Introduction to Generalised Linear Models for Ecologists

Dr Niamh Mimnagh

niamh@prstats.org

https://github.com/niamhmimnagh/GLME01---Introduction-to-Generalised-Linear-Models-for-Ecologists

PR STATS

Course Outline

- A recap on the normal model
- Models for binary data
- Models for binomial data
- Models for multinomial data
- Models for count data
- Models for overdispersion data
- Models for zero-inflated data
- Bayesian models
- Models for grouped data

Why Do We Model Data?

1. To understand relationships between variables
2. To predict outcomes for new or future data
3. To test hypotheses about ecological processes
4. To simplify complex systems into interpretable components

Examples:

- Predicting species richness based on elevation, temperature or rainfall.
- Estimating plant biomass from soil nitrogen or sunlight exposure.
- Modelling bird abundance based on land-use type or proximity to water.

What is a Normal Model?

- Y_i is a response variable associated with observational or experimental unit i.
- We assume it comes from a certain probability distribution with probability mass function/probability density function f and vector of parameters θ
- In general, one of the parameters in θ is the mean of the distribution
- We also have predictors x_i that we are interested in studying
- We can link these predictors to a parameter of interest, typically the mean of the distribution

What is a Normal Model?

For the normal model we typically write for each observation:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_{2i} x_{2i} + \ \ldots + \epsilon_i$$
$$\epsilon \sim N(0, \sigma^2)$$

Each β coefficient represents the expected mean change in y for a 1-unit increase in its predictor, provided all other predictors are fixed.

We can show that, from the equation:

$$y_i = \beta_0 + \beta_1 x_{1i} + \ \ldots + \beta_p x_{pi} + \epsilon_i$$

The expected value of Y_i is:

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

And the variance is:

$$Var(Y_i) = \sigma^2$$

Key Assumptions: Independence

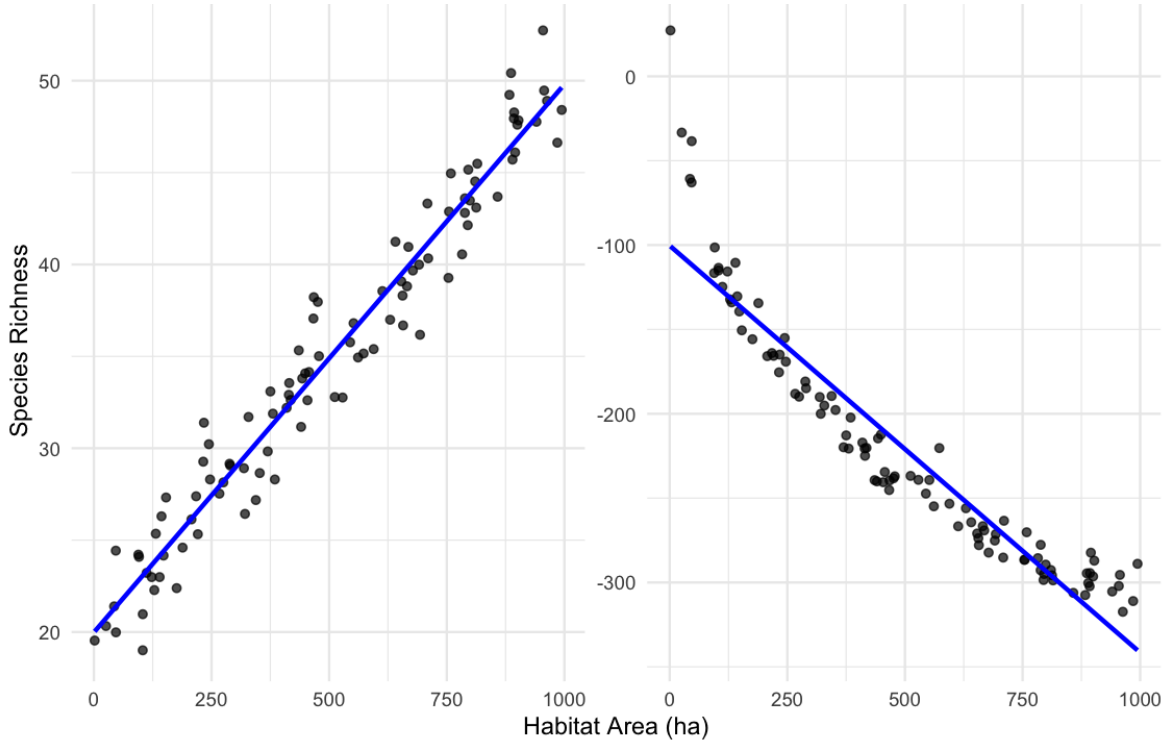
- The observations are assumed to be independent of each other. No observation should influence another.

- Independence is determined from the context of the data.

Independent	**Dependent**
One measurement taken per individual	Multiple measurements taken per individual
Measurements taken across random locations	Spatially correlated locations
Individuals in shared environments	Individuals in independent environments

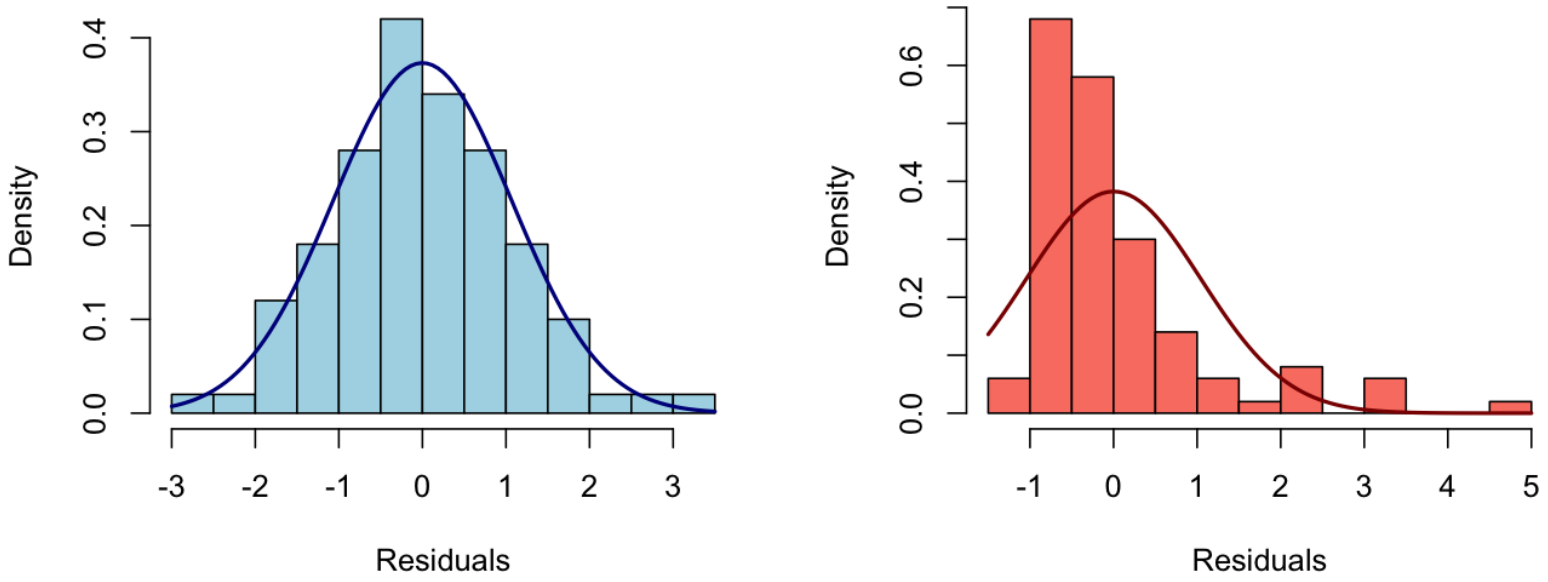
Key Assumptions: Linearity

- The relationship between the predictors and the response variable is assumed to be linear.



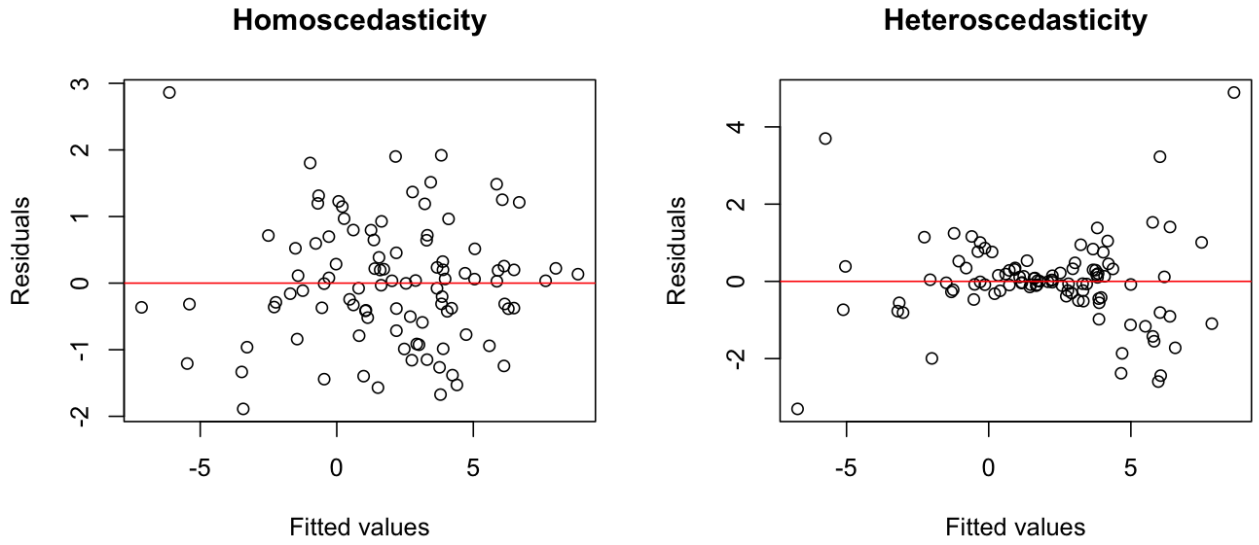
Key Assumptions: Normality of Residuals

- The residuals should be approximately normally distributed.



Key Assumptions: Homoscedasticity

- The residuals should have constant variance (the spread of the errors should be roughly equal for all predicted values).



Example: Animal Movement

- Our example investigates factors influencing average daily movement distance of tracked animals (km).
- Lets say we have 50 forest sites surveyed using GPS collars.
- Response: Average daily movement distance (km) - continuous, approximately normal.
- Predictors:
  - Habitat area (ha) - continuous
  - Canopy cover (%) - continuous

$$Movement_i = \beta_0 + \beta_1 HabitatArea_i + \beta_2 CanopyCover_i + \epsilon_i$$

Coding Demo

Limitations of Linear Models

- Linear models are appropriate for continuous, normally distributed outcomes only.
- They are not suitable for many types of data that are commonly encountered in ecological studies, including:
  - Binary data
  - Counts
  - Proportions

Model Formulation

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_{2i} x_{2i} + \ ... + \epsilon_i$$
$$\epsilon \sim N(0, \sigma^2)$$

$$Y_i \sim Normal(\mu_i, \sigma^2)$$
$$\mu_i = \beta_o + \beta_1 x_i$$

What about the linear model assumptions?

They are built into this model!

Generalised Linear Models

Y can be assumed to have any distribution

So why do we use the normal distribution so often?

A 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)
- The exponential family class of distributions (Fisher, 1934)
- Probit analysis: a binomial distribution with the probit link (Bliss, 1935)
- Logistic regression: a binomial distribution wit the logit link (Berkson, 1944; Dyke and Patterson, 1952)
- 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 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

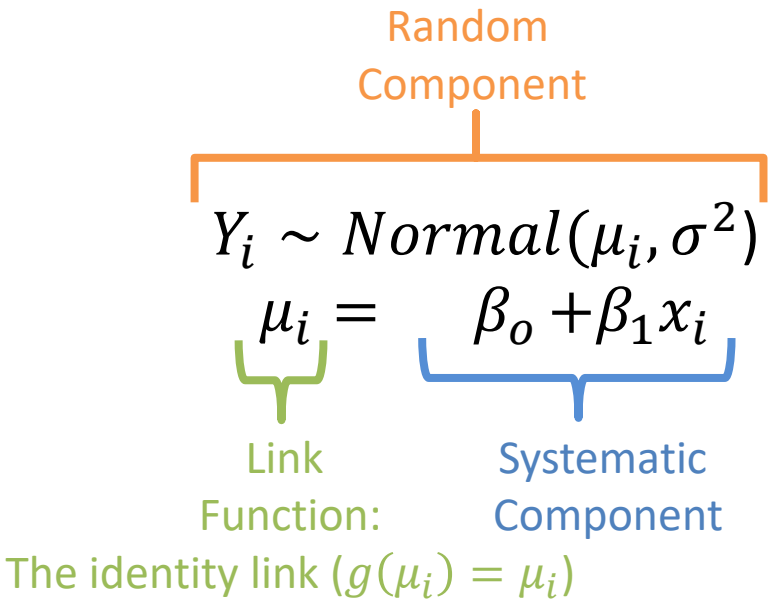
GLMs

GLMs extend linear models to accommodate non-normal response distributions.

Structure:

1. Random component: a distribution belonging to the exponential family
2. Systematic component: a linear predictor
3. Link function: a function that links the mean to the linear predictor

GLMs


Random
Component
$Y_i \sim Normal(\mu_i, \sigma^2)$
$\mu_i = \beta_o + \beta_1 x_i$
Link
Function:
The identity link ($g(\mu_i) = \mu_i$)
Systematic
Component

GLMs

We have independent random variables $Y_i, i = 1, \ldots, n$

The linear predictor can be written as $\boldsymbol{\eta} = \boldsymbol{X\beta}$ where $\boldsymbol{X}$ is the $n \times (p+1)$ design (model) matrix and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ is the vector of model coefficients

The link function $g(.)$ relates the mean μ_i to η_i, i.e. $g(\mu_i) = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$

Note:

$$\boldsymbol{X} = \mathbb{X}$$
$$\boldsymbol{\beta} = \vec{\beta} = \beta$$

Exponential Family of Distributions

- GLMs are based on distributions from the exponential family:
  - Normal
  - Binomial
  - Poisson
- Each has a specific mean-variance relationship.
- Binary response → Binomial (e.g., species presence)
- Count data → Poisson (e.g., number of individuals)

Link Functions

- A link function connects the mean of the response to the linear predictor.
  - Identity: $g(\mu) = \mu$
  - Logit: $g(\mu) = \log(\frac{\mu}{1-\mu})$
  - Log: $g(\mu) = \log(\mu)$

Choosing the Right Model

- Pick a GLM based on:
  - Nature of response variable
  - Shape of data distribution
  - Mean-variance relationship
- Plot the data first!

Example: Presence/Absence Data

$$Presence \sim Elevation + Habitat$$

- Binary response → Binomial family
- Logit link: log-odds of presence

R code:

```
glm(Presence ~ Elevation + Habitat, family = binomial,
data = data)
```

Example: Count Data

$$Abundance \sim Rainfall + Temperature$$

- Count response → Poisson family
- Log link: log of expected count

R code:

```
glm(Abundance ~ Rainfall + Temperature, family = poisson, data = data)
```

Interpreting GLM Coefficients

- Interpretation depends on the link:
  - If we use a logit link, we interpret based on a change in the log-odds (odds ratios on the response scale).
  - If we use a log link, we interpret based on a multiplicative effect (rate ratios on the response scale).
  - If we use the identity link, we interpret based on an additive effect (direct change in the mean response).

GLMs in R

Basic syntax:

```
glm(response ~ predictors, family = ..., data = ...)
```

Common families:

- `binomial(link = 'logit')`
- `poisson(link = 'log')`
- `gaussian(link = 'identity')`
- `Gamma(link='log')`

Coding Demo

Model Assumptions

- The key assumptions of the model are:
  - Observations are independent
  - Variance is constant (homoscedasticity)
  - Residuals are approximately normally distributed.
- Once a model is fitted, running model diagnostics are an important step to assess whether the model assumptions are reasonable for inference and prediction.

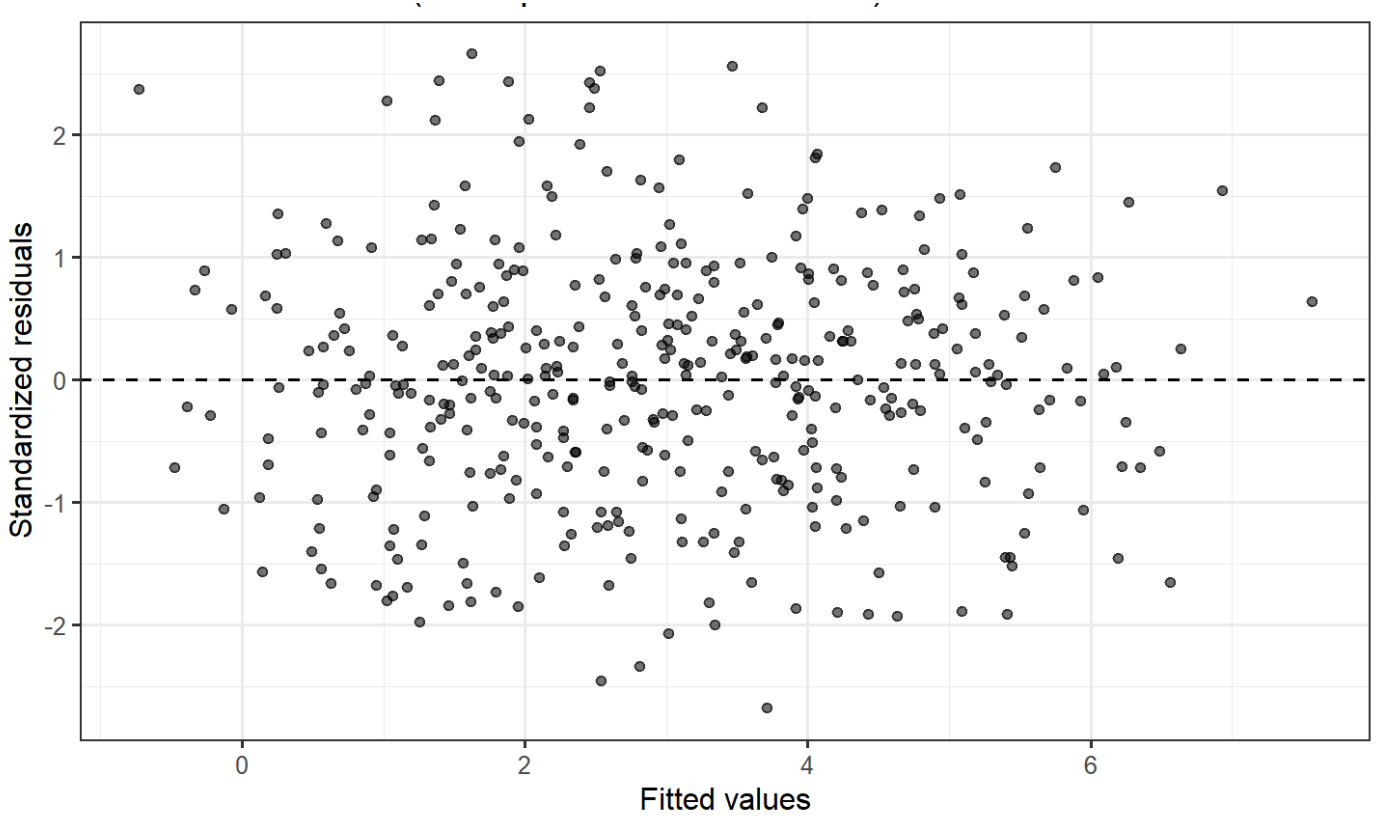
Residuals

- A residual is the difference between an observed outcome and the model's prediction for that observation; it is the model's error for that point.
- Raw residuals $e_i = y_i - \hat{y}_i$ show the basic discrepancy in the original units, which is useful for spotting patterns (curvature, heteroscedasticity) against fitted values or predictors.
- Standardised residuals $r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$ rescale by the expected residual variability and adjust for leverage h_{ii}, making them unitless and comparable across points; values around ±2 to ±3 can flag potential outliers under Normal-error assumptions.
- Raw residuals help you understand the size and direction of errors in real units, while standardised residuals provide fair, apples-to-apples comparison for outlier detection and influence checks.

Residuals vs Fitted

- This plot checks whether the mean structure is appropriately linear and whether residuals are centered at zero with roughly constant variance.
- It plots the raw or standardised residuals against the fitted values, add a horizontal reference line at zero, and (optionally) overlay a light smooth to highlight trends.
- A random cloud around zero with a roughly constant spread supports a correct mean structure and homoscedastic errors.
- Systematic curves suggest nonlinearity or omitted terms; a funnel shape indicates heteroscedasticity; bands, stripes, or group-specific patterns point to missing interactions or random effects.

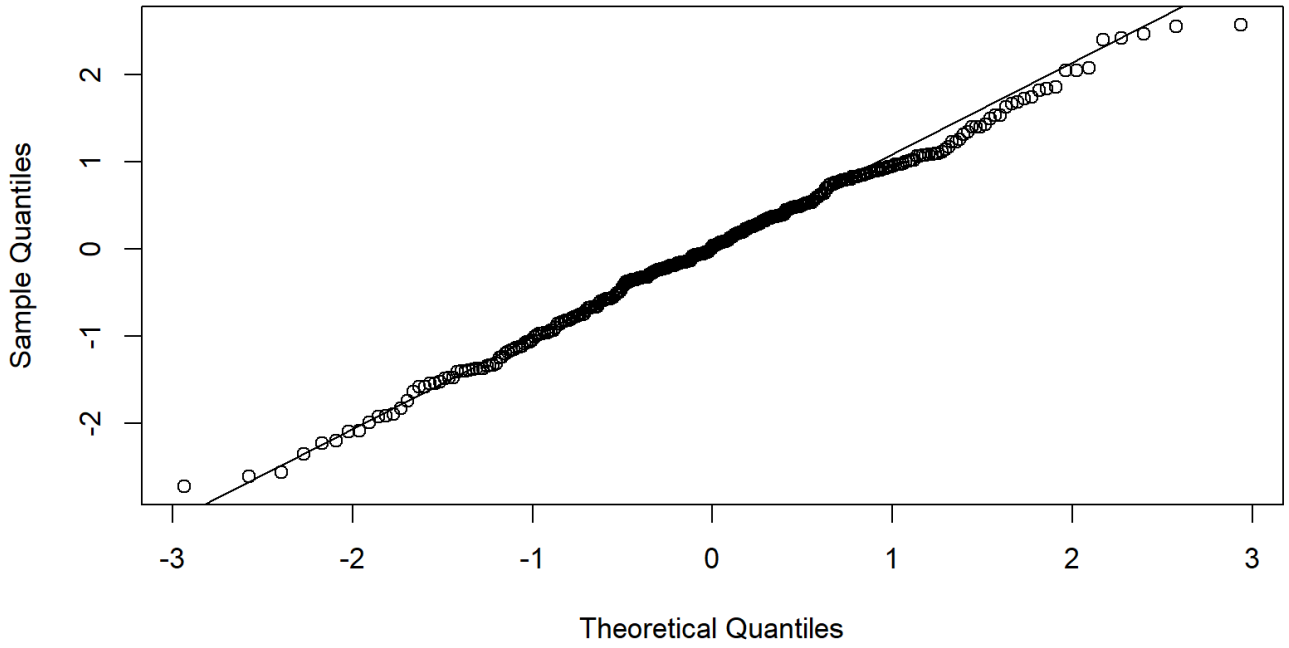
Residuals vs Fitted



Normal Q-Q Plot

- A QQ plot checks whether residuals are approximately Normal - an assumption that matters most for confidence intervals in linear models.
- We order the (standardised) residuals and plot them against theoretical Normal quantiles, add a 45° reference line, and, if possible, include a simulation envelope to gauge expected sampling variation.
- A near-straight line indicates residuals are close to Normal; an S-shape suggests heavy or light tails, and a systematic bend indicates skewness. Points peeling away at the ends often flag outliers; minor deviations are usually acceptable for large samples.

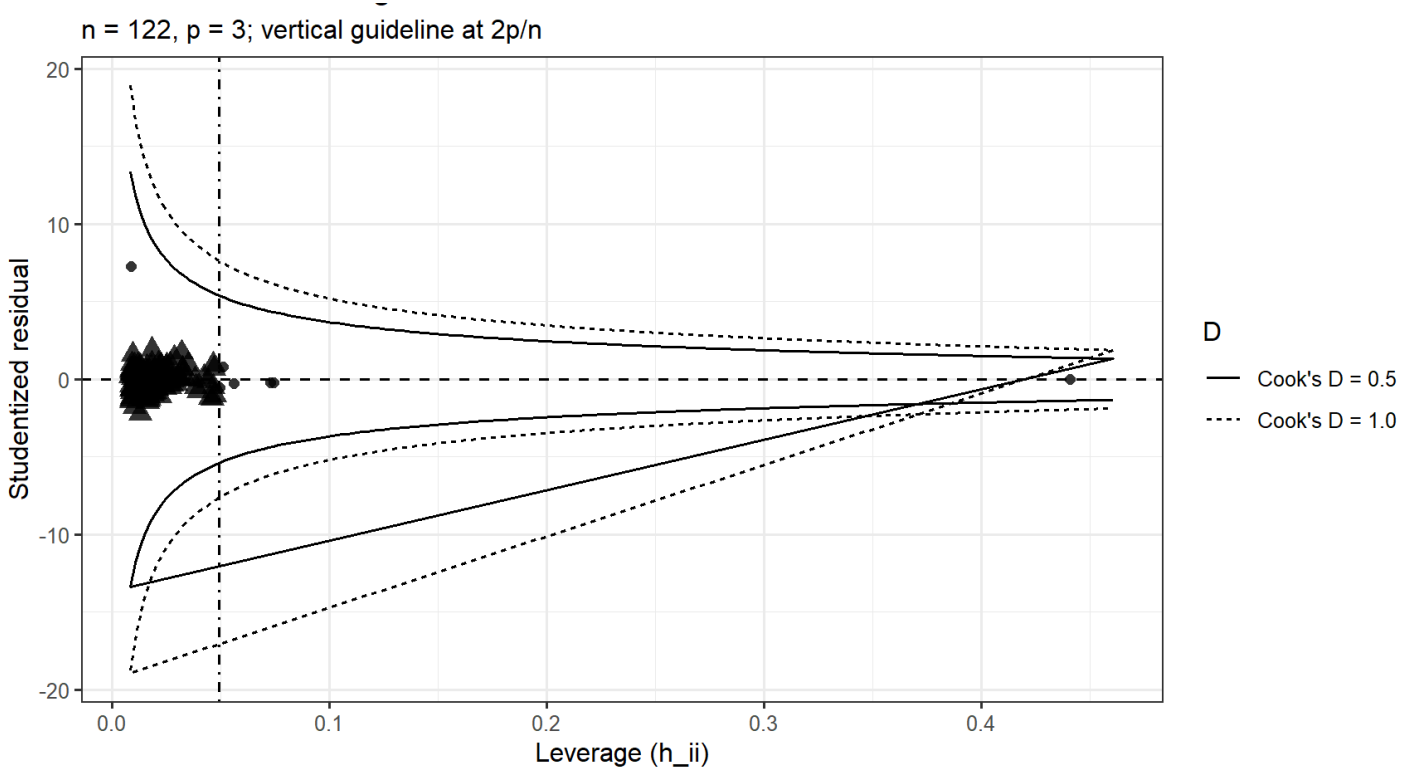
Normal Q-Q Plot



Residuals vs Leverage

- This plot checks the combination of outlier-ness in the response (via large residuals) and unusual predictor configurations (via leverage).
- Leverage h_{ii} measures how far an observation's predictor values are from the centre of the design; high-leverage points can pull the fitted line substantially even if their residual is small.
- Points with leverage roughly greater than 2p/n may strongly affect coefficients and should be investigated.
- When such points appear, check for data errors, assess whether they represent the target population, and compare fits with and without them; consider robust methods or modelling the source of extremeness.

Residuals vs Leverage



Influence

- An outlier is a point with an unusual response y given its predictors X, typically flagged by a large absolute studentised residual $| ti |$
- A high-leverage point has unusual predictor values (large h_{ii}); even with a small residual it can pull the fit toward itself.
- A point is influential if removing it changes the estimated coefficients or predictions materially.
- Treat influence as a diagnostic, not a deletion rule: check for data errors, assess whether the point is in-scope, report its impact, and consider robust models or stratification if it reflects a real subgroup.
- Useful rules of thumb: $Cook's\ Distance \gtrsim 1$

Cook's Distance

- Cook's distance measures how much all fitted coefficients change jointly when observation i is removed; it summarizes the observation's overall influence on the model.

$$D_i = \frac{r_i^2}{p} \frac{h_{ii}}{1 - h_{ii}}$$

- Large residuals combined with high leverage produces large influence.
- As a rule of thumb, values exceeding 4/n warrant investigation.
- When D_i is large, check data quality, assess whether the point is in-scope for your question, and report a sensitivity refit with and without it.

Collinearity and VIF

- Collinearity occurs when one predictor is a linear combination of others (or highly correlated with them), which inflates standard errors, makes coefficients unstable, and can even flip signs with small changes to the data.
- The Variance Inflation Factor for predictor x_j is defined as

$$VIF_j = \frac{1}{1-R_j^2},$$

- where R_j^2 comes from regressing x_j on all the other predictors; it quantifies how much the variance of β_j is inflated relative to the case of uncorrelated predictors.
- As rough guidelines, VIF>5suggests notable collinearity and VIF>10 suggests serious collinearity, but these are heuristics—use subject-matter judgment and consider sample size and study goals.

Model Comparison: R^2 for Linear Models

- In ordinary linear regression, the R^2 (coefficient of determination) tells us:
- How much of the variability in the response variable is explained by the model.

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}}$$

- The value ranges between 0 and 1.
- If $R^2 = 0.65 \rightarrow$ model explains 65% of the variation
- This is easy to interpret when using models with continuous outcomes

Practical Workflow

1. Fit model →
2. Core plots (Residuals–Fitted, Q-Q) →
3. Influence (Cook's Distance, DFFITS, DFBETAs).
4. check VIF →
5. Address issues (transform, add terms, GLS/mixed, robust SEs).
6. Report what you checked and decisions made.