CHAPTER 2
Binomial Data

2.1 Challenger Disaster Example

In January 1986, the space shuttle Challenger exploded shortly after launch. An investigation was launched into the cause of the crash and attention focused on the rubber O-ring seals in the rocket boosters. At lower temperatures, rubber becomes more brittle and is a less effective sealant. At the time of the launch, the temperature was 31°F. Could the failure of the O-rings have been predicted? In the 23 previous shuttle missions for which data exists, some evidence of damage due to blow by and erosion was recorded on some O-rings. Each shuttle had two boosters, each with three O-rings. For each mission, we know the number of O-rings out of six showing some damage and the launch temperature. This is a simplification of the problem—see Dalal, Fowlkes, and Hoadley (1989) for more details.

Let's start our analysis with R. For help in obtaining R and installing the necessary add-on packages and datasets, please see Appendix B. First we load the data. To do this, you will first need to load the faraway package using the library command as seen in here. You will need to do this in every session that you run examples from this book. If you forget, you will receive a warning message about the data not being found. We then plot the proportion of damaged O-rings against temperature in Figure 2.1:

```
> library(faraway)
> data(orings)
> plot (damage/6 ~ temp, orings, xlim=c(25,85), ylim =
c(0,1),
  xlab="Temperature",  ylab="Prob of damage")
```

We are interested in how the probability of failure in a given O-ring is related to the launch temperature and predicting that probability when the temperature is 31°F. A naive approach, based on linear models, simply fits a line to this data:

```
> lmod <- lm(damage/6 ~ temp, orings)
> abline(lmod)
```

The fit is shown in Figure 2.1. There are several problems with this approach. Most obviously from the plot, it can predict probabilities greater than one or less than zero. One might suggest truncating predictions outside the range to zero or one as appropriate, but it does not seem credible that these probabilities would be exactly zero or one, in this particular example or many others.

We might consider the number of damage incidents to be binomially distributed. For a linear model, we require the errors to be approximately normally distributed for accurate inference. However, for a binomial with only six trials, the normal approx-

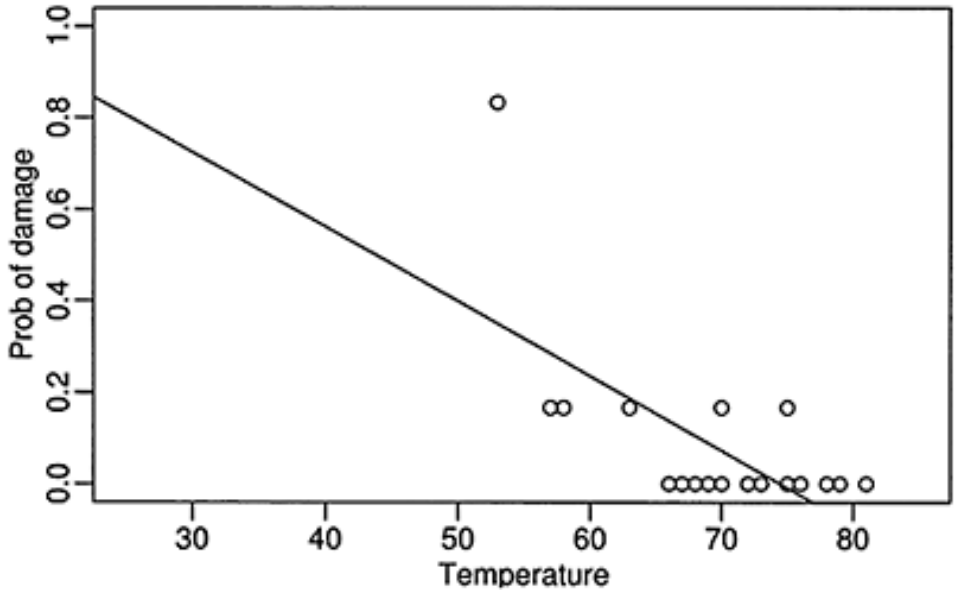

Figure 2.1 *Damage to O-rings in 23 space shuttle missions as a function of launch temperature. Least squares fit line is shown.*

imation is too much of a stretch. Furthermore, the variance of a binomial variable is not constant which violates another crucial assumption of the linear model.

The standard linear model is clearly not directly suitable here. Although, we could use transformation and weighting to correct some of these problems, it is better to develop a model that is directly suited for binomial data.

2.2 Binomial Regression Model

Suppose the response variable Y_i for $i=1,\ldots, n_i$ is binomially distributed $B(n_i, p_i)$ so that:

$$P(Y_i = y_i) = \binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i - y_i}$$

We further assume that the Y_i are independent. The individual trials that compose the response Y_i are all subject to the same q predictors $(x_{i1},\ldots, x_{iq})$. The group of trials is known as a *covariate class*. We need a model that describes the relationship of $x_1,\ldots, x_q$ to p. Following the linear model approach, we construct a *linear predictor*:

$\eta_i=\beta_0+\beta_1 x_{i1}+\ldots+\beta_q x_{iq}$

Since the linear predictor can accommodate quantitative and qualitative predictors with the use of dummy variables and also allows for transformations and combinations of the original predictors, it is very flexible and yet retains interpretability. This notion that we can express the effect of the predictors on the response solely through the linear predictor is important. The idea can be extended to models for other types of response and is one of the defining features of the wider class of generalized linear models (GLMs) discussed in Chapter 6.

We have already seen above that setting $\eta_i=p_i$ is not appropriate because we require $0\leq p_i\leq 1$. Instead we shall use a *link function* g such that $\eta i=g(p_i)$. For this application, we shall need g to be monotone and be such that $0\leq g^{-1}(\eta)\leq 1$ for any η. There are three common choices:

1. Logit: $\eta=\log(p/(1-p))$.
2. Probit: $\eta=\Phi^{-1}(p)$ where Φ^{-1} is the inverse normal cumulative distribution function.
3. Complementary log-log: $\eta=\log(-\log(1-p))$.

The idea of the link function is also one of the central ideas of generalized linear models. It is used to link the linear predictor to the mean of the response in the wider class of models.

We will compare these three choices of link function later, but first we estimate the parameters of the model. We shall use the method of maximum likelihood; see Appendix A for a brief introduction to this method. The log-likelihood is given by:

$$l(\beta)=\sum_{i=1}^{n}\left[y_i\eta_i-n_i\log(1+e_i^{\eta})+\log\binom{n_i}{y_i}\right]$$

We can maximize this to obtain the maximum likelihood estimates $\hat{\beta}$ and use the standard theory to obtain approximate standard errors. An algorithm to perform the maximization will be discussed in Chapter 6.

We use R to estimate the regression parameters for the Challenger data:

```
> logitmod <- glm(cbind(damage,6-damage) ~ temp,
family=binomial, orings)
> summary(logitmod)
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-0.953  -0.735  -0.439  -0.208   1.957
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  11.6630     3.2963    3.54    4e-04
temp         -0.2162     0.0532   -4.07  4.8e-05
(Dispersion parameter for binomial family taken to be
1)
    Null deviance: 38.898  on 22 degrees of freedom
Residual deviance: 16.912  on 21 degrees of freedom
AIC: 33.67
Number of Fisher Scoring iterations: 6
```

For binomial response data, we need two pieces of information about the response values—y and n. In R, one way of doing this is to form a two-column matrix with the first column representing the number of "successes" y and the second column the number of "failures" $n–y$. We have specified that the response is binomially distributed. The default choice of link is the logit—other choices need to be specifically stated as we shall see shortly. This default choice is sometimes called *logistic regression*. The regression coefficients are given in the output — $\hat{\beta}_0 = 11.6$ and $\hat{\beta}_1 = -0.216$ along with their respective standard errors. The rest of the output will be explained shortly.

We show the logit fit to the data as seen in Figure 2.2:

```
> plot (damage/6 ~ temp, orings, xlim=c(25,85),
ylim=c(0,1),
  xlab="Temperature", ylab="Prob of damage")
> x <- seq(25,85,1)
> lines(x,ilogit(11.6630-0.2162*x))
```

Notice how the logit fit tends asymptotically toward zero and one at high and low temperatures, respectively. The fitted values never actually reach zero or one, so the model never predicts anything to completely certain or completely impossible. Now

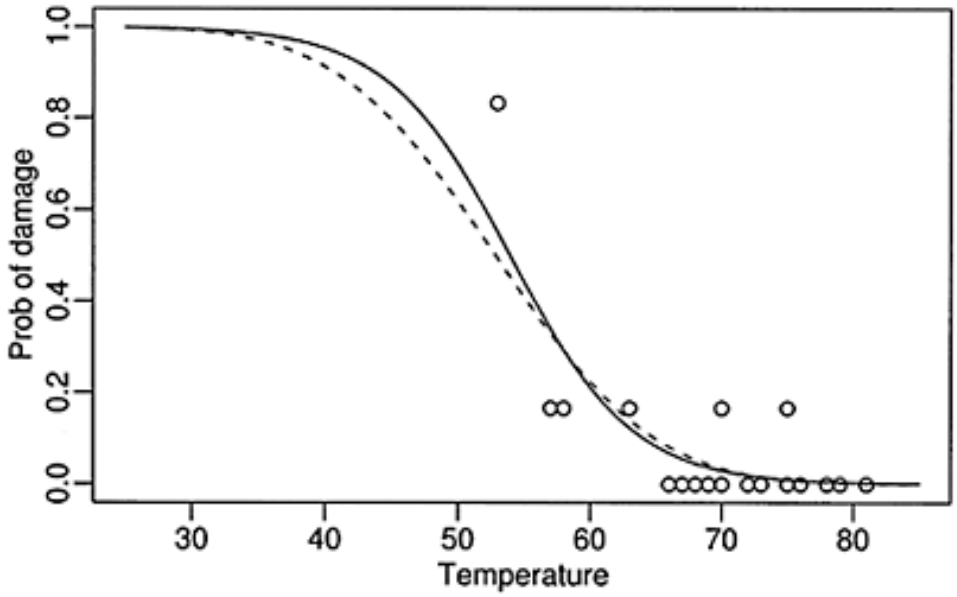

Figure 2.2 *Logit (solid line) and probit (dashed line) fits to the Challenger data*

compare this to the probit fit:

```
> probitmod <- glm(cbind(damage,6-damage) ~ temp,
  family=binomial(link=probit), orings)
```

```
> summary(probitmod)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   5.5915     1.7105    3.27   0.0011
temp         -0.1058     0.0266   -3.98  6.8e-05
(Dispersion parameter for binomial family taken to be
1)
    Null deviance: 38.898 on 22 degrees of freedom
Residual deviance: 18.131 on 21 degrees of freedom
AIC: 34.89
```

Although the coefficients seem quite different, the fits are similar, particularly in the range of the data, as seen in Figure 2.2:

```
> lines(x, pnorm(5.5915-0.1058*x), lty=2)
```

We can predict the response at 31°F for both models:

```
> ilogit (11.6630-0.2162*31)
[1] 0.99304
> pnorm(5.5915-0.1058*31)
[1] 0.9896
```

We see a very high probability of damage with either model although we still need to develop some inferential techniques before we leap to conclusions.

2.3 Inference

Consider two models, a larger model with l parameters and likelihood L_L and a smaller model with s parameters and likelihood L_S where the smaller model represents a linear subspace (a linear restriction on the parameters) of the larger model. Likelihood methods suggest the likelihood ratio statistic:

$$2\log\frac{L_L}{L_S} \qquad (2.1)$$

as an appropriate test statistic for comparing the two models. Now suppose we choose a saturated larger model—such a model typically has as many parameters as cases and has fitted values $\hat{p}_i = y_i/n_i$. In such a case, the test statistic becomes:

$$D = 2\sum_{i=1}^{n}\{y_i \log y_i/\hat{y}_i + (n_i - y_i)\log(n_i - y_i)/(n_i - \hat{y}_i)\}$$

where $\hat{y}_i$ are the fitted values from the smaller model. Now since the saturated model fits as well as any model can fit, the *deviance* D measures how close the (smaller) model comes to perfection. Thus deviance is a measure of goodness of fit. In the output for the

models above, the Residual deviance is the deviance for the current model while the Null deviance is the deviance for a model with no predictors and just an intercept term.

Provided that Y is truly binomial and that the n_i are relatively large, the deviance is approximately χ^2 distributed with $n–l$ degrees of freedom if the model is correct. Thus we can use the deviance to test whether the model is an adequate fit. For the logit model of the Challenger data, we may compute:

```
> pchisq(deviance(logitmod),
df.residual(logitmod),lower=FALSE)
[1] 0.71641
```

Since this p-value is well in excess of 0.05, we may conclude that this model fits sufficiently well. Of course, this does not mean that this model is correct or that a simpler model might not also fit adequately. Even so, for the null model:

```
> pchisq(38.9,22,lower=FALSE)
[1] 0.014489
```

we see that the fit is inadequate, so we cannot ascribe the response to simple variation not dependent on any predictor. Note that a χ^2_d variable has mean d and standard deviation $\sqrt{2d}$ so that it is often possible to quickly judge whether a deviance is large or small without explicitly computing the p-value. If the deviance is far in excess of the degrees of freedom, the null hypothesis can be rejected.

The χ^2 distribution is only an approximation that becomes more accurate as the n_i increase. For the case, n_i=1, when y_i=0 or 1, in other words, a binary response, the deviance reduces to:

$$-2\sum_{i=1}^{n}\{\hat{p}_i\text{logit}(\hat{p}_i)+\log(1-\hat{p}_i)$$

For a deviance to measure fit, it has to compare the fitted values $\hat{p}_i$ to the data y_i, but here we have only a function of $\hat{p}_i$. Thus this deviance does not assess goodness of fit and furthermore, it is not even approximately χ^2 distributed. Other methods must be used to judge goodness of fit for binary data—for example, the Hosmer-Lemeshow test described in Hosmer and Lemeshow (2000).

The approximation is very poor for small n_i. Although it is not possible to say exactly how large ni should be for an adequate approximation, $n_i \geq 5$ has often been suggested. Permutation or bootstrap methods might be considered as an alternative.

We can also use the deviance to compare two nested models. The test statistic in (2.1) becomes DS—DL. This test statistic is asymptotically distributed χ^2_{l-s}, assuming that the smaller model is correct and the distributional assumptions hold. We can use this to test the significance of temperature by computing the difference in the deviances between the model with and without temperature. The model without temperature is just the null model and the difference in degrees of freedom or parameters is one:

```
> pchisq(38.9-16.9,1,lower=FALSE)
[1] 2.7265e-06
```

Since the *p*-value is so small, we conclude that the effect of launch temperature is statistically significant. An alternative to this test is the *z*-value, which is $\hat{\beta}/se(\hat{\beta})$, here –4.07 with a *p*-value of 4.8e-05. In contrast to the normal (Gaussian) linear model, these two statistics are not identical. In this particular example, there is no practical difference, but in some cases, especially with sparse data, the standard errors can be overestimated and so the *z*-value is too small and the significance of an effect could be missed. This is known as the Hauck-Donner effect—see Hauck and Donner (1977). So the deviance-based test is preferred.

Again, there are concerns with the accuracy of the approximation, but the test involving differences of deviances is generally more accurate than the goodness of fit test involving a single deviance.

Confidence intervals for the regression parameters may be constructed using normal approximations for the parameter estimates. A 100(1—α)% confidence interval for β_i would be:

$$\hat{\beta}_i \pm z^{\alpha/2} se(\hat{\beta}_i)$$

where $z^{\alpha/2}$ is a quantile from the normal distribution. Thus a 95% confidence interval for β_1 in our model would be:

```
> c(-0.2162-1.96*0.0532,-0.2162+1.96*0.0532)
[1] -0.32047 -0.11193
```

It is also possible to construct a profile likelihood-based confidence interval:

```
> library(MASS)
> confint(logitmod)
Waiting for profiling to be done...
                2.5 %   97.5 %
(Intercept)  5.57543 18.73812
temp        -0.33267 -0.12018
```

It is important to load the MASS package or the default confint method for ordinary linear models will be used (which will not be quite right). The profile likelihood method is generally preferable for the same Hauck-Donner reasons discussed above although it is more work to compute.

Although we have only computed results for the logit link, the same methods would apply for the probit or any other link.

2.4 Tolerance Distribution

Suppose that students answers questions on a test and that a specific student has an aptitude T. A particular question might have difficulty d_i and the student will get the answer correct only if $T>d_i$. Now if we consider d_i fixed and $T\sim N(\mu,\sigma 2)$, then the probability that a randomly selected student will get the answer wrong is:

$p_i=P(T\leq d_i)=\Phi((d_i-\mu)/\sigma)$

So

$\Phi^{-1}(p_i)=-\mu/\sigma+d_i/\sigma$

If we set $\beta_0=-\mu/\sigma$ and $\beta_1=1/\sigma$, we now have a probit regression model. So we see that the probit link can be naturally motivated by the existence of a normally distributed *tolerance distribution* T. The term arose from toxicity studies where the aptitude of the subject would be replaced with the tolerance of the insect.

The logit model arises from a logistically distributed tolerance distribution. The logistic and normal density are very similar in the mid-range, but differ more in a relative sense in the tails. The complementary log-log is similarly associated with an extreme value distribution.

2.5 Interpreting Odds

Odds are sometimes a better scale than probability to represent chance. They arose as a way to express the payoffs for bets. An *evens* bet means that the winner gets paid an equal amount to that staked. A 3–1 *against* bet would pay \$3 for every \$1 bet while a 3–1 *on* bet would pay only \$1 for every \$3 bet. If these bets are *fair* in the sense that a bettor would break even in the long-run average, then we can make a correspondence to probability. Let p be the probability and o be the odds, where we represent 3–1 against as 1/3 and 3–1 on as 3, then the following relationships hold:

$$\frac{p}{1-p}=o \qquad\qquad p=\frac{o}{1+o}$$

One mathematical advantage of odds is that they are unbounded above which makes them more convenient for some modeling purposes.

Odds also form the basis of a subjective assessment of probability. Some probabilities are determined from considerations of symmetry or long-term frequencies, but such information is often unavailable. Individuals may determine their subjective probability for events by considering what odds they would be prepared to offer on the outcome. Under this theory, other potential persons would be allowed to place bets for or against the event occurring. Thus the individual would be forced to make an honest assessment of probability to avoid financial loss.

If we have two covariates x_1 and x_2, then the logistic regression model is:

$$\log(\text{odds})=\log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1 x_1+\beta_2 x_2$$

Now β_1 can be interpreted as follows: a unit increase in x_1 with x_2 held fixed increases the log-odds of success by β_1 or increases the odds of success by a factor of exp β_1. Of course, the usual interpretational difficulties regarding causation apply as in standard regression. No such simple interpretation exists for other links such as the probit.

An alternative notion to odds-ratio is relative risk. Suppose the probability of "success" in the presence of some condition is p_1 and p_2 in its absence. The relative risk is P_1/P_2. For rare outcomes, the relative risk and the o dds ratio will be very similar, but for larger probabilities, there may be substantial differences. There is some debate over which is the more intuitive way of expressing the effect of some condition.

Consider the data shown in Table 2.1 from a study on infant respiratory disease, namely the proportions of children developing bronchitis or pneumonia in their first year of life by type of feeding and sex, which may be found in Payne (1987):

	Bottle Only	Some Breast with Supplement	Breast Only
Boys	77/458	19/147	47/494
Girls	48/384	16/127	31/464

Table 2.1 *Incidence of respiratory disease in infants to the age of 1 year.*

We can recover the layout above with the proportions as follows:

```
> data(babyfood)
> xtabs(disease/(disease+nondisease)~sex+food,
babyfood)
      food
sex     Bottle   Breast    Suppl
  Boy 0.16812 0.095142 0.12925
Girl 0.12500 0.066810 0.12598
```

Fit and examine the model:

```
> mdl <- glm(cbind(disease, nondisease) ~ sex+food,
family=binomial,
babyfood)
> summary(mdl)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.613      0.112  -14.35  < 2e-16
sexGirl       -0.313      0.141   -2.22    0.027
foodBreast    -0.669      0.153   -4.37  1.2e-05
foodSuppl     -0.173      0.206   -0.84    0.401
(Dispersion parameter for binomial family taken to be
1)
    Null deviance: 26.37529  on 5  degrees of freedom
Residual deviance:  0.72192  on 2  degrees of freedom
AIC: 40.24
```

The χ^2 approximation can be expected to be accurate here due to the large covariate class sizes. Is there a sex-by-food interaction? Notice that a model with the interaction effect would be saturated with deviance and degrees of freedom zero, so we can look at the residual deviance of this model to test for an interaction effect. A deviance of 0.72 is not at all large for two degrees of freedom, so we may conclude that there is no evidence of an interaction effect. This means that we may interpret the main effects separately.

We can test for the significance of the main effects:

```
> dropl(mdl,test="Chi")
Single term deletions
Model:
cbind(disease, nondisease) ~ sex + food
       Df Deviance  AIC  LRT Pr (Chi)
<none>          0.7 40.2
sex     1       5.7 43.2  5.0    0.026
food    2      20.9 56.4 20.2  4.2e-05
```

The drop1 function tests each predictor relative to the full. We see that both predictors are significant in this sense. Now consider the interpretation of the coefficients, starting with the effect of breast feeding:

```
> exp(-0.669)
[1] 0.51222
```

We see that breast feeding reduces the odds of respiratory disease to 51% of that for bottle feeding. We could compute a confidence interval by figuring the standard error on the odds scale; however, we get better coverage properties by computing the interval on the log-odds scale and then transforming the endpoints as follows:

```
> exp(c (-0.669-1.96*0.153, -0.669+1.96*0.153))
[1] 0.37951 0.69134
```

Notice that the interval is asymmetric about the estimated effect of 0.512. Confidence intervals can also be computed using profile likelihood methods:

```
> library(MASS)
> exp(confint(mdl))
Waiting for profiling to be done...
              2.5 %  97.5 %
(Intercept) 0.15920 0.24743
sexGirl     0.55362 0.96292
foodBreast  0.37819 0.68952
foodSuppl   0.55552 1.24643
```

which gives a slightly wider interval. This latter result is usually more reliable although it makes little difference for this data.

As an aside, note that for small values of ε, we have:

$\log(x(1+\varepsilon))=\log x+\log(1+\varepsilon) \approx \log x+\varepsilon$

This approximation is reasonable for values $-0.25<\varepsilon<0.25$. So, for example, given the observed supplement coefficient of –0.173, we can approximate the reduction in odds as about 17% relative to bottle feeding. The exact figure is:

```
> 1exp(-0.173)
[1] 0.15886
```

that is about 16%. So the approximation is only good for a quick sense of the effect, but an exact calculation is necessary for results that will be presented to others.

Here we see that breast-fed and to a lesser extent supplement-fed babies are less vulnerable to respiratory disease. We also see that boys are more vulnerable than girls. We should be careful about making any general conclusions from this data without knowing how it was collected. In particular, the decision to breast feed is almost certainly related to other socioeconomic factors and we would need to investigate whether it is these rather than the breast feeding that is responsible for the reduction in the incidence of respiratory disease.

2.6 Prospective and Retrospective Sampling

In *prospective* sampling, the predictors are fixed and then the outcome is observed. In other words, in the infant respiratory disease example shown in Table 2.1, we would select a sample of newborn girls and boys whose parents had chosen a particular method of feeding and then monitor them for their first year. This is also called a *cohort study*.

In *retrospective* sampling, the outcome is fixed and then the predictors are observed. Typically, we would find infants coming to a doctor with a respiratory disease in the first year and then record their sex and method of feeding. We would also obtain a sample of respiratory disease-free infants and record their information. How these samples are obtained is important—we require that the probability of inclusion in the study is independent of the predictor values. This is also called a *case-control study*.

Since the question of interest is how the predictors affect the response, prospective sampling seems to be required. Let's focus on just boys who are breast or bottle fed. The data we need is:

```
> babyfood[c(1,3),]
disease nondisease sex   food
1    77        381 Boy Bottle
3    47        447 Boy Breast
```

- Given the infant is *breast* fed, the log-odds of having a respiratory disease are log47/447=–2.25
- Given the infant is *bottle* fed, the log-odds of having a respiratory disease are log 77/381=–1.60

The difference between these two log-odds, $\Delta=-1.60-(-2.25)=0.65$, represents the increased risk of respiratory disease incurred by bottle feeding relative to breast feeding. This is the log-odds ratio.

Now suppose that this had been a retrospective study—we could compute the log-odds of feeding type given respiratory disease status and then find the difference. Notice that this would give the same result because:

Δ=log77/47–log381/447=log77/381–log47/447=0.65

This shows that a retrospective design is as effective as a prospective design for estimating Δ.

Retrospective designs are cheaper, faster and more efficient, so it is convenient that the same result may be obtained from the prospective study. This manipulation is not possible for other links. The downside to retrospective studies is that they are typically less reliable than prospective studies. Retrospective studies rely on historical records which may be of unknown accuracy and completeness. They may also rely on the memory of the subject which may be unreliable.

In most practical situations, we will also need to account for the effects of covariates X. Let π_0 be the probability that an individual is included in the study if they do *not* have the disease, while let π_1 be the probability of inclusion if they do have the disease. For a prospective study, $\pi_0=\pi_1$ because we have no knowledge of the outcome, while for a retrospective study typically π_1 is much greater than π_0. Suppose that for given x, $p^*(x)$ is the conditional probability that an individual has the disease given that he or she was included in the study, while $p(x)$ is the unconditional probability that he or she has the disease as we would obtain from a prospective study. Now by Bayes theorem:

$$p^*(x) = \frac{\pi_1 p(x)}{\pi_1 p(x) + \pi_0(1 - p(x))}$$

which can be rearranged to show that:

$$\text{logit}(p^*(x)) = \log \frac{\pi_1}{\pi_0} + \text{logit}(p(x))$$

So the only difference between the retrospective and the prospective study would be the difference in the intercept: $\log(\pi_1/\pi_0)$. Generally π_1/π_0 would not be known, so we would not be able to estimate β_0, but knowledge of the other β would be most important since this can be used to assess the *relative* effect of the covariates. We could not, however, estimate the absolute effect. This does not work for other links such as the probit.

2.7 Choice of Link Function

We must choose a link function to specify a binomial regression model. It is usually not possible to make this choice based on the data alone. For regions of moderate p, that is not close to zero or one, the link functions we have proposed are quite similar and so a very large amount of data would be necessary to distinguish between them. Larger differences are apparent in the tails, but for very small p, one needs a very large amount

of data to obtain just a few successes, making it expensive to distinguish between link functions in this region. So usually, the choice of link function is made based on assumptions derived from physical knowledge or simple convenience. We now look at some of the advantages and disadvantages of the three proposed link functions and what motivates the choice.

Bliss (1935) analyzed some data on the numbers of insects dying at different levels of insecticide concentration. We fit all three link functions:

```
> data (bliss)
> bliss
  dead  alive conc
1    2     28    0
2    8     22    1
3   15     15    2
4   23      7    3
5   27      3    4
> modl <- glm(cbind(dead, alive) ~ conc,
family=binomial, data=bliss)
> modp <- glm(cbind(dead, alive) ~ conc,
family=binomial(link=probit),
  data=bliss)
> modc <- glm(cbind(dead, alive) ~ conc,
family=binomial(link=cloglog),
  data=bliss)
```

We start by considering the fitted values:

```
> fitted(modl)
       1         2         3         4         5
0.089172  0.238323  0.500000  0.761677  0.910828
```

or from predict (modl, type="response"). These are constructed using linear predictor, η:

```
> coef(modl)[1]+coef(modl)[2]*bliss$conc
[1] -2.3238 -1.1619  0.0000  1.1619  2.3238
```

Alternatively, these values may be obtained from modl$linear.predictors or predict (modl). The fitted values are then:

```
> ilogit (modl$lin)
       1         2         3         4         5
0.089172  0.238323  0.500000  0.761677  0.910828
```

Notice the need to distinguish between predictions in the scale of the response and the link. Now compare the logit, probit and complementary log-log fits:

```
> cbind(fitted(modl),fitted(modp),fitted(modc))
      [,1]     [,2]    [,3]
1 0.089172 0.084242 0.12727
```

```
2 0.238323 0.244873 0.24969
3 0.500000 0.498272 0.45459
4 0.761677 0.752396 0.72177
5 0.910828 0.914411 0.93277
```

These are not very different, but now look at a wider range:

```
> x <- seq(-2,8,0.2)
> pl <- ilogit(modl$coef[1]+modl$coef[2]*x)
> pp <- pnorm(modp$coef[1]+modp$coef[2]*x)
> pc <- 1-exp(-exp((modc$coef[1]+modc$coef[2]*x)))
> plot(x,pl,type="1",ylab="Probability",xlab="Dose")
> lines(x,pp,lty=2)
> lines(x,pc,lty=5)
```

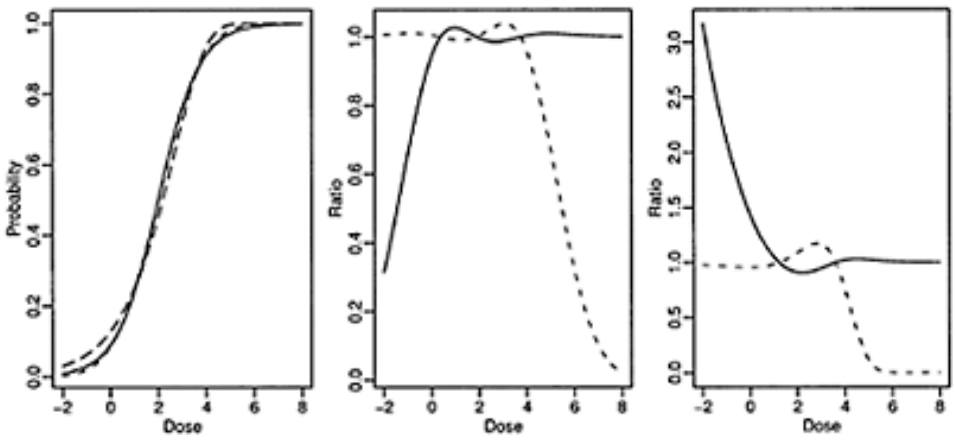

Figure 2.3 *Probit, logit and complementary log-log compared. The fitted probabilities are shown on the left. The logit fit is shown with a solid line, the probit is shown by a dotted line and the complementary log-log by a dashed line. In the central plot, the ratio of probit to logit probabilities in both tails is shown. The lower tail ratio is given by the solid line while the upper tail ratio is given by the dotted line. In the plot on the right the same information is shown for the ratio of the complementary log-log to the logit. The data range from 0 to 4. We see that the links are similar in this range*

and only begin to diverge as we extrapolate.

The lines in the left panel of Figure 2.3 do not seem very different, but look at the relative differences:

```
> matplot(x,cbind(pp/pl,(1-pp)/(1-
pl)),type="1",xlab="Dose",ylab="Ratio")
> matplot(x,cbind(pc/pl,(1-pc)/(1-
pl)),type="1",xlab="Dose",ylab="Ratio")
```

as they appear in the second and third panels of Figure 2.3. We see that the probit and logit differ substantially in the tails. The same phenomenon is observed for the complementary log-log. This is problematic since the former plot indicates it would be difficult to distinguish between the two using the data we have. This is an issue in trials of potential carcinogens and other substances that must be tested for possible harmful effects on humans. Some substances are highly poisonous in that their effects become immediately obvious at doses that might normally be experienced in the environment. It is not difficult to detect such substances. However, there are other substances whose harmful effects only become apparent at large dosages where the observed probabilities are sufficiently larger than zero to become estimable without immense sample sizes. In order to estimate the probability of a harmful effect at a low dose, it would be necessary to select an appropriate link function and yet the data for high dosages will be of little help in doing this. As Paracelsus (1493–1541) said, "All substances are poisons; there is none which is not a poison. The right dose differentiates a poison."

A good example of this problem is asbestos. Information regarding the harmful effects of asbestos derives from historical studies of workers in industries exposed to very high levels of asbestos dust. However, we would like to know the risk to individuals exposed to low levels of asbestos dust such as those found in old buildings. It is virtually impossible to accurately determine this risk. We cannot accurately measure exposure or outcome. This is not to argue that nothing should be done, but that decisions should be made in recognition of the uncertainties.

In summary, the default choice is the logit link. There are three advantages: it leads to simpler mathematics due the intractability of Φ; it is easier to interpret using odds and it allows easier analysis of retrospectively sampled data.

2.8 Estimation Problems

Estimation using the Fisher scoring algorithm, described in Section 6.2, is usually fast. However, difficulties can sometimes arise. When convergence fails, it is sometimes due to a problem exhibited by the following dataset. Urinary androsterone (androgen) and etiocholanolone (estrogen) values were recorded from 26 healthy males by Margolese (1970). The data were also analyzed by Hand (1981). We start by plotting the data as shown in Figure 2.4:

```
> data(hormone)
> plot(estrogen ~
androgen,data=hormone,pch=as.character(orientation))
```

We now fit a binomial model to see if the orientation can be predicted from the two hormone values. Notice that when the response is binary, we can use it directly as the response variable in the glm function:

```
> modl <- glm(orientation ~ estrogen + androgen,
hormone, family=binomial)
Warning messages:
1: Algorithm did not converge in: glm.fit(x = X, y = Y,
   weights = weights, start = start, etastart =
etastart,
2: fitted probabilities numerically 0 or 1 occurred in:
   glm.fit(x = X, y = Y, weights = weights, start =
start,
   etastart = etastart,
```

We see that there were problems with the convergence. A look at the summary reveals further evidence:

```
> summary(modl)
Coefficients:
            Estimate Std. Error  z value Pr(>|z|)
(Intercept)    -84.5   136095.1 -0.00062        1
estrogen       -90.2    75911.0 -0.00119        1
```

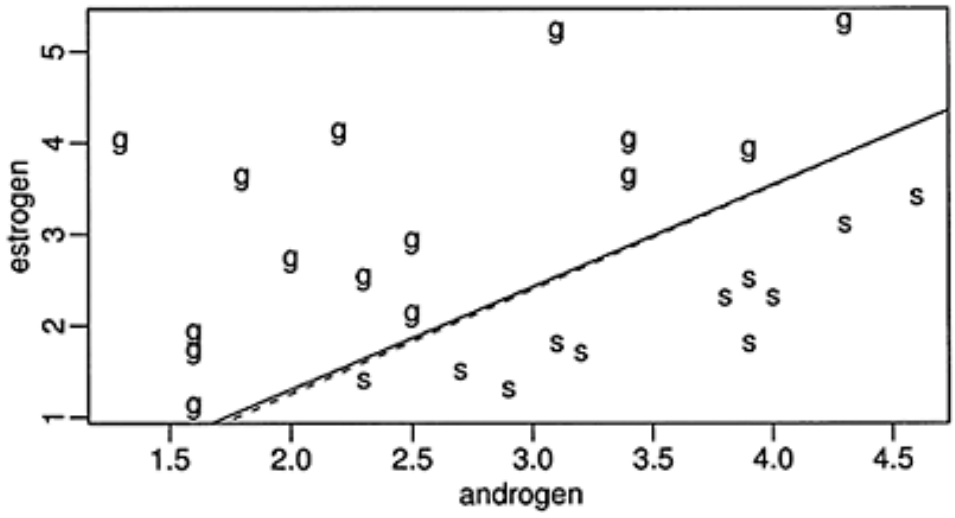

Figure 2.4 *Levels of androgen and estrogen for 15 homosexual (g) and 11 heterosexual (s) males. Solid line shows predictions from g 1m fit that*

correspond to p=1/2. The dotted line is equivalent from brlr.

```
androgen       100.9    92755.6  0.00109       1
(Dispersion parameter for binomial family taken to be
1)
    Null deviance: 3.5426e+01  on 25  degrees of
freedom
Residual deviance: 2.3229e-09  on 23  degrees of
freedom
AIC: 6
Number of Fisher Scoring iterations: 25
```

Notice that the residual deviance is extremely small indicating a very good fit and yet none of the predictors are significant due to the high standard errors. We see that the maximum default number of iterations (25) has been reached. A look at the data reveals the reason for this. We see that the two groups are *linearly separable* so that a perfect fit is possible. We can compute the line separating the groups by finding the line that corresponds to p=1/2 which is when the logit is zero:

```
> abline(-84.5/90.2,100.9/90.2)
```

We suffer from an embarrassment of riches in this example—we can fit the data perfectly. Unfortunately, this results in unstable estimates of the parameters and their standard errors and would (probably falsely) suggest that perfect predictions can be made. An alternative fitting approach might be considered in such cases called *exact logistic regression*. See Cox (1970) and the work of Cyrus Mehta, for example: Mehta and Patel (1995). Currently, there are no comprehensive packages for such exact methods in R, although it is available in products such as LogExact©.

An alternative to exact methods is the bias reduction method of Firth (1993). For the MLE, $E\hat{\beta} \neq \beta$ and indeed a sensible unbiased estimator would be difficult to obtain. Firth's method removes the $O(1/n)$ term from the asymptotic bias of estimated coefficients. These estimates have the advantage of always being finite:

```
> library(brlr)
> modb <- brlr(orientation ~ estrogen + androgen,
hormone,
  family=binomial)
> summary(modb)
Coefficients:
            Value   Std. Error t value
(Intercept) -3.650   2.910     -1.254
estrogen    -3.586   1.499     -2.393
androgen     4.074   1.621      2.513
Deviance: 3.70
Penalized deviance:  4.184
Residual df: 23
```

We can see that this results in significant predictors which we expect given Figure 2.4. Although the fit appears, judging from the coefficients, to be different from the glm result, it is effectively very close as we can see by plotting the line corresponding to p=1/2:

```
> abline(-3.65/3.586,4.074/3.586,lty=2)
```

Instability in parameter estimation will also occur in datasets that approach linear separability. Care will be needed in such cases.

2.9 Goodness of Fit

The deviance is one measure of how well the model fits the data, but there are alternatives. The Pearson's X^2 statistic takes the general form:

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed counts and E_i are the expected counts for case i. For a binomial response, we count the number of successes for which 0_i=y_i while $E_i = n_i\hat{p}_i$ and failures for which O_i=n_i–y_i and $E_i = n_i(1-\hat{p}_i)$ which results in:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i(1-\hat{p}_i)}$$

If we define *Pearson residuals* as:

$$r_i^P = (y_i - n_i\hat{p}_i)/\sqrt{\text{var}\,\hat{y}_i}$$

which can be viewed as a type of standardized residual, then $X^2 = \sum_{i=1}^{n}(r_i^P)^2$. So the Pearson's X^2 is analogous to the residual sum of squares used in normal linear models.

The Pearson X^2 will typically be close in size to the deviance and can be used in the same manner. Alternative versions of the hypothesis tests described above might use the X^2 in place of the deviance with the same approximate null distributions.

However, some care is necessary because the model is fit to minimize the deviance and not the Pearson's X^2. This means that it is possible, although unlikely, that the X^2 could increase as a predictor is added to the model. X^2 can be computed like this:

```
> modl <- glm(cbind(dead,alive) ~ conc,
family=binomial, data=bliss)
> sum(residuals(modi,type="pearson")^2)
[1] 0.36727
> deviance(modl)
[1] 0.37875
```

As can be seen, there is little difference here between X^2 and the deviance.

The proportion of variance explained or R^2 is a popular measure of fit for normal linear models. We might consider applying the same concept to binomial regression models by using the proportion of deviance explained. However, a better statistic is due to Naglekerke (1991):

$$R^2 = \frac{1-(\hat{L}_0/\hat{L})^{2/n}}{1-\hat{L}_o^{2/n}} = \frac{1-\exp((D-D_{null})/n)}{1-\exp(-D_{null}/n)}$$

where n is the number of binary observations and $\hat{L}_0$ is the maximized likelihood under the null. The numerator can be seen as a ratio of the relative likelihood with the $1/n$ power having the effect of a geometric mean on the observations. The denominator simply normalizes so that $0 \le R^2 \le 1$. For example, for the Bliss insect data, the R^2 is:

```
> (1-exp((modl$dev-modl$null)/150))/(1-exp(-
modl$null/150))
[1] 0.99532
```

Notice that we have used n=150 as there are 5 covariate class with 30 observations each. We can see that this is a very good fit.

2.10 Prediction and Effective Doses

Sometimes we wish to predict the outcome for given values of the covariates. For binary data this will mean estimating the probability of success. For given covariates x_0, $\hat{\eta} = x_0\hat{\beta}$ with variance given by $x_0^T(X^TWX)^{-1}x_0$. Approximate confidence intervals may be obtained using a normal approximation. To get an answer in the probability scale, it will be necessary to transform back using the inverse of the link function. We predict the response for the insect data:

```
> data(bliss)
> modl <- glm(cbind(dead, alive) ~ conc,
family=binomial,data=bliss)
> lmodsum <- summary(modl)
```

We show how to predict the response at dose of 2.5:

```
> x0 <- c(1, 2.5)
> eta0 <- sum(x0*coef(modl))
> ilogit(eta0)
[1] 0.64129
```

A 64% predicted chance of death at this dose—now compute a 95% confidence interval (CI) for this probability. First, extract the variance matrix of the coefficients:

```
> (cm <- lmodsum$cov.unsealed)
```

```
             (Intercept)       conc
(Intercept)    0.174630 -0.065823
conc          -0.065823  0.032912
```

The standard error on the logit scale is then:

```
> se <- sqrt(t(x0) %*% cm %*% x0)
```

so the CI on the probability scale is:

```
> ilogit(c(eta0-1.96*se,eta0+1.96*se))
[1] 0.53430 0.73585
```

A more direct way of obtaining the same result is:

```
> predict(modi,newdata=data.frame(conc=2.5),se=T)
$fit
[1] 0.58095
$se.fit
[1] 0.2263
> ilogit(c(0.58095-1.96*0.2263, 0.58095+1.96*0.2263))
[1] 0.53430 0.73585
```

Note that in contrast to the linear regression situation, there is no distinction possible between confidence intervals for a future observation and those for the mean response. Now we try predicting the response probability at the low dose of –5:

```
> x0 <- c(1,-5)
> se <- sqrt(t(x0) %*% cm %*% x0)
> eta0 <- sum(x0*1mod$coef)
> ilogit(c(eta0-1.96*se, eta0+1.96*se))
[1] 2.3577e-05 3.6429e-03
```

This is not a wide interval in absolute terms, but in relative terms, it certainly is. The upper limit is about 100 times larger than the lower limit.

Logistic regression models have been widely used for classification purposes. Depending on whether $\hat{p}$ is greater or less than 0.5, the case may be classified as a success or failure. In cases where the losses due to misclassification are not symmetrical, such as in disease diagnosis, critical values other than 0.5 should be used. Another example is in credit scoring. When financial institutions decide whether to make a loan, it is helpful to estimate the probability that a given borrower will default. A logistic regression model is one way in which this probability can be estimated using past financial data.

When there is a single (continuous) covariate or when other covariates are held fixed, we sometimes wish to estimate the value of x corresponding to a chosen p. For example we may wish to determine which dose, x, will lead to a probability of success p. ED50 stands for the *effective dose* for which there will be a 50% chance of success. When the

objective is to kill the subjects or determine toxicity, as when using insecticides, the term LD50 would be used. LD stands for *lethal dose*. Other percentiles are also of interest. For a logit link, we can set p=1/2 and then solve for x to find:

$$\widehat{\text{ED50}} = -\hat{\beta}_0/\hat{\beta}_1$$

Using the Bliss data, the LD50 is:

```
> (ld50 <- -lmod$coef[1]/lmod$coef[2])
(Intercept)
          2
```

To determine the standard error, we can use the delta method. The general expression for the variance of $g(\hat{\theta})$ for multivariate θ is given by

$$\text{var } g(\hat{\theta}) \approx g'(\hat{\theta})^T \text{var } \hat{\theta} g'(\hat{\theta})$$

which, in this example, works out as:

```
> dr <- c(-1/lmod$coef[2],lmod$coef[1]/lmod$coef[2]^2)
> sqrt(dr %*% lmodsum$cov.un %*% dr)[,]
[1] 0.17844
```

So the 95% CI is given by:

```
> c(2-1.96*0.178, 2+1.96*0.178)
[1] 1.6511 2.3489
```

Other levels may be considered—the effective dose x_p for probability of success p is:

$$x_p = \frac{\text{logit}(p) - \beta_0}{\beta_1}$$

So, for example:

```
> ed90 <- (logit(0.9)-lmod$coef[1])/lmod$coef[2]
> ed90
(Intercept)
     3.8911
```

More conveniently, we may use the dose. p function in the MASS package:

```
> library(MASS)
> dose.p(lmod,p=c(0.5,0.9))
            Dose      SE
p = 0.5:2.0000 0.17844
p = 0.9:3.8911 0.34499
```

2.11 Overdispersion

If the binomial GLM model specification is correct, we expect that the residual deviance will be approximately distributed χ^2 with the appropriate degrees of freedom. Sometimes, we observe a deviance that is much larger than would be expected if the model were correct. We must then determine which aspect of the model specification is incorrect.

The most common explanation is that we have the wrong structural form for the model. We have not included the right predictors or we have not transformed or combined them in the correct way. We have a number of ways of determining the importance of potential additional predictors and diagnostics for determining better transformations—see Section 6.4. Suppose, however, that we are able to exclude this explanation. This is difficult to achieve, but when we have only one or two predictors, it is feasible to explore the model space quite thoroughly and be sure that there is not a plausible superior model formula.

Another common explanation for a large deviance is the presence of a small number of outliers. Fortunately, these are easily checked using diagnostic methods explained more fully in Section 6.4. When larger numbers of points are identified as outliers, they become unexceptional, and we might more reasonably conclude that there is something amiss with the error distribution.

Sparse data can also lead to large deviances. In the extreme case of a binary response, the deviance is not even approximately χ^2. In situations where the group sizes are simply small, the approximation is poor. Because we cannot judge the fit using the deviance, we shall exclude this case from further consideration in this section.

Having excluded these other possibilities, we might explain a large deviance by deficiencies in the random part of the model. A binomial distribution for Y arises when the probability of success p is independent and identical for each trial within the group. If the group size is m, then var $Y=mp(1-p)$ if the binomial assumptions are correct. However, the assumptions are broken, the variance may be greater. This is *overdispersion.* In rarer cases, the variance is less and *underdispersion* results.

There are two main ways that overdispersion can arise—the independent or identical assumptions can be violated. We look at the constant p assumption first. It is easy to see how there may be some unexplained heterogeneity within a group that might lead to some variation in p. For example, in the shuttle disaster case study of Section 2.1, the position of the O-ring on the booster rocket may have some effect on the failure probability. Yet this variable was not recorded and so we cannot include it as a predictor. Heterogeneity can also result from clustering. Suppose a population is divided into clusters, so that when you take a sample, you actually get a sample of clusters. This would be common in epidemiological applications.

Let the sample size be m, the cluster size be k and the number of clusters be $l=m/k$. Let the number of successes in cluster i be $Z_i \sim B(k, p_i)$. Now suppose that p_i is a random variable such that $Ep_i=p$ and var $p_i=\tau^2 p(1-p)$. Let the total number of successes be $Y=Z_{1+}\ldots+Z_l$. Then:

$$EY = \sum EZ_i = \sum_{i=1}^{l} kp = mp$$

as in the standard case, but:

$$\text{var } Y = \sum \text{var } Z_i = \sum \{E(\text{var } (Z_i|p_i)) + \text{var } (E(Z_i|p_i))\} = 1 + (k-1)\tau^2 mp(1-p)$$

So Y is overdispersed since $1+(k–1)\tau^2 \geq 1$. Notice that in the sparse case, m=1, and this problem cannot arise.

Overdispersion can also result from dependence between trials. If the response has a common cause, say a disease is influenced by genes, the responses will tend to be positively correlated. For example, subjects in human or animal trials may be influenced in their response by other subjects. If the food supply is limited, the probability of survival of an animal may be increased by the death of others. This circumstance would result in underdispersion.

The simplest approach for modeling overdispersion is to introduce an additional dispersion parameter, σ^2. In the standard binomial case $\sigma^2 = \phi = 1$. We now let σ^2 vary and estimate using the data. Notice the similarity to linear regression. The dispersion parameter may be estimated using:

$$\hat{\sigma}^2 = \frac{X^2}{n-p}$$

Using the deviance in place of the Pearson's X^2 is not recommended as it may not be consistent. The estimation of β is unaffected since σ^2 does not change the mean response but:

$$\hat{\text{var}}\hat{\beta} = \hat{\sigma}^2 (X^T \hat{W} X)^{-1}$$

So we need to scale up the standard errors by a factor of $\hat{\sigma}$.

We cannot use the difference in deviances when comparing models, because the test statistic will be distributed $\sigma^2\chi^2$. Since σ^2 is not known and must be estimated in the overdispersion situation, an F-statistic must be used:

$$F = \frac{(D_{small} - D_{large})/(df_{small} - df_{large})}{\hat{\sigma}^2}$$

This statistic is only an approximately F distributed, in contrast to the Gaussian case.

This dispersion parameter method is only appropriate when the covariate classes are roughly equal in size. If not, more sophisticated methods should be used. One such approach uses the beta-binomial distribution where we assume that p follows a beta distribution. This approach is discussed in Williams (1982) and Crowder (1978) and can be implemented using the aod package in R.

In Manly (1978), an experiment is reported where boxes of trout eggs were buried at five different stream locations and retrieved at four different times, specified by the number of weeks after the original placement. The number of surviving eggs was recorded. The box was not returned to the stream. The data is also analyzed by Hinde and Demetrio (1988). We can construct a tabulation of the data by:

```
> data(troutegg)
> ftable(xtabs(cbind(survive,total)
location+period,troutegg))
```

```
                 survive total
location period
1        4            89    94
         7            94    98
         8            77    86
         11          141   155
2        4           106   108
         7            91   106
         8            87    96
         11          104   122
3        4           119   123
         7           100   130
         8            88   119
         11           91   125
4        4           104   104
         7            80    97
         8            67    99
       11      111    132
5      4         49     93
       7         11    113
       8         18     88
       11         0    138
```

Notice that in one case, all the eggs survive, while in another, none of the eggs survive. We now fit a binomial GLM for the two main effects:

```
> bmod <- glm(cbind(survive,total-survive) ~
location+period,
  family=binomial,troutegg)
> bmod
Coefficients:
(Intercept)   location2     location3     location4   l
ocation5
      4.636      -0.417        -1.242        -
0.951      -4.614
    period7     period8      period11
     -2.170      -2.326        -2.450
Degrees of Freedom:19 Total (i.e. Null); 12 Residual
Null Deviance:    1020
Residual Deviance:64.5 AIC: 157
```

The deviance of 64.5 on 12 degrees of freedom seems to show that this model does not fit. Before we conclude that there is overdispersion, we need to eliminate other potential explanations. With about 100 eggs in each box, we have no problem with sparseness, but we do need to check for outliers and look at the model formula. A half-normal plot of the residuals is a good way to check for outliers:

- `halfnorm(residuals(bmod))`

The half-normal plot is shown in the left panel of Figure 2.5. No single outlier is apparent. Perhaps one can discern a larger number of residuals which seem to follow a more dispersed distribution than the rest.

We can also check whether the predictors are correctly expressed by plotting the *empirical logits*. These are defined as:

$$\log\left(\frac{y+1/2}{m-y+1/2}\right)$$

The halves are added to prevent infinite values for groups consisting of all successes or failures. We now construct an interaction plot of the empirical logits:

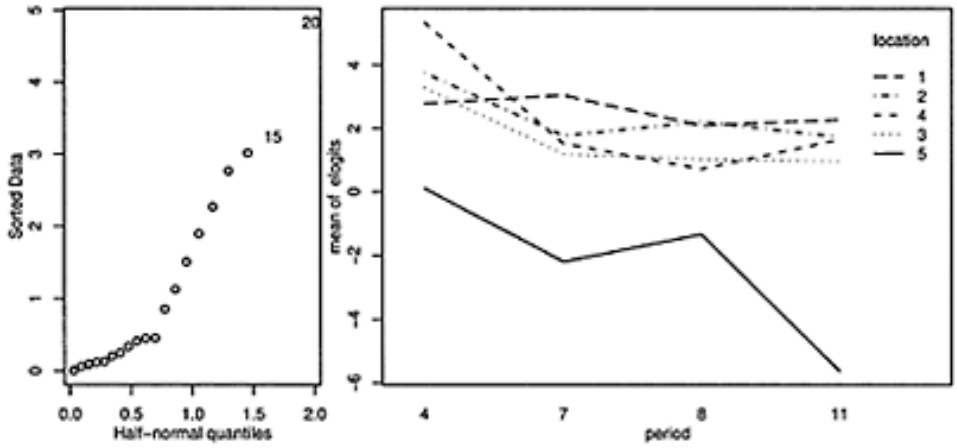

Figure 2.5 *Diagnostic plots for the trout egg model. A half-normal plot of the residuals is shown on the left and an interaction plot of the empirical logits is shown on the right.*

```
> elogits <-
log((troutegg$survive+0.5)/(troutegg$total-
  troutegg$survive+0.5))
>
with(troutegg,interaction.plot(period,location,elogits)
)
```

Interaction plots are always difficult to interpret conclusively, but there is no obvious sign of large interactions. So there is no evidence that the linear model is inadequate. We do not have any outliers and the functional form of the model appears to be suitable, but the deviance is still larger than should be expected. Having eliminated these more obvious causes as the source of the problem, we may now put the blame on overdispersion.

Possible reasons for the overdispersion include inhomogeneous trout eggs, variation in the experimental procedures or unknown variables affecting survival.

We can estimate the dispersion parameter as:

```
> (sigma2 <- sum(residuals(bmod,type="pearson")^2) /12)
[1] 5.3303
```

We see that this is substantially larger than one as it would be in the standard binomial GLM. We can now make F-tests on the predictors using:

```
> drop1(bmod,scale=sigma2,test="F")
Single term deletions
scale:  5.3303
         Df  Deviance AIC  F value   Pr(F)
<none>             64 157
location 4       914 308     39.5 8.1e-07
period   3       229 182     10.2  0.0013
Warning message:
F test assumes quasibinomial family in:
dropl.glm(bmod, scale = sigma2, test = "F")
```

We see that both terms are clearly significant. It is necessary to specify the scale argument using the estimated value of σ^2. If this argument is omitted, the deviance will be used in the estimation of the dispersion parameter. For this particular dataset, it makes very little difference, but in some cases, using the deviance to estimate the dispersion gives inconsistent results. The warning message reminds us that the use of free dispersion parameter results in a model that is no longer a true binomial GLM, but rather what is known as a *quasi-binomial* GLM. More on such models may be found in Section 7.4.

No goodness of fit test is possible because we have a free dispersion parameter. We can use the dispersion parameter to scale up the estimates of the standard error as in:

```
> summary (bmod, dispersion=sigma2 )
Coefficients:
            Estimate Std.   Error z value Pr(>|z|)
(Intercept)    4.636        0.649    7.14  9.5e-13
location2     -0.417        0.568   -0.73    0.463
location3     -1.242        0.507   -2.45    0.014
location4     -0.951        0.528   -1.80    0.072
location5     -4.614        0.578   -7.99  1.4e-15
period7       -2.170        0.550   -3.94  8.1e-05
period8       -2.326        0.561   -4.15  3.4e-05
period11      -2.450        0.540   -4.53  5.8e-06
```

2.12 Matched Case-Control Studies

In a case-control study, we try to determine the effect of certain risk factors on the outcome. We understand that there are other confounding variables that may affect the

outcome. One approach to dealing with these is to measure or record them, include them in the logistic regression model as appropriate and thereby control for their effect. But this method requires that we model these confounding variables with the correct functional form. This may be difficult. Also, making an appropriate adjustment is problematic when the distribution of the confounding variables is quite different in the cases and controls. So we might consider an alternative where the confounding variables are explicitly adjusted for in the design.

In a *matched case-control study,* we match each case (diseased person, defective object, success, etc.) with one or more controls that have the same or similar values of some set of potential confounding variables. For example, if we have a 56-year-old, Hispanic male case, we try to match him with some number of controls who are also 56-year-old Hispanic males. This group would be called a *matched set*. Obviously, the more confounding variables one specifies, the more difficult it will be to make the matches. Loosening the matching requirements, for example, accepting controls who are 50-60 years old might be necessary. Matching also gives us the possibility of adjusting for confounders that are difficult to measure. For example, suppose we suspect an environmental effect on the outcome. However, it is difficult to measure exposure, particularly when we may not know which substances are relevant. We could match subjects based on their place of residence or work. This would go some way to adjusting for the environmental effects.

Matched case-control studies also have some disadvantages apart from the difficulties of forming the matched sets. One loses the possibility of discovering the effects of the variables used to determine the matches. For example, if we match on sex, we will not be able to investigate a sex effect. Furthermore, the data will likely be far from a random sample of the population of interest. So although relative effects may be found, it may be difficult to generalize to the population.

Sometimes, cases are rare but controls are readily available. A1: M design has M controls for each case. M is typically small and can even vary in size from matched set to matched set due to difficulties in finding matching controls and missing values. Each additional control yields a diminished return in terms of increased efficiency in estimating risk factors—it is usually not worth exceeding M=5.

For individual i in the j^{th} matched set, we also observe a covariate vector x_{ij} which will include the risk factors of interest plus any other variables that we may wish to adjust for, but were unable for various reasons to include among the criteria used to match the sets. It is important that the decision to include a subject in the study be independent of the risk factors as in the unmatched case-control studies. Suppose we have n matched sets and that we take i=0 to represent the case and i=1,…, M to represent the controls. We propose a logistic regression model of the following form:

$$\text{logit}(p_j(x_{ij}))=\alpha_j+\beta^T x_{ij}$$

The α_j models the effect of the confounding variables in the j^{th} matched set. Given a matched set j of M+1 subjects known to have one case and M controls, the conditional probability of the observed outcome, or, in other words, that subject i=0 is the case and the rest are controls is:

$$\frac{\exp \beta^T x_{0j}}{\sum_{i=0}^{M} \exp \beta^T x_{ij}}$$

Notice that α_j cancels out in this expression. We may then form the conditional likelihood for the model by taking the product over all the matched sets:

$$L(\beta) = \prod_{j=1}^{n} \{1 + \sum_{i=1}^{M} \exp[\beta^T (x_{ij} - x_{0j})]\}^{-1}$$

We may now employ standard likelihood methods to make inference—see Breslow (1982) for details. The likelihood takes the same form as that used for the proportional hazards model used in survival analysis. This is convenient because we may use software developed for those models as we demonstrate below. Since the as are not estimated, we cannot make predictions about individuals, but only make statements about the relative risks as measured by the βs. This same restriction also applies to the unmatched model, so this is nothing new.

In Le (1998), a matched case-control study is presented concerning the association between x-rays and childhood acute myeloid leukemia. The sets are matched on age, race and county of residence. For the most part, there is only one control for each case, but there are a few instances of two controls. We start with a look at the data:

```
> data(amlxray)
> head(amlxray)
    ID disease Sex downs age Mray MupRay MlowRay Fray
Cray CnRay
1
7004       1   F    no    0  no     no      no    no   n
o     1
2
7004       0   F    no   0   no     no      no    no   n
o     1
3
7006       1   M    no   6   no     no      no    no  ye
s     3
4
7006       0   M    no   6   no     no      no    no  ye
s     2
5
7009       1   F    no   8   no     no      no    no   n
o     1
6
7009       0   F    no   8   no     no      no    no   n
o     1
```

Only the age is presented here as one of the matching variables. In the three sets shown here, we see that both subjects have the same age and the first is the case and the second is the control. The other variables are risk factors of interest.

Down syndrome is known to be a risk factor. There are only seven such subjects in the dataset:

```
> amlxray[amlxray$downs=="yes",1:4]
      ID disease Sex downs
7   7010       1   M   yes
17  7018       1   F   yes
78  7066       1   F   yes
88  7077       1   M   yes
173 7146       1   F   yes
196 7176       1   F   yes
210 7189       1   F   yes
```

We see that all seven subjects are cases. If we include this variable in the regression, its coefficient is infinite. Given this and the prior knowledge, it is simplest to exclude all these subjects and their associated matched subjects:

```
> (ii <- which(amlxray$downs=="yes"))
[1]   7  17  78  88 173 196 210
> ramlxray <- amlxray[-c(ii,ii+1),]
```

The variables Mray, MupRay and MlowRay record whether the mother has ever had an x-ray, ever had an upper body x-ray and ever had a lower body x-ray, respectively. These variables are closely associated, so we will pick just Mray for now and investigate the others more closely if indicated. We will also use CnRay, a four-level ordered factor grouping the number of x-rays that the child has received in preference to Cray which merely indicates whether the child has ever had an x-ray.

The clogit function fits a conditional logit model. Since the likelihood is identical with that from a proportional hazards model, it may be found in the survival package. The matched sets must designated by the strata function:

```
> library (survival)
> cmod <- clogit(disease ~
Sex+Mray+Fray+CnRay+strata(ID),ramlxray)
> summary(cmod)
          coef exp(coef) se(coef)      z      p
SexM     0.156      1.17    0.386  0.405 0.6900
Mrayyes  0.228      1.26    0.582  0.391 0.7000
Frayyes  0.693      2.00    0.351  1.974 0.0480
CnRay.L  1.941      6.96    0.621  3.127 0.0018
CnRay.Q -0.248      0.78    0.582 -0.426 0.6700
CnRay.C -0.580      0.56    0.591 -0.982 0.3300
        exp(coef) exp(-coef) lower .95 upper .95
SexM         1.17      0.855     0.549      2.49
Mrayyes      1.26      0.796     0.401      3.93
Frayyes      2.00      0.500     1.005      3.98
CnRay.L      6.96      0.144     2.063     23.51
CnRay.Q      0.78      1.281     0.249      2.44
CnRay.C      0.56      1.786     0.176      1.78
```

```
Rsquare= 0.089   (max possible= 0.499)
Likelihood ratio test= 20.9  on 6 df,  p=0.00192
Wald test            = 14.5 on 6 df, p=0.0246
Score (logrank) test = 18.6 on 6 df, p=0.0049
```

The overall tests for significance of the predictors indicate that at least some of the variables are significant. We see that Sex and whether the mother had an x-ray are not significant. There seems little point in investigating the other x-ray variables associated with the mother. An x-ray on the father is marginally significant. However, the x-ray on the child has the clearest effect. Because this is an ordered factor, we have used linear, quadratic and cubic contrasts. Only the linear effect is significant.

The second table of coefficients gives us information helpful for interpreting the size of the effects. We see that the father having had an x-ray doubles the odds of the disease. The interpretation of the number of x-rays of the child is more difficult to interpret because of the coding. Since we have found only a linear effect, we convert CnRay to the numerical values 1–4 using unclass. We also drop the insignificant predictors:

```
> cmodr <- clogit(disease ~
Fray+unclass(CnRay)+strata(ID),ramlxray)
> summary(cmodr)
                coef exp(coef) se(coef)    z       p
Frayyes        0.670      1.96    0.344 1.95 0.05100
unclass(CnRay) 0.814      2.26    0.237 3.44 0.00058
               exp(coef) exp(-coef) lower .95 upper .95
Frayyes             1.96      0.512     0.996      3.84
unclass(CnRay)      2.26      0.443     1.419      3.59
```

The codes for Cnray are 1=none, 2=1 or 2 x-rays, 3=3 or 4 x-rays and 4=5 or more x-rays. We see that the odds of the disease increase by a factor of 2.26 as we move between adjacent categories. Notice that the father's x-ray variable is now just insignificant in this regression underlining its borderline status.

An incorrect analysis of this data ignores the matching structure and simply uses a binomial GLM:

```
> gmod <- glm(disease ~
Fray+unclass(CnRay),family=binomial, ramlxray)
> summary(gmod)
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.162      0.301   -3.86 0.00011
Frayyes           0.500      0.308    1.63 0.10405
unclass(CnRay)    0.601      0.177    3.39 0.00071
```

The results are somewhat different.

Although we have found an effect due to x-rays of the child, we cannot conclude the effect is causal. After all, subjects only have x-rays when something is wrong, so it is quite possible that the x-rays are linked to some unknown causal factor.

Other examples of matched data may be found in Section 4.3.

Further Reading: See books by Collett (2003), Hosmer and Lemeshow (2000), Cox (1970), Harrell (2001), Menard (2002), Christensen (1997) and Kleinbaum and Klein (2002).

Exercises

1. The question concerns data from a case-control study of esophageal cancer in Ileet-Vilaine, France. The data is distributed with R and may be obtained along with a description of the variables by:

```
> data(esoph)
> help(esoph)
```

   (a) Fit a binomial GLM with interactions between all three predictors. Use backward elimination to simplify the model as far as is reasonable.
   (b) All three factors are ordered and so special contrasts have been used appropriate for ordered factors involving linear, quadratic and cubic terms. Further simplification of the model is possible by eliminating some of these terms. Use the unclass function to convert some or all factors to a numerical representation and show how the model may be simplified.
   (c) Does your final model fit the data? Is the test you make accurate for this data?
   (d) Check for outliers in your final model.
   (e) What is the predicted effect of moving one category higher in alcohol consumption?
   (f) Compute a 95% confidence interval for this predicted effect.
   (g) Bearing in mind that this is a case-control study, what can be said about the predicted probability that a 25-year-old who does not smoke or drink will get esophageal cancer?

2. The dataset wbcd comes from a study of breast cancer in Wisconsin. There are 681 cases of potentially cancerous tumors of which 238 are actually malignant. Determining whether a tumor is really malignant is traditionally determined by an invasive surgical procedure. The purpose of this study was to determine whether a new procedure called fine needle aspiration, which draws only a small sample of tissue, could be effective in determining tumor status.

   (a) Fit a binomial regression with Class as the response and the other nine variables as predictors. Report the residual deviance and associated degrees of freedom. Can this information be used to determine if this model fits the data? Explain.
   (b) Use AIC as the criterion to determine the best subset of variables. (Use the step function.)
   (c) Use the reduced model to predict the outcome for a new patient with predictor variables 1, 1, 3, 2, 1, 1, 4, 1, 1 (same order as above). Give a confidence interval for your prediction.

(d) Suppose that a cancer is classified as benign if $p>0.5$ and malignant if $p<0.5$. Compute the number of errors of both types that will be made if this method is applied to the current data with the reduced model.
(e) Suppose we change the cutoff to 0.9 so that $p<0.9$ is classified as malignant and $p>0.9$ as benign. Compute the number of errors in this case. Discuss the issues in determining the cutoff.
(f) It is usually misleading to use the same data to fit a model and test its predictive ability. To investigate this, split the data into two parts—assign every third observation to a test set and the remaining two thirds of the data to a training set. Use the training set to determine the model and the test set to assess its predictive performance. Compare the outcome to the previously obtained results.

3. The National Institute of Diabetes and Digestive and Kidney Diseases conducted a study on 768 adult female Pima Indians living near Phoenix. The purpose of the study was to investigate factors related to diabetes. The data may be found in the the dataset pima.

(a) Perform simple graphical and numerical summaries of the data. Can you find any obvious irregularities in the data? If you do, take appropriate steps to correct the problems.
(b) Fit a model with the result of the diabetes test as the response and all the other variables as predictors. Can you tell whether this model fits the data?
(c) What is the difference in the odds of testing positive for diabetes for a woman with a BMI at the first quartile compared with a woman at the third quartile, assuming that all other factors are held constant? Give a confidence interval for this difference.
(d) Do women who test positive have higher diastolic blood pressures? Is the diastolic blood pressure significant in the regression model? Explain the distinction between the two questions and discuss why the answers are only apparently contradictory.
(e) Perform diagnostics on the regression model, reporting any potential violations and any suggested improvements to the model.
(f) Predict the outcome for a woman with predictor values 1, 99, 64, 22, 76, 27, 0.25, 25 (same order as in the dataset). Give a confidence interval for your prediction.

4. Aflatoxin B1 was fed to lab animals at various doses and the number responding with liver cancer recorded. The data may be found in the dataset af latoxin.

(a) Build a model to predict the occurrence of liver cancer. Compute the ED50 level.
(b) Discuss the extrapolation properties of your chosen model for low doses.

5. A study was conducted to determine the effectiveness of a new teaching method in economics. The data may be found in the dataset spector. Write a report on how well the new method works.
6. Incubation temperature can affect the sex of turtles. An experiment was conducted with three independent replicates for each temperature and the number of male and female turtles born was recorded and can be found in the turtle dataset. Check for evidence of overdispersion in a binomial model for the sex of the turtle.

7. The infert dataset from the survival package presents data from a study of infertility after spontaneous and induced abortion. Analyze and report on the factors related to infertility based on this data.