MATH3701 Statistical Modelling
School of Mathematics, University of Leeds

Supplementary Proof: Derivation of IRLS for GLMs

This note derives the Iteratively Reweighted Least Squares (IRLS) algorithm for maximum likelihood estimation in a generalised linear model. The result is used in Chapter 4 of the notes.

Setup

Consider a GLM with independent observations $y_1, \ldots, y_n$ where y_i has an exponential family distribution with canonical parameter θ_i and known dispersion ϕ. The three model components are:

- **Random:** $\mathbb{E}[Y_i] = \mu_i = b'(\theta_i)$, $\mathrm{Var}[Y_i] = \phi\, b''(\theta_i)$.
- **Systematic:** linear predictor $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$, where $\mathbf{x}_i$ is the $p \times 1$ covariate vector for observation i.
- **Link:** $\eta_i = g(\mu_i)$, so $\mu_i = g^{-1}(\eta_i)$.

The log-likelihood and score equations

The log-likelihood is (dropping terms not involving $\boldsymbol{\beta}$)

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

Differentiating with respect to β_j via the chain rule $\partial/\partial\beta_j = (\partial/\partial\mu_i) \cdot (\partial\mu_i/\partial\eta_i) \cdot x_{ij}$,

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi\, b''(\theta_i)} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot x_{ij}.$$

Using $b''(\theta_i) = V(\mu_i)$ (the variance function) and writing $\dot{\mu}_i = \partial\mu_i/\partial\eta_i = 1/g'(\mu_i)$, the score equations $\partial\ell/\partial\beta_j = 0$ become

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)\, \dot{\mu}_i}{\phi\, V(\mu_i)}\, x_{ij} = 0, \qquad j = 1, \ldots, p.$$

In matrix form, with $\mathbf{X}$ the $n \times p$ design matrix,

$$\mathbf{X}^T \mathbf{W} (\mathbf{z} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}, \tag{1}$$

where we have introduced the *working weights* and *working response* as follows.

Working weights and working response

Define the $n \times n$ diagonal weight matrix

$$\mathbf{W} = \operatorname{diag}(w_1, \ldots, w_n), \qquad w_i = \frac{\dot{\mu}_i^2}{\phi\, V(\mu_i)},$$

and the *working (adjusted dependent) variable*

$$z_i = \eta_i + (y_i - \mu_i)\, g'(\mu_i) = \eta_i + \frac{y_i - \mu_i}{\dot{\mu}_i}.$$

Lemma 1. *With the above definitions, the score equations (1) hold.*

Proof. Observe that $w_i(z_i - \mathbf{x}_i^T\boldsymbol{\beta}) = w_i(z_i - \eta_i) = \frac{\dot{\mu}_i^2}{\phi V(\mu_i)} \cdot \frac{y_i - \mu_i}{\dot{\mu}_i} = \frac{(y_i - \mu_i)\dot{\mu}_i}{\phi V(\mu_i)}$, which is exactly the i-th summand in the score equation. □

The IRLS update

Equation (1) has the form of a *weighted normal equation*. If $\mathbf{W}$ and $\mathbf{z}$ were constants (not depending on $\boldsymbol{\beta}$), the solution would be

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z},$$

the weighted least squares estimator of $\boldsymbol{\beta}$ in the regression of $\mathbf{z}$ on $\mathbf{X}$ with weights $\mathbf{W}$.

However, both w_i and z_i depend on $\mu_i = g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta})$ and hence on $\boldsymbol{\beta}$. The IRLS algorithm resolves this by iterating:

1. Choose initial values $\hat{\mu}_i^{(0)}$ (e.g. $\hat{\mu}_i^{(0)} = y_i$ with small adjustment if on boundary).
2. At iteration t, compute $\eta_i^{(t)} = g(\hat{\mu}_i^{(t)})$, weights $w_i^{(t)}$, and working responses $z_i^{(t)}$.
3. Update the parameter estimate:

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

4. Set $\hat{\mu}_i^{(t+1)} = g^{-1}(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}^{(t+1)})$ and return to step 2.
5. Repeat until $\|\hat{\boldsymbol{\beta}}^{(t+1)} - \hat{\boldsymbol{\beta}}^{(t)}\| < \varepsilon$ for some tolerance $\varepsilon > 0$.

Under regularity conditions, IRLS converges to the MLE $\hat{\boldsymbol{\beta}}$ and is equivalent to Fisher scoring.

Asymptotic distribution of $\hat{\boldsymbol{\beta}}$

At convergence, the MLE $\hat{\boldsymbol{\beta}}$ satisfies, asymptotically,

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}_p\big(\boldsymbol{\beta},\, \mathcal{J}(\boldsymbol{\beta})^{-1}\big)\,,$$

where $\mathcal{J}(\boldsymbol{\beta}) = \mathbf{X}^T\mathbf{W}\mathbf{X}/\phi$ is the Fisher information matrix evaluated at the true parameter. Standard errors for $\hat{\beta}_j$ are the square roots of the diagonal elements of $(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}$, with $\hat{\mathbf{W}}$ evaluated at $\hat{\boldsymbol{\beta}}$.

Remark 1. The IRLS algorithm is implemented in `R`'s `glm` function. Each call to `glm` runs IRLS internally; the number of IRLS iterations is reported in `summary(fit)$iter`.