# Statistical Modelling in Immunology

Dr Rafael de Andrade Moral
Associate Professor of Statistics, Maynooth University

rafael.deandrademoral@mu.ie
https://rafamoral.github.io

# Outline

We will cover the following topics:

- A recap on the normal model
- Generalized linear models and extensions
- Generalized additive models for location, scale, and shape
- Mixed models

# The Normal Model

# The Normal Model: A Recap

- $Y_i$ is a response variable associated with experimental or observational unit $i$
- We *assume* it comes from a certain probability distribution with pmf/pdf $f$ and vector of parameters $\boldsymbol{\theta}$
- In general, one of the parameters in $\boldsymbol{\theta}$ is the mean of the distribution
- We also have predictors $x_i$ we are interested in studying
- We may link it to a parameter of interest, typically the mean of the distribution

# The Normal Model: A Recap

- For the normal model, we typically write for each observation $y_i$:

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

  where $\varepsilon_i \sim \mathsf{N}(0, \sigma^2)$
- Each $\beta$ coefficient represents the expected mean change in $y$ for a 1-unit increase in its associated predictor
- We can show that, from the equation above, the expected value of $Y_i$ is

$$\mathsf{E}[Y_i] = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$$

  and the variance is

$$\mathsf{Var}(Y_i) = \sigma^2$$

- Therefore, we are assuming the variance is constant

# The Normal Model: A Recap

- The error notation has its advantages...
- However, let's switch things up a bit

# Statistical Modelling

## What is Statistical Modelling?

# Statistical Modelling

## What is a Statistical Model?

# Statistical Modelling

It's all about Probability!

# Statistical Modelling

Building blocks

1. Response variable ($Y$)
2. Probability distribution
3. Parameters of interest

# Statistical Modelling

Building blocks

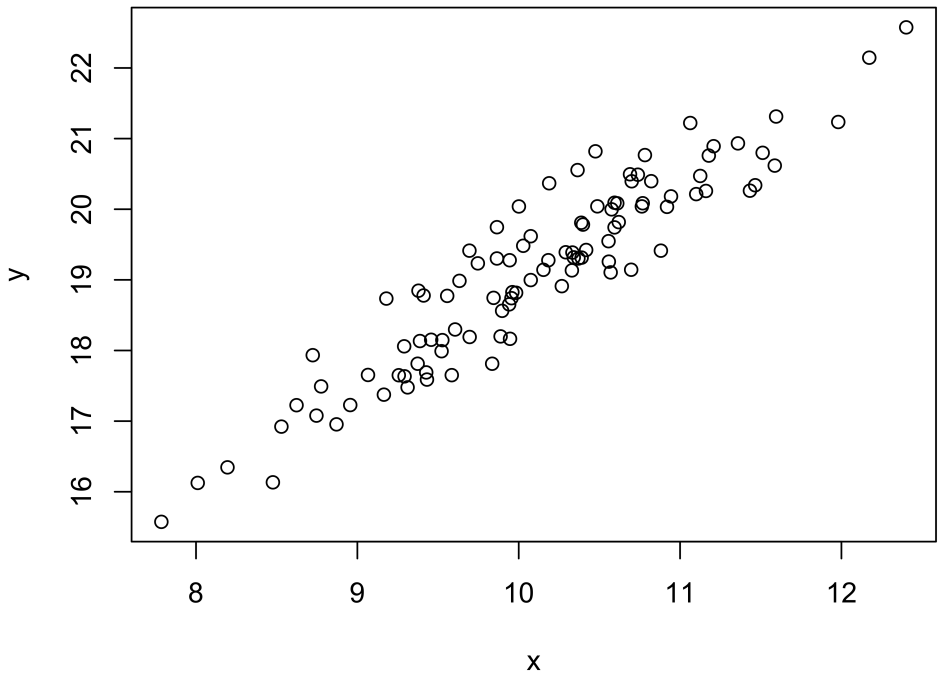
1. Response variable ($Y$)
2. Probability distribution
3. Parameters of interest $\leftarrow$ covariates / predictors

# Statistical Modelling

- $Y_i$ is a response variable associated with experimental or observational unit $i$
- We *assume* it comes from a certain probability distribution with pmf/pdf $f$ and vector of parameters $\boldsymbol{\theta}$
- Very often one of the parameters in $\boldsymbol{\theta}$ is the mean of the distribution (or a function of the mean)
  - e.g.[1] for the normal distribution, $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$, where $\mu$ is the mean of the distribution
  - e.g.[2] for the Poisson distribution, $\boldsymbol{\theta} = \mu$, where $\mu$ is the mean of the distribution
- We also have predictors $x_i$ we are interested in studying
- We may link these predictors to any parameter of interest, but we typically do it for the **mean**

# Statistical Modelling

Example



# Statistical Modelling

Example

$$\begin{aligned} Y_i &\sim \mathsf{N}(\mu_i, \sigma^2) \\ \mu_i &= \beta_0 + \beta_1 x_i \end{aligned}$$

# Statistical Modelling

Example

$$\begin{aligned} Y_i &\sim \mathsf{N}(\mu_i, \sigma^2) \\ \mu_i &= \beta_0 + \beta_1 x_i \end{aligned}$$

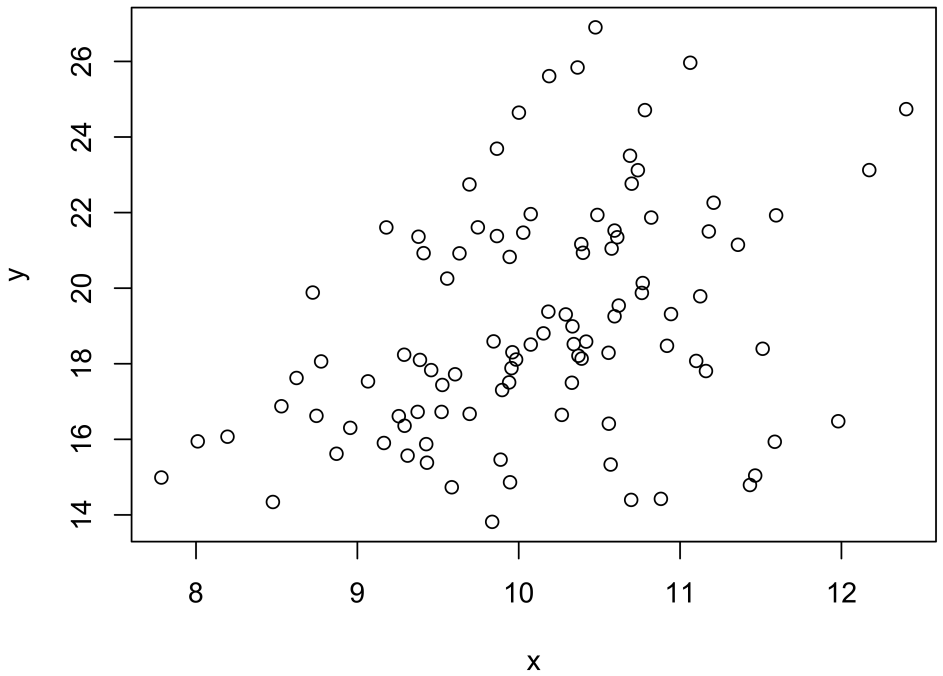
What about

- normality of residuals?
- homogeneity of variances?

This *only* makes sense for the model above!

# Statistical Modelling

What if you had



# Statistical Modelling

Maybe then you'd assume

$$\begin{aligned} Y_i &\sim \mathsf{N}(\mu_i, \sigma_i^2) \\ \mu_i &= \beta_0 + \beta_1 x_i \\ \log \sigma_i^2 &= \gamma_0 + \gamma_1 x_i \end{aligned}$$

No homogeneity of variances here!

# Statistical Modelling

$Y$ can be assumed to have *any* distribution

# Statistical Modelling

$Y$ can be assumed to have *any* distribution

Why is the normal distribution used so often then?

# GLMs

# A Brief History of GLMs

- Multiple linear regression: a normal model with the identity link (Legendre, Gauss, Galton, 19th century)
- Analysis of variance (ANOVA): a normal model with the identity link (Fisher, 1918)
- Analysis of dilution assays: a binomial model with the complementary log-log link (Fisher, 1922)
- The exponential family class of distributions (Fisher, 1934)
- Probit analysis: a binomial distribution with the probit link (Bliss, 1935)
- Logistic regression: a binomial distribution with the logit link (Berkson, 1944; Dyke and Patterson, 1952)
- Item analysis: a Bernoulli distribution with the logit link (Rasch, 1960)
- Log-linear models: a Poisson distribution with the log link (Birch, 1963)
- Regression for survival data: an exponential distribution with the inverse or log links (Feigl and Zelen, 1965; Zippin and Armitage, 1966; Gasser, 1967)
- Inverse polynomials: a gamma distribution with the inverse link (Nelder, 1966)

# A Brief History of GLMs

*J. R. Statist. Soc.* A,
(1972), **135**, *Part* 3, p. 370



## Generalized Linear Models

By J. A. NELDER and R. W. M. WEDDERBURN

*Rothamsted Experimental Station, Harpenden, Herts*

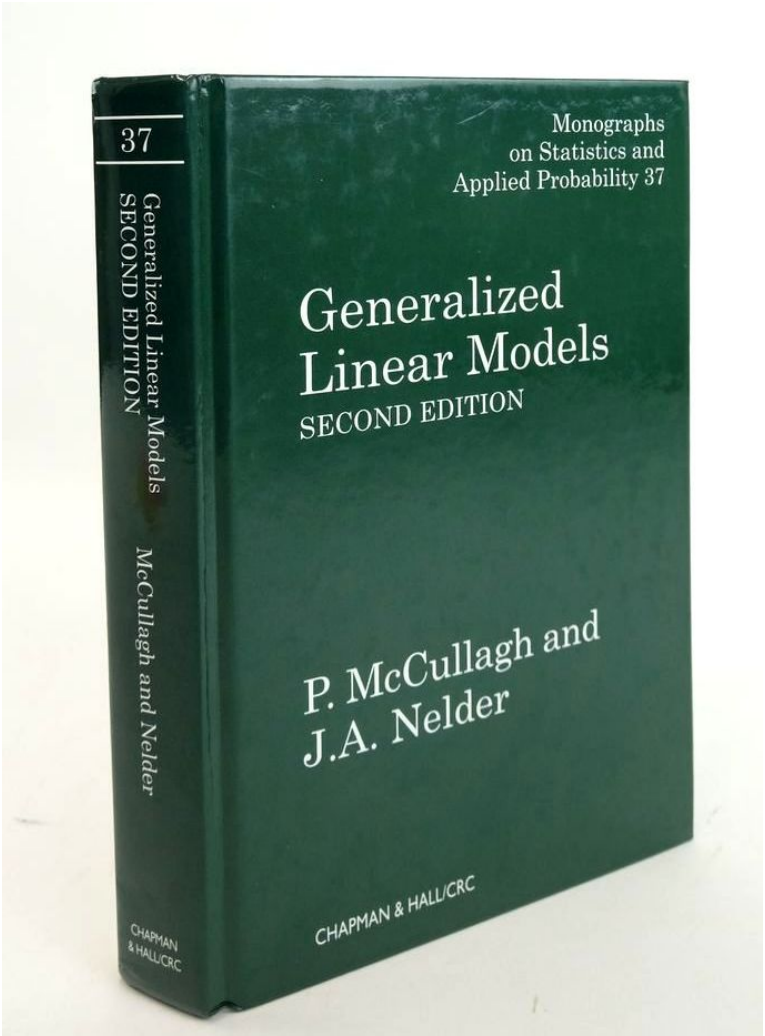
SUMMARY

The technique of iterative weighted linear regression can be used to obtain maximum likelihood estimates of the parameters with observations distributed according to some exponential family and systematic effects that can be made linear by a suitable transformation. A generalization of the analysis of variance is given for these models using log-likelihoods. These generalized linear models are illustrated by examples relating to four distributions; the Normal, Binomial (probit analysis, etc.), Poisson (contingency tables) and gamma (variance components).

The implications of the approach in designing statistics courses are discussed.

*Keywords*: ANALYSIS OF VARIANCE; CONTINGENCY TABLES; EXPONENTIAL FAMILIES; INVERSE POLYNOMIALS; LINEAR MODELS; MAXIMUM LIKELIHOOD; QUANTAL RESPONSE; REGRESSION; VARIANCE COMPONENTS; WEIGHTED LEAST SQUARES

# A Brief History of GLMs



# The Generalized Linear Model

- The generalized linear model can be defined using three components:

1. The random component: *a distribution belonging to the exponential family*
2. The systematic component: *a linear predictor*
3. The link function: *a function that links the mean to the linear predictor*

# The Generalized Linear Model

- The generalized linear model can be defined using three components:

1. The random component: *a distribution belonging to the exponential family*
2. The systematic component: *a linear predictor*
3. The link function: *a function that links the mean to the linear predictor*

- More specifically, we have *independent* random variables $Y_i, i = 1, \ldots, n$
- The linear predictor can be written as $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ where $\mathbf{X}$ is the $n \times (p+1)$ design (or model) matrix and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ is the vector of model coefficients
- The link function $g(\cdot)$ relates the mean $\mu_i$ to $\eta_i$, i.e. $g(\mu_i) = \eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$, where $\mathbf{x}_i$ is the $i-$th row of $\mathbf{X}$

# The Simple Linear Regression Model as a GLM

1. Random component: $Y_i \sim \mathsf{N}(\mu_i, \sigma^2)$
2. Systematic component: $\eta_i = \beta_0 + \beta_1 x_{1i}$
3. Link function: $g(\mu_i) = \mu_i$, the *identity* link

# The Simple Linear Regression Model as a GLM

1. Random component: $Y_i \sim \mathsf{N}(\mu_i, \sigma^2)$
2. Systematic component: $\eta_i = \beta_0 + \beta_1 x_{1i}$
3. Link function: $g(\mu_i) = \mu_i$, the *identity* link

- More simply put:

$$\begin{aligned} Y_i &\sim \mathsf{N}(\mu_i, \sigma^2) \\ \mu_i &= \beta_0 + \beta_1 x_{1i} \end{aligned}$$

# The Simple Linear Regression Model as a GLM

1. Random component: $Y_i \sim \mathsf{N}(\mu_i, \sigma^2)$
2. Systematic component: $\eta_i = \beta_0 + \beta_1 x_{1i}$
3. Link function: $g(\mu_i) = \mu_i$, the *identity* link

- More simply put:

$$\begin{aligned} Y_i &\sim \mathsf{N}(\mu_i, \sigma^2) \\ \mu_i &= \beta_0 + \beta_1 x_{1i} \end{aligned}$$

- Example: IFNG gene expression data

# Half-Normal Plots with a Simulated Envelope

- One way of empirically checking whether a GLM (or any[1] statistical model, in fact), is to use a graphical technique called *half-normal plot with a simulated envelope*
- It consists in plotting ordered residuals in absolute value versus expected order statistics of the half-normal distribution, and adding a simulated envelope based on the fitted model
- The envelope serves as a guide to what expect if the observed data are a plausible realisation of the fitted model
- In R it is implemented as the `hnp` package
- If the majority of points lie within the simulated envelope, the model can be considered to be well-fitted to the data

[1] as long as response variables can be simulated from it

# GAMs

# GAMs

- The polynomial and spline regression models can be regarded as special cases of a more general type of regression model known as a *generalized additive model* (GAM)
- Example of a normal GAM:

$$\begin{aligned} Y_i &\sim \mathsf{N}(\mu_i, \sigma^2) \\ \mu_i &= f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) \end{aligned}$$

- $f_p(\cdot)$ are *smooth* functions

# GAMLSS

# GAMLSS

- Generalized Additive Models for Location, Scale, and Shape (GAMLSS; Rigby and Stasinopoulos, 2005) are a very flexible semi-parametric modelling framework

$$Y \sim f(\mu, \sigma, \nu, \tau)$$

- Includes many distributions with up to 4 parameters
- Allows for distributional regression, i.e. all parameters can be modelled with covariates
- Spline and loess smoothing, as well as random effects allowed
- Unified framework for model diagnostics

# GAMLSS

- The `gamlss` package is the main implementation
- Extra features available through companion packages
- The package `gamlss.add` includes extra features, such as the inclusion of regression trees in the linear predictor for any parameter of interest

# Mixed Models

# Mixed Models

- Mixed models are a broad class of models that are applied to data that consist of sub-groups, or *clusters*
- Mixed models have many other names in the literature and in practice
- You may have heard/seen: *multilevel models*, *hierarchical models*, *conditional models*, *random effects models*
- In practice, these are all synonyms, but each name highlights a different part of the construction of a mixed model
- The defining feature of multilevel models is that they are *models of models*[2]

[2] Andrews, M. (2021). Doing Data Science in R: An Introduction for Social Scientists. SAGE Publishing, London, UK.

# Fixed vs Random effects

- It is important to understand the conceptual and practical differences between *fixed* versus *random* effects.
- We typically use fixed effects when we assume there is a systematic difference between levels of a factor, or a trend over time or space.
- We typically use random effects when we assume that what we observe is a random sample from a population.
  - reflect design
  - accommodate extra-variability

# Example: Gene expression data

- Expression data from 559 genes $(g_i, i = 1, \ldots, 559)$, 805 subjects $(j = 1, \ldots, 805)$, and 7 stimuli $(s_k, k = 1, \ldots, 7)$
- Random effects model:

$$\begin{aligned} Y_{ijk}|g_i &\sim \mathsf{N}(\mu_{ijk}, \sigma^2) \\ g_i &\sim \mathsf{N}(0, \sigma_g^2) \\ \mu_{ijk} &\sim \beta_0 + g_i \end{aligned}$$

# Example: Gene expression data

- Expression data from 559 genes $(g_i, i = 1, \ldots, 559)$, 805 subjects $(j = 1, \ldots, 805)$, and 7 stimuli $(s_k, k = 1, \ldots, 7)$
- Mixed effects model:

$$\begin{aligned} Y_{ijk}|g_i &\sim \mathsf{N}(\mu_{ijk}, \sigma^2) \\ g_i &\sim \mathsf{N}(0, \sigma_g^2) \\ \mu_{ijk} &\sim s_k + g_i \end{aligned}$$

# Example: Gene expression data

- Expression data from 559 genes $(g_i, i = 1, \ldots, 559)$, 805 subjects $(j = 1, \ldots, 805)$, and 7 stimuli $(s_k, k = 1, \ldots, 7)$
- Random effects per stimulus:

$$\begin{aligned} Y_{ijk}|g_i &\sim \mathsf{N}(\mu_{ijk}, \sigma^2) \\ g_{ik} &\sim \mathsf{N}(0, \sigma_k^2) \\ \mu_{ijk} &\sim s_k + g_{ik} \\ \mathsf{Corr}(g_{ik}, g_{ik'}) &= 0 \end{aligned}$$