Chapter 7
Binary Data Models

7.1 Introduction

In this chapter we consider the modeling of binary data. Such data are ubiquitous in many fields. Binary data present a number of distinct challenges, and so we devote a separate chapter to their modeling, though we lean heavily on the methods introduced in Chap. 6 on general regression modeling. It is perhaps surprising that the simplest form of outcome can pose difficulties in analysis, but a major problem is the lack of information contained within a variable that can take one of only two values. This can lead to a number of problems, for example, in assessing model fit. Another major complication arises because models for probabilities are generally nonlinear, which can lead to curious behavior of estimators in the presence of confounders. Difficulties in interpretation also arise, even when independent regressors are added to the model.

The outline of this chapter is as follows. We give some motivating examples in Sect. 7.2, and in Sect. 7.3, describe the genesis of the binomial model, which is a natural candidate for the analysis of binary data. Generalized linear models for binary data are examined in Sect. 7.4. The binomial model has a variance determined by the mean, with no additional parameter to accommodate excess-binomial variation, and so Sect. 7.5 describes methods for dealing with such variation. For reasons that will become apparent, we will focus on logistic regression models, beginning with a detailed description in Sect. 7.6. This section includes discussions of estimation from likelihood, quasi-likelihood, and Bayesian perspectives. Conditional likelihood and "exact" inference are the subject of Sect. 7.7. Assessing the adequacy of binary models is discussed in Sect. 7.8. Summary measures that exhibit nonobvious behavior are the subject of Sect. 7.9. Case-control studies are a common design, which offer interesting inferential challenges with respect to inference, and are described in Sect. 7.10. Concluding comments appear in Sect. 7.11. Section 7.12 gives references to more in-depth treatments of binary modeling and to source materials.

J. Wakefield, *Bayesian and Frequentist Regression Methods*, Springer Series in Statistics, DOI 10.1007/978-1-4419-0925-1_7,

7.2 Motivating Examples

7.2.1 Outcome After Head Injury

We will illustrate methods for binary data using the data first encountered in Sect. 1.3.2. The binary response is outcome after head injury (dead/alive), with four discrete covariates: pupils (good/poor), coma score (depth of coma, low/high), hematoma present (no/yes), and age (categorized as 1–25, 26–54, $\geq$55). These data were presented in Table 1.1, but it is difficult to discern patterns from this table. In general, cross-classified data such as these may be explored by looking at marginal and conditional tables of counts or frequencies. Figure 7.1 displays conditional frequencies, with panel (a) corresponding to low coma score and panel (b) to high coma score. These plots suggest that the probability of death increases with age, that a low coma score is preferable to a high coma score, and that good pupils are beneficial. The association with the hematoma variable is less clear. The sample sizes are lost in these plots, which makes interpretation more difficult.

7.2.2 Aircraft Fasteners

Montgomery and Peck (1982) describe a study in which the compressive strength of fasteners used in the construction of aircraft was examined. Table 7.1 gives the total number of fasteners tested and the number of failures at a range of pressure loads. We see that the proportion failing increases with load. For these data we will aim to find a curve to adequately model the relationship between the probability of fastener failure and load pressure.

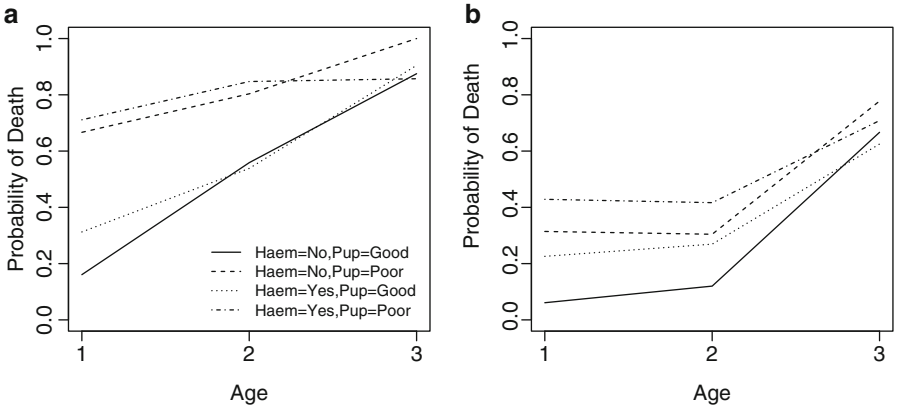

Fig. 7.1 Probability of death after head injury as a function of age, hematoma score, and pupils: Panels (**a**) and (**b**) are for low and high coma scores, respectively

Table 7.1 Number of aircraft fastener failures at specified pressure loads

Load (psi)	Failures	Sample size	Proportion failing
2,500	10	50	0.20
2,700	17	70	0.24
2,900	30	100	0.30
3,100	21	60	0.35
3,300	18	40	0.45
3,500	43	85	0.51
3,700	54	90	0.60
3,900	33	50	0.66
4,100	60	80	0.75
4,300	51	65	0.78

7.2.3 Bronchopulmonary Dysplasia

We describe data from van Marter et al. (1990) and subsequently analyzed by Pagano and Gauvreau (1993) on the absence/presence of bronchopulmonary dysplasia (BPD) as a function of birth weight (in grams) for $n = 223$ babies. BPD is a chronic lung disease that affects premature babies. In this study, BPD was defined as a function of both oxygen requirement and compatible chest radiograph, with 147 of the babies having neither characteristic by day 28 of life. We take as illustrative aim the prediction of BDP using birth weight, the rationale being that if a good predictive model can be found, then measures could be taken to decrease the probability of BPD. There are a number of caveats that should be attached to this analysis. First, these data are far from a random sample of births, as they are sampled from intubated infants with weights less than 1,751 g (so that all of the babies are of low birth weight). In general, an estimate of the incidence of BPD is difficult to tie down, in part, because of changes in the definition of the condition. Allen et al. (2003) provide a discussion of this issue and report that, of preterm infants with birth weights less than 1,000 g, 30% develop BPD. Second, a number of additional covariates would be available in a serious attempt at prediction, including gender and the medication used by the mothers.

Figure 7.2 displays the BPD indicator, plotted as short vertical lines at 0 and 1, as a function of birth weight. Visual assessment suggests that children with lower birth weight tend to have an increased chance of BPD. It is hard to discern the shape of the association from the raw binary data alone, however, since one is trying to compare the distributions of zeros and ones, which is difficult. This example is distinct from the aircraft fasteners because the latter contained multiple responses at each x value. Binning on the basis of birthweight and plotting the proportions with BPD in each bin would provide a more informative plot.

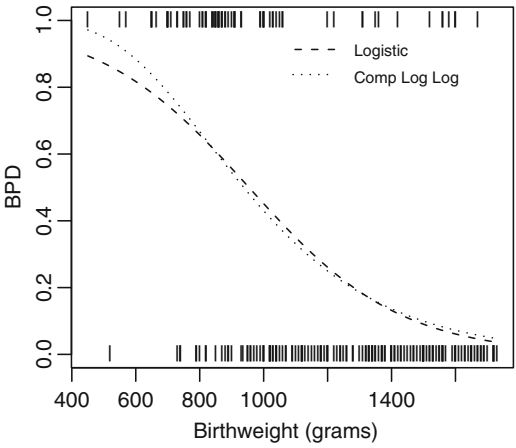

Fig. 7.2 Indicator of bronchopulmonary dysplasia (BPD), as a function of birth weight. The *short vertical lines* at 0 and 1 indicate the observed birth weights for non-BPD and BPD infants, respectively. The *dashed curve* corresponds to a logistic regression fit and the *dotted curve* to a complementary log–log regression fit

7.3 The Binomial Distribution

7.3.1 Genesis

In the following we will refer to the basic sampling unit as an individual. Let Z denote the *Bernoulli* random variable with

$$\Pr(Z = z \mid p) = p^z(1-p)^{1-z},$$

$z = 0, 1$, and

$$p = \Pr(Z = 1 \mid p),$$

for $0 < p < 1$. For concreteness, we will call the $Z = 1$ outcome a positive response. A random variable taking two values *must* have a Bernoulli distribution, and all moments are determined as functions of p. In particular, $\text{var}(Z \mid p) = p(1-p)$ so that there is no concept of underdispersion or overdispersion for a Bernoulli random variable.

Suppose there are N individuals, and let Z_j denote the outcome for the jth individual, $j = 1, \ldots, N$. Also let $Y = \sum_{j=1}^{N} Z_j$ be the total number of individuals with a positive outcome, and suppose that each has equal probabilities, that is, $p = p_1 = \ldots = p_N$. Under the assumption that the Bernoulli random variables are *independent*,

$$Y \mid p \sim \text{Binomial}(N, p)$$

so that

$$\Pr(Y = y \mid p) = \binom{N}{y} p^y(1-p)^{1-y}, \tag{7.1}$$

for $y = 0, 1, \ldots, N$.

Constant $p = p_j$, $j = 1, \ldots, N$, over the N individuals is not necessary for Y to follow a binomial distribution. Suppose that individual j has probability p_j drawn at random from a distribution with mean $\overline{p}$. In this case,

$$\mathrm{E}[Z_j] = \mathrm{E}\left[\mathrm{E}(Z_j \mid p_j)\right] = \overline{p}$$

and

$$Y \mid \overline{p} \sim \text{Binomial}(N, \overline{p}). \tag{7.2}$$

Crucial to this derivation is the assumption that p_j are *independent* draws from the distribution with mean $\overline{p}$, which means that the Z_j are also independent for $j = 1, \ldots, N$. Alternative scenarios are described in the context of overdispersion in Sect. 7.5.

We give a second derivation of the binomial distribution. Suppose $Y_j \mid \lambda_j \sim_{ind} \text{Poisson}(\lambda_j)$, $j = 1, 2$ are independent Poisson random variables with rates λ_j. Then,

$$Y_1 \mid Y_1 + Y_2, p \sim \text{Binomial}(Y_1 + Y_2, p),$$

with $p = \lambda_1/(\lambda_1 + \lambda_2)$ (Exercise 7.3).

7.3.2 Rare Events

Suppose that $Y \mid p \sim \text{Binomial}(N, p)$ and that $p \to 0$ and $N \to \infty$, with $\lambda = Np$ fixed (or tending to a constant). Then Exercise 7.1 shows that, in the limit, $Y \mid \lambda \sim \text{Poisson}(\lambda)$. Approximating the binomial distribution with a Poisson has a number of advantages. Computationally, the Poisson model can be more stable than the binomial model. Also, $\lambda > 0$ can be modeled via a loglinear form which provides a more straightforward interpretation than the logistic form, $\log[p/(1-p)]$. The following example illustrates one use of this result for obtaining a closed-form distribution when counts are summed.

Example: Lung Cancer and Radon

In Sect. 1.3.3 we introduced the lung cancer dataset, with Y_i being the number of cases in area i. A possible model for these data is

$$Y_i \mid \theta_i \sim \text{Poisson}(E_i\theta_i), \tag{7.3}$$

where E_i is the expected number of cases based on the age and gender breakdown of area i and θ_i is the relative risk associated with the area, for $i = 1, \ldots, n$.

A formal derivation of this model is as follows (see Sect. 6.5 for a related discussion). Let Y_{ij} be the disease counts in area i and age-gender stratum j and

N_{ij} the associated population, $i = 1, \ldots, n, j = 1, \ldots, J$. In the Minnesota study, we have $J = 36$, corresponding to male/female and 18 age bands: 0–4, 5–9,..., 80–84, 85+. We only have access to the total counts in the area, Y_i, and so we require a model for this sum. One potential model is $Y_{ij} \mid p_{ij} \sim \text{Binomial}(N_{ij}, p_{ij})$, with p_{ij} the probability of lung cancer diagnosis in area i, stratum j. With binomial Y_{ij}, the distribution of $Y_i = \sum_{j=1}^{J} Y_{ij}$ is a convolution, which is unfortunately awkward to work with. For example, for $J = 2$,

$$\Pr(y_i \mid p_{i1}, p_{i2})$$
$$= \sum_{y_{i1}=l_i}^{u_i} \binom{N_{i1}}{y_{i1}} \binom{N_{i2}}{y_i - y_{i1}} p_{i1}^{y_{i1}} (1 - p_{i1})^{N_{i1} - y_{i1}} p_{i2}^{y_i - y_{i1}} (1 - p_{i2})^{N_{i2} - y_i + y_{i1}}$$

where $l_i = \max(0, y_i - N_{i2}), u_i = \min(N_{i1}, y_i)$, gives the range of admissible values that y_{i1} can take, given the margins $Y_i, N_i - Y_{i1} - Y_{i2}, N_{i1}, N_{i2}$. Lung cancer is statistically rare, and so we can use the Poisson approximation to give $Y_{ij} \mid p_{ij} \sim \text{Poisson}(N_{ij} p_{ij})$. The distribution of the sum Y_i is then straightforward:

$$Y_i \mid p_{i1}, \ldots, p_{iJ} \sim \text{Poisson}\left(\sum_{j=1}^{J} N_{ij} p_{ij} \right). \tag{7.4}$$

There are insufficient data to estimate the $n \times J$ probabilities p_{ij}, and so it is common to assume $p_{ij} = \theta_i \times q_j$, where q_j are a set of known reference stratum-specific rates and θ_i is an area-specific term that summarizes the deviation of the risks in area i from the reference rates. Therefore, this model assumes that the effect on risk of being in area i is the same across stratum. Usually, the q_j are assumed known. Consequently, (7.4) simplifies to $Y_i \mid \theta_i \sim \text{Poisson}\left(\theta_i \sum_{j=1}^{J} N_{ij} q_j\right)$, and substituting the expected numbers $E_i = \sum_{j=1}^{J} N_{ij} q_j$ produces model (7.3).

7.4 Generalized Linear Models for Binary Data

7.4.1 Formulation

Let $Z_{ij} = 0/1$ denote the absence/presence of the binary characteristic of interest in each of the $j = 1, \ldots, N_i$ trials, with $i = 1, \ldots, n$ different "conditions." Let $Y_i = \sum_{j=1}^{N_i} Z_{ij}$ denote the number of positive responses and $N = \sum_{i=1}^{n} N_i$ the total number of trials. Further, suppose there are k explanatory variables recorded for each condition, and let $\boldsymbol{x}_i = [1, x_{i1}, \ldots, x_{ik}]$ denote the row vector of dimension $1 \times (k + 1)$ for $i = 1, \ldots, n$. We now wish to model the probability of a positive response $p(\boldsymbol{x}_i)$, as a function of $\boldsymbol{x}_i$, in order to identify structure within the data.

We might naively model the observed proportion via the linear model

$$\frac{Y_i}{N_i} = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i,$$

for $i = 1, \ldots, n$. There are a number of difficulties with such an approach. First, the observed proportions must lie in the range $[0, 1]$, while the modeled probability $\boldsymbol{x}_i\boldsymbol{\beta}$ is unrestricted. We could attempt to put constraints on the parameters in order to alleviate this drawback, but this is inelegant and soon becomes cumbersome with multiple explanatory variables. The resultant inference is also difficult due to the restricted ranges. The second difficulty is that we saw in Sect. 5.6.4 that in the usual linear model framework, an appropriate mean–variance model is crucial for well-calibrated inference (unless sandwich estimation is turned to). A linear model is usually associated with error terms with constant variance, but this is not appropriate here since

$$\text{var}\left(\frac{Y_i}{N_i}\right) = \frac{p(\boldsymbol{x}_i)[1 - p(\boldsymbol{x}_i)]}{N_i}$$

so that the variance changes with the mean. The generalized linear model, introduced and discussed in Sect. 6.3, can rectify these deficiencies. For sums of binary variables, the binomial model is a good starting point.

The binomial model is a member of the exponential family, specifically $Y \mid p \sim \text{Binomial}(N, p)$, that is, (7.1), translates to

$$p(y \mid p) = \exp\left[y \log\left(\frac{p}{1-p}\right) + N \log(1-p)\right], \tag{7.5}$$

which provides the stochastic element of the model. For the deterministic part, we specify a monotonic, differentiable link function:

$$g\left[\, p(\boldsymbol{x}) \,\right] = \boldsymbol{x}\boldsymbol{\beta}. \tag{7.6}$$

The exponential family is appealing from a statistical standpoint since correct specification of the mean function leads to consistent inference, since the score function is linear in the data (this function is given for the logistic model in (7.12)). With a GLM, the computation is also usually straightforward (Sect. 6.5.2). Nonlinear models can also be considered, however, if warranted by the application. For example, Diggle and Rowlingson (1994) considered modeling disease risk as a function of distance x from a point source of pollution. These authors desired a model for which disease risk returned to baseline as $x \to \infty$ and suggested a model for the odds of the form

$$\frac{\Pr(Z=1 \mid x)}{\Pr(Z=0 \mid x)} = \beta_0 \left[1 + \beta_1 \exp(-\beta_2 x^2)\right],$$

with β_0 corresponding to baseline odds, β_1 corresponding to the excess odds at $x = 0$ (i.e., at the point source), and β_2 determining the speed at which the odds decline to baseline. Such nonlinear models are computationally more difficult to fit but produce consistent parameter estimates, if combined with an exponential family.

7.4.2 Link Functions

From (7.5) we see that the so-called *canonical* link is the logit $\theta = \log[p/(1-p)]$. We will see that *logistic regression models* of the form

$$\log\left[\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})}\right] = \boldsymbol{x}\boldsymbol{\beta} \tag{7.7}$$

offer a number of advantages in terms of computation and inference. This link function is by far the most popular in practice, and so Sect. 7.6 is dedicated to logistic regression modeling.

Other link functions that may be used for binomial data include the *probit*, *complimentary log–log*, and *log–log* links. The probit link is

$$\Phi^{-1}\left[\, p(\boldsymbol{x}) \,\right] = \boldsymbol{x}\boldsymbol{\beta},$$

where $\Phi[\cdot]$ is the distribution function of a standard normal random variable. This link function generally produces similar inference to the logistic link function. The logistic and probit link functions are symmetric in the sense that $g(p) = -g(1-p)$.

The complementary log–log link function is

$$\log\left\{-\log\left[1 - p(\boldsymbol{x})\right]\right\} = \boldsymbol{x}\boldsymbol{\beta}, \tag{7.8}$$

to give

$$p(\boldsymbol{x}) = 1 - \exp\left[-\exp(\boldsymbol{x}\boldsymbol{\beta})\right],$$

which is not symmetric. Hence, the log–log link model

$$-\log\left\{-\log\left[p(\boldsymbol{x})\right]\right\} = \boldsymbol{x}\boldsymbol{\beta}$$

with

$$p(\boldsymbol{x}) = \exp[-\exp(-\boldsymbol{x}\boldsymbol{\beta})]$$

may also be used and will not produce the same inference as (7.8). If $g_{\text{CLL}}(\cdot)$ and $g_{\text{LL}}(\cdot)$ represent the complementary log–log and log–log links, respectively, then the two are related via $g_{\text{CLL}}(p) = -g_{\text{LL}}(1-p)$.

7.5 Overdispersion

Overdispersion is a phenomena that occurs frequently in applications and, in the binomial data context, describes a situation in which the variance $\text{var}(Y_i \mid p_i)$ exceeds the binomial variance $N_i p_i(1-p_i)$.

Often overdispersion occurs due to clustering in the population from which the individuals were drawn. To motivate a variance model, suppose for simplicity that the N_i individuals for whom we measure outcomes in trial i are actually broken into C_i clusters of size k_i so that $N_i = C_i \times k_i$. These clusters may correspond to families, geographical areas, genetic subgroups, etc. Within the cth cluster, the number of positive responders Y_{ic} has distribution $Y_{ic} \mid p_{ic} \sim_{ind} \text{binomial}(k_i, p_{ic})$, where each p_{ic} is drawn independently from some distribution, for $c = 1, \ldots, C_i$. Let P_{ic} represent a random variable with

$$\text{E}[P_{ic}] = p_i$$

$$\text{var}(P_{ic}) = \tau_i^2 p_i(1-p_i),$$

where the variance is written in this form for convenience (as we see shortly). In the following we will use expressions for iterated expectation, variance, and covariance, as described in Appendix B. Then, letting $Y_i = \sum_{c=1}^{C_i} Y_{ic}$,

$$\text{E}[Y_i] = \text{E}\left[\sum_{c=1}^{C_i} Y_{ic}\right] = \sum_{c=1}^{C_i} \text{E}_{P_{ic}}\left[\text{E}(Y_{ic} \mid P_{ic})\right] = \sum_{c=1}^{C_i} \text{E}_{P_{ic}}[k_i P_{ic}] = N_i p_i.$$

Turning to the variance,

$$\text{var}(Y_i) = \text{var}\left(\sum_{c=1}^{C_i} Y_{ic}\right) = \sum_{c=1}^{C_i} \text{var}(Y_{ic}),$$

since the counts are independent, as each p_{ic} is drawn independently. Continuing with this calculation and exploiting the iterated variance formula,

$$\begin{aligned}
\text{var}(Y_i) &= \sum_{c=1}^{C_i} \{\text{E}\left[\text{var}(Y_{ic} \mid p_{ic})\right] + \text{var}\left(\text{E}[Y_{ic} \mid p_{ic}]\right)\} \\
&= \sum_{c=1}^{C_i} \left\{\text{E}_{P_{ic}}[k_i P_{ic}(1-P_{ic})] + \text{var}_{P_{ic}}(k_i P_{ic})\right\} \\
&= \sum_{c=1}^{C_i} \left\{k_i p_i - k_i \left[\text{var}(P_{ic}) + \text{E}[P_{ic}]^2\right] + k_i^2 \tau_i^2 p_i(1-p_i)\right\} \\
&= N_i p_i(1-p_i) \times \left[1 + (k_i - 1)\tau_i^2\right] \\
&= N_i p_i(1-p_i)\sigma_i^2.
\end{aligned}$$

Hence, the within-trial clustering has induced excess-binomial variation. Suppose each cluster is of size $k_i = 1$ (i.e., $C_i = N_i$); then we recover the binomial case (7.2). The above derivation requires $1 \leq \sigma_i^2 \leq k_i \leq N_i$, since $0 \leq \sigma_i^2 \leq 1$ (McCullagh and Nelder 1989, Sect. 4.5.1). If we were to assume a second moment model with a common $\sigma_i^2 = \sigma^2$ to give

$$\text{var}(Y_i) = N_i p_i(1-p_i)\sigma^2 \tag{7.9}$$

then the constraint becomes $\sigma^2 \leq N_i$, which is unfavorable, but will rarely be a problem in practice.

If we have a single cluster, that is, $C_i = 1$, then $k_i = N_i$ and

$$\text{var}(Y_i) = N_i p_i(1-p_i) \times \left[1 + (N_i - 1)\tau_i^2\right]. \tag{7.10}$$

Suppose Z_{ij}, $j = 1, \ldots, N_i$ are the binary outcomes within-trial i so that $Y_i = \sum_{j=1}^{N_i} Z_{ij}$. Then, for the case of a single cluster ($C_i = 1$),

$$\begin{aligned}
\text{cov}(Z_{ij}, Z_{ik}) &= \text{E}[\text{cov}(Z_{ij}, Z_{ik} \mid p_{i1})] + \text{cov}(\text{E}[Z_{ij} \mid p_{ij}], \text{E}[Z_{ik} \mid p_{ik}]) \\
&= \text{cov}_{P_{i1}}(P_{i1}, P_{i1}) \\
&= \text{var}(P_{i1}) = \tau_i^2 p_i(1-p_i),
\end{aligned}$$

so that τ_i^2 is the correlation between any two outcomes in trial i.

We now discuss a closely related scenario in which we start by assuming that outcomes within a trial have correlation τ_i^2. Then (Exercise 7.4),

$$\text{var}(Y_i) = N_i p_i(1-p_i) \times \left[1 + (N_i - 1)\tau_i^2\right]. \tag{7.11}$$

Notice that, unlike the derivation leading to (7.10), underdispersion can occur if $\tau_i^2 < 0$. The equality of (7.10) and (7.11) shows that the effect of either a random response probability or positively correlated outcomes within a trial is

indistinguishable marginally (unless one is willing to make assumptions about the within-trial distribution, but such assumptions are uncheckable).

Inferentially, two approaches are suggested. We could specify the first two moments only and use quasi-likelihood. This route is taken in Sect. 7.6.3. Alternatively, one can assume a specific distributional form and then proceed with parametric inference, as we now illustrate.

The most straightforward way to model overdispersion parametrically is to assume the binomial probability arises from a conjugate beta model. This model is

$$\begin{aligned} Y_i \mid q_i &\sim \text{Binomial}(N_i, q_i) \\ q_i &\sim \text{Beta}(a_i, b_i), \end{aligned}$$

where we can parameterize as $a_i = dp_i$, $b_i = d(1 - p_i)$ so that

$$\begin{aligned} p_i &= \frac{a_i}{d} \\ \text{var}(p_i) &= \frac{p_i(1-p_i)}{d+1}. \end{aligned}$$

An obvious choice of mean model is the linear logistic model

$$p_i = \frac{\exp(\boldsymbol{x}_i\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i\boldsymbol{\beta})}.$$

Notice that $d = 0$ corresponds to the binomial model. Integration over the random effects results in the beta-binomial marginal model:

$$\Pr(Y_i{=}y_i) = \binom{N_i}{y_i} \frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\Gamma(b_i)} \frac{\Gamma(a_i + y_i)\Gamma(b_i + N_i - y_i)}{\Gamma(a_i + b_i + N_i)}, \quad y_i = 0, 1, \ldots, N_i.$$

The marginal moments are

$$\begin{aligned} \text{E}[Y_i] &= N_i p_i = N_i \left(\frac{a_i}{a_i + b_i} \right) \\ \text{var}(Y_i) &= N_i p_i (1 - p_i) \left(\frac{a_i + b_i + N_i}{a_i + b_i + 1} \right), \end{aligned}$$

confirming that there is no overdispersion when $N_i = 1$. This variance is also equal to (7.10), with the assumption of constant τ_i^2 on recognizing that $\tau^2 = (a_i + b_i + 1)^{-1} = 1/(d+1)$. Unfortunately, the log-likelihood $l(\boldsymbol{\beta}, d)$ is not easy to deal with due to the gamma functions. More seriously, the beta-binomial distribution

is not of exponential family form and does not possess the consistency properties of distributions within this family.

Liang and McCullagh (1993) discuss the modeling of overdispersed binary data. In particular, they suggest plotting residuals

$$\frac{y_i - N_i\widehat{p}_i}{\sqrt{N_i\widehat{p}_i(1-\widehat{p}_i)}}$$

against N_i in order to see whether there is any association, which may help to choose between models (7.9) and (7.10).

7.6 Logistic Regression Models

7.6.1 Parameter Interpretation

We write the probability of $Y = 1$ as $p(\boldsymbol{x})$ to emphasize the dependence on covariates $\boldsymbol{x}$. Model (7.7) is equivalent to saying that the *odds* of a positive outcome may be modeled in a multiplicative fashion, that is,

$$\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})} = \exp(\boldsymbol{x}\boldsymbol{\beta}) = \exp(\beta_0)\prod_{j=1}^{k}\exp(x_j\beta_j).$$

Less intuition is evident on the probability scale for which

$$p(\boldsymbol{x}) = \frac{\exp(\boldsymbol{x}\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}\boldsymbol{\beta})}.$$

The transformation used here is known as the expit transform (and is the inverse of the logit transform). The expression for the probability makes it clear that we have enforced $0 < p(\boldsymbol{x}) < 1$.

For clarity, we discuss interpretation in the situation in which $p(\boldsymbol{x})$ is the probability of a disease, given exposure $\boldsymbol{x}$. Consider first the logistic regression model in the case where the exposures have no effect on the probability of disease:

$$\log\left[\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})}\right] = \beta_0.$$

In this case, β_0 is the log odds of disease for all levels of the exposures $\boldsymbol{x}$. Equivalent statements are that $\exp(\beta_0)$ is the odds of disease and $\exp(\beta_0)/[1+\exp(\beta_0)]$ is the probability of disease, regardless of the levels of $\boldsymbol{x}$.

Now consider the situation of a single exposure x for an individual with probability $p(x)$ and

$$\log\left[\frac{p(x)}{1-p(x)}\right] = \beta_0 + \beta_1 x.$$

The parameter $\exp(\beta_0)$ is the odds of disease at exposure $x = 0$, that is, the odds for an unexposed individual. The parameter $\exp(\beta_1)$ is the odds ratio for a unit increase in x. For example, if $\exp(\beta_1) = 2$, the odds of disease double for a unit increase in exposure. If x is a binary exposure, coded as 0/1, then $\exp(\beta_1)$ is the ratio of odds when going from unexposed to exposed:

$$\frac{p(1)/[1-p(1)]}{p(0)/[1-p(0)]} = \frac{\exp(\beta_0+\beta_1)}{\exp(\beta_0)} = \exp(\beta_1).$$

For a rare disease, the odds ratio and relative risk, which is given by $p(x)/p(x-1)$ for a univariate exposure, are approximately equal, with the relative risks being easier to interpret (see Sect. 7.10.2 for a more detailed discussion).

Logistic regression models may be defined for multiple factors and continuous variables in an exactly analogous fashion to the multiple linear models considered in Chap. 5. We simply include on the right-hand side of (7.6) the relevant design matrix and associated parameters. This is a benefit of the GLM framework in which we have linearity on some scale, though, with noncanonical link functions, parameter interpretation is usually more difficult.

The logistic model may be derived in terms of the so-called *tolerance distributions*. Let $U(x)$ denote an underlying continuous measure of the disease state at exposure x. We observe a binary version, $Y(x)$, of this variable which is related to $U(x)$ via

$$Y(x) = \begin{cases} 0 & \text{if } U(x) \le c \\ 1 & \text{if } U(x) > c, \end{cases}$$

for some threshold c. Suppose that the continuous measure follows a logistic distribution: $U(x) \sim \text{logistic}\,[\mu(x), 1]$. This distribution is given by

$$p(u \mid \mu, \sigma) = \frac{\exp\{(u-\mu)/\sigma\}}{\sigma\{1+\exp[(u-\mu)/\sigma]\}^2}, \quad -\infty < u < \infty.$$

The logistic distribution function, for the case $\sigma = 1$, is

$$\Pr[U(x) < u] = \frac{\exp(u-\mu)}{1+\exp(u-\mu)}, \quad -\infty < u < \infty.$$

From this model for $U(x)$, we can obtain the probability of the discrete outcome as

$$p(x) = \Pr[Y(x) = 1] = \Pr[U(x) > c] = \frac{\exp(\mu(x)-c)}{1+\exp(\mu(x)-c)},$$

which is equivalent to

$$\log\left[\frac{p(x)}{1-p(x)}\right] = \mu(x) - c.$$

So far we have not specified how the exposure x changes the distribution of the continuous latent variable $U(x)$. We assume that the effect of exposure to x is to move the location of the underlying variable $U(x)$ in a linear fashion via $\mu(x) = a + bx$, but while keeping the variance constant. We then obtain

$$\log\left[\frac{p(x)}{1-p(x)}\right] = \beta_0 + \beta_1 x,$$

where $\beta_0 = a - c$ and $\beta_1 = b$, that is, a linear logistic regression model.

The probit and complementary log–log links may similarly be derived from normal and extreme-value[1] tolerance distributions, respectively.

7.6.2 Likelihood Inference for Logistic Regression Models

We consider the logistic regression model

$$\log\left[\frac{p_i(\boldsymbol{\beta})}{1-p_i(\boldsymbol{\beta})}\right] = \boldsymbol{x}_i\boldsymbol{\beta},$$

where $\boldsymbol{x}_i$ is a $1 \times (k+1)$ vector of covariates measured on the ith individual and $\boldsymbol{\beta}$ is the $(k+1) \times 1$ vector of associated parameters. We write $p_i(\boldsymbol{\beta})$ to emphasize that the probability of a positive response is a function of $\boldsymbol{\beta}$. For the general binomial model the log-likelihood is

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i \log p_i(\boldsymbol{\beta}) + \sum_{i=1}^{n} (N_i - Y_i) \log\left[1 - p_i(\boldsymbol{\beta})\right],$$

with score function

$$\boldsymbol{S}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\partial p_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \frac{[Y_i - N_i p(\widehat{\boldsymbol{\beta}})]}{p(\boldsymbol{\beta})[1 - p(\widehat{\boldsymbol{\beta}})]}. \tag{7.12}$$

Letting $\boldsymbol{\mu}$ represent the $n \times 1$ vector with ith element $\mu_i = N_i p_i(\boldsymbol{\beta})$ allows (7.12) to be rewritten as

$$\boldsymbol{S}(\boldsymbol{\beta}) = \boldsymbol{D}^{\mathrm{T}} \boldsymbol{V}^{-1} \left[\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})\right], \tag{7.13}$$

where $\boldsymbol{D}$ is the $n \times (k+1)$ matrix with (i,j)th element $\partial \mu_i / \partial \beta_j$, $i = 1, \ldots, n$, $j = 0, \ldots, k$ and $\boldsymbol{V}$ is the $n \times n$ diagonal matrix with ith diagonal element $N_i p(\boldsymbol{x}_i)\left[1 - p(\boldsymbol{x}_i)\right]$. From Sect. 6.5.1,

[1] u has an extreme-value distribution if its distribution function is of the form $F(u) = 1 - \exp\{-\exp[(u-\mu)/\sigma]\}$.

$$\boldsymbol{I}_n(\boldsymbol{\beta})^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \;\rightarrow_d\; \mathrm{N}_{k+1}(\,\mathbf{0}, \mathbf{I}_{k+1}),$$

where $I_n(\boldsymbol{\beta}) = \boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{D}$. For the logistic model,

$$\frac{\partial \mu_i}{\partial \beta_j} = x_{ij} N_i p_i (1 - p_i)$$
$$V_{ii} = N_i p_i (1 - p_i).$$

Consequently, the score takes a particularly simple form:

$$\boldsymbol{S}(\boldsymbol{\beta}) = \boldsymbol{x}^{\mathrm{T}} \left[\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})\right].$$

Hence, at the maximum, $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Y} = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})$ so that selected sums of the outcomes (as defined by the design matrix) are preserved. In addition, element (j, j') of $I_n(\boldsymbol{\beta})$ takes the form

$$\sum_{i=1}^{n} x_{ij} x_{ij'} N_i p_i (1 - p_i).$$

We now turn to hypothesis testing and consider a model with $0 < q \leq k$ parameters and fitted probabilities $\widehat{\boldsymbol{p}}$. The log-likelihood is

$$l(\widehat{\boldsymbol{p}}) = \sum_{i=1}^{n} \left[y_i \log \widehat{p}_i + (N_i - y_i) \log(1 - \widehat{p}_i)\right],$$

with the maximum attainable value occurring at $\widetilde{p}_i = y_i / N_i$. The deviance is

$$\begin{aligned} D &= 2\left[l(\widetilde{\boldsymbol{p}}) - l(\widehat{\boldsymbol{p}})\right] \\ &= 2\sum_{i=1}^{n} \left[y_i \log\left(\frac{y_i}{\widehat{y}_i}\right) + (N_i - y_i) \log\left(\frac{N_i - y_i}{N_i - \widehat{y}_i}\right)\right], \end{aligned} \tag{7.14}$$

where $\widetilde{\boldsymbol{p}}$ is the vector of probabilities, $\widetilde{p}_i$, $i = 1, \ldots, n$. Notice that the deviance will be small when $\widehat{y}_i$ is close to y_i. The above form may also be derived directly from (6.22) under a binomial model. If n, the number of parameters in the saturated model (which, recall, is the number of conditions considered and not the total number of trials which is given by N), is fixed, then under the hypothesized model that produced $\widehat{\boldsymbol{p}}$, $D \;\rightarrow_d\; \chi^2_{n-q}$. The important emphasis here is on *fixed* n. The outcome after head injury dataset provides an example in which this assumption is valid since there are $n = 2 \times 2 \times 2 \times 3 = 24$ binomial trials being carried out at each combination of the levels of coma score, pupils, hematoma, and age.

When n is not fixed, the above result on the *absolute fit* is not relevant, but the *relative fit* may be assessed by comparing the difference in deviances. Specifically, consider nested models with q_j parameters under H_j, $j = 0, 1$. Further, the estimated probabilities and fitted values under hypothesis H_j will be denoted $\widehat{\boldsymbol{p}}_j$ and $\widehat{y}^{(j)}$, $j = 0, 1$, respectively. Then the reduction in deviance is

$$\begin{aligned} D_0 - D_1 &= 2\left\{l(\widetilde{\boldsymbol{p}}) - l(\widehat{\boldsymbol{p}}_0) - [l(\widetilde{\boldsymbol{p}}) - l(\widehat{\boldsymbol{p}}_1)]\right\} \\ &= 2\left[l(\widehat{\boldsymbol{p}}_1) - l(\widehat{\boldsymbol{p}}_0)\right] \\ &= 2\sum_{i=1}^{n}\left[y_i \log\left(\frac{\widehat{y}_i^{(1)}}{\widehat{y}_i^{(0)}}\right) + (N_i - y_i)\log\left(\frac{N_i - \widehat{y}_i^{(1)}}{N_i - \widehat{y}_i^{(0)}}\right)\right]. \end{aligned}$$

Under H_0, $D_0 - D_1 \to_d \chi^2_{q_1 - q_0}$.

When the denominators N_i are small, the deviance should not be used, as we now illustrate in the case of $N_i = 1$. Suppose that $Y_i \mid p_i \sim_{ind} \text{Bernoulli}(p_i)$, with a logistic model, $\text{logit}(p_i) = \boldsymbol{x}_i\boldsymbol{\beta}$, for $i = 1, \ldots, n$. We fit this model using maximum likelihood, resulting in estimates $\widehat{\boldsymbol{\beta}}$ and fitted probabilities $\widehat{\boldsymbol{p}}$. In this case, (7.14) becomes

$$\begin{aligned} D &= -2\sum_{i=1}^{n} y_i \log\left(\frac{\widehat{p}_i}{1-\widehat{p}_i}\right) - 2\sum_{i=1}^{n} y_i \log(1-\widehat{p}_i) \\ &= -2\boldsymbol{y}^{\mathrm{T}}\boldsymbol{x}\widehat{\boldsymbol{\beta}} - 2\sum_{i=1}^{n}\log(1-\widehat{p}_i) \\ &= -2\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} - 2\sum_{i=1}^{n}\log(1-\widehat{p}_i) \end{aligned}$$

since $y \log y = (1-y)\log(1-y) = 0$. At the MLE, $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{x}^{\mathrm{T}}\widehat{\boldsymbol{p}}$ so that

$$D = -2\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}^{\mathrm{T}}\widehat{\boldsymbol{p}} - 2\sum_{i=1}^{n}\log(1-\widehat{p}_i)$$

and the deviance is a function only of $\widehat{\boldsymbol{\beta}}$. In other words, D is a deterministic function of $\widehat{\boldsymbol{\beta}}$ only and cannot be used as a goodness of fit statistic. With small N_i, this is a problem for any link function.

An alternative goodness of fit measure for a model with q parameters is the Pearson statistic, as introduced in Sect. 6.5.3:

$$X^2 = \sum_{i=1}^{n}\frac{(Y_i - N_i\widehat{p}_i)^2}{N_i\widehat{p}_i(1-\widehat{p}_i)}, \tag{7.15}$$

with $X^2 \to_d \chi^2_{n-q}$ under the null and under the assumption of fixed n. The Pearson statistic also has problems with small N_i. For example, for the model $Y_i \mid p \sim_{ind} \text{Bernoulli}(p)$, $\widehat{p} = \overline{y}$ and

$$X^2 = \sum_{i=1}^{n}\frac{(y_i - \overline{y})^2}{\overline{y}(1-\overline{y})} = n,$$

which is not a useful goodness of fit measure (McCullagh and Nelder 1989, Sect. 4.4.5). The deviance also has problems under this Bernoulli model (Exercise 7.5).

7.6.3 *Quasi-likelihood Inference for Logistic Regression Models*

As we saw in Sect. 6.6, an extremely simple and appealing manner of dealing with overdispersion is to assume the model

$$\begin{aligned} \mathrm{E}[Y_i \mid \boldsymbol{\beta}] &= N_i p_i(\boldsymbol{\beta}) \\ \mathrm{var}(Y_i \mid \boldsymbol{\beta}) &= \alpha N_i p_i(\boldsymbol{\beta}) \left[1 - p_i(\boldsymbol{\beta})\right], \end{aligned}$$

with $\mathrm{cov}(Y_i, Y_j \mid \boldsymbol{\beta}) = 0$, for $i \neq j$. Under this model, due to the proportionality of the variance model, the maximum quasi-likelihood estimator satisfies the score function (7.12), since the value of α is irrelevant to finding the root of the estimating equation. Hence, the quasi-likelihood estimator $\widehat{\boldsymbol{\beta}}$ corresponds to the MLE. Interval estimates and tests are altered, however. In particular, asymptotic confidence intervals are derived from the variance–covariance $\widehat{\alpha}(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{D})^{-1}$. An obvious estimator of α is provided by the method of moments, which corresponds to the Pearson statistic (7.15) divided by $n - k - 1$. This estimator is consistent if the first two moments are correctly specified. The reference χ^2 distribution under the null is also perturbed, as in (6.27).

7.6.4 *Bayesian Inference for Logistic Regression Models*

A Bayesian approach to inference combines the likelihood $L(\boldsymbol{\beta})$ with a prior $\pi(\boldsymbol{\beta})$, with a multivariate normal distribution being the obvious choice. For the binomial model there is no conjugate distribution for general regression models. In simple situations with a small number of discrete covariates, one could specify beta priors with known parameters for each combination of levels and obtain analytic posteriors, but there would be no linkage between the different groups, that is, no transfer of information. With multivariate normal priors, computation may be carried out using INLA (Sect. 3.7.4), though this approximation strategy may be inaccurate if the binomial denominators are small (Fong et al. 2010). An alternative is provided by MCMC (Sect. 3.8).

As discussed in Sect. 7.6.3, it is common to encounter excess-binomial variation. This may be dealt with in a Bayesian context via the introduction of random effects. The beta-binomial described in Sect. 7.5 provides one possibility. An alternative, more flexible formulation would assume the two-stage model:

Stage One: The likelihood:

$$Y_i \mid \boldsymbol{\beta}, b_i \sim_{ind} \text{Binomial}\left[N, p(\boldsymbol{x}_i)\right]$$
$$\log\left[\frac{p(\boldsymbol{x}_i)}{1-p(\boldsymbol{x}_i)}\right] = \boldsymbol{x}_i\boldsymbol{\beta} + b_i$$

Stage Two: The random effects distribution:

$$b_i \mid \sigma_0^2 \sim_{iid} \text{N}(0, \sigma_0^2).$$

The parameter σ_0^2 controls the amount of overdispersion, though not in a simple fashion. A Bayesian approach adds priors on $\boldsymbol{\beta}$ and σ_0^2. This model is discussed further in Sect. 9.13.

Example: Outcome After Head Injury

Parameter estimation, whether via likelihood or Bayes, is straightforward for these data *given* a particular model. The difficult task in this problem is deciding upon a model. If prediction is all that is required, then Bayesian model averaging provides one possibility, and this is explored for these data in Chap. 12.

In exploratory mode, we illustrate some approaches to model selection. In Sect. 4.8, approaches to variable selection were reviewed and critiqued. In particular, the hierarchy principle, in which all interactions are accompanied by their constituent main effect, was discussed. Even applying the hierarchy principle here, there are still 167 models with $k = 4$ variables.

We begin by applying forward selection (obeying the hierarchy principle), beginning with the null model and using AIC as the selection criteria. This leads to a model with all main effects and the three two-way interactions `H.P`, `H.A`, and `P.A`. Since there are $n = 24$ fixed cells here we can assess the overall fit. The deviance associated with the model selected via forward selection is 13.6 on 13 degrees of freedom which indicates a good fit. Applying backward elimination produces a model with all main effects and five two-way interactions, the three selected using forward selection and, in addition, `H.C` and `C.A`. This model has a deviance of 7.0 on 10 degrees of freedom, so the overall fit is good.

Carrying out an exhaustive search over all 167 models using AIC as the criterion leads to the model selected with backward selection (i.e., main effects plus five two-way interactions). Using BIC as the criteria leads to a far simpler model with the main effects `H`, `C`, and `A` only. It is often found that BIC picks simpler models.

We consider inference for the model:

```
1+H+P+C+A2+A3+H.P+H.A2+H.A3+P.A2+P.A3,          (7.16)
```

Table 7.2 Likelihood and Bayesian estimates and uncertainty measures for model (7.16) applied to the head injury data

	MLE	Std. err.	Post. mean	Post S.D.
`1`	−1.39	0.26	−1.37	0.26
`H`	1.03	0.35	1.02	0.35
`P`	2.05	0.30	2.04	0.29
`C`	−1.52	0.17	−1.53	0.17
`A2`	1.20	0.33	1.18	0.32
`A3`	3.69	0.48	3.68	0.47
`H.P`	−0.55	0.34	−0.55	0.34
`H.A2`	−0.39	0.36	−0.38	0.36
`H.A3`	−1.32	0.53	−1.29	0.52
`P.A2`	−0.57	0.37	−0.56	0.36
`P.A3`	−1.35	0.49	−1.33	0.48

that is, the model with main effects for hematoma (`H`), pupils (`P`), coma score (`C`), and age (with `A2` and `A3` representing the second and third levels) and with interactions between hematoma and pupils (`H.P`), hematoma and age (`H.A2` and `H.A3`), and pupils and age (`P.A2` and `P.A3`).

The MLEs and standard errors are given in Table 7.2, along with Bayesian posterior means and standard deviations. The prior on the intercept was taken as flat, and for the 10 log odds ratios, independent normal priors $\mathrm{N}(0, 4.70^2)$ were taken, which correspond to 95% intervals for the odds ratios of [0.0001,10000], that is, very weak prior information was incorporated. The INLA method was used for computation. The original scale of the parameters is given in the table, which is not ideal for interpretation, but makes sense for comparison of results since the sampling distributions and posteriors are close to normal. The first thing to note is that inference from the two approaches is virtually identical. This is not surprising, given the relatively large counts and weak priors.

The pupil and age variables and their interaction at the highest age level are clearly very important. The high coma score parameter is large and negative, and since the coma variable is not involved in any interactions, we can say that having a high coma score reduces the odds of death by $\exp(-1.52) = 0.22$.

The observed and fitted probabilities are displayed in Fig. 7.3 with different line types joining the observed probabilities (as in Fig. 7.1). The vertical lines join the fitted to the observed probabilities, with the same line type as the observed probabilities with which they are associated. There are no clear badly fitting cells.

Example: Aircraft Fasteners

Let Y_i be the number of fasteners failing at pressure x_i, and assume $Y_i \mid p_i \sim_{ind} \text{Binomial}(n_i, p_i)$, $i = 1, \ldots, n$, with the logistic model $\text{logit}(p_i) = \beta_0 + \beta_1 x_i$. This specification yields likelihood

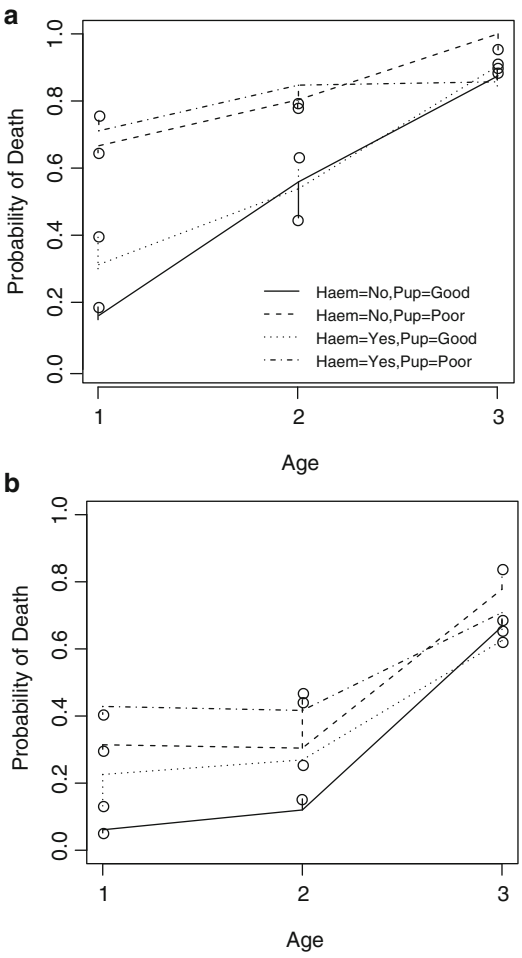

Fig. 7.3 Probability of death after head injury as a function of age, hematoma score, and pupils. Panels (**a**) and (**b**) are for low and high coma scores, respectively. The *open circles* are the fitted values. The observed values are joined by different line types. The residuals $y/n - \widehat{p}$ are shown as *vertical lines* of the same line type

$$L(\boldsymbol{\beta}) = \exp\left(\beta_0 \sum_{i=1}^{n} y_i + \beta_1 \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} n_i \log\left[1 + \exp(\beta_0 + \beta_1 x_i)\right]\right) \tag{7.17}$$

where $\boldsymbol{\beta} = [\beta_0, \beta_1]^{\mathrm{T}}$. The MLEs and variance–covariance matrix are

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} -5.34 \\ 0.0015 \end{bmatrix}, \ \widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}}) = \begin{bmatrix} 2.98 \times 10^{-1} & -8.50 \times 10^{-5} \\ -8.50 \times 10^{-5} & 2.48 \times 10^{-8} \end{bmatrix}. \tag{7.18}$$

The solid line in Fig. 7.4 is the fitted curve $\widehat{p}(x)$ corresponding to the MLE. The fit appears good. For comparison we also fit models with complementary log–log and log–log link functions, as described in Sect. 7.4.2. Figure 7.4 shows the fit from

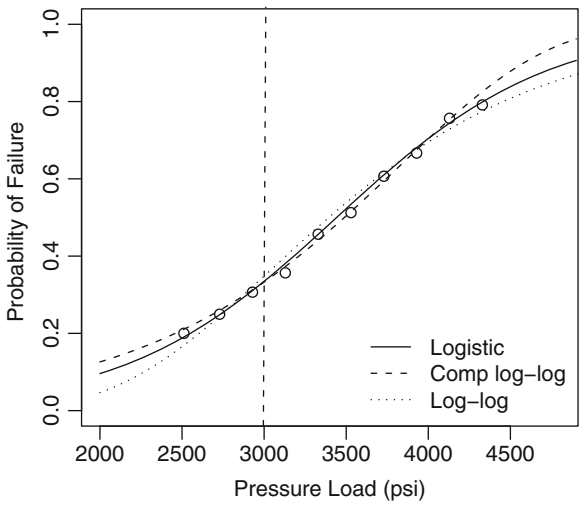

Fig. 7.4 Fitted curves for the aircraft fasteners data under three different link functions

these models. The residual deviance from logistic, complementary log–log, and log–log links are 0.37, 0.69, and 1.7, respectively. These values are not comparable via likelihood ratio tests since the models are not nested. AIC (Sect. 4.8.2) can be used for such comparisons, but the approximations inherent in the derivation are more accurate for nested models (Ripley 2004 and Sect. 10.6.4). The differences are so small here that we would not make any conclusions on the basis of these numbers. Since the number of x categories is not fixed in this example, we cannot formally examine the absolute fit of the models. In Fig. 7.6, we see that residual plots for these three models indicate that the logistic fit is preferable.

A 95% confidence interval for the odds ratio corresponding to a 500 psi increase in pressure load is

$$\exp\left[500 \times \widehat{\beta}_1 \pm 1.96 \times 500\sqrt{\text{var}(\widehat{\beta}_1)}\right] = [1.86, 2.53]. \tag{7.19}$$

We now present a Bayesian analysis. For these abundant data and without any available prior information, the improper uniform prior $\pi(\boldsymbol{\beta}) \propto 1$ is assumed. The posterior is therefore proportional to (7.17). We use a bivariate Metropolis–Hastings random walk MCMC algorithm (Sect. 3.8.2) to explore the posterior. A bivariate normal proposal was used, with variance–covariance matrix proportional to the asymptotic variance–covariance matrix, $\widehat{\text{var}}(\widehat{\boldsymbol{\beta}})$, (7.18). This matrix was multiplied by four to give an acceptance ratio of around 30%. Panels (a) and (b) of Fig. 7.5 show histograms of the dependent samples from the posterior $\beta_0^{(s)}$ and $\beta_1^{(s)}$, $s = 1, \ldots, S = 500$, and panel (c) the bivariate posterior. The posterior median for $\boldsymbol{\beta}$ is $[-5.36, 0.0015]$, and a 95% posterior interval for the odds ratio corresponding to a 500 psi increase in pressure is identical to the asymptotic likelihood interval (7.19).

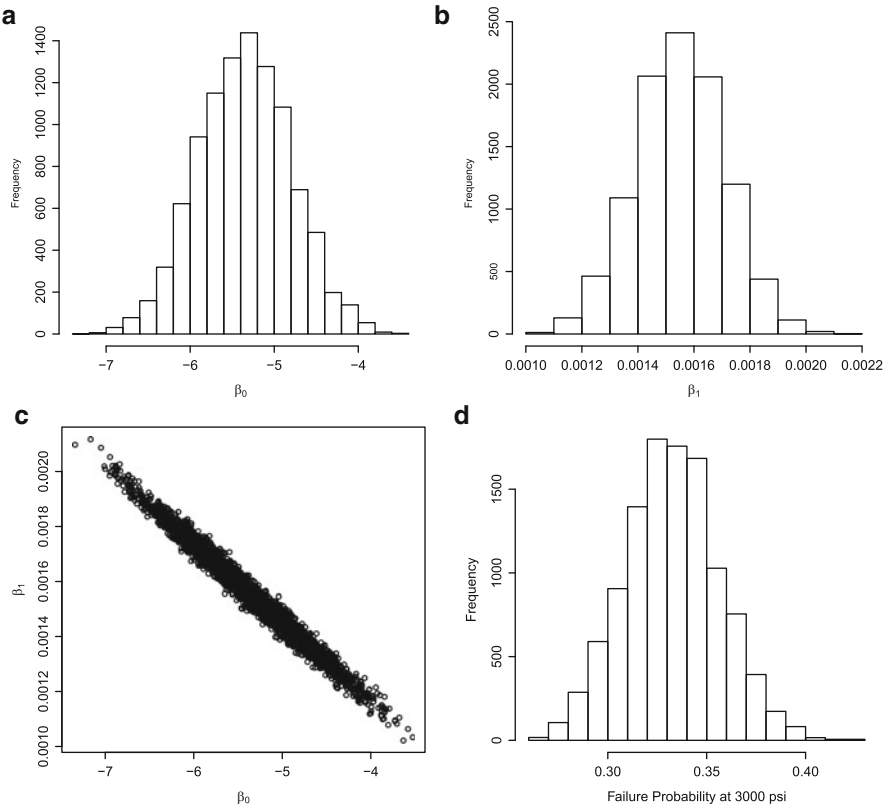

Fig. 7.5 Posterior summaries for the aircraft fasteners data: (**a**) $p(\beta_0|\boldsymbol{y})$, (**b**) $p(\beta_1|\boldsymbol{y})$, (**c**) $p(\beta_0, \beta_1|\boldsymbol{y})$, (**d**) $p(\exp(\theta)/[1+\exp(\theta)]|\boldsymbol{y})$, where $\theta = \beta_0 + \beta_1\tilde{x}$, that is, the posterior for the probability of failure at a load of $\tilde{x} = 3{,}000$ psi

We now imagine that it is of interest to give an interval estimate for the probability of failure at $\widetilde{x} = 3{,}000$ psi (which is indicated as a dashed vertical line on Fig. 7.4). An asymptotic 95% confidence interval for $\theta = \beta_0 + \beta_1\widetilde{x}$ is

$$\widehat{\theta} \pm 1.96 \times \sqrt{\text{var}(\widehat{\theta})},$$

where

$$\begin{aligned} \widehat{\theta} &= \widehat{\beta}_0 + \widetilde{x}\widehat{\beta}_1 \\ \text{var}(\widehat{\theta}) &= \text{var}(\widehat{\beta}_0) + 2\widetilde{x}\text{cov}(\widehat{\beta}_0, \widehat{\beta}_1) + \widetilde{x}^2\text{var}(\widehat{\beta}_1). \end{aligned}$$

Taking the expit transform of the endpoints of the confidence interval on the linear predictor scale leads to a 95% interval of [0.29,0.38]. Substitution of the posterior

Table 7.3 A generic 2×2 table

	$Y=0$	$Y=1$	
$X=0$	y_{00}	y_{01}	$y_{0\cdot}$
$X=1$	y_{10}	y_{11}	$y_{1\cdot}$
	$y_{\cdot 0}$	$y_{\cdot 1}$	$y_{\cdot\cdot}$

samples $\boldsymbol{\beta}^{(s)}$ to give expit$(\theta^{(s)})$, $s = 1, \ldots, S$ results in a 95% interval which is again identical to the frequentist interval.

7.7 Conditional Likelihood Inference

In Sect. 2.4.2, conditional likelihood was introduced as a procedure that could be used for eliminating nuisance parameters. In this chapter, conditional likelihood will be used for discrete data, which we denote $\boldsymbol{y}$. Suppose the distribution for the data can be represented as,

$$p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{\phi}) \propto p(\boldsymbol{t}_1 \mid \boldsymbol{t}_2, \boldsymbol{\lambda}) p(\boldsymbol{t}_2 \mid \boldsymbol{\lambda}, \boldsymbol{\phi}), \tag{7.20}$$

where $\boldsymbol{\lambda}$ is a parameter of interest and $\boldsymbol{\phi}$ is a nuisance parameter. Then inference for $\boldsymbol{\lambda}$ may be based on the *conditional likelihood*

$$L_c(\boldsymbol{\lambda}) = p(\boldsymbol{t}_1 \mid \boldsymbol{t}_2, \boldsymbol{\lambda}).$$

Perhaps the most popular use of conditional likelihood leads to Fisher's exact test. Consider the 2×2 layout of data shown in Table 7.3 with

$$\begin{aligned} y_{01} \mid p_0 &\sim \text{Binomial}(y_{0\cdot}, p_0) \\ y_{11} \mid p_1 &\sim \text{Binomial}(y_{1\cdot}, p_1), \end{aligned}$$

which we combine with the logistic regression model:

$$\begin{aligned} \log\left(\frac{p_0}{1-p_0}\right) &= \beta_0 \\ \log\left(\frac{p_1}{1-p_1}\right) &= \beta_0 + \beta_1. \end{aligned}$$

Here,

$$\exp(\beta_1) = \frac{p_1/(1-p_1)}{p_0/(1-p_0)}$$

is the odds of a positive response in the $X = 1$ group, divided by the odds of a positive response in the $X = 0$ group, that is, the odds ratio. This setup gives likelihood

$$\Pr(y_{01}, y_{11} \mid \beta_0, \beta_1) = \binom{y_{0\cdot}}{y_{01}} \binom{y_{1\cdot}}{y_{11}} \frac{e^{y_{11}\beta_1}}{(1+e^{\beta_0+\beta_1})^{y_{1\cdot}}} \frac{e^{y_{\cdot 1}\beta_0}}{(1+e^{\beta_0})^{y_{0\cdot}}}. \tag{7.21}$$

Now $[y_{01}, y_{11}]$ implies the distribution of $[y_{11}, y_{\cdot 1}]$, so we can write

$$\Pr(y_{11}, y_{\cdot 1} \mid \beta_0, \beta_1) = \binom{y_{0\cdot}}{y_{\cdot 1} - y_{11}} \binom{y_{1\cdot}}{y_{11}} \frac{e^{y_{11}\beta_1}}{(1+e^{\beta_0+\beta_1})^{y_{1\cdot}}} \frac{e^{y_{\cdot 1}\beta_0}}{(1+e^{\beta_0})^{y_{0\cdot}}}.$$

We now show that by conditioning on the column totals, in addition to the row totals, we obtain a distribution that depends only on the parameter of interest β_1. Consider

$$\Pr(y_{11} \mid y_{\cdot 1}, \beta_0, \beta_1) = \frac{\Pr(y_{11}, y_{\cdot 1} \mid \beta_0, \beta_1)}{\Pr(y_{\cdot 1} \mid \beta_0, \beta_1)},$$

where the marginal distribution is obtained by summing over the possible values that y_{11} can take, that is,

$$\begin{aligned}\Pr(y_{\cdot 1} \mid \beta_0, \beta_1) &= \sum_{u=u_0}^{u_1} \Pr(u, y_{\cdot 1} \mid \beta_0, \beta_1) \\ &= \sum_{u=u_0}^{u_1} \binom{y_{0\cdot}}{y_{\cdot 1} - u} \binom{y_{1\cdot}}{u} \frac{e^{u\beta_1}}{(1+e^{\beta_0+\beta_1})^{y_{1\cdot}}} \frac{e^{y_{\cdot 1}\beta_0}}{(1+e^{\beta_0})^{y_{0\cdot}}}\end{aligned}$$

where $u_0 = \max(0, y_{\cdot 1} - y_{0\cdot})$ and $u_1 = \min(y_{1\cdot}, y_{\cdot 1})$ ensure that the marginals are preserved. With respect to (7.20), $\boldsymbol{\lambda} \equiv \beta_1$, $\boldsymbol{\phi} \equiv \beta_0$, $\boldsymbol{t}_1 \equiv y_{11}$, and $\boldsymbol{t}_2 \equiv y_{\cdot 1}$. Accordingly, the conditional distribution takes the form

$$\Pr(y_{11} \mid y_{\cdot 1}, \beta_1) = \frac{\binom{y_{0\cdot}}{y_{\cdot 1} - y_{11}} \binom{y_{1\cdot}}{y_{11}} e^{y_{11}\beta_1}}{\sum_{u=u_0}^{u_1} \binom{y_{0\cdot}}{y_{\cdot 1} - u} \binom{y_{1\cdot}}{u} e^{u\beta_1}}, \tag{7.22}$$

an *extended hypergeometric* distribution. We have removed the conditioning on β_0 since this distribution depends on β_1 only (which was the point of this derivation). Inference for β_1 may be based on the conditional likelihood (7.22). In particular, the conditional MLE may be determined, though unfortunately no closed form exists.

Conventionally, estimates of β_0 and β_1 would be determined from the product of binomial likelihoods, (7.21). Unless the samples are small, the conditional and

unconditional MLEs (and associated variances) will be in close agreement, but for small samples, the conditional MLE is preferred due to the following informal argument. Consider the original 2×2 data in Table 7.3. If we knew $y_{\cdot 1}$, then this alone would not help us to estimate β_1, *but* the precision of conclusions about β_1 will depend on this column total, and we should therefore condition on the observed value. This is to ensure that we attach to the conclusions the precision actually achieved and not that to be achieved hypothetically in a particular situation that has in fact not occurred. For further discussion, see Cox and Snell (1989, p. 27–29).

To derive the conditional MLE, first consider the conditional likelihood

$$L_c(\beta_1) = \frac{c(y_{11})e^{y_{11}\beta_1}}{\sum_{u=u_0}^{u_1} c(u)e^{u\beta_1}}$$

where

$$c(u) = \binom{y_{0\cdot}}{y_{\cdot 1} - u}\binom{y_{1\cdot}}{u}.$$

The (conditional) score is

$$S_c(\beta_1) = \frac{\partial}{\partial \beta_1}\log L_c(\beta_1) = y_{11} - \frac{\sum_{u=u_0}^{u_1} c(u)ue^{\widehat{\beta}_1 u}}{\sum_{u=u_0}^{u_1} c(u)e^{\widehat{\beta}_1 u}}. \tag{7.23}$$

The extended hypergeometric distribution is a member of the exponential family (Exercise 7.6) and

$$\mathrm{E}[S_c(\beta_1)] = \frac{\partial}{\partial \beta_1}\log L_c(\beta_1)\Big|_{\widehat{\beta}_1} = 0,$$

at the MLE. Consequently, from (7.23), we can use the equation $\mathrm{E}[Y_{11} \mid \widehat{\beta}_1] = y_{11}$ to solve for $\widehat{\beta}_1$. Asymptotic inference is based on

$$I_c(\beta_1)^{1/2}\left(\widehat{\beta}_1 - \beta_1\right) \to_d \mathrm{N}(0,1), \tag{7.24}$$

where the (conditional) information is

$$\begin{aligned} I_c(\beta_1) &= -\frac{\partial^2}{\partial \beta_1^2}\log L_c(\beta_1) = \frac{\sum_{u=u_0}^{u_1} c(u)u^2 e^{\widehat{\beta}_1 u}}{\sum_{u=u_0}^{u_1} c(u)e^{\widehat{\beta}_1 u}} - \left(\frac{\sum_{u=u_0}^{u_1} c(u)ue^{\widehat{\beta}_1 u}}{\sum_{u=u_0}^{u_1} c(u)e^{\widehat{\beta}_1 u}}\right)^2 \\ &= \mathrm{var}(Y_{11} \mid \beta_1). \end{aligned}$$

It is straightforward to test the null hypothesis $H_0 : \beta_1 = 0$ using the conditional likelihood. When $\beta_1 = 0$, the distribution (7.22) is hypergeometric, and so

Table 7.4 Data on tumor appearance within rats

		Tumor		
		Absent $Y = 0$	Present $Y = 1$	
Control	$X = 0$	13	19	32
Treated	$X = 1$	2	21	23
		15	40	55

$$\Pr(y_{11} \mid y_{\cdot 1}, \beta_1 = 0) = \frac{\binom{y_{0\cdot}}{y_{\cdot 1} - y_{11}} \binom{y_{1\cdot}}{y_{11}}}{\binom{y_{\cdot\cdot}}{y_{\cdot 1}}}. \tag{7.25}$$

The comparison of the observed y_{11} with the tail of this distribution is known as *Fisher's exact test* (Fisher 1935). Various possibilities are available to obtain a two-sided significance level, the simplest being to double the one-sided p-value. An alternative is provided by summing all probabilities less than the observed table. Confidence intervals for β_1 may be obtained from (7.24), or by inverting the test. See Agresti (1990, Sects. 3.5 and 3.6) for further discussion; in particular, the problems of the discreteness of the sampling distribution are discussed.

Example: Tumor Appearance Within Mice

We illustrate the application of conditional likelihood using data reported by Essenberg (1952) and presented in Table 7.4. To examine the carcinogenic effects of tobacco, 36 albino mice were placed in an enclosed chamber which was filled with the smoke of one cigarette every 12 h per day. Another group of mice were kept in an alternative chamber without smoke. After 1 year, autopsies were carried out on those mice that had survived for at least the first 2 months of the experiment. The data in Table 7.4 give the numbers of mice with and without tumors in the "control" and "treated" groups.

For these data, the permissible values of y_{11} lie between $u_0 = \max(0, 40 - 32) = 8$ and $u_1 = \min(23, 40) = 23$. Under $H_0 : \beta_1 = 0$, the probabilities of $y_{11} = 21, 22, 23$, from (7.25), are 0.00739, 0.00091, and 0.00005, which sum to 0.00834, the one-sided p-value. The simplest version of the two-sided p-value is therefore 0.0167, which would lead to rejection of H_0 under the usual threshold of 0.05. Summing the probabilities of more extreme tables gives a p-value of 0.0130.

Denoting by $\widehat{\beta}_1^u$ the (unconditional) MLE of the log odds ratio, we have

$$\widehat{\beta}_1^u = \log\left(\frac{21 \times 13}{2 \times 19}\right) = \log(7.18) = 1.97,$$

with asymptotic standard error

$$\sqrt{\widehat{\text{var}}(\widehat{\beta}_1^u)} = \sqrt{\frac{1}{2} + \frac{1}{21} + \frac{1}{13} + \frac{1}{19}} = 0.82,$$

to give asymptotic 95% confidence interval for the odds ratio of

$$\exp(1.97 \pm 1.96 \times 0.82) = [1.44, 35.8]\,.$$

The Wald test p-value of 0.0166 is very close to that obtained from Fisher's exact test. The conditional MLE is

$$\widehat{\beta}_1 = \log(6.95) = 1.93$$

with conditional standard error

$$\sqrt{\text{var}(\widehat{\beta}_1)} = 0.61,$$

illustrating the extra precision gained by conditioning on $y_{\cdot 1}$. The conditional asymptotic 95% confidence interval for the odds ratio based on (7.24) is

$$\exp(1.93 \pm 1.96 \times 0.61) = [2.11, 22.9]\,.$$

7.8 Assessment of Assumptions

In general, residual analysis is subjective, and though one might be able to conclude that a model is inadequate, concluding adequacy is much more difficult. Unfortunately, for logistic regression models with binary data, the assessment is even more tentative. Even when the model is true, little can be said about the moments and distribution of the residuals.

We briefly review Pearson and deviance residuals as defined for GLMs in Sect. 6.9. Pearson residuals are defined as $e_i^\star = (Y_i - \widehat{\mu}_i)/\sqrt{\widehat{\text{var}}(Y_i)}$, and for $Y_i \mid p_i \sim \text{Binomial}(n_i, p_i)$, we obtain

$$e_i^\star = \frac{y_i - n_i\widehat{p}_i}{[n_i\widehat{p}_i(1-\widehat{p}_i)]^{1/2}},$$

$i = 1, \ldots, n$. Pearson's statistic is

$$X^2 = \sum_{i=1}^{n} \frac{(Y_i - n_i\widehat{p}_i)^2}{n_i\widehat{p}_i(1-\widehat{p}_i)} = \sum_{i=1}^{n} (e_i^\star)^2,$$

showing the link between the measures of local and absolute fit.

Deviance residuals are defined as

$$e_i^\star = \text{sign}(y_i - \widehat{\mu}_i)\sqrt{D}_i,$$

$i = 1, \ldots, N$. Note that the deviance $D = \sum_{i=1}^{n}(e_i^\star)^2$ where D is given by (7.14). For binary Y_i and a particular value of $\widehat{p}_i$, the residuals can only take one of two possible values, which is clearly a problem (this is illustrated later, in Fig. 7.8).

Few analytical results are available for the case of a binomial model, but, if the model is correct, both the Pearson and deviance residuals are asymptotically normally distributed. Hence, they may be put to many of the same uses as residual defined with respect to the normal linear regression model (as described in Sect. 5.11.3). For example, residuals may be plotted against covariates x and examined for outlying values. Interpretation is more difficult, however, as one must examine the appropriateness of the link function as well as the linearity assumption. A normal QQ plot of residuals can indicate outlying observations.

Empirical logits $\log[(y_i + 0.5)/(N_i - y_i + 0.5)]$ are useful for examining the adequacy of the logistic linear model. The addition of 0.5 removes problems when $y_i = 0$ or N_i. This adjustment is optimal; see Cox and Snell (1989, Sect. 2.1.6) for details. The mean–variance relationship can be examined by plotting residuals versus fitted values. In particular, different overdispersion models may be compared, as discussed in Sect. 7.5.

Example: Aircraft Fasteners

In this example, the denominators are relatively large (ranging between 40 and 100 for each of the 10 trials), and so the residuals are informative. Figure 7.6 shows Pearson residuals plotted against pressure load for each of three different link functions. On the basis of these plots, the logistic model looks the most reasonable since there are runs of positive and negative residuals associated with the other two link functions, signifying mean model misspecification.

Example: Outcome After Head Injury

The binary response in this example is cross-classified with respect to factors with 2 or 3 levels. We saw in Fig. 7.3 that the fit of model (7.16) appeared reasonable, though the distances $\frac{y_i}{n_i} - \widehat{p}_i$ that are displayed as vertical lines are not standardized, making interpretation difficult. Figure 7.7 gives a normal QQ plot of the Pearson residuals, and there are no obvious causes for concern with no outlying points.

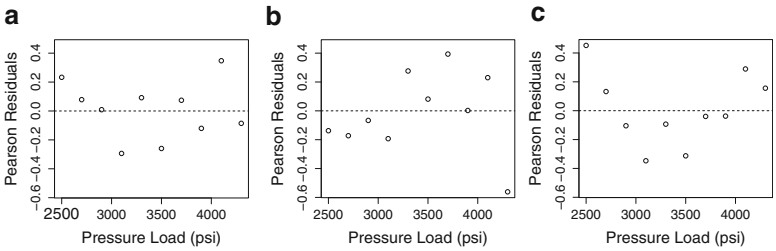

Fig. 7.6 Pearson residuals versus pressure load for the aircraft fasteners data for (**a**) logistic link model, (**b**) complementary log–log link model, and (**c**) log–log link model

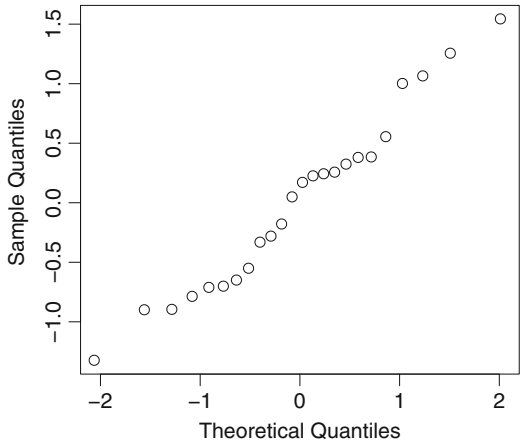

Fig. 7.7 QQ plot of Pearson residuals for the head injury data

Example: BPD and Birth Weight

We fit a logistic regression model

$$\Pr(Y = 1 \mid x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}, \tag{7.26}$$

with $Y = 0/1$ corresponding to absence/presence of BPD and x to birth weight. The curve arising from fitting this model is shown in Fig. 7.2, along with the curve from the use of the complementary log–log link. We might question whether either of these curves is adequate, since they are relatively inflexible, with forms determined by two parameters only. The Pearson residuals from the two models are plotted versus birth weight in Fig. 7.8. The binary nature of the response is evident in these plots, and assessing whether the models are adequate is not possible from this plot. In Chap. 11, we return to these data and fit flexible nonlinear models.

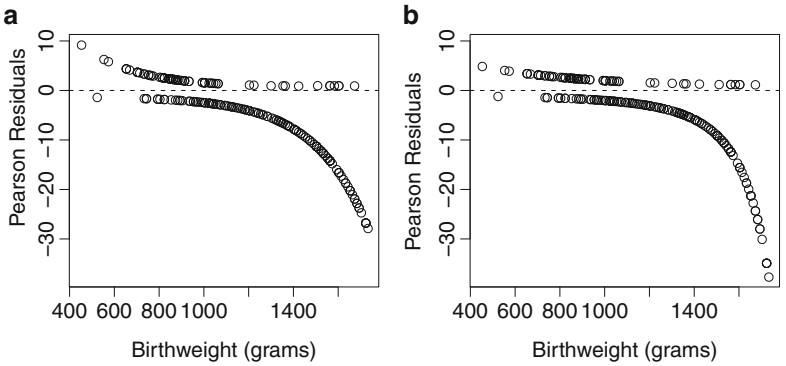

Fig. 7.8 Pearson residuals versus birth weight for the BPD data: (**a**) logistic model, (**b**) complementary log–log model

7.9 Bias, Variance, and Collapsibility

We begin by summarizing some of the results of Sect. 5.9 in which the bias and variance of estimators were examined for the linear model. Consider the models:

$$\mathrm{E}[Y \mid x, z] = \beta_0 + \beta_1 x + \beta_2 z \tag{7.27}$$

$$\mathrm{E}[Y \mid x] = \beta_0^\star + \beta_1^\star x. \tag{7.28}$$

First, suppose that x and z are orthogonal. Roughly speaking, if z is related to Y, then fitting model (7.27) will lead to a reduction in the variance of $\widehat{\beta}_1$, and $\mathrm{E}[\widehat{\beta}_1] = \mathrm{E}[\widehat{\beta}_1^\star]$ so that bias is not an issue. When x and z are not orthogonal, then fitting model (7.28) will lead to bias in the estimation of β_1 since $\beta_1^\star$ reflects not only x but also the effect of z through its association with x.

In this section we discuss these issues with respect to logistic regression models. To this end, consider the *logistic models*:

$$\mathrm{E}[Y|x, z] = \frac{\exp(\beta_0 + \beta_1 x + \beta_2 z)}{1 + \exp(\beta_0 + \beta_1 x + \beta_2 z)} \tag{7.29}$$

$$\mathrm{E}[Y|x] = \frac{\exp(\beta_0^* + \beta_1^* x)}{1 + \exp(\beta_0^* + \beta_1^* x)} = \mathrm{E}_{z|x}\left[\frac{\exp(\beta_0 + \beta_1 x + \beta_2 z)}{1 + \exp(\beta_0 + \beta_1 x + \beta_2 z)}\right]. \tag{7.30}$$

The last equation indicates that determining the effects of omission of z will be very hard to determine due to the nonlinearity of the logistic function. As we illustrate shortly though, even if x and z are orthogonal, $\mathrm{E}[\beta_1] \neq \mathrm{E}[\beta_1^\star]$. Linear models for the probabilities are more straightforward to understand, but, as discussed previously, since probabilities are constrained $[0, 1]$, such models are rarely appropriate for binary data.

Table 7.5 Illustration of Simpson's paradox for the case of non-orthogonal x and z

		$z=0$		$z=1$		Marginal	
		$Y=0$	$Y=1$	$Y=0$	$Y=1$	$Y=0$	$Y=1$
Control	$x=0$	8	2	9	21	17	23
Treatment	$x=1$	18	12	2	8	20	20
	Odds Ratio	1.6		1.7		0.7	

We now discuss the marginalization of effect measures. Roughly speaking, if an effect measure is constant across strata (subtables) and equal to the measure calculated from the marginal table, it is known as *collapsible*. Non-collapsibility is sometimes referred to as Simpson's paradox (Simpson 1951) in the statistics literature. As in Greenland et al. (1999), we include the case of orthogonal x and z in Simpson's paradox, though first illustrate with a case in which x and z are non-orthogonal.

Consider the data in Table 7.5 in which $x = 0/1$ represents a control/treatment which is applied in two strata $z = 0/1$, with a binary response $Y = 0/1$ being recorded. In both z strata, the treatment appears beneficial with odds ratios of 1.6 and 1.7. However, when the data are collapsed over strata, the marginal association is reversed to give an odds ratio of 0.7 so that the treatment appears detrimental.

Mathematically, the paradox is relatively simple to understand. Let

$$p_{xz} = \Pr(Y = 1 \mid X = x, Z = z)$$
$$p_x^* = \Pr(Y = 1 \mid X = x)$$

be the conditional and marginal probabilities of a response and $q_x = \Pr(Z = 1 \mid X = x)$ summarize the relationship between x and z, for $x, z = 0, 1$. The "paradox" reflects the fact that it is possible to have

$$p_{00} < p_{10} \quad \text{and} \quad p_{01} < p_{11},$$

that is, the probability of a positive response being greater under $X = 1$ for both strata, but

$$p_{00}(1 - q_0) + p_{01}q_0 = p_0^* > p_1^* = p_{10}(1 - q_1) + p_{11}q_1$$

so that the marginal probability of a positive response is greater under $x = 0$ than under $x = 1$. For the data of Table 7.5,

$$p_{00} = \frac{2}{10} = 0.20, \quad p_{10} = \frac{13}{30} = 0.43, \quad p_{01} = \frac{21}{30} = 0.7, \quad p_{11} = \frac{8}{10} = 0.8,$$

and

$$p_0^* = \frac{23}{40} = 0.58, \quad p_1^* = \frac{20}{40} = 0.50,$$

Table 7.6 Illustration of Simpson's paradox for the case of orthogonal x and z

		$z=0$		$z=1$		Marginal	
		$Y=0$	$Y=1$	$Y=0$	$Y=1$	$Y=0$	$Y=1$
Control	$x=0$	95	5	10	90	105	95
Treatment	$x=1$	90	10	5	95	95	105
	Odds ratio	2.1		2.1		1.2	

with

$$q_0 = \frac{30}{40}, \quad q_1 = \frac{10}{40}.$$

It is important to realize that the paradox has nothing to do with the absolute values of the counts. Reversal of the association (as measured by the odds ratio) cannot occur if $q_0 = q_1$ (i.e., if there is no confounding), but the odds ratio is still non-collapsible, as the next example illustrates.

We now consider the situation in which $q_0 = q_1$. Such a balanced situation would occur, by construction, in a randomized clinical trial in which (say) equal numbers of $x = 0$ and $x = 1$ groups receive the treatment. We illustrate in Table 7.6 in which there are 100 patients in each of the four combinations of x and z. In each of the z stratum, we see an odds ratio for the treatment as compared to the control of 2.1. We do not see a reversal in the direction of the association but rather an attenuation toward the null, with the marginal association being 1.2.

We emphasize that the marginal estimator is not a biased estimate, but is rather estimating a different quantity, the averaged or marginal association. A second point to emphasize is that, as we have just illustrated, collapsibility and confounding are different issues and should not be confused. In particular, it is possible to have confounding present without non-collapsibility, as discussed in Greenland et al. (1999).

Another issue that we briefly discuss is the effect of stratification on the variance of an estimator. As discussed at the start of this section, if x and z are orthogonal but z is associated with y, then including z in a linear model will increase the precision of the estimator of the association between y and x. We illustrate numerically that this is not the case in the logistic regression context, again referring to the data in Table 7.6. Let p_{xz} represent the probability of disease for treatment group x and strata z. In the conditional analysis we fit the model

$$\log\left(\frac{p_{xz}}{1-p_{xz}}\right) = \begin{cases} \beta_0 & \text{for } x=0, z=0 \\ \beta_0 + \beta_{\mathrm{x}} & \text{for } x=1, z=0 \\ \beta_0 + \beta_{\mathrm{z}} & \text{for } x=0, z=1 \\ \beta_0 + \beta_{\mathrm{x}} + \beta_{\mathrm{z}} & \text{for } x=1, z=1, \end{cases}$$

where we have not included an interaction between x and z. This results in $\exp(\beta_{\mathrm{x}}) = \exp(0.75) = 2.1$, as expected from Table 7.6, with standard error 0.40.

Now suppose we ignore the stratum information and let $p_x^\star$ be the probability of disease for treatment group x. We fit the model

$$\log\left(\frac{p_x^\star}{1-p_x^\star}\right) = \begin{cases} \beta_0^\star & \text{for } x = 0 \\ \beta_0^\star + \beta_{\mathrm{x}}^\star & \text{for } x = 1 \end{cases}$$

This gives $\exp(\beta_{\mathrm{x}}^\star) = \exp(0.20) = 1.2$, again as expected from Table 7.6, but with standard error 0.20 which is a reduction from the conditional model and is in stark contrast to the behavior we saw with the linear model.

In any cross-classified table the summary we observe is an "averaged" measure, where the average is with respect to the population underlying that table. Consider the right-hand 2×2 set of counts in Table 7.6, in which we had equal numbers in each strata (which mimics a randomized trial). The odds ratio comparing treatment to control is 1.2 here and is the effect averaged across strata (and any other variables that were unobserved). Such measures are relevant to what are sometimes referred to as *population* contrasts. Depending on the context, we will often wish to include additional covariates in order to obtain effect measures most relevant to particular subgroups (or subpopulations). The issues here have much in common with marginal and conditional modeling as discussed in the context of dependent data in Chaps. 8 and 9.

We emphasize that, as mentioned above, the difference between population and subpopulation-specific estimates should not be referred to as "bias" since different quantities are being estimated. As a final note, the discussion in this section has centered on logistic regression models, but the same issues hold for other nonlinear summary measures.

7.10 Case-Control Studies

In this section we discuss a very popular design in epidemiology, the case-control study. In the econometrics literature, this design is known as *choice-based sampling*.

7.10.1 The Epidemiological Context

Cohort (prospective) studies investigate the causes of disease by proceeding in the natural way from cause to effect. Specifically, individuals in different exposure groups of interest are enrolled, and then one observes whether they develop the disease or not over some time period. In contrast, case-control (retrospective) studies proceed from effect to cause. Cases and disease-free controls are identified, and then the exposure status of these individuals is determined. Table 7.7 demonstrates the simplest example in which there is a single binary exposure, with y_{ij} representing

Table 7.7 Generic 2×2 table for a binary exposure and binary disease outcome

		Not diseased $Y = 0$	Diseased $Y = 1$	
Unexposed	$X = 0$	y_{00}	y_{01}	n_0
Exposed	$X = 1$	y_{10}	y_{11}	n_1
		m_0	m_1	n

the number of individuals in exposure group i, $i = 0, 1$ and disease group j, $j = 0, 1$. In a cohort study, n_0 and n_1, the numbers of unexposed and exposed individuals, are fixed by design, and the random variables are the number of unexposed cases y_{01} and the number of exposed cases y_{11}.

There are a number of strong motivations for carrying out a case-control study. Since many diseases are rare, a cohort study has to generally contain a large number of participants to demonstrate an association between a risk factor and disease because few individuals will develop the disease (unless the effect of the exposure of interest is very strong). It may be difficult to assemble a full picture of the disease across subgroups (as defined by covariates) within a cohort study because the cohort is assembled at a particular time, the start of the study. As the study proceeds, certain subgroups, for example, the young, disappear. In this case it will not be possible to investigate a calendar time/age interaction, that is, the effect of calendar time at different age groups. Finally, the disease may take a long time to develop (this is true, for example, for most cancers), and so the study may need to run for a long period.

The case-control study provides a way of overcoming these difficulties. With reference to Table 7.7, m_0 and m_1, the numbers of controls and case, are fixed by design, and the random variables are the number of exposed controls y_{10} and the number of exposed cases, y_{11}.

A case-control study is not without its drawbacks. Probabilities of disease given exposure status are no longer directly estimable without external information, as we will discuss in more detail shortly. Most importantly, the study participants must be selected very carefully. The probability of selection for the study, for both cases and controls, must not depend on exposure status; otherwise, *selection bias* will be introduced; this bias can arise in many subtle ways. The great benefit of case-control studies is that we can still estimate the strength of the relationship between exposure and disease, a topic we discuss in-depth in the next section.

7.10.2 Estimation for a Case-Control Study

Consider the situation in which we have a binary response Y taking the values 0/1 corresponding to disease-free/diseased and exposures contained in a $(k + 1) \times 1$ vector $\boldsymbol{x}$. The exposures can be a mix of continuous and discrete variables. In the case-control scenario, we select individuals on the basis of their disease status y, and the random variables are the exposures $\boldsymbol{X}$.

In a cohort study with a binary endpoint, a logistic regression disease model is the most common choice for analysis, with form

$$\Pr(Y = 1 \mid \boldsymbol{x}) = p(\boldsymbol{x}) = \frac{\exp\left(\beta_0 + \sum_{j=1}^{k} x_j \beta_j\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^{k} x_j \beta_j\right)}. \tag{7.31}$$

The *relative risk* of individuals having exposures $\boldsymbol{x}$ and $\boldsymbol{x}^\star$ is defined as

$$\text{Relative risk } = \frac{\Pr(Y = 1 \mid \boldsymbol{x})}{\Pr(Y = 1 \mid \boldsymbol{x}^\star)}$$

and is an easily interpretable quantity that epidemiologists are familiar with. As already mentioned in Sect. 7.6.1, for rare diseases, the relative risk is well approximated by the odds ratio

$$\frac{\Pr(Y = 1 \mid \boldsymbol{x}) / \Pr(Y = 0 \mid \boldsymbol{x})}{\Pr(Y = 1 \mid \boldsymbol{x}^\star) / \Pr(Y = 0 \mid \boldsymbol{x}^\star)}.$$

With respect to the logistic regression model (7.31),

$$\frac{p(\boldsymbol{x}) / [1 - p(\boldsymbol{x})]}{p(\boldsymbol{x}^\star) / [1 - p(\boldsymbol{x}^\star)]} = \exp\left[\sum_{j=1}^{k} \beta_j (x_j - x_j^\star)\right],$$

so that, in particular, $\exp(\beta_j)$ represents the increase in the odds of disease associated with a unit increase in x_j, with all other covariates held fixed (Sect. 7.6.1). The parameter β_0 represents the baseline log odds of disease, corresponding to the odds when all of the exposures are set equal to zero.

We now turn to interpretation in a case-control study. We first introduce an indicator variable Z which represents the event that an individual was selected for the study ($Z = 1$) or not ($Z = 0$). Let $\pi_y = \Pr(Z = 1 \mid Y = y)$ denote the probabilities of selection, given response y, $y = 0, 1$. Typically, π_1 is much greater than π_0, since cases are rarer than non-cases. Now consider the probability that a person is diseased, given exposures $\boldsymbol{x}$ and selection for the study:

$$\Pr(Y = 1 \mid Z = 1, \boldsymbol{x}) = \frac{\Pr(Z = 1 \mid Y = 1, \boldsymbol{x}) \Pr(Y = 1 \mid \boldsymbol{x})}{\Pr(Z = 1 \mid \boldsymbol{x})}. \tag{7.32}$$

The denominator may be simplified to

$$\begin{aligned} \Pr(Z = 1 \mid \boldsymbol{x}) &= \sum_{y=0}^{1} \Pr(Z = 1 \mid Y = y, \boldsymbol{x}) \Pr(Y = y \mid \boldsymbol{x}) \\ &= \sum_{y=0}^{1} \Pr(Z = 1 \mid Y = y) \Pr(Y = y \mid \boldsymbol{x}), \end{aligned}$$

where we have made the crucial assumption that

$$\Pr(Z = 1 \mid Y = y, \boldsymbol{x}) = \Pr(Z = 1 \mid Y = y) = \pi_y,$$

for $y = 0, 1$, that is, the selection probabilities depend only on the disease status and *not* on the exposures (i.e., there is no selection bias). If we take a random sample of cases and controls, this assumption is valid. Substitution in (7.32), and assuming a logistic regression model, gives

$$\begin{aligned}
\Pr(Y = 1 \mid Z = 1, \boldsymbol{x}) &= \frac{\pi_1 \exp(\boldsymbol{x}\boldsymbol{\beta})/[1 + \exp(\boldsymbol{x}\boldsymbol{\beta})]}{\pi_1 \exp(\boldsymbol{x}\boldsymbol{\beta})/[1 + \exp(\boldsymbol{x}\boldsymbol{\beta})] + \pi_0/[1 + \exp(\boldsymbol{x}\boldsymbol{\beta})]} \\
&= \frac{\pi_1 \exp\left(\beta_0 + \sum_{j=1}^{k} x_j\beta_j\right)}{\pi_0 + \pi_1 \exp\left(\beta_0 + \sum_{j=1}^{k} x_j\beta_j\right)} \\
&= \frac{\exp\left(\beta_0^\star + \sum_{j=1}^{k} x_j\beta_j\right)}{1 + \exp\left(\beta_0^\star + \sum_{j=1}^{k} x_j\beta_j\right)},
\end{aligned}$$

where $\beta_0^\star = \beta_0 + \log \pi_1/\pi_0$. Hence, we see that the probabilities of disease in a case-control study also follow a logistic model but with an altered intercept. In the usual case, $\pi_1 > \pi_0$ so that the intercept is increased to account for the over-sampling of cases. Unless information on π_0 and π_1 is available, we cannot obtain estimates of $\Pr(Y = 1 \mid \boldsymbol{x})$ (the incidence for different exposure groups).

This derivation shows that assuming a logistic model in the cohort context implies that the disease frequency within the case-control sample also follows a logistic model, but does not illuminate how inference may be carried out. Suppose there are m_0 controls and m_1 cases. Since the exposures are random in a case-control context, the likelihood is of the form

$$L(\boldsymbol{\theta}) = \prod_{y=0}^{1} \prod_{j=1}^{m_y} p(\boldsymbol{x}_{yj} \mid y, \boldsymbol{\theta}),$$

where $\boldsymbol{x}_{yj}$ is the set of covariates for individual j in disease group y, and it appears that we are faced with the unenviable task of specifying forms, depending on parameters $\boldsymbol{\theta}$, for the distribution of covariates in the control and case populations. In a seminal paper, Prentice and Pyke (1979) showed that asymptotic likelihood inference for the odds ratio parameters was identical irrespective of whether the data are collected prospectively or retrospectively. The proof of this result hinges on assuming a logistic disease model, depending on parameters $\boldsymbol{\beta}$, with additional nuisance parameters being estimated via nonparametric maximum likelihood. Great care is required in this context because unless the sample space for $\boldsymbol{x}$ is finite (i.e., the covariates are all discrete with a fixed number of categories), the dimension of the nuisance parameter increases with the sample size.

To summarize, when data are collected from a case-control study, a likelihood-based analysis with a logistic regression model may proceed with asymptotic inference, acting as if the data were collected in a cohort fashion, except that the intercept is no longer interpretable as the baseline log odds of disease.

7.10.3 Estimation for a Matched Case-Control Study

A common approach in epidemiological studies is to "match" the controls to the cases on the basis of known confounders. By choosing controls to be similar to cases, one "controls" for the confounding variables. This provides efficiency gains since the controls are more similar to the cases with respect to confounders, which increases power. It also removes the need to model the disease-confounder relationship.

In a *frequency-matched* design, the cases are grouped into broad strata (e.g., 10-year age bands), and controls are matched on the basis of these variables. In an *individually matched* study, controls are matched exactly, usually upon multiple variables, for example, age, gender, time of diagnosis, and area of residence. For both forms of matching, the nonrandom selection of controls must be acknowledged in the analysis by including a parameter for each matching set in the logistic model.

For matched data, let $j = 1, \ldots, J$ index the matched sets, and Y_{ij} and $\boldsymbol{x}_{ij}$ denote the responses and covariate vector of additional variables (i.e., beyond the matching variables) for individual i, with $i = 1, \ldots, m_{1j}$ representing the cases and $i = m_{1j} + 1, \ldots, m_{1j} + m_{0j}$ the controls. Hence, for $j = 1, \ldots, J$,

$$\begin{aligned} y_{ij} &= 1 \quad \text{for} \quad i = 1, \ldots, m_{1j} \\ y_{ij} &= 0 \quad \text{for} \quad i = m_{1j} + 1, \ldots, m_{1j} + m_{0j}, \end{aligned}$$

and there are $m_1 = \sum_{j=1}^{J} m_{1j}$ cases and $m_0 = \sum_{j=1}^{J} m_{0j}$ controls in total.

The disease model is

$$\log \left[\frac{p_j(\boldsymbol{x}_{ij})}{1 - p_j(\boldsymbol{x}_{ij})} \right] = \alpha_j + \boldsymbol{x}_{ij}\boldsymbol{\beta} \tag{7.33}$$

where

$$p_j(\boldsymbol{x}_{ij}) = \Pr(Y_{ij} = 1 \mid \boldsymbol{x}_{ij}, \text{ stratum } j)$$

for $i = 1, \ldots, m_{0j} + m_{1j}$, $j = 1, \ldots, J$. In terms of inference, the key distinction between the two matching situations is that in the frequency matching situation, the number of matching strata J is fixed. In this case, the result outlined in Sect. 7.10.2 can be extended so that the matched data can be analyzed as if they were gathered prospectively, though the intercept parameters α_j are no longer interpretable as log odds ratios describing the association between disease and the variables defining stratum j. For the same reason, it is not possible to estimate interactions between stratum variables and exposures of interest. Calculations in Breslow and Day (1980)

show that, in terms of efficiency gains, it is usually not worth exceeding 5 controls per case and 3 will often be sufficient. Exercise 7.8 considers the analysis of a particular set of data to illustrate the benefits of case-control sampling and matching.

For individually matched data, for simplicity, suppose there are M controls for each case so that $m_{1j} = 1$ and $m_{0j} = M$ for all j. Hence, $m_1 = J$ and $m_0 = MJ = Mm_1$. Also let $n = m_1$ represent the number of cases so that $m_0 = Mn$ is the number of controls. The likelihood contribution of the jth stratum is

$$p(\boldsymbol{x}_{1j} \mid Y_{1j} = 1) \prod_{i=2}^{M+1} p(\boldsymbol{x}_{ij} \mid Y_{ij} = 0), \tag{7.34}$$

but care is required for inference because the number of nuisance parameters, $\alpha_1, \ldots, \alpha_n$, is equal to the number of cases/matching sets, n, and so increases with sample size.

To overcome this violation of the usual regularity conditions, a conditional likelihood may be constructed. Specifically, for each j, one conditions on the collection of $M + 1$ covariate vectors within each matching set. The conditional contribution is the probability that subject $i = 1$ is the case, given it could have been any of the $M + 1$ subjects within that matching set. The numerator is (7.34), and the denominator is this expression but evaluated under the possibility that each of the $i = 1, \ldots, M + 1$ individuals could have been the case. Hence, the jth contribution to the conditional likelihood is

$$\frac{p(\boldsymbol{x}_{1j} \mid Y_{1j} = 1) \prod_{i=2}^{M+1} p(\boldsymbol{x}_{ij} \mid Y_{ij} = 0)}{\sum_{R_j} p(\boldsymbol{x}_{\pi(1),j} \mid Y_{1j} = 1) \prod_{i=2}^{M+1} p(\boldsymbol{x}_{\pi(i),j} \mid Y_{ij} = 0)}$$

where R_j is the set of $M + 1$ permutations, $[\boldsymbol{x}_{\pi(1),j}, \ldots, \boldsymbol{x}_{\pi(M+1),j}]$ of $[\boldsymbol{x}_{1j}, \ldots, \boldsymbol{x}_{M+1,j}]$. Applying Bayes theorem to each term,

$$p(\boldsymbol{x}_{ij} \mid Y = y) = \frac{p(Y = y \mid \boldsymbol{x}_{ij}) p(\boldsymbol{x}_{ij})}{p(Y = y)},$$

and taking the product across matching sets, we obtain

$$L_c(\boldsymbol{\beta}) = \prod_{j=1}^{n} \frac{p(Y_{1j} = 1 \mid \boldsymbol{x}_{1j}) \prod_{i=2}^{M+1} p(Y_{ij} = 0 \mid \boldsymbol{x}_{ij})}{\sum_{R_j} p(Y_{1j} = 1 \mid \boldsymbol{x}_{\pi(1),j}) \prod_{i=2}^{M+1} p(Y_{ij} = 0 \mid \boldsymbol{x}_{\pi(i),j})}.$$

Substitution of the logistic disease model (7.33) yields the conditional likelihood

$$\begin{aligned} L_c(\boldsymbol{\beta}) &= \prod_{j=1}^{n} \frac{\exp(\boldsymbol{x}_{1j}\boldsymbol{\beta})}{\sum_{i=1}^{M+1} \exp(\boldsymbol{x}_{ij}\boldsymbol{\beta})} \\ &= \prod_{j=1}^{n} \left(1 + \sum_{i=2}^{M+1} \exp\left[(\boldsymbol{x}_{ij} - \boldsymbol{x}_{1j})\boldsymbol{\beta} \right] \right)^{-1} \end{aligned}$$

Table 7.8 Notation for a matched-pair case-control study with n controls and n cases and a single exposure

		Not diseased $Y = 0$	Diseased $Y = 1$
Unexposed	$X = 0$	m_{00}	m_{01}
Exposed	$X = 1$	m_{10}	m_{11}
		n	n

with the α_j terms having canceled out, as was required. For further details, see Cox and Snell (1989) and Prentice and Pyke (1979, Sect. 6). As an example, if $M = 2$ (two controls per case), the conditional likelihood is

$$L_c(\boldsymbol{\beta}) = \prod_{j=1}^{n} \frac{\exp(\boldsymbol{x}_{1j}\boldsymbol{\beta})}{\exp(\boldsymbol{x}_{1j}\boldsymbol{\beta}) + \exp(\boldsymbol{x}_{2j}\boldsymbol{\beta}) + \exp(\boldsymbol{x}_{3j}\boldsymbol{\beta})}$$
$$= \prod_{j=1}^{n} \left(1 + \sum_{i=2}^{3} \exp\left[(\boldsymbol{x}_{ij} - \boldsymbol{x}_{1j})\boldsymbol{\beta}\right]\right)^{-1}.$$

The importance of the use of conditional likelihood can be clearly demonstrated in the matched-pairs situation, in which there is one control per case. Suppose that the data are as summarized in Table 7.8 so that there is a single exposure only. There are m_{00} concordant pairs in which neither case nor control is exposed and m_{11} concordant pairs in which both are exposed. Exercise 7.12 shows that the unconditional MLE of the odds ratio is $(m_{10}/m_{01})^2$, the square of the ratio of discordant pairs. In contrast, the estimate based on the appropriate conditional likelihood is m_{10}/m_{01}. Hence, the unconditional estimator is the square of the correct conditional estimator.

A further caveat to the use of individually matched case-control data is that it is more difficult to generalize inference to a specific population under this design because the manner of selection is far from that of a random sample.

7.11 Concluding Remarks

The analysis of binomial data is difficult unless the denominators are large because there is so little information in a single Bernoulli outcome. In addition, the models for probabilities are typically nonlinear. Logistic regression models are the obvious candidate for analysis, but the interpretation of odds ratios is not straightforward, unless the outcome of interest is rare. The effect of omitting variables is also nonobvious. The fact that the linear logistic model is a GLM does offer advantages in terms of consistency, however, and the logit being the canonical link gives simplifications in terms of computation.

The use of conditional likelihood in individually matched case-control studies in practice is uncontroversial, but its theoretical underpinning is not completely convincing (since the conditioning statistic is not ancillary). Fisher's exact test is historically popular, but, as discussed in Sect. 4.2, frequentist hypothesis testing can be difficult to implement in practice since p-values need to be interpreted in the context of the sample size. For Fisher's exact, the discreteness of the test statistic can also be problematic. Exercise 7.11 provides an alternative approach based on Bayes factors. The latter do not suffer from the discreteness of the sampling distribution (since one only uses the observed data and not other hypothetical realizations).

7.12 Bibliographic Notes

Robinson and Jewell (1991) examine the effects of omission of variables in logistic regression models and contrast the implications with the linear model case. Greenland et al. (1999) is a wide-ranging discussion on collapsibility and confounding. A seminal book on the design and analysis of case-control studies is Breslow and Day (1980). There is no Bayesian analog of the Prentice and Pyke (1979) result showing the equivalence of odds ratio estimation for prospective and retrospective sampling, though Seaman and Richardson (2004) show the equivalence in restricted circumstances. Simplified estimation based on nonparametric maximum likelihood has also been established for other outcome-dependent sampling schemes such as two-phase sampling; see, for example, White (1982) and Breslow and Chatterjee (1999). Again, no equivalent Bayesian approaches are available. A fully Bayesian approach in a case-control setting would require the modeling of the covariate distributions for each of the cases and controls, which is, in general, a difficult process and seems unnecessary given that there is no direct interest in these distributions. Hence, the nonparametric maximum likelihood procedure seems preferable, though a hybrid approach in which one simply combines the prospective likelihood with a prior would seem practically reasonable if one has prior information and/or one is worried about asymptotic inference.

Rice (2008) shows the equivalence between conditional likelihood and random effects approaches to the analysis of matched-pairs case-control data. In general, conditional likelihood does not have a Bayesian interpretation, though Bayesian analyses have been carried out in the individually matched case-control situation by combining a prior with the conditional likelihood. This approach avoids the difficulty of specifying priors over nuisance parameters with dimension equal to the number of matching sets (Diggle et al. 2000).

Fisher's exact test has been discussed extensively in the statistics literature; see, for example, Yates (1984). Altham (1969) published an intriguing result showing that Fisher's exact test is equivalent to a Bayesian analysis. Specifically, let $p_{00}, p_{10}, p_{01}, p_{11}$ denote the underlying probabilities in a 2×2 table with entries $\boldsymbol{y} = [y_{00}, y_{10}, y_{01}, y_{11}]^{\mathrm{T}}$ (see Table 7.3), and suppose the prior on these probabilities is (improper) Dirichlet with parameters (0,1,1,0). Then the posterior probability

$\Pr(p_{11}p_{22}/p_{12}p_{21} < 1 \mid \boldsymbol{y})$ equals the Fisher's exact test p-value for testing $H_0 : p_{00}p_{11} = p_{10}p_{01}$ versus $H_1 : p_{00}p_{11} < p_{10}p_{01}$. Hence, the prior (slightly) favors a negative association between rows and columns, which is related to the fact that conditioning on the margins (as is done in Fisher's exact test) does lead to a small loss of information.

7.13 Exercises

7.1 Suppose $Z \mid p \sim \text{Bernoulli}(p)$.

(a) Show that the moment-generating function (Appendix D) of Z is $M_Z(t) = 1 - p + p\exp(t)$. Hence, show that the moment-generating function of $Y = \sum_{i=1}^{n} Z_i$ is

$$M_Y = [1 - p + p\exp(t)]^n,$$

which is the moment-generating function of a binomial random variable.

(b) Suppose $Y \mid \lambda \sim \text{Poisson}(\lambda)$. Show that the cumulant-generating function (Appendix D) of Y is

$$\lambda[\exp(t) - 1].$$

(c) From part (a), obtain the form of the cumulant-generating function of Y. Suppose that $p \to 0$ and $n \to \infty$ in such a way that $\mu = np$ remains fixed. By considering the limiting form of the cumulant-generating function of Y, show that in this situation, the limiting distribution of Y is Poisson with mean μ.

7.2 Before the advent of GLMs, the arc sine variance stabilizing transformation was used for the analysis of binomial data. Suppose that $Y \mid p \sim \text{Binomial}(N, p)$ with N large. Using a Taylor series expansion, show that the random variable

$$W = \arcsin\left[\sqrt{(Y/N)}\right]$$

has approximate first two moments:

$$\text{E}[W] \approx \arcsin(\sqrt{p}) - \frac{1 - 2p}{8\sqrt{Np(1-p)}}$$

$$\text{var}(W) \approx \frac{1}{4N}.$$

7.3 Suppose $Z_j \mid \lambda_j \sim_{ind} \text{Poisson}(\lambda_j)$, $j = 1, 2$ are independent Poisson random variables with rates λ_j. Show that

$$Z_1 \mid Z_1 + Z_2, p \sim \text{Binomial}(Z_1 + Z_2, p),$$

with $p = \lambda_1/(\lambda_1 + \lambda_2)$.

7.4 Consider n Bernoulli trials with Z_{ij}, $j = 1, \ldots, N_i$ the outcomes within-trial i with $Y_i = \sum_{j=1}^{N_i} Z_{ij}$, $i = 1, \ldots, n$. By writing

$$\text{var}(Y_i) = \sum_{j=1}^{N_i} \text{var}(Z_{ij}) + \sum_{j=1}^{N_i} \sum_{j \neq k} \text{cov}(Z_{ij}, Z_{ik}),$$

show that

$$\text{var}(Y_i) = N_i p_i (1 - p_i) \times [1 + (N_i - 1)\tau_i^2].$$

7.5 With respect to Sect. 7.6.2, show that for Bernoulli data the Pearson statistic is $X^2 = n$. Find the deviance in this situation and comments on its usefulness as a test of goodness of fit.

7.6 Show that the extended hypergeometric distribution (7.22) is a member of the exponential family (Sect. 6.3), that is, the distribution can be written in the form

$$\Pr(y_{11} \mid \theta, \alpha) = \exp\left(\frac{y_{11}\theta - b(\theta)}{\alpha} + c(y, \alpha)\right)$$

for suitable choices of α, $b(\cdot)$, and $c(\cdot, \cdot)$.

7.7 In this question, a simulation study to investigate the impact on inference of omitting covariates in logistic regression will be performed, in the situation in which the covariates are independent of the exposure of interest. Let x be the covariate of interest and z another covariate. Suppose the true (adjusted) model is $Y_i \mid x_i, z_i \sim_{iid} \text{Bernoulli}(p_i)$, with

$$\log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_i + \beta_2 z_i. \tag{7.35}$$

A comparison with the unadjusted model $Y_i \mid x_i \sim_{iid} \text{Bernoulli}(p_i^\star)$, where

$$\log\left(\frac{p_i^\star}{1 - p_i^\star}\right) = \beta_0^\star + \beta_1^\star x_i, \tag{7.36}$$

for $i = 1, \ldots, n = 1{,}000$ will be made. Suppose x is binary with $\Pr(X{=}1) = 0.5$ and $Z \sim_{iid} \text{N}(0, 1)$ with x and z independent. Combinations of the parameters $\beta_1 = 0.5, 1.0$ and $\beta_2 = 0.5, 1.0, 2.0, 3.0$, with $\beta_0 = -2$ in all cases, will be considered.

For each combination of parameters, compare the results from the two models, (7.35) and (7.36), with respect to:

(a) $\text{E}[\widehat{\beta}_1]$ and $\text{E}[\widehat{\beta}_1^\star]$, as compared to β_1
(b) The standard errors of $\widehat{\beta_1}$ and $\widehat{\beta}_1^\star$
(c) The coverage of 95% confidence intervals for β_1 and $\beta_1^\star$
(d) The probability of rejecting $H_0 : \beta_1 = 0$ in model (7.35) and the probability of rejecting $H_0 : \beta_1^* = 0$ in model (7.36). These probabilities correspond to the powers of the tests. Calculate these probabilities using Wald tests.

Table 7.9 *Left table*: leprosy cases and non-cases versus presence/absence of BCG scar. *Right table*: leprosy cases and controls versus presence/absence of BCG scar

BCG scar	Cases	Non-cases
Present	101	46,028
Absent	159	34,594

BCG scar	Cases	Controls
Present	101	554
Absent	159	446

Based on the results, summarize the effect of omitting a covariate that is independent of the exposure of interest, in particular in comparison with the linear model case (as discussed in Sect. 5.9).

7.8 This question illustrates the benefits of case-control and matched case-control sampling, taking data from Fine et al. (1986) and following loosely the presentation of Clayton and Hills (1993). Table 7.9 gives data from a cross-sectional survey carried out in Northern Malawi. The aim of this study was to investigate whether receiving a bacillus Calmette-Guérin (BCG) vaccination in early childhood (which protects against tuberculosis) gives any protection against leprosy. Let $X = 0/1$ denote absence/presence of BCG scar, $Y = 0/1$ denote leprosy-free/leprosy, and $p_x = \Pr(Y = 1 \mid X = x)$, $x = 0, 1$:

(a) Fit the logistic model

$$\log\left(\frac{p_x}{1-p_x}\right) = \beta_0 + \beta_1 x$$

to the case/non-case data in the left half of Table 7.9. Report your findings in terms of an estimate of the odds ratio $\exp(\beta_1)$ along with an associated standard error.

(b) Now consider the case/control data in the right half of Table 7.9 (these data were simulated from the full dataset). Fit the logistic model

$$\log\left(\frac{p_x}{1-p_x}\right) = \beta_0^\star + \beta_1 x$$

to the case/control data, and again report your findings in terms of the odds ratio $\exp(\beta_1)$ along with an associated standard error. Hence, use this example to describe the benefits, in terms of efficiency, of a case-control study.

(c) In this example, the population data are known and consequently the sampling fractions of cases and controls are also known. Hence, reconstruct an estimate of β_0, using the results from the case-control analysis.

(d) Next the benefits of matching will be illustrated. BCG vaccination was gradually introduced into the study region, and so older people are less likely to have been vaccinated but also more likely to have developed leprosy. Therefore, age is a potential confounder in this study.

Let $z = 0, 1, \ldots, 6$ denote age represented as a factor and $p_{xz} = \Pr(Y = 1 \mid X = x,)$, for $x = 0, 1$, denote the probability of leprosy

Table 7.10 *Left table*: leprosy cases and non-cases as a function of presence/absence of BCG scar and age. *Right table*: leprosy cases and matched controls as a function of presence/absence of BCG scar and age

	BCG scar					BCG scar			
	Cases		Non-cases			Cases		Controls	
Age	Absent	Present	Absent	Present	Age	Absent	Present	Absent	Present
0–4	1	1	7,593	11,719	0–4	1	1	3	5
5–9	11	14	7,143	10,184	5–9	11	14	48	52
10–14	28	22	5,611	7,561	10–14	28	22	67	133
15–19	16	28	2,208	8,117	15–19	16	28	46	130
20–24	20	19	2,438	5,588	20–24	20	19	50	106
25–29	36	11	4,356	1,625	25–29	36	11	126	62
30–34	47	6	5,245	1,234	30–34	47	6	174	38

for an individual with BCG status x and in age strata z. To adjust for age, fit the logistic model

$$\log\left(\frac{p_{xz}}{1-p_{xz}}\right) = \beta_0 + \beta_1 x + \beta_z z$$

to the data in the left half of Table 7.10. This model assumes a common odds ratio across age strata. Report your findings in terms of the odds ratio $\exp(\beta_1)$ and associated standard error.

(e) If it were possible to sample controls from the non-cases in the left half of Table 7.10, the age distribution would be highly skewed toward the young, which would lead to an inefficient analysis. As an alternative, the right half of Table 7.10 gives a simulated frequency-matched case-control study with 4 controls per case within each age strata. Analyze these data using the logistic model

$$\log\left(\frac{p_{xz}}{1-p_{xz}}\right) = \beta_0^\star + \beta_1 x + \beta_z^\star z,$$

and report your findings in terms of $\exp(\beta_1)$ and its associated standard error. Comment on the accuracy of inference as compared to the analysis using the complete data.

7.9 Table 7.11 gives data from a toxicological experiment in which the number of beetles that died after 5 h exposure to gaseous carbon disulphide at various doses.

(a) Fit complementary log–log, probit, and logit link models to these data using likelihood methods.
(b) Summarize the association for each model in simple terms.
(c) Examine residuals and report the model that you believe provides the best fit to these data, along with your reasoning.

Table 7.11 Number of beetle deaths as a function of log dose, from Bliss (1935)

Log dose	No. beetles	No. killed
1.691	59	6
1.724	60	13
1.755	62	18
1.784	56	28
1.811	63	52
1.837	59	53
1.861	62	61
1.884	60	60

Table 7.12 Death penalty verdict by race of victim and defendant

Defendant's race	Victim's race	Death penalty Yes	No
White	White	19	132
	Black	0	9
Black	White	11	52
	Black	06	97

(d) Fit your favored model with a Bayesian approach using (improper) flat priors. Is there a substantive difference in the conclusions, as compared to the likelihood analysis?

7.10 Table 7.12 contains data from Radelet (1981) on death penalty verdict, cross-classified by defendant's race and victim's race.

(a) Fit a logistic regression model that includes factors for both defendant's race and victim's race. Estimate the odds ratios associated with receiving the death penalty if Black as compared to if White, for the situations in which the victim was White and in which the victim was Black.
(b) Fit a logistic regression model to the marginal 2×2 table that collapses across victim's race, and hence, estimate the odds ratio associated with receiving the death penalty if Black versus if White.
(c) Discuss the results of the two parts, in relation to Simpson's paradox. In particular, discuss the paradox in terms understandable to a layperson.

7.11 Suppose $Y_i \mid p_i \sim \text{Binomial}(N_i, p_i)$ for $i = 0, 1$ and that interest focuses on $H_0 : p_0 = p_1 = p$ versus $H_1 : p_0 \neq p_1$:

(a) Consider the Bayes factor (Sect. 3.10)

$$\text{BF} = \frac{\Pr(y_0, y_1 \mid H_0)}{\Pr(y_0, y_1 \mid H_1)}$$

with the priors: $p \sim \text{Be}(a_0, b_0)$ under H_0 and $p_i \sim \text{Be}(a_1, b_1)$, for $i = 0, 1$, under H_1. Obtain a closed-form expression for the Bayes factor.
(b) Calculate the Bayes factor for the tumor data given in Table 7.4 using uniform priors, that is, $a_0 = a_1 = b_0 = b_1 = 1$.
(c) Based on the Bayes factor, would you reject H_0? Why?

(d) Using the same priors as in the previous part, evaluate the posterior probability that $\Pr(p_0 > p_1 \mid y_0, y_1)$. Based on this probability, what would you conclude about equality of p_0 and p_1? Is your conclusion in agreement with the previous part?
[Hint: Obtaining samples from the posteriors $p(p_i \mid y_i)$, for $i = 0, 1$, is a simple way of obtaining the posterior of interest in the final part.]

7.12 This question derives unconditional and conditional estimators for the case of a matched-pairs case-control design with n pairs and a binary exposure. The notation is given in Table 7.8, and the logistic model in the jth matching set is

$$\Pr(Y = 1 \mid x, j) = \frac{\exp(\alpha_j + x\beta)}{1 + \exp(\alpha_j + x\beta)},$$

for $x = 0, 1$ and $j = 1, \ldots, n$.

(a) Show that the unconditional maximum likelihood estimator of β is the square of the ratio of the discordant pairs, $(m_{10}/m_{01})^2$.
(b) Show, by considering the distribution of m_{10} given the total $m_{10} + m_{01}$, that the estimate based on the appropriate conditional likelihood is m_{10}/m_{01}.