Lecture 1: Introduction to Generalised Linear Models

MATH3823 Generalised Linear Models

Richard P Mann

MATH3823 Generalised Linear Models

Reading

Course notes: Chapter 1 (Introduction)

`www.richardpmann.com/MATH3823`

Course Overview

What we will cover:

1. Revision of linear models with normal errors
2. Introduction to generalised linear models (GLMs)
3. Logistic regression models
4. Loglinear models and contingency tables

Key idea: GLMs extend normal linear models to handle responses that are *not necessarily normal.*

Prerequisites: Comfort with R programming and basic statistical computation.

The Fundamental Question

Goal: Describe how a response variable Y depends on p explanatory variables $\mathbf{x} = (x_1, x_2, \ldots, x_p)$.

Normal linear model assumption:

$$Y \mid \mathbf{x} \sim \mathcal{N}(\mu(\mathbf{x}), \sigma^2)$$

The problem: Many real-world responses are:

- Binary (success/failure, yes/no)
- Counts (number of events)
- Proportions (bounded between 0 and 1)
- Positive continuous (times, costs)

$\Rightarrow$ The normal distribution is often inappropriate.

Motivating Example: Beetle Mortality

Dose-response experiment: Beetles exposed to carbon disulphide gas.

Dose (x_i)	Beetles (m_i)	Killed (y_i)
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

Question: How does mortality rate depend on dose?

Why Linear Regression Fails

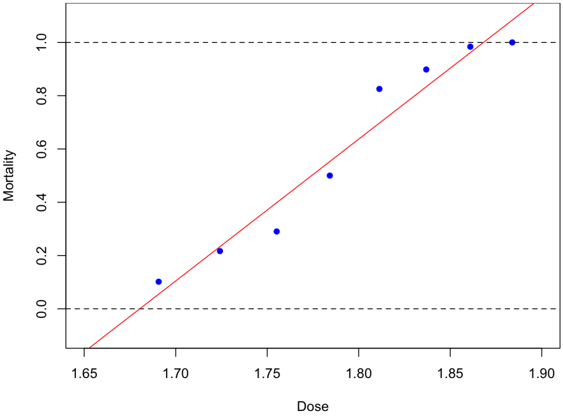
Fitted linear model:

$$\hat{p} = \hat{\alpha} + \hat{\beta}x$$

where $p = y/m$ is the mortality proportion.

Problems:

- Predicts $p > 1$ at high doses
- Predicts $p < 0$ at low doses
- Mortality is *bounded* $[0, 1]$



A Better Approach: The Logistic Model

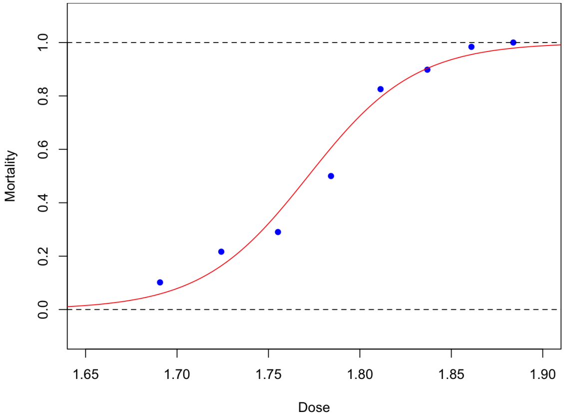
Logistic function:

$$p = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}$$

Properties:

- S-shaped (sigmoid) curve
- Always between 0 and 1
- Fitted via maximum likelihood

This is an example of a **Generalised Linear Model**.



Revision: The Linear Model

For n paired observations (x_i, y_i):

$$y_i = \alpha + \beta x_i + \epsilon_i$$

Assumptions:

- Errors ϵ_i are independent
- $\mathbb{E}[\epsilon_i] = 0$
- $\mathrm{Var}[\epsilon_i] = \sigma^2$ (constant variance)

Goal: Estimate α, β, and σ^2 from data.

Least Squares Estimation

Residual Sum of Squares:

$$\mathrm{RSS}(\alpha, \beta) = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

Minimize RSS to obtain the **least squares estimators**:

$$\hat{\beta} = \frac{s_{xy}}{s_x^2}, \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

where:

- $\bar{x}, \bar{y}$ are sample means
- $s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$ is sample covariance
- $s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$ is sample variance of x

Properties of Least Squares Estimators

Unbiasedness:

$$\mathbb{E}[\hat{\alpha}] = \alpha, \qquad \mathbb{E}[\hat{\beta}] = \beta$$

Fitted values and residuals:

- Fitted values: $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$
- Residuals: $r_i = y_i - \hat{y}_i$

Error variance estimation:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} r_i^2$$

This is an unbiased estimator: $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$.

Types of Variables

Quantitative variables:

- **Continuous:** Real-valued measurements (height, weight, time)
- **Count (discrete):** Non-negative integers (number of events)

Qualitative (categorical) variables:

- **Ordinal:** Ordered categories (mild/moderate/severe)
- **Nominal:** Unordered categories
  - **Binary/Dichotomous:** Two categories (yes/no, male/female)
  - **Polytomous:** Multiple categories (blood type, eye color)

$\Rightarrow$ Variable type determines the appropriate modeling approach.

Looking Ahead: The GLM Framework

Three components of a GLM:

1. **Random component:** Distribution of Y (exponential family)

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

2. **Systematic component:** Linear predictor

$$\eta = \sum_{j=1}^{p} \beta_j x_j = \mathbf{x}'\boldsymbol{\beta}$$

3. **Link function:** Connects mean to linear predictor

$$\eta = g(\mu), \qquad \mu = g^{-1}(\eta)$$

Normal linear regression is a special case with $g(\mu) = \mu$ (identity link).

Summary

Key points from today:

- Linear regression assumes normal errors, which is often inappropriate
- GLMs extend linear models to non-normal responses
- The beetle example shows why we need bounded response models
- Least squares estimation minimizes RSS
- Variable classification guides model choice

Next lecture: Normal linear models in detail — matrix formulation and model fitting in R.