Chapter 2

Generalised Linear Models

2.1 The Exponential family

A probability distribution is said to be a member of the *exponential family* if its probability density function (or probability function, if discrete) can be written in the form

$$f_Y(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right) \tag{1}$$

The parameter θ is called the *natural* or *canonical* parameter. The parameter ϕ is usually assumed known. If it is unknown then it is often called the *nuisance* parameter.

The density (1) can be thought of as a likelihood resulting from a single observation y. Then

$$\begin{array}{rl} & \log f_Y(y;\theta,\phi) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi) \\ \Rightarrow & u(\theta) = \frac{\partial}{\partial\theta}\log f_Y(y;\theta,\phi) = \frac{y - \frac{\partial}{\partial\theta}b(\theta)}{a(\phi)} = \frac{y - b'(\theta)}{a(\phi)} \\ \Rightarrow & H(\theta) = \frac{\partial^2}{\partial\theta^2}\log f_Y(y;\theta,\phi) = -\frac{\frac{\partial^2}{\partial\theta^2}b(\theta)}{a(\phi)} = -\frac{b''(\theta)}{a(\phi)} \\ \Rightarrow & \mathcal{I}(\theta) = E[-H(\theta)] = \frac{b''(\theta)}{a(\phi)}. \end{array}$$

From the properties of the score function in Section 1.7.3, we know that $E[U(\theta)] = 0$. Therefore

$$E\left[\frac{Y - b'(\theta)}{a(\phi)}\right] = 0 \quad \Rightarrow \quad E[Y] = b'(\theta).$$

Furthermore,

$$Var[U(\theta)] = Var\left[\frac{Y - b'(\theta)}{a(\phi)}\right] = \frac{Var[Y]}{a(\phi)^2}$$

as $b'(\theta)$ and $a(\phi)$ are constants (not random variables). Now, we also know from Section 1.7.4 that $Var[U(\theta)] = \mathcal{I}(\theta)$. Therefore,

$$Var[Y] = a(\phi)^2 Var[U(\theta)] = a(\phi)^2 \mathcal{I}(\theta) = a(\phi) b''(\theta).$$

and hence the mean and variance of a random variable with probability density function (or probability function) of the form (1), are $b'(\theta)$ and $a(\phi)b''(\theta)$ respectively.

We often denote the mean by μ, so $\mu = b'(\theta)$. The variance is the product of two functions; $b''(\theta)$ depends on the canonical parameter θ (and hence μ) only and is called the *variance function* ($V(\mu) \equiv b''(\theta)$); $a(\phi)$ is sometimes of the form $a(\phi) = \sigma^2/w$ where w is a known *weight* and σ^2 is called the *dispersion parameter* or *scale parameter*.

♡ **Example** 2.1. **Normal distribution,** $Y \sim \mathbf{N}(\mu, \sigma^2)$

$$\begin{array}{rcl} f_Y(y;\mu,\sigma^2) & = & \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right) \qquad y \in \mathcal{R}; \quad \mu \in \mathcal{R} \\ & = & \exp\left(\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{1}{2}\left[\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right]\right) \end{array}$$

This is in the form (1), with $\theta = \mu$, $b(\theta) = \frac{1}{2}\theta^2$, $a(\phi) = \sigma^2$ and

$$c(y,\phi) = -\frac{1}{2}\left[\frac{y^2}{a(\phi)} + \log(2\pi a[\phi])\right].$$

Therefore

$$\begin{array}{l} E(Y) = b'(\theta) = \theta = \mu \\ Var(Y) = a(\phi)b''(\theta) = \sigma^2 \\ V(\mu) = 1. \end{array}$$

♡ **Example** 2.2. **Poisson distribution,** $Y \sim$ **Poisson**(λ)

$$\begin{array}{rcl} f_Y(y;\lambda) & = & \frac{\exp(-\lambda)\lambda^y}{y!} \qquad y \in \{0,1,\ldots\}; \quad \lambda \in \mathcal{R}_+ \\ & = & \exp\left(y\log\lambda - \lambda - \log y!\right) \end{array}$$

This is in the form (1), with $\theta = \log\lambda$, $b(\theta) = \exp\theta$, $a(\phi) = 1$ and $c(y,\phi) = -\log y!$. Therefore

$$\begin{array}{l} E(Y) = b'(\theta) = \exp\theta = \lambda \\ Var(Y) = a(\phi)b''(\theta) = \exp\theta = \lambda \\ V(\mu) = \mu. \end{array}$$

♡ **Example** 2.3. **Bernoulli distribution,** $Y \sim$ **Bernoulli**(p)

$$\begin{aligned} f_Y(y;p) &= p^y(1-p)^{1-y} \qquad y \in \{0,1\}; \quad p \in (0,1) \\ &= \exp\left(y \log \tfrac{p}{1-p} + \log(1-p)\right) \end{aligned}$$

This is in the form (1), with $\theta = \log \frac{p}{1-p}$, $b(\theta) = \log(1+\exp\theta)$, $a(\phi) = 1$ and $c(y,\phi) = 0$. Therefore

$$\begin{aligned} &E(Y) = b'(\theta) = \tfrac{\exp\theta}{1+\exp\theta} = p \\ &Var(Y) = a(\phi)b''(\theta) = \tfrac{\exp\theta}{(1+\exp\theta)^2} = p(1-p) \\ &V(\mu) = \mu(1-\mu). \end{aligned}$$

♡ **Example** 2.4. **Binomial distribution,** $Y \sim$ **Binomial**(n,p) Here, n is assumed known (as usual) and the random variable Y is taken as the *proportion* of successes, so

$$\begin{aligned} f_Y(y;p) &= \binom{n}{ny} p^{ny}(1-p)^{n(1-y)} \qquad y \in \left\{0, \tfrac{1}{n}, \tfrac{2}{n}, \ldots, 1\right\}; p \in (0,1) \\ &= \exp\left(\frac{y \log \frac{p}{1-p} + \log(1-p)}{\frac{1}{n}} + \log\binom{n}{ny}\right) \end{aligned}$$

This is in the form (1), with $\theta = \log \frac{p}{1-p}$, $b(\theta) = \log(1+\exp\theta)$, $a(\phi) = \frac{1}{n}$ and $c(y,\phi) = \log\binom{n}{ny}$. Therefore

$$\begin{aligned} &E(Y) = b'(\theta) = \tfrac{\exp\theta}{1+\exp\theta} = p \\ &Var(Y) = a(\phi)b''(\theta) = \tfrac{1}{n}\tfrac{\exp\theta}{(1+\exp\theta)^2} = \tfrac{p(1-p)}{n} \\ &V(\mu) = \mu(1-\mu). \end{aligned}$$

Here, we can write $a(\phi) \equiv \sigma^2/w$ where the scale parameter $\sigma^2 = 1$ and the weight w is n, the binomial denominator.

2.2 Components of a generalised linear model

2.2.1 The random component

In practical applications, we often distinguish between a *response* variable and a group of *explanatory* variables. The aim is to determine the pattern of dependence of the response variable on the explanatory variables. We denote the n observations of the response by $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$. In a generalised linear model (g.l.m.), these are assumed to be observations of *independent* random variables $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$, which take the same distribution from the exponential family. In other words, the functions a, b and c and usually the scale parameter ϕ are the same for all observations, but the canonical parameter θ may differ. Therefore, we write

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

and the joint density for $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ is

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta},\boldsymbol{\phi}) &= \textstyle\prod_{i=1}^n f_{Y_i}(y_i;\theta_i,\phi_i) \\ &= \exp\left(\textstyle\sum_{i=1}^n \frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + \sum_{i=1}^n c(y_i,\phi_i)\right) \end{aligned} \qquad (2)$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T$ is the collection of canonical parameters and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)^T$ is the collection of nuisance parameters (where they exist).

Note that for a particular sample of observed responses, $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$, (2) is the likelihood function for $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$.

2.2.2 The systematic (or structural) component

Associated with each y_i is a vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ of values of p explanatory variables. In a generalised linear model, the distribution of the response variable Y_i depends on $\mathbf{x}_i$ through the *linear predictor* η_i where

$$\begin{aligned} \eta_i &= \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \\ &= \textstyle\sum_{j=1}^p x_{ij}\beta_j \\ &= \mathbf{x}_i^T\boldsymbol{\beta} \\ &= [\mathbf{X}\boldsymbol{\beta}]_i, \qquad i = 1, \ldots, n \end{aligned} \qquad (3)$$

where, as with a linear model,

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a vector of fixed but unknown parameters describing the dependence of Y_i on $\mathbf{x}_i$. The four ways of describing the linear predictor in (3) are equivalent, but the most economical is the matrix form

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}. \qquad (4)$$

Again, we call the $n \times p$ matrix $\mathbf{X}$ the *design matrix*. The ith row of $\mathbf{X}$ is $\mathbf{x}_i^T$, the explanatory data corresponding to the ith observation of the response. The jth column of $\mathbf{X}$ contains the n observations of the jth explanatory variable.

2.2.3 The link function

For specifying the pattern of dependence of the response variable on the explanatory variables, the canonical parameters $\theta_1, \ldots, \theta_n$ in (2) are not of direct interest. Furthermore, we have already specified that the distribution of Y_i should depend on $\mathbf{x}_i$ through the linear predictor η_i. It is the parameters $\beta_1, \ldots, \beta_p$ of the linear predictor which are of primary interest.

The link between the distribution of $\mathbf{Y}$ and the linear predictor $\boldsymbol{\eta}$ is provided by the *link function* g,

$$\eta_i = g(\mu_i) \qquad i = 1, \ldots, n$$

where $\mu_i \equiv E(Y_i),\ i = 1, \ldots, n$. Hence, the dependence of the distribution of the response on the explanatory variables is established as

$$g(E[Y_i]) = g(\mu_i) = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta} \qquad i = 1, \ldots, n$$

In principle, the link function g can be any one-to-one differentiable function. However, we note that η_i can in principle take any value in $\mathcal{R}$ (as we make no restriction on possible values taken by explanatory variables or model parameters). However, for some exponential family distributions μ_i is restricted. For example, for the Poisson distribution $\mu_i \in \mathcal{R}_+$; for the Bernoulli distribution $\mu_i \in (0,1)$. If g is not chosen carefully, then there may exist a possible $\mathbf{x}_i$ and $\boldsymbol{\beta}$ such that $\eta_i \neq g(\mu_i)$ for any possible value of μ_i. Therefore, 'sensible' choices of link function map the set of allowed values for μ_i onto $\mathcal{R}$.

Recall that for a random variable Y with a distribution from the exponential family, $E(Y) = b'(\theta)$. Hence, for a generalised linear model

$$\mu_i = E(Y_i) = b'(\theta_i) \qquad i = 1, \ldots, n.$$

Therefore

$$\theta_i = b^{'-1}(\mu_i) \qquad i = 1, \ldots, n$$

and as $g(\mu_i) = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$, then

$$\theta_i = b^{'-1}(g^{-1}[\mathbf{x}_i^T \boldsymbol{\beta}]) \qquad i = 1, \ldots, n. \tag{5}$$

Hence, we can express the joint density (2) in terms of the coefficients $\boldsymbol{\beta}$, and for observed data $\mathbf{y}$, this is the likelihood $f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\phi})$ for $\boldsymbol{\beta}$. As $\boldsymbol{\beta}$ is our parameter of real interest (describing the dependence of the response on the explanatory variables) this likelihood will play a crucial role.

Note that considerable simplification is obtained in (5) if the functions g and $b^{'-1}$ are identical. Then

$$\theta_i = \mathbf{x}_i^T \boldsymbol{\beta} \qquad i = 1, \ldots, n.$$

and the resulting likelihood is

$$f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\phi}) = \exp\left(\sum_{i=1}^{n} \frac{y_i \mathbf{x}_i^T \boldsymbol{\beta} - b(\mathbf{x}_i^T \boldsymbol{\beta})}{a(\phi_i)} + \sum_{i=1}^{n} c(y_i, \phi_i) \right).$$

The link function

$$g(\mu) \equiv b^{'-1}(\mu)$$

is called the *canonical* link function. Under the canonical link, the canonical parameter is equal to the linear predictor.

Canonical link functions

Distribution	Normal	Poisson	Bernoulli Binomial
$b(\theta)$	$\frac{1}{2}\theta^2$	$\exp\theta$	$\log(1+\exp\theta)$
$b'(\theta) \equiv \mu$	θ	$\exp\theta$	$\dfrac{\exp\theta}{1+\exp\theta}$
$b^{'-1}(\mu) \equiv \theta$	μ	$\log\mu$	$\log\dfrac{\mu}{1-\mu}$
Link	$g(\mu) = \mu$	$g(\mu) = \log\mu$	$g(\mu) = \log\dfrac{\mu}{1-\mu}$
	Identity link	Log link	Logistic link (Logit link)

2.2.4 The linear model

Clearly the linear model considered in Section 1.9 is also a generalised linear model. We assume $Y_1, \ldots, Y_n$ are independent normally distributed random variables. The normal distribution is a member of the exponential family.

Furthermore, the explanatory variables enter a linear model through the linear predictor

$$\eta_i = \mathbf{x}_i^T \boldsymbol{\beta} \qquad i = 1, \dots, n.$$

Finally, the link between $E(\mathbf{Y}) = \boldsymbol{\mu}$ and the linear predictor $\boldsymbol{\eta}$ is through the (canonical) identity link function

$$\mu_i = \eta_i \qquad i = 1, \dots, n.$$

2.3 Maximum likelihood estimation

The regression coefficients $\beta_1, \dots, \beta_p$ describe the pattern by which the response depends on the explanatory variables. We use the observed data $y_1, \dots, y_n$ to *estimate* this pattern of dependence.

Recall the maximum likelihood estimate (m.l.e.) $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is the value of $\boldsymbol{\beta}$ which maximises $f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\phi})$ as a function of $\boldsymbol{\beta}$. Recall also that the m.l.e. has 'good' properties. It is intuitively sensible and asymptotically normal and unbiased.

As usual, we maximise the log likelihood function which, from (2), can be written

$$\log f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\phi}) = \sum_{i=1}^{n} \frac{y_i \theta_i - b(\theta_i)}{a(\phi_i)} + \sum_{i=1}^{n} c(y_i, \phi_i) \tag{6}$$

and depends on $\boldsymbol{\beta}$ through

$$\begin{aligned} &\mu_i = b'(\theta_i) \qquad i = 1, \dots, n \\ &g(\mu_i) = \eta_i \qquad i = 1, \dots, n \\ &\eta_i = \mathbf{x}_i^T \boldsymbol{\beta} = \textstyle\sum_{i=1}^{p} x_{ij} \beta_j \qquad i = 1, \dots, n. \end{aligned}$$

To find $\hat{\boldsymbol{\beta}}$, we consider the scores

$$u_k(\boldsymbol{\beta}) = \frac{\partial}{\partial \beta_k} \log f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\phi}) \qquad k = 1, \dots, p$$

and then

$$\begin{aligned} &u_k(\hat{\boldsymbol{\beta}}) = 0 \qquad k = 1, \dots, p \\ &\Rightarrow \quad \mathbf{u}(\hat{\boldsymbol{\beta}}) = \mathbf{0}. \end{aligned}$$

Now from (6)

$$\begin{aligned} u_k(\boldsymbol{\beta}) &= \frac{\partial}{\partial \beta_k} \log f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\phi}) \\ &= \frac{\partial}{\partial \beta_k} \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + \frac{\partial}{\partial \beta_k} \sum_{i=1}^{n} c(y_i, \phi_i) \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial \beta_k} \left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} \right] \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial \theta_i} \left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} \right] \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_k} \qquad k = 1, \dots, p. \\ &= \sum_{i=1}^{n} \frac{y_i - b'(\theta_i)}{a(\phi_i)} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_k} \qquad k = 1, \dots, p. \end{aligned}$$

where

$$\begin{aligned} \frac{\partial \theta_i}{\partial \mu_i} &= \left[\frac{\partial \mu_i}{\partial \theta_i} \right]^{-1} = \frac{1}{b''(\theta_i)} \\ \frac{\partial \mu_i}{\partial \eta_i} &= \left[\frac{\partial \eta_i}{\partial \mu_i} \right]^{-1} = \frac{1}{g'(\mu_i)} \\ \frac{\partial \eta_i}{\partial \beta_k} &= \frac{\partial}{\partial \beta_k} \sum_{j=1}^{p} x_{ij}\beta_j = x_{ik}. \end{aligned}$$

Therefore

$$\begin{aligned} u_k(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \frac{y_i - b'(\theta_i)}{a(\phi_i)} \frac{x_{ik}}{b''(\theta_i) g'(\mu_i)} \\ &= \sum_{i=1}^{n} \frac{y_i - \mu_i}{Var(Y_i)} \frac{x_{ik}}{g'(\mu_i)} \qquad k = 1, \dots, p(7) \end{aligned}$$

which depends on $\boldsymbol{\beta}$ through $\mu_i \equiv E(Y_i)$ and $Var(Y_i)$, $i = 1, \dots, n$.

In theory, we solve the p simultaneous equations $u_k(\hat{\boldsymbol{\beta}}) = 0,\ k = 1, \dots, p$ to evaluate $\hat{\boldsymbol{\beta}}$. In practice, these equations are usually non-linear and have no analytic solution. Therefore, we rely on numerical methods to solve them.

First, we note that the Hessian and Fisher information matrices can be derived directly from (7).

$$[\mathbf{H}(\boldsymbol{\beta})]_{jk} = \frac{\partial^2}{\partial \beta_j \partial \beta_k} \log f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\phi}) = \frac{\partial}{\partial \beta_j} u_k(\boldsymbol{\beta}).$$

Therefore

$$\begin{aligned} [\mathbf{H}(\boldsymbol{\beta})]_{jk} &= \frac{\partial}{\partial \beta_j} \sum_{i=1}^{n} \frac{y_i - \mu_i}{Var(Y_i)} \frac{x_{ik}}{g'(\mu_i)} \\ &= \sum_{i=1}^{n} \frac{-\frac{\partial \mu_i}{\partial \beta_j}}{Var(Y_i)} \frac{x_{ik}}{g'(\mu_i)} \\ &\qquad + \sum_{i=1}^{n} (y_i - \mu_i) \frac{\partial}{\partial \beta_j} \left[\frac{x_{ik}}{Var(Y_i) g'(\mu_i)} \right] \end{aligned}$$

and

$$\begin{aligned} [\mathcal{I}(\boldsymbol{\beta})]_{jk} &= \sum_{i=1}^{n} \frac{\frac{\partial \mu_i}{\partial \beta_j}}{Var(Y_i)} \frac{x_{ik}}{g'(\mu_i)} \\ &\qquad - \sum_{i=1}^{n} (E[Y_i] - \mu_i) \frac{\partial}{\partial \beta_j} \left[\frac{x_{ik}}{Var(Y_i) g'(\mu_i)} \right] \\ &= \sum_{i=1}^{n} \frac{\frac{\partial \mu_i}{\partial \beta_j}}{Var(Y_i)} \frac{x_{ik}}{g'(\mu_i)} \\ &= \sum_{i=1}^{n} \frac{x_{ij} x_{ik}}{Var(Y_i) g'(\mu_i)^2}. \end{aligned}$$

Hence we can write

$$\mathcal{I}(\boldsymbol{\beta}) = \mathbf{X}^T\mathbf{W}\mathbf{X} \tag{8}$$

where

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

$$\mathbf{W} = \mathrm{diag}(\mathbf{w}) = \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & w_n \end{pmatrix}$$

and

$$w_i = \frac{1}{Var(Y_i)g'(\mu_i)^2} \qquad i = 1, \ldots, n.$$

The Fisher information matrix $\mathcal{I}(\boldsymbol{\theta})$ depends on $\boldsymbol{\beta}$ through $\boldsymbol{\mu}$ and $Var(Y_i),\ i = 1, \ldots, n$.

We notice that the score in (7) may now be written as

$$\begin{aligned} u_k(\boldsymbol{\beta}) &= \textstyle\sum_{i=1}^n (y_i - \mu_i)x_{ik}w_i g'(\mu_i) \\ &= \textstyle\sum_{i=1}^n x_{ik}w_i z_i \qquad k = 1, \ldots, p \end{aligned}$$

where

$$z_i = (y_i - \mu_i)g'(\mu_i) \qquad i = 1, \ldots, n.$$

Therefore

$$\mathbf{u}(\boldsymbol{\theta}) = \mathbf{X}^T\mathbf{W}\mathbf{z}. \tag{9}$$

One possible method to solve the p simultaneous equations $\mathbf{u}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ that give $\hat{\boldsymbol{\beta}}$ is the (multivariate) Newton-Raphson method. [Recall that the univariate Newton-Raphson method obtains a solution to $u(\beta) = 0$ by iteratively updating β^i, the current estimate of the solution to $\beta^{i+1} = \beta^i - [u(\beta^i)/u'(\beta^i)]$.]

If $\boldsymbol{\beta}^t$ is the current estimate of $\hat{\boldsymbol{\beta}}$ then the next estimate is

$$\boldsymbol{\beta}^{i+1} = \boldsymbol{\beta}^i - \mathbf{H}(\boldsymbol{\beta}^i)^{-1}\mathbf{u}(\boldsymbol{\beta}^i). \tag{10}$$

In practice, an alternative to Newton-Raphson replaces $\mathbf{H}(\boldsymbol{\theta})$ in (10) with $E[\mathbf{H}(\boldsymbol{\theta})] \equiv -\mathcal{I}(\boldsymbol{\theta})$. Therefore, if $\boldsymbol{\beta}^i$ is the current estimate of $\hat{\boldsymbol{\beta}}$ then the next estimate is

$$\boldsymbol{\beta}^{i+1} = \boldsymbol{\beta}^i + \mathcal{I}(\boldsymbol{\beta}^i)^{-1}\mathbf{u}(\boldsymbol{\beta}^i). \tag{11}$$

The resulting iterative algorithm is called *Fisher scoring.* Notice that if we substitute (8) and (9) into (11) we get

$$\begin{aligned}\boldsymbol{\beta}^{i+1} &= \boldsymbol{\beta}^{i} + [\mathbf{X}^T\mathbf{W}^i\mathbf{X}]^{-1}\mathbf{X}^T\mathbf{W}^i\mathbf{z}^i \\ &= [\mathbf{X}^T\mathbf{W}^i\mathbf{X}]^{-1}[\mathbf{X}^T\mathbf{W}^i\mathbf{X}\boldsymbol{\beta}^i + \mathbf{X}^T\mathbf{W}^i\mathbf{z}^i] \\ &= [\mathbf{X}^T\mathbf{W}^i\mathbf{X}]^{-1}\mathbf{X}^T\mathbf{W}^i[\mathbf{X}\boldsymbol{\beta}^i + \mathbf{z}^i] \\ &= [\mathbf{X}^T\mathbf{W}^i\mathbf{X}]^{-1}\mathbf{X}^T\mathbf{W}^i[\boldsymbol{\eta}^i + \mathbf{z}^i]\end{aligned}$$

where $\boldsymbol{\eta}^i$, $\mathbf{W}^i$ and $\mathbf{z}^i$ are all functions of $\boldsymbol{\beta}^i$.

Note that this is a weighted least squares equation, that is $\boldsymbol{\beta}^{i+1}$ minimises the weighted sum of squares

$$(\boldsymbol{\eta} + \mathbf{z} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{W}(\boldsymbol{\eta} + \mathbf{z} - \mathbf{X}\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i \left(\eta_i + z_i - \mathbf{x}_i^T\boldsymbol{\beta}\right)^2$$

as a function of $\boldsymbol{\beta}$ where $w_1, \ldots, w_n$ are the weights and $\boldsymbol{\eta}+\mathbf{z}$ is called the *adjusted dependent variable.* Therefore , the Fisher scoring algorithm proceeds as follows.

1. Choose an initial estimate $\boldsymbol{\beta}^i$ for $\hat{\boldsymbol{\beta}}$ at i=0.

2. Evaluate $\boldsymbol{\eta}^i$, $\mathbf{W}^i$ and $\mathbf{z}^i$ at $\boldsymbol{\beta}^i$.

3. Calculate $\boldsymbol{\beta}^{i+1} = [\mathbf{X}^T\mathbf{W}^i\mathbf{X}]^{-1}\mathbf{X}^T\mathbf{W}^i[\boldsymbol{\eta}^i + \mathbf{z}^i]$.

4. If $||\boldsymbol{\beta}^{i+1} - \boldsymbol{\beta}^i|| >$ some prespecified (small) tolerance then set $i \rightarrow i+1$ and go to 2.

5. Use $\boldsymbol{\beta}^{i+1}$ as the solution for $\hat{\boldsymbol{\beta}}$.

As this algorithm involves iteratively minimising a weighted sum of squares, it is sometimes known as *iteratively (re)weighted least squares.*

Notes

1. Recall that the canonical link function is $g(\mu) = b^{'-1}(\mu)$ and with this link $\eta_i = g(\mu_i) = \theta_i$. Then

$$\frac{1}{g'(\mu_i)} = \frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) \qquad i = 1, \ldots, n.$$

Therefore $Var(Y_i)g'(\mu_i) = a(\phi_i)$ which does not depend on $\boldsymbol{\beta}$, and hence

$$\frac{\partial}{\partial \beta_j}\left[\frac{x_{ik}}{Var(Y_i)g'(\mu_i)}\right] = 0$$

for all $j = 1, \ldots, p$. It follows that $\mathbf{H}(\boldsymbol{\theta}) = -\mathcal{I}(\boldsymbol{\theta})$ and, for the canonical link, Newton-Raphson and Fisher scoring are equivalent.

2. The linear model is a generalised linear model with identity link, $\eta_i = g(\mu_i) = \mu_i$ and $Var(Y_i) = \sigma^2$ for all $i = 1, \ldots, n$. Therefore $w_i = [Var(Y_i)g'(\mu_i)^2]^{-1} = \sigma^{-2}$ and $z_i = (y_i - \mu_i)g'(\mu_i) = y_i - \eta_i$, $i = 1, \ldots, n$.

   Hence $\mathbf{z} + \boldsymbol{\eta} = \mathbf{y}$ and $\mathbf{W} = \sigma^{-2}\mathbf{I}$, neither of which depend on $\boldsymbol{\beta}$. Hence, the Fisher scoring algorithm converges in a single iteration to the usual least squares estimate.

3. Estimation of an unknown scale parameter σ^2 is discussed later. A common (to all i) σ^2 has no effect on $\hat{\boldsymbol{\beta}}$.

2.4 Inference

Recall from Section 1.7.5 that the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ is asymptotically normally distributed with mean $\boldsymbol{\beta}$ (it is unbiased) and variance covariance matrix $\mathcal{I}(\boldsymbol{\theta})^{-1}$. For 'large enough n' we treat this distribution as an approximation.

Therefore, standard errors (estimated standard deviations) are given by

$$s.e.(\hat{\beta}_i) = [\mathcal{I}(\hat{\boldsymbol{\beta}})^{-1}]_{ii}^{\frac{1}{2}} = [(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}]_{ii}^{\frac{1}{2}} \qquad i = 1, \ldots, p.$$

where the diagonal matrix $\hat{\mathbf{W}} = \text{diag}(\hat{\mathbf{w}})$ is evaluated at $\hat{\boldsymbol{\beta}}$, that is $\hat{w}_i = (\hat{Var}(Y_i)g'(\hat{\mu}_i)^2)^{-1}$ where $\hat{\mu}_i$ and $\hat{Var}(Y_i)$ are evaluated at $\hat{\boldsymbol{\beta}}$ for $i = 1, \ldots, n$. Furthermore, if $Var(Y_i)$ depends on an unknown scale parameter, then this too must be estimated in the standard error.

The asymptotic distribution of the maximum likelihood estimator can be used to provide approximate large sample confidence intervals. We can find h such that

$$P\left(-h \le \frac{\hat{\beta}_i - \beta_i}{[\mathcal{I}(\boldsymbol{\theta})^{-1}]_{ii}^{\frac{1}{2}}} \le h\right) = \alpha.$$

Therefore

$$P\left(\hat{\beta}_i - h[\mathcal{I}(\boldsymbol{\theta})^{-1}]_{ii}^{\frac{1}{2}} \le \beta_i \le \hat{\beta}_i + h[\mathcal{I}(\boldsymbol{\theta})^{-1}]_{ii}^{\frac{1}{2}}\right) = \alpha.$$

The endpoints of this interval cannot be evaluated because they also depend on the unknown parameter vector $\boldsymbol{\beta}$. However, if we replace $\mathcal{I}(\boldsymbol{\theta})$ by its m.l.e. $\mathcal{I}(\hat{\boldsymbol{\beta}})$ we obtain the approximate large sample $100\alpha\%$ confidence interval

$$[\hat{\beta}_i - s.e.(\hat{\beta}_i)h\,,\; \hat{\beta}_i + s.e.(\hat{\beta}_i)h].$$

For $\alpha = 0.9, 0.95, 0.99$, $h = 1.64, 1.96, 2.58$.

2.5 Comparing generalised linear models

2.5.1 The generalised likelihood ratio test

If we have a set of competing generalised linear models which might explain the dependence of the response on the explanatory variables, we will want to determine which of the models is most appropriate. Recall that we have three main requirements of a statistical model; plausibility, parsimony and goodness of fit, of which parsimony and goodness of fit are statistical issues.

As with linear models, we proceed by comparing models pairwise using a generalised likelihood ratio test. This kind of comparison is restricted to situations where one of the models, H_0, is *nested* in the other, H_1. Then the asymptotic distribution of the log likelihood ratio statistic under H_0 is a chi-squared distribution with known degrees of freedom.

For generalised linear models, 'nested' means that H_0 and H_1 are

1. based on the same exponential family distribution, and
2. have the same link function, but
3. the explanatory variables present in H_0 are a subset of those present in H_1.

We will assume that model H_1 contains p linear parameters and model H_0 a subset of $q < p$ of these. Without loss of generality, we can think of H_1 as the model

$$\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j \qquad i = 1, \dots , n$$

and H_0 is the same model with

$$\beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = 0.$$

Then model H_0 is a special case of model H_1, where certain coefficients are set equal to zero, and therefore $\Theta^{(0)}$, the set of values of the canonical parameter $\boldsymbol{\theta}$ allowed by H_0, is a subset of $\Theta^{(1)}$, the set of values allowed by H_1.

Now, the log likelihood ratio statistic for a test of H_0 against H_1 is

$$\begin{aligned} L_{01} &\equiv 2\log\left(\frac{\max_{\boldsymbol{\theta}\in\Theta^{(1)}} f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})}{\max_{\boldsymbol{\theta}\in\Theta^{(0)}} f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})}\right) \\ &= 2\log f_{\mathbf{Y}}(\mathbf{y};\hat{\boldsymbol{\theta}}^{(1)}) - 2\log f_{\mathbf{Y}}(\mathbf{y};\hat{\boldsymbol{\theta}}^{(0)})(12) \end{aligned}$$

where $\hat{\boldsymbol{\theta}}^{(1)}$ and $\hat{\boldsymbol{\theta}}^{(0)}$ follow from $b'(\hat{\theta}_i) = \hat{\mu}_i$, $g(\hat{\mu}_i) = \hat{\eta}_i$, $i = 1, \ldots, n$ where $\hat{\boldsymbol{\eta}}$ for each model is the linear predictor evaluated at the corresponding maximum likelihood estimate for $\boldsymbol{\beta}$. Here, we assume that $a(\phi_i)$, $i = 1, \ldots, n$ are known; unknown $a(\phi)$ is discussed in Section 2.6.

Recall that we reject H_0 in favour of H_1 when L_{01} is 'too large' (the observed data are much more probable under H_1 than H_0). To determine a threshold value k for L_{01}, beyond which we reject H_0, we set the size of the test α and use the result of Section 1.8.1 that, because H_0 is nested in H_1, L_{01} has an asymptotic chi-squared distribution with $p-q$ degrees of freedom. Usually $\alpha = 0.05$ and hence we reject H_0 in favour of H_1 when L_{01} is greater than the 95% point of the χ^2_{p-q} distribution.

Note that setting up our model selection procedure in this way is consistent with our desire for parsimony. The simpler model is H_0, and we do not reject H_0 in favour of the more complex model H_1 unless the data provide convincing evidence for H_1 over H_0, that is unless H_1 provides fits the data significantly better.

2.5.2 Scaled deviance and the saturated model

Consider a model where $\boldsymbol{\beta}$ is n-dimensional, and therefore $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$. Assuming that $\mathbf{X}$ is invertible, then this model places no constraints on the linear predictor $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$. It can take any value in $\mathcal{R}^n$. Correspondingly the means $\boldsymbol{\mu}$ and the canonical parameters $\boldsymbol{\theta}$ are unconstrained. The model is of dimension n and can be parameterised equivalently using $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, $\boldsymbol{\mu}$ or $\boldsymbol{\theta}$. Such a model is called the *saturated* model.

As the canonical parameters $\boldsymbol{\theta}$ are unconstrained, we can calculate their maximum likelihood estimates $\hat{\boldsymbol{\theta}}$ directly from their likelihood (2) (without first having to calculate $\hat{\boldsymbol{\beta}}$)

$$\log f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + \sum_{i=1}^{n} c(y_i, \phi_i). \tag{13}$$

We obtain $\hat{\boldsymbol{\theta}}$ by first differentiating with respect to $\theta_1, \ldots, \theta_n$ to give

$$\frac{\partial}{\partial\theta_k}\log f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta}) = \frac{y_k - b'(\theta_k)}{a(\phi_k)} \qquad k = 1, \ldots, n.$$

Therefore $b'(\hat{\theta}_k) = y_k,\ k = 1, \dots, n$, and it follows immediately that $\hat{\mu}_k = y_k,\ k = 1, \dots, n$. Hence the saturated model fits the data perfectly, as the *fitted values* $\hat{\mu}_k$ and observed values y_k are the same for every observation $k = 1, \dots, n$.

The saturated model is rarely of any scientific interest in its own right. It is highly parameterised, having as many parameters as there are observations. This goes against our desire for parsimony in a model. However, every other model is necessarily nested in the saturated model, and a test comparing a model H_0 against the saturated model H_S can be interpreted as a goodness of fit test. If the saturated model, which fits the observed data perfectly, does not provide a significantly better fit than model H_0, we can conclude that H_0 is an acceptable fit to the data.

The log likelihood ratio statistic for a test of H_0 against H_S is, from (12)

$$L_{0s} = 2 \log f_{\mathbf{Y}}(\mathbf{y}; \hat{\boldsymbol{\theta}}^{(s)}) - 2 \log f_{\mathbf{Y}}(\mathbf{y}; \hat{\boldsymbol{\theta}}^{(0)})$$

where $\hat{\boldsymbol{\theta}}^{(s)}$ follows from $b'(\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{\mu}} = \mathbf{y}$ and $\hat{\boldsymbol{\theta}}^{(0)}$ is a function of the corresponding maximum likelihood estimate for $\boldsymbol{\beta} = (\beta_1, \dots, \beta_q)^T$. Under H_0, L_{0s} has an asymptotic chi-squared distribution with $n - q$ degrees of freedom. Therefore, if L_{0s} is 'too large' (in practice, larger than the 95% point of the χ^2_{n-q} distribution) then we reject H_0 as a plausible model for the data, as it does not fit the data adequately.

The *degrees of freedom* of model H_0 is defined to be the degrees of freedom for this test, $n - q$, the number of observations minus the number of linear parameters of H_0. We call L_{0s} the *scaled deviance* (S-Plus calls it the *residual deviance*) of model H_0.

From (12) and (13) we can write the deviance of model H_0 as

$$L_{0s} = 2 \sum_{i=1}^{n} \frac{y_i[\hat{\theta}_i^{(s)} - \hat{\theta}_i^{(0)}] - [b(\hat{\theta}_i^{(s)}) - b(\hat{\theta}_i^{(0)})]}{a(\phi_i)}. \tag{14}$$

which can be calculated using the observed data, provided that $a(\phi_i),\ i = 1, \dots, n$ is known.

Notes

1. The log likelihood ratio statistic (12) for testing H_0 against a non-saturated alternative H_1 can be written as

$$\begin{aligned} L_{01} &= 2 \log f_{\mathbf{Y}}(\mathbf{y}; \hat{\boldsymbol{\theta}}^{(1)}) - 2 \log f_{\mathbf{Y}}(\mathbf{y}; \hat{\boldsymbol{\theta}}^{(0)}) \\ &= [2 \log f_{\mathbf{Y}}(\mathbf{y}; \hat{\boldsymbol{\theta}}^{(s)}) - 2 \log f_{\mathbf{Y}}(\mathbf{y}; \hat{\boldsymbol{\theta}}^{(0)})] - [2 \log f_{\mathbf{Y}}(\mathbf{y}; \hat{\boldsymbol{\theta}}^{(s)}) - 2 \log f_{\mathbf{Y}}(\mathbf{y}; \hat{\boldsymbol{\theta}}^{(1)})] \\ &= L_{0s} - L_{1s}. \end{aligned} \tag{15}$$

Therefore the log likelihood ratio statistic for comparing two nested models is the difference of their deviances. Furthermore, as $p - q = (n-q) - (n-p)$, the degrees of freedom for the test is the difference in degrees of freedom of the two models.

2. The asymptotic theory used to derive the distribution of the log likelihood ratio statistic under H_0 does not really apply to the goodness of fit test (comparison with the saturated model). However, for binomial or Poisson data, we can proceed as long as the relevant binomial or Poisson distributions are likely to be reasonably approximated by normal distributions (*i.e.* for binomials with large denominators or Poissons with large means). However, for Bernoulli data, we cannot use the scaled deviance as a goodness of fit statistic in this way.

3. An alternative goodness of fit statistic for a model H_0 is Pearson's X^2 given by
$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i^{(0)})^2}{\hat{Var}(Y_i)}. \tag{16}$$
X^2 is small when the squared differences between observed and fitted values (scaled by variance) is small. Hence, large values of X^2 correspond to poor fitting models. In fact, X^2 and L_{0s} are asymptotically equivalent and under H_0, X^2, like L_{0s}, has an asymptotic chi-squared distribution with $n-q$ degrees of freedom. However, the asymptotics associated with X^2 are often more reliable for small samples, so if there is a discrepancy between X^2 and L_{0s}, it is usually safer to base a test of goodness of fit on X^2.

4. Although the deviance for a model is expressed in (14) in terms of the maximum likelihood estimates of the canonical parameters, it is more usual to express it in terms of the maximum likelihood estimates $\hat{\mu}_i,\ i = 1, \ldots, n$ of the mean parameters. For the saturated model, these are just the observed values $y_i,\ i = 1, \ldots, n$, and for the model of interest, H_0, we call them the *fitted values*. Hence, for a particular generalised linear model, the scaled deviance function describes how discrepancies between the observed and fitted values are penalised.

♡ **Example** 2.5. $Y_i \sim$ **Poisson**$(\lambda_i),\ i = 1, \ldots, n$ Recall from 2.1 that $\theta = \log\lambda$, $b(\theta) = \exp\theta$, $\mu = b'(\theta) = \exp\theta$ and $Var(Y) = a(\phi)V(\mu) = 1 \cdot \mu$. Therefore, by (14) and (16)

$$\begin{aligned} L_{0s} &= 2\textstyle\sum_{i=1}^{n} y_i[\log\hat{\mu}_i^{(s)} - \log\hat{\mu}_i^{(0)}] - [\hat{\mu}_i^{(s)} - \hat{\mu}_i^{(0)}] \\ &= 2\textstyle\sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\hat{\mu}_i^{(0)}}\right) - y_i + \hat{\mu}_i^{(0)} \end{aligned}$$

and

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i^{(0)})^2}{\hat{\mu}_i^{(0)}}.$$

♡ **Example** 2.6. $Y_i \sim$ **Binomial**$(n_i, p_i),\ i = 1, \dots, n$ Recall from Section 2.1 that $\theta = \log \frac{p}{1-p}$, $b(\theta) = \log(1 + \exp\theta)$, $\mu = b'(\theta) = \frac{\exp\theta}{1+\exp\theta}$ and $Var(Y) = a(\phi)V(\mu) = \frac{1}{n} \cdot \mu(1-\mu)$. Therefore, by (14) and (16)

$$\begin{aligned} L_{0s} &= 2\textstyle\sum_{i=1}^n n_i y_i \left[\log \frac{\hat{\mu}_i^{(s)}}{1-\hat{\mu}_i^{(s)}} - \log \frac{\hat{\mu}_i^{(0)}}{1-\hat{\mu}_i^{(0)}}\right] \\ &\qquad + n_i[\log(1-\hat{\mu}_i^{(s)}) - \log(1-\hat{\mu}_i^{(0)})] \\ &= 2\textstyle\sum_{i=1}^n n_i y_i \log\left(\frac{y_i}{\hat{\mu}_i^{(0)}}\right) + n_i(1-y_i)\log\left(\frac{1-y_i}{1-\hat{\mu}_i^{(0)}}\right) \end{aligned}$$

and

$$X^2 = \sum_{i=1}^{n} \frac{n_i(y_i - \hat{\mu}_i^{(0)})^2}{\hat{\mu}_i^{(0)}(1-\hat{\mu}_i^{(0)})}.$$

Bernoulli data are binomial with $n_i = 1,\ i = 1, \dots, n$.

2.6 Models with unknown $a(\phi)$

Thus far, the theory of Section 2.5 has assumed that $a(\phi)$ is known. This is the case for both the Poisson distribution ($a(\phi) = 1$) and the binomial distribution ($a(\phi) = 1/n$). Neither the scaled deviance (14) nor Pearson X^2 statistic (16) can be evaluated unless $a(\phi)$ is known. Therefore, when $a(\phi)$ is not known, we cannot use the scaled deviance as a measure of goodness of fit, or to compare models using (15). For such models, there is no equivalent goodness of fit test, but we can develop a test for comparing nested models.

Here assume that $a(\phi_i) = \sigma^2/m_i,\ i = 1, \dots, n$ where σ^2 is a common unknown scale parameter and $m_1, \dots, m_n$ are known weights. (A normal generalised linear model takes this form, if we assume that $Var(Y_i) = \sigma^2,\ i = 1, \dots, n$, in which case $m_i = 1,\ i = 1, \dots, n$). Under this assumption

$$\begin{aligned} L_{0s} &= \tfrac{2}{\sigma^2}\textstyle\sum_{i=1}^n m_i y_i[\hat{\theta}_i^{(s)} - \hat{\theta}_i^{(0)}] - m_i[b(\hat{\theta}_i^{(s)}) - b(\hat{\theta}_i^{(0)})] \\ &= \tfrac{1}{\sigma^2} D_{0s} \qquad\qquad (17) \end{aligned}$$

where D_{0s} is defined to be twice the sum above, which can be calculated using the observed data. We call D_{0s} the *deviance* of the model.

In order to test nested models H$_0$ and H$_1$ as set up in Section 2.5.1 we calculate the test statistic

$$\begin{aligned} F &= \tfrac{L_{01}/(p-q)}{L_{1s}/(n-p)} = \tfrac{(L_{0s}-L_{1s})/(p-q)}{L_{1s}/(n-p)} \\ &= \tfrac{\left(\frac{1}{\sigma^2}D_{0s}-\frac{1}{\sigma^2}D_{1s}\right)/(p-q)}{\frac{1}{\sigma^2}D_{1s}/(n-p)} = \tfrac{(D_{0s}-D_{1s})/(p-q)}{D_{1s}/(n-p)}. \end{aligned} \tag{18}$$

This statistic does not depend on the unknown scale parameter σ^2, so can be calculated using the observed data. Asymptotically, if H$_0$ is true, we know that L_{01} has a χ^2_{p-q} distribution and L_{1s} has a χ^2_{n-p} distribution. furthermore, L_{01} and L_{1s} are independent (not proved here) so F has an asymptotic F distribution with $p-q$ degrees of freedom in the numerator and $n-p$ degrees of freedom in the denominator. Hence, we compare nested generalised linear models by calculating F and rejecting H$_0$ in favour of H$_1$ if F is too large (in practice, greater than the 95% point of the relevant F distribution).

The dependence of the maximum likelihood equations $\mathbf{u}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ on σ^2 (where $\mathbf{u}$ is given by (7)) can be eliminated by multiplying through by σ^2. However, inference based on the maximum likelihood estimates, as described in Section **??** does require knowledge of σ^2. This is because asymptotically $Var(\hat{\boldsymbol{\beta}})$ is the Fisher information matrix $\mathcal{I}(\boldsymbol{\beta}) = \mathbf{X}^T\mathbf{W}\mathbf{X}$, and this depends on $w_i = \frac{1}{Var(Y_i)g'(\mu_i)^2}$ where $Var(Y_i) = a(\phi_i)b''(\theta_i) = \sigma^2 b''(\theta_i)/m_i$ here.

Therefore, to calculate standard errors and confidence intervals, we need to supply an estimate $\hat{\sigma}^2$ of σ^2. Generally, we do not use the maximum likelihood estimate. Instead, we notice that, from (17), $L_{0s} = D_{0s}/\sigma^2$, and we know that asymptotically, if model H$_0$ is an adequate fit, L_{0s} has a χ^2_{n-q} distribution. Hence

$$E(L_{0s}) = E\left(\frac{1}{\sigma^2}D_{0s}\right) = n-q \quad \Rightarrow \quad E\left(\frac{1}{n-q}D_{0s}\right) = \sigma^2.$$

Therefore the deviance of a model divided by its degrees of freedom is an asymptotically unbiased estimator of the scale parameter σ^2. Hence $\hat{\sigma}^2 = D_{0s}/(n-q)$.

An alternative estimator of σ^2 is based on the Pearson X^2 statistic. As $Var(Y) = a(\phi)V(\mu) = \sigma^2 V(\mu)/m$ here, then from (16)

$$X^2 = \frac{1}{\sigma^2}\sum_{i=1}^n \frac{m_i(y_i - \hat{\mu}_i^{(0)})^2}{V(\hat{\mu}_i)}. \tag{19}$$

Again, if H$_0$ is an adequate fit, X^2 has an chi-squared distribution with $n-q$ degrees of freedom, so

$$\hat{\sigma}^2 = \frac{1}{n-q}\sum_{i=1}^n \frac{m_i(y_i - \hat{\mu}_i^{(0)})^2}{V(\hat{\mu}_i)}$$

is an alternative unbiased estimator of σ^2. This estimator tends to be more reliable in small samples.

♡ **Example** 2.7. $Y_i \sim$ **Normal**(μ_i, σ^2), $i = 1, \ldots, n$ Recall from Section 2.1 that $\theta = \mu$, $b(\theta) = \theta^2/2$, $\mu = b'(\theta) = \theta$ and $Var(Y) = a(\phi)V(\mu) = \sigma^2 \cdot 1$, so $m_i = 1$, $i = 1, \ldots, n$. Therefore, by (17)

$$\begin{aligned} D_{0s} &= 2\sum_{i=1}^n y_i[\hat{\mu}_i^{(s)} - \hat{\mu}_i^{(0)}] - [\tfrac{1}{2}\hat{\mu}_i^{(s)^2} - \tfrac{1}{2}\hat{\mu}_i^{(0)^2}] \\ &= \sum_{i=1}^n [y_i - \hat{\mu}_i^{(0)}]^2 \qquad (20) \end{aligned}$$

which is just the residual sum of squares for model H_0. Therefore, we estimate σ^2 for a normal g.l.m. by its residual sum of squares for the model divided by its degrees of freedom. From (19), the estimate for σ^2 based on X^2 is identical.

We compare normal generalised linear models, using the F statistic (18) where the deviances D_{0s} and D_{1s} for the two nested models being compared are their residual sums of squares.

2.7 Residuals

Recall that for linear models, we define the residuals to be the differences between the observed and fitted values $y_i - \hat{\mu}_i^{(0)}$, $i = 1, \ldots, n$. From (20) we notice that both the scaled deviance and Pearson X^2 statistic for a normal g.l.m. are the sum of the squared residuals divided by σ^2. We can generalise this to define residuals for other other generalised linear models in a natural way.

For any g.l.m. we define the *Pearson residuals* to be

$$r_i^P = \frac{y_i - \hat{\mu}_i^{(0)}}{V\hat{a}r(Y_i)^{\frac{1}{2}}} \qquad i = 1, \ldots, n.$$

Then, from (16), X^2 is the sum of the squared Pearson residuals.

For any g.l.m. we define the *deviance residuals* to be

$$r_i^D = \text{sign}(y_i - \hat{\mu}_i^{(0)}) \left[\frac{y_i[\hat{\theta}_i^{(s)} - \hat{\theta}_i^{(0)}] - [b(\hat{\theta}_i^{(s)}) - b(\hat{\theta}_i^{(0)})]}{a(\phi_i)} \right]^{\frac{1}{2}}$$
$$i = 1, \ldots, n,$$

where $\text{sign}(x) = 1$ if $x > 0$ and -1 if $x < 0$. Then, from (14), the scaled deviance, L_{0s}, is the sum of the squared deviance residuals.

When $a(\phi) = \sigma^2/m$ and σ^2 is unknown, as in Section 2.6 the residuals are based on (17) and (19), and the expressions above need to be multiplied through by σ^2 to eliminate dependence on the unknown scale parameter. Therefore, for a normal g.l.m. the Pearson and deviance residuals are both equal to the usual residuals, $y_i - \hat{\mu}_i^{(0)}$, $i = 1, \dots, n$.

Residual plots are most commonly of use in normal linear models, where they provide an essential check of the model assumptions. This kind of check is less important for a model without an unknown scale parameter as the scaled deviance provides a useful overall assessment of fit which takes into account most aspects of the model.

However, when data have been collected in serial order, a plot of the deviance or Pearson residuals against the order may again be used as a check for potential serial correlation.

Otherwise, residual plots are most useful when a model fails to fit (scaled deviance is too high). Then, examining the residuals may give an indication of the reason(s) for lack of fit. For example, there may be a small number of outlying observations.

A plot of deviance or Pearson residuals against the linear predictor should produce something that looks like a random scatter. If not, then this may be due to incorrect link function, wrong scale for an explanatory variable, or perhaps a missing polynomial term in an explanatory variable.