

36-711 Fall 2003: Maximum Likelihood II

Brian Junker

November 18, 2003

Slide 1

- Newton's Method and Scoring for MLE's
- Aside on WLS/GLS
- Application to Exponential Families
- Application to Generalized Linear Models
- Application to Nonlinear Least Squares
- Application to Robust Regression

Newton's Method and Scoring for MLE's

Slide 2

- When carrying out *Newton-Raphson* to maximize $\ell_n(\theta)$, the natural iterates are:

$$\hat{\theta}_n^{(j+1)} = \hat{\theta}_n^{(j)} - [\nabla^2 \ell_n(\hat{\theta}_n^{(j)})]^{-1} \nabla \ell_n(\hat{\theta}_n^{(j)})$$

- Sometimes the observed information has a simpler form, in which case one may use

$$\hat{\theta}_n^{(j+1)} = \hat{\theta}_n^{(j)} + I_n(\hat{\theta}_n^{(j)})^{-1} \nabla \ell_n(\hat{\theta}_n^{(j)})$$

Maximizing using this quasi-Newton method is called *Fisher Scoring*.

Slide 3

Aside on WLS/GLS

Suppose

$$y = X\beta + \epsilon, \quad \epsilon \sim N(0, \Sigma)$$

We can convert this to an ordinary least squares problem via

$$\Sigma^{-1/2}y = \Sigma^{-1/2}X\beta + \epsilon^*, \quad \epsilon^* \sim N(0, I_{p \times p})$$

whose solution is

$$\hat{\beta} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} y$$

- When Σ is diagonal, this is called *weighted least squares (WLS)*
- When Σ is general, this is called *generalized least squares (GLS)*

We will repeatedly apply this idea in the examples below.

Slide 4

Application to Exponential Families

Let $y = (y_1, \dots, y_n)^T$ denote iid data to be modelled with an exponential family model. Recall that an exponential family model has the form

$$f(y_i|\theta) = g(y_i)e^{\beta(\theta) + \gamma(\theta)^T k(y_i)}$$

where $\theta_{p \times 1}$ are the p original parameters, $\gamma(\theta)_{r \times 1}$ are the r *natural parameters*, and $k(y_i)_{r \times 1}$ are the r *sufficient statistics* for γ . Then the likelihood for y is

$$\begin{aligned} L_n(\theta) &= \prod_{i=1}^n f(y_i|\theta) = \prod_{i=1}^n g(y_i)e^{\sum_{i=1}^n \beta(\theta) + \sum_{i=1}^n \gamma(\theta)^T k(y_i)} \\ &= G(y)e^{B(\theta) + \gamma(\theta)^T K(y)} \end{aligned}$$

and the log-likelihood is

$$\ell_n(\theta) = \log G(y) + B(\theta) + \gamma(\theta)^T K(y)$$

Comparing Newton-Raphson with Fisher Scoring

In order to apply Newton-Raphson to find $\hat{\theta}_n$, we need to compute $\nabla \ell_n(\theta)$ and $\nabla^2 \ell_n(\theta)$. A simple form for $\nabla \ell_n(\theta)$ follows from

$$\begin{aligned}\nabla \ell_n(\theta) &= \nabla B(\theta) + \nabla \gamma(\theta) K(y) \\ 0 &= E[\nabla \ell_n(\theta)] = \nabla B(\theta) + \nabla \gamma(\theta) \mu(\theta)\end{aligned}$$

so that

$$\nabla \ell_n(\theta) = \nabla \gamma(\theta) [K(y) - \mu(\theta)]$$

where $\mu(\theta)_{r \times 1} = E_\theta[K(y)_{r \times 1}]$ and $\nabla \gamma(\theta) = [\partial \gamma_j / \partial \theta_i]_{p \times r} = J_\gamma(\theta)^T$. From the first expression for $\nabla \ell_n(\theta)$ above it is also easy to see that

$$-\nabla^2 \ell_n(\theta) = \begin{cases} -\nabla^2 B(\theta), & \text{if } \exists A_{r \times p} \text{ s.t. } \gamma(\theta) = A\theta \\ (\text{messy}) & \text{otherwise} \end{cases}$$

In the first case, we see that Newton-Raphson and Fisher Scoring are really the same thing: $I_n(\theta) = E[-\nabla^2 \ell_n(\theta)] = E[-\nabla^2 B(\theta)] = -\nabla^2 B(\theta)$.

Slide 5

Fisher Scoring when Newton is Ugly

Using the form $\nabla \ell_n(\theta) = \nabla \gamma(\theta) [K(y) - \mu(\theta)]$, we have

$$\begin{aligned}I_n(\theta) &= E_\theta[-\nabla^2 \ell_n(\theta)] = \text{Var}_\theta(\nabla \ell_n(\theta)) \\ &= \nabla \gamma(\theta) \Sigma(\theta) \nabla \gamma(\theta)^T = \nabla \mu(\theta) \Sigma(\theta)^{-1} \nabla \mu(\theta)^T\end{aligned}$$

where $\Sigma(\theta) = \text{Var}_\theta(K(y))$ and the last equality follows from

$$\begin{aligned}\nabla \mu(\theta)^T &= \nabla \int K(y) L_n(y|\theta) d\nu(y) = \int K(y) [\nabla L_n(y|\theta)]^T d\nu(y) \\ &= \int K(y) [\nabla \ell_n(\theta)]^T L_n(\theta) d\nu(y) = \int K(y) [K(y) - \mu(\theta)]^T \nabla \gamma(\theta)^T L_n(\theta) d\nu(y) \\ &= \int [K(y) - \mu(\theta)] [K(y) - \mu(\theta)]^T L_n(\theta) d\nu(y) \cdot \nabla \gamma(\theta)^T = \Sigma(\theta) \nabla \gamma(\theta)^T\end{aligned}$$

This shows that $I_n(\theta)$ can be expressed in terms of the first and second moments of $K(y)$, which may be simpler than working with $-\nabla^2 \ell_n(\theta)$.

Slide 6

Slide 7

Some details

Fisher scoring, $\hat{\theta}^{(j+1)} = \hat{\theta}^{(j)} + I_n(\hat{\theta}^{(j)})^{-1} \nabla \ell_n(\hat{\theta}^{(j)})$, may be expressed as

$$\hat{\theta}^{(j+1)} = \hat{\theta}^{(j)} + \left\{ \nabla \mu(\hat{\theta}^{(j)}) \Sigma(\theta^{(j)})^{-1} \nabla \mu(\theta^{(j)})^T \right\}^{-1} \nabla \gamma(\hat{\theta}^{(j)}) [K(y) - \mu(\hat{\theta}^{(j)})]$$

and, after applying our identity $\nabla \mu(\theta)^T = \Sigma(\theta) \nabla \gamma(\theta)^T$, we get

$$\hat{\theta}^{(j+1)} = \hat{\theta}^{(j)} + \left\{ \nabla \mu(\hat{\theta}^{(j)}) \Sigma(\theta^{(j)})^{-1} \nabla \mu(\theta^{(j)})^T \right\}^{-1} \nabla \mu(\hat{\theta}^{(j)}) \Sigma(\theta^{(j)})^{-1} [K(y) - \mu(\hat{\theta}^{(j)})]$$

which again uses just $K(y)$ and its first two moments $\mu(\theta)$ and $\Sigma(\theta)$. This suggests the following *iteratively reweighted least squares (IRLS)* algorithm:

- Compute the WLS/GLS solution $\hat{\beta}$ for $\tilde{y} = \tilde{X}\beta + \tilde{\epsilon}$, $\epsilon \sim N(0, \Sigma)$:

$$\hat{\beta} = (\tilde{X}^T \Sigma^{-1} \tilde{X})^{-1} \tilde{X}^T \Sigma^{-1} \tilde{y}$$

where $\tilde{y} = [K(y) - \mu(\hat{\theta}^{(j)})]$, $\tilde{X} = \nabla \mu(\hat{\theta}^{(j)})^T$, and $\Sigma = \Sigma^{-1}(\hat{\theta}^{(j)})$;

- Let $\hat{\theta}^{(j+1)} = \hat{\theta}^{(j)} + \hat{\beta}$;
- Repeat until converged.

Slide 8

Application to Generalized Linear Models (GLM's)

Examples:

- Loglinear (Multinomial and Poisson) models for tables of counts
- Poisson regression models
- Logistic and probit regression models
- Normal linear regression

The essential assumptions are:

$$L_n(\theta) = G(y) e^{B(\theta) + \gamma^T(\theta)y}$$

$$\mu(\theta) = E_\theta[Y] = \begin{bmatrix} q(x_1^T \theta) \\ q(x_2^T \theta) \\ \vdots \\ q(x_n^T \theta) \end{bmatrix} \equiv q(X\theta)$$

where X is a model matrix with rows $x_1^T, x_2^T, \dots, x_n^T$, and $q^{-1}(\cdot)$ is called the *link function* for the model.

Example: Logistic Regression

We assume

$$y_i|x_i \sim \text{Bin}\left(1, \frac{e^{x_i^T \theta}}{1 + e^{x_i^T \theta}}\right) \quad (y_i \in \{0, 1\})$$

Then

$$\begin{aligned} L_n(\theta) &= \prod_{i=1}^n \left(\frac{e^{x_i^T \theta y_i}}{1 + e^{x_i^T \theta}} \right) = \left(\prod_{i=1}^n \frac{1}{1 + e^{x_i^T \theta}} \right) e^{\sum_{i=1}^n x_i^T \theta y_i} \\ &= e^{B(\theta) + y^T X \theta} \end{aligned}$$

Slide 9

Since $\gamma(\theta) = X\theta$ is linear in θ , Newton's method and Scoring will be the same.

Also note that

$$\mu(\theta) = E[y] = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix} = \begin{bmatrix} q(x_1^T \theta) \\ q(x_2^T \theta) \\ \vdots \\ q(x_n^T \theta) \end{bmatrix} \equiv q(X\theta)$$

where

$$q(t) = \frac{e^t}{1 + e^t} \quad \text{and so} \quad q^{-1}(p) = \log \frac{p}{1-p} = \text{logit}(p)$$

Fisher Scoring for GLM's

$$\begin{aligned} \nabla \mu(\theta)^T &= \left[\frac{\partial q(x_i^T \theta)}{\partial \theta_j} \right] = [q'(x_i^T \theta) x_{ij}] = \begin{bmatrix} q'(x_1^T \theta) & 0 & \cdots & 0 \\ 0 & q'(x_2^T \theta) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & q'(x_n^T \theta) \end{bmatrix} X \\ &\equiv Q'(X\theta)X \end{aligned}$$

Slide 10

so that

$$\begin{aligned} \nabla \ell_n(\theta) &= \nabla \gamma(\theta)[K(y) - \gamma(\theta)] = \nabla \mu(\theta) \Sigma^{-1}(\theta)[y - \mu(\theta)] \\ &= [Q'(X\theta)X]^T D^{-1}[y - \mu(\theta)] = X^T [Q'(X\theta)]^T D^{-1}[y - \mu(\theta)] \end{aligned}$$

where $D_{n \times n}$ is the diagonal matrix with diagonal elements $d_{ii} = \sigma^2 = \text{Var}(y_i)$.

Also,

$$\begin{aligned} I_n(\theta) &= \nabla \mu(\theta) \Sigma^{-1} \nabla \mu(\theta)^T = [Q'(X\theta)X]^T D^{-1} [Q'(X\theta)X] \\ &= X^T [Q'(X\theta)]^T D^{-1} [Q'(X\theta)] X \end{aligned}$$

Slide 11

This leads to an IRLS algorithm that operates almost directly on X and $\mu(\theta) = q(X\theta)$:

$$\hat{\theta}^{(j+1)} - \hat{\theta}^{(j)} = (\tilde{X}^T D^{-1} \tilde{X})^{-1} \tilde{X}^T D^{-1} \tilde{y}$$

where $\tilde{y} = y - \mu(\hat{\theta}^{(j)})$ and $\tilde{X} = Q'(X\theta^{(j)})X$.

Example: Logistic Regression (cont'd)

$$\mu_i(\theta) = E[y_i] = q(x_i^T \theta) = e^{x_i^T \theta} / (1 + e^{x_i^T \theta}) \equiv p_i$$

$$\sigma_i^2 = \text{Var}(y_i) = p_i(1 - p_i)$$

$$q'(t) = \left[\frac{e^t}{1 + e^t} \right]' = \frac{1}{(1 + e^t)^2} = [1 - q(t)]^2$$

What do $\nabla \ell_n(\theta)$ and $\nabla^2 \ell_n(\theta)$ look like in this case?

Slide 12

Application to Nonlinear Least Squares

Basic assumptions:

$$Y_i \stackrel{\text{indep}}{\sim} N(\mu_i(\phi), \sigma^2 / w_i)$$

$$\mu_i(\phi) = q(X_i; \phi)$$

where the function $q(\cdot)$, the design matrix X with rows x_i^T and the weights w_i are all known in advance. We wish to estimate $\theta = (\phi, \sigma^2)$

Examples:

- *Exponential model:* $q(x; \alpha, \beta) = \alpha e^{\beta x}$; $\phi = (\alpha, \beta)$.
- *Logistic model:* $q(x; \alpha, \beta, \gamma) = \alpha / (1 + \gamma e^{-\beta x})$; $\phi = (\alpha, \beta, \gamma)$.
- *Gompertz model:* $q(x; \alpha, \beta, \gamma) = \alpha e^{-\gamma e^{-\beta x}}$; $\phi = (\alpha, \beta, \gamma)$.

Slide 13

A Sketch of Fisher Scoring

Since $L_n(\theta)$ is a normal likelihood,

$$\ell_n(\theta) = \frac{-1}{n} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n w_i [y_i - \mu_i(\phi)]^2$$

and so

$$\begin{aligned} \nabla \ell_n(\theta) &= \begin{bmatrix} \frac{1}{\sigma^2} \sum_{i=1}^n w_i [y_i - \mu_i(\phi)] \nabla \mu_i(\phi) \\ \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n w_i [y_i - \mu_i(\phi)]^2 \end{bmatrix} \\ \nabla^2 \ell_n(\theta) &= \begin{bmatrix} \frac{1}{\sigma^2} \sum_{i=1}^n w_i \nabla \mu_i(\phi) \nabla \mu_i(\phi)^T & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix} \end{aligned}$$

where the matrices have been partitioned into parts relevant to ϕ and to σ^2 .

Scoring yields

$$\begin{aligned} \hat{\phi}^{(j+1)} &= \hat{\phi}^{(j)} + \left[\sum_{i=1}^n w_i \nabla \mu_i(\phi^{(j)}) \nabla \mu_i(\phi^{(j)})^T \right]^{-1} \left[\sum_{i=1}^n w_i [y_i - \mu_i(\phi^{(j)})] \nabla \mu_i(\phi^{(j)}) \right] \\ \hat{\sigma}^{2(j+1)} &= \hat{\sigma}^{2(j)} - \hat{\sigma}^{2(j)} + \frac{1}{n} \sum_{i=1}^n w_i [y_i - \mu_i(\phi^{(j)})]^2 \end{aligned}$$

Slide 14

Application to Robust Regression

Main idea:

$$E[y_i | x_i] = \mu_i(\phi) = x_i^T \phi, \quad y_i \stackrel{\text{indep}}{\sim} p(y_i | x_i, \phi, \sigma)$$

Least Squares: Minimize $S(\phi) = \sum_{i=1}^n \left(\frac{y_i - \mu_i(\phi)}{\sigma} \right)^2$

Robust Regression: Minimize $S(\phi) = \sum_{i=1}^n \rho \left(\frac{y_i - \mu_i(\phi)}{\sigma} \right)$

Examples:

$$\begin{aligned} &\bullet \rho(t) = t^2 \\ &\bullet \rho(t) = |t| \\ &\bullet \rho(t) = \begin{cases} t^2/2, & |t| < k \\ k|t| - k^2/2, & |t| \geq k \end{cases} \\ &\bullet \rho(t) = \begin{cases} t^2 & |t| < k \\ k^2 & |t| \geq k \end{cases} \\ &\bullet \rho(t) = \log \cosh^2(t) \dots \end{aligned}$$

Typically $\rho(t)$ is *symmetric, even*, and has a *unique antimode* $\rho(0) = 0$.

Slide 15

Two common approaches to estimating robust regression models

- **Scoring Approach:** Replace the model

$$X_i \sim \frac{c}{\sigma} e^{-\frac{1}{2} \left(\frac{y_i - \mu_i}{\sigma} \right)^2}, \quad c = \frac{1}{\sqrt{2\pi}}$$

with the model

$$X_i \sim \frac{c}{\sigma} e^{-\rho \left(\frac{y_i - \mu_i}{\sigma} \right)}, \quad c^{-1} = \int_{-\infty}^{\infty} e^{-\rho(y)} dy$$

Apply the Scoring idea. This yields an IRLS algorithm like the one for nonlinear least-squares.

Slide 16

- **Iterative weighting with influence function:** Observe that if $S(\phi) = \sum_{i=1}^n \rho \left(\frac{y_i - \mu_i(\phi)}{\sigma} \right)$, then

$$\begin{aligned} \nabla S(\phi) &= \sum_{i=1}^n \rho' \left(\frac{y_i - \mu_i}{\sigma} \right) X_i / \sigma = \sum_{i=1}^n w_i \left(\frac{y_i - \mu_i}{\sigma} \right) X_i \\ &= \frac{1}{\sigma} X^T W (y - X\phi) \end{aligned}$$

where W is a diagonal matrix with diagonal elements

$$w_i = \rho' \left(\frac{y_i - \mu_i}{\sigma} \right) \left/ \left(\frac{y_i - \mu_i}{\sigma} \right) \right.$$

w_i gives values of the “influence function” at each i . This leads to another IRLS-like procedure:

1. Compute $\hat{\sigma}$ as a robust estimate of the residual standard deviation σ (for example, $2\text{IQR} \approx 3\hat{\sigma}$, or take $\hat{\sigma} = \text{med}|resid - \text{med}(resid)|/0.6745$).
2. Use this $\hat{\sigma}$ to calculate W , and obtain $\phi^{(j+1)}$ using WLS.