

Generalised Linear Models

The concept of generalised linear models was originally proposed by Nelder & Wedderburn (1972) and developed further in McCulloch & Nelder (1989). A *general* linear model is a linear model that is general in the sense that it can combine both categorical and continuous explanatory variables (Dobson 2002). General linear models were introduced at the end of the 'statistics for free' course on linear models. Some statistical packages (e.g. R, S-Plus, Genstat) have also adopted a general approach to specifying linear models. For example, in R, rather than having a separate command for different types of regression, ANOVA, ANCOVA *etc*, all of these statistical models can be specified using the **lm** or **aov** functions (depending if you want a summary of the parameter estimates or an ANOVA table). However, these general linear models are all based on the normal distribution. For the analysis to be valid the errors in the data must be approximately normally distributed or made that way by transformation of the response variable. The approach is to mould the data to fit the normal distribution. But, there are other distributions besides the normal and Nelder and Wedderburn proposed an alternative approach: leave the data untransformed and search for a distribution that fits the data. So, *generalised* linear models are general linear models that can use not just the normal distribution but a whole family of distributions. Generalised linear models (GLMs or GLIMs for short) are more complex than the normal models that use least squares methods but most of the concepts involved are analogous to those from least squares regression and ANOVA and this can aid in a general understanding. Some data are clearly non-normal - binary response data (alive/dead, present/absent, male/female *etc.*) are probably the clearest example, - and GLMs are most useful in these situations. However, another potential advantage over using transformations and assuming a normal error distribution is that, as we shall see below, GLMs have more components than normal linear models and this can offer greater flexibility in some circumstances. McCullagh & Nelder (1989) give an example of an analysis of some hypothetical data which are Poisson distributed and have multiplicative effects, i.e. the analysis produces a significant interaction between explanatory variables. To analyse the data with normal linear models it would have to be transformed. The following transformations could be used to achieve the following:

$\sqrt{y}$: approximately constant variance
 $y^{2/3}$: approximately normal errors
 $\ln(y)$: additive effects

A single transformation changes many properties of the data even if we want to change these different properties in different ways. We may not be able to find a good compromise and it is possible that a transformation may even solve one problem while creating another. In contrast, GLMs can allow you to use their different components to address different aspects of

the data in different ways. What are these different components - there are three main components to GLMs:

- The structure of the error distribution (random component)

- The linear predictor (systematic component)
- The link function

We begin by looking at some of the available error distributions. There are many of these distributions that come from the so-called exponential family. However, in this course we will concentrate on three common non-normal types of data: i) proportions (or percentages); ii) binary responses; and iii) frequency or count data. The first two are both handled with the binomial distribution and the third with the Poisson distribution. We begin with a brief comparison of these distributions with the more familiar normal distribution.

Models for Data and Error Distributions

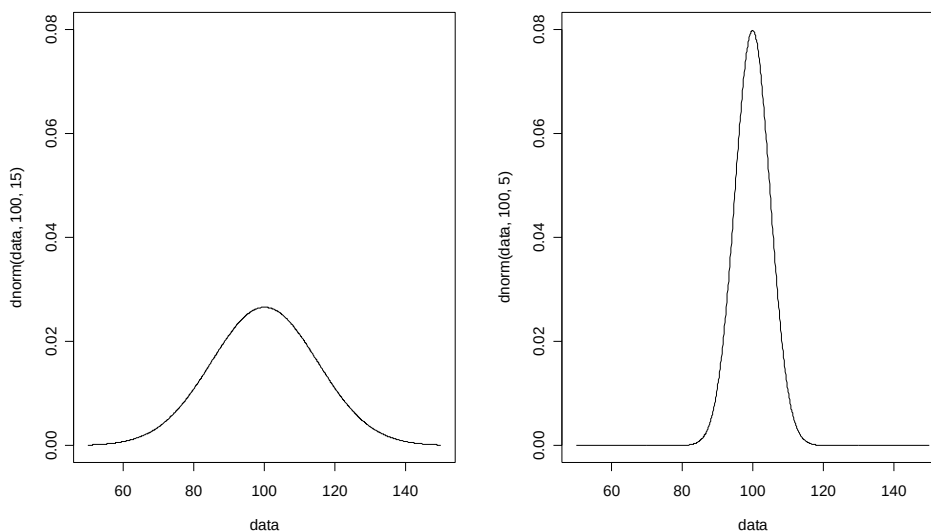
The Normal Distribution

This is the most familiar continuous distribution. It is an excellent model for a lot of continuous variables, e.g. height, weight, length of limbs within a population etc. It is defined by two parameters (a two-parameter distribution): the mean, μ (mu), and the standard deviation, σ (sigma). In fact, the variance, σ^2 , is more usually specified. This may be written:

$$X \sim N(\mu, \sigma^2) \quad (1)$$

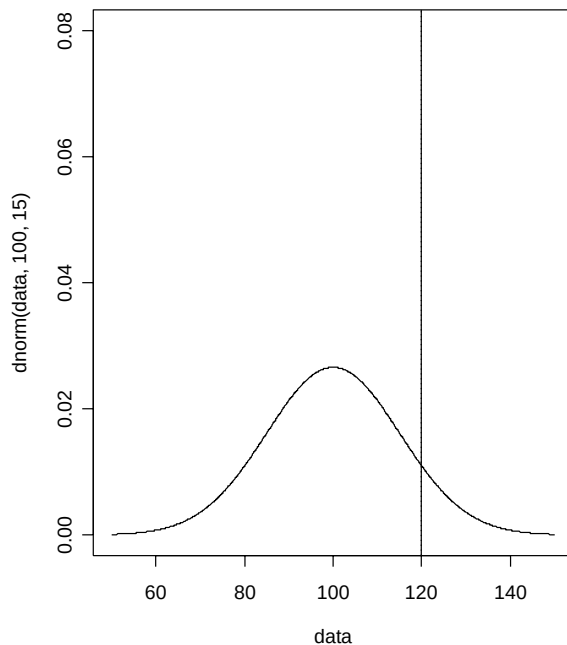
By changing these two parameters we can get a huge range of normal distributions. For example:

```
par(mfrow=c(1,2))  
data <- seq(50,150,0.1)  
plot(data, dnorm(data, 100, 15), type = "l", ylim=c(0,0.08))  
plot(data, dnorm(data, 100, 5), type = "l", ylim=c(0,0.08))
```



When dealing with a continuous distribution we are normally not interested in the probability of obtaining any particular value. For example, we are not normally interested in the probability of a person measuring 175.357 cm in height! Instead we are more usually interested in where a particular values lies within the overall distribution. For example, if the data above were the height of children, then for a child measuring 120 cm how unusual is this height for her age group? Let's add the child's value to the overall distribution:

```
par(mfrow=c(1,1))
data <- seq(50,150,0.1)
plot(data, dnorm(data, 100, 15), type = "l", ylim=c(0,0.08))
abline(v = 120, lty = 3)
```



She's certainly taller than average but would we class her height as statistically significant?

```
pnorm(120, mean = 100, sd=15, lower.tail = FALSE)
```

```
0.09121122
```

In fact we would not, 9.1 percent of children of that age are 120 cm or taller.

Discrete Distributions

Discrete data are those which can only take whole number values. For example, you cannot have 3.675 slugs. There are several discrete distributions that can be used to model different types of discrete data.

The Bernoulli Trial

Not all data can be modelled with the normal distribution. A Bernoulli trial deals with a probabilistic event that can only have two outcomes. For example, when you throw a coin it can land a head or a tail; or when you throw a dice it can land a six or a non-six. As a more ecological example, your experimental organism may die or survive. Traditionally one of these outcomes is labelled a success and the probability of success, p , is often what we want to estimate. In this case the more trials we have the more confidence we can have in our estimate of p .

The Binomial distribution

This is once again a two-parameter distribution. This time we are dealing with a sequence of independent Bernoulli trials. So, our two parameters are the number of trials, n , and the probability of success, p , in any given trial. It follows that the probability of failure, q , must be $1 - p$. An important assumption of this model is that the probability of success does not change from trial to trial. The data of interest is the total number of successes, x . For example, we might be interested in the probability of obtaining three sixes if we rolled a dice six times. The formula for the Binomial distribution is:

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad (2)$$

where:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \quad (3)$$

This formula allows you to calculate the exact probability of obtaining x successes in n trials. The first part of the formula gives the number of ways of obtaining x successes from n trials - this can be surprisingly large. The second part gives the probability of obtaining any one of those combinations. For example, imagine we rolled a dice 5 times ($n = 5$) and success would be throwing a 6, then what is the probability of throwing three sixes? First, let's look at the number of ways there are of throwing three sixes out of five throws. For example, you might throw 6, 6, 6, -, - (where - is a non-six). Or, you might throw, 6, -, 6 -, 6. In total, following formula 3 above, there are 10 ways of throwing three sixes from five throws. The probability of throwing,

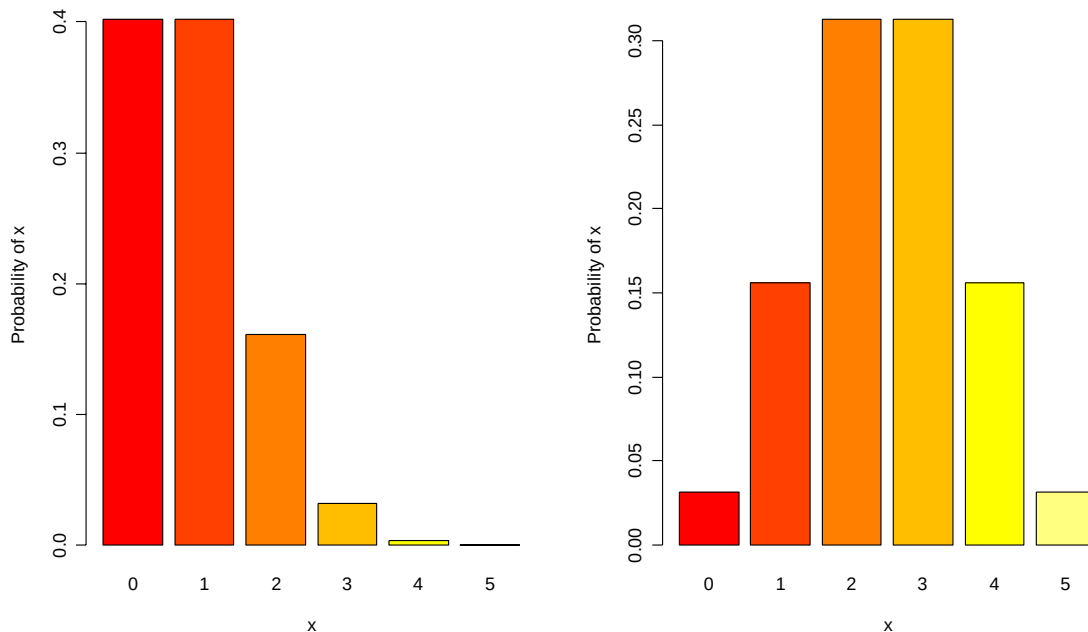
6, 6, 6, -, - is: $p \times p \times p \times q \times q$, or, $p^3 q^2$

Also, the probability of throwing,

6, -, 6, -, 6 is: $p \times q \times p \times q \times p$, or, $p^3 q^2$

Therefore, the order of the successes and failures does not affect the probability. Notice, that $p^3 q^2$ can be written as $p^3 (1-p)^2$, which can in turn be replaced with, $p^x (1-p)^{n-x}$. Let's have a look at some Binomial distributions with different values of p and n .

```
par(mfrow=c(1,2))
xvals <- seq(0, 5, 1) ; xvals
p <- numeric(6) ; p
for(i in 1:6)
  p[i] <- dbinom(xvals[i], 5, 1/6)
  p ; sum(p)
barplot(p, names=levels(factor(xvals)), xlab="x", ylab="Probability of x")
p2 <- numeric(6) ; p2
for(i in 1:6)
  p2[i] <- dbinom(xvals[i], 5, 1/2)
  p2 ; sum(p2)
barplot(p2, names=levels(factor(xvals)), xlab="x", ylab="Probability of x")
```



When p equals 0.5 the distribution is much more symmetrical. You can also see that it might be possible to make a good normal approximation to the second distribution but not to the first.

The Poisson Distribution

Sometimes we do not know the number of trials but we do know the number of successes. If the number of trials is assumed to be very large, i.e. an event had many opportunities to occur, and the number of successes is relatively small, then we can use the Poisson distribution, which is an approximation to the Binomial distribution. The Poisson distribution has only one parameter, the mean number of successes, μ (mu). Don't forget that the assumption of independence of trials still holds. Examples of a good Poisson approximation could be the number of diseased trees in one hectare sample plots in a forest. If the incidence of disease is relatively low (most trees are healthy) and a one

hectare plot contains a lot of trees it is much easier to just count the number of diseased trees rather than all the healthy ones as well. In this case a Poisson approximation to the exact binomial distribution would be very good. The classic example comes from Prussian cavalry where records were kept of the number of individuals killed by horse kicks in fourteen units in each of 20 successive years. In this case the number of failures (where a horse did not kick a man to death) would be impossible to quantify. The formula for calculating each term of the Poisson distribution is:

$$P(x) = \frac{e^{-\mu} \mu^x}{x!}$$

And, because,

$$P(x) = P(x-1) \frac{\mu}{x}$$

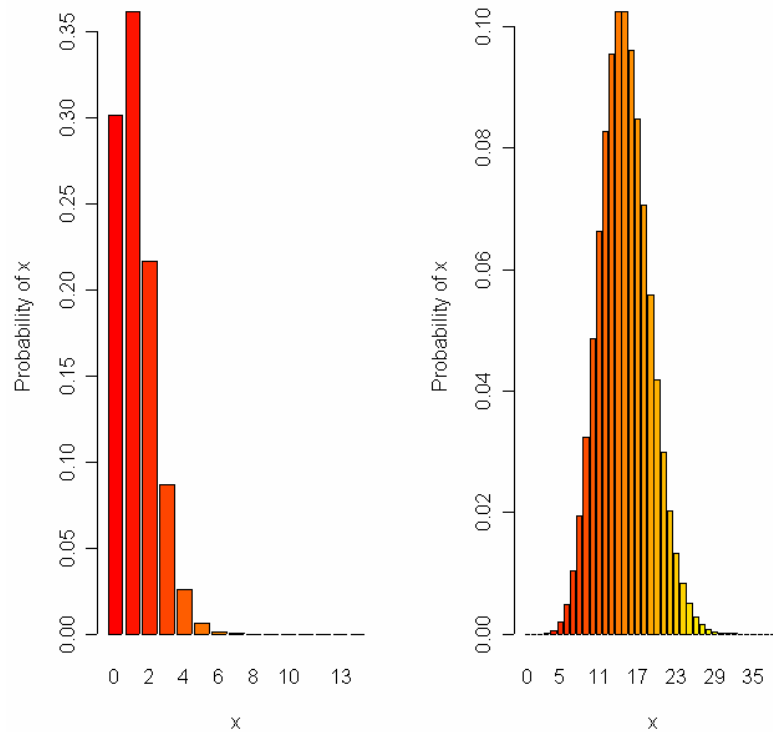
then successive probabilities can be calculated by taking the preceeding probability, multiplying by the mean and dividing by x , starting with the zero term as follows:

$$P(0) = e^{-\mu}$$

Notice that the only thing we know is the mean number of times that something happens. From this we can calculate the exact probability of that same event never occurring, occurring once, occurring twice etc. An important property of this distribution is that the variance, σ^2 is equal to the mean, μ (μ).

Let's have a look at some Poisson distributions:

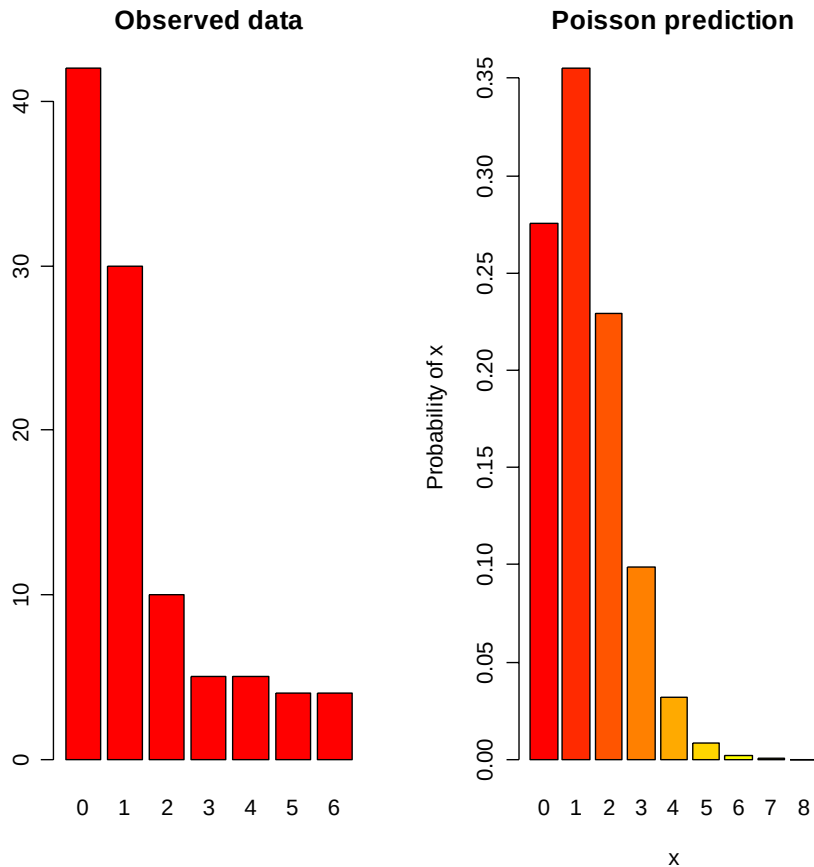
```
par(mfrow=c(1,2))
xvals1 <- seq(0, 14, 1) ; xvals1
xvals2 <- seq(0, 39, 1) ; xvals2
p <- numeric(15) ; p
for(i in 1:15)
  p[i] <- dpois(xvals1[i], 1.2)
  p ; sum(p)
barplot(p, names=levels(factor(xvals1)), xlab="x", ylab="Probability of x")
p2 <- numeric(40) ; p2
for(i in 1:40)
  p2[i] <- dpois(xvals2[i], 15)
  p2 ; sum(p2)
barplot(p2, names=levels(factor(xvals2)), xlab="x", ylab="Probability of x")
```



Notice that as the mean of the Poisson distribution increases the distribution becomes more like the normal. The Poisson distribution is incredibly useful as a model of random events in either space or time. Let's look at some hypothetical data on the numbers of diseased trees obtained from one hectare forest plots described above.

```
data <- c(rep(0:7,c(42,30,10,5,5,4,4,0))) ; data
freq.data<-table(data)
mean(data)
par(mfrow=c(1,2))
barplot(freq.data, main="Observed data")
xvals1 <- seq(0, 8, 1) ; xvals1
p <- numeric(9) ; p
for(i in 1:9)
  p[i] <- dpois(xvals1[i], mean(data))
p ; sum(p)
barplot(p, names=levels(factor(xvals1)), xlab="x", ylab="Probability of x",
main="Poisson prediction")
mean(data)

[1] 1.29
```



There is clearly a mis-match between the observed data and the corresponding Poisson distribution, i.e. a Poisson distribution with the same mean. The data has more zeros and more high values than we would expect. This is very common in ecological data and is called 'over-dispersion'. A simple test for this is to calculate the mean/variance ratio. Remember that in the Poisson distribution this ratio should be one.

```
var(data)/mean(data)
```

```
[1] 2.134445
```

In this case the ratio is 2.13 and indicates that our data deviates from a Poisson distribution. If we want to formally test for this we can compare the observed values to those expected using a Chi-squared goodness-of-fit test. First we need the expected values. We have the probabilities from the vector, `p`, above so we multiply by the total number of observations (100) to get the expected number in each category.

```
p*100
```

```
[1] 27.527078309 35.509931019 22.903905507 9.848679368
3.176199096 0.819459367 0.176183764 0.032468151
0.005235489
```

Number of observations	Observed	Expected	$(O - E)^2 / E$
0	42	27.53	7.61
1	30	35.51	0.85
2	10	22.90	7.27
>2	18	14.06	1.10

Notice that one restriction of the Chi-squared test is that categories should not have an expected values of less than five. When this occurs categories can be pooled until the expected value is large enough, in our case all values greater than two are pooled. We can use R as a calculator to perform the Chi-squared test:

```
obs <- c(42, 30, 10, 18)
exp <- c(27.53, 35.51, 22.9, 14.06)
deviations <- ((obs-exp)^2)/exp ; deviations
test.stat <- sum(deviations) ; test.stat

[1] 7.6055539 0.8549732 7.2668122 1.1040967
[1] 16.83144
```

The actual test statistic is the sum of the deviations from expected values. In this case the value is 16.83. We can test the significance against a Chi-squared distribution (see below) with $k-p-1$ degrees of freedom where p is the number of parameter estimated and k is the number of categories of observation. In our case we have estimated one parameter, the Poisson mean μ , so we have $4 - 1 - 1 = 2$ degrees of freedom.

```
1 - pchisq(16.83, 2)

[1] 0.0002215195
```

The deviation from the Poisson distribution is highly significant (notice that because `pchisq` gives us the lower tail of the chi-squared distribution we use `1 - pchisq` to get the upper tail, i.e. the probability of obtaining a value at least as high as 16.83 by chance alone). Biologically this tells us that diseased trees do not occur at random. Remember the assumption behind the Poisson distribution is a series of independent Bernoulli trials (each trial here is a tree) in which the probability of success does not vary from trial to trial. A moments thought reveals that these assumptions are unlikely to be met in the case of diseases. Healthy trees adjacent to diseased trees are much more likely to become infected generating the aggregated pattern that we saw. Tests like this are often used to test for randomness. Aggregated distributions can instead be modelled using the negative binomial distribution (a two-parameter distribution with a mean and aggregation parameter).

The Central Limit Theorem

The theorem states that even for highly non-normal populations repeated experiments to

obtain the sample mean will result in a distribution of sample means that is approximately normal. The approximation improves as the sample size on which the mean is based increases. This means that if your data are means based on relatively large sample sizes they are likely to be normally distributed even though the individual observations may be highly non-normal. The normal distribution is incredibly important and useful in statistics not just because it provides a good model for data like height or weight but because other distributions can often be approximated with a corresponding normal distribution. For example, for the binomial distribution, if $np \geq 5$ and $nq \geq 5$ then a normal distribution provides a usable approximation. In practise, as we have already seen, these conditions are most likely when n is large and p is close to 0.5. For the Poisson distribution if μ is at least 30 then a normal distribution with mean 30 and standard deviation 30 will provide a good approximation. The approximation gets better as μ gets bigger. You can see this from the two Poisson distributions given above. The take home message is that you should not automatically assume that proportion and count data will always be non-normal - under circumstances like those given above you may be able to analyse these data assuming a normal distribution.

Linear Predictors

The linear predictor, the systematic (as opposed to random) component, represents the explanatory variables (and their interactions) in the model. The model relates each observed value of y to a predicted value. The predicted value is obtained by transformation of the value emerging from the linear predictor (see the link function below). The linear predictor, η (eta), is a linear sum of the effects of the explanatory variables, x_j :

$$\eta_i = \sum_1^p x_{ij} \beta_j =$$

where i indexes the observations and the x are the values of the p different explanatory variables. The β s are the parameters to be estimated from the data (see McCulloch & Nelder 1989, p.26). There are as many terms in the linear predictor as there are parameters to be estimated. For a simple linear regression the linear predictor is the familiar:

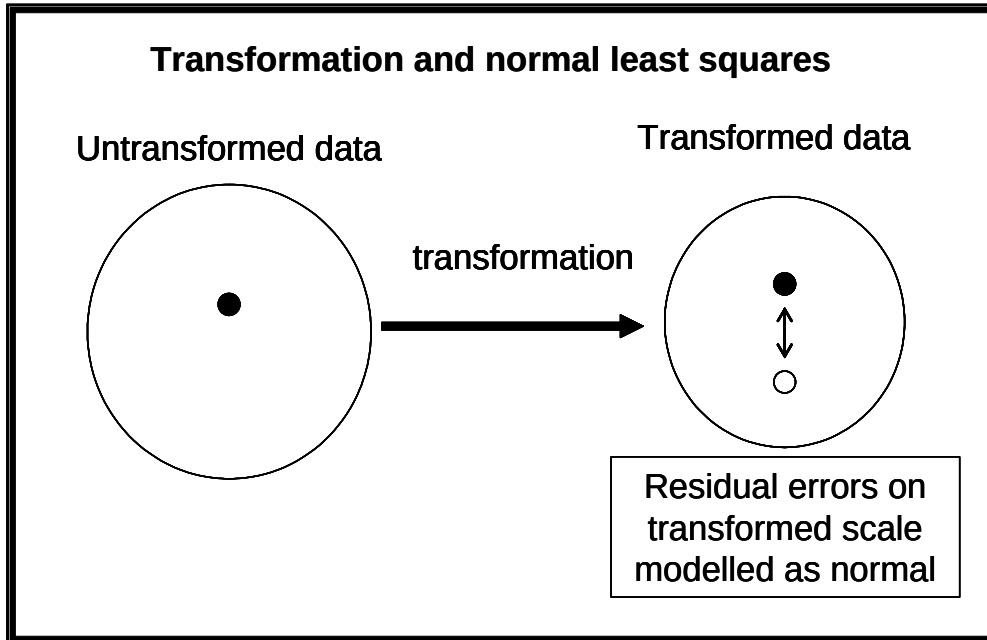
$$y = a + bx$$

where a is the intercept and b is the slope. The linear predictor is the sum of the two terms. For a one-way ANOVA with two treatments the linear predictor is the sum of the means for each factor level. A covariate adds an extra parameter to the linear predictor and an interaction term between two continuous variables also adds one (more generally interaction terms adds as many parameters as the degrees of freedom that the interaction has).

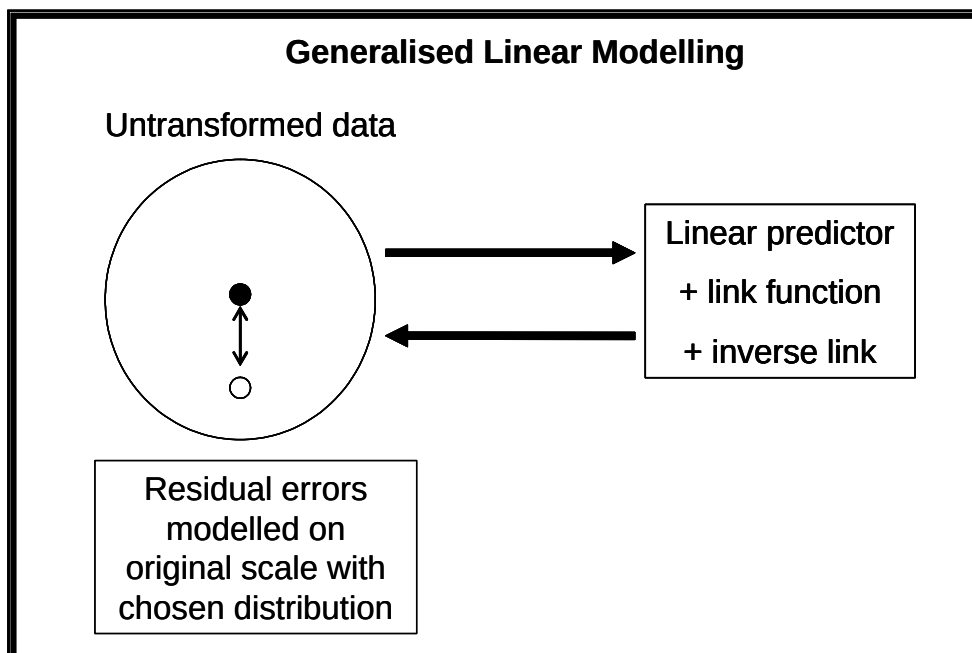
Link Functions

The link function is probably the trickiest aspect of GLMs to understand. The link function is the GLM alternative to transforming the data to fit the normal distribution.

Instead, the data is left alone and a different error distribution can be used in conjunction with a link function. The following diagram is adapted from Crosbie & Hinch (1985). When using normal least squares, if residual plots show them to be non-normal, the data must be transformed until the residuals become approximately normally distributed.



With GLMs the data is left untransformed. Instead, the link function is combined with an inverse link to produce fitted values back on the original scale of measurement. These residuals errors may be modelled with a normal error distribution, or more likely with a non-normal distribution such as the Poisson.



(η) to the

random component (the response variable with its error distribution) by linking the values emerging from the linear predictor back to the data on the original scale of measurement:

$$\eta = g(\mu)$$

where the link function is symbolised as $g(\cdot)$. Predicted values are generated by the GLM evaluating the linear predictor for each value of the response variables including any transformation involved. The fitted value is produced by back-transforming the value emerging from the linear predictor by applying the inverse of the link function to get back to the original scale of measurement. Residual differences between the observed and predicted values are then modelling using the selected error distribution (or 'family' in R) on the original scale of measurement. It may be helpful to think about a very simple situation: imagine we were performing a simple linear regression with some data that was normally distributed and where no transformation was required (obviously, we would not really need GLMs in a case like this and would just use least squares, but for the sake of illustration imagine we used maximum likelihood with a normal error structure and a link function that leaves things unchanged). The linear predictor is the familiar $y = a + bx$. When no transformation is employed the GLM equivalent is the identity link, which also does nothing. In this case we select the normal error distribution and the errors are modelled accordingly. Now imagine a more complex example with some Poisson distributed count data. In this case the link function would now be the log link and the fitted values would be generated by the inverse link, taking the exponent (anti-logging). The residuals would be generated by comparing these predicted counts to the observed data and the residuals modelled using the Poisson distribution.

Maximum Likelihood

Given a set of data and having chosen a model (we could try and compare several) maximum likelihood is a method for determining which parameters of the model produce the best model fit and make the data most likely to be observed. For a simple linear regression of data with a normal error structure we use least squares to minimise the sum of the squared deviations of the observed and expected values. With different error structures and link functions there are alternative forms of maximum likelihood calculation. However, for normal data the maximum likelihood estimate is the least squares estimate.

Crawley (2002) illustrates maximum likelihood with the following example. Suppose we threw three dice and added up the total, what would the expected value be? We can simulate one roll of the di in R as:

```
ceiling(6 * runif(1))
```

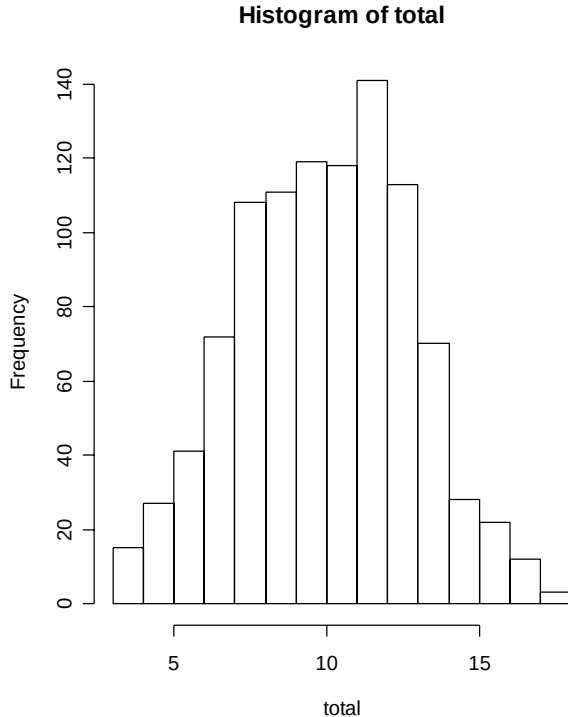
The total score off all 6 faces of a di is $1 + 2 + 3 + 4 + 5 + 6 = 21$, so the expectation for a single roll is $21 \times 1/6 = 3.5$. We can simulate three rolls and add up the total by:

```
sum(ceiling(6 * runif(3)))
```

The total score from three dice is $3 \times 21 = 63$. So the expected score is 10.5. Finally, we can repeat the whole process many times and graph the results as:

```
total <- numeric(1000)
for (i in 1:1000) total[i] <- sum(ceiling(6 * runif(3)))
hist(total)
```

The maximum score from three dice is $3 \times 6 = 18$ and we can see that our histogram extends up to this maximum with the most likely outcome around 10 and 11.



However, what if we did not know the probability of each outcome ($1/6$ in this case) and the data was used to determine the value of the probability, p ? The maximum likelihood estimate of p can be defined as:

$$E(x) = pT = p \times 63$$

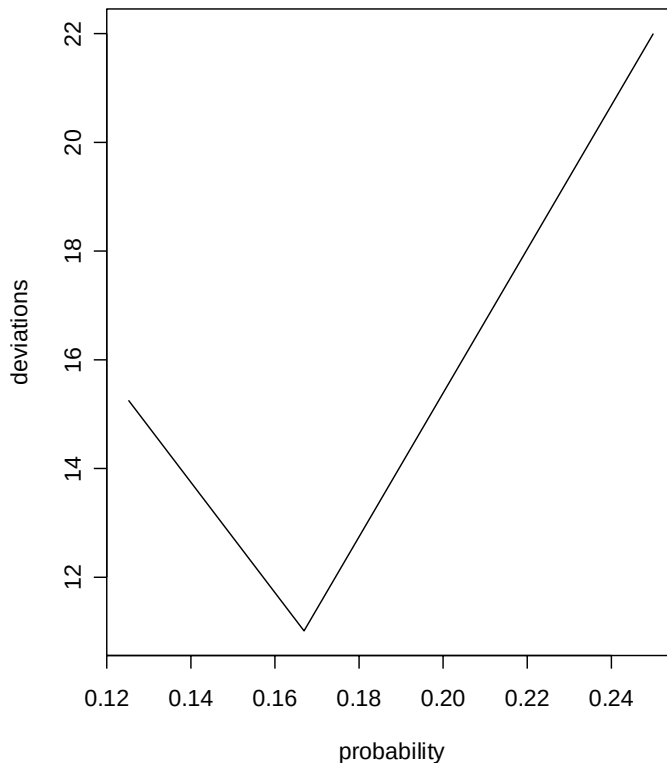
where T is the total from three dice, 63. Given the observed data and our model, $E(x) = 63p$, what value of p makes the data most likely to be observed? Suppose we ran the experiment three times and got the data:

10, 12, 5, 14

We can compare some different possible values of p . If $p = 1/4$ then our expectation would be $63 \times 1/4 = 15.75$ with residual differences of -5.75, -3.75, -10.75 and -1.75. with an absolute sum of 22. All of the observed data are less than the expected

value. If p was $1/8$ the expectation would be $63 \times 1/8 = 7.875$ and the deviations of 2.125, 4.125, -2.875 and 6.125 come to an absolute sum of 15.25. Finally, if $p = 1/6$ then the expectation, given above, is 10.5 and the sum of the absolute deviations is 11. If we take the sum of the absolute deviations as a measure of the goodness of fit of the model then we can plot the results as:

```
probability <- c(1/4, 1/6, 1/8)
deviations <- c(22, 11, 15.25)
plot(deviations ~ probability, type = "l")
```



The value of p that minimises the deviations and makes the data most likely to be observed is $p = 1/6$.

For a binomial example: imagine we carry out a single trial and find 5 parasitised animals in a sample of 9. The maximum likelihood estimator is:

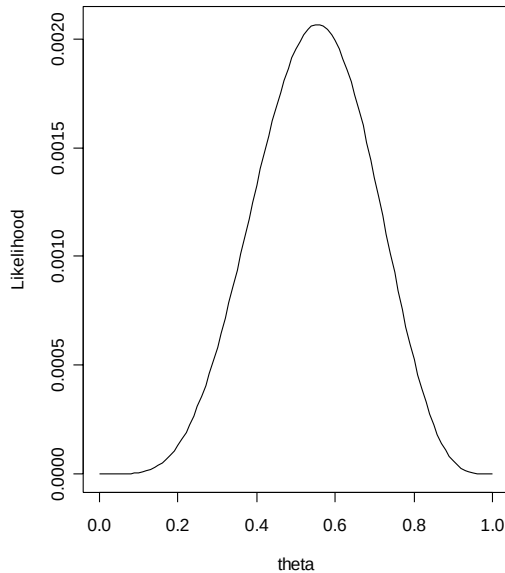
$$P(5) = \left(\frac{9!}{(9-5)!5!} \right) \theta^5 (1-\theta)^{(9-5)}$$

where we are trying to estimate θ which simplifies to:

$$L(\theta) = \theta^5 (1 - \theta)^{(9-5)}$$

We can plot the likelihood against theta:

```
theta <- seq(0, 1, 0.01)
par(mfrow = c(1, 1))
plot(theta, theta^5 * (1 - theta)^4, type = "l", ylab = "Likelihood")
```



The maximum likelihood occurs at 0.55 which agrees with an intuitive estimate of $5 / 9$.

For a Poisson process the maximum likelihood estimator of μ for r events (counts) observed in n trials is also the mean and hence again agrees with our intuition(see, e.g., Crawley 2002 p.507):

$$\mu = \frac{\sum r}{n}$$

Parameter Estimation

Maximum likelihood estimation is performed by an iterative process of weighted (sometimes written as re-weighted) least squares. The process takes the data as initial estimates of the fitted values and then repeats attempting to minimise the residual errors - the differences between the observed and fitted values. The regression is not carried out on y , as in normal least squares, but on a combination of the linear predictor and the link function applied to y . The weights are functions of the fitted values. The parameter estimates are adjusted on each iteration until changes are small. Alternatively the estimates may 'fail to converge' and R will stop the iterative process (the default is after ten cycles) and give a message to this effect.

Deviance and Hypothesis Testing

In normal models the fit of a model is measured by the sum of squares (SS), the sum of the squared (to get rid of negative values) differences between the observed and predicted

(fitted) values. In GLMs the equivalent is deviance. Deviance is defined as, -2 times the difference in the log likelihood between the fitted model and the saturated model, i.e. a model that fits the data perfectly. Deviance is estimated in different ways for different distributions following these formulas:

Error distribution	Deviance
Normal	$\sum (y - \bar{y})^2$
Binomial	$2 \sum y \ln \left(\frac{y}{\mu} \right) + (n - y) \ln \left[\frac{n - y}{n - \mu} \right]$
Poisson	$2 \sum y \ln \left(\frac{y}{\mu} \right) - (y - \mu)$

where y is the observed data, $\bar{y}$ the mean of y , μ are fitted values predicted by maximum likelihood and n is the binomial denominator.

In binomial GLMs the estimation of the deviance can be illustrated with this simple example from Crawley (2002).

In binomial GLMs the estimation of the deviance can be illustrated with this simple example from Crawley (2002). For proportion data the deviance is:

$$2 \sum y \log \left[\frac{y}{\hat{y}} \right] + (n - y) \log \left[\frac{n - y}{n - \hat{y}} \right]$$

where, y is the observed number of successes, $\hat{y}$ is the predicted fitted value and n the binomial denominator (the total from which we observed y successes). Imagine an experiment with two treatments, each with two replicates, where we got 3 out of 6 and 4 out of 10 for one treatment and 5 out of 9 and 8 out of 12 for the other. First, calculate the average of two proportions by summing the successes and dividing by the total number of attempts:

$$3 + 4 + 5 + 8 = 20 / 6 + 10 + 9 + 12 = 37 = 0.5405$$

Next we calculate the four expected counts, $\hat{y}$:

```
y <- c(3, 4, 5, 8)
n <- c(6, 10, 9, 12)
n*20/37
```

```
[1] 3.243243 5.405405 4.864865 6.486486
```

The total deviance, calculated in two portions, is:

```
p1 <- 2*(3*log(3/3.243)+3*log(3/(6-3.243))+4*log(4/5.405)+6*log(6/(10-5.405)))
p2 <- 2*(5*log(5/4.865)+4*log(4/(9-4.865))+8*log(8/6.486)+4*log(4/(12-6.486)))
total.deviance <- p1+p2 ; total.deviance
```

```
[1] 1.629673
```

To get the residual deviance we need to calculate a new set of fitted values based on the two treatment means. The mean proportion in the first treatment was:

```
mean1 <- 7/16 ; mean1
mean2 <- 13/21 ; mean2
```

```
[1] 0.4375
```

```
[1] 0.6190476
```

so the fitted values are:

```
p <- c(7/16, 7/16, 13/21, 13/21)
p*n
```

```
[1] 2.625000 4.375000 5.571429 7.428571
```

and the residual deviance, again calculated in two parts, is:

```
p1 <- 2*(3*log(3/2.625)+3*log(3/(6-2.625))+4*log(4/4.375)+6*log(6/(10-4.375)))
p2 <- 2*(5*log(5/5.571)+4*log(4/(9-5.571))+8*log(8/7.429)+4*log(4/(12-7.429)))
residual.deviance <- p1+p2 ; residual.deviance
```

```
[1] 0.4201974
```

We can now check our calculated deviance against that from R:

```
treatment <- factor(c(1, 1, 2, 2))
pq <- cbind(y, n-y)
treat.model <- glm(pq ~ treatment, family = binomial) ; summary(treat.model)
total.dev.model <- glm(pq ~ 1, family = binomial) ; summary(total.dev.model)
```

```
Call: glm(formula = pq ~ treatment, family = binomial)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.2513	0.5040	-0.499	0.618
treatment2	0.7368	0.6752	1.091	0.275

Residual deviance: **0.4206** on 2 degrees of freedom

```
Call: glm(formula = pq ~ 1, family = binomial)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.1625	0.3299	0.493	0.622

Residual deviance: **1.6297** on 3 degrees of freedom

We will look at the output in detail below but it shows us that the residual deviance and the total deviance (also called residual in this case, but from a model where we have only fitted the grand mean, parameter 1) match our calculated values. The calculations get more complex for more complex examples but this simple example gives us an idea of what is going on inside R and where the outputted deviances come from.

In GLMs with Poisson errors the deviance is:

$$2 \sum O \ln \left(\frac{O}{E} \right)$$

where O is the observed count and E is the expected count based on the statistical model. Imagine you had two counts from each of two areas: 3 and 6 in Location one and 4 and 7 in location two. Are the areas different? The expected count is the overall mean:

```
mean(c(3, 6, 4, 7))
```

```
[1] 5
```

And the total deviance can then be calculated as:

$$2 \times \left\{ 3 \times \ln \left(\frac{3}{5} \right) + 6 \times \ln \left(\frac{6}{5} \right) + 4 \times \ln \left(\frac{4}{5} \right) + 7 \times \ln \left(\frac{7}{5} \right) \right\}$$

```
2 * (3 * log(3/5) + 6 * log(6/5) + 4 * log(4/5) + 7 * log(7/5))
```

```
[1] 2.048368
```

The means for the two locations are $9/2 = 4.5$ and $11/2 = 5.5$ respectively. The residual deviance after fitting location should therefore be:

$$2 \times \left\{ 3 \times \ln \left(\frac{3}{4.5} \right) + 6 \times \ln \left(\frac{6}{4.5} \right) + 4 \times \ln \left(\frac{4}{5.5} \right) + 7 \times \ln \left(\frac{7}{5.5} \right) \right\}$$

```
2 * (3 * log(3/4.5) + 6 * log(6/4.5) + 4 * log(4/5.5) + 7 * log(7/5.5))
```

```
[1] 1.848033
```

The treatment deviance is the difference:

2.048368 - 1.848033

[1] 0.200335

We can perform the same analysis in R using a GLM with Poisson errors as follows:

```
y <- c(3, 6, 4, 7);
location <- factor(c(1,1,2,2));
model <- glm(y ~ location, poisson); summary(model)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.5041	0.3331	4.516	6.32e-06 ***
location2	0.2007	0.4492	0.447	0.655

Null deviance: 2.0484 on 3 degrees of freedom
Residual deviance: 1.8480 on 2 degrees of freedom

The total deviance of the null model and the residual deviance of the model including location match. Calculations for more elaborate examples are more complex as they are iterative estimations but these two simple example hopefully serves to demystify deviance.

With normal least squares we test hypotheses by performing variance ratio tests to derive F-values. The variance (mean square) is the sums of squares divided by the degrees of freedom. The variance of the term in question is divided by the variance of the appropriate error term and compared to an F-value with the appropriate degrees of freedom for the numerator and denominator in the variance ratio test.

Hypothesis testing in GLMs works differently. We will see exactly how it works when we come to real examples later on but the general process is as follows. We begin by fitting a saturated model. This accounts for all the deviance in the model but has no explanatory power. In normal least squares it would be like fitting a variable that coded for every individual data point - you would account for all the sums of squares but it would be of no use for explanation. You then fit one or a series of GLMs. Each GLM has a deviance which is calculated by the maximum likelihood methods explained above. The deviance is assumed to be distributed approximately as chi-square so the deviance values can be tested directly as chi-square values. We only need consider the degrees of freedom for the explanatory variable in question (e.g. 1 d.f. for a regression). Because we compare the likelihoods of whole models when using GLMs it is usual to compare models that differ in the term of interest. Effects of individual variables may be reported in the text when writing up results, giving the chi-square value and the degrees of freedom. Alternatively, we can draw up analysis of deviance tables that compare a series of models in terms of their likelihood ratios (with the degrees of freedom again of course) rather than an ANOVA table with its F-tests. One complexity comes if the dispersion parameter is not approximately equal to one, so-called over- or under-dispersion, when we return to

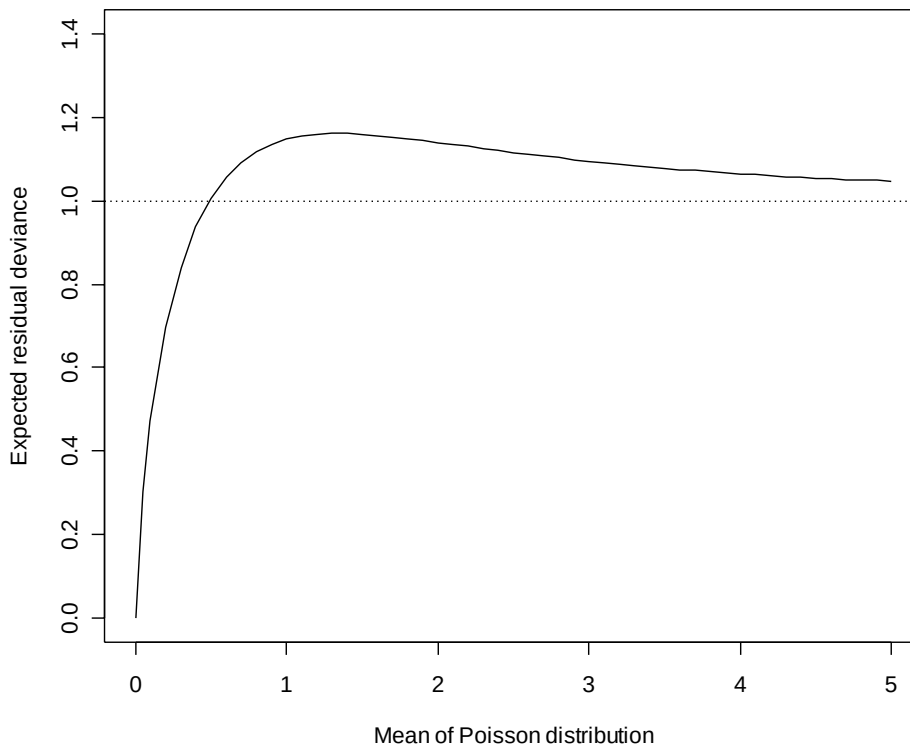
using F-tests because the approximation to the chi-square distribution has been shown not to hold (see below).

The Chi-square Distribution

The chi-square distribution is a one-parameter distribution characterised by the degrees of freedom where the mean is equal to the degrees of freedom ν , new, and the variance to 2ν . We will come across it a lot as deviance is approximately distributed as chi-square. One of the assumptions of GLMs is that the residual deviance is approximately equal to the degrees of freedom. That is, the ratio, the so-called dispersion parameter, is approximately one.

The following graph is taken from Venables & Ripley (2002) where they emphasize that the assumption of a dispersion parameter of 1 is only approximate, especially at lower values of np for the binomial and lower means of the Poisson (see page 208).

```
x <- 0:100
plik <- function(lambda)
  sum(dpois(x, lambda) * 2 * ( (lambda - x) + x * log(pmax(1, x) / lambda)))
lambda <- c(1e-8, 0.05, seq(0.1, 5, 0.1))
plot(lambda, sapply(lambda, plik), type = "l", ylim = c(0, 1.4), ylab = "Expected
residual deviance", xlab = "Mean of Poisson distribution")
abline(h = 1, lty = 3)
```



Quasi-Likelihood

Following the invention of the GLM framework by Nelder & Wedderburn (1972), Wedderburn (1974) went on to demonstrate an alternative approach of quasi-maximum likelihood. In this approach, rather than assuming a distribution (with their varying parameters and properties), the modelling is done based only on the mean and variance of the data. This approach is useful when the mean/variance relationship assumed by some of the GLMs proves not to hold, that is when the dispersion parameter described above is different from the assumed value of one. In these cases we can use QML as a way to deal with the so-called over- (or under-) dispersion by employing R's quasi-binomial and quasi-poisson functions.

Analogies between Normal Least Squares and Generalised Linear Models

The broader suite of GLMs is more complex to understand than normal least squares methods. However, most of the features of GLMs have analogies in normal least squares methods (in fact the NLS methods can be seen as special cases for the normal distribution) as follows:

Normal Least Squares	Generalised Linear Models
Sums of squares and variance (MS)	deviance
Least squares	Iterated weighted maximum likelihood
Normal error distribution	Alternative error distributions from the exponential family
Transformations	Link functions
Variance ratio F-tests	Deviance ratio Chi-squared tests

Further Reading

Detailed descriptions of GLMs can be found in McCullagh & Nelder (1989) and McCulloch & Searle (2001). Dobson (2002) provides an introduction. The application of GLMs using S-Plus (and R) is described in Crawley (2002).

References

- Crawley, M. J. 1993. GLIM for Ecologists. Blackwell Scientific Publications, London.
- Crawley, M. J. 2002. Statistical computing: An introduction to data analysis using S-Plus. John Wiley and Sons Ltd, Chichester.
- Crosbie, S. F., and G. N. Hinch. 1985. An intuitive explanation of generalised linear models. New Zealand Journal of Agricultural Research **28**:19 - 29.
- Dobson, A., J. 2002. An Introduction to Generalized Linear Models (second edition). Chapman & Hall, Boca Raton, Florida.
- McConway, K. J., M. C. Jones, and P. C. Taylor. 1999. Statistical Modelling using GENSTAT. Arnold, London.
- McCullagh, P., and J. A. Nelder. 1989. Generalized Linear Models. Chapman & Hall, London.
- McCulloch, C. E., and S. R. Searle. 2001. Generalized, Linear and Mixed Models. John Wiley & Sons, New York.
- Nelder, J. A., and R. W. M. Wedderburn. 1972. Generalized Linear Models. Journal of the Royal Statistical Society A **135**:370-384.
- Venables, W. N., and B. D. Ripley. 2002. Modern applied statistics with S. Springer-Verlag, Berlin.
- Wedderburn, R. W. M. 1974. Quasi-likelihood functions, GLM's and the gauss-Newton method. Biometrika **61**:439-447.