Chapter 9
Models for Proportions: Binomial GLMs

We believe no statistical model is ever *final; it is simply a placeholder until a better model is found.*
Singer and Willett [22, p. 105]

9.1 Introduction and Overview

Chapters 5–8 develop the theory of GLMs in general. This chapter focuses on one specific GLM: the binomial GLM. The binomial GLM is the most commonly used of all GLMs. It is used to model proportions, where the proportions are obtained as the number of 'positive' cases out of a total number of independent cases. We first compile important information about the binomial distribution (Sect. 9.2), then discuss the common link functions used for binomial GLMs (Sect. 9.3), and the threshold interpretation of the link function (Sect. 9.4). We then discuss model interpretation in terms of odds (Sect. 9.5), and how binomial GLMs can be used to estimate the median effective dose ED50 (Sect. 9.6). The issue of overdispersion is then discussed (Sect. 9.8), followed by a warning about a potential problem with parameter estimation in binomial GLMs (Sect. 9.9). Finally, we explain why goodness-of-fit tests are not appropriate for binary data (Sect. 9.10).

9.2 Modelling Proportions

The outcome of many studies is a proportion y of a total number m: the proportion of individuals having a disease; the proportion of voters who vote in favour of a particular election candidate; the proportion of insects that die after being exposed to different doses of a poison. For all these examples, a binomial distribution may be an appropriate response distribution. In each case, the m individuals in each group are assumed to be independent, and each individual can be classified into one of two possible outcomes.

The binomial distribution has already been established as an EDM (Example 5.3), and binomial GLMs used in examples in previous chapters to

P. K. Dunn, G. K. Smyth, *Generalized Linear Models with Examples in R*,
Springer Texts in Statistics, https://doi.org/10.1007/978-1-4419-0118-7_9

develop the theory of GLMs. Useful information about the binomial distribution appears in Table 5.1 (p. 221). The probability function for a binomial EDM is

$$\mathcal{P}(y;\mu,m) = \binom{m}{my}\mu^{my}(1-\mu)^{m(1-y)} \tag{9.1}$$

where m is known and $\phi = 1$, and where $y = 0, 1/m, 2/m, \ldots 1$, and the expected proportion is $0 < \mu < 1$. To use the binomial distribution in a GLM, the prior weights w are set to the group totals m. The unit deviance for the binomial distribution is

$$d(y,\mu) = 2\left\{y\log\frac{y}{\mu} + (1-y)\log\frac{1-y}{1-\mu}\right\}.$$

When $y = 0$ or $y = 1$, the limit form of the unit deviance (5.14) is used. The residual deviance is $D(y,\hat{\mu}) = \sum_{i=1}^{n} m_i d(y_i,\hat{\mu}_i)$. By the saddlepoint approximation, $D(y,\hat{\mu}) \sim \chi^2_{n-p'}$ for a model with p' parameters in the linear predictor. The saddlepoint approximation is adequate if $\min\{m_i y_i\} \geq 3$ and $\min\{m_i(1-y_i)\} \geq 3$ (Sect. 7.5).

A binomial GLM is denoted GLM(binomial; link), and is specified in R using `family=binomial()` in the `glm()` call. Binomial responses may be specified in the `glm()` formula in one of three ways:

1. The response can be supplied as the observed proportions y_i, when the sample sizes m_i are supplied as the `weights` in the call to `glm()`.
2. The response can be given as a two-column array, the columns giving the numbers of successes and failures respectively in each group of size m_i. The prior weights `weights` do not need to be supplied (R computes the weights m as the sum of the number of successes and failures for each row).
3. The response can be given as a factor (when the first factor level corresponds to failures, and all others levels to successes) or as a logicals (either `TRUE`, which is treated as the success, or `FALSE`). The prior weights `weights` do not need to be supplied in this specification (and are set to one by default). This specification is useful if the data have one row for each observation (see Example 9.1). In this form, the responses are binary and the model is a Bernoulli GLM (see Example 4.6). While many of the model statistics are the same (Problem 9.14), there are some limitations with using this form (Sect. 9.10).

For binomial GLMs, the use of quantile residuals [5] is strongly recommended for diagnostic analysis (Sect. 8.3.4.2).

Example 9.1. An experiment running turbines for various lengths of time [19, 20] recorded the proportion of turbine wheels y_i out of a total of m_i turbines developing fissures (narrow cracks) (Table 9.1; Fig. 9.1; data set: `turbines`). A suitable model for the data may be a binomial GLM.

Table 9.1 The number of turbine wheels developing fissures and the number of hours they are run (Example 9.1)

Case i	Hours x_i	Turbines m_i	Prop. of fissures y_i	No. of fissures $m_i y_i$	Case i	Hours x_i	Turbines m_i	Prop. of fissures y_i	No. of fissures $m_i y_i$
1	400	39	0.0000	0	7	3000	42	0.2143	9
2	1000	53	0.0755	4	8	3400	13	0.4615	6
3	1400	33	0.0606	2	9	3800	34	0.6471	22
4	1800	73	0.0959	7	10	4200	40	0.5250	21
5	2200	30	0.1667	5	11	4600	36	0.5833	21
6	2600	39	0.2308	9					

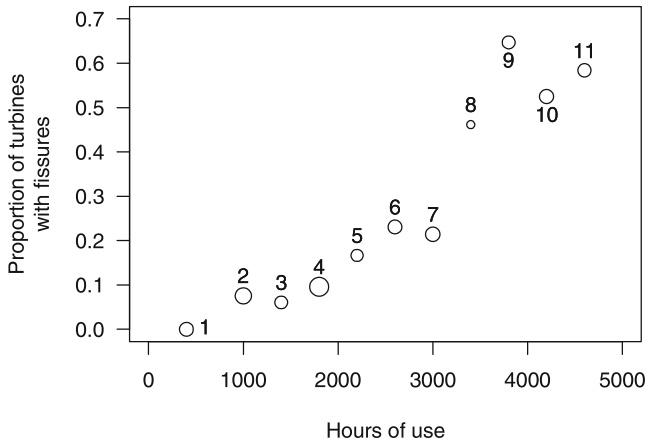

Fig. 9.1 The proportion of turbine wheels developing fissures plotted against the number of hours of use. Larger plotting symbols indicate proportions based on larger sample sizes. The numbers beside the points refer to the case number (Example 9.1)

For these data, the first and second forms of specifying the response are appropriate and equivalent:

```
> library(GLMsData); data(turbines)
> tur.m1 <- glm( Fissures/Turbines ~ Hours, family=binomial,
                 weights=Turbines, data=turbines)
> tur.m2 <- glm( cbind(Fissures, Turbines-Fissures) ~ Hours,
                 family=binomial, data=turbines)
> coef(tur.m1); coef(tur.m2)
  (Intercept)         Hours
-3.9235965551  0.0009992372
  (Intercept)         Hours
-3.9235965551  0.0009992372
```

To use the third form of data entry, the data would need to be rearranged so that each individual turbine was represented in its own line, hence having $\sum_{i=1}^{n} m_i = 432$ rows. □

9.3 Link Functions

Specific link functions are required for binomial GLMs to ensure that $0 < \mu < 1$. Numerous suitable choices are available. Three link functions are commonly used with the binomial distribution:

1. The *logit* (or logistic) link function, which is the canonical link function for the binomial distribution and the default link function in R:

$$\eta = \log \frac{\mu}{1-\mu} = \text{logit}(\mu). \tag{9.2}$$

   (R uses natural logarithms.) This link function is specified in R using `link="logit"`. A binomial GLM with a logit link function is often called a *logistic regression model*.
2. The *probit link function*: $\eta = \Phi^{-1}(\mu) = \text{probit}(\mu)$, where $\Phi(\cdot)$ is the CDF for the normal distribution. This link function is specified in R as `link="probit"`.
3. The *complementary log-log link function*: $\eta = \log\{-\log(1-\mu)\}$. This link function is specified in R as `link="cloglog"`.

In practice, the logit and probit link functions are very similar (Fig. 9.2). In addition, both are symmetric about $\mu = 0.5$, whereas the complementary log-log link function is not.

Two other less common link functions permitted in R for binomial GLMs are the `"cauchit"` and `"log"` links. The `"cauchit"` link function is based on the Cauchy distribution (see Sect. 9.4), but is rarely used in practice. The

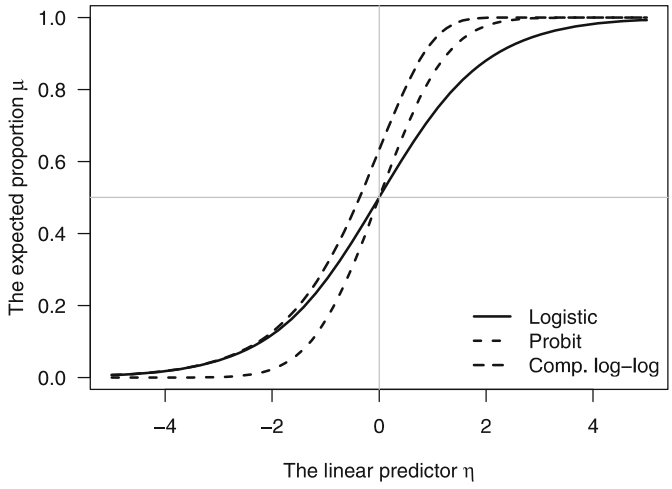

Fig. 9.2 Common link functions used with the binomial distribution: the logit, probit, and complementary log-log link functions (Sect. 9.3)

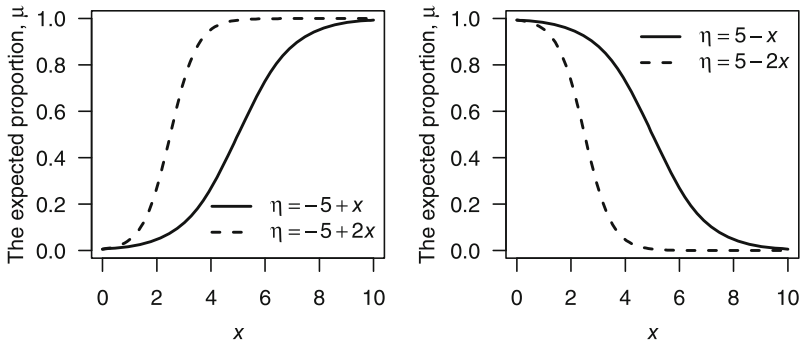

Fig. 9.3 The relationships between x and the predicted proportions μ for various linear predictors η using the logit link function, where $\text{logit}(\mu) = \eta$ (Sect. 9.3)

`"log"` link function is sometimes used for modelling risk ratios or relative risks. It is an approximation to the logit link when μ is small [16].

To understand the relationship between the explanatory variables and μ, consider the case of one explanatory variable where $\eta = \beta_0 + \beta_1 x$. Figure 9.3 shows the corresponding relationships between x and μ using the logit link function.

Example 9.2. For the turbine data (data set: `turbines`), we can fit binomial GLMs using the three common link functions, using the hours run-time as the explanatory variable:

```
> tr.logit  <- glm( Fissures/Turbines ~ Hours, data=turbines,
                    family=binomial, weights=Turbines)
> tr.probit <- update( tr.logit, family=binomial(link="probit") )
> tr.cll    <- update( tr.logit, family=binomial(link="cloglog") )
> tr.array <- rbind( coef(tr.logit), coef(tr.probit), coef(tr.cll))
> tr.array <- cbind( tr.array, c(deviance(tr.logit),
                     deviance(tr.probit), deviance(tr.cll)) )
> colnames(tr.array) <- c("Intercept", "Hours","Residual dev.")
> rownames(tr.array) <- c("Logit","Probit","Comp log-log")
> tr.array
             Intercept        Hours Residual dev.
Logit        -3.923597 0.0009992372     10.331466
Probit       -2.275807 0.0005783211      9.814837
Comp log-log -3.603280 0.0008104936     12.227914
```

The residual deviances are similar for the logit and probit GLMs, and slightly larger for the complementary log-log link function. The coefficients from the three models are reasonably different. However, the models themselves are very similar, as we can see by plotting the models. To do so, first set up a vector of values for the run-time:

```
> newHrs <- seq( 0, 5000, length=100)
```

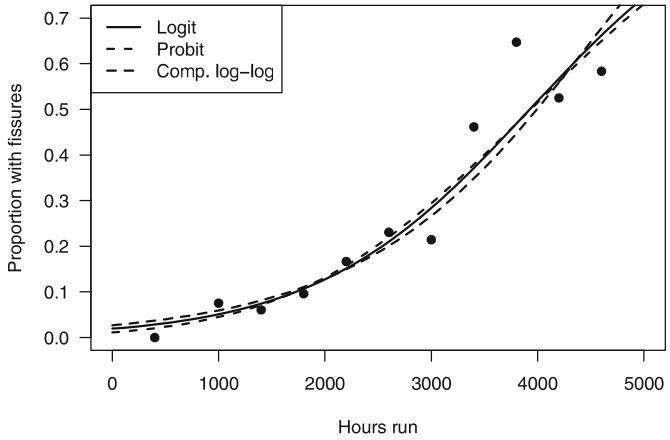

Fig. 9.4 The turbines data, showing the fitted binomial GLMs, using logistic, probit and complementary log-log link functions (Example 9.2)

Now, make predictions from these values using each model:

```
> newdf <- data.frame(Hours=newHrs)
> newP.logit  <- predict( tr.logit,  newdata=newdf, type="response")
> newP.probit <- predict( tr.probit, newdata=newdf, type="response")
> newP.cll    <- predict( tr.cll,    newdata=newdf, type="response")
```

The type of prediction is set as `"response"` because, by default, `predict()` returns the predictions on the linear predictor scale (that is, $\hat{\eta}$ is returned rather than $\hat{\mu}$). Now, plot these predictions using `lines()`, then add a legend (Fig. 9.4):

```
> plot( Fissures/Turbines ~ Hours, data=turbines, pch=19, las=1,
        xlim=c(0, 5000), ylim=c(0, 0.7),
        xlab="Hours run", ylab="Proportion with fissures")
> lines( newP.logit  ~ newHrs, lty=1, lwd=2)
> lines( newP.probit ~ newHrs, lty=2, lwd=2)
> lines( newP.cll    ~ newHrs, lty=4, lwd=2)
> legend("topleft", lwd=2, lty=c(1, 2, 4),
         legend=c("Logit","Probit","Comp. log-log"))
```

All three models produce similar predictions, which is not unusual. □

9.4 Tolerance Distributions and the Probit Link

The link functions can be understood using a threshold interpretation. In what follows, we show how the threshold interpretation applies for the probit link function, using the `turbines` data as the example.

Assume each individual turbine has a different tolerance beyond which it will develop fissures. As part of the natural variation in turbines, this

tolerance varies from turbine to turbine (but is fixed for any one turbine). Denote this tolerance level as t_i for turbine i; note that t_i is a continuous variable. Assume that t_i follows a normal distribution with mean tolerance τ_i, so that

$$\begin{cases} t_i \sim N(\tau_i, \sigma^2) \\ \tau_i = \beta_0' + \beta_1' x_i, \end{cases} \tag{9.3}$$

where x_i is the number of hours that turbine i is run. In this context, the normal distribution in (9.3) is called the *tolerance distribution*.

The variable of interest is whether or not the turbines develop fissures. Assume that turbines develop fissures if the tolerance level t_i of turbine i is less than some fixed tolerance threshold T. In other words, define

$$y_i = \begin{cases} 1 \text{ if } t_i \leq T, \text{ and the turbine develops fissures} \\ 0 \text{ if } t_i > T, \text{ and the turbine does not develop fissures.} \end{cases}$$

Then, the probability that turbine i develops fissures is

$$\mu_i = \Pr(y_i = 1) = \Pr(t_i \leq T) = \Phi\left(\frac{T - \tau_i}{\sigma}\right), \tag{9.4}$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution. We can write

$$\frac{T - \tau_i}{\sigma} = \frac{T - \beta_0' - \beta_1' x_i}{\sigma} = \beta_0 + \beta_1 x_i$$

with $\beta_0 = (T - \beta_0')/\sigma$ and $\beta_1 = -\beta_1'/\sigma$. Then (9.4) becomes

$$g(\mu_i) = \beta_0 + \beta_1 x_i$$

where $g()$ is the probit link function.

Other choices of the tolerance distribution lead to other link functions by a similar process (Table 9.2). The logit link function emerges as the link function when the logistic distribution is used as the tolerance distribution (Problem 9.4). The complementary log-log link function emerges as the link function when the extreme value (or Gumbel) distribution is used as the tolerance distribution. The cauchit link function assumes the threshold distribution is the Cauchy distribution. The logistic and normal tolerance distributions are both symmetric, and usually give similar results except for probabilities near zero or one. In contrast, the extreme value distribution is not symmetric, so the complementary log-log link function often gives somewhat different results than using the logit and probit link functions (Fig. 9.2). In principle, the CDF for any continuous distribution can be used as a basis for the link function.

Table 9.2 Tolerance distributions leading to link functions for binomial GLMs (Sect. 9.3)

Link function	Tolerance distribution	Distribution function
Logit	Logistic	$\mathcal{F}(y) = \exp(y)/\{1+\exp(y)\}$
Probit	Normal	$\mathcal{F}(y) = \Phi(y)$
Complementary log-log	Extreme value	$\mathcal{F}(y) = 1-\exp\{-\exp(y)\}$
Cauchit	Cauchy	$\mathcal{F}(y) = \{\arctan(y)+0.5\}/\pi$

9.5 Odds, Odds Ratios and the Logit Link

Using the logit link function with the binomial distribution produces a useful interpretation. To understand this interpretation, the concept of *odds* first must be understood. If event A has probability μ of occurring, then the *odds* of event A occurring is the ratio of the probability that A occurs to the probability that A does not occur: $\mu/(1-\mu)$. For example, if the probability that a turbine develops fissures is 0.6, the *odds* that a turbine develops fissures is $0.6/(1-0.6) = 1.5$. This means that the probability of observing fissures is 1.5 times greater than the probability of *not* observing a fissure (that is, $1.5 \times 0.4 = 0.6$). Clearly, using the logit link function in a binomial GLM is equivalent to modelling the logarithm of the odds (or the 'log-odds').

The binomial GLM using the logit function can be written as

$$\log(\text{odds}) = \beta_0 + \beta_1 x$$
$$\text{or equivalently} \quad \text{odds} = \exp(\beta_0)\{\exp(\beta_1)\}^x.$$

As x increases by one unit, the log-odds increase by linearly by an amount β_1. Alternatively, if x increases by one unit, the odds increase by a *factor* of $\exp(\beta_1)$. These interpretations in terms of the odds have intuitive appeal, and for this reason the logit link function is often preferred for the link function.

Example 9.3. For the turbines data (data set: `turbines`), the fitted logistic regression model (Example 9.1) has coefficients:

```
> coef(tr.logit)
  (Intercept)         Hours
-3.9235965551  0.0009992372
```

In this model, increasing `Hours` by one increases the odds of a turbine developing fissures by $\exp(0.0009992) = 1.001$. In this case, the interpretation is more useful if we consider increasing `Hours` by 1000 h. This increases the odds of a turbine developing fissures by $\exp(1000 \times 0.0009992) = 2.716$ times. Using the logistic regression model `tr.logit` assumes that the relationship between the run-time and the log-odds is approximately linear (Fig. 9.5):

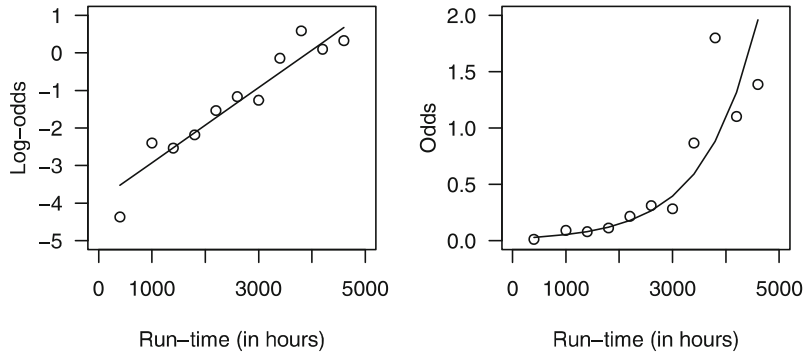

Fig. 9.5 The log-odds plotted against the run-time (left panel) and the odds plotted against the run-time (right panel) for the binomial logistic GLM fitted to the turbine data (Example 9.3)

```
> LogOdds <- predict( tr.logit ); odds <- exp( LogOdds )
> plot( LogOdds ~ turbines$Hours, type="l", las=1,
        xlim=c(0, 5000), ylim=c(-5, 1),
        ylab="Log-odds", xlab="Run-time (in hours)" )
> my <- turbines$Fissures; m <- turbines$Turbines
> EmpiricalOdds <- (my + 0.5)/(m - my + 0.5)   # To avoid log of zeros
> points( log(EmpiricalOdds) ~ turbines$Hours)
> #
> plot( odds ~ turbines$Hours, las=1, xlim=c(0, 5000), ylim=c(0, 2),
        type="l", ylab="Odds", xlab="Run-time (in hours)")
> points( EmpiricalOdds ~ turbines$Hours)
```

Note the use of empirical log-odds, adding 0.5 to both the numerator and denominator of the odds, so that the log-odds can be computed even when $y = 0$. □

Logistic regression models are often fitted to data sets that include factors as explanatory variables. In these situations, the concept of the *odds ratio* is useful to define. Consider the binomial GLM with systematic component

$$\log \frac{\mu}{1-\mu} = \text{log-odds} = \beta_0 + \beta_1 x,$$

where x is a dummy variable taking the values 0 or 1. From this equation, we see that the odds of observing a success when $x = 0$ is $\exp(\beta_0)$, and the odds of observing a success when $x = 1$ is $\exp(\beta_0 + \beta_1) = \exp(\beta_0)\exp(\beta_1)$. The ratio of these two odds is $\exp(\beta_1)$. This means that the odds of a success occurring when $x = 1$ is $\exp(\beta_1)$ times greater than when $x = 0$. This ratio is called the *odds ratio*, often written OR. When a number of factors are fitted as explanatory variables, each of the corresponding regression parameters β_j can be interpreted as odds ratios in a similar manner.

Table 9.3 The germination of two types of seeds for two root extracts. The number of seeds germinating *my* from *m* seeds planted is shown (Table 9.4)

O. aegyptiaco 75 seeds				*O. aegyptiaco* 73 seeds			
Bean extracts		Cucumber extracts		Bean extracts		Cucumber extracts	
my	*m*	*my*	*m*	*my*	*m*	*my*	*m*
10	39	5	6	8	16	3	12
23	62	53	74	10	30	22	41
23	81	55	72	8	28	15	30
26	51	32	51	23	45	32	51
17	39	46	79	0	4	3	7
		10	13				

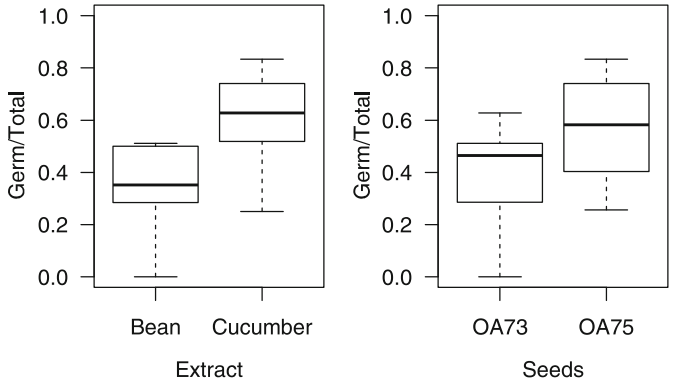

Fig. 9.6 The germination data: germination proportions plotted against extract type (left panel) and seed type (right panel) (Example 9.4)

Example 9.4. A study [3] of seed germination used two types of seeds and two types of root stocks (Table 9.3; data set: `germ`). A plot of the data (Fig. 9.6) shows possible relationships between the proportions of seeds germinating and both factors:

```
> data(germ); str(germ)
'data.frame':        21 obs. of  4 variables:
 $ Germ   : int  10 23 23 26 17 5 53 55 32 46 ...
 $ Total  : int  39 62 81 51 39 6 74 72 51 79 ...
 $ Extract: Factor w/ 2 levels "Bean","Cucumber": 1 1 1 1 1 2 2 2 2 2 ...
 $ Seeds  : Factor w/ 2 levels "OA73","OA75": 2 2 2 2 2 2 2 2 2 2 ...
> plot( Germ/Total ~ Extract, data=germ, las=1, ylim=c(0, 1) )
> plot( Germ/Total ~ Seeds,   data=germ, las=1, ylim=c(0, 1) )
```

The model with both factors as explanatory variables can be fitted:

```
> gm.m1 <- glm(Germ/Total ~ Seeds + Extract, family=binomial,
               data=germ, weights=Total)
> printCoefmat(coef(summary(gm.m1)))
```

```
                Estimate Std. Error z value  Pr(>|z|)
(Intercept)     -0.70048    0.15072 -4.6475 3.359e-06 ***
SeedsOA75        0.27045    0.15471  1.7482   0.08044 .
ExtractCucumber  1.06475    0.14421  7.3831 1.546e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Recall the R output means that the R variable `Seeds` takes the value one for `OA75` and is zero for `OA73`. Likewise the R variable `Extract` takes the value one for `Cucumber` and is zero for `Bean`.

Note that

```
> exp( coef(gm.m1) )
    (Intercept)       SeedsOA75 ExtractCucumber
      0.4963454       1.3105554       2.9001133
```

This means that the odds of seed germination occurring using cucumber extracts is 2.900 times the odds of seed germination occurring using bean extracts. Similarly, the odds of seed germination occurring using *O. aegyptiaco* 75 seeds are 1.311 times the odds of seed germination occurring using *O. aegyptiaco* 73 seeds.

These data are explored later also (Example 9.8), where the interaction term is considered. □

9.6 Median Effective Dose, ED50

Binomial GLMs are commonly used to examine the relationship between the dose d of a drug or poison and the proportion y of insects (or plants, or animals) that survive. These models are called *dose–response models*. Associated with these experiments is the concept of the median effective dose, ED50: the dose of poison affecting 50% of the insects. Different fields use different names for similar concepts, such as median lethal dose LD50 or median lethal concentration LC50. Here, for simplicity, we use ED50 to refer to any of these quantities. The ED50 concept can be applied to other contexts also. By definition, $\mu = 0.5$ at the ED50.

For a binomial GLM using a logit link function, $\eta = \text{logit}(\mu) = 0$ when $\mu = 0.5$. Writing the linear predictor as $\eta = \beta_0 + \beta_1 d$ where d is the dose, then solving for the dose d shows that $\text{ED50} = -\hat{\beta}_0/\hat{\beta}_1$. More generally, the dose effective on any proportion ρ of the population, denoted $\text{ED}(\rho)$, is estimated by

$$\text{ED}(\rho) = \frac{g(\rho) - \beta_0}{\beta_1},$$

where $g()$ refers to the link function used in fitting the model. In Problem 9.2, formulae are developed for computing ED50 for the probit and complementary log-log link functions.

The function `dose.p()` in the R package **MASS** (which comes with R distributions) conveniently returns $\widehat{\text{ED}}(\rho)$ and the corresponding estimated standard error. The first input to `dose.p()` is the `glm()` object, and the second input identifies the two coefficients of importance: the coefficient for the intercept and for the dose (in that order). By default, these are assumed to be the first and second coefficients. The third input is ρ; by default $\rho = 0.5$, and so $\widehat{\text{ED}}50$ is returned by default.

Example 9.5. Consider the turbine data again (data set: `turbines`). The ED50 corresponds to the run time for which 50% of turbines would be expected to experience fissures:

```
> library(MASS)    # MASS  comes with R
> ED50s <- cbind("Logit"     = dose.p(tr.logit),
                 "Probit"    = dose.p(tr.probit),
                 "C-log-log" = dose.p(tr.cll))
> ED50s
            Logit   Probit C-log-log
p = 0.5: 3926.592 3935.197  3993.575
```

Running the turbines for approximately 3927 h would produce fissures in about 50% of the turbines (using the logistic link function model). All three link functions produce similar estimates of ED50, which seems reasonable based on Fig. 9.4 (p. 338). □

9.7 The Complementary Log-Log Link in Assay Analysis

A common problem in biology is to determine the proportion of cells or organisms of interest amongst a much larger population. For example, does a sample of tissue contain infective bacteria, and how many? Or what is the frequency of adult stem cells in a sample of tissue?

Suppose the presence of active particles can be detected by undertaking an assay. For example, the presence of bacteria might be detected by incubating the sample on an agar plate, and observing whether a bacterial culture grows. Or the presence of stem cells might be detected by transplanting cells into a host animal, and observing whether a new growth occurs. However, the same response is observed, more or less, regardless of the number of active particles in the original sample. A single stem cell would result in a new growth. When a growth is observed, we cannot determine directly whether there was one stem cell or many to start with.

Dilution assays are an experimental technique to estimate the frequency of active cells. The idea is to dilute the sample down to the point where some assays yield a positive result (so at least one active particles is present) and some yield a negative result (so no active particles are present).

The fundamental property of limiting dilution assays is that each assay results in a positive or negative result. Write μ_i for the probability of a

positive result given that the expected number of cells in the culture is d_i. If m_i independent cultures are conducted at dose d_i, then the number of positive results follows a binomial distribution.

Write λ for the proportion of active cells in the cell population, so that the expected number of active cells in the culture is λd_i. If the cells behave independently (that is, if there are no community effects amongst the cells), and if the cell dose is controlled simply by dilution, then the actual number of cells in each culture will vary according to a Poisson distribution. A culture will give a negative result only if there are no active cells in the assay. The Poisson probability formula tells us that this occurs with probability

$$1 - \mu_i = \exp(-\lambda d_i).$$

This formula can be linearized by taking logarithms of both sides, as

$$\log(1 - \mu_i) = -\lambda d_i \tag{9.5}$$

or, taking logarithms again,

$$\log\{-\log(1 - \mu_i)\} = \log\lambda + \log d_i. \tag{9.6}$$

This last formula is the famous complementary log-log transformation from Mather [18].

The proportion of active cells can be estimated by fitting a binomial GLM with a complementary log-log link:

$$g(\mu_i) = \beta_0 + \log d_i \tag{9.7}$$

where $\log d_i$ is an offset and $g()$ is the complementary log-log link function. The estimated proportion of active cells is then $\hat{\lambda} = \exp(\hat{\beta}_0)$.

In principle, a GLM could also have be fitted using (9.5) as a link-linear predictor, in this case with a log-link. However (9.6) is superior, because it leads to a GLM (9.7) without any constraints on the coefficient β_0.

As usual, a confidence interval is given by

$$\hat{\beta}_0 \pm z_{\alpha/2}\text{se}(\hat{\beta}_0)$$

where $\text{se}(\hat{\beta}_0)$ is the standard error of the estimate and $z_{\alpha/2}$ is the critical value of the normal distribution, e.g., $z = 1.96$ for a 95% confidence interval. To get back to the active cell frequency simply exponentiate and invert the estimate and the confidence interval: $1/\hat{\lambda} = \exp(-\hat{\beta}_0)$. Confidence intervals can be computed for $1/\lambda$, representing the number of cells required on average to obtain one responding cell.

The dilution assay model assumes that a single active cell is sufficient to achieve a positive result, so it is sometimes called the *single-hit* model (though other assumptions are possible [25]). One way to check this model is

Table 9.4 The average number of cells in each assay in which cells were transplanted in host mice, the number of assays at that cell number, and the number of assays giving a positive outcome, a milk gland outgrowth (Example 9.6)

Number of cells per assay	Number of assays	Number of outgrowths
15	38	3
40	6	6
60	17	13
90	8	6
125	12	9

to fit a slightly larger model in which the offset coefficient is not set to one:

$$g(\mu_i) = \beta_0 + \beta_1 \log d_i.$$

The correctness of the single-hit model can then be checked [10] by testing the null hypothesis H_0: $\beta_1 = 1$.

Example 9.6. Shackleton et al. [21] demonstrated the existence of adult mammary stem cells. They showed, for the first time, that a complete mammary milk producing gland could be produced in mice from a single cell. After a series of steps, they were able to purify a population of cells that was highly enriched for mammary stem cells, although stem cells were still a minority of the total.

The data (Table 9.4; data set: `mammary`) relate to a number of assays in which cells were transplanted into host mice. A positive outcome here consists of seeing a milk gland outgrowth, evidence that the sample of cells included as least one stem cell. The data give the average number of cells in each assay, the number of assays at that cell number, and the number of assays giving a positive outcome.

```
> data(mammary); mammary
  N.Cells N.Assays N.Outgrowths
1      15       38            3
2      40        6            6
3      60       17           13
4      90        8            6
5     125       12            9
> y <- mammary$N.Outgrowths / mammary$N.Assays
> fit <- glm(y~offset(log(N.Cells)), family=binomial(link="cloglog"),
             weights=N.Assays, data=mammary)
> coef(summary(fit))
             Estimate Std. Error   z value      Pr(>|z|)
(Intercept) -4.163625  0.1744346 -23.86925 6.391454e-126
> frequency <- 1/exp(coef(fit)); frequency
(Intercept)
   64.30418
```

The mammary stem cell frequency is estimated to be about 1 in 64 cells. A 95% confidence interval is computed as follows:

```
> s <- summary(fit)
> Estimate <- s$coef[, "Estimate"]
> SE <- s$coef[, "Std. Error"]
> z <- qnorm(0.05/2, lower.tail=FALSE)
> CI <- c(Lower=Estimate+z*SE, Estimate=Estimate, Upper=Estimate-z*SE)
> CI <- 1/exp(CI); round(CI, digits=1)
   Lower Estimate    Upper
    45.7     64.3     90.5
```

The frequency of stem cells is between 1/46 and 1/91. There is no evidence of any deviation from the single-hit model:

```
> fit1 <- glm(y~log(N.Cells), family=binomial(link="cloglog"),
              weights=N.Assays, data=mammary)
> anova(fit, fit1, test="Chisq")
Analysis of Deviance Table

Model 1: y ~ offset(log(N.Cells))
Model 2: y ~ log(N.Cells)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         4     16.852
2         3     16.205  1   0.6468   0.4213
```

□

9.8 Overdispersion

For a binomial distribution, $\text{var}[y] = \mu(1-\mu)$. However, in practice the amount of variation in the data can exceed $\mu(1-\mu)$, even for ostensibly binomial-like data. This is called *overdispersion*. Underdispersion also occurs, but is less common.

Overdispersion has serious consequences for the GLM. It means that standard errors returned by the GLM are underestimated, and tests on the explanatory variables will generally appear to be more significant that warranted by the data, leading to overly complex models.

Overdispersion is detected by conducting a goodness-of-fit test, as described in Sect. 7.4. If the residual deviance and Pearson statistics are much greater than the residual degrees of freedom, then there is evidence of lack of fit. Lack of fit may be caused by an inadequate model, for example because important explanatory variables are missing from the model. However, if all relevant or possible explanatory variables are already included in the model, and the data has been checked for outliers that might inflate the residuals, but lack of fit remains, then overdispersion is the alternative interpretation.

Overdispersion means that the binomial model is incorrect in some respect. Overdispersion can arise from two major causes. The probabilities μ_i

are not constant between observations, even when all the explanatory variables are unchanged. Alternatively the m_i cases, of which observation y_i is a proportion, are not independent.

The first type of overdispersion can be modelled by a hierarchical model. Suppose that $m_i y_i$ follows a binomial distribution with m_i cases and success probability p_i. Suppose that the p_i is itself a random variable, with mean μ_i. Then

$$\mathrm{E}[y_i] = \mu_i$$

but

$$\mathrm{var}[y_i] > \mu_i(1-\mu_i)/m_i.$$

The greater the variability of p_i the greater the degree of overdispersion. A commonly-used model is to assume that p_i follows a beta distribution [3]. This leads to a beta-binomial model for y_i in which

$$\mathrm{var}[y_i] = \phi_i \mu_i(1-\mu_i)/m_i, \tag{9.8}$$

where ϕ_i depends on m_i and the parameters of the beta distribution.

More generally, overdispersion arises when the m_i Bernoulli cases, that make up observation y_i, are positively correlated. For example, positive cases may arrive in clusters rather than as individual cases. Writing ρ for the correlation between the Bernoulli trials leads to the same variance as the beta-binomial model (9.8) with $\phi_i = 1 + (m_i - 1)\rho$. If the m_i are approximately equal, or if ρ is inversely proportional to $m_i - 1$, then the ϕ_i will be approximately equal. In this case, both overdispersion models lead to variances

$$\mathrm{var}[y_i] = \phi \mu_i(1-\mu_i)/m_i,$$

which are larger but proportional to the variances under the binomial model. Note that overdispersion cannot arise for binary data with $m_i = 1$.

This reasoning leads to the idea of quasi-binomial models (Sect. 8.10). Quasi-binomial models keep the same variance function $V(\mu) = \mu(1-\mu)$ as binomial GLMs, but allow a general positive dispersion ϕ instead of assuming $\phi = 1$. The dispersion parameter is usually estimated by the Pearson estimator (Sect. 6.8.5). Quasi-binomial models do not correspond to any EDM, but the quasi-likelihood theory of Sect. 8.10 provides reassurance that the model will still yield consistent estimators provided that the variance function represents the correct mean–variance relationship. In particular, quasi-binomial models will give consistent estimators of the model coefficients under the beta-binomial or correlation models described above when the m_i are roughly equal. Even when the m_i are not equal, a quasi-binomial model is likely still preferable to assuming $\phi = 1$ when overdisperion is present.

The parameter estimates for binomial and quasi-binomial GLMs are identical (since the estimates $\hat{\beta}_j$ do not depend on ϕ), but the standard errors are different. The effect of using the quasi-binomial model is to inflate the standard error of the parameter estimates by $\sqrt{\phi}$, so confidence intervals and statistics for testing hypotheses tests will change.

A quasi-binomial model is fitted in R using `glm()` by using `family=quasibinomial()`. As for `family=binomial()`, the default link function for the `quasibinomial()` family is the `"logit"` link, while `"probit"`, `"cloglog"`, `"cauchit"`, and `"log"` are also permitted. Since the quasi-binomial model is not based on a probability model, the AIC is undefined.

Example 9.7. Machine turbines operate more or less independently, so it seems reasonable to suppose that independence between Bernoulli trials might hold for the turbines data (data set: `turbines`). Indeed neither the residual deviance nor the Pearson statistics show any evidence of overdispersion (using model `tr.logit` fitted in Example 9.1):

```
> c(Df = df.residual( tr.logit ),
    Resid.Dev = deviance( tr.logit ),
    Pearson.X2 = sum( resid(tr.logit, type="pearson")^2 ))
       Df  Resid.Dev Pearson.X2
 9.000000  10.331466   9.250839
```

Neither goodness-of-fit statistic is appreciably larger than the residual degrees of freedom. This data set does contain two small values of $m_i y_i$, but these are too few to change the conclusion even if the residuals for these observations were underestimated. □

Example 9.8. Example 9.4 (p. 341) discussed the seed germination for two types of seeds and two types of root stocks (data set: `germ`). Since seeds are usually planted together in common plots, it is highly possible that they might interact or be affected by common causes; in other words we might well expect seeds to be positively correlated, leading to overdispersion. We start by fitting a binomial GLM with `Extract` and `Seed` and their interaction as explanatory variables:

```
> gm.m1 <- glm( Germ/Total ~ Extract * Seeds, family=binomial,
                weights=Total, data=germ )
> anova(gm.m1, test="Chisq")
             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                              20     98.719
Extract       1   55.969        19     42.751 7.364e-14 ***
Seeds         1    3.065        18     39.686   0.08000 .
Extract:Seeds 1    6.408        17     33.278   0.01136 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> df.residual(gm.m1)
[1] 17
```

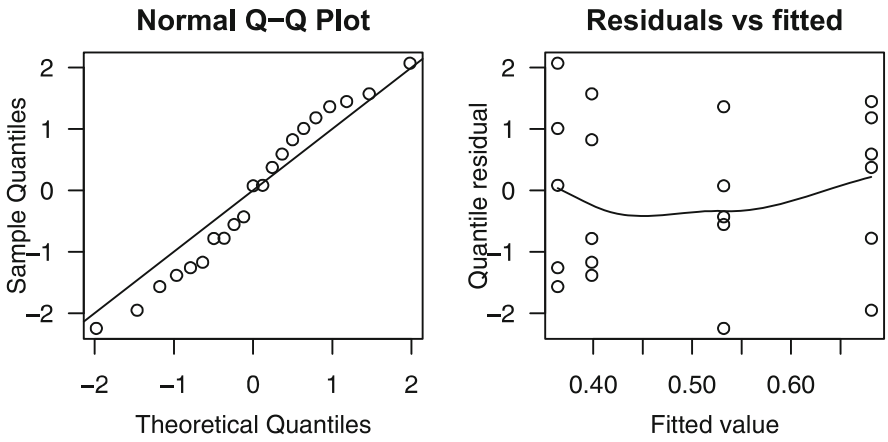

Fig. 9.7 Diagnostic plots after fitting a binomial GLM to the seed germination data (Example 9.8)

Despite the fact that the maximal possible explanatory model has been fitted, overdispersion is clearly present; the residual deviance is much larger than the residual degrees of freedom:

```
> c( deviance(gm.m1), df.residual(gm.m1) )
[1] 33.27779 17.00000
```

The Pearson statistic tells the same story:

```
> sum( resid(gm.m1, type="pearson")^2 ) # Pearson.X2
[1] 31.65114
```

There are no large residuals present that would suggest outliers (Fig. 9.7):

```
> library(statmod)
> qres <- qresid(gm.m1); qqnorm(qres, las=1); abline(0, 1)
> scatter.smooth( qres~fitted(gm.m1), las=1, main="Residuals vs fitted",
      xlab="Fitted value", ylab="Quantile residual")
```

The chi-square approximation to the goodness-of-fit statistics seems good enough. The data includes one observation (number 16) with $my = 0$ and other with $m - my = 1$ (number 6), but neither has a large enough residual to be responsible for the apparent overdispersion:

```
> qres[c(6, 16)]
[1]  1.180272 -1.172095
```

Finally, this a designed experiment, with nearly equal numbers of observations in each combination of the experimental factors `Extract` and `Seeds`, so influential observations cannot be an issue.

Having ruled out all alternative explanations, we accept that overdispersion is present and fit a quasi-binomial model:

```
> gm.od <- update(gm.m1, family=quasibinomial)
> anova(gm.od, test="F")
             Df Deviance Resid. Df Resid. Dev       F    Pr(>F)
NULL                           20     98.719
Extract       1   55.969        19     42.751 30.0610 4.043e-05 ***
Seeds         1    3.065        18     39.686  1.6462   0.21669
Extract:Seeds 1    6.408        17     33.278  3.4418   0.08099 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that F-tests are used for comparisons between quasi-binomial models. This follows because the dispersion ϕ is estimated (using the Pearson estimator by default). The quasi-binomial analysis of deviance suggests that only `Extract` is significant in the model, so germination frequency differs by root stock but not by seed type, unlike the binomial GLM which showed a significant `Extract` by `Seeds` interaction.

The binomial and quasi-binomial GLMs give identical coefficient estimates, but the standard errors from the quasi-binomial GLM are $\sqrt{\phi}$ times those from the binomial model:

```
> sqrt(summary(gm.od)$dispersion)
[1] 1.36449
> beta <- coef(summary(gm.m1))[,"Estimate"]
> m1.se <- coef(summary(gm.m1))[,"Std. Error"]
> od.se <- coef(summary(gm.od))[,"Std. Error"]
> data.frame(Estimate=beta, Binom.SE=m1.se,
             Quasi.SE=od.se, Ratio=od.se/m1.se)
                            Estimate  Binom.SE  Quasi.SE   Ratio
(Intercept)                -0.4122448 0.1841784 0.2513095 1.36449
ExtractCucumber             0.5400782 0.2498130 0.3408672 1.36449
SeedsOA75                  -0.1459269 0.2231659 0.3045076 1.36449
ExtractCucumber:SeedsOA75   0.7781037 0.3064332 0.4181249 1.36449
```

□

9.9 When Wald Tests Fail

Standard errors and Wald tests experience special difficulties when the fitted values from binomial GLMs are very close to zero or one. When the linear predictor includes factors, sometimes in practice there is a factor level for which the y_i are either all zero or all one. In this situation, the fitted values estimated by the model will also be zero or one for this level of the factor. This situation inevitably causes problems for standard errors and Wald tests, because at least one of the coefficients in the linear predictor must tend to infinity as the fitted model converges.

Suppose for example that the logit link function is used, so the fitted values are related to the linear predictor by

$$\hat{\mu} = \frac{\exp(\hat{\eta})}{1 + \exp(\hat{\eta})}. \tag{9.9}$$

Suppose also that the model includes just one explanatory variable x, so $\eta = \beta_0 + \beta_1 x$. The only way for $\hat{\mu}$ to be zero or one is for $\hat{\eta}$ to be $\pm\infty$. If $\hat{\mu} \to 0$, then $\hat{\eta} \to -\infty$, which implies $\hat{\beta}_0 \to -\infty$ and/or $\hat{\beta}_1 x \to -\infty$. In other words, one or both of the parameters must approach $\pm\infty$. If $\hat{\mu} \to 1$, then $\hat{\eta} \to \infty$ and a similar situation exists. The phenomenon is the same for other link functions.

When parameter estimates approach $\pm\infty$, the standard errors for those parameters must also approach $\pm\infty$, and Wald test statistics, which are ratios of coefficients to standard errors (Sect. 7.2.1), become very unreliable [23, p. 197]. In particular, the standard errors often tend to infinity faster than the coefficients themselves, meaning that the Wald statistic tends to zero, regardless of the true significance of the variable. This is called the *Hauck–Donner effect* [7].

Despite the problems with Wald tests, the likelihood ratio and score test usually remain quite serviceable in these situations, even when fitted values are zero or one. This is because the problem of infinite parameters is removable, in principle, by re-parametrising the model, and likelihood ratio and score tests are invariant to reparameterization. Wald tests are very susceptible to infinite parameters in the model because they are dependent on the particular parameterization used.

Example 9.9. A study [17] of the habitats of the noisy miner (a small but aggressive native Australian bird) recorded whether noisy miners were detected in various two hectare transects in buloke woodland patches (data set: `nminer`). Part of this data frame was discussed in Example 1.5 (p. 14), where models were fitted for the *number* of noisy miners.

Here we consider fitting a binomial GLM to model the presence of noisy miners in each buloke woodland patch (`Miners`). More specifically, we study whether the presence of noisy miners is impacted by whether or not the number of eucalypts exceeds 15 or not:

```
> data(nminer); Eucs15 <- nminer$Eucs>15
> m1 <- glm(Miners ~ Eucs15, data=nminer, family=binomial)
> printCoefmat(coef(summary(m1)))
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.84730    0.48795 -1.7364  0.08249 .
Eucs15TRUE    20.41337 3242.45694  0.0063  0.99498
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Wald test results indicate that the explanatory variable is not significant: $P = 0.995$. Note the large standard error for `Eucs15`. Compare to the likelihood ratio test results:

```
> anova(m1, test="Chisq")
      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     30     42.684
Eucs15  1    18.25        29     24.435 1.937e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The likelihood ratio test results indicate that the explanatory variable is highly significant: $P \approx 0$. Similarly, the score test results indicate that `Miners` is highly significant also:

```
> m0 <- glm(Miners ~ 1, data=nminer, family=binomial)
> z.score <- glm.scoretest(m0, Eucs15)
> P.score <- 2*(1-pnorm(abs(z.score))); c(z.score, P.score)
[1] 3.7471727820 0.0001788389
```

Despite the Wald test results, a plot of `Miners` against `Eucs15` (Fig. 9.8) shows an obvious relationship: in woodland patches with more than 15 eucalypts, noisy miners were *always* observed:

```
> plot( factor(Miners, labels=c("No","Yes")) ~ factor(Eucs15), las=1,
    ylab="Noisy miners present?", xlab="Eucalypts > 15", data=nminer)
> plot( Miners ~ Eucs, pch=ifelse(Eucs15, 1, 19), data=nminer, las=1)
> abline(v=15.5, col="gray")
```

The situation is exactly as described in the text, and an example of the Hauck–Donner effect. This means that the Wald test results are not trustworthy. When the number of eucalypts exceeds 15, all woodland patches in the sample have noisy miners, so $\hat{\mu} \to 1$. This is achieved as $\hat{\beta}_1 \to \infty$. The fitted probability when `Eucs15` is `TRUE` is one to computer precision:

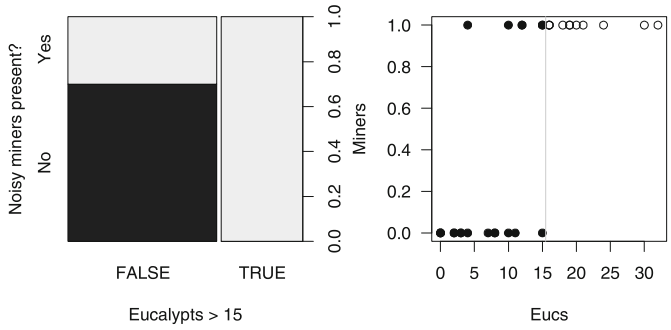

Fig. 9.8 The presence of noisy miners. Left panel: the presence of noisy miners as a function of whether 15 eucalypts are observed or not; right panel: the presence of noisy miners as a function of the number of eucalypts, showing the division at 15 eucalypts (Example 9.9)

```
> tapply(fitted(m1), Eucs15, mean)
FALSE  TRUE
  0.3   1.0
```

In this situation, the score or likelihood ratio tests must be used instead of the Wald test. □

9.10 No Goodness-of-Fit for Binary Responses

When $m_i = 1$ for all i, the binomial responses y_i are all 0 or 1; that is, the data are binary. In this case the residual deviance and Pearson goodness-of-fit statistics are determined entirely by the fitted values. This means that there is no concept of residual variability, and goodness-of-fit tests are not meaningful. For binary data, likelihood ratio tests and score tests should be used, making sure that p' is much smaller than n.

Example 9.10. In the `nminer` example in the previous section, the residual deviance is less than the residual degrees of freedom. This might be thought to suggest underdispersion, but it has no meaning. The size of the residual deviance is determined only by the sizes of the fitted values, and how far they are from zero and one. □

9.11 Case Study

An experiment [8, 13] exposed batches of insects to various deposits (in mg) of insecticides (Table 9.5; data set: `deposit`). The proportion of insects y killed after six days of exposure in each batch of size m is potentially a function of the dose of insecticide and the type of insecticide. The data are available in the R package **GLMsData**:

Table 9.5 The number of insects killed $z_i = y_i m_i$ out of a total of m_i insects, after three days exposure to different deposits of insecticides (Sect. 9.11)

	Amount of deposit (in mg)											
	2.00		2.64		3.48		4.59		6.06		8.00	
Insecticide	z_i	m_i	z_i	m_i	z_i	m_i	z_i	m_i	z_i	m_i	z_i	m_i
A	3	50	5	49	19	47	19	38	24	29	35	50
B	2	50	14	49	20	50	27	50	41	50	40	50
C	28	50	37	50	46	50	48	50	48	50	50	50

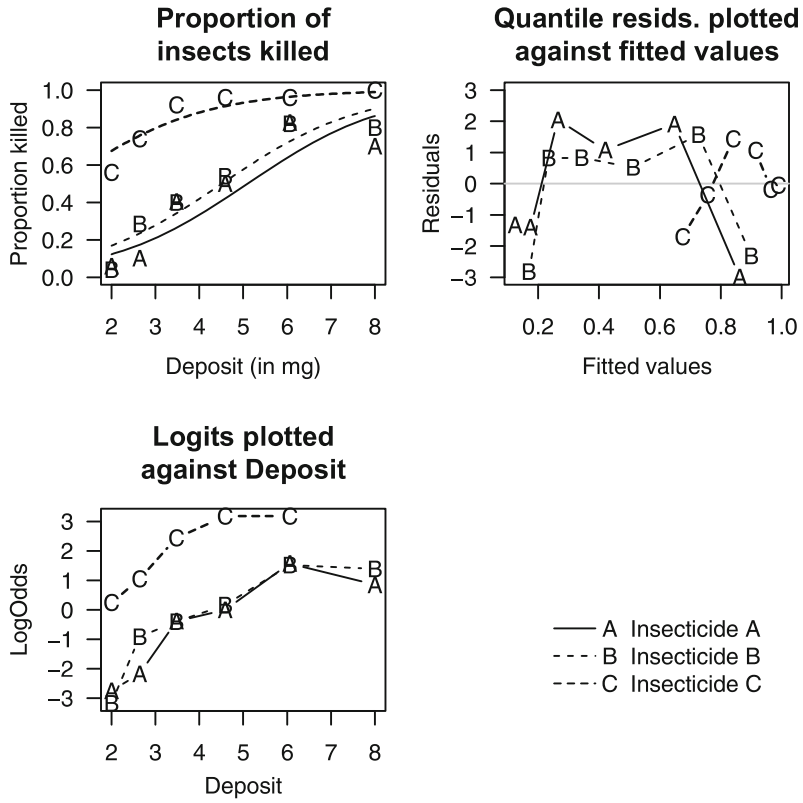

Fig. 9.9 The insecticide data. Top left panel: the data, showing the fitted model `ins.m1`; top right panel: a plot of the quantile residuals against the fitted values; bottom panel: the log-odds plotted against the deposit (Sect. 9.11)

```
> data(deposit); str(deposit)
'data.frame':        18 obs. of  4 variables:
 $ Killed     : int  3 5 19 19 24 35 2 14 20 27 ...
 $ Number     : int  50 49 47 38 29 50 50 49 50 50 ...
 $ Insecticide: Factor w/ 3 levels "A","B","C": 1 1 1 1 1 1 2 2 2 2 ...
 $ Deposit    : num  2 2.64 3.48 4.59 6.06 8 2 2.64 3.48 4.59 ...
```

A plot of the data (Fig. 9.9, p. 355, top left panel) shows insecticides A and B appear to have similar effects, while insecticide C appears different from A and B. The amount of deposit clearly is significant:

```
> deposit$Prop <- deposit$Killed / deposit$Number
> plot( Prop ~ Deposit, type="n", las=1, ylim=c(0, 1),
     data=deposit, main="Proportion of\ninsects killed",
     xlab="Deposit (in mg)", ylab="Proportion killed")
> points( Prop ~ Deposit, pch="A", subset=(Insecticide=="A"), data=deposit)
> points( Prop ~ Deposit, pch="B", subset=(Insecticide=="B"), data=deposit)
> points( Prop ~ Deposit, pch="C", subset=(Insecticide=="C"), data=deposit)
```

A model using the deposit amount and the type of insecticide as explanatory variables seems sensible:

```
> ins.m1 <- glm(Killed/Number ~ Deposit + Insecticide,
       family = binomial, weights = Number, data = deposit)
> coef(ins.m1)
 (Intercept)      Deposit InsecticideB InsecticideC
  -3.2213638    0.6316762    0.3695267    2.6880162
```

The fitted lines are shown in the top left panel of Fig. 9.9:

```
> newD <- seq( min(deposit$Deposit), max(deposit$Deposit), length=100)
> newProp.logA <- predict(ins.m1, type="response",
     newdata=data.frame(Deposit=newD, Insecticide="A") )
> newProp.logB <- predict(ins.m1, type="response",
     newdata=data.frame(Deposit=newD, Insecticide="B") )
> newProp.logC <- predict(ins.m1, type="response",
     newdata=data.frame(Deposit=newD, Insecticide="C") )
> lines( newProp.logA ~ newD, lty=1); lines( newProp.logB ~ newD, lty=2)
> lines( newProp.logC ~ newD, lty=3)
```

Before evaluating this model, we pause to demonstrate the estimation of ED50. The function `dose.p()` requires the name of the model, and the location of the coefficients that refer to the intercept and the slope. For insecticide A:

```
> dose.p(ins.m1, c(1, 2))
             Dose        SE
p = 0.5: 5.099708 0.2468085
```

For other insecticides, the intercept term is not contained in a single parameter. However, consider fitting an equivalent model:

```
> ins.m1A <- update( ins.m1, .~. - 1)   # Do not fit a constant term
> coef( ins.m1A )
     Deposit InsecticideA InsecticideB InsecticideC
   0.6316762   -3.2213638   -2.8518371   -0.5333477
```

Fitting the model without β_0 forces R to fit a model with separate intercept terms for each insecticide. Then, being careful to give the location of the intercept term first:

```
> ED50s <- cbind( dose.p(ins.m1A, c(2, 1)),    dose.p(ins.m1A, c(3, 1)),
                  dose.p(ins.m1A, c(4, 1)) )
> colnames(ED50s) <- c("Insect. A","Insect. B", "Insect. C"); ED50s
         Insect. A Insect. B Insect. C
p = 0.5:  5.099708  4.514714 0.8443372
```

Returning now to the diagnostic analysis of the model, close inspection of the top left panel in Fig. 9.9 shows model `ins.m1` is inadequate. The pattern in the residuals is easier to see in the top right panel:

```
> library(statmod)     # For  qresid()
> plot( qresid(ins.m1) ~ fitted(ins.m1), type="n", las=1, ylim=c(-3, 3),
     main="Quantile resids. plotted\nagainst fitted values",
     xlab="Fitted values", ylab="Residuals")
> abline(h = 0, col="grey")
> points( qresid(ins.m1) ~ fitted(ins.m1), pch="A", type="b", lty=1,
           subset=(deposit$Insecticide=="A") )
> points( qresid(ins.m1) ~ fitted(ins.m1), pch="B", type="b", lty=2,
           subset=(deposit$Insecticide=="B") )
> points( qresid(ins.m1) ~ fitted(ins.m1), pch="C", type="b", lty=3,
           subset=(deposit$Insecticide=="C"))
```

For each insecticide, the proportions are under-estimated at the lower and higher values of deposit. Plotting the log-odds against the deposit shows the relationship is not linear on the log-odds scale (Fig. 9.9, bottom panel):

```
> LogOdds <- with(deposit, log(Prop/(1-Prop)) )
> plot( LogOdds ~ Deposit, type="n", xlab="Deposit", data=deposit,
     main="Logits plotted\nagainst Deposit", las=1)
> points( LogOdds ~ Deposit, pch="A", type="b", lty=1,
           data=deposit, subset=(Insecticide=="A") )
> points( LogOdds ~ Deposit, pch="B", type="b", lty=2,
           data=deposit, subset=(Insecticide=="B") )
> points( LogOdds ~ Deposit, pch="C", type="b", lty=3,
           data=deposit, subset=(Insecticide=="C") )
```

As suggested earlier (Sect. 9.2), the *logarithm* of the dose is commonly used in dose–response models, so we try such a model (Fig. 9.10, top left panel):

```
> deposit$logDep <- log( deposit$Deposit )
> ins.m2 <- glm(Killed/Number ~ logDep + Insecticide - 1,
       family = binomial, weights = Number, data = deposit)
```

The ED50 estimates are on the log-scale for this model:

```
> ED50s <- cbind( dose.p(ins.m2, c(2, 1)),    dose.p(ins.m2, c(3, 1)),
                  dose.p(ins.m2, c(4, 1)) )
> colnames(ED50s) <- c("Insect. A","Insect. B", "Insect. C"); exp(ED50s)
         Insect. A Insect. B Insect. C
p = 0.5:  4.688232  4.154625  1.753202
```

The ED50 estimates are quite different from those computed using model `ins.m1A`.

While model `ins.m2` is an improvement over model `ins.m1`, proportions are still under-estimated for all types at the lower and higher values of deposit (Fig. 9.10, top right panel).

Plotting the log-odds against the logarithm of `Deposit` indicates that the log-odds are not constant, but are perhaps quadratic (Fig. 9.10, bottom panel; code not shown). Because of this, we try this model:

```
> ins.m3 <- glm(Killed/Number ~ poly(logDep, 2) + Insecticide,
       family = binomial, weights = Number, data = deposit)
```

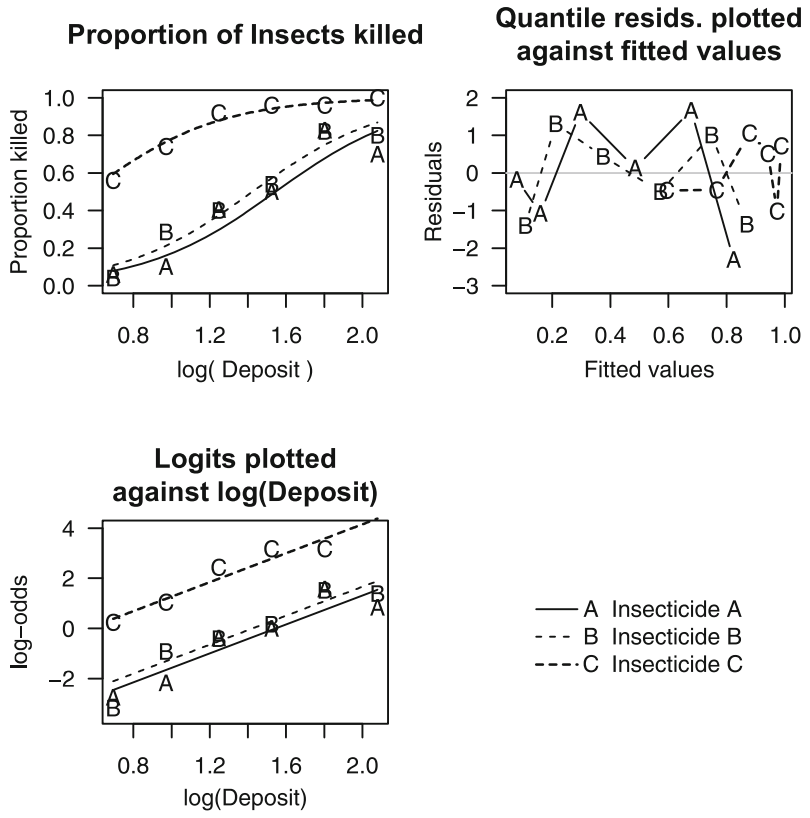

Fig. 9.10 The binomial GLMs for the insecticide data using the *logarithm of deposit* as an explanatory variable in model `ins.m2`. Top left panel: the log-odds against the logarithm of deposit showing the fitted models; top right panel: the quantile residuals plotted against the fitted values; bottom panel: the log-odds plotted against the logarithm of deposit (Sect. 9.11)

Now compare the two models involving `logDep`:

```
> anova( ins.m2, ins.m3, test="Chisq")
Analysis of Deviance Table

Model 1: Killed/Number ~ logDep + Insecticide - 1
Model 2: Killed/Number ~ poly(logDep, 2) + Insecticide
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        14     23.385
2        13     15.090  1   8.2949 0.003976 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

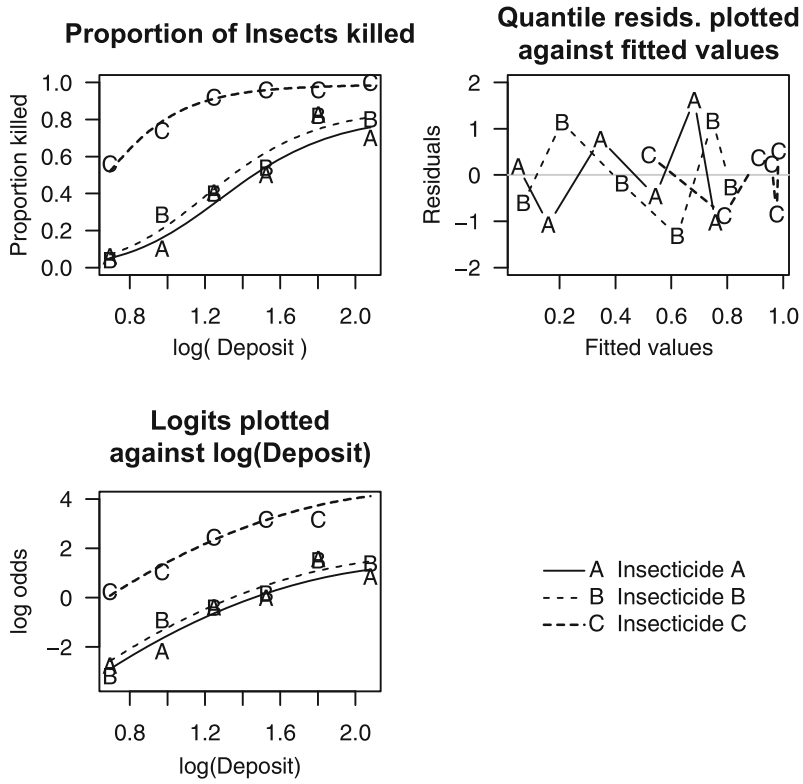

Fig. 9.11 The binomial GLMs for the insecticide data using the *square of the logarithm of deposit* as an explanatory variable in model `ins.m3`. Top left panel: the log-odds against the logarithm of deposit showing the fitted models; top right panel: the quantile residuals plotted against the fitted values; bottom panel: the log-odds plotted against the logarithm of deposit (Sect. 9.11)

This quadratic model is a statistically significantly improvement; the plotted lines appear much better (Fig. 9.11):

```
> newD <- seq( min(deposit$logDep), max(deposit$logDep), length=200)
> newProp4.logA <- predict(ins.m3, type="response",
     newdata=data.frame(logDep=newD, Insecticide="A") )
> newProp4.logB <- predict(ins.m3, type="response",
     newdata=data.frame(logDep=newD, Insecticide="B") )
> newProp4.logC <- predict(ins.m3, type="response",
     newdata=data.frame(logDep=newD, Insecticide="C") )
> lines( newProp4.logA ~ newD, lty=1); lines( newProp4.logB ~ newD, lty=2)
> lines( newProp4.logC ~ newD, lty=3)
```

The ED50 for this quadratic model cannot be computed using `dose.p` (because of the quadratic term in `logDep`), but can be found using simple algebra (Problem 9.3).

The structural changes to the model show that the model now is adequate (diagnostic plots not shown). No evidence exists to support overdispersion:

```
> c( deviance( ins.m3 ), df.residual( ins.m3 ) )
[1] 15.09036 13.00000
```

However, the saddlepoint approximation is probably not satisfactory and so this conclusion may not be entirely trustworthy:

```
> c( min( deposit$Killed ), min( deposit$Number - deposit$Killed) )
[1] 2 0
```

9.12 Using R to Fit GLMs to Proportion Data

Binomial GLMs are fitted in R using `glm()` with `family=binomial()`. The link functions `"logit"` (the default), `"probit"`, `"cloglog"` (the complementary log-log), `"log"` and `"cauchit"` are permitted. The response for a binomial GLM can be supplied in one of three ways:

- `glm( y ~ x, weights=m, family=binomial)`, where `y` are the observed proportions of successes in `m` trials.
- `glm( cbind(success, fail) ~ x, family=binomial)`, where `success` is a column of the number of successes, and `fail` is a column of the corresponding number of failures.
- `glm( fac ~ x, family=binomial)`, where `fac` is a factor. The first level denotes failure and all other levels denote successes, or where `fac` consists of logicals (either `TRUE`, which is treated as the success, or `FALSE`). Each individual in the study is represented by one row. This fits a Bernoulli GLM.

9.13 Summary

Chapter 9 considers fitting binomial GLMs. Proportions may be modelled using the binomial distribution (Sect. 9.2) where μ is the expected proportion where $0 < \mu < 1$, and $y = 0, 1/m, 2/m, \ldots, 1$. The prior weights are $w = m$. The residual deviance is suitably described by a $\chi^2_{n-p'}$ distribution if $\min\{m_i\mu_i\} \geq 3$ and $\min\{m_i(1-\mu_i)\} \geq 3$.

Commonly-used link functions are the logit (the canonical link function), probit and complementary log-log link functions (Sects. 9.3 and 9.4). Using the logistic link function enables an interpretation in terms of odds $\mu/(1-\mu)$ and odds ratios (OR) (Sect. 9.5).

The median effective dose (ED50) is the value of the covariates when the expected proportion is $\mu = 0.5$ (Sect. 9.6).

Overdispersion is observed when the variation in the data is greater than expected under the binomial model (Sect. 9.8). If overdispersion is observed, a quasi-binomial model may be fitted, which assumes $V(\mu) = \phi\mu(1-\mu)$. Overdispersion causes the estimates of the standard error to be underestimated and confidence intervals for parameters to be too narrow (Sect. 9.8).

For binomial GLMs, the Wald tests can fail in circumstances where one or more of the regression parameters tend to $\pm\infty$ (Sect. 9.9).

Problems

Selected solutions begin on p. 539.

9.1. Suppose the proportion y has the binomial distribution so that $z \sim \text{Bin}(\mu, m)$ where $z = my$ is the number of successes. Show that the transformation $y^* = \sin^{-1}\sqrt{y}$ produces approximately constant variance, by first expanding the transformation about μ using a Taylor series. (HINT: Follow the steps outlined in Sect. 5.8.)

9.2. Suppose that a given dose–response experiment records the dose of poison d and proportion y of insects out of m that are killed at each dose, such that the model has the systematic component $g(\eta) = \beta_0 + \beta_1 d$.

1. Show that the ED50 for such a model using a probit link function is $\text{ED50} = -\beta_0/\beta_1$.
2. Show that the ED50 for such a model using the complementary log-log link function is $\text{ED50} = \{\log(\log 2) - \beta_0\}/\beta_1$.
3. Show that the ED50 for such a model using the logarithmic link function is $\text{ED50} = (\log 0.5 - \beta_0)/\beta_1$.

9.3. Consider a binomial GLM using a logistic link function with systematic component $\text{logit}(\mu) = \beta_0 + \beta_1 \log x + \beta_2 (\log x)^2$.

1. For this model, deduce a formula for estimating the ED50.
2. Use this result to estimate the ED50 for the three insecticides using model `ins.m3` fitted in Sect. 9.11.

9.4. In Sect. 9.3 (p. 336), the probit binomial GLM was developed as a threshold model. Here consider using the *logistic distribution* with mean μ and variance σ^2 as the tolerance distribution. The logistic distribution has the probability function

$$\mathcal{P}(y; \mu, \sigma^2) = \frac{\pi \exp\{-(y-\mu)\pi/(\sigma\sqrt{3})\}}{\sigma\sqrt{3}\left[1 + \exp\{-(y-\mu)\pi/(\sigma\sqrt{3})\}\right]^2}$$

for $-\infty < y < \infty$, $-\infty < \mu < \infty$ and $\sigma > 0$.

Table 9.6 The logistic regression model fitted to data relating hypertension to sleep apnoea-hypopnoea (Problem 9.5)

Variable	$\hat{\beta}_j$	$\text{se}(\hat{\beta}_j)$
Intercept	−6.949	0.377
Age	0.805	0.0444
Sex	0.161	0.113
Body mass index	0.332	0.0393
Apnoea-hypopnoea index	0.116	0.0204

1. Show that the logistic distribution is not an EDM.
2. Determine the CDF for the logistic distribution.
3. Plot the density function and CDF for the logistic distribution with mean 0 and variance 1. Also plot the same graphs for the normal distribution with mean 0 and variance 1. Comment on the similarities and differences between the two probability functions.
4. Using the logistic distribution as the tolerance distribution, show that the threshold model in Sect. 9.4 corresponds to a binomial GLM with a logistic link function.

9.5. In a study [14] of the relationship between hypertension and sleep apnoea-hypopnoea (breathing difficulties while sleeping), a logistic regression model was fitted. The dependent variable was the presence of hypertension. The independent variables were dichotomized as follows: Age: 0 for 10 years or under, and 1 otherwise; sex: 0 for females, and 1 for males; body mass index: 0 if under $5\,\text{kg/m}^2$, and 1 otherwise; apnoea-hypopnoea index: 0 if fewer than ten events per hour of sleep, and 1 otherwise. Age, sex and body mass index are extraneous variables. The fitted model is summarized in Table 9.6.

1. Write down the fitted model.
2. Use a Wald test to test if $\beta_j = 0$ for each independent variable. Which variables seems important in the model?
3. Find 95% confidence intervals for each regression parameter.
4. Compute and interpret the odds ratios for each independent variable.
5. Predict the mean probability of observing hypertension in 30 year-old males with a BMI of $6\,\text{kg/m}^2$ who have an apnoea-hypopnoea index value of 5.

9.6. A study of stress and aggression in youth [15] measured the 'role stress' (an additive index from survey responses) and adolescent aggression levels (1 if the subject had engaged in at least one aggressive act as a youth, and 0 otherwise) in non-Hispanic whites. The response variable was aggression as an adult (1 if the subject had engaged in at least one aggressive act, and 0 otherwise). The fitted model is summarized in Table 9.7. (A number of other extraneous variables are also fitted, such as marital status and illicit drug use, but are not displayed in the table.)

Table 9.7 Two binomial GLMs fitted to the aggression data (Problem 9.6)

	Males		Females	
Variable	$\hat{\beta}_j$	$\text{se}(\hat{\beta}_j)$	$\hat{\beta}_j$	$\text{se}(\hat{\beta}_j)$
Intercept	0.45	0.40	−0.22	0.53
Role stress, RS	0.04	0.08	0.26	0.06
Adolescent aggression, AA	0.25	0.15	0.82	0.19
Interaction, $RS.AA$	0.23	0.17	−0.22	0.11
Residual deviance	57.40		121.67	
p'	13		13	
n	1323		1427	

1. Write down the two fitted models (one for males, one for females).
2. Use a Wald statistic to test if $\beta_j = 0$ for the interaction terms for both the male and female models. Comment.
3. The residual deviances for the fitted logistic regression models without the interaction term are 53.40 (males) and 117.82 (females). Use a likelihood ratio test to determine if the interaction terms are necessary in the models. Compare with the results of the Wald test.
4. Find 95% confidence intervals for both interaction terms.
5. Compute and interpret the odds ratios for AA.
6. Is overdispersion likely to be a problem for the models shown in the table?
7. Suppose a logistic GLM was fitted to the data with role stress, adolescent aggression, gender (G) and all the extraneous variables fitted to the model. Do you think the regression parameter for the three-way interaction $RS.AA.G$ would be different from zero? Explain.

9.7. After the explosion of the space shuttle *Challenger* on January 28, 1986, a study was conducted [1, 4] to determine if previously-collected data about the ambient air temperature at the time of launch could have been used to foresee potential problems with the launch (Table 4.1; data set: `shuttles`). In Example 4.2, a model was proposed for these data.

1. Plot the data.
2. Fit and interpret the proposed model.
3. Perform a diagnostic analysis.
4. On the day of the *Challenger* launch, the forecast temperature was 31°F. What is the predicted probability of an O-ring failure?
5. What would the ED50 mean in this context? What would be a more sensible ED for this context?

9.8. An experiment [11] studied the survival of mice after receiving a test dose of culture with five different doses of antipneumococcus serum (in cc) (Table 9.8; data set: `serum`).

Table 9.8 The number of mice surviving exposure to pneumococcus after receiving a dose of antipneumococcus (Problem 9.8)

Dose (in cc)	Total number of mice	Number of survivors
0.000625	40	7
0.00125	40	18
0.0025	40	32
0.005	40	35
0.01	40	38

Table 9.9 The number of tobacco budworm moths (*Heliothis virescens*) out of 20 that were killed when exposed for three days to pyrethroid *trans*-cypermethrin (Problem 9.9)

	Pyrethroid dose (in µg)					
Gender	1	2	4	8	16	32
Male	1	4	9	13	18	20
Female	0	2	6	10	12	16

1. Fit and interpret a logistic regression model to the data with systematic component `Survivors/Number ~ 1 + log(Dose)`.
2. Examine the diagnostics from the above model.
3. Plot the data with the fitted lines, and the corresponding 95% confidence intervals.
4. Estimate the ED50.
5. Interpret your fitted model using the threshold interpretation for the link function.

9.9. The responses of the tobacco budworm *Heliothis virescens* to doses of pyrethroid *trans*-cypermethrin were recorded (Table 9.9; data set: `budworm`) [2, 23] from a small experiment. Twenty male and twenty female moths were exposed at each of six doses of the pyrethroid, and the number killed was recorded.

1. Plot survival proportions against dose, distinguishing male and female moths. Explain why using the logarithms of dose as a covariate is sensible given the values used for the pyrethroid dose.
2. Fit a binomial GLM to the data, ensuring a diagnostic analysis. Begin by fitting a model with a systematic component of the form `1 + log2(Dose) * Gender`, and show that the interaction term is not significant. Hence refit the model with systematic component `1 + log2(Dose) + Gender`.
3. Plot the fitted lines on the plot of the data (distinguishing between males and females) and comment on the suitability of the model.
4. Determine the odds ratio for comparing the odds of a male moth dying to the odds to a female moth dying.

Table 9.10 The gender of candidates in the 1992 British general election; M means males and F means females (Problem 9.10)

	Cons		Labour		Lib-Dem		Greens		Other	
Region	M	F	M	F	M	F	M	F	M	F
South East	101	8	84	25	81	28	42	15	86	27
South West	45	3	36	12	35	13	21	6	61	11
Great London	76	8	57	27	63	19	37	13	93	21
East Anglia	19	1	16	4	16	4	6	4	23	8
East Midlands	39	3	35	7	36	6	8	3	19	7
Wales	36	2	34	4	30	8	7	0	44	10
Scotland	63	9	67	5	51	21	14	6	87	17
West Midlands	50	8	43	15	49	9	11	4	30	5
Yorks and Humbers	51	3	45	9	42	12	22	3	22	6
North West	65	8	57	16	61	12	17	5	75	20
North	32	4	34	2	32	4	7	1	6	3

5. Determine if there is any evidence of a difference in the mortality rates between the male and female moths.
6. Determine estimates of the ED50 for both genders.
7. Determine the 90% confidence interval for the gender effect.

9.10. The *Independent* newspaper tabulated the gender of all candidates running for election in the 1992 British general election (Table 9.10; data set: `belection`) [6].

1. Plot the proportion of female candidates against the Party, and comment.
2. Plot the proportion of female candidates against the Region, and comment.
3. Find a suitable binomial GLM, ensuring a diagnostic analysis.
4. Is overdispersion evident?
5. Interpret the fitted model.
6. Estimate and interpret the odds of a female candidate running for the Conservative and Labour parties. Then compute the odds ratio of the Conservative party fielding a female candidate to the odds of the Labour party fielding a female candidate.
7. Determine if the saddlepoint approximation is likely to be suitable for these data.

9.11. A study [9, 12] of patients treated for nonmetastatic sarcoma obtained data on the gender of the patients, the presence of lymphocytic infiltration and any asteoid pathology. The treatment was considered a success if patients were disease-free for 3 years (Table 9.11). Here, consider the effect of lymphocytic infiltration on the proportion of success.

1. Plot the proportion of successes against gender. Then plot the proportion of successes against the presence or absence of lymphocytic infiltration. Comment on the relationships.

Table 9.11 The nonmetastatic sarcoma data (Problem 9.11)

Lymphotic infiltration	Gender	Osteoid pathology	Group size m	Number of successes my
Absent	Female	Absent	3	3
Absent	Female	Present	2	2
Absent	Male	Absent	4	4
Absent	Male	Present	1	1
Present	Female	Absent	5	5
Present	Female	Present	5	3
Present	Male	Absent	9	5
Present	Male	Present	17	6

2. Fit the binomial GLM using the gender and presence or absence of lymphocytic infiltration as explanatory variables. Show that the Wald test results indicate that the effect of lymphocytic infiltration is not significant.
3. Show that the likelihood ratio test indicates that the effect of lymphocytic infiltration is significant.
4. Show that the score test also indicates that the effect of lymphocytic infiltration is significant.
5. Explain the results from the three tests.

9.12. Chromosome aberration assays are used to determine whether or not a substance induces structural changes in chromosomes. One study [24] compared the results of two substances at various doses (Table 9.12). A large number of cells were sampled at each dose to see how many were aberrant.

1. Fit a binomial GLM to determine if there is evidence of a difference between the two substances.
2. Use the dose and the logarithm of dose as an explanatory variable in separate GLMs, and compare. Which is better, and why?
3. Compute the 95% confidence interval for the dose regression parameter, and interpret.
4. Why would estimation of the ED50 be inappropriate?

9.13. A study [17] of the habitats of the noisy miner (a small but aggressive native Australian bird; data set: `nminer`) recorded whether noisy miners were present in various two hectare transects in buloke woodland patches (`Miners`), and considered the following potential explanatory variables: the number of eucalypt trees (`Eucs`); the number of buloke trees (`Bulokes`); the area of contiguous remnant patch vegetation in which each site was located (`Area`); whether the area was grazed (`Grazed`: `1` means yes); whether shrubs were present in the transect (`Shrubs`: `1` means yes); and the number of pieces of fallen timber (`Timber`). Part of this data frame was discussed in Example 1.5 (p. 14), where models were fitted for the *number* of noisy miners.

Table 9.12 The number of aberrant cells for different doses of two substances (Problem 9.12)

Substance	Dose (in mg/ml)	No. cell samples	No. cells aberrant	Substance	Dose (in mg/ml)	No. cell samples	No. cells aberrant
A	0	400	3	B	0.0	400	5
A	20	200	5	B	62.5	200	2
A	100	200	14	B	125.0	200	2
A	200	200	4	B	250.0	200	4
				B	500.0	200	7

Fit a suitable logistic regression model for predicting the *presence* of noisy miners in two hectare transects in buloke woodland patches, ensuring an appropriate diagnostic analysis. Also estimate the number of eucalypt trees in which there is a greater than 90% chance of finding noisy miners.

9.14. In Example 9.4, data [3] were introduced regarding the germination of seeds, using two types of seeds and two types of root stocks (Table 9.3). An alternative way of entering the data is to record whether or not each individual seed germinates or not (data set: `germBin`).

1. Fit the equivalent model to that fitted in Example 9.4, but using data prepared as in the data file `germBin`. This model is based on using a Bernoulli distribution.
2. Show that both the Bernoulli and binomial GLMs produce the same values for the parameter estimates and standard errors.
3. Show that the two models produce different values for the residual deviance, but the same values for the deviance.
4. Show that the two models produce similar results from the sequential likelihood-ratio tests.
5. Compare the log-likelihoods for the binomial and Bernoulli distributions. Comment.
6. Explain why overdispersion cannot be detected in the Bernoulli model.

References

[1] Chatterjee, S., Handcock, M.S., Simonoff, J.S.: A Casebook for a First Course in Statistics and Data Analysis. John Wiley and Sons, New York (1995)
[2] Collett, D.: Modelling Binary Data. Chapman and Hall, London (1991)
[3] Crowder, M.J.: Beta-binomial anova for proportions. Applied Statistics **27**(1), 34–37 (1978)

[4] Dala, S.R., Fowlkes, E.B., Hoadley, B.: Risk analysis of the space shuttle: pre-Challenger prediction of failure. Journal of the American Statistical Association **84**(408), 945–957 (1989)

[5] Dunn, P.K., Smyth, G.K.: Randomized quantile residuals. Journal of Computational and Graphical Statistics **5**(3), 236–244 (1996)

[6] Hand, D.J., Daly, F., Lunn, A.D., McConway, K.Y., Ostrowski, E.: A Handbook of Small Data Sets. Chapman and Hall, London (1996)

[7] Hauck Jr., W.W., Donner, A.: Wald's test as applied to hypotheses in logit analysis. Journal of the American Statistical Association **72**, 851–853 (1977)

[8] Hewlett, P.S., Plackett, T.J.: Statistical aspects of the independent joint action of poisons, particularly insecticides. II Examination of data for agreement with hypothesis. Annals of Applied Biology **37**, 527–552 (1950)

[9] Hirji, K.F., Mehta, C.R., Patel, N.R.: Computing distributions for exact logistic regression. Journal of the American Statistical Association **82**(400), 1110–1117 (1987)

[10] Hu, Y., Smyth, G.K.: ELDA: Extreme limiting dilution analysis for comparing depleted and enriched populations in stem cell and other assays. Journal of Immunological Methods **347**, 70–78 (2009)

[11] Irwin, J.O., Cheeseman, E.A.: On the maximum-likelihood method of determining dosage-response curves and approximations to the median-effective dose, in cases of a quantal response. Supplement to the Journal of the Royal Statistical Society **6**(2), 174–185 (1939)

[12] Kolassa, J.E., Tanner, M.A.: Small-sample confidence regions in exponential families. Biometrics **55**(4), 1291–1294 (1999)

[13] Krzanowski, W.J.: An Introduction to Statistical Modelling. Arnold, London (1998)

[14] Lavie, P., Herer, P., Hoffstein, V.: Obstructive sleep apnoea syndrome as a risk factor for hypertension: Population study. British Medical Journal **320**(7233), 479–482 (2000)

[15] Liu, R.X., Kaplan, H.B.: Role stress and aggression among young adults: The moderating influences of gender and adolescent aggression. Social Psychology Quarterly **67**(1), 88–102 (2004)

[16] Lumley, T., Kronmal, R., Ma, S.: Relative risk regression in medical research: Models, contrasts, estimators, and algorithms. UW Biostatistics Working Paper Series 293, University of Washington (2006)

[17] Maron, M.: Threshold effect of eucalypt density on an aggressive avian competitor. Biological Conservation **136**, 100–107 (2007)

[18] Mather, K.: The analysis of extinction time data in bioassay. Biometrics **5**(2), 127–143 (1949)

[19] Myers, R.H., Montgomery, D.C., Vining, G.G.: Generalized Linear Models with Applications in Engineering and the Sciences. Wiley Series in Probability and Statistics. Wiley, Chichester (2002)

[20] Nelson, W.: Applied Life Data Analysis. John Wiley and Sons, New York (1982)

[21] Shackleton, M., Vaillant, F., Simpson, K.J., Sting, J., Smyth, G.K., Asselin-Labat, M.L., Wu, L., Lindeman, G.J., Visvader, J.E.: Generation of a functional mammary gland from a single stem cell. Nature **439**, 84–88 (2006)

[22] Singer, J.D., Willett, J.B.: Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence. Oxford University Press, New York (2003)

[23] Venables, W.N., Ripley, B.D.: Modern Applied Statistics with S, fourth edn. Springer-Verlag, New York (2002). URL http://www.stats.ox.ac.uk/pub/MASS4

[24] Williams, D.A.: Tests for differences between several small proportions. Applied Statistics **37**(3), 421–434 (1988)

[25] Xie, G., Roiko, A., Stratton, H., Lemckert, C., Dunn, P., Mengersen, K.: Guidelines for use of the approximate beta-Poisson dose-response models. Risk Analysis **37**, 1388–1402 (2017)