Generalized Linear Models

Exponential Family of Distributions

A random variable Y has a distribution in the exponential family if its density takes the form:

$$f(Y;\theta,\phi) = exp\left\{\frac{Y\theta - b(\theta)}{a(\phi)} + c(Y,\phi)\right\}$$

for some specific functions $a(\cdot)$, $b(\cdot)$, $c(\cdot)$.

$\theta-$ canonical (natural) parameter (parameter of interest).

$\phi-$ scale (dispersion) parameter (nuisance parameter).

$a(\phi)$ often has the form $a(\phi) = \frac{\phi}{w}$ for some known weight w.

This family includes several important distributions such as normal, binomial, poisson, gamma, inverse gaussian......

Remarks:

We here assume the support of $f(Y;\theta,\phi) = \{y : f(y;\theta,\phi) > 0\}$ does not depend on θ.

- $f(Y;\theta,\phi) = exp\left\{\frac{Y \cdot \theta - b(\theta)}{a(\phi)} + c(Y,\phi)\right\}$

- Loglikelihood: $\ell(\theta) = ln\{f(Y;\theta)\}$

- Score: $U(\theta) = \frac{\partial \ell}{\partial \theta}$

- Observed information: $J(\theta) = -\frac{\partial^2 \ell}{\partial\theta\partial\theta^T}$

- Expected information: $I(\theta) = -E(\frac{\partial^2 \ell}{\partial\theta\partial\theta^T}) > 0$

- Equalities:

$$E\{U(\theta)\} = E\left(\frac{\partial \ell}{\partial \theta}\right) = 0$$

$$\begin{aligned} Cov\{U(\theta)\} &= Cov\left(\frac{\partial \ell}{\partial \theta}\right) \\ &= E\left(\frac{\partial \ell}{\partial \theta}\frac{\partial \ell}{\partial \theta^T}\right) = -E\left(\frac{\partial^2 \ell}{\partial\theta\partial\theta^T}\right) = I(\theta) \end{aligned}$$

Mean and Variance in the Exponential Family:

$$E(Y) = \mu = b'(\theta)$$

$$var(Y) = a(\phi)b''(\theta) = a(\phi)v(\mu) = \phi a^{-1}v(\mu)$$

Why?

$$\frac{\partial \ell}{\partial \theta} = \frac{1}{a(\phi)}[Y - b'(\theta)] \rightarrow E(Y) = b'(\theta)$$

$$\frac{\partial^2 \ell}{\partial \theta^2} = -\frac{b''(\theta)}{a(\phi)} \rightarrow var(Y) = a(\phi)b''(\theta)$$

Mean/Variance Relationship:

The exponential family assumes the variance $v(\mu)$ is a function of the mean μ.

Example 1 (Normal):

$$
\begin{aligned}
f(Y;\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y-\mu)^2}{2\sigma^2}} \\
&= exp\{\frac{Y\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{Y^2}{2\sigma^2} - \frac{1}{2} ln 2\pi\sigma^2\}
\end{aligned}
$$

where $\theta = \mu$, $b(\theta) = \frac{\theta^2}{2}$, $\phi = \sigma^2$, $c(Y,\phi) = -\frac{Y^2}{2\sigma^2} - \frac{1}{2} ln 2\pi\sigma^2$,

$E(Y) = \mu$, $var(Y) = \sigma^2 = \phi \cdot 1$, $v(\mu) = 1$.

Example 2 (Binomial):

$Z \sim Bin(m, p)$, $Y = \frac{Z}{m}$.

$$
\begin{aligned}
f(Y; p) &= \begin{pmatrix} m \\ Z \end{pmatrix} p^Z (1-p)^{m-Z} \\
&= \begin{pmatrix} m \\ mY \end{pmatrix} p^{mY} (1-p)^{m(1-Y)} \\
&= exp\{\frac{Y ln(\frac{p}{1-p}) + ln(1-p)}{\frac{1}{m}} + ln \begin{pmatrix} m \\ mY \end{pmatrix}\}
\end{aligned}
$$

where

$\theta = ln(\frac{p}{1-p})$, $b(\theta) = -ln(1-p) = ln(1+e^\theta)$, $a = m$, $\phi = 1$,

$E(Y) = \mu = p$, $Var(Y) = \frac{1}{m} p(1-p) = \frac{1}{m}\mu(1-\mu)$, $v(\mu) = \mu(1-\mu)$.

Example 3 (Poisson):

$$f(Y;\mu) = \frac{\mu^Y e^{-\mu}}{Y!} = exp^{\{Y ln\mu - \mu - lnY!\}}$$

$\theta = ln\mu$, $b(\theta) = \mu$, $a = 1$, $\phi = 1$

$$\begin{aligned} E(Y) &= \mu \\ Var(Y) &= \mu \\ v(\mu) &= \mu \end{aligned}$$

Generalized Linear Models (GLMs)

Outcome: $Y \sim f(Y;\theta) \in$ exponential family

Covariates: $X_1, \cdots, X_p$

Objective: To model the relationship between the mean of Y and $X's$.

Three components of a GLM: random component, systematic component, link between μ and η.

1. Random component:

Observations $\mathbf{Y} = (Y_1, \cdots, Y_n)^T$ are independent, and follow a distribution in the canonical exponential family:

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

$\Rightarrow$

$$E(Y_i) = \mu_i$$

2. Systematic component:

A linear predictor $\boldsymbol{\eta}$ is specified as a linear function of a set of covariates $\mathbf{X}_1, \cdots, \mathbf{X}_p$.

$$\eta_i = \mathbf{X}_i^T \boldsymbol{\beta} \Rightarrow \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$

where $\boldsymbol{\eta} = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} \mathbf{X}_1^T \\ \vdots \\ \mathbf{X}_n^T \end{pmatrix}$.

3. Link between $\boldsymbol{\mu}$ and $\boldsymbol{\eta}$:

$$g(\mu_i) = \eta_i \iff g(\boldsymbol{\mu}) = \boldsymbol{\eta}$$

where $g(\cdot)$ is a monotonic differential function.

$g(\cdot)$ is called the link function.

How do GLMs extend classical linear models?

1. The distribution of Y may come from a distribution in the exponential family other than normal.

2. The mean of Y is related to $\mathbf{X^T}\boldsymbol{\beta}$ via a link function $g(\cdot)$ other than the identity function, i.e., a transformed mean of Y is equal to $\mathbf{X}^T\boldsymbol{\beta}$

$$g(\mu_i) = \mathbf{X_i^T}\boldsymbol{\beta}$$

Example 1: (Linear model)

$$Y_i \overset{i.i.d}{\sim} N(\mu_i, \sigma^2)$$

$$\mu_i = \mathbf{X}_i^T \boldsymbol{\beta}$$

- Random component: $Y_i \sim N(\mu_i, \sigma^2)$
- Systematic component: $\eta_i = \mathbf{X}_i^T \boldsymbol{\beta}$
- Link: $\mu_i = \eta_i$, $-\infty < \mu_i < \infty$, $g(\mu_i) = \mu_i =$ identity link.

Example 2: (Logistic regression) – for binary/proportion data

$m_i Y_i \sim Binomial(m_i, \mu_i)$

$ln\left(\frac{\mu_i}{1-\mu_i}\right) = \mathbf{X}_i^T \boldsymbol{\beta}$

- Random component: $m_i Y_i \sim Binomial(m_i, \mu_i)$

- Systematic component: $\eta_i = \mathbf{X}_i^T \boldsymbol{\beta}$

- Link: $g(\mu_i) = ln\left(\frac{\mu_i}{1-\mu_i}\right) =$ logit link.

Interpretation of β_k in example 2 (often, not always):

β_k: $ln(OR)$ for one unit increase in X_k given the other $X's$ are held constant.

Other link functions of interest:

- Probit: $\Phi^{-1}(\mu_i) = \mathbf{X}_i^T \boldsymbol{\beta}$. Note $ln\left(\frac{\mu}{1-\mu}\right) \approx 1.70\Phi^{-1}(\mu)$.

- Complementary log log: $ln\{-ln(1-p)\} = \mathbf{X}_i^T \boldsymbol{\beta}$

Example 3: (Poisson regression) – Log linear model for count data.

$Y_i \sim Poisson(\mu_i)$

$ln\mu_i = \mathbf{X}_i^T\boldsymbol{\beta}$

- Random component: $Y_i \sim Poisson(\mu_i)$

- Systematic component: $\eta_i = \mathbf{X}_i^T\boldsymbol{\beta}$

- Link: $g(\mu_i) = ln\,(\mu_i) =$ log link.

In many studies, $Y_i \sim Poisson(N_i, \lambda_i)$, where N_i =person-year, λ_i =incidence rate.

$$ln\lambda_i = \mathbf{X}_i^T \boldsymbol{\beta}$$

$$ln\mu_i = ln(N_i) + \mathbf{X}_i^T \boldsymbol{\beta}$$

$ln(N_i)$ is an offset.

$\beta_k = lnRR$ for one unit increase in X_k given the other $X's$ are held constant.

Example of Logistic Regression (Low birth weight study)

A case-control study of mother's weight (in pounds) on the risk of delivering a low birth weight baby.

$$\text{Outcome: } \begin{cases} 0 & : \text{ birth weight } \geq 2500g \quad \text{(normal)} \\ 1 & : \text{ birth weight } < 2500g \quad \text{(abnormal)} \end{cases}$$

Covariates:

- LWT: mother's weight at the last menstrual period.
- AGE: mother's age.

μ: probability of delivering a low birth weight baby in the case-control sample.

Logistic model:

$$logit(\mu) = ln\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1 LWT + \beta_2 AGE$$

Using the ML method,

$$logit(\mu) = ln\left(\frac{\mu}{1-\mu}\right) = 1.74 - 0.01 \cdot LWT - 0.04 \cdot AGE$$

-0.01	$ln(\widehat{OR})$ associated with $1lb$ increase in LWT at any given age.
$e^{-0.01} = 0.99$	$\widehat{OR}$ associated with $1lb$ increase in LWT at any given age.
$e^{-0.01\times 10} = 0.90$	$\widehat{OR}$ associated with $10lb$ increase in LWT at any given age.

Canonical Link

Recall GLM:

- $Y_1, \cdots, Y_n$ indep, and $f(Y) = exp\{\frac{Y\theta - b(\theta)}{a(\phi)} + c(Y, \phi)\}$.

- $\mu = E(Y) = b'(\theta)$, $var(Y) = a(\phi)b''(\theta) = a(\phi)v(\mu)$.

- $g(\mu_i) = \eta_i = \mathbf{X}_i^T \boldsymbol{\beta}$.

Canonical Link:

$g(\cdot)$ is a canonical link if $g(\cdot)$ satisfies $\theta_i = \eta_i$, i.e., $g(\cdot) = b'^{-1}(\cdot)$.

Properties of Canonical Link:

- $g'(\mu) = \frac{1}{v(\mu)}$

- $\mathbf{X^T Y} = \sum_{i=1}^{n} \mathbf{X_i Y_i}$ is a sufficient statistics for $\boldsymbol{\beta}$.

Examples:

Distribution	Model	Canonical link	Remarks
Normal	$\mu = \mathbf{X}^T\boldsymbol{\beta}$	$g(\mu) = \mu$	$\theta = \mu$
Binomial	$ln(\frac{\mu}{1-\mu}) = \mathbf{X}^T\boldsymbol{\beta}$	$g(\mu) = ln(\frac{\mu}{1-\mu})$	$\theta = ln(\frac{\mu}{1-\mu})$
Poisson	$ln(\mu) = \mathbf{X}^T\boldsymbol{\beta}$	$g(\mu) = ln(\mu)$	$\theta = ln(\mu)$
Gamma	$\frac{1}{\mu} = \mathbf{X}^T\boldsymbol{\beta}$	$g(\mu) = \frac{1}{\mu}$	$\theta = \frac{1}{\mu}$

Estimation of β in GLMs

Data: n independent observations $(Y_i, \mathbf{X}_i)$ where Y_i is an outcome and $\mathbf{X}_i$ is a $(p+1) \times 1$ covariate vector.

Pdf of Y_i:

$$f(Y_i) = exp\{\frac{Y_i\theta_i - b(\theta_i)}{\phi a_i^{-1}} + c(Y_i, \phi)\}$$

GLM:

$$g(\mu_i) = \mathbf{X}_i^T \boldsymbol{\beta}$$

where $\mu_i = E(Y_i) = b'(\theta_i)$.

Q: MLE of $\boldsymbol{\beta}$?

loglikelihood of $\boldsymbol{\beta}$:

$$\ell(\boldsymbol{\beta}) \;=\; \sum_{i=1}^{n} \ell_i(Y_i; \boldsymbol{\beta}, \phi) = \sum_{i=1}^{n} \{\frac{Y_i\theta_i - b(\theta_i)}{\phi a_i^{-1}} + c(Y_i, \phi)\}$$

$$\Rightarrow \theta_i \xrightarrow{\mu_i = b'(\theta_i)} \mu_i \xrightarrow{\eta_i = g(\mu_i)} \eta_i \xrightarrow{\eta_i = \mathbf{X}_i^T\boldsymbol{\beta}} \boldsymbol{\beta}$$

Score Equation of $\boldsymbol{\beta}$:

$$\begin{aligned} U(\boldsymbol{\beta}) &= \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \boldsymbol{\beta}} \\ &= \frac{1}{\phi}\sum_{i=1}^{n} \frac{1}{a_i^{-1} v(\mu_i)[g'(\mu_i)]^2} g'(\mu_i)(Y_i - \mu_i)\mathbf{X}_i \end{aligned}$$

Denote $w_i = \{a_i^{-1} v(\mu_i)[g'(\mu_i)]^2\}^{-1}$, then

$$U(\boldsymbol{\beta}) = \frac{1}{\phi} \sum_{i=1}^{n} w_i g'(\mu_i)(Y_i - \mu_i)\mathbf{X}_i.$$

Denote

$$W = \begin{pmatrix} w_1 & & \\ & \ddots & \\ & & w_n \end{pmatrix}, \Delta = \begin{pmatrix} g'(\mu_1) & & \\ & \ddots & \\ & & g'(\mu_n) \end{pmatrix}$$

$$U(\boldsymbol{\beta}) = \frac{1}{\phi}\mathbf{X}^{\mathbf{T}}\mathbf{W}\boldsymbol{\Delta}(\mathbf{Y} - \boldsymbol{\mu})$$

Note $U(\boldsymbol{\beta})$ is a nonlinear eq in $\boldsymbol{\beta}$.

Special case: Canonical Link

Recall

$$g'(\mu) = \frac{1}{v(\mu)} \Longrightarrow w_i = a_i v(\mu)$$

$U(\boldsymbol{\beta}) = \frac{1}{\phi}\sum_{i=1}^{n} a_i (Y_i - \mu_i)\mathbf{X}_i$

If $a_i = 1$ (linear, logistic, Poisson)

$$\begin{aligned} & U(\boldsymbol{\beta}) = \frac{1}{\phi}\mathbf{X}^T(\mathbf{Y} - \boldsymbol{\mu}) = 0 \\ \Longleftrightarrow \quad & \mathbf{X}^T\mathbf{Y} = \mathbf{X}^T\boldsymbol{\mu} \end{aligned}$$

Note:

1. $\mathbf{X^T Y}$ is sufficient statistics for $\boldsymbol{\beta}$.

2. $\mathbf{X^T}\mu$ is its expectation.

For example:

Linear	$\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = 0$
Logistic	$\sum \mathbf{X}_i^T (Y_i - \frac{e^{\mathbf{X}_i^T \boldsymbol{\beta}}}{1+e^{\mathbf{X}_i^T \boldsymbol{\beta}}}) = 0$
Poisson	$\sum \mathbf{X}_i^T (Y_i - e^{\mathbf{X}_i^T \boldsymbol{\beta}}) = 0$

Observed information:

$$J(\boldsymbol{\beta}) = -\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}$$

Recall

$$\begin{aligned}
\frac{\partial \ell}{\partial \boldsymbol{\beta}} &= \frac{1}{\phi} \sum w_i g'(\mu_i)(Y_i - \mu_i)\mathbf{X}_i \\
J(\beta) = \frac{\partial \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} &= \frac{1}{\phi} \sum \mathbf{X}_i w_i g'(\mu_i) \frac{\partial \mu_i}{\partial \boldsymbol{\beta}^T} \\
& \quad - \frac{1}{\phi} \sum \mathbf{X}_i (Y_i - \mu_i) \frac{\partial (w_i g'(\mu_i))}{\partial \boldsymbol{\beta}^T} \\
&= \frac{1}{\phi} \sum \mathbf{X}_i {w_i}^T - \frac{1}{\phi} \sum \mathbf{X}_i (Y_i - \mu_i) \frac{\partial (w_i g'(\mu_i))}{\partial \boldsymbol{\beta}^T} \\
&= \frac{1}{\phi} \mathbf{X}^{\mathbf{T}} \mathbf{W} \mathbf{X} - \frac{1}{\phi} \sum \mathbf{X}_i (Y_i - \mu_i) \frac{\partial (w_i g'(\mu_i))}{\partial \boldsymbol{\beta}^T}
\end{aligned}$$

Expected Information:

$$I(\boldsymbol{\beta}) = E\{J(\boldsymbol{\beta})\} = \frac{1}{\phi}\mathbf{X^T W X} = \frac{1}{\phi}\sum_{i=1}^{n} \mathbf{X_i^T W_i X_i}$$

Canonical Link:

$$I(\boldsymbol{\beta}) = J(\boldsymbol{\beta})$$

Why?

Large Sample Properties:

Under some regularity conditions,

(1). $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$.

(2). $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N\{0, I_o^{-1}(\boldsymbol{\beta})\}$, where $I_o(\boldsymbol{\beta}) = \lim_{n\to\infty} \frac{1}{n} I(\boldsymbol{\beta})$.

(3). For large n, $\widehat{\boldsymbol{\beta}} \sim N\{\boldsymbol{\beta}, I^{-1}(\boldsymbol{\beta})\}$ approximately.

Example 1 (Linear regression):

$Y_i \sim N(\mu_i, \sigma^2)$, $\mu_i = \mathbf{X}_i^T \boldsymbol{\beta}$, $v(\mu_i) = 1$.

Score equation:

$$\mathbf{X}^T(\mathbf{Y} - \boldsymbol{\mu}) = 0 \Longleftrightarrow \mathbf{X}^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = 0$$

$$\Longrightarrow \widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y} = (\sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{X}_i)^{-1} \sum_{i=1}^{n} \mathbf{X}_i^T Y_i$$

Information matrix:

$$I(\boldsymbol{\beta}) = \frac{1}{\sigma^2}(\mathbf{X}^{\mathbf{T}}\mathbf{X})$$

$$\widehat{\boldsymbol{\beta}} \sim N\{\boldsymbol{\beta}, I^{-1}(\boldsymbol{\beta})\} = N\{\boldsymbol{\beta}, \sigma^2(X^T X)^{-1}\}$$

Example 2 (Logistic regression):

$m_iY_i \sim Bin(m_i, \mu_i)$, $log(\frac{\mu_i}{1-\mu_i}) = \mathbf{X}_i^T\boldsymbol{\beta}$, $v(\mu_i) = \mu_i(1-\mu_i)$, $a_i = m_i$, $\phi = 1$.

<u>Score equation</u>:

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{X}_i m_i(Y_i - \mu_i) = \sum \mathbf{X}_i m_i(Y_i - \frac{e^{\mathbf{X}_i^T\boldsymbol{\beta}}}{1 + e^{\mathbf{X}_i^T\boldsymbol{\beta}}})$$

<u>Information matrix</u>:

$$I(\boldsymbol{\beta}) = \mathbf{X^TWX} = \sum_i \mathbf{X_i^TW_iX_i}$$

where $\mathbf{W} = diag\{m_i v(\mu_i)\} = diag\{m\mu_{ii}(1-\mu_i)\}$

<u>Note:</u> $\widehat{\boldsymbol{\beta}} \sim N\{\boldsymbol{\beta}, (\mathbf{X^TWX})^{-1}\}$ approximately for large n.

Example 3 (Poisson regression):

$Y_i \sim Poisson(\mu_i)$, $ln(\mu_i) = \mathbf{X}_i^T \boldsymbol{\beta}$, $v(\mu_i) = \mu_i$, $a_i = 1$, $\phi = 1$.

Score equation:

$$U(\boldsymbol{\beta}) = \sum \mathbf{X}_i (Y_i - \mu_i) = \sum \mathbf{X}_i (Y_i - e^{\mathbf{X}_i^T \boldsymbol{\beta}})$$

information matrix:

$$I(\boldsymbol{\beta}) = \mathbf{X^T W X} = \sum_i \mathbf{X_i^T W_i X_i},$$

where $\mathbf{W} = diag(\mu_i)$.

How to solve the score equations?

1. Newton-Raphson method

Let $\beta^{[k]}$ denote the kth approximation for the MLE $\widehat{\boldsymbol{\beta}}$. The $(k+1)$th approximation is given by

$$\begin{aligned}
& \boldsymbol{\beta}^{[k+1]} = \boldsymbol{\beta}^{[k]} - \left(\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\Big|_{\beta=\boldsymbol{\beta}^{[k]}}\right)^{-1} U(\boldsymbol{\beta}^{[k]}) \\
\iff\ & \boldsymbol{\beta}^{[k+1]} = \boldsymbol{\beta}^{[k]} - \left(\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\Big|_{\beta=\boldsymbol{\beta}^{[k]}}\right)^{-1} \frac{\partial \ell}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{[k]}} \\
\iff\ & \boldsymbol{\beta}^{[k+1]} = \boldsymbol{\beta}^{[k]} + J^{-1}(\boldsymbol{\beta}^{[k]}) U(\boldsymbol{\beta}^{[k]})
\end{aligned}$$

Iterations proceed until convergence.

2. Fisher - Scoring method

$$\boldsymbol{\beta}^{[k+1]} = \boldsymbol{\beta}^{[k]} - \left(E[\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}] \right)^{-1} |_{\beta=\boldsymbol{\beta}^{[k]}} U(\boldsymbol{\beta}^{[k]})$$

$$\iff \boldsymbol{\beta}^{[k+1]} = \boldsymbol{\beta}^{[k]} - \left(E[\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}] \right)^{-1} |_{\beta=\boldsymbol{\beta}^{[k]}} \frac{\partial \ell}{\partial \boldsymbol{\beta}} |_{\boldsymbol{\beta}=\boldsymbol{\beta}^{[k]}}$$

$$\iff \boldsymbol{\beta}^{[k+1]} = \boldsymbol{\beta}^{[k]} + I^{-1}(\boldsymbol{\beta}^{[k]}) U(\boldsymbol{\beta}^{[k]})$$

Iterations proceed until convergence.

Notes:

1. The Fisher scoring is often simpler than the Newton-Raphson and is widely used.

2. For canonical links, these two methods are identical.

Rationale of the Newton-Raphson method:

The MLE $\widehat{\boldsymbol{\beta}}$ satisfies $U(\widehat{\boldsymbol{\beta}}) = 0$.
Let $\boldsymbol{\beta}^{[k]}$ be the kth approximation of $\widehat{\boldsymbol{\beta}}$.

$$\Rightarrow \quad 0 = U(\widehat{\boldsymbol{\beta}}) \approx U(\boldsymbol{\beta}^{[k]}) + \frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}|_{\beta=\boldsymbol{\beta}^{[k]}}(\widehat{\beta} - \boldsymbol{\beta}^{[k]})$$

$$\Rightarrow \quad \frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}|_{\beta=\boldsymbol{\beta}^{[k]}}(\widehat{\beta} - \boldsymbol{\beta}^{[k]}) \approx -U(\boldsymbol{\beta}^{[k]})$$

$$\Rightarrow \quad \widehat{\beta} \approx \boldsymbol{\beta}^{[k]} - (\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T})^{-1}|_{\beta=\boldsymbol{\beta}^{[k]}} U(\boldsymbol{\beta}^{[k]})$$

Recall Fisher - Scoring:

$$
\begin{aligned}
& \boldsymbol{\beta}^{[k+1]} = \boldsymbol{\beta}^{[k]} + I^{-1}(\boldsymbol{\beta}^{[k]})U(\boldsymbol{\beta}^{[k]}) \\
& I(\boldsymbol{\beta}^{[k]})\boldsymbol{\beta}^{[k+1]} = I(\boldsymbol{\beta}^{[k]})\boldsymbol{\beta}^{[k]} + U(\boldsymbol{\beta}^{[k]}) \\
\Rightarrow\ & \frac{1}{\phi}(\mathbf{X}^T\mathbf{W}^{[k]}\mathbf{X})\boldsymbol{\beta}^{[k+1]} = \frac{1}{\phi}(\mathbf{X}^T\mathbf{W}^{[k]}\mathbf{X})\boldsymbol{\beta}^{[k]} \\
& +\frac{1}{\phi}(\mathbf{X}^T\mathbf{W}^{[k]}\boldsymbol{\Delta}^{[k]}(Y-\mu^{[k]}) \\
\Rightarrow\ & (\mathbf{X}^T\mathbf{W}^{[k]}\mathbf{X})\boldsymbol{\beta}^{[k+1]} = \mathbf{X}^T\mathbf{W}^{[k]}\{\mathbf{X}\boldsymbol{\beta}^{[k]} + \boldsymbol{\Delta}^{[k]}(Y-\mu^{[k]})\} \\
\Rightarrow\ & (\mathbf{X}^T\mathbf{W}^{[k]}\mathbf{X})\boldsymbol{\beta}^{[k+1]} = \mathbf{X}^T\mathbf{W}^{[k]}\boldsymbol{Z}^{[k]}
\end{aligned}
$$

where we have the working vector:

$$
\begin{cases}
\mathbf{Z}^{[k]} & = & \mathbf{X}\boldsymbol{\beta}^{[k]} + \boldsymbol{\Delta}^{[k]}(Y-\mu^{[k]}) \\
\mathbf{Z}_i^{[k]} & = & \mathbf{X}_i\boldsymbol{\beta}^{[k]} + g'(\mu_i^{[k]})(Y-\mu_i^{[k]})
\end{cases}
$$

Recall iterative score equations:

$$(\mathbf{X}^T\mathbf{W}^{[k]}\mathbf{X})\boldsymbol{\beta}^{[k+1]} = \mathbf{X}^T\mathbf{W}^{[k]}\boldsymbol{Z}^{[k]}$$

where

$$\begin{cases} \text{working vector :} & \mathbf{Z} = \quad \boldsymbol{\eta} + \boldsymbol{\Delta}(\mathbf{Y} - \boldsymbol{\mu}) \\ \text{working dependent variable:} & Z_i = \eta_i + g'(\mu_i)(Y_i - \mu_i) \end{cases}$$

This has the same form as the normal equations for using weighted least squares to fit a linear model with

a dependent variable: $\quad Z_i = \mathbf{X}_i^T\boldsymbol{\beta} + g'(\mu_i) \cdot (Y_i - \mu_i)$

covariates: $\quad \mathbf{X}_i$

weights: $\quad w_i = [a_i^{-1}v(\mu_i)(g'(\mu_i)^2]^{-1}$

Note:

1. Unlike the standard linear models, Z_i and w_i depend on β.

2. An alternative way to construct the working vector:

$$\begin{aligned} g(Y_i) &\approx g(\mu_i) + g'(\mu_i)(Y_i - \mu_i) \\ &= \mathbf{X}_i^T \boldsymbol{\beta} + g'(\mu_i)(Y_i - \mu_i) \\ &= Z_i \end{aligned}$$

3. Iterative reweighted least squares (IWLS): The score equations $\mathrm{U}(\boldsymbol{\beta}) = 0$ can be solved by iteratively fitting $Z_i = \mathbf{X}_i^T \boldsymbol{\beta} + \epsilon_i$, where $\epsilon_i \sim N(0, w_i^{-1})$. Z_i and w_i are updated at each iteration.

IWLS algorithm for GLMs

1. Calculate an initial value for the linear predictor.

$$\eta^{(0)} \;=\; \mathbf{X}^T \boldsymbol{\beta}^{(0)} = g(Y)$$

For example:

* In Poisson regression, $\eta_i^{(0)} = ln(Y_i + \frac{1}{2})$.

* In logistic regression where $y_i \sim bin(m_i, p_i)$, $\eta_i^{(0)} = ln(\frac{Y_i + \frac{1}{2}}{m_i - Y_i + \frac{1}{2}})$ which is called the empirical logit.

2. Calculate fitted values, variances and link derivatives.

$$\begin{aligned} \mu^{(0)} &= g^{-1}\{\eta^{(0)}\} \\ v^{(0)} &= v\{\mu^{(0)}\} \\ \Delta^{(0)} &= diag\{g'(\mu_i^{(0)})\} \end{aligned}$$

3. Calculate working weights and working vectors.

$$W^{(0)} = diag(w_i^{(0)})$$

where $w_i^{(0)} = \frac{a_i}{v(\mu_i^{(0)})\{g'(\mu_i^{(0)})\}^2}$.

$$Z^{(0)} = \eta^{(0)} + \Delta^{(0)}(Y - \mu^{(0)}).$$

4. Solve for $\beta^{(1)}$: WLS regression of $Z^{(0)}$ on $\mathbf{X}$ with weights $w^{(0)}$.

5. Set $\eta^{(1)} = \mathbf{X}^T\beta^{(1)}$ and go to 2. Iterations continue until convergence, i.e. $||\beta^{(k+1)} - \beta^{(k)}|| < 10^{-6}$.

6. $c\widehat{o}v(\widehat{\beta}) = \frac{1}{\phi}(\mathbf{X}^{\mathbf{T}}\widehat{\mathbf{W}}\mathbf{X})^{-1}$ at convergence.

Deviance

Recall GLM:

$$g(\mu_i) = \mathbf{X}_i^T \boldsymbol{\beta}, \tag{1}$$

where $\boldsymbol{\beta}$ is $p' \times 1$ and $p' << n$.

Deviance:

$$D = 2\phi\{\ell(\widehat{\boldsymbol{\beta}}_s) - \ell(\widehat{\boldsymbol{\beta}}_M)\}$$

where ϕ is the scale parameter, $\widehat{\boldsymbol{\beta}}_s$ is the MLE under the saturated model, and $\widehat{\boldsymbol{\beta}}_M$ is the MLE under (1).

Scaled Deviance:

$$D^* = \frac{D}{\phi} = 2\{\ell(\widehat{\boldsymbol{\beta}}_s) - \ell(\widehat{\boldsymbol{\beta}}_m)\}$$

Remarks:

1. The saturated model is the model with the number of parameters equal to the number of observations, i.e., $\widehat{\boldsymbol{\beta}_s}$ is an $n \times 1$vector.

2. The saturated model gives the best fit to the data and yields the largest loglikelihood function.

3. It is uninformative since it does not summarize the data but simply repeats them in full.

4. The deviance measures the discrepancy of the fitted model from the observed data, i.e., how good the model fits the data.

$$\begin{aligned} \text{large } D &\Rightarrow \text{poor fit} \\ \text{small } D &\Rightarrow \text{good fit} \end{aligned}$$

Deviances of Common Distributions

1. Normal with σ^2 known:

$$
\begin{aligned}
D &= 2\phi\{\ell(\widehat{\boldsymbol{\beta}}_s) - \ell(\widehat{\boldsymbol{\beta}}_m)\} \\
&= 2\sigma^2\{-\frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - Y_i)^2 \\
&\quad -[-\frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \widehat{\mu}_i)^2]\} \\
D &= \sum_{i=1}^{n}(Y_i - \widehat{\mu}_i)^2 \\
&= RSS
\end{aligned}
$$

2. Binomial: $Y_i \sim B(m_i, p_i)$

$$
\begin{aligned}
D &= 2\phi\{\ell(\widehat{\boldsymbol{\beta}}_s) - \ell(\widehat{\boldsymbol{\beta}}_m)\} \\
&= 2\{\sum_{i=1}^{n}[Y_i \ln(\frac{Y_i}{m_i}) + (m_i - Y_i)\ln(\frac{m_i - Y_i}{m_i}) + \ln\begin{pmatrix} m_i \\ Y_i \end{pmatrix}] \\
&\quad - \sum_{i=1}^{n}[Y_i \ln(\frac{\widehat{\mu}_i}{m_i}) + (m_i - Y_i)\ln(\frac{m_i - \widehat{\mu}_i}{m_i}) + \ln\begin{pmatrix} m_i \\ Y_i \end{pmatrix}]\} \\
D &= 2\sum_{i=1}^{n}\{Y_i \ln(\frac{Y_i}{\widehat{\mu}_i}) + (m_i - Y_i)\ln(\frac{m_i - Y_i}{m_i - \widehat{\mu}_i})\} \\
&= 2\sum_{i=1}^{n}\ln(\frac{O_i}{E_i})
\end{aligned}
$$

where $\widehat{\mu}_i = m_i \widehat{p}_i = m_i g^{-1}(\mathbf{X}_i^T \widehat{\boldsymbol{\beta}})$.

Relationship between LR stat. and Deviance stat.

Full Model:

$$g(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \beta_{k+1} X_{k+1} + \cdots + \beta_p X_p$$

Reduced Model:

$$g(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

Hypothesis:

$$\begin{aligned} H_0 &: \beta_{k+1} = \cdots = \beta_p = 0 \\ H_1 &: \neq 0 \text{ some where} \end{aligned}$$

$$\chi^2_{LR} = D^*_{n-(k+1)}(\widehat{reduced}) - D^*_{n-(p+1)}(\widehat{full}) \sim \chi^2(p-k)$$

Why?

$$
\begin{aligned}
\chi^2_{LR} &= 2[\ell(\widehat{full}) - \ell(\widehat{reduced})] \\
&= 2[\ell(\widehat{\beta}) - \ell(\widehat{\beta}^0)] \\
&= 2[\ell(\widehat{\beta}_s) - \ell(\widehat{\beta}^0)] - 2[\ell(\widehat{\beta}_s) - \ell(\widehat{\beta})] \\
&= D^*(\widehat{reduced}) - D^*(\widehat{full})
\end{aligned}
$$

The $d.f.$ of $D^*(\widehat{reduced})$ is $n-(k+1)$ while the $d.f.$ of $D^*(\widehat{full})$ is $n-(p+1)$. So χ^2_{LR} has $p-k$ degree of freedom.

Goodness-of-fit Statistics

1. Deviance Statistic:

$$D = 2\phi\{\ell(\widehat{\boldsymbol{\beta}}_s) - \ell(\widehat{\boldsymbol{\beta}}_M)\}$$

Normal: $D=\sum_{i=1}^{n}(Y_i - \widehat{\mu}_i)^2$

Binomial: $D = 2\sum_{i=1}^{n}\{Y_i \ln(\frac{Y_i}{\widehat{\mu}_i}) + (m_i - Y_i)\ln(\frac{m_i - Y_i}{m_i - \widehat{\mu}_i})\}$

Poisson: $D = 2\sum_{i=1}^{n}\{Y_i \ln(\frac{Y_i}{\widehat{\mu}_i}) - (Y_i - \widehat{\mu}_i)\}$

For grouped data, $D \sim \chi^2(n-p)$ approximately.

2. Pearson Statistic:

$$X^2 = \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{v(\widehat{\mu}_i)}$$

Normal: $X^2 = \sum_{i=1}^{n} (Y_i - \widehat{\mu}_i)^2 = D$

Binomial: $X^2 = 2 \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{m_i \widehat{\mu}_i (1 - \widehat{\mu}_i)}\}$

Poisson: $X^2 = 2 \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i}$

For grouped data, $X^2 \sim \chi^2(n-p)$ approximately.

Remarks:

1. X^2 is a quadratic approximation of D.

2. $X^2 = \sum w_i (Z_i - \mathbf{X}_i^T \widehat{\boldsymbol{\beta}})^2$ at convergence.

3. For two nested models,

$$\begin{aligned} D_R - D_F &\rightarrow \chi^2 \\ X_R^2 - X_F^2 &\rightarrow \chi^2 \quad ? \end{aligned}$$

Quasi-likelihood Functions

Question:

1. Given the first two moments of Y: $E(Y) = \mu$, $Var(Y) = a^{-1}\phi v(\mu)$.

2. Assume $g(\mu) = \mathbf{X}^T\boldsymbol{\beta}$.

How to construct a likelihood of μ or $\boldsymbol{\beta}$?

Answer: Quasi-likelihood function.

Quasi-Score:

$$U = \frac{Y - \mu}{a^{-1}\phi v(\mu)}$$

Remarks:

(1). For an exponential family distribution, $U = \frac{\partial \ell}{\partial \mu}$. (Why?)

(2). Properties of U:

$$\begin{cases} E(U) = 0 \\ E(U^2) = -E(\frac{\partial U}{\partial \mu}) = \frac{1}{a^{-1}\phi v(\mu)} \end{cases}$$

Why are these properties important?

Log Quasi-likelihood (QL) [Wedderburn, 1974]:

$$Q(\mu; Y) = \int_Y^{\mu} \frac{Y - u}{a^{-1}\phi v(u)} du$$

Remarks:

(1) For fixed ϕ, $Q(\mu; Y)$ behaves like a log likelihood function.

(2) Some special cases (see the next page).

Table 9.1. *Quasi-likelihoods associated with some simple variance functions*

Variance function $V(\mu)$	*Quasi-likelihood* $Q(\mu; y)$	*Canonical parameter* θ	*Distribution name*	*Range restrictions*
1	$-(y-\mu)^2/2$	μ	Normal	—
μ	$y \log \mu - \mu$	$\log \mu$	Poisson	$\mu > 0,\ y \geq 0$
μ^2	$-y/\mu - \log \mu$	$-1/\mu$	Gamma	$\mu > 0,\ y > 0$
μ^3	$-y/(2\mu^2) + 1/\mu$	$-1/(2\mu^2)$	Inverse Gaussian	$\mu > 0,\ y > 0$
μ^ζ	$\mu^{-\zeta}\left(\frac{\mu y}{1-\zeta} - \frac{\mu^2}{2-\zeta}\right)$	$\frac{1}{(1-\zeta)\mu^{\zeta-1}}$	—	$\mu > 0,\ \zeta \neq 0, 1, 2$
$\mu(1-\mu)$	$y \log\left(\frac{\mu}{1-\mu}\right) + \log(1-\mu)$	$\log\left(\frac{\mu}{1-\mu}\right)$	Binomial/m	$0 < \mu < 1,\ 0 \leq y \leq 1$
$\mu^2(1-\mu)^2$	$(2y-1)\log\left(\frac{\mu}{1-\mu}\right) - \frac{y}{\mu} - \frac{1-y}{1-\mu}$	—	—	$0 < \mu < 1,\ 0 < y < 1$
$\mu + \mu^2/k$	$y \log\left(\frac{\mu}{k+\mu}\right) + k \log\left(\frac{k}{k+\mu}\right)$	$\log\left(\frac{\mu}{k+\mu}\right)$	Negative binomial	$\mu > 0,\ y \geq 0$

(3) For many common distributions, a QL is identical to a log-likelihood function. However, it is likely that a QL exists but a log likelihood does not, since one may specify an arbitrary variance function.

(4) $\frac{\partial Q(\mu;Y)}{\partial\mu} = \frac{Y-\mu}{a^{-1}\phi v(\mu)}$

Quasi-Deviance

Quasi-Deviance:

$$D(\mu; Y) = -2\phi Q(\mu; Y) = 2\int_{\mu}^{Y} \frac{Y-u}{a^{-1}v(u)}du$$

Remarks:

(1). $D(\mu; Y) \geq 0$.

(2). $D(\mu; Y)$ does not depend on ϕ.

Quasi-score function for β:

$$\mathbf{U}(\boldsymbol{\beta}) = \frac{1}{\phi}\mathbf{D}^T V^{-1}(Y-u)$$

where $\mathbf{D} = \frac{\partial \mu}{\partial \boldsymbol{\beta}^T}$. Why?

Summary

1. $E(Y) = \mu$, $var(Y) = a^{-1}\phi v(u)$

2. Model: $g(\mu) = \mathbf{X}^T\boldsymbol{\beta}$

3. Quasi-loglikelihood: $Q(\mu; Y) = \int_Y^\mu \frac{Y-u}{a^{-1}\phi v(u)} du$

4. $\mathbf{U}(\boldsymbol{\beta}) = \frac{\partial Q(\mu;Y)}{\partial \boldsymbol{\beta}}$

5. Quasi-information matrix for $\boldsymbol{\beta}$:

$$\mathbf{I}(\boldsymbol{\beta}) = -E\{\frac{\partial \mathbf{U}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\} = \frac{1}{\phi}\mathbf{D}^T V^{-1}\mathbf{D}$$

where $\mathbf{D} = \frac{\partial \mu}{\partial \boldsymbol{\beta}^T}$.

Data: Y_i, $\mathbf{X}_i$ $(i = 1, \cdots, n)$

Quasi-Loglikelihood : $QL = \sum_{i=1}^{n} Q_i(\mu_i; Y_i) = \sum_{i=1}^{n} \int_{Y_i}^{\mu_i} \frac{Y_i - u}{a^{-1}\phi v(u)} du$

Quasi-Deviance : $QDev = \sum_{i=1}^{n} d_i(\mu_i; Y_i) = 2 \sum_{i=1}^{n} \int_{\mu_i}^{Y_i} \frac{Y_i - u}{a^{-1} v(u)} du$

Quasi-Score : $\mathbf{U}(\boldsymbol{\beta}) = \frac{1}{\phi} \sum_{i=1}^{n} \mathbf{D}_i^T V_i^{-1} (Y_i - \mu_i)$

Quasi-Information : $\mathbf{I}(\boldsymbol{\beta}) = \frac{1}{\phi} \sum_{i=1}^{n} \mathbf{D}_i^T V_i^{-1} \mathbf{D}_i$

Theorem [Wedderburn 1974] :

Under some regularity conditions, the maximum quasi-likelihood estimator $\widehat{\beta}$, which satisfies $U(\widehat{\beta}) = 0$, is consistent and asymptotically normal. i.e.

(1). $\widehat{\boldsymbol{\beta}} \overset{p}{\rightarrow} \boldsymbol{\beta}$

(2). $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{d}{\rightarrow} N\{0, \mathbf{I}^{-1}(\boldsymbol{\beta})\}$

Questions:

(1). If the variance function $V(\mu)$ is misspecified, is the solution $\widehat{\boldsymbol{\beta}}$ of $\mathbf{U}(\widehat{\boldsymbol{\beta}}) = 0$ consistent? asymptotically normal?

(2). Does $\widehat{\boldsymbol{\beta}}$ depend on ϕ?

Estimation of ϕ:

1. $\widehat{\phi} = \frac{Pearson\ \chi^2}{n-p} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(Y_i-\widehat{\mu}_i)^2}{a_i^{-1}v(\widehat{\mu}_i)}$

2. $\widehat{\phi} = \frac{Deviance}{n-p} = \frac{1}{n-p}\sum_{i=1}^{n} d_i$

Usually the first estimator is recommended.

Example (Wedderburn, 1974)

Outcome: the percentage of leaf blotch

Covariates: 10 varieties of barley, 9 sites. (see table)

Model:

$$logit(\mu) = \mathbf{X}^T \boldsymbol{\beta}$$

variance function:

(1). $\mu(1-\mu)$

(2). $\mu^2(1-\mu)^2$

Table 9.2. *Incidence of* R. secalis *on the leaves of ten varieties of barley grown at nine sites: response is the percentage of leaf affected*

	Variety										
Site	1	2	3	4	5	6	7	8	9	10	*Mean*
1	0.05	0.00	0.00	0.10	0.25	0.05	0.50	1.30	1.50	1.50	0.52
2	0.00	0.05	0.05	0.30	0.75	0.30	3.00	7.50	1.00	12.70	2.56
3	1.25	1.25	2.50	16.60	2.50	2.50	0.00	20.00	37.50	26.25	11.03
4	2.50	0.50	0.01	3.00	2.50	0.01	25.00	55.00	5.00	40.00	13.35
5	5.50	1.00	6.00	1.10	2.50	8.00	16.50	29.50	20.00	43.50	13.36
6	1.00	5.00	5.00	5.00	5.00	5.00	10.00	5.00	50.00	75.00	16.60
7	5.00	0.10	5.00	5.00	50.00	10.00	50.00	25.00	50.00	75.00	27.51
8	5.00	10.00	5.00	5.00	25.00	75.00	50.00	75.00	75.00	75.00	40.00
9	17.50	25.00	42.50	50.00	37.50	95.00	62.50	95.00	95.00	95.00	61.50
Mean	4.20	4.77	7.34	9.57	14.00	21.76	24.17	34.81	37.22	49.33	20.72

Source: Wedderburn (1974) taken from an unpublished Aberystwyth Ph.D thesis by J.F. Jenkyn.

Fig 1. Pearson residuals plotted against the linear predictor $\widehat{\eta} = \log(\frac{\widehat{\pi}}{1-\widehat{\pi}})$ for the 'binomial' fit to the leaf-blotch data using variance function $\widehat{\pi}(1-\widehat{\pi})$.

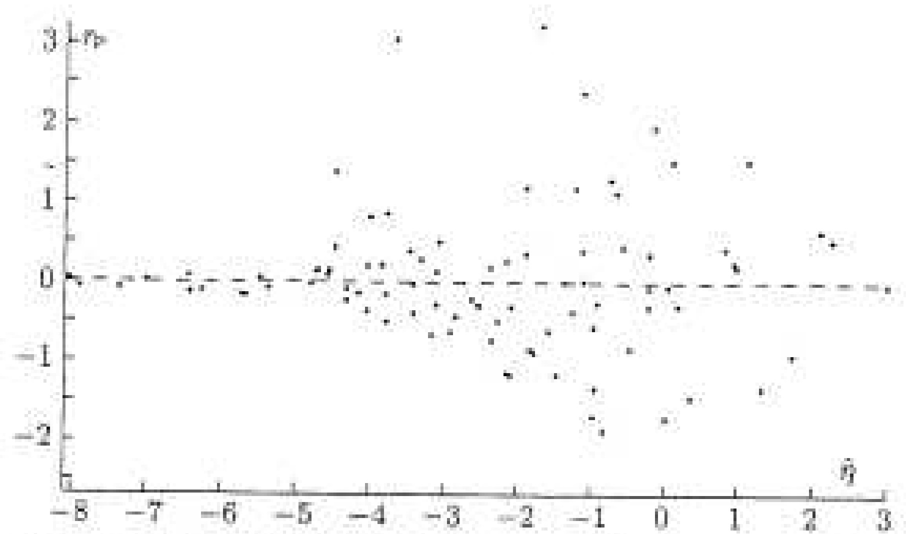

Fig. 9.1a. *Pearson residuals plotted against the linear predictor* $\hat{\eta} = \log(\hat{\pi}/(1-\hat{\pi}))$ *for the 'binomial' fit to the leaf-blotch data.*

Fig 2. Pearson residuals using variance function $\widehat{\pi}^2(1-\widehat{\pi})^2$ plotted against the linear predictor $\widehat{\eta} = \log(\frac{\widehat{\pi}}{1-\widehat{\pi}})$ for the leaf-blotch data .

Variety									
1	2	3	4	5	6	7	8	9	10
0.00	-0.47	0.08	0.95	1.35	1.33	2.34	3.26	3.14	3.89
(0.00)	(0.47)	(0.47)	(0.47)	(0.47)	(0.47)	(0.47)	(0.47)	(0.47)	(0.47)

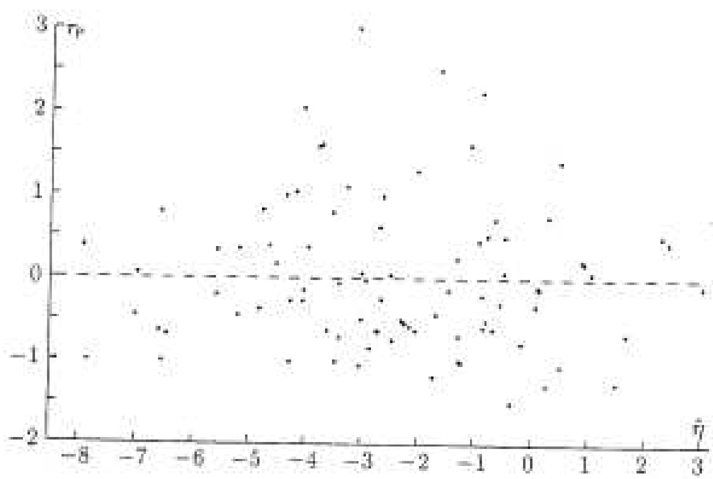

Fig. 9.2. *Pearson residuals using the variance function $\pi^2(1-\pi)^2$ plotted against the linear predictor $\hat{\eta}$ for the leaf-blotch data.*

Estimating Equations

Estimating Function:

$e(Y;\boldsymbol{\theta})$ is an estimating function for $\boldsymbol{\theta}(p \times 1)$ if $E[e(Y;\boldsymbol{\theta})] = 0$ for all θ;

or $E[e(Y;\boldsymbol{\theta})|\mathbf{A}] = 0$ for all $\boldsymbol{\theta}$ (more generally), where $\mathbf{A}$ is some covariate vector.

e.g.

1. $e(Y;\boldsymbol{\theta}) = Y - \mu(\boldsymbol{\theta})$

2. $e(Y;\boldsymbol{\theta}) = \mathbf{X}^T\{Y - \mu(\boldsymbol{\theta})\}$

3. AR(1) model:

$$Y_t = \theta Y_{t-1} + \epsilon_t$$

where $\epsilon_t \sim N(0, \sigma^2)$. Then

$$e(Y_t; \theta) = Y_t - \theta Y_{t-1}.$$

4. Martingale (survival analysis):

$$E[e(Y_t; \boldsymbol{\theta}) | \mathbf{A}_t] = 0$$

where $\mathbf{A}_t = Y_1, \cdots, Y_{t-1}$, the past history. Then

$$e(Y_t; \boldsymbol{\theta}) = \text{element of the cox model score equation.}$$

Question

For each observation, one can construct an estimating equation. How to combine these n equations optimally into p equations?

Optimal Estimating Equation (Function)

Let

$$\begin{aligned}\mathbf{D}_i &= -E[\frac{\partial e_i(Y_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}|\mathbf{A}_i] \\ V_i &= var(e_i(Y_i;\theta)|\mathbf{A}_i),\end{aligned}$$

Then within the class of estimating function $c_i(\boldsymbol{\theta})e_i(\boldsymbol{\theta})$, the optimal estimating equation is

$$\mathbf{U}(\boldsymbol{\theta};y) = \sum_i \mathbf{D}_i^T V_i^{-1} e_i(y;\boldsymbol{\theta}) = \mathbf{D}^T \mathbf{V}^{-1}\mathbf{e}$$

where $\mathbf{D} = (\mathbf{D}_1,\cdots,\mathbf{D}_n)^T$, $\mathbf{V} = diag(V_1,\cdots,V_n)$, and $\mathbf{e} = (e_1,\cdots,e_n)^T$.

Optimal Linear Estimating Equation (Function)

Within the class of linear estimating functions:

$$\mathrm{s}(\theta; \mathbf{Y}) = \mathbf{H}^T(\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\theta}))$$

where $\mathbf{Y} = (Y_1, \cdots, Y_n)^T$ and $\mathbf{H}$: $n \times p$.

Then $\mathrm{U}(\boldsymbol{\theta}; \mathbf{Y}) = \mathbf{D}^T\mathbf{V}^{-1}\{\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\theta})\}$ is optimal.

What is the optimality criterion?

Let $\widehat{\boldsymbol{\theta}}$ satisfy $\mathrm{s}(\widehat{\boldsymbol{\theta}}; \mathbf{Y}) = 0$, and $\widehat{\boldsymbol{\theta}}_0$ satisfy $\mathrm{U}(\widehat{\boldsymbol{\theta}}_0; \mathbf{Y}) = 0$,

then $cov(\widehat{\boldsymbol{\theta}}) - cov(\widehat{\boldsymbol{\theta}}_0) \geq 0$ (nonnegative definite).

Examples of $\mathrm{U}(\theta; \mathrm{Y}) = \mathbf{D}^T \mathbf{V}^{-1} \mathrm{e}$

Example 1 (GLM) :

$$\mathbf{e}(\boldsymbol{\beta}; \mathbf{Y}) = \mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\theta})$$

$$\boldsymbol{\theta} = \boldsymbol{\beta}, \mathbf{D} = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}^T}, \mathbf{V} = diag\{\phi a_i^{-1} v(\mu_i)\}.$$

Then the optimal linear estimating equation

$$\mathbf{U}(\beta; Y) = \mathbf{D}^T \mathbf{V}^{-1} \{\mathbf{Y} - \mu(\boldsymbol{\beta})\}$$

is the QL score equation. It is also the regular score equation for common GLMs.

Example 2 (AR1 Model) :

$$Y_t = \theta Y_{t-1} + \epsilon_t$$

where $\epsilon_t \sim N(0, \sigma^2)$.

$$e(Y_t; \theta) = Y_t - \theta Y_{t-1}.$$

$D_t =$?

$V_t =$?

$U(\boldsymbol{\beta}; \boldsymbol{Y}) =$?

Why is $\mathbf{U}(\boldsymbol{\theta}; \mathbf{Y})$ optimal?

For simplicity, we here concentrate on linear estimating equations.

Recall $\mathbf{s}(\boldsymbol{\theta}; \mathbf{Y}) = \mathbf{H}^T\{\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\theta})\}$, where $\mathbf{H}$ may depend on $\boldsymbol{\theta}$.

$$\mathbf{U}(\boldsymbol{\theta}; \mathbf{Y}) = \mathbf{D}^T\mathbf{V}^{-1}\{\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\theta})\}$$

Noting $0 = \mathrm{s}(\widehat{\boldsymbol{\theta}}) \approx \mathrm{s}(\boldsymbol{\theta}) - \mathbf{H}^T\mathbf{D}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$, we have

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} = (\mathbf{H}^T\mathbf{D})^{-1}\mathbf{H}^T\{\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\theta})\}.$$

Therefore,

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) &= (\frac{1}{n}\mathbf{H}^T\mathbf{D})^{-1}\frac{1}{\sqrt{n}}\mathbf{H}^T\{\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\theta})\} + o_p(1) \\ &\overset{d}{\to} N(0, \boldsymbol{\Sigma})\end{aligned}$$

where $\boldsymbol{\Sigma} = lim_{n\to\infty}(\frac{1}{n}\mathbf{H}^T\mathbf{D})^{-1}(\frac{1}{n}\mathbf{H}^T\mathbf{V}\mathbf{H})[\frac{1}{n}(\mathbf{H}^T\mathbf{D})^{-1}]^T$ is a sandwich estimator.

Suppose we have another estimator $\widehat{\boldsymbol{\theta}}_0$, which satisfies

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}) \xrightarrow{d} N(0, \boldsymbol{\Sigma}_0),$$

and $\boldsymbol{\Sigma}_0 = (\mathbf{D}^T \mathbf{V}^{-1} \mathbf{D})^{-1}$.

We need to show

$$\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_0 \geq 0.$$

This is equivalent to

$$(\mathbf{H}^T\mathbf{D})^{-1}(\mathbf{H}^T\mathbf{V}\mathbf{H})(\mathbf{D}^T\mathbf{H})^{-1} - (\mathbf{D}^T\mathbf{V}^{-1}\mathbf{D})^{-1} \geq 0.$$

It is sufficient to show

$$\mathbf{\Sigma}_0^{-1} - \mathbf{\Sigma}^{-1} \geq 0,$$

i.e.

$$\begin{aligned} & \mathbf{D}^T\mathbf{V}^{-1}\mathbf{D} - \mathbf{D}^T\mathbf{H}(\mathbf{H}^T\mathbf{V}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{D} \\ = \;& \mathbf{D}^T\{\mathbf{V}^{-1} - \mathbf{H}(\mathbf{H}^T\mathbf{V}\mathbf{H})^{-1}\mathbf{H}^T\}\mathbf{D} \end{aligned}$$

* $\{\mathbf{V} - \mathbf{H}(\mathbf{H}^T\mathbf{V}\mathbf{H})^{-1}\mathbf{H}^T\}$ is non-negative definite. (Why?)

Questions:

1. If the data $\mathbf{Y}$ have a likelihood $\ell(\boldsymbol{\theta})$, would the linear estimating equation $U(\boldsymbol{\theta};\mathbf{Y}) = \mathbf{D}^T\mathbf{V}^{-1}\{\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\theta})\}$ always be the optimal for $\boldsymbol{\theta}$? (i.e. equals to the score equation $\frac{\partial \ell}{\partial \boldsymbol{\theta}}$?)

2. Can you give a counter example, if no?