8 Obtaining MLEs and information

8.1 Review of numerical ML estimation

We aim at maximizing $l(\beta)$ w.r. to $\beta \in \mathcal{R}^p$.

Score vector	$U(\beta) = \begin{pmatrix} \partial l/\partial\beta_1 \\ \vdots \\ \partial l/\partial\beta_p \end{pmatrix}$
Hessian matrix	$H(\beta) = (\partial^2 l/\partial\beta_j\partial\beta_k),\ 1 \le j,k \le p$
Observed information	$I_{\text{obs}}(\beta) = -H(\beta)$
Information	$I(\beta) = E(I_{\text{obs}}(\beta))$
ML estimating equations	$U(\hat\beta) = 0,$ root $\hat\beta$ is MLE

Newton-Raphson iteration:

Expand score function by a multivariate Taylor expansion,

$$0 = U(\hat\beta) = U(\beta) + H(\beta)(\hat\beta - \beta) + O(\parallel \hat\beta - \beta \parallel^2) \tag{8.1}$$

$$\hat\beta = \beta - H^{-1}(\beta)U(\beta) + O(\cdots) \leftarrow \text{negligible under regularity conditions}$$

This motivates the Newton-Raphson iteration

Initialization $\hat\beta_{(0)}$ = starting value

= (for GLM) regular least squares estimate for β from multiple regression

Updating step

$$\hat\beta_{(l+1)} = \hat\beta_{(l)} - H^{-1}(\hat\beta_{(l)})U(\hat\beta_{(l)}) \tag{8.2}$$

For acceleration of convergence, often use acceleration factors for updating (Levenberg-Marquardt).

An important variant is *Fisher Scoring*, where we replace $I_{obs} = -H$ by I, the expected information:

$$\hat\beta_{(l+1)} = \hat\beta_{(l)} + I^{-1}(\hat\beta_{(l)})U(\hat\beta_{(l)}). \tag{8.3}$$

According to the second Bartlett identity, approximately,

$$\text{cov}(U(\beta)) = I(\beta).$$

Assume the limiting information $I_\infty(\beta) = \lim_{n\to\infty} n^{-1} I(\beta)$ is well-defined. By the law of large numbers, then also $n^{-1} I_{obs}(\beta) \to I_\infty(\beta)$ in probability. Then, using the score statistic (2.6),

$$nI_{obs}^{-1}(\beta) \to_P I_\infty^{-1}(\beta), \quad n^{-1/2} I^{1/2}(\beta) \to I_\infty^{1/2}(\beta), \quad I^{-1/2}(\beta) U(\beta) \to_D N_p(0, I_p).$$

We conclude, applying Slutsky's theorem, and ignoring terms of smaller order, from (8.1), that

$$\begin{aligned} \sqrt{n}(\hat{\beta} - \beta) &= \sqrt{n}[-H^{-1}(\beta) U(\beta)] \\ &= \sqrt{n} I_{obs}(\beta)^{-1} U(\beta) \\ &= [nI_{obs}(\beta)^{-1}]\ [n^{-1/2} I^{1/2}(\beta)]\ [I^{-1/2}(\beta) U(\beta)] \\ &\to_D N_p(0, I_\infty^{-1}(\beta)). \end{aligned} \tag{8.4}$$

This result provides asymptotic normality for the MLEs and justifies the finite sample approximations

$$\hat{\beta} \sim_{approx.} N_p(\beta, I^{-1}(\hat{\beta})), \quad \hat{\beta} \sim_{approx.} N_p(\beta, I_{obs}^{-1}),$$

which are used for finite-sample inference, corresponding to the Wald statistic (2.7).

8.2 Application to GLMs

Define $n \times n$-diagonal *weight matrix*

$$W = \{\text{diag}(g'(\mu_1)^2 V(\mu_1)\, \phi, \ldots, g'(\mu_n)^2 V(\mu_n)\, \phi)\}^{-1} = (w_{ij})_{1 \le i,j \le n}.$$

where $\eta_i = g(\mu_i)$, $\eta = g(\mu)$, $\mu = g^{-1}(\eta)$, and $V(\mu_i)$ is the *variance function*, ϕ is a constant. Further define the $n \times p$ *design matrix*

$$X = \begin{bmatrix} 1 & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

In the GLM, the log likelihood for one observation is $l(y, \theta, \phi) = \{(y\theta - b(\theta))/\phi + c(y, \phi)\}$, and the r-th element of the score vector $U = (u_1, \ldots, u_p)^\top$ (for one observation and in terms of β) is

$$u_r = \frac{\partial l}{\partial \beta_r} = \frac{\partial l}{\partial \theta} \frac{\partial \theta}{\partial \mu} \frac{\partial \mu}{\partial \eta} \frac{\partial \eta}{\partial \beta_r} \tag{8.5}$$

by the chain rule applied to the composite mapping

$$\beta \to \eta \to \mu \to \theta \to l, \qquad \text{where}$$

$$\eta = X\beta, \;\; \mu = g^{-1}(\eta), \;\; \theta = b'^{-1}(\mu) \qquad \text{in the exponential family.}$$

Identifying the derivatives, denoting by x_r the r-th column of the design matrix,

$$\begin{aligned} \frac{\partial \eta}{\partial \beta_r} &= x_r \\ \frac{\partial l}{\partial \theta} &= \frac{y - b'(\theta)}{\phi} = \frac{y - \mu}{\phi}, \;\; \text{since } \mu = b'(\theta) \\ \frac{\partial \theta}{\partial \mu} &= \frac{\partial b'^{-1}(\mu)}{\partial \mu} = \frac{1}{b''(b'^{-1}(\mu))} = \frac{1}{V(\mu)}, \end{aligned} \tag{8.6}$$

noting $V(\mu) = V(\theta(\mu)) = b''(\theta(\mu)) = b''(b'^{-1}(\mu))$ and therefore

$$u_r = \frac{y - \mu}{\phi V(\mu)} \frac{\partial \mu}{\partial \eta} x_r = (y - \mu)\, W \frac{d\eta}{d\mu} x_r. \tag{8.7}$$

8.3 Estimating equation

$$\sum_{i=1}^{n} u_{ri} = \sum_{i=1}^{n} (\frac{\partial l_i}{\partial \beta_r}) = 0$$

$$\Rightarrow \sum_{i=1}^{n} w_{ii}(y_i - \mu_i)\,\frac{d\eta_i}{d\mu_i}\,x_{ir} = 0$$

Hessian: $H = (h_{rs})_{1\le r,s\le p}$ with elements

$$\begin{aligned} h_{rs} &= \sum_{i=1}^{n} \frac{\partial^2 l_i}{\partial\beta_r \partial\beta_s} = \sum_{i=1}^{n} \frac{\partial u_{ri}}{\partial\beta_s} \\ &= \sum_{i=1}^{n} (y_i - \mu_i)\,\frac{\partial}{\partial\beta_s}\,(w_{ii}\frac{d\eta_i}{d\mu_i})\,x_{ir} + \sum_{i=1}^{n} w_{ii}\,\frac{d\eta_i}{d\mu_i}\,x_{ir}\,\frac{\partial}{\partial\beta_s}\,(y_i - \mu_i) \end{aligned}$$

$$H = X^\top (y-\mu)\,\left[\frac{\partial}{\partial\beta_s}\,W\,\frac{d\eta}{d\mu}\right] \;-\; X^\top W\,X, \tag{8.8}$$

since

$$\frac{\partial}{\partial\beta_s}\,(y_i - \mu_i) = -\frac{\partial\mu_i}{\partial\eta_i}\,\frac{\partial\eta_i}{\partial\beta_s} = -\frac{\partial\mu_i}{\partial\eta_i}\,x_{is}. \tag{8.9}$$

From $E\,y_i = \mu_i$, we obtain

$$I = -E(H) = X^\top W\,X \tag{8.10}$$

For the case of a canonical link function, $\theta = \eta$, for one observation, using $\partial\theta/\partial\eta \equiv \mathrm{id}$,

$$\theta = \eta \;\;\Rightarrow_{(8.5)}\;\; u_r = \frac{\partial l}{\partial\beta_r} = \frac{\partial l}{\partial\theta}\,\frac{\partial\theta}{\partial\mu}\,\frac{\partial\mu}{\partial\eta}\frac{\partial\eta}{\partial\beta_r} = \frac{\partial l}{\partial\theta}\,\frac{\partial\eta}{\partial\beta_r} = x_r\frac{y-\mu}{\phi}$$

$$h_{rs} = \sum_{i=1}^{n} \frac{\partial^2 l_i}{\partial\beta_r\partial\beta_s} = \sum_{i=1}^{n} \frac{\partial u_{ri}}{\partial\beta_s} =_{(8.9)} \frac{1}{\phi}\sum_{i=1}^{n}(-\frac{\partial\mu_i}{\partial\eta_i}x_{is}x_{ir}) = (-X^\top W\,X)_{rs} \tag{8.11}$$

as

$$w_{ii} = \frac{1}{g'(\mu_i)^2\phi V(\mu_i)} = \frac{1}{\phi}\,\frac{\partial\theta_i}{\partial\mu_i}\,(\frac{\partial\mu_i}{\partial\eta_i})^2 = \frac{1}{\phi}\frac{\partial\mu_i}{\partial\eta_i}$$

in the canonical case, using $\frac{\partial\theta_i}{\partial\mu_i} = 1/V(\mu_i)$, by (8.6). This implies

$$I = I_{\mathrm{obs}} = X^\top W\,X. \tag{8.12}$$

Conclusion: For the canonical link ($\theta = \eta$), , $I_{\text{obs}} = I$ and Fisher scoring is the same as Newton-Raphson with observed information. For any link, the Fisher information in a GLM is $I = X^\top W X$.

9 Iterated weighted least squares

Define auxiliary variables

$$Z = \eta + (y - \mu)\frac{d\eta}{d\mu} = \begin{bmatrix} \eta_1 + (y_1 - \mu_1)g'(\mu_1) \\ \vdots \\ \eta_n + (y_n - \mu_n)g'(\mu_n) \end{bmatrix} \quad n \times 1 \text{ - vector}$$

Iterated weighted least squares algorithm to obtain MLE $\hat{\beta}$:

1. Starting value for β, Z, W: $\hat{\beta}_{(0)}$, often obtained by regular (unweighted) multiple regressions: $Z_{(0)} = g(Y)$; $W_{(0)} = I_{n\times n}$, where $I_{n\times n}$ is the identity matrix.

2. Updating step: Given $\hat{\beta}_{(l)}$, obtain $\hat{\beta}_{(l+1)}$ by a weighted multiple linear regression step:

$$\begin{aligned} \hat{\beta}_{(l+1)} &= \text{argmin}_\beta \sum_{i=1}^{n} \{Z_{(l)i} - \sum_{j=1}^{p} x_{ij}\beta_{(l)j}\}^2 W_{(l)ii} \\ &= \text{argmin}_\beta (Z_{(l)} - X\beta)^\top W_{(l)} (Z_{(l)} - X\beta), \end{aligned} \tag{9.1}$$

   then update

$$\begin{aligned} \hat{\eta}_{(l+1)} &= X\hat{\beta}_{(l+1)}, \\ \hat{\mu}_{(l+1)} &= g^{-1}(\hat{\eta}_{(l+1)}), \\ W_{(l+1)} &= \{\text{diag}(g'(\hat{\mu}_{1(l+1)})^2 V(\hat{\mu}_{1(l+1)})\hat{\phi}, \ldots, g'(\hat{\mu}_{n(l+1)})^2 V(\hat{\mu}_{n(l+1)})\hat{\phi})\}^{-1}, \\ Z_{(l+1)} &= \hat{\eta}_{(l+1)} + (y - \hat{\mu}_{(l+1)})\frac{d\eta}{d\mu}|_{\mu=\hat{\mu}_{(l+1)}}. \end{aligned}$$

   Once β is updated, then update η, μ, W and Z.

3. Stopping criterion: Stop the iteration, if

$$\frac{\|\hat{\beta}_{(\ell+1)} - \hat{\beta}_{(\ell)}\|}{\|\hat{\beta}_{(\ell)}\|} \leq \varepsilon$$

for a suitably chosen small constant ε, or if the number of iterations is larger than a pre-specified number N.

Motivation: Note that for

$$Z = \eta + (y - \mu)g'(\mu) \approx g(y),$$

one has

$$Z - \eta = (y - \mu)g'(\mu)$$

and

$$(Z - \eta)^2 \frac{1}{g'(\mu)^2 V(\mu)\phi} = \frac{(y - \mu)^2}{V(\mu)\phi} = P,$$

where the l.h.s. can be rewritten as $(Z - X\beta)^\top W(Z - X\beta)$. Therefore, weighted least squares aims at minimizing the Pearson distance P. Note that $\hat{\phi}$ may or may not be updated.
Analyzing this iteration, we obtain

$$\hat{\beta}_{(l+1)} = (X^\top W_{(l)} X)^{-1} X^\top W_{(l)} Z_{(l)} \tag{9.2}$$

as the weighted least squares solution from normal equations.
Inserting $Z_{(l)}$, noting that by (8.7)

$$U(\hat{\beta}) = X^\top W(y - \hat{\mu}) \frac{\partial \eta}{\partial \mu},$$

$$\begin{aligned}
\hat{\beta}_{(l+1)} &= (X^\top W_{(l)} X)^{-1} X^\top W_{(l)} (X\hat{\beta}_{(l)} + (y - \hat{\mu}_{(l)}) \frac{d\eta}{d\mu}|_{\mu=\hat{\mu}_{(l)}}) \\
&= \hat{\beta}_{(l)} + (X^\top W_{(l)} X)^{-1} X^\top W_{(l)} (y - \hat{\mu}_{(l)}) \frac{d\eta}{d\mu}|_{\mu=\hat{\mu}_{(l)}} \\
&= \hat{\beta}_{(l)} + (X^\top W_{(l)} X)^{-1} U(\hat{\beta}_{(l)}) = \hat{\beta}_{(l)} + I^{-1} U(\hat{\beta}_{(l)}).
\end{aligned}$$

Result: For ML estimation in the GLM, Fisher scoring and iterated weighted least squares are equivalent. For the canonical link, Fisher scoring and Newton-Raphson iteration with observed information are equivalent.

Notes: (1) Inference is usually based on expected information

$$E(I_{obs}) = I = X^\top W X,$$

and on the approximate distribution

$$\hat{\beta} \sim N_p(\beta, (X^\top W X)^{-1}), \quad W = W(\beta) \tag{9.3}$$

for large samples, from which we obtain simultaneous inference or inference for individual components as described previously.

(2) The linearization through IWLS is a key device to extend arguments and tools that are available for (weighted) least squares to GLMs.

(3) An underlying regularity condition for the asymptotics to work is:

$$I_\infty = \lim_{n\to\infty} \frac{1}{n}(X^\top W X)$$

exists and is invertible. This being an asymptotic condition, it has no direct bearing on the finite sample situation.

Example for inference: To test

$$H_0: \ A\beta = \zeta, \ \text{ where } \quad A \text{ is } q \times p, \ \beta \in \mathcal{R}^p \ \text{ and } \ \zeta \in \mathcal{R}^q,$$

note that by (9.3), approximately,

$$A\hat{\beta} \sim N_q(A\beta, A(X^\top W X)^{-1} A^\top),$$

and that under H_0 we have, approximately,

$$(A\hat{\beta} - \zeta)^\top [A(X^\top W X)^{-1} A^\top]^{-1} (A\hat{\beta} - \zeta) \sim \chi^2_q,$$

if $A(X^\top W X)^{-1} A^\top$ is of full rank q. Can also construct confidence ellipsoid for parameter vector β, and for $\zeta = A\beta$, the latter based on the spectral decomposition of $[A(X^T W X)^{-1} A^T]$ (rather than that of $[X^T W X]^{-1}$).

Computational note: Implementing weighted least squares of Z on X with weight matrix W requires nothing more than a simple least squares step. Setting

$$X' = W^{1/2} X, \quad Z' = W^{1/2} Z, \tag{9.4}$$

the weighted least squares solution $\hat{\beta}$ is the same as the least squares solution for regressing Z' on predictors X'. Note that the *hat matrix* H of the transformed regression is defined by $W^{1/2}\hat{Z} = HW^{1/2}Z$ for

$$H = W^{1/2} X (X^T W X)^{-1} X^T W^{1/2}. \tag{9.5}$$

Note that when applying (9.4), one has to do this for each iteration step as the matrices W are updated for each iteration. The hat matrix H to be used for diagnostics is the one obtained at the last iteration.