Exponential Dispersion Family

Definition 1.1. An **exponential dispersion density** takes the form:

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

Here θ_i is the *canonical parameter*, w_i is an subject-specific weight, ϕ is the *dispersion parameter*, $b(\cdot)$ is the *cumulant function*, and $c(y_i,\phi)$ is the log partition function. ■

Result 1.1 (Exponential Dispersion Properties).

- The log likelihood contribution of y_i is:

$$\ell(\theta_i,\phi) = \frac{y_i\theta_i - b(\theta_i)}{w_i\phi} + c(y_i,\phi).$$

- The score contribution of y_i:

$$s_i(\theta_i,\phi) = \frac{\partial \ell_i}{\partial \theta_i} = \frac{y_i - \dot{b}(\theta_i)}{w_i\phi}.$$

- The information contribution of y_i:

$$\mathcal{I}_{\theta_i\theta_i} = -E\left(\frac{\partial^2 \ell_i}{\partial \theta_i^2}\right) = \frac{\ddot{b}(\theta_i)}{w_i\phi}.$$

- The mean $E[y_i]$ of an exponential dispersion model is the first derivative of the cumulant function:

$$\mu_i = \dot{b}(\theta_i).$$

- The variance of an exponential dispersion model is a function of the mean:

$$\text{Var}(y_i) = w_i\phi\ddot{b}(\theta_i) = w_i\phi\ddot{b}\circ\dot{b}^{-1}(\mu_i) \equiv w_i\phi\nu(\mu_i).$$

  Here $\nu(\mu_i) = \ddot{b}\circ\dot{b}^{-1}(\mu_i)$ is the *variance function.*

♣

Generalized Linear Models

Definition 2.1. In a **generalized linear model** (GLM), a regression function is specified for the conditional mean:

$$E\big(y_i|\boldsymbol{x}_i\big) \equiv \mu_i = h(\eta_i).$$

Here $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$ is the *linear predictor* and h is the *activation function.* The inverse g of h is the *link function*:

$$g(\mu_i) = \eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

The activation function h and linear predictor η_i imply a model for the canonical parameter θ_i via:

$$\dot{b}(\theta_i) = \mu_i = h(\eta_i) \implies \theta_i = \dot{b}^{-1} \circ h(\eta_i).$$

■

2.1 Miscellaneous Relations

Proposition 2.1.

$$\dot{h}(\eta_i) = \frac{1}{\dot{g}(\mu_i)}.$$

♦

Proof.

$$1 = \frac{\partial}{\partial\eta_i}\eta_i = \frac{\partial}{\partial\eta_i}g \circ\ h(\eta_i) = \dot{g}\{h(\eta_i)\}\dot{h}(\eta_i) \implies \dot{h}(\eta_i) = \frac{1}{\dot{g}\{h(\eta_i)\}}.$$

■

Definition 2.2. The canonical parameter is related to the linear predictor via:

$$\theta_i = \dot{b}^{-1} \circ h(\eta_i).$$

If $g = \dot{b}^{-1}$ st $h = \dot{b}$, then:

$$\theta_i = \dot{b}^{-1} \circ \dot{b}(\eta_i) = \eta_i.$$

This choice of g is referred to as the **canonical link**. Under the canonical link, the canonical parameter is exactly the linear predictor. ■

Proposition 2.2. Under the canonical link $h(\cdot) = \dot{b}(\cdot)$:

$$\nu(\mu_i)\dot{g}(\mu_i) = 1.$$

♦

Proof. Recall $\dot{b}(\theta_i) = \mu_i$ and $\ddot{b}(\theta_i) = \nu(\mu_i)$. Under the canonical link $h = \dot{b}$ and $\theta_i = \eta_i$, thus:

$$\nu(\mu_i) = \ddot{b}(\theta_i) = \dot{h}(\eta_i) = \frac{1}{\dot{g}(\mu_i)}.$$

Conclude $\nu(\mu_i)\dot{g}(\mu_i) = 1$. ■

Proposition 2.3.

$$\frac{\partial\theta_i}{\partial\mu_i} = \frac{1}{\ddot{b}(\theta_i)} = \frac{1}{\nu(\mu_i)}.$$

♦

Proof. Since $\dot{b}(\theta_i) = \mu_i$:

$$\ddot{b}(\theta_i)\frac{\partial\theta_i}{\partial\mu_i} = \frac{\partial\mu_i}{\partial\mu_i} = 1.$$

■

2.2 Properties of GLMs

Result 2.1 (GLM Properties).

- The score for $\boldsymbol{\beta}$ is:

$$\boldsymbol{S}_\beta = \sum_{i=1}^{n} \frac{y_i - \dot{b}(\theta_i)}{w_i\phi\nu(\mu_i)}\frac{\boldsymbol{x}_i}{\dot{g}(\mu_i)}.$$

- The score for ϕ is:

$$S_\phi = \sum_{i=1}^{n}(-)\frac{y_i\theta_i - b(\theta_i)}{w_i\phi^2} + \dot{c}(y_i, \phi).$$

- The information for $\boldsymbol{\beta}$ is:

$$\mathcal{I}_{\beta\beta'} = \sum_{i=1}^{n} \frac{\boldsymbol{x}_i\boldsymbol{x}_i'}{w_i\phi\nu(\mu_i)\dot{g}^2(\mu_i)}.$$

- The information for ϕ is:

$$\mathcal{I}_{\phi\phi} = -2\sum_{i=1}^{n}\frac{y_i\theta_i - b(\theta_i)}{w_i\phi^3} + \frac{1}{2}\ddot{c}(y_i,\phi).$$

- The cross information between $\boldsymbol{\beta}$ and ϕ is:

$$\mathcal{I}_{\beta\phi} = \mathbf{0}.$$

♣

Proof. The model log likelihood is:

$$\ell(\boldsymbol{\beta},\phi) = \sum_{i=1}^{n}\ell_i(\boldsymbol{\beta},\phi) = \sum_{i=1}^{n}\frac{y_i\theta_i - b(\theta_i)}{w_i\phi} + c(y_i,\phi).$$

The score for $\boldsymbol{\beta}$ is:

$$\boldsymbol{S}_\beta = \frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n}\frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n}\frac{y_i - \dot{b}(\theta_i)}{w_i\phi}\cdot\frac{1}{\ddot{b}(\theta_i)}\cdot\dot{h}(\eta_i)\cdot\boldsymbol{x}_i.$$

Since $\ddot{b}(\theta_i) = \nu(\mu_i)$:

$$\boldsymbol{S}_\beta = \sum_{i=1}^{n}\boldsymbol{s}_i(\boldsymbol{\beta},\phi) = \sum_{i=1}^{n}\frac{y_i - \dot{b}(\theta_i)}{w_i\phi\nu(\mu_i)}\frac{\boldsymbol{x}_i}{\dot{g}(\mu_i)}.$$

The score for ϕ is:

$$S_\phi = \frac{\partial \ell}{\partial \phi} = \sum_{i=1}^{n}\frac{\partial \ell_i}{\partial \phi} = \sum_{i=1}^{n}(-)\frac{y_i\theta_i - b(\theta_i)}{w_i\phi^2} + \dot{c}(y_i,\phi).$$

The observed information for $\boldsymbol{\beta}$ is:

$$-\mathcal{J}_{\beta\beta'} = \frac{\partial^2 \ell}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} = \sum_{i=1}^{n}\left(\frac{\partial \boldsymbol{s}_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \boldsymbol{\beta}'} + \frac{\partial \boldsymbol{s}_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \boldsymbol{\beta}'}\right).$$

Evaluating the first derivative within the sum:

$$\frac{\partial \boldsymbol{s}_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \boldsymbol{\beta}} = -\frac{\ddot{b}(\theta_i)}{w_i\phi\nu(\mu_i)}\frac{\boldsymbol{x}_i}{\dot{g}(\mu_i)}\cdot\frac{1}{\ddot{b}(\theta_i)}\cdot\frac{1}{\dot{g}(\mu_i)}\boldsymbol{x}_i' = \frac{-\boldsymbol{x}_i\boldsymbol{x}_i'}{w_i\phi\nu(\mu_i)\dot{g}^2(\mu_i)}.$$

Observe that the second derivative within the sum is of the form:

$$\big(y_i - \dot{b}(\theta_i)\big)\frac{\boldsymbol{x}_i}{w\phi}\frac{\partial}{\partial \mu_i}\frac{1}{\nu(\mu_i)\dot{g}(\mu_i)}\cdot\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \boldsymbol{\beta}'}$$

Upon taking the expectation, this term vanishes due to the leading factor of $y_i - \dot{b}(\theta_i)$. Therefore, the Fisher information for $\boldsymbol{\beta}$ is:

$$\mathcal{I}_{\beta\beta'} = -E\left(\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right) = \sum_{i=1}^{n} \frac{\boldsymbol{x}_i \boldsymbol{x}_i'}{w_i \phi \nu(\mu_i) \dot{g}^2(\mu_i)}.$$

The observed information for ϕ is:

$$-\mathcal{J}_{\phi\phi} = \frac{\partial^2 \ell}{\partial \phi^2} = \sum_{i=1}^{n} 2\frac{y_i\theta_i - b(\theta_i)}{w_i \phi^3} + \ddot{c}(y_i, \phi).$$

The Fisher information for ϕ is:

$$\mathcal{I}_{\phi\phi} = -E\left(\frac{\partial^2 \ell}{\partial \phi^2}\right) = -2\sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{w_i \phi^3} + \frac{1}{2}\ddot{c}(y_i, \phi).$$

The observed information between $\boldsymbol{\beta}$ and ϕ is:

$$-\mathcal{J}_{\beta\phi} = \frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \phi} = \sum_{i=1}^{n} (-)\frac{y_i - \dot{b}(\theta_i)}{w_i \phi^2 \nu(\mu_i)} \frac{\boldsymbol{x}_i}{\dot{g}(\mu_i)}.$$

The Fisher information between $\boldsymbol{\beta}$ and ϕ is:

$$\mathcal{I}_{\beta\phi} = -E\left(\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \phi}\right) = \mathbf{0}.$$

■

Remark 2.1. Since $\hat{\boldsymbol{\beta}}$ is asymptotically independent of ϕ, a consistent estimate of $\boldsymbol{\beta}$ is obtained by solving the score equations $\boldsymbol{S}_\beta$ for β with any consistent estimator $\hat{\phi}$ substituted for the unknown dispersion parameter ϕ. ♦

Result 2.2. Define the following $n \times n$ matrices:

$$\boldsymbol{\Delta} = \text{diag}\Big\{\dot{g}(\mu_i)\Big\}$$
$$\boldsymbol{W} = \text{diag}\left\{\frac{1}{w_i \phi \nu(\mu_i) \dot{g}^2(\mu_i)}\right\}$$
$$\boldsymbol{\Sigma} = \text{diag}\Big\{\text{Var}(y_i)\Big\} = \text{diag}\Big\{w_i \phi \nu(\mu_i)\Big\}$$

These matrices are related through:

$$\boldsymbol{W}^{-1} = \boldsymbol{\Delta\Sigma\Delta}.$$

Using these forms, the score for $\boldsymbol{\beta}$ is expressible as:

$$\boldsymbol{S}_\beta = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{\Delta}(\boldsymbol{y} - \boldsymbol{\mu}).$$

The information for $\boldsymbol{\beta}$ is expressible as:

$$\mathcal{I}_{\beta\beta'} = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}.$$

♣

Result 2.3. Suppose $\hat{\boldsymbol{\beta}}^{(r)}$ is the current estimate of $\boldsymbol{\beta}$, and define the *working response vector* as:

$$\boldsymbol{z}^{(r)} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(r)} + \boldsymbol{\Delta}^{(r)}\left(\boldsymbol{y} - \boldsymbol{\mu}^{(r)}\right).$$

The $(r+1)$st estimate of $\boldsymbol{\beta}$, as given by the Newton-Raphson iteration, is identically the weighted least squares (WLS) estimator for regression of $\boldsymbol{z}^{(r)}$ on $\boldsymbol{X}$ using weights $\boldsymbol{W}^{(r)}$. That is:

$$\hat{\boldsymbol{\beta}}^{(r+1)} = \left(\boldsymbol{X}'\boldsymbol{W}^{(r)}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{W}^{(r)}\boldsymbol{z}^{(r)}.$$

♣

Proof. The Newton-Raphson iteration for updating $\boldsymbol{\beta}$ is:

$$\hat{\boldsymbol{\beta}}^{(r+1)} = \hat{\boldsymbol{\beta}}^{(r)} + \mathcal{I}^{-1}_{\beta\beta'}\left(\hat{\boldsymbol{\beta}}^{(r)}\right)\boldsymbol{S}_{\beta}\left(\hat{\boldsymbol{\beta}}^{(r)}\right).$$

Writing out the score and information:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{(r+1)} &= \hat{\boldsymbol{\beta}}^{(r)} + (\boldsymbol{X}'\boldsymbol{W}^{(r)}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{(r)}\boldsymbol{\Delta}^{(r)}\left(\boldsymbol{y} - \boldsymbol{\mu}^{(r)}\right) \\
&= (\boldsymbol{X}'\boldsymbol{W}^{(r)}\boldsymbol{X})^{-1}\left[(\boldsymbol{X}'\boldsymbol{W}^{(r)}\boldsymbol{X})\hat{\boldsymbol{\beta}}^{(r)} + \boldsymbol{X}'\boldsymbol{W}^{(r)}\boldsymbol{\Delta}^{(r)}\left(\boldsymbol{y} - \boldsymbol{\mu}^{(r)}\right)\right] \\
&= (\boldsymbol{X}'\boldsymbol{W}^{(r)}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{(r)}\left[\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(r)} + \boldsymbol{\Delta}^{(r)}\left(\boldsymbol{y} - \boldsymbol{\mu}^{(r)}\right)\right].
\end{aligned}$$

■

2.3 Deviance

Definition 2.3. Let $\ell(\boldsymbol{\mu}, \phi; \boldsymbol{y})$ denote the log likelihood as a function of the mean vector $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_n)$ and the dispersion parameter ϕ. If $\hat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\mu}} = h(\boldsymbol{X}\hat{\boldsymbol{\beta}})$, then $\ell(\hat{\boldsymbol{\mu}}, \phi; \boldsymbol{y})$ is the realized log likelihood. The maximum attainable log likelihood is $\ell(\boldsymbol{y}, \phi; \boldsymbol{y})$. Let $\hat{\theta}_i$ denote the canonical parameter for the ith observation under the MLE, and let $\tilde{\theta}_i$ denote the canonical parameter for the model that maximizes the log likelihood. The **scaled deviance** is:

$$D = -2\big\{\ell(\hat{\boldsymbol{\mu}}, \phi; \boldsymbol{y}) - \ell(\boldsymbol{y}, \phi; \boldsymbol{y})\big\} = \frac{2}{\phi}\sum_{i=1}^{n} w_i^{-1}\left[y_i\big(\hat{\theta}_i - \tilde{\theta}_i\big) - \big\{b(\hat{\theta}_i) - b(\tilde{\theta}_i)\big\}\right].$$

■

Result 2.4. The Pearson χ^2 statistic for GLMs is:

$$T = \sum_{i=1}^{n}\left\{\frac{y_i - \mu_i}{\sqrt{\mathrm{Var}(y_i)}}\right\}^2 = \sum_{i=1}^{n}\frac{(y_i - \mu_i)^2}{w_i\phi\nu(\mu_i)} \xrightarrow{\mathcal{L}} \chi^2_{n-p},$$

where $p = \dim(\boldsymbol{\beta})$. Setting $T \overset{\text{Set}}{=} E\{\chi^2_{n-p}\} = (n-p)$ and solving for ϕ gives a method of moments estimator for ϕ:

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

♣

2.4 Quasi Likelihood

Definition 2.4. The **log quasi likelihood** of an observation y_i with mean μ_i:

$$q_i = q(\mu_i) = \int_{y_i}^{\mu_i}\frac{y_i - u}{\phi\nu(\mu_i)}du.$$

■

Remark 2.2. The use of quasi likelihood allows for specification of GLMs with non-standard mean-variance relationships. ♦