Lecture 8: GLM Estimation — Maximum Likelihood

MATH3823 Generalised Linear Models

Richard P Mann

MATH3823 Generalised Linear Models

Reading

Course notes: Chapter 4, Sections 4.1–4.2

`www.richardpmann.com/MATH3823`

The Estimation Problem

Given:

- Observations $y_1, \dots, y_n$
- Covariates $\mathbf{x}_1, \dots, \mathbf{x}_n$
- A GLM specification (distribution, link)

Goal: Estimate the parameter vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$

Method: Maximum likelihood estimation (MLE)

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} L(\boldsymbol{\beta}; \mathbf{y})$$

The Log-Likelihood Function

For independent observations from exponential family:

$$L(\boldsymbol{\beta}; \mathbf{y}, \phi) = \prod_{i=1}^{n} f(y_i; \theta_i, \phi)$$

Log-likelihood:

$$\ell(\boldsymbol{\beta}; \mathbf{y}, \phi) = \sum_{i=1}^{n} \left\{ \frac{y_i \theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right\}$$

Key: θ_i depends on $\boldsymbol{\beta}$ through:

$$\theta_i \leftarrow \mu_i = g^{-1}(\eta_i) = g^{-1}(\mathbf{x}_i' \boldsymbol{\beta})$$

The Simple Case: i.i.d. Observations

If all $\theta_i = \theta$ (same for all observations):

$$\ell(\theta; \mathbf{y}, \phi) = \frac{n(\bar{y}\theta - b(\theta))}{\phi} + \text{const}$$

Score equation:

$$\frac{\partial \ell}{\partial \theta} = \frac{n(\bar{y} - b'(\theta))}{\phi} = 0$$

Solution:

$$\boxed{b'(\hat{\theta}) = \bar{y}}$$

Since $\mathbb{E}[Y] = b'(\theta)$, the MLE satisfies $\hat{\mu} = \bar{y}$.

Example: Poisson i.i.d. Case

For $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$:

- $b(\theta) = e^\theta$, so $b'(\theta) = e^\theta$
- Score equation: $e^{\hat{\theta}} = \bar{y}$
- Solution: $\hat{\theta} = \log \bar{y}$
- Therefore: $\hat{\lambda} = \bar{y}$

The MLE is the sample mean!

This makes intuitive sense: best estimate of the rate is the observed average.

The General Case: Non-identical θ_i

When θ_i varies with covariates:

$$\ell(\boldsymbol{\beta}; \mathbf{y}, \phi) = \sum_{i=1}^{n} \frac{y_i \theta_i(\boldsymbol{\beta}) - b(\theta_i(\boldsymbol{\beta}))}{\phi} + \text{const}$$

Problem: There is no closed-form solution in general.

Exception: Normal linear model has

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

For other GLMs: Need iterative methods (Newton-Raphson, Fisher Scoring).

The Score Function

Definition: The score function is the gradient of the log-likelihood:

$$\mathbf{U}(\boldsymbol{\beta}) = \frac{\partial \ell}{\partial \boldsymbol{\beta}} = \begin{pmatrix} \frac{\partial \ell}{\partial \beta_1} \\ \vdots \\ \frac{\partial \ell}{\partial \beta_p} \end{pmatrix}$$

Properties:

- $\mathbb{E}[\mathbf{U}(\boldsymbol{\beta})] = \mathbf{0}$ at true parameter value
- MLE satisfies $\mathbf{U}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$

Fisher Information

Observed Fisher information:

$$\mathbf{I}(\boldsymbol{\beta}) = -\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \quad (p \times p \text{ matrix})$$

Expected Fisher information:

$$\mathbf{J}(\boldsymbol{\beta}) = \mathbb{E}[\mathbf{I}(\boldsymbol{\beta})] = \mathbb{E}\left[-\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right]$$

Key result:

$$\mathbf{J}(\boldsymbol{\beta}) = \mathbb{E}[\mathbf{U}(\boldsymbol{\beta})\mathbf{U}(\boldsymbol{\beta})'] = \mathrm{Var}[\mathbf{U}(\boldsymbol{\beta})]$$

Newton-Raphson Method

Iterative algorithm to find $\hat{\beta}$:

Starting from $\boldsymbol{\beta}^{(0)}$, update:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \mathbf{I}^{-1}(\boldsymbol{\beta}^{(t)})\mathbf{U}(\boldsymbol{\beta}^{(t)})$$

Intuition:

- $\mathbf{U}(\boldsymbol{\beta})$: direction of steepest ascent
- $\mathbf{I}^{-1}(\boldsymbol{\beta})$: adjusts step size based on curvature

Convergence: Stop when $\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\| < \epsilon$.

Fisher Scoring

Replace observed with expected information:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \mathbf{J}^{-1}(\boldsymbol{\beta}^{(t)})\mathbf{U}(\boldsymbol{\beta}^{(t)})$$

Advantages over Newton-Raphson:

- $\mathbf{J}$ is guaranteed positive definite
- More stable convergence
- Equivalent to iteratively reweighted least squares (IRLS)

For canonical links: Newton-Raphson and Fisher Scoring are identical.

Asymptotic Properties of the MLE

Proposition

Under regularity conditions, as $n \to \infty$:

1. $\hat{\boldsymbol{\beta}}$ is consistent: $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$
2. $\hat{\boldsymbol{\beta}}$ is asymptotically unbiased: $\mathbb{E}[\hat{\boldsymbol{\beta}}] \approx \boldsymbol{\beta}$
3. $\hat{\boldsymbol{\beta}}$ is asymptotically normal:
$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} \mathcal{N}_p(\boldsymbol{\beta}, \mathbf{J}^{-1}(\boldsymbol{\beta}))$$

Implication: Standard errors from $\sqrt{\mathrm{diag}(\hat{\mathbf{J}}^{-1})}$

The Saturated Model

Definition: A model with as many parameters as observations ($p = n$).

Properties:

- Fits data perfectly: $\hat{\mu}_i = y_i$
- Maximum possible likelihood
- Not useful for prediction, but useful as a **benchmark**

For exponential family:

$$\hat{\theta}_i = (b')^{-1}(y_i)$$

Summary

Key points:

- MLEs maximize the log-likelihood $\ell(\boldsymbol{\beta}; \mathbf{y})$
- For i.i.d. case: $\hat{\mu} = \bar{y}$
- General case requires iterative methods
- Newton-Raphson uses observed information
- Fisher Scoring uses expected information (more stable)
- MLEs are asymptotically normal: $\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \mathbf{J}^{-1})$
- Saturated model provides a benchmark for comparison

Next lecture: Model deviance and residual analysis.