CHAPTER 5

Logistic Regression

In introducing generalized linear models for binary data in Chapter 4 we highlighted logistic regression. This is the most important model for categorical response data, being commonly used for a wide variety of applications.

Early uses of logistic regression were in biomedical studies, for instance, to model whether subjects have a particular condition such as lung cancer. The past 25 years have seen much use in social science research, for modeling opinions and behavior decisions, and in business applications. In *credit-scoring*, logistic regression is used to model the probability that a subject is credit worthy. For instance, the probability that a subject pays a bill on time may use predictors such as the size of the bill, annual income, occupation, mortgage and debt obligations, percentage of bills paid on time in the past, and other aspects of an applicant's credit history. Another area of increasing application is genetics, such as to estimate quantitative trait loci effects by modeling the probability that an offspring inherits an allele of one type instead of another type as a function of phenotypic values on various traits for that offspring.

In this chapter we study logistic regression more closely. Section 5.1 discusses parameter interpretation. In Section 5.2 we present inferential methods for those parameters. Sections 5.3 and 5.4 generalize the model to multiple predictors, which may be quantitative and/or qualitative. Finally, in Section 5.5 we apply GLM fitting methods to specify and solve likelihood equations for logistic regression.

5.1 INTERPRETING PARAMETERS IN LOGISTIC REGRESSION

For a binary response variable Y and an explanatory variable X, let $\pi(x) = P(Y = 1|X = x) = 1 - P(Y = 0|X = x)$. The logistic regression model is

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}. \tag{5.1}$$

Equivalently, the *logit* (log odds) has the linear relationship

$$\text{logit}[\pi(x)] = \log \frac{\pi(x)}{1 - \pi(x)} = \alpha + \beta x. \tag{5.2}$$

Categorical Data Analysis, Third Edition. Alan Agresti.

5.1.1 Interpreting β: Odds, Probabilities, and Linear Approximations

How can we interpret β in (5.2)? Its sign determines whether $\pi(x)$ is increasing or decreasing as x increases. The rate of climb or descent increases as $|\beta|$ increases; as $\beta \to 0$ the curve flattens to a horizontal straight line. When $\beta = 0$, Y is independent of X. For quantitative x with $\beta > 0$, the curve for $\pi(x)$ has the shape of the cdf of the logistic distribution (recall Section 4.2.5). Since the logistic density is symmetric, $\pi(x)$ approaches 1 at the same rate that it approaches 0.

Exponentiating both sides of (5.2) shows that the odds are an exponential function of x. This provides a basic interpretation for the magnitude of β: The odds multiply by e^{β} for every 1-unit increase in x. In other words, e^{β} is an odds ratio, the odds at $X = x + 1$ divided by the odds at $X = x$.

Many scientists are not familiar with odds or logits, so the interpretation of a multiplicative effect of e^{β} on the odds scale or an additive effect of β on the logit scale is not helpful to them. A simpler slope interpretation uses a linearization argument (Berkson 1951). Since it has a curved rather than a linear appearance, the logistic regression function (5.1) 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, shown in Figure 5.1, describes the instantaneous rate of change at that point. Calculating $\partial\pi(x)/\partial x$ with (5.1) yields a fairly complex function of the parameters and x, but it simplifies to the form $\beta\pi(x)[1 - \pi(x)]$.

For instance, the line tangent to the curve at x for which $\pi(x) = \frac{1}{2}$ has slope $\beta\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \beta/4$; when $\pi(x) = 0.9$ or 0.1, it has slope 0.09β. The slope approaches 0 as $\pi(x)$ approaches 1.0 or 0. The steepest slope occurs at x for which $\pi(x) = \frac{1}{2}$; that x value is $x = -\alpha/\beta$. [To check that $\pi(x) = \frac{1}{2}$ at this point, substitute $-\alpha/\beta$ for x in (5.1), or substitute $\pi(x) = \frac{1}{2}$

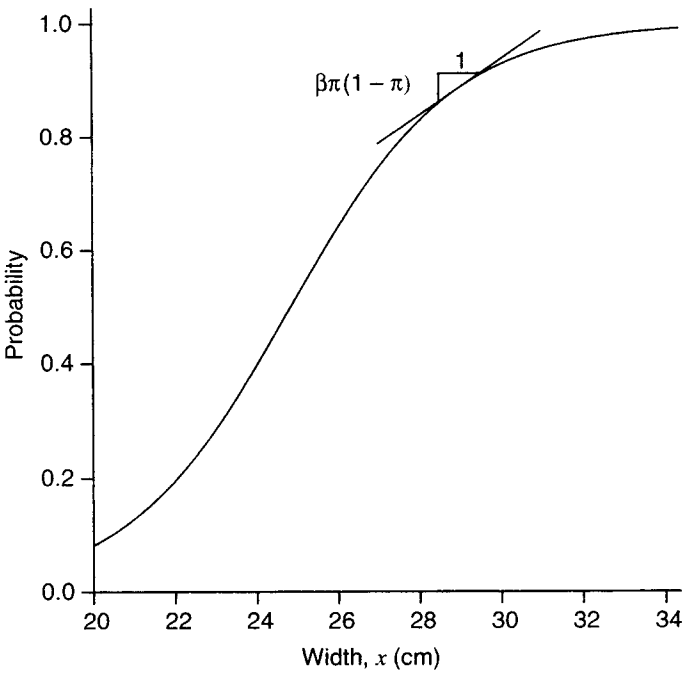

Figure 5.1 Linear approximation to logistic regression curve.

in (5.2) and solve for x.] This x value is sometimes called the *median effective level*. In toxicology studies it is called LD_{50} (LD = lethal dose), the dose with a 50% chance of a lethal result.

From this linear approximation, near x where $\pi(x) = \frac{1}{2}$, a change in x of $1/\beta$ corresponds to a change in $\pi(x)$ of roughly $(1/\beta)(\beta/4) = \frac{1}{4}$; that is, $1/\beta$ approximates the distance between x values where $\pi(x) = 0.50$ and where $\pi(x) = 0.25$ or 0.75 (in reality, 0.27 and 0.73). The linear approximation works better for smaller changes in x, however. Since the rate of change varies according to the value of x, a summary of them is the average of $\beta\pi(x_i)[1 - \pi(x_i)]$ for the subjects in the sample.

An alternative way to interpret the effect reports the values of $\pi(x)$ at certain x values, such as at the minimum and maximum values. To do this, we substitute the values for x into formula (5.1) for $\pi(x)$. It is more resistant to outliers on x to report the $\pi(x)$ values at the quartiles of x than at the extremes. The change in $\pi(x)$ over the middle half of x values, from the lower quartile to the upper quartile, is a useful summary of the effect. It can be compared with the corresponding change over the middle half of values of other quantitative predictors.

The intercept parameter α is not usually of particular interest. However, by centering the predictor about 0 [i.e., replacing x by $(x - \bar{x})$], α becomes the logit at $x = \bar{x}$, and thus $e^{\alpha}/(1 + e^{\alpha}) = \pi(\bar{x})$. As in ordinary regression, centering is also helpful in complex models containing quadratic or interaction terms to reduce correlations among model parameter estimates.

5.1.2 Looking at the Data

In practice, these interpretations use formula (5.1) with ML estimates substituted for parameters. Before fitting the model and making such interpretations, look at the data to check that the logistic regression model is appropriate. Since y takes only values 0 and 1, it is difficult to check this by an ordinary scatterplot of observed (x, y) values.

It can be helpful to plot sample proportions or logits against x. Let n_i denote the number of observations at setting i of x. Of them, let y_i denote the number of "1" outcomes, with $p_i = y_i/n_i$. Sample logit (also called *empirical logit*) i is $\log[p_i/(1 - p_i)] = \log[y_i/(n_i - y_i)]$. The scatterplot of sample logits should be roughly linear. The sample logit is not finite when $y_i = 0$ or n_i. An ad hoc adjustment adds a positive constant to the number of outcomes of the two types. The adjustment

$$\log \frac{y_i + \frac{1}{2}}{n_i - y_i + \frac{1}{2}}$$

is the least-biased estimator of this form for the true logit (see Note 5.2).

When x is continuous and all $n_i = 1$, or when x is essentially continuous and all n_i are small, this is unsatisfactory. We could group the data with nearby x values into categories before calculating sample proportions and sample logits. A better approach that does not require choosing arbitrary categories uses a smoothing mechanism to reveal trends. One such smoothing approach fits a *generalized additive model* (to be introduced in Section 7.4.9), which replaces the linear predictor of a GLM by a smooth function. A plot of this fit reveals whether severe discrepancies occur from the S-shaped trend predicted by logistic regression.

5.1.3 Example: Horseshoe Crab Mating Revisited

To illustrate logistic regression, we reanalyze the horseshoe crab data introduced in Section 4.3.2. The binary response is whether a female crab has any male crabs residing nearby (satellites). For crab i, let $y_i = 1$ if she has at least one satellite and $y_i = 0$ if she has none. Here, we use as a predictor the female crab's carapace width.

Figure 5.2 plots the data against x = width. The scatterplot consists of a set of points with $y_i = 1$ and a second set of points with $y_i = 0$. The numbered symbols indicate the number of observations at each point. It appears that $y_i = 1$ tends to occur relatively more often at higher x values; in fact, all crabs with width >29 cm have satellites. The positive effect of width is also suggested by the grouping of the data used to investigate adequacy of Poisson regression models in Section 4.3.3 (Table 4.4). 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 shows eight dots representing the sample proportions of female crabs having satellites plotted against the mean widths for the eight categories. Figure 5.2 also shows a curve based on smoothing the data using the generalized additive modeling method, assuming a binomial response and logit link. This curve shows a roughly increasing trend and is more informative than viewing the binary data alone. It suggests that an S-shaped regression function may describe this relationship relatively well. Since the eight plotted sample proportions and the GAM smoothing curve both suggest an increasing trend, we proceed with fitting the logistic regression model with linear width predictor.

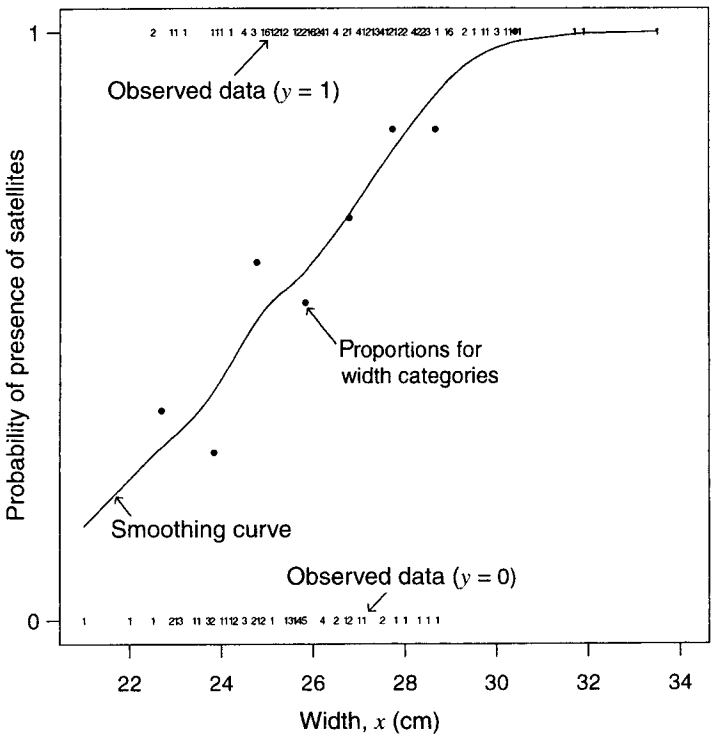

Figure 5.2 Whether satellites are present (1 = yes, 0 = no) by width of female crab, with smoothing fit of generalized additive model.

Table 5.1 Output (Based on SAS) for Logistic Regression Model with Horseshoe Crab Data

```
        Criteria For Assessing Goodness Of Fit
Criterion              DF           Value
Deviance              171        194.4527
Pearson Chi-Square    171        165.1434
Log Likelihood                   -97.2263

                    Std     Likelihood-Ratio     Wald
Parameter  Estimate Error    95% Conf Limits    Chi-Sq  P>ChiSq
Intercept -12.3508 2.6287  -17.8097  -7.4573    22.07   <.0001
width       0.4972 0.1017    0.3084   0.7090    23.89   <.0001
```

We defer to Section 5.5 the details about ML fitting. Software (see the text website) reports output such as Table 5.1 exhibits. Let $\pi(x)$ denote the probability that a female horseshoe crab of width x has a satellite. The ML fit is

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

Substituting $x = 26.3$ cm, the mean width level in this sample, $\hat{\pi}(x) = 0.674$. The estimated probability equals $\frac{1}{2}$ when $x = -\hat{\alpha}/\hat{\beta} = 12.351/0.497 = 24.8$.

Figure 5.3 plots $\hat{\pi}(x)$ from the logistic fit against width, again superimposing the sample proportions that we viewed in Figure 5.2. The curve seems to follow reasonably well the trend in those proportions.

The estimated odds of a satellite multiply by $\exp(\hat{\beta}) = \exp(0.497) = 1.64$ for each 1-cm increase in width; that is, there is a 64% increase. To convey the effect less technically, we could report the incremental rate of change in the probability of a satellite.

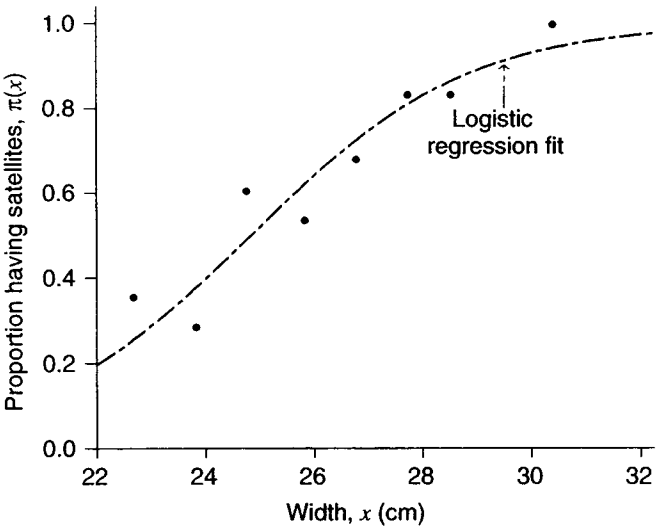

Figure 5.3 Logistic regression fitted curve and sample proportions of satellites, by width of female crab.

At the mean width, $\hat{\pi}(x) = 0.674$, and $\hat{\pi}(x)$ increases by about $\hat{\beta}[\hat{\pi}(x)(1 - \hat{\pi}(x))] = 0.497(0.674)(0.326) = 0.11$ for a 1-cm increase in width. Or, we could report $\hat{\pi}(x)$ at the quartiles of x. The lower quartile, median, and upper quartile for width are 24.9, 26.1, and 27.7; $\hat{\pi}(x)$ at those values equals 0.51, 0.65, and 0.81, increasing by 0.30 over the x values for the middle half of the sample.

The latter summary is useful for comparing the effects of predictors having different units. For instance, with the female crab's weight as the predictor, $\text{logit}[\hat{\pi}(x)] = -3.695 + 1.815x$. A 1-kg increase in weight is not comparable to a 1-cm increase in width, so $\hat{\beta} = 1.815$ for $x =$ weight is not comparable to $\hat{\beta} = 0.497$ for $x =$ width. The quartiles for weight are 2.00, 2.35, and 2.85; $\hat{\pi}(x)$ at those values are 0.48, 0.64, and 0.81, increasing by 0.33 over the middle half of the sampled weights. The effect is similar to that of width, which is not surprising as these predictors are very highly correlated.

5.1.4 Logistic Regression with Retrospective Studies

Another property of logistic regression relates to situations in which the explanatory variable X rather than the response variable Y is random. This occurs with retrospective sampling designs, such as case–control biomedical studies. For samples of subjects having $y = 1$ (cases) and having $y = 0$ (controls), the value of X is observed. Evidence exists of an association if the distribution of X values differs between cases and controls. In retrospective studies, we can estimate odds ratios. Effects in the logistic regression model refer to odds ratios. We can fit such models and estimate effects in case–control studies (Breslow and Powers 1978, Prentice and Pyke 1979).

Here is a justification for this. Let Z indicate whether a subject is sampled (1 = yes, 0 = no). Let $\rho_1 = P(Z = 1|y = 1)$ denote the probability of sampling a case, and let $\rho_0 = P(Z = 1|y = 0)$ denote the probability of sampling a control. Even though the conditional distribution of Y given $X = x$ is not sampled, we need a model for $P(Y = 1|z = 1, x)$, assuming that $P(Y = 1|x)$ follows the logistic model. By Bayes' theorem,

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

Now, suppose that $P(Z = 1|y, x) = P(Z = 1|y)$ for $y = 0$ and 1; that is, for each y, the sampling probabilities do not depend on x. For instance, often x refers to exposure of some type, such as whether someone has been a smoker. Then, for cases and for controls, the probability of being sampled is the same for smokers and nonsmokers. Under this assumption, substituting ρ_1 and ρ_0 in (5.3) and dividing numerator and denominator by $P(Y = 0|x)$, we get

$$P(Y = 1|z = 1, x) = \frac{\rho_1 \exp(\alpha + \beta x)}{\rho_0 + \rho_1 \exp(\alpha + \beta x)}.$$

Then, dividing numerator and denominator by ρ_0 and using $\rho_1/\rho_0 = \exp[\log(\rho_1/\rho_0)]$ yields

$$\text{logit}[P(Y = 1|z = 1, x)] = \alpha^* + \beta x$$

with $\alpha^* = \alpha + \log(\rho_1/\rho_0)$. The logistic regression model holds with the same effect parameter β as in the model for $P(Y = 1|x)$. If the sampling rate for cases is greater than

that for controls, the intercept estimated is larger than the one estimated with a prospective study.

With case–control studies, it is not possible to estimate β in binary-response models with links other than the logit. Unlike the odds ratio, the effect for the conditional distribution of X given y does not then equal that for Y given x. This is an important advantage of the logit link and is one reason why logistic regression models are so popular in biomedical studies.

Many case–control studies employ matching. Each case is matched with one or more control subjects. The controls are like the case on key characteristics such as age. The model and subsequent analysis should take the matching into account. In Section 11.2.5 we discuss logistic regression for matched case–control studies.

5.1.5 Logistic Regression Is Implied by Normal Explanatory Variables

Regardless of the sampling mechanism, logistic regression may or may not describe a relationship well. In one special case, it necessarily holds. Given that $Y = i$, suppose that X has $N(\mu_i, \sigma^2)$ distribution, $i = 0, 1$. Then, by Bayes' theorem, $P(Y = 1|X = x)$ satisfies the logistic model with $\beta = (\mu_1 - \mu_0)/\sigma^2$ (Cornfield 1962). Thus, when a population is a mixture of two types of subjects, one type with $y = 1$ that is approximately normally distributed on X and the other type with $y = 0$ that is approximately normal on X with similar variance, the logistic regression function approximates well the curve for $\pi(x)$.

The result extends to a vector of explanatory variables having multivariate normal distributions in each case (Exercise 5.30 and Section 15.1.1). If the distributions are normal but with different variances, the model applies but having a quadratic term. In that case, the relationship is nonmonotone, with $\pi(x)$ increasing and then decreasing, or the reverse.

5.2 INFERENCE FOR LOGISTIC REGRESSION

By standard results, ML estimators of logistic regression model parameters have large-sample normal distributions. Inference can use the (Wald, likelihood-ratio, score) triad of methods introduced in Section 1.3.3.

5.2.1 Inference About Model Parameters and Probabilities

For the logistic model with a single predictor,

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

significance tests focus on H_0: $\beta = 0$, the hypothesis of independence. The Wald test uses the log likelihood at $\hat{\beta}$, with test statistic $z = \hat{\beta}/SE$ or its square; under H_0, z^2 is asymptotically χ^2_1. The likelihood-ratio test uses twice the difference between the maximized log likelihood at $\hat{\beta}$ and at $\beta = 0$ and also has an asymptotic χ^2_1 null distribution. The score test uses the log likelihood at $\beta = 0$ through the derivative of the log likelihood (i.e., the score function) at that point. The test statistic compares the sufficient statistic for β to its null expected value, suitably standardized [$N(0,1)$ or χ^2_1]. In Section 5.3.5 we present this test.

A confidence interval for β results from inverting a test of H_0: $\beta = \beta_0$. The interval is the set of β_0 for which the chi-squared test statistic is no greater than $\chi^2_1(\alpha) = z^2_{\alpha/2}$. For the Wald approach, this means $[(\hat{\beta} - \beta_0)/SE]^2 \leq z^2_{\alpha/2}$, so the interval is $\hat{\beta} \pm z_{\alpha/2}(SE)$.

For summarizing the relationship, other characteristics may have greater importance than β, such as $\pi(x)$ at various x values. For fixed $x = x_0$, $\text{logit}[\hat{\pi}(x_0)] = \hat{\alpha} + \hat{\beta}x_0$ has a large-sample SE given by the estimated square root of

$$\text{var}(\hat{\alpha} + \hat{\beta}x_0) = \text{var}(\hat{\alpha}) + x_0^2\,\text{var}(\hat{\beta}) + 2x_0\,\text{cov}(\hat{\alpha}, \hat{\beta}).$$

A 95% confidence interval for $\text{logit}[\pi(x_0)]$ is $(\hat{\alpha} + \hat{\beta}x_0) \pm 1.96(SE)$. Substituting each endpoint into the inverse transformation $\pi(x_0) = \exp(\text{logit})/[1 + \exp(\text{logit})]$ gives a corresponding interval for $\pi(x_0)$.

5.2.2 Example: Inference for Horseshoe Crab Mating Data

We illustrate logistic regression inferences with the model for the probability that a horseshoe crab has a satellite, with crab width as the predictor. Table 5.1 showed the fit and standard errors. The statistic $z = \hat{\beta}/SE = 0.497/0.102 = 4.89$ provides strong evidence of a positive width effect ($P < 0.0001$). The equivalent Wald chi-squared statistic, $z^2 = 23.89$, has df = 1. The maximized log likelihoods equal -112.88 under H_0: $\beta = 0$ and -97.23 for the full model. The likelihood-ratio statistic equals $-2[-112.88 - (-97.23)] = 31.31$, with df = 1. This provides even stronger evidence than the Wald test.

The Wald 95% confidence interval for β is $0.497 \pm 1.96(0.102)$, or (0.298, 0.697). Table 5.1 reports a likelihood-ratio confidence interval of (0.308, 0.709), based on the profile likelihood function. The confidence interval for the effect on the odds per 1-cm increase in width equals $(e^{0.308}, e^{0.709}) = (1.36, 2.03)$. We infer that a 1-cm increase in width has at least a 36% increase and at most a doubling in the odds of a satellite.

Most software for logistic regression also can report estimates and confidence intervals for $\pi(x)$ (for examples, see the text website). Consider this for crabs of width $x = 26.5$, which is near the mean width. The estimated logit is $-12.351 + 0.497(26.5) = 0.826$, and $\hat{\pi}(x) = 0.695$. Software reports

$$\widehat{\text{var}}(\hat{\alpha}) = 6.91023, \quad \widehat{\text{var}}(\hat{\beta}) = 0.01035, \quad \widehat{\text{cov}}(\hat{\alpha}, \hat{\beta}) = -0.26685,$$

from which

$$\widehat{\text{var}}\{\text{logit}[\hat{\pi}(x)]\} = 6.91023 + x^2(0.01035) + 2x(-0.26685).$$

At $x = 26.5$ this is 0.0356, so the 95% confidence interval for $\text{logit}[\pi(26.5)]$ equals $0.826 \pm (1.96)\sqrt{0.0356}$, or (0.456, 1.196). This translates to the interval (0.61, 0.77) for the probability of satellites (e.g., $\exp(0.456)/[1 + \exp(0.456)] = 0.61$). Since $\text{corr}(\hat{\alpha}, \hat{\beta})$ is near 1.0, for better computational precision, fit the model using predictor $x^* = x - 26.5$, so that $\hat{\alpha}$ and its SE are the estimated logit and its SE. Figure 5.4 plots the confidence bands around the prediction equation for $\pi(x)$ as a function of x. Hauck (1983) gave alternative bands for which the confidence coefficient applies simultaneously to all possible predictor values.

We could ignore the model fit and simply use sample proportions (i.e., the saturated model) to estimate such probabilities. Six female crabs in the sample had $x = 26.5$, and

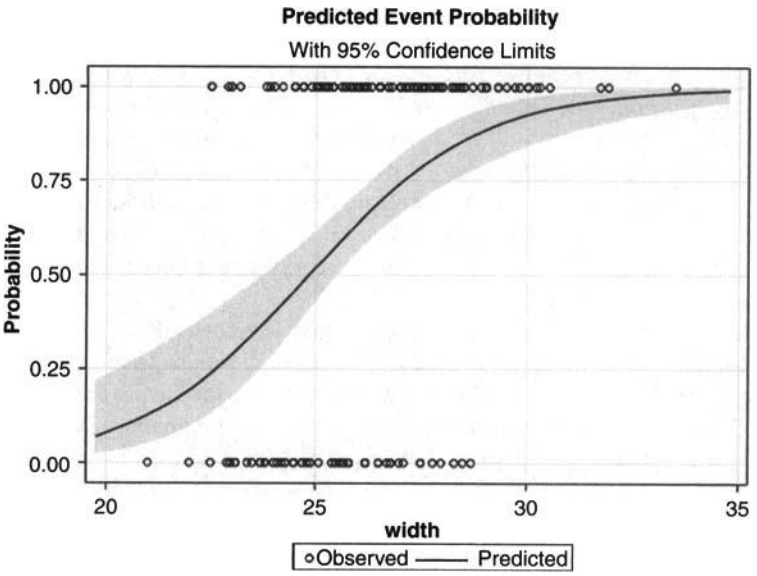

Figure 5.4 Prediction equation and 95% confidence bands (from SAS PROC LOGISTIC) for probability of satellite as a function of width.

four of them had satellites. The sample proportion estimate at $x = 26.5$ is $\hat{\pi} = 4/6 = 0.67$, similar to the model-based estimate. The 95% score confidence interval (Section 1.4.2) based on these six observations alone equals (0.30, 0.90).

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 saturated model has a separate parameter for every distinct value of x. For instance, at $x = 26.5$, software reports $SE = 0.04$ for the model-based estimate 0.695, whereas the SE is $\sqrt{\hat{\pi}(1-\hat{\pi})/n} = \sqrt{(0.67)(0.33)/6} = 0.19$ for the sample proportion of 0.67 with only 6 observations. The 95% confidence intervals are (0.61, 0.77) using the model versus (0.30, 0.90) using the sample proportion. 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, the model is not *exactly* the true relationship between $\pi(x)$ and x. However, if it approximates the true probabilities decently, its estimator still tends to be 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. Section 5.3.10 discusses this advantage of using models.

5.2.3 Checking Goodness of Fit: Grouped and Ungrouped Data

In practice, there is no guarantee that a certain logistic regression model fits the data well. For any type of binary data, one way to detect lack of fit uses a likelihood-ratio test to compare the model to more complex ones. A more complex model might contain a nonlinear effect. Models with multiple predictors would consider interaction. 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. This is simplest when the explanatory variables are solely categorical, as we'll illustrate in Section 5.4.3. At each setting of x, multiplying the estimated probabilities of the two outcomes by the number of subjects at that setting yields estimated expected frequencies for $y = 0$ and $y = 1$. These are *fitted values*. The test of the model compares the observed counts and fitted values using a Pearson X^2 or likelihood-ratio G^2 statistic. For a fixed number of settings, as the fitted counts increase, X^2 and G^2 have limiting chi-squared null distributions. The degrees of freedom, called the *residual* df for the model, subtract the number of parameters in the model from the number of parameters in the saturated model (i.e., the number of settings of x).

The reason for the restriction to categorical predictors for a global test of fit relates to the distinction that we mentioned in Section 4.5.3 between grouped and ungrouped data for binomial models. The saturated model differs in the two cases. An asymptotic chi-squared distribution for the deviance results as $n \to \infty$ with a fixed number of parameters in that model and hence a fixed number of settings of predictor values (i.e., *grouped* data).

5.2.4 Example: Model Goodness of Fit for Horseshoe Crab Data

We illustrate with a goodness-of-fit analysis for the model using $x =$ width to predict the probability that a female crab has a satellite. One way to check it compares it to a more complex model, such as the model containing a quadratic term or linear spline. With width centered at 0 by subtracting its mean of 26.3, the quadratic model has fit

$$\text{logit}[\hat{\pi}(x)] = 0.618 + 0.533(x - \bar{x}) + 0.040(x - \bar{x})^2.$$

The quadratic estimate has $SE = 0.046$. There is not much evidence to support adding that term. The likelihood-ratio statistic for testing that the true coefficient of x^2 is 0 equals 0.83 (df = 1).

We next evaluate overall goodness of fit. Width takes 66 distinct values for the 173 crabs, with few observations at most widths. We can view the data as a 66×2 contingency table. The two cells in each row count the number of crabs with satellites and the number of crabs without satellites, at that width. The chi-squared theory for X^2 and G^2 applies when the number of levels of x is fixed, and the number of observations at each level grows. Although we grouped the data using the distinct width values rather than using 173 separate binary responses, this theory is violated here 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 with continuous or nearly continuous predictors do not have approximate chi-squared distributions. Normal approximations can be more appropriate (see Section 10.6.4 for references), but no such method has become as popular as methods presented next.

5.2.5 Checking Goodness of Fit with Ungrouped Data by Grouping

As just noted, with ungrouped data or with continuous or nearly continuous predictors, X^2 and G^2 do not have limiting chi-squared distributions. They are still useful for comparing models, as done above for checking a quadratic term. Also, we can apply them in an

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

Width (cm)	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

approximate manner to grouped observed and fitted values for a partition of the space of x values.

Table 5.2 uses the groupings of Table 4.4, giving an 8×2 table. In each width category, the fitted value for a "yes" response is the sum of the estimated probabilities $\hat{\pi}(x)$ for all crabs having width in that category; the fitted value for a "no" response is the sum of $1 - \hat{\pi}(x)$ for those crabs. The fitted values are then much larger. Then, X^2 and G^2 have better validity, although the chi-squared theory still is not perfect because $\pi(x)$ is not constant in each category. Their values are $X^2 = 5.3$ and $G^2 = 6.2$. Table 5.2 has eight binomial samples, one for each width setting; the model has two parameters, so $df = 8 - 2 = 6$. Neither X^2 nor G^2 shows evidence of lack of fit ($P > 0.4$). Thus, we can feel more comfortable about using the model for the original ungrouped data.

As the number of explanatory variables increases, this strategy loses effectiveness. Simultaneous grouping of values for each variable can produce a contingency table with a large number of cells, most of which have very small counts.

Regardless of the number of explanatory variables, we can partition observed and fitted values according to the estimated probabilities of success using the original ungrouped data. One common approach forms the groups in the partition so they have approximately equal size. With 10 groups, the first pair of observed counts and corresponding fitted counts refers to the $n/10$ observations having the highest estimated probabilities, the next pair refers to the $n/10$ observations having the second decile of estimated probabilities, and so on. Each group has an observed count of subjects with each outcome and a fitted value for each outcome. The fitted value for an outcome is the sum of the estimated probabilities for that outcome for all observations in that group.

This construction is the basis of a test due to Hosmer and Lemeshow (1980). They proposed a Pearson statistic comparing the observed and fitted counts for this partition. Let y_{ij} denote the binary outcome for observation j in group i of the partition, $i = 1, \ldots, g$, $j = 1, \ldots, n_i$. Let $\hat{\pi}_{ij}$ denote the corresponding fitted probability for the model fitted to the ungrouped data. Their statistic equals

$$\sum_{i=1}^{g} \frac{\left(\sum_j y_{ij} - \sum_j \hat{\pi}_{ij}\right)^2}{\left(\sum_j \hat{\pi}_{ij}\right)\left[1 - \left(\sum_j \hat{\pi}_{ij}\right)/n_i\right]}.$$

When many observations have the same estimated probability, there is some arbitrariness in forming the groups, and different software may report somewhat different values. This

statistic does not have a limiting chi-squared distribution, because the observations in a group do not share a common success probability and thus are not identical trials. However, Hosmer and Lemeshow noted that when the number of distinct patterns of covariate values equals the sample size, the null distribution is approximated by chi-squared with df $= g - 2$.

For the logistic regression fitted to the horseshoe crab data with continuous width predictor, the Hosmer–Lemeshow statistic with $g = 10$ groups equals 3.5, with df $= 8$. It also indicates a decent fit.

In summary, the X^2 and G^2 goodness-of-fit tests work well when n is large relative to the number of distinct covariate patterns, whereas the Hosmer–Lemeshow test works well when the number of distinct covariate patterns is large. Unfortunately, like other proposed global fit statistics, the Hosmer–Lemeshow statistic does not have good power for detecting particular types of lack of fit (Hosmer et al. 1997). One example is when the correct model has an interaction between a binary and continuous covariate but the chosen model has only the continuous covariate. Tsiatis (1980) suggested an alternative goodness-of-fit test that partitions values for the explanatory variables into a set of regions and adds an indicator variable to the model for each region. The test statistic compares the fit of this model to the simpler one, testing that the extra parameters are not needed. Alternatively, one could use a bootstrap method to evaluate fit. Azzalini et al. (1989) used the parametric bootstrap to evaluate the distance between the logistic model fit and a nonparametric smoothing of the data (to be introduced in Section 7.4.2); the bootstrap simulations estimated the proportion of times that a likelihood-ratio form of statistic is larger than observed. In any case, 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 to a more complex one is more useful from a scientific perspective, since it searches for lack of fit of a particular type.

For any approach to checking fit, when the fit is poor, diagnostic measures describe the influence of individual observations on the model fit and highlight reasons for the inadequacy. We discuss these in Section 6.2.1.

5.2.6 Wald Inference Can Be Suboptimal

Wald, likelihood-ratio, and score methods of inference usually give similar results for large samples. Each method of inference can also produce small-sample confidence intervals and tests. We defer discussion of this until Sections 7.3, 16.5, and 16.6.

Although these methods usually give similar results, the Wald method has two disadvantages compared with the likelihood-ratio and score methods. First, its results depend on the scale for the parameterization. To illustrate, suppose that Y has a bin(n, π) distribution. For the model, logit$(\pi) = \alpha$, consider testing H_0: $\alpha = 0$ (i.e., $\pi = 0.50$). From Section 3.1.6, the asymptotic variance of $\hat{\alpha} = \text{logit}(\hat{\pi})$ (with $\hat{\pi} = y/n$) is $[n\pi(1-\pi)]^{-1}$. The Wald chi-squared test statistic is $[\text{logit}(\hat{\pi})]^2[n\hat{\pi}(1-\hat{\pi})]$. On the proportion scale, the Wald statistic is $(\hat{\pi} - 0.50)^2[n/\hat{\pi}(1-\hat{\pi})]$. These are not the same. For example, when $\hat{\pi}$ is near 0 or 1 (so $|\hat{\alpha}|$ is large), the ratio of the Wald statistic on the logit scale to the Wald statistic on the proportion scale approaches 0 as n increases. Evaluations reveal that the logit-scale statistic tends to be too conservative and the proportion-scale statistic tends to be too liberal.

This behavior of the Wald statistic for the logit reflects another disadvantage. When a true effect is relatively large, the Wald test is not as powerful as the likelihood-ratio and score test and can even show aberrant behavior (Hauck and Donner 1977). For the single-binomial case just described, for example, suppose $n = 25$. We would regard $y = 24$ as

stronger evidence against H_0 than $y = 23$, yet the logit Wald statistic equals 9.7 when $y = 24$ and 11.0 when $y = 23$. For comparison, the likelihood-ratio statistics are 26.3 and 20.7.

More generally, Hauck and Donner showed that for fixed sample size, the Wald statistic for testing H_0: $\beta = 0$ in the logistic model eventually starts decreasing and actually converges toward 0 as $\hat{\beta}$ grows unboundedly. A similar result holds for logistic models with multiple predictors.

5.3 LOGISTIC MODELS WITH CATEGORICAL PREDICTORS

Like ordinary regression, logistic regression extends to include qualitative explanatory variables, often called *factors*, as first noted by Dyke and Patterson (1952). We use indicator variables to do this.

5.3.1 ANOVA-Type Representation of Factors

For simplicity, we first consider a single factor X, with I categories. In row i of the $I \times 2$ table, let y_i be the number of outcomes in the first column (successes) out of n_i trials. We treat y_i as binomial with parameter π_i.

The logistic regression model with a single factor as a predictor is

$$\log \frac{\pi_i}{1 - \pi_i} = \alpha + \beta_i. \tag{5.4}$$

The higher β_i is, the higher the value of π_i. The right-hand side of (5.4) resembles the model formula for means in one-way ANOVA.

As in ANOVA, the factor has as many parameters $\{\beta_i\}$ as categories. Unless we delete α from the model, one β_i is redundant. One β_i can be set to 0, say, $\beta_I = 0$ for the last category. If the values do not satisfy this, we can recode so that it is true. For instance, set $\tilde{\beta}_i = \beta_i - \beta_I$ and $\tilde{\alpha} = \alpha + \beta_I$, which satisfy $\tilde{\beta}_I = 0$. Then

$$\text{logit}(\pi_i) = \alpha + \beta_i = (\tilde{\alpha} - \beta_I) + (\tilde{\beta}_i + \beta_I) = \tilde{\alpha} + \tilde{\beta}_i,$$

where the newly defined parameters satisfy the constraint. When $\beta_I = 0$, α equals the logit in row I, and β_i is the difference between the logits in rows i and I. Thus, β_i equals the log odds ratio for that pair of rows.

For *any* $\{\pi_i > 0\}$, $\{\beta_i\}$ exist such that model (5.4) holds. The model has as many parameters (I) as binomial observations and is *saturated*. When a factor has *no* effect, $\beta_1 = \beta_2 = \cdots = \beta_I$. Since this is equivalent to $\pi_1 = \cdots = \pi_I$, this case corresponds to statistical independence of X and Y.

5.3.2 Indicator Variables Represent a Factor

An equivalent expression of model (5.4) uses indicator variables. Let $x_i = 1$ for observations in row i and $x_i = 0$ otherwise, $i = 1, \ldots, I - 1$. The model is

$$\text{logit}(\pi_i) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{I-1} x_{I-1}.$$

This accounts for parameter redundancy by not forming an indicator variable for category I. The constraint $\beta_I = 0$ corresponds to this choice of indicator variables. The category to exclude for an indicator variable is arbitrary. Some software sets $\beta_1 = 0$; this corresponds to a model with indicator variables for categories 2 through I, but not category 1.

Another way to impose constraints sets $\sum_i \beta_i = 0$. When X has $I = 2$ categories, then $\beta_1 = -\beta_2$. This results from *effect coding* for an indicator variable, $x = 1$ in category 1 and $x = -1$ in category 2.

The same substantive results about estimable effects occur for any coding scheme. For model (5.4), regardless of the constraint for $\{\beta_i\}$, the linear predictor values $\{\hat{\alpha} + \hat{\beta}_i\}$ and hence $\{\hat{\pi}_i\}$ are the same. The differences $\hat{\beta}_a - \hat{\beta}_b$ for pairs (a, b) of categories of X are identical and represent estimated log odds ratios. Thus, $\exp(\hat{\beta}_a - \hat{\beta}_b)$ is the estimated odds of success in category a of X divided by the estimated odds of success in category b of X. Reparameterizing a model may change parameter estimates but does not change the model fit or the effects of interest.

The value β_i or $\hat{\beta}_i$ for a single category is irrelevant. Different constraint systems result in different values. For a binary predictor, for instance, using indicator variables with reference value $\beta_2 = 0$, the log odds ratio equals $\beta_1 - \beta_2 = \beta_1$; by contrast, for effect coding with ± 1 indicator variable and hence $\beta_1 + \beta_2 = 0$, the log odds ratio equals $\beta_1 - \beta_2 = \beta_1 - (-\beta_1) = 2\beta_1$. A parameter or its estimate makes sense only by comparison with one for another category.

5.3.3 Example: Alcohol and Infant Malformation Revisited

We return now to Table 3.8 from the study of maternal alcohol consumption and child's congenital malformations, shown again in Table 5.3. For model (5.4), we treat malformation status as the response and alcohol consumption as an explanatory factor. Regardless of the constraint for $\{\beta_i\}$, the model is saturated and $\{\hat{\alpha} + \hat{\beta}_i\}$ are the sample logits, reported in Table 5.3. For instance,

$$\text{logit}(\hat{\pi}_1) = \hat{\alpha} + \hat{\beta}_1 = \log(48/17,066) = -5.87.$$

For the coding that constrains $\beta_5 = 0$, $\hat{\alpha} = -3.61$ and $\hat{\beta}_1 = -2.26$. For the coding $\beta_1 = 0$, $\hat{\alpha} = -5.87$. Table 5.3 shows that except for the slight reversal between the first and second categories of alcohol consumption, the sample logits and hence the sample proportions of malformation cases increase as alcohol consumption increases.

Table 5.3 Sample Logits and Proportion of Malformation for Table 3.8, with Fitted Proportions for Linear Logit Model

Alcohol	Malformation		Sample	Proportion Malformed	
Consumption	Present	Absent	Logit	Observed	Fitted
0	48	17,066	−5.87	0.0028	0.0026
<1	38	14,464	−5.94	0.0026	0.0030
1–2	5	788	−5.06	0.0063	0.0041
3–5	1	126	−4.84	0.0079	0.0091
≥6	1	37	−3.61	0.0263	0.0231

The simpler model with all $\beta_i = 0$ specifies independence. For it, $\hat{\alpha}$ equals the logit for the overall sample proportion of malformations, which is $\log(93/32,481) = -5.86$. To test H_0: independence (df = 4), the Pearson statistic (3.10) is $X^2 = 12.1$ ($P = 0.02$), and the likelihood-ratio statistic (3.11) is $G^2 = 6.2$ ($P = 0.19$). These provide mixed signals. Table 5.3 has a mixture of very small, moderate, and extremely large counts. Even though $n = 32,574$, the null sampling distributions of X^2 or G^2 may not be close to chi-squared. The P-values using the exact conditional distributions of X^2 and G^2 (Section 16.5.2) are 0.03 and 0.13. These are closer, but still give differing evidence. In any case, these statistics ignore the ordinality of alcohol consumption. The sample suggests that malformations may tend to be more likely with higher alcohol consumption. The first two proportions are similar and the next two are also similar, however, and either of the last two proportions changes substantially with the addition or deletion of one malformation case.

5.3.4 Linear Logit Model for $I \times 2$ Contingency Tables

Model (5.4) is invariant to the ordering of categories, so it treats the explanatory factor as nominal. For ordered factor categories, other models are more parsimonious, yet more complex than the independence model. For instance, let $(x_1, x_2, \ldots, x_I)$ be scores that describe distances between categories of X. When we expect a monotone effect of X on Y, it is natural to fit the *linear logit model*

$$\text{logit}(\pi_i) = \alpha + \beta x_i. \tag{5.5}$$

The independence model is the special case $\beta = 0$.

The near-monotone increase in the sample logits in Table 5.3 indicates that the linear logit model may fit better than the independence model. As measured, alcohol consumption groups a naturally continuous variable. With scores ($x_1 = 0, x_2 = 0.5, x_3 = 1.5, x_4 = 4.0, x_5 = 7.0$), the last score being somewhat arbitrary, Table 5.4 shows results. The estimated multiplicative effect of a unit increase in daily alcohol consumption on the odds of malformation is $\exp(0.317) = 1.37$. Table 5.3 shows the observed and fitted proportions of malformation. The model seems to fit well, as statistics comparing observed and fitted counts are $G^2 = 1.95$ and $X^2 = 2.05$, with df = 3.

Table 5.4 Software Output (Based on SAS) for Linear Logit Model Fitted to Table 5.3 on Infant Malformation and Alcohol Consumption

Criteria For Assessing Goodness Of Fit		
Criterion	DF	Value
Deviance	3	1.9487
Pearson Chi-Square	3	2.0523
Log Likelihood		−635.5968

Parameter	Estimate	Std Error	Likelihood-Ratio 95% Conf	Limits	Wald Chi-Sq	Pr>ChiSq
Intercept	−5.9605	0.1154	−6.1930	−5.7397	2666.41	<.0001
alcohol	0.3166	0.1254	0.0187	0.5236	6.37	0.0116

5.3.5 Cochran–Armitage Trend Test

Armitage (1955) and Cochran (1954) were among the first to emphasize the importance of utilizing ordered categories in a contingency table. For $I \times 2$ tables with ordered rows and I independent bin(n_i, π_i) variates $\{y_i\}$, they proposed a trend statistic for testing independence by partitioning the Pearson statistic for that hypothesis. They used a linear probability model,

$$\pi_i = \alpha + \beta x_i, \tag{5.6}$$

fitted by ordinary least squares. The null hypothesis of independence is H_0: $\beta = 0$. Let $\bar{x} = \sum_i n_i x_i / n$. Let $p_i = y_i / n_i$, and let $p = (\sum_i y_i)/n$ denote the overall proportion of successes. The prediction equation is

$$\hat{\pi}_i = p + b(x_i - \bar{x}),$$

where

$$b = \frac{\sum_i n_i (p_i - p)(x_i - \bar{x})}{\sum_i n_i (x_i - \bar{x})^2}.$$

Denote the Pearson statistic for testing independence by $X^2(I)$. We express $X^2(I)$ in terms of variation among the I sample proportions by

$$X^2(I) = \frac{1}{p(1-p)} \sum_i n_i (p_i - p)^2.$$

Reported by Fisher (1934) and attributed to A. E. Brandt and G. W. Snedecor, this is referred to as the *Brandt–Snedecor formula*. It generalizes the equality in 2×2 tables between X^2 and the square of the pooled two-sample z-statistic (3.12). Cochran (1954) noted that this Pearson formula decomposes into

$$X^2(I) = z^2 + X^2(L),$$

where

$$X^2(L) = \frac{1}{p(1-p)} \sum_i n_i (p_i - \hat{\pi}_i)^2,$$

$$z^2 = \frac{b^2}{p(1-p)} \sum_i n_i (x_i - \bar{x})^2 = \left[\frac{\sum_i (x_i - \bar{x}) y_i}{\sqrt{p(1-p) \sum_i n_i (x_i - \bar{x})^2}} \right]^2. \tag{5.7}$$

When the linear probability model holds, $X^2(L)$ is asymptotically chi-squared with df $= I - 2$. It tests the fit of the model. The statistic z^2, with df $= 1$, tests H_0: $\beta = 0$ for the linear trend (5.6) in the proportions. The test of independence using this statistic is called the *Cochran–Armitage trend test*.

This statistic relates to the correlation-based statistic M^2 introduced in (3.16) in Section 3.4.1 to test for a linear trend in an $I \times J$ table; namely, $z^2 = nr^2 = [n/(n-1)]M^2$. See Yates (1948) and Mantel (1963). When $I = 2$, then $X^2(L) = 0$ and $z^2 = X^2(I)$.

The Cochran–Armitage trend test seems unrelated to the linear logit model. However, this test statistic is equivalent to the score statistic for testing H_0: $\beta = 0$ in that model. In fact, Tarone and Gart (1980) showed that the score test for a binary linear trend model does not depend on the link function. Thus, this trend test is locally asymptotically efficient for both linear and logistic alternatives for $P(Y = 1)$. See Cox (1958a) for related remarks. Gross (1981) showed that when the linear logit model holds but we use an incorrect set of scores, the local asymptotic relative efficiency for testing independence using the statistic with those scores equals the square of the Pearson correlation between the true and the incorrect scores.

5.3.6 Example: Alcohol and Infant Malformation Revisited

For Table 5.3 on alcohol consumption and infant malformation, $X^2(I) = 12.08$. Using the scores (0, 0.5, 1.5, 4.0, 7.0) as in the linear logit model, the Cochran–Armitage trend test has $z^2 = 6.57$ (P-value $= 0.010$). The test suggests strong evidence of a positive slope. In addition,

$$X^2(I) = 12.08 = 6.57 + 5.51,$$

where $X^2(L) = 5.51$ (df $= 3$) shows only slight evidence of departure of the proportions from linearity. The trend test result is nearly identical to the test using $M^2 = (n-1)r^2$ based on the sample correlation of $r = 0.0142$ for $n = 32,573$. For the chosen scores, the correlation seems weak. However, r has limited use as a descriptive measure for tables that are highly discrete and unbalanced.

The Cochran–Armitage trend test (i.e., the score test) usually gives results similar to the Wald or likelihood-ratio test of H_0: $\beta = 0$ in the linear logit model. The asymptotics work well even for quite small n when $\{n_i\}$ are equal and $\{x_i\}$ are equally spaced. With Table 5.3, the Wald statistic equals $(\hat{\beta}/SE)^2 = (0.3166/0.1254)^2 = 6.37$ ($P = 0.012$) and the likelihood-ratio statistic equals 4.25 ($P = 0.039$). Here, however, the highly unbalanced counts suggest that it is best not to use the Wald approach for testing or for interval estimation. The profile likelihood 95% confidence interval of (0.02, 0.52) for β reported in Table 5.4 is preferable to the Wald interval of $0.317 \pm 1.96(0.125) = (0.07, 0.56)$. The sample size in the last row is relatively small, and the single "present" observation in that row is highly influential. P-values depend dramatically on whether that observation is included in the analysis (Exercise 5.10).

5.3.7 Using Directed Models Can Improve Inferential Power

When contingency tables have ordered categories, in Section 3.4 we showed that tests that utilize the ordering can have improved power. Testing independence against a linear trend alternative in a linear logit model is a way to do this. In this section we present the reason for these power improvements.

In an $I \times 2$ contingency table for I binomial variates with parameters $\{\pi_i\}$, H_0: independence states $\text{logit}(\pi_i) = \alpha$. The ordinary G^2 and X^2 statistics of Section 3.2.1 refer to the general alternative,

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

which is saturated. They test H_0: $\beta_1 = \beta_2 = \cdots = \beta_I$ in that model, with df $= (I - 1)$. Their general alternative treats both classifications as nominal. Denote these test statistics as $G^2(I)$ and $X^2(I)$. Note that $G^2(I)$ is the likelihood-ratio statistic $G^2(M_0|M_1) = -2(L_0 - L_1)$ for comparing the saturated model M_1 with the independence (I) model M_0.

Ordinal test statistics refer to narrower, often more relevant, alternatives. With ordered rows, an example is a test of H_0: $\beta = 0$ in the linear logit model, $\text{logit}(\pi_i) = \alpha + \beta x_i$. The likelihood-ratio statistic $G^2(I|L) = G^2(I) - G^2(L)$ compares the linear logit model and the independence model. When a test statistic focuses on a single parameter, such as β in that model, it has df $= 1$. Now, df equals the mean of the chi-squared distribution. A large test statistic with df $= 1$ falls farther out in its right-hand tail than a comparable value of $X^2(I)$ or $G^2(I)$ with df $= (I - 1)$. Thus, it has a smaller P-value.

5.3.8 Noncentral Chi-Squared Distribution and Power for Narrower Alternatives

To compare power of $G^2(I|L)$ and $G^2(I)$, it is necessary to compare their nonnull sampling distributions. When H_0 is false, their distributions are approximately *noncentral chi-squared*. This distribution, introduced by R. A. Fisher in 1928, arises from the following construction: If $Z_i \sim N(\mu_i, 1)$, $i = 1, \ldots, \nu$, and if $Z_1, \ldots, Z_\nu$ are independent, $\sum_i Z_i^2$ has the noncentral chi-squared distribution with df $= \nu$ and *noncentrality parameter* $\lambda = \sum_i \mu_i^2$. Its mean is $\nu + \lambda$ and its variance is $2(\nu + 2\lambda)$. The ordinary (central) chi-squared distribution, which occurs when H_0 is true, has $\lambda = 0$.

Let $X^2_{\nu,\lambda}$ denote a noncentral chi-squared random variable with df $= \nu$ and noncentrality λ. A fundamental result for chi-squared analyses is that, for fixed λ,

$$P[X^2_{\nu,\lambda} > \chi^2_\nu(\alpha)] \text{ increases as } \nu \text{ decreases.}$$

That is, the power for rejecting H_0 at a fixed α-level increases as the df of the test decreases (Das Gupta and Perlman 1974). For fixed ν, the power equals α when $\lambda = 0$, and it increases as λ increases. The inverse relation between power and df suggests that focusing the noncentrality on a statistic having a small df value can improve power.

Suppose that an explanatory variable has, at least approximately, a linear effect on $\text{logit}[P(Y = 1)]$. To test independence with reasonable power, it is then sensible to use a statistic based on the linear logit model, using the likelihood-ratio statistic $G^2(I|L)$, the Wald statistic $z = \hat{\beta}/SE$, and the Cochran–Armitage (score) statistic. When is $G^2(I|L)$ more powerful than $G^2(I)$? The statistics satisfy

$$G^2(I) = G^2(I|L) + G^2(L),$$

where $G^2(L)$ tests goodness of fit of the linear logit model. When the linear logit model holds, $G^2(L)$ has an asymptotic chi-squared distribution with df $= I - 2$; then if $\beta \neq 0$, $G^2(I)$ and $G^2(I|L)$ both have approximate noncentral chi-squared distributions with the same noncentrality. Whereas df $= I - 1$ for $G^2(I)$, df $= 1$ for $G^2(I|L)$. Thus, $G^2(I|L)$ is more powerful, because it uses fewer degrees of freedom.

When the linear logit model does not hold, $G^2(I)$ has greater noncentrality than $G^2(I|L)$, the discrepancy increasing as the model fits more poorly. However, when the model approximates reality fairly well, usually $G^2(I|L)$ is still more powerful. That test's df value of 1 more than compensates for its loss in noncentrality. The closer the true relationship is to the linear logit, the more nearly $G^2(I|L)$ captures the same noncentrality as $G^2(I)$, and the

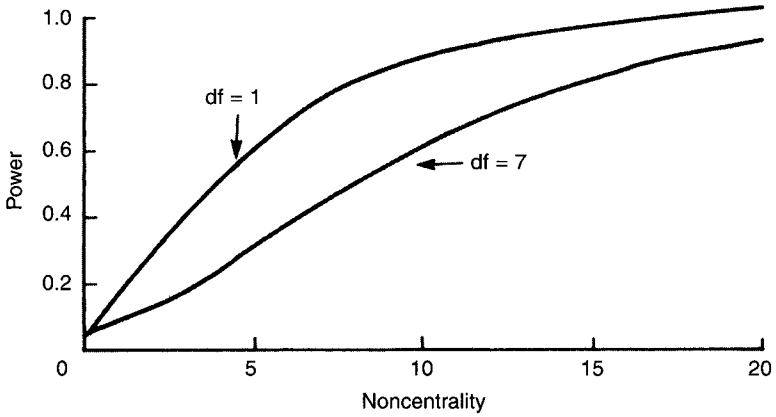

Figure 5.5 Power and noncentrality, for df = 1 and df = 7, when $\alpha = 0.05$.

more powerful it is compared with $G^2(I)$. To illustrate, Figure 5.5 plots power as a function of noncentrality when df = 1 and 7. When the noncentrality of a test having df = 1 is at least about half that of a test having df = 7, the test with df = 1 is more powerful. The linear logit model then helps detect a key component of an association. As Mantel (1963) argued in a similar context, "that a linear regression is being tested does not mean that an assumption of linearity is being made. Rather it is that test of a linear component of regression provides power for detecting any progressive association which may exist."

The improved power for the linear trend statistic results from sacrificing power in other cases. The $G^2(I)$ test can have greater power than $G^2(I|L)$ when the linear logit model describes the true relationship very poorly.

5.3.9 Example: Skin Damage and Leprosy

Table 5.5 refers to an experiment on the use of sulfones and streptomycin drugs in the treatment of leprosy. The degree of infiltration at the start of the experiment measures a type of skin damage. The response is the change in the overall clinical condition of the patient after 48 weeks of treatment. We use response scores (−1, 0, 1, 2, 3). The question of interest is whether subjects with high infiltration changed differently from those with low infiltration.

Table 5.5 Change in Clinical Condition by Degree of Infiltration

	Degree of Infiltration		Proportion
Clinical Change	High	Low	High
Worse	1	11	0.08
Stationary	13	53	0.20
Slight improvement	16	42	0.28
Moderate improvement	15	27	0.36
Marked improvement	7	11	0.39

Source: Reprinted with permission from the Biometric Society (Cochran 1954).

The test $G^2(I) = 7.28$ (df = 4) does not show much evidence of association ($P = 0.12$), but it ignores that the clinical change response variable is ordinal. It seems natural to compare the mean change for the two infiltration levels. Cochran (1954) and Yates (1948) noted that this analysis is identical to a trend test treating the binary variable as the response. In fact, the sample proportion of high infiltration increases monotonically as the clinical change improves. The test of H_0: $\beta = 0$ in the linear logit model has $G^2(I|L) = 6.65$, with df = 1 ($P = 0.01$). It gives strong evidence of more positive clinical change at the higher level of infiltration. Using the ordering by decreasing df from 4 to 1 pays a strong dividend. In addition, $G^2(L) = 0.63$ with df = 3 suggests that the linear trend model fits well.

5.3.10 Model Smoothing Improves Precision of Estimation

Using directed alternatives can improve not only *test power*, but also *estimation* of cell probabilities and summary measures. In generic form, let $\boldsymbol{\pi}$ be true cell probabilities in a contingency table, let $\boldsymbol{p}$ denote sample proportions, and let $\hat{\boldsymbol{\pi}}$ denote model-based ML estimates of $\boldsymbol{\pi}$.

When $\boldsymbol{\pi}$ satisfies a certain model, both $\hat{\boldsymbol{\pi}}$ for that model and $\boldsymbol{p}$ are consistent estimators of $\boldsymbol{\pi}$. The model-based estimator $\hat{\boldsymbol{\pi}}$ is better, as its true asymptotic standard error cannot exceed that of $\boldsymbol{p}$. This happens because of model parsimony: The unsaturated model, on which $\hat{\boldsymbol{\pi}}$ is based, has fewer parameters than the saturated model, on which $\boldsymbol{p}$ is based. In fact, model-based estimators are also more efficient in estimating functions $g(\boldsymbol{\pi})$ of cell probabilities. For any differentiable function g,

$$\text{asymp. var}[\sqrt{n}g(\hat{\boldsymbol{\pi}})] \leq \text{asymp. var}[\sqrt{n}g(\boldsymbol{p})].$$

In Section 16.2.3 we show formulas. The result holds more generally than for categorical data models (Altham 1984), a reason that statisticians prefer parsimonious models.

In reality, of course, a chosen model is unlikely to hold exactly. However, when the model approximates $\boldsymbol{\pi}$ well, unless n is extremely large, $\hat{\boldsymbol{\pi}}$ is still better than $\boldsymbol{p}$. Although $\hat{\pi}_i$ is biased, it has smaller variance than p_i, and $\text{MSE}(\hat{\pi}_i) < \text{MSE}(p_i)$ when its variance plus squared bias is smaller than $\text{var}(p_i)$. In Section 3.3.8, for example, we showed that independence-model estimates of cell probabilities in two-way tables can be much better than sample proportions even when that model does not hold.

5.4 MULTIPLE LOGISTIC REGRESSION

Like ordinary regression, logistic regression extends to models with multiple explanatory variables, which can be a mixture of quantitative and qualitative (Cox 1958). The model for $\pi(\boldsymbol{x}) = P(Y = 1)$ at values $\boldsymbol{x} = (x_1, \ldots, x_p)$ of p predictors is

$$\text{logit}[\pi(\boldsymbol{x})] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p. \tag{5.8}$$

The alternative formula, directly specifying $\pi(\boldsymbol{x})$, is

$$\pi(\boldsymbol{x}) = \frac{\exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}{1 + \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}. \tag{5.9}$$

For qualitative predictors, we use indicator variables for its categories.

The parameter β_j refers to the effect of x_j on the log odds that $Y = 1$, adjusting for the other x_k. For instance, $\exp(\beta_j)$ is the multiplicative effect on the odds of a 1-unit increase in x_j, when we can keep fixed the levels of other x_k.

5.4.1 Logistic Models for Multiway Contingency Tables

When all variables are categorical, a multiway contingency table displays the data. We illustrate ideas with binary predictors X and Z. We treat the sample size at given combinations of X and Z as fixed and regard the two counts on Y at each setting as binomial, with different binomials treated as independent. We let indicator variables x and z take value 1 in the first category and 0 in the second. The model

$$\text{logit}[P(Y = 1)] = \alpha + \beta_1 x + \beta_2 z \tag{5.10}$$

has main effects for X and Z but assumes an absence of interaction. The effect of one factor is the same at each level of the other.

At a fixed level of Z, the effect on the logit of changing categories of X is

$$[\alpha + \beta_1(1) + \beta_2 z] - [\alpha + \beta_1(0) + \beta_2 z] = \beta_1. \tag{5.11}$$

This logit difference equals the difference of log odds, which is the log odds ratio between X and Y, fixing Z. Thus, $\exp(\beta_1)$ is the conditional odds ratio between X and Y. Adjusting for Z, the odds of success when $x = 1$ equal $\exp(\beta_1)$ times the odds when $x = 0$. This conditional odds ratio is the same at each level of z; that is, there is *homogeneous XY association* (Section 2.3.5). The lack of an interaction term implies a common odds ratio for the partial tables. When $\beta_1 = 0$, that common odds ratio equals 1. Then X and Y are independent in each partial table, or *conditionally independent, given Z* (Section 2.3.4).

Additivity on the logit scale is the usual definition of no interaction for categorical variables. However, it could instead be defined as additivity on some other scale, such as with probit or identity link. Interaction can occur on one scale when there is none on another scale. In some applications, a particular definition may be natural. For instance, theory might assume an underlying normal distribution for Y and predict that the probit is an additive function of predictor effects.

A factor with I categories needs $I - 1$ indicator variables. With I categories for X and K categories for Z, model (5.10) extends to

$$\text{logit}[P(Y = 1)] = \alpha + \beta_1^X x_1 + \cdots + \beta_{I-1}^X x_{I-1} + \beta_1^Z z_1 + \cdots + \beta_{K-1}^Z x_{K-1},$$

where, for example, $z_k = 1$ for observations in category k of Z and $z_k = 0$ otherwise, $k = 1, \ldots, K - 1$. This equation represents effects of X with parameters $\{\beta_i^X\}$ and effects of Z with parameters $\{\beta_k^Z\}$. The X and Z superscripts are merely labels and do not represent powers. This model form applies for any number of categories for X and Z. The parameter β_k^Z, for example, denotes the effect on the logit of classification in category k of Z instead of category K.

An alternative representation of such factors resembles the way that ANOVA factorial models often express them. The equivalent model formula is

$$\text{logit}[P(Y = 1)] = \alpha + \beta_i^X + \beta_k^Z. \tag{5.12}$$

For each factor, one parameter is redundant. Fixing one at 0, such as $\beta_I^X = \beta_K^Z = 0$, represents the category not having its own indicator variable. Conditional independence between X and Y, given Z, corresponds to $\beta_1^X = \beta_2^X = \cdots = \beta_I^X$, whereby $P(Y = 1)$ does not change as i changes, for fixed k.

5.4.2 Example: AIDS and AZT Use

Table 5.6 is from a study 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 HIV were randomly assigned either to receive AZT immediately or to wait until their T cells showed severe immune weakness. Table 5.6 cross-classifies the veterans' race, whether they received AZT immediately, and whether they developed AIDS symptoms during the 3-year study.

In model (5.10), we identify X with AZT treatment ($x = 1$ for immediate AZT use, $x = 0$ otherwise) and Z with race ($z = 1$ for whites, $z = 0$ for blacks), for predicting the probability that AIDS symptoms developed. Thus, α is the log odds of developing AIDS symptoms for black subjects without immediate AZT use, β_1 is the increment to the log odds for those with immediate AZT use, and β_2 is the increment to the log odds for white subjects. Table 5.7 shows output. The estimated odds ratio between immediate AZT use and development of AIDS symptoms equals $\exp(-0.7195) = 0.487$. For each race, the estimated odds of symptoms are half as high for those who took AZT immediately. The Wald confidence interval for this effect is $\exp[-0.720 \pm 1.96(0.279)] = (0.28, 0.84)$. Similar results occur for the likelihood-based interval, as shown.

The hypothesis of conditional independence of AZT treatment and development of AIDS symptoms, controlling for race, is H_0: $\beta_1 = 0$ in (5.10). The likelihood-ratio statistic comparing the model with the simpler model having $\beta_1 = 0$ equals 6.87 (df = 1), showing evidence of association ($P = 0.01$). The Wald statistic $(\hat{\beta}_1/SE)^2 = (-0.7195/0.279)^2 = 6.65$, shown in the output, provides similar results.

Table 5.6 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

Source: The New York Times, Feb. 15, 1991.

Table 5.7 Software Output (Based on SAS) for Logistic Model with AIDS Symptoms Data

```
       Goodness-of-Fit Statistics
Criterion      DF       Value      Pr > ChiSq
Deviance        1      1.3835        0.2395
Pearson         1      1.3910        0.2382

              Analysis of Maximum Likelihood Estimates
Parameter      Estimate      Std Error      Wald Chi-Sq      Pr > ChiSq
Intercept       −1.0736        0.2629          16.6705          <.0001
azt             −0.7195        0.2790           6.6507          0.0099
race             0.0555        0.2886           0.0370          0.8476
Obs      race      azt      y       n       pi_hat       lower       upper
 1         1        1      14     107      0.14962     0.09897     0.21987
 2         1        0      32     113      0.26540     0.19668     0.34774
 3         0        1      11      63      0.14270     0.08704     0.22519
 4         0        0      12      55      0.25472     0.16953     0.36396

Profile Like. CI for Odds Ratios
Effect      Estimate       95% Conf Limits
azt            0.487       0.279       0.835
race           1.057       0.605       1.884
```

Table 5.8 shows parameter estimates for three ways of defining factor parameters in (5.12): (1) setting the last parameter equal to 0, (2) setting the first parameter equal to 0, and (3) having parameters sum to zero. This corresponds to setting up indicator variables for each category except the last in scheme (1), for each category except the first in scheme (2). In scheme (3), there is also a reference category, and for other categories the indicator is 1 for an observation in the category, -1 for an observation in the reference category, and 0 otherwise. For each coding scheme, at a given combination of AZT use and race, the estimated probability of developing AIDS symptoms is the same. For instance, the intercept estimate plus the estimate for immediate AZT use plus the estimate for being white is -1.738 for each scheme, so the estimated probability that white veterans with immediate AZT use develop AIDS symptoms equals $\exp(-1.738)/[1 + \exp(-1.738)] = 0.15$. The bottom of Table 5.7 shows point and interval estimates of the probabilities. Figure 5.6

Table 5.8 Parameter Estimates for Logistic Model Fitted to Table 5.6 on AIDS and AZT Use

	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
No	0.000	0.720	0.360
Race White	0.055	0.000	0.028
Black	0.000	−0.055	−0.028

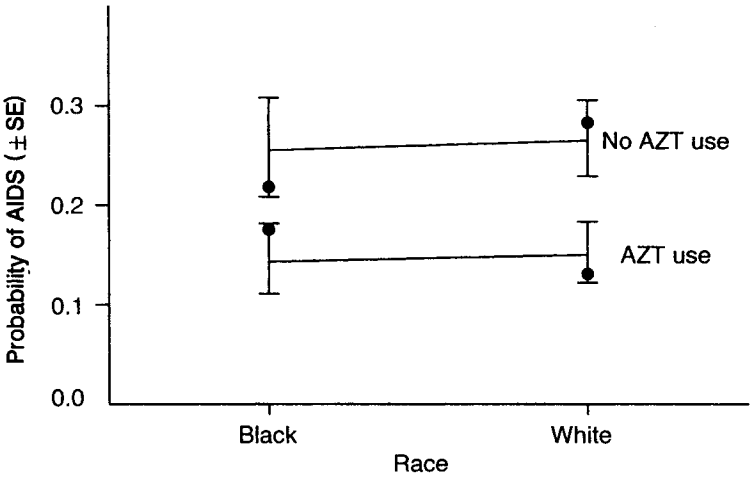

Figure 5.6 Estimated effects of AZT use and race on probability of developing AIDS symptoms (dots are sample proportions).

shows a graphical representation of the sample proportions (the four dots) and the point estimates plus and minus a standard error.

For each coding scheme, $\beta_1^X - \beta_2^X$ is identical and represents the conditional log odds ratio of X with the response, given Z. Here, $\exp(\hat{\beta}_1^X - \hat{\beta}_2^X) = \exp(-0.720) = 0.49$ estimates the common odds ratio between immediate AZT use and AIDS symptoms, for each race.

5.4.3 Goodness of Fit as a Likelihood-Ratio Test

The likelihood-ratio statistic $G^2(M_0|M_1) = -2(L_0 - L_1)$ tests whether certain model parameters are zero, given that M_1 holds, by comparing the log likelihood L_1 for the fitted model M_1 with L_0 for a simpler model M_0. The goodness-of-fit statistic $G^2(M)$ is a special case in which $M_0 = M$ and M_1 is the saturated model. In testing whether M fits, we test whether *all* parameters in the saturated model but not in M equal zero. The asymptotic df is the difference in the number of parameters in the two models, which is the number of binomials modeled minus the number of parameters in M.

We illustrate by checking the fit of model (5.10) for the AIDS data. For its fit, white veterans with immediate AZT use had estimated probability 0.150 of developing AIDS symptoms during the study. Since 107 white veterans took AZT, the fitted value is $107(0.150) = 16.0$ for developing symptoms and $107(0.850) = 91.0$ for not developing them. Similarly, we can obtain fitted values for all eight cells in Table 5.6. The goodness-of-fit statistics comparing these with the cell counts are $G^2 = 1.38$ and $X^2 = 1.39$. The model has four binomials, one at each combination of AZT use and race. Since it has three parameters, residual df $= 4 - 3 = 1$. The small G^2 and X^2 values suggest that the model fits decently ($P > 0.2$).

For model (5.10), the odds ratio between X and Y is the same at each level of Z. The goodness-of-fit test checks this structure. That is, the test also provides a test of homogeneous odds ratios. For Table 5.6, homogeneity is plausible. Since residual df $= 1$, the more complex model that adds an interaction term and permits the two odds ratios to differ is saturated.

5.4.4 Model Comparison by Comparing Deviances

Let L_S denote the maximized log likelihood for the saturated model. As discussed in Section 4.5.4, the likelihood-ratio statistic for comparing models M_1 and M_0 is

$$G^2(M_0|M_1) = -2(L_0 - L_1) = -2(L_0 - L_S) - [-2(L_1 - L_S)] = G^2(M_0) - G^2(M_1).$$

The test statistic comparing two models is identical to the difference in G^2 goodness-of-fit statistics (deviances) for the two models. To illustrate, consider H_0: $\beta_2 = 0$ for the race effect with the AIDS data. The likelihood-ratio statistic equals 0.04, suggesting that the simpler model is adequate. But this equals $G^2(M_0) - G^2(M_1) = 1.42 - 1.38$, where M_0 is the simpler model with $\beta_2 = 0$.

The model comparison statistic often has an approximate chi-squared null distribution even when separate $G^2(M_i)$ do not. For instance, when at least one predictor is continuous or a contingency table has very small fitted values, the sampling distribution of $G^2(M_i)$ may be far from chi-squared. Nonetheless, if df for the comparison statistic is modest (as in comparing two models that differ by at most a few parameters), the null distribution of $G^2(M_0|M_1)$ is approximately chi-squared.

5.4.5 Example: Horseshoe Crab Satellites Revisited

For the horseshoe crab data, we next use both the female crab's carapace width and color as predictors of Y = whether the crab has at least one satellite (1 = yes, 0 = no). Color has five categories: light, medium light, medium, medium dark, dark. It 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 as qualitative. The four categories use three indicator variables. The model for the probability that the crab has at least one satellite is

$$\text{logit}[P(Y = 1)] = \alpha + \beta_1 c_1 + \beta_2 c_2 + \beta_3 c_3 + \beta_4 x, \tag{5.13}$$

where x = width in centimeters, 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 $c_1 = c_2 = c_3 = 0$. Table 5.9 shows the ML parameter estimates. For instance, for dark crabs, $\text{logit}[\hat{P}(Y = 1)] = -12.715 + 0.468x$; by contrast, for medium-light crabs, $c_1 = 1$, and $\text{logit}[\hat{P}(Y = 1)] = (-12.715 + 1.330) + 0.468x = -11.385 + 0.468x$. At the average width of 26.3 cm, $\hat{P}(Y = 1) = 0.399$ for dark crabs and 0.715 for medium-light crabs. The exponentiated difference between two color parameter estimates is an odds ratio comparing those colors. For instance, the difference for 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. At width $x = 26.3$, the odds equal $0.715/0.285 = 2.51$ for a medium-light crab and $0.399/0.601 = 0.66$ for a dark crab, for which $2.51/0.66 = 3.8$.

Table 5.9 Software Output (Based on SAS) for Model with Width and Color Predictors of Whether Horseshoe Crab Has Satellites

Criteria For Assessing Goodness Of Fit		
Criterion	DF	Value
Deviance	168	187.4570
Pearson Chi-Square	168	168.6590
Log Likelihood		−93.7285

Parameter	Estimate	Standard Error	Likelihood-Ratio 95% Confidence	Limits	Chi-Square	Pr>ChiSq
intercept	−12.7151	2.7618	−18.4564	−7.5788	21.20	<.0001
c1	1.3299	0.8525	−0.2738	3.1354	2.43	0.1188
c2	1.4023	0.5484	0.3527	2.5260	6.54	0.0106
c3	1.1061	0.5921	−0.0279	2.3138	3.49	0.0617
width	0.4680	0.1055	0.2713	0.6870	19.66	<.0001

To test whether color contributes significantly to model (5.13), we test H_0: $\beta_1 = \beta_2 = \beta_3 = 0$. This 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 (5.13) to L_0 for the simpler model. The test statistic $-2(L_0 - L_1) = 7.0$ has df = 3, the difference between the numbers of parameters in the two models. The chi-squared P-value of 0.07 provides slight evidence of a color effect.

The model assumes a lack of interaction between color and width in their effects. Width has the coefficient of 0.468 for all colors, so the shapes of the curves relating width to $P(Y = 1)$ are identical. Figure 5.7 displays the fitted model. Any one curve equals any

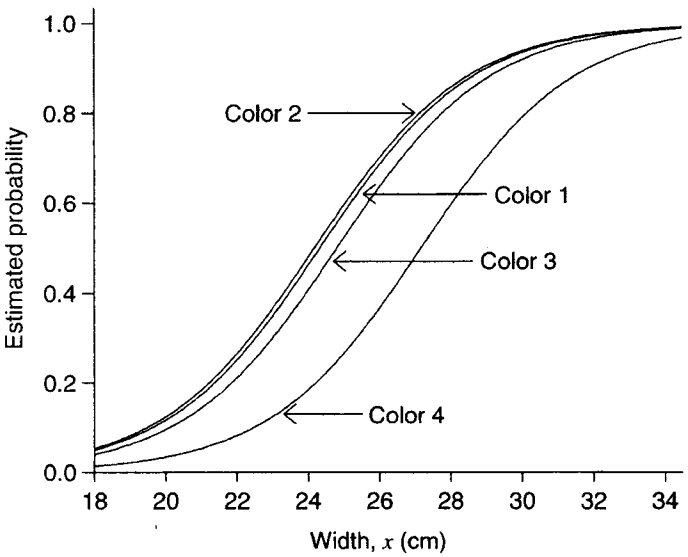

Figure 5.7 Logistic regression model using additive width and color predictors of whether horseshoe crab has satellites.

other curve shifted to the right or left. The parallelism of curves in the horizontal dimension implies that any two curves never cross. At all width values, color 4 (dark) has a lower estimated probability of a satellite than the other colors. There is a noticeable positive effect of width.

The more complex model allowing color × width interaction has three additional terms, the cross-products of width with the color indicator variables. Fitting this model is equivalent to fitting logistic regression with width predictor separately for crabs of each color. Each color then has a different-shaped curve relating width to $P(Y = 1)$, so a comparison of two colors varies according to the width value. The likelihood-ratio statistic comparing the models with and without the interaction terms equals 4.4, with df = 3. The evidence of interaction is weak ($P = 0.22$).

5.4.6 Quantitative Treatment of Ordinal Predictor

Color has ordered categories, from lightest to darkest. A simpler model yet treats this predictor as quantitative. Color may have a linear effect, for a set of monotone scores. To illustrate, for scores $c = (1, 2, 3, 4)$ for the color categories, the model

$$\text{logit}[P(Y = 1)] = \alpha + \beta_1 c + \beta_2 x \tag{5.14}$$

has $\hat{\alpha} = -10.071$, $\hat{\beta}_1 = -0.509$ ($SE = 0.224$) and $\hat{\beta}_2 = 0.458$ ($SE = 0.104$). This shows 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 $\exp(-0.509) = 0.60$.

The likelihood-ratio statistic comparing this fit to the more complex model (5.13) having a separate parameter for each color equals 1.66 (df = 2). This statistic tests that the simpler model (5.14) is adequate, given that model (5.13) holds. It tests that when plotted against the color scores, the color parameters in (5.13) follow a linear trend. The simplification seems permissible ($P = 0.44$).

The color parameter estimates in the qualitative-color model (5.13) are (1.33, 1.40, 1.11, 0), the 0 value for the dark category reflecting its lack of an indicator variable. Although these values do not depart significantly from a linear trend, the first three are quite similar compared with the last one. Thus, another potential color scoring for model (5.14) is (1, 1, 1, 0); that is, score = 0 for dark-colored crabs, and score = 1 otherwise. The likelihood-ratio statistic comparing model (5.14) with these binary scores to model (5.13) equals 0.50 (df = 2), showing that this simpler model is also adequate. Its fit is

$$\text{logit}[\hat{P}(Y = 1)] = -12.980 + 1.300c + 0.478x, \tag{5.15}$$

with standard errors 0.526 and 0.104. 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 quantitative-color model with 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. A much larger sample is needed to determine which color scoring is most appropriate. With moderate-sized samples, it's not unusual for quite different models to be consistent with the data.

5.4.7 Probability-Based and Standardized Interpretations

Although it is natural to interpret logistic regression model parameters as effects on a log odds, some find it difficult to understand odds or odds ratio effects. The simpler interpretation using the instantaneous rate of change in the probability (Section 5.1.1) applies also to multiple predictors. Consider a setting of predictors at which $\hat{P}(Y = 1) = \hat{\pi}$. Then, adjusting for the other predictors, as a function of a quantitative predictor x_j, $\hat{\pi}$ has instantaneous rate of change of $\hat{\beta}_j\hat{\pi}(1 - \hat{\pi})$. For instance, at predictor settings at which $\hat{\pi} = 0.50$ for fit (5.15), the approximate effect of a 1-cm increase in width is $(0.478)(0.50)(0.50) = 0.12$. This is considerable, since a 1-cm change in width is less than half a standard deviation.

We could summarize the effect of x_j on the probability scale by averaging the instantaneous rates for the sample. Let x_{ij} denote the value of x_j for subject i and let $\hat{\pi}(x_{i1}, \ldots, x_{ip})$ denote the estimate of $P(Y = 1)$ at the explanatory variable values for subject i. This summary is

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_j\hat{\pi}(x_{i1}, \ldots, x_{ip})[1 - \hat{\pi}(x_{i1}, \ldots, x_{ip})].$$

Alternatively, to describe the effect of x_j in a simpler manner not depending on its units, we could set the other predictors at their sample means and compute the estimated probabilities at the smallest and largest x_j values. These are sensitive to outliers, however, so we could instead use the upper and lower quartiles of x_j. For the fit (5.15) with binary color, the sample means are 26.3 for x and 0.873 for c. The lower and upper quartiles of x are 24.9 and 27.7. At $x = 24.9$ and $c = \bar{c}$, $\hat{\pi} = 0.51$. At $x = 27.7$ and $c = \bar{c}$, $\hat{\pi} = 0.80$. The change in $\hat{\pi}$ from 0.51 to 0.80 over the middle 50% of the range of width values reflects a strong width effect. Since c takes only values 0 and 1, we could instead report this effect separately for each. Also, when an explanatory variable is an indicator, it makes sense to report the estimated probabilities at its two values rather than at quartiles, which could be identical. At $\bar{x} = 26.3$, $\hat{\pi} = 0.40$ when $c = 0$ and $\hat{\pi} = 0.71$ when $c = 1$. This color effect, differentiating dark crabs from others, is also substantial.

Table 5.10 summarizes the logistic parameter estimates and some probability comparison effects. It also shows results of the extension of model (5.15), permitting interaction. The

Table 5.10 Summary of Effects in Model (5.15) with Crab Width and Color (Treated as Binary) as Predictors of Presence of Satellites

Variable	Estimate	*SE*	Comparison	Change in Probability
No interaction model				
Intercept	−12.980	2.727		
Color (0 = dark, 1 = other)	1.300	0.526	(1, 0) at $\bar{x}$	$0.31 = 0.71 - 0.40$
Width, x (cm)	0.478	0.104	(UQ, LQ) at $\bar{c}$	$0.29 = 0.80 - 0.51$
Interaction model				
Intercept	−5.854	6.694		
Color (0 = dark, 1 = other)	−6.958	7.318		
Width, x (cm)	0.200	0.262	(UQ, LQ) at $c = 0$	$0.13 = 0.43 - 0.30$
Width × color	0.322	0.286	(UQ, LQ) at $c = 1$	$0.29 = 0.84 - 0.55$

estimated width effect is then greater for the lighter-colored crabs. However, the interaction is not significant.

To compare effects of quantitative predictors having different units, it can also be helpful to report standardized coefficients. One approach fits the model to standardized predictors, replacing each x_j by $(x_j - \bar{x}_j)/s_{x_j}$. Then, each regression coefficient represents the effect of a standard deviation change in a predictor, adjusting for the other variables. Equivalently, for each j the standardized coefficient results from multiplying the unstandardized estimate $\hat{\beta}_j$ by s_{x_j}. For example, for fit (5.15) with binary color, the standard deviation of width is 2.109 cm. The standardized estimate for the effect of width for that model is $0.478(2.109) = 1.01$. When we replace width by weight (with standard deviation 0.577 kg) in the model, the unstandardized estimate 1.729 corresponds to the standardized estimate $1.729(0.577) = 1.00$. The unstandardized estimates 0.478 and 1.729 are quite different, but width and weight (standardized) have similar effects, conditional on whether or not a crab is dark.

Since the standard logistic cdf has standard deviation $\pi/\sqrt{3}$, some software (e.g., PROC LOGISTIC in SAS) defines a standardized estimate by multiplying the unstandardized estimate by $s_{x_j}\sqrt{3}/\pi$. Such a standardized estimate represents the effect on the location of an underlying latent response variable (in standard deviations units) for a standard deviation change in a predictor, adjusting for the other variables. For example, for fit (5.15) with binary color, this standardized estimate for the effect of width is $0.478(2.109)\sqrt{3}/\pi = 0.556$. A standard deviation change in width, conditional on a color, corresponds to a 0.556 standard deviation shift upwards in the distribution of the latent logistic response variable.

5.4.8 Estimating an Average Causal Effect

In many applications the explanatory variable of primary interest specifies two groups to be compared while adjusting for the other explanatory variables in the model. Let $j = 1$ identify this binary group variable, with the groups denoted by $x_1 = 0$ and $x_1 = 1$. For the logistic regression model, an alternative to the log odds ratio $\hat{\beta}_1$ as an effect summary is the estimated *average causal effect*,

$$\frac{1}{n}\sum_i \left[\hat{\pi}(x_{i1} = 1, x_{i2}, \ldots x_{ip}) - \hat{\pi}(x_{i1} = 0, x_{i2}, \ldots x_{ip})\right].$$

For each observation i, we find the fitted probability for the given values of $x_{i2}, \ldots x_{ip}$ (1) if that observation were in group 1 and (2) if that observation were in group 0, and average the differences among all n observations. This estimates the difference between the overall proportions of "successes" if all subjects in the study were in group 1 compared with all being in group 0. It is usually not adequate to use a linear probability model (i.e., identity link function) for the full data set, by which such a difference would be constant across subjects, but nonetheless this is a useful summary for cases in which this difference is relatively stable.

We illustrate using Table 5.6 from the randomized study of AZT use and AIDS. In Section 5.4.2 we summarized the effect of AZT use by the estimated conditional odds ratio of $\exp(\hat{\beta}_1) = 0.487$. Alternatively, from the probability estimates shown in Table 5.7, the difference between those not receiving AZT and those receiving AZT in the estimated proportion developing AIDS symptoms was $0.2654 - 0.1496 = 0.1158$ for whites and

$0.2547 - 0.1427 = 0.1120$ for blacks. Weighted by the sample sizes of whites and blacks, the estimated average causal effect is $(220/338)(0.1158) + (118/338)(0.1120) = 0.1145$. In fact, this is similar to the ML estimate of $\hat{\beta}_1 = 0.1152$ for the corresponding linear probability model.

For categorical predictors, Copas and Eguchi (2010) showed how to obtain a standard error for an estimated average causal effect that applies for the logistic model. They also presented a nonparametric standard error for an estimate that, instead of being model-based, is a weighted average of the differences of the sample proportions at the various levels of the explanatory variables. The main theme of their article, however, was adjusting inferences for the fact that many models may be consistent with the data. The average causal effect is often a relevant measure regardless of the form of the true relationship.

Estimating an average causal effect is natural for experimental studies. It has also received much attention for nonrandomized studies since the fundamental article by Rubin (1974) and later work using methods to adjust for different propensities of a subject to be in one group or the other (e.g., see Section 6.4.11).

5.5 FITTING LOGISTIC REGRESSION MODELS

The mechanics of ML estimation and model fitting for logistic regression are special cases of the GLM fitting results of Section 4.6. With n subjects, we treat the n binary responses as independent. Let $\boldsymbol{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ip})$ denote setting i of the values of p explanatory variables and a coefficient $x_{i0} = 1$ for an intercept term, $i = 1, \ldots, N$. When explanatory variables are continuous, a different setting may occur for each subject, in which case $N = n$. This also happens when the data file consists of ungrouped data. The logistic regression model (5.8), treating the intercept α as a regression parameter β_0 for an explanatory variable that always equals 1, is

$$\pi(\boldsymbol{x}_i) = \frac{\exp\left(\sum_{j=0}^{p} \beta_j x_{ij}\right)}{1 + \exp\left(\sum_{j=0}^{p} \beta_j x_{ij}\right)}. \tag{5.16}$$

5.5.1 Likelihood Equations for Logistic Regression

When more than one observation occurs at a fixed $\boldsymbol{x}_i$ value, it is sufficient to record the number of observations n_i and the number of successes. We then let y_i refer to this success count rather than to an individual binary response. Then $\{Y_1, \ldots, Y_N\}$ are independent binomials with $E(Y_i) = n_i \pi(\boldsymbol{x}_i)$, where $n_1 + \cdots + n_N = n$. Their joint probability mass function is proportional to the product of N binomial functions,

$$\begin{aligned}
&\prod_{i=1}^{N} \pi(\boldsymbol{x}_i)^{y_i}[1 - \pi(\boldsymbol{x}_i)]^{n_i - y_i} \\
&= \left\{\prod_{i=1}^{N} \exp\left[\log\left(\frac{\pi(\boldsymbol{x}_i)}{1 - \pi(\boldsymbol{x}_i)}\right)^{y_i}\right]\right\}\left\{\prod_{i=1}^{N}[1 - \pi(\boldsymbol{x}_i)]^{n_i}\right\} \\
&= \left\{\exp\left[\sum_{i=1}^{N} y_i \log\frac{\pi(\boldsymbol{x}_i)}{1 - \pi(\boldsymbol{x}_i)}\right]\right\}\left\{\prod_{i=1}^{N}[1 - \pi(\boldsymbol{x}_i)]^{n_i}\right\}.
\end{aligned}$$

For model (5.16), the ith logit is $\sum_j \beta_j x_{ij}$, so the exponential term here equals $\exp\left[\sum_i y_i\left(\sum_j \beta_j x_{ij}\right)\right] = \exp\left[\sum_j \left(\sum_i y_i x_{ij}\right)\beta_j\right]$. Also, since $[1 - \pi(\boldsymbol{x}_i)] = \left[1 + \exp\left(\sum_j \beta_j x_{ij}\right)\right]^{-1}$, the log likelihood equals

$$L(\boldsymbol{\beta}) = \sum_j \left(\sum_i y_i x_{ij}\right)\beta_j - \sum_i n_i \log\left[1 + \exp\left(\sum_j \beta_j x_{ij}\right)\right]. \tag{5.17}$$

This depends on the binomial counts only through the sufficient statistics for the model parameters, $\left\{\sum_i y_i x_{ij}\right\}$, $j = 0, 1, \ldots, p$.

The likelihood equations result from setting $\partial L(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = \mathbf{0}$. Since

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_i y_i x_{ij} - \sum_i n_i x_{ij} \frac{\exp\left(\sum_k \beta_k x_{ik}\right)}{1 + \exp\left(\sum_k \beta_k x_{ik}\right)},$$

the likelihood equations are

$$\sum_i y_i x_{ij} - \sum_i n_i \hat{\pi}_i x_{ij} = 0, \quad j = 0, 1, \ldots, p, \tag{5.18}$$

where $\hat{\pi}_i = \exp\left(\sum_k \hat{\beta}_k x_{ik}\right)/\left[1 + \exp\left(\sum_k \hat{\beta}_k x_{ik}\right)\right]$ is the ML estimate of $\pi(\boldsymbol{x}_i)$. We observed these equations as a special case of those for binomial GLMs in (4.28) (but there y_i is the *proportion* of successes). The equations are nonlinear and require iterative solution.

Let $\boldsymbol{X}$ denote the matrix of values of $\{x_{ij}\}$, with N rows for the binomial observations and a column for each parameter. The likelihood equations (5.18) have form

$$\boldsymbol{X}^T \boldsymbol{y} = \boldsymbol{X}^T \hat{\boldsymbol{\mu}}, \tag{5.19}$$

where $\hat{\mu}_i = n_i \hat{\pi}_i$. This equation illustrates the fundamental result for GLMs with canonical link, shown in equation (4.51), that the likelihood equations equate the sufficient statistics to their expected values.

5.5.2 Asymptotic Covariance Matrix of Parameter Estimators

The ML estimators $\hat{\boldsymbol{\beta}}$ have a large-sample normal distribution with covariance matrix equal to the inverse of the information matrix. The observed information matrix has elements

$$-\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_a \partial\beta_b} = \sum_i \frac{x_{ia}x_{ib}n_i \exp\left(\sum_j \beta_j x_{ij}\right)}{\left[1 + \exp\left(\sum_j \beta_j x_{ij}\right)\right]^2} = \sum_i x_{ia}x_{ib}n_i\pi_i(1 - \pi_i). \tag{5.20}$$

This is not a function of $\{y_i\}$, so the observed and expected information are identical. This happens for all GLMs that use canonical links (Section 4.6.5).

The estimated covariance matrix is the inverse of the matrix having elements (5.20), substituting $\hat{\boldsymbol{\beta}}$. This has form

$$\widehat{\text{cov}}(\hat{\boldsymbol{\beta}}) = \{\boldsymbol{X}^T \mathbf{Diag}[n_i\hat{\pi}_i(1 - \hat{\pi}_i)]\boldsymbol{X}\}^{-1}, \tag{5.21}$$

where $\mathbf{Diag}[n_i\hat{\pi}_i(1-\hat{\pi}_i)]$ denotes the $N \times N$ diagonal matrix having $\{n_i\hat{\pi}_i(1-\hat{\pi}_i)\}$ on the main diagonal. This is the special case of the GLM covariance matrix (4.31) with estimated diagonal weight matrix $\hat{\boldsymbol{W}}$ having elements $\hat{w}_i = n_i\hat{\pi}_i(1-\hat{\pi}_i)$. The square roots of the main diagonal elements of (5.21) are estimated standard errors of $\hat{\boldsymbol{\beta}}$.

5.5.3 Distribution of Probability Estimators

Using $\widehat{\text{cov}}(\hat{\boldsymbol{\beta}})$, we can conduct Wald inference about $\boldsymbol{\beta}$ and related effects such as odds ratios. We can also construct confidence intervals for response probabilities $\pi(\boldsymbol{x})$ at particular settings $\boldsymbol{x}^T = (x_0, x_1, \ldots, x_p)$.

The estimated variance of $\text{logit}[\hat{\pi}(\boldsymbol{x})] = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ is $\boldsymbol{x}^T\widehat{\text{cov}}(\hat{\boldsymbol{\beta}})\boldsymbol{x}$. For large samples, $\text{logit}[\hat{\pi}(\boldsymbol{x})] \pm z_{\alpha/2}\sqrt{\boldsymbol{x}^T\widehat{\text{cov}}(\hat{\boldsymbol{\beta}})\boldsymbol{x}}$ is a confidence interval for the true logit. The endpoints invert to a corresponding interval for $\pi(\boldsymbol{x})$ using the transform $\pi = \exp(\text{logit})/[1+\exp(\text{logit})]$.

5.5.4 Newton–Raphson Method Applied to Logistic Regression

Section 4.6.1 introduced the Newton–Raphson iterative method, which applies in a straightforward manner to logistic regression. Let

$$u_j^{(t)} = \left.\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j}\right|_{\boldsymbol{\beta}^{(t)}} = \sum_i (y_i - n_i\pi_i^{(t)})x_{ij},$$

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

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

$$\pi_i^{(t)} = \frac{\exp\left(\sum_{j=1}^p \beta_j^{(t)}x_{ij}\right)}{1+\exp\left(\sum_{j=1}^p \beta_j^{(t)}x_{ij}\right)}. \tag{5.22}$$

We use $\boldsymbol{u}^{(t)}$ and $\boldsymbol{H}^{(t)}$ with formula (4.45) to obtain the next value $\boldsymbol{\beta}^{(t+1)}$, which in this context is

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \left\{\boldsymbol{X}^T\mathbf{Diag}\left[n_i\pi_i^{(t)}(1-\pi_i^{(t)})\right]\boldsymbol{X}\right\}^{-1}\boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{\mu}^{(t)}), \tag{5.23}$$

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 $\boldsymbol{\beta}^{(0)}$, (5.22) yields $\boldsymbol{\pi}^{(0)}$, and for $t > 0$ the iterations proceed as just described using (5.23) and (5.22). In the limit, $\boldsymbol{\pi}^{(t)}$ and $\boldsymbol{\beta}^{(t)}$ converge to the ML estimates $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\beta}}$ (Walker and Duncan 1967). The $\boldsymbol{H}^{(t)}$ matrices converge to $\hat{\boldsymbol{H}} = -\boldsymbol{X}^T\mathbf{Diag}[n_i\hat{\pi}_i(1-\hat{\pi}_i)]\boldsymbol{X}$. By (5.21) the estimated asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is a by-product of the Newton–Raphson method, namely $-\hat{\boldsymbol{H}}^{-1}$.

From the argument in Section 4.6.4, $\boldsymbol{\beta}^{(t+1)}$ has the iterative reweighted least-squares form $(\boldsymbol{X}^T\boldsymbol{V}_t^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}_t^{-1}\boldsymbol{z}^{(t)}$, where $z^{(t)}$ has elements

$$z_i^{(t)} = \log\frac{\pi_i^{(t)}}{1-\pi_i^{(t)}} + \frac{y_i - n_i\pi_i^{(t)}}{n_i\pi_i^{(t)}\left(1-\pi_i^{(t)}\right)}, \tag{5.24}$$

and where $\boldsymbol{V}_t$ is a diagonal matrix with elements $\{1/n_i\pi_i^{(t)}(1-\pi_i^{(t)})\}$. In this expression, $z^{(t)}$ is the linearized form of the logit link function for the sample data, evaluated at $\boldsymbol{\pi}^{(t)}$ [see (4.49)]. From Section 3.1.6 the elements of $\boldsymbol{V}_t$ are estimated asymptotic variances of the sample logits. The ML estimate is the limit of a sequence of weighted least-squares estimates, where the weight matrix changes at each cycle.

The log likelihood is concave, so there is no danger of iterative methods converging to a local maximum. However, in some cases at least one estimate may be infinite, as discussed in Section 6.5.

NOTES

Section 5.1: Interpreting Parameters in Logistic Regression

5.1 **Logistic books**: Books focusing on logistic regression include Collett (2003), Cox and Snell (1989), and Hosmer and Lemeshow (2000).

5.2 **Bias reduction**: Haldane (1956) recommended adding $\frac{1}{2}$ to each count in estimating a logit. With this modification, the bias is on the order of only $1/n_i^2$, for large n_i. See also Firth (1993a), Gart and Zweifel (1967), and Exercise 16.8. For bias reduction in logistic regression and GLMs, see Cordeiro and McCullagh (1991) and Firth (1993a).

5.3 **LD_{50}**: Paige et al. (2011) summarized confidence intervals for LD_{50} and proposed small-sample intervals using saddlepoint approximations.

5.4 **Retrospective logistic**: For discussion of logistic regression with retrospective studies, see Anderson (1972), Breslow (1996), Breslow and Day (1980, p. 203), Breslow and Powers (1978), Carroll et al. (1995), Farewell (1979), Ghosh and Mukherjee (2010), Mantel (1973), Neuhaus and Jewell (1990b), Piegorsch et al. (1994), Prentice (1976a), Prentice and Pyke (1979), Roeder et al. (1996), and Umbach and Weinberg (1997). Scott and Wild (2001) considered case–control studies with complex sampling designs, and Bhadra et al. (2012) incorporated longitudinal information on exposure history. Qin and Liang (2011) considered a mixture model to handle situations in which some controls are contaminated. See Section 7.2.3 for Bayesian literature.

5.5 **Design**: Khuri et al. (2006) reviewed articles about design problems for binary response experiments. Issues include choosing settings for a predictor to optimize a criterion for estimating parameter values, and estimating the setting at which the response probability equals some fixed value. The nonconstant variance makes this challenging. Zocchi and Atkinson (1999) considered multinomial logistic models.

Section 5.2: Inference for Logistic Regression

5.6 **Fitting/checking**: Albert and Anderson (1984), Berkson (1944, 1951, 1953, 1955), Cox (1958a), Hodges (1958), and Walker and Duncan (1967) discussed ML estimation for logistic regression, although Berkson argued for the computationally simpler minimum logit chi-squared. For adjustments with complex sample surveys, see Hosmer and Lemeshow (2000, Sec. 6.4) and LaVange et al. (2001). Grouping values to check model fit extends to any GLM (Pregibon 1982). Hosmer et al. (1997) compared various ways to do this. Presnell and Boos

(2004) proposed a general likelihood-based method for detecting model misspecification. See also Capanu and Presnell (2008).

Section 5.3: Logistic Models with Categorical Predictors

5.7 **Trend tests**: Extensions of the trend test include handling of correlated binary data by Corcoran et al. (2001) and stratified $I \times J$ tables by Mantel (1963). Williams (2005) surveyed trend tests for proportions and counts.

Section 5.4: Multiple Logistic Regression

5.8 **Standardizing**: Menard (2004) discussed several approaches to standardizing logistic regression coefficients. He noted that merely standardizing predictors, as was done in Section 5.4.7, is adequate for comparing influences of predictors.

5.9 **Quasi-variances**: For multipredictor models such as (5.12), tables that contain factor-level estimates $\{\hat{\beta}_i^X\}$ and their *SE* values but not their covariance matrix permit comparison of each category to the baseline (having estimate 0) but not to other categories. Firth and De Menezes (2004) showed how to construct *quasi-variances* $\{q_k\}$ such that the *SE* of $\hat{\beta}_a^X - \hat{\beta}_b^X$ is approximately $\sqrt{q_a^2 + q_b^2}$.

EXERCISES

Applications

5.1 An article about the contributions of star players in the National Basketball Association (by M. L. Jones and R. J. Parker, *Chance*: **23**, 39–45, 2010) reported prediction equations for the probability π of a win in a game for a player, using as predictors *ortg* = player's offensive rating in the game, which is the number of points produced per hundred possessions, *drtg* = player's defensive rating in the game, which is the number of points allowed per hundred possessions (the lower the better), and *home*, which indicates whether the game was played at home (1 = yes, 0 = no). For LeBron James using data from the 2008-2009 season,

$$\text{logit}(\hat{\pi}) = 1.379 + 0.119(ortg) - 0.139(drtg) + 3.393(home).$$

a. Over the season, James's quartiles (lower, median, upper) were (108.7, 123.2, 136.1) for *ortg* and (91.7, 99.5, 107.7) for *drtg*. Summarize the *ortg* effect for James by comparing $\hat{\pi}$ at its upper and lower quartiles. Do this at the median level of *drtg*, separately for home and away games. Repeat for the *drtg* effect, and compare.

b. Summarize the *home* effect by (i) comparing $\hat{\pi}$ for home and away games, at the median levels of *ortg* and *drtg*, (ii) interpreting its coefficient in the fitted logistic equation.

5.2 For a study using logistic regression to determine characteristics associated with remission in cancer patients, Table 5.11 shows the most important explanatory variable, a labeling index (LI) that measures proliferative activity of cells after

a patient receives an injection of tritiated thymidine. It represents the percentage of cells that are "labeled." The response measured whether the patient achieved remission. Software reports Table 5.12 for a logistic regression model using LI to estimate $\pi = P(\text{remission})$.

a. Show how software obtained $\hat{\pi} = 0.068$ when LI = 8.

b. Show that $\hat{\pi} = 0.50$ when LI = 26.0.

c. Show that the rate of change in $\hat{\pi}$ is 0.009 when LI = 8 and 0.036 when LI = 26.

d. The lower quartile and upper quartile for LI are 14 and 28. Show that $\hat{\pi}$ increases by 0.42, from 0.15 to 0.57, between those values.

e. For a unit increase in LI, show that the estimated odds of remission multiply by 1.16.

f. Explain how to obtain the confidence interval reported for the odds ratio. Interpret.

g. Construct a Wald test for the effect. Interpret.

Table 5.11 Data for Exercise 5.2 on Cancer Remission

LI	Number of Cases	Number of Remissions	LI	Number of Cases	Number of Remissions	LI	Number of Cases	Number of Remissions
8	2	0	18	1	1	28	1	1
10	2	0	20	3	2	32	1	0
12	3	0	22	2	1	34	1	1
14	3	0	24	1	0	38	3	2
16	3	0	26	1	1			

Source: Data reprinted with permission from E. T. Lee, *Comput. Prog. Biomed.* **4**: 80–92, 1974.

Table 5.12 Software Output (Based on SAS) for Exercise 5.2

```
                          Intercept    Intercept and
            Criterion      Only         Covariates
            -2 Log L      34.372         26.073

Parameter   Estimate   Standard Error   Chi-Square   Pr > ChiSq
Intercept    -3.7771       1.3786         7.5064       0.0061
li            0.1449       0.0593         5.9594       0.0146

                    Odds Ratio Estimates
   Effect   Point Estimate   95% Wald Confidence Limits
   li            1.156        1.029          1.298

                Estimated Covariance Matrix
            Variable    Intercept      li
            Intercept    1.900616   -0.07653
            li          -0.07653     0.003521

    Obs   li   remiss   n   pi_hat    lower     upper
     1     8      0     2   0.06797   0.01121   0.31925
     2    10      0     2   0.08879   0.01809   0.34010
```

h. Conduct a likelihood-ratio test for the effect, showing how to construct the test statistic using the $-2 \log L$ values reported.

i. Show how software obtained the confidence interval for π reported at LI = 8. [*Hint:* Use the reported covariance matrix.]

5.3 The text website has a data file (created from data at `www.basketball-reference.com`) showing, for each game in the 2010–2011 season of the National Basketball Association in which Rajon Rondo of the Boston Celtics played, x = the number of assists he recorded and y = whether the Celtics won (1 = yes). Using software, (**a**) show that the logistic model fitted to these data gives $\text{logit}[\hat{P}(Y = 1)] = -2.235 + 0.294x$; (**b**) show that $\hat{P}(Y = 1)$ increases from 0.21 to 0.99 over the observed range of x from 3 to 24; and (**c**) construct a significance test and confidence interval about the effect in the conceptual population that these games represent.

5.4 Table 5.13 summarizes logistic regression results from a study[1] of how family transitions relate to first home purchase by young married households. The response variable is whether the subject owns a home (1 = yes, 0 = no).

a. Interpret the effects that seem to be significant.

b. Fill in the blanks: Adjusting for the other explanatory variables, each additional child had the effect of multiplying the estimated odds of owning a home by _____; that is, the estimated odds increase by _____%. A $10,000 increase in earnings had the effect of multiplying the estimated odds of owning a home by _____ if the earnings add to husband's income and by _____ for wife's income.

Table 5.13 Results of Logistic Regression for Probability of Home Ownership

Variable	Estimate	Std. Error
Intercept	−2.870	—
Husband earnings ($10,000)	0.569	0.088
Wife earnings ($10,000)	0.306	0.140
Number of years married	−0.039	0.042
Married in 2 years (1 = yes)	0.224	0.304
Working wife in 2 years (1 = yes)	0.373	0.283
Number of children	0.220	0.101
Add child in 2 years (1 = yes)	0.271	0.140
Head's education (no. years)	−0.027	0.032
Parents' home ownership (1 = yes)	0.387	0.176

5.5 Consider the fit of model (5.2) for the horseshoe crabs using x = width.

a. Show that (i) at the mean width (26.3), the estimated odds of a satellite equal 2.07; (ii) at $x = 27.3$, the estimated odds equal 3.40; and (iii) since $\exp(\hat{\beta}) = 1.64$, $3.40 = (1.64)2.07$, and the odds increase by 64%.

[1] From J. Henretta, *Social Forces* **66**: 520–536, 1987.

b. Based on the 95% confidence interval for β, show that for x near where $\pi = 0.50$, the rate of increase in the probability of a satellite per 1-cm increase in x falls between about 0.07 and 0.17.

5.6 For the 23 space shuttle flights before the *Challenger* mission disaster in 1986, Table 5.14 shows the temperature at the time of the flight and whether at least one primary O-ring suffered thermal distress.

a. Use logistic regression to model the effect of temperature on the probability of thermal distress. Plot a figure of the fitted model, and interpret.

b. Estimate the probability of thermal distress at 31°F, the temperature at the place and time of the *Challenger* flight.

c. Construct a confidence interval for the effect of temperature on the odds of thermal distress, and test the statistical significance of the effect.

Table 5.14 Data for Exercise 5.6 on Challenger Space-Shuttle Disaster[a]

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

[a]Ft, flight number; Temp, temperature (°F); TD, thermal distress (1, yes; 0, no).
Source: Data based on Table 1 in *J. Am. Statist. Assoc.* **84**: 945–957, 1989, by S. R. Dalal, E. B. Fowlkes, and B. Hoadley. Reprinted with permission from *J. Am. Statist. Assoc.*

5.7 Refer to Table 4.2. Using scores (0, 2, 4, 5) for snoring, fit the logistic regression model. Interpret using fitted probabilities, linear approximations, and effects on the odds. Analyze the goodness of fit.

5.8 Hastie and Tibshirani (1990, p. 282) described a study to determine risk factors for kyphosis, severe forward flexion of the spine following corrective spinal surgery. The age in months at the time of the operation for the 18 subjects for whom kyphosis was present were 12, 15, 42, 52, 59, 73, 82, 91, 96, 105, 114, 120, 121, 128, 130, 139, 139, 157 and for 22 of the subjects for whom kyphosis was absent were 1, 1, 2, 8, 11, 18, 22, 31, 37, 61, 72, 81, 97, 112, 118, 127, 131, 140, 151, 159, 177, 206.

a. Fit a logistic regression model using age as a predictor of whether kyphosis is present. Test whether age has a significant effect.

b. Plot the data. Note the difference in dispersion on age at the two levels of kyphosis. Fit the model $\text{logit}[\pi(x)] = \alpha + \beta_1 x + \beta_2 x^2$. Test the significance of the squared age term, plot the fit, and interpret. (See also Exercise 5.30 and Section 7.4.3.)

5.9 For Table 5.5 on treating leprosy, the Pearson test of independence has $X^2(I) = 6.88$ ($P = 0.14$). For equally spaced scores, the Cochran–Armitage trend test has $z^2 = 6.67$ ($P = 0.01$). Interpret, and explain why the P-values differ so. Analyze the data using a linear logit model. Test independence using the Wald and

likelihood-ratio tests, and compare results to the Cochran–Armitage test. Check the fit of the model, and interpret.

5.10 Refer to Table 5.3 on infant malformation and alcohol consumption.

a. Repeat the trend test of Section 5.3.5 after deleting the single case in the last row. Comment on that observation's influence.

b. Repeat the trend test using alcohol consumption scores (1, 2, 3, 4, 5) instead of (0.0, 0.5, 1.5, 4.0, 7.0). Compare results, noting the potential sensitivity to the choice of scores for highly unbalanced data.

5.11 A study used the 1998 Behavioral Risk Factors Social Survey to consider factors associated with women's use of oral contraceptives in the United States. Table 5.15 summarizes effects for a logistic regression model for the probability of using oral contraceptives. Each predictor uses an indicator variable, and the table lists the category having indicator outcome 1. Interpret effects. Construct and interpret a confidence interval for the conditional odds ratio between contraceptive use and education.

Table 5.15 Data for Exercise 5.11 on Oral Contraceptive Use

Variable	Coding = 1 if:	Estimate	*SE*
Age	35 or younger	−1.320	0.087
Race	White	0.622	0.098
Education	≥1 year college	0.501	0.077
Marital status	Married	−0.460	0.073

Source: Data courtesy of Debbie Wilson, College of Pharmacy, University of Florida.

5.12 For the horseshoe crab data, available at `www.stat.ufl.edu/~aa/cda/cda.html`, fit a logistic regression model for the probability of a satellite, using color alone as the predictor.

a. Treat color as nominal. Explain why this model is saturated. Express its parameter estimates in terms of the sample logits for each color.

b. Conduct a likelihood-ratio test that color has no effect.

c. Fit a model that treats color as quantitative. Interpret the fit, and test that color has no effect.

d. Test the goodness of fit of the model in part (c). Interpret.

5.13 For model (5.15) with binary color c and width x, **(a)** describe the effect of width by finding the estimated probabilities of a satellite at its lower and upper quartiles, separately for $c = 1$ and $c = 0$, and **(b)** describe the effect of color by its average causal effect.

5.14 Refer to the prediction equation $\text{logit}(\hat{\pi}) = -10.071 - 0.509c + 0.458x$ for model (5.14) using quantitative color and width. The means and standard deviations are $\bar{c} = 2.44$ and $s = 0.80$ for color, and $\bar{x} = 26.30$ and $s = 2.11$ for width. For standardized predictors [e.g., $x = (\text{width} - 26.30)/2.11$], explain why the estimated coefficients

of c and x equal -0.41 and 0.97. Interpret these by comparing the partial effects of a standard deviation increase in each predictor on the odds. Describe the color effect by estimating the change in $\hat{\pi}$ between the first and last color categories at the sample mean width.

5.15 For Table 2.6, we fitted a logistic model, treating death penalty as the response (1 = yes) and defendant's race (1 = white) and victims' race (1 = white) as indicator predictors. Table 5.16 shows results.

a. Interpret parameter estimates. Which group is most likely to have the yes response? Find the estimated probability in that case.

b. Interpret 95% confidence intervals for conditional odds ratios.

c. Test the effect of defendant's race, controlling for victims' race, using a (i) Wald test and (ii) likelihood-ratio test. Interpret.

d. Test the goodness of fit of the model. Interpret.

Table 5.16 Software Output (Based on SAS) for Exercise 5.15 on the Death Penalty

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value
Deviance	1	0.3798
Pearson Chi-Square	1	0.1978
Log Likelihood		−209.4783

Parameter	Estimate	Standard Error	Likelihood Ratio 95% Conf Limits		Chi-Square
Intercept	−3.5961	0.5069	−4.7754	−2.7349	50.33
def	−0.8678	0.3671	−1.5633	−0.1140	5.59
vic	2.4044	0.6006	1.3068	3.7175	16.03

LR Statistics

Source	DF	Chi-Square	Pr > ChiSq
def	1	5.01	0.0251
vic	1	20.35	<.0001

5.16 Model the effects of victim's race and defendant's race for Table 2.12. Interpret.

5.17 In a 2011 article in *North Carolina Law Review*, M. Radelet and G. Pierce reported a logistic prediction equation for death penalty verdicts in North Carolina. Let Y denote whether a subject convicted of murder received the death penalty (1 = yes), for defendant's race h ($h = 1$, black; $h = 2$, white), victim's race i ($i = 1$, black; $i = 2$, white), and number of additional factors j ($j = 0, 1, 2$). For the model

$$\text{logit}[P(Y = 1)] = \alpha + \beta_h^D + \beta_i^V + \beta_j^F$$

they reported $\hat{\alpha} = -5.26$, $\hat{\beta}_1^D = 0.00$, $\hat{\beta}_2^D = 0.17$, $\hat{\beta}_1^V = 0.00$, $\hat{\beta}_2^V = 0.91$, $\hat{\beta}_0^F = 0.00$, $\hat{\beta}_1^F = 2.02$, $\hat{\beta}_2^F = 3.98$.

a. Estimate the probability of receiving the death penalty for the group most likely to receive it.

b. If, instead, parameters used constraints $\beta_2^D = \beta_2^V = \beta_2^F = 0$, report the estimates.

c. If, instead, parameters used constraints $\sum_h \beta_h^D = \sum_i \beta_i^V = \sum_j \beta_j^F = 0$, report the estimates.

5.18 In a study designed to evaluate whether an educational program makes sexually active adolescents more likely to obtain condoms, adolescents were randomly assigned to two experimental groups. The educational program, involving a lecture and videotape about transmission of HIV, was provided to one group but not the other. Table 5.17 summarizes results of a logistic regression model for factors observed to influence teenagers to obtain condoms.

a. Find the parameter estimates for the fitted model, using (1, 0) indicator variables for the first three predictors. Based on the corresponding confidence interval for the log odds ratio, determine the standard error for the group effect.

b. Explain why either the estimate of 1.38 for the odds ratio for gender or the corresponding confidence interval is incorrect. Show that if the reported interval is correct, 1.38 is actually the *log* odds ratio, and the estimated odds ratio equals 3.98.

Table 5.17 Data for Exercise 5.18 on Obtaining Condoms

Variable	Odds Ratio	95% Confidence Interval
Group (education vs. none)	4.04	(1.17, 13.9)
Gender (males vs. females)	1.38	(1.23.12.88)
SES (high vs. low)	5.82	(1.87, 18.28)
Lifetime number of partners	3.22	(1.08, 11.31)

Source: V. I. Rickert et al., *Clin. Pediatr.* **31**: 205–210, 1992.

5.19 Table 5.18 shows estimated effects for a logistic regression model with squamous cell esophageal cancer ($Y = 1$, yes; $Y = 0$, no) as the response. Smoking status (S) equals 1 for at least one pack per day and 0 otherwise, alcohol consumption (A) equals the average number of alcoholic drinks consumed per day, and race (R) equals 1 for blacks and 0 for whites. To describe the $R \times S$ interaction, construct the prediction equation when $R = 1$ and again when $R = 0$. Find the fitted YS conditional odds ratio for each case. Similarly, construct the prediction equation when $S = 1$ and again when $S = 0$. Find the fitted YR conditional odds ratios. Note that for each association, the coefficient of $R \times S$ is the difference between the log

Table 5.18 Data for Exercise 5.19 on Esophageal Cancer

Variable	Effect	P-value
Intercept	−7.00	< 0.01
Alcohol use (A)	0.10	0.03
Smoking (S)	1.20	< 0.01
Race (R)	0.30	0.02
Race × smoking ($R \times S$)	0.20	0.04

odds ratios at the two fixed levels for the other variable. Explain why the coefficient of S represents the log odds ratio between Y and S for whites. To what hypotheses do the P-values for R and S refer?

5.20 A survey of high school students on Y = whether the subject has driven a motor vehicle after consuming a substantial amount of alcohol (1 = yes), s = sex (1 = female), r = race (1 = black; 0 = white), and g = grade ($g_1 = 1$, grade 9; $g_2 = 1$, grade 10; $g_3 = 1$, grade 11; $g_1 = g_2 = g_3 = 0$, grade 12) has prediction equation

$$\text{logit}[\hat{P}(Y = 1)] = -0.88 - 0.40s - 0.72r - 2.22g_1 - 1.43g_2 - 0.58g_3 + 0.74(r \times g_1) + 0.38(r \times g_2) + 0.01(r \times g_3).$$

Carefully interpret effects. Explain the interaction by describing the race effect at each grade and the grade effect for each race.

5.21 The Gallup Poll reported in March 2010 that the percentage believing that news reports exaggerate the seriousness of global warming is 66% for Republicans and 22% for Democrats. By contrast, in 1998 the corresponding percentages were 34% and 23%. Considered as results for a three-way table cross-classifying opinion by political party and year, do these data seem to display interaction? In what sense?

5.22 A table at the text website refers to a sample of subjects randomly selected for an Italian study on the relation between income and whether one possesses a travel credit card. At each level of annual income in millions of lira (the Italian currency at the time of the study), the table indicates the number of subjects sampled and the number possessing at least one travel credit card. Analyze these data.

5.23 A research article in the *British Medical Journal* (by C. de Oliveira et al., 2010, vol. 340) showed results from the Scottish Health Survey, indicating that over a period of about 8 years, cardiovascular disease events occurred for 308 of 8481 subjects who reported brushing their teeth at least twice a day, for 188 of 2850 subjects who reported brushing once a day, and for 59 of 538 subjects who reported brushing less than once a day. Analyze these data.

5.24 Are people with more social ties less likely to get colds? Use logistic models to analyze the $2 \times 2 \times 2 \times 2$ contingency table on p. 1943 of the article by S. Cohen et al., *J. Am. Med. Assoc.* **277** (24).

Theory and Methods

5.25 For logistic regression model (5.1), show that $\partial\pi(x)/\partial x = \beta\pi(x)[1 - \pi(x)]$.

5.26 For logistic model (5.1), when $\pi(x)$ is small, explain why you can interpret $\exp(\beta)$ approximately as $\pi(x + 1)/\pi(x)$.

5.27 Prove that the logistic regression curve (5.1) has the steepest slope where $\pi(x) = \frac{1}{2}$. Generalize to model (5.8).

5.28 The calibration problem is that of estimating x at which $\pi(x) = \pi_0$ for some fixed π_0 such as 0.50. For the linear logit model, argue that a confidence interval is the set of x values for which

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

An alternative approach inverts a likelihood-ratio test.

5.29 A study for several professional sports of the effect of a player's draft position d ($d = 1, 2, 3, \ldots$) of selection from the pool of potential players in a given year on the probability π of eventually being named an all star used the model $\text{logit}(\pi) = \alpha + \beta \log d$ (S. M. Berry, *Chance*, **14**(2): 53–57, 2001).

a. Show that $\pi/(1-\pi) = e^{\alpha}d^{\beta}$. Show that e^{α} = odds for the first draft pick.

b. In the United States, Berry reported $\hat{\alpha} = 2.3$ and $\hat{\beta} = -1.1$ for pro basketball and $\hat{\alpha} = 0.7$ and $\hat{\beta} = -0.6$ for pro baseball. This suggests that in basketball a first draft pick is more crucial and picks with high d are relatively less likely to be all-stars. Explain why.

5.30 For the population having $Y = j$, suppose X has a $N(\mu_j, \sigma^2)$ distribution, $j = 0, 1$.

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

b. Suppose that $(X|Y = j)$ is $N(\mu_j, \sigma_j^2)$ with $\sigma_0 \neq \sigma_1$. Show that the logistic model holds with a quadratic term (Anderson 1975). [Exercise 5.8 showed that a quadratic term is helpful when x values have quite different dispersion at $y = 0$ and $y = 1$. This result also suggests that to test equality of means of normal distributions when the variances differ, we can fit a quadratic logistic regression with the two groups as the response and test the linear and quadratic terms together; see O'Brien (1988).]

c. Suppose that $(X|Y = j)$ has an exponential family density $f(x;\theta_j) = a(\theta_j)b(x)\exp[xQ(\theta_j)]$. Show that $P(Y = 1|x)$ satisfies the logistic model, with effect of x equal to $[Q(\theta_1) - Q(\theta_0)]$.

d. For multiple predictors, suppose that $(\boldsymbol{X}|Y = j)$ has a multivariate $N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ distribution, $j = 0, 1$. Show that $P(Y = 1|\boldsymbol{x})$ satisfies logistic regression with effect parameters $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$ (Cornfield 1962, Warner 1963).

5.31 Suppose that $\pi(x) = F(x)$ for some strictly increasing cdf F. Explain why a monotone transformation of x exists such that the logistic regression model holds. Generalize to alternative link functions.

5.32 For an $I \times 2$ contingency table, consider logistic model (5.4).

a. Given $\{\pi_i > 0\}$, show how to find $\{\beta_i\}$ satisfying $\beta_I = 0$.

b. Prove that $\beta_1 = \beta_2 = \cdots = \beta_I$ is the independence model. Find its likelihood equation, and show that $\hat{\alpha} = \text{logit}[(\sum_i y_i)/(\sum_i n_i)]$.

5.33 For a multinomial distribution, let $\gamma = \sum_i b_i \pi_i$, and suppose that $\pi_i = f_i(\theta) > 0$, $i = 1, \ldots, I$. For sample proportions $\{p_i\}$, let $S = \sum_i b_i p_i$. Let $T = \sum_i b_i \hat{\pi}_i$, where $\hat{\pi}_i = f_i(\hat{\theta})$, for the ML estimator $\hat{\theta}$ of θ.

a. Show that $\text{var}(S) = [\sum_i b_i^2 \pi_i - (\sum_i b_i \pi_i)^2]/n$.

b. Using the delta method, show $\text{var}(T) \approx [\text{var}(\hat{\theta})][\sum_i b_i f_i'(\theta)]^2$.

c. By computing the information for $L(\theta) = \sum_i n_i \log[f_i(\theta)]$, show that $\text{var}(\hat{\theta})$ is approximately $[n \sum_i (f_i'(\theta))^2 / f_i(\theta)]^{-1}$.

d. Asymptotically, show that $\text{var}[\sqrt{n}(T - \gamma)] \leq \text{var}[\sqrt{n}(S - \gamma)]$. [*Hint*: Show that $\text{var}(T)/\text{var}(S)$ is a squared correlation between two random variables, where with probability π_i the first equals b_i and the second equals $f_i'(\theta)/f_i(\theta)$.]

5.34 Construct the log-likelihood function for the model $\text{logit}[\pi(x)] = \alpha + \beta x$ with independent binomial outcomes of y_0 successes in n_0 trials at $x = 0$ and y_1 successes in n_1 trials at $x = 1$. Derive the likelihood equations, and show that $\hat{\beta}$ is the sample log odds ratio.

5.35 A study has n_i independent binary observations $\{y_{i1}, \ldots, y_{in_i}\}$ when $X = x_i$, $i = 1, \ldots, N$, with $n = \sum_i n_i$. Consider the model $\text{logit}(\pi_i) = \alpha + \beta x_i$, where $\pi_i = P(Y_{ij} = 1)$.

a. Show that the kernel of the likelihood function is the same 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. [*Hint:* The number of parameters differs.] 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.

d. Suppose that each $n_i = 1$. Show that the deviance depends on $\hat{\pi}_i$ but not y_i. Hence, it is not useful for checking model fit (see also Exercise 4.18).

5.36 Suppose that Y has a $\text{bin}(n, \pi)$ distribution. For the model, $\text{logit}(\pi) = \alpha$, consider testing H_0: $\alpha = 0$ (i.e., $\pi = 0.50$). Let $\hat{\pi} = y/n$.

a. Compare the estimated *SE* for the Wald test and the *SE* using the null value 0.50 for π, for two possible denominators in the test statistic $[\text{logit}(\hat{\pi})/SE]^2$. Show that the ratio of the Wald statistic to the statistic with null *SE* equals $4\hat{\pi}(1 - \hat{\pi})$. What is the implication about performance of the Wald test if $|\alpha|$ is large and $\hat{\pi}$ tends to be near 0 or 1?

b. How does the comparison of tests change with the scale $[(\hat{\pi} - 0.5)/SE]^2$, where *SE* is now the estimated or null *SE* of $\hat{\pi}$? [Analogous results apply for inference about the Poisson mean versus the log mean; see also Mantel (1987a) and Section 5.2.6.]

5.37 Find the likelihood equations for model (5.10) with two binary predictors. Show that they imply that the fitted values and the sample counts are identical in the marginal two-way tables.

5.38 Consider the likelihood equations (5.18) for a logistic regression model. Using the equation resulting from the intercept parameter, show that the overall sample proportion of successes equals the sample mean of the fitted success probabilities.

5.39 Consider the linear logit model (5.5) for an $I \times 2$ table, with y_i a bin(n_i, π_i) variate.

a. Show that the log likelihood is

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{I} y_i(\alpha + \beta x_i) - \sum_{i=1}^{I} n_i \log[1 + \exp(\alpha + \beta x_i)].$$

b. Show that the sufficient statistic for β is $\sum_i y_i x_i$, and explain why this is essentially the variable utilized in the Cochran–Armitage test. (That test is a score test of H_0: $\beta = 0$.)

c. Letting $S = \sum_i y_i$, show that the likelihood equations are

$$S = \sum_i n_i \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)},$$

$$\sum_i y_i x_i = \sum_i n_i x_i \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}.$$

d. Let $\{\hat{\mu}_i = n_i \hat{\pi}_i\}$. Explain why $\sum_i \hat{\mu}_i = \sum_i y_i$ and

$$\sum_i x_i \frac{y_i}{S} = \sum_i x_i \frac{\hat{\mu}_i}{\sum_a \hat{\mu}_a}.$$

Explain why this implies that the mean score on x across the rows in the first column is the same for the model fit as for the observed data. (They are also identical for the second column.)

5.40 Let Y_i be bin(n_i, π_i) at x_i, and let $p_i = y_i/n_i$. For binomial GLMs with logit link:

a. For p_i near π_i, show that

$$\log \frac{p_i}{1 - p_i} \approx \log \frac{\pi_i}{1 - \pi_i} + \frac{p_i - \pi_i}{\pi_i(1 - \pi_i)}.$$

b. Show that $z_i^{(t)}$ in (5.24) is a linearized version of the ith sample logit, evaluated at approximation $\pi_i^{(t)}$ for $\hat{\pi}_i$.

c. Verify the formula (5.21) for $\widehat{\text{cov}}(\hat{\boldsymbol{\beta}})$.