MA50260 Statistical Modelling

Lecture 8: Introduction to GLM

Ilaria Bussoli

March 1, 2024

Motivation

While the (normal) linear regression model

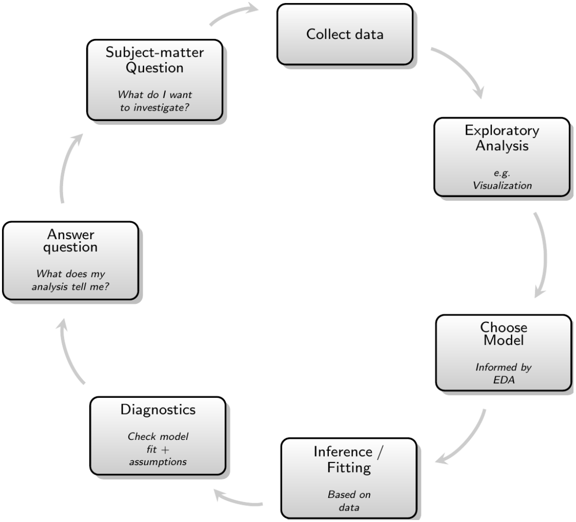
$$Y_i \sim \mathrm{Normal}(\beta_1 x_{i,1} + \cdots + \beta_p x_{i,p}\,,\, \sigma^2), \qquad i = 1, \ldots, n,$$

is very useful, it cannot handle

- Non-normality of the residuals;
- Y bounded by nature;
- Residual variance changes across observations;
- A non-linear relationship between Y_i and $x_{i,1} \ldots, x_{i,p}$.

The rest of this course will introduce generalisations / extensions of the linear model.

Philosophy of Statistical Modelling



Types of Response Variables

Let's focus on a more refined classification:

- **Continuous** $\rightarrow$ Normal
- **Count (bounded)** $\rightarrow$ Binomial
- **Count (unbounded)** $\rightarrow$ Poisson
- **Binary** $\rightarrow$ Bernoulli
- **Time-to-Event** $\rightarrow$ Exponential, Gamma
- **Categorical** $\rightarrow$ Categorical

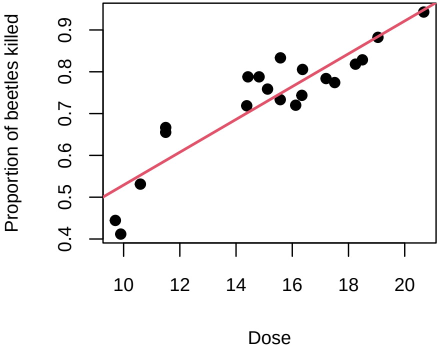
Exercise

Which distribution should we choose in the following cases?

1. Amounts of Rainfall
2. Number of hospital beds occupied
3. Wingspan of an albatross
4. Age of cancer incidence
5. Number of insurance claims

Motivating Example: Beetles

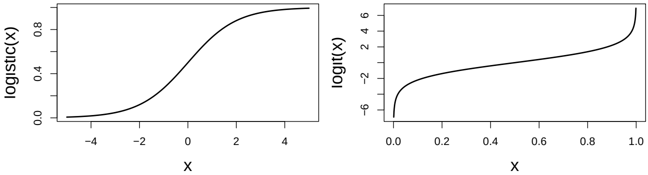
We study the effect of dose of an insecticide on beetle mortality.



Logit Transformation

We wish to map values from $(0, 1)$ to $(-\infty, \infty)$, and we use the **logit** transformation

$$\begin{aligned} \text{logistic}(x) &= \frac{\exp(x)}{1+\exp(x)}, \qquad x \in (-\infty, \infty), \\ \text{logit}(x) &= \log\left(\frac{x}{1-x}\right), \qquad x \in (0, 1). \end{aligned}$$



We could have also considered the **probit** transformation

$$\text{probit}(x) = \Phi^{-1}(x), \qquad x \in (0, 1).$$

Logistic Regression

The number of beetles killed is likely to be binomially distributed

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

We include the logit transformation in our model and define

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_1 + \beta_2 x_i.$$

Logistic Regression

The number of beetles killed is likely to be binomially distributed

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

We include the logit transformation in our model and define

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_1 + \beta_2 x_i.$$

We thus obtain a **logistic regression model** with

- The **linear predictor** $\eta_i = \beta_1 + \beta_2 x_i$.
- The **link function** $\log\left(\frac{p_i}{1-p_i}\right) = \eta_i$ between the mean and the predictor.
- The **distribution** of the observations, $Y_i \sim \mathrm{Binomial}(m_i, p_i)$.

Generalized Linear Models (GLMs)

A GLM is generally defined by three components:

- The **linear predictor** $\eta_i = \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p} = \mathbf{x}_i^{\mathrm{T}} \underline{\beta}$. This is known as the **systematic component**.

Generalized Linear Models (GLMs)

A GLM is generally defined by three components:

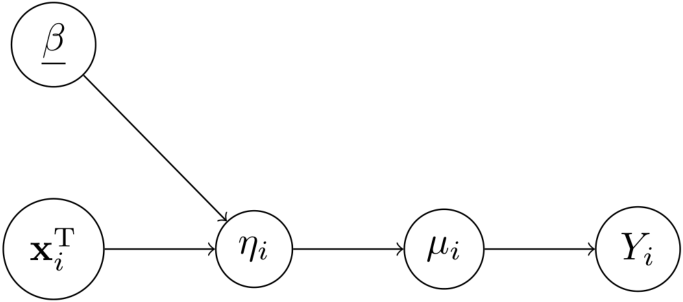
- The **linear predictor** $\eta_i = \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p} = \mathbf{x}_i^{\mathrm{T}} \underline{\beta}$. This is known as the **systematic component**.
- $g(\mu_i) = \eta_i$ – a **link function** mapping the linear predictor to the mean of the distribution.

Generalized Linear Models (GLMs)

A GLM is generally defined by three components:

- The **linear predictor** $\eta_i = \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p} = \mathbf{x}_i^{\mathrm{T}} \underline{\beta}$. This is known as the **systematic component**.
- $g(\mu_i) = \eta_i$ – a **link function** mapping the linear predictor to the mean of the distribution.
- **Probability distribution** $Y_i \sim F(\mu_i)$ from **the exponential family**. The distribution may also have a parameter ϕ. This is termed the **random component**.

Illustration



Link functions

Name	Form
identity	$\mu_i = \eta_i$
logarithmic	$\log(\mu_i) = \eta_i$
reciprocal	$1/\mu_i = \eta_i$
square	$\mu_i^2 = \eta_i$
logit	$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \eta_i$
probit	$\Phi^{-1}(\mu_i) = \eta_i$
complementary log-log	$\log[-\log(1-\mu_i)] = \eta_i$

Example: For $Y_i \sim \mathrm{Normal}(\mathbf{x}_i^{\mathrm{T}}\underline{\beta}, \sigma^2)$, we have $\mu_i = \eta_i$.

Example

For the beetle data set, we used

$$\eta_i = \log\left(\frac{p_i}{1-p_i}\right).$$

This is sometimes called the proportional odds model.

Consider the odds

$$\frac{p(x)}{1-p(x)} = \exp(\eta) = \exp(\beta_1 + \beta_2 x).$$

Suppose we compare two groups with

$$\eta = \eta(s, x) = \gamma_s + \beta_1 + \beta_2 x$$

and we thus have

$$\frac{p(s,x)}{1-p(s,x)} \div \frac{p(s',x)}{1-p(s',x)} \;=\; \exp(\gamma_s - \gamma_{s'}).$$