, p$$

or in matrix form: $X^t W^r \mathbf{Z}^r = X^t W^r X \boldsymbol{\beta}^{r+1}$.

These equations correspond to the normal equations of a weighted least squares estimation with response Z_i^r, covariates $\mathbf{x_1}, \ldots, \mathbf{x_p}$ and weights $(W_i^r)^{-1}$. Therefore we speak of the IWLS (iterated weighted least square).

IWLS algorithm

Step 1: Let $\boldsymbol{\beta}^r$ the current estimate of $\hat{\boldsymbol{\beta}}$, determine

- $\hat{\eta}_i^r := \mathbf{x_i}^t \boldsymbol{\beta}^r \quad i = 1, \ldots, n$ (current linear predictors)
- $\hat{\mu}_i^r := g^{-1}(\hat{\eta}_i^r)$ (current fitted means)
- $\hat{\theta}_i^r := h^{-1}(\hat{\mu}_i^r)$
- $V_i^r := a(\phi) \cdot b''(\theta_i)|_{\theta_i = \hat{\theta}_i^r}$
- $Z_i^r := \hat{\eta}_i^r + (y_i - \hat{\mu}_i^r) \left(\frac{d\eta_i}{d\mu_i}|_{\eta_i = \hat{\eta}_i^r} \right)$ (adjusted dependent variable)
- $W_i^r := \left[V_i^r \left(\frac{d\eta_i}{d\mu_i}|_{\eta_i = \hat{\eta}_i^r} \right)^2 \right]^{-1}$

Step 2: Regress Z_i^r on $x_{i1}, \ldots, x_{ip}$ with weights $(W_i^r)^{-1}$ to obtain new estimate $\boldsymbol{\beta}^{r+1}$ and continue with step 1 until $||\boldsymbol{\beta}^r - \boldsymbol{\beta}^{r+1}||$ sufficiently small.

Remarks

1) Z_i^r is the linearized form of the link function at y_i, since

$$g(y_i) \approx \underbrace{g(\mu_i^r)}_{\eta_i^r} + (y_i - \mu_i^r)\underbrace{g'(\mu_i^r)}_{\frac{d\eta_i^r}{d\mu_i^r}}$$

$\Rightarrow Z_i^r \approx g(y_i)$ up to the first order

2) $Var(Z_i^r) \approx \underbrace{Var(Y_i - \mu_i^r)}_{V_i} \cdot \left(\frac{d\eta_i^r}{d\mu_i^r}\right)^2 = (W_i^r)^{-1}$

if η_i^r, μ_i^r are considered fixed and known.

3) Often one can use the data as starting values, i.e.

$$\hat{\mu}_i^0 = y_i \quad \Rightarrow \quad \hat{\eta}_i^0 = g(\hat{\mu}_i^0)$$

If $Y_i = 0$ in the binomial case one needs to change the start values to avoid log(0).

References

McCullagh, P. and J. Nelder (1989). *Generalized linear models*. Chapman & Hall.

Rao, C. (1973). *Linear Statistical Inference and Its Applications, 2nd Ed.* New York: Wiley.