5 Modelling proportions - the logistic model

5.1 The binomial distribution

The Binomial distribution Bin(m, p) is the distribution of the number of successes in m independent trials, where p is the probability of success in each trial. For example, tossing a fair coin 100 times would result in a Bin$(100, 0.5)$ distribution for the number of 'heads'. The term "success" need not correspond to a favourable outcome; it is merely the language traditionally used by statisticians in connection with this model. For example, "success" might correspond to death.

If y has a Bin(m, p) distribution, its pmf is

$$f(y) = \left(\begin{array}{c} m \\ y \end{array} \right) p^y (1-p)^{m-y}. \tag{5.1}$$

The special case with $m = 1$ is the *Bernoulli* distribution. This is the distribution of the probability of success in a single trial. Suppose we conduct m independent Bernoulli trails, each having success probability p. Let B_j denote the result of the jth Bernoulli trial, where $B_j = 1$ if the jth trial results in success, and let $B_j = 0$ otherwise. Then the total number of successes has a Binomial distribution:

$$\begin{aligned} y &= \sum_{j=1}^{m} B_j \\ &\sim \text{Bin}(m, p). \end{aligned} \tag{5.2}$$

We consider data where each observation y_i originates as a sum of m_i i.i.d. Bernoulli random variables as in (5.2), where for each i the success probability p_i may differ due to the influence of covariates x_i. Thus

$$y_i \sim \text{Bin}(m_i, p_i), \qquad \text{for } i = 1, \ldots, n, \tag{5.3}$$

where p_i depends on x_i.

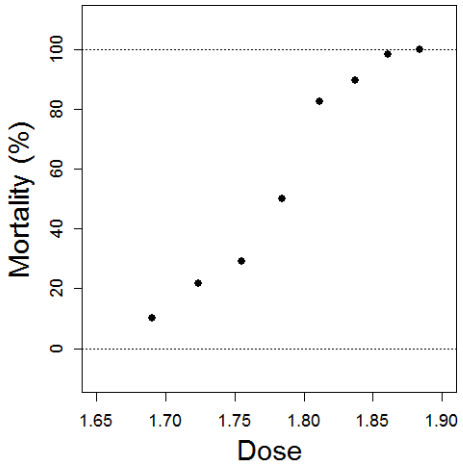
5.2 Application: dose–response experiments

A variable dose of some reagent is administered to each study subject, and the occurrence of a specific response is recorded. This is a *dose–response* experiment, one of the first uses of regression models for Bernoulli (or Binomial) responses.

For example, the Table below gives the number of beetles killed y_i out of a total number m_i that were exposed to a dose x_i of gaseous carbon disulphide, for $n = 8$ dose levels

$i = 1, \ldots, 8$ (Dobson: pp.109 in 1st edn; pp.119 in 2nd edn; pp.127 in 3rd edn). The proportion killed $p_i = y_i/m_i$ at each dose level i is plotted on the right.

Dose x_i	No. of beetles m_i	No. killed y_i
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60



Now equation (5.3) motivates modelling the beetle data as $y_i \sim \text{Bin}(m_i, p_i)$, where $\eta_i = \alpha + \beta x_i$.

Common choices for the link function, $g(\mu) = \eta_i$, where η_i takes values in $(-\infty, \infty)$, include:

(a)

$$\begin{aligned} g(\mu_i) &= \Phi^{-1}\left(\frac{\mu_i}{m_i}\right) \\ &= \Phi^{-1}(p_i), \end{aligned}$$

where Φ^{-1} is the inverse Standard Normal cumulative distribution function. Commonly called *probit analysis*; historically very popular.

(b)

$$\begin{aligned} g(\mu_i) &= \log\frac{\mu_i}{m_i - \mu_i} = \log\frac{m_i p_i}{m_i - m_i p_i} \\ &= \log\frac{p_i}{1 - p_i} = \text{logit}(p_i) \end{aligned}$$

This is referred to as *logistic regression*. In practice, this is more popular than probit analysis as it fits more neatly in the generalized linear model framework as it is the canonical link function.

(c)

$$\begin{aligned} g(\mu_i) &= \log\{-\log(1 - \mu_i/m_i)\} \\ &= \log\{-\log(1 - p_i). \end{aligned}$$

This is the *complementary log-log* or *cloglog* transformation. Note that $g(\mu_i) \neq -g(m_i - \mu_i)$, so allows for asymmetric treatment of $p_i > \frac{1}{2}$ versus $p_i < \frac{1}{2}$. Due to this asymmetry, in practice we might consider whether to model "successes" or "failures".

5.3 Residuals

The Pearson residuals for Binomial data take the form:

$$e_i^P = \frac{y_i - m_i\hat{p}_i}{\sqrt{m_i\hat{p}_i(1-\hat{p}_i)}}.$$

For large m_i, the usual Normal approximation to the Binomial means that the Pearson residuals are approximately $N(0,1)$ distributed.

5.4 Deviance

The deviance is

$$D = 2\sum_{i=1}^{n}\left\{y_i \log\left(\frac{y_i}{m_i\hat{p}_i}\right) + (m_i - y_i)\log\left(\frac{m_i - y_i}{m_i(1-\hat{p}_i)}\right)\right\}, \tag{5.4}$$

which is approximately χ^2_{n-r} distributed if the model is "correct", where r is the number of degrees of freedom in the model (i.e. the number of columns of the design matrix). This formula can be remembered as:

$$D = 2\sum_{j=0}^{1}\sum_{i=1}^{n} o_{ji}\log\frac{o_{ji}}{e_{ji}}$$

where o_{ji} denotes the observed value and e_{ji} denotes the expected value in cell (j,i) of the $2 \times n$ table of successes and failures:

	i			
	1	2	...	n
Failure $(j=0)$	$m_1 - y_1$	$m_2 - y_2$		$m_n - y_n$
Success $(j=1)$	y_1	y_2		y_n

Another goodness-of-fit statistic is the Pearson chi-squared statistic:

$$X^2 = \sum_{j=0}^{1}\sum_{i=1}^{n}\frac{(o_{ji} - e_{ji})^2}{e_{ji}}.$$

This is asymptotically equivalent to the deviance (5.4) (proof is by Taylor series expansion; omitted). Thus, asymptotically, X^2 is also approximately χ^2_{n-r} distributed. Both approximations can be poor if the expected frequencies are small, but X^2 copes slightly better with this problem. See Dobson, p.136 for more details.

5.5 Overdispersion

Examination of residuals and deviances may indicate that a model is not an adequate fit to the data. One possible reason is *overdispersion*. This can occur for any error distribution where the variance is linked to the mean — e.g. Binomial, Poisson. Overdispersion that occurs with these distributions is called *extra-Binomial* or *extra-Poisson* variation.

Recall that if $y_i \sim \text{Bin}(m_i, p_i)$, $\text{Var}(y_i) = m_i p_i(1 - p_i)$. Overdispersion occurs if observations which have been modelled by a $\text{Bin}(m_i, \widehat{p}_i)$ distribution have substantially greater variation than $m_i\widehat{p}_i(1 - \widehat{p}_i)$. This will lead to a value of D substantially greater than the expected value of $n - r$. This can occur if the model is missing appropriate explanatory variables or has the wrong link function, or if the y_i are not independent.

One solution is to include an extra parameter τ in the model so that $\text{Var}(y_i) = m_i p_i(1 - p_i)\tau$. For more details, see Section 7.7 of Dobson or chapter 6 of Collett (1991) "Modelling Binary data", Chapman & Hall.

The `glm` function in `R` allows for *extra-Binomial* or *extra-Poisson* variation through setting `family=quasibinomial()` or `family=quasipoisson()`.