Chapter 1
Tools for Assessing Goodness-of-fit of GLMs: Case Studies in Entomology

Darshana Jayakumari, John Hinde, Jochen Einbeck and Rafael A. Moral

Abstract In this chapter, we discuss the analysis of data that typically arise from entomological studies using generalized linear models. We focus on techniques that can be used to assess model goodness-of-fit, which is an important step in statistical modelling to ensure the reliability of the inferences made. Specifically, we demonstrate the utility of half-normal plots with a simulated envelope as a complementary tool for assessing model assumptions. We illustrate the concepts with two examples, one involving count responses and another involving continuous responses.

1.1 Introduction

Given a scientific hypothesis, an experiment or observational study can be carried out to collect data that may confirm or provide evidence against the said hypothesis. Statistical models represent an attempt to explain patterns of variation found in a response variable through the use of specific distributional assumptions and predictor variables. These patterns change depending on the nature of the response. A useful model should be able to capture most of the relevant variation in the data and distinguish between a true signal and noise, while at the same time maintaining parsimony.

Consider the following multiple linear regression model:

$$\begin{aligned} \mathbf{Y} &\sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{I}_n \sigma^2), \\ \boldsymbol{\mu} &= \mathbf{X}\boldsymbol{\beta}, \end{aligned} \tag{1.1}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ is the vector of responses of dimension n, $\boldsymbol{\mu}$ is the vector of means, $\mathbf{X}$ is the $n \times p$ design matrix, $\boldsymbol{\beta}$ is the vector of regression coefficients of dimension p, $\mathbf{I}_n$ is the $n \times n$ identity matrix and σ^2 is the variance parameter. This model makes three main assumptions: (i) the response variable is assumed to be normally distributed; (ii) the means are allowed to differ across the Y_is, according to the specified linear model, however their variances are assumed to be the same (σ^2 for all Y_i); and (iii) the responses are assumed to be independent (which under the assumption of a normal distribution is implied from the diagonal covariance matrix with $\mathrm{Cov}(Y_i, Y_j) = 0$, for $i \neq j$).

When fitting model (1.1) to real data, it is important to assess whether the aforementioned assumptions are met before treating any inferential results as reliable. Assumption (i) can be checked in many different ways. We may assess it through hypothesis testing using, for example, the Shapiro-Wilk test of normality based on a standardised version of the model residuals [21]. We may also carry out graphical assessments, that include quantile-quantile plots, which

Darshana Jayakumari
Maynooth University, Ireland e-mail: darshana.jayakumari.2021@mumail.ie

John Hinde
University of Galway, Galway, Ireland e-mail: john.hinde@universityofgalway.ie

Jochen Einbeck
Durham University, Durham, England e-mail: jochen.einbeck@durham.ac.uk

Rafael A. Moral
Maynooth University, Maynooth, Ireland e-mail: rafael.deandrademoral@mu.ie

will be discussed in detail in Section 1.5. Assumption (ii) can also be checked through formal hypothesis tests, such as the Bartlett test for variance homogeneity [23], and graphically, by looking at a plot of residuals versus fitted values (see Figure 1.3 for examples). Frequently, assumption (iii) will be deemed to be true or not based on the design of the experiment or observational study, without resorting to formal hypothesis testing. Tests are available, but typically they are onl;y carried out when analysing longitudinal or time-series data, which may be assumed to be correlated. For these cases, calculating the empirical auto-correlation and partial auto-correlation functions is especially helpful.

Nevertheless, all assumptions discussed above can be summarised by a single distributional assumption, which is denoted by (1.1). If the distributional assumption in (1.1) is true, then the observed data must be a plausible realisation of the estimated model $\mathrm{N}(\hat{\boldsymbol{\mu}}, \mathbf{I}_n\hat{\sigma}^2)$. In other words, theoretically we should be able to generate the observed values of the response variable we obtained in the experiment or observational study by simulating from our fitted model. Goodness-of-fit assessment methods that rely on simulation (such as half-normal plots with a simulation envelope) are based on this principle. They involve simulating multiple times from a fitted model and comparing the results obtained with the observed response or a function of the response. This is especially useful in the context of more general models (e.g. based on a wider family of probability distributions), for which assumptions (i) and (ii) do not hold (and consequently carrying out tests for normality and homogeneity of variances would be pointless in this case). Note that here we confine our attention to settings where the assumption (iii) of independent responses is plausible by virtue of the form of data collection; more general data structures may require model extensions, such as the mixed modelling framework for multiple components of variation.

In entomological studies, it is often the case that non-normal continuous data and discrete data (counts and proportions) are collected. For these types of data, model (1.1) would not be suitable for analysis, and different distributional assumptions would be required. For instance, to analyse count data, the Poisson model is one of many alternatives; to analyse proportion data, the binomial model could be used; and gamma and inverse Gaussian models are able to flexibly accommodate right-skewed data with positive support. These are all examples of generalized linear models, for which many different extensions are also available.

There is a plethora of modelling options available for the analysis of entomological data. In many cases, more than one distribution can suitably accommodate the variability in the data. In Section 1.2, we provide a general definition and overview of generalized linear models. Later, in Sections 1.3, 1.4 and 1.5, we present an overview of goodness-of-fit assessment tools and techniques. We conclude by illustrating their use with real datasets in Section 1.6. All analyses and figures in this chapter are generated using R [19].

1.2 Generalized Linear Models

The modelling of non-normal data could involve, as alternatives, distributions belonging to the exponential family (EF) of distributions. The EF includes discrete (e.g. Poisson, negative binomial, and binomial), as well as continuous (e.g. gamma and inverse Gaussian) distributions, representing flexible alternatives to the normal distribution when modelling discrete, or continuous and strictly positive, or skewed data in entomology.

The probability density function (pdf) of a random variable Y whose distribution belongs to the EF of distributions can be written, in the canonical form, as

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right\}, \tag{1.2}$$

where $b(\cdot)$ and $c(\cdot)$ are functions of the dispersion parameter ϕ, the canonical parameter θ and the data and dispersion, respectively.

The generalized linear model (GLM) is an extension to the classical linear model where the response variable is assumed to follow a distribution which belongs to the EF. By developing an inferential framework that encompassed distributions belonging to the EF, [17] allowed for fitting GLMs using a unified estimation process.

The GLM consists of three components:

1. Random component: this is the assumed distribution for the response variable, which belongs to the EF. This component is termed ‘random’ because it is a probability distribution that is used to model the variability in the data.

2. Systematic component: takes the form of a linear predictor, which consists of a linear combination of the predictor variables and unknown parameters. It may be written as

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta},$$

where $\mathbf{X}_{n\times p}$ is the design matrix and $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients.
3. Link function: this *links* the random component to the systematic component through a monotonic and differentiable function $g(\cdot)$. Typically we aim to link the parameter corresponding to the mean $\boldsymbol{\mu}$ of the distribution to the linear predictor through the transformation $\boldsymbol{\eta} = g(\boldsymbol{\mu})$. As link functions are invertible, we have that the mean $\boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta}) = g^{-1}(\mathbf{X}\boldsymbol{\beta})$ and so is determined by the linear predictor. The link function that transforms the mean $\boldsymbol{\mu}$ to the natural parameter $\boldsymbol{\theta}$ is known as the canonical link. Note that using the canonical link corresponds to specifying a linear predictor for the canonical parameter, which may, or may not, be desirable depending on the substantive questions of interest. Table 1.1 presents six commonly used EF distributions with their canonical link functions, the dispersion (or scale) parameter, and the form of the variance function (discussed in subsequent sections).

Distribution	Representation	$g(\mu)$	$V(\mu)$	ϕ
Normal (Gaussian)	$\mathrm{N}(\mu,\sigma^2)$	μ	1	σ^2
Gamma	$\mathrm{Gamma}(\mu,\alpha)$	μ^{-1}	μ^2	α^{-1}
Inverse Gaussian	$\mathrm{IG}(\mu,\sigma^2)$	μ^{-2}	μ^3	σ^2
Poisson	$\mathrm{Pois}(\mu)$	$\log(\mu)$	μ	1
Negative binomial	$\mathrm{NB}(\mu,k)$	$\log\left(\frac{\mu}{\mu+k}\right)$	$\mu\left(\frac{\mu}{k}+1\right)$	k^{-1}
Binomial	$\mathrm{Binom}(m,\pi)$	$\log\left(\frac{\mu}{m-\mu}\right)$	$\frac{\mu}{m}(m-\mu)$	1

Table 1.1 Representation, canonical link functions ($g(\mu)$), variance functions ($V(\mu)$), and dispersion parameter (ϕ) for six commonly used distributions within the generalized linear modelling framework. Here and throughout this chapter, log denotes the natural logarithm (i.e. log base e).

The estimation of GLMs is typically done using the maximum likelihood (ML) method. For the normal model, ML estimates are equivalent to the ones obtained via ordinary least squares (which aims to minimise the sum of squared differences between observed and fitted values). Under the assumed distribution, the ML estimates are the ones that maximise the likelihood function, that is the likelihood of the observed data being generated by that distribution. Let $f(y_i;\boldsymbol{\beta},\phi)$ be the probability density or mass function distribution function for observation i, where $\boldsymbol{\beta}$ is the vector of regression parameters to be estimated and ϕ is the dispersion parameter of the assumed distribution. The likelihood function for a single observation is defined as $L(\boldsymbol{\beta},\phi;y_i) = f(y_i;\boldsymbol{\beta},\phi)$. Assuming observations are independent, the overall likelihood for the full sample is given by the joint probability density or mass function

$$L(\boldsymbol{\beta},\phi;\mathbf{y}) = \prod_{i=1}^{n} f(y_i;\boldsymbol{\beta},\phi).$$

To avoid numerical problems when working with likelihood functions, it is commonplace to work with the log-likelihood $l(\boldsymbol{\beta},\phi;\mathbf{y}) = \log L(\boldsymbol{\beta},\phi;\mathbf{y}) = \sum_{i=1}^{n}\log f(y_i;\boldsymbol{\beta},\phi)$ instead. Since logarithms are monotonic functions, the parameter values that maximise $l(\cdot)$ will also maximise $L(\cdot)$. Therefore, the ML estimates will be the parameter values that maximise $l(\cdot)$, i.e.

$$\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, l(\boldsymbol{\theta};\mathbf{y}),$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top,\phi)^\top$.

The maximised log-likelihood value is used within different goodness-of-fit criteria, and in some cases it can itself be used as a goodness-of-fit measure. When comparing model fits, a higher value of the log-likelihood would indicate a better reproduction of the observed data. However, a saturated model, for instance, would reproduce every single observation; while the log-likelihood would be larger than for a less complex model, saturated models overfit the data

and are not flexible enough to generate predictions for other sets of predictor values. Therefore, selecting a model is a task that involves balancing flexibility and explainability.

The residual deviance measure is defined as the difference between the log-likelihood of the saturated model and log-likelihood of the fitted model (also called 'current' model), scaled by a factor of two, i.e., it conveys how far the model is from fully reproducing the observed data. (Strictly speaking the deviance here is the *scaled deviance* where the basic deviance is ϕ times this and for the EF gives a fitting criteria that does not involve ϕ and plays the same role as the residual sum of squares for the normal model. More recent usages tend to blur this distinction and, of course, for models where $\phi = 1$ it is not an issue, but in other models the user needs to take care as to what version if being reported.) In GLM theory, the deviance is an important measure, because by subtracting the residual (scaled) deviances between two nested models, we obtain a statistic called 'likelihood-ratio' (since the difference between (scaled) deviances is equivalent to a ratio between model likelihoods in the natural scale). It can be proven that likelihood ratios, under the null hypothesis that the simplest model is most adequate to explain the data, for a fixed or known value of the scale parameter ϕ, asymptotically follow a χ^2 distribution with the number of degrees of freedom equal to the difference between the number of estimated parameters between the models being compared.

Now we briefly present the most commonly used EF models when analysing entomological data, for which we provide examples of their probability density or mass functions in Figure 1.2.

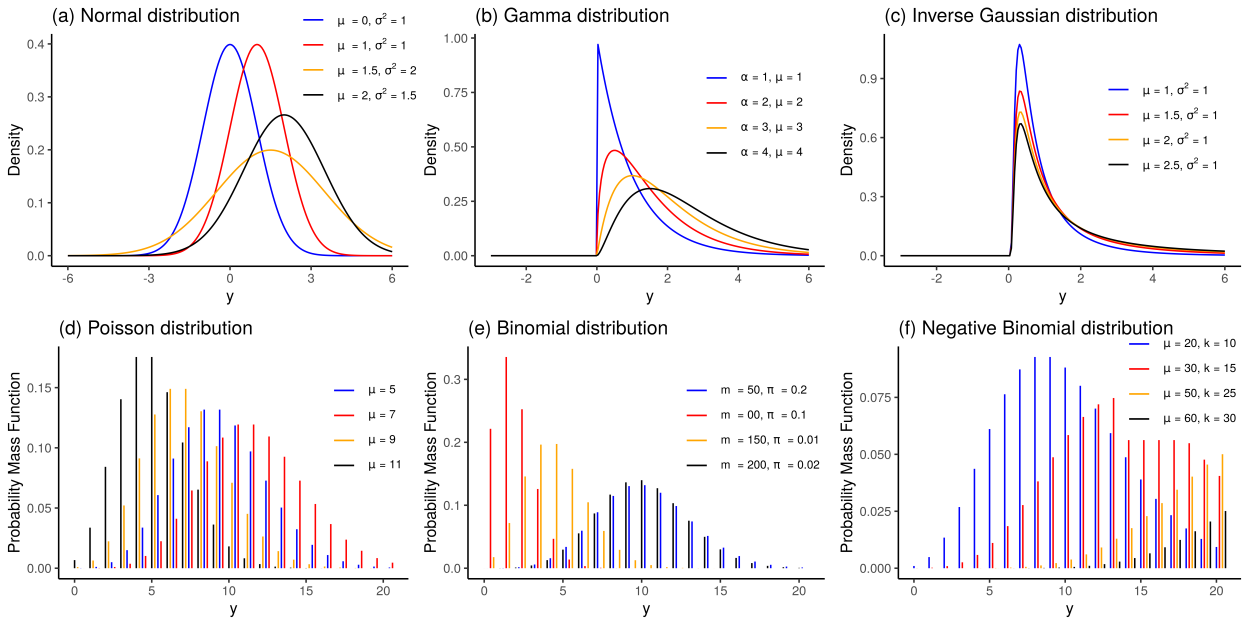

Fig. 1.1 Probability density/mass functions for six distributions belonging to the exponential family. The top row shows the density of the continuous distributions for different sets of parameter values. The bottom row shows the probability mass function of discrete distributions.

1.2.1 The Normal model

The normal distribution is the most commonly used distribution in statistics and is widely used to model real life events[28]. It is a continuous and symmetric probability distribution, with a probability density function (pdf) given by:

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, \quad -\infty < y < \infty.$$

It is also known as the 'Gaussian' distribution as it was introduced by Carl Friedrich Gauss in 1809. The distribution has two parameters (μ, σ) and is denoted by $\mathrm{N}(\mu, \sigma^2)$ where $\mu \in (-\infty, \infty)$ is the mean and $\sigma^2 > 0$ is the variance. The density function of the normal distribution resembles a bell shape and is known as the 'bell curve' and is centred around the mean μ. When $\mu = 0$ and $\sigma^2 = 1$ the distribution is known as the 'standard normal' distribution. It can be shown that approximately 68.2% of the area under the curve is contained in the interval $(\mu \pm \sigma)$; approximately 95.4% of the area under the curve is contained in the interval $(\mu \pm 2\sigma)$; while approximately 99.7% of the area under the curve is contained in the interval $(\mu \pm 3\sigma)$, see Figure 1.2.1. The effect of different means and variances can be seen in Figure 1.2(a).

The Central Limit Theorem, a key theorem in statistics, states that for a sufficiently large sample of independent and identically distributed variables, the sampling distribution of the mean is approximated by a normal distribution, no matter the shape of the population distribution. This makes it possible for other distributions to be approximated by a normal distribution, which makes it easier to solve complex problems by using the properties of the normal distribution.

Referring to Equation 1.2 and rewriting the pdf of the normal distribution we obtain:

$$f(y) = \exp\left\{\frac{y\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right\}, \quad -\infty < y < \infty;$$

which gives the canonical parameter $\theta = \mu$, the dispersion parameter $\phi = \sigma^2$, $b(\theta) = \theta^2/2$, and $c(y, \phi) = -\frac{1}{2}\left(\frac{y^2}{\phi} + \log(2\pi\phi)\right)$.

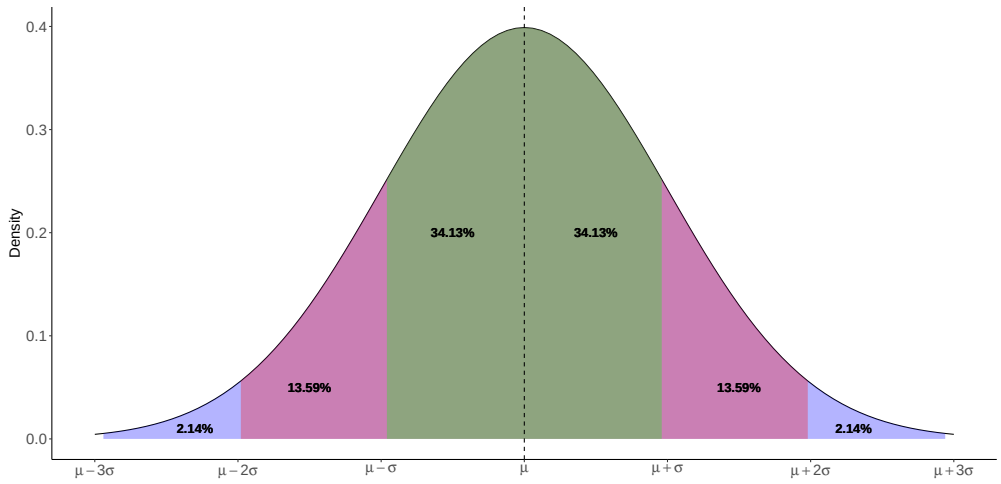

Fig. 1.2 The normal distribution pdf curve. We have that 68.2% of the area under the curve falls between $\mu \pm \sigma$, around 95.4% of the area falls between $\mu \pm 2\sigma$, and approximately 99.7% of the area under the curve falls between $\mu \pm 3\sigma$.

1.2.2 The Gamma model

The gamma distribution is generally seen in examples where the outcome is strictly positive and skewed. Real life examples of the gamma distribution include the modelling of rain fall [4], faults in the equipment maintenance [25], and insect populations [16]. It can be parameterised in different ways. It is common to use a shape parameter $\alpha > 0$ and a scale parameter $\beta > 0$, such that if $Y \sim \text{Gamma}(\alpha, \beta)$ the pdf is given by

$$f(y;\alpha,\beta) = \frac{y^{\alpha-1}e^{-\beta y}\beta^{\alpha}}{\Gamma(\alpha)}, \quad y > 0, \tag{1.3}$$

where $\Gamma(\cdot)$ is the gamma function, defined as

$$\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1}e^{-y}dy.$$

The gamma distribution may also be considered a generalised form of other distributions. For instance, when $\alpha = 1$, it is reduced to the exponential distribution with parameter β, i.e. a $\text{Gamma}(1,\beta)$ distribution is equivalent to $\text{Exponential}(\beta)$. Moreover, a $\text{Gamma}(\nu/2, 1/2)$ distribution is equivalent to a χ^2_ν distribution.

For modelling purposes a more useful parameterisation is to use the mean μ. We have that the expected value and variance are

$$\text{E}(Y) = \frac{\alpha}{\beta}, \ \text{Var}(\mathbf{Y}) = \frac{\alpha}{\beta^2}.$$

Therefore, we may reparameterise the pdf (1.3) using $\mu = \alpha/\beta$, to obtain the following exponential family pdf in canonical form:

$$f(y) = \exp\left\{\left(\frac{-y}{\mu} - \log(\mu)\right)\alpha + \alpha\log\alpha + (\alpha-1)\log y - \log(\Gamma(\alpha))\right\}, \quad y > 0.$$

It is clear from above that the canonical parameter $\theta = -1/\mu$, and the dispersion parameter $\phi = 1/\alpha$. This gives $b(\theta) = \log(-1/\theta)$, and

$$c(y,\phi) = \frac{1}{\phi}\log\left(\frac{1}{\phi}\right) + \left(\frac{1}{\phi} - 1\right)\log y + \log\left(\Gamma\left(\frac{1}{\phi}\right)\right), \quad y > 0.$$

See Figure 1.2(b) for different shapes of the gamma distribution.

1.2.3 The Inverse Gaussian model

The inverse Gaussian distribution is a continuous distribution similar to the gamma distribution, but with a sharper peak and greater skewness [24]. It has a single mode and a long tail in the density function which helps modelling data sets with extreme values. The name inverse Gaussian is related to the fact that its cumulant generating function is the inverse of the Gaussian distribution's. The pdf of an $\text{IG}(\mu,\sigma^2)$ distribution is given by

$$f(y) = \sqrt{\frac{1}{2\pi\sigma^2 y^3}}\, e^{-\frac{(y-\mu)^2}{2\mu^2\sigma^2 y}}, \quad y > 0,$$

where $\mu \in (-\infty,\infty)$ is the mean and $\phi = \sigma^2 > 0$ is the dispersion parameter. We have that $\text{E}(Y) = \mu$ and $\text{Var}(Y) = \mu^3\sigma^2$.

We may re-write the pdf of the inverse Gaussian distribution in the canonical exponential family form as

$$f(y) = \exp\left\{\left(-\frac{y}{2\mu^2} + \frac{1}{\mu}\right)\sigma^2 - \frac{1}{2}\left(\log(2\pi\sigma^2 y^3) + \frac{1}{\sigma^2 y}\right)\right\}, \quad y > 0,$$

where we identify the canonical parameter $\theta = -1/2\mu^2$, $b(\theta) = \sqrt{-2\theta}$, and

$$c(y,\phi) = \frac{-1}{2}\left(\log 2\pi y^3\phi + \frac{1}{y\phi}\right).$$

For different shapes of the inverse Gaussian distribution, see Figure 1.2(c).

1.2.4 The Poisson model

Unlike a Normal distribution which is a continuous distribution, the Poisson distribution is a discrete probability distribution that is used to model data in the form of counts for a specified interval of time (or space). Examples of data that can be modelled using a Poisson distribution include the number of eggs laid by insects over a specified period of time, or the number of insects counted per m^2.

If Y has a Poisson distribution, we may write $Y \sim \mathrm{P}(\mu)$, and write its probability mass function (pmf) is given by

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

where $\mu > 0$ is the mean parameter. A property of the Poisson distribution is that the mean is equal to the variance, i.e. $\mathrm{E}(Y) = \mathrm{Var}(Y) = \mu$, known as the 'equi-dispersion' property. A consequence of this is that the Poisson model is not able to appropriately model data for which the variance is greater than the mean (a phenomenon known as 'over-dispersion' [14]). Real entomological data often exhibits over-dispersion, and therefore extensions of the Poisson model would be more suitable for analysis, such as the quasi-Poisson, negative binomial and Poisson-normal models [7].

Re-writing the pmf in the canonical exponential family form we obtain:

$$f(y) = \exp(y\log\mu - \mu - \log(y!)), \quad y \in \{0,1,2,\ldots\}.$$

We can identify the canonical parameter $\theta = \log\mu$, as well as the dispersion parameter $\phi = 1$. Moreover, $b(\theta) = \mu$, and $c(y,\phi) = -\log(y!)$. See Figure 1.2(d) for different shapes of the Poisson pmf.

1.2.5 The Negative Binomial model

The negative binomial distribution, also known as the 'Pascal' distribution, is the distribution of the number of failures before the first $k \in \{1,2,\ldots\}$ successes in a sequence of independent Bernoulli trials[1] with probability of success $0 \leq \pi \leq 1$. This distribution can also be expressed as a sum of k independent geometric random variables, since the geometric distribution describes the number of failures before the first success in a sequence of independent Bernoulli trials. The pmf of the negative Binomial distribution is given by

$$f(y|\mu,k) = \binom{y+k-1}{y}\left(\frac{\mu}{\mu+k}\right)^n\left(\frac{k}{\mu+k}\right)^k, \quad y \in \{0,1,2,\ldots\}$$

where $\mathrm{E}(Y) = \mu$ and $\phi = k^{-1}$ is the dispersion parameter. In this parameterisation the negative binomial distribution is in the EF and has a quadratic variance function, given by $\mathrm{Var}(Y) = \mu + \mu^2\phi$; we will refer to this parameterisation of the negative binomial distribution as negbin-quad.

The negative binomial distribution can also be viewed as arising from a two-stage model with a Poisson distribution where the parameter is assumed to follow a gamma distribution, reflecting additional heterogeneity over the observed counts. By considering different parameterisations of the gamma distribution we obtain different parametric forms of the negative binomial. For fixed values of the mean μ they are the same distribution but when we consider allowing the mean to vary (as in regression models) they exhibit different mean-variance behaviour. In particular, there is one

[1] A Bernoulli trial is an experiment with only two possible outcomes: 'success' or 'failure'.

variant with a linear variance function $\text{Var}(Y) = \mu + \mu\psi$, referred to here as negbin-lin; but note that in this form the resulting negative binomial model is not in the EF.

Depending on the data, either parameterisation can be the best choice to accommodate the extra variability. Note that these variance functions are inflated with respect to the Poisson distribution, and we have that $\text{Var}(Y) > \text{E}(Y)$, with a limiting case of equi-dispersion obtained when the additional parameter ϕ, or ψ, is 0 (note that these correspond to a gammma distribution with zero variance, i.e. a degenerate constant distribution, hence the reduction to the equi-dispersed Poisson distribution). Therefore, the negative binomial distribution can be considered a one-parameter extension of the Poisson distribution for modelling overdispersed counts. See Figure 1.2(e) for examples of pmf shapes for the negbin-quad distribution.

1.2.6 The Binomial model

The number of successes out of $m \in \{0,1,2,\ldots\}$ independent Bernoulli trials with same probability of success $0 \leq \pi \leq 1$ follows binomial distribution, denoted as $\text{Binom}(m,\pi)$. The values assumed by a binomial random variable are discrete and bounded between 0 and m, i.e. they can be referred to as 'discrete proportions'. In entomology, there are many cases where this type of data arises, such as in experiments measuring the proportion of viable eggs, sex ratios, or dose-response experiments where the focus is on mortality (or survival) of insects. The pmf of the binomial distribution is given by

$$f(y) = \binom{m}{y} \pi^y (1-\pi)^{m-y}, \quad y \in \{0,1,\ldots,m\}.$$

We have that $\text{E(Y)} = m\pi$ and $\text{Var(Y)} = m\pi(1-\pi)$.

Re-writing the pmf of the binomial distribution in the canonical exponential family form we obtain

$$f(y) = \exp\left\{ y \log\left(\frac{\pi}{1-\pi}\right) + \log\binom{m}{y} - m\log(1-\pi) \right\}, \quad y \in \{0,1,\ldots,m\}.$$

We can identify the canonical parameter $\theta = \log\left(\frac{\pi}{1-\pi}\right) = \text{logit}(\pi)$ and the dispersion parameter $\phi = 1$. We also have $b(\theta) = -m\log\left(\frac{1}{1+e^\theta}\right)$, and $c(y,\phi) = \log\binom{m}{y}$.

The binomial model is naturally under-dispersed, since $\text{Var}(Y) = m\pi(1-\pi) = \text{E}(Y)(1-\pi) < \text{E}(Y)$. In many applications, we may find that this mean-variance relationship does not hold, and the variability in the data is greater than accommodated by the standard binomial model. In such cases, extensions of the binomial model can be used, such as the quasi-binomial, beta-binomial and logistic-normal models [9]. See Figure 1.2(f) for different shapes of the binomial distribution pmf.

1.3 Residuals

Here we provide a brief introduction to different types of residuals used when assessing goodness-of-fit and performing model selection.

Residuals can be considered as an information metric that gives an idea about how well the specified model fits the data. They are based on the deviation between fitted/predicted values from observed values. Since they can be used to detect outliers and abnormalities in the data, they are considered an integral part of exploratory analysis.

When working with the classical linear model (Eq. 1.1), different types of residuals can be used to verify the assumptions of linearity, independence, homogeneity of variances and normality, through either visual displays or formal hypothesis testing. In this model we have the fundamental decomposition of an observation y_i into its fitted value, $\hat{y}_i$, and the residual $r_i = y_i - \hat{y}_i$ with

$$y_i = \hat{y}_i + (y_i - \hat{y}_i) = \hat{y}_i + r_i$$

and moreover the vectors $\hat{\mathbf{y}}$ and $\mathbf{r}$ are orthogonal. Hence, a very useful and basic form of diagnostic display is to plot residuals ($\mathbf{r}$) versus fitted values ($\hat{\mathbf{y}}$). This plot can help to assess whether the assumptions of variance homogeneity and linearity are met. An ideal plot would show the residuals distributed randomly around zero, with no trend (Figure 1.3(a)). When the variance is not constant, one would see changes in variability throughout the plot, such as the example in Figures 1.3(b) and 1.3(d), where the variance changes proportionately with the fitted values. When the linearity assumption is not met, the residuals versus fitted values plot will show a trend or curve, rather than points distributed randomly around zero, such as the examples in Figures 1.3(c) and 1.3(d).

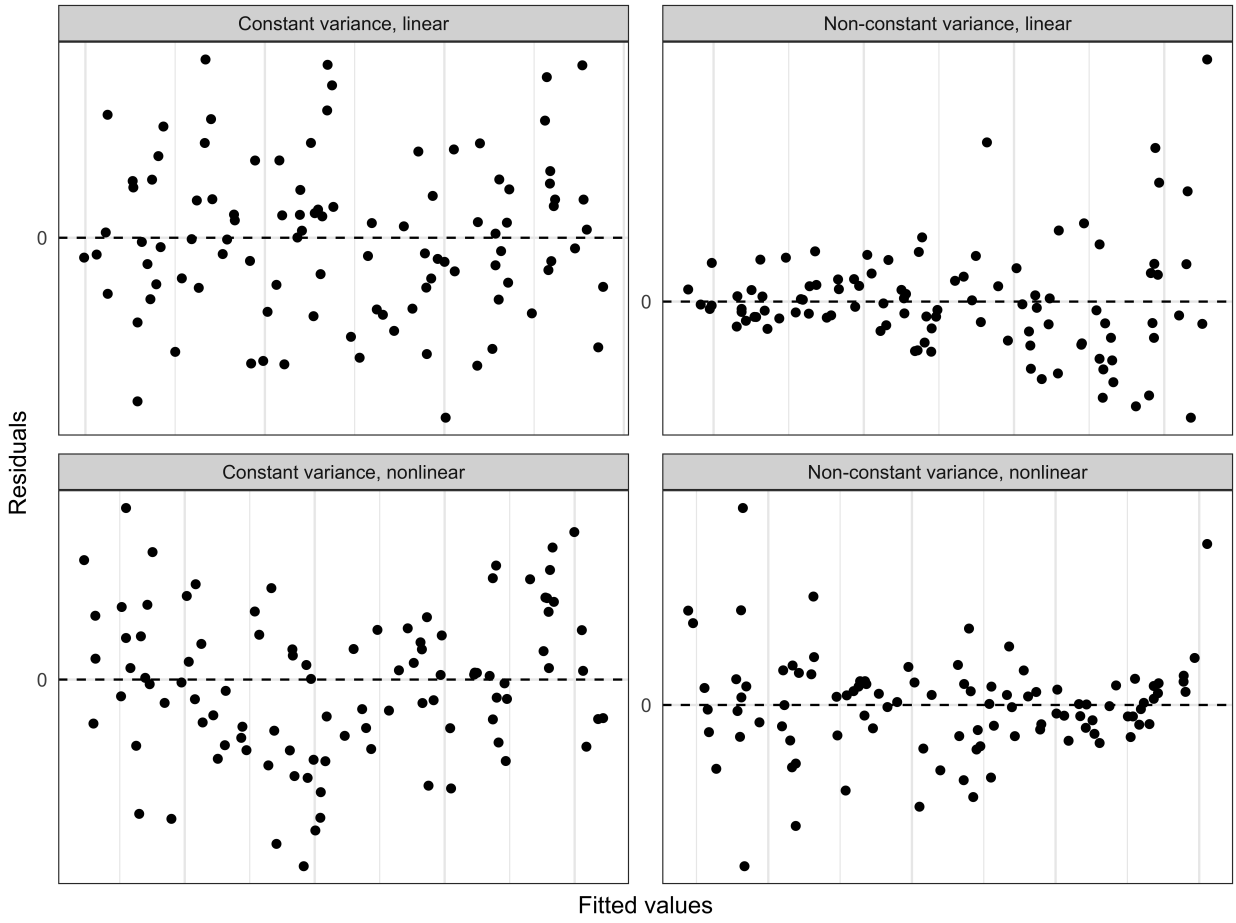

Fig. 1.3 Examples of patterns expected when looking at 'residuals versus fitted values' plots for four different scenarios. (a) Constant variance and linearity assumptions are met; (b) variance is not constant, but linearity assumption is met; (c) plot shows a trend, therefore linearity assumption is not met, however the variance seems to be constant; (d) neither the constant variance nor linearity assumptions are met.

Naturally, when working with other generalized linear models that are not the normal model, we would not expect the assumption of constant variance to hold, since some of the variance functions are proportional to the mean (Table 1.1). However, the residuals and the fitted values still form the basic building blocks of quantities of interest and useful displays. We now introduce the most commonly used residual types. Note that the type of residuals used for the model selection process should depend on the models considered, and the nature of the response variable.

1.3.1 Raw Residuals

The *raw residuals* (r_i) are defined as the difference between the observed data and fitted/predicted values:

$$r_i = y_i - \hat{\mu}_i \; ; \quad i = 0, 1, 2, \ldots, n \tag{1.4}$$

where y_i is the observed value and $\hat{\mu}_i$ is the fitted value. Large $|r_i|$ shows a higher discrepancy between the observed data and the predicted value, which may indicate that either observation y_i is an outlier under the distributional assumption or that the model does not have a good fit, if that is the case for many observations. Very small $|r_i|$ for many observations could indicate overfitting. However, this will depend on the scale of the response variable. Moreover, in classical linear regression the raw residuals should follow a normal distribution with zero mean and variance $\sigma^2 > 0$. But in non-normal scenarios the raw residuals behave differently, and may be asymmetric, and have non-constant variance. This is why it is better and commonplace to use scaled versions of the residuals.

1.3.2 Pearson Residuals

The 'Pearson residuals' r_i^P are defined as,

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}, \tag{1.5}$$

which are the raw residuals scaled by the estimated standard deviation. This formulation addresses the problem of non-constant variance; therefore under well-fitted models r_i^P should display a constant variance behaviour when plotted against the fitted values.

The Pearson statistic X^2 is calculated by summing the squared Pearson residuals, i.e. $X^2 = \sum_{i=1}^{n} (r_i^P)^2$. It can be shown that $1/phi$ times this statistic , asymptotically, follows a χ^2 distribution with $n-p$ degrees of freedom [15]. This statistic can be used to test the goodness-of-fit of a GLM when the dispersion parameter ϕ is known. This is especially useful for the Poisson and binomial GLMs, for which $\phi = 1$, fixed. Under a well-fitted Poisson or binomial GLM, we would expect X^2 to be similar to $n-p$. If $X^2 >> n-p$ this may be an indication that the variability in the data is larger than expected by the model, and therefore extended models that accommodate extra-variability would be more appropriate for analysis.

1.3.3 Deviance Residuals

The *deviance residuals* r_i^D are associated with the concept of deviance D, which is a measure that considers the departure of the fitted model from the saturated model. The saturated model has as many as parameters as observations, and reproduces all observed values exactly, i.e. $\hat{\mu}_i = \hat{y}_i = y_i$. This model clearly overfits the data, however it is useful to quantify how far the fitted (or current) model is from reproducing the data exactly. We may write

$$D = 2(l^* - l),$$

where l^* and l are the maximised log-likelihoods of the saturated and current models, respectively. The deviance can be represented as the sum of the deviance measures from each data point, such that $D = \sum d_i^2$, where d_i is the i^{th} component of deviance. The deviance residual r_i^D is then given by

$$r_i^D = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}.$$

Under a well-fitting model, the distribution of the deviance residuals approximates a normal distribution, and they are a common choice for likelihood-based methods [18].

1.4 Influence measures

There are certain data points in a data set which upon deletion would bring a significant difference in the model estimates. These data points can be identified by using suitable metrics known as *influence measures*. We now describe a few types of influence measures that can be used in model diagnostics.

1.4.1 DFBETA

Outliers are the data points that are considered to have a larger impact on the model estimates compared to other data points. In most of the cases outliers are considered as an outcome of instrumental errors or other types of error incurred during the data collection process, and depending on the circumstances could be removed prior to analysis. But in some instances outliers can be of high influence to the model estimates. DFBETA is a measure of influence that can be used to find influential outliers in the dataset. It can be calculated for parameter j and observation i as:

$$\text{DFBETA}_{ij} = \hat{\beta}_j - \hat{\beta}_{(i)j}$$

where $\hat{\beta}_j$ is the $j-$th coefficient from the regression calculated using all of the data and $\beta_{(i)j}$ is the $j-$th coefficient from the regression calculated without the $i-$th observation. The standardised version of is given by:

$$\text{DFBETAS}_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{(i)j}}{\text{SE}(\hat{\beta}_j)} \tag{1.6}$$

As a rule of thumb, it is considered that a data point is influential when $\text{DFBETAS} \geq 2/\sqrt{n}$.

1.4.2 DFFIT

DFFIT is an other useful measure to study the influence of a data point in a model which is very similar to DFBETA, however it looks at the influence on the predicted response rather than the estimated regression coefficients. The formula for DFFIT is given by:

$$\text{DFFIT}_i = \hat{y}_i - \hat{y}_{(i)i}$$

where $\hat{y}_i$ is the predicted value for point i and $\hat{y}_{(i)i}$ is the predicted value when the $i-$th observation is removed. The standardised version of DFFIT is denoted as DFFITS and is given by:

$$\text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{(i)i}}{s_{(i)}\sqrt{h_{ii}}}$$

where $s_{(i)}$ is the standard error for the fitted value $\hat{y}_i$ estimated without including y_i in the model, and h_{ii} is the $i-$th element of the diagonal of the 'hat matrix' $\mathbf{H}$. For the normal model it is calculated as

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top,$$

and is thus named since $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$. In other generalized linear models, the hat matrix is of a more complicated form, but has similar properties. The h_{ii} values are also known as *leverage* measures, which are another type of important influence metric. Higher leverage values indicate a higher chance that a specific data point is influential.

1.4.3 Cook's distance

Cook's distance D^{Cook} is generally used to find influential outliers in regression analysis and is given by:

$$D_i^{\text{Cook}} = \sum_{i=1}^{n} \frac{(\hat{y}_i - \hat{y}_{i(i)})^2}{ps^2},$$

where

$$s^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n-p}$$

is the mean squared error of the regression model, with p the number of regression coefficients. A data point is suspected to be influential when $D^{\text{Cook}} > p/n$.

1.5 Half-Normal Plots with a Simulated Envelope

A quantile-quantile plot, *q-q plot*, is a graphical method used to compare the distributions of two samples by plotting the quantiles against each other. This is mainly employed in cases where we assume a distribution for a response variable and would like to check if that is a reasonable assumption. The quantiles of the theoretical distribution are plotted against the quantiles of the data and if the distributional assumption is reasonable, the points would fall on, or close to, the identity line $y = x$ (see Figure 1.5(a)).

A normal q-q plot compares the observed data against a normal distribution. Figure 1.5 shows two q-q plots: in Figure 1.5(a) the theoretical model considered is the normal model and the data is also simulated from the normal distribution. The observed behaviour is similar to a $y = x$ plot which means the assumed model is reasonable. But from Figure 1.5(b) we conclude the assumed distribution is not a good approximation for the data.

Half-normal plots can be considered as an extension to the q-q plot where the ordered absolute value of a model diagnostic (e.g. residuals, leverage, Cook's distance, etc.) is plotted against the expected order statistics of the half normal distribution; these are particularly useful in smaller samples where the full normal plot can be rather sparse and natural sampling variation can be potentially misleading. It is a graphical method now primarily used to identify outliers and assess distributional assumptions. The half-normal plot was orginally introduced by Daniel [5] for the analysis of factorial experiments, especially those involving un-replicated designs. The paper also introduces 'guardrails' for giving better interpretation of the results. A major revision to this method was proposed by [29]. Since then, several alternative methods were introduced to reduce subjectivity.

If Y follows a Normal distribution, then $|Y|$ follows a half-normal distribution [29], with pdf given by:

$$f(x) = \sqrt{\frac{2}{\pi\sigma^2}}\, e^{-\frac{y^2}{2\sigma^2}}, \quad y \geq 0$$

The expected ordered statistics of the half-normal distribution, hereby referred to as 'half-normal scores', are approximated by:

$$\Phi^{-1}\left(\frac{i+n-\frac{1}{8}}{2n+\frac{1}{2}}\right)$$

where i is the $i-$th order statistic, $1 \leq i \leq n$, and $\Phi^{-1}(\cdot)$ is the inverse of the cumulative distribution function of the normal distribution [6].

The application of half-normal plots as a model selection method is implemented with an added simulated envelope as suggested by [2] to aid interpretability. To construct the plot, first, model diagnostics are calculated from the fitted model, the absolute value is taken and they are then sorted from minimum to maximum. Second, 99, or more, simulated realisations of the response variable are created from the fitted model, that is using the same model matrix and distributional assumptions and parameter values as given by the fitted model. The next step is to fit the same model to these simulated responses and re-calculate the same model diagnostics, take absolute values and sort them. The final step is to form the envelope by computing the percentiles of interest for each order statistic from the set of (99)

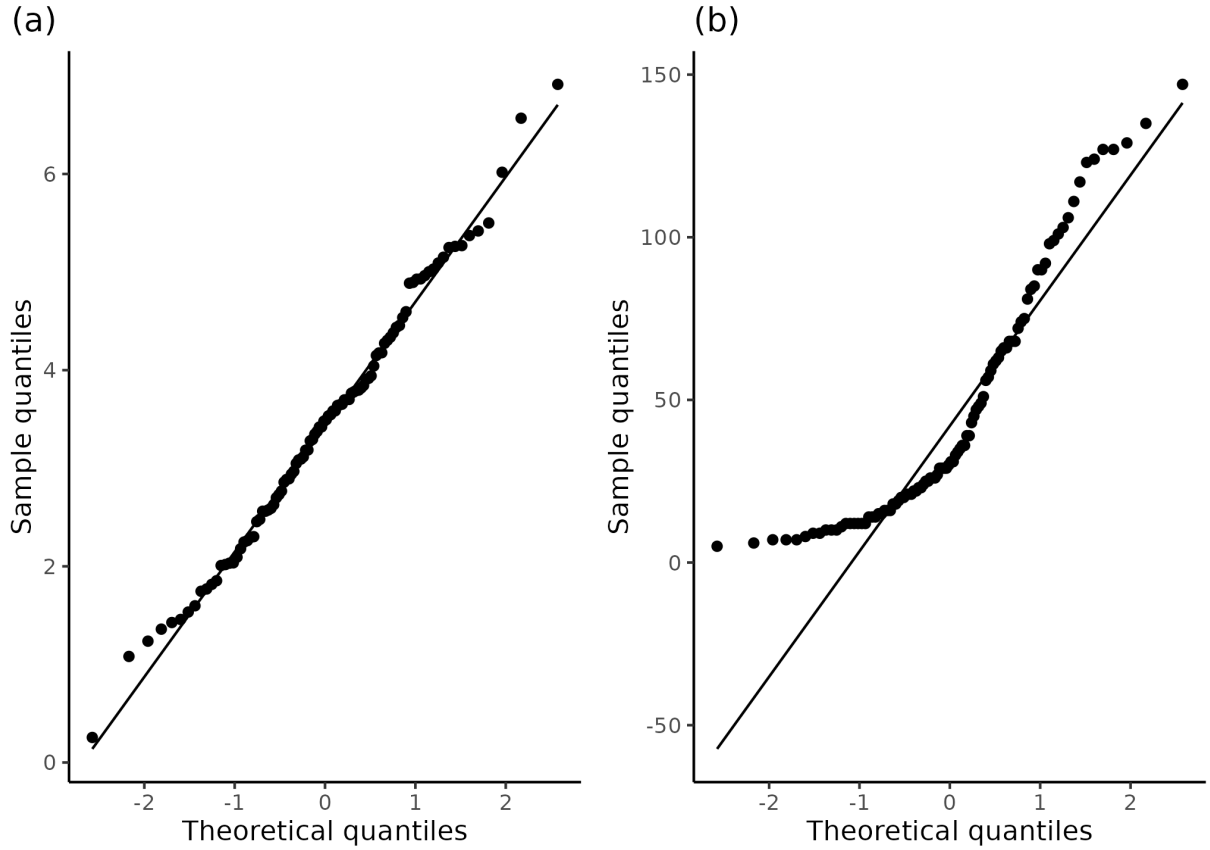

Fig. 1.4 Quantile-quantile plots showing (a) agreement between an assumed distribution and the sample distribution, and (b) disagreement (i.e. the assumption is not a reasonable one).

simulated values together with the original real data one. Typically the chosen percentiles are 2.5% and 97.5%. For this case, up to 5% of the points (order statistic values) may fall outside of the envelope to indicate a well-fitted model. If much more than 5% of the points lie outside of the envelope, it means that the observed data is not a plausible realisation of the fitted model.

We present two sample half-normal plots with a simulated envelope for model residuals in Figure 1.5. In Figure 1.5(a) all the residual points fall inside the simulated envelope, which means that the model fits the data well, i.e. the data is a plausible realisation of the assumed probability distribution. In Figure 1.5(b), however, most residual points falls outside the envelope, which means that the model is not a good fit for the data.

The half-normal plot with a simulated envelope is simple to interpret, however if the estimation procedure for a model is time-consuming, it might be computationally expensive to produce.

1.6 Examples

1.6.1 Biological control of ticks

To study the efficacy of biologically controlling ticks using nematodes in grasslands, an experiment was set up in a randomised complete block design with six blocks in a field of *Megathyrsus maximus* grass, in the state of Goiás, Brazil [10]. The field was divided into six groups of two plots each, totalling twelve plots. Within each group (block), one

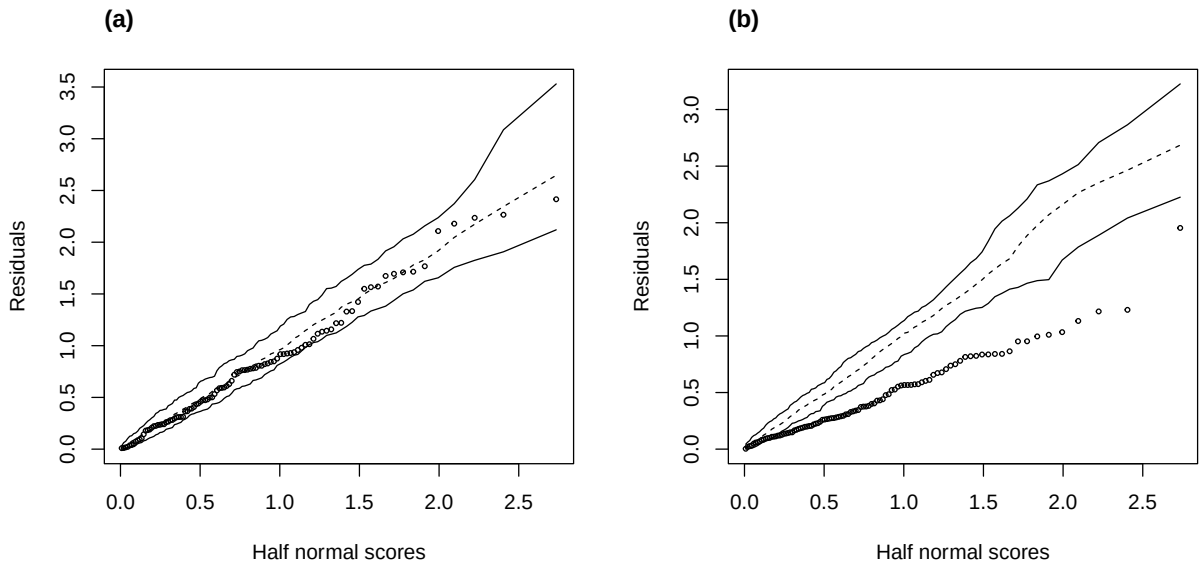

Fig. 1.5 Half-normal plot with a simulated envelope for model residuals for (a) a case where the model fits the data well, and (b) a case where with poor goodness-of-fit.

plot was treated with the entomopathogenic nematode *Heterorhabditis bacteriophara* by introducing infected *Tenebrio molitor* larvae one week before the experiment commenced, while the other plot received no treatment (control). A day before the experiment commenced, six females of the tick *Rhipicephalus microplus* were placed in each of the twelve plots. The total number of ticks in each plot was observed after 1, 2, 3, and 4 weeks (Figure 1.6)

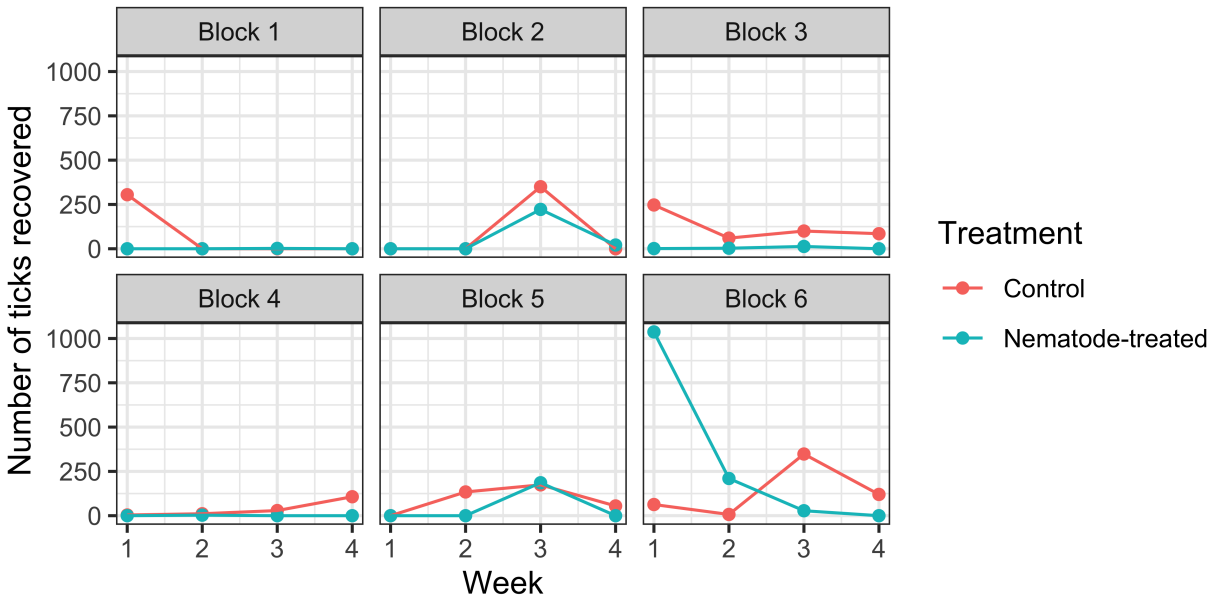

Fig. 1.6 Number of ticks of the species *Rhipicephalus microplus* recovered in each of twelve plots in a field of *Megathyrsus maximus* grass. In six of these plots, the entomopathogenic nematode *Rhipicephalus microplus* was introduced a week prior to commencement of the experiment.

It seems that the control plots had larger numbers of ticks when compared to the nematode-treated plots, apart from the first and second week in block 6, where over 1,000 ticks were recovered after one week of experiment. This type of behaviour occurs in field experiments involving arthropods, where population sizes and reproduction rates vary between plots. In this particular plot the ticks reproduced rapidly and their population exploded after one week. However, there is an exponential decline after 1 week, which could reflect successful control of the tick via the introduction of the entomopathogenic nematode.

Since the response variable consists of counts, the normal distribution is not suitable for analysis. The Poisson model is a reasonable starting point, since it is suitable to analyse count data. We observe, however, that the variance is much greater than the mean for all plots over time (Table 1.2). This indicates extra-variability, or over-dispersion, and therefore extensions to the Poisson model could be more appropriate to analyse this dataset.

Treatment	Week	Mean	Variance
Control	1	103.2	18831.77
	2	35.3	2847.07
	3	166.8	23544.17
	4	61.2	2730.17
Nematode-treated	1	173.0	179159.20
	2	36.0	7268.40
	3	75.2	10187.37
	4	3.5	73.50

Table 1.2 Mean and variance of the number of recovered *Rhipicephalus microplus* ticks in plots treated or not with the entomopathogenic nematode *Heterorhabditis bacteriophara*, over four weeks of observation.

Firstly, we ignore the time dependence between observations made on the same plot at different occasions. We fit the normal, Poisson and negative binomial models to the ticks data, using the same linear predictor, which included the effects of block, treatment, week, and an interaction between treatment and week. The normal model assumes variance homogeneity, and a quick glance at the residuals versus fitted values plot (Figure 1.7) reveals that this assumption is not met. Moreover, since the lower bound of the response variable is zero, there is an obvious lower bound for the residuals as well. The lack-of-fit of the normal model is confirmed by the half-normal plot with a simulated envelope for the raw residuals (Figure 1.8(a)). The Poisson model fit is also not adequate according to the half-normal plot in Figure 1.8(b), and this is due to the extra-variability in the data. The negative binomial model, however, seems to fit the data well ((Figure 1.8(c)).

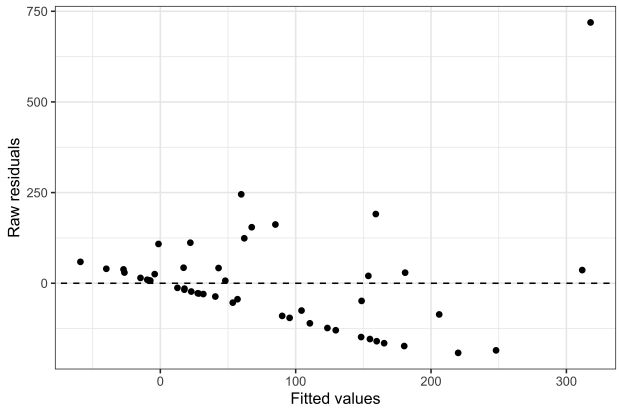

Fig. 1.7 Residuals versus fitted values for the normal model fitted to the ticks data.

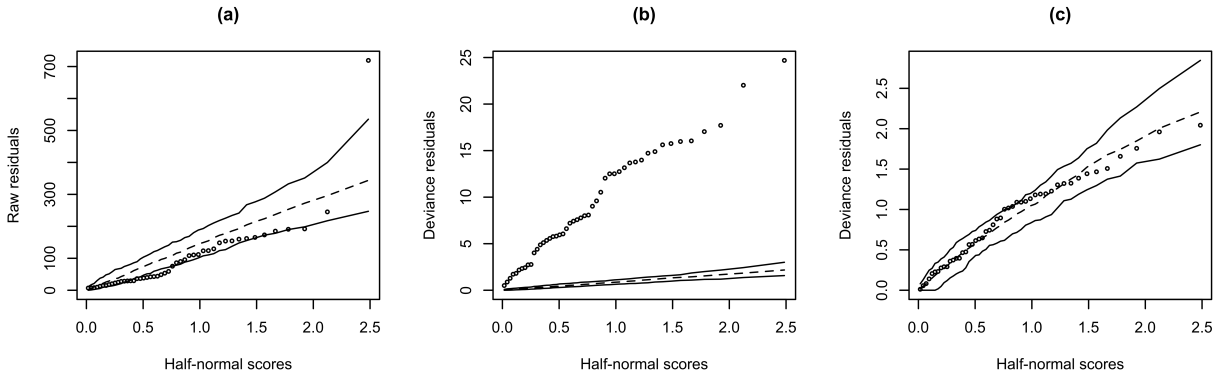

Fig. 1.8 Half-normal plots with a simulated envelope for the (a) normal model using raw residuals, (b) Poisson and (c) negative binomial models using deviance residuals, fitted to the ticks data.

The importance in assessing goodness-of-fit before drawing inferential conclusions from a statistical model is enhanced when we look at the results presented in Table 1.3. According to the normal model, there are no significant effects of time (weeks) or treatment (the entomopathogenic nematode) on the number of ticks retrieved from the field, and therefore would lead researchers to conclude that the nematode is inefficient in controlling the tick population size. This lack of significance is due to the large variability in the data overall, which results in an overestimation of the standard errors. The Poisson model, on the other hand, detects a significant interaction between time and treatment. This is due to the assumption of equi-dispersion, which results in the underestimation of the overall variability of the data. Finally, the negative binomial model, which suitably incorporates the over-dispersion in the data, yields inferential results confirming that the nematode is indeed efficient in controlling the pest and shows the interaction to be unnecessary, indeed there is no evidence of any time effect.

Model	Source	d.f.	Test statistic	$p-$value
Normal	Week	3, 35	1.29	0.29
	Treatment	1, 35	0.16	0.69
	Week $\times$ Treatment	3, 35	0.52	0.67
Poisson	Week	3	1444.10	< 0.01
	Treatment	1	57.14	< 0.01
	Week $\times$ Treatment	3	638.18	< 0.01
Negative binomial	Week	3	3.76	0.29
	Treatment	1	4.72	0.03
	Week $\times$ Treatment	3	4.98	0.17

Table 1.3 Test statistics and associated $p-$values for the effects in the linear predictor of the models fitted to the ticks data. F test statistics are used for the normal model and likelihood-ratio χ^2 statistics are used for the Poisson and negative binomial models.

1.6.2 Sustainable management of parasitic nematodes using bioagents – the 'plant height' data

The use of microbial agents as pesticides has been shown to be a more sustainable approach than chemical pesticides on agricultural pests. [20] carried out an experiment where they assessed the effectiveness of using a filamentous fungi of the order Hypocreales, namely two strains of *Pochonia chlamydosporia* and three strains of *Purpureocillium lilacinum* as potential bioagents against plant parasitic nematodes. Seeds of the common bean cultivar "IAC Milênio" were treated with suspensions prepared using each fungal strain, as well as a negative control that used only Arabic

gum. The seeds were planted and ten potted plants were used for each treatment as observational units. After 45 days, the height of the plants was measured in cm. This experiment was repeated three times, totalling 30 plants receiving each treatment.

The box plots in Figure 1.6.2 show that the plant height is very homongeneous across treatments, especially for experiments 1 and 2. However, for experiment 3 it appears that plants treated with *P. lilacinum* strain ESALQ2593 were slightly taller, suggesting a significant interaction between experiments and treatments.

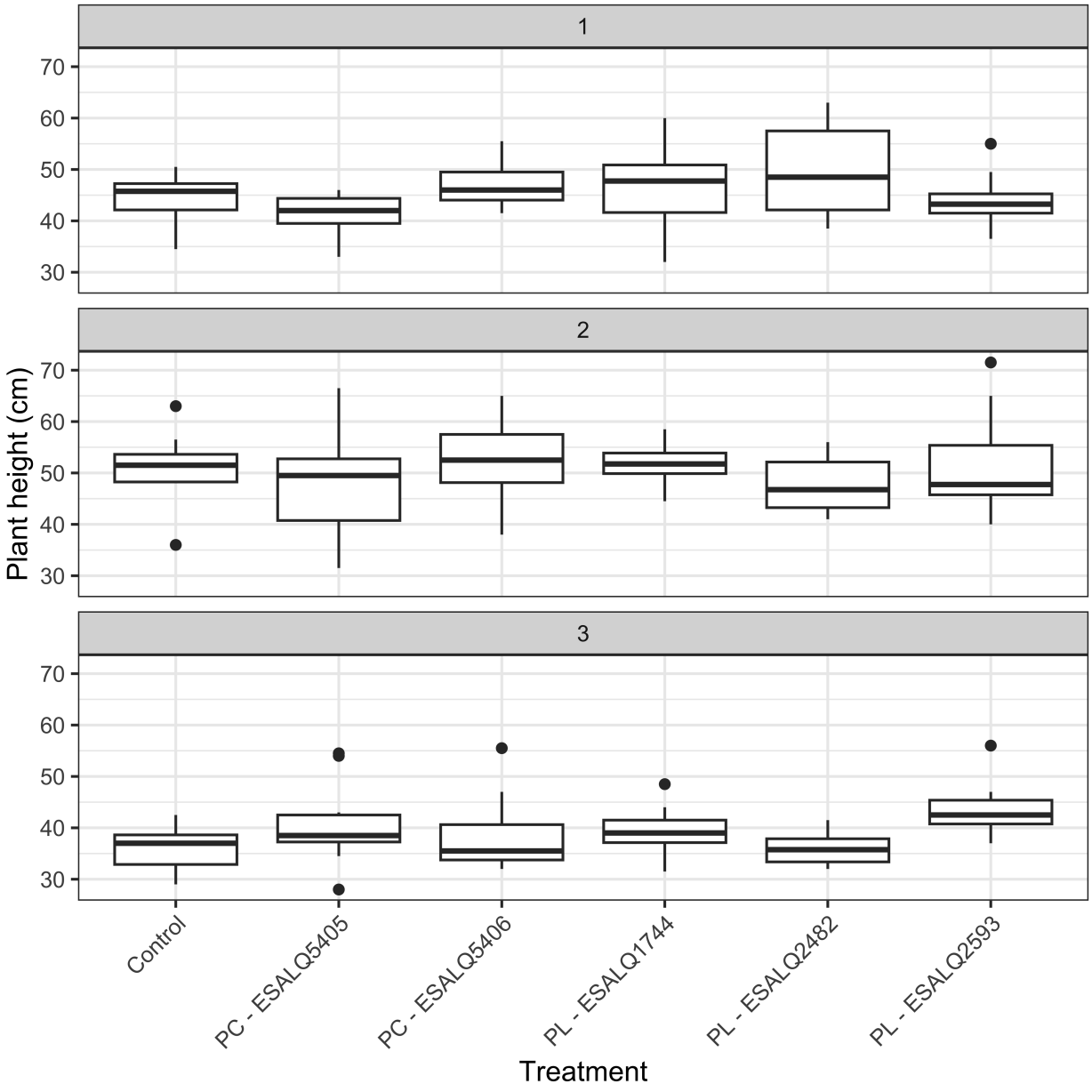

Fig. 1.9 Box plots of the height for the ten plants within each experiment (each panel numbered 1, 2 and 3) according to each treatment, which included a negative control, two strains of *Pochonia chlamydosporia* (PC - ESALQ5405 and PC - ESALQ5406), and three strains of *Purpureocillium lilacinum* (PL - ESALQ1744, PL - ESALQ2482 and PL - ESALQ2593).

We fitted normal, gamma and inverse Gaussian models to the plant height data, all including the effects of experiment, treatment, and the two-way interaction between experiment and treatment in the linear predictor. We used the

canonical link functions, i.e. identity for the normal model, inverse for the gamma model, and $g(\mu) = 1/\mu^2$ for the inverse Gaussian model. From Table 1.4, we observe that while the gamma and inverse Gaussian models agree with respect to the significance of the two-way interaction between experiment and treatment, the normal model yields a p-value larger than 0.05 for this effect (although close to the 5% significance threshold). The half-normal plots with a simulated envelope indicate, however, that the normal model is an inadequate representation of the data (Figure 1.6.2), and therefore inferential results from this model should not be taken into account. On the other hand, the gamma and inverse Gaussian models seem to fit the data well.

Model	Source	d.f.	Test statistic	$p-$value
Normal	Experiment	2, 162	42.82	< 0.01
	Treatment	5, 162	1.13	0.35
	Experiment $\times$ Treatment	10, 162	1.81	0.06
Gamma	Experiment	2, 162	45.41	< 0.01
	Treatment	5, 162	1.16	0.33
	Experiment $\times$ Treatment	10, 162	2.09	0.03
Inverse Gaussian	Experiment	2, 162	46.04	< 0.01
	Treatment	5, 162	1.14	0.34
	Experiment $\times$ Treatment	10, 162	2.25	0.02

Table 1.4 F test statistics and associated $p-$values for the effects in the linear predictor of the models fitted to the plant height data.

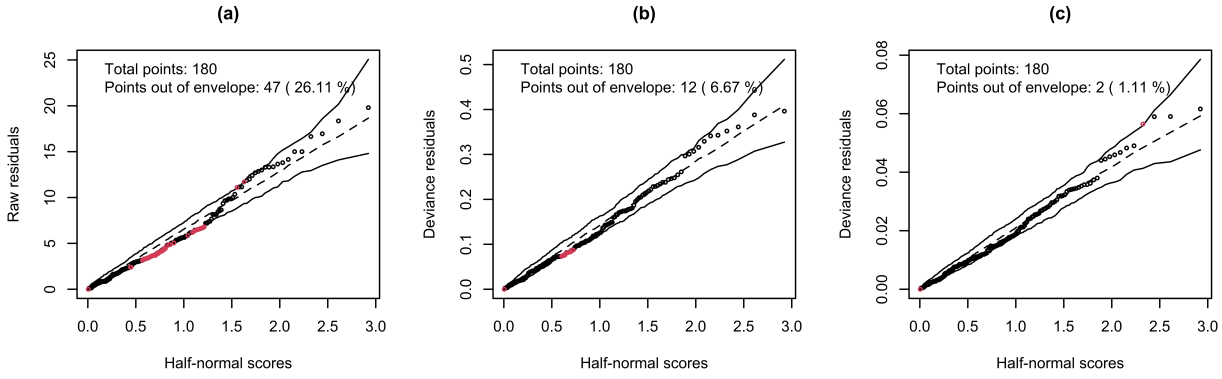

Fig. 1.10 Half-normal plots with a simulated envelope for the (a) normal, (b) gamma, and (c) inverse Gaussian models fitted to the plant height data.

1.7 Discussion

In this chapter, we aimed to present the generalized linear modelling framework with a specific focus on the use of diagnostic analyses to assess model goodness-of-fit, specifically through the use of the half-normal plot with a simulated envelope. We demonstrated that the normal model is not the most suitable option for analysis of discrete data, which is commonly found in entomological studies. In terms of software, although we used R throughout the chapter, there are other implementations of the models and techniques presented here through SPSS, Python, SAS, among others. The focus here has been on a single response variable; multivariate extensions to jointly model responses of interest are more complicated, and subject of active research.

Acknowledgements DJ is funded by Science Foundation Ireland (SFI) under Grant Number SFI 18/CRT/6049.

References

1. Akinkunmi, M. (2019). Poisson Distribution. In: Introduction to Statistics Using R. Synthesis Lectures on Mathematics & Statistics. Springer, Cham. https://doi.org/10.1007/978-3-031-02419-112
2. Atkinson, A.C. (1985) Plots, transformations and regression; an introduction to graphical methods of diagnostic regression analysis. Oxford: Clarendon Press. 282 p.
3. Balakrishnan, N., and M. Hamada. "Analyzing unreplicated factorial experiments: A review with some new proposals." University of Waterloo Institute for Improvement in Quality and Productivity Research Report (1994).
4. Coe, R., and R. D. Stern. "Fitting models to daily rainfall data." Journal of Applied Meteorology and Climatology 21.7 (1982): 1024-1031.
5. Daniel, Cuthbert. "Use of half-normal plots in interpreting factorial two-level experiments." Technometrics 1.4 (1959): 311-341.
6. Moral, RA, Hinde, J and Demétrio, CGB. "Half-normal plots and overdispersed models in R: the hnp package." Journal of Statistical Software 81.10 (2017).
7. DE
8. Dobson, Annette J., and Adrian G. Barnett. An introduction to generalized linear models. Chapman and Hall/CRC, 2018
9. Fatoretto, MaÃra Blumer, et al. "Overdispersed fungus germination data: statistical analysis using R." Biocontrol Science and Technology 28.11 (2018): 1034-1053.
10. Filgueiras, Marcos Daniel Gomes, et al. "From the laboratory to the field: efficacy of entomopathogenic nematodes to control the cattle tick." Pest Management Science (2022).
11. Folks, J. Leroy, and Raj S. Chhikara. "The inverse Gaussian distribution and its statistical application - a review." Journal of the Royal Statistical Society: Series B (Methodological) 40.3 (1978): 263-275.
12. Gardner, William, Edward P. Mulvey, and Esther C. Shaw. "Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models." Psychological bulletin 118.3 (1995): 392.
13. Haslett, John, et al. "Modelling excess zeros in count data: A new perspective on modelling approaches." International Statistical Review 90.2 (2022): 216-236.
14. Hinde, John, and Clarice GB Demétrio. "Overdispersion: models and estimation." Computational statistics & data analysis 27.2 (1998): 151-170.
15. JØrgensen, Bent. "Generalized linear models." Encyclopedia of environmetrics (2006).
16. Matis, James H., W. L. Rubink, and M. Makela. "Use of the gamma distribution for predicting arrival times of invading insect populations." Environmental entomology 21.3 (1992): 436-440.
17. Nelder, John Ashworth, and Robert WM Wedderburn. "Generalized linear models." Journal of the Royal Statistical Society: Series A (General) 135.3 (1972): 370-384.
18. Pierce, Donald A., and Daniel W. Schafer. "Residuals in generalized linear models." Journal of the American Statistical Association 81.396 (1986): 977-986.
19. R Core Team (2023) R: A language and environment for statistical computing. The R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/
20. Silva, Daniela Milanez, et al. "Production of Purpureocillium lilacinum and Pochonia chlamydosporia by Submerged Liquid Fermentation and Bioactivity against Tetranychus urticae and Heterodera glycines through Seed Inoculation." Journal of Fungi 8.5 (2022): 511.
21. Shapiro, Samuel Sanford, and Martin B. Wilk. "An analysis of variance test for normality (complete samples)." Biometrika 52.3/4 (1965): 591-611
22. Shapiro, Samuel S., and R. S. Francia. "An approximate analysis of variance test for normality." Journal of the American statistical Association 67.337 (1972): 215-216.
23. Snedecor, George W., and William G. Cochran. "Statistical Methods, eight edition." Iowa state University press, Ames, Iowa 1191 (1989).
24. Tseng, Kuan-Wei. (2012). Introduction to the Inverse Gaussian Distribution.
25. van Noortwijk, Jan M. "A survey of the application of gamma processes in maintenance." Reliability Engineering & System Safety 94.1 (2009): 2–21.
26. Ver Hoef JM, Boveng PL (2007) Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data? Ecology 88(11):2766–2772.
27. Wald, Abraham. "Foundations of a general theory of sequential decision functions." Econometrica, Journal of the Econometric Society (1947): 279-313
28. Weisstein, Eric W. "Normal Distribution." From MathWorld–A Wolfram Web Resource. https://mathworld.wolfram.com/NormalDistribution.html
29. Zahn, Douglas A. "An empirical study of the half-normal plot." Technometrics 17.2 (1975): 201-211.