Lecture 5: GLM Theory — The Exponential Family

MATH3823 Generalised Linear Models

Richard P Mann

MATH3823 Generalised Linear Models

Reading

Course notes: Chapter 3, Sections 3.1–3.3

`www.richardpmann.com/MATH3823`

Why Generalise?

Normal linear models assume:

$$Y_i \sim \mathcal{N}(\mu_i, \sigma^2), \quad \mu_i = \mathbf{x}_i'\boldsymbol{\beta}$$

But many responses are:

- **Counts:** $Y_i \in \{0, 1, 2, \ldots\}$ (Poisson)
- **Binary:** $Y_i \in \{0, 1\}$ (Bernoulli)
- **Proportions:** $Y_i/m_i \in [0, 1]$ (Binomial)
- **Positive continuous:** $Y_i > 0$ (Gamma, Exponential)

GLMs: A unified framework for all of these.

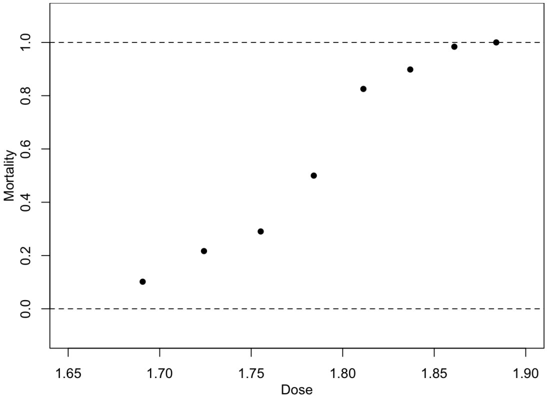
Motivating Example 1: Beetle Mortality

Data: Number killed (Y_i) out of m_i beetles at dose x_i

Model:

$$Y_i \sim \text{Binomial}(m_i, p_i)$$

Goal: Model how p_i depends on dose x_i.



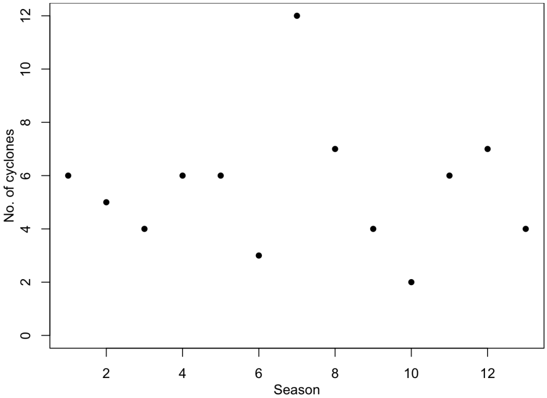
Motivating Example 2: Cyclone Counts

Data: Number of tropical cyclones (Y_i) in season i

Model:

$$Y_i \sim \text{Poisson}(\lambda_i)$$

Goal: Model how λ_i depends on climate variables.



The Three Components of a GLM

A GLM has three parts:

1. **Random Component:** Distribution of Y from the exponential family
2. **Systematic Component:** Linear predictor $\eta = \mathbf{X}\boldsymbol{\beta}$
3. **Link Function:** Connects mean to linear predictor via $\eta = g(\mu)$

Today: Focus on the **random component** — the exponential family.

Next lecture: Moments, then systematic component and link function.

The Exponential Family

Definition: A distribution belongs to the **exponential family** if its pdf/pmf can be written as:

$$\boxed{f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right\}}$$

Components:

- θ = **canonical (natural) parameter**
- $\phi > 0$ = **scale (dispersion) parameter**
- $b(\cdot)$ = function determining the distribution
- $c(\cdot,\cdot)$ = normalizing function

Key question: How do we write common distributions in this form?

Poisson Distribution

PMF:

$$f(y;\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}, \quad y = 0, 1, 2, \ldots$$

Rewrite using $\lambda^y = e^{y \log \lambda}$:

$$f(y) = \exp\{y \log \lambda - \lambda - \log y!\}$$

Matching to exponential family form:

- $\theta = \log \lambda$ (canonical parameter)
- $b(\theta) = e^\theta = \lambda$
- $\phi = 1$ (no dispersion parameter)
- $c(y, \phi) = -\log y!$

Binomial Distribution

PMF:

$$f(y; m, p) = \binom{m}{y} p^y (1-p)^{m-y}, \quad y = 0, 1, \ldots, m$$

Rewrite:

$$\begin{aligned} f(y) &= \exp\left\{ y \log p + (m-y)\log(1-p) + \log \binom{m}{y} \right\} \\ &= \exp\left\{ y \log \frac{p}{1-p} + m\log(1-p) + \log \binom{m}{y} \right\} \end{aligned}$$

The key step: factor out to get y times a function of p.

Binomial: Exponential Family Components

From the rewritten form:

$$f(y) = \exp\left\{ y \log \frac{p}{1-p} + m \log(1-p) + \log \binom{m}{y} \right\}$$

Matching to exponential family:

- $\theta = \log \frac{p}{1-p} = \text{logit}(p)$ (canonical parameter)
- $b(\theta) = m \log(1 + e^{\theta})$
- $\phi = 1$
- $c(y, \phi) = \log \binom{m}{y}$

Useful identity:

$$p = \frac{e^{\theta}}{1 + e^{\theta}}, \qquad 1 - p = \frac{1}{1 + e^{\theta}}$$

Normal Distribution

PDF:

$$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$$

Expand and rewrite:

$$\begin{aligned} f(y) &= \exp\left\{-\frac{y^2 - 2y\mu + \mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right\} \\ &= \exp\left\{\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right\} \end{aligned}$$

Group terms: $y\theta - b(\theta)$ in the numerator, rest in $c(y,\phi)$.

Normal: Exponential Family Components

Matching to exponential family form:

- $\theta = \mu$ (the canonical parameter equals the mean)
- $b(\theta) = \theta^2/2 = \mu^2/2$
- $\phi = \sigma^2$ (dispersion parameter)
- $c(y, \phi) = -\frac{y^2}{2\phi} - \frac{1}{2}\log(2\pi\phi)$

Special property of Normal:

- The canonical parameter θ is just μ
- This is why normal linear models use $\mu = \mathbf{x}'\boldsymbol{\beta}$ directly

Summary: Exponential Family Forms

Distribution	θ	$b(\theta)$	ϕ	$c(y, \phi)$
Poisson(λ)	$\log \lambda$	e^θ	1	$-\log y!$
Binomial(m, p)	logit(p)	$m \log(1 + e^\theta)$	1	$\log \binom{m}{y}$
Normal(μ, σ^2)	μ	$\theta^2/2$	σ^2	$-\frac{y^2}{2\phi} - \frac{1}{2}\log(2\pi\phi)$

Key observation: Different distributions share the same unified form.

Why the Exponential Family?

Advantages of the exponential family form:

1. **Unified theory:** Same formulas work for all members
2. **Moments from derivatives:** Mean and variance can be found from $b(\theta)$
3. **Maximum likelihood:** Score equations have nice form
4. **Sufficient statistics:** Data can be summarized efficiently

Members: Normal, Poisson, Binomial, Gamma, Exponential, Geometric, Inverse Gaussian, ...

Not members: Student's t, Uniform (with unknown bounds)

Preview: Moments from $b(\theta)$

A remarkable result:

$$\boxed{\mathbb{E}[Y] = b'(\theta), \qquad \text{Var}[Y] = b''(\theta) \cdot \phi}$$

Example (Poisson): $b(\theta) = e^\theta$

- $b'(\theta) = e^\theta = \lambda \quad \Rightarrow \mathbb{E}[Y] = \lambda$ ✓
- $b''(\theta) = e^\theta = \lambda \quad \Rightarrow \text{Var}[Y] = \lambda$ ✓

Next lecture: We'll prove this result and verify it for all distributions.

Summary

Key points:

- GLMs extend linear models to non-normal responses
- GLMs have three components: random, systematic, link
- The random component uses the exponential family
- Exponential family form: $f(y) = \exp\{(y\theta - b(\theta))/\phi + c(y, \phi)\}$
- Poisson: $\theta = \log\lambda$, $b(\theta) = e^\theta$
- Binomial: $\theta = \text{logit}(p)$, $b(\theta) = m\log(1 + e^\theta)$
- Normal: $\theta = \mu$, $b(\theta) = \theta^2/2$

Next lecture: Deriving moments from $b(\theta)$, then the systematic component and link function.