Chapter 3

Logit Models for Binary Data

We now turn our attention to regression models for dichotomous data, including logistic regression and probit analysis. These models are appropriate when the response takes one of only two possible values representing success and failure, or more generally the presence or absence of an attribute of interest.

分成两个的，叉状分枝的

3.1 Introduction to Logistic Regression

We start by introducing an example that will be used to illustrate the analysis of binary data. We then discuss the stochastic structure of the data in terms of the Bernoulli and binomial distributions, and the systematic structure in terms of the logit transformation. The result is a generalized linear model with binomial response and link logit.

3.1.1 The Contraceptive Use Data

Table 3.1, adapted from Little (1978), shows the distribution of 1607 currently married and fecund women interviewed in the Fiji Fertility Survey of 1975, classified by current age, level of education, desire for more children, and contraceptive use.

In our analysis of these data we will view current use of contraception as the response or dependent variable of interest and age, education and desire for more children as predictors. Note that the response has two categories: use and non-use. In this example all predictors are treated as categorical

TABLE 3.1: Current Use of Contraception Among Married Women by Age, Education and Desire for More Children Fiji Fertility Survey, 1975

Age	Education	Desires More Children?	Contraceptive Use		Total
			No	Yes	
<25	Lower	Yes	53	6	59
		No	10	4	14
	Upper	Yes	212	52	264
		No	50	10	60
25–29	Lower	Yes	60	14	74
		No	19	10	29
	Upper	Yes	155	54	209
		No	65	27	92
30–39	Lower	Yes	112	33	145
		No	77	80	157
	Upper	Yes	118	46	164
		No	68	78	146
40–49	Lower	Yes	35	6	41
		No	46	48	94
	Upper	Yes	8	8	16
		No	12	31	43
Total			1100	507	1607

variables, but the techniques to be studied can be applied more generally to both discrete factors and continuous variates.

The original dataset includes the date of birth of the respondent and the date of interview in month/year form, so it is possible to calculate age in single years, but we will use ten-year age groups for convenience. Similarly, the survey included information on the highest level of education attained and the number of years completed at that level, so one could calculate completed years of education, but we will work here with a simple distinction between lower primary or less and upper primary or more. Finally, desire for more children is measured as a simple dichotomy coded yes or no, and therefore is naturally a categorical variate.

The fact that we treat all predictors as discrete factors allows us to summarize the data in terms of the numbers using and not using contraception in each of sixteen different groups defined by combinations of values of the pre-

dictors. For models involving discrete factors we can obtain exactly the same results working with grouped data or with individual data, but grouping is convenient because it leads to smaller datasets. If we were to incorporate continuous predictors into the model we would need to work with the original 1607 observations. Alternatively, it might be possible to group cases with identical covariate patterns, but the resulting dataset may not be much smaller than the original one.

The basic aim of our analysis will be to describe the way in which contraceptive use varies by age, education and desire for more children. An example of the type of research question that we will consider is the extent to which the association between education and contraceptive use is affected by the fact that women with upper primary or higher education are younger and tend to prefer smaller families than women with lower primary education or less.

3.1.2 The Binomial Distribution

We consider first the case where the response y_i is binary, assuming only two values that for convenience we code as one or zero. For example, we could define

$$y_i = \begin{cases} 1 & \text{if the } i\text{-th woman is using contraception} \\ 0 & \text{otherwise.} \end{cases}$$

We view y_i as a realization of a random variable Y_i that can take the values one and zero with probabilities π_i and $1-\pi_i$, respectively. The distribution of Y_i is called a *Bernoulli* distribution with parameter π_i, and can be written in compact form as

$$\Pr\{Y_i = y_i\} = \pi_i^{y_i}(1-\pi_i)^{1-y_i}, \tag{3.1}$$

for $y_i = 0, 1$. Note that if $y_i = 1$ we obtain π_i, and if $y_i = 0$ we obtain $1-\pi_i$.

It is fairly easy to verify by direct calculation that the expected value and variance of Y_i are

$$\begin{aligned} \mathrm{E}(Y_i) &= \mu_i = \pi_i, \text{ and} \\ \mathrm{var}(Y_i) &= \sigma_i^2 = \pi_i(1-\pi_i). \end{aligned} \tag{3.2}$$

Note that the mean and variance depend on the underlying probability π_i. Any factor that affects the probability will alter not just the mean but also the variance of the observations. This suggest that a linear model that allows

the predictors to affect the mean but assumes that the variance is constant will not be adequate for the analysis of binary data.

Suppose now that the units under study can be classified according to the factors of interest into k groups in such a way that all individuals in a group have identical values of all covariates. In our example, women may be classified into 16 different groups in terms of their age, education and desire for more children. Let n_i denote the number of observations in group i, and let y_i denote the number of units who have the attribute of interest in group i. For example, let

$$y_i = \text{number of women using contraception in group } i.$$

We view y_i as a realization of a random variable Y_i that takes the values $0, 1, \ldots, n_i$. If the n_i observations in each group are *independent*, and they all have the same probability π_i of having the attribute of interest, then the distribution of Y_i is *binomial* with parameters π_i and n_i, which we write

$$Y_i \sim B(n_i, \pi_i).$$

The probability distribution function of Y_i is given by

$$\Pr\{Y_i = y_i\} = \left(\begin{array}{c} n_i \\ y_i \end{array} \right) \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i} \tag{3.3}$$

for $y_i = 0, 1, \ldots, n_i$. Here $\pi_i^{y_i}(1 - \pi_i)^{n_i - y_i}$ is the probability of obtaining y_i successes and $n_i - y_i$ failures in some specific order, and the combinatorial coefficient is the number of ways of obtaining y_i successes in n_i trials.

The mean and variance of Y_i can be shown to be

$$\begin{aligned} \mathrm{E}(Y_i) &= \mu_i = n_i \pi_i, \text{ and} \\ \mathrm{var}(Y_i) &= \sigma_i^2 = n_i \pi_i (1 - \pi_i). \end{aligned} \tag{3.4}$$

The easiest way to obtain this result is as follows. Let Y_{ij} be an indicator variable that takes the values one or zero if the j-th unit in group i is a success or a failure, respectively. Note that Y_{ij} is a Bernoulli random variable with mean and variance as given in Equation 3.2. We can write the number of successes Y_i in group i as a sum of the individual indicator variables, so $Y_i = \sum_j Y_{ij}$. The mean of Y_i is then the sum of the individual means, and by independence, its variance is the sum of the individual variances, leading to the result in Equation 3.4. Note again that the mean and variance depend

on the underlying probability π_i. Any factor that affects this probability will affect both the mean and the variance of the observations.

From a mathematical point of view the grouped data formulation given here is the most general one; it includes individual data as the special case where we have n groups of size one, so $k = n$ and $n_i = 1$ for all i. It also includes as a special case the other extreme where the underlying probability is the same for all individuals and we have a single group, with $k = 1$ and $n_1 = n$. Thus, all we need to consider in terms of estimation and testing is the binomial distribution.

From a practical point of view it is important to note that if the predictors are discrete factors and the outcomes are independent, we can use the Bernoulli distribution for the individual zero-one data or the binomial distribution for grouped data consisting of counts of successes in each group. The two approaches are equivalent, in the sense that they lead to exactly the same likelihood function and therefore the same estimates and standard errors. Working with grouped data when it is possible has the additional advantage that, depending on the size of the groups, it becomes possible to test the goodness of fit of the model. In terms of our example we can work with 16 groups of women (or fewer when we ignore some of the predictors) and obtain exactly the same estimates as we would if we worked with the 1607 individuals.

In Appendix B we show that the binomial distribution belongs to Nelder and Wedderburn's (1972) exponential family, so it fits in our general theoretical framework.

3.1.3 The Logit Transformation

The next step in defining a model for our data concerns the systematic structure. We would like to have the probabilities π_i depend on a vector of observed covariates $\mathbf{x}_i$. The simplest idea would be to let π_i be a linear function of the covariates, say

$$\pi_i = \mathbf{x}_i'\boldsymbol{\beta}, \tag{3.5}$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients. Model 3.5 is sometimes called the *linear probability model*. This model is often estimated from individual data using ordinary least squares (OLS).

One problem with this model is that the probability π_i on the left-hand-side has to be between zero and one, but the linear predictor $\mathbf{x}_i'\boldsymbol{\beta}$ on the right-hand-side can take any real value, so there is no guarantee that the

predicted values will be in the correct range unless complex restrictions are imposed on the coefficients.

A simple solution to this problem is to *transform* the probability to remove the range restrictions, and model the transformation as a linear function of the covariates. We do this in two steps.

First, we move from the probability π_i to the *odds*

$$\text{odds}_i = \frac{\pi_i}{1-\pi_i},$$

defined as the ratio of the probability to its complement, or the ratio of favorable to unfavorable cases. If the probability of an event is a half, the odds are one-to-one or even. If the probability is 1/3, the odds are one-to-two. If the probability is very small, the odds are said to be long. In some contexts the language of odds is more natural than the language of probabilities. In gambling, for example, odds of $1:k$ indicate that the fair payoff for a stake of one is k. The key from our point of view is that the languages are equivalent, i.e. one can easily be translated into the other, but odds can take any positive value and therefore have no ceiling restriction.

Second, we take logarithms, calculating the *logit* or log-odds

$$\eta_i = \text{logit}(\pi_i) = \log \frac{\pi_i}{1-\pi_i}, \tag{3.6}$$

which has the effect of removing the floor restriction. To see this point note that as the probability goes down to zero the odds approach zero and the logit approaches $-\infty$. At the other extreme, as the probability approaches one the odds approach $+\infty$ and so does the logit. Thus, logits map probabilities from the range $(0,1)$ to the entire real line. Note that if the probability is 1/2 the odds are even and the logit is zero. Negative logits represent probabilities below one half and positive logits correspond to probabilities above one half. Figure 3.1 illustrates the logit transformation.

Logits may also be defined in terms of the binomial mean $\mu_i = n_i\pi_i$ as the log of the ratio of expected successes μ_i to expected failures $n_i - \mu_i$. The result is exactly the same because the binomial denominator n_i cancels out when calculating the odds.

In the contraceptive use data there are 507 users of contraception among 1607 women, so we estimate the probability as $507/1607 = 0.316$. The odds are 507/1100 or 0.461 to one, so non-users outnumber users roughly two to one. The logit is $\log(0.461) = -0.775$.

The logit transformation is one-to-one. The inverse transformation is sometimes called the *antilogit*, and allows us to go back from logits to prob-

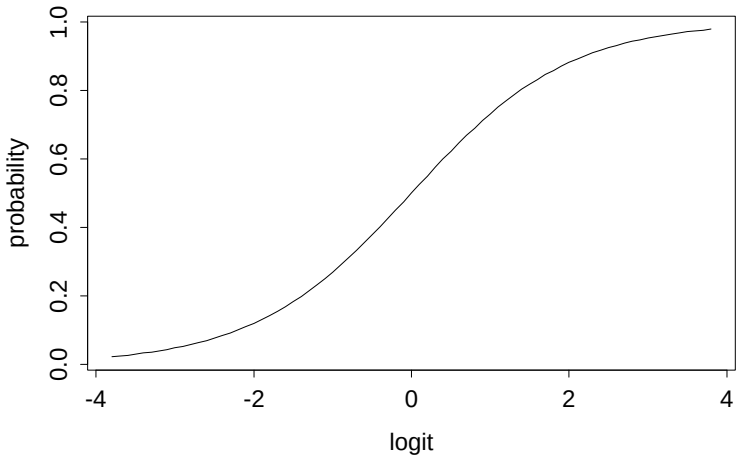

FIGURE 3.1: The Logit Transformation

abilities. Solving for π_i in Equation 3.6 gives

$$\pi_i = \text{logit}^{-1}(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}}. \tag{3.7}$$

In the contraceptive use data the estimated logit was -0.775. Exponentiating this value we obtain odds of $\exp(-0.775) = 0.461$ and from this we obtain a probability of $0.461/(1 + 0.461) = 0.316$.

We are now in a position to define the logistic regression model, by assuming that the *logit* of the probability π_i, rather than the probability itself, follows a linear model.

3.1.4 The Logistic Regression Model

Suppose that we have k independent observations $y_1, \ldots, y_k$, and that the i-th observation can be treated as a realization of a random variable Y_i. We assume that Y_i has a binomial distribution

$$Y_i \sim B(n_i, \pi_i) \tag{3.8}$$

with binomial denominator n_i and probability π_i. With individual data $n_i = 1$ for all i. This defines the stochastic structure of the model.

Suppose further that the *logit* of the underlying probability π_i is a linear function of the predictors

$$\boxed{\text{logit}(\pi_i) = \mathbf{x}_i'\boldsymbol{\beta},} \tag{3.9}$$

where $\mathbf{x}_i$ is a vector of covariates and $\boldsymbol{\beta}$ is a vector of regression coefficients. This defines the systematic structure of the model.

The model defined in Equations 3.8 and 3.9 is a generalized linear model with binomial response and link logit. Note, incidentally, that it is more natural to consider the distribution of the response Y_i than the distribution of the implied error term $Y_i - \mu_i$.

The regression coefficients $\boldsymbol{\beta}$ can be interpreted along the same lines as in linear models, bearing in mind that the left-hand-side is a logit rather than a mean. Thus, β_j represents the change in the *logit* of the probability associated with a unit change in the j-th predictor holding all other predictors constant. While expressing results in the logit scale will be unfamiliar at first, it has the advantage that the model is rather simple in this particular scale.

Exponentiating Equation 3.9 we find that the odds for the i-th unit are given by

$$\frac{\pi_i}{1-\pi_i} = \exp\{\mathbf{x}_i'\boldsymbol{\beta}\}. \tag{3.10}$$

This expression defines a multiplicative model for the odds. For example if we were to change the j-th predictor by one unit while holding all other variables constant, we would multiply the odds by $\exp\{\beta_j\}$. To see this point suppose the linear predictor is $\mathbf{x}_i'\boldsymbol{\beta}$ and we increase x_j by one, to obtain $\mathbf{x}_i'\boldsymbol{\beta} + \beta_j$. Exponentiating we get $\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}$ times $\exp\{\beta_j\}$. Thus, the exponentiated coefficient $\exp\{\beta_j\}$ represents an *odds ratio*. Translating the results into multiplicative effects on the odds, or odds ratios, is often helpful, because we can deal with a more familiar scale while retaining a relatively simple model.

Solving for the probability π_i in the logit model in Equation 3.9 gives the more complicated model

$$\pi_i = \frac{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{1+\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}. \tag{3.11}$$

While the left-hand-side is in the familiar probability scale, the right-hand-side is a non-linear function of the predictors, and there is no simple way to express the effect on the probability of increasing a predictor by one unit while holding the other variables constant. We can obtain an approximate answer by taking derivatives with respect to x_j, which of course makes sense only for continuous predictors. Using the quotient rule we get

$$\frac{\mathrm{d}\pi_i}{\mathrm{d}x_{ij}} = \beta_j \pi_i (1-\pi_i).$$

Thus, the effect of the j-th predictor on the probability π_i depends on the coefficient β_j *and* the value of the probability. Analysts sometimes evaluate this product setting π_i to the sample mean (the proportion of cases with the attribute of interest in the sample). The result approximates the effect of the covariate near the mean of the response.

In the examples that follow we will emphasize working directly in the logit scale, but we will often translate effects into odds ratios to help in interpretation.

Before we leave this topic it may be worth considering the linear probability model of Equation 3.5 one more time. In addition to the fact that the linear predictor $\mathbf{x}_i'\boldsymbol{\beta}$ may yield values outside the $(0, 1)$ range, one should consider whether it is reasonable to assume linear effects on a probability scale that is subject to floor and ceiling effects. An incentive, for example, may increase the probability of taking an action by ten percentage points when the probability is a half, but couldn't possibly have that effect if the baseline probability was 0.95. This suggests that the assumption of a linear effect across the board may not be reasonable.

In contrast, suppose the effect of the incentive is 0.4 in the logit scale, which is equivalent to approximately a 50% increase in the odds of taking the action. If the original probability is a half the logit is zero, and adding 0.4 to the logit gives a probability of 0.6, so the effect is ten percentage points, just as before. If the original probability is 0.95, however, the logit is almost three, and adding 0.4 in the logit scale gives a probability of 0.97, an effect of just two percentage points. An effect that is constant in the logit scale translates into varying effects on the probability scale, adjusting automatically as one approaches the floor of zero or the ceiling of one. This feature of the transformation is clearly seen from Figure 3.1.

3.2 Estimation and Hypothesis Testing

The logistic regression model just developed is a generalized linear model with binomial errors and link logit. We can therefore rely on the general theory developed in Appendix B to obtain estimates of the parameters and to test hypotheses. In this section we summarize the most important results needed in the applications.

3.2.1 Maximum Likelihood Estimation

Although you will probably use a statistical package to compute the estimates, here is a brief description of the underlying procedure. The likelihood

function for n independent binomial observations is a product of densities given by Equation 3.3. Taking logs we find that, except for a constant involving the combinatorial terms, the log-likelihood function is

$$\log L(\boldsymbol{\beta}) = \sum \{y_i \log(\pi_i) + (n_i - y_i) \log(1 - \pi_i)\},$$

where π_i depends on the covariates $\mathbf{x}_i$ and a vector of p parameters $\boldsymbol{\beta}$ through the logit transformation of Equation 3.9.

At this point we could take first and expected second derivatives to obtain the score and information matrix and develop a Fisher scoring procedure for maximizing the log-likelihood. As shown in Appendix B, the procedure is equivalent to iteratively re-weighted least squares (IRLS). Given a current estimate $\hat{\boldsymbol{\beta}}$ of the parameters, we calculate the linear predictor $\hat{\boldsymbol{\eta}} = \mathbf{x}_i'\hat{\boldsymbol{\beta}}$ and the fitted values $\hat{\boldsymbol{\mu}} = \text{logit}^{-1}(\boldsymbol{\eta})$. With these values we calculate the working dependent variable $\mathbf{z}$, which has elements

$$z_i = \hat{\eta}_i + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i(n_i - \hat{\mu}_i)} n_i,$$

where n_i are the binomial denominators. We then regress $\mathbf{z}$ on the covariates calculating the weighted least squares estimate

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z},$$

where $\mathbf{W}$ is a diagonal matrix of weights with entries

$$w_{ii} = \hat{\mu}_i(n_i - \hat{\mu}_i)/n_i.$$

(You may be interested to know that the weight is inversely proportional to the estimated variance of the working dependent variable.) The resulting estimate of $\boldsymbol{\beta}$ is used to obtain improved fitted values and the procedure is iterated to convergence.

Suitable initial values can be obtained by applying the link to the data. To avoid problems with counts of 0 or n_i (which is always the case with individual zero-one data), we calculate empirical logits adding $1/2$ to both the numerator and denominator, i.e. we calculate

$$z_i = \log \frac{y_i + 1/2}{n_i - y_i + 1/2},$$

and then regress this quantity on $\mathbf{x}_i$ to obtain an initial estimate of $\boldsymbol{\beta}$.

The resulting estimate is consistent and its large-sample variance is given by

$$\text{var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \tag{3.12}$$

where $\mathbf{W}$ is the matrix of weights evaluated in the last iteration.

Alternatives to maximum likelihood estimation include weighted least squares, which can be used with grouped data, and a method that minimizes Pearson's chi-squared statistic, which can be used with both grouped and individual data. We will not consider these alternatives further.

3.2.2 Goodness of Fit Statistics

Suppose we have just fitted a model and want to assess how well it fits the data. A measure of discrepancy between observed and fitted values is the *deviance* statistic, which is given by

$$D = 2\sum\{y_i \log(\frac{y_i}{\hat{\mu}_i}) + (n_i - y_i)\log(\frac{n_i - y_i}{n_i - \hat{\mu}_i})\}, \tag{3.13}$$

where y_i is the observed and $\hat{\mu}_i$ is the fitted value for the i-th observation. Note that this statistic is twice a sum of 'observed times log of observed over expected', where the sum is over both successes and failures (i.e. we compare both y_i and $n_i - y_i$ with their expected values). In a perfect fit the ratio observed over expected is one and its logarithm is zero, so the deviance is zero.

In Appendix B we show that this statistic may be constructed as a likelihood ratio test that compares the model of interest with a saturated model that has one parameter for each observation.

With grouped data, the distribution of the deviance statistic as the group sizes $n_i \to \infty$ for all i, converges to a chi-squared distribution with $n - p$ d.f., where n is the number of *groups* and p is the number of parameters in the model, including the constant. Thus, for reasonably large groups, the deviance provides a goodness of fit test for the model. With individual data the distribution of the deviance does not converge to a chi-squared (or any other known) distribution, and cannot be used as a goodness of fit test. We will, however, consider other diagnostic tools that can be used with individual data.

An alternative measure of goodness of fit is *Pearson's chi-squared statistic*, which for binomial data can be written as

$$\chi^2_P = \sum_i \frac{n_i(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i(n_i - \hat{\mu}_i)}. \tag{3.14}$$

Note that each term in the sum is the squared difference between observed and fitted values y_i and $\hat{\mu}_i$, divided by the variance of y_i, which is $\mu_i(n_i -$

$\mu_i)/n_i$, estimated using $\hat{\mu}_i$ for μ_i. This statistic can also be derived as a sum of 'observed minus expected squared over expected', where the sum is over both successes and failures.

With grouped data Pearson's statistic has approximately in large samples a chi-squared distribution with $n-p$ d.f., and is asymptotically equivalent to the deviance or likelihood-ratio chi-squared statistic. The statistic can not be used as a goodness of fit test with individual data, but provides a basis for calculating residuals, as we shall see when we discuss logistic regression diagnostics.

3.2.3 Tests of Hypotheses

Let us consider the problem of testing hypotheses in logit models. As usual, we can calculate Wald tests based on the large-sample distribution of the m.l.e., which is approximately normal with mean $\boldsymbol{\beta}$ and variance-covariance matrix as given in Equation 3.12.

In particular, we can test the hypothesis

$$H_0 : \beta_j = 0$$

concerning the significance of a single coefficient by calculating the ratio of the estimate to its standard error

$$z = \frac{\hat{\beta}_j}{\sqrt{\hat{\text{var}}(\hat{\beta}_j)}}.$$

This statistic has approximately a standard normal distribution in large samples. Alternatively, we can treat the square of this statistic as approximately a chi-squared with one d.f.

The Wald test can be use to calculate a confidence interval for β_j. We can assert with $100(1-\alpha)\%$ confidence that the true parameter lies in the interval with boundaries

$$\hat{\beta}_j \pm z_{1-\alpha/2}\sqrt{\hat{\text{var}}(\hat{\beta}_j)},$$

where $z_{1-\alpha/2}$ is the normal critical value for a two-sided test of size α. Confidence intervals for effects in the logit scale can be translated into confidence intervals for odds ratios by exponentiating the boundaries.

The Wald test can be applied to tests hypotheses concerning several coefficients by calculating the usual quadratic form. This test can also be inverted to obtain confidence regions for vector-value parameters, but we will not consider this extension.

For more general problems we consider the likelihood ratio test. A key to construct these tests is the deviance statistic introduced in the previous subsection. In a nutshell, the likelihood ratio test to compare two nested models is based on the *difference* between their deviances.

To fix ideas, consider partitioning the model matrix and the vector of coefficients into two components

$$\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2) \qquad \text{and} \qquad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}$$

with p_1 and p_2 elements, respectively. Consider testing the hypothesis

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0},$$

that the variables in $\mathbf{X}_2$ have no effect on the response, i.e. the joint significance of the coefficients in $\boldsymbol{\beta}_2$.

Let $D(\mathbf{X}_1)$ denote the deviance of a model that includes only the variables in $\mathbf{X}_1$ and let $D(\mathbf{X}_1 + \mathbf{X}_2)$ denote the deviance of a model that includes all variables in $\mathbf{X}$. Then the difference

$$\chi^2 = D(\mathbf{X}_1) - D(\mathbf{X}_1 + \mathbf{X}_2)$$

has approximately in large samples a chi-squared distribution with p_2 d.f. Note that p_2 is the difference in the number of parameters between the two models being compared.

The deviance plays a role similar to the residual sum of squares. In fact, in Appendix B we show that in models for normally distributed data the deviance *is* the residual sum of squares. Likelihood ratio tests in generalized linear models are based on scaled deviances, obtained by dividing the deviance by a scale factor. In linear models the scale factor was σ^2, and we had to divide the RSS's (or their difference) by an estimate of σ^2 in order to calculate the test criterion. With binomial data the scale factor is one, and there is no need to scale the deviances.

The Pearson chi-squared statistic in the previous subsection, while providing an alternative goodness of fit test for grouped data, cannot be used in general to compare nested models. In particular, differences in deviances have chi-squared distributions but differences in Pearson chi-squared statistics do not. This is the main reason why in statistical modelling we use the deviance or likelihood ratio chi-squared statistic rather than the more traditional Pearson chi-squared of elementary statistics.

3.3 The Comparison of Two Groups

We start our applications of logit regression with the simplest possible example: a two by two table. We study a binary outcome in two groups, and introduce the odds ratio and the logit analogue of the two-sample t test.

3.3.1 A 2-by-2 Table

We will use the contraceptive use data classified by desire for more children, as summarized in Table 3.2

TABLE 3.2: Contraceptive Use by Desire for More Children

Desires	Using	Not Using	All
i	y_i	$n_i - y_i$	n_i
Yes	219	753	972
No	288	347	635
All	507	1100	1607

We treat the counts of users y_i as realizations of independent random variables Y_i having binomial distributions $B(n_i, \pi_i)$ for $i = 1, 2$, and consider models for the logits of the probabilities.

3.3.2 Testing Homogeneity

There are only two possible models we can entertain for these data. The first one is the *null* model. This model assumes homogeneity, so the two groups have the same probability and therefore the same logit

$$\text{logit}(\pi_i) = \eta.$$

The m.l.e. of the common logit is -0.775, which happens to be the logit of the sample proportion $507/1607 = 0.316$. The standard error of the estimate is 0.054. This value can be used to obtain an approximate 95% confidence limit for the logit with boundaries $(-0.880, -0.669)$. Calculating the antilogit of these values, we obtain a 95% confidence interval for the overall probability of using contraception of $(0.293, 0.339)$.

The deviance for the null model happens to be 91.7 on one d.f. (two groups minus one parameter). This value is highly significant, indicating that this model does not fit the data, i.e. the two groups classified by desire for more children do not have the same probability of using contraception.

The value of the deviance is easily verified by hand. The estimated probability of 0.316, applied to the sample sizes in Table 3.2, leads us to expect 306.7 and 200.3 users of contraception in the two groups, and therefore 665.3 and 434.7 non-users . Comparing the observed and expected numbers of users and non-users in the two groups using Equation 3.13 gives 91.7.

You can also compare the observed and expected frequencies using Pearson's chi-squared statistic from Equation 3.14. The result is 92.6 on one d.f., and provides an alternative test of the goodness of fit of the null model.

3.3.3 The Odds Ratio

The other model that we can entertain for the two-by-two table is the *one-factor* model, where we write

$$\text{logit}(\pi_i) = \eta + \alpha_i,$$

where η is an overall logit and α_i is the effect of group i on the logit. Just as in the one-way anova model, we need to introduce a restriction to identify this model. We use the reference cell method, and set $\alpha_1 = 0$. The model can then be written

$$\text{logit}(\pi_i) = \begin{cases} \eta & i = 1 \\ \eta + \alpha_2 & i = 2 \end{cases}$$

so that η becomes the logit of the reference cell, and α_2 is the effect of level two of the factor compared to level one, or more simply the difference in logits between level two and the reference cell. Table 3.3 shows parameter estimates and standard errors for this model.

TABLE 3.3: Parameter Estimates for Logit Model of Contraceptive Use by Desire for More Children

Parameter	Symbol	Estimate	Std. Error	z-ratio
Constant	η	-1.235	0.077	-16.09
Desire	α_2	1.049	0.111	9.48

The estimate of η is, as you might expect, the logit of the observed proportion using contraception among women who desire more children, $\text{logit}(219/972) = -1.235$. The estimate of α_2 is the difference between the logits of the two groups, $\text{logit}(288/635) - \text{logit}(219/972) = 1.049$.

Exponentiating the additive logit model we obtain a multiplicative model for the odds:

$$\frac{\pi_i}{1-\pi_i} = \begin{cases} e^{\eta} & i = 1 \\ e^{\eta}e^{\alpha_2} & i = 2 \end{cases}$$

so that e^{η} becomes the odds for the reference cell and e^{α_2} is the ratio of the odds for level 2 of the factor to the odds for the reference cell. Not surprisingly, e^{α_2} is called the *odds ratio*.

In our example, the effect of 1.049 in the logit scale translates into an odds ratio of 2.85. Thus, the odds of using contraception among women who want no more children are nearly three times as high as the odds for women who desire more children.

From the estimated logit effect of 1.049 and its standard error we can calculate a 95% confidence interval with boundaries $(0.831, 1.267)$. Exponentiating these boundaries we obtain a 95% confidence interval for the odds ratio of $(2.30, 3.55)$. Thus, we conclude with 95% confidence that the odds of using contraception among women who want no more children are between two and three-and-a-half times the corresponding odds for women who want more children.

The estimate of the odds ratio can be calculated directly as the cross-product of the frequencies in the two-by-two table. If we let f_{ij} denote the frequency in cell i, j then the estimated odds ratio is

$$\frac{f_{11}f_{22}}{f_{12}f_{21}}.$$

The deviance of this model is zero, because the model is saturated: it has two parameters to represent two groups, so it has to do a perfect job. The reduction in deviance of 91.7 from the null model down to zero can be interpreted as a test of

$$H_0 : \alpha_2 = 0,$$

the significance of the effect of desire for more children.

An alternative test of this effect is obtained from the m.l.e of 1.049 and its standard error of 0.111, and gives a z-ratio of 9.47. Squaring this value we obtain a chi-squared of 89.8 on one d.f. Note that the Wald test is similar, but not identical, to the likelihood ratio test. Recall that in linear models the two tests were identical. In logit models they are only asymptotically equivalent.

The logit of the observed proportion $p_i = y_i/n_i$ has large-sample variance

$$\text{var}(\text{logit}(p_i)) = \frac{1}{\mu_i} + \frac{1}{n_i - \mu_i},$$

which can be estimated using y_i to estimate μ_i for $i = 1, 2$. Since the two groups are independent samples, the variance of the difference in logits is the sum of the individual variances. You may use these results to verify the Wald test given above.

3.3.4 The Conventional Analysis

It might be instructive to compare the results obtained here with the conventional analysis of this type of data, which focuses on the sample proportions and their difference. In our example, the proportions using contraception are 0.225 among women who want another child and 0.453 among those who do not. The difference of 0.228 has a standard error of 0.024 (calculated using the pooled estimate of the proportion). The corresponding z-ratio is 9.62 and is equivalent to a chi-squared of 92.6 on one d.f.

Note that the result coincides with the Pearson chi-squared statistic testing the goodness of fit of the null model. In fact, Pearson's chi-squared and the conventional test for equality of two proportions are one and the same.

In the case of two samples it is debatable whether the group effect is best measured in terms of a difference in probabilities, the odds-ratio, or even some other measures such as the relative difference proposed by Sheps (1961). For arguments on all sides of this issue see Fleiss (1973).

3.4 The Comparison of Several Groups

Let us take a more general look at logistic regression models with a single predictor by considering the comparison of k groups. This will help us illustrate the logit analogues of one-way analysis of variance and simple linear regression models.

3.4.1 A k-by-Two Table

Consider a cross-tabulation of contraceptive use by age, as summarized in Table 3.4. The structure of the data is the same as in the previous section, except that we now have four groups rather than two.

The analysis of this table proceeds along the same lines as in the two-by-two case. The null model yields exactly the same estimate of the overall logit and its standard error as before. The deviance, however, is now 79.2 on three d.f. This value is highly significant, indicating that the assumption of a common probability of using contraception for the four age groups is not tenable.

TABLE 3.4: Contraceptive Use by Age

Age i	Using y_i	Not Using $n_i - y_i$	Total n_i
<25	72	325	397
25–29	105	299	404
30–39	237	375	612
40–49	93	101	194
Total	507	1100	1607

3.4.2 The One-Factor Model

Consider now a one-factor model, where we allow each group or level of the discrete factor to have its own logit. We write the model as

$$\text{logit}(\pi_i) = \eta + \alpha_i.$$

To avoid redundancy we adopt the reference cell method and set $\alpha_1 = 0$, as before. Then η is the logit of the reference group, and α_i measures the difference in logits between level i of the factor and the reference level. This model is exactly analogous to an analysis of variance model. The model matrix $\mathbf{X}$ consists of a column of ones representing the constant and $k-1$ columns of dummy variables representing levels two to k of the factor.

Fitting this model to Table 3.4 leads to the parameter estimates and standard errors in Table 3.5. The deviance for this model is of course zero because the model is saturated: it uses four parameters to model four groups.

TABLE 3.5: Estimates and Standard Errors for Logit Model of Contraceptive Use by Age in Groups

Parameter		Symbol	Estimate	Std. Error	z-ratio
Constant		η	−1.507	0.130	−11.57
Age	25–29	α_2	0.461	0.173	2.67
	30–39	α_3	1.048	0.154	6.79
	40–49	α_4	1.425	0.194	7.35

The baseline logit of −1.51 for women under age 25 corresponds to odds of 0.22. Exponentiating the age coefficients we obtain odds ratios of 1.59, 2.85 and 4.16. Thus, the odds of using contraception increase by 59% and

185% as we move to ages 25–29 and 30–39, and are quadrupled for ages 40–49, all compared to women under age 25.

All of these estimates can be obtained directly from the frequencies in Table 3.4 in terms of the logits of the observed proportions. For example the constant is $\text{logit}(72/397) = -1.507$, and the effect for women 25–29 is logit(105/404) minus the constant.

To test the hypothesis of no age effects we can compare this model with the null model. Since the present model is saturated, the difference in deviances is exactly the same as the deviance of the null model, which was 79.2 on three d.f. and is highly significant. An alternative test of

$$H_0 : \alpha_2 = \alpha_3 = \alpha_4 = 0$$

is based on the estimates and their variance-covariance matrix. Let $\boldsymbol{\alpha} = (\alpha_2, \alpha_3, \alpha_4)'$. Then

$$\hat{\boldsymbol{\alpha}} = \begin{pmatrix} 0.461 \\ 1.048 \\ 1.425 \end{pmatrix} \qquad \text{and} \qquad \hat{\text{var}}(\hat{\boldsymbol{\alpha}}) = \begin{pmatrix} 0.030 & 0.017 & 0.017 \\ 0.017 & 0.024 & 0.017 \\ 0.017 & 0.017 & 0.038 \end{pmatrix},$$

and the Wald statistic is

$$W = \hat{\boldsymbol{\alpha}}' \, \hat{\text{var}}^{-1}(\hat{\boldsymbol{\alpha}}) \, \hat{\boldsymbol{\alpha}} = 74.4$$

on three d.f. Again, the Wald test gives results similar to the likelihood ratio test.

3.4.3 A One-Variate Model

Note that the estimated logits in Table 3.5 (and therefore the odds and probabilities) increase monotonically with age. In fact, the logits seem to increase by approximately the same amount as we move from one age group to the next. This suggests that the effect of age may actually be linear in the logit scale.

To explore this idea we treat age as a variate rather than a factor. A thorough exploration would use the individual data with age in single years (or equivalently, a 35 by two table of contraceptive use by age in single years from 15 to 49). However, we can obtain a quick idea of whether the model would be adequate by keeping age grouped into four categories but representing these by the *mid-points* of the age groups. We therefore consider a model analogous to simple linear regression, where

$$\text{logit}(\pi_i) = \alpha + \beta x_i,$$

where x_i takes the values 20, 27.5, 35 and 45, respectively, for the four age groups. This model fits into our general framework, and corresponds to the special case where the model matrix $\mathbf{X}$ has two columns, a column of ones representing the constant and a column with the mid-points of the age groups, representing the linear effect of age.

Fitting this model gives a deviance of 2.40 on two d.f. , which indicates a very good fit. The parameter estimates and standard errors are shown in Table 3.6. Incidentally, there is no explicit formula for the estimates of the constant and slope in this model, so we must rely on iterative procedures to obtain the estimates.

TABLE 3.6: Estimates and Standard Errors for Logit Model of Contraceptive Use with a Linear Effect of Age

Parameter	Symbol	Estimate	Std. Error	z-ratio
Constant	α	-2.673	0.233	-11.46
Age (linear)	β	0.061	0.007	8.54

The slope indicates that the logit of the probability of using contraception increases 0.061 for every year of age. Exponentiating this value we note that the odds of using contraception are multiplied by 1.063—that is, increase 6.3%—for every year of age. Note, by the way, that $e^{\beta} \approx 1 + \beta$ for small $|\beta|$. Thus, when the logit coefficient is small in magnitude, 100β provides a quick approximation to the percent change in the odds associated with a unit change in the predictor. In this example the effect is 6.3% and the approximation is 6.1%.

To test the significance of the slope we can use the Wald test, which gives a z statistic of 8.54 or equivalently a chi-squared of 73.9 on one d.f. Alternatively, we can construct a likelihood ratio test by comparing this model with the null model. The difference in deviances is 76.8 on one d.f. Comparing these results with those in the previous subsection shows that we have captured most of the age effect using a single degree of freedom.

Adding the estimated constant to the product of the slope by the mid-points of the age groups gives estimated logits at each age, and these may be compared with the logits of the observed proportions using contraception. The results of this exercise appear in Figure 3.2. The visual impression of the graph confirms that the fit is quite good. In this example the assumption of linear effects on the logit scale leads to a simple and parsimonious model. It would probably be worthwhile to re-estimate this model using the individual

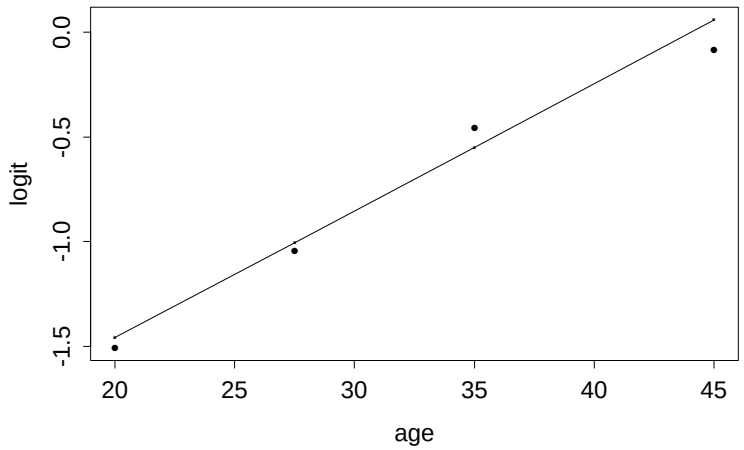

FIGURE 3.2: Observed and Fitted Logits for Model of Contraceptive Use with a Linear Effect of Age

ages.

3.5 Models With Two Predictors

We now consider models involving two predictors, and discuss the binary data analogues of two-way analysis of variance, multiple regression with dummy variables, and analysis of covariance models. An important element of the discussion concerns the key concepts of main effects and interactions.

3.5.1 Age and Preferences

Consider the distribution of contraceptive use by age and desire for more children, as summarized in Table 3.7. We have a total of eight groups, which will be indexed by a pair of subscripts i, j, with $i = 1, 2, 3, 4$ referring to the four age groups and $j = 1, 2$ denoting the two categories of desire for more children. We let y_{ij} denote the number of women using contraception and n_{ij} the total number of women in age group i and category j of desire for more children.

We now analyze these data under the usual assumption of a binomial error structure, so the y_{ij} are viewed as realizations of independent random variables $Y_{ij} \sim B(n_{ij}, \pi_{ij})$.

TABLE 3.7: Contraceptive Use by Age and Desire for More Children

Age i	Desires j	Using y_{ij}	Not Using $n_{ij} - y_{ij}$	All n_{ij}
<25	Yes	58	265	323
	No	14	60	74
25–29	Yes	68	215	283
	No	37	84	121
30–39	Yes	79	230	309
	No	158	145	303
40–49	Yes	14	43	57
	No	79	58	137
Total		507	1100	1607

3.5.2 The Deviance Table

There are five basic models of interest for the systematic structure of these data, ranging from the null to the saturated model. These models are listed in Table 3.8, which includes the name of the model, a descriptive notation, the formula for the linear predictor, the deviance or goodness of fit likelihood ratio chi-squared statistic, and the degrees of freedom.

Note first that the null model does not fit the data: the deviance of 145.7 on 7 d.f. is much greater than 14.1, the 95-th percentile of the chi-squared distribution with 7 d.f. This result is not surprising, since we already knew that contraceptive use depends on desire for more children and varies by age.

TABLE 3.8: Deviance Table for Models of Contraceptive Use by Age (Grouped) and Desire for More Children

Model	Notation	$\text{logit}(\pi_{ij})$	Deviance	d.f.
Null	ϕ	η	145.7	7
Age	A	$\eta + \alpha_i$	66.5	4
Desire	D	$\eta + \beta_j$	54.0	6
Additive	$A + D$	$\eta + \alpha_i + \beta_j$	16.8	3
Saturated	AD	$\eta + \alpha_i + \beta_j + (\alpha\beta)_{ij}$	0	0

Introducing age in the model reduces the deviance to 66.5 on four d.f. The difference in deviances between the null model and the age model provides a test for the *gross* effect of age. The difference is 79.2 on three d.f.,

and is highly significant. This value is exactly the same that we obtained in the previous section, when we tested for an age effect using the data classified by age only. Moreover, the estimated age effects based on fitting the age model to the three-way classification in Table 3.7 would be exactly the same as those estimated in the previous section, and have the property of reproducing exactly the proportions using contraception in each age group.

This equivalence illustrate an important property of binomial models. All information concerning the gross effect of age on contraceptive use is contained in the marginal distribution of contraceptive use by age. We can work with the data classified by age only, by age and desire for more children, by age, education and desire for more children, or even with the individual data. In all cases the estimated effects, standard errors, and likelihood ratio tests based on differences between deviances will be the same.

The deviances themselves will vary, however, because they depend on the context. In the previous section the deviance of the age model was zero, because treating age as a factor reproduces exactly the proportions using contraception by age. In this section the deviance of the age model is 66.5 on four d.f. and is highly significant, because the age model does not reproduce well the table of contraceptive use by both age and preferences. In both cases, however, the difference in deviances between the age model and the null model is 79.2 on three d.f.

The next model in Table 3.8 is the model with a main effect of desire for more children, and has a deviance of 54.0 on six d.f. Comparison of this value with the deviance of the null model shows a gain of 97.1 at the expense of one d.f., indicating a highly significant *gross* effect of desire for more children. This is, of course, the same result that we obtained in Section 3.3, when we first looked at contraceptive use by desire for more children. Note also that this model does not fit the data, as it own deviance is highly significant.

The fact that the effect of desire for more children has a chi-squared statistic of 91.7 with only one d.f., whereas age gives 79.2 on three d.f., suggests that desire for more children has a stronger effect on contraceptive use than age does. Note, however, that the comparison is informal; the models are not nested, and therefore we cannot construct a significance test from their deviances.

3.5.3 The Additive Model

Consider now the two-factor additive model, denoted $A + D$ in Table 3.8. In this model the logit of the probability of using contraception in age group i

and in category j of desire for more children is

$$\text{logit}(\pi_{ij}) = \eta + \alpha_i + \beta_j,$$

where η is a constant, the α_i are age effects and the β_j are effects of desire for more children. To avoid redundant parameters we adopt the reference cell method and set $\alpha_1 = \beta_1 = 0$. The parameters may then be interpreted as follows:

- η is the logit of the probability of using contraception for women under 25 who want more children, who serve as the reference cell,
- α_i for $i = 2, 3, 4$ represents the *net* effect of ages 25–29, 30–39 and 40–49, compared to women under age 25 in the same category of desire for more children,
- β_2 represents the *net* effect of wanting no more children, compared to women who want more children in the same age group.

The model is additive in the logit scale, in the usual sense that the effect of one variable does not depend on the value of the other. For example, the effect of desiring no more children is β_2 in all four age groups. (This assumption must obviously be tested, and we shall see that it is not consistent with the data.)

The deviance of the additive model is 16.8 on three d.f. With this value we can calculate three different tests of interest, all of which involve comparisons between nested models.

- As we move from model D to $A + D$ the deviance decreases by 37.2 while we lose three d.f. This statistic tests the hypothesis $H_0 : \alpha_i = 0$ for all i, concerning the *net* effect of age after adjusting for desire for more children, and is highly significant.
- As we move from model A to $A + D$ we reduce the deviance by 49.7 at the expense of one d.f. This chi-squared statistic tests the hypothesis $H_0 : \beta_2 = 0$ concerning the *net* effect of desire for more children after adjusting for age. This value is highly significant, so we reject the hypothesis of no net effects.
- Finally, the deviance of 16.8 on three d.f. is a measure of goodness of fit of the additive model: it compares this model with the saturated model, which adds an interaction between the two factors. Since the deviance exceeds 11.3, the one-percent critical value in the chi-squared

distribution for three d.f., we conclude that the additive model fails to fit the data.

Table 3.9 shows parameter estimates for the additive model. We show briefly how they would be interpreted, although we have evidence that the additive model does not fit the data.

TABLE 3.9: Parameter Estimates for Additive Logit Model of Contraceptive Use by Age (Grouped) and Desire for Children

Parameter		Symbol	Estimate	Std. Error	z-ratio
Constant		η	−1.694	0.135	−12.53
Age	25–29	α_2	0.368	0.175	2.10
	30–39	α_3	0.808	0.160	5.06
	40–49	α_4	1.023	0.204	5.01
Desire	No	β_2	0.824	0.117	7.04

The estimates of the α_j's show a monotonic effect of age on contraceptive use. Although there is evidence that this effect may vary depending on whether women desire more children, on average the odds of using contraception among women age 40 or higher are nearly three times the corresponding odds among women under age 25 in the same category of desire for another child.

Similarly, the estimate of β_2 shows a strong effect of wanting no more children. Although there is evidence that this effect may depend on the woman's age, on average the odds of using contraception among women who desire no more children are more than double the corresponding odds among women in the same age group who desire another child.

3.5.4 A Model With Interactions

We now consider a model which includes an interaction of age and desire for more children, denoted AD in Table 3.8. The model is

$$\text{logit}(\pi_{ij}) = \eta + \alpha_i + \beta_j + (\alpha\beta)_{ij},$$

where η is a constant, the α_i and β_j are the main effects of age and desire, and $(\alpha\beta)_{ij}$ is the interaction effect. To avoid redundancies we follow the reference cell method and set to zero all parameters involving the first cell, so that $\alpha_1 = \beta_1 = 0$, $(\alpha\beta)_{1j} = 0$ for all j and $(\alpha\beta)_{i1} = 0$ for all i. The remaining parameters may be interpreted as follows:

η is the logit of the reference group: women under age 25 who desire more children.

α_i for $i = 2, 3, 4$ are the effects of the age groups 25–29, 30–39 and 40–49, compared to ages under 25, for women who want another child.

β_2 is the effect of desiring no more children, compared to wanting another child, for women under age 25.

$(\alpha\beta)_{i2}$ for $i = 2, 3, 4$ is the *additional* effect of desiring no more children, compared to wanting another child, for women in age group i rather than under age 25. (This parameter is also the *additional* effect of age group i, compared to ages under 25, for women who desire no more children rather than those who want more.)

One way to simplify the presentation of results involving interactions is to combine the interaction terms with one of the main effects, and present them as effects of one factor within categories or levels of the other. In our example, we can combine the interactions $(\alpha\beta)_{i2}$ with the main effects of desire β_2, so that

$\beta_2 + (\alpha\beta)_{i2}$ is the effect of desiring no more children, compared to wanting another child, for women in age group i.

Of course, we could also combine the interactions with the main effects of age, and speak of age effects which are specific to women in each category of desire for more children. The two formulations are statistically equivalent, but the one chosen here seems demographically more sensible.

To obtain estimates based on this parameterization of the model we have to define the columns of the model matrix as follows. Let a_i be a dummy variable representing age group i, for $i = 2, 3, 4$, and let d take the value one for women who want no more children and zero otherwise. Then the model matrix $\mathbf{X}$ should have a column of ones to represent the constant or reference cell, the age dummies a_2, a_3 and a_4 to represent the age effects for women in the reference cell, and then the dummy d and the products a_2d, a_3d and a_4d, to represent the effect of wanting no more children at ages < 25, 25–29, 30–39 and 40–49, respectively. The resulting estimates and standard errors are shown in Table 3.10.

The results indicate that contraceptive use among women who desire more children varies little by age, increasing up to age 35–39 and then declining somewhat. On the other hand, the effect of wanting no more children

TABLE 3.10: Parameter Estimates for Model of Contraceptive Use With an Interaction Between Age (Grouped) and Desire for More Children

Parameter		Estimate	Std. Error	z-ratio
Constant		−1.519	0.145	−10.481
Age	25–29	0.368	0.201	1.832
	30–39	0.451	0.195	2.311
	40–49	0.397	0.340	1.168
Desires	<25	0.064	0.330	0.194
No More	25–29	0.331	0.241	1.372
at Age	30–39	1.154	0.174	6.640
	40–49	1.431	0.353	4.057

increases dramatically with age, from no effect among women below age 25 to an odds ratio of 4.18 at ages 40–49. Thus, in the older cohort the odds of using contraception among women who want no more children are four times the corresponding odds among women who desire more children. The results can also be summarized by noting that contraceptive use for spacing (i.e. among women who desire more children) does not vary much by age, but contraceptive use for limiting fertility (i.e among women who want no more children) increases sharply with age.

3.5.5 Analysis of Covariance Models

Since the model with an age by desire interaction is saturated, we have essentially reproduced the observed data. We now consider whether we could attain a more parsimonious fit by treating age as a variate and desire for more children as a factor, in the spirit of covariance analysis models.

Table 3.11 shows deviances for three models that include a linear effect of age using, as before, the midpoints of the age groups. To emphasize this point we use X rather than A to denote age.

The first model assumes that the logits are linear functions of age. This model fails to fit the data, which is not surprising because it ignores desire for more children, a factor that has a large effect on contraceptive use.

The next model, denoted $X + D$, is analogous to the two-factor additive model. It allows for an effect of desire for more children which is the same at all ages. This common effect is modelled by allowing each category of desire for more children to have its own constant, and results in two parallel lines. The common slope is the effect of age within categories of desire for

TABLE 3.11: Deviance Table for Models of Contraceptive Use by Age (Linear) and Desire for More Children

Model	Notation	$\text{logit}(\pi_{ij})$	Deviance	d.f.
One Line	X	$\alpha + \beta x_i$	68.88	6
Parallel Lines	$X + D$	$\alpha_j + \beta x_i$	18.99	5
Two Lines	XD	$\alpha_j + \beta_j x_i$	9.14	4

more children. The reduction in deviance of 39.9 on one d.f. indicates that desire for no more children has a strong effect on contraceptive use after controlling for a linear effect of age. However, the attained deviance of 19.0 on five d.f. is significant, indicating that the assumption of two parallel lines is not consistent with the data.

The last model in the table, denoted XD, includes an interaction between the linear effect of age and desire, and thus allows the effect of desire for more children to vary by age. This variation is modelled by allowing each category of desire for more children to have its own slope in addition to its own constant, and results in two regression lines. The reduction in deviance of 9.9 on one d.f. is a test of the hypothesis of parallelism or common slope $H_0 : \beta_1 = \beta_2$, which is rejected with a P-value of 0.002. The model deviance of 9.14 on four d.f. is just below the five percent critical value of the chi-squared distribution with four d.f., which is 9.49. Thus, we have no evidence against the assumption of two straight lines.

Before we present parameter estimates we need to discuss briefly the choice of parameterization. Direct application of the reference cell method leads us to use four variables: a dummy variable always equal to one, a variable x with the mid-points of the age groups, a dummy variable d which takes the value one for women who want no more children, and a variable dx equal to the product of this dummy by the mid-points of the age groups. This choice leads to parameters representing the constant and slope for women who want another child, and parameters representing the *difference* in constants and slopes for women who want no more children.

An alternative is to simply report the constants and slopes for the two groups defined by desire for more children. This parameterization can be easily obtained by omitting the constant and using the following four variables: d and $1 - d$ to represent the two constants and dx and $(1 - d)x$ to represent the two slopes. One could, of course, obtain the constant and slope for women who want no more children from the previous parameterization

simply by adding the main effect and the interaction. The simplest way to obtain the standard errors, however, is to change parameterization.

In both cases the constants represent effects at age zero and are not very meaningful. To obtain parameters that are more directly interpretable, we can center age around the sample mean, which is 30.6 years. Table 3.12 shows parameter estimates obtained under the two parameterizations discussed above, using the mid-points of the age groups minus the mean.

TABLE 3.12: Parameter Estimates for Model of Contraceptive Use With an Interaction Between Age (Linear) and Desire for More Children

Desire	Age	Symbol	Estimate	Std. Error	z-ratio
More	Constant	α_1	−1.1944	0.0786	−15.20
	Slope	β_1	0.0218	0.0104	2.11
No More	Constant	α_2	−0.4369	0.0931	−4.69
	Slope	β_2	0.0698	0.0114	6.10
Difference	Constant	$\alpha_2 - \alpha_1$	0.7575	0.1218	6.22
	Slope	$\beta_2 - \beta_1$	0.0480	0.0154	3.11

Thus, we find that contraceptive use increases with age, but at a faster rate among women who want no more children. The estimated slopes correspond to increases in the odds of two and seven percent per year of age for women who want and do not want more children, respectively. The difference of the slopes is significant by a likelihood ratio test or by Wald's test, with a z-ratio of 3.11.

Similarly, the effect of wanting no more children increases with age. The odds ratio around age 30.6—which we obtain by exponentiating the difference in constants–is 2.13, so not wanting more children at this age is associated with a doubling of the odds of using contraception. The difference in slopes of 0.048 indicates that this differential increases five percent per year of age.

The parameter estimates in Table 3.12 may be used to produce fitted logits for each age group and category of desire for more children. In turn, these can be compared with the empirical logits for the original eight groups, to obtain a visual impression of the nature of the relationships studied and the quality of the fit. The comparison appears in Figure 3.3, with the solid line representing the linear age effects (the dotted lines are discussed below). The graph shows clearly how the effect of wanting no more children increases with age (or, alternatively, how age has much stronger effects among limiters

than among spacers).

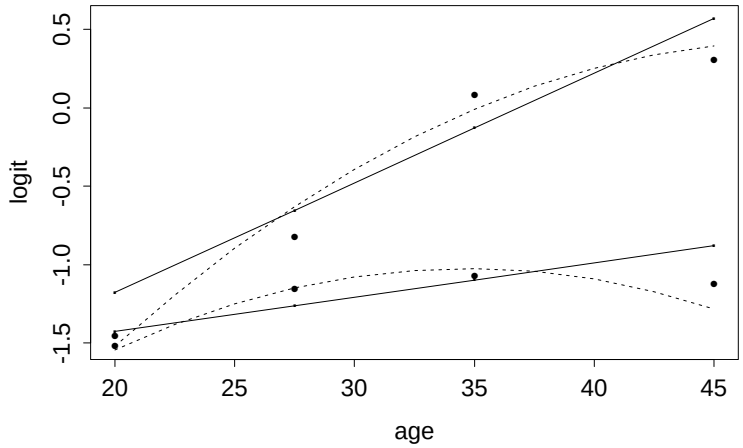

FIGURE 3.3: Observed and Fitted Logits for Models of Contraceptive Use With Effects of Age (Linear and Quadratic), Desire for More Children and a Linear Age by Desire Interaction.

The graph also shows that the assumption of linearity of age effects, while providing a reasonably parsimonious description of the data, is somewhat suspect, particularly at higher ages. We can improve the fit by adding higher-order terms on age. In particular

- Introducing a quadratic term on age yields an excellent fit, with a deviance of 2.34 on three d.f. This model consists of two parabolas, one for each category of desire for more children, but with the same curvature.

- Adding a quadratic age by desire interaction further reduces the deviance to 1.74 on two d.f. This model allows for two separate parabolas tracing contraceptive use by age, one for each category of desire.

Although the linear model passes the goodness of fit test, the fact that we can reduce the deviance by 6.79 at the expense of one d.f. indicates significant curvature. The dotted line in Figure 3.3 shows the intermediate model, where the curvature by age is the same for the two groups. While the fit is much better, the overall substantive conclusions do not change.

3.6 Multi-factor Models: Model Selection

Let us consider a full analysis of the contraceptive use data in Table 3.1, including all three predictors: age, education and desire for more children.

We use three subscripts to reflect the structure of the data, so π_{ijk} is the probability of using contraception in the (i, j, k)-th group, where $i = 1, 2, 3, 4$ indexes the age groups, $j = 1, 2$ the levels of education and $k = 1, 2$ the categories of desire for more children.

3.6.1 Deviances for One and Two-Factor Models

There are 19 basic models of interest for these data, which are listed for completeness in Table 3.13. Not all of these models would be of interest in any given analysis. The table shows the model in abbreviated notation, the formula for the linear predictor, the deviance and its degrees of freedom.

Note first that the null model does not fit the data. The assumption of a common probability of using contraception for all 16 groups of women is clearly untenable.

Next in the table we find the three possible one-factor models. Comparison of these models with the null model provides evidence of significant *gross* effects of age and desire for more children, but not of education. The likelihood ratio chi-squared tests are 91.7 on one d.f. for desire, 79.2 on three d.f. for age, and 0.7 on one d.f. for education.

Proceeding down the table we find the six possible two-factor models, starting with the additive ones. Here we find evidence of significant *net effects* of age and desire for more children after controlling for one other factor. For example the test for an effect of desire net of age is a chi-squared of 49.7 on one d.f., obtained by comparing the additive model $A + D$ on age and desire the one-factor model A with age alone. Education has a significant effect net of age, but not net of desire for more children. For example the test for the net effect of education controlling for age is 6.2 on one d.f., and follows from the comparison of the $A + E$ model with A. None of the additive models fits the data, but the closest one to a reasonable fit is $A + D$.

Next come the models involving *interactions* between two factors. We use the notation ED to denote the model with the main effects of E and D as well as the $E \times D$ interaction. Comparing each of these models with the corresponding additive model on the same two factors we obtain a test of the interaction effect. For example comparing the model ED with the additive model $E + D$ we can test whether the effect of desire for more children varies

TABLE 3.13: Deviance Table for Logit Models of Contraceptive Use by Age, Education and Desire for More Children

Model	$\text{logit}(\pi_{ijk})$	Dev.	d.f.
Null	η	165.77	15
One Factor			
*A*ge	$\eta+\alpha_i$	86.58	12
*E*ducation	$\eta+\beta_j$	165.07	14
*D*esire	$\eta+\gamma_k$	74.10	14
Two Factors			
$A+E$	$\eta+\alpha_i+\beta_j$	80.42	11
$A+D$	$\eta+\alpha_i+\gamma_k$	36.89	11
$E+D$	$\eta+\beta_j+\gamma_k$	73.87	13
AE	$\eta+\alpha_i+\beta_j+(\alpha\beta)_{ij}$	73.03	8
AD	$\eta+\alpha_i+\gamma_k+(\alpha\gamma)_{ik}$	20.10	8
ED	$\eta+\beta_j+\gamma_k+(\beta\gamma)_{jk}$	67.64	12
Three Factors			
$A+E+D$	$\eta+\alpha_i+\beta_j+\gamma_k$	29.92	10
$AE+D$	$\eta+\alpha_i+\beta_j+\gamma_k+(\alpha\beta)_{ij}$	23.15	7
$AD+E$	$\eta+\alpha_i+\beta_j+\gamma_k+(\alpha\gamma)_{ik}$	12.63	7
$A+ED$	$\eta+\alpha_i+\beta_j+\gamma_k+(\beta\gamma)_{jk}$	23.02	9
$AE+AD$	$\eta+\alpha_i+\beta_j+\gamma_k+(\alpha\beta)_{ij}+(\alpha\gamma)_{ik}$	5.80	4
$AE+ED$	$\eta+\alpha_i+\beta_j+\gamma_k+(\alpha\beta)_{ij}+(\beta\gamma)_{jk}$	13.76	6
$AD+ED$	$\eta+\alpha_i+\beta_j+\gamma_k+(\alpha\gamma)_{ik}+(\beta\gamma)_{jk}$	10.82	6
$AE+AD+ED$	$\eta+\alpha_i+\beta_j+\gamma_k+(\alpha\beta)_{ij}+(\alpha\gamma)_{ik}+(\beta\gamma)_{jk}$	2.44	3

with education. Making these comparisons we find evidence of interactions between age and desire for more children ($\chi^2 = 16.8$ on three d.f.), and between education and desire for more children ($\chi^2 = 6.23$ on one d.f.), but not between age and education ($\chi^2 = 7.39$ on three d.f.).

All of the results described so far could be obtained from two-dimensional tables of the type analyzed in the previous sections. The new results begin

to appear as we consider the nine possible three-factor models.

3.6.2 Deviances for Three-Factor Models

The first entry is the *additive* model $A+E+D$, with a deviance of 29.9 on ten d.f. This value represents a significant improvement over any of the additive models on two factors. Thus, we have evidence that there are significant net effects of age, education and desire for more children, considering each factor after controlling the other two. For example the test for a net effect of education controlling the other two variables compares the three-factor additive model $A+E+D$ with the model without education, namely $A+D$. The difference of 6.97 on one d.f. is significant, with a P-value of 0.008. However, the three-factor additive model does not fit the data.

The next step is to add *one interaction* between two of the factors. For example the model $AE + D$ includes the main effects of A, E and D and the $A \times E$ interaction. The interactions of desire for more children with age and with education produce significant gains over the additive model ($\chi^2 = 17.3$ on three d.f. and $\chi^2 = 6.90$ on one d.f., respectively), whereas the interaction between age and education is not significant ($\chi^2 = 6.77$ with three d.f.). These tests for interactions differ from those based on two-factor models in that they take into account the third factor. The best of these models is clearly the one with an interaction between age and desire for more children, $AD+E$. This is also the first model in our list that actually passes the goodness of fit test, with a deviance of 12.6 on seven d.f.

Does this mean that we can stop our search for an adequate model? Unfortunately, it does not. The goodness of fit test is a joint test for all terms omitted in the model. In this case we are testing for the AE, ED and AED interactions simultaneously, a total of seven parameters. This type of omnibus test lacks power against specific alternatives. It is possible that one of the omitted terms (or perhaps some particular contrast) would be significant by itself, but its effect may not stand out in the aggregate. At issue is whether the remaining deviance of 12.6 is spread out uniformly over the remaining d.f. or is concentrated in a few d.f. If you wanted to be absolutely sure of not missing anything you might want to aim for a deviance below 3.84, which is the five percent critical value for one d.f., but this strategy would lead to over-fitting if followed blindly.

Let us consider the models involving *two interactions* between two factors, of which there are three. Since the AD interaction seemed important we restrict attention to models that include this term, so we start from $AD + E$, the best model so far. Adding the age by education interaction

AE to this model reduces the deviance by 6.83 at the expense of three d.f. A formal test concludes that this interaction is not significant. If we add instead the education by desire interaction ED we reduce the deviance by only 1.81 at the expense of one d.f. This interaction is clearly not significant. A model-building strategy based on *forward selection* of variables would stop here and choose $AD + E$ as the best model on grounds of parsimony and goodness of fit.

An alternative approach is to start with the saturated model and impose progressive simplification. Deleting the *three-factor interaction* yields the model $AE + AD + ED$ with three two-factor interactions, which fits the data rather well, with a deviance of just 2.44 on three d.f. If we were to delete the AD interaction the deviance would rise by 11.32 on three d.f., a significant loss. Similarly, removing the AE interaction would incur a significant loss of 8.38 on 3 d.f. We can, however, drop the ED interaction with a non-significant increase in deviance of 3.36 on one d.f. At this point we can also eliminate the AE interaction, which is no longer significant, with a further loss of 6.83 on three d.f. Thus, a *backward elimination* strategy ends up choosing the same model as forward selection.

Although you may find these results reassuring, there is a fact that both approaches overlook: the AE and DE interactions are jointly significant! The change in deviance as we move from $AD+E$ to the model with three two-factor interactions is 10.2 on four d.f., and exceeds (although not by much) the five percent critical value of 9.5. This result indicates that we need to consider the more complicated model with all three two-factor interactions. Before we do that, however, we need to discuss parameter estimates for selected models.

3.6.3 The Additive Model: Gross and Net Effects

Consider first Table 3.14, where we adopt an approach similar to multiple classification analysis to compare the gross and net effects of all three factors. We use the reference cell method, and include the omitted category for each factor (with a dash where the estimated effect would be) to help the reader identify the baseline.

The gross or unadjusted effects are based on the single-factor models A, E and D. These effects represent overall differences between levels of each factor, and as such they have descriptive value even if the one-factor models do not tell the whole story. The results can easily be translated into odds ratios. For example not wanting another child is associated with an increase in the odds of using contraception of 185%. Having upper primary or higher

TABLE 3.14: Gross and Net Effects of Age, Education and Desire for More Children on Current Use of Contraception

Variable and category		Gross effect	Net effect
Constant		–	−1.966
Age	<25	–	–
	25–29	0.461	0.389
	30–39	1.048	0.909
	40–49	1.425	1.189
Education			
	Lower	–	–
	Upper	-0.093	0.325
Desires More			
	Yes	–	–
	No	1.049	0.833

education rather than lower primary or less appears to reduce the odds of using contraception by almost 10%.

The net or adjusted effects are based on the three-factor additive model $A+E+D$. This model assumes that the effect of each factor is the same for all categories of the others. We know, however, that this is not the case—particularly with desire for more children, which has an effect that varies by age—so we have to interpret the results carefully. The net effect of desire for more children shown in Table 3.14 represents an average effect across all age groups and may not be representative of the effect at any particular age. Having said that, we note that desire for no more children has an important effect net of age and education: on the average, it is associated with an increase in the odds of using contraception of 130%.

The result for education is particularly interesting. Having upper primary or higher education is associated with an increase in the odds or using contraception of 38%, compared to having lower primary or less, after we control for age and desire for more children. The gross effect was close to zero. To understand this result bear in mind that contraceptive use in Fiji occurs mostly among older women who want no more children. Education has no effect when considered by itself because in Fiji more educated women are likely to be younger than less educated women, and thus at a stage of their lives when they are less likely to have reached their desired family size,

even though they may want fewer children. Once we adjust for their age, calculating the net effect, we obtain the expected association. In this example age is said to act as a *suppressor* variable, masking the association between education and contraceptive use.

We could easily add columns to Table 3.14 to trace the effects of one factor after controlling for one or both of the other factors. We could, for example, examine the effect of education adjusted for age, the effect adjusted for desire for more children, and finally the effect adjusted for both factors. This type of analysis can yield useful insights into the confounding influences of other variables.

3.6.4 The Model with One Interaction Effect

Let us now examine parameter estimates for the model with an age by desire for more children interaction $AD + E$, where

$$\text{logit}(\pi_{ijk}) = \eta + \alpha_i + \beta_j + \gamma_j + (\alpha\gamma)_{ik}.$$

The parameter estimates depend on the restrictions used in estimation. We use the reference cell method, so that $\alpha_1 = \beta_1 = \gamma_1 = 0$, and $(\alpha\gamma)_{ik} = 0$ when either $i = 1$ or $k = 1$.

In this model η is the logit of the probability of using contraception in the reference cell, that is, for women under 25 with lower primary or less education who want another child. On the other hand β_2 is the effect of upper primary or higher education, compared to lower primary or less, for women in any age group or category of desire for another child. The presence of an interaction makes interpretation of the estimates for age and desire somewhat more involved:

α_i represents the effect of age group i, compared to age < 25, for women who want more children.

γ_2 represents the effect of wanting no more children, compared to desiring more, for women under age 25.

$(\alpha\gamma)_{i2}$, the interaction term, can be interpreted as the *additional* effect of wanting no more children among women in age group i, compared to women under age 25.

It is possible to simplify slightly the presentation of the results by combining the interactions with some of the main effects. In the present example, it is convenient to present the estimates of α_i as the age effects for women who

TABLE 3.15: The Estimates

Variable	Category	Symbol	Estimate	Std. Err	z-ratio
Constant		η	-1.803	0.180	-10.01
Age	25–29	α_2	0.395	0.201	1.96
	30–39	α_3	0.547	0.198	2.76
	40–49	α_4	0.580	0.347	1.67
Education	Upper	β_2	0.341	0.126	2.71
Desires	<25	γ_2	0.066	0.331	0.20
no more	25–29	$\gamma_2 + (\alpha\gamma)_{22}$	0.325	0.242	1.35
at age	30–39	$\gamma_2 + (\alpha\gamma)_{32}$	1.179	0.175	6.74
	40–49	$\gamma_2 + (\alpha\gamma)_{42}$	1.428	0.354	4.04

want another child, and to present $\gamma_2 + (\alpha\gamma)_{i2}$ as the effect of not wanting another child for women in age group i.

Calculation of the necessary dummy variables proceeds exactly as in Section 3.5. This strategy leads to the parameter estimates in Table 3.15.

To aid in interpretation as well as model criticism, Figure 3.4 plots observed logits based on the original data in Table 3.1, and fitted logits based on the model with an age by desire interaction.

The graph shows four curves tracing contraceptive use by age for groups defined by education and desire for more children. The curves are labelled using L and U for lower and upper education, and Y and N for desire for more children. The lowest curve labelled LY corresponds to women with lower primary education or less who want more children, and shows a slight increase in contraceptive use up to age 35–39 and then a small decline. The next curve labelled UY is for women with upper primary education or more who also want more children. This curve is parallel to the previous one because the effect of education is additive on age. The constant difference between these two curves corresponds to a 41% increase in the odds ratio as we move from lower to upper primary education. The third curve, labelled LN, is for women with lower primary education or less who want no more children. The distance between this curve and the first one represents the effect of wanting no more children at different ages. This effect increases sharply with age, reaching an odds ratio of four by age 40–49. The fourth curve, labelled UN, is for women with upper primary education or more who want no more children. The distance between this curve and the previous one is the effect of education, which is the same whether women want more children or not, and is also the same at every age.

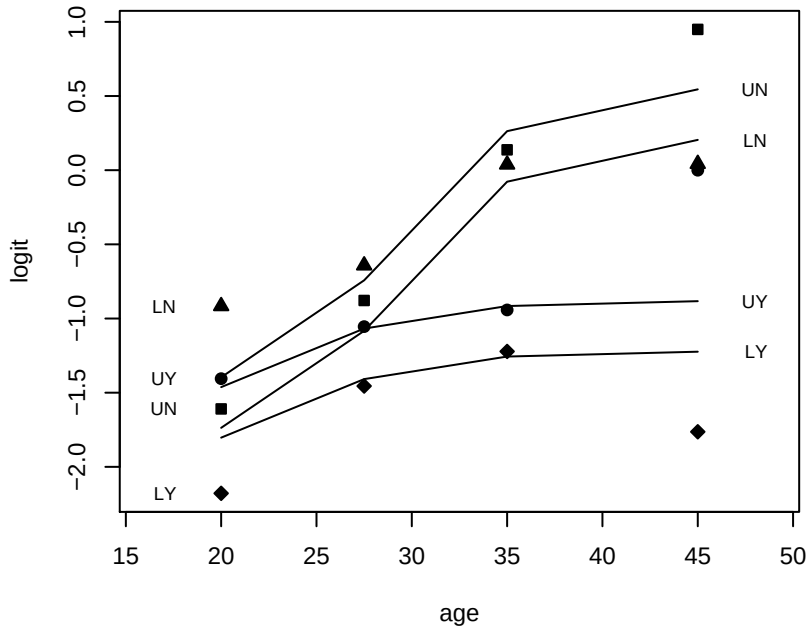

FIGURE 3.4: Logit Model of Contraceptive Use By Age, Education and Desire for Children, With Age by Desire Interaction

The graph also shows the observed logits, plotted using different symbols for each of the four groups defined by education and desire. Comparison of observed and fitted logits shows clearly the strengths and weaknesses of this model: it does a fairly reasonable job reproducing the logits of the proportions using contraception in each group *except* for ages 40–49 (and to a lesser extend the group < 25), where it seems to underestimate the educational differential. There is also some indication that this failure may be more pronounced for women who want more children.

3.6.5 Best Fitting and Parsimonious Models

How can we improve the model of the last section? The most obvious solution is to move to the model with all three two-factor interactions, $AE + AD + ED$, which has a deviance of 2.44 on three d.f. and therefore fits the data

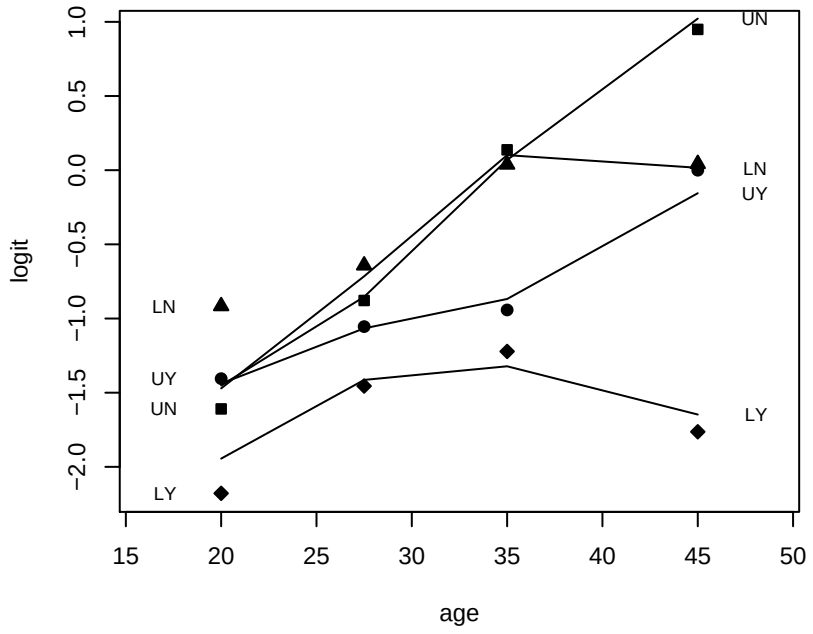

FIGURE 3.5: Observed and Fitted Logits of Contraceptive Use Based on Model with Three Two-Factor Interactions

extremely well. This model implies that the effect of each factor depends on the levels of the other two, but not on the combination of levels of the other two. Interpretation of the coefficients in this model is not as simple as it would be in an additive model, or in a model involving only one interaction. The best strategy in this case is to plot the fitted model and inspect the resulting curves.

Figure 3.5 shows fitted values based on the more complex model. The plot tells a simple story. Contraceptive use for spacing increases slightly up to age 35 and then declines for the less educated but continues to increase for the more educated. Contraceptive use for limiting increases sharply with age up to age 35 and then levels off for the less educated, but continues to increase for the more educated. The figure shows that the effect of wanting no more children increases with age, and appears to do so for both educational groups in the same way (look at the distance between the LY and LN curves, and

between the UY and UN curves). On the other hand, the effect of education is clearly more pronounced at ages 40–49 than at earlier ages, and also seems slightly larger for women who want more children than for those who do not (look at the distance between the LY and UY curves, and between the LN and UN curves).

One can use this knowledge to propose improved models that fit the data without having to use all three two-factor interactions. One approach would note that all interactions with age involve contrasts between ages 40–49 and the other age groups, so one could collapse age into only two categories for purposes of modelling the interactions. A simplified version of this approach is to start from the model $AD + E$ and add one d.f. to model the larger educational effect for ages 40–49. This can be done by adding a dummy variable that takes the value one for women aged 40–49 who have upper primary or more education. The resulting model has a deviance of 6.12 on six d.f., indicating a good fit. Comparing this value with the deviance of 12.6 on seven d.f. for the $AD + E$ model, we see that we reduced the deviance by 6.5 at the expense of a single d.f. The model $AD + AE$ includes all three d.f. for the age by education interaction, and has a deviance of 5.8 on four d.f. Thus, the total contribution of the AE interaction is 6.8 on three d.f. Our one-d.f. improvement has captured roughly 90% of this interaction.

An alternative approach is to model the effects of education and desire for no more children as smooth functions of age. The logit of the probability of using contraception is very close to a linear function of age for women with upper primary education who want no more children, who could serve as a new reference cell. The effect of wanting more children could be modelled as a linear function of age, and the effect of education could be modelled as a quadratic function of age. Let L_{ijk} take the value one for lower primary or less education and zero otherwise, and let M_{ijk} be a dummy variable that takes the value one for women who want more children and zero otherwise. Then the proposed model can be written as

$$\text{logit}(\pi_{ijk}) = \alpha + \beta x_{ijk} + (\alpha_E + \beta_E x_{ijk} + \gamma_E x_{ijk}^2)L_{ijk} + (\alpha_D + \beta_D x_{ijk})M_{ijk}.$$

Fitting this model, which requires only seven parameters, gives a deviance of 7.68 on nine d.f. The only weakness of the model is that it assumes equal effects of education on use for limiting and use for spacing, but these effects are not well-determined. Further exploration of these models is left as an exercise.

3.7 Other Choices of Link

All the models considered so far use the logit transformation of the probabilities, but other choices are possible. In fact, any transformation that maps probabilities into the real line could be used to produce a generalized linear model, as long as the transformation is one-to-one, continuous and differentiable.

In particular, suppose $F(.)$ is the cumulative distribution function (c.d.f.) of a random variable defined on the real line, and write

$$\pi_i = F(\eta_i),$$

for $-\infty < \eta_i < \infty$. Then we could use the inverse transformation

$$\eta_i = F^{-1}(\pi_i),$$

for $0 < \pi_i < 1$ as the link function.

Popular choices of c.d.f.'s in this context are the normal, logistic and extreme value distributions. In this section we motivate this general approach by introducing models for binary data in terms of latent variables.

3.7.1 A Latent Variable Formulation

Let Y_i denote a random variable representing a binary response coded zero and one, as usual. We will call Y_i the *manifest* response. Suppose that there is an unobservable continuous random variable Y_i^* which can take any value in the real line, and such that Y_i takes the value one if an only if Y_i^* exceeds a certain threshold θ. We will call Y_i^* the *latent* response. Figure 3.6 shows the relationship between the latent variable and the response when the threshold is zero.

The interpretation of Y_i and Y_i^* depends on the context. An economist, for example, may view Y_i as a binary choice, such as purchasing or renting a home, and Y_i^* as the difference in the utilities of purchasing and renting. A psychologist may view Y_i as a response to an item in an attitude scale, such as agreeing or disagreeing with school vouchers, and Y_i^* as the underlying attitude. Biometricians often view Y_i^* as a dose and Y_i as a response, hence the name dose-response models.

Since a positive outcome occurs only when the latent response exceeds the threshold, we can write the probability π_i of a positive outcome as

$$\pi_i = \Pr\{Y_i = 1\} = \Pr\{Y_i^* > \theta\}.$$

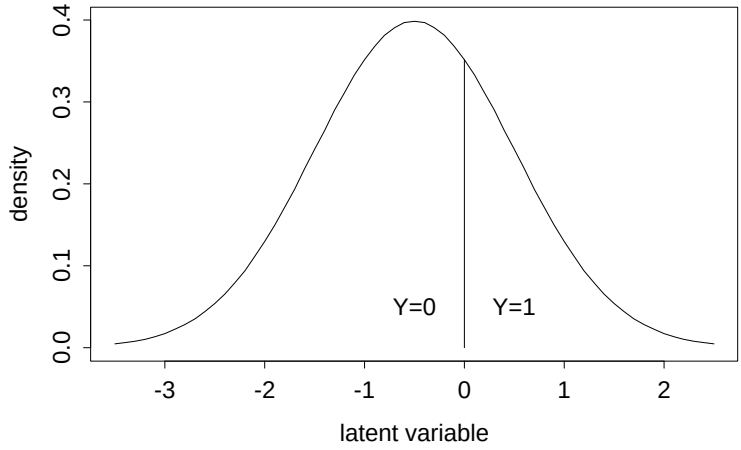

FIGURE 3.6: Latent Variable and Manifest Response

As often happens with latent variables, the location and scale of Y_i^* are arbitrary. We can add a constant a to both Y_i^* and the threshold θ, or multiply both by a constant c, without changing the probability of a positive outcome. To identify the model we take the threshold to be zero, and standardize Y_i^* to have standard deviation one (or any other fixed value).

Suppose now that the outcome depends on a vector of covariates $\mathbf{x}$. To model this dependence we use an ordinary linear model for the *latent* variable, writing

$$Y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + U_i, \tag{3.15}$$

where $\boldsymbol{\beta}$ is a vector of coefficients of the covariates $\mathbf{x}_i$ and U_i is the error term, assumed to have a distribution with c.d.f. $F(u)$, not necessarily the normal distribution.

Under this model, the probability π_i of observing a positive outcome is

$$\begin{aligned} \pi_i &= \Pr\{Y_i > 0\} \\ &= \Pr\{U_i > -\eta_i\} \\ &= 1 - F(-\eta_i), \end{aligned}$$

where $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$ is the linear predictor. If the distribution of the error term U_i is symmetric about zero, so $F(u) = 1 - F(-u)$, we can write

$$\pi_i = F(\eta_i)$$

This expression defines a generalized linear model with Bernoulli response and link

$$\eta_i = F^{-1}(\pi_i). \tag{3.16}$$

In the more general case where the distribution of the error term is not necessarily symmetric, we still have a generalized linear model with link

$$\eta_i = -F^{-1}(1 - \pi_i). \tag{3.17}$$

We now consider some specific distributions.

3.7.2 Probit Analysis

The obvious choice of an error distribution is the normal. Assuming that the error term has a standard normal distribution $U_i \sim N(0, 1)$, the results of the previous section lead to

$$\pi_i = \Phi(\eta_i),$$

where Φ is the standard normal c.d.f. The inverse transformation, which gives the linear predictor as a function of the probability

$$\eta_i = \Phi^{-1}(\pi_i),$$

is called the *probit*.

It is instructive to consider the more general case where the error term $U_i \sim N(0, \sigma^2)$ has a normal distribution with variance σ^2. Following the same steps as before we find that

$$\begin{aligned}\pi_i &= \Pr\{Y_i^* > 0\} \\ &= \Pr\{U_i > -\mathbf{x}_i'\boldsymbol{\beta}\} = \Pr\{U_i/\sigma > -\mathbf{x}_i'\boldsymbol{\beta}/\sigma\} \\ &= 1 - \Phi(-\mathbf{x}_i'\boldsymbol{\beta}/\sigma) = \Phi(\mathbf{x}_i'\boldsymbol{\beta}/\sigma),\end{aligned}$$

where we have divided by σ to obtain a standard normal variate, and used the symmetry of the normal distribution to obtain the last result.

This development shows that we cannot identify $\boldsymbol{\beta}$ and σ separately, because the probability depends on them only through their ratio $\boldsymbol{\beta}/\sigma$. This is another way of saying that the scale of the latent variable is not identified. We therefore take $\sigma = 1$, or equivalently interpret the β's in units of standard deviation of the latent variable.

As a simple example, consider fitting a probit model to the contraceptive use data by age and desire for more children. In view of the results in Section 3.5, we introduce a main effect of wanting no more children, a linear effect

TABLE 3.16: Estimates for Probit Model of Contraceptive Use With a Linear Age by Desire Interaction

Parameter	Symbol	Estimate	Std. Error	z-ratio
Constant	α_1	−0.7297	0.0460	−15.85
Age	β_1	0.0129	0.0061	2.13
Desire	$\alpha_2 - \alpha_1$	0.4572	0.0731	6.26
Age × Desire	$\beta_2 - \beta_1$	0.0305	0.0092	3.32

of age, and a linear age by desire interaction. Fitting this model gives a deviance of 8.91 on four d.f. Estimates of the parameters and standard errors appear in Table 3.16

To interpret these results we imagine a latent continuous variable representing the woman's motivation to use contraception (or the utility of using contraception, compared to not using). At the average age of 30.6, not wanting more children increases the motivation to use contraception by almost half a standard deviation. Each year of age is associated with an increase in motivation of 0.01 standard deviations if she wants more children and 0.03 standard deviations more (for a total of 0.04) if she does not. In the next section we compare these results with logit estimates.

A slight disadvantage of using the normal distribution as a link for binary response models is that the c.d.f. does not have a closed form, although excellent numerical approximations and computer algorithms are available for computing both the normal probability integral and its inverse, the probit.

3.7.3 Logistic Regression

An alternative to the normal distribution is the standard logistic distribution, whose shape is remarkably similar to the normal distribution but has the advantage of a closed form expression

$$\pi_i = F(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}},$$

for $-\infty < \eta_i < \infty$. The standard logistic distribution is symmetric, has mean zero, and has variance $\pi^2/3$. The shape is very close to the normal, except that it has heavier tails. The inverse transformation, which can be obtained solving for η_i in the expression above is

$$\eta_i = F^{-1}(\pi_i) = \log \frac{\pi_i}{1 - \pi_i},$$

our good old friend, the *logit*.

Thus, coefficients in a logit regression model can be interpret not only in terms of log-odds, but also as effects of the covariates on a latent variable that follows a linear model with logistic errors.

The logit and probit transformations are almost linear functions of each other for values of π_i in the range from 0.1 to 0.9, and therefore tend to give very similar results. Comparison of probit and logit coefficients should take into account the fact that the standard normal and the standard logistic distributions have different variances. Recall that with binary data we can only estimate the ratio $\boldsymbol{\beta}/\sigma$. In probit analysis we have implicitly set $\sigma = 1$. In a logit model, by using a standard logistic error term, we have effectively set $\sigma = \pi/\sqrt{3}$. Thus, coefficients in a logit model should be standardized dividing by $\pi/\sqrt{3}$ before comparing them with probit coefficients.

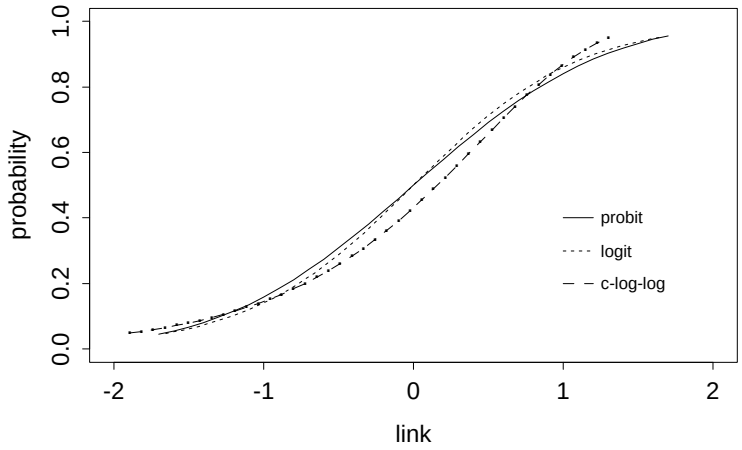

FIGURE 3.7: The Standardized Probit, Logit and C-Log-Log Links

Figure 3.7 compares the logit and probit links (and a third link discussed below) after standardizing the logits to unit variance. The solid line is the probit and the dotted line is the logit divided by $\pi/\sqrt{3}$. As you can see, they are barely distinguishable.

To illustrate the similarity of these links in practice, consider our models of contraceptive use by age and desire for more children in Tables 3.10 and 3.16. The deviance of 9.14 for the logit model is very similar to the deviance of 8.91 for the probit model, indicating an acceptable fit. The Wald tests of individual coefficients are also very similar, for example the test for the effect of wanting no more children at age 30.6 is 6.22 in the logit model and 6.26

in the probit model. The coefficients themselves look somewhat different, but of course they are not standardized. The effect of wanting no more children at the average age is 0.758 in the logit scale. Dividing by $\pi/\sqrt{3}$, the standard deviation of the underlying logistic distribution, we find this effect equivalent to an increase in the latent variable of 0.417 standard deviations. The probit analysis estimates the effect as 0.457 standard deviations.

3.7.4 The Complementary Log-Log Transformation

A third choice of link is the complementary log-log transformation

$$\eta_i = \log(-\log(1-\pi_i)),$$

which is the inverse of the c.d.f. of the extreme value (or log-Weibull) distribution, with c.d.f.

$$F(\eta_i) = 1 - e^{-e^{\eta_i}}.$$

For small values of π_i the complementary log-log transformation is close to the logit. As the probability increases, the transformation approaches infinity more slowly that either the probit or logit.

This particular choice of link function can also be obtained from our general latent variable formulation if we assume that $-U_i$ (note the minus sign) has a standard extreme value distribution, so the error term itself has a *reverse* extreme value distribution, with c.d.f.

$$F(U_i) = e^{-e^{-U_i}}.$$

The reverse extreme value distribution is asymmetric, with a long tail to the right. It has mean equal to Euler's constant 0.577 and variance $\pi^2/6 = 1.645$. The median is $-\log\log 2 = 0.367$ and the quartiles are -0.327 and 1.246.

Inverting the reverse extreme value c.d.f. and applying Equation 3.17, which is valid for both symmetric and asymmetric distributions, we find that the link corresponding to this error distribution is the complementary log-log.

Thus, coefficients in a generalized linear model with binary response and a complementary log-log link can be interpreted as effects of the covariates on a latent variable which follows a linear model with reverse extreme value errors.

To compare these coefficients with estimates based on a probit analysis we should standardize them, dividing by $\pi/\sqrt{6}$. To compare coefficients with logit analysis we should divide by $\sqrt{2}$, or standardize both c-log-log and logit coefficients.

Figure 3.7 compares the c-log-log link with the probit and logit after standardizing it to have mean zero and variance one. Although the c-log-log link differs from the other two, one would need extremely large sample sizes to be able to discriminate empirically between these links.

The complementary log-log transformation has a direct interpretation in terms of hazard ratios, and thus has practical applications in terms of hazard models, as we shall see later in the sequel.

3.8 Regression Diagnostics for Binary Data

Model checking is just as important in logistic regression and probit analysis as it is in classical linear models. The raw materials are again the residuals, or differences between observed and fitted values. Unlike the case of linear models, however, we now have to make allowance for the fact that the observations have different variances. There are two types of residuals in common use.

3.8.1 Pearson Residuals

A very simple approach to the calculation of residuals is to take the difference between observed and fitted values and divide by an estimate of the standard deviation of the observed value. The resulting residual has the form

$$p_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i(n_i - \hat{\mu}_i)/n_i}}, \tag{3.18}$$

where $\hat{\mu}_i$ is the fitted value and the denominator follows from the fact that $\text{var}(y_i) = n_i\pi_i(1-\pi_i)$.

The result is called the Pearson residual because the square of p_i is the contribution of the i-th observation to Pearson's chi-squared statistic, which was introduced in Section 3.2.2, Equation 3.14.

With grouped data the Pearson residuals are approximately normally distributed, but this is not the case with individual data. In both cases, however, observations with a Pearson residual exceeding two in absolute value may be worth a closer look.

3.8.2 Deviance Residuals

An alternative residual is based on the deviance or likelihood ratio chi-squared statistic. The deviance residual is defined as

$$d_i = \sqrt{2[y_i \log(\frac{y_i}{\hat{\mu}_i}) + (n_i - y_i)\log(\frac{n_i - y_i}{n_i - \hat{\mu}_i})]}, \tag{3.19}$$

with the same sign as the raw residual $y_i - \hat{y}_i$. Squaring these residuals and summing over all observations yields the deviance statistic. Observations with a deviance residual in excess of two may indicate lack of fit.

3.8.3 Studentized Residuals

The residuals defined so far are not fully standardized. They take into account the fact that different observations have different variances, but they make no allowance for additional variation arising from estimation of the parameters, in the way studentized residuals in classical linear models do.

Pregibon (1981) has extended to logit models some of the standard regression diagnostics. A key in this development is the weighted *hat* matrix

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2},$$

where $\mathbf{W}$ is the diagonal matrix of iteration weights from Section 3.2.1, with entries $w_{ii} = \mu_i(n_i - \mu_i)/n_i$, evaluated at the m.l.e.'s. Using this expression it can be shown that the variance of the raw residual is, to a first-order approximation,

$$\text{var}(y_i - \hat{\mu}_i) \approx (1 - h_{ii})\text{var}(y_i),$$

where $h_i i$ is the leverage or diagonal element of the weighted hat matrix. Thus, an internally studentized residual can be obtained dividing the Pearson residual by the square root of $1 - h_{ii}$, to obtain

$$s_i = \frac{p_i}{\sqrt{1 - h_{ii}}} = \frac{y_i - \hat{\mu}_i}{\sqrt{(1 - h_{ii})\hat{\mu}_i(n_i - \hat{\mu}_i)/n_i}}.$$

A similar standardization can be applied to deviance residuals. In both cases the standardized residuals have the same variance only approximately because the correction is first order, unlike the case of linear models where the correction was exact.

Consider now calculating jack-knifed residuals by omitting one observation. Since estimation relies on iterative procedures, this calculation would

be expensive. Suppose, however, that we start from the final estimates and do only one iteration of the IRLS procedure. Since this step is a standard weighted least squares calculation, we can apply the standard regression updating formulas to obtain the new coefficients and thus the predictive residuals. Thus, we can calculate a jack-knifed residual as a function of the standardized residual using the same formula as in linear models

$$t_i = s_i \sqrt{\frac{n-p-1}{n-p-s_i^2}}$$

and view the result as a one-step approximation to the true jack-knifed residual.

3.8.4 Leverage and Influence

The diagonal elements of the hat matrix can be interpreted as leverages just as in linear models. To measure actual rather than potential influence we could calculate Cook's distance, comparing $\hat{\boldsymbol{\beta}}$ with $\hat{\boldsymbol{\beta}}_{(i)}$, the m.l.e.'s of the coefficients with and without the i-th observation. Calculation of the later would be expensive if we iterated to convergence. Pregibon (1981), however, has shown that we can use the standard linear models formula

$$D_i = s_i^2 \frac{h_{ii}}{(1-h_{ii})p},$$

and view the result as a one-step approximation to Cook's distance, based on doing one iteration of the IRLS algorithm towards $\hat{\boldsymbol{\beta}}_{(i)}$ starting from the complete data estimate $\hat{\boldsymbol{\beta}}$.

3.8.5 Testing Goodness of Fit

With grouped data we can assess goodness of fit by looking directly at the deviance, which has approximately a chi-squared distribution for large n_i. A common rule of thumb is to require all expected frequencies (both expected successes $\hat{\mu}_i$ and failures $n_i - \hat{\mu}_i$) to exceed one, and 80% of them to exceed five.

With individual data this test is not available, but one can always group the data according to their covariate patterns. If the number of possible combinations of values of the covariates is not too large relative to the total sample size, it may be possible to group the data and conduct a formal goodness of fit test. Even when the number of covariate patterns is large, it is possible that a few patterns will account for most of the observations. In this

case one could compare observed and fitted counts at least for these common patterns, using either the deviance or Pearson's chi-squared statistic.

Hosmer and Lemeshow (1980, 1989) have proposed an alternative procedure that can be used with individual data even if there are no common covariate patterns. The basic idea is to use predicted probabilities to create groups. These authors recommend forming ten groups, with predicted probabilities of 0–0.1, 0.1–0.2, and so on, with the last group being 0.9–1. One can then compute expected counts of successes (and failures) for each group by summing the predicted values (and their complements), and compare these with observed values using Pearson's chi-squared statistic. Simulation studies show that the resulting statistic has approximately in large samples the usual chi-squared distribution, with degrees of freedom equal to $g - 2$, where g is the number of groups, usually ten. It seems reasonable to assume that this result would also apply if one used the deviance rather than Pearson's chi-squared.

Another measure that has been proposed in the literature is a pseudo-R^2, based on the proportion of deviance explained by a model. This is a direct extension of the calculations based on RSS's for linear models. These measures compare a given model with the null model, and as such do not necessarily measure goodness of fit. A more direct measure of goodness of fit would compare a given model with the saturated model, which brings us back again to the deviance.

Yet another approach to assessing goodness of fit is based on prediction errors. Suppose we were to use the fitted model to predict 'success' if the fitted probability exceeds 0.5 and 'failure' otherwise. We could then crosstabulate the observed and predicted responses, and calculate the proportion of cases predicted correctly. While intuitively appealing, one problem with this approach is that a model that fits the data may not necessarily predict well, since this depends on how predictable the outcome is. If prediction was the main objective of the analysis, however, the proportion classified correctly would be an ideal criterion for model comparison.