BIOSTAT 651
Notes #5: GLM: Estimation

- Lecture Topics:
  - Parameter estimation
  - Iterative methods
- Text (Dobson & Barnett, 3rd Ed.): Chapter 4

Exponential Family: Recap

- Suppose that Y_i arises from an exponential family with parameters θ_i and ϕ, where ϕ is known
  - density:

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

  - link function:

$\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$ $\qquad$ $\eta_i = g(\mu_i)$

  - moments:

$$\begin{aligned} E[Y_i] &\equiv \mu_i = b'(\theta_i) \\ V(Y_i) &= b''(\theta_i)a(\phi) = \frac{\partial \mu_i}{\partial \theta_i} a(\phi) = v(\mu_i)a(\phi) \\ v(\mu_i) &\equiv \frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) \end{aligned}$$

GLM: Canonical Link

- The function $g(\cdot)$ is a *canonical* link if $\theta_i = \eta_i$
- For canonical link,
  - $g(\mu_i) = \theta_i$
  - note: we already showed that $\mu_i = b'(\theta_i)$
  - therefore, $g(\cdot)$ and $b'(\cdot)$ are inverse functions

    $$g(b'(x)) = b'(g(x)) = x$$
    $$g^{-1}(x) = b'(x)$$
    $$b'^{-1}(x) = g(x)$$

    where the -1 refers to *inverse* as opposed to reciprocal
- Note that $g(\cdot)$ and $b'(\cdot)$ are one-to-one in the settings of our interest

GLM: Variance Function

- Calculating the variance function:
  - recall that $\mu_i = b'(\theta_i)$
  - $v(\mu_i) = b''(\theta_i)$
- Under the canonical link function: $v(\mu_i) = 1/g'(\mu_i)$

$$
\begin{aligned}
v(\mu_i) &= \frac{\partial \mu_i}{\partial \theta_i} \\
&= \left\{ \frac{\partial \theta_i}{\partial \mu_i} \right\}^{-1} \\
&= \left\{ \frac{\partial \eta_i}{\partial \mu_i} \right\}^{-1} \\
&= \frac{1}{g'(\mu_i)}
\end{aligned}
$$

Examples: Variance Function

- e.g., $Y_i \sim$ Normal:

$$\begin{aligned} g(\mu_i) &= \mu_i \\ g'(\mu_i) &= 1 \\ v(\mu_i) &= 1 \end{aligned}$$

- e.g., Logistic:

$$\begin{aligned} g(\mu_i) &= \log\left\{\frac{\mu_i}{1-\mu_i}\right\} \\ g'(\mu_i) &= \frac{1}{\mu_i(1-\mu_i)} \\ v(\mu_i) &= \mu_i(1-\mu_i) \end{aligned}$$

- e.g., Poisson:

$$\begin{aligned} g(\mu_i) &= \log(\mu_i) \\ g'(\mu_i) &= \frac{1}{\mu_i} \\ v(\mu_i) &= \mu_i \end{aligned}$$

GLMs with Canonical Link

Response	Distribution	η_i	$v(\mu_i)$
continuous	Normal	μ_i	1
$0, 1$	Bernoulli	$\log\left\{\frac{\mu_i}{1-\mu_i}\right\}$	$\mu_i(1-\mu_i)$
$0, 1, 2, \ldots$	Poisson	$\log(\mu_i)$	μ_i

Maximum Likelihood: GLM

- Likelihood:

$$L_i = \exp\left\{ \frac{Y_i\theta_i - b(\theta_i)}{a(\phi)} \right\}$$

- Log likelihood:

$$\ell_i = \frac{Y_i\theta_i - b(\theta_i)}{a(\phi)}$$

- Score function:
  - with ϕ treated as a *nuisance parameter*, the focus is on $\boldsymbol{\beta}$
  - therefore, work with

$$U_i(\boldsymbol{\beta}) = \frac{\partial \ell_i}{\partial \boldsymbol{\beta}}$$

  - although we could derive U_i from first principles, it is often easier to employ the *chain rule* ...

Maximum Likelihood: GLM (continued)

- Recall: Chain Rule for differentiation:

$$\frac{\partial}{\partial x} f(g(x)) \quad = \quad \frac{\partial f}{\partial g}\frac{\partial g}{\partial x}$$

- Applied to our setting,

$$U_i(\boldsymbol{\beta}) \quad = \quad \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \boldsymbol{\beta}}$$

- Computing each of the partial derivatives,

$$\begin{aligned}
\frac{\partial \ell_i}{\partial \theta_i} &= \frac{Y_i - b'(\theta_i)}{a(\phi)} \\
\frac{\partial \theta_i}{\partial \mu_i} &= \frac{1}{b''(\theta_i)} = \frac{1}{v(\mu_i)} \\
\frac{\partial \mu_i}{\partial \eta_i} &= \left\{\frac{\partial \eta_i}{\partial \mu_i}\right\}^{-1} = \frac{1}{g'(\mu_i)} \\
\frac{\partial \eta_i}{\partial \boldsymbol{\beta}} &= \mathbf{x}_i
\end{aligned}$$

Score Function: GLM

- Combining these results,

$$U(\boldsymbol{\beta}) \quad = \quad \sum_{i=1}^{n} \left\{ \frac{Y_i - \mu_i}{a(\phi)} \right\} \frac{1}{v(\mu_i) g'(\mu_i)} \, \mathbf{x}_i$$

- A more compact representation,

$$U(\boldsymbol{\beta}) \quad = \quad \frac{1}{a(\phi)} \mathbf{X}^T \mathbf{V}^{-1} \boldsymbol{\Delta}^{-1} (\mathbf{Y} - \boldsymbol{\mu})$$

where we have

$$\begin{aligned} \mathbf{V} &= \operatorname{diag}\{v(\mu_1), \ldots, v(\mu_n)\} \\ \boldsymbol{\Delta} &= \operatorname{diag}\{g'(\mu_1), \ldots, g'(\mu_n)\} \end{aligned}$$

- If the canonical link is used, then

$$U(\boldsymbol{\beta}) \quad = \quad \frac{1}{a(\phi)} \mathbf{X}^T (\mathbf{Y} - \boldsymbol{\mu})$$

Connection to Moment Estimator

- Therefore, under the canonical link, we could compute $\widehat{\boldsymbol{\beta}}$ as the solution to

$$\mathbf{X}^T(\mathbf{Y} - \boldsymbol{\mu}) \quad = \quad \mathbf{0}$$

- Note: $\widehat{\boldsymbol{\beta}}$ could also be viewed as a Method-of-Moments (MoM) estimator
  - i.e., equate the sufficient statistic for $\boldsymbol{\beta}$, namely $\mathbf{X}^T\mathbf{Y}$, with its mean:

$$\mathbf{X}^T\mathbf{Y} \quad = \quad E[\mathbf{X}^T\mathbf{Y}] = \mathbf{X}^T\boldsymbol{\mu}$$

Score Function: Normal Response

- Note: we've now derived the general form of the score function for any exponential family (with canonical link function)
- Now, suppose $Y_i \sim$ Normal with constant variance,

$$\begin{aligned} \theta_i &= \eta_i = \mu_i \\ \mu_i &= \mathbf{x}_i^T \boldsymbol{\beta} \\ a(\phi) &= \sigma^2 \end{aligned}$$

- Score function could then be written as

$$U(\boldsymbol{\beta}) = \frac{1}{a(\phi)} \mathbf{X}^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

Score Functions: Canonical Link

- To obtain the score function for other distributions, we just need expressions for $\boldsymbol{\mu}$ and $a(\phi)$
- e.g., Logistic regression:

$$\begin{aligned}
\eta_i = \mathbf{x}_i^T \boldsymbol{\beta} &= \log\left\{\frac{\mu_i}{1-\mu_i}\right\} \\
\mu_i &= \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \\
U(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \mathbf{x}_i \left(Y_i - \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \right)
\end{aligned}$$

- e.g., Poisson regression:

$$\begin{aligned}
\eta_i = \mathbf{x}_i^T \boldsymbol{\beta} &= \log(\mu_i) \\
\mu_i &= e^{\mathbf{x}_i^T \boldsymbol{\beta}}
\end{aligned}$$

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{x}_i \left(Y_i - e^{\mathbf{x}_i^T \boldsymbol{\beta}} \right)$$

Information Matrix: Canonical Link

- We need to compute the information matrix for *inference*, and even for *parameter estimation* itself
- Observed information (canonical link):

$$J(\boldsymbol{\beta}) = \frac{-\partial U(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} = -\sum_{i=1}^{n} \frac{\partial U_i}{\partial \boldsymbol{\beta}^T}$$

- Applying the chain rule again,

$$\frac{\partial U_i}{\partial \boldsymbol{\beta}^T} = \frac{\partial U_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \boldsymbol{\beta}^T}$$

with each of the partial derivatives given by

$$\begin{aligned} \frac{\partial U_i}{\partial \mu_i} &= -\frac{1}{a(\phi)} \mathbf{x}_i \\ \frac{\partial \mu_i}{\partial \eta_i} &= v(\mu_i) \\ \frac{\partial \eta_i}{\partial \boldsymbol{\beta}^T} &= \mathbf{x}_i^T \end{aligned}$$

Information Matrix: Can. Link (continued)

- Combining results on the preceding slide,

$$
\begin{aligned}
J(\boldsymbol{\beta}) &= \frac{1}{a(\phi)} \sum_{i=1}^{n} \mathbf{x}_i \; v(\mu_i) \; \mathbf{x}_i^T \\
&= \frac{1}{a(\phi)} \mathbf{X}^T \mathbf{V} \mathbf{X},
\end{aligned}
$$

  where $\mathbf{V} = \text{diag}\{v(\mu_1), \ldots, v(\mu_n)\}$

- e.g., $Y_i \sim$ Bernoulli, $v(\mu_i) = \mu_i(1 - \mu_i)$,

$$
J(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \mu_i (1 - \mu_i)
$$

- e.g., $Y_i \sim$ Poisson, $v(\mu_i) = \mu_i$,

$$
J(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \mu_i
$$

Information Matrix: Non-Canonical Link (Added)

- Recall: Score function (with non-canonical link)

$$U(\boldsymbol{\beta}) \quad = \quad \sum_{i=1}^{n} \left\{ \frac{Y_i - \mu_i}{a(\phi)} \right\} \frac{1}{v(\mu_i) g'(\mu_i)} \; \mathbf{x}_i$$

- Convenient to use the expected information,

$$\begin{aligned} I_i &= V(U_i) \\ &= E\left[(Y_i - \mu_i)^2 \right] \frac{1}{a(\phi)^2 g'(\mu_i)^2 v(\mu_i)^2} \mathbf{x}_i \mathbf{x}_i^T \\ &= \frac{1}{a(\phi) v(\mu_i) g'(\mu_i)^2} \mathbf{x}_i \mathbf{x}_i^T \end{aligned}$$

Information Matrix: Non-Canonical Link (Added)

- Expected Information

$$I_i = \frac{1}{a(\phi)v(\mu_i)g'(\mu_i)^2}\mathbf{x}_i\mathbf{x}_i^T$$

- We then obtain

$$\begin{aligned} I(\boldsymbol{\beta}) &= \sum_{i=1}^{n} I_i \\ &= \mathbf{X}^T \left\{a(\phi)\boldsymbol{\Delta}\mathbf{V}\boldsymbol{\Delta}\right\}^{-1} \mathbf{X} \end{aligned}$$

where we have

$$\begin{aligned} \mathbf{V} &= \text{diag}\{v(\mu_1), \ldots, v(\mu_n)\} \\ \boldsymbol{\Delta} &= \text{diag}\{g'(\mu_1), \ldots, g'(\mu_n)\} \end{aligned}$$

Computing MLEs

- Closed-form version of $\widehat{\boldsymbol{\beta}}$ generally only for the Normal model with identity link

  in all other cases, iterative methods are required ...

- We will now study:
  - Newton-Raphson method
  - Fisher scoring
  - role of WLS

Newton-Raphson: MLE

- Applying Newton-Raphson to solve the score equation,
  - initial value: $\widehat{\boldsymbol{\beta}}_0$; often set to $\mathbf{0}$
  - update step:

$$\widehat{\boldsymbol{\beta}}_{(j+1)} = \widehat{\boldsymbol{\beta}}_j + J^{-1}(\widehat{\boldsymbol{\beta}}_j)U(\widehat{\boldsymbol{\beta}}_j)$$

  - stopping criterion: $||\widehat{\boldsymbol{\beta}}_{(j+1)} - \widehat{\boldsymbol{\beta}}_j|| < \xi$
- Procedure is somewhat sensitive to the choice of starting value

Fisher Scoring

- Same general idea as Newton-Raphson, but replace $J(\boldsymbol{\beta})$ with its expectation, $I(\boldsymbol{\beta})$
  - update step:

$$\widehat{\boldsymbol{\beta}}_{(j+1)} = \widehat{\boldsymbol{\beta}}_j + I^{-1}(\widehat{\boldsymbol{\beta}}_j)U(\widehat{\boldsymbol{\beta}}_j)$$

- Lacks optimality properties of N-R method
  - generally takes longer to converge
  - more robust to poor choice of $\widehat{\boldsymbol{\beta}}_{(0)}$
- Note: for GLM with canonical link, Newton-Raphson and Fisher Scoring are equivalent

IRWLS in GLM: Canonical Link

- Consider the $(j+1)$th Fisher Scoring iterate,

$$\widehat{\boldsymbol{\beta}}_{(j+1)} = \widehat{\boldsymbol{\beta}}_j + I^{-1}(\widehat{\boldsymbol{\beta}}_j)U(\widehat{\boldsymbol{\beta}}_j)$$

- Multiply both sides by $I(\cdot)$,

$$I(\widehat{\boldsymbol{\beta}}_j)\widehat{\boldsymbol{\beta}}_{(j+1)} = I(\widehat{\boldsymbol{\beta}}_j)\widehat{\boldsymbol{\beta}}_j + U(\widehat{\boldsymbol{\beta}}_j)$$

- Written more in terms of the observed data,

$$\mathbf{X}^T\mathbf{V}_j\mathbf{X}\widehat{\boldsymbol{\beta}}_{(j+1)} = \mathbf{X}^T\mathbf{V}_j\left\{\mathbf{X}\widehat{\boldsymbol{\beta}}_j + \mathbf{V}_j^{-1}(\mathbf{Y} - \boldsymbol{\mu}_j)\right\}$$

- Setting $\boldsymbol{\eta}_j = \mathbf{X}\boldsymbol{\beta}_j$,

$$\mathbf{X}^T\mathbf{V}_j\mathbf{X}\widehat{\boldsymbol{\beta}}_{(j+1)} = \mathbf{X}^T\mathbf{V}_j\left\{\boldsymbol{\eta}_j + \mathbf{V}^{-1}(\mathbf{Y} - \boldsymbol{\mu}_j)\right\}$$

- Set $\mathbf{Z}_j = \boldsymbol{\eta}_j + \mathbf{V}_j^{-1}(\mathbf{Y} - \boldsymbol{\mu}_j)$, then solving,

$$\widehat{\boldsymbol{\beta}}^{(j+1)} = (\mathbf{X}^T\mathbf{V}_j\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}_j\mathbf{Z}_j$$

  - amounts to WLS estimator, with covariate $\mathbf{X}$, weight matrix $\mathbf{V}_j$ and response vector $\mathbf{Z}_j$

IRWLS

- Algorithm is known as *Iteratively Reweighted Least Squares* (IRWLS)
  - need initial estimate: e.g., $\widehat{\boldsymbol{\beta}}_0 = \mathbf{0}$
  - then, compute $\widehat{\mu}_{i,0}$, $V(\widehat{\mu}_{i,0})$, $\widehat{\eta}_{i,0} = \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}_0$
  - update weight matrix, $\mathbf{V}_0$, and response, $\mathbf{Z}_0 = \widehat{\boldsymbol{\eta}}_0 + \{\mathbf{V}_0\}^{-1}(\mathbf{Y} - \widehat{\boldsymbol{\mu}}_0)$
  - finally, update parameter estimate:

$$\widehat{\boldsymbol{\beta}}_1 = (\mathbf{X}^T \mathbf{V}_0 \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}_0 \mathbf{Z}_0$$

  - iterate until convergence obtained

IRWLS in GLM: Non-canonical Link (Added)

- Use the same algorithm with

$$
\begin{aligned}
U(\boldsymbol{\beta}) &= \frac{1}{a(\phi)}\mathbf{X}^T\mathbf{V}^{-1}\boldsymbol{\Delta}^{-1}(\mathbf{Y}-\boldsymbol{\mu}) \\
I(\boldsymbol{\beta}) &= \mathbf{X}^T\left\{a(\phi)\boldsymbol{\Delta}\mathbf{V}\boldsymbol{\Delta}\right\}^{-1}\mathbf{X}
\end{aligned}
$$

  where we have

$$
\begin{aligned}
\mathbf{V} &= \mathrm{diag}\{v(\mu_1),\ldots,v(\mu_n)\} \\
\boldsymbol{\Delta} &= \mathrm{diag}\{g'(\mu_1),\ldots,g'(\mu_n)\}
\end{aligned}
$$

IRWLS: Issues

- Q: Why did we switch from MLE to weighted least squares?