Exponential Dispersion Family

Definition 1.1.1. An **exponential dispersion density** takes the form:

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

where θ_i is the *canonical parameter*, ϕ is the *dispersion parameter*, $b(\cdot)$ is the *cumulant function*, and $c(y_i, \phi)$ is the log partition function. ■

Proposition 1.1.1 (Exponential Dispersion Properties).

- The log likelihood contribution of y_i is:

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

- The score contribution of y_i:

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

- The information contribution of y_i:

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

- 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:

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

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

♦

Generalized Linear Models

Definition 1.2.1. 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 μ_i is the conditional mean, $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$ is the *linear predictor*, and h is the *activation function*. The *link function* g is the inverse of the activation function. The link function applied to the conditional mean returns the linear predictor:

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

The conditional mean model implies a model for the canonical parameter:

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

■

2.1 Miscellaneous Relations

Proposition 1.2.1. The derivative of the activation function is the reciprocal of the derivative of the link function:

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

♦

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

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

If the link function g is selected to coincide with the inverse of the derivative of the cumulant function $g = \dot{b}^{-1}$, such that $h = \dot{b}$, 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 1.2.2. Under the canonical link $g = \dot{b}^{-1}$:

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

♦

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

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

■

Proposition 1.2.3. For any link function:

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

In the case of the canonical link: $\partial_\mu\theta_i = \dot{g}(\mu_i)$. ♦

Proof. By implicit differentiation of $\dot{b}(\theta_i) = \mu_i$:

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

■

2.2 Properties of GLMs

Proposition 1.2.4 (GLM Properties). Suppose $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$.

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

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

- The score for ϕ is:

$$\mathcal{U}_\phi = \sum_{i=1}^{n}\left\{\frac{\partial c(y_i,\phi)}{\partial\phi} - \frac{y_i\theta_i - b(\theta_i)}{\phi^2}\right\}.$$

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

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

- The information for ϕ is:

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

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

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

♦

Proof. The model log likelihood is:

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

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

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

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

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

The score for ϕ is:

$$\mathcal{U}_\phi = \frac{\partial\ell}{\partial\phi} = \sum_{i=1}^{n}\frac{\partial\ell_i}{\partial\phi} = \sum_{i=1}^{n}\left\{\frac{\partial c(y_i,\phi)}{\partial\phi} - \frac{y_i\theta_i - b(\theta_i)}{\phi^2}\right\}.$$

Towards finding the information for $\boldsymbol{\beta}$, observe that $\mathcal{U}_\beta$ is expressed as a function of θ_i and μ_i, each of which is a function of $\boldsymbol{\beta}$. Now, the Hessian for $\boldsymbol{\beta}$ is:

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

Evaluating the first derivative within the sum:

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

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

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

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:

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

The Hessian for ϕ is:

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

The Fisher information for ϕ is:

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

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

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

Due to the factor of $\{y_i - \dot{b}(\theta_i)\}$, the Fisher information between $\boldsymbol{\beta}$ and ϕ vanishes:

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

■

Remark 1.2.1. Since $\hat{\boldsymbol{\beta}}$ and $\hat{\phi}$ are asymptotically independent, a consistent estimate of $\boldsymbol{\beta}$ is obtained by solving the score equations $\mathcal{U}_\beta$ for β with ϕ replaced by any consistent estimator $\hat{\phi}$. ♦

Proposition 1.2.5. Define the following $n \times n$ characteristic matrices:

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

These matrices are related via:

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

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

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

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

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

♦

Corollary 1.2.1. Under the canonical link $\nu(\mu_i)g(\mu_i) = 1$, hence:

$$\boldsymbol{W} = \text{diag}\left\{\frac{1}{\phi\dot{g}(\mu_i)}\right\}, \qquad \boldsymbol{W\Delta} = \phi^{-1}\boldsymbol{I}, \qquad \boldsymbol{W} = \boldsymbol{\Sigma}^{-1}.$$

♣

Result 1.2.1 (Iteratively Reweighted Least Squares). Suppose $\hat{\boldsymbol{\beta}}^{(r)}$ is the current estimate of $\boldsymbol{\beta}$, and define the *working response vector* as:

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

The Newton-Raphson update for $\boldsymbol{\beta}$ is the weighted least squares (WLS) estimator for regression of $\tilde{\boldsymbol{y}}^{(r)}$ on $\boldsymbol{X}$ using weights $\boldsymbol{W}^{(r)}$. That is:

$$\hat{\boldsymbol{\beta}}^{(r+1)} = \left(\boldsymbol{X}'\boldsymbol{W}^{(r)}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{W}^{(r)}\tilde{\boldsymbol{y}}^{(r)}. \tag{1.2.1}$$

♣

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

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

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 1.2.3. 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:

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

■

Result 1.2.2. The Pearson χ^2 statistic for GLMs is:

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

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:

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\nu(\hat{\mu}_i)}. \tag{1.2.2}$$

♣

2.4 Quasi Likelihood

Definition 1.2.4. The **log quasi likelihood** of an observation y_i with mean μ_i:

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

■

Discussion 1.2.1. The use of quasi likelihood allows for specification of GLMs with non-standard mean-variance relationships. The derivative of q_i is the quasi score:

$$u_i = \frac{\partial q_i}{\partial \mu_i} = \frac{y_i - \mu_i}{\phi\nu(\mu_i)}.$$

Observe that the quasi score has expectation zero and variance:

$$\text{Var}(u_i) = \text{Var}\left(\frac{y_i - \mu_i}{\phi\nu(\mu_i)} \right) = \frac{1}{\phi\nu(\mu_i)},$$

which coincides with the negative expected Hessian:

$$-E\left(\frac{\partial^2 q_i}{\partial \mu_i^2} \right) = \frac{1}{\phi\nu(\mu_i)}.$$

The regression parameters are estimated by solving the estimating equations:

$$\mathcal{U}_\beta = \frac{\partial \boldsymbol{\mu}'}{\partial \boldsymbol{\beta}} \text{diag}\{\text{Var}(y_i)\} \frac{\boldsymbol{y} - \boldsymbol{\mu}}{\phi} \overset{\text{Set}}{=} 0,$$

where the dispersion parameter ϕ is estimated as in (1.2.2). ♠

Inference

Example 1.3.1. Consider a GLM with linear predictor:

$$\eta_i = \boldsymbol{x}_i'\boldsymbol{\alpha} + \boldsymbol{z}_i'\boldsymbol{\beta}.$$

Suppose the hypothesis of interest is $H_0 : \boldsymbol{\beta} = \boldsymbol{0}$. Let $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ denote solutions to the score equations under the full model:

$$\mathcal{U}_\alpha(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{\Delta}(\boldsymbol{y} - \boldsymbol{\mu}) \overset{\text{Set}}{=} \boldsymbol{0},$$
$$\mathcal{U}_\beta(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \boldsymbol{Z}'\boldsymbol{W}\boldsymbol{\Delta}(\boldsymbol{y} - \boldsymbol{\mu}) \overset{\text{Set}}{=} \boldsymbol{0}.$$

Let $\tilde{\boldsymbol{\alpha}}$ a solution to the score equation under the null model:

$$\mathcal{U}_\alpha(\boldsymbol{\alpha}, \boldsymbol{\beta} = \boldsymbol{0}) = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{\Delta}(\boldsymbol{y} - \boldsymbol{\mu}) \overset{\text{Set}}{=} \boldsymbol{0},$$

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

$$\mathcal{I}_{\beta\beta'|\alpha} = \boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z} - \boldsymbol{Z}'\boldsymbol{W}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{Z}. \tag{1.3.3}$$

The Wald statistic for evaluating $H_0 : \boldsymbol{\beta} = \boldsymbol{0}$ is:

$$T_W = \hat{\boldsymbol{\beta}}'\mathcal{I}_{\beta\beta'|\alpha}\hat{\boldsymbol{\beta}} \tag{1.3.4}$$

Let $\tilde{\mathcal{U}}_\beta$ denote the score

$$\mathcal{U}_\beta = \boldsymbol{Z}'\boldsymbol{W}\boldsymbol{\Delta}(\boldsymbol{y} - \boldsymbol{\mu})$$

for $\boldsymbol{\beta}$ evaluated at $\boldsymbol{\alpha} = \tilde{\boldsymbol{\alpha}}$ and $\boldsymbol{\beta} = \boldsymbol{0}$. The score statistic for evaluating $H_0 : \boldsymbol{\beta} = \boldsymbol{0}$ is:

$$T_S = \tilde{\mathcal{U}}_\beta'\mathcal{I}_{\beta\beta|\alpha}^{-1}\tilde{\mathcal{U}}. \tag{1.3.5}$$

Under H_0, T_W and T_S are asymptotically $\chi^2_{\dim(\boldsymbol{\beta})}$. ♠

Example 1.3.2. Consider again a GLM with linear predictor:

$$\eta_i = \boldsymbol{x}_i'\boldsymbol{\alpha} + \boldsymbol{z}_i'\boldsymbol{\beta}.$$

A score statistic for testing $H_0 : \boldsymbol{\beta} = \boldsymbol{0}$ may be constructed using the working vector:

$$\tilde{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{\Delta}(\boldsymbol{y} - \boldsymbol{\mu}).$$

The expectation and variance of $\tilde{\boldsymbol{y}}$ under H_0 are:

$$E_0(\tilde{\boldsymbol{y}}) = \boldsymbol{X}\boldsymbol{\alpha}, \qquad \text{Var}_0(\tilde{\boldsymbol{y}}) = \boldsymbol{\Delta}\boldsymbol{\Sigma}\boldsymbol{\Delta} = \boldsymbol{W}^{-1}.$$

The GLS estimator of $\boldsymbol{\alpha}$ is:

$$\tilde{\boldsymbol{\alpha}} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\tilde{\boldsymbol{y}}.$$

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

$$\mathcal{U}_{\beta} = \boldsymbol{Z}'\boldsymbol{W}\boldsymbol{\Delta}(\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{Z}'\boldsymbol{W}\big(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\alpha}\big).$$

Evaluating the score at $\boldsymbol{\alpha} = \tilde{\boldsymbol{\alpha}}$ gives:

$$\tilde{\mathcal{U}}_{\beta} = \boldsymbol{Z}'\boldsymbol{Q}\tilde{\boldsymbol{y}},$$

where $\boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}$. The variance of $\tilde{\mathcal{U}}_{\beta}$ is:

$$\mathrm{Var}(\tilde{\mathcal{U}}_{\beta}) = \boldsymbol{Z}'\boldsymbol{Q}\boldsymbol{W}^{-1}\boldsymbol{Q}\boldsymbol{Z} = \boldsymbol{Z}'\boldsymbol{Q}\boldsymbol{Z},$$

which is exactly the efficient information (1.3.3). The working vector score statistic:

$$\begin{aligned} T_S &= \tilde{\mathcal{U}}_{\beta}'(\boldsymbol{Z}'\boldsymbol{Q}\boldsymbol{Z})^{-1}\tilde{\mathcal{U}}_{\beta} = \tilde{\boldsymbol{y}}\boldsymbol{Q}\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Q}\boldsymbol{Z})^{-1}\boldsymbol{Z}\boldsymbol{Q}\tilde{\boldsymbol{y}} \\ &= (\tilde{\boldsymbol{y}} - \boldsymbol{X}\tilde{\boldsymbol{\alpha}})'\boldsymbol{W}\boldsymbol{Z}\big(\mathcal{I}_{\beta\beta|\alpha}\big)^{-1}\boldsymbol{Z}\boldsymbol{W}(\tilde{\boldsymbol{y}} - \boldsymbol{X}\tilde{\boldsymbol{\alpha}}) \end{aligned}$$

coincides with the standard score statistic in (1.3.5). ♠

Examples

4.1 Normal Distribution

Example 1.4.1 (Canonical Link)**.** The normal distribution takes the form:

$$f(y_i|\mu_i,\sigma^2) = (2\pi\sigma^2)^{-1/2}\exp\left\{-\frac{(y_i-\mu_i)^2}{2\sigma^2}\right\} = \exp\left\{\frac{y_i\mu_i - \mu_i^2/2}{\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2) - \frac{y_i^2}{2\sigma^2}\right\}.$$

Identify σ^2 as the dispersion parameter and $\theta = \mu$ as the canonical parameter. The cumulant function and variance functions are:

$$b(\theta) = \frac{\theta^2}{2}, \qquad \nu(\mu) = \ddot{b}(\theta) = 1.$$

The canonical activation function is the identity:

$$h(\eta) = \dot{b}(\eta) = \eta.$$

Thus $g(\mu) = \mu$ and $\dot{g}(\mu) = 1$.

The characteristic matrices are:

$$\boldsymbol{\Delta} = \boldsymbol{I}, \qquad \boldsymbol{W} = (\sigma^2)^{-1}\boldsymbol{I}, \qquad \boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}.$$

♠

4.2 Bernoulli Distribution

Example 1.4.2 (Canonical Link)**.** The Bernoulli distribution takes the form:

$$f(y_i|\mu_i) = \mu_i^{y_i}(1-\mu_i)^{1-y_i} = \exp\left\{y_i \ln\left(\frac{\mu_i}{1-\mu_i}\right) + \ln(1-\mu_i)\right\}.$$

Identify:

$$\theta = \ln\left(\frac{\mu}{1-\mu}\right),$$

as the canonical parameter, which is related to μ via:

$$\mu = \frac{e^\theta}{1+e^\theta}, \qquad 1-\mu = \frac{1}{1+e^\theta}.$$

The dispersion parameter is $\phi \equiv 1$.

Expressed in terms of the canonical parameter, the density takes the form:

$$f(y_i|\theta_i) = \exp\left\{y_i\theta_i - \ln(1+e^{\theta_i})\right\}.$$

Identify $b(\theta) = \ln(1+e^\theta)$ as the cumulant function, whose derivatives are:

$$\dot{b}(\theta) = \frac{e^\theta}{1+e^\theta}, \qquad \ddot{b}(\theta) = \frac{e^\theta(1+e^\theta) - e^\theta(e^\theta)}{(1+e^\theta)^2)} = \frac{e^\theta}{1+e^\theta}\cdot\frac{1}{1+e^\theta} = \mu(1-\mu).$$

The canonical activation function is:

$$h(\eta) = \dot{b}(\eta) = \frac{e^\eta}{1+e^\eta},$$

with inverse:

$$g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right) = \ln(\mu) - \ln(1-\mu).$$

The derivative of the canonical link is:

$$\dot{g}(\mu) = \frac{1}{\mu} + \frac{1}{1-\mu} = \frac{1}{\mu(1-\mu)}$$

The characteristic matrices are:

$$\boldsymbol{\Delta} = \text{diag}\big\{\mu_i(1-\mu_i)\big\}, \qquad \boldsymbol{W} = \text{diag}\left\{\frac{1}{\mu_i(1-\mu_i)}\right\}, \qquad \boldsymbol{\Sigma} = \boldsymbol{W}^{-1}.$$

♠

Example 1.4.3 (Probit Link)**.** Consider the Bernoulli distribution with the Probit link:

$$g(\mu) = \Phi^{-1}(\mu) = \eta.$$

The first derivative is:

$$\dot{g}(\mu) = \frac{1}{\phi\big\{\Phi^{-1}(\mu)\big\}} = \frac{1}{\phi(\eta)},$$

where ϕ denotes the standard normal density. The characteristic matrices become:

$$\boldsymbol{\Delta} = \text{diag}\left\{\frac{1}{\phi(\eta_i)}\right\}, \qquad \boldsymbol{W} = \text{diag}\left\{\frac{\phi^2(\eta_i)}{\mu_i(1-\mu_i)}\right\}, \qquad \boldsymbol{\Sigma} = \text{diag}\big\{\mu_i(1-\mu_i)\big\}.$$

Note that:

$$\mu_i(1-\mu_i) = \Phi(\eta_i)\big\{1-\Phi(\eta_i)\big\} = \Phi(\eta_i)\Phi(-\eta_i).$$

♠

4.3 Poisson Distribution

Example 1.4.4 (Canonical Link)**.** The Poisson distribution takes the form:

$$f(y_i|\mu_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} = \exp\big\{y_i\ln(\mu_i) - \mu_i - \ln(y_i!)\big\}.$$

Identify $\theta = \ln(\mu)$ as the canonical parameter, with inverse $\mu = e^{\theta}$. The dispersion parameter is $\phi \equiv 1$. Expressing the density in terms of the canonical parameter:

$$f(y_i|\theta_i) = \exp\big\{y_i\theta_i - e^{\theta_i} - \ln(y_i!)\big\}.$$

Identify $b(\theta) = e^{\theta}$ as the cumulant function, with derivatives:

$$\dot{b}(\theta) = \ddot{b}(\theta) = e^{\theta}.$$

The canonical activation function is $h(\eta) = \dot{b}(\eta) = e^{\eta}$, with inverse $g(\mu) = \ln(\mu)$. The derivative of the canonical link is $\dot{g}(\mu) = \mu^{-1}$. The characteristic matrices are:

$$\boldsymbol{\Delta} = \left\{\frac{1}{\mu_i}\right\}, \qquad \boldsymbol{W} = \text{diag}\big\{\mu_i\big\}, \qquad \boldsymbol{\Sigma} = \boldsymbol{W}^{-1}.$$

♠