Handout on Exponential Families and GLMs

Richard L. Smith

April 9, 2019

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 famility 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.

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 $b''(\theta)a(\phi) = \frac{\mu^3}{\lambda}$.

4 Consequences for GLMs

The Binomial is a bit of a special case because the main parameter p is restricted to [0,1] and the transformation $\theta = \log \frac{p}{1-p}$ is unique to that case. However the other four cases are defined by specific forms of the variance function $V(\mu)$:

- $V(\mu)$ is constant: normal (Gaussian) case
- $V(\mu) = \mu$: Poisson
- $V(\mu) = \mu^2$: Gamma
- $V(\mu) = \mu^3$: Inverse Gaussian

These relationships may be valid regardless of the underlying models that they are derived from: for example, $V(\mu) = \mu$ is derived from the Poisson model but the underlying mean-variance relationship may be valid without the assumption of a Poisson distribution, or even any discrete distribution. Therefore, for any particular example, it is legitimate to explore different functions $V(\mu)$ to find which one best fits the data.

5 The Delta Method for Variances

This is a totally different topic but it's mentioned on p. 181 so I thought I should explain.

Suppose Y is a random variable with mean μ and variance σ^2. Suppose, however, we are interested in the mean and variance, not of Y itself, but some transformation $g(Y)$.

We may write

$$g(Y) - g(\mu) \approx (Y - \mu)g'(\mu). \tag{8}$$

Based on the approximation (8), we deduce

$$\begin{aligned} \mathrm{E}\{g(Y)\} &= g(\mu), & (9)\\ \mathrm{Var}\{g(Y)\} &= \{g'(\mu)\}^2\sigma^2. & (10) \end{aligned}$$

For example (the case used on p. 181 of the text), based on $g(y) = e^y$, we deduce that the mean and standard deviation of e^Y are approximately e^μ and σe^μ.

For a formal asymptotic derivation of this approximation, suppose $Y_n,\ n = 1, 2, ...$, is a sequence of random variables and μ and σ^2 are limiting quantities such that $\sqrt{n}(Y_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ where $\mathcal{N}$ denotes the normal distribution and $\xrightarrow{d}$ denotes convergence in distribution. We may write $g(Y_n) - g(\mu) = g'(\tilde{\mu}_n)$ where $\tilde{\mu}_n$ lies between μ and Y_n. Since $Y_n \xrightarrow{p} \mu$ (convergence in probability) and we assume g' is continuous, we also deduce $\tilde{\mu}_n \xrightarrow{p} \mu$ and hence $g'(\tilde{\mu}_n) \xrightarrow{p} g'(\mu)$. Therefore

$$\sqrt{n}\{g(Y_n) - g(\mu)\} = g'(\tilde{\mu}_n) \cdot \sqrt{n}(Y_n - \mu)$$

which is the product of two random variables, one of which converges in probability to $g'(\mu)$, and the second of which converges in distribution to $\mathcal{N}(0, \sigma^2)$. A result known as *Slutsky's Theorem* then shows that the product converges in distribution to the product of the limits, in other words, $\mathcal{N}(0, \{g'(\mu)\}^2\sigma^2)$. In practice, we often assume (9) and (10) hold as approximations without checking the formal convergence.

6 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 $\mathbf{s}(\boldsymbol{\beta}) = \begin{pmatrix} s_1(\boldsymbol{\beta}) & \dots & s_p(\boldsymbol{\beta}) \end{pmatrix}$.

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

$$\boldsymbol{\beta}^{(\ell+1)} \;=\; \boldsymbol{\beta}^{(\ell)} - H(\boldsymbol{\beta}^{(\ell)})^{-1}\mathbf{s}(\boldsymbol{\beta}^{(\ell)}). \tag{11}$$

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 (11).

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.

$$w_i = \frac{1}{V(\mu_i)} \left(\frac{d\eta_i}{d\mu_i}\right)^{-2}\Big|_{\mu_i=\hat{\mu}_i^{(\ell)}}. \tag{12}$$

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

$$z_i = x_i^T \boldsymbol{\beta}^{(\ell)} + t_i = \eta_i^{(\ell)} + \left(\frac{d\eta_i}{d\mu_i}\right)(y_i - \hat{\mu}_i^{(\ell)}) \tag{13}$$

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

The conclusion is: define adjusted observations z_i by (13), weights w_i by (12); 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.