CHAPTER 2. THE EXPONENTIAL FAMILY

The development of the theory of the generalized linear model is based upon the *exponential family* of distributions. This formalization recharacterizes familiar functions into a formula that is more useful theoretically and demonstrates similarity between seemingly disparate mathematical forms. The name refers to the manner in which all of the terms in the expression for these PDFs and PMFs are moved into the exponent to provide common notation. This does not imply some restrictive relationship with the well-known exponential probability density function.

Quasi-likelihood models (McCullagh, 1983; Wedderburn, 1974), which we describe later, replace this process and only require the stipulation of the first two moments. This allows the separation of the mean and variance functions, and the estimation is accomplished by employing a quasi-likelihood function. This approach has the advantage of accommodating situations in which the data are found or assumed not to be independent and identically distributed (henceforth iid).

Justification

Fisher (1934) developed the idea that many commonly applied probability mass functions and probability density functions are really just special cases of a more general classification he called the exponential family. The basic idea is to identify a general mathematical structure to the function in which uniformly labeled subfunctions characterize individual differences. The label "exponential family" comes from the convention that subfunctions are contained within the exponent component of the natural exponential function (i.e., the irrational number $e = 2.718281\ldots$ raised to some specified power). This is not a rigid restriction as any subfunction that is not in the exponent can be placed there by substituting its natural logarithm.

The primary payoff to reparameterizing a common and familiar function into the exponential form is that the isolated subfunctions quite naturally produce a small number of statistics that compactly summarize even large datasets without any loss of information. Specifically, the exponential family form readily yields sufficient statistics for the unknown parameters. A sufficient statistic for some parameter is one that contains all the information available in a given dataset about that parameter. For example, if we are interested in estimating the true range, $[a, b]$, for some uniformly distributed random variable: $X_i \in [a, b] \; \forall X_i$, then a sufficient statistic is the vector containing the first and last order statistics: $[x_{(1)}, x_{(n)}]$ from the sample of size n (i.e., the smallest and largest of the sampled values). No other elements of the data and no other statistic that we could construct from the data would provide further information about the limits. Therefore, $[x_{(1)}, x_{(n)}]$ provides "sufficient" information about the unknown parameters from the given data.

It has been shown (Barndorff-Nielsen, 1978, p. 114) that exponential family probability functions have all of their moments. The nth moment of a random variable about an arbitrary point, a, is $\mu_n = \mathbb{E}[(X - a)^n]$, and if a is equal to the expected value of X, then this is called the nth *central* moment. The first moment is the arithmetic mean of the random variable X, and the second moment along with the square of the first can be used to produce the following variance: $\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. While we are often interested only in the first two moments, the infinite moment property is very useful in assessing higher-order properties in more complex settings. In general, it is straightforward to calculate the moment generating function and the cumulant generating function for exponential family forms. These are simply functions that provide any desired moment or cumulant (logged moments) with quick calculations.

Throughout this monograph, we describe in detail the form, elements, characteristics, and examples of the most common probability density functions: Poisson, binomial, normal, gamma, negative binomial, and multinomial. Extensions and adaptions are more briefly described.

Two important classes of probability density functions are not members of the exponential family and therefore are not featured in this volume. The Student's t and the uniform distribution cannot be put into the form of Equation 2.1. Also, in general, a probability function in which the parameterization is dependent on the bounds, such as the uniform distribution, is not a member of the exponential family. Even if a probability function is not an exponential family member, it can sometimes qualify under particular circumstances. The Weibull probability density function (useful for modeling failure times), $f(y|\gamma, \beta) = \frac{\gamma}{\beta} y^{\gamma-1} \exp(-y^{\gamma}/\beta)$ for $x \geq 0, \gamma, \beta > 0$, is not an exponential family form since it cannot be rewritten in the required form Equation 2.2. However, if γ is known (or we are willing to assign a fixed value), then the Weibull PDF reduces to an exponential family form.

In the final chapter, we provide a brief introduction and description of other less common extensions to exponential forms that are designed to deal with certain data challenges. These include quasi-likelihood forms, zero-inflated models, generalized linear mixed-effects models, fractional regression, and Tobit models.

Some widely used members of the exponential family that facilitate generalized linear models but are *not* discussed here include beta, curved normal, Dirichlet, Pareto, and inverse gamma. The theoretical focus of this monograph is intended to provide readers with an understanding necessary to successfully encounter these and other distributional forms.

Derivation of the Exponential Family Form

Suppose we consider a one-parameter conditional probability density function or probability mass function for the random variable Z of the form $f(z|\zeta)$. This is read as "*f* of *z* given *zeta*." This function or, more specifically, this family of PDFs or PMFs is classified as an exponential family if it can be written in the following form:

$$f(z|\zeta) = \exp\left[t(z)u(\zeta)\right] r(z)s(\zeta), \tag{2.1}$$

where r and t are real-valued functions of z that do not depend on ζ, s and u are real-valued functions of ζ that do not depend on z, and $r(z) > 0, s(\zeta) > 0\ \forall z, \zeta$.

Furthermore, Equation 2.1 can easily be rewritten according to

$$f(z|\zeta) = \exp\left[\underbrace{t(z)u(\zeta)}_{\substack{\text{interaction}\\\text{component}}} + \underbrace{\log(r(z)) + \log(s(\zeta))}_{\text{additive component}}\right]. \tag{2.2}$$

The second part of the right-hand side of the equation is labeled the "additive component" because the summed components are distinct and additive with regard to z and ζ. The first part of the right-hand side is labeled the "interaction component" because it is reminiscent of the interaction specification of two parameters in a standard linear model. In other words, it is the component that reflects the product-indistinguishable relationship between z and ζ. It should be noted that the interaction component must specify $t(z)u(\zeta)$ in a strictly multiplicative manner. So a term such as $-\frac{1}{\beta}y^{\gamma}$, as seen in the exponent of the Weibull PDF, disqualifies this PDF from the exponential family classification.

In addition, the exponential structure of Equation 2.2 is preserved under random sampling such that the joint density function of independent, identically distributed (iid) random variables is given by the following: $\boldsymbol{Z} = \{Z_1, Z_2, \ldots, Z_n\}$ is

$$f(\boldsymbol{z}|\zeta) = \exp\left[u(\zeta)\sum_{i=1}^{n} t(z_i) + \sum_{i=1}^{n} \log(r(z_i)) + n\log(s(\zeta))\right]. \tag{2.3}$$

This means that the joint distribution of a systematic random sample of variates with exponential family marginal distributions is also an exponential family form. While the following chapters develop the theory of generalized linear models with Equation 2.2 for simplicity, the joint density function, Equation 2.3, is the more appropriate form since multiple data are used in all practical work. Fortunately, there is no loss of generality since the joint density function is also an exponential family form. If it makes the exposition easier to follow, picture Equation 2.2 with subscript i as an index of the data: $f(z_i|\zeta) = \exp\left[t(z_i)u(\zeta) + \log(r(z_i)) + \log(s(\zeta))\right]$.

Canonical Form

The canonical form is a handy simplification that greatly facilitates moment calculations as shown in Chapter 3. It is a one-to-one transformation (i.e., the inverse function of this function returns the same unique value) of terms of the probability function that reduces the complexity of the symbolism and reveals structure. It turns out to be much easier to work with an exponential family form when the format of the terms in the function says something directly about the structure of the data.

If $t(z) = z$ in Equation 2.2, then we say that this PDF or PMF is in its canonical form for the random variable Z. Otherwise, we can make the simple transformation $y = t(z)$ to force a canonical form. Similarly, if $u(\zeta) = \zeta$ in Equation 2.2, then this PDF or PMF is in its canonical form for the parameter ζ. Again, if not, we can force a canonical form by transforming $\theta = u(\zeta)$ and call θ the canonical parameter.

In many cases, it is not necessary to perform these transformations as the canonical form already exists or the transformed functions are tabulated for various exponential families of distributions. The final form after these transformations is the following general expression:

$$f(y|\theta) = \exp\left[y\theta - b(\theta) + c(y)\right]. \tag{2.4}$$

Note that the only term with both y and θ is a multiplicative term. McCullagh and Nelder (1989, p. 30) call $b(\theta)$ the "cumulant function," but $b(\theta)$ is also often called a "normalizing constant" because it is the only nonfunction of the data and can therefore be manipulated to ensure that Equation 2.4 sums or integrates to 1. This is a minor point here as all of the commonly applied forms of Equation 2.4 are well behaved in this respect. More important, $b(\theta)$ will play a key role in calculating the moments of the distribution. In addition, the form of θ, the *canonical link* between the original form and the θ parameterized form, is also important. The canonical link is used to generalize the linear model by connecting the linear-additive component of the nonnormal outcome variable.

The form of Equation 2.4 is not unique in that linear transformations can be applied to exchange values of y and θ between the additive component and the interaction component. In general, however, common families of PDFs and PMFs are typically parameterized in a standard form that minimizes the number of interaction terms. Also, it will sometimes be helpful to use Equation 2.4 expressed as a joint distribution of the data, particularly when working with the likelihood function (Chapter 3). This is just

$$f(\boldsymbol{y}|\theta) = \exp\left[\sum_{i=1}^{n} y_i\theta - nb(\theta) + \sum_{i=1}^{n} c(y_i)\right]. \tag{2.5}$$

The canonical form is used in each of the developed examples in this monograph. There is absolutely no information gained or lost by this treatment; rather, the form of Equation 2.5 is an equivalent form to Equation 2.3 where certain structures such as θ and $b(\theta)$ are isolated for theoretical consideration. As will be shown, these terms are the key to generalizing the linear model.

To add more intuition to the exponential family form Equation 2.5, consider how likelihood is constructed and used. A likelihood function is just the joint distribution of an observed set of data under the iid assumption for a given PDF or PMF: $f(\boldsymbol{X}|\theta) = f(\boldsymbol{X}_1|\theta) \times f(\boldsymbol{X}_2|\theta) \times \cdots \times f(\boldsymbol{X}_n|\theta)$. Fisher's notational sleight of hand was to note that once we observe the data, they are *known* even if θ is unknown. So notationally, use $L(\theta|\boldsymbol{X})$ for this product since we want the unknown value of θ *that is mostly likely to have generated* **X**. Returning to Equation 2.5, we see that there are three subcomponents of the function: one that interacts with the data and the parameter, one that is only a function of θ (multiplied by n, however), and finally one that is a function of the data only. If we care about the θ that mostly likely generated the data, then the latter is not going to be consequential. The first subcomponent shows how different values of $\boldsymbol{y}$ weight θ in the calculation, and $b(\theta)$ is a *parametric* statement about how we should treat θ in the maximum likelihood estimation process. In combination, these two demonstrate how data and distributional assumptions determine the final model that we produce. Obviously, some of these statements are slightly vague because we have yet to derive how the components of the exponential family form produce parameter estimates and predicted values.

Multiparameter Models

Up until now, only single-parameter forms have been presented. If generalized linear models were confined to single-parameter density functions, they would be quite restrictive. Suppose now that there are k parameters specified. A k-dimensional parameter vector, rather than just a scalar θ, is now easily incorporated into the exponential family form of Equation 2.4:

$$f(y|\boldsymbol{\theta}) = \exp\left[\sum_{j=1}^{k}\big(y\theta_j - b(\theta_j)\big) + c(y)\right]. \tag{2.6}$$

Here the dimension of $\boldsymbol{\theta}$ can be arbitrarily large but is often as small as two, as in the normal ($\boldsymbol{\theta} = \{\mu, \sigma^2\}$) or the gamma ($\boldsymbol{\theta} = \{\alpha, \beta\}$).

In the following examples, several common probability functions are rewritten in exponential family form with the intermediate steps shown (for the most part). It is actually not strictly necessary to show the process since the number of PDFs and PMFs of interest is relatively small. However, there is great utility in seeing the steps both as an instructional exercise and as a starting point for other distributions of interest not covered herein. Also, in each case, the $b(\theta)$ term is derived. The importance of doing this will be apparent in Chapter 3.

Example 2.1: Poisson Distribution Exponential Family Form
The Poisson distribution is often used to model counts such as the number of arrivals, deaths, or failures in a given time period. The Poisson distribution assumes that for short time intervals, the probability of an arrival is fixed and proportional to the length of the interval. It is indexed by only one (necessarily positive) parameter, which is both the mean and variance.

Given the random variable, Y, distributed Poisson with expected number of occurrences per interval μ, we can rewrite the familiar Poisson PMF in the following manner:

$$f(y|\mu) = \frac{e^{-\mu}\mu^y}{y!} = \exp\Big[\underbrace{y\log(\mu)}_{y\theta} - \underbrace{\mu}_{b(\theta)} \underbrace{-\log(y!)}_{c(y)}\Big].$$

In this example, the three components from Equation 2.4 are labeled by the underbraces. The interaction component, $y\log(\mu)$, clearly identifies $\theta = \log(\mu)$ as the canonical link. Also, $b(\theta)$ is simply μ. Therefore, the $b(\theta)$ term parameterized by θ (i.e., the canonical form) is obtained by taking the inverse of the $\theta = \log(\mu)$ to solve for μ. This produces $\mu = b(\theta) = \exp(\theta)$. Obviously, the Poisson distribution is a simple parametric form in this regard. ‖

Example 2.2: Binomial Distribution Exponential Family Form

The binomial distribution summarizes the outcome of multiple binary outcome (Bernoulli) trials such as flipping a coin. This distribution is particularly useful for modeling counts of success or failures given a number of independent trials such as votes received given an electorate, international wars given country dyads in a region, or bankruptcies given company starts.

Suppose now that Y is distributed binomial (n, p), where Y is the number of "successes" in a known number of n trials given a probability of success p. We can rewrite the binomial PMF in exponential family form as follows:[3]

$$\begin{aligned} f(y|n,p) &= \binom{n}{y} p^y (1-p)^{n-y} \\ &= \exp\left[\log\binom{n}{y} + y\log(p) + (n-y)\log(1-p)\right] \\ &= \exp\Bigg[\underbrace{y\log\left(\frac{p}{1-p}\right)}_{y\theta} - \underbrace{(-n\log(1-p))}_{b(\theta)} + \underbrace{\log\binom{n}{y}}_{c(y)}\Bigg]. \end{aligned}$$

From the first term in the exponent, we can see that the canonical link for the binomial distribution is $\theta = \log\left(\frac{p}{1-p}\right)$, so substituting the inverse of the canonical link function into $b(\theta)$ produces (with modest algebra) the following:

$$b(\theta) = [-n\log(1-p)]\Big|_{\theta=\log\left(\frac{p}{1-p}\right)} = n\log\left(1+\exp(\theta)\right).$$

So the expression for the $b(\theta)$ term in terms of the canonical parameter is $b(\theta) = n\log(1+\exp(\theta))$. In this example, n was treated as a known quantity or simply ignored as a nuisance parameter. Suppose instead that p was known and we developed the exponential family PMF with n as the parameter of interest:

$$\begin{aligned} f(y|n,p) &= \exp\left[\log\binom{n}{y} + y\log(p) + (n-y)\log(1-p)\right] \qquad (2.7)\\ &= \exp\left[\log(n!) - \log((\text{n-y})!) - \log(\text{y}!) + \ldots\right]. \end{aligned}$$

However, we cannot separate n and y in $\log((n-y)!)$ and they are not in product form, so this is not an exponential family PMF in this context. ‖

Example 2.3: Normal Distribution Exponential Family Form

The normal distribution is without question the workhorse of social science data analysis. Given its simplicity in practice and well-understood theoretical foundations, this is not surprising. The linear model (typically estimated with ordinary least squares [OLS]) is based on normal distribution theory, and as we shall see in Chapter 4, this comprises a very simple special case of the generalized linear model.

Often, we need to explicitly treat nuisance parameters instead of ignoring them or assuming they are known as was done in the binomial example above. The most important case of a two-parameter exponential family is when the second parameter is a scale parameter. Suppose ψ is such a scale parameter, possibly modified by the function $a(\psi)$, and then Equation 2.4 is rewritten:

$$f(y|\theta) = \exp\left[\frac{y\theta - b(\theta)}{a(\psi)} + c(y,\psi)\right]. \qquad (2.8)$$

When a given PDF or PMF does not have a scale parameter, then $a(\psi) = 1$, and Equation 2.8 reduces to Equation 2.4. In addition, Equation 2.8 can be put into the more general form of Equation 2.6 if we define $\boldsymbol{\theta} = \{\theta, a(\psi)^{-1}\}$ and rearrange. However, this form would no longer remind us of the important role the scale parameter plays.

The Gaussian normal distribution fits this class of exponential families. The subclass is called a location-scale family and has the attribute that it is fully specified by two parameters: a centering or location parameter and a dispersion parameter. It can be rewritten as follows:

$$\begin{aligned} f(y|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2)\right] \\ &= \exp\left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y^2 - 2y\mu + \mu^2)\right] \\ &= \exp\left[(\underbrace{y\mu}_{y\theta} - \underbrace{\frac{\mu^2}{2}}_{b(\theta)})/\underbrace{\sigma^2}_{a(\psi)} + \underbrace{\frac{-1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)}_{c(y,\psi)}\right]. \end{aligned}$$

Note that the μ parameter (the mean) is already in canonical form ($\theta = \mu$), so $b(\theta)$ is simply $b(\theta) = \frac{\theta^2}{2}$. This treatment assumes that μ is the parameter of interest and σ^2 is the nuisance parameter, but we might want to look at the opposite situation. However, in this treatment, μ is not considered a scale parameter. Treating σ^2 as the variable of interest produces

$$\begin{aligned} f(y|\mu,\sigma^2) &= \exp\left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y^2 - 2y\mu - \mu^2)\right] \\ &= \exp\left[\underbrace{\frac{1}{\sigma^2}}_{\theta}\underbrace{\left(y\mu - \frac{1}{2}y^2\right)}_{z} + \underbrace{\frac{-1}{2}\left(\log(2\pi\sigma^2) - \frac{\mu^2}{\sigma^2}\right)}_{b(\theta)}\right]. \end{aligned}$$

Now the canonical link is $\theta = \frac{1}{\sigma^2}$. So $\sigma^2 = \theta^{-1}$, and we can calculate the new $b(\theta)$:

$$b(\theta) = -\frac{1}{2}\left(\log(2\pi\sigma^2) - \frac{\mu^2}{\sigma^2}\right) = -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\theta) - \mu^2\theta.$$

‖

Example 2.4: Gamma Distribution Exponential Family Form

The gamma distribution is particularly useful for modeling terms that are required to be nonnegative such as variance components. Furthermore, the gamma distribution has two important special cases: the χ^2 distribution is gamma $(\frac{\rho}{2}, \frac{1}{2})$ for ρ degrees of freedom, and the exponential distribution is gamma$(1, \beta)$, both of which arise quite often in applied settings.

Assume Y is now distributed gamma indexed by two parameters: the shape parameter and the inverse-scale (rate) parameter. The gamma distribution is most commonly written as $f(y|\alpha, \beta) = \frac{1}{\Gamma(\alpha)}\beta^{\alpha}y^{\alpha-1}e^{-\beta y}, y, \alpha, \beta > 0$. For our purposes, a more convenient form is produced by transforming $\alpha = \delta, \beta = \delta/\mu$. The exponential family form of the gamma is produced by

$$
\begin{aligned}
f(y|\mu, \delta) &= \left(\frac{\delta}{\mu}\right)^{\delta} \frac{1}{\Gamma(\delta)} y^{\delta-1} \exp\left[\frac{-\delta y}{\mu}\right] \\
&= \exp\left[\delta \log(\delta) - \delta \log(\mu) - \log(\Gamma(\delta)) + (\delta - 1)\log(y) - \frac{\delta y}{\mu}\right] \\
&= \exp\Bigg[(\underbrace{-\frac{1}{\mu}y}_{\theta y} - \underbrace{\log(\mu)}_{b(\theta)})/\underbrace{\frac{1}{\delta}}_{a(\psi)} \\
&\qquad + \underbrace{\delta \log(\delta) + (\delta - 1)\log(y) - \log(\Gamma(\delta))}_{c(y,\psi)}\Bigg].
\end{aligned}
$$

From the first term in the last equation above, the canonical link for the gamma family variable μ is $\theta = -\frac{1}{\mu}$. So $b(\theta) = \log(\mu) = \log\left(-\frac{1}{\theta}\right)$ with the restriction $\theta < 0$. Therefore, $b(\theta) = -\log(-\theta)$. ‖

Example 2.5: Negative Binomial Distribution Exponential Family Form
The binomial distribution measures the number of successes in a given number of fixed trials, whereas the negative binomial distribution measures the number of failures before the rth success.[4] An important application of the negative binomial distribution is in survey research design. If the researcher knows the value of p from previous surveys, then the negative binomial can provide the number of subjects to contact to get the desired number of responses for analysis.

If Y is distributed negative binomial with success probability p and a goal of r successes, then the PMF in exponential family form is produced by

$$
\begin{aligned}
f(y|r, p) &= \binom{r + y - 1}{y} p^r (1 - p)^y \\
&= \exp\left[\underbrace{y \log(1 - p)}_{y\theta} + \underbrace{r \log(p)}_{b(\theta)} + \underbrace{\log\binom{r + y - 1}{y}}_{c(y)}\right].
\end{aligned}
$$

The canonical link is easily identified as $\theta = \log(1-p)$. Substituting this into $b(\theta)$ and applying some algebra gives $b(\theta) = r\log(1-\exp(\theta))$. ‖

Example 2.6: Multinomial Distribution Exponential Family Form

The multinomial distribution generalizes the binomial by allowing more than $k = 2$ nominal choices or events to occur. The set of possible outcomes for an individual i is a $k-1$ length vector of all zeros except for a single 1 identifying the chosen response: $\mathbf{Y}_i = [\mathbf{Y}_{i2}, \mathbf{Y}_{i3}, \ldots, \mathbf{Y}_{i(k)}]$. The first category $k = 1$ is left out of this vector and is called the reference (or baseline) category, and all model inferences are comparative to this category. Therefore, an individual picking the reference category will have all zeros in the $\mathbf{Y}_i$ vector.

We want to estimate the $k-1$ length of categorical probabilities $(\pi_1, \ldots, \pi_k)$ for a sample size of n, $g^{-1}(\boldsymbol{\theta}) = \boldsymbol{\mu} = [\pi_1, \pi_2, \ldots, \pi_{k-1}]$, from the dataset consisting of the $n \times (k-1)$ outcome matrix $\mathbf{Y}$ and the $n \times p$ matrix of p covariates $\boldsymbol{X}$ including a leading column of 1s. The estimates are provided with a logit (or a probit) link function, giving for each of the $k-1$ categories the probability that the ith individual picks category r:

$$p(\mathbf{Y}_i = r|\boldsymbol{X}) = \frac{\exp(\boldsymbol{X}_i\boldsymbol{\beta}_r)}{1 + \sum_{s=1}^{k-1}\exp(\boldsymbol{X}\boldsymbol{\beta}_s)}$$

where $\boldsymbol{\beta}_r$ is the coefficient vector for the rth category (logit version).

For simplicity of notation, consider $k = 3$ possible outcomes, without loss of generality, and drop the indexing by individuals. If there are n individuals in the data picking from these three categories, then the intuitive PMF that shows similarity to the binomial case is given by

$$\begin{aligned} f(\mathbf{Y}|n, \boldsymbol{\mu}) &= \frac{n}{Y_1!Y_2!(n - Y_1 - Y_2)!}\pi_1^{Y_1}\pi_2^{Y_2}(1-\pi_1-\pi_2)^{n-Y_1-Y_2} \\ &= \exp\Bigg[\Bigg(\underbrace{\left(\frac{\mathbf{Y}_1}{n}, \frac{\mathbf{Y}_2}{n}\right)}_{\mathbf{Y}}\underbrace{\left(\log\left(\frac{\pi_1}{1-\pi_1-\pi_2}\right), \log\left(\frac{\pi_2}{1-\pi_1-\pi_2}\right)\right)'}_{\boldsymbol{\theta}} \\ &\quad - \underbrace{(-\log(1-\pi_1-\pi_2))}_{b(\theta)}\Bigg)n + \underbrace{\log\left(\frac{n}{Y_1!Y_2!(n-Y_1-Y_2)!}\right)}_{c(y)}\Bigg]. \end{aligned}$$

Note that this exponential form has two-dimensional structure for $\mathbf{Y}$ and $\boldsymbol{\theta}$, which is an important departure from the previous examples. The two-dimensional link function that results from this form is

$$\boldsymbol{\theta} = (\theta_1, \theta_2) = g(\pi_1, \pi_2) = \left(\log\left(\frac{\pi_1}{1-\pi_1-\pi_2}\right), \log\left(\frac{\pi_2}{1-\pi_1-\pi_2}\right)\right).$$

We can therefore interpret the results in the following way for a single respondent:

$$\theta_1 = \log\left[\frac{p(\text{choice 1})}{p(\text{reference category})}\right] = \log\left[\frac{\pi_1}{1-\pi_1-\pi_2}\right] = \boldsymbol{X}_i\boldsymbol{\beta}_1$$

$$\theta_2 = \log\left[\frac{p(\text{choice 2})}{p(\text{reference category})}\right] = \log\left[\frac{\pi_2}{1-\pi_1-\pi_2}\right] = \boldsymbol{X}_i\boldsymbol{\beta}_2.$$

With minor algebra, we can solve for the inverse of the canonical link function:

$$\pi_1 = (1-\pi_2)\frac{\exp(\theta_1)}{1+\exp(\theta_1)}$$

$$\pi_2 = (1-\pi_1)\frac{\exp(\theta_2)}{1+\exp(\theta_2)}.$$

This allows us to rewrite $b(\theta)$ in terms of the two-dimensional canonical link function $\boldsymbol{\theta}$: $b(\theta) = -\log\left(1-(1-\pi_2)\frac{\exp(\theta_1)}{1+\exp(\theta_1)} - (1-\pi_1)\frac{\exp(\theta_2)}{1+\exp(\theta_2)}\right)$ which reveals multinomial structure in this simplified case. ‖

We have now shown that some of the most useful and popular PMFs and PDFs can easily be represented in the exponential family form. The payoff for this effort is yet to come, but it can readily be seen that if $b(\theta)$ has particular theoretical significance, then isolating it as we have in the θ parameterization is helpful. This is exactly the case as $b(\theta)$ is the engine for producing moments from the exponential family form through some basic likelihood theory. The reparameterization of commonly used PDFs and PMFs into the exponential family form highlights some well-known but not necessarily intuitive relationships between parametric forms. For instance, virtually all introductory statistics texts explain that the normal distribution is the limiting form for the binomial distribution. Setting the first and second derivatives of the $b(\theta)$ function in these forms equal to each other gives the appropriate asymptotic reparameterization: $\mu = np, \sigma^2 = np(1-p)$.