CHAPTER 5

Models for Binary Data

For binary responses, analysts usually assume a binomial distribution for the random component of a generalized linear model (GLM). From its exponential dispersion representation (4.6) in Section 4.1.2, the binomial natural parameter is the log odds, the so-called *logit*. The canonical link function for binomial GLMs is the logit, for which the model itself is referred to as *logistic regression*. This is the most important model for binary response data and has been used for a wide variety of applications. Early uses were in biomedical studies, for instance to model the effects of smoking, cholesterol, and blood pressure on the presence or absence of heart disease. The past 25 years have seen of substantial use in social science research for modeling opinions (e.g., favor or oppose legalization of same-sex marriage) and behaviors, in marketing applications for modeling consumer decisions (e.g., a choice between two products), and in finance for modeling credit-related outcomes (e.g., whether a credit card bill is paid on time).

In this chapter we focus on logistic regression and other models for binary response data. Section 5.1 presents some link functions and a latent variable model that motivates particular cases. Section 5.2 shows properties of logistic regression models and interprets its parameters. In Section 5.3 we apply GLM methods to specify likelihood equations and then conduct inference based on the logistic regression model. Section 5.4 covers model fitting. In Section 5.5 we find the deviance for binomial GLMs and discuss ways of checking the model fit. In Section 5.6 we present alternatives to logistic regression, such as the model using the *probit* link. Section 5.7 illustrates the models with two examples.

5.1 LINK FUNCTIONS FOR BINARY DATA

In this chapter, we distinguish between two sample size measures: a measure n_i for the number of Bernoulli trials that constitute a particular binomial observation, and a measure N for the number of binomial observations. We assume that $y_1, \ldots, y_N$ are

Foundations of Linear and Generalized Linear Models, First Edition. Alan Agresti.

independent binomial proportions, with $n_i y_i \sim \text{bin}(n_i, \pi_i)$. That is, y_i is the *proportion* of "successes" out of n_i independent Bernoulli trials, and $E(y_i) = \pi_i$ does not depend on n_i. Let $\boldsymbol{n} = (n_1, \ldots, n_N)$ denote the binomial sample sizes. The overall number of binary observations is $n = \sum_{i=1}^{N} n_i$.

5.1.1 Ungrouped versus Grouped Binary Data

Data files for binary data have two possible formats. For *ungrouped data*, $\boldsymbol{n} = (1, \ldots, 1)$. The data file takes this form when each observation y_i results from a single Bernoulli trial, and thus equals 0 or 1. Large-sample methods for statistical inference then apply as $N \to \infty$.

For *grouped data*, sets of observations have the same value for each explanatory variable. Most commonly this happens when all explanatory variables are categorical. Then, n_i refers to the number of observations at setting i of the explanatory variables, $i = 1, \ldots, N$. For example, in a dose–response study of the effect of various dosages of a drug on the probability of an adverse outcome, $\{n_i\}$ record the number of observations at the various dosages. For grouped data, the number N of combinations of the categorical predictors is fixed, and large-sample methods for inference and model checking apply as each $n_i \to \infty$. Under such *small-dispersion asymptotics*, as we obtain more data, the variance for each binomial observation decreases.

A grouped-data file for binary data can be converted to ungrouped form. The same maximum likelihood (ML) estimates $\hat{\boldsymbol{\beta}}$ and standard errors occur, with the same large-sample normal distributions; however, other summary measures of fit, such as the deviance, change. We will see that the grouped-data format is useful for checking model fit. An ungrouped-data file can be converted to grouped-data form only when multiple subjects share the same values for explanatory variables.

5.1.2 Latent Variable Threshold Model for Binary GLMs

A latent variable model called a *threshold model* provides motivation for families of GLMs. We express this model in terms of ungrouped data. The model assumes (1) there is an unobserved continuous response y_i^* for subject i satisfying $y_i^* = \sum_j \beta_j x_{ij} + \epsilon_i$, where $\{\epsilon_i\}$ are independent from a distribution with mean 0 and having cdf F, and (2) there is a threshold τ such that we observe $y_i = 0$ if $y_i^* \le \tau$ and $y_i = 1$ if $y_i^* > \tau$. See Figure 5.1. Then

$$P(y_i = 1) = P(y_i^* > \tau) = P\left(\sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i > \tau\right)$$

$$= 1 - P\left(\epsilon_i \le \tau - \sum_{j=1}^{p} \beta_j x_{ij}\right) = 1 - F\left(\tau - \sum_{j=1}^{p} \beta_j x_{ij}\right).$$

The data contain no information about τ, so without loss of generality we take $\tau = 0$. Likewise, an equivalent model results if we multiply all parameters by any positive

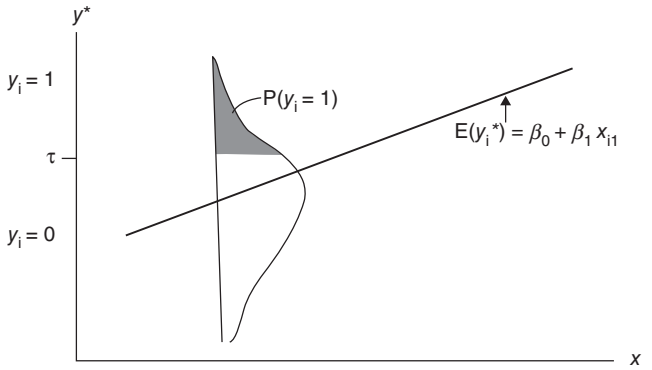

Figure 5.1 Threshold latent variable model, for which we observe $y_i = 1$ when underlying latent variable $y_i^* > \tau$.

constant, so we can take F to have a standard form with fixed variance, such as the standard normal cdf.

For the most common models, F corresponds to a pdf that is symmetric around 0, so $F(z) = 1 - F(-z)$ and

$$P(y_i = 1) = F\left(\sum_{j=1}^{p} \beta_j x_{ij}\right), \quad \text{and} \quad F^{-1}[P(y_i = 1)] = \sum_{j=1}^{p} \beta_j x_{ij}. \tag{5.1}$$

That is, models for binary data naturally take the link function to be the inverse of the standard cdf for a family of continuous distributions for a latent variable.

5.1.3 Probit, Logistic, and Linear Probability Models

When F is the standard normal cdf, the link function F^{-1} is called the *probit link* and model (5.1) is called the *probit model*. We discuss probit models in Section 5.6. A model that has a similar fit but a simpler form of link function uses the standard cdf of the *logistic distribution*,

$$F(z) = e^z/(1 + e^z).$$

Like the standard normal, the standard logistic distribution is defined over the entire real line and has a bell-shaped density function with mean 0. The model (5.1) is then the *logistic regression model*. Its link function F^{-1} is the logit.

Occasionally the identity link function is used, corresponding to F^{-1} for a uniform cdf. The model for the binomial parameter π_i for observation i,

$$\pi_i = \sum_{j=1}^{p} \beta_j x_{ij},$$

is called the *linear probability model*. This model has the awkward aspect that the linear predictor must fall between 0 and 1 for the model to generate legitimate probability values. Because of this and because in practice an S-shaped curve for which π_i very gradually approaches 0 and 1 is more plausible, linear probability models are not commonly used.

5.2 LOGISTIC REGRESSION: PROPERTIES AND INTERPRETATIONS

Next we present properties and interpretations of model parameters for logistic regression. The model has two formulations.

Logistic regression model formulas:

$$\pi_i = \frac{\exp\left(\sum_{j=1}^{p} \beta_j x_{ij}\right)}{1 + \exp\left(\sum_{j=1}^{p} \beta_j x_{ij}\right)} \quad \text{or} \quad \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{j=1}^{p} \beta_j x_{ij}. \qquad (5.2)$$

5.2.1 Interpreting β: Effects on Probabilities and on Odds

For a single quantitative x with $\beta > 0$, the curve for π_i has the shape of the cdf of a logistic distribution. Since the logistic density is symmetric, as x_i changes, π_i approaches 1 at the same rate that it approaches 0. With multiple explanatory variables, since $1 - \pi_i = [1 + \exp(\sum_j \beta_j x_{ij})]^{-1}$, π_i is monotone in each explanatory variable according to the sign of its coefficient. The rate of climb or descent increases as $|\beta_j|$ increases. When $\beta_j = 0$, y is conditionally independent of x_j, given the other explanatory variables.

How do we interpret the magnitude of β_j? For a quantitative explanatory variable, a straight line drawn tangent to the curve at any particular value describes the instantaneous rate of change in π_i at that point. Specifically,

$$\frac{\partial \pi_i}{\partial x_{ij}} = \beta_j \frac{\exp\left(\sum_j \beta_j x_{ij}\right)}{\left[1 + \exp\left(\sum_j \beta_j x_{ij}\right)\right]^2} = \beta_j \pi_i (1 - \pi_i).$$

The slope is steepest (and equals $\beta_j/4$) at a value of x_{ij} for which $\pi_i = 1/2$, and the slope decreases toward 0 as π_i moves toward 0 or 1.

How do we interpret β_j for a qualitative explanatory variable? Consider first a single binary indicator x. The model, $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$, then describes a 2×2 contingency table. For it,

$$\text{logit}[P(y = 1 \mid x = 1)] - \text{logit}[P(y = 1 \mid x = 0)] = [\beta_0 + \beta_1(1)] - [\beta_0 + \beta_1(0)] = \beta_1.$$

It follows that e^{β_1} is the *odds ratio* (Yule 1900, 1912),

$$e^{\beta_1} = \frac{P(y = 1 \mid x = 1)/[1 - P(y = 1 \mid x = 1)]}{P(y = 1 \mid x = 0)/[1 - P(y = 1 \mid x = 0)]}.$$

With multiple explanatory variables, exponentiating both sides of the equation for the logit shows that the odds $\pi_i/(1 - \pi_i)$ are an exponential function of x_j. The odds multiply by e^{β_j} per unit increase in x_j, adjusting for the other explanatory variables in the model. For example, e^{β_j} is a conditional odds ratio—the odds at $x_j = u + 1$ divided by the odds at $x_j = u$, adjusting for the other $\{x_k\}$.

It is simpler to understand the effects presented on a probability scale than as odds ratios. To summarize the effect of a quantitative explanatory variable, we could compare $P(y = 1)$ at extreme values of that variable, with other explanatory variables set at their means. This type of summary is sensible when the distribution of the data indicate that such extreme values can occur at mean values for the other explanatory variables. With a continuous variable, however, this summary can be sensitive to an outlier. So the comparison could instead use its quartiles, thus showing the change in $P(y = 1)$ over the middle half of the explanatory variable's range of observations. The data can more commonly support such a comparison.

5.2.2 Logistic Regression with Case-Control Studies

In case-control studies, y is known, and researchers look into the past to observe $\boldsymbol{x}$ as the random variable. For example, for cases of a particular type of cancer ($y = 1$) and disease-free controls ($y = 0$), a study might observe $x =$ whether the person has been a significant smoker. For 2×2 tables, we just observed that e^{β} is the odds ratio with y as the response. But, from Bayes' theorem,

$$\begin{aligned} e^{\beta} &= \frac{P(y = 1 \mid x = 1)/P(y = 0 \mid x = 1)}{P(y = 1 \mid x = 0)/P(y = 0 \mid x = 0)} \\ &= \frac{P(x = 1 \mid y = 1)/P(x = 0 \mid y = 1)}{P(x = 1 \mid y = 0)/P(x = 0 \mid y = 0)}. \end{aligned}$$

So it is possible to estimate the odds ratio in retrospective studies that sample x, for given y. More generally, with logistic regression we can estimate effects in studies for which the research design reverses the roles of x and y as response and explanatory, and the effect parameters still have interpretations as log odds ratios.

Here is a formal justification: let z indicate whether a subject is sampled ($1 =$ yes, $0 =$ no). Even though the conditional distribution of y given $\boldsymbol{x}$ is not sampled, we need a model for $P(y = 1 \mid z = 1, \boldsymbol{x})$, assuming that $P(y = 1 \mid \boldsymbol{x})$ follows the logistic model. By Bayes' theorem,

$$P(y = 1 \mid z = 1, \boldsymbol{x}) = \frac{P(z = 1 \mid y = 1, \boldsymbol{x})P(y = 1 \mid \boldsymbol{x})}{\sum_{j=0}^{1}[P(z = 1 \mid y = j, \boldsymbol{x})P(y = j \mid \boldsymbol{x})]}. \tag{5.3}$$

Now, suppose that $P(z = 1 \mid y, \boldsymbol{x}) = P(z = 1 \mid y)$ for $y = 0$ and 1; that is, for each y, the sampling probabilities do not depend on $\boldsymbol{x}$. For instance, for cases and for controls, the probability of being sampled is the same for smokers and nonsmokers. Under this assumption, substituting $\rho_1 = P(z = 1 \mid y = 1)$ and $\rho_0 = P(z = 1 \mid y = 0)$ in Equation (5.3) and dividing the numerator and denominator by $P(y = 0 \mid \boldsymbol{x})$,

$$P(y = 1 \mid z = 1, \boldsymbol{x}) = \frac{\rho_1 \exp\left(\sum_j \beta_j x_j\right)}{\rho_0 + \rho_1 \exp\left(\sum_j \beta_j x_j\right)}.$$

Then, letting $\beta_0^* = \beta_0 + \log(\rho_1/\rho_0)$,

$$\text{logit}[P(y = 1 \mid z = 1, \boldsymbol{x})] = \beta_0^* + \beta_1 x_1 + \cdots.$$

The logistic regression model holds with the same effect parameters as in the model for $P(y = 1 \mid \boldsymbol{x})$. With a case-control study we can estimate those effects but we cannot estimate the intercept term, because the data do not supply information about the relative numbers of $y = 1$ and $y = 0$ observations.

5.2.3 Logistic Regression is Implied by Normal Explanatory Variables

Regardless of the sampling design, suppose the explanatory variables are continuous and have a normal distribution, for each response outcome. Specifically, given y, suppose $\boldsymbol{x}$ has an $N(\boldsymbol{\mu}_y, \boldsymbol{V})$ distribution, $y = 0, 1$. Then, by Bayes' theorem, $P(y = 1 \mid \boldsymbol{x})$ satisfies the logistic regression model with $\boldsymbol{\beta} = \boldsymbol{V}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ (Warner 1963).

For example, in a health study of senior citizens, suppose $y =$ whether a person has ever had a heart attack and $x =$ cholesterol level. Suppose those who have had a heart attack have an approximately normal distribution on x and those who have not had one also have an approximately normal distribution on x, with similar variance. Then, the logistic regression function approximates well the curve for $P(y = 1 \mid x)$. The effect is greater when the groups' mean cholesterol levels are farther apart. If the distributions are normal but with different variances, the logistic model applies, but having a quadratic term (Exercise 5.1).

5.2.4 Summarizing Predictive Power: Classification Tables and ROC Curves

A *classification table* cross-classifies the binary response y with a prediction $\hat{y}$ of whether $y = 0$ or 1 (see Table 5.1). For a model fit to ungrouped data, the prediction for observation i is $\hat{y}_i = 1$ when $\hat{\pi}_i > \pi_0$ and $\hat{y}_i = 0$ when $\hat{\pi}_i \leq \pi_0$, for a selected cutoff π_0. Common cutoffs are (1) $\pi_0 = 0.50$, (2) the sample proportion of $y = 1$ outcomes, which each $\hat{\pi}_i$ equals for the model containing only an intercept term. Rather than using $\hat{\pi}_i$ from the model fitted to the dataset that includes y_i, it is better to make the prediction with the "leave-one-out" cross-validation approach, which bases $\hat{\pi}_i$ on the

Table 5.1 A Classification Table

	Prediction $\hat{y}$	
y	0	1
0		
1		

Cell counts in such tables yield estimates of sensitivity = $P(\hat{y} = 1 \mid y = 1)$ and specificity = $P(\hat{y} = 0 \mid y = 0)$.

model fitted to the other $n - 1$ observations. For a particular cutoff, summaries of the predictive power from the classification table are estimates of

$$\text{sensitivity} = P(\hat{y} = 1 \mid y = 1) \quad \text{and} \quad \text{specificity} = P(\hat{y} = 0 \mid y = 0).$$

A disadvantage of a classification table is that its cell entries depend strongly on the cutoff π_0 for predictions. A more informative approach considers the estimated sensitivity and specificity for all the possible π_0. The sensitivity is the *true positive rate* (tpr), and $P(\hat{y} = 1 \mid y = 0) = (1 - \text{specificity})$ is the *false positive rate* (fpr). A plot of the true positive rate as a function of the false positive rate as π_0 decreases from 1 to 0 is called a *receiver operating characteristic* (ROC) curve. When π_0 is near 1, almost all predictions are $\hat{y}_i = 0$; then, the point (fpr, tpr) $\approx$ (0, 0). When π_0 is near 0, almost all predictions are $\hat{y}_i = 1$; then, (fpr, tpr) $\approx$ (1, 1). For a given specificity, better predictive power corresponds to higher sensitivity. So, the better the predictive power, the higher the ROC curve and the greater the area under it. A ROC curve usually has a concave shape connecting the points (0, 0) and (1, 1), as illustrated by Figure 5.2.

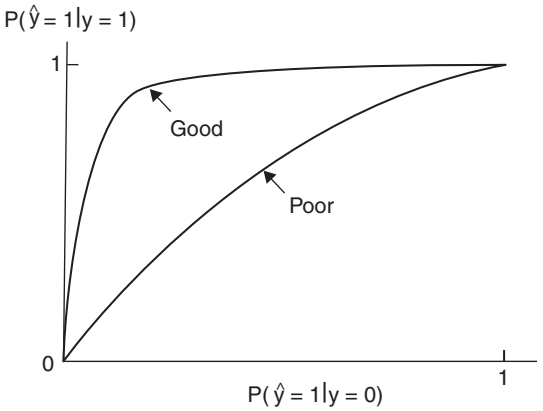

Figure 5.2 ROC curves for a binary GLM having good predictive power and for a binary GLM having poor predictive power.

The area under a ROC curve equals a measure of predictive power called the *concordance index* (Hanley and McNeil 1982). Consider all pairs of observations (i, j) for which $y_i = 1$ and $y_j = 0$. The concordance index c is the proportion of the pairwise predictions that are concordant with the outcomes, having $\hat{\pi}_i > \hat{\pi}_j$. A pair having $\hat{\pi}_i = \hat{\pi}_j$ contributes $\frac{1}{2}$ to the count of such pairs. The "no effect" value of $c = 0.50$ occurs when the ROC curve is a straight line connecting the points (0, 0) and (1, 1).

5.2.5 Summarizing Predictive Power: Correlation Measures

An alternative measure of predictive power is the correlation between the observed responses $\{y_i\}$ and the model's fitted values $\{\hat{\mu}_i\}$. This generalization of the multiple correlation for linear models is applicable for any GLM (Section 4.6.4). In logistic regression with ungrouped data, $\text{corr}(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ is the correlation between the N binary $\{y_i\}$ observations (1 or 0 for each) and the estimated probabilities. The highly discrete nature of y constrains the range of possible correlation values. A related measure estimates $\text{corr}(\boldsymbol{y}^*, \hat{\boldsymbol{\mu}})$ for the latent continuous variable for the underlying latent variable model. The square of this measure is an R^2 analog (McKelvey and Zavoina 1975) that divides the estimated variance of $\hat{y}^*$ by the estimated variance of y^*, where $\hat{y}_i^* = \sum_j \hat{\beta}_j x_{ij}$ is the same as the estimated linear predictor. The estimated variance of y^* equals the estimated variance of $\hat{y}^*$ plus the variance of ϵ in the latent variable model. For the probit latent model with standard normal error, $\text{var}(\epsilon) = 1$. For the corresponding logistic model, $\text{var}(\epsilon) = \pi^2/3 = 3.29$, the variance of the standard logistic distribution.

Such correlation measures are useful for comparing fits of different models for the same data. They can distinguish between models when the concordance index does not. For instance, with a single explanatory variable, c takes the same value for every link function that gives a monotone relationship of the same sign between x and $\hat{\pi}$.

5.3 INFERENCE ABOUT PARAMETERS OF LOGISTIC REGRESSION MODELS

The mechanics of ML estimation and model fitting for logistic regression are special cases of the GLM fitting results of Sections 4.1 and 4.5. From (4.10), the likelihood equations for a GLM are

$$\sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{\text{var}(y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 1, 2, \ldots, p.$$

For a GLM for binary data, $n_i y_i \sim \text{bin}(n_i, \pi_i)$ with $\pi_i = \mu_i = F(\sum_j \beta_j x_{ij}) = F(\eta_i)$ for some standard cdf F. Thus, $\partial \mu_i / \partial \eta_i = f(\eta_i)$ where f is the pdf corresponding to F.

Since the binomial proportion y_i has $\text{var}(y_i) = \pi_i(1-\pi_i)/n_i$, the likelihood equations are

$$\sum_{i=1}^{N} \frac{n_i(y_i - \pi_i)x_{ij}}{\pi_i(1-\pi_i)} f(\eta_i) = 0, \quad j = 1, 2, \ldots, p.$$

That is, in terms of $\boldsymbol{\beta}$,

$$\sum_{i=1}^{N} \frac{n_i \left[y_i - F\left(\sum_j \beta_j x_{ij}\right)\right] x_{ij} f\left(\sum_j \beta_j x_{ij}\right)}{F\left(\sum_j \beta_j x_{ij}\right)\left[1 - F\left(\sum_j \beta_j x_{ij}\right)\right]} = 0, \quad j = 1, 2, \ldots, p. \tag{5.4}$$

5.3.1 Logistic Regression Likelihood Equations

For logistic regression models for binary data,

$$F(z) = \frac{e^z}{1+e^z}, \quad f(z) = \frac{e^z}{(1+e^z)^2} = F(z)[1 - F(z)].$$

The likelihood equations then simplify to

$$\sum_{i=1}^{N} n_i(y_i - \pi_i)x_{ij} = 0, \quad j = 1, \ldots, p. \tag{5.5}$$

Let $\boldsymbol{X}$ denote the $N \times p$ model matrix of values of $\{x_{ij}\}$. Let $\boldsymbol{s}$ denote the binomial vector of "success" totals with elements $s_i = n_i y_i$. The likelihood equations (5.5) have the form

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{s} = \boldsymbol{X}^{\mathrm{T}}E(\boldsymbol{s}).$$

This equation illustrates the fundamental result for GLMs with canonical link function, shown in Equation 4.27, that the likelihood equations equate the sufficient statistics to their expected values.

5.3.2 Covariance Matrix of Logistic Parameter Estimators

The ML estimator $\hat{\boldsymbol{\beta}}$ has a large-sample normal distribution around $\boldsymbol{\beta}$ with covariance matrix equal to the inverse of the information matrix. From (4.13), the information matrix for a GLM has the form $\boldsymbol{\mathcal{J}} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X}$, where $\boldsymbol{W}$ is the diagonal matrix with elements

$$w_i = (\partial\mu_i/\partial\eta_i)^2/\text{var}(y_i).$$

For binomial observations, $\mu_i = \pi_i$ and $\text{var}(y_i) = \pi_i(1-\pi_i)/n_i$. For the logistic regression model, $\eta_i = \log[\pi_i/(1-\pi_i)]$, so that $\partial\eta_i/\partial\pi_i = 1/[\pi_i(1-\pi_i)]$. Thus, $w_i = n_i\pi_i(1-\pi_i)$, and for large samples, the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\widehat{\text{var}}(\hat{\boldsymbol{\beta}}) = \{\boldsymbol{X}^{\text{T}}\hat{\boldsymbol{W}}\boldsymbol{X}\}^{-1} = \{\boldsymbol{X}^{\text{T}}\mathbf{Diag}[n_i\hat{\pi}_i(1-\hat{\pi}_i)]\boldsymbol{X}\}^{-1}, \tag{5.6}$$

where $\hat{\boldsymbol{W}} = \mathbf{Diag}[n_i\hat{\pi}_i(1-\hat{\pi}_i)]$ denotes the $N \times N$ diagonal matrix having $\{n_i\hat{\pi}_i(1-\hat{\pi}_i)\}$ on the main diagonal. "Large samples" here means a large number of Bernoulli trials, that is, large N for ungrouped data and large $n = \sum_i n_i$ for grouped data, in each case with p fixed. The square roots of the main diagonal elements of Equation (5.6) are estimated standard errors of $\hat{\boldsymbol{\beta}}$.

5.3.3 Statistical Inference: Wald Method is Suboptimal

For statistical inference for logistic regression models, we can use the Wald, likelihood-ratio, or score methods introduced in Section 4.3. For example, to test H_0: $\beta_j = 0$, the Wald chi-squared ($df = 1$) uses $(\hat{\beta}_j/SE_j)^2$, whereas the likelihood-ratio chi-squared uses the difference between the deviances for the simpler model with $\beta_j = 0$ and the full model.

These methods usually give similar results for large sample sizes. However, the Wald method has two disadvantages. First, its results depend on the scale for parameterization. To illustrate, for the null model, $\text{logit}(\pi) = \beta_0$, consider testing H_0: $\beta_0 = 0$ (i.e., $\pi = 0.50$) when ny has a $\text{bin}(n, \pi)$ distribution. From the delta method, the asymptotic variance of $\hat{\beta}_0 = \text{logit}(y)$ is $[n\pi(1-\pi)]^{-1}$. The Wald chi-squared test statistic, which uses the ML estimate of the asymptotic variance, is $(\hat{\beta}_0/SE)^2 = [\text{logit}(y)]^2[ny(1-y)]$. On the proportion scale, the Wald test statistic is $(y-0.50)^2/[y(1-y)/n]$. These are not the same. Evaluations reveal that the logit-scale statistic is too conservative[1] and the proportion-scale statistic is too liberal. A second disadvantage is that when a true effect in a binary regression model is very large, the Wald test is less powerful than the other methods and can show aberrant behavior. For this single-binomial example, suppose $n = 25$. Then, $y = 24/25$ is stronger evidence against H_0: $\pi = 0.50$ than $y = 23/25$, yet the logit Wald statistic equals 9.7 when $y = 24/25$ and 11.0 when $y = 23/25$. For comparison, the likelihood-ratio statistics are 26.3 and 20.7. As the true effect in a binary regression model increases, for a given sample size the information decreases so quickly that the standard error grows faster than the effect.[2] The Wald method fails completely when $\hat{\beta}_j = \pm\infty$, a case we discuss in Section 5.4.2.

5.3.4 Conditional Logistic Regression to Eliminate Nuisance Parameters

The total number of binary observations is $n = \sum_{i=1}^{N} n_i$ for grouped data and $n = N$ for ungrouped data. ML estimators of the p parameters of the logistic regression

[1] When H_0 is true, the probability a test of nominal size α rejects H_0 is *less* than α.
[2] See Davison (2003, p. 489), Hauck and Donner (1977), and Exercise 5.7.

model and standard methods of inference perform well when n is large compared with p. Sometimes n is small. Sometimes p grows as n grows, as in highly stratified data in which each stratum has its own model parameter. In either case, improved inference results from using *conditional maximum likelihood*. This method reduces the parameter space, eliminating nuisance parameters from the likelihood function by conditioning on their sufficient statistics. Inference based on the conditional likelihood can use large-sample asymptotics or small-sample distributions.

We illustrate with a simple case: logistic regression with a single binary explanatory variable x and small n. For subject i in an ungrouped data file,

$$\text{logit}[P(y_i = 1)] = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, N, \tag{5.7}$$

where $x_i = 1$ or $x_i = 0$. Usually the log odds ratio β_1 is the parameter of interest, and β_0 is a nuisance parameter. For the exponential dispersion family (4.1) with $a(\phi) = 1$, the kernel of the log-likelihood function is $\sum_i y_i\theta_i$. For the logistic model, this is

$$\sum_{i=1}^{N} y_i\theta_i = \sum_{i=1}^{N} y_i(\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^{N} y_i + \beta_1 \sum_{i=1}^{N} x_i y_i.$$

The sufficient statistics are $\sum_i y_i$ for β_0 and $\sum_i x_i y_i$ for β_1. The grouped form of the data is summarized with a 2×2 contingency table. Denote the two independent binomial "success" totals in the table by s_1 and s_2, having bin(n_1, π_1) and bin(n_2, π_2) distributions, as Table 5.2 shows. To conduct conditional inference about β_1 while eliminating β_0, we use the distribution of $\sum_i x_i y_i = s_1$, conditional on $\sum_i y_i = s_1 + s_2$.

Consider testing H_0: $\beta_1 = 0$, which corresponds to H_0: $\pi_1 = \pi_2$. Under H_0, let $\pi = e^{\beta_0}/(1 + e^{\beta_0})$ denote the common value. We eliminate β_0 by finding $P(s_1 = t \mid s_1 + s_2 = v)$. By the independence of the binomial variates and the fact that their sum is also binomial, under H_0

$$P(s_1 = t, s_2 = u) = \binom{n_1}{t}\pi^t(1-\pi)^{n_1-t}\binom{n_2}{u}\pi^u(1-\pi)^{n_2-u}, \; t = 0, \ldots, n_1, \; u = 0, \ldots, n_2$$

$$P(s_1 + s_2 = v) = \binom{n_1 + n_2}{v}\pi^v(1-\pi)^{n_1+n_2-v}, \quad v = 0, 1, \ldots, n_1 + n_2.$$

Table 5.2 A 2×2 Table for Binary Response and Explanatory Variables

	y		
x	1	0	Total
1	s_1	$n_1 - s_1$	n_1
0	s_2	$n_2 - s_2$	n_2

So the conditional probability is

$$P(s_1 = t \mid s_1 + s_2 = v) = \frac{\binom{n_1}{t}\pi^t(1-\pi)^{n_1-t}\binom{n_2}{v-t}\pi^{v-t}(1-\pi)^{n_2-(v-t)}}{\binom{n_1+n_2}{v}\pi^v(1-\pi)^{n_1+n_2-v}}$$

$$= \frac{\binom{n_1}{t}\binom{n_2}{v-t}}{\binom{n_1+n_2}{v}}, \quad \max(0, v-n_2) \le t \le \min(n_1, v).$$

This is the *hypergeometric distribution*. To test H_0: $\beta_1 = 0$ against H_1: $\beta_1 > 0$, the P-value is $P(s_1 \ge t \mid s_1 + s_2)$, for observed value t for s_1. This probability does not depend on β_0. We can find it exactly rather than rely on a large-sample approximation. This test was proposed by R. A. Fisher (1935) and is called *Fisher's exact test* (see Exercise 5.31).

The conditional approach has the limitation of requiring sufficient statistics for the nuisance parameters. Reduced sufficient statistics exist only with GLMs that use the canonical link. Thus, the conditional approach works for the logistic model but not for binary GLMs that use other link functions. Another limitation is that when some explanatory variables are continuous, the $\{y_i\}$ values may be completely determined by the given sufficient statistics, making the conditional distribution degenerate.

5.4 LOGISTIC REGRESSION MODEL FITTING

We can use standard iterative methods to solve the logistic regression likelihood equations (5.5). In certain cases, however, some or all ML estimates may be infinite or may not even exist.

5.4.1 Iterative Fitting of Logistic Regression Models

The Newton–Raphson iterative method (Section 4.5.1) is equivalent to Fisher scoring, because the logit link is the canonical link. Using expressions (4.8) and the inverse of Equation (5.6), in terms of the binomial "success" counts $\{s_i = n_i y_i\}$, let

$$u_j^{(t)} = \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j}\bigg|_{\beta^{(t)}} = \sum_i (s_i - n_i \pi_i^{(t)}) x_{ij}$$

$$h_{ab}^{(t)} = \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_a \, \partial \beta_b}\bigg|_{\beta^{(t)}} = -\sum_i x_{ia} x_{ib} n_i \pi_i^{(t)}(1 - \pi_i^{(t)}).$$

Here $\boldsymbol{\pi}^{(t)}$, approximation t for $\hat{\boldsymbol{\pi}}$, is obtained from $\boldsymbol{\beta}^{(t)}$ through

$$\pi_i^{(t)} = \frac{\exp\left(\sum_{j=1}^p \beta_j^{(t)} x_{ij}\right)}{1 + \exp\left(\sum_{j=1}^p \beta_j^{(t)} x_{ij}\right)}. \tag{5.8}$$

We use $\boldsymbol{u}^{(t)}$ and $\boldsymbol{H}^{(t)}$ with formula (4.23) to obtain the next value, $\boldsymbol{\beta}^{(t+1)}$, which in this context is

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \left\{\boldsymbol{X}^{\mathrm{T}}\mathbf{Diag}\left[n_i\pi_i^{(t)}(1-\pi_i^{(t)})\right]\boldsymbol{X}\right\}^{-1}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{s}-\boldsymbol{\mu}^{(t)}), \tag{5.9}$$

where $\mu_i^{(t)} = n_i\pi_i^{(t)}$. This is used to obtain $\boldsymbol{\pi}^{(t+1)}$, and so forth.

With an initial guess $\boldsymbol{\beta}^{(0)}$, Equation (5.8) yields $\boldsymbol{\pi}^{(0)}$, and for $t > 0$ the iterations proceed as just described using Equations (5.9) and (5.8). In the limit, $\boldsymbol{\pi}^{(t)}$ and $\boldsymbol{\beta}^{(t)}$ converge to the ML estimates $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\beta}}$, except for certain data configurations for which at least one estimate is infinite or does not exist (Section 5.4.2). The $\boldsymbol{H}^{(t)}$ matrices converge to $\widehat{\boldsymbol{H}} = -\boldsymbol{X}^{\mathrm{T}}\mathbf{Diag}[n_i\hat{\pi}_i(1-\hat{\pi}_i)]\boldsymbol{X}$. By Equation (5.6) the estimated asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is a by-product of the model fitting, namely $-\widehat{\boldsymbol{H}}^{-1}$.

From Section 4.5.4, $\boldsymbol{\beta}^{(t+1)}$ has the iteratively reweighted least squares form $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_t^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_t^{-1}\boldsymbol{z}^{(t)}$, where $\boldsymbol{z}^{(t)}$ has elements

$$z_i^{(t)} = \log\frac{\pi_i^{(t)}}{1-\pi_i^{(t)}} + \frac{s_i - n_i\pi_i^{(t)}}{n_i\pi_i^{(t)}\left(1-\pi_i^{(t)}\right)},$$

and where $\boldsymbol{V}_t = (\boldsymbol{W}^{(t)})^{-1}$ is a diagonal matrix with elements $\{1/[n_i\pi_i^{(t)}(1-\pi_i^{(t)})]\}$. In this expression, $\boldsymbol{z}^{(t)}$ is the linearized form of the logit link function for the sample data, evaluated at $\boldsymbol{\pi}^{(t)}$ (see (4.25)). The limit $\widehat{\boldsymbol{V}}$ of $\boldsymbol{V}_t$ has diagonal elements that estimate the variances of the approximate normal distributions[3] of the sample logits for large $\{n_i\}$, by the delta method.

5.4.2 Infinite Parameter Estimates in Logistic Regression

The Hessian matrix for logistic regression models is negative-definite, and the log-likelihood function is concave. ML estimates exist and are finite except when a hyperplane separates the set of explanatory variable values having $y = 0$ from the set having $y = 1$ (Albert and Anderson 1984).

For example, with a single explanatory variable and six observations, suppose $y = 1$ at $x = 1, 2, 3$ and $y = 0$ at $x = 4, 5, 6$ (see Figure 5.3). For the model $\mathrm{logit}(\pi_i) = \beta_0 + \beta_1 x_i$ with observations in increasing order on x, the likelihood equations (5.5) are $\sum_i \hat{\pi}_i = \sum_i y_i$ and $\sum_i x_i\hat{\pi}_i = \sum_i x_i y_i$, or

$$\sum_{i=1}^{6}\hat{\pi}_i = 3 \quad \text{and} \quad \sum_{i=1}^{6} i\hat{\pi}_i = (1+2+3)1 + (4+5+6)0 = 6.$$

A solution is $\hat{\pi}_i = 1$ for $i = 1, 2, 3$ and $\hat{\pi}_i = 0$ for $i = 4, 5, 6$. Any other set of $\{\hat{\pi}_i\}$ having $\sum_i \hat{\pi}_i = 3$ would have $\sum_i i\hat{\pi}_i > 6$, so this is the unique solution. By letting $\hat{\beta}_1 \to -\infty$ and, for fixed $\hat{\beta}_1$, letting $\hat{\beta}_0 = -3.5\hat{\beta}_1$ so that $\hat{\pi} = 0.50$ at $x = 3.5$, we can

[3]The actual variance does not exist, because with positive probability the sample proportion $y_i = 1$ or 0 and the sample logit $= \pm\infty$.

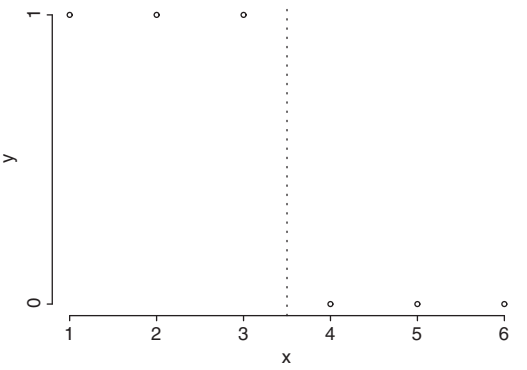

Figure 5.3 Complete separation of explanatory variable values, such as $y = 1$ when $x < 3.5$ and $y = 0$ when $x > 3.5$, causes an infinite ML effect estimate.

generate a sequence with ever-increasing value of the likelihood function that comes successively closer to satisfying these equations and giving a perfect fit.

In practice, software may fail to recognize when an ML estimate is actually infinite. After a certain number of cycles of iterative fitting, the log-likelihood looks flat at the working estimate, because the log-likelihood approaches a limiting value as the parameter value grows unboundedly. So, convergence criteria are satisfied, and software reports estimated. Because the log-likelihood is so flat and because the variance of $\hat{\beta}_j$ comes from its curvature as described by the negative inverse of the matrix of second partial derivatives, software typically reports huge standard errors.

```
------------------------------------------------------------------------
> x <- c(1,2,3,4,5,6);  y <- c(1,1,1,0,0,0)  # complete separation
> fit <- glm(y ~ x, family = binomial(link = logit))
> summary(fit)
Coefficients:
            Estimate  Std. Error  z value Pr(>|z|)
(Intercept)   165.32   407521.43        0        1 # x estimate is
x             -47.23   115264.41        0        1 # actually -infinity

Number of Fisher Scoring iterations: 25   # unusually large
> logLik(fit)
'log Lik.' -1.107576e-10 (df=2) # maximized log-likelihood = 0
------------------------------------------------------------------------
```

The space of explanatory variable values is said to have *complete separation* when a hyperplane can pass through that space such that on one side of that hyperplane $y_i = 0$ for all observations, whereas on the other side $y_i = 1$ always, as in Figure 5.3. There is then *perfect discrimination*, as we can predict the sample outcomes perfectly by knowing the explanatory variable values. In practice, we have an indication of complete separation when the fitted prediction equation perfectly predicts the response outcome for the entire dataset; that is, $\hat{\pi}_i = 1.0$ (to many decimal places) whenever

$y_i = 1$ and $\hat{\pi}_i = 0.0$ whenever $y_i = 0$. A related indication is that the reported maximized log-likelihood value is 0 to many decimal places. Another warning signal is standard errors that seem unnaturally large.

A weaker condition that causes at least one estimate to be infinite, called *quasi-complete separation*, occurs when a hyperplane separates explanatory variable values with $y_i = 1$ and with $y_i = 0$, but cases exist with both outcomes on that hyperplane. For example, this toy example of six observations has quasi-complete separation if we add two observations at $x = 3.5$, one with $y = 1$ and one with $y = 0$. Quasi-complete separation is more likely to happen with qualitative predictors than with quantitative predictors. If any category of a qualitative predictor has either no cases with $y = 0$ or no cases with $y = 1$, quasi-complete separation occurs when that variable is entered as a factor in the model (i.e., using an indicator variable for that category). With quasi-complete separation, there is not perfect discrimination for all observations. The maximized log-likelihood is then strictly less than 0. However, a warning signal is again reported standard errors that seem unnaturally large.

What inference can you conduct when the data have complete or quasi-complete separation? With an infinite estimate, you can still compute likelihood-ratio tests. The log-likelihood has a maximized value at the infinite estimate for a parameter, so you can compare it with the value when the parameter is equated to some fixed value such as zero. Likewise, you can invert the test to construct a confidence interval. If $\hat{\beta} = \infty$, for example, a 95% profile likelihood confidence interval has the form (L, ∞), where L is such that the likelihood-ratio test of H_0: $\beta = L$ has P-value $= 0.05$. With quasi-complete separation, some parameter estimates and *SE* values may be unaffected, and even Wald inference methods are available with them.

Alternatively, you can make some adjustment so that all estimates are finite. Some approaches smooth the data, thus producing finite estimates. The Bayesian approach is one way to do that (Section 10.3). A related way maximizes a *penalized likelihood* function. This adds a term to the ordinary log-likelihood function such that maximizing the amended function smooths the estimates by shrinking them toward 0 (Section 11.1.7).

5.5 DEVIANCE AND GOODNESS OF FIT FOR BINARY GLMS

For grouped or ungrouped binary data, one way to detect lack of fit uses a likelihood-ratio test to compare the model with more complex ones. If more complex models do not fit better, this provides some assurance that the model chosen is reasonable. Other approaches to detecting lack of fit search for *any* way that the model fails, using global statistics such as the deviance or Pearson statistics.

5.5.1 Deviance and Pearson Goodness-of-Fit Statistics

From Section 4.4.3, for binomial GLMs the deviance is the likelihood-ratio statistic comparing the model to the unrestricted (saturated model) alternative. The saturated

model has the perfect fit $\tilde{\pi}_i = y_i$. The likelihood-ratio statistic comparing this to the ML model fit $\hat{\pi}_i$ for all i is

$$-2\log\left\{\left[\prod_{i=1}^{N}\hat{\pi}_i^{n_iy_i}(1-\hat{\pi}_i)^{n_i-n_iy_i}\right]\Big/\left[\prod_{i=1}^{N}\tilde{\pi}_i^{n_iy_i}(1-\tilde{\pi}_i)^{n_i-n_iy_i}\right]\right\}$$
$$= 2\sum_i n_iy_i\log\frac{n_iy_i}{n_i\hat{\pi}_i} + 2\sum_i(n_i-n_iy_i)\log\frac{n_i-n_iy_i}{n_i-n_i\hat{\pi}_i}.$$

At setting i of the explanatory variables, n_iy_i is the number of successes and $(n_i - n_iy_i)$ is the number of failures, $i = 1, \ldots, N$. Thus, the deviance is a sum over the $2N$ success and failure totals at the N settings, having the form

$$D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = 2\sum \text{observed} \times \log(\text{observed}/\text{fitted}).$$

This has the same form as the deviance (4.16) for Poisson loglinear models with intercept term. In either case, we denote it by G^2.

For naturally grouped data (e.g., solely categorical explanatory variables), the data file can be expressed in grouped or in ungrouped form. The deviance differs[4] in the two cases. For grouped data, the saturated model has a parameter at each setting for the explanatory variables. For ungrouped data, by contrast, it has a parameter for each subject.

For grouped data, a Pearson statistic also summarizes goodness of fit. It is the sum over the $2N$ cells of successes and failures,

$$X^2 = \sum\frac{(\text{observed} - \text{fitted})^2}{\text{fitted}}$$
$$= \sum_{i=1}^{N}\frac{(n_iy_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i} + \sum_{i=1}^{N}\frac{[(n_i-n_iy_i)-(n_i-n_i\hat{\pi}_i)]^2}{n_i(1-\hat{\pi}_i)}$$
$$= \sum_{i=1}^{N}\frac{(n_iy_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)} = \sum_{i=1}^{N}\frac{(y_i-\hat{\pi}_i)^2}{\hat{\pi}_i(1-\hat{\pi}_i)/n_i}. \qquad (5.10)$$

In the form of Equation (5.10), this statistic is a special case of the score statistic for GLMs introduced in (4.17), having variance function in the denominator.

5.5.2 Chi-Squared Tests of Fit and Model Comparisons

When the data are grouped, the deviance G^2 and Pearson X^2 are goodness-of-fit test statistics for testing H_0 that the model truly holds. Under H_0, they have limiting chi-squared distributions as the overall sample size n increases, by $\{n_i\}$ increasing

[4]Exercise 5.17 shows a numerical example.

(i.e., small-dispersion asymptotics). Grouped data have a fixed number of settings N of the explanatory variables and hence a fixed number of parameters for the saturated model, so the df for the chi-squared distribution is the difference between the numbers of parameters in the two models, $df = N - p$. The X^2 statistic results[5] from summing the terms up to second-order in a Taylor series expansion of G^2, and $(X^2 - G^2)$ converges in probability to 0 under H_0. As n increases, the X^2 statistic converges to chi-squared more quickly than G^2 and has a more trustworthy P-value when some expected success or failure totals are less than about five.

The chi-squared limiting distribution does not occur for ungrouped data. In fact, G^2 and X^2 can be uninformative about lack of fit (Exercises 5.14 and 5.16). The chi-squared approximation is also poor with grouped data having a large N with relatively few observations at each setting, such as when there are many explanatory variables or one of them is nearly continuous in measurement (e.g., a person's age). For ungrouped data, G^2 and X^2 can be applied in an approximate manner to grouped observed and fitted values for a partition of the space of $\boldsymbol{x}$ values (Tsiatis 1980) or for a partition of the estimated probabilities of success (Hosmer and Lemeshow 1980). However, a large value of any global fit statistic merely indicates *some* lack of fit but provides no insight about its nature. The approach of comparing the working model with a more complex one is more useful from a scientific perspective, since it investigates lack of fit of a particular type.

Although the deviance is not useful for testing model fit when the data are ungrouped or nearly so, it remains useful for comparing models. For either grouped or ungrouped data, we can compare two nested models using the difference of deviances (Section 4.4.3). Suppose model M_0 has p_0 parameters and the more complex model M_1 has $p_1 > p_0$ parameters. Then the difference of deviances is the likelihood-ratio test statistic for comparing the models. If model M_0 holds, this difference has an approximate chi-squared distribution with $df = p_1 - p_0$. One can also compare the models using the Pearson comparison statistic (4.18).

5.5.3 Residuals: Pearson, Deviance, and Standardized

After a preliminary choice of model, such as with a global goodness-of-fit test or by comparing pairs of models, we obtain further insight by switching to a microscopic mode of analysis. With grouped data, it is useful to form residuals to compare observed and fitted proportions.

For observation i with sample proportion y_i and model fitted proportion $\hat{\pi}_i$, the Pearson residual (4.20) is

$$e_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\widehat{\text{var}}(y_i)}} = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)/n_i}}.$$

Equivalently, this divides the raw residual $(n_i y_i - n_i \hat{\pi}_i)$ comparing the observed and fitted number of successes by the estimated binomial standard deviation of $n_i y_i$. From

[5]For details, see Agresti (2013, p. 597).

Equation (5.10) these residuals satisfy

$$X^2 = \sum_{i=1}^{N} e_i^2,$$

for the Pearson statistic for testing the model fit. An alternative *deviance residual*, introduced for GLMs in (4.21), uses components of the deviance.

As explained in Section 4.4.6, the Pearson residuals have standard deviations less than 1. The standardized residual divides $(y_i - \hat{\pi}_i)$ by its estimated standard error. This uses the leverage $\hat{h}_{ii}$ from the diagonal of the GLM estimated hat matrix

$$\widehat{\boldsymbol{H}}_{\boldsymbol{W}} = \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\widehat{\boldsymbol{W}}^{1/2},$$

in which the weight matrix $\widehat{\boldsymbol{W}}$ is diagonal with element $\hat{w}_{ii} = n_i\hat{\pi}_i(1 - \hat{\pi}_i)$. For observation i, the standardized residual is

$$r_i = \frac{e_i}{\sqrt{1 - \hat{h}_{ii}}} = \frac{y_i - \hat{\pi}_i}{\sqrt{[\hat{\pi}_i(1 - \hat{\pi}_i)(1 - \hat{h}_{ii})]/n_i}}.$$

Compared with the Pearson and deviance residuals, it has the advantages of having an approximate $N(0, 1)$ distribution when the model holds (with large n_i) and appropriately recognizing redundancies in the data (Exercise 5.12). Absolute values larger than about 2 or 3 provide evidence of lack of fit.

Plots of residuals against explanatory variables or linear predictor values help to highlight certain types of lack of fit. When fitted success or failure totals are very small; however, just as X^2 and G^2 lose relevance, so do residuals. As an extreme case, for ungrouped data, $n_i = 1$ at each setting. Then y_i can equal only 0 or 1, and a residual can take only two values. One must then be cautious about regarding either outcome as extreme, and a single residual is essentially uninformative. When $\hat{\pi}_i$ is near 1, for example, residuals are necessarily either small and positive or large and negative. Plots of residuals also then have limited use. For example, suppose an explanatory variable x has a strong positive effect. Then, necessarily for small values of x, an observation with $y_i = 1$ will have a relatively large positive residual, whereas for large x an observation with $y_i = 0$ will have a relatively large negative residual. When raw residuals are plotted against fitted values, the plot consists merely of two nearly parallel lines of points. (Why?) When explanatory variables are categorical, so data can have grouped or ungrouped form, it is better to compute residuals and the deviance for the grouped data.

5.5.4 Influence Diagnostics for Logistic Regression

Other regression diagnostic tools also help in assessing fit. These include analyses that describe an observation's influence on parameter estimates and fit statistics.

However, a single observation can have a much greater influence in ordinary least squares regression than in logistic regression, because ordinary regression has no bound on the distance of y_i from its expected value. Also, the estimated hat matrix $\widehat{\boldsymbol{H}}_{\boldsymbol{W}}$ for a binary GLM depends on the fit as well as the model matrix $\boldsymbol{X}$. Points that have extreme predictor values need not have high leverage. In fact, the leverage can be relatively small if $\hat{\pi}_i$ is close to 0 or 1.

Several measures describe the effect of removing an observation from the dataset (Pregibon 1981; Williams 1987). These include the change in X^2 or G^2 goodness-of-fit statistics and analogs of influence measures for ordinary linear models, such as Cook's distance ($r_i^2[\hat{h}_{ii}/p(1-\hat{h}_{ii})]$) using the leverage and standardized residual.

5.6 PROBIT AND COMPLEMENTARY LOG–LOG MODELS

In this section we present two alternatives to the logistic regression model for binary responses. Instead of using the logistic distribution for the cdf inverted to get the link function, one uses the normal distribution and the other uses a skewed distribution.

5.6.1 Probit Models: Interpreting Effects

The binary-response model that takes the link function to be the inverse of the standard normal cdf Φ is called the *probit model*. For the binomial parameter π_i for observation i, the model is

$$\Phi^{-1}(\pi_i) = \sum_{j=1}^{p} \beta_j x_{ij}, \quad \text{or} \quad \pi_i = \Phi\left(\sum_{j=1}^{p} \beta_j x_{ij}\right).$$

For the probit model, the instantaneous rate of change in π_i as predictor j changes, adjusting for the other predictors, is $\partial\pi_i/\partial x_{ij} = \beta_j \phi(\sum_j \beta_j x_{ij})$, where $\phi(\cdot)$ is the standard normal density function. The rate is highest when $\sum_j \beta_j x_{ij} = 0$, at which $\pi_i = \frac{1}{2}$ and the rate equals $0.40\beta_j$. As a function of predictor j, the probit response curve for π_i (or for $1-\pi_i$, when $\beta_j < 0$) has the appearance of a normal cdf with standard deviation $1/|\beta_j|$. By comparison, in logistic regression the rate of change at $\pi_i = \frac{1}{2}$ is $0.25\beta_j$, and the logistic curve for π_i as a function of predictor j has standard deviation $\pi/|\beta_j|\sqrt{3}$ (for $\pi = 3.14\ldots$). The rates of change at $\pi_i = \frac{1}{2}$ are the same for the cdf's corresponding to the probit and logistic curves when the logistic β_j is $0.40/0.25 = 1.60$ times the probit β_j. The standard deviations for the response curves are the same when the logistic β_j is $\pi/\sqrt{3} = 1.81$ times the probit β_j. When both models fit well, ML parameter estimates in logistic regression are about 1.6–1.8 times those in probit models. Although probit model parameters are on a different scale than logistic model parameters, the probability summaries of effects are similar.

Another way to interpret parameters in probit models uses effects in the latent variable threshold model of Section 5.1.2. Since $y_i^* = \sum_j \beta_j x_{ij} + \epsilon_i$ where $\epsilon_i \sim N(0,1)$

has cdf Φ, a 1-unit increase in x_{ij} corresponds to a change of β_j in $E(y_i^*)$, adjusted for the other explanatory variables. We interpret the magnitude of β_j in terms of the conditional standard deviation of 1 for y_i^*, so β_j represents a fraction or multiple of a standard deviation increase. Summary measures of model predictive power include the area under the ROC curve and corr$(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$, as described in Sections 5.2.4 and 5.2.5.

5.6.2 Probit Model Fitting

The likelihood equations for a probit model substitute Φ and ϕ in the general equations (5.4) for GLMs for binary data. The estimated large-sample covariance matrix of $\hat{\boldsymbol{\beta}}$ has the GLM form (4.14),

$$\widehat{\text{var}}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^{\text{T}}\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1},$$

where $\widehat{\boldsymbol{W}}$ is the diagonal matrix with estimates of $w_i = (\partial\mu_i/\partial\eta_i)^2/\text{var}(y_i)$. Since $\mu_i = \pi_i = \Phi(\eta_i) = \Phi(\sum_j \beta_j x_{ij})$,

$$\hat{w}_i = n_i \left[\phi\left(\sum_{j=1}^{p} \hat{\beta}_j x_{ij}\right)\right]^2 \Bigg/ \left\{\Phi\left(\sum_{j=1}^{p} \hat{\beta}_j x_{ij}\right)\left[1 - \Phi\left(\sum_{j=1}^{p} \hat{\beta}_j x_{ij}\right)\right]\right\}.$$

We can solve the likelihood equations using the Fisher scoring algorithm for GLMs or the Newton–Raphson algorithm. They both yield the ML estimates but the Newton–Raphson algorithm gives slightly different standard errors because it inverts the observed information matrix to estimate the covariance matrix, whereas Fisher scoring uses expected information. These differ for link functions other than the canonical link.

5.6.3 Log–Log and Complementary Log–Log Link Models

The logit and probit links are symmetric about 0.50, in the sense that

$$\text{link}(\pi_i) = -\text{link}(1 - \pi_i).$$

To illustrate,

$$\text{logit}(\pi_i) = \log[\pi_i/(1 - \pi_i)] = -\log[(1 - \pi_i)/\pi_i] = -\text{logit}(1 - \pi_i).$$

This means that the response curve for π_i has a symmetric appearance about the point where $\pi_i = 0.50$. Logistic models and probit models are inappropriate when this is badly violated.

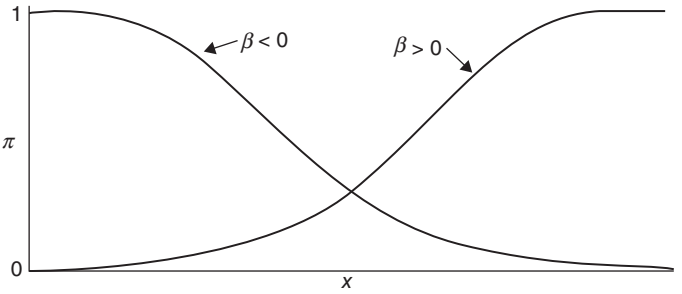

Figure 5.4 GLM for binary data using complementary log–log link function.

A different shape of response curve is given by the model

$$\pi_i = 1 - \exp\left[-\exp\left(\sum_{j=1}^{p} \beta_j x_{ij}\right)\right]. \tag{5.11}$$

With a single explanatory variable, this has the shape shown in Figure 5.4. The curve is asymmetric, π_i approaching 0 slowly but approaching 1 rather sharply. For this model,

$$\log[-\log(1-\pi_i)] = \sum_{j=1}^{p} \beta_j x_{ij}.$$

The link function for this GLM is called the *complementary log–log* link, since the log–log link applies to the complement of π_i.

A related model to Equation (5.11) is

$$\pi_i = \exp\left[-\exp\left(-\sum_{j=1}^{p} \beta_j x_{ij}\right)\right].$$

In GLM form it uses the *log–log* link function.

$$-\log[-\log(\pi_i)] = \sum_{j=1}^{p} \beta_j x_{ij}.$$

For it, π_i approaches 0 sharply but approaches 1 slowly. When the log–log model holds for the probability of a success, the complementary log–log model holds for the probability of a failure, but with a reversal in sign of $\{\hat{\beta}_j\}$.

The log–log link is a special case of an inverse cdf link using the cdf of the *Type I extreme-value* distribution (also called the *Gumbel* distribution). The cdf equals

$$F(x) = \exp\{-\exp[-(x-a)/b]\}$$

for parameters $b > 0$ and $-\infty < a < \infty$. The distribution has mode a, mean $a + 0.577b$, and standard deviation $1.283b$, and is highly skewed to the right. The term *extreme value* refers to its being the limit distribution of the maximum of a sequence of independent and identically distributed continuous random variables.

Models with log–log link can be fitted by using the Fisher scoring algorithm for GLMs. How do we interpret effects in such models? Consider the complementary log–log link model (5.11) with a single explanatory variable x. As x increases, the curve is monotone increasing when $\beta > 0$. The complement probability at $x + 1$ equals the complement probability at x raised to the $\exp(\beta)$ power. We illustrate in the following example.

How can we evaluate the suitability of various possible link functions for a dataset? Measures such as the deviance and AIC provide some information. It is challenging to provide graphical portrayals of relations, especially for ungrouped data, since only $y = 1$ and $y = 0$ values appear on the graph. Plotting response proportions for grouped data can be helpful, as illustrated in the following example. Smoothing methods presented in Section 11.3 are also helpful for portraying the effects.

5.7 EXAMPLES: BINARY DATA MODELING

In this section we analyze two datasets. The first illustrates a logistic regression analysis for which one parameter has an infinite ML estimate. The second is a classic dose–response example from Bliss (1935), the first article to use ML fitting of a probit model.

5.7.1 Example: Risk Factors for Endometrial Cancer Grade

Heinze and Schemper (2002) described a study about endometrial cancer that analyzed how y = histology of 79 cases (0 = low grade for 30 patients, 1 = high grade for 49 patients) relates to three risk factors: x_1 = neovasculation (1 = present for 13 patients, 0 = absent for 66 patients), x_2 = pulsatility index of arteria uterina (ranging from 0 to 49), and x_3 = endometrium height (ranging from 0.27 to 3.61). Table 5.3 shows some of the data.

For these data, consider the main effects model

$$\text{logit}[P(y_i = 1)] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}.$$

Table 5.3 Part of Endometrial Cancer Dataset[a]

HG	NV	PI	EH	HG	NV	PI	EH	HG	NV	PI	EH
0	0	13	1.64	0	0	16	2.26	0	0	8	3.14
...											
1	1	21	0.98	1	0	5	0.35	1	1	19	1.02

Source: Data courtesy of Ella Asseryanis, Georg Heinze, and Michael Schemper. Complete data ($n = 79$) are in the file `Endometrial.dat` at `www.stat.ufl.edu/~aa/glm/data`.

[a]HG = histology grade, NV = neovasculation, PI = pulsatility index, EH = endometrium height.

When $x_{i1} = 0$ both response outcomes occur, but for all 13 patients having $x_{i1} = 1$ the outcome is $y_i = 1$, so there is quasi-complete separation. The ML estimate $\hat{\beta}_1 = \infty$.

```
------------------------------------------------------------------------
> Endometrial
   NV PI   EH HG
1   0 13 1.64  0
2   0 16 2.26  0
...
79  1 19 1.02  1
> attach(Endometrial)
> table(NV,HG) # quasi-complete separation: When NV=1, no HG=0 cases
          HG
   NV    0   1
     0  49  17
     1   0  13
> fit <- glm(HG ~ NV + PI + EH, family=binomial) # logit default link
> summary(fit)
Coefficients:
           Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  4.305        1.637    2.629   0.0086
NV          18.186     1715.751    0.011   0.9915 # 18.186 and 1715.751
PI          -0.042        0.044   -0.952   0.3413 # should be infinity
EH          -2.903        0.846   -3.433   0.0006
---
    Null deviance: 104.903  on 78  degrees of freedom
Residual deviance:  55.393  on 75  degrees of freedom
> logLik(fit) # not exactly 0 because separation is quasi, not complete
'log Lik.' -27.69663 (df=4)
------------------------------------------------------------------------
```

Despite $\hat{\beta}_1 = \infty$, inference is possible[6] about β_1. The likelihood-ratio statistic for H_0: $\beta_1 = 0$ equals 9.36 with $df = 1$ and has P-value = 0.002. The 95% profile likelihood confidence interval for β_1 is (1.28, ∞).

```
------------------------------------------------------------------------
> deviance(glm(HG ~ PI + EH, family=binomial)) - deviance(fit)
[1] 9.357643 # likelihood-ratio (LR) stat. with df=1 for H0: beta1 = 0

> library(ProfileLikelihood)
> xx <- profilelike.glm(HG~1+PI+EH,data=Endometrial,family=binomial,
+ profile.theta="NV",method="ML",lo.theta=-5,hi.theta=10,length=500,
+ round=3)
> profilelike.plot(theta=xx$theta, profile.lik.norm=xx$profile.lik.norm,
+ round=2)
> profilelike.summary(k=6.82,theta=xx$theta,
+ profile.lik.norm=xx$profile.lik.norm)
$LI.norm # LR = 6.82 gives 2log(6.82)=3.84 = 95 chi-sq percentile
[1]  1.283 10.000 # 10 was initial upper bound, correct upper limit
# is infinity but numerical instability occurs for beta1 values above 10
------------------------------------------------------------------------
```

[6]We present other inferences for β_1 using Bayesian methods in Section 10.3.2 and using penalized likelihood methods in Section 11.1.8.

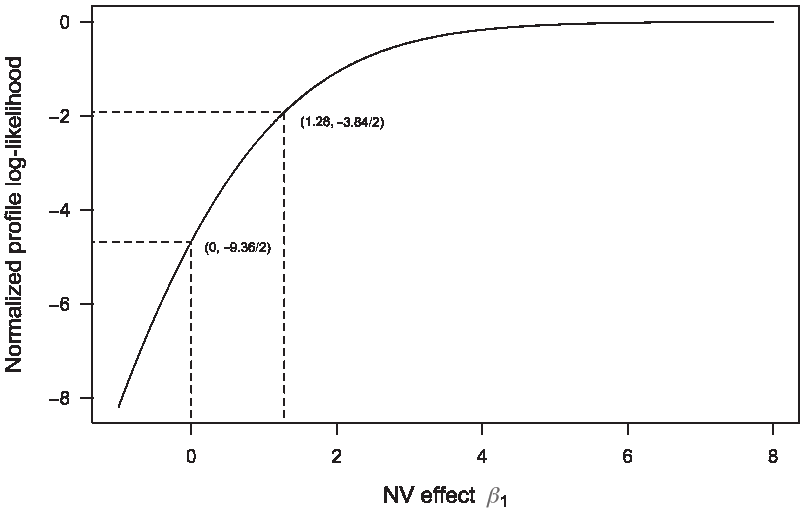

Figure 5.5 Normalized profile log-likelihood function $L(\beta_1) - L(\hat{\beta}_1)$ for NV effect in main-effects logistic model. Double the log-likelihood increases by 9.36 between $\beta_1 = 0$ and $\hat{\beta}_1 = \infty$ and by 3.84 between $\beta_1 = 1.28$ and $\hat{\beta}_1 = \infty$ (the 95% profile likelihood confidence interval). Figure constructed by Alessandra Brazzale with `cond` R package for higher-order likelihood-based conditional inference for logistic models.

We can conclude that $\beta_1 > 0$ (despite what the Wald P-value shows on the R output!) and that the effect is substantial. Figure 5.5 shows the normalized profile log-likelihood function for β_1.

The other ML estimates are not affected by the quasi-complete separation. Most of the predictive power is provided by the EH predictor: $\text{corr}(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 0.745$ for the full model and 0.692 for the model with EH as the sole predictor; the areas under the ROC curves are 0.907 and 0.895. More complex models (not shown here) do not provide an improved fit.

5.7.2 Example: Dose–Response Study

From a dose–response study, Table 5.4 reports, in grouped-data form, the number of adult flour beetles that died after 5 hours of exposure to gaseous carbon disulfide at various dosages. Figure 5.6 plots the proportion killed against $x = \log_{10}(\text{dose})$. The proportion jumps up at about $x = 1.81$, and it is close to 1 above there.

To let the response curve take the shape of a normal cdf, Bliss (1935) used the probit model. The ML fit is

$$\Phi^{-1}(\hat{\pi}_i) = -34.96 + 19.74x_i, \quad i = 1, \ldots, 8.$$

Now $\hat{\pi} = 0.50$ when $\hat{\beta}_0 + \hat{\beta}_1 x = 0$, which for this fit is at $x = 34.96/19.74 = 1.77$. The fit corresponds to a normal cdf with $\mu = 1.77$ and $\sigma = 1/19.74 = 0.05$. Figure 5.6

Table 5.4 Beetles Killed after Exposure to Carbon Disulfide

			Fitted Number Dead		
Log Dosage	Number of Beetles	Number Dead	Comp. Log–Log	Probit	Logit
1.691	59	6	5.6	3.4	3.5
1.724	60	13	11.3	10.7	9.8
1.755	62	18	21.0	23.5	22.5
1.784	56	28	30.4	33.8	33.9
1.811	63	52	47.8	49.6	50.1
1.837	59	53	54.1	53.3	53.3
1.861	62	61	61.1	59.7	59.2
1.884	60	60	59.9	59.2	58.7

Source: Data file `Beetles2.dat` at text website, reprinted from Bliss (1935) with permission of John Wiley & Sons, Inc.

shows the fit. As x increases from 1.691 to 1.884, $\hat{\pi}$ increases from 0.058 to 0.987. For a 0.10-unit increase in x, such as from 1.70 to 1.80, the estimated conditional distribution of the latent variable y^* shifts up by $0.10(19.74) \approx 2$ standard deviations. The following R code enters the data in grouped-data form:

```
-------------------------------------------------------------------------
> logdose <- c(1.691, 1.724, 1.755, 1.784, 1.811, 1.837, 1.861, 1.884)
> dead <- c(6, 13, 18, 28, 52, 53, 61, 60)  # numbers dead
> n <- c(59, 60, 62, 56, 63, 59, 62, 60)    # binomial sample sizes
> alive <- n - dead                         # numbers not dead
> data <- matrix(append(dead,alive),ncol=2) # matrix of binomial counts
> fit.probit <- glm(data ~ logdose, family=binomial(link=probit))
> summary(fit.probit)
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -34.956       2.649   -13.20    <2e-16
```

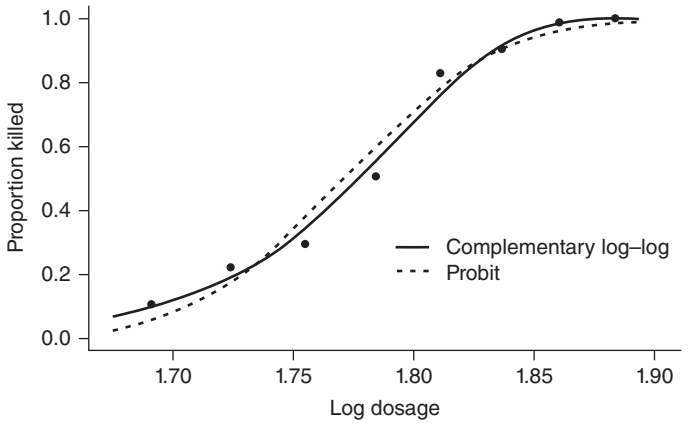

Figure 5.6 Proportion of dead beetles versus log dosage of gaseous carbon disulfide, with fits of probit and complementary log–log models.

```
logdose       19.741       1.488    13.27    <2e-16
---
    Null deviance: 284.202 on 7 degrees of freedom
Residual deviance:   9.987 on 6 degrees of freedom # (df = N-p = 8-2)
AIC: 40.185
> sum(resid(fit.probit, type="pearson")^2) # Pearson chi-squared, df=6
[1] 9.368992
> 1 - pchisq(9.368992, 6) # P-value for Pearson goodness-of-fit test
[1] 0.1538649
-------------------------------------------------------------------------
```

The deviance $G^2 = 9.99$ and Pearson $X^2 = 9.37$ (the sum of the squared Pearson residuals) have $df = 8 - 2 = 6$ and show slight evidence of lack of fit (P-value $= 0.15$ for X^2). The ML estimates are the same for grouped and ungrouped data, but the goodness-of-fit statistics apply only for the grouped data. The following R code shows the fit for the ungrouped data (file `Beetles.dat` at the text website).

```
-------------------------------------------------------------------------
Beetles <- read.table("Beetles.dat",header=TRUE)
> Beetles # ungrouped data at www.stat.ufl.edu/~aa/glm/data/Beetles.dat
          x  y # y=1 for dead, y=0 for alive
1     1.691  1
2     1.691  1
...
481   1.884  1
> attach(Beetles)
> fit.probit2 <- glm(y ~ x, family=binomial(link=probit))
> summary(fit.probit2)
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -34.956       2.649   -13.20    <2e-16
x             19.741       1.488    13.27    <2e-16
---
    Null deviance: 645.44 on 480 degrees of freedom  # very different
Residual deviance: 371.23 on 479 degrees of freedom  # from grouped data
-------------------------------------------------------------------------
```

To summarize predictive power, for the ungrouped data $\text{corr}(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 0.696$. The following code shows this and shows how to use an R package to construct the ROC curve for the model fit. For that curve, shown in Figure 5.7, the estimated concordance index $c = 0.901$.

```
-------------------------------------------------------------------
> cor(y, fitted(fit.probit2))
[1] 0.696391
> library(ROCR) # to construct ROC curve
> pred <- prediction(fitted(fit.probit2), y)
> perf <- performance(pred, "tpr", "fpr")
> plot(perf)
> performance(pred, "auc")
[1] 0.9010852     # concordance index = area under ROC curve (auc)
-------------------------------------------------------------------
```

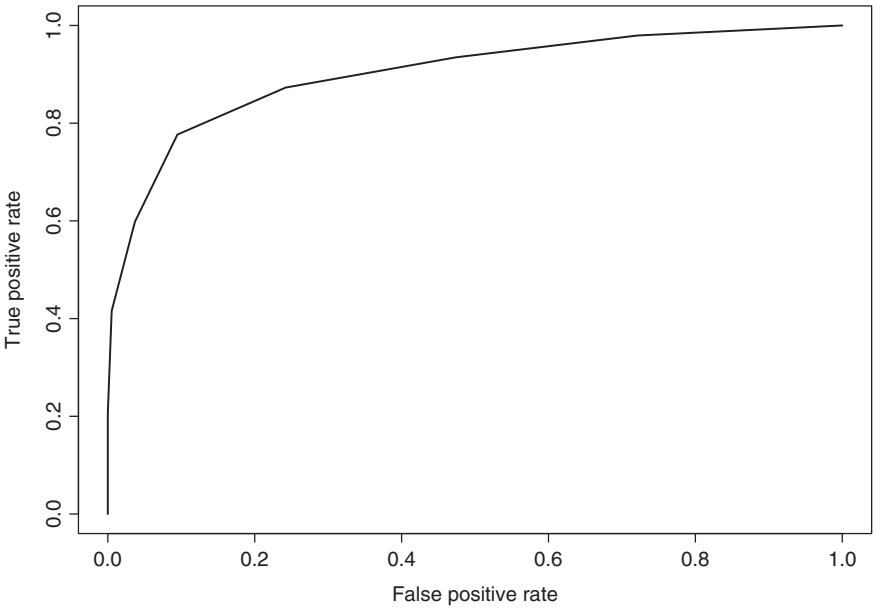

Figure 5.7 ROC curve for probit model fitted to beetle mortality data.

For comparison, we fit the corresponding logistic model. The ratio of $\hat{\beta}_1$ estimates for logit/probit is $34.29/19.74 = 1.74$. At dosage x_i with n_i beetles, $n_i\hat{\pi}_i$ is the fitted death count. Table 5.4 reports the fitted values for the grouped data. The logistic and probit models fit similarly.

```
------------------------------------------------------------------------
> fit.logit <- glm(data ~ logdose, family = binomial(link=logit))
> summary(fit.logit) # grouped data
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -60.740       5.182   -11.72    <2e-16
logdose       34.286       2.913    11.77    <2e-16
---
    Null deviance: 284.202  on 7  degrees of freedom
Residual deviance:  11.116  on 6  degrees of freedom
AIC: 41.314
------------------------------------------------------------------------
```

The model with complementary log–log link has $\log[-\log(1-\hat{\pi}_i)] = -39.52 + 22.015x_i$. At dosage $x = 1.70$, the fitted probability of survival is $\exp\{-\exp[-39.52 + 22.015(1.70)]\} = 0.885$, whereas at $x = 1.80$ it is 0.330 and at $x = 1.90$ it is 4×10^{-5}. The probability of survival at dosage $x + 0.10$ equals the probability of survival at dosage x raised to the $e^{0.10(22.015)} = 9.04$ power. For instance, $0.330 = (0.885)^{9.04}$. Table 5.4 shows the fitted values for the grouped data, and Figure 5.6 shows the fit, which seems adequate (deviance $G^2 = 3.51, df = 6$). The code also shows the use

of the `confint` function for obtaining profile-likelihood confidence intervals for the model parameters.

```
------------------------------------------------------------------------
> fit.cloglog <- glm(data ~ logdose, family = binomial(link=cloglog))
> summary(fit.cloglog) # grouped data
            Estimate  Std. Error  z value   Pr(>|z|)
(Intercept)  -39.522       3.236    -12.21   <2e-16
logdose       22.015       1.797     12.25   <2e-16
---
    Null deviance: 284.2024  on 7  degrees of freedom
Residual deviance:   3.5143  on 6  degrees of freedom
AIC: 33.712
> sum(resid(fit.cloglog, type="pearson")^2) # Pearson chi-squared stat.
[1] 3.35924
> confint(fit.cloglog) # profile likelihood confidence intervals
               2.5 %    97.5 %
(Intercept)  -46.140   -33.499
logdose       18.669    25.689
------------------------------------------------------------------------
```

By contrast, the log–log link yields a very poor fit. To use $-\log[-\log(\pi_i)]$ instead of $\log[-\log(\pi_i)]$ as the link function, corresponding to the inverse of the extreme-value cdf, we take the negative of the estimates reported here in the output for the model object called *fit.loglog*.

```
------------------------------------------------------------------------
> data2 <- matrix(append(alive,dead),ncol=2) # reverse for log-log link
> fit.loglog <- glm(data2 ~ logdose, family=binomial(link=cloglog))
> summary(fit.loglog) # much poorer fit than complementary log-log link
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)   37.661       2.949    12.77    <2e-16
logdose      -21.583       1.680   -12.85    <2e-16
---
    Null deviance: 284.202  on 7  degrees of freedom
Residual deviance:  27.573  on 6  degrees of freedom # grouped data
AIC: 57.771
------------------------------------------------------------------------
```

The models with different link functions are not nested, so we cannot compare them with likelihood-ratio tests. The AIC values for the grouped data are 41.3 for the logit link, 40.2 for the probit model, 33.7 for the complementary log–log link, and 57.8 for the log–log link, showing a clear preference for the complementary log–log link. By contrast, the ROC curve is identical for the four link functions. The corr($\boldsymbol{y}$, $\hat{\boldsymbol{\mu}}$) values for the ungrouped data are 0.684 for the log–log link, 0.696 for the probit link, 0.697 for the logit link, and 0.701 for the complementary log–log link.

Next, we perform a residual analysis for the complementary log–log link model applied to the grouped data. According to the standardized residuals, no observation exhibits lack of fit. Finally, Figure 5.8 plots the sample proportions dead, the fitted

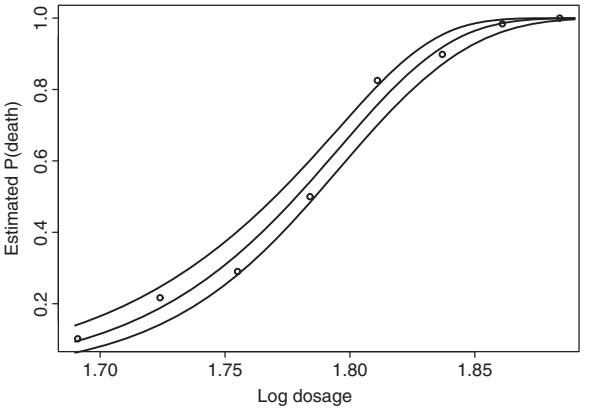

Figure 5.8 Plot of sample proportions, fitted complementary log–log model, and model-based confidence intervals for probability of death as function of log dosage.

values for the model, and 95% pointwise confidence bands for the true probabilities (assuming the model).

```
-------------------------------------------------------------------------
> pearson.res <- resid(fit.cloglog, type="pearson") # Pearson residuals
> std.res <- rstandard(fit.cloglog, type="pearson") # standardized res.
> cbind(logdose, dead/n, fitted(fit.cloglog), pearson.res, std.res)
  logdose                 pearson.res    std.res  # grouped data
1   1.691  0.102  0.096         0.153      0.177
2   1.724  0.217  0.188         0.568      0.669
3   1.755  0.290  0.338        -0.790     -0.922
4   1.784  0.500  0.542        -0.627     -0.704
5   1.811  0.825  0.757         1.268      1.486
6   1.837  0.898  0.918        -0.565     -0.702
7   1.861  0.984  0.986        -0.125     -0.149
8   1.884  1.000  0.999         0.228      0.237
> plot(logdose, dead/n)
> lines(logdose, fitted(fit.cloglog))
> fv <- predict(fit.cloglog, se.fit = TRUE)
>  U <- fv$fit + 1.96*fv$se.fit;  L <- fv$fit - 1.96*fv$se.fit
> lines(logdose, 1 - exp(-exp(U)));  lines(logdose, 1 - exp(-exp(L)))
-------------------------------------------------------------------------
```

CHAPTER NOTES

Section 5.1: Link Functions for Binary Data

5.1 **Other link functions**: Other link functions for binary data include the inverse cdf of a t distribution (the probit being the limit as $df \to \infty$); a log-gamma link (Genter and Farewell 1985), for which probit, complementary log–log and log–log are special cases; a family of link functions that includes the logit (Pregibon 1980); and extensions with shape parameters that modify the logistic curve in extreme probability regions (Aranda-Ordaz 1981; Stukel 1988).

Section 5.3: Inference about Parameters of Logistic Regression Models

5.2 **Conditional logistic**: For more details about case-control studies and conditional logistic regression, see Breslow and Day (1980, Chapter 7). For more on "exact" inference using conditional distributions with logistic models, see Mehta and Patel (1995). Fisher's exact test extends to $r \times c$ tables and to stratified tables (Agresti 1992).

5.3 **Propensity scores**: Rosenbaum and Rubin (1983) proposed methods of comparing $E(y)$ for two groups in observational studies while adjusting for possibly confounding variables $\boldsymbol{x}$. They defined the *propensity* as the probability of being in one group, as a function of $\boldsymbol{x}$. They used logistic regression to estimate how propensity depends on $\boldsymbol{x}$. Their method takes into account differing distributions of the groups on $\boldsymbol{x}$ by using the estimated propensity to match samples from the groups or to subclassify subjects into intervals of propensity scores or to adjust directly by entering the propensity in the model.

Section 5.6: Probit and Complementary Log–Log Models

5.4 **Binary GLM history**: The probit model was presented by Bliss (1935) and popularized in three editions of Finney (1971). Logistic regression was proposed by Berkson (1944) as a model that has similar fit as a probit model but has closed form for the link function. Yates (1955) proposed the complementary log–log link. The logistic model became more popular following publication of an influential article (1958) and text (1970) by D. R. Cox, because of its direct interpretation in terms of odds ratios, validity in case-control studies, and availability of the conditional approach to eliminate nuisance parameters.

EXERCISES

5.1 For the population having value y on a binary response, suppose x has an $N(\mu_y, \sigma^2)$ distribution, $y = 0, 1$.

a. Using Bayes' theorem, show that $P(y = 1 \mid x)$ satisfies the logistic regression model with $\beta_1 = (\mu_1 - \mu_0)/\sigma^2$.

b. Suppose that $(x \mid y) \sim N(\mu_y, \sigma_y^2)$ with $\sigma_0 \neq \sigma_1$. Show that the logistic model holds with a quadratic term (Anderson 1975).

c. Suppose that $(x \mid y)$ has natural exponential family density

$$f(x; \theta_y) = h(x) \exp[x\theta_y - b(\theta_y)].$$

Show that $P(y = 1 \mid x)$ satisfies the logistic model with $\beta_1 = (\theta_1 - \theta_0)$.

5.2 Refer to Note 1.5. For a logistic model, show that the average estimated rate of change in the response probability as a function of explanatory variable j, adjusting for the others, satisfies $\frac{1}{n}\sum_i(\partial\hat{\pi}_i/\partial x_{ij}) = \hat{\beta}_j \frac{1}{n}\sum_i[\hat{\pi}_i(1 - \hat{\pi}_i)]$.

5.3 Construct the ROC curves for (**a**) the toy example in Section 5.4.2 with complete separation and (**b**) the dataset ($n = 8$) that adds two observations at $x = 3.5$, one with $y = 1$ and one with $y = 0$. In each case, report the area

under the curve and summarize predictive power. For contrast, construct a toy dataset with $n = 8$ for which the area under the ROC curve equals 0.50.

5.4 From the likelihood equation (5.5) for a logistic regression intercept parameter, show that the overall sample proportion of successes equals the sample mean of the fitted success probabilities. Is this true for other binary GLMs?

5.5 Suppose that $n_i y_i$ has a bin(n_i, π_i) distribution. Consider a binary GLM $\pi_i = F(\sum_j \beta_j x_{ij})$ with F the standard cdf of some family of continuous distributions. Find w_i in $w_i = (\partial\mu_i/\partial\eta_i)^2/\text{var}(y_i)$ and hence $\text{var}(\hat{\boldsymbol{\beta}})$.

5.6 Explain how expression (5.6) for $\widehat{\text{var}}(\hat{\boldsymbol{\beta}})$ in logistic regression suggests that the standard errors of $\{\hat{\beta}_j\}$ tend to be smaller as you obtain more data. Answer this for (**a**) grouped data with $\{n_i\}$ increasing, (**b**) ungrouped data with N increasing.

5.7 Assuming the model $\text{logit}[P(y_i = 1)] = \beta x_i$, you take all n observations at x_0. Find $\hat{\beta}$ and the large-sample $\text{var}(\hat{\beta})$. For the Wald test, explain why the chi-squared noncentrality is $\beta^2/\text{var}(\hat{\beta})$, and evaluate it as $\beta \to \infty$. Explain how this illustrates that the Wald test in logistic regression has poor behavior when the effect is strong.

5.8 For a $2 \times 2 \times \ell$ contingency table that cross classifies y with a binary treatment variable x and an adjustment factor z, specify a logistic model with a lack of interaction between x and z. Construct the likelihood function, and explain the conditioning required to generate an exact conditional test for the effect of x. Explain how you would form a P-value for a one-sided alternative of a positive effect of x.

5.9 To use conditional logistic regression to test H_0: $\beta_1 = 0$ against H_1: $\beta_1 < 0$ for the toy example in Section 5.4.2, find the conditional distribution of $\sum_i x_i y_i$, given $\sum_i y_i$. Find the exact small-sample P-value.

5.10 The calibration problem is that of estimating x_0 at which $P(y = 1) = \pi_0$ for some fixed π_0 such as 0.50. For the logistic model with a single explanatory variable, explain why a confidence interval for x_0 is the set of x values for which

$$|\hat{\beta}_0 + \hat{\beta}_1 x - \text{logit}(\pi_0)|/[\text{var}(\hat{\beta}_0) + x^2\text{var}(\hat{\beta}_1) + 2x\,\text{cov}(\hat{\beta}_0, \hat{\beta}_1)]^{1/2} < z_{\alpha/2}.$$

How could you invert a likelihood-ratio test to form an interval?

5.11 Construct the log-likelihood function for the model $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$ with independent binomial proportions of y_1 successes in n_1 trials at $x_1 = 0$ and y_2

successes in n_2 trials at $x_2 = 1$. Derive the likelihood equations, and show that $\hat{\beta}_1$ is the sample log odds ratio.

5.12 Refer to the previous exercise. Denote the cell counts in the 2×2 table by $\{n_{ij}\}$. For the case $\beta_1 = 0$ (the *independence model*), the fitted values in the cells of that table are $\{\hat{\mu}_{ij} = n_{i+}n_{+j}/n\}$. These have a common value for the four $|n_{ij} - \hat{\mu}_{ij}|$.

a. Construct the Pearson residuals. Explain why all four may differ in absolute value.

b. The standardized residuals in this case are

$$r_{ij} = (n_{ij} - \hat{\mu}_{ij})/\sqrt{\hat{\mu}_{ij}[1 - (n_{i+}/n)][1 - (n_{+j}/n)]}.$$

Show that all four are identical in absolute value, thus appropriately recognizing that residual $df = 1$ for the independence model.

5.13 Suppose the logistic model holds in which x is uniformly distributed between 0 and 100, and $\text{logit}(\pi_i) = -2.0 + 0.04x_i$. Randomly generate 100 independent observations from this model. Plot the residuals against x and against the fitted values. Why do residual plots for binary data have this appearance?

5.14 Let $n_i y_i$ be a $\text{bin}(n_i, \pi_i)$ variate for group i, $i = 1, \ldots, N$, with $\{y_i\}$ independent. Consider the null model, for which $\pi_1 = \cdots = \pi_N$. Show that $\hat{\pi} = (\sum_i n_i y_i)/(\sum_i n_i)$. When all $n_i = 1$, for testing goodness of fit of the null model in the $N \times 2$ table, show that $X^2 = N$.

5.15 Let y_i be a $\text{bin}(1, \pi_i)$ variate, $i = 1, \ldots, N$. For the model $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$, show that the deviance depends on $\hat{\pi}_i$ but not y_i. Hence, it is not useful for checking model fit. (This exercise and the previous one show that goodness-of-fit statistics are uninformative for ungrouped data.)

5.16 A study has n_i independent binary observations $\{y_{i1}, \ldots, y_{in_i}\}$ at x_i, $i = 1, \ldots, N$, with $n = \sum_i n_i$. Consider the model $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$, where $\pi_i = P(y_{ij} = 1)$.

a. Show that the kernel of the likelihood function is the same if treating the data as n Bernoulli observations or N binomial observations.

b. For the saturated model, explain why the likelihood function is different for these two data forms. Hence, the deviance reported by software depends on the form of data entry.

c. Explain why the difference between deviances for two unsaturated models does not depend on the form of data entry.

5.17 Use the following toy data to illustrate comments in Section 5.5 about grouped versus ungrouped binary data in the effect on the deviance:

x	Number of trials	Number of successes
0	4	1
1	4	2
2	4	4

Denote by M_0 the null model $\text{logit}(\pi_i) = \beta_0$ and by M_1 the model $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$.

a. Create a data file in two ways, entering the data as (i) ungrouped data: $n_i = 1$, $i = 1, \ldots, 12$, (ii) grouped data: $n_i = 4$, $i = 1, 2, 3$. Fit M_0 and M_1 for each data file. Show that the deviances for M_0 and M_1 differ for the two forms of data entry. Why is this?

b. Show that the difference between the deviances for M_0 and M_1 is the same for each form of data entry. Why is this? (Thus, the data file format does not matter for inference, but it does matter for goodness-of-fit testing.)

5.18 Refer to the deviance comparison statistic $G^2(M_0 \mid M_1)$ introduced in Section 4.4.3. For a sequence of s nested binary response models $M_1, \ldots, M_s$, model M_s is the most complex. Let v denote the difference in residual *df* between M_1 and M_s.

a. Explain why for $j < k$, $G^2(M_j \mid M_k) \leq G^2(M_j \mid M_s)$.

b. Assume model M_j, so that M_k also holds when $k > j$. For all $k > j$, as $n \to \infty$, explain why $P[G^2(M_j \mid M_k) > \chi^2_v(\alpha)] \leq \alpha$.

c. Gabriel (1966) suggested a simultaneous testing procedure in which, for each pair of models, the critical value for differences between G^2 values is $\chi^2_v(\alpha)$. The final model accepted must be more complex than any model rejected in a pairwise comparison. Since part (**b**) is true for all $j < k$, argue that Gabriel's procedure has type I error probability no greater than α.

5.19 In a football league, for matches involving teams a and b, let π_{ab} be the probability that a defeats b. Suppose $\pi_{ab} + \pi_{ba} = 1$ (i.e., ties cannot occur). Bradley and Terry (1952) proposed the model

$$\log(\pi_{ab}/\pi_{ba}) = \beta_a - \beta_b.$$

For $a < b$, let N_{ab} denote the number of matches between teams a and b, with team a winning n_{ab} times and team b winning n_{ba} times.

a. Find the log-likelihood, treating n_{ab} as a binomial variate for N_{ab} trials. Show that sufficient statistics are $\{n_{a+}\}$, so that "victory totals" determine the estimated ranking of teams.

b. Generalize the model to allow a "home-team advantage," with a team's chance of winning possibly increasing when it plays at its home city. Interpret parameters.

5.20 Let y_i, $i = 1, \ldots, N$, denote N independent binary random variables.

a. Derive the log-likelihood for the probit model $\Phi^{-1}[\pi(\mathbf{x}_i)] = \sum_j \beta_j x_{ij}$.

b. Show that the likelihood equations for the logistic and probit regression models are

$$\sum_{i=1}^{N} (y_i - \hat{\pi}_i) z_i x_{ij} = 0, \qquad j = 1, \ldots, p,$$

where $z_i = 1$ for the logistic case and $z_i = \phi(\sum_j \hat{\beta}_j x_{ij})/\hat{\pi}_i(1 - \hat{\pi}_i)$ for the probit case.

5.21 An alternative latent variable model results from early applications of binary response models to toxicology studies (such as Table 5.4) of the effect of dosage of a toxin on whether a subject dies, with an unobserved *tolerance distribution*. For a randomly selected subject, let x_i denote the dosage level and let $y_i = 1$ if the subject dies. Suppose that the subject has a latent tolerance threshold T_i for the dosage, with $(y_i = 1)$ equivalent to $(T_i \leq x_i)$. Let $F(t) = P(T \leq t)$.

a. For fixed dosage x_i, explain why $P(y_i = 1 \mid x_i) = F(x_i)$.

b. Suppose F belongs to the normal parametric family, for some μ and σ. Explain why the model has the form

$$\Phi^{-1}(\pi_i) = \beta_0 + \beta_1 x_i$$

and relate β_0 and β_1 to μ and σ.

5.22 Consider the choice between two options, such as two product brands. Let U_y denote the *utility* of outcome y, for $y = 0$ and $y = 1$. Suppose $U_y = \beta_{y0} + \beta_{y1} x + \epsilon_y$, using a scale such that ϵ_y has some standardized distribution. A subject selects $y = 1$ if $U_1 > U_0$ for that subject.

a. If ϵ_0 and ϵ_1 are independent $N(0, 1)$ random variables, show that $P(y = 1)$ satisfies the probit model.

b. If ϵ_y are independent extreme-value random variables, with cdf $F(\epsilon) = \exp[-\exp(-\epsilon)]$, show that $P(y = 1)$ satisfies the logistic regression model (McFadden 1974).

5.23 When $\Phi^{-1}(\pi_i) = \beta_0 + \beta_1 x_i$, explain why the response curve for π_i [or for $1 - \pi_i$, when $\beta_1 < 0$] has the appearance of a normal cdf with mean $\mu = -\beta_0/\beta_1$ and standard deviation $\sigma = 1/|\beta_1|$. By comparison, explain why the logistic regression curve for π_i has mean $\mu = -\beta_0/\beta_1$ and standard deviation $\pi/|\beta_1|\sqrt{3}$. What does this suggest about relative magnitudes of estimates in logistic and probit models?

5.24 Consider binary GLM $F^{-1}(\pi_i) = \beta_0 + \beta_1 x_i$, where F is a cdf corresponding to a pdf f that is symmetric around 0. Show that x_i at which $\pi_i = 0.50$ is $x_i = -\beta_0/\beta_1$. Show that the rate of change in π_i when $\pi_i = 0.50$ is $\beta_1 f(0)$, and

find this for the logit and probit links. What does this suggest about relative magnitudes of estimates in logistic and probit models?

5.25 For the model $\log[-\log(1-\pi_i)] = \beta_0 + \beta_1 x_i$, find x_i at which $\pi_i = \frac{1}{2}$. Show that the greatest rate of change of π occurs at $x = -\beta_0/\beta_1$, and find π at that point. Give the corresponding result for the model with log–log link, and compare with the logistic and probit models.

5.26 In a study of the presence of tumors in animals, suppose $\{y_i\}$ are independent counts that satisfy a Poisson loglinear model, $\log(\mu_i) = \sum_j \beta_j x_{ij}$. However, the observed response merely indicates whether each y_i is positive, $z_i = I(y_i > 0)$, for the indicator function I. Show that $\{z_i\}$ satisfy a binary GLM with complementary log–log link (Dunson and Herring 2005).

5.27 Suppose $y = 0$ at $x = 10, 20, 30, 40$ and $y = 1$ at $x = 60, 70, 80, 90$. Using software, what do you get for estimates and standard errors when you fit the logistic regression model (**a**) to these data? (**b**) to these eight observations and two observations at $x = 50$, one with $y = 1$ and one with $y = 0$? (**c**) to these eight observations and observations at $x = 49.9$ with $y = 1$ and at $x = 50.1$ with $y = 0$? In cases (**a**) and (**b**), explain why actually the ML estimate $\hat{\beta} = \infty$. Why does software report such a large SE for $\hat{\beta}$? In case (**a**), what is the reported maximized log-likelihood value. Why?

5.28 For the logistic model (5.7) for a 2×2 table, give an example of cell counts corresponding to (**a**) complete separation and $\hat{\beta}_1 = \infty$, (**b**) quasi-complete separation and $\hat{\beta}_1 = \infty$, (**c**) non-existence of $\hat{\beta}_1$.

5.29 You plan to study the relation between x = age and y= whether belong to a social network such as Facebook (1 = yes). A priori, you predict that $P(y = 1)$ is currently between about 0.80 and 0.90 at $x = 18$ and between about 0.20 and 0.30 at $x = 65$. If the logistic regression model describes this relation well, what is a plausible range of values for the effect β_1 of x in the model?

5.30 In one of the first studies of the link between lung cancer and smoking[7], Richard Doll and Austin Bradford Hill collected data from 20 hospitals in London, England. Each patient admitted with lung cancer in the preceding year was queried about their smoking behavior. For each of the 709 patients admitted, they recorded the smoking behavior of a noncancer patient at the same hospital of the same gender and within the same 5-year grouping on age. A smoker was defined as a person who had smoked at least one cigarette a day for at least a year. Of the 709 cases having lung cancer, 688 reported being smokers. Of the 709 controls, 650 reported being smokers. Specify a relevant logistic regression model, explain what can be estimated and what cannot (and why), and conduct a statistical analysis.

[7]See *British Med. J.*, Sept. 30, 1950, pp. 739–748.

5.31 To illustrate Fisher's exact test, Fisher (1935) described the following experiment: a colleague of his claimed that, when drinking tea, she could distinguish whether milk or tea was added to the cup first (she preferred milk first). To test her claim, Fisher asked her to taste eight cups of tea, four of which had milk added first and four of which had tea added first. She knew there were four cups of each type and had to predict which four had the milk added first. The order of presenting the cups to her was randomized. For the 2×2 table relating what was actually poured first to the guess of what was poured first, explain how to use Fisher's exact test to evaluate whether her ability to distinguish the order of pouring was better than with random guessing. Find the P-value if she guesses correctly for three of the four cups that had milk poured first.

5.32 For the horseshoe crab dataset (`Crabs.dat` at the text website) introduced in Section 4.4.3, let $y = 1$ if a female crab has at least one satellite, and let $y = 0$ if a female crab does not have any satellites. Fit a main-effects logistic model using color and weight as explanatory variables. Interpret and show how to conduct inference about the color and weight effects. Next, allow interaction between color and weight in their effects on y, and test whether this model provides a significantly better fit.

5.33 The dataset `Crabs2.dat` at the text website collects several variables that may be associated with y = whether a female horseshoe crab is monandrous (eggs fertilized by a single male crab) or polyandrous (eggs fertilized by multiple males). A probit model that uses as explanatory variables *Fcolor* = the female crab's color (1 = dark, 3 = medium, 5 = light) and *Fsurf* = her surface condition (values 1, 2, 3, 4, 5 with lower values representing worse) has the output shown. Interpret the parameter estimates and the inferential results. Approximately what values would you expect for the ML estimate of the Fsurf effect and its *SE* if you fitted the corresponding logistic model?

```
-------------------------------------------------------------
                Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)      -0.3378      0.1217   -2.775  0.005522
factor(Fcolor)3   0.4797      0.1065    4.504  6.66e-06
factor(Fcolor)5   0.1651      0.1158    1.426  0.153902
Fsurf            -0.1360      0.0376   -3.619  0.000296
---
    Null deviance: 1633.8  on 1344  degrees of freedom
Residual deviance: 1587.8  on 1341  degrees of freedom
-------------------------------------------------------------
```

5.34 Refer to the previous exercise. Download the file from the text website. Using *year* of observation, *Fcolor*, *Fsurf*, *FCW* = female's carapace width, *AMCW* = attached male's carapace width, *AMcolor* = attached male's color, and *AMsurf* = attached male's surface condition, conduct a logistic model-building process, including descriptive and inferential analyses. Prepare a

report summarizing this process (with edited software output as an appendix), also interpreting results for your chosen model.

5.35 *The New York Times* reported results of a study on the effects of AZT in slowing the development of AIDS symptoms (February 15, 1991). Veterans whose immune symptoms were beginning to falter after infection with HIV were randomly assigned to receive AZT immediately or wait until their T cells showed severe immune weakness. During the 3-year study, of those who received AZT, 11 of 63 black subjects and 14 of 107 white subjects developed AIDS symptoms. Of those who did not receive AZT, 12 of 55 black subjects and 32 of 113 white subjects developed AIDS symptoms. Use model building, including checking fit and interpreting effects and inference, to analyze these data.

5.36 Download the data for the example in Section 5.7.1. Fit the main effects model. What does your software report for $\hat{\beta}_1$ and its *SE*? How could you surmise from the output that actually $\hat{\beta}_1 = \infty$?

5.37 Refer to the previous exercise. For these data, what, if anything, can you learn about potential interactions for pairs of the explanatory variables? Conduct the likelihood-ratio test of the hypothesis that all three interaction terms are 0.

5.38 Table 5.5 shows data, the file `SoreThroat.dat` at the text website, from a study about y = whether a patient having surgery experienced a sore throat on waking (1 = yes, 0 = no) as a function of d = duration of the surgery (in minutes) and t = type of device used to secure the airway (1 = tracheal tube, 0 = laryngeal mask airway). Use a model-building strategy to select a GLM for binary data. Interpret parameter estimates and conduct inference about the effects.

Table 5.5 Data for Exercise 5.38 on Surgery and Sore Throats

Patient	*d*	*t*	*y*	Patient	*d*	*t*	*y*	Patient	*d*	*t*	*y*
1	45	0	0	13	50	1	0	25	20	1	0
2	15	0	0	14	75	1	1	26	45	0	1
3	40	0	1	15	30	0	0	27	15	1	0
4	83	1	1	16	25	0	1	28	25	0	1
5	90	1	1	17	20	1	0	29	15	1	0
6	25	1	1	18	60	1	1	30	30	0	1
7	35	0	1	19	70	1	1	31	40	0	1
8	65	0	1	20	30	0	1	32	15	1	0
9	95	0	1	21	60	0	1	33	135	1	1
10	35	0	1	22	61	0	0	34	20	1	0
11	75	0	1	23	65	0	1	35	40	1	0
12	45	1	1	24	15	1	0				

Source: Data from Collett (2005) with permission of John Wiley & Sons, Inc.