Exercises 2: Generalized linear models

Exponential families

We say that a distribution $f(y \mid \theta, \phi)$ is in an exponential family if we can write its PDF or PMF in the form

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

for some known functions a, b and c. We refer to θ as the canonical parameter of the family, and to ϕ as the dispersion parameter.

(A) Starting from the "standard" form of each PDF/PMF, show that the following distributions are in an exponential family, and find the corresponding b, c, θ, and $a(\phi)$.

- $Y \sim N(\mu, \sigma^2)$ for known σ^2.
- $Y = Z/N$ where $Z \sim \text{Binom}(N, P)$ for known N.
- $Y \sim \text{Poisson}(\lambda)$

(B) We want to characterize the mean and variance of an exponential family, but to do this simply, we need a preliminary lemma (that holds for all distributions, not just the exponential family). Define the *score* $s(\theta)$ as the gradient of the log likelihood:

$$s(\theta) = \frac{\partial}{\partial \theta} \log L(\theta)\,, \quad L(\theta) = \sum_{i=1}^{n} f(y_i; \theta)\,;$$

we've written this in multivariate form for the sake of generality, but of course it just involves an ordinary partial derivative (w.r.t. θ) in case where θ is one-dimensional. Let's also define $H(\theta)$ as the Hessian matrix, i.e. the matrix of second partial derivatives of the log likelihood:

$$H(\theta) = \frac{\partial}{\partial \theta^T} s(\theta) = \frac{\partial^2}{\partial \theta \partial \theta^T} \log L(\theta)$$

While we think of the score as a function of θ, clearly (like the likelihood) it also depends on the data. So a natural question is: what can we say about the *distribution* of the score over different random realizations of the data under the true data-generating process, i.e. at the true θ? It turns out we can say the following, sometimes referred to as the score equations:

$$\begin{aligned} E\{s(\theta)\} &= 0 \\ \text{var}\{s(\theta)\} &= -E\left\{H(\theta)\right\} \end{aligned}$$

where the mean and variance are taken under the true θ. **Prove the score equations.** Hints: prove the first equation first. You can assume that it's OK to switch the order of differentiation and integration (i.e. that any necessary technical conditions are met). To prove the second equation, differentiate both sides of the first equation with respect to θ^T and switch the order of differentiation and integration again. Expand out and simplify.

(C) Use the score equations you just proved to show that, if $Y \sim f(y; \theta, \phi)$ is in an exponential family, then

$$\begin{aligned} E\{Y\} &= b'(\theta) \\ \text{var}\{Y\} &= a(\phi)b''(\theta) \end{aligned}$$

Thus the variance of Y is a product of two terms. One of these terms, $b''(\theta)$, depends only on the canonical parameter θ, and hence on the mean, since you showed that $E\{Y\} = b'(\theta)$. The other, $a(\phi)$, is independent of θ. Note that the most common form of a is $a(\phi) = \phi/w$, where ϕ is called a dispersion parameter and where w is a known prior weight that can vary from one observation to another.

(D) To convince yourself that your result in (C) is correct, use these results to compute the mean and variance of the $N(\mu, \sigma^2)$ distribution.

Generalized linear models

Suppose we observe data like in the typical regression setting: that is, pairs $\{y_i, x_i\}$ where y_i is a scalar response for case i, and x_i is a p-vector of predictors or features for that same case i. We say that the y_i's follow a *generalized linear model* (GLM) if two conditions are met. First, the PDF (or PMF, if discrete) can be written as:

$$f(y_i; \theta_i, \phi) = \exp\left\{ \frac{y_i\theta_i - b(\theta_)}{\phi/w_i} + c(y_i; \phi/w_i) \right\}$$

where the weights w_i are all known. This is referred to as a the stochastic or random component of the model. Second, for some known invertible function g we have

$$g(\mu_i) = x_i^T \beta$$

where $\mu_i = E(Y_i; \theta_i, \phi)$. This is the systematic component of the model, and g is referred to as a link function, since it links the mean of the response μ_i with the *linear predictor* $\eta_i = x_i^T \beta$.

(A) Deduce from your results above that, in a GLM,

$$\theta_i = (b')^{-1}\left\{g^{-1}(x_i^T\beta)\right\}$$
$$\mathrm{var}\{Y_i\} = \frac{\phi}{w_i}V(\mu_i)$$

for some function V that you should specify in terms of the building blocks of the exponential family model. V is often referred to as a the *variance function*, since it explicitly relates the mean and the variance in a GLM.

(B) Take two special cases.

(1) Suppose that Y is a Poisson GLM, i.e. that the stochastic component of the model is a Poisson distribution. Show that $V(\mu) = \mu$.

(2) Suppose that $Y = Z/N$ is a Binomial GLM, i.e. that the stochastic component of the model is a Binomial distribution $Z \sim \mathrm{Binom}(N, P)$. Show that $V(\mu) = \mu(1-\mu)$.

(C) To specify a GLM we must choose the link function $g(\mu_i)$. Recall that g links the predictors with the mean of the response: $g(\mu_i) = x_i^T\beta$. Since you've shown that

$$\theta_i = (b')^{-1}\left\{g^{-1}(x_i^T\beta)\right\}$$

a particular simple choice of link function is one where $g^{-1} = b'$, or equivalently $g(\mu) = (b')^{-1}(\mu)$. This is known as the *canonical link*, in which case the canonical parameter simplifies to

$$\theta_i = (b')^{-1}\left\{b'(x_i^T\beta)\right\} = x_i^T\beta\,.$$

So under the canonical link $g(\mu) = b'^{-1}(\mu)$, we have the model

$$f(y_i;\beta,\phi)\exp\left\{\frac{y_i x_i^T\beta - b(x_i^T\beta)}{\phi/w_i} + c(y_i;\phi/w_i)\right\}$$

Now return to the two special cases from the previous problem.

(1) Suppose that Y is a Poisson GLM, i.e. that the stochastic component of the model is a Poisson distribution. Show that the canonical link is $g(\mu) = \log\mu$.

(2) Suppose that $Y = Z/N$ is a Binomial GLM, i.e. that the stochastic component of the model is a Binomial distribution $Z \sim \mathrm{Binom}(N, P)$. Show that the canonical link is $g(\mu) = \log\{\mu/(1-\mu)\}$.