An Introduction to Generalized Linear Models

1.1 Why Generalized Linear Models?

In a linear regression model we assume that the responses $Y_1, \ldots, Y_n$ are independent and satisfy $E(Y_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$, where $\boldsymbol{x}_i$ is an observed vector of predictors (covariates) associated with the ith observation and $\boldsymbol{\beta}$ is an unknown vector of regression coefficients. If we assume in addition that the Y_is are normally distributed with common, but unknown variance $\sigma^2 > 0$, i.e.,

$$Y_i \sim \text{independent } N(\boldsymbol{x}_i^T \boldsymbol{\beta},\, \sigma^2), \quad i = 1, \ldots, n,$$

then the usual tests and confidence regions based on the t and F distributions provide exact inferences about $\boldsymbol{\beta}$.

If we drop the normality assumption, but assume still that $Y_1, \ldots, Y_n$ are independent with $E(Y_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$ and $\text{Var}(Y_i) = \sigma^2$, abbreviated

$$Y_i \sim \text{independent } (\boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma^2), \quad i = 1, \ldots, n,$$

then under reasonable regularity conditions the standard tests and confidence intervals are approximately correct when the sample size n is large; in other words, the standard methods are *robust* to violation of the normality assumption.

However, in many situations even these weakened assumptions are untenable, i.e., either

- $\mu(\boldsymbol{x}) := E(Y|\boldsymbol{x})$ is not a linear function of $\boldsymbol{x}$, and/or
- $\text{Var}(Y_i)$ is not constant in i (heterogeneity of variance).

Possible solutions to these problems include:

– weighted least squares, if $\text{Var}(Y_i) = a_i^2 \sigma^2$, where $a_1^2, \ldots, a_n^2$ are known constants;

– transformations of $\boldsymbol{x}$ to correct nonlinearity; and

– transformations of Y if either $\mu(\boldsymbol{x}) = E(Y|\boldsymbol{x})$ is nonlinear or $\sigma^2(\boldsymbol{x}) := \text{Var}(Y|\boldsymbol{x})$ is not constant in $\boldsymbol{x}$ (variance stabilizing transformations).

Transformation of the response, i.e., replacing Y by $Y^* = h(Y)$, for some monotonic, non-affine function $h(\cdot)$, may work to address nonlinearity or heterogeneity of variance. However,

- we are now modeling $\mu^* = E(Y^*) = E[h(Y)]$ rather than $\mu = E(Y)$. In general, for non-affine $h(\cdot)$,

  $$\mu^* = E[h(Y)] \neq h(E[Y]) = h(\mu) \implies \mu \neq h^{-1}(\mu^*),$$

  so we cannot use $h^{-1}(\hat{\mu}^*)$ to directly estimate μ.

- there may not be a single transformation that achieves both linearity and homogeneity of variance;

- for some responses, especially binary responses, transformation of Y may not be useful at all.

Example (Space Shuttle Challenger) The space shuttle Challenger crashed on Jan-uary 28, 1986, due to the failure of O-rings in its solid-rocket boosters. Data available from 23 shuttle flights prior to the crash included

- the number of primary O-rings (out of 6) experiencing "thermal distress", and

- the temperature at launch time.

Let

$$\pi = \Pr(\text{at least one O-ring experiences thermal stress}) = \Pr(Y = 1),$$

where

$$Y = I(\text{at least one O-ring experiences thermal stress}).$$

We are interested in modeling π as a function of $x =$ temperature. We will revisit this example in more detail in Chapter 2 (Example 2.1, p. 2-2). □

For binary responses, $\mu = E(Y) = \Pr(Y = 1) = \pi$, so $0 \leq \mu \leq 1$ regardless of the value of $\boldsymbol{x}$. Thus, the model $\mu = \boldsymbol{x}^T\boldsymbol{\beta}$ is not tenable unless $\boldsymbol{x}$ is restricted to a bounded domain, since $\beta_j x_j$ is unbounded if $\beta_j \neq 0$ and x_j is unbounded. Also, if $\mu = \pi$ depends on $\boldsymbol{x}$, then so does

$$\text{Var}(Y) = \pi(1 - \pi),$$

so homogeneity of variance is ruled out in all but trivial cases. This heterogeneity cannot be repaired by transformation of Y, since the transformed response still takes only two values. More precisely, for any monotone transformation $h(\cdot)$,

$$h(Y) = (b - a)Y + a, \quad a = h(0), \quad b = h(1),$$

so any transformation of a binary Y is effectively an affine transformation and hence does nothing to address nonlinearity or heterogeneity of variance.

1.2 Defining a GLM

1.2.1 Systematic Component

The systematic component of a GLM relates the vector of predictors $\boldsymbol{x} = (x_1, \ldots, x_p)$ to the mean response $\mu = E(Y)$, and consists of two parts:

- the *linear predictor*: $\eta = \boldsymbol{x}^T\boldsymbol{\beta}$;
- the *link function*: $g(\mu) = \eta$, where g is a smooth, monotonic function.

Example An ordinary linear model assumes $\mu = \boldsymbol{x}^T\ \boldsymbol{\beta}$, so it uses the *identity link*, $g(\mu) = \mu$. □

Example For binary responses, $\mu = \pi := P(Y\!=1)$, and the logistic regression model assumes that

$$\log\left(\frac{\pi}{1-\pi}\right) = \boldsymbol{x}^T\boldsymbol{\beta}.$$

Thus, logistic regression assumes the so-called *logit* link,

$$g(\mu) = \text{logit}(\mu) = \log\left(\frac{\mu}{1-\mu}\right).$$

The inverse of the logit link function is sometimes called the *expit* function,

$$g^{-1}(\eta) = \text{expit}(\eta) = \frac{e^\eta}{1+e^\eta}.$$ □

Remark 1.1 It may seem more natural to think of the model as specifying $\mu = h(\eta)$, where $h = g^{-1}$ is the inverse link function, but GLMs are almost always defined in terms of the link function rather than the inverse link, so we may as well get used to it. □

1.2.2 Random (or Stochastic) Component

The stochastic component of a GLM specifies the distributional form of the responses. It is assumed that

- $Y_1, \ldots, Y_n$ are independent; and
- Y_i has density $f(\cdot;\theta_i,\phi_i)$, $\phi_i = \phi/a_i$, $\phi > 0$, where $a_1, \ldots, a_n$ are known *prior weights*,[1] and f has the form

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

Remark 1.2
- More precisely, f is a density with respect to some fixed measure m, usually either Lebesgue measure (for continuous responses) or counting measure (for discrete responses).

[1] To my knowledge, the use of the word "prior" here has no Bayesian connotation; it is simply used to distinguish these weights from the "iterative weights" to be encountered later.

- For fixed ϕ, $f(\cdot;\theta,\phi)$ defines a one-parameter exponential family with canonical parameter θ. When ϕ is also included as a parameter, it is called an *exponential dispersion* family (Jørgensen, 1997).
- The *dispersion parameter* ϕ may be known or unknown, depending on the particular exponential dispersion family.
- McCullagh and Nelder (1989) call $b(\theta)$ the *cumulant function*. This is not exactly the cumulant generating function of $f(\cdot;\theta,\phi)$, but it is related. There is more on this in the exercises. □

Let

$$\ell(\theta,\phi) = \log f(y;\theta,\phi) = \frac{1}{\phi}[y\theta - b(\theta)] + c(y;\phi).$$

Then

$$\frac{\partial \ell}{\partial \theta} = \frac{1}{\phi}[y - b'(\theta)], \qquad \frac{\partial^2 \ell}{\partial \theta^2} = -\frac{1}{\phi} b''(\theta),$$

and from the first two Bartlett identities,[2]

$$E_{\theta,\phi}\left(\frac{\partial \ell}{\partial \theta}\right) = 0 \qquad \text{and} \qquad E_{\theta,\phi}\left[\left(\frac{\partial \ell}{\partial \theta}\right)^2\right] = E_{\theta,\phi}\left(-\frac{\partial^2 \ell}{\partial \theta^2}\right),$$

we see that

$$E_{\theta,\phi}\left\{\frac{1}{\phi}[Y - b'(\theta)]\right\} = \frac{1}{\phi}\Big[E_{\theta,\phi}(Y) - b'(\theta)\Big] = 0 \implies E_{\theta,\phi}(Y) = b'(\theta), \tag{1.1}$$

and

$$\frac{1}{\phi^2}\underbrace{E_{\theta,\phi}\{[y - b'(\theta)]^2}_{=\mathrm{Var}_{\theta,\phi}(Y)} = \frac{1}{\phi} b''(\theta) \implies \mathrm{Var}_{\theta,\phi}(Y) = \phi\, b''(\theta). \tag{1.2}$$

Remark 1.3 All of these calculations can be justified rigorously, and we may revisit the details later in an exercise. For now, *assume* that the domain of b is an interval, that the required derivatives of b exist, and that differentiating across the integral can be justified. □

Remark 1.4

- Assuming that $\mathrm{Var}_{\theta,\phi}(Y) > 0$ for all θ of interest, equation (1.2) and the assumption $\phi > 0$ together imply that $b''(\theta) > 0$ for all θ in the (interval) domain of b, so that $b'(\theta)$ is strictly increasing and $b(\theta)$ is strictly convex.
- Since b' is strictly increasing, it is invertible and from (1.1),

  $$\mu = b'(\theta) \implies \theta = (b')^{-1}(\mu). \tag{1.3}$$

  Of course the systematic component of the model specifies $g(\mu) = \eta = \boldsymbol{x}^T\boldsymbol{\beta}$, so

  $$\mu = g^{-1}(\eta) = g^{-1}(\boldsymbol{x}^T\boldsymbol{\beta}),$$

[2] Obtained by differentiating with respect to θ across the integral in the identity $\int f(y;\theta,\phi)\,dm(y) \equiv 1$

and thus by (1.3),

$$\theta = (b')^{-1}\big(g^{-1}(\eta)\big) = (b')^{-1}\big(g^{-1}(\boldsymbol{x}^T\boldsymbol{\beta})\big). \tag{1.4}$$

This means that the random and systematic components of the model specify the distribution of $Y_1, \dots, Y_n$ up to the vector of unknown regression parameters $\boldsymbol{\beta}$ and a possibly unknown dispersion parameter ϕ.

- The function $(b')^{-1}(\mu)$ is called the *canonical link* for the model. If we choose $g = (b')^{-1}$, then by (1.4),

  $$\theta = (b')^{-1}\big(g^{-1}(\eta)\big) = (b')^{-1}\big(b'(\eta)\big) = \eta.$$

  Thus, by using the canonical link function we are assuming that the canonical parameter θ is equal to the linear predictor $\eta = \boldsymbol{x}^T\boldsymbol{\beta}$. We will see later that use of the canonical link simplifies some calculations, though this fact alone should not be taken as justification for using the canonical link in a particular application.

- Substituting (1.3) into (1.2), we have

  $$\mathrm{Var}(Y) = \phi\, b''\big((b')^{-1}(\mu)\big) = \phi\, V(\mu),$$

  where the function

  $$V(\mu) := b''\big((b')^{-1}(\mu)\big) \tag{1.5}$$

  is called the *variance function* of the model.

- If g is the canonical link, $g(\mu) = (b')^{-1}(\mu)$, then by the inverse function theorem,

  $$g'(\mu) = \frac{1}{b''\big((b')^{-1}(\mu)\big)} = \frac{1}{V(\mu)}.$$

  Thus the derivative of the canonical link is the reciprocal of the variance function. □

Example ($\boldsymbol{N(\mu, \sigma^2)}$)

$$\begin{aligned} f(y) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\} \\ &= \exp\left\{\frac{1}{\sigma^2}\Big(y\mu - \tfrac{1}{2}\mu^2\Big) - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right\} \end{aligned}$$

$$\Longrightarrow \left\{ \begin{array}{cc} \theta = \mu & \phi = \sigma^2 \\ b(\theta) = \tfrac{1}{2}\theta^2 & c(y;\phi) = -\dfrac{1}{2}\left(\dfrac{y^2}{\phi} + \log(2\pi\phi)\right) \end{array} \right.$$

Note that in this example $b'(\theta) = \theta$, the identity function. The inverse of the identity function is the identity, so the canonical link for the normal family is the identity function. Of course, $b''(\theta) \equiv 1 \implies V(\mu) \equiv 1$.

Example (Poisson($\boldsymbol{\mu}$)) Here

$$f(y) = \frac{\mu^y}{y!}e^{-\mu} = \exp\big\{ y\log\mu - \mu - \log(y!) \quad = \exp\big\{ \big(y\theta - e^{\theta}\big) - \log(y!) \quad , \qquad y = 0, 1, 2, \ldots$$

which has the exponential dispersion form with

$$\begin{aligned} \theta &= \log\mu & \phi &= 1 \\ b(\theta) &= e^{\theta} & c(y;\phi) &= -\log(y!). \end{aligned}$$

Obviously the canonical link is the log function, and since $\mathrm{Var}(Y) = \mu$, the variance function is the identity, i.e.,

$$\begin{aligned} (b')^{-1}(\mu) &= \log\mu \qquad \text{(canonical link)} \\ V(\mu) &= b''\big((b')^{-1}(\mu)\big) = \mu. \end{aligned}$$

Note that ϕ is known and fixed for the Poisson model, as it is for the binomial. □

See Table 2.1 of McCullagh and Nelder (1989, p. 30) for a list of the five standard exponential dispersion families, their canonical links, variance functions, and so forth.

1.3 The GLM Likelihood

Recall that in a GLM, $Y_1, \ldots, Y_n$ are independent with

$$Y_i \sim f(\cdot;\theta_i,\phi_i), \quad \phi_i = \phi/a_i,$$

where

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

Further,

θ_i is a function of μ_i: $\mu_i = b'(\theta_i) \implies \theta_i = (b')^{-1}(\mu_i)$;

μ_i is a function of η_i: $g(\mu_i) = \eta_i \implies \mu_i = g^{-1}(\eta_i)$; and

η_i is a function of β: $\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$.

So each θ_i is a function of the unknown $\boldsymbol{\beta}$, while ϕ, which may be known or unknown, is free of $\boldsymbol{\beta}$. Thus, we must estimate β and possibly ϕ.

1.3.1 Likelihood Equations for the Regression Coefficients

The contribution of the ith observation to the log-likelihood is

$$\ell_i(\boldsymbol{\beta},\phi) = \log f(y_i;\theta_i,\phi/a_i) = \frac{y_i\theta_i - b(\theta_i)}{\phi/a_i} + c(y_i;\phi/a_i),$$

and the full log-likelihood is $\ell = \sum_{i=1}^n \ell_i$. Note that ϕ does not depend on $\boldsymbol{\beta}$. By the chain rule,

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta_j} + \frac{\partial \ell_i}{\partial \phi}\underbrace{\frac{\partial \phi}{\partial \beta_j}}_{=0} = \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}.$$

Let $V(\mu) = b''\big((b')^{-1}(\mu)\big)$ be the variance function, so that $\mathrm{Var}(Y) = \phi\, V(\mu)$. Then

$$\begin{aligned}
&\frac{\partial \ell_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{\phi/a_i} = \frac{1}{\phi}\, a_i(y_i - \mu_i),\\
&\mu_i = b'(\theta_i) \implies \frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{\partial \mu_i/\partial \theta_i} = \frac{1}{b''(\theta_i)} = \frac{1}{b''\big((b')^{-1}(\mu_i)\big)} = \frac{1}{V(\mu_i)},\\
&g(\mu_i) = \eta_i \implies \frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{\partial \eta_i/\partial \mu_i} = \frac{1}{g'(\mu_i)},
\end{aligned}$$

and

$$\eta_i = x_i^T\boldsymbol{\beta} = \sum_{j=1}^p x_{ij}\beta_j \implies \frac{\partial \eta_i}{\partial \beta_j} = x_{ij}.$$

Thus

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{1}{\phi}\frac{a_i(y_i - \mu_i)}{V(\mu_i)g'(\mu_i)}\, x_{ij} \implies \frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial \ell_i}{\partial \beta_j} = \frac{1}{\phi}\sum_{i=1}^n \frac{a_i(y_i - \mu_i)}{V(\mu_i)g'(\mu_i)}\, x_{ij}. \tag{1.6}$$

Setting $\partial\ell/\partial\beta_j$ equal to 0, the positive factor $1/\phi$ can be cancelled, and since $\partial\ell/\partial\beta_j$ otherwise only depends on $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ (through the μ_is), the MLE of $\boldsymbol{\beta}$ can be calculated by solving the system of equations

$$\sum_{i=1}^n \frac{a_i(y_i - \mu_i)}{V(\mu_i)g'(\mu_i)}\, x_{ij} = 0, \quad j = 1, \ldots, p,$$

for $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$. These are the likelihood equations for $\boldsymbol{\beta}$.

Remark 1.5 If g is the canonical link, $g(\mu) = (b')^{-1}(\mu)$, then $\theta_i = \eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$ and $V(\mu)g'(\mu) \equiv 1$. Thus the likelihood equations simplify to $\sum_{i=1}^n a_i(y_i - \mu_i)x_{ij} = 0$ and the log-likelihood has the form

$$\begin{aligned}
\ell &= \sum_{i=1}^n \left\{ \frac{y_i\theta_i - b(\theta_i)}{\phi/a_i} + c(y_i; \phi/a_i) \right\}\\
&= \frac{1}{\phi}\sum_{i=1}^n a_i y_i \boldsymbol{x}_i^T\boldsymbol{\beta} - \frac{1}{\phi}\sum_{i=1}^n a_i b(\boldsymbol{x}_i^T\boldsymbol{\beta}) + \sum_{i=1}^n c(y_i; \phi/a_i)\\
&= \frac{1}{\phi}\boldsymbol{\beta}^T X^T A\boldsymbol{y} - \frac{1}{\phi}\sum_{i=1}^n a_i b(\boldsymbol{x}_i^T\boldsymbol{\beta}) + \sum_{i=1}^n c(y_i; \phi/a_i),
\end{aligned}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, $A = \operatorname{diag}(a_1, \ldots, a_n)$, and X is the design matrix

$$X = (x_{ij})_{\substack{1 \le i \le n \\ 1 \le j \le p}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T \in \mathbb{R}^{n \times p}$$

Thus, if ϕ is known, then by the Factorization Theorem the statistic $X^T A \boldsymbol{y} = \sum_{i=1}^n a_i y_i \boldsymbol{x}_i$ is sufficient for $\boldsymbol{\beta}$. □

1.3.2 Fisher Information

Assuming that ϕ is known, the *observed Fisher information* is $\mathscr{J}(\hat{\boldsymbol{\beta}})$, where $\hat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$ and $\mathscr{J}(\boldsymbol{\beta})$ is the negative-Hessian of the log-likelihood, i.e.,

$$\mathscr{J}(\boldsymbol{\beta}) = -\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = \left(-\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} \right)_{1 \le i,j \le p}. \tag{1.7}$$

To calculate $\mathscr{J}(\boldsymbol{\beta})$, note that

$$\frac{\partial^2 \ell_i}{\partial \beta_j \partial \beta_k} = \frac{\partial}{\partial \beta_k} \left(\frac{1}{\phi} \frac{a_i (y_i - \mu_i)}{V(\mu_i) g'(\mu_i)} x_{ij} \right) = \frac{\partial}{\partial \mu_i} \left(\frac{1}{\phi} \frac{a_i (y_i - \mu_i)}{V(\mu_i) g'(\mu_i)} x_{ij} \right) \cdot \frac{\partial \mu_i}{\partial \beta_k},$$

$$\frac{\partial}{\partial \mu_i} \left(\frac{1}{\phi} \frac{a_i (y_i - \mu_i)}{V(\mu_i) g'(\mu_i)} x_{ij} \right) = \frac{-1}{\phi} \frac{a_i}{V(\mu_i) g'(\mu_i)} x_{ij} + \frac{1}{\phi} a_i (y_i - \mu_i) x_{ij} \frac{\partial}{\partial \mu_i} \left(\frac{1}{V(\mu_i) g'(\mu_i)} \right),$$

and

$$\frac{\partial \mu_i}{\partial \beta_k} = \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_k} = \frac{1}{\partial \eta_i / \partial \mu_i} \frac{\partial \eta_i}{\partial \beta_k} = \frac{1}{g'(\mu_i)} x_{ik},$$

so that

$$\begin{aligned} \mathscr{J}(\boldsymbol{\beta})_{j,k} &= -\sum_{i=1}^n \frac{\partial^2 \ell_i}{\partial \beta_j \partial \beta_k} \\ &= \frac{1}{\phi} \sum_{i=1}^n \left\{ \frac{a_i}{V(\mu_i) g'(\mu_i)^2} - \frac{a_i (y_i - \mu_i)}{g'(\mu_i)} \frac{\partial}{\partial \mu_i} \left(\frac{1}{V(\mu_i) g'(\mu_i)} \right) \right\} x_{ij} x_{ik}. \end{aligned} \tag{1.8}$$

Note that the terms on the righthand side of this equation depend on $\boldsymbol{\beta}$ through μ_i.

The *(expected) Fisher information* is $\mathscr{I}(\boldsymbol{\beta}) = E_{\boldsymbol{\beta}}[\mathscr{J}(\boldsymbol{\beta})]$. Since only the y_i in (1.8) are random, and $E[(Y_i - \mu_i)] = 0$,

$$\mathscr{I}(\boldsymbol{\beta})_{jk} = \frac{1}{\phi} \sum_{i=1}^n \frac{a_i}{V(\mu_i) g'(\mu_i)^2} x_{ij} x_{ik}. \tag{1.9}$$

Letting

$$W = W(\boldsymbol{\beta}) = \operatorname{diag}\left\{ \frac{a_i}{V(\mu_i) g'(\mu_i)^2} : i = 1, \ldots, n \right\}, \tag{1.10}$$

we have the matrix expression

$$\mathscr{I}(\boldsymbol{\beta}) = \frac{1}{\phi} X^T W X. \tag{1.11}$$

Remark 1.6 If g is the canonical link for f, then $V(\mu)g'(\mu) \equiv 1$, so that the partial derivative on the righthand side of (1.8) is zero and (1.8) reduces to (1.9). Thus, the observed and expected Fisher information matrices are equal when the canonical link is used, and in this case

$$\mathscr{J}(\boldsymbol{\beta})_{jk} = \mathscr{I}(\boldsymbol{\beta})_{jk} = \frac{1}{\phi}\sum_{i=1}^{n} a_i V(\mu_i) x_{ij} x_{ik}. \qquad \square$$

For unknown ϕ, we calculate

$$\begin{aligned}
\frac{\partial \ell}{\partial \phi} &= \frac{-1}{\phi^2}\sum_{i=1}^{n} a_i[y_i\theta_i - b(\theta_i)] + \sum_{i=1}^{n} \frac{\partial}{\partial \phi} c(y_i; \phi/a_i), \\
\frac{\partial^2 \ell}{\partial \phi^2} &= \frac{2}{\phi^3}\sum_{i=1}^{n} a_i[y_i\theta_i - b(\theta_i)] + \sum_{i=1}^{n} \frac{\partial^2}{\partial \phi^2} c(y_i; \phi/a_i),
\end{aligned} \tag{1.12}$$

and

$$\frac{\partial^2 \ell}{\partial \phi \partial \beta_j} = \frac{-1}{\phi^2}\sum_{i=1}^{n} \frac{a_i(y_i - \mu_i)}{V(\mu_i)g'(\mu_i)} x_{ij}, \quad j = 1, \ldots, p. \tag{1.13}$$

From (1.13) is it obvious that

$$E\left(-\frac{\partial^2 \ell}{\partial \phi \partial \beta_j}\right) = 0,$$

so that the full Fisher information matrix has the block-diagonal form

$$\mathscr{I}(\boldsymbol{\beta}, \phi) = \begin{pmatrix} \mathscr{I}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \phi) & \mathbf{0} \\ \mathbf{0} & \mathscr{I}_{\phi}(\boldsymbol{\beta}, \phi) \end{pmatrix} \tag{1.14}$$

where $\mathscr{I}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \phi)$ is given by (1.11) and $\mathscr{I}_{\phi}(\boldsymbol{\beta}, \phi) = E\big(-\partial^2\ell/\partial\phi^2\big)$.

Note also that if $\hat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$ and ϕ is any value of ϕ, then

$$-\frac{\partial^2 \ell}{\partial \phi \partial \beta_j}\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = \frac{1}{\phi}\underbrace{\left(\frac{\partial \ell}{\partial \beta_j}\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}\right)}_{=\mathbf{0}} = \mathbf{0},$$

so that

$$\mathscr{J}(\hat{\boldsymbol{\beta}}, \phi) = \begin{pmatrix} \mathscr{J}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}, \phi) & \mathbf{0} \\ \mathbf{0} & \mathscr{J}_{\phi}(\hat{\boldsymbol{\beta}}, \phi) \end{pmatrix},$$

where $\mathscr{J}_{\phi}(\boldsymbol{\beta}, \phi) = -\partial^2\ell/\partial\phi^2$. Thus the observed Fisher information (evaluated at the MLE of $\boldsymbol{\beta}$) is also block diagonal.

Remark 1.7 Parameters $\boldsymbol{\beta}$ and ϕ are said to be *orthogonal* when the (expected) Fisher information has this block diagonal form. Large sample theory implies that the MLE of $(\boldsymbol{\beta},\ \phi)$ satisfies

$$\begin{pmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\phi}\end{pmatrix} \sim \mathrm{AN}\left(\begin{pmatrix}\boldsymbol{\beta}\\ \phi\end{pmatrix}, \mathscr{I}(\boldsymbol{\beta},\phi)^{-1}\right) \quad \text{as } n \to \infty,$$

where AN denotes "asymptotically normal". Since (1.14) implies that

$$\mathscr{I}(\boldsymbol{\beta},\phi)^{-1} = \begin{pmatrix}\mathscr{I}_{\boldsymbol{\beta}}(\boldsymbol{\beta},\phi)^{-1} & \mathbf{0}\\ \mathbf{0} & \mathscr{I}_{\phi}(\boldsymbol{\beta},\phi)^{-1}\end{pmatrix},$$

it follows in particular that $\hat{\boldsymbol{\beta}}$ and $\hat{\phi}$ are asymptotically independent. By (1.11), $\mathscr{I}_{\boldsymbol{\beta}}(\boldsymbol{\beta},\phi) = X^TWX/\phi$, so

$$\hat{\boldsymbol{\beta}} \sim \mathrm{AN}\big(\boldsymbol{\beta}, \phi(X^TWX)^{-1}\big). \qquad \square$$

1.4 Computation of Estimators

1.4.1 Newton's Method

Recall that to solve $h(x) = 0$ for univariate x, the Newton-Raphson algorithm iteratively computes

$$x^{(k+1)} = x^{(k)} - \frac{1}{h'(x^{(k)})}\, h(x^{(k)}),$$

beginning from a starting point $x^{(0)}$ and iterating until "convergence". The algorithm is derived by replacing $h(x)$ by its first order Taylor series expansion about $x^{(k)}$ (i.e., the tangent line to h at $(x^{(k)}, h(x^{(k)})))$ in the equation $h(x) = 0$ and solving for x (see Figure 1.1):

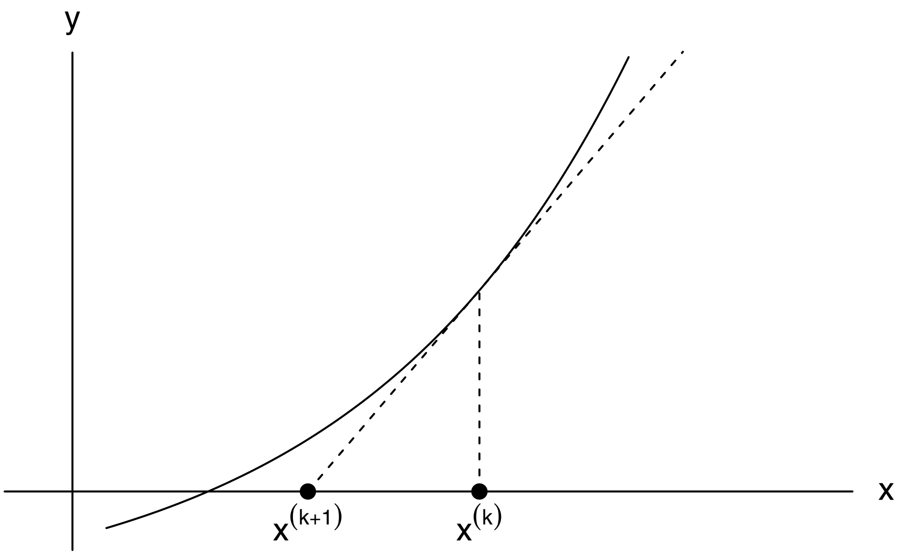
$$h(x) \approx h(x^{(k)}) + h'(x^{(k)})(x - x^{(k)}) = 0 \implies x = x^{(k)} - \frac{1}{h'(x^{(k)})}\, h(x^{(k)}).$$

To find a (local) maximum or minimum of a function $g(x)$, *Newton's method* applies the Newton-Raphson algorithm to find a zero of the derivative $g'(x)$.

For maximum likelihood estimation of a parameter vector $\boldsymbol{\theta}$ (no relation to the canonical parameter of a GLM), it is usually most convenient to work with the log-likelihood. Maximizing the log-likelihood by Newton's method is equivalent to solving the likelihood equations $\partial\ell/\partial\boldsymbol{\theta} = \mathbf{0}$ using the multivariate form of the Newton-Raphson algorithm, and leads to the following update formula:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \quad \frac{\partial^2\ell}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}}\Bigg)^{-1} \quad \frac{\partial\ell}{\partial\boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}}\Bigg) = \boldsymbol{\theta}^{(k)} + \mathscr{J}\big(\boldsymbol{\theta}^{(k)}\big)^{-1} S\big(\boldsymbol{\theta}^{(k)}\big)$$

where $\mathscr{J}(\boldsymbol{\theta})$ is the observed Fisher information evaluated at $\boldsymbol{\theta}$ and $S(\boldsymbol{\theta}) = \partial\ell/\partial\boldsymbol{\theta}$ is the *score function.*



@

Figure 1.1: Illustration of the Newton-Raphson algorithm in the univariate case.

In a GLM, to compute $\hat{\boldsymbol{\beta}}$ via Newton's method (recall that ϕ does not enter into this calculation), we iterate via

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \mathscr{J}\big(\boldsymbol{\beta}^{(k)}\big)^{-1} S(\boldsymbol{\beta}^{(k)}),$$

where $\mathscr{J}(\boldsymbol{\beta}) = -\partial^2 \ell / \partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T$ is given by (1.11) and $S(\boldsymbol{\beta}) = \partial\ell/\partial\boldsymbol{\beta}$. Note that, although both $\mathscr{J}\big(\boldsymbol{\beta}^{(k)}\big)^{-1}$ and $S(\boldsymbol{\beta}^{(k)})$ depend on ϕ, it is canceled in their product, so that the updating formula above does not depend on ϕ.[4]

1.4.2 Iteratively Reweighted Least Squares

Fisher scoring is a “quasi-Newton” algorithm that replaces the observed Fisher information, $\mathscr{J}(\boldsymbol{\theta}^{(k)})$, by the expected Fisher information, $\mathscr{I}(\boldsymbol{\theta}^{(k)})$. The expected Fisher information can be easier/faster to compute than the observed information matrix, and the Fisher scoring algorithm is generally more numerically stable than the basic Newton-Raphson algorithm described above (see Osborne, 1992, for further details and refinements of these algorithms).

To compute $\hat{\boldsymbol{\beta}}$ by Fisher scoring, we iterate via

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \mathscr{I}\big(\boldsymbol{\beta}^{(k)}\big)^{-1} S\big(\boldsymbol{\beta}^{(k)}\big).$$

[4] For convenience, we have written these expressions as if ϕ were known, but computation of $\hat{\boldsymbol{\beta}}$ is the same whether ϕ is known or unknown.

From (1.11) and (1.6) we see that

$$\begin{aligned}\mathscr{I}(\boldsymbol{\beta})^{-1}S(\boldsymbol{\beta}) &= \left(\cancel{\frac{1}{\phi}}X^TWX\right)^{-1}\cancel{\frac{1}{\phi}}\sum_{i=1}^n \frac{a_i(y_i-\mu_i)}{V(\mu_i)g'(\mu_i)}\,\boldsymbol{x}_i\\ &= \left(X^TWX\right)^{-1}\sum_{i=1}^n \boldsymbol{x}_i\frac{a_i}{V(\mu_i)g'(\mu_i)^2}g'(\mu_i)(y_i-\mu_i)\\ &= \left(X^TWX\right)^{-1}X^TW\boldsymbol{u},\end{aligned}$$

where W is given by (1.10) and $\boldsymbol{u}=(u_1,\ldots,u_n)^T$ with $u_i=g'(\mu_i)(y_i-\mu_i)$. (Note that W and $\boldsymbol{u}$ depend on $\boldsymbol{\beta}$ through $\boldsymbol{\mu}=(\mu_1,\ldots,\mu_n)^T$.) Thus the Fisher scoring algorithm for $\hat{\boldsymbol{\beta}}$ has the form

$$\boldsymbol{\beta}^{(k+1)}=\boldsymbol{\beta}^{(k)}+\left(X^TW^{(k)}X\right)^{-1}X^TW^{(k)}\boldsymbol{u}^{(k)}.$$

Left-multiplying both sides by $X^TW^{(k)}X$ we obtain

$$X^TW^{(k)}X\boldsymbol{\beta}^{(k+1)}=X^TW^{(k)}\underbrace{X\boldsymbol{\beta}^{(k)}}_{=\boldsymbol{\eta}^{(k)}}+X^TW^{(k)}\boldsymbol{u}^{(k)}=X^TW^{(k)}(\boldsymbol{\eta}^{(k)}+\boldsymbol{u}^{(k)}).$$

Letting

$$z_i^{(k)}=\eta_i^{(k)}+u_i^{(k)}=g\big(\mu_i^{(k)}\big)+g'\big(\mu_i^{(k)}\big)\left(y_i-\mu_i^{(k)}\right) \tag{1.15}$$

and $\boldsymbol{z}^{(k)}=(z_1^{(k)},\ldots,z_n^{(k)})^T$, we have

$$X^TW^{(k)}X\boldsymbol{\beta}^{(k+1)}=X^TW^{(k)}\boldsymbol{z}^{(k)}, \tag{1.16}$$

or equivalently

$$\boldsymbol{\beta}^{(k+1)}=(X^TW^{(k)}X)^{-1}X^TW^{(k)}\boldsymbol{z}^{(k)}. \tag{1.17}$$

From either (1.16) or (1.17) we see that $\boldsymbol{\beta}^{(k+1)}$ is obtained by least squares regression of $z^{(k)}$ on X with diagonal weight matrix $W^{(k)}$. For this reason, this algorithm for computing $\hat{\boldsymbol{\beta}}$ is called *iteratively reweighted least squares (IRLS)* (or *iteratively weighted least squares* (IWLS)). The vector $\boldsymbol{z}^{(k)}$, which plays the role of the response in the update from $\boldsymbol{\beta}^{(k)}$ to $\boldsymbol{\beta}^{(k+1)}$, is called the *working response vector*, and the diagonal elements of $W^{(k)}$ are called the *iterative weights.*

IRLS Algorithm

1. Initialize $\boldsymbol{\mu}^{(0)}$, $\boldsymbol{z}^{(0)}$, and $W^{(0)}$.
2. Update $\boldsymbol{\beta}^{(k)}$ using $X^TW^{(k)}X\boldsymbol{\beta}^{(k+1)}=X^TW^{(k)}\boldsymbol{z}^{(k)}$.
3. Update $\boldsymbol{\mu}^{(k)}$, $\boldsymbol{z}^{(k)}$ and $W^{(k)}$ using $\boldsymbol{\beta}^{(k+1)}$.
4. Repeat steps 2 and 3 until convergence, determined by monitoring either the change in the log-likelihood, the value of $\boldsymbol{\beta}^{(k)}$, or both. In case these values to not converge, stop after some maximum number of iterates and return an error.

Because $E(Y_i) = \mu_i$, a simple way to start the IRLS algorithm is to set with $\mu_i^{(0)} = y_i$ so that

$$z_i^{(0)} = g(y_i) + g'(y_i)\big(y_i - y_i\big) = g(y_i)$$

and

$$W^{(0)} = \operatorname{diag}\left\{\frac{a_i}{V(y_i)g'(y_i)^2} : i = 1, \ldots, n\right\}.$$

However, in some cases $g(y_i)$ is undefined and some modification of these starting values is required. For example, in logistic regression, $g(y_i) = \operatorname{logit}(y_i) = \log\big(y_i/(1-y_i)\big)$ is undefined at $y_i = 0$ and $y_i = 1$, and in this case we might define

$$\mu_i^{(0)} = \begin{cases} 1-\delta, & \text{if } y_i = 1, \\ \delta, & \text{if } y_i = 0, \end{cases}$$

for some small $\delta > 0$.

1.4.3 Estimation of the Dispersion Parameter

If ϕ is unknown, then once $\hat{\boldsymbol{\beta}}$ is computed, the MLE of ϕ can be computed via Newton's algorithm:

$$\phi^{(k+1)} = \phi^{(k)} - \left.\frac{\partial^2 \ell}{\partial \phi^2}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, \phi=\phi^{(k)}}\Bigg)^{-1} \left.\frac{\partial \ell}{\partial \phi}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, \phi=\phi^{(k)}}\Bigg) = \phi^{(k)} + \frac{S_\phi\big(\hat{\boldsymbol{\beta}}, \phi^{(k)}\big)}{\mathscr{I}_\phi\big(\hat{\boldsymbol{\beta}}, \phi^{(k)}\big)}.$$

However, ϕ is usually estimated by a "moment estimator" as follows. Recall that $\operatorname{Var}(Y_i) = \phi_i V(\mu_i)$, i.e.,

$$E\big[(Y_i - \mu_i)^2\big] = \frac{\phi}{a_i} V(\mu_i) \implies \phi = E\left[\frac{a_i(Y_i - \mu_i)^2}{V(\mu_i)}\right].$$

Thus, if $\boldsymbol{\beta}$ and hence $\mu_1, \ldots, \mu_n$ were known, then

$$\frac{1}{n}\sum_{i=1}^{n} \frac{a_i(y_i - \mu_i)^2}{V(\mu_i)}$$

would be an unbiased estimator of ϕ. Since β is unknown, we use the MLE $\hat{\beta}$ to calculate $\hat{\mu}_1, \ldots, \hat{\mu}_n$ and we take

$$\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n} \frac{a_i(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}.$$

In the special case of a normal model with identity link, this reduces to the usual unbiased estimate of $\phi = \sigma^2$.

1.5 Deviance

The deviance is a measure of fit for a GLM that is somewhat analogous the the residual sum of squares in ordinary linear regression. Let

$$\ell(\boldsymbol{\mu}, \phi; \boldsymbol{y}) = \sum_{i=1}^{n} \log f(y_i; \theta_i, \phi/a_i) = \frac{1}{\phi} \sum_{i=1}^{n} a_i[y_i\theta_i - b(\theta_i)] + \sum_{i=1}^{n} c(y_i; \phi/a_i),$$

where $\theta_i = (b')^{-1}(\mu_i)$. For the GLM with

$$\eta_i = x_i^T \beta \qquad \text{and} \qquad g(\mu_i) = \eta_i, \qquad i = 1, \ldots, n,$$

let $\hat{\beta}$ be the MLE and let

$$\hat{\eta}_i = x_i^T \hat{\beta}, \qquad \hat{\mu}_i = g^{-1}(\hat{\eta}_i), \qquad \hat{\theta}_i = (b')^{-1}(\hat{\mu}_i), \qquad \text{and} \qquad \tilde{\theta}_i = (b')^{-1}(y_i).$$

Definition 1.1 With the notation above, the *deviance* for the fitted GLM model is

$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = 2[\ell(\boldsymbol{y}, \phi; \boldsymbol{y}) - \ell(\hat{\boldsymbol{\mu}}, \phi; \boldsymbol{y})] \cdot \phi = \sum_{i=1}^{n} 2a_i \Big\{ y_i(\tilde{\theta}_i - \hat{\theta}_i) - \big[b(\tilde{\theta}_i) - b(\hat{\theta}_i)\big] \Big\}, \tag{1.18}$$

and the *scaled deviance* is

$$D^*(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = \frac{1}{\phi} D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}). \tag{1.19}$$

□

Remark 1.8

- $D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})$ does not depend on ϕ.
- For binomial and Poisson models, $\phi = 1$, so $D^* = D$.
- $\ell(\boldsymbol{y}, \phi; \boldsymbol{y})$ is the log-likelihood for the *saturated model* with as many mean parameters as observations.
- In some cases, $\tilde{\theta}_i = (b')^{-1}(y_i)$ may not be defined, but we will procede formally as if it were. When deriving the form of the deviance for a particular model, these gaps can be filled in by determining the actual value $f(y_i; y_i, \phi/a_i)/f(y_i; \hat{\mu}_i, \phi/a_i)$ and taking its logarithm. □

Another measure of fit is *Pearson's* χ^2 *statistic*,

$$X^2(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} \frac{a_i(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}, \tag{1.20}$$

and the *scaled Pearson* χ^2 *statistic* is

$$X^{*2}(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = \frac{1}{\phi} X^2(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} \frac{a_i(y_i - \hat{\mu}_i)^2}{\phi V(\hat{\mu}_i)}. \tag{1.21}$$

Recall that for a $N(\mu, \sigma^2)$ distribution, we have $\theta = \mu$, $\phi = \sigma^2$, $b(\theta) = \frac{1}{2}\theta^2$, and $V(\mu) = 1$, so that for a normal GLM,

$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} 2a_i \Big\{ y_i(y_i - \hat{\mu}_i) - \tfrac{1}{2}\big(y_i^2 - \hat{\mu}_i^2\big) \Big\} = \sum_{i=1}^{n} a_i (y_i - \hat{\mu}_i)^2 = X^2(\boldsymbol{y}; \hat{\boldsymbol{\mu}})$$

and of course

$$D^*(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = X^{*2}(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} \frac{a_i (y_i - \hat{\mu}_i)^2}{\sigma^2}$$

as well.

It was shown in a homework problem that for small ϕ, a distribution of the exponential dispersion form is approximately normal. Thus, when ϕ is small, or more commonly when $a_1, \ldots, a_n$ are large, we expect that $D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) \approx X^2(\boldsymbol{y}; \hat{\boldsymbol{\mu}})$.

Example 1.8 Suppose $Y \sim \frac{\text{Bin}(m,\pi)}{m}$. Then $\theta = \log\big(\mu/(1-\mu)\big)$, $b(\theta) = \log(1 + e^\theta) = -\log(1-\mu)$, $V(\mu) = \mu(1-\mu)$, $\phi = 1$, and $a = m$, so for independent binomial proportions,[5]

$$\begin{aligned} D^*(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) &= D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) \\ &= 2\sum_{i=1}^{n} m_i \left\{ y_i \left[\log\left(\frac{y_i}{1-y_i}\right) - \log\left(\frac{\hat{\mu}_i}{1-\hat{\mu}_i}\right) \right] - \Big[-\log(1-y_i) + \log(1-\hat{\mu}_i) \Big] \right\} \\ &= 2\sum_{i=1}^{n} m_i \left\{ \underbrace{y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right)}_{:=0 \text{ if } y_i = 0} + \underbrace{(1-y_i)\log\left(\frac{1-y_i}{1-\hat{\mu}_i}\right)}_{:=0 \text{ if } y_i = 1} \right\}, \end{aligned}$$

while

$$X^{*2}(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = X^2(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} \frac{m_i (y_i - \hat{\mu}_i)^2}{\hat{\mu}_i (1 - \hat{\mu}_i)}.$$

As a point of interest, note that following the standard mnemonic "observed minus expected squared over expected" for the $2n$ cells of success and failure counts yields

$$\sum_{i=1}^{n} \left\{ \frac{(m_i y_i - m_i \hat{\mu}_i)^2}{m_i \hat{\mu}_i} + \frac{[(m_i - m_i y_i) - m_i(1-\hat{\mu}_i)]^2}{m_i(1-\hat{\mu}_i)} \right\} = \sum_{i=1}^{n} m_i (y_i - \hat{\mu}_i)^2 \left\{ \frac{1}{\hat{\mu}_i} + \frac{1}{1-\hat{\mu}_i} \right\} = X^2(\boldsymbol{y}; \hat{\boldsymbol{\mu}}). \quad \square$$

Example 1.9 For the Poisson distribution with mean μ, recall that $\theta = \log \mu$, $b(\theta) = e^\theta = \mu$, $V(\mu) = \mu$, and $\phi = 1$, so that[6]

$$\begin{aligned} D^*(\boldsymbol{y}; \boldsymbol{\mu}) = D(\boldsymbol{y}; \boldsymbol{\mu}) &= 2\sum_{i=1}^{n} \big\{ y_i[\log y_i - \log \hat{\mu}_i] - (y_i - \hat{\mu}_i) \\ &= 2\sum_{i=1}^{n} \big\{ \underbrace{y_i \log(y_i/\hat{\mu}_i)}_{:=0 \text{ if } y_i = 0} - (y_i - \mu_i) \end{aligned}$$

[5] Check that taking $0 \log 0 = 0$ yields the correct result for $y_i = 0$ and $y_i = 1$.

[6] Again, check that taking $0 \log 0 = 0$ yields the correct result for $y_i = 0$.

and

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

Note that $X^2(\boldsymbol{y};\hat{\boldsymbol{\mu}})$ again has the "observed minus expected squared over expected" form. □

1.5.1 Analysis of Deviance

Suppose that $M_0 \subset M_1$ are nested GLMs, and $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}_1$ are the corresponding MLEs with $\hat{\boldsymbol{\mu}}_0$ and $\hat{\boldsymbol{\mu}}_1$ defined in the obvious way. Assuming that ϕ is known, the likelihood ratio test statistic for testing H_0: M_0 is the true model versus H_a: M_1 is the true model is then

$$T_{\mathrm{LR}} = -2\big[\ell(\hat{\boldsymbol{\mu}}_0,\phi;\boldsymbol{y}) - \ell(\hat{\boldsymbol{\mu}}_1,\phi;\boldsymbol{y})\big] = \frac{D(\boldsymbol{y};\hat{\boldsymbol{\mu}}_0) - D(\boldsymbol{y};\hat{\boldsymbol{\mu}}_1)}{\phi} = D^*(\boldsymbol{y};\hat{\boldsymbol{\mu}}_0) - D^*(\boldsymbol{y};\hat{\boldsymbol{\mu}}_1),$$

and

$$\frac{D(\boldsymbol{y};\hat{\boldsymbol{\mu}}_0) - D(\boldsymbol{y};\hat{\boldsymbol{\mu}}_1)}{\phi} \xrightarrow[\mathrm{H}_0]{\mathrm{d}} \chi^2_r \quad \text{as } n \to \infty,$$

where r is the difference in the number of free parameters between models M_0 and M_1, or equivalently, the number of independent constraints imposed by H_0 on the coefficients of M_1. Thus we reject model M_0 in favor of M_1 if

$$\frac{D(\boldsymbol{y};\hat{\boldsymbol{\mu}}_0) - D(\boldsymbol{y};\hat{\boldsymbol{\mu}}_1)}{\phi} > \chi^2_{r,\alpha}.$$

Comparison of nested GLMs by likelihood ratio tests is often called the *analysis of deviance.*

Remark 1.9 Note that although the distribution of the difference in scaled deviances for two nested models is often well approximated by a chisquare distribution, the same cannot necessarily be said for the individual scaled deviances themselves. As you might expect from our earlier discussion, this typically requires small dispersion, i.e., large values of $a_1, \ldots, a_n$, as in a binomial model in which all the sample sizes are large. □

In models in which ϕ is unknown, it is usually estimated by the moment estimator

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{a_i(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} = \frac{1}{n-p} X^2(\boldsymbol{y};\hat{\boldsymbol{\mu}}),$$

where $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}_1$ is calculated for the larger of the two models and p is the number of free regression parameters in the larger model. Note that as long as $\hat{\phi}$ is consistent for ϕ under the null model, Slutsky's theorem implies that

$$\frac{D(\boldsymbol{y};\hat{\boldsymbol{\mu}}_0) - D(\boldsymbol{y};\hat{\boldsymbol{\mu}}_1)}{\hat{\phi}} = \underbrace{\frac{\phi}{\hat{\phi}}}_{\xrightarrow[\mathrm{H}_0]{\mathrm{Pr}} 1} \cdot \underbrace{\frac{D(\boldsymbol{y};\hat{\boldsymbol{\mu}}_0) - D(\boldsymbol{y};\hat{\boldsymbol{\mu}}_1)}{\phi}}_{\xrightarrow[\mathrm{H}_0]{\mathrm{d}} \chi^2_r} \xrightarrow[\mathrm{H}_0]{\mathrm{d}} \chi^2_r.$$

Nevertheless, in cases in which ϕ is estimated as above, the F-distribution generally provides a better approximation to the null distribution of $[D(\boldsymbol{y};\hat{\boldsymbol{\mu}}_0)-D(\boldsymbol{y};\hat{\boldsymbol{\mu}}_1)]/\hat{\phi}$, and so it is common to refer the test statistic

$$\frac{\big[D(\boldsymbol{y};\hat{\boldsymbol{\mu}}_0)-D(\boldsymbol{y};\hat{\boldsymbol{\mu}}_1)\big]/r}{\hat{\phi}} \quad \text{to the } F_{r,n-p} \text{ distribution, where } r \text{ and } p \text{ are as above.}$$

Of course for Poisson and binomial models $\phi = 1$ is known and the χ^2_r reference distribution is used.

1.6 Residuals for Generalized Linear Models

1.6.1 Review of Residuals for Ordinary Linear Models

Consider the linear model

$$\boldsymbol{z} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim (\boldsymbol{0}, \sigma^2 I).$$

The ith raw residual is $e_i = z_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ and a *naive* (or *unadjusted*) residual is

$$r_i = \frac{e_i}{\hat{\sigma}} = \frac{z_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}}{\hat{\sigma}},$$

where $\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\boldsymbol{z}$ and $\hat{\sigma}^2 = (n-p)^{-1}(\boldsymbol{z}-X\hat{\boldsymbol{\beta}})^T(\boldsymbol{z}-X\hat{\boldsymbol{\beta}}) = (n-p)^{-1}\sum_{i=1}^n(z_i-\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})^2$ is the usual MSE ("mean square due to error").

We call r_i a naive residual because e_i is divided by an estimate of the standard deviation of ϵ_i rather than of e_i. Note that

$$\mathrm{Var}(\boldsymbol{z} - X\hat{\boldsymbol{\beta}}) = \mathrm{Var}\big(\underbrace{[I - X(X^TX)^{-1}X^T]}_{\text{symmetric, idempotent}}\boldsymbol{z}\big) = \sigma^2[I - \underbrace{X(X^TX)^{-1}X^T}_{=:H}]$$

Letting $H = (h_{ij})_{i,j=1}^n = X(X^TX)^{-1}X^T$, the ith *standardized residual* is given by

$$r_i' = \frac{z_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}}{\sqrt{\hat{\sigma}^2(1-h_{ii})}} = \frac{r_i}{\sqrt{1-h_{ii}}}.$$

Recall that h_{ii} is called the *leverage* of the ith observation.

Remark 1.10 Unfortunately, the nomenclature for residuals is not at all "standardized." Many authors refer to the standardized residual above as a "Studentized" residual, since it involves division by an estimate of the standard deviation of the numerator, rather than the actual standard deviation. Others, however, reserve the term "Studentized residual" to refer to the modification of the above "standardized" residual in which the $\hat{\sigma}^2$ is replaced by $\hat{\sigma}^2_{(i)}$, the estimate of σ^2 computed by leaving out the ith observation. This terminology also appears to have been adopted in R, which is one reason that it is used here as well. It is also common to refer to the "standardized" residuals above as "internally Studentized", and to the Studentized residuals as "externally Studentized", but there are many other terminologies in use.

In a normal linear model, the ith (externally) Studentized residual can be viewed as the likelihood ratio test statistic for a mean-shift-outlier alternative to the linear model, in which the mean for the ith observation is a free parameter. From this it is easy to see that the ith Studentized residual follows Student's t distribution on $n-p-1$ degrees of freedom, and this is the best reason I know of for referring to it as the Studentized residual. Note that the standardized residual does not follow a t distribution.

The primary implication of the lack of standardization of this terminology is that you must check each individual author/software-package's definitions to know what sort of residuals you are looking at. □

In weighted least squares regression, we assume that

$$\boldsymbol{z} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim (\boldsymbol{0}, \sigma^2 A^{-1}),$$

where $A = \operatorname{diag}(a_1, \ldots, a_n)$ is a diagonal matrix of known positive weights $a_1, \ldots, a_n$. In this case, the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ is the weighted least squares estimator

$$\hat{\boldsymbol{\beta}} = (X^T A X)^{-1} X^T A \boldsymbol{z}, \tag{1.22}$$

and it is easy to show that

$$\operatorname{Var}\big(A^{1/2}(\boldsymbol{z} - X\hat{\boldsymbol{\beta}})\big) = \sigma^2[I - A^{1/2}X(X^TAX)^{-1}X^TA^{1/2}],$$

or equivalently, that

$$\operatorname{Var}(\boldsymbol{z} - X\hat{\boldsymbol{\beta}}) = \sigma^2 A^{-1/2}[I - A^{1/2}X(X^TAX)^{-1}X^TA^{1/2}]A^{-1/2}.$$

Thus, in this case the ith naive residual is

$$r_i = \frac{z_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}}{\sqrt{\hat{\sigma}^2/a_i}}, \tag{1.23}$$

while the ith standardized residual is

$$r_i' = \frac{z_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}}{\sqrt{\frac{\hat{\sigma}^2}{a_i}(1 - h_{ii})}}, \tag{1.24}$$

where

$$H = (h_{ij})_{i,j=1}^n = A^{1/2}X(X^TAX)^{-1}X^TA^{1/2} \tag{1.25}$$

and

$$\hat{\sigma}^2 = \frac{1}{n-p}(\boldsymbol{z} - X\hat{\boldsymbol{\beta}})^T A(\boldsymbol{z} - X\hat{\boldsymbol{\beta}}) = \frac{1}{n-p}\sum_{i=1}^n a_i(z_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})^2.$$

1.6.2 Residuals for GLMs

Recall that IRLS algorithm takes

$$\boldsymbol{\beta}^{(k+1)} = (X^T W^{(k)} X)^{-1} X^T W^{(k)} \boldsymbol{z}^{(k)},$$

where

$$W = \mathrm{diag}(w_1, \ldots, w_n), \quad \text{with} \quad w_i = \frac{a_i}{V(\mu_i) g'(\mu_i)^2},$$

is the diagonal matrix of *iterative weights* and

$$z_i = g(\mu_i) + g'(\mu_i)(y_i - \mu_i) = \eta_i + g'(\mu_i)(y_i - \mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} + g'(\mu_i)(y_i - \mu_i), \quad i = 1, \ldots, n$$

are the *working responses*. Upon convergence,

$$\hat{\boldsymbol{\beta}} = (X^T \hat{W} X)^{-1} X^T \hat{W} \hat{\boldsymbol{z}}.$$

Notice that this has the same form as the weighted least square estimator of (1.22), except that the weight matrix $\hat{W}$ and the response vector $\hat{\boldsymbol{z}}$ depend on $\hat{\boldsymbol{\beta}}$.

Thus, by analogy with weighted least squares (see (1.23), (1.24), and (1.25)), the ith *(scaled) Pearson residual* is given by[7]

$$r_{P,i} = \frac{\hat{z}_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}}{\sqrt{\phi / \hat{w}_i}} = \frac{g'(\hat{\mu}_i)(y_i - \hat{\mu}_i)}{\sqrt{\frac{\phi}{a_i} V(\hat{\mu}_i) g'(\hat{\mu}_i)^2}} = \frac{y_i - \hat{\mu}_i}{\sqrt{\frac{\phi}{a_i} V(\hat{\mu}_i)}},$$

and the ith *standardized Pearson residual* is

$$r'_{P,i} = \frac{r_{P,i}}{\sqrt{1 - \hat{h}_{ii}}},$$

where

$$\hat{H} = (\hat{h}_{ij})_{i,j=1}^n = \hat{W}^{1/2} X (X^T \hat{W} X)^{-1} X^T \hat{W}^{1/2}. \tag{1.26}$$

Note that

$$\sum_{i=1}^n r_{P,i}^2 = \frac{1}{\phi} \sum_{i=1}^n \frac{a_i (y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} = \frac{1}{\phi} X^2(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = X^{*2}(\boldsymbol{y}, \hat{\boldsymbol{\mu}}).$$

Let $d_i = 2\phi[\log f(y_i; y_i, \phi/a_i) - \log f(y_i; \hat{\mu}_i, \phi/a_i)]$ be the contribution of the ith observation to the deviance. Then the ith *scaled deviance residual* is defined to be

$$r_{D,i} = \mathrm{sign}(y_i - \hat{\mu}_i) \sqrt{d_i / \phi} = \mathrm{sign}(y_i - \hat{\mu}_i) \sqrt{2 \left[\log f(y_i; y_i, \phi) - \log f(y_i; \hat{\mu}_i, \phi)\right]},$$

and the ith *standardized deviance residual* is

$$r'_{D,i} = \frac{r_{D,i}}{\sqrt{1 - \hat{h}_{ii}}},$$

where $\hat{h}_{ii}$ is given by (1.26).

[7] Here we assume that link function g is increasing. Note also that the analogy with weighted least-squares is useful not so much for deriving the form of the (scaled) Pearson residual $r_{P,i}$ as for deriving the form of $\hat{h}_{ii}$ used in the standardized Pearson residual $r'_{P,i}$.

Remark 1.11

- For small values of ϕ/a_i, we expect

$$r_{D,i} \approx \frac{y_i - \hat{\mu}_i}{\sqrt{\frac{\phi}{a_i} V(\hat{\mu}_i)}} \dot{\sim} N(0, 1 - h_{ii}),$$

  so in this case $r'_{P,i}$ and $r'_{D,i}$ should be roughly $N(0,1)$, and absolute values larger than 2 or 3 suggest an outlying observation.

  However, if ϕ/a_i is not "small", then the normal approximation may be very poor indeed (e.g., for binary data), and such rules of thumb may not be very useful.

- Of course, as in ordinary linear regression, the residuals (standardized or not) are *dependent.*

- In standardizing these residuals, we have ignored the fact that the iterative weights $\hat{w}_1, \dots, \hat{w}_n$ are random (which means that the variance formulas from weighted least squares, where the weights are nonrandom, are not exactly correct).

- Most or all of the ingredients needed to compute these residuals are readily available from the last iteration of the IRLS algorithm.

- Several other residuals are common in the GLM world, including the *Anscombe* residuals and several proposed generalizations of the "Studentized" residual, which are mean to mimic the definition and/or distribution of the (externally) Studentized residual in a linear model. □

Example For a binomial GLM, where $Y_i \sim \text{Bin}(m_i, \mu_i)/m_i$, the ith (scaled) Pearson residual is

$$r_{P,i} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i(1 - \hat{\mu}_i)/m_i}}$$

while the ith (scaled) Deviance residual is

$$r_{D,i} = \text{sign}(y_i - \hat{\mu}_i) \sqrt{2m_i \left\{ y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (1 - y_i) \log\left(\frac{1 - y_i}{1 - \hat{\mu}_i}\right) \right\}}. \qquad \square$$

Example For a Poisson GLM, $\phi = 1$ and the (scaled) Pearson and deviance residuals are

$$r_{P,i} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

and

$$r_{D,i} = \text{sign}(y_i - \hat{\mu}_i) \sqrt{2 \left\{ y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i) \right\}}. \qquad \square$$