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 Consumption	Malformation		Sample Logit	Proportion Malformed	
	Present	Absent		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

                          Std    Likelihood-Ratio     Wald
Parameter   Estimate      Error   95% Conf Limits     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
```