Handout on Exponential Families and GLMs

Richard L. Smith

September 8, 2020

Updated from version of August 30. Changes are in red.

1 Background: Two formulas for likelihood functions

We derive two formulas that are used later to calculate means and variances of exponential family densities.

Suppose $f(y;\theta)$ is the density of a random variable Y depending on (scalar) parameter θ. Let $\ell(\theta;Y)$ be the log likelihood function based on a single observation Y. Assume ℓ is at least twice differentiable with respect to θ, with first two derivatives $\ell' = \frac{d\ell}{d\theta}$ and $\ell'' = \frac{d^2\ell}{d\theta^2}$. Then:

$$\begin{aligned} \mathrm{E}\left\{\ell'(\theta;Y)\right\} &= 0, & (1)\\ \mathrm{E}\left[\left\{\ell'(\theta;Y)\right\}^2\right] &= -\mathrm{E}\left\{\ell''(\theta;Y)\right\}. & (2) \end{aligned}$$

Proof of (1). We have

$$\begin{aligned} 1 &= \int f(y;\theta)dy,\\ 0 &= \frac{d}{d\theta}\int f(y;\theta)dy\\ &= \int \frac{d}{d\theta} f(y;\theta)dy\\ &= \int \frac{d}{d\theta}\left\{\log f(y;\theta)\right\} f(y;\theta)dy \qquad (3)\\ &= \mathrm{E}\left\{\ell'(\theta;Y)\right\}. \end{aligned}$$

Proof of (2). Continuing the same argument by differentiating (3),

$$\begin{aligned} 0 &= \frac{d}{d\theta}\left[\int \frac{d}{d\theta}\left\{\log f(Y;\theta)\right\} f(y;\theta)dy\right]\\ &= \int \frac{d^2}{d\theta^2}\left\{\log f(y;\theta)\right\} f(y;\theta)dy + \int \frac{d}{d\theta}\left\{\log f(y;\theta)\right\}\frac{d}{d\theta} f(y;\theta)dy\\ &= \int \frac{d^2}{d\theta^2}\left\{\log f(y;\theta)\right\} f(y;\theta)dy + \int \left[\frac{d}{d\theta}\left\{\log f(y;\theta)\right\}\right]^2 f(y;\theta)dy\\ &= \mathrm{E}\left\{\ell''(\theta;Y)\right\} + \mathrm{E}\left[\left\{\ell'(\theta;Y)\right\}^2\right]. \end{aligned}$$

Remarks

1. The above glosses over some technical details, in particular, justifying the interchange of the differentiation and integration operators. This can be problematic under certain circumstances, in particular, when the range of integration is itself dependent on θ. This sort of issue is not a problem in exponential families.

2. If Y is a discrete random variable (the two best-known examples are Binomial and Poisson), the same proof holds but with the integrals replaced by sums over the possible values of y. Note that we always assume $f(y;\theta)$ is continuous and at least twice differentiable in θ, but differentiability with respect to y is not required.

3. For simplicity, the derivation here assumes θ is one-dimensional but the same result holds in multidimensions. In particular, all the partial derivatives of the log likelihood function have expectation zero, while the covariance matrix of all the first-order partial derivatives is minus the expectation of the matrix of second-order derivatives. The latter quantity is known as the *Fisher Information Matrix.*

2 Exponential Families

An exponential family is defined by the formula

$$f(y;\theta,\phi) \quad = \quad \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \tag{4}$$

where:

- Y is a discrete or continuous random variable; if Y is discrete, then $f(y;\theta,\phi)$ is the probability mass function evaluated at a particular value y; if Y is continuous, $f(y;\theta,\phi)$ is the probability density function;

- θ is the main parameter of the exponential family; in all our examples, θ itself is a scalar parameter, though in GLMs it typically depends on additional parameters through the link function (defined later) and covariates;

- ϕ is an additional parameter usually known as the *dispersion parameter*; $a(\phi)$ is an arbitrary function of ϕ and $c(y,\phi)$ is also arbitrary but (the key point) it cannot depend on θ.

For an exponential family density of the form (4), we have (note that dashes still refer to differentiation with respect to θ, not ϕ),

$$\begin{aligned}
\ell(y;\theta,\phi) &= \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi), \\
\ell'(y;\theta,\phi) &= \frac{y - b'(\theta)}{a(\phi)}, \\
\ell''(y;\theta,\phi) &= \frac{b''(\theta)}{a(\phi)}.
\end{aligned}$$

Applying the formulas of Section 1, we deduce

$$\begin{aligned} \mathrm{E}\left\{\frac{Y-b'(\theta)}{a(\phi)}\right\} &= 0, \\ \mathrm{E}\left[\left\{\frac{Y-b'(\theta)}{a(\phi)}\right\}^2\right] &= -\frac{b''(\theta)}{a(\phi)}, \end{aligned}$$

and hence

$$\mathrm{E}\{Y\} = b'(\theta), \tag{5}$$

$$\mathrm{Var}\{Y\} = b''(\theta)a(\phi). \tag{6}$$

Note that we often write μ for $b'(\theta)$, the mean of the random variable Y.

3 Examples of Exponential Families

3.1 Normal

For the normal or Gaussian density, we have

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

This is of the form (4) if we define $\theta=\mu,\ \phi=\sigma^2,\ b(\theta)=\frac{\theta^2}{2},\ a(\phi)=\phi,\ c(y,\phi)=-\frac{y^2}{2\phi}-\frac{1}{2}\log(2\pi\phi)$. Here $b'(\theta)=\theta,\ b''(\theta)=1$, so the mean is $\theta=\mu$ and the variance is $\phi=\sigma^2$. This of course agrees with well-known results for the normal distribution.

3.2 Poisson

This is an example of a discrete RV with

$$\begin{aligned} f(y;\mu) &= \frac{\mu^y e^{-\mu}}{y!},\ y=0,1,2,...., \\ &= \exp\left\{y\log\mu-\mu-\log(y!)\right\}. \end{aligned}$$

In this case we identify θ with $\log\mu$, we can set $\phi\equiv 1$, and $c(y,\phi)=-\log(y!)$. So $b(\theta)=b'(\theta)=b''(\theta)=e^\theta=\mu$, so the mean and variance are both μ.

3.3 Binomial

I have thought about how to do this in a consistent way but I think it makes most sense for the GLM framework if we define y to the the *proportion* of successes (as Faraway does on p. 155, but

not when he first introduces the Binomial model as an example of an exponential family on p. 152). In this case,

$$\begin{aligned} f(y;p,n) &= \binom{n}{ny} p^{ny}(1-p)^{n-ny},\ y = 0, 1/n, 2/n, ..., 1 \\ &= \exp\left\{ ny \log\left(\frac{p}{1-p}\right) + n\log(1-p) + \log\binom{n}{ny} \right\} \end{aligned}$$

We can define $\phi = n,\ a(\phi) = \frac{1}{\phi},\ \theta = \log\left(\frac{p}{1-p}\right),\ p = \frac{e^\theta}{1+e^\theta},\ 1-p = (1+e^\theta)^{-1}$ and therefore $b(\theta) = -\log(1-p) = \log(1+e^\theta)$, $b'(\theta) = \frac{e^\theta}{1+e^\theta} = 1 - \frac{1}{1+e^\theta} = p$, $b''(\theta) = \frac{e^\theta}{(1+e^\theta)^2} = p(1-p)$ so the mean is $b'(\theta) = p$ and the variance is $b''(\theta)a(\phi) = \frac{p(1-p)}{n}$ in accordance with well-known formulas.

Alternatively, if the ith observation y_i is derived from a binomial distribution with parameters (n_i, p_i) with n_i known, we may define $\phi = 1$ and let w_i be the *weight* $1/n_i$. The variance in this case is $p_i(1-p_i)/n_i = w_i p_i(1-p_i)/\phi$. This makes everything consistent with the formulas on pages 154 and 157 of Faraway.

3.4 Gamma

The most usual way to write the Gamma density is $\frac{\beta^\alpha y^{\alpha-1} e^{-\beta y}}{\Gamma(\alpha)}$ with mean $\frac{\alpha}{\beta}$ and variance $\frac{\alpha}{\beta^2}$. Here we write $\nu = \alpha,\ \mu = \frac{\alpha}{\beta}$ so

$$\begin{aligned} f(y;\mu,\nu) &= \frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^\nu y^{\nu-1} e^{-y\nu/\mu} \\ &= \exp\left\{ -\frac{y\nu}{\mu} - \nu\log\mu + (\nu-1)\log y + \nu\log\nu - \log\Gamma(\nu) \right\} \end{aligned}$$

and in this case the mean is μ and the variance is $\frac{\mu}{\nu}$.

I think the simplest way to handle this is to define $\theta = \frac{1}{\mu},\ \phi = \frac{1}{\nu},\ a(\phi) = -\phi$ (nothing in the preceding general theory said $a(\phi)$ had to be positive). In this case we write

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

so $b(\theta) = \log\theta,\ b'(\theta) = \frac{1}{\theta} = \mu,\ b''(\theta) = -\frac{1}{\theta^2} = -\mu^2$ and the variance function is $b''(\theta)a(\phi) = \frac{\mu^2}{\nu}$ consistent with what we already knew about the gamma distribution.

3.5 Inverse Gaussian

The last of our basic catalog of exponential families is the *Inverse Gaussian*, for which

$$\begin{aligned} f(y;\mu,\lambda) &= \left(\frac{\lambda}{2\pi y^3}\right)^{1/2} \exp\left\{ -\frac{\lambda(y-\mu)^2}{2\mu^2 y} \right\} \\ &= \exp\left\{ -\frac{\lambda y}{2\mu^2} + \frac{\lambda}{\mu} - \frac{\lambda}{2y} + \frac{1}{2}\log\left(\frac{\lambda}{2\pi y^3}\right) \right\} \end{aligned} \tag{7}$$

where $y,\ \mu,\ \lambda > 0$. In this case we define $\phi = \frac{1}{\lambda},\ a(\phi) = -\phi,\ \theta = (2\mu^2)^{-1},\ b(\theta) = (2\theta)^{1/2},\ c(y,\phi) = -\frac{\lambda}{2y} + \frac{1}{2}\log\left(\frac{\lambda}{2\pi y^3}\right)$ and rewrite (7) in the form (4). So $b'(\theta) = (2\theta)^{-1/2} = \mu$, $b''(\theta) = -(2\theta)^{-3/2} = -\mu^3$ and the variance function is $b''(\theta)a(\phi) = \frac{\mu^3}{\lambda}$.

[*Side Comment.* If you are wondering *why* this is a pdf, define $F(y) = \Phi\left(\sqrt{\frac{\lambda}{y}}\left(\frac{y}{\mu} - 1\right)\right) + e^{2\lambda/\mu}\Phi\left(-\sqrt{\frac{\lambda}{y}}\left(\frac{y}{\mu} + 1\right)\right)$. Here $\Phi(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-t^2/2}dt$ is the standard normal CDF, with derivative $\Phi'(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}$. You can easily verify that $F(y) \to 0$ as $y \downarrow 0$ and $F(y) \to 1$ as $y \uparrow +\infty$. Also, this takes some algebraic manipulation but you can also show $\frac{dF(y)}{dy} = f(y;\mu,\lambda)$ with the same μ and λ. Therefore, F is a legitimate CDF and f is its density. This and other properties of the Inverse Gaussian distribution can be found on a very well-written wikipedia page, https://en.wikipedia.org/wiki/Inverse_Gaussian_distribution.]

4 Link Functions and Variances

In addition to having an exponential family, the specification of a GLM involves a "link function" that relates the mean μ to the covariates. Specifically, if μ_i is the mean of the i'th observation y_i, then we assume

$$g(\mu_i) \quad = \quad \sum_{j=0}^{p} x_{ij}\beta_j$$

for *some* function $g(\mu_i)$. In principle, g can be any monotone differentiable function. However, in the case of an exponential family with canonical parameter θ and associated $b(\theta)$, it is particularly natural to define g so that $g(\mu) = \theta$. Since $\mu = b'(\theta)$, this also means that g is the inverse of the function b'. In that case, g is called the *canonical link function.*

There is some confusion over the function that Faraway writes $V(\mu)$ or $V(\mu_i)$ if we are writing specifically about the ith observation. In different places, this is described as either the variance of y_i or the portion of the variance that depends specifically on μ_i — the two could be different if ϕ is also unknown or if there are variable weights w_i. I think it's most consistent if you define $V(\mu_i)$ to be the same as $b''(\theta_i)$ — this makes sense, because there is a one to one relationship between μ_i and θ_i. This is consistent with Table 1 below, and it's also the way $V(\mu)$ is defined in [1], which is usually considered the authoritative reference book on GLMs. However in the case where you do have variable weights w_i, it may also make sense to define a function $V_i(\mu_i) = V(\mu_i)/w_i$, where $V(\mu_i)$ is still defined to be $b''(\theta_i)$.

Three examples:

1. Normal distribution: $b(\theta) = \frac{\theta^2}{2},\ b'(\theta) = \theta$ so $g(\mu) = \mu$, the identity link function. If we write the variance of Y as $V(\mu)\phi$ with $\phi = \sigma^2$ the dispersion parameter, then $V(\mu) = 1$.

2. Poisson distribution: $b(\theta) = b'(\theta) = e^\theta$, so the canonical link satisfies $g(e^\theta) = \theta$, which leads to $g(\mu) = \log\mu$. Also, in this case the variance function is $V(\mu) = \mu$.

3. Binomial distribution: $b(\theta) = \log(1 + e^\theta),\ b'(\theta) = \frac{e^\theta}{1+e^\theta}$, so the equation $\frac{e^\theta}{1+e^\theta} = \mu$ solves to $\theta = \log\frac{\mu}{1-\mu}$. In this case $V(\mu) = \mu(1-\mu)$.

Calculations for the gamma and inverse Gaussian distributions are similar, leading to the following table:

Model	Link function $g(\mu)$	$V(\mu)$
Normal	μ	1
Poisson	$\log \mu$	μ
Binomial	$\log(\mu/(1-\mu))$	$\mu(1-\mu)$
Gamma	$1/\mu$	μ^2
Inverse Gaussian	$1/\mu^2$	μ^3

Table 1: Canonical link and variance functions for the five previous models.

However, use of the canonical link is not mandatory, see, e.g. section 4.2 (p. 68 of the text), which gives three alternative link functions for binomial regression.

5 Estimating a GLM

Suppose the i'th observation has $a_i(\phi) = \frac{\phi}{w_i}$, then the log likelihood is

$$\ell(\beta; y_1, ..., y_n) \quad = \quad \sum_i \left\{ w_i \cdot \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right\}. \tag{8}$$

Note that the mean of y_i is $\mu_i = b'(\theta_i)$ and the variance is $b''(\theta_i)\phi/w_i = V(\mu_i)\phi/w_i = V_i(\mu_i)\phi$.

We have

$$\frac{\partial \ell}{\partial \beta_j} \quad = \quad \frac{1}{\phi} \sum_i w_i \left(y_i \frac{\partial \theta_i}{\partial \beta_j} - b'(\theta_i) \frac{\partial \theta_i}{\partial \beta_j} \right).$$

By the chain rule, $\frac{\partial \theta_i}{\partial \beta_j} = \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j}$. But $\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i)$, so

$$\frac{\partial \ell}{\partial \beta_j} \quad = \quad \frac{1}{\phi} \sum_i \left(\frac{y_i - b'(\theta_i)}{b''(\theta_i)/w_i} \cdot \frac{\partial \mu_i}{\partial \beta_j} \right).$$

But $b'(\theta_i) = \mu_i$, $b''(\theta_i) = V(\mu_i)$, so the equation $\frac{\partial \ell}{\partial \beta_j} = 0$ reduces to

$$\sum_i \frac{w_i}{\phi} \left(\frac{y_i - \mu_i}{V(\mu_i)} \cdot \frac{\partial \mu_i}{\partial \beta_j} \right) \quad = \quad 0. \tag{9}$$

However, solving (9) is equivalent to minimizing

$$\sum_i w_i \frac{(y_i - \mu_i)^2}{V(\mu_i)} \quad = \quad \sum_i \frac{(y_i - \mu_i)^2}{V_i(\mu_i)}. \tag{10}$$

In other words, finding the maximum likelihood estimators for a GLM is equivalent to solving the *weighted least squares* problem (10). This observation is critical in defining the algorithm used to find these estimators.

6 Deviances

Recall the formula (p. 157 of text)

$$\begin{aligned}\frac{D}{\phi} &= \frac{2}{\phi}\sum w_i\left\{y_i(\tilde{\theta}_i-\hat{\theta}_i)-b(\tilde{\theta}_i)+b(\hat{\theta}_i)\right\}\\ &= \sum\frac{2}{a_i(\phi)}\left\{y_i(\tilde{\theta}_i-\hat{\theta}_i)-b(\tilde{\theta}_i)+b(\hat{\theta}_i)\right\}\end{aligned}\tag{11}$$

where the second version follows from the formula $w_i=\frac{\phi}{a_i(\phi)}$ given on p. 153. This version is important because, in the derivation given above, $a_i(\phi)$ was negative. Here, $\frac{D}{\phi}$ is called the *scaled deviance*; D on its own is called the deviance. In (11), $\tilde{\theta}_i$ is the estimate of θ_i under the full or saturated model (in the normal case this corresponds to $\mu_i=\theta_i=y_i$) and $\hat{\theta}_i$ is the estimate under whatever model we are considering.

We derive the deviance for the five cases given previously.

1. Normal distribution. Here $\phi=a(\phi)=\sigma^2,\ b(\theta)=\frac{\theta^2}{2}$. Then

$$\begin{aligned}\frac{D}{\phi} &= \frac{2}{\sigma^2}\sum\left\{y_i(y_i-\hat{\mu}_i)-\frac{y_i^2}{2}+\frac{\hat{\mu}_i^2}{2}\right\}\\ &= \sum\frac{(y_i-\hat{\mu}_i)^2}{\sigma^2}.\end{aligned}$$

2. Poisson distribution. Here $\theta=\log\mu,\ b(\theta)=e^\theta,\ \phi=1$.

$$\begin{aligned}D &= 2\sum\{y_i(\log y_i-\log\hat{\mu}_i)-y_i+\hat{\mu}_i\}\\ &= 2\sum\left\{y_i\log\frac{y_i}{\hat{\mu}_i}-(y_i-\hat{\mu}_i)\right\}.\end{aligned}$$

3. Binomial distribution. Here $w_i=n_i,\ \theta_i=\log\frac{\mu_i}{1-\mu_i},\ b(\theta)=\log(1+e^\theta)=-\log(1-\mu_i)$ so

$$\begin{aligned}D &= 2\sum n_i\left\{y_i\left(\log\frac{y_i}{1-y_i}-\log\frac{\hat{\mu}_i}{1-\hat{\mu}_i}\right)+\log(1-y_i)-\log(1-\hat{\mu}_i)\right\}\\ &= 2\sum n_i\left\{y_i\log\frac{y_i}{\mu_i}+(1-y_i)\log\frac{1-y_i}{1-\hat{\mu}_i}\right\}\end{aligned}$$

which is the same as the formula given on page 157 though Faraway formulated it with y_i the ith count rather than the ith proportion. (In fact, my formula looks a bit more general than Faraway's because it allows for the possibility of different n_i, but Faraway could have done that as well.)

4. Gamma distribution. Here we use the second form of (11) in which $a(\phi)=-\phi$ as I had it; also $\theta=\frac{1}{\mu}$ and $b(\theta)=\log\theta=-\log\hat{\mu}$ so

$$\begin{aligned}\frac{D}{\phi} &= -\frac{2}{\phi}\sum\left\{y_i\left(\frac{1}{y_i}-\frac{1}{\hat{\mu}_i}\right)+\log y_i-\log\hat{\mu}_i\right\}\\ &= \frac{2}{\phi}\sum\left(\frac{y_i-\hat{\mu}_i}{\hat{\mu}_i}-\log\frac{y_i}{\hat{\mu}_i}\right\}.\end{aligned}$$

5. Inverse Gaussian distribution. Again $a(\phi) = -\phi$ and $\theta = \frac{1}{2\mu^2}$, $b(\theta) = (2\theta)^{1/2} = \frac{1}{\mu}$, so

$$\begin{aligned} \frac{D}{\phi} &= -\frac{2}{\phi}\sum\left\{y_i\left(\frac{1}{2y_i^2} - \frac{1}{2\hat{\mu}_i^2}\right) - \frac{1}{y_i} + \frac{1}{\hat{\mu}_i}\right\} \\ &= \frac{1}{\phi}\sum\frac{(y_i - \hat{\mu}_i)^2}{y_i\hat{\mu}_i^2}. \end{aligned}$$

7 Derivation of the Formulas on Page 155

The derivation of these formulas was given on pp. 40–43 of McCullagh and Nelder [1], but this is still not easy to follow, so I'm giving my own derivation here.

The objective is defined on the second last formula of page 154; we seek values of $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 & \dots & \beta_p \end{pmatrix}$, and hence $\eta_i = \sum_j x_{ij}\beta_j$ and $\mu_i = g^{-1}(\eta_i)$, to solve the equations

$$s_j = \sum_i \frac{(y_i - \mu_i)}{V(\mu_i)}\frac{\partial\mu_i}{\partial\beta_j} = 0,\ j = 1, ..., p.$$

Define $\mathbf{s}$ to be the vector with elements $s_1, ..., s_p$. To emphasize the dependence on $\boldsymbol{\beta}$, we also write this as $\mathbf{s}(\boldsymbol{\beta}) = \begin{pmatrix} s_1(\boldsymbol{\beta}) & \dots & s_p(\boldsymbol{\beta}) \end{pmatrix}$.

If there are fixed but variable weights w_i, then I think the same formula holds if you replace $V(\mu_i)$ by $V_i(\mu_i) = V(\mu_i)/w_i$ — and consistently use $V_i(\mu_i)$ in place of $V(\mu_i)$ in the whole of the following derivation. This case doesn't seem to be addressed explicitly in either Faraway or [1], but I think everything remains self-consistent if this one change to the notation is made.

We also let $H(\boldsymbol{\beta})$ be the vector of partial derivatives of $\mathbf{s}(\boldsymbol{\beta})$, with entries $h_{jk}(\boldsymbol{\beta})$, $1 \le j \le p$, $1 \le k \le p$ where

$$h_{jk}(\boldsymbol{\beta}) = \frac{\partial s_j(\boldsymbol{\beta})}{\partial\beta_k}.$$

The idea is this. Suppose our current guess (ℓ'th iteration) for the solution $\boldsymbol{\beta}$ is $\boldsymbol{\beta}^{(\ell)}$ and the true solution is $\boldsymbol{\beta}^0$, for which $\mathbf{s}(\boldsymbol{\beta}^{(0)}) = 0$. Then a first-order Taylor expansion gives

$$\begin{aligned} \mathbf{s}(\boldsymbol{\beta}^{(\ell)}) &= \mathbf{s}(\boldsymbol{\beta}^{(\ell)}) - \mathbf{s}(\boldsymbol{\beta}^{(0)}) \\ &= H(\bar{\boldsymbol{\beta}})(\boldsymbol{\beta}^{(\ell)} - \boldsymbol{\beta}^{(0)}) \end{aligned}$$

where $\bar{\boldsymbol{\beta}}$ is somewhere on the straight line joining $\boldsymbol{\beta}^{(\ell)}$ and $\boldsymbol{\beta}^{(0)}$. However, since $\bar{\boldsymbol{\beta}}$ is unknown, we approximate it by substituting $\boldsymbol{\beta}^{(\ell)}$. This suggests the next stage of the iteration

$$\boldsymbol{\beta}^{(\ell+1)} = \boldsymbol{\beta}^{(\ell)} - H(\boldsymbol{\beta}^{(\ell)})^{-1}\mathbf{s}(\boldsymbol{\beta}^{(\ell)}). \tag{12}$$

So far, this is essentially the method of *Newton-Raphson iteration*, which is one of the best-known algorithms for optimization. However, at this point, rather than perform an exact Newton-Raphson method, we make several approximations and simplifications of the formula (12).

Step 1. Since $\eta_i = \sum_j x_{ij}\beta_j$ we can write $x_{ij} = \frac{\partial \eta_i}{\partial \beta_j} = \frac{d\eta}{d\mu} \cdot \frac{\partial \mu_i}{\partial \beta_j}$ by the chain rule, so $\frac{\partial \mu_i}{\partial \beta_j} = \left(\frac{d\eta_i}{d\mu_i}\right)^{-1} x_{ij}$. We also have $0 = \frac{\partial^2 \eta}{\partial \beta_j \partial \beta_k} = \left(\frac{d\eta_i}{d\mu_i}\right)^2 \frac{\partial^2 \mu}{\partial \beta_j \partial \beta_k}$. Also, *we ignore the partial derivatives of V* — in other words, we act as though $V(\mu_i)$ were known and constant through a small neighborhood of the true μ_i. This is a critical step in the argument, which is difficult to explain beyond the intuition that it simplifies the algorithm without sacrificing much in terms of convergence. McCullagh and Nelder call this step *Fisher scoring*, citing a 1935 paper by R.A. Fisher, the great British statistician who was responsible for many of the developments in statistical methodology during the first half of the twentieth century. Fisher's paper was itself written as a discussion of a paper by Bliss, whose data on deaths in insects due to different concentrations of inseticide we have already seen in this course.

So if we put these items together,

$$\frac{\partial s_j}{\partial \beta_k} = \sum_i \left[-\frac{\partial \mu_i}{\partial \beta_k} \cdot \frac{1}{V(\mu_i)} \cdot \frac{\partial \mu_i}{\partial \beta_j} + \frac{y_i - \mu_i}{V(\mu_i)} \frac{\partial^2 \mu_i}{\partial \beta_j \partial \beta_k} - \frac{y_i - \mu_i}{V^2(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} \frac{\partial V_i}{\partial \beta_k} \right]$$

but we saw the second term is 0 and we are going to ignore the third term, so the (j,k) entry of H is

$$\begin{aligned} h_{jk} &\approx -\sum_i \frac{1}{V(\mu_i)} \left(\frac{d\eta_i}{d\mu_i}\right)^{-2} x_{ij} x_{ik} \\ &= -\sum_i w_i x_{ij} x_{ik} \end{aligned}$$

where the definition of the weight w_i is as given in step 2 on page 155, i.e.

$$w_i = \frac{1}{V(\mu_i)} \left(\frac{d\eta_i}{d\mu_i}\right)^{-2} \Big|_{\mu_i = \hat{\mu}_i^{(\ell)}}. \tag{13}$$

Hence $H = -X^T W X$ where W is the diagonal matrix with entries $w_1, ..., w_n$.

Step 2. We can also write

$$\begin{aligned} s_j &= \sum_i \frac{y_i - \mu_i}{V(\mu_i)} \Big|_{\mu_i = \hat{\mu}_i^{(\ell)}} \left(\frac{d\eta_i}{d\mu_i}\right)^{-1} x_{ij} \\ &= \sum_i w_i x_{ij} \cdot \left(\frac{d\eta_i}{d\mu_i}\right) (y_i - \hat{\mu}_i^{(\ell)}) \end{aligned}$$

so $\mathbf{s}(\boldsymbol{\beta})$ is of form $X^T W \mathbf{t}$ where $\mathbf{t} = \begin{pmatrix} t_1 & \dots & t_n \end{pmatrix}$ and $t_i = \left(\frac{d\eta_i}{d\mu_i}\right)(y_i - \hat{\mu}_i^{(\ell)})$.

Step 3. The iteration now reduces to

$$\begin{aligned} \boldsymbol{\beta}^{(\ell+1)} &= \boldsymbol{\beta}^{(\ell)} + (X^T W X)^{-1} X^T W \mathbf{t} \\ &= (X^T W X)^{-1} \left(X^T W X \boldsymbol{\beta}^{(\ell)} + X^T W \mathbf{t} \right) \\ &= (X^T W X)^{-1} X^T W \mathbf{z} \end{aligned}$$

where the vector $\mathbf{z}$ has entries z_i defined by

$$z_i \quad = \quad x_i^T \boldsymbol{\beta}^{(\ell)} + t_i \quad = \quad \eta_i^{(\ell)} + \left(\frac{d\eta_i}{d\mu_i} \right) (y_i - \hat{\mu}_i^{(\ell)}) \tag{14}$$

which is exactly the formula given on step 1 on page 155.

The conclusion is: define adjusted observations z_i by (14), weights w_i by (13); then the next iteration of $\boldsymbol{\beta}$ is defined by solving the linear regression equation for observations z_i with weights w_i.

Side note about notation: Since $\eta = g(\mu)$ with known link function g, we may also write $\frac{d\eta}{d\mu} = g'(\mu)$ which I find easier to comprehend, but to be consistent with Faraway's notation (and that of McCullagh and Nelder), I have kept that here. Where I have written $\frac{d\eta_i}{d\mu_i}$, this is to be understood the same as $g'(\hat{\mu}_i^{(\ell)})$, in other words, the value of the partial derivative $\frac{d\eta}{d\mu}$ when η and μ are both evaluated at the ith observation on the ℓ'th iteration.

References

[1] P. McCullagh and J.A. Nelder. *Generalized Linear Models, Second Edition.* Chapman and Hall, London, 1989.