a. For each age, compute the sample coronary death rates per 1000 person-years, for nonsmokers and smokers. To compare them, take their ratio and describe its dependence on age.

b. Fit a main-effects model for the log rates having four parameters for age and one for smoking. In discussing lack of fit, show that this model assumes a constant ratio of nonsmokers' to smokers' coronary death rates over levels of age.

c. Based on (**a**), explain why it is sensible to add a quantitative interaction of age and smoking. Assign scores to the levels of age, and add a term based on the product of age and smoking. For this model, show that the log of the ratio of coronary death rates changes linearly with age. Fit the model, and interpret.

4.17. For rate data, the Poisson GLM with identity link is

$$\frac{\mu}{t} = \alpha + \beta x.$$

a. Since the model has form $\mu = \alpha t + \beta tx$, argue that it is equivalent to a Poisson GLM for the response totals at the various levels of x, using identity link with t and tx as explanatory variables and no intercept or offset terms.

b. Fit this model to the grouped data in Table 4.3, using average width scores. Compare results, including interpretations, goodness of fit, and residual analyses, to those obtained with the log link.

CHAPTER 5

Logistic Regression

Let's now take a closer look at the statistical modeling of binary response variables, for which the response measurement for each subject is a "success" or "failure." Binary data are perhaps the most common form of categorical data, and the methods of this chapter are of fundamental importance. The most popular model for binary data is *logistic regression*. Section 4.2.3 introduced this model as a generalized linear model (GLM) for a binomial random component. This chapter studies the application of logistic regression in greater detail.

The first section discusses interpretation of the logistic regression model. Section 5.2 presents statistical inference for the model parameters, and Section 5.3 presents ways of checking the model fit. Section 5.4 shows how to handle qualitative predictors in the model, and Section 5.5 discusses the extension of the model for multiple explanatory variables. Section 5.6 discusses determination of the sample size needed to obtain adequate inferential power. Finally, Section 5.7 presents small-sample, exact inference for logistic regression.

5.1 INTERPRETING THE LOGISTIC REGRESSION MODEL

For a binary response Y and a quantitative explanatory variable X, let $\pi(x)$ denote the "success" probability when X takes value x. This probability is the parameter for the binomial distribution. The logistic regression model has linear form for the logit of this probability,

$$\text{logit}[\pi(x)] = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x. \qquad (5.1.1)$$

The formula implies that $\pi(x)$ increases or decreases as an S-shaped function of x (recall Figure 4.2).

An alternative formula for logistic regression refers directly to the success probability. This formula uses the exponential function $\exp(x) = e^x$, in the form

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}. \qquad (5.1.2)$$

This section shows ways of interpreting these model formulas.

5.1.1 Linear Approximation Interpretations

The parameter β determines the rate of increase or decrease of the S-shaped curve. The sign of β indicates whether the curve ascends or descends, and the rate of change increases as $|\beta|$ increases. When the model holds with $\beta = 0$, the right-hand side of (5.1.2) simplifies to a constant. Then, $\pi(x)$ is identical at all x, so the curve becomes a horizontal straight line. The binary response Y is then independent of X.

Figure 5.1 shows the S-shaped appearance of the logistic regression model for $\pi(x)$, as fitted for the example in the following subsection. Since it has a curved rather than linear appearance, the function (5.1.2) implies that the rate of change in $\pi(x)$ per unit change in x varies. A straight line drawn tangent to the curve at a particular x value, such as shown in Figure 5.1, describes the rate of change at that point. For logistic regression parameter β, that line has slope equal to $\beta\pi(x)[1 - \pi(x)]$. For instance, the line tangent to the curve at x for which $\pi(x) = .5$ has slope $\beta(.5)(.5) = .25\beta$; by contrast, when $\pi(x) = .9$ or .1, it has slope $.09\beta$. The slope approaches 0 as the probability approaches 1.0 or 0.

The steepest slope of the curve occurs at x for which $\pi(x) = .5$; that x value is $x = -\alpha/\beta$. (One can check that $\pi(x) = .5$ at this point by substituting $-\alpha/\beta$ for x in (5.1.2), or by substituting $\pi(x) = .5$ in (5.1.1) and solving for x.) This x value

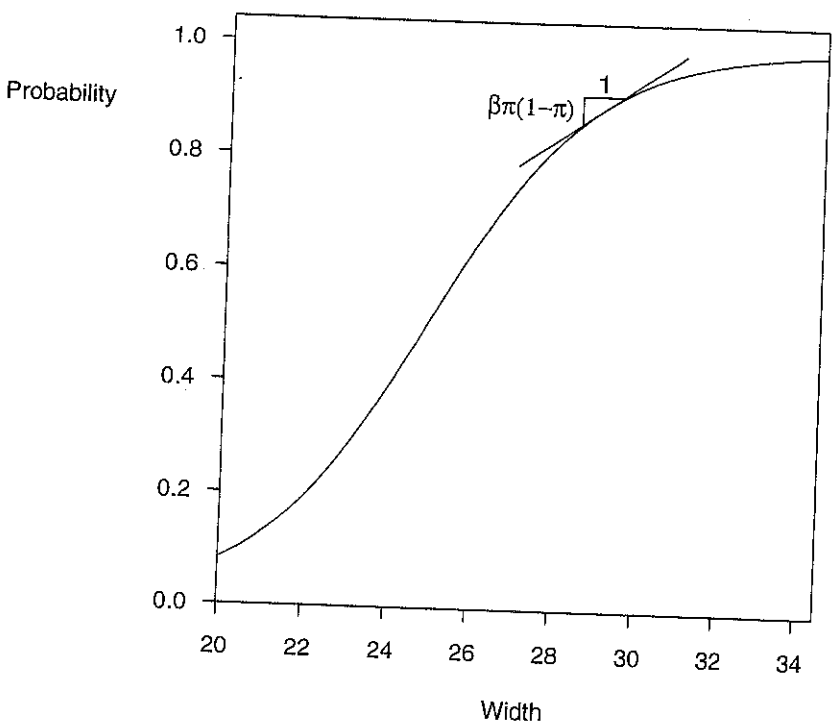

Figure 5.1 Linear approximation to logistic regression curve.

is sometimes called the *median effective level* and is denoted EL_{50}. It represents the level at which each outcome has a 50% chance.

5.1.2 Horseshoe Crabs Revisited

Maximum likelihood (ML) computations for fitting logistic regression models are complex, but are easy to perform using statistical software. To illustrate the model, we reanalyze the horseshoe crab data introduced in Section 4.3.2. We use the binary response of whether a female crab has any satellites present; that is, $Y = 1$ if a female crab has at least one satellite, and $Y = 0$ if she has no satellite. We first use the female crab's width as the sole predictor. Section 5.5 discusses models having additional predictors.

Figure 5.2 shows a plot of the data. It consists of a set of points at the level $Y = 1$ and a second set of points at the level $Y = 0$. The numbered symbols indicate the number of observations at each point. It appears that $Y = 1$ tends to occur relatively more often at higher x values. Since Y takes only values 0 and 1, however, it is difficult to determine whether a logistic regression model is reasonable

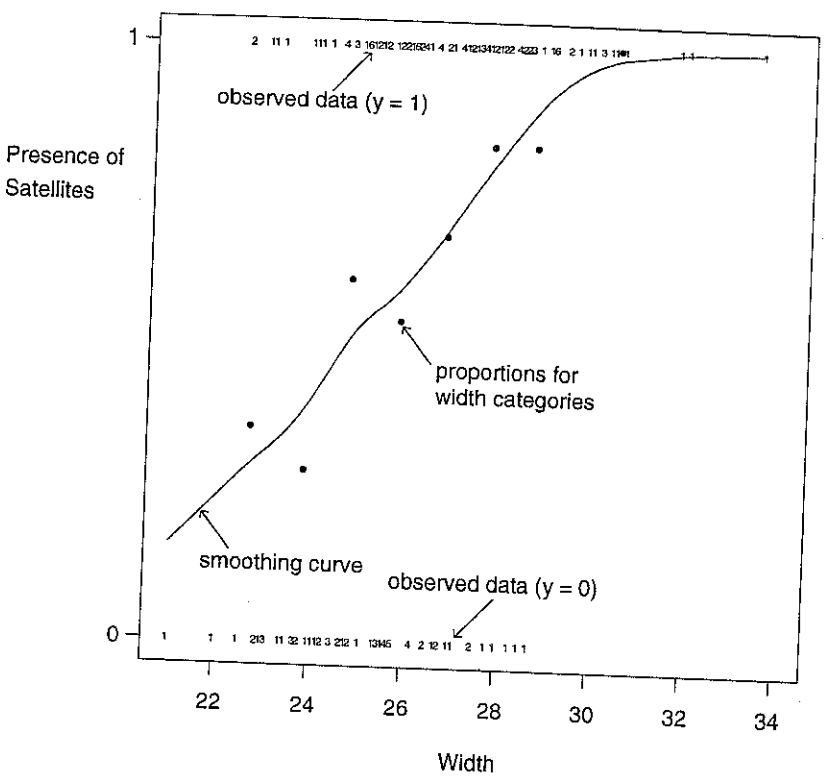

Figure 5.2 Whether satellites are present ($Y = 1$, yes; $Y = 0$, no), by width of female crab.

Table 5.1 Relation Between Width of Female Crab and Existence of Satellites, and Predicted Values for Logistic Regression Model

Width	Number Cases	Number Having Satellites	Sample Proportion	Predicted Probability	Predicted Number Crabs with Satellites
< 23.25	14	5	.36	.26	3.64
23.25–24.25	14	4	.29	.38	5.31
24.25–25.25	28	17	.61	.49	13.78
25.25–26.25	39	21	.54	.62	24.23
26.25–27.25	22	15	.68	.72	15.94
27.25–28.25	24	20	.83	.81	19.38
28.25–29.25	18	15	.83	.87	15.65
> 29.25	14	14	1.00	.93	13.08

by plotting Y against x. Better information results from grouping the width values into categories and calculating a sample proportion of crabs having satellites for each category. This reveals whether the true proportions follow approximately the trend required by this model. Consider the grouping used to investigate adequacy of Poisson regression models in Section 4.4.2, shown again in Table 5.1. In each of the eight width categories, we computed the sample proportion of crabs having satellites and the mean width for the crabs in that category. Figure 5.2 also contains eight dots representing the sample proportions of female crabs having satellites plotted against the mean widths for the eight categories.

Alternatively, some software can smooth the data, revealing a general trend without assuming a particular functional form for the relationship. Smoothing methods based on *generalized additive models* do this by providing even more general structural form than GLMs. For instance, they find possibly complex functions of the explanatory variables that serve as the "best" predictors of a certain type. Figure 5.2 also shows a curve based on smoothing the data using this method. The eight plotted sample proportions and this smoothing curve both show a roughly increasing trend, so we proceed with fitting models that imply such trends.

For the ungrouped data from Table 4.2, let $\pi(x)$ denote the probability that a female horseshoe crab of width x has a satellite. The simplest model to interpret is the linear probability model, $\pi(x) = \alpha + \beta x$. For these data, some predicted values for this GLM fall outside the legitimate range for a binomial parameter, so ML fitting fails. Ordinary least squares fitting yields $\hat{\pi}(x) = -1.766 + 0.092x$. The predicted probability of a satellite increases by .092 for each 1-cm increase in width. This model provides a simple interpretation and realistic predictions over most of the width range, but it is inadequate for extreme values. For instance, at the maximum width in this sample of 33.5, its predicted probability equals $-1.766 + 0.092(33.5) = 1.3$.

The ML parameter estimates for the logistic regression model are $\hat{\alpha} = -12.351$ and $\hat{\beta} = 0.497$. The predicted probability of a satellite is the sample analog of (5.1.2),

$$\hat{\pi} = \frac{\exp(-12.351 + 0.497x)}{1 + \exp(-12.351 + 0.497x)}.$$

Since $\hat{\beta} > 0$, the predicted probability $\hat{\pi}$ is higher at larger width values. At the minimum width in this sample of 21.0 cm, the predicted probability is $\hat{\pi} = \exp(-12.351 + 0.497(21.0))/[1 + \exp(-12.351 + 0.497(21.0))] = .129$; at the maximum width of 33.5 cm, the predicted probability equals $\exp(-12.351 + 0.497(33.5))/[1 + \exp(-12.351 + 0.497(33.5))] = .987$. The median effective level is the width at which the predicted probability equals .5, which is $x = EL_{50} = -\hat{\alpha}/\hat{\beta} = 12.351/0.497 = 24.8$. Figure 5.1 plots the predicted probabilities as a function of width.

At the sample mean width of 26.3 cm, the predicted probability of a satellite equals .674. The incremental rate of change in the fitted probability at that point is $\hat{\beta}\hat{\pi}(1 - \hat{\pi}) = 0.497(0.674)(0.326) = .11$. For female crabs near the mean width, the estimated probability of a satellite increases at the rate of .11 per cm increase in width. The predicted rate of change is greatest at the x value (24.8) at which $\hat{\pi} = .5$; there, the predicted probability increases at the rate of $(0.497)(0.5)(0.5) = .12$ per cm increase in width. Unlike the linear probability model, the logistic regression model permits the rate of change to vary as x varies.

To further describe the fit, for each category of width Table 5.1 reports the predicted number of crabs having satellites (i.e., the fitted values). To get these, one adds the predicted probabilities for all crabs in the category; for instance, the predicted probabilities for the 14 crabs with widths below 23.25 cm sum to 3.64. The average predicted probability for female crabs in a given width category equals the fitted value divided by the number of female crabs in that category. For the first width category, $3.64/14 = .26$ is the average predicted probability. Table 5.1 reports the fitted values and the average predicted probabilities in grouped fashion. An eyeball comparison of these to the sample counts of crabs having satellites and the sample proportions suggests that the model fits decently. Section 5.3 presents objective criteria for making this comparison.

5.1.3 Odds Ratio Interpretation

Another interpretion of the logistic regression model uses the *odds* and the *odds ratio*. For model (5.1.1), the odds of response 1 (i.e., the odds of a "success") are

$$\frac{\pi(x)}{1 - \pi(x)} = \exp(\alpha + \beta x) = e^{\alpha}(e^{\beta})^{x}. \tag{5.1.3}$$

This exponential relationship provides an interpretation for β: The odds increase multiplicatively by e^{β} for every one-unit increase in x. That is, the odds at level $x + 1$ equal the odds at x multiplied by e^{β}. When $\beta = 0$, $e^{\beta} = 1$, and the odds do not change as x changes.

For the female horseshoe crabs, the estimated odds of a satellite multiply by $\exp(\hat{\beta}) = \exp(0.497) = 1.64$ for each centimeter increase in width; that is, there is a 64% increase. To illustrate, the mean width value of $x = 26.3$ has a predicted probability of a satellite equal to .674, and odds $= .674/.326 = 2.07$. At $x = 27.3 = 26.3 + 1.0$, one can check that the predicted probability equals .773, and odds $= .773/.227 = 3.40$. But this is a 64% increase; that is, $3.40 = 2.07(1.64)$.

The logarithm of the odds, which is the logit transform of $\pi(x)$, has the linear relationship (5.1.1). This is the logit expression of the model, originally introduced in Section 4.2.3. It states that the logit increases by β units for every 1-unit change in x. Most of us do not think naturally on a logit scale, so this interpretation has limited use.

5.1.4 Logistic Regression with Case-Control Studies*

Another property of the logistic regression model relates to situations in which the explanatory variable X rather than the response variable Y is random. Most commonly, this occurs with retrospective sampling designs, such as case-control studies (Section 2.3.4). For samples of subjects having $Y = 1$ ("cases") and having $Y = 0$ ("controls"), the value of X is observed. Evidence exists of an association between X and Y if the distribution of X values differs between cases and controls.

Many biomedical studies, particularly epidemiological studies, use case-control designs. For retrospective data, Section 2.3.4 noted that one can estimate odds ratios. Logistic regression parameters refer to odds and odds ratios. Thus, one can fit such models to retrospective data, and one can estimate effects in case-control studies. This is not true of other models for binary responses, since the odds ratio is not their natural measure for describing effects. This provides an important advantage of the logit link over links such as the probit and is a major reason why the logit model has surpassed the others in popularity. Section 9.2.3 discusses the application of logistic regression to case-control studies that match each case with a single control.

Regardless of the sampling mechanism, the logistic regression model may or may not describe a relationship well. In one special case, it does necessarily hold. Suppose that the distribution of X for all subjects having $Y = 1$ is normal $N(\mu_1, \sigma)$, and suppose the distribution of X for all subjects having $Y = 0$ is normal $N(\mu_0, \sigma)$; that is, with different mean but the same standard deviation. Then one can show that $\pi(x)$ satisfies the logistic regression curve, with β having the same sign as $\mu_1 - \mu_0$. When a population consists of a mixture of two types of subjects, one set with $Y = 1$ having a bell-shaped distribution on X and the other set with $Y = 0$ having another bell-shaped distribution with similar spread, then the logistic regression function (5.1.2) approximates well the curve for $\pi(x)$. If the distributions are bell-shaped but with highly different spreads, then a model containing also a quadratic term (i.e., both x and x^2) often fits well. In that case, the relationship is nonmonotone, with $\pi(x)$ increasing and then decreasing, or the reverse (Problem 5.5).

5.2 INFERENCE FOR LOGISTIC REGRESSION

We have studied how the fit of a logistic regression model helps us describe the effects of a predictor on a binary response variable. We next present statistical inference for the model parameters, to help judge the significance and size of the effects. Widely available software reports the parameter estimates and their standard errors, as well as other information about the model fit.

5.2.1 Confidence Intervals for Effects

A large-sample confidence interval for the parameter β in the logistic regression model, $\text{logit}[\pi(x)] = \alpha + \beta x$, is

$$\hat{\beta} \pm z_{\alpha/2}(ASE).$$

Exponentiating the endpoints of this interval yields one for e^{β}, the multiplicative effect on the odds of a 1-unit increase in X.

To illustrate, we continue our logistic regression analysis of the horseshoe crab data. The estimated effect of width in the fitted equation for the probability of a satellite is $\hat{\beta} = 0.497$, with $ASE = 0.102$. A 95% confidence interval for β is $0.497 \pm 1.96(0.102)$, or $(0.298, 0.697)$. The confidence interval for the effect on the odds per centimeter increase in width equals $(e^{.298}, e^{.697}) = (1.35, 2.01)$. We infer that each centimeter increase in width has at least a 35% increase and at most a doubling in the odds that a female crab has a satellite.

Section 5.1.1 noted that simpler interpretations result from linear approximations to the logistic regression curve. The term $\beta\pi(1 - \pi)$ approximates the change in the probability per unit change in x. For instance, at $\pi = .5$, the estimated rate of change is $0.25\hat{\beta} = .124$. A 95% confidence interval for 0.25β equals 0.25 times the endpoints of the interval for β, or $[0.25(.298), 0.25(.697)] = (.074, .174)$. If the logistic regression model holds, for values of x near the width at which $\pi = .5$, the rate of increase in the probability of a satellite per centimeter increase in width falls between about .07 and .17.

5.2.2 Significance Testing

We next discuss significance tests for the effect of X on the binary response. For the logistic regression model, the null hypothesis $H_0 : \beta = 0$ states that the probability of success is independent of X.

For large samples, the test statistic

$$z = \frac{\hat{\beta}}{ASE}$$

has a standard normal distribution when $\beta = 0$. One can refer z to the standard normal table to get a one-sided or two-sided P-value in the usual manner. Equivalently, for the two-sided alternative $\beta \neq 0$, $(\hat{\beta}/ASE)^2$ is a Wald statistic (Section 4.4.1) having a large-sample chi-squared distribution with $df = 1$.

Though the Wald test works well for very large samples, the likelihood-ratio test (Section 4.4.1) is more powerful and reliable for sample sizes used in practice. The test statistic compares the maximum L_0 of the log-likelihood function when $\beta = 0$ (i.e., when $\pi(x)$ is forced to be identical at all x values) to the maximum L_1 of the log-likelihood function for unrestricted β. The test statistic, $-2(L_0 - L_1)$, also has a large-sample chi-squared distribution with $df = 1$. Most software for logistic

regression reports the maximized log-likelihoods L_0 and L_1 and the likelihood-ratio statistic derived from those maxima.

For the horseshoe crab data, the statistic $z = \hat{\beta}/ASE = 0.497/0.102 = 4.9$ shows strong evidence of a positive effect of width on the presence of satellites ($P < .0001$). The equivalent Wald chi-squared statistic, $z^2 = 23.9$, has $df = 1$. The maximized log likelihoods equal $L_0 = -112.88$ under $H_0 : \beta = 0$ and $L_1 = -97.23$ for the full model. The likelihood-ratio statistic equals $-2(L_0 - L_1) = 31.3$, with $df = 1$. This provides even stronger evidence than the Wald statistic of a width effect.

5.2.3 Distribution of Probability Estimates*

The estimated probability that $Y = 1$ at a fixed setting x of X equals

$$\hat{\pi}(x) = \frac{\exp(\hat{\alpha} + \hat{\beta}x)}{1 + \exp(\hat{\alpha} + \hat{\beta}x)}. \tag{5.2.1}$$

Most software for logistic regression can report such estimates as well as confidence intervals for the true probabilities.

We illustrate by estimating the probability of a satellite for female crabs of width $x = 26.5$, which is near the mean width. The logistic regression fit yields the estimate, $\hat{\pi}(26.5) = \exp(-12.351 + 0.497(26.5))/[1 + \exp(-12.351 + 0.497(26.5))] = .695$. From software, a 95% confidence interval for the true probability is (.61, .77).

One can construct confidence intervals for probabilities using the covariance matrix of the model parameter estimates. The term $\hat{\alpha} + \hat{\beta}x$ in the exponents of the prediction equation (5.2.1) is the estimated linear predictor in the logit transform (5.1.1) of $\pi(x)$. This estimated logit has large-sample *ASE* given by the estimated square root of

$$\text{Var}(\hat{\alpha} + \hat{\beta}x) = \text{Var}(\hat{\alpha}) + x^2\,\text{Var}(\hat{\beta}) + 2x\,\text{Cov}(\hat{\alpha}, \hat{\beta}).$$

A 95% confidence interval for the true logit is $(\hat{\alpha} + \hat{\beta}x) \pm 1.96ASE$. Substituting the endpoints of this interval for $\alpha + \beta x$ in the exponents of (5.2.1) gives a corresponding interval for the probability.

For instance, at $x = 26.5$, the predicted logit is $-12.351 + 0.497(26.5) = 0.825$. Software reports $\hat{\text{V}}\text{ar}(\hat{\alpha}) = 6.910$, $\hat{\text{V}}\text{ar}(\hat{\beta}) = 0.01035$, $\hat{\text{C}}\text{ov}(\hat{\alpha}, \hat{\beta}) = -0.2668$, from which the estimated variance of this predicted logit equals $(6.910) + (26.5)^2(0.01035) + 2(26.5)(-0.2668) = 0.038$. The 95% confidence interval for the true logit equals $0.825 \pm (1.96)\sqrt{0.038}$, or (0.44, 1.21). From (5.2.1), this translates to the interval

$$\left(\frac{\exp(0.44)}{1 + \exp(0.44)}, \frac{\exp(1.21)}{1 + \exp(1.21)}\right) = (.61, .77)$$

for the probability of satellites at width 26.5 cm.

One could ignore the model fit and simply use sample proportions to estimate such probabilities. Six crabs in the sample had width 26.5, and four of them had satellites.

The sample proportion estimate at $x = 26.5$ is $p = \frac{4}{6} = .67$, similar to the model-based estimate. From inverting small-sample tests using the binomial distribution, a 95% confidence interval based on these six observations alone equals (.22, .96).

When the logistic regression model truly holds, the model-based estimator of a probability is considerably better than the sample proportion. The model has only two parameters to estimate, whereas the nonmodel-based approach has a separate parameter for every distinct value of X. For instance, at $x = 26.5$, software reports an $ASE = 0.04$ for the model-based estimate .695, whereas the estimated standard error is $\sqrt{p(1-p)/n} = \sqrt{(.67)(.33)/6} = 0.19$ for the sample proportion of .67 based on only 6 observations. The 95% confidence intervals are (.61, .77) versus (.22, .96). Instead of using only 6 observations, the model uses the information that all 173 observations provide in estimating the two model parameters. The result is a much more precise estimate.

Reality is a bit more complicated. In practice, any model will not exactly represent the true relationship between $\pi(x)$ and x. Thus, as the sample size increases, the model-based estimator may not converge exactly to the true value of the probability. This does not imply, however, that the sample proportion is actually a better estimator in practice. If the model approximates the true probabilities decently, its estimator still tends to be much closer than the sample proportion to the true value. The model smooths the sample data, somewhat dampening the observed variability. The resulting estimators tend to be better unless each sample proportion is based on an extremely large sample.

In summary, if the logistic regression model approximates well the true dependence of $\pi(x)$ on x, point and interval estimates of $\pi(x)$ based on it are quite useful. The next section shows how to investigate whether the model does in fact provide an adequate fit to the sample data.

5.3 MODEL CHECKING

So far, we have used logistic regression for description and inference about the effects of predictors on binary responses. There is no guarantee, however, that a particular model of this form is appropriate or that it provides a good fit to the data. This section discusses ways of checking the model fit.

Fitted logistic regression models provide predicted probabilities that $Y = 1$. At each setting of the explanatory variables, one can multiply the predicted probability by the number of subjects to obtain a fitted count. Similarly, one can obtain the fitted count for $Y = 0$ at each setting. The test of the null hypothesis that the model holds compares the fitted and observed counts using a Pearson X^2 or likelihood-ratio G^2 test statistic.

For a fixed number of settings, when most fitted counts equal at least about 5, X^2 and G^2 have approximate chi-squared distributions. The degrees of freedom, called the *residual df* for the model, equal the number of sample logits (i.e., the number of settings of explanatory variables) minus the number of model parameters. As usual, large X^2 or G^2 values provide evidence of lack of fit, and the P-value is the right-

tail probability above the observed value. When the fit is poor, residuals and other diagnostic measures describe the influence of individual observations on the model fit and highlight reasons for the inadequacy.

5.3.1 Goodness of Fit for Models with Continuous Predictors

We first present a goodness-of-fit analysis for the model using x = width to predict the probability $\pi(x)$ that a female crab has a satellite,

$$\text{logit}[\pi(x)] = \alpha + \beta x. \tag{5.3.1}$$

Width takes 66 distinct values for the 173 crabs, with few observations at most widths. One could regard the data as a 66×2 contingency table, in which the two cells in each row give the counts of the number of crabs with satellites and the number of crabs without satellites, at that width. The cell counts in this table are small, as are the fitted counts.

The large-sample theory for X^2 and G^2 applies for a fixed number of cells when the fitted counts are large. This theory is violated for the 66×2 table in two ways. First, most fitted counts are very small. Second, when more data are collected, additional width values would occur, so the contingency table would contain more cells rather than a fixed number. Because of this, X^2 and G^2 for logistic regression models fitted with continuous or nearly-continuous predictors do not have approximate chi-squared distributions. These indices of fit are more properly applied when the explanatory variables are categorical, and relatively few fitted counts are small.

To check the adequacy of logistic regression for these data, we compare the observed and fitted values in the grouped form of Table 5.1, shown again in Table 5.2, which is an 8×2 table. In each width category, the fitted value for response "yes" is the sum of the predicted probabilities $\hat{\pi}(x)$ for all crabs having width in that category; the fitted value for response "no" is the sum of $1 - \hat{\pi}(x)$ for those crabs. The fitted values displayed in this form are much larger than in the original 66×2 table, and chi-squared statistics for testing the model have better validity. Substituting the 16

Table 5.2 Grouping of Observed and Fitted Values for Fit of Logistic Regression Model to Horseshoe Crab Data

Width	Number Yes	Number No	Fitted Yes	Fitted No
< 23.25	5	9	3.64	10.36
23.25-24.25	4	10	5.31	8.69
24.25-25.25	17	11	13.78	14.22
25.25-26.25	21	18	24.23	14.77
26.25-27.25	15	7	15.94	6.06
27.25-28.25	20	4	19.38	4.62
28.25-29.25	15	3	15.65	2.35
> 29.25	14	0	13.08	0.92

grouped observed counts and fitted values into the standard chi-squared statistics,

$$X^2 = \sum \frac{(\text{observed} - \text{fitted})^2}{\text{fitted}} = 5.3,$$

and

$$G^2 = 2\sum(\text{observed}) \log\left(\frac{\text{observed}}{\text{fitted}}\right) = 6.2.$$

Table 5.2 has 8 sample logits, one for each width setting; the logistic regression model (5.3.1) has two parameters, so $df = 8 - 2 = 6$. Neither X^2 nor G^2 shows evidence of lack of fit ($P > .4$). Thus, we can feel more comfortable about the use of the model in Sections 5.1 and 5.2 for the original ungrouped data.

A simpler but more approximate method for obtaining goodness-of-fit statistics fits the logistic regression model directly to the observed counts in the 8×2 table. To treat width as quantitative, we assign scores to its categories, such as the mean widths $\{22.69, 23.84, 24.77, 25.84, 26.79, 27.74, 28.67, 30.41\}$ for the crabs in each category. The logit prediction equation then equals $\text{logit}[\hat{\pi}(x)] = -11.51 + 0.465x$, which yields a set of predicted probabilities and fitted values. For this fit, $X^2 = 5.0$ and $G^2 = 6.0$, based on $df = 6$. Results are similar to the statistics using fitted values based on predicted probabilities at the individual width values.

When explanatory variables are continuous, it is difficult to analyze lack of fit without some type of grouping. As the number of explanatory variables increases, however, simultaneous grouping of values for each variable can produce a contingency table with a large number of cells, many of which have small counts. An alternative way of grouping forms observed and fitted values based on a partitioning of predicted probabilities. One can regard the grouping in Table 5.2 as having been done in this way. For the model fitted, the 14 crabs in the first width category are the ones with the smallest predicted probabilities of a satellite; the 14 crabs in the second width category have higher predicted probabilities than the crabs in the first category, but smaller predicted probabilities than crabs in the other categories, and so forth.

Regardless of how many predictors are in the model, one can partition observed and fitted values according to the predicted probabilities. One common approach forms the groups in the partition so they have approximately equal size. To form 10 groups, for instance, one pair of observed and fitted counts refers to the $n/10$ observations having the highest predicted probabilities, another pair refers to the $n/10$ observations having the second decile of predicted probabilities, and so forth. In practice, it is usually not possible to form groups of exactly equal size because sets of observations have the same predicted probability, and all observations having the same predicted probability are kept in the same group. For each group, the fitted value for an outcome is the sum of the predicted probabilities for that outcome for all observations in that group.

This construction is the basis of a test due to Hosmer and Lemeshow (1989, p. 140). Their Pearson-like statistic does not actually have a chi-squared distribution,

but simulations have shown that its distribution is roughly approximated by chi-squared with $df = g - 2$, where g denotes the number of groups. We applied their test with $g = 10$ approximately equally sized groups for the logistic regression model fitted to the ungrouped data (Table 4.2). The Hosmer–Lemeshow statistic equals 3.5, based on $df = 8$, indicating a decent fit.

One can also detect lack of fit by using a likelihood-ratio test to compare the working model to more complex ones, as we will illustrate in Section 5.5. For instance, we might consider more complex models containing nonlinear effects (such as quadratic terms) for quantitative predictors or interaction terms. If we do not find a more complex model that provides a better fit, this provides some assurance that our fitted model is reasonable. This approach is more useful from a scientific perspective. A large goodness-of-fit statistic simply indicates there is *some* lack of fit, but provides no insight about its nature. Comparing a model to a more complex model, on the other hand, indicates whether lack of fit exists of a particular type.

5.3.2 Goodness of Fit and Likelihood-Ratio Model Comparison Tests

Sections 4.4.1 and 4.5.2 introduced the likelihood-ratio statistic $-2(L_0 - L_1)$ for testing whether certain parameters in a model equal zero. The test compares the maximized log likelihood (L_1) for the model to the maximized log likelihood (L_0) for the simpler model that deletes those parameters. Denote the fitted model by M_1 and the simpler model for which those parameters equal zero by M_0.

The goodness-of-fit statistic G^2 for testing the fit of a logistic regression model M is the special case of the likelihood-ratio statistic in which $M_0 = M$ and M_1 is the most complex model possible. That complex model has a separate parameter for each logit, and provides a perfect fit to the sample logits. It is called the *saturated model*. In testing whether M fits, we test whether *all* parameters that are in the saturated model but not in M equal zero. Denote this statistic for testing the fit of M by $G^2(M)$. In GLM terminology, it is called the *deviance* of the model (Section 4.5.3). Let L_S denote the maximized log likelihood for the saturated model. Then, for instance, the deviances for models M_0 and M_1 are $G^2(M_0) = -2(L_0 - L_S)$ and $G^2(M_1) = -2(L_1 - L_S)$.

Denote the likelihood-ratio statistic for testing M_0, given that M_1 holds, by $G^2(M_0 \mid M_1)$. This statistic for comparing these models equals

$$G^2(M_0 \mid M_1) = -2(L_0 - L_1) = -2(L_0 - L_S) - [-2(L_1 - L_S)] = G^2(M_0) - G^2(M_1),$$

the difference in G^2 goodness-of-fit statistics for the two models. That is, the likelihood-ratio statistic for comparing two models is simply the difference in the deviances of those models. This statistic is large when M_0 fits poorly compared to M_1. It is a large-sample chi-squared statistic, with df equal to the difference between the residual df values for the two models.

We illustrate this comparison for two models fitted to the grouped crab data. Denote the logistic regression model (5.3.1) with width as the sole predictor by M_1 and the simpler model having only an intercept parameter as M_0. That simpler model posits independence of width and having a satellite, and the G^2 goodness-of-fit

statistic for testing it is simply the G^2 statistic (2.4.3) for testing independence in a two-way contingency table. For the observed counts in the 8×2 Table 5.2, it equals $G^2(M_0) = 34.0$, based on $df = 7$. Since the fit of the model with width as a predictor has $G^2(M_1) = 6.0$ with $df = 6$, the comparison statistic for the two models is $G^2(M_0 \mid M_1) = G^2(M_0) - G^2(M_1) = 34.0 - 6.0 = 28.0$, based on $df = 7 - 6 = 1$. In fact, this equals the likelihood-ratio statistic $-2(L_0 - L_1)$ for testing that $\beta = 0$ in the logistic regression model fitted to the grouped data of Table 5.2.

5.3.3 Residuals for Logit Models

Goodness-of-fit statistics such as G^2 and X^2 are summary indicators of the overall quality of fit. Additional diagnostic analyses are necessary to describe the nature of any lack of fit. Residuals comparing observed and fitted counts are useful for this purpose.

Let y_i denote the number of "successes" for n_i trials at the ith setting of the explanatory variables. Let $\hat{\pi}_i$ denote the predicted probability of success for the model fit. Then $n_i\hat{\pi}_i$ is the fitted number of successes. For a GLM with binomial random component, the Pearson residual (4.4.1) for the fit at setting i is

$$e_i = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{[n_i\hat{\pi}_i(1 - \hat{\pi}_i)]}}. \tag{5.3.2}$$

Each residual divides the difference between an observed count and its fitted value by the estimated binomial standard deviation of the observed count.

The Pearson statistic for testing the model fit satisfies

$$X^2 = \sum e_i^2. \tag{5.3.3}$$

Each squared Pearson residual is a component of X^2. When the binomial index n_i is large, the Pearson residual e_i has an approximate normal distribution. When the model holds, it has an approximate expected value of zero but a smaller variance than a standard normal variate. If the number of model parameters is small compared to the number of sample logits, Pearson residuals are treated like standard normal deviates, with absolute values larger than 2 indicating possible lack of fit.

Table 5.3 shows Pearson residuals for two logistic regression models fitted to the grouped crab data—model (5.3.1) with width as the predictor and the model having only an intercept term. The latter model, which is (5.3.1) with $\beta = 0$, treats the response as independent of width. Some residuals for that model are large, and they show an increasing trend. This trend disappears for the model with a linear effect of width.

Graphical displays are also useful for showing lack of fit. One can compare observed and fitted proportions by plotting them against each other, or by plotting both of them against explanatory variables. For Table 5.2, Figure 5.3 plots both the observed proportions and the predicted probabilities of a satellite against width. The

Table 5.3 Residuals for Logistic Regression Models Fitted to Grouped Crab Data

Width	Number Cases	Number Yes	Fitted[a] Yes	Pearson[a] Residual	Fitted Yes	Pearson Residual	Adjusted Residual
< 23.25	14	5	8.98	−2.22	3.85	0.69	0.85
23.25–24.25	14	4	8.98	−2.78	5.50	−0.82	−0.93
24.25–25.25	28	17	17.96	−0.38	13.97	1.14	1.35
25.25–26.25	39	21	25.02	−1.34	24.21	−1.06	−1.24
26.25–27.25	22	15	14.12	0.39	15.80	−0.38	−0.42
27.25–28.25	24	20	15.40	1.96	19.16	0.43	0.49
28.25–29.25	18	15	11.55	1.70	15.46	−0.31	−0.36
> 29.25	14	14	8.98	2.80	13.05	1.01	1.14

[a]Independence model, other fitted values and residuals refer to model (5.3.1) with width predictor.

fit seems decent, which is not surprising since the formal goodness-of-fit tests showed no evidence of lack of fit.

We have noted that X^2 and G^2 are invalid when fitted values are very small. Similarly, residuals have limited meaning in that case. When explanatory variables are continuous, often $n_i = 1$ at many settings. Then, y_i can equal only 0 or 1, and

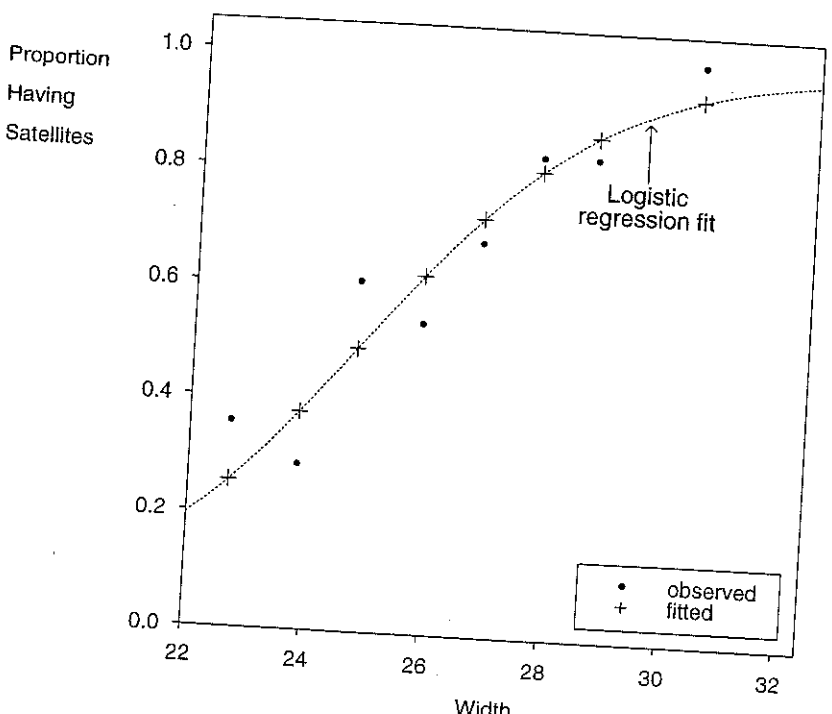

Figure 5.3 Observed and fitted proportions of satellites by width of female crab.

e_i can assume only two values. One must then be cautious about regarding either outcome as "extreme," and a single residual is usually uninformative.

5.3.4 Diagnostic Measures of Influence*

As in ordinary regression modeling, some observations may have much, perhaps too much, influence in determining parameter estimates. The fit could be quite different if they were deleted. An observation is more likely to have a large influence when it takes an extreme value on one or more of the explanatory variables. It may be informative to report the fit of the model after deleting one or two observations, if the fit with them seems misleading.

Several measures describe various aspects of influence. Many of them relate to the effect on certain characteristics of removing the observation from the data set. These measures are algebraically related to an observation's *leverage*, its element from the diagonal of the so-called *hat matrix*. (Roughly speaking, the hat matrix is a matrix that, when applied to the sample logits, yields the predicted logit values for the model.) The greater an observation's leverage, the greater its potential influence. Formulas for leverages and the diagnostic measures of influence are complex, so we do not reproduce them here. Most software for logistic regression produces these diagnostics. Influence measures for each observation include the following:

1. For each parameter in the model, the change in the parameter estimate when the observation is deleted. This change, divided by its standard error, is called *Dfbeta*.
2. A measure of the change in a joint confidence interval for the parameters produced by deleting the observation. This confidence interval displacement diagnostic is denoted by c.
3. The change in X^2 or G^2 goodness-of-fit statistics when the observation is deleted.

For each measure, the larger the value, the greater the observation's influence. We illustrate them using the logistic regression model with width as a predictor for the grouped crab data. Table 5.4 contains the *Dfbeta* measure for the coefficient of width, the confidence interval diagnostic c, the change in X^2, and the change in G^2. None of their values reveal any highly influential observations. By contrast, Table 5.4 also contains the changes in X^2 and G^2 for deleting observations from the fit of the independence model ((5.3.1) with $\beta = 0$). At the low and high ends of the width values, several of these changes are large. The severe influence of these values partly reflects the poor fit of the model that does not permit width to have an effect on the response.

One can also use leverage values to construct an adjustment to the Pearson residual e_i that is slightly larger in absolute value and *does* have an approximate standard normal distribution when the model holds. For observation i with leverage h_i, the

Table 5.4 Diagnostic Measures for Logistic Regression Models Fitted to Grouped Crab Data

Width	*Dfbeta*	*c*	Pearson Diff.	Likelihood-Ratio Diff.	Pearson[a] Diff.	Likelihood-Ratio[a] Diff.
< 23.25	−0.54	0.38	0.73	0.70	5.36	5.09
23.25–24.25	0.37	0.25	0.87	0.89	8.39	8.00
24.25–25.25	−0.43	0.71	1.82	1.83	0.17	0.17
25.25–26.25	−0.02	0.58	1.55	1.52	2.33	2.27
26.25–27.25	−0.09	0.04	0.17	0.17	0.18	0.18
27.25–28.25	0.21	0.08	0.24	0.25	4.45	4.95
28.25–29.25	−0.17	0.04	0.13	0.13	3.21	3.58
> 29.25	0.55	0.34	1.29	2.24	8.51	13.11

[a]Independence model, other values refer to model (5.3.1) with width predictor.

adjusted residual has form

$$\frac{e_i}{\sqrt{1-h_i}} = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{[n_i\hat{\pi}_i(1-\hat{\pi}_i)(1-h_i)]}}.$$

These serve the same purpose as the adjusted residuals defined in Section 2.4.5 for detecting patterns of dependence in two-way contingency tables and in Section 4.4.3 for describing lack of fit in Poisson regression models.

Table 5.3 contains adjusted residuals for the logistic regression model with width predictor fitted to the grouped crab data. Though slightly larger than the Pearson residuals, they show a similar pattern and do not suggest any lack of fit.

5.4 LOGIT MODELS FOR QUALITATIVE PREDICTORS

Logistic regression, like ordinary regression, extends to models incorporating multiple explanatory variables. Moreover, some or all of those explanatory variables can be qualitative, rather than quantitative. This section shows the use of dummy variables for including qualitative predictors, often called *factors*, and Section 5.5 presents the general form of multiple logistic regression models.

5.4.1 Dummy Variables in Logit Models

Suppose that a binary response Y has two binary predictors, X and Z. Denote the two levels for each variable by (0, 1). For the $2 \times 2 \times 2$ contingency table, the model for the probability π that $Y = 1$,

$$\text{logit}(\pi) = \alpha + \beta_1 x + \beta_2 z, \tag{5.4.1}$$

has separate main effects for the two predictors. It assumes an absence of interaction, the effect of one factor being the same at each level of the other factor.

The variables x and z in this model are *dummy variables* that indicate categories for the predictors. At a fixed level z of Z, the effect on the logit of changing from $x = 0$ to $x = 1$ is

$$= [\alpha + \beta_1(1) + \beta_2 z] - [\alpha + \beta_1(0) + \beta_2 z] = \beta_1.$$

This difference between two logits equals the difference of log odds, which equals the log of the odds ratio between X and Y, at a fixed level of Z. Thus, $\exp(\beta_1)$ describes the conditional odds ratio between X and Y. Controlling for Z, the odds of "success" at $x = 1$ equal $\exp(\beta_1)$ times the odds of success at $x = 0$. This conditional odds ratio is the same at each level z of Z. The lack of an interaction term in this model implies a common value of the odds ratio for the partial tables at the two levels of Z. The model satisfies homogeneous association (Sections 3.1.5, 3.2.3, 3.2.4).

Conditional independence exists between X and Y, controlling for Z, if $\beta_1 = 0$, in which case the common odds ratio equals 1. The simpler model

$$\text{logit}(\pi) = \alpha + \beta_2 z \tag{5.4.2}$$

then applies to the three-way table. One can test whether $\beta_1 = 0$ using a Wald statistic or a likelihood-ratio statistic comparing the two models.

5.4.2 AZT and AIDS Example

We illustrate models with qualitative predictors using Table 5.5, based on a study described in the *New York Times* (Feb. 15, 1991) on the effects of AZT in slowing the development of AIDS symptoms. In the study, 338 veterans whose immune systems were beginning to falter after infection with the AIDS virus were randomly assigned either to receive AZT immediately or to wait until their T cells showed severe immune weakness. Table 5.5 is a $2 \times 2 \times 2$ cross classification of the veterans' race, whether they received AZT immediately, and whether they developed AIDS symptoms during the three-year study.

In model (5.4.1), we identify X with AZT treatment ($x = 1$ for those who took AZT immediately and $x = 0$ otherwise) and Z with race ($z = 1$ for whites and $z = 0$ for blacks), for predicting the probability that AIDS symptoms developed. The ML estimate of the effect of AZT is $\beta_1 = -0.720$ ($ASE = 0.279$). The estimated

Table 5.5 Development of AIDS Symptoms by AZT Use and Race

		Symptoms	
Race	AZT Use	Yes	No
White	Yes	14	93
	No	32	81
Black	Yes	11	52
	No	12	43

odds ratio between immediate AZT use and development of AIDS symptoms equals $\exp(-0.720) = 0.49$. For each race, the estimated odds of developing symptoms are half as high for those who took AZT immediately.

The hypothesis of conditional independence of AZT treatment and the development of AIDS symptoms, controlling for race, is $H_0 : \beta_1 = 0$. The likelihood-ratio statistic $-2(L_0 - L_1)$ based on comparing models (5.4.2) and (5.4.1) equals 6.9, based on $df = 1$, showing evidence of association ($P = .01$). The Wald statistic $(\hat{\beta}_1/ASE)^2 = (-0.720/0.279)^2 = 6.6$ provides similar results ($P = .01$).

We next analyze the goodness of fit of model (5.4.1). For its fit, white veterans with immediate AZT use had predicted probability .150 of developing AIDS symptoms during the study. Since 107 white veterans took AZT, the fitted number developing symptoms is $107(.150) = 16.0$, and the fitted number not developing symptoms is $107(.850) = 91.0$. Similarly, one can obtain fitted values for all eight cells in Table 5.5. Substituting these and the cell counts into the usual goodness-of-fit statistics, we obtain $G^2 = 1.4$ and $X^2 = 1.4$. The model has four sample logits, one for each binomial response distribution at the four combinations of AZT use and race. The model has three parameters, so the residual $df = 4 - 3 = 1$. The small values for G^2 and X^2 suggest that the model fits decently ($P > .2$). Further analysis suggests that an even simpler model may be adequate, since the effect of race is not significant.

5.4.3 ANOVA-Type Representation of Factors

A factor having two levels requires only a single dummy variable. A factor having I levels requires $I - 1$ dummy variables, as shown in Section 5.5.1.

An alternative representation of factors in logistic regression models resembles the way ANOVA models ordinarily express them. The model formula

$$\text{logit}(\pi) = \alpha + \beta_i^X + \beta_k^Z \tag{5.4.3}$$

represents the effects of X through parameters $\{\beta_i^X\}$ and the effects of Z through parameters $\{\beta_k^Z\}$. (The X and Z superscripts are simply labels, and do not represent powers.) Model form (5.4.3) applies for any number of levels for X and Z. Each factor has as many parameters as it has levels, but one is redundant. For instance, if X has I levels, it has $I - 1$ nonredundant parameters; β_i^X denotes the effect on the logit of being classified in level i of X. Conditional independence between X and Y, given Z, corresponds to $\beta_1^X = \beta_2^X = \cdots = \beta_I^X$.

One can account for redundancies in parameters in (5.4.3) by setting the parameter for the last category equal to zero. When X and Z have two categories, as in Table 5.5, the parameterization in model (5.4.3) then corresponds to that in model (5.4.1) with $\beta_1^X = \beta_1$ and $\beta_2^X = 0$, and with $\beta_1^Z = \beta_2$ and $\beta_2^Z = 0$. For model (5.4.3) fitted to Table 5.5, Table 5.6 shows parameter estimates for three ways of defining parameters: (1) the approach just described that sets the last parameter (e.g., β_2^X) equal to 0, (2) an analogous approach for which the first parameter equals 0, and (3) an approach whereby each factor's parameters sum to zero. For the second approach, model (5.4.3) corresponds to model (5.4.1) with the dummy variable $x = 0$ in

Table 5.6 Parameter Estimates for Logit Model Fitted to Table 5.5

	Definition of Parameters		
Parameter	Last = zero	First = zero	Sum = zero
Intercept	−1.074	−1.738	−1.406
AZT—yes	−0.720	0.000	−0.360
AZT—no	0.000	0.720	0.360
Race—W	0.055	0.000	0.028
Race—B	0.000	−0.055	−0.028

category 1 and $x = 1$ in category 2 of AZT use. For the third approach, when a factor has two levels, one estimate is the negative of the other (e.g., $\hat{\beta}_1^X = -\hat{\beta}_2^X$). This results from "effect coding" for a dummy variable, such as $x = 1$ in category 1 and $x = -1$ in category 2.

For any of the three coding schemes, the differences $\beta_1^X - \beta_2^X$ and $\beta_1^Z - \beta_2^Z$ are identical and represent the conditional log odds ratios of X and Z with the response, given the other variable. For instance, $\exp(\hat{\beta}_1^X - \hat{\beta}_2^X) = \exp(-0.720) = 0.49$ refers to the estimated common odds ratio between immediate AZT use and development of symptoms, for each race. The estimate of a parameter for a single category of a factor is irrelevant; different ways of handling parameter redundancies result in different values for that estimate. An estimate makes sense only by comparison with one for another category. Exponentiating a difference between estimates for two categories determines the odds ratio relating to the effect of classification in one category rather than the other.

Similarly, different parameter coding schemes yield the same estimated probabilities. The sum of the intercept estimate and the estimates for given factor levels is identical for each scheme. For instance, from Table 5.6, the intercept estimate plus the estimate for immediate AZT use plus the estimate for being white is −1.738 for each scheme, leading to a predicted probability that white veterans with immediate AZT use develop AIDS symptoms equal to $\exp(-1.738)/[1 + \exp(-1.738)] = .15$.

5.4.4 Logit Models for $2 \times 2 \times K$ Contingency Tables

An important special case of logit models with qualitative predictors occurs when X is a binary classification of two groups, and Z is a control variable with K levels. For instance, X might refer to two experimental treatments and Z might refer to several locations for conducting the experiment. Model (5.4.3) then refers to a $2 \times 2 \times K$ contingency table that cross classifies X, Y, and Z. In this model, conditional independence exists between X and Y, controlling for Z, if $\beta_1^X = \beta_2^X$, in which case the common X-Y odds ratio $\exp(\beta_1^X - \beta_2^X)$ for the K partial tables equals 1. One can then absorb the common value of β_i^X into the α term, yielding the simpler model

$$\text{logit}(\pi) = \alpha + \beta_k^Z \tag{5.4.4}$$

for the three-way table. When $K = 2$, this is equivalent to model (5.4.2).

Given that model (5.4.3) holds, one can test conditional independence of X and Y by the likelihood-ratio statistic $-2(L_0 - L_1)$, comparing that model to the simpler model (5.4.4) not having the X main effect, as illustrated above in Section 5.4.2. Section 3.2.1 presented the Cochran–Mantel–Haenszel (*CMH*) test of this same hypothesis for $2 \times 2 \times K$ tables. That test performs well when the association between X and Y is similar in each partial table. In fact, the *CMH* statistic is the efficient score statistic (Section 4.5.2) for testing X-Y conditional independence in model (5.4.3), which assumes a common odds ratio for the partial tables. The likelihood-ratio test is an alternative to the *CMH* procedure for testing X-Y conditional independence in $2 \times 2 \times K$ tables. For model (5.4.3), the ML estimate $\exp(\hat{\beta}_1^X - \hat{\beta}_2^X)$ of the common X-Y odds ratio for the K partial tables is an alternative to the Mantel–Haenszel estimate (3.2.2).

For Table 5.5, we noted that the likelihood-ratio test of conditional independence of immediate AZT use and AIDS symptom development has test statistic equal to 6.9, with $df = 1$, and the ML estimate of the conditional odds ratio equals $\exp(\hat{\beta}_1^X - \hat{\beta}_2^X) = 0.49$. The *CMH* statistic (3.2.1) equals 6.8 with $df = 1$, and the Mantel–Haenszel estimate (3.2.2) of a common odds ratio equals 0.49. Similarity of results among likelihood-ratio, Wald, and *CMH* (efficient score) tests usually happens when the sample size is large relative to the number of strata.

For model (5.4.3) with binary X, the odds ratio between X and Y is the same at each level of Z. In testing goodness of fit for this model, we are testing that this structure holds. That is, the goodness-of-fit tests also provide tests of the hypothesis of homogeneous odds ratios between X and Y at the K levels of Z. These large-sample tests have $df = K - 1$ and are alternatives to the Breslow–Day test presented in Section 3.2.4. For Table 5.5, for instance, the goodness-of-fit statistics equal $G^2 = 1.38$ and $X^2 = 1.39$ and the Breslow–Day statistic equals 1.39, all based on $df = 1$. They all indicate that homogeneity is plausible.

5.5 MULTIPLE LOGISTIC REGRESSION

The logistic regression model and other GLMs, like ordinary regression models for normal data, generalize to allow for several explanatory variables. The predictors can be quantitative, qualitative, or of both types.

Denote a set of k predictors for a binary response Y by $X_1, X_2, \ldots, X_k$. Model (5.1.1) for the logit of the probability π that $Y = 1$ generalizes to

$$\text{logit}(\pi) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k. \tag{5.5.1}$$

The parameter β_i refers to the effect of X_i on the log odds that $Y = 1$, controlling the other Xs. For instance, $\exp(\beta_i)$ is the multiplicative effect on the odds of a 1-unit increase in X_i, at fixed levels of the other Xs.

5.5.1 Horseshoe Crab Example Using Color and Width Predictors

We continue our analysis of the horseshoe crab data of Table 4.2 (Section 4.3.2) by including both the female crab's width and color as predictors. Color has five

categories: light, medium light, medium, medium dark, dark. Color is a surrogate for age, older crabs tending to be darker. The sample contained no light crabs, so our models use only the other four categories.

We first treat color in a qualitative manner by using three dummy variables to represent the four categories. The model is

$$\text{logit}(\pi) = \alpha + \beta_1 c_1 + \beta_2 c_2 + \beta_3 c_3 + \beta_4 x, \tag{5.5.2}$$

where x denotes width and

$c_1 = 1$ for medium light color, and 0 otherwise,
$c_2 = 1$ for medium color, and 0 otherwise,
$c_3 = 1$ for medium dark color, and 0 otherwise.

The crab color is dark (category 4) when $c_1 = c_2 = c_3 = 0$. The ML estimates of the parameters are

$$\begin{aligned}
\text{Intercept: } \hat{\alpha} &= -12.715, \; ASE = 2.762 \\
c_1\text{: } \hat{\beta}_1 &= 1.330, \; ASE = 0.852 \\
c_2\text{: } \hat{\beta}_2 &= 1.402, \; ASE = 0.548 \\
c_3\text{: } \hat{\beta}_3 &= 1.106, \; ASE = 0.592 \\
\text{width: } \hat{\beta}_4 &= 0.468, \; ASE = 0.106.
\end{aligned}$$

For instance, for dark crabs, $c_1 = c_2 = c_3 = 0$, and the prediction equation is $\text{logit}(\hat{\pi}) = -12.715 + 0.468x$; by contrast, for medium-light crabs, $c_1 = 1$, and $\text{logit}(\hat{\pi}) = (-12.715 + 1.330) + 0.468x = -11.385 + 0.468x$.

The model assumes a lack of interaction between color and width in their effects on the response. Width has the same effect (coefficient 0.468) for all colors, so the shapes of the curves relating width to $\pi = P(Y = 1)$ are identical. For each color, a 1-cm increase in width has a multiplicative effect of $\exp(0.468) = 1.60$ on the odds that $Y = 1$. Figure 5.4 displays the fitted model. Any one curve is simply any other curve shifted to the right or to the left. The parallelism of curves in the horizontal dimension implies that two curves never cross. At all width values, for instance, color 4 (dark) has a lower predicted probability of a satellite than the other colors.

The positive effect of width on the odds of having satellites is similar to the effect seen in Section 5.1.2 for the simpler model excluding color. One can calculate predicted probabilities of a satellite using prediction equations for the probabilities, extending the approach of Section 5.1.2. To illustrate, for a medium-light crab of average width (26.3 cm), the predicted probability is

$$\frac{\exp[-11.385 + 0.468(26.3)]}{1 + \exp[-11.385 + 0.468(26.3)]} = .715.$$

By comparison, a dark crab of average width has predicted probability $\exp[-12.715 + 0.468(26.3)]/\{1 + \exp[-12.715 + 0.468(26.3)]\} = .399$.

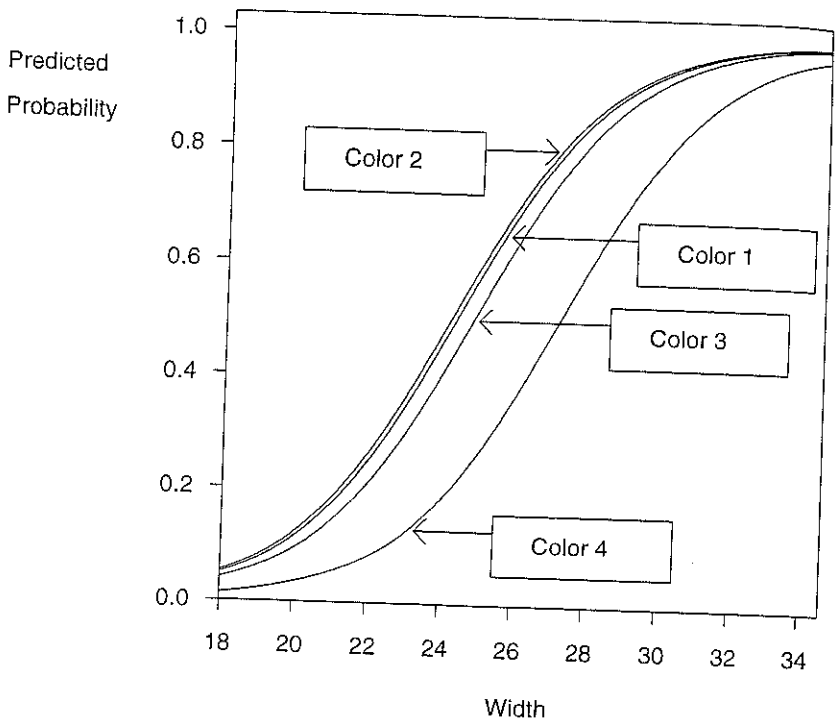

Figure 5.4 Logistic regression model using width and color predictors.

The exponentiated difference between two color parameter estimates is an odds ratio comparing those colors. For instance, the difference in color parameter estimates between medium-light crabs and dark crabs equals 1.330; at any given width, the estimated odds that a medium-light crab has a satellite are $\exp(1.330) = 3.8$ times the estimated odds for a dark crab. Using the probabilities just calculated at width 26.3, the odds equal $.715/.285 = 2.51$ for a medium-light crab, and $.399/.601 = 0.66$ for a dark crab, for which $2.51/0.66 = 3.8$. The color estimates indicate that, in this sample, dark crabs are less likely than crabs of other colors to have satellites.

5.5.2 Model Comparison

One can use the likelihood-ratio method to test hypotheses about parameters in multiple logistic regression models. For instance, to test whether color makes a significant contribution to model (5.5.2), we test $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$. This hypothesis states that, controlling for width, the probability of a satellite is independent of color. We compare the maximized log-likelihood L_1 for the full model (5.5.2) to the maximized log-likelihood L_0 for the simpler model in which those parameters equal 0, using test statistic $-2(L_0 - L_1) = 7.0$. The chi-squared $df = 3$, the difference between the

numbers of parameters in the two models. The P-value of .07 provides slight evidence of a color effect.

More generally, one can compare maximized log-likelihoods for any pair of models such that one is a special case of the other. One such comparison checks whether the model requires interaction terms. The test analyzes whether a better fitting model results from adding the interaction of color and width to model (5.5.2). This more complex model allows a separate width effect for each color. It has three additional terms, the cross-products of width with the color dummy variables. Fitting this model is equivalent to fitting the logistic regression model with width as the predictor separately for the crabs of each of the four colors. Each color then has a different-shaped curve relating width to the probability of a satellite, so a comparison of two colors varies according to the level of width. The likelihood-ratio statistic comparing the models with and without the interaction terms equals 4.4, based on $df = 3$. The evidence of interaction is not strong ($P = .22$).

The reduced model (5.5.2) has the advantage of simpler interpretations. In fact, this model fits adequately according to formal goodness-of-fit tests. For instance, the Hosmer–Lemeshow test with ten groups of predicted probabilities has a test statistic equal to 3.7, based on $df = 8$.

5.5.3 Quantitative Treatment of Ordinal Predictor

Color has a natural ordering of categories, from lightest to darkest. One can construct a simpler model yet by treating this ordinal predictor in a quantitative manner. Color may have a linear effect, for a set of monotone scores assigned to its categories.

To illustrate, we assign scores $c = \{1, 2, 3, 4\}$ to the color categories and fit the model

$$\text{logit}(\pi) = \alpha + \beta_1 c + \beta_2 x, \tag{5.5.3}$$

The prediction equation is

$$\text{logit}(\hat{\pi}) = -10.071 - 0.509c + 0.458x.$$

The color and width estimates have *ASE* values of 0.224 and 0.104, showing strong evidence of an effect for each. At a given width, for every one-category increase in color darkness, the estimated odds of a satellite multiply by $\exp(-0.509) = 0.60$. For instance, the estimated odds of a satellite for medium colored crabs are 60% of those for medium-light crabs.

A likelihood-ratio test compares the fit of this model to the more complex model (5.5.2) that has a separate parameter for each color. The test statistic equals $-2(L_0 - L_1) = 1.7$, based on $df = 2$. This statistic tests that the simpler model (5.5.3) holds, given that model (5.5.2) is adequate. It tests that the color parameters in (5.5.2), when plotted against the color scores, follow a linear trend. The simplification seems permissible ($P = .44$).

The estimates of the color parameters in the model (5.5.2) that treats color as qualitative are $(1.33, 1.40, 1.11, 0)$, the 0 value for the dark category reflecting the lack of a dummy variable for that category. Though these values do not depart significantly from a linear trend, the first three are quite similar compared to the last one. This suggests that another potential color scoring for model (5.5.3) is $\{1, 1, 1, 0\}$; that is, score = 0 for dark-colored crabs, and score = 1 otherwise. The likelihood-ratio statistic comparing model (5.5.3) with these binary scores to model (5.5.2) equals 0.5, based on $df = 2$, showing that this simpler model is also adequate ($P = .78$). This model has a width estimate of 0.478 ($ASE = 0.104$) and a color estimate of 1.300 ($ASE = 0.525$). At a given width, the estimated odds that a lighter-colored crab has a satellite are $\exp(1.300) = 3.7$ times the estimated odds for a dark crab.

In summary, the qualitative-color model, the ordinal model with color scores $\{1, 2, 3, 4\}$, and the model with binary color scores $\{1, 1, 1, 0\}$ all suggest that dark crabs are least likely to have satellites. It would require a much larger sample size to determine which of the two color scorings is more appropriate. It is advantageous to treat ordinal predictors in a quantitative manner when such models fit well. The advantages include that the model is simpler and easier to interpret, and that tests of the effect of the ordinal predictor are generally more powerful when it has a single parameter rather than several parameters.

5.5.4 Model Selection with Several Predictors

The horseshoe crab data set in Table 4.2 has four predictors: color (four categories), spine condition (three categories), weight, and width of the carapace shell. We next fit a logistic regression model using all these predictors.

Several model selection procedures exist, no one of which is "best." Cautions that apply to ordinary regression modeling of normal data hold for any generalized linear model. For instance, a model with several predictors has the potential for *multicollinearity*: strong correlations among predictors making it seem that no one variable is important when all the others are in the model. A variable may seem to have little effect simply because it "overlaps" considerably with other predictors in the model.

To illustrate, suppose we started by fitting a model containing main effects for the four predictors, treating color and spine condition as qualitative (factors). A likelihood-ratio test that the probability of a satellite is jointly independent of these four predictors simultaneously tests that all their parameters equal zero. The likelihood-ratio statistic based on comparing the main-effects model to the null model that has only an intercept term equals $-2(L_0 - L_1) = 40.6$ with $df = 7$. This statistic has P-value $< .0001$, extremely strong evidence that at least one predictor has an effect. Table 5.7 shows the parameter estimates and their *ASE* values. Even though the overall test is highly significant, the results in this table are not encouraging. The estimates for the quantitative predictors, weight and width, are only slightly larger than their *ASE* values. The estimates for the qualitative predictors compare each level to the final category as a baseline; that is, they set up dummy variables for the first three colors and for the first two spine conditions. For color, the largest difference

Table 5.7 Parameter Estimates for Main Effects Model with Horseshoe Crab Data

Parameter	Estimate	*ASE*
Intercept	−9.273	3.838
Color(1)	1.609	0.936
Color(2)	1.506	0.567
Color(3)	1.120	0.593
Spine(1)	−0.400	0.503
Spine(2)	−0.496	0.629
Weight	0.826	0.704
Width	0.263	0.195

between estimates for two levels is between the first and fourth, which is less than two standard errors; for spine condition, the largest difference between estimates for two levels is between the second and third, which is less than a standard error.

The very small P-value for the overall test, yet the lack of significance shown in Table 5.7, is a warning signal of potential multicollinearity. Section 5.2 showed strong evidence of a width effect on the presence of satellites, yet, controlling for weight, color, and spine condition, little evidence exists of a partial width effect. Graphical exploration reveals, however, a strong linear component for the relationship between width and weight. The sample correlation between them equals 0.887. It does not make much sense to analyze an effect of width while controlling for weight, since weight naturally increases as width does.

For practical purposes, width and weight serve equally well as predictors, but it is redundant to use them both. In further analysis, we use width alone together with color and spine condition as predictors. Denote these predictors by W, C, and S. For simplicity, we symbolize various models by the highest order terms in the model, regarding C and S in the models as factors. For instance, $C + S + W$ denotes a model with main effects of the sort shown in Table 5.7, whereas $C + S * W$ denotes a model that also has an interaction between S and W.

5.5.5 Backward Elimination of Predictors

Table 5.8 summarizes results of fitting and comparing several logistic regression models. The *deviance* of a model is the G^2 test of goodness of fit based on comparing the model to the saturated model (Sections 4.5.3 and 5.3.2); the difference of deviances between two models is the likelihood-ratio statistic $-2(L_0 - L_1)$ for comparing them. To select a model, we use a *backward elimination procedure*, starting with a complex model and successively taking out terms. At each stage, we eliminate the term in the model that has the largest P-value when we test that its parameters equal zero. We test only the highest-order terms for each variable. It is inappropriate, for instance, to remove a main effect term if the model contains higher-order interactions involving that term.

Table 5.8 Results of Fitting Several Logistic Regression Models to Horseshoe Crab Data

Model	Predictors	Deviance	*DF*	Models Compared	Difference	$(Y, \hat{\pi})$ Correlation
(1)	$C*S*W$	170.44	152	—	—	0.526
(2)	$C*S+C*W+S*W$	173.68	155	(2)-(1)	3.2 ($df=3$)	
(3a)	$C*S+S*W$	177.34	158	(3a)-(2)	3.7 ($df=3$)	
(3b)	$C*W+S*W$	181.56	161	(3b)-(2)	7.9 ($df=6$)	
(3c)	$C*S+C*W$	173.69	157	(3c)-(2)	0.0 ($df=2$)	
(4a)	$S+C*W$	181.64	163	(4a)-(3c)	8.0 ($df=6$)	
(4b)	$W+C*S$	177.61	160	(4b)-(3c)	3.9 ($df=3$)	
(5)	$C+S+W$	186.61	166	(5)-(4b)	9.0 ($df=6$)	
(6a)	$C+S$	208.83	167	(6a)-(5)	22.2 ($df=1$)	
(6b)	$S+W$	194.42	169	(6b)-(5)	7.8 ($df=3$)	
(6c)	$C+W$	187.46	168	(6c)-(5)	0.8 ($df=2$)	0.452
(7a)	C	212.06	169	(7a)-(6c)	24.5 ($df=1$)	0.285
(7b)	W	194.45	171	(7b)-(6c)	7.0 ($df=3$)	0.402
(8)	C=dark + W	187.96	170	(8)-(6c)	0.5 ($df=2$)	0.447
(9)	None	225.76	172	(9)-(8)	37.8 ($df=2$)	0.000

Note: C = color, S = spine condition, W = width

We begin with the most complex model, symbolized by $C*S*W$, listed as model (1) in Table 5.8. This model predicts the logit of the probability of a satellite using main effects for each term as well as the three two-factor interactions and the three-factor interaction. Removing the three-factor interaction term yields the simpler model $C*S+C*W+S*W$ containing two-factor interactions and main effects, model (2) in Table 5.8. To compare the fits, we test the hypothesis that the simpler model holds against the alternative that the more complex one holds. The likelihood-ratio statistic comparing the two models equals the difference in deviances, or 3.2 with $df=3$. This does not suggest that the three-factor term is needed ($P=.36$), thank goodness, so we continue the simplification process.

The next stage considers the three models that remove a two-factor interaction. Of these models, $C*S+C*W$ gives essentially the same fit as the more complex model, so we drop the $S*W$ interaction from the model. Next, we consider dropping one of the other two-factor interactions. The model $S+C*W$ (i.e., dropping the $C*S$ interaction but maintaining the S main effect) has an increased deviance of 8.0 on $df=6$ ($P=.24$); the model $W+C*S$, dropping the $C*W$ interaction, has an increased deviance of 3.9 on $df=3$ ($P=.27$). Neither increase is important, suggesting that we can drop either one of them and proceed. In either case, dropping next the remaining interaction also seems permissible. For instance, dropping the $C*S$ interaction from model $W+C*S$, leaving the main-effects model $C+S+W$, increases the deviance by 9.0 on $df=6$ ($P=.17$).

The working model now has the main effects alone. The next stage considers dropping one of them. Table 5.8 shows little consequence from removing S. Both remaining variables (C and W) then have nonnegligible effects. For instance, removing C increases the deviance (comparing models (7b) and (6c)) by 7.0 on $df=3$ ($P=.07$). The analysis in Section 5.5.3 revealed a noticeable difference between dark crabs (category 4) and the others. The simpler model that has a single dummy variable for color, equaling 0 for dark crabs and 1 otherwise, fits essentially as well (the deviance difference between models (8) and (6c) equals 0.5, with $df=2$). Further simplification results in large increases in deviance and is unjustified. As a final step, one can use the methods of Section 5.3 to check further the fit of this model.

Computerized variable selection procedures should be used with caution. When one considers a large number of terms for potential inclusion in a model, one or two of them that are not really important may look impressive simply due to chance. For instance, when all the true effects are weak, the largest sample effect may substantially overestimate its true effect. In addition, it often makes sense to include certain variables of special interest in a model and report their estimated effects even if they are not statistically significant at some level.

5.5.6 A Correlation Summary of Predictive Power*

It can be informative to compare various GLMs fitted to a data set in terms of their predictive power. The correlation R between the observed responses $\{Y_i\}$ and the model's fitted values $\{\hat{\mu}_i\}$ describes this. For least squares regression (i.e., a GLM with normal random component), R represents the *multiple correlation* between the response variable and the predictors; then R^2 describes the proportion of variation in Y that is explained by the predictors.

For logistic regression, R is the correlation between binary $Y=(0,1)$ observations on the response and the predicted probabilities $\hat{\pi}$. For such models, R is a crude index of predictive power, and it does not have the nice properties it has for normal GLMs. For instance, R is not guaranteed to be nondecreasing as the model gets more complex. Also, like any correlation measure, its value can depend strongly on the range of observed values of explanatory variables. Nevertheless, R is useful for comparing fits of different models to the same data set.

Table 5.8 shows correlations between observed responses and estimated probabilities for a few of the models fitted to the horseshoe crab data. Width alone has $R=.402$, and adding color to the model increases R to .452. The simpler model that uses color merely to indicate whether a crab is dark does essentially as well, with $R=.447$. This model has 85% as high a correlation as the complex model containing color, spine condition, width, and all their two-way and three-way interactions ($R=.526$). Little is lost and much is gained by using the simpler model to describe the data.

To compare effects of quantitative predictors having different units in multiple logistic regression models, it can be helpful to report standardized coefficients. One can do this by fitting the model to standardized versions of the predictors, where each measurement on a predictor is replaced by its z-score, (measurement − mean)/(standard deviation). Then, each reported regression coefficient represents the effect of a one standard deviation change in a predictor, controlling for the other variables. See Problem 5.30.

5.6 SAMPLE SIZE AND POWER FOR LOGISTIC REGRESSION*

The major aim of many studies is to determine whether a particular variable X has an effect on a binary response. The study design should take into account the sample size N needed to provide a good chance of detecting an effect of a given size.

5.6.1 Sample Size for Comparing Two Proportions

When X also is binary, the study often refers to comparing two groups (i.e., the two levels of X) on the binary response. To test the hypothesis that the "success" probabilities π_1 and π_2 for the groups are identical, one might conduct a chi-squared test for the 2×2 table that cross-classifies group by response, rejecting the null hypothesis if the P-value is smaller than some fixed level, α. To determine sample size, one needs to specify the probability β of failing to detect a difference between π_1 and π_2 of some fixed size considered to be practically important. For this size of effect, β is the probability of failing to "reject H_0" at the α level. Then, α is the probability of Type I error and β is the probability of Type II error. The *power* of the test equals $1 - \beta$.

A study using equal sample sizes for each group requires approximately

$$N_1 = N_2 = \frac{(z_{\alpha/2} + z_\beta)^2[\pi_1(1-\pi_1) + \pi_2(1-\pi_2)]}{(\pi_1 - \pi_2)^2}$$

in each group to achieve these error probabilities. This formula requires values for π_1 and π_2 and for the error probabilities α and β. To illustrate, for testing $H_0 : \pi_1 = \pi_2$ at the .05 level, suppose one wants a probability of Type II error equal to .10 if π_1 and π_2 are truly about .20 and .30. Then $\alpha = .05$, $\beta = .10$, the z-scores are $z_{.025} = 1.96$ and $z_{.10} = 1.28$, and the required sample sizes equal

$$N_1 = N_2 = \frac{(1.96 + 1.28)^2[(.2)(.8) + (.3)(.7)]}{(.2 - .3)^2} = 389.$$

This formula also provides the sample sizes needed for a comparable confidence interval for $\pi_1 - \pi_2$. Then, α is the error probability for the interval and β equals the probability that the confidence interval indicates a plausible lack of effect, in the sense that it contains the value zero. For instance, based on the above calculation with $\alpha = .05$ and $\beta = .10$, one needs about 400 subjects in each group for a 95% confidence interval to have only a .10 chance of containing 0 when actually $\pi_1 = .20$ and $\pi_2 = .30$.

This sample-size formula is approximate and may underestimate slightly the actual required values. It is adequate for most practical work, though, since one normally has only rough conjectures for π_1 and π_2. Fleiss (1981) provided more precise formulas.

The null hypothesis for the test comparing two proportions in a 2×2 table corresponds to one for a parameter in a logistic regression model having form

$$\text{logit}(\pi) = \alpha^* + \beta^* x. \qquad (5.6.1)$$

(We use the * notation so as not to confuse the parameters with the probabilities of Type I and Type II error.) If we set $x = 1$ for Group 1 and $x = 0$ for Group 2, the logit of the success probability equals $\alpha^* + \beta^*$ for Group 1 and α^* for Group 2. Identical probabilities corresponds to identical odds and identical logits, or $\beta^* = 0$. Thus, this example relates to sample size determination for a simple logistic regression model.

5.6.2 Sample Size with a Quantitative Predictor

For models of form (5.6.1) in which x is quantitative rather than a qualitative indicator, the sample size needed to achieve a certain power for testing $H_0 : \beta^* = 0$ depends on the distribution of the x values. One needs to guess the probability of success $\bar{\pi}$ at the mean of x. The size of the effect is the odds ratio θ comparing the probability of success at that point to the probability of success one standard deviation above the mean of x. Let $\lambda = \log(\theta)$. F. Y. Hsieh (*Statist. Medic., 8*: 795–802 (1989)) provided the approximate sample-size formula for a one-sided test,

$$N = \frac{[z_\alpha + z_\beta \exp(-\lambda^2/4)]^2(1 + 2\bar{\pi}\delta)}{\bar{\pi}\lambda^2},$$

where

$$\delta = \frac{1 + (1+\lambda^2)\exp(5\lambda^2/4)}{1 + \exp(-\lambda^2/4)}.$$

We illustrate for modeling the dependence of the probability of severe heart disease on x = cholesterol level for a middle-aged population. Consider the test of independence ($\beta^* = 0$) against the alternative $\beta^* > 0$ of increasing risk as cholesterol increases. Suppose previous studies have suggested that the probability of severe heart disease at an average level of cholesterol is about .08. Suppose we want the test to be sensitive to a 50% increase in this probability (i.e., to .12), for a standard deviation increase in cholesterol. The odds of severe heart disease at the mean cholesterol level equal $.08/.92 = 0.087$, and the odds one standard deviation above the mean equal $.12/.88 = 0.136$. The odds ratio equals $\theta = 0.136/0.087 = 1.57$, and $\lambda = \log(1.57) = 0.450$, $\lambda^2 = 0.202$. For a $\beta = .10$ chance of a Type II error in an $\alpha = .05$-level test, $z_\alpha = z_{.05} = 1.645$, $z_\beta = z_{.10} = 1.28$. Thus,

$$\delta = \frac{1 + (1.202)\exp(5 \times 0.202/4)}{1 + \exp(-0.202/4)} = \frac{2.548}{1.951} = 1.306,$$

and

$$N = \frac{[1.645 + 1.28\exp(-0.202/4)]^2(1 + 2(.08)(1.306))}{(.08)(0.202)} = 612.$$

The value N decreases as $\bar{\pi}$ gets closer to .5 and as $|\lambda|$ gets farther from the null value of 0. Its derivation assumes that the explanatory variable has approximately a normal distribution.

5.6.3 Sample Size in Multiple Logistic Regression

A multiple logistic regression model requires larger sample sizes to detect partial effects. Let R denote the multiple correlation between the predictor X of interest and the others in the model. One divides the above formula for N by $(1 - R^2)$. In that formula, $\bar{\pi}$ denotes the probability at the mean value of all the explanatory variables, and the odds ratio refers to the effect of the predictor of interest at the mean level of the others.

We illustrate by continuing the previous example. Consider a test for the effect of cholesterol on severe heart disease, while controlling for blood pressure level. If the correlation between cholesterol and blood pressure levels is .40, we need a sample size of roughly $612/[1 - (.40)^2] = 729$ for detecting the stated partial effect of cholesterol.

These formulas provide, at best, rough ballpark indications of sample size. In most applications, one has only a crude guess for indices such as $\bar{\pi}$ and R, and the explanatory variable may be far from normally distributed.

5.7 EXACT INFERENCE FOR LOGISTIC REGRESSION*

ML estimators of model parameters work best when the sample size is large compared to the number of parameters in the model. When the sample size is small, or when there are many parameters relative to the sample size, improved inference results using the method of *conditional maximum likelihood*.

5.7.1 Conditional Maximum Likelihood Inference

The conditional ML method bases inference for the primary parameters of interest on a *conditional likelihood* function that eliminates the other parameters. The technique uses a conditional probability distribution defined over data sets in which the values of certain "sufficient statistics" for the other parameters are fixed. This distribution is defined for potential samples that provide the same information about the other parameters that occurs in the observed sample. The distribution and the related conditional likelihood function depend only on the parameters of interest. The conditional ML method applies to any GLM that uses the *canonical link* (Section 4.1.3); for instance, the logit link with binomial data and the log link with Poisson data.

For binary data, conditional likelihood methods are especially useful when a logistic regression model contains a large number of "nuisance" parameters. For instance, models for matched-pairs data in matched case-control studies contain a separate parameter for each matched pair, so the number of parameters increases with the sample size. More accurate inference for the parameters of primary interest results from eliminating the nuisance parameters from the likelihood function. Section 9.2 discusses conditional ML methods for this type of model.

Conditional likelihood methods are also useful for small samples. One can perform *exact* inference for a parameter by using the conditional likelihood function that eliminates all the other parameters. Since that conditional likelihood does not involve unknown parameters, one can calculate probabilities such as P-values exactly rather than use crude approximations. This section introduces conditional ML methods for small-sample exact inference.

5.7.2 Exact Inference for Contingency Tables

Consider first the logistic regression model for a single explanatory variable,

$$\text{logit}[\pi(x)] = \alpha + \beta x.$$

When X can take only two values, the model applies to 2×2 tables for which the two rows are the levels of X and the two columns are the levels of Y. The usual sampling model treats the responses on Y at the separate x values as independent binomial variates. The row totals, which are the "numbers of trials" for those binomial variates, are naturally fixed.

For this model, the hypothesis of independence is $H_0 : \beta = 0$. The unknown parameter α refers to the relative number of response outcomes of the two types. One can eliminate α from the likelihood by conditioning also on the column marginal totals. These "sufficient statistics" for α represent the information that the data provide about α. Fixing both sets of marginal totals yields hypergeometric probabilities that do not depend on any unknown parameters. The resulting exact conditional test that $\beta = 0$ in 2×2 tables is simply Fisher's exact test (Section 2.6.1).

When X has I ordered levels, one can apply this logit model to the $I \times 2$ table by assigning scores to the rows. Again, conditioning on the column totals yields a conditional likelihood free of α, which one can use for exact inference about β. The exact test of $\beta = 0$ is an alternative to the large-sample trend test discussed in Section 2.5.5 (which is the "efficient score" test for the logit model) or the large-sample Wald or likelihood-ratio tests for β.

Next, suppose the model also contains a second explanatory factor, Z, having K levels. If Z is qualitative, a relevant model is

$$\text{logit}[\pi(x)] = \alpha + \beta x + \beta_k^Z.$$

When X is binary, this is the model presented in Section 5.4.4 for $2 \times 2 \times K$ contingency tables. The test of $H_0 : \beta = 0$ refers to the effect of X, controlling for Z. The exact test eliminates the other parameters by conditioning on the marginal totals in each of the partial tables. Section 3.3.1 discussed this exact test of conditional independence between X and Y, controlling for Z.

Computations for exact inference in logistic regression models are highly complex, but software is available. In fact, the manual for the software LogXact (Cytel Software) is a good source for discussion of this approach. When the sample size is small, the exact tests and related confidence intervals are more reliable than the ordinary large-sample ML inferences.

Exact methods are also useful when ordinary ML methods report an infinite parameter estimate. In such cases, exact methods can still yield P-values for tests

and confidence intervals in which one endpoint is finite. An infinite estimate would occur in logistic regression model (5.1.1) for the effect of width on the presence of a satellite, for instance, if $Y = 0$ whenever width is 26.0 or less, and $Y = 1$ whenever width is at least 26.0. Generally, suppose there is no overlap in the sets of explanatory variable values having the two response outcomes, in the sense that a line (or plane) can pass through the data such that $Y = 1$ on one side of the line and $Y = 0$ on the other side. There is then *perfect discrimination*, as one can predict the sample outcomes perfectly by knowing the predictor values (except possibly at a boundary point). In such cases, an ML parameter estimate for the logistic regression model is infinite. Software may not always detect this, and may instead report a very large estimate with an extremely large standard error.

5.7.3 Diarrhea Example

Table 5.9 refers to a sample of 2493 patients having stays in a hospital. The response is whether they suffered from an acute form of diarrhea during their stay. The three predictors are age (1 for over 50 years old, 0 for under 50), length of stay in hospital (1 for more than one week, 0 for less than one week), and exposure to an antibiotic called Cephalexin (1 for yes, 0 for no). We discuss estimation of the effect of Cephalexin, controlling for age and length of stay, using a model containing only main effect terms.

The sample size is large, yet relatively few cases occurred of acute diarrhea. Moreover, all subjects having exposure to Cephalexin were also diarrhea cases. Such "boundary" situations in which none or all responses fall in one category cause infinite ML estimates of some model parameters. An ML estimate of ∞ for the Cephalexin effect means that the likelihood function continually increases as the parameter estimate for Cephalexin increases indefinitely. Software for logistic regression may or may not indicate this. Some software reports "no convergence," whereas other packages simply report a large value for the estimated effect of Cephalexin with an extremely large standard error (e.g., one procedure reports a ML estimate of 27.0 based on $ASE = 70{,}065.8$).

Table 5.9 Example for Exact Conditional Logistic Regression

Cephalexin	Age	Length of Stay	Cases of Diarrhea	Sample Size
0	0	0	0	385
0	0	1	5	233
0	1	0	3	789
0	1	1	47	1081
1	1	1	5	5

Source: Based on study by Dr. E. Jaffe and Dr. V. Chang, Cornell Medical Center, reported in manual for *LogXact* (Cambridge, MA: Cytel Software, 1993, p. 7-5).

To study the effect of Cephalexin, we conduct an exact analysis, conditioning on sufficient statistics for the other predictors. Though the estimate of the parameter for the effect of Cephalexin is infinite, a 95% confidence interval for the true value is $(2.95, \infty)$. Exponentiating yields a confidence interval of $(19, \infty)$ for the odds ratio between exposure to Cephalexin and the response. This suggests that subjects taking Cephalexin have odds of developing diarrhea at least 19 times the odds for subjects not taking it, controlling for age and length of stay in the hospital. Assuming that the main-effects model is valid, Cephalexin appears to have a strong effect. Similarly, we obtain $P < .0001$ for testing that the parameter for Cephalexin equals zero.

Though the confidence interval is wide, it provides information not available through ordinary ML methods. Results must be qualified somewhat, because there were no Cephalexin cases at the first three combinations of levels of age and length of stay. In fact, the first three rows of Table 5.9 make no contribution to the analysis (Problem 5.39). The data actually provide evidence about the effect of Cephalexin only for older subjects having a long stay.

PROBLEMS

5.1. For the 23 space shuttle flights that occurred before the Challenger mission disaster in 1986, Table 5.10 shows the temperature (°F) at the time of the flight and whether at least one primary O-ring suffered thermal distress.

Table 5.10

Ft	Temp	TD	Ft	Temp	TD	Ft	Temp	TD
1	66	0	9	57	1	17	70	0
2	70	1	10	63	1	18	81	0
3	69	0	11	70	1	19	76	0
4	68	0	12	78	0	20	79	0
5	67	0	13	67	0	21	75	1
6	72	0	14	53	1	22	76	0
7	73	0	15	67	0	23	58	1
8	70	0	16	75	0			

Note: Ft = flight no., Temp = temperature, TD = thermal distress (1 = yes, 0 = no).
Source: Data based on Table 1 in S. R. Dalal, E. B. Fowlkes, and B. Hoadley, *J. Amer. Statist. Assoc*, *84*: 945–957 (1989). Reprinted with permission of the American Statistical Association.

a. Use logistic regression to model the effect of temperature on the probability of thermal distress. Interpret the model fit.

b. Calculate the predicted probability of thermal distress at 31°, the temperature at the time of the Challenger flight. At what temperature does the predicted probability equal .5? At that temperature, give a linear approximation for the change in the predicted probability per degree increase in temperature.