2021 Advanced Regression Methods for Independent Data

BIOSTAT/STAT 570

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington
jonno@uw.edu

CHAPTER 7: BINARY DATA MODELS

Outline

Introduction and Motivation

Overview

In this chapter we consider the modeling of binary data.

Such data are ubiquitous in many fields, and perhaps counter-intuitively[1] many aspects of their modeling are tricky to think about.

Binary data present a number of distinct challenges and so we devote a separate chapter to their modeling, though leaning heavily on the methods introduced in Chapter 6 on general regression modeling.

[1] Because a binary random variable can only be 0 or 1, how hard can that be?!

Overview

A major problem is the lack of information contained within a variable that can only take one of two values.

This can lead to a number of problems, for example:

- In assessing model fit.
- Numerical problems with rare events.
- Because models for probabilities are generally nonlinear which can lead to curious behavior of estimators in the presence of confounders.
- Difficulties in interpretation even arise when independent regressors are added to the model.

Example: Outcome After Head Injury

We will illustrate methods for binary data using the data first encountered in the introductory chapter.

The binary response is outcome after head injury (dead/alive), with four discrete covariates:

- pupils (good/poor);
- coma score (depth of coma, low/high);
- haematoma present (no/yes); and
- age (categorized as 1–25, 26–54, $\geqslant$ 55).

Example: Outcome After Head Injury

Figure 1 displays conditional frequencies.

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 haematoma variable is less clear.

The sample sizes are lost in these plots, which makes interpretation more difficult.

We could look at these associations on a logit scale, since this is a common link, and then consider associations by analogy with the linear model (e.g., linearity, interactions,...).

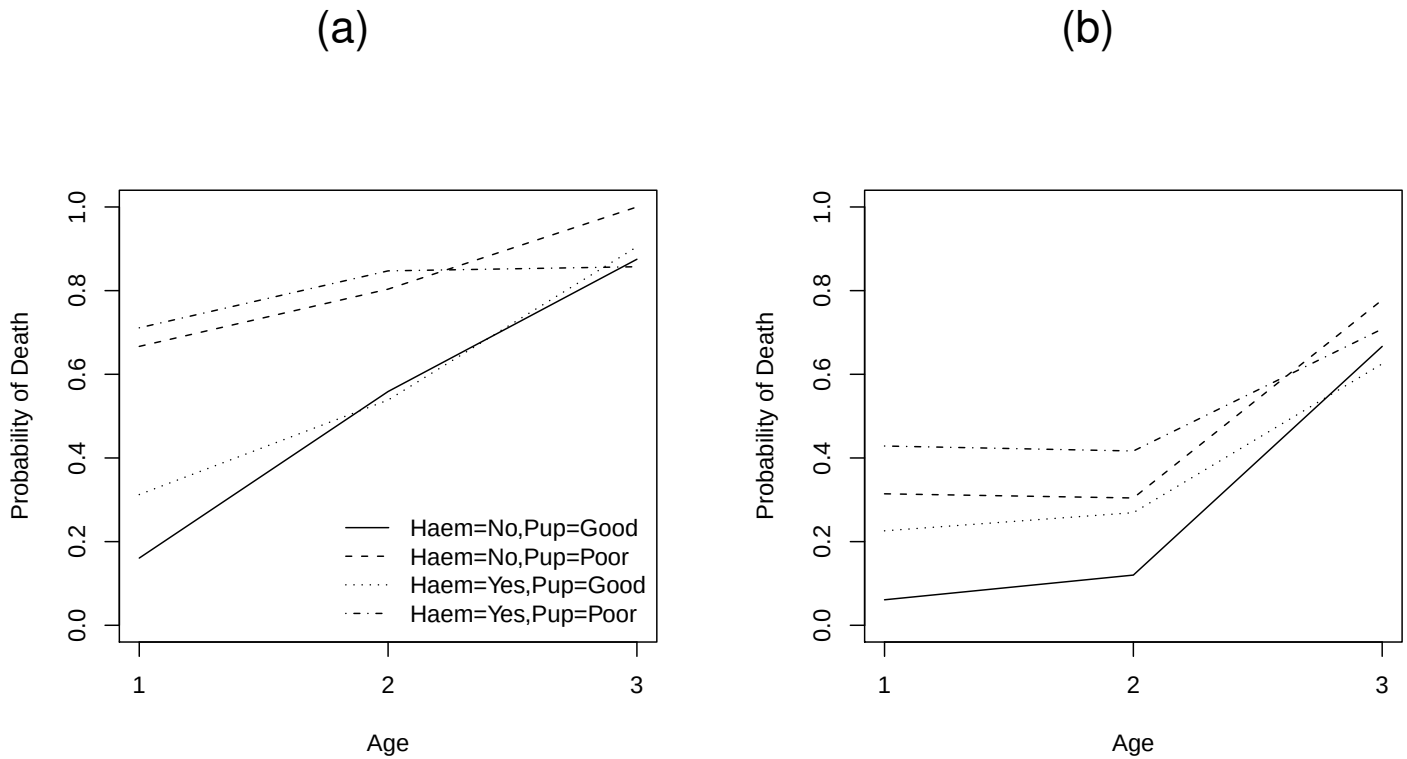

FIGURE 1: Probability of death after head injury as a function of age, haematoma score and pupils: panels (a) and (b) are for low and high coma scores, respectively.

Example: Aircraft Fasteners

Table 1 gives the total number of fasteners tested and the number failure at a range of pressure loads.

Load (psi)	Failures	Sample Size	Proportion Failing
2500	10	50	0.20
2700	17	70	0.24
2900	30	100	0.30
3100	21	60	0.35
3300	18	40	0.45
3500	43	85	0.51
3700	54	90	0.60
3900	33	50	0.66
4100	60	80	0.75
4300	51	65	0.78

Table 1: Number of aircraft fastener failures at specified pressure loads.

Example: Bronchopulmonary Dysplasia

These data concern the absence/presence of bronchopulmonary dysplasia (BPD) as a function of birth weight (in grams) for $n = 223$ babies.

Figure 2 displays the BPD indicator, plotted as short vertical lines at 0 and 1, as a function of birthweight.

Visual assessment suggests that children with lower birthweight tend to have an increased chance of BPD.

This example is distinct from the aircraft fasteners because the latter contained multiple responses at each x value.

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 – we could bin into groups of birthweight, and examine how the proportions in each bin change with group.

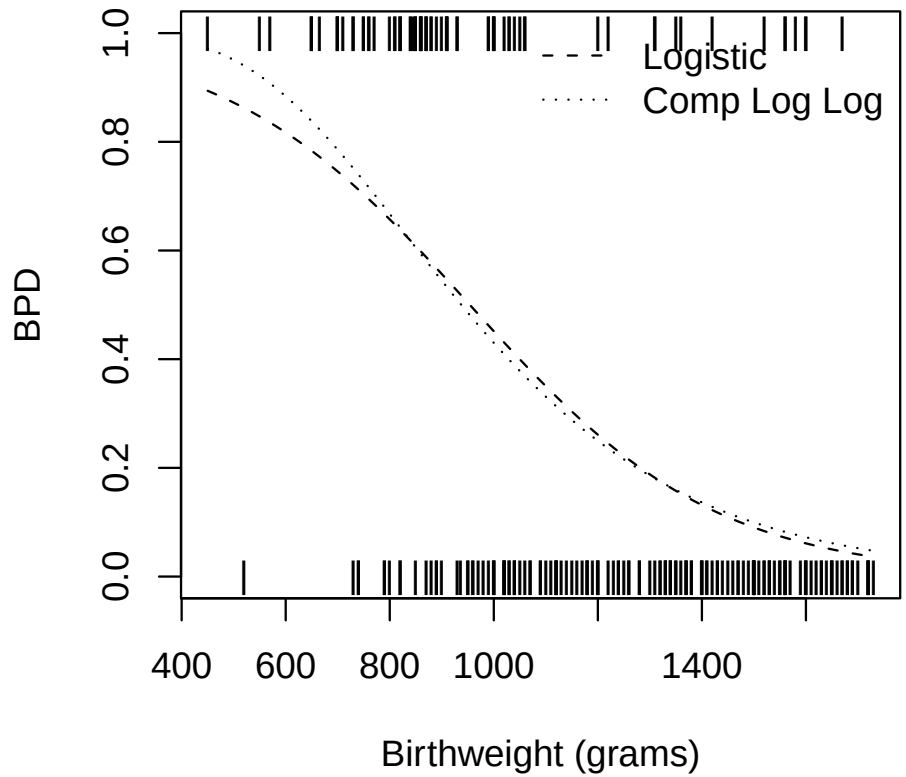

FIGURE 2: BPD, as a function of birthweight. The short vertical lines at 0 and 1 indicate the observed birthweights for non-BPD and BPD infants, respectively. The dashed curve corresponds to a logistic regression fit, and the dotted curve to a complementary log-log regression fit.

The Binomial Distribution

The Binomial Distribution

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 binary variable has to follow a Bernoulli distribution, there are no other options, but when we consider multiple binary variables, things get more interesting.

The Binomial Distribution

For a Bernoulli random variable, and all moments are 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.

The probability distribution of Z gives the frequencies under hypothetical repeated sampling.

Consider the scenario in which the repeated sampling is:

- Draw $p \sim_{iid} f(\overline{p})$, a distribution with mean $\overline{p}$.
- Draw $Z|p \sim_{iid} \text{Bernoulli}(p)$.

The (marginal) distribution of Z is $\text{Bernoulli}(\overline{p})$.

They key is independence – no clustering (which leads to overdispersion) or restrictions (for example, via a hypergeometric, to give underdispersion).

The Binomial Distribution

Suppose there are N individuals, and let Z_j denote the outcome for the j-th 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 have equal probabilities, i.e., $p = p_1 = \cdots = p_N$.

Under the assumption that the Bernoulli random variables are independent,

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

so that

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

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

The Binomial Distribution

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,

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

and

$$Y \mid \overline{p} \sim \text{Binomial}(N, \overline{p}). \tag{2}$$

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 $j = 1, \ldots, N$.

If we were to draw a single p from the distribution with mean $\overline{p}$ and draw Bernoulli random variables with the same p, then an overdispersed distribution will result.

The Binomial Distribution

We give a second derivation of the binomial distribution. Suppose

$$Y_j \mid \lambda_j \sim_{ind} \text{Poisson}(\lambda_j),$$

are independent Poisson random variables with rates λ_j, $j = 1, 2$.

Then,

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

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

This can be useful in applied work, if one wishes to compare the ratio of rates from a Poisson process.

Rare Events

Suppose that

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

and that $p \to 0$ and $N \to \infty$, in such a way that $\mathrm{E}[Y] = Np$ approaches a constant λ.

Then, in the limit,

$$Y \mid \lambda \sim \text{Poisson}(\lambda).$$

Rare Events

Approximating the binomial distribution with a Poisson has a number of advantages:

- Computationally, the Poisson model can be more stable than the Poisson model.
- $\lambda > 0$ can be modeled via a loglinear form which provides a more straightforward interpretation than the logistic form, $\log[p/(1-p)]$ – relative rates are easier to think about than odds ratios.
- Poissons sum whereas binomials in general do not.

The following example illustrates one use of this result, for obtaining a closed form distribution when counts are summed.

Example: Lung Cancer and Radon

A possible model for these data is the Poisson model

$$Y_i \mid \theta_i \sim \text{Poisson}(E_i\theta_i), \tag{3}$$

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.

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 aggregate count.

Example: Lung Cancer and Radon

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 $Y_i, N_i - Y_{i1} - Y_{i2}, N_{i1}, N_{i2}$.

Example: Lung Cancer and Radon

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:

$$Y_i \mid p_{i1}, \ldots, p_{iJ} \sim \text{Poisson}\left(\sum_{j=1}^{J} N_{ij} p_{ij} \right). \tag{4}$$

Example: Lung Cancer and Radon

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.

Consequently, (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 (3).

GLMs for Binary Data

Generalized Linear Models (GLMs) for Binary Data

Let $Z_{ij} = 0/1$ denote the absence/presence of the binary characteristic of interest in each of $j = 1, \ldots, N_i$ trials, with each trial $i = 1, \ldots, n$, corresponding to different "conditions" (corresponding to covariate specifications).

Further, let the number of positive responses be denoted,

$$Y_i = \sum_{j=1}^{N_i} Z_{ij}.$$

Suppose there are k covariates/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 $i = 1, \ldots, n$.

GLMs for Binary Data

We wish to model the probability of a positive response

$$p(\boldsymbol{x}_i) = \Pr(Z_{ij} = 1 | \boldsymbol{x}_i), \qquad j = 1, \ldots, N_i, \quad i = 1, \ldots, n,$$

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.

The observed proportions must lie in the range $[0, 1]$, while the modeled probability $\boldsymbol{x}_i \boldsymbol{\beta}$ is unrestricted.

Putting constraints on the parameters is inelegant, causes difficulties for inference, and soon becomes cumbersome with multiple explanatory variables.

Also, 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

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

under a binomial model, so that the variance changes with the mean.

Many other binary data models can be envisaged, beyond the binomial, but it's hard to think of a situation when the variance isn't a function of the mean.

A GLM can rectify these deficiencies.

For sums of independent binary variables the binomial model is a good starting point.

The binomial model is a member of the exponential family, specifically

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

can be written,

$$p(y \mid p) = \exp\left[y \log\left(\frac{p}{1-p} \right) + N \log(1-p) \right], \tag{5}$$

which provides the stochastic element of the model.

For the deterministic part we specify a monotonic, differentiable link function:

$$g\left[\, p(\boldsymbol{x}) \,\right] = \boldsymbol{x}\beta. \tag{6}$$

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.

With a GLM the computation is also usually straightforward.

Non-linear 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.

They desired a model for which disease risk returned to baseline (and not zero) as $x \to \infty$ and so suggested a model of the form

$$p(x) = \beta_0 \left[1 + \beta_1 \exp(-\beta_2 x^2)\right],$$

with β_0 corresponding to baseline risk, β_1 to the excess risk at $x = 0$ (i.e., at the point source), and with β_2 determining the speed at which the risk declines to baseline.

Such nonlinear-models are computationally more difficult to fit but produce consistent parameter estimation, if combined with an exponential family for the response.

Link Functions

From (5) we see that the so-called canonical link is the logit

$$\theta = \log\left(\frac{p}{1-p}\right).$$

We will see that so-called logistic regression models of the form

$$\log\left(\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})}\right) = \boldsymbol{x}\boldsymbol{\beta}, \tag{7}$$

offer a number of advantages in terms of computation and inference.

Link Functions

Other link functions that may be used for binomial data include the probit, complimentary log-log and log-log links.

The probit link is,

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

where $\Phi[\cdot]$ is the distribution function of a standard normal random variable.

The probit link function generally produces similar inference to the logistic link function.

The logistic and probit link functions are symmetric in the sense that

$$g(p) = -g(1-p).$$

Link Functions

The complementary log-log link function is

$$\log\{-\log[1-p(\boldsymbol{x})]\} = \boldsymbol{x}\beta, \tag{8}$$

to give

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

which is not symmetric.

Hence, the log-log link model

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

with

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

may also be used and will not produce the same inference as (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 $g_{\text{CLL}}(p) = -g_{\text{LL}}(1-p)$.

Overdispersion

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 $N_i = C_i \times k_i$, $i = 1, \ldots, n$.

This set-up, with fixed cluster sizes, is a little unrealistic, but it shows the effect of within-trial clustering, and can be extended to unequal cluster sizes.

These clusters may correspond to families, geographical areas, genetic subgroups, etc.

Overdispersion

Within the c-th 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 cluster $c = 1, \ldots, C_i$.

Let P_{ic} represent a random variable with

$$\begin{aligned} \mathrm{E}[P_{ic}] &= p_i \\ \mathrm{var}(P_{ic}) &= \tau_i^2 p_i (1 - p_i), \end{aligned}$$

where the variance is written in this form for convenience (as we see shortly).

Letting $Y_i = \sum_{c=1}^{C_i} Y_{ic}$, we have,

$$\begin{aligned}
\mathrm{E}[Y_i] &= \mathrm{E}\left[\sum_{c=1}^{C_i} Y_{ic}\right] \\
&= \sum_{c=1}^{C_i} \mathrm{E}_{P_{ic}}\left[\mathrm{E}(Y_{ic} \mid p_{ic})\right] \\
&= \sum_{c=1}^{C_i} \mathrm{E}_{P_{ic}}[k_i P_{ic}] \\
&= N_i p_i,
\end{aligned}$$

which corresponds to the marginal expectation (averaging over the uncertainty in P_{ic}).

Overdispersion

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.

Using iterated variance,

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

The marginal variance,

$$\text{var}(Y_i) = N_i p_i (1 - p_i) \times \underbrace{\left[1 + (k_i - 1)\tau_i^2\right]}_{=\sigma_i^2 \geqslant 1},$$

shows that 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 (2), because binary data cannot display overdispersion.

The above derivation requires $1 \leqslant \sigma_i^2 \leqslant k_i \leqslant N_i$, since $0 \leqslant \tau_i^2 \leqslant 1$ (McCullagh and Nelder, 1989, Section 4.5.1).

If we were to assume a second moment model with a common σ_i^2 to give

$$\text{var}(Y_i) = N_i p_i (1 - p_i)\sigma^2 \tag{9}$$

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

If we have a single cluster, i.e. $C_i = 1$, then $k_i = N_i$ and

$$\text{var}(Y_i) = N_i p_i (1 - p_i) \times \left[1 + (N_i - 1)\tau_i^2\right]. \tag{10}$$

Overdispersion

Recall, 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 ($C_i = 1$),

$$\begin{aligned}
\text{cov}(Z_{ij}, Z_{ik}) &= \text{E}[\underbrace{\text{cov}(Z_{ij}, Z_{ik} \mid p_{i1})}_{=0}] + \text{cov}(\text{E}[Z_{ij} \mid p_{i1}], \text{E}[Z_{ik} \mid p_{i1}]) \\
&= \text{cov}_{P_{i1}}(P_{i1}, P_{i1}) \\
&= \text{var}(P_{i1}) \\
&= \tau_i^2 p_i (1 - p_i),
\end{aligned}$$

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

Overdispersion

If we start with unequal cluster sizes, i.e., $N_i = \sum_{c=1}^{C_i} N_{ic}$, then the above derivations can be repeated and give:

$$\begin{aligned} \mathrm{E}[Y_i] &= N_i p_i, \\ \mathrm{var}(Y_i) &= N_i p_i (1 - p_i) \times \left[1 + \frac{1}{N_i} \sum_{c=1}^{C_i} N_{ic}(N_{ic} - 1)\tau_i^2 \right] \end{aligned}$$

We now discuss a closely-related scenario in which we start by assuming that outcomes within a trial have correlation τ_i^2.

Overdispersion

With a common correlation τ_i^2,

$$\begin{aligned} \text{var}(Y_i) &= \sum_{j=1}^{N_i} \text{var}(Z_{ij}) + 2\sum_{j<j'} \text{cov}(Z_{ij}, Z_{ij'}) \\ &= N_i p_i(1-p_i) + 2\frac{N_i(N_i-1)}{2} p_i(1-p_i)\tau_i^2 \\ &= N_i p_i(1-p_i) \times \left[1 + (N_i - 1)\tau_i^2\right] \end{aligned} \tag{11}$$

Notice that, unlike the derivation leading to (10), underdispersion can occur if $\tau_i^2 < 0$.

The equality of equations (10) and (11) show that the effect of either

- a random response probability, or
- positively correlated outcomes within a trial,

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

Overdispersion

Inferentially, two approaches are suggested:

- We could specify the first two moments only and use quasi-likelihood.
- Alternatively, one can assume a specific distributional form and then proceed with parametric inference, as we now illustrate.

Overdispersion

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

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

An obvious choice of mean model is the linear logistic model

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

Notice that $d = 0$ corresponds to the binomial model.

Overdispersion

Integration over the random effects results in the beta-binomial marginal model:

$$\Pr(Y_i = y_i) = \binom{N_i}{y_i} \frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\Gamma(b_i)} \frac{\Gamma(a_i + y_i)\Gamma(b_i + N_i - y_i)}{\Gamma(a_i + b_i + N_i)},$$

$y_i = 0, 1, \ldots, N_i$.

The marginal moments are

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

confirming that there is no overdispersion when $N_i = 1$.

This variance is also equal to (10), with the assumption of constant τ_i^2, on recognizing that $\tau^2 = (a_i + b_i + 1)^{-1} = 1/(d + 1)$.

Overdispersion

Unfortunately, the likelihood $L(\beta, d)$ is not easy to deal with analytically, 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

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

against N_i in order to see whether there is any association, which may help to choose between models (9) and (10).

Logistic Regression Models

Logistic Regression Models: Parameter Interpretation

We write the probability of $Y = 1$ as $p(\boldsymbol{x})$, to emphasize the dependence on covariates $\boldsymbol{x}$.

Model (7) is equivalent to saying that the odds of a positive outcome may be modeled in a multiplicative fashion, i.e.,

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

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 $0 < p(\boldsymbol{x}) < 1$.

Logistic Regression Models: Parameter Interpretation

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

$$\frac{\exp(\beta_0)}{1+\exp(\beta_0)}$$

is the probability of disease, regardless of the levels of $\boldsymbol{x}$.

Logistic Regression Models: Parameter Interpretation

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 doubles 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:

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

Logistic Regression Models: Parameter Interpretation

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

Logistic regression models may be defined for multiple factors and continuous variables in an exactly analogous fashion to multiple linear models.

We simply include on the right hand side of (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 non-canonical link functions parameter interpretation is usually more difficult.

The logistic model may be derived in terms of 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

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

for some threshold c.

Suppose that the continuous measure follow a logistic distribution:

$$U(x) \sim \text{Logistic}\,(\,[\mu(x), 1\,)\,.$$

This distribution is,

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

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

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

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 $U(x)$.

Logistic Regression Models

If we assume that the effect of exposure to x is to move the location of the underlying variable $U(x)$ in a linear fashion, i.e.,

$$\mu(x) = a + bx,$$

but while keeping the variance constant, we obtain

$$\text{logit } p(x) = \beta_0 + \beta_1 x,$$

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

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

Likelihood Inference for Logistic Regression Models

We consider the logistic regression model

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

where $\boldsymbol{x}_i$ is a $1 \times (k+1)$ vector of covariates measured on the i-th individual and β is the $(k+1) \times 1$ vector of associated parameters.

We write $p_i(\beta)$ to emphasize that the probability of a positive response is a function of β.

For the general binomial model, the log-likelihood is

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

with score function

$$\boldsymbol{S}(\beta) = \sum_{i=1}^{n} \frac{\partial p_i(\beta)}{\partial \beta} \frac{[Y_i - N_i p(\widehat{\beta})]}{p(\widehat{\beta})[1-p(\widehat{\beta})]}. \tag{12}$$

Likelihood Inference for Logistic Regression Models

Letting $\boldsymbol{\mu}$ represent the $n \times 1$ vector with i-th element $\mu_i = N_i p_i(\boldsymbol{\beta})$ allows the score (12) to be rewritten as

$$\boldsymbol{S}(\boldsymbol{\beta}) = \boldsymbol{D}(\boldsymbol{\beta})^{\mathsf{T}} \boldsymbol{V}(\boldsymbol{\beta})^{-1} \left[\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})\right], \tag{13}$$

where

- $\boldsymbol{D}(\boldsymbol{\beta})$ 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 i-th diagonal element $N_i p(\boldsymbol{x}_i)\left[1 - p(\boldsymbol{x}_i)\right]$.

Likelihood Inference for Logistic Regression Models

Then:

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

where $I_n(\boldsymbol{\beta}) = \boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{D}$.

For the logistic model,

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

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,

$$\boldsymbol{x}^{\mathsf{T}}\boldsymbol{Y} = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})$$

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 $0 < q \leqslant k$ parameters and fitted probabilities $\widehat{\boldsymbol{p}}$.

Likelihood Inference for Logistic Regression Models

The log-likelihood is

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

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

The deviance is

$$\begin{aligned} D &= 2 \left[\ell(\widetilde{\boldsymbol{p}}) - \ell(\widehat{\boldsymbol{p}}) \right] \\ &= 2 \sum_{i=1}^{n} \left[y_i \log \left(\frac{y_i}{\widehat{y}_i} \right) + (N_i - y_i) \log \left(\frac{N_i - y_i}{N_i - \widehat{y}_i} \right) \right], \end{aligned} \tag{14}$$

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.

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 provide an example of when 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 levels of coma score, pupils, haematoma and age (we illustrate, via an analyses of these data, shortly).

Likelihood Inference for Logistic Regression Models

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 differences in deviance.

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

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

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

Likelihood Inference for Logistic Regression Models

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,

$$\text{logit}(p_i) = \boldsymbol{x}_i \beta,$$

for $i = 1, \ldots, n$.

We fit this model using maximum likelihood, resulting in estimates $\widehat{\beta}$ and fitted probabilities $\widehat{\boldsymbol{p}}$.

In this case (14) becomes, since $y_i \log y_i = (1 - y_i) \log(1 - y_i) = 0$,

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

At the MLE, $\boldsymbol{x}^\top \boldsymbol{y} = \boldsymbol{x}^\top \hat{\boldsymbol{p}}$, so that

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

and the deviance is a function only of $\hat{\beta}$, i.e., regardless of the data, it is only a function of the parameter estimate.

In other words, D is a deterministic function of $\hat{\beta}$, 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:

$$X^2 = \sum_{i=1}^{n} \frac{(Y_i - N_i\hat{p}_i)^2}{N_i\hat{p}_i(1 - \hat{p}_i)}, \tag{15}$$

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, Section 4.4.5).

Quasi-Likelihood Inference for Logistic Regression Models

An extremely simple and appealing manner of dealing with overdispersion is to assume the model

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

with

$$\mathrm{cov}(Y_i, Y_j \mid \beta) = 0,$$

for $i \neq j$.

Quasi-Likelihood Inference for Logistic Regression Models

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 $\widehat{\alpha}(\boldsymbol{D}^{\mathsf{T}}\boldsymbol{V}^{-1}\boldsymbol{D})^{-1}$.

An obvious estimator of α is provided by the method of moments, which corresponds to the Pearson statistic (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.

Bayesian Inference for Logistic Regression Models

A Bayesian approach to inference combines the likelihood $L(\boldsymbol{\beta})$ with a prior $\pi(\boldsymbol{\beta})$ – 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, i.e., no transfer of information.

Bayesian Inference for Logistic Regression Models

With multivariate normal priors, computation may be carried out using INLA though this approximation strategy may be inaccurate if the binomial denominators are very small.

An alternative is provided by MCMC.

It is common to encounter excess-binomial variation.

This may be dealt with in a Bayesian context parametrically via the use of a beta-binomial likelihood, or more generally through the introduction of random effects.

Bayesian Inference for Logistic Regression Models

A two-stage random effects model is:

Stage One: The likelihood:

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

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 β and σ_0^2.

Example: Outcome After Head Injury

Parameter estimation, whether via likelihood or Bayes is straightforward for these data given a particular model – the difficult task is deciding on a model.

In exploratory mode, we illustrate some approaches to model selection – applying the hierarchy principle, 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`, `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.

Example: Outcome After Head Injury

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.

Example: Outcome After Head Injury

We consider inference for the model:

$$\texttt{1} + \texttt{H} + \texttt{P} + \texttt{C} + \texttt{A2} + \texttt{A3} + \texttt{H.P} + \texttt{H.A2} + \texttt{H.A3} + \texttt{P.A2} + \texttt{P.A3}, \qquad (16)$$

i.e., the model with main effects for:

- haematoma (`H`),
- pupils (`P`),
- coma score (`C`) and
- age (with `A2` and `A3` representing the second and third levels).

Interctations between

- haematoma and pupils (`H.P`),
- haematoma and age (`H.A2` and `H.A3`) and
- pupils and age (`P.A2` and `P.A3`).

Example: Outcome After Head Injury

The MLEs and standard errors are given in Table 2, along with Bayesian posterior means and standard deviations.

The prior on the intercept was taken as flat and for the ten log odds ratios independent normal priors $N(0, 4.70^2)$ were taken, which correspond to 95% intervals for the odds ratios of [0.0001,10000], i.e. 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.

Example: Outcome After Head Injury

	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

Table 2: Likelihood and Bayesian estimates and uncertainty measures for model (16) applied to the head injury data.

Example: Outcome After Head Injury

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 being having a high coma score reduces the odds of death by $\exp(-1.52) = 0.22$.

The observed and fitted probabilities are displayed in Figure 3 with different line types joining the observed probabilities (as in Figure 1).

The vertical lines join the fitted to the observed probabilities, with the same line type as the observed probabilities with which are associated. There are no clear badly fitting cells.

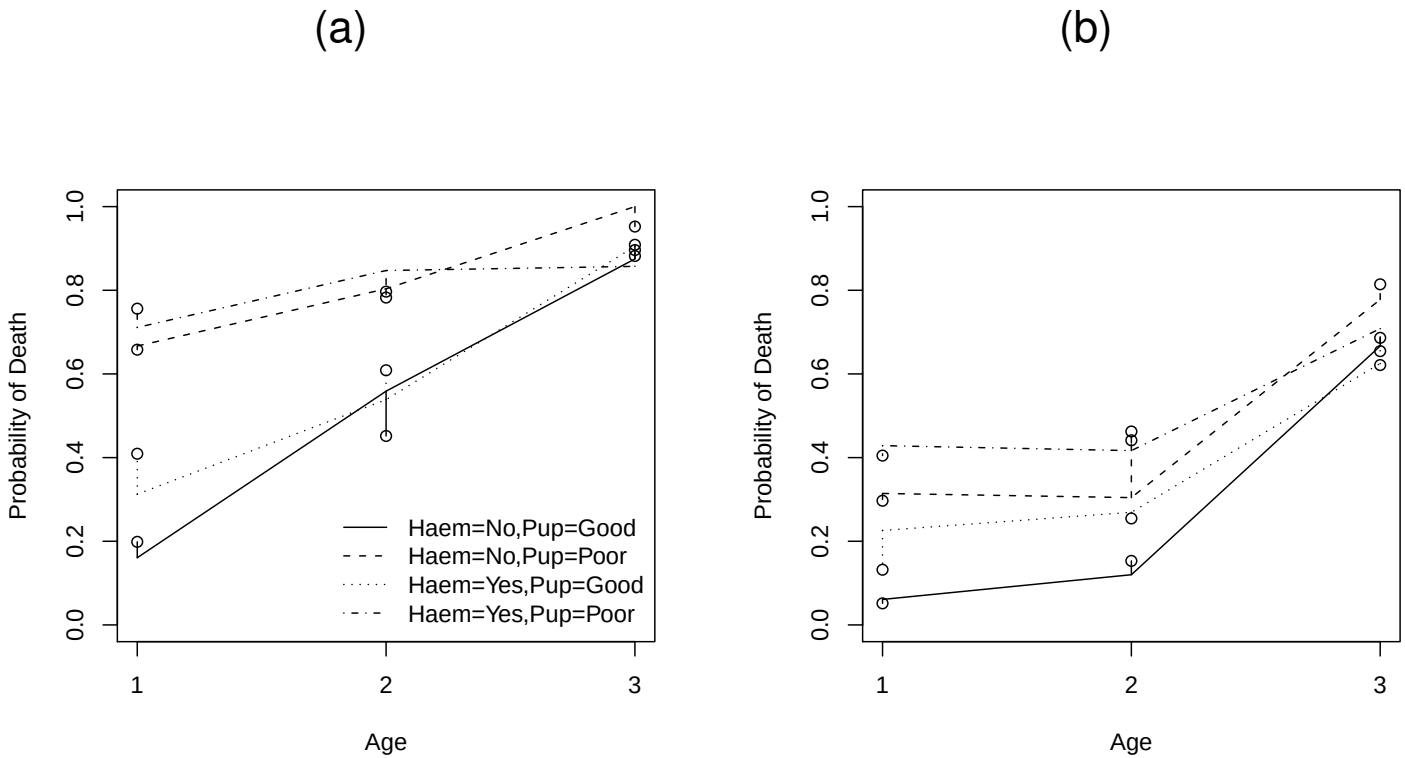

FIGURE 3: Probability of death after head injury as a function of age, haematoma score and pupils. Panels (a) and (b) are for low and high coma scores, respectively. The open circles are the fitted values. The observed values are joined by different line types. The residuals $y/n - \widehat{p}$ are shown as vertical lines of the same line type.

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

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \exp\left(\beta_0 \sum_{i=1}^{n} y_i + \beta_1 \sum_{i=1}^{n} x_i y_i - n_i \log\left[1 + \exp(\beta_0 + \beta_1 x_i)\right] \right) \tag{17}$$

where $\boldsymbol{\beta} = [\beta_0, \beta_1]$.

The MLEs and variance covariance matrix are

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \begin{bmatrix} -5.34 \\ 0.0015 \end{bmatrix}, \\ \widehat{\text{var}}(\widehat{\boldsymbol{\beta}}) &= \begin{bmatrix} 0.54^2 & -0.99 \times 0.54 \times 0.00016 \\ -0.99 \times 0.54 \times 0.00016 & 0.00016^2 \end{bmatrix}. \end{aligned}$$

Example: Aircraft Fasteners

The solid line in Figure 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.

Figure 4 also shows the fit from these models.

Example: Aircraft Fasteners

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 can be used for such comparisons but the approximations inherent in the derivation are more accurate for nested models (Ripley, 2004).

The differences are so small here that we would not make any conclusions on the basis of these numbers.

Example: Aircraft Fasteners

Since the number of x categories is not fixed in this example we cannot formally examine the absolute fit of the models.

In Figure 6 we see that residual plots for these three models indicates the logistic fit is best.

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

$$\exp\left[500 \times \widehat{\beta}_1 \pm 1.96 \times 500\sqrt{\text{var}(\widehat{\beta}_1)}\right] = [1.86, 2.53]. \qquad (18)$$

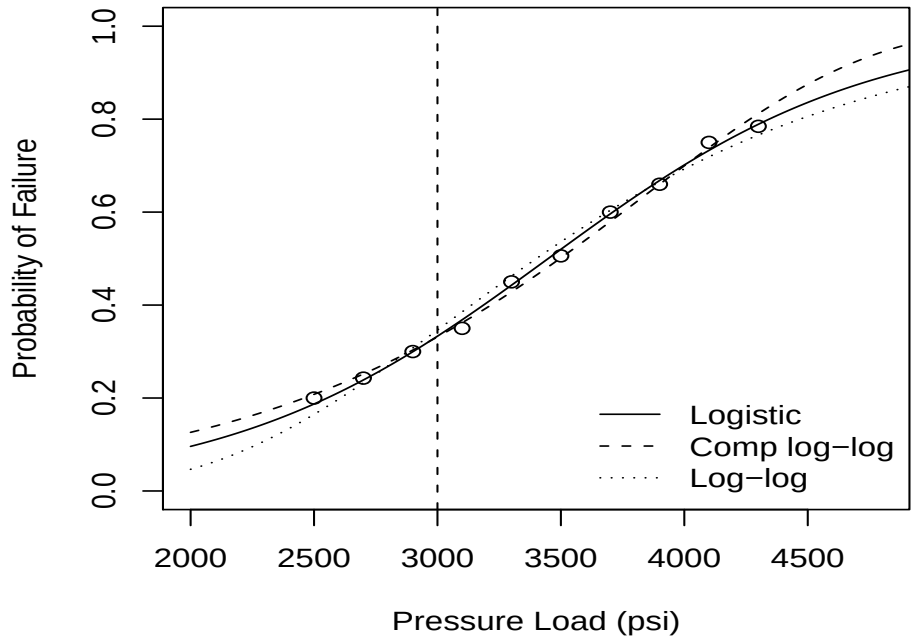

FIGURE 4: Fitted curves for the aircraft fasteners data under three different link functions.

Example: Aircraft Fasteners

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 (17). We use a bivariate Metropolis-Hastings random walk MCMC algorithm to explore the posterior.

INLA would be easier!

Example: Aircraft Fasteners

A bivariate normal proposal was used, with variance-covariance matrix proportional to the asymptotic variance covariance matrix, $\widehat{\text{var}}(\hat{\boldsymbol{\beta}})$, (18). This matrix was multiplied by four to give an acceptance ratio of around 30%.

Panels (a) and (b) of Figure 5 shows 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 pis increase in pressure is identical to the asymptotic likelihood interval (18).

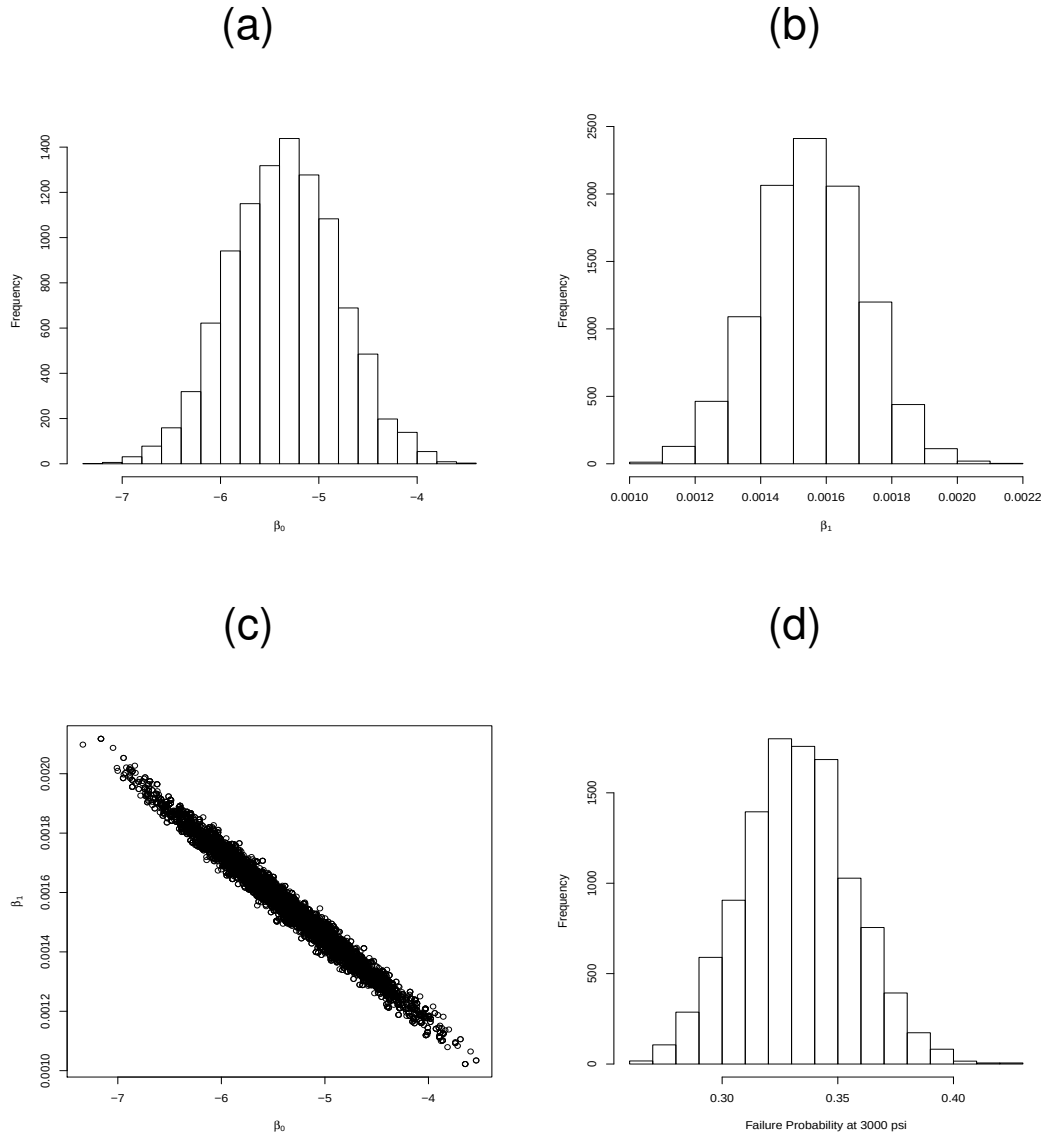

FIGURE 5: Posteriors for the aircraft fasteners data: (a) $p(\beta_0|\boldsymbol{y})$, (b) $p(\beta_1|\boldsymbol{y})$, (c) $p(\beta_0, \beta_1|\boldsymbol{y})$, (d) $p(\exp(\theta)/[1+\exp(\theta)]|\boldsymbol{y})$, where $\theta = \beta_0 + \beta_1\tilde{x}$.

Example: Aircraft Fasteners

We imagine that it is of interest to give an interval estimate for the probability of failure at $\tilde{x} = 3000$ psi (which is indicated as a dashed vertical line on Figure 4).

An asymptotic 95% confidence interval for $\theta = \beta_0 + \beta_1 \tilde{x}$ is

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

where

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

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 samples $\beta^{(s)}$, to give expit$(\theta^{(s)})$, $s = 1, \ldots, S$ results in the 95% interval of [0.29,0.38], which is again identical to the frequentist interval.

Conditional Likelihood Inference

Conditional Likelihood Inference

Suppose the distribution for the data can be represented as,

$$p(\boldsymbol{y} \mid \lambda, \phi) \propto p(\boldsymbol{t}_1 \mid \boldsymbol{t}_2, \lambda) p(\boldsymbol{t}_2 \mid \lambda, \phi), \tag{19}$$

where λ is a parameter of interest and ϕ is a nuisance parameter.

Then inference for λ may be based on the conditional likelihood

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

Perhaps the most popular use of conditional likelihood leads to Fisher's exact test.

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

Table 3: A generic 2×2 table.

Conditional Likelihood Inference

Consider the 2×2 lay out of data shown in Table 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, i.e. the odds ratio.

This set up gives likelihood

$$\Pr(y_{01}, y_{11} \mid \beta_0, \beta_1) = \begin{pmatrix} y_{0\cdot} \\ y_{01} \end{pmatrix} \begin{pmatrix} y_{1\cdot} \\ y_{11} \end{pmatrix} \frac{e^{y_{11}\beta_1}}{(1+e^{\beta_0+\beta_1})^{y_{1\cdot}}} \frac{e^{y_{\cdot 1}\beta_0}}{(1+e^{\beta_0})^{y_{0\cdot}}} \tag{20}$$

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

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

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, i.e.

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

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 (19), $\boldsymbol{\lambda} \equiv \beta_1$, $\boldsymbol{\phi} \equiv \beta_0$, $\boldsymbol{t}_1 \equiv y_{11}$, $\boldsymbol{t}_2 \equiv y_{\cdot 1}$.

Accordingly, the conditional distribution takes the form

$$\Pr(y_{11} \mid y_{\cdot 1}, \beta_1) = \frac{\binom{y_{0\cdot}}{y_{\cdot 1} - y_{11}} \binom{y_{1\cdot}}{y_{11}} e^{y_{11}\beta_1}}{\sum_{u=u_0}^{u_1} \binom{y_{0\cdot}}{y_{\cdot 1} - u} \binom{y_{1\cdot}}{u} e^{u\beta_1}}, \tag{21}$$

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 (21). In particular, the conditional MLE may be determined, though unfortunately no closed form exists.

Conditional Likelihood Inference

Conventionally, estimates of β_0 and β_1 would be determined from the product of binomial likelihoods, (20).

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 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).

Conditional Likelihood Inference

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

$$c(u) = \begin{pmatrix} y_{0\cdot} \\ y_{\cdot 1} - u \end{pmatrix} \begin{pmatrix} y_{1\cdot} \\ u \end{pmatrix}.$$

The (conditional) score is

$$S_c(\beta_1) = \frac{\partial}{\partial \beta_1} \log L_c(\beta_1) = y_{11} - \frac{\sum_{u=u_0}^{u_1} c(u)ue^{\hat{\beta}_1 u}}{\sum_{u=u_0}^{u_1} c(u)e^{\hat{\beta}_1 u}}. \tag{22}$$

The extended hypergeometric distribution is a member of the exponential family and, since the second term in (22) is the mean of this distribution,

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

at the MLE.

Consequently, from (22), we can equate $\mathrm{E}[Y_{11} \mid \widehat{\beta}_1] = y_{11}$ and solve for $\widehat{\beta}_1$.

Asymptotic inference is based on

$$I_c(\beta_1)^{1/2}\left(\widehat{\beta}_1 - \beta_1\right) \to_d \mathrm{N}(0,1), \tag{23}$$

where the (conditional) information is

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

It is straightforward to test the null hypothesis $H_0 : \beta_1 = 0$ using the conditional likelihood.

When $\beta_1 = 0$ the distribution (21) is hypergeometric and so

$$\Pr(y_{11} \mid y_{\cdot 1}, \beta_1 = 0) = \frac{\binom{y_{0\cdot}}{y_{\cdot 1} - y_{11}} \binom{y_{1\cdot}}{y_{11}}}{\binom{y_{\cdot\cdot}}{y_{\cdot 1}}}. \tag{24}$$

The comparison of the observed y_{11} with the tail of this distribution is known as Fisher's exact test.

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 (23) or by inverting the test. See Agresti (1990, Sections 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 presented in Table 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 for 12 hours per day.

Another group of mice were kept in an alternative chamber without smoke. After one 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 4 give the numbers of mice with and without tumors in the “control” and “treated” groups.

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

TABLE 4: Data on tumor appearance within rats.

Example: Tumor Appearance Within Mice

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 (24) are 0.00739, 0.00091, 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.

Example: Tumor Appearance Within Mice

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

$$\widehat{\beta}_1^u = \log\left(\frac{21 \times 13}{2 \times 91}\right) = \log(7.18) = 1.97,$$

with asymptotic standard error

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

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.

Example: Tumor Appearance Within Mice

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

Assessment of Assumptions

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.

Recall, 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

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

$i = 1, \ldots, n$.

Pearson's chi 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 measures of the 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 the deviance $D = \sum_{i=1}^{n}(e_i^\star)^2$ where D is given by (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 Figure 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.

Assessment of Assumptions

Hence, they may be put to many of the same uses as residual defined with respect to the normal linear regression model.

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, Section 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.

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 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 misspecification.

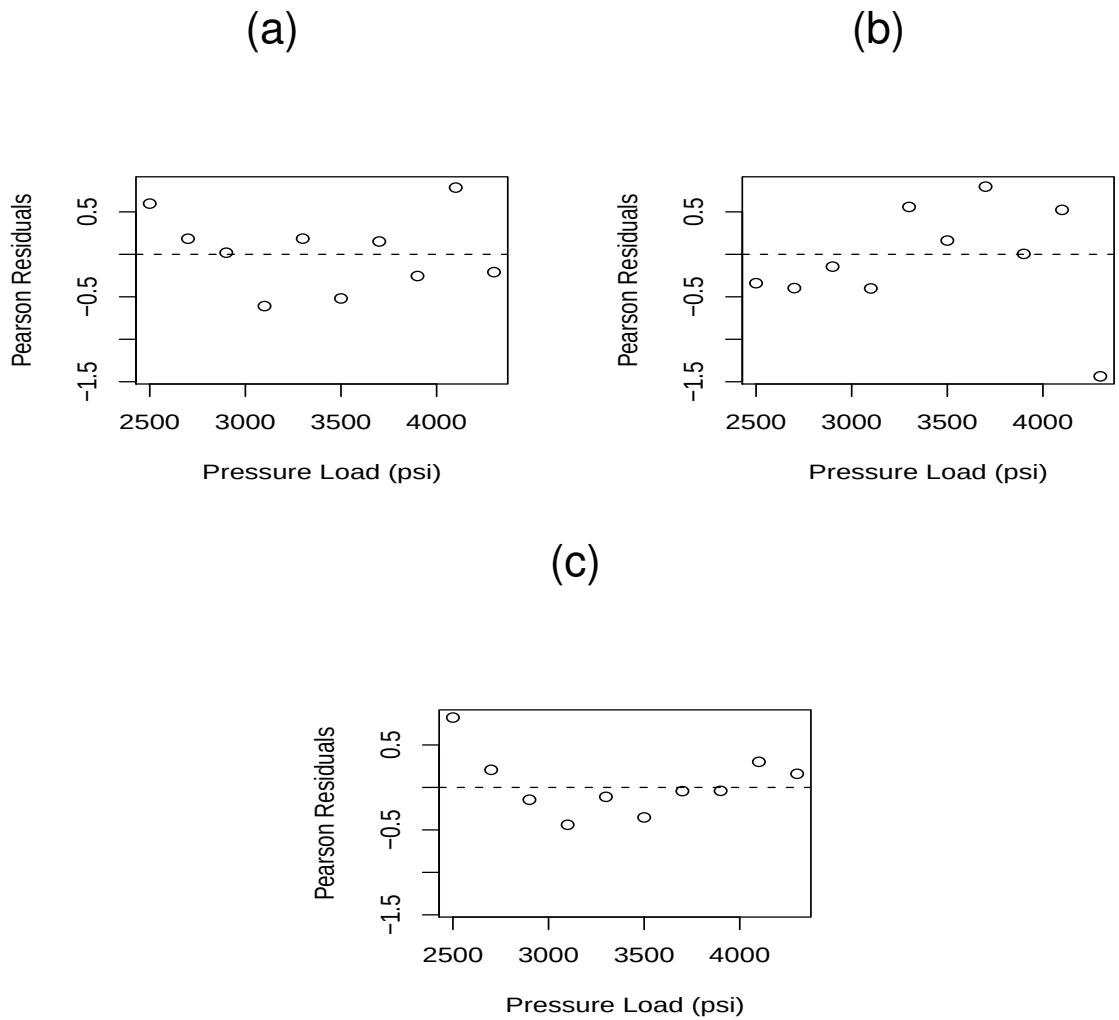

FIGURE 6: Pearson residuals versus pressure load for the aircraft fasteners data for: (a) logistic link model, (b) complementary log-log link model, (c) log-log link model.

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 Figure 3 that the fit of model (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 gives a normal QQ plot of the Pearson residuals and there are no causes for concern with no outlying points.

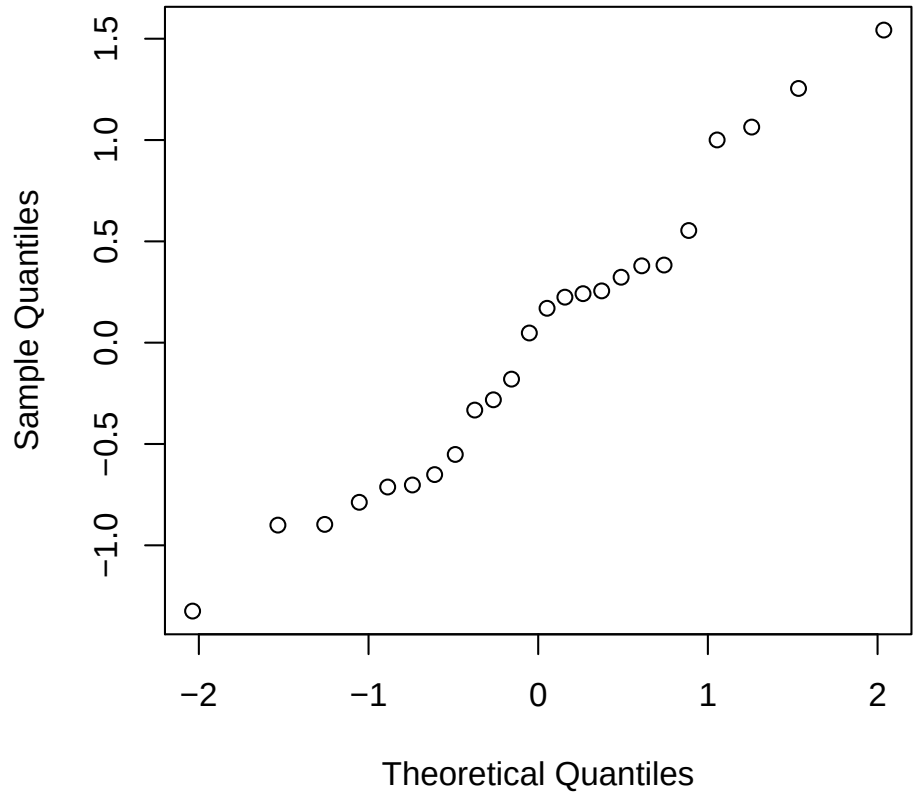

FIGURE 7: QQ plot of Pearson residuals for the head injury data.

Example: BPD and Birthweight

We fit a logistic regression model

$$\Pr(Y = 1 \mid x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}, \tag{25}$$

with $Y = 0/1$ corresponding to absence/presence of BPD and x to birthweight.

The curve arising from fitting this model is shown on Figure 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 birthweight in Figure 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.

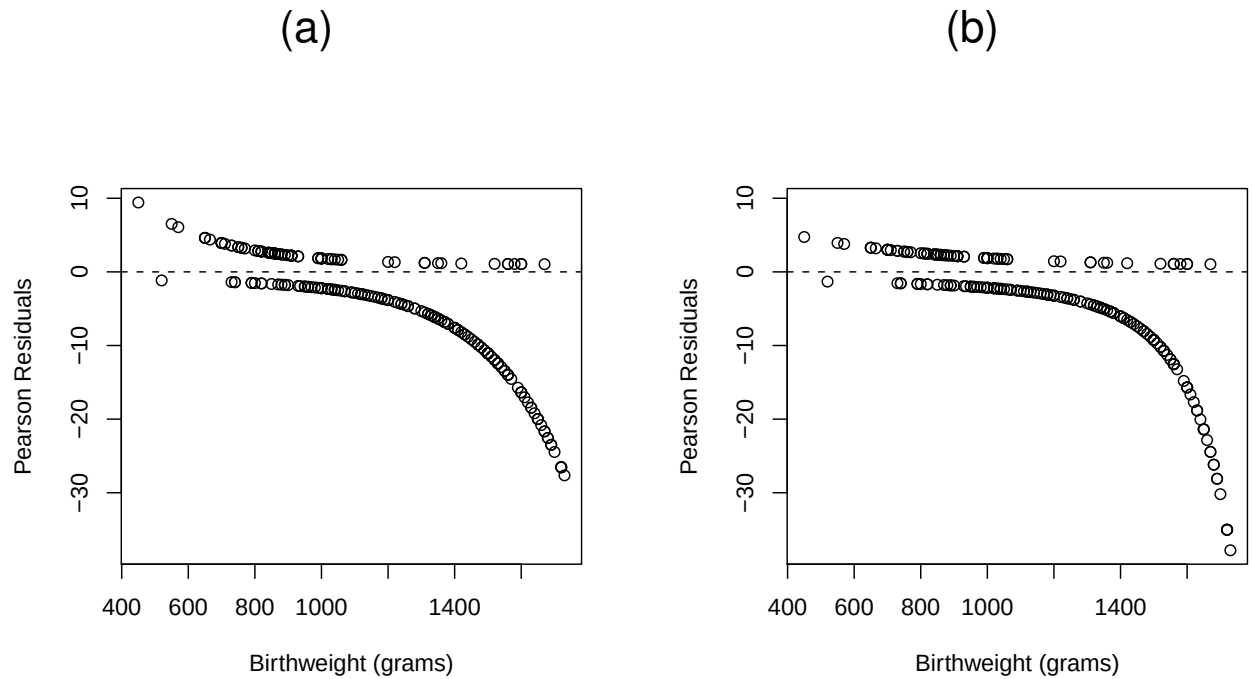

FIGURE 8: Pearson residuals versus birthweight for the BPD data: (a) logistic model, (b) complementary log-log model.

Inference for Arbitrary Functions

Inference for Arbitrary Functions

Suppose that from a frequentist procedure we obtain estimates $\widehat{\theta}_n$, with associated variance-covariance $\boldsymbol{V}/n$ for a p-dimensional vector of parameters $\boldsymbol{\theta}$.

We have

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \to_d \mathrm{N}_p(\mathbf{0}, \boldsymbol{V}),$$

where $\boldsymbol{\theta}_0$ is the true value.

We wish to make inference for a function of interest $g(\boldsymbol{\theta}) : \mathbb{R}^p \to \mathbb{R}$ and we have already discussed the delta method by which we use

$$\sqrt{n}(g(\widehat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta}_0)) \to_d \mathrm{N}(\mathbf{0}, \boldsymbol{D}^{\mathsf{T}}\boldsymbol{V}\boldsymbol{D}),$$

where $\boldsymbol{D} = [\partial g/\partial\theta_1, \ldots, \partial g/\partial\theta_p]^{\mathsf{T}}$ is the gradient of $g(\cdot)$.

It can be time-consuming to implement this method, particularly if we have many functions of interest, and so we describe an alternative.

Inference for Arbitrary Functions

The method can be summarized via the algorithm:

1. Simulate parameter estimates $\widehat{\theta}^{(s)}$ from $\mathrm{N}_p(\widehat{\theta}, \boldsymbol{V}/n)$, for $s = 1, \ldots, S$.
2. For each sample, calculate $g(\widehat{\theta}^{(s)})$ and then calculate the mean squared error (MSE)

$$\widehat{\mathrm{MSE}} = \frac{1}{S} \sum_{s=1}^{S} \left(g(\widehat{\theta}^{(s)}) - g(\widehat{\theta}_n) \right)^2 .$$

3. Construct the confidence interval

$$g(\widehat{\theta}_n) \pm z_{1-/\alpha/2} \sqrt{\widehat{\mathrm{MSE}}}$$

where $z_{1-/\alpha/2}$ is the $1 - \alpha/2$ quantile of an $\mathrm{N}(0, 1)$ random variable.

This method can be justified as a type of parametric bootstrap (Mandel, 2013) in which we replace the simulation of new datasets steps with the simulation of new parameter estimates.

Inference for Arbitrary Functions

We can also justify via a Bayesian argument in which we take as likelihood the sampling distribution of the estimator and an improper prior – the inferential interpretation is obviously different under the two justifications.

Specifically,

$$\begin{aligned} \widehat{\boldsymbol{\theta}}|\boldsymbol{\theta} &\sim \mathsf{N}_p(\boldsymbol{\theta}, \boldsymbol{V}/n) \\ \pi(\boldsymbol{\theta}) &\propto 1, \end{aligned}$$

gives the posterior,

$$\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}} \sim \mathsf{N}_p(\widehat{\boldsymbol{\theta}}, \boldsymbol{V}/n).$$

We can simulate to obtain samples from the posterior $g(\boldsymbol{\theta})|\widehat{\boldsymbol{\theta}}$.

With a Bayesian slant the algorithm is:

1. Simulate posterior samples $\boldsymbol{\theta}^{(s)}$ from $\mathsf{N}_p(\widehat{\boldsymbol{\theta}}, \boldsymbol{V}/n)$, for $s = 1, \ldots, S$.
2. For each sample, calculate $g(\boldsymbol{\theta}^{(s)})$ and then these samples can be used as samples from the posterior $g(\boldsymbol{\theta})|\widehat{\boldsymbol{\theta}}$.

Bias, Variance and Collapsibility

Bias, Variance and Collapsibility

We recap results on the the bias and variance of estimators for the linear model.

Consider the models:

$$\mathrm{E}[Y \mid x, z] = \beta_0 + \beta_1 x + \beta_2 z \tag{26}$$

$$\mathrm{E}[Y \mid x, z] = \beta_0^\star + \beta_1^\star x. \tag{27}$$

First, suppose that x and z are orthogonal.

Roughly speaking, if z is related to Y, then fitting model (26) 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 (27) 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

$$\mathsf{E}[Y|x,z] = \frac{\exp(\beta_0+\beta_1 x+\beta_2 z)}{1+\exp(\beta_0+\beta_1 x+\beta_2 z)} \tag{28}$$

$$\mathsf{E}[Y|x,z] = \frac{\exp(\beta_0^*+\beta_1^* x)}{1+\exp(\beta_0^*+\beta_1^* x)} = \mathsf{E}_{z|x}\left[\frac{\exp(\beta_0+\beta_1 x+\beta_2 z)}{1+\exp(\beta_0+\beta_1 x+\beta_2 z)}\right]. \tag{29}$$

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, $\mathsf{E}[\beta_1] \neq \mathsf{E}[\beta_1^\star]$.

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.

Noncollapsibility is sometimes referred to as Simpson's Paradox in the statistics literature.

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 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 ratio 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.

		$z=0$		$z=1$		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	

Table 5: Illustration of Simpson's paradox for the case of non-orthogonal x and z.

Mathematically, the paradox is relatively simple to understand.

Let

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

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, $x, z = 0, 1$.

The "paradox" reflects the fact that it is possible to have

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

i.e. 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 $x = 0$. For the data of Table 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

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

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 situation of balance 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 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 towards the null, with the marginal association being 1.2.

Dipping a Toe in the Causal Water

Suppose $z = 0/1$ corresponds to female/male.

For females the treatment effect is

$$\Pr(Y = 1|x = 1, z = 0) - \Pr(Y = 1|x = 0, z = 0) = 12/30 - 2/10 = 2/10.$$

For males the treatment effect is

$$\Pr(Y = 1|x = 1, z = 1) - \Pr(Y = 1|x = 0, z = 1) = 8/10 - 9/30 = 5/10.$$

So positive risk difference in each.

Marginally:

$$\Pr(Y = 1|x = 1) - \Pr(Y = 1|x = 0) = 20/40 - 23/40 = -3/40,$$

a negative risk difference.

This shows that confounding distorts all outcome measures (so its not collapsibility).

Dipping a Toe in the Causal Water

From Pearl et al (2016, Section 3.2):

Average Causal Effect (ACE) $= \Pr(Y = 1|\text{do}(X = 1)) - \Pr(Y = 1|\text{do}(X = 0))$

where

$$\Pr(Y = 1|\text{do}(X = x)) = \sum_z \Pr(Y = 1|X = x, Z = z) \times \Pr(Z = z).$$

For the example:

$$\begin{aligned}
\text{ACE} &= \Pr(Y = 1|\text{do}(X = 1)) - \Pr(Y = 1|\text{do}(X = 0)) \\
&= \Pr(Y = 1|X = 1, Z = 0)\Pr(Z = 0) + \Pr(Y = 1|X = 1, Z = 1)\Pr(Z = 1) \\
&- \Pr(Y = 1|X = 0, Z = 0)\Pr(Z = 0) - \Pr(Y = 1|X = 0, Z = 1)\Pr(Z = 1) \\
&= [12/30 \times 1/2 + 8/10 \times 1/2] - [2/10 \times 1/2 + 21/30 \times 1/2] \\
&= 0.60 - 0.45 = 0.15
\end{aligned}$$

		$z=0$		$z=1$		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	

Table 6: Illustration of Simpson's paradox for the case of orthogonal x and z.

Note: Risk difference is not distorted here since linear. It's 0.05 in both conditional tables and in the marginal.

Bias, Variance and Collapsibility

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 noncollapsibility.

Bias, Variance and Collapsibility

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 6.

Let p_{xz} represent the probability of disease for treatment group x and strata z. We may fit the model

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

where we have not included an interaction between x and z.

This results in $\exp(\beta_x) = \exp(0.75) = 2.1$, as expected from Table 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

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

This gives $\exp(\beta_x^\star) = \exp(0.20) = 1.2$, again as expected from Table 6, but with standard error 0.20 which is a reduction from the full 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.

So the right hand 2×2 set of counts in Table 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 and is the averaged 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 sub-populations).

We emphasize that, as mentioned above, the difference between population and sub-population 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.

Case-Control Studies

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*.

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 demonstrates the simplest example in which there is a single binary exposure, with y_{ij} representing 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}.

		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

Table 7: Generic 2×2 table for a binary exposure and binary disease outcome.

Case-Control Studies

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, e.g., 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 most cancers, for example) and so the study may need to run for a long period.

Case-Control Studies

The case-control study provides a way of overcoming these difficulties.

With reference to Table 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 shortly.

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

$$\Pr(Y = 1 \mid \boldsymbol{x}) = p(\boldsymbol{x}) = \frac{\exp\left(\beta_0 + \sum_{j=1}^{p} x_j \beta_j\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^{k} x_j \beta_j\right)}. \tag{30}$$

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, 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 (30),

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

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.

The parameter β_0 represents the baseline odds of disease, corresponding to the odds when all of the exposures are set equal to zero.

We now show how 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:

$$\Pr(Y = 1 \quad | \quad Z = 1, \boldsymbol{x}) = \frac{\Pr(Z = 1 \mid Y = 1, \boldsymbol{x}) \Pr(Y = 1 \mid \boldsymbol{x})}{\Pr(Z = 1 \mid \boldsymbol{x})} \tag{31}$$

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, that 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 (31), 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.

Case-Control Studies

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

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

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 β, with additional nuisance parameters that are 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 may proceed with asymptotic inference acting as it the data were collected in a cohort fashion, except that the intercept is no longer interpretable as the baseline log odds of disease.

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 these variables.

This provides efficiency gains, since the controls are more similar to the cases, 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 (for example, ten year age bands) and controls are matched on the basis of these variables.

Estimation for a Matched Case-Control Study

In an individually matched study controls are matched exactly, usually upon multiples variable, for example, age, gender, time of diagnosis and area of residence.

For both forms of matching the non-random selection of controls must be acknowledged in the analysis by including a parameter for each matching set in the logistic model.

Estimation for a Matched Case-Control Study

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 $i = 1, \ldots, m_{1j}$ denoting 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 $m_0 = \sum_{j=1}^{J} n_{0j}$ controls in total.

Estimation for a Matched Case-Control Study

The disease model is

$$\log\left(\frac{p_j(\boldsymbol{x}_{ij})}{1-p_j(\boldsymbol{x}_{ij})}\right) = \alpha_j + \boldsymbol{x}_{ij}\boldsymbol{\beta} \tag{32}$$

where

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

for $i = 1, \ldots, m_{0j} + m_{1j}$, $j = 1, \ldots, J$.

Estimation for a Matched Case-Control Study

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 previously 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.

Estimation for a Matched Case-Control Study

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 $m_0 = JM = Mm_1$.

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 j-th stratum is

$$p(\boldsymbol{x}_{1j} \mid Y_{1j} = 1) \prod_{i=2}^{M+1} p(\boldsymbol{x}_{ij} \mid Y_{ij} = 0) \tag{33}$$

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.

Estimation for a Matched Case-Control Study

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.

We need to consider

$$p(\boldsymbol{x}_j | \boldsymbol{Y}_j, x_{j+}) = \frac{p(\boldsymbol{x}_j | \boldsymbol{Y}_j)}{p(x_{j+} | \boldsymbol{Y}_j)}$$

Estimation for a Matched Case-Control Study

The numerator is (33) 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 j-th 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

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

Estimation for a Matched Case-Control Study

Substitution of the logistic disease model (32) yields the conditional likelihood

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

with the α_j terms having canceled out, as was required.

Estimation for a Matched Case-Control Study

For further details, see Prentice and Pyke (1979, Section 6). As an example, if $M = 2$ (two controls per case), the conditional likelihood is

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

Estimation for a Matched Case-Control Study

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 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.

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} – the unconditional estimator is the square of the correct conditional estimator.

Estimation for a Matched Case-Control Study

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

Table 8: Notation for a matched pairs case control study with n controls and n cases and a single exposure.

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

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 often 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.

Collapsibility of non-linear measures is tricky to think about.

The effect of omitting variables is also non-obvious.

The fact that the linear logistic model is a GLM does offer advantages in terms of consistency, however, and logistic model allow comparable summary parameters across cross-sectional, cohort and case-control studies.

Concluding Remarks

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) but it is generally accepted as a reasonable approach.

Fisher's exact test is historically popular but frequentist hypothesis testing can be difficult to implement in practice since p-values need to be interpreted as a function of the sample size.

For Fisher's exact, the discreteness of the test statistic can also be problematic.

Final Comment: Keep a balanced view of estimating functions, likelihood and Bayes approaches, they all haver their merits/challenges, and good data analysis is the most important factor in an analysis.

References

Agresti, A. (2990). *Categorical Data Analysis*. John Wiley & Sons.

Breslow, N. and Day, N. (1980). *Statistical Methods in Cancer Research, Volume 1- The Analysis of Case-Control Studies*. Scientific Publications No. 32. Lyon: International Agency for Research on Cancer.

Cox, D. and Snell, E. (1989). *The Analysis of Binary Data*. Chapman and Hall/CRC, Boca Raton, second edition.

Diggle, P. and Rowlingson, B. (1994). A conditional approach to point process modelling of raised incidence. *Journal of the Royal Statistical Society, Series A*, **157**, 433–440.

Liang, K.-Y. and McCullagh, P. (1993). Case studies in binary dispersion. *Biometrics*, **49**, 623–630.

Mandel, M. (2013). Simulation-based confidence intervals for functions with complicated derivatives. *The American Statistician*, **67**, 76–81.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, Second Edition*. Chapman and Hall, London.

Prentice, R. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, **66**, 403–411.

Ripley, B. (2004). Selecting amongst large classes of models. In N. Adams, M. Crowder, D. Hand, and D. Stephens, editors, *Methods and Models in Statistics: In Honor of Professor John Nelder, FRS*, pages 155–170. Imperial College Press, London.