CHAPTER 6

Building, Checking, and Applying Logistic Regression Models

Having studied the basics of fitting and interpreting logistic regression models, we now turn our attention to building and applying them. With several explanatory variables, there are many potential models. In Section 6.1 we discuss strategies for model selection. After choosing a preliminary model, model checking addresses whether systematic lack of fit exists. Section 6.2 covers diagnostics, such as residuals, for model checking. Section 6.3 presents ways of summarizing the predictive power of a model.

In practice, an important application is comparing two groups on a binary response, while adjusting for possibly confounding variables. In Section 6.4 we present the Cochran–Mantel–Haenszel test, a popular way to do this by forming strata for levels of control variables. We then present ways of summarizing the effect, with application to meta-analyses.

Infinite estimates of logistic regression model parameters can occur with certain data configurations. Section 6.5 discusses ways to detect and deal with them. Section 6.6 covers power and sample size determination for logistic regression.

6.1 STRATEGIES IN MODEL SELECTION

Model selection for logistic regression faces the same issues as for ordinary regression. The selection process becomes harder as the number of explanatory variables increases, because of the rapid increase in possible effects and interactions. There are two competing goals: The model should be complex enough to fit the data well. On the other hand, ideally it should be relatively simple to interpret, smoothing rather than overfitting the data. Complications can arise because of the binary nature of the response variable, such as infinite ML parameter estimates for some models when one response outcome is much more common than the other.

Most research studies are designed to answer certain questions. Those questions guide the choice of model terms. Confirmatory analyses then use a restricted set of models. For instance, a study hypothesis about an effect may be tested by comparing models with and

Categorical Data Analysis, Third Edition. Alan Agresti.

without that effect. For studies that are exploratory rather than confirmatory, a search among possible models may provide clues about the dependence structure and raise questions for future research. In either case, it is helpful first to study the effect of each predictor by itself using graphics (incorporating smoothing) for a continuous predictor or conditional distributions within a contingency table for a discrete predictor. This gives you a feel for the marginal effects.

6.1.1 How Many Explanatory Variables Can Be in the Model?

Unbalanced data, with relatively few responses of one type, limit the number of predictors for which we can effectively estimate effects. One guideline based on a Monte Carlo study (Peduzzi et al. 1996) suggested that when there are fewer than 10 outcomes of each type per predictor, impacts can include severely biased parameter estimates, poor standard error estimates, and error rates for Wald tests and confidence intervals far from the nominal level. If $y = 1$ only 30 times out of $n = 1000$, for instance, this guideline implies that ideally the model should contain no more than three predictors.

This is merely one guideline and does *not* mean that you should never consider models that violate it. Many data sets now have large numbers of variables relative to the sample size. With certain strategies presented in Chapter 7, such as penalized likelihood methods that can shrink many estimates to 0, it is possible to have very many predictors. Likewise, you should not use such a guideline to justify being overly ambitious. For example, if you have 1000 outcomes of each type, you are not usually well served by a model with 100 predictors.

Many model selection procedures exist, no one of which is always best. Cautions that apply to ordinary regression hold for any generalized linear model. For instance, a model with several explanatory variables may exhibit *multicollinearity*—correlations among them making it seem that no one variable is important when all the others are in the model. A variable may seem to have little effect because it overlaps considerably with the other explanatory variables in the model, itself being predicted well by the others. Deleting such a redundant variable can be helpful, for instance, to reduce standard errors of other estimated effects.

6.1.2 Example: Horseshoe Crab Mating Data Revisited

The horseshoe crab data set in Table 4.3 has four explanatory variables: color (four categories), spine condition (three categories), weight, and width of the shell. We now fit a logistic regression model using all these to predict whether the female crab has male satellites nearby ($y = 1$).

We start by fitting a model containing all the main effects,

$$\begin{aligned}\text{logit}[P(Y = 1)] = \alpha + \beta_1 \text{weight} + \beta_2 \text{width} + \beta_3 c_1 \\ + \beta_4 c_2 + \beta_5 c_3 + \beta_6 s_1 + \beta_7 s_2,\end{aligned}$$

treating color (c_i) and spine condition (s_j) as qualitative (factors), with indicator variables for the first three colors and the first two spine conditions. Table 6.1 shows results. A likelihood-ratio test that Y is jointly independent of these predictors simultaneously tests H_0: $\beta_1 = \cdots = \beta_7 = 0$. The test statistic equals 40.56 with df = 7 ($P < 0.0001$). This shows extremely strong evidence that at least one predictor has an effect.

Table 6.1 Software Output (Based on SAS) from Fitting Model with All Main Effects to Horseshoe Crab Data

Testing Global Null Hypothesis: BETA = 0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	40.5565	7	< .0001

Analysis of Maximum Likelihood Estimates

Parameter	Estimate	Std Error	Chi-Square	Pr > ChiSq
Intercept	−9.2734	3.8378	5.8386	0.0157
weight	0.8258	0.7038	1.3765	0.2407
width	0.2631	0.1953	1.8152	0.1779
color 1	1.6087	0.9355	2.9567	0.0855
color 2	1.5058	0.5667	7.0607	0.0079
color 3	1.1198	0.5933	3.5624	0.0591
spine 1	−0.4003	0.5027	0.6340	0.4259
spine 2	−0.4963	0.6292	0.6222	0.4302

Although the overall test is highly significant, the Table 6.1 results are discouraging. The estimates for weight and width are only slightly larger than their *SE* values. The estimates for the factors compare each category to the final one as a baseline. For color, only one effect is clearly significant; for spine condition, the largest difference is less than a standard error.

The small *P*-value for the overall test, yet the lack of significance for individual effects, is a warning sign of multicollinearity. In Section 5.2.2 we showed strong evidence of a width effect. Adjusting for weight, color, and spine condition, little evidence remains of a partial width effect. However, weight and width have a strong correlation (0.887). For practical purposes they are equally good predictors, but it is nearly redundant to use them both. Our further analysis uses width (W) with color (C) and spine condition (S) as explanatory variables. For simplicity, we symbolize models by their highest-order terms, regarding C and S as factors. For instance, $(C + S + W)$ denotes a model with main effects, whereas $(C + S * W)$ denotes a model that has those main effects plus an $S \times W$ interaction. It is not usually sensible to consider a model with interaction that does not also contain the main effects that make up that interaction.

6.1.3 Stepwise Procedures: Forward Selection and Backward Elimination

In exploratory studies, an algorithmic method for searching among models can be informative if we use results cautiously. Goodman (1971a) proposed methods analogous to forward selection and backward elimination in ordinary regression.

Forward selection adds terms sequentially. At each stage it selects the term giving the greatest improvement in fit. The minimum *P*-value for testing the term in the model is a sensible criterion, since reductions in deviance for different terms may have different df values. A point of diminishing returns occurs in adding predictors, when new predictors are so correlated with ones already used that they do not improve predictive power. The process stops when further additions do not significantly improve the fit. A stepwise variation of this procedure retests, at each stage, terms added at previous stages to see if they are still significant.

Backward elimination begins with a complex model and sequentially removes terms. At each stage, it selects the term whose removal has the least damaging effect on the model (e.g., largest P-value). The process stops when any further deletion leads to a significantly poorer fit. With either approach, for qualitative predictors with more than two categories, the process should consider the entire variable at any stage rather than just individual indicator variables. Add or drop the entire variable rather than just one of its indicators. Otherwise, the result depends on the choice of baseline for the indicator coding. The same remark applies to interactions containing that variable.

Some statisticians prefer backward elimination over forward selection, feeling it safer to delete terms from an overly complex model than to add terms to an overly simple one. Forward selection can stop prematurely because a particular test in the sequence has low power. Neither strategy necessarily yields a meaningful model. Use variable selection procedures with caution! Various studies have shown their limitations and pitfalls (e.g., Steyerberg et al. 2001). When you evaluate many terms, one or two that are not truly important may look impressive merely due to chance. For instance, when all the true effects are weak, the largest sample effect is likely to overestimate substantially its true effect. It is best to use such algorithms in an informal manner. This includes the interpretation of P-values used as cutoff points, since the distribution of the minimum or maximum P-value evaluated over a set of predictors is not the same as that of a P-value for a preselected variable.

Some software has additional options for selecting a model. One approach attempts to determine the best model with some fixed number of terms, according to some criterion. If such a method and backward and forward selection procedures yield quite different models, this is an indication that such results are of dubious use. Another such indication would be when a quite different model results from applying a given procedure to a bootstrap sample of the same size from the sample distribution.

Finally, statistical significance should not be the sole criterion for inclusion of a term in a model, and true significance can be difficult to judge in any case (Westfall and Young 1993). It is sensible to include a variable that is central to the purposes of the study and report its estimated effect even if it is not statistically significant. Keeping it in the model may help reduce bias in estimated effects of other predictors and may make it possible to compare results with other studies where the effect is significant, perhaps because of a larger sample size. Algorithmic selection procedures are no substitute for careful thought in guiding the formulation of models.

6.1.4 Example: Backward Elimination for Horseshoe Crab Data

Table 6.2 summarizes results of fitting and comparing several logistic models to the horseshoe crab data with predictors width, color, and spine condition. The deviance (G^2) test of fit compares the model to the saturated model. As noted in Sections 5.2.4 and 5.2.5, this is not approximately chi-squared when a predictor is continuous, as width is. However, the deviance difference between two models that differ by a modest number of parameters is relevant. That difference is the likelihood-ratio statistic $-2(L_0 - L_1)$ comparing the models, and it has an approximate null chi-squared distribution.

To select a model, we use backward elimination, at each stage testing only the highest-order terms for each variable. It is inappropriate, for instance, to remove a main effect term if the model has interactions involving that term.

Table 6.2 Results of Fitting Several Logistic Regression Models to Horseshoe Crab Data

Model	Predictors[a]	Deviance G^2	df	AIC	Models Compared	Deviance Difference	Corr. $R(y, \hat{\mu})$
1	$(C * S * W)$	170.44	152	212.4	—	—	
2	$(C * S + C * W + S * W)$	173.68	155	209.7	(2)–(1)	3.2 (df = 3)	
3a	$(C * S + S * W)$	177.34	158	207.3	(3a)–(2)	3.7 (df = 3)	
3b	$(C * W + S * W)$	181.56	161	205.6	(3b)–(2)	7.9 (df = 6)	
3c	$(C * S + C * W)$	173.69	157	205.7	(3c)–(2)	0.0 (df = 2)	
4a	$(S + C * W)$	181.64	163	201.6	(4a)–(3c)	8.0 (df = 6)	
4b	$(W + C * S)$	177.61	160	203.6	(4b)–(3c)	3.9 (df = 3)	
5	$(C + S + W)$	186.61	166	200.6	(5)–(4b)	9.0 (df = 6)	0.456
6a	$(C + S)$	208.83	167	220.8	(6a)–(5)	22.2 (df = 1)	0.314
6b	$(S + W)$	194.42	169	202.4	(6b)–(5)	7.8 (df = 3)	0.402
6c	$(C + W)$	187.46	168	197.5	(6c)–(5)	0.8 (df = 2)	0.452
7a	(C)	212.06	169	220.1	(7a)–(6c)	24.5 (df = 1)	0.285
7b	(W)	194.45	171	198.5	(7b)–(6c)	7.0 (df = 3)	0.402
8	$(C = \text{dark} + W)$	187.96	170	194.0	(8)–(6c)	0.5 (df = 2)	0.447
9	None	225.76	172	227.8	(9)–(8)	37.8 (df = 2)	0.000

[a] C, color; S, spine condition; W, width.

We begin with the most complex model, symbolized by $(C * S * W)$, model 1 in Table 6.2. This model uses main effects for each term as well as the three two-factor interactions and the three-factor interaction. It allows a separate width effect at each CS combination. (In fact, at some of those combinations y outcomes of only one type occur, which implies that those effects are not estimable.) The likelihood-ratio statistic comparing this model to the simpler model $(C * S + C * W + S * W)$ removing the three-factor interaction term equals 3.2 (df = 3). This suggests that the three-factor term is not needed ($P = 0.36$), thank goodness, so we continue the simplification process.

At the next stage we compare the model $(C * S + C * W + S * W)$ to the simpler model $C + S + W$ containing only main effects. The likelihood-ratio statistic comparing the model is the change in deviance, $186.61 - 173.68 = 12.9$ (df $= 166 - 155 = 11$). This suggests that two-factor interactions terms are not needed either ($P = 0.30$). Table 6.2 also shows results for intermediate models, and a backward process dropping a term at a time also results in eliminating all the three-factor terms.

At the next stage we consider dropping a main effect term. Table 6.2 shows little consequence of removing S. Both remaining variables (C and W) then have nonnegligible effects. For instance, removing C increases the deviance (comparing models 7b and 6c) by 7.0 on df = 3 ($P = 0.07$). The analysis in Section 5.4.6 revealed a noticeable difference between dark crabs (category 4) and the others. The simpler model that has a single indicator variable for color, equaling 0 for dark crabs and 1 otherwise, fits essentially as well. Further simplification results in large increases in deviance and is unjustified.

6.1.5 Model Selection and the "Correct" Model

In selecting a model from a set of candidates, we are mistaken if we think that there is a "correct" one. Any model is a simplification of reality. For instance, width does not have

exactly a linear effect on the probability of satellites, whether we use the logit link or the identity link.

What is the logic of testing the fit of a model when we know that it does not truly hold? A simple model that fits adequately has the advantages of model parsimony. If a model has relatively little bias, describing reality well, it tends to provide more accurate estimates of the quantities of interest.[1]

Other criteria besides significance tests can help select a good model in terms of estimating quantities of interest. We next introduce the best known of such criteria.

6.1.6 AIC: Minimizing Distance of the Fit from the Truth

The *Akaike information criterion* (AIC) judges a model by how close its fitted values tend to be to the true mean values, in terms of a certain expected value. Even though a simple model is farther from the true relationship than is a more complex model, it may be preferred because it tends to provide better estimates of certain characteristics, such as cell probabilities. Thus, the optimal model is the one that tends to have fit closest to the true values.

Akaike defined closeness in terms of a Kullback–Leibler measure of distance. Let $p(\boldsymbol{y})$ denote the probability (or density) of the data under the true model and $p_M(\boldsymbol{y})$ the probability under the chosen model. The distance measure is $E\{\log[p(\boldsymbol{y})/p_M(\boldsymbol{y})]\}$, where the expected value is taken relative to the true distribution. For categorical data, this measure resembles G^2 in form. With a sample, this criterion selects the model that minimizes

$$\text{AIC} = -2\,(\text{maximized log likelihood} - \text{number of parameters in model}).$$

This penalizes a model for having many parameters. With models for categorical Y, this ordering is equivalent to one based on an adjustment of the deviance, $[G^2 - 2(\text{df})]$, by twice its residual df.

With many potential predictors, we can use the AIC to aid in variable selection. Out of a set of candidate models, we identify the one with smallest AIC. However, models with similar AIC values are also of interest. For instance, we would consider also more parsimonious models that have AIC relatively close to the minimum value.

We illustrate AIC for model selection using the models that Table 6.2 lists. That table also shows the AIC values. Of models using the three basic variables, AIC is smallest (AIC = 197.5) for $C + W$, having main effects of color and width. The simpler model having an indicator variable for whether a crab is dark fares better yet (AIC = 194.0). Either model seems reasonable. We should balance the lower AIC for the simpler model against its having been suggested by the fit of model $C + W$.

An alternative *Bayesian information criterion* (BIC) penalizes more severely for the number of parameters in the model. It replaces 2 by $\log(n)$ as the multiple of the number of parameters, so the selected model is no more complex than the one selected with AIC. Compared with AIC, BIC gravitates less quickly toward more complex models as n increases. It is derived based on a Bayesian argument for determining which of a set of models has highest posterior probability. Differences between BIC values for two models

[1] We discussed the parsimony issue, with examples, in Sections 3.3.8, 5.2.2, and 5.3.10.

relate to a Bayes factor comparing them. It has the property of selecting the "correct model" with probability converging to 1 as $n \to \infty$. However, this is based on the Bayesian structure that provides justification for this approach, and its relevance is unclear when applied with frequentist methods. Also, in practice we do not regard any one model as "correct," so the AIC approach of choosing the model that is closest to reality seems sensible.

For the horseshoe crab mating data, from Table 6.2, AIC = 197.5 for model $(C + W)$ and AIC = 198.5 for model (W). By contrast, BIC = 213.2 for model $(C + W)$ and BIC = 204.8 for model (W), thus differing from AIC by preferring the simpler model.

6.1.7 Example: Using Causal Hypotheses to Guide Model Building

Although selection procedures are helpful exploratory tools, the model-building process should utilize theory and common sense. Often, a time ordering among the variables suggests possible casual relationships. Analyzing a certain sequence of models helps to investigate those relationships (Goodman 1973).

We illustrate with Table 6.3, from a British study that employed a random sample survey. A sample of men and women who had petitioned for divorce and an independent sample of married people were asked: (a) "Before you married your (former) husband/wife, had you ever made love with anyone else?"; (b) "During your (former) marriage, (did you have) have you had any affairs or brief sexual encounters with another man/woman?" The $2 \times 2 \times 2 \times 2$ table has variables G = gender, E = extramarital sex report (yes or no), P = premarital sex report, and M = marital status.

The time points at which responses on the four variables occur suggests the following ordering of the variables:

$$\underset{\text{gender}}{G} \to \underset{\substack{\text{premarital}\\ \text{sex}}}{P} \to \underset{\substack{\text{extramarital}\\ \text{sex}}}{E} \to \underset{\substack{\text{marital}\\ \text{status}}}{M}$$

Any of these is an explanatory variable when a variable listed to its right is the response. Figure 6.1 shows one possible causal structure. In this figure, a variable at the tip of an arrow is a response for a model at some stage. The explanatory variables have arrows pointing toward the response, directly or indirectly.

We first treat P as a response. Figure 6.1 predicts that G has a direct effect on P, so the model of independence of these variables is inadequate. At the second stage, E is the

Table 6.3 Marital Status by Report of Pre- and Extramarital Sex (PMS and EMS)

		Gender							
		Women				Men			
	PMS:	Yes		No		Yes		No	
Marital Status	EMS:	Yes	No	Yes	No	Yes	No	Yes	No
Divorced		17	54	36	214	28	60	17	68
Still married		4	25	4	322	11	42	4	130

Source: G. N. Gilbert, *Modelling Society*. London: George Allen & Unwin, 1981. Reprinted with permission from Unwin Hyman Ltd.

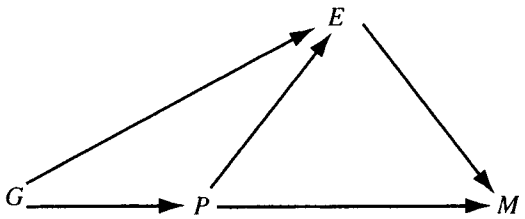

Figure 6.1 Causal diagram for Table 6.3.

response. Figure 6.1 predicts that P and G have direct effects on E. It also suggests that G has an indirect effect on E, through its effect on P. These effects on E can be analyzed using the logistic model for E with additive G and P effects. If G has only an indirect effect on E, the model with P alone as a predictor is adequate; that is, at a given level of P, E and G are conditionally independent. At the third stage, M is the response. Figure 6.1 predicts that E has a direct effect on M, P has direct effects and indirect effects through its effects on E, and G has indirect effects through its effects on P and E. This suggests the logistic model for M having additive E and P effects. For this model, G and M are independent, given P and E.

Table 6.4 shows results. The first stage, having P as the response, shows strong evidence of a GP association. The sample odds ratio for their marginal table is 0.27; the estimated odds of premarital sex for females are 0.27 times that for males. The second stage has E as the response. Only weak evidence occurs that G had a direct as well as an indirect effect on E, as G^2 drops by 2.9 (df = 1) after adding G to a model already containing P as a predictor. For this model, the estimated EP conditional odds ratio is 3.6.

The third stage has M as the response. Figure 6.1 specifies the logistic model with main effects of E and P, but it fits poorly. The model that allows an $E \times P$ interaction in their effects on M but assumes conditional independence of G and M fits much better (G^2 decrease of 13.0, df = 1). The model that also has a main effect for G fits slightly better yet. Either model is more complicated than Figure 6.1 predicted, since the effects of E on M vary according to the level of P. However, some preliminary thought about causal relationships suggested a model similar to one giving a good fit. We leave it to the reader to estimate and interpret effects for the third stage.

Table 6.4 Goodness of Fit of Various Models for Table 6.3[a]

Stage	Response Variable	Potential Explanatory	Actual Explanatory	G^2	df
1	P	G	None	75.3	1
			(G)	0.0	0
2	E	G, P	None	48.9	3
			(P)	2.9	2
			$(G + P)$	0.0	1
3	M	G, P, E	$(E + P)$	18.2	5
			$(E * P)$	5.2	4
			$(E * P + G)$	0.7	3

[a]P, premarital sex; E, extramarital sex; M, marital status; G, gender.

6.1.8 Alternative Strategies, Including Model Averaging

In practice, many models can be consistent with the data. If, as stated in Section 6.1.5, no one of them is "correct," it is logically inconsistent to choose one model based on its fitting the data well and then make subsequent inferences acting as if the model is fixed. This can result in a tendency to underestimate uncertainty and to exaggerate significance. Copas and Eguchi (2010) discussed this issue. They noted that an increasingly popular way of dealing with this is Bayesian model averaging: Identify a set of plausible models, specify prior probabilities for them, and base inference on a weighting according to posterior model probabilites. Copas and Eguchi proposed an alternative approach that identifies statistically equivalent models (that are consistent with the data) and constructs an "envelope likelihood" that reflects the model uncertainty. For estimation of a particular measure, this approach typically generates wider limits that more appropriately reflect the uncertainty.

As computing power continues to explode, enormous data sets are more common, in applications as diverse as genomic investigations and credit scoring by financial institutions. Many applications have huge numbers of potential explanatory variables, making model selection much more difficult. We discuss special issues for such cases in Section 7.5.

In summary, although the focus of this section has been "model selection," it is often not sensible to have the goal of picking a single model. Also, we should keep in mind the selection uncertainty when we make inferences based on a model, and also realize the tentative nature of using the same data in making those inferences that were used to select a model.

6.2 LOGISTIC REGRESSION DIAGNOSTICS

In Section 5.2.3 we introduced statistics for checking model fit in a global sense. After selecting a preliminary model, we obtain further insight by switching to a microscopic mode of analysis. In contingency tables, for instance, the pattern of lack of fit revealed in cell-by-cell comparisons of observed and fitted counts may suggest a better model or may indicate a segment of the population for which a generally good-fitting model fails.

6.2.1 Residuals: Pearson, Deviance, and Standardized

With categorical predictors, it is useful to form residuals to compare observed and fitted counts. Let y_i denote the binomial outcome for n_i trials at setting i of the explanatory variables, $i = 1, \ldots, N$. Let $\hat{\pi}_i$ denote the model estimate of $P(Y = 1)$. Then $\hat{\mu}_i = n_i\hat{\pi}_i$ is the fitted number of successes.

For a GLM with binomial random component, for observation i the Pearson residual (4.41) is

$$e_i = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{\widehat{\text{var}}(Y_i)}} = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{[n_i\hat{\pi}_i(1-\hat{\pi}_i)]}}. \tag{6.1}$$

This divides the raw residual $(y_i - \hat{\mu}_i)$ by the estimated binomial standard deviation of y_i. The Pearson statistic for testing the model fit satisfies

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

An alternative residual uses components of the G^2 fit statistic. This is the *deviance residual*, introduced for GLMs in (4.42). For a binomial GLM, this is

$$\sqrt{d_i} \times \text{sign}(y_i - n_i\hat{\pi}_i), \tag{6.2}$$

where

$$d_i = 2\left(y_i \log \frac{y_i}{n_i\hat{\pi}_i} + (n_i - y_i)\log \frac{n_i - y_i}{n_i - n_i\hat{\pi}_i}\right).$$

As explained in Section 4.5.6, these and the $\{e_i\}$ are less variable than $N(0, 1)$.

A standardized version of the Pearson residual divides it by its estimated standard error. As noted in Section 4.5.6, this is larger than the Pearson residual, with adjustment that uses the leverage from an estimated hat matrix. For observation i with leverage $\hat{h}_i$, the standardized residual is

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

It has the advantages compared with the Pearson and deviance residuals of having an approximate $N(0, 1)$ distribution when the model holds and appropriately recognizing redundancies (as noted for 2×2 tables in Section 3.3.1 and in Section 6.2.3 below). Absolute values larger than roughly 2 or 3 provide evidence of lack of fit. It takes larger values to be noteworthy when relatively more of them are inspected.

Plots of residuals against explanatory variables or linear predictor values may detect a type of lack of fit. When fitted values are very small, however, just as X^2 and G^2 lose relevance, so do residuals. When explanatory variables are continuous, often $n_i = 1$ at each setting. Then y_i can equal only 0 or 1, and e_i can assume only two values. One must then be cautious about regarding either outcome as extreme, and a single residual is usually uninformative (see Exercise 6.32). Plots of residuals also then have limited use. Figure 6.2 illustrates, plotting for the horseshoe crab data the standardized residuals against width for the model (5.13) fitted in Section 5.4.5 having width and color as predictors. Width has a strong positive effect, so necessarily for small width values an observation of $y = 1$ will have a relatively large positive residual whereas for large width values an observation of $y = 0$ will have a relatively large negative residual. When plotted against fitted values, a plot of the raw residuals consists merely of two parallel lines of points. The deviance itself is then completely uninformative (Exercise 5.35). When data can be grouped into sets of observations having common predictor values, it is better to compute residuals for the grouped data than for individual subjects.

6.2.2 Example: Heart Disease and Blood Pressure

A sample of male residents of Framingham, Massachusetts, aged 40 through 59, were classified on several factors, including systolic blood pressure. The response variable is whether they developed coronary heart disease during a six-year follow-up period. Table 6.5 shows results.

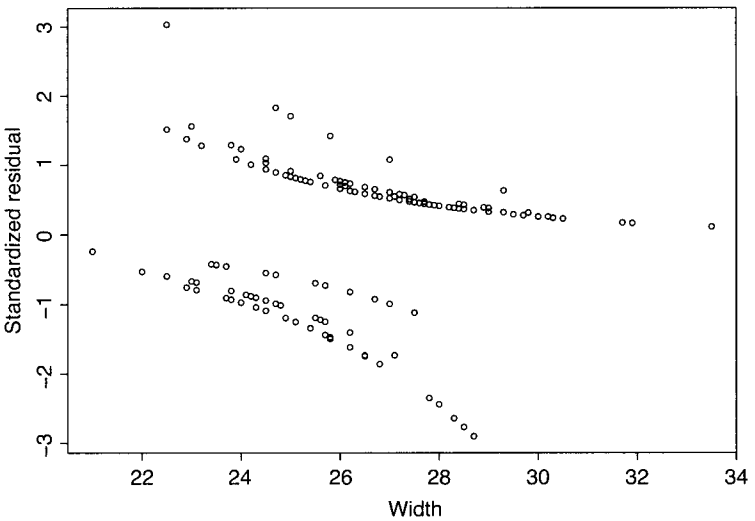

Figure 6.2 Plot of standardized residuals against width, for model predicting horseshoe crab satellites using width and color predictors.

Let π_i be the probability of heart disease for blood pressure category i. The table shows the fit and the standardized residuals for two logistic regression models. The first model,

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

treats the response as independent of blood pressure. Some residuals for that model are large. This is not surprising, since the model fits poorly ($G^2 = 30.02$, $X^2 = 33.38$, $\text{df} = 7$).

A plot of the residuals for the independence model shows an increasing trend. This suggests the linear logit model,

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

Table 6.5 Presence of Heart Disease by Blood Pressure, with Fit of Logistic Models and Standardized Residuals

Systolic Pressure (mmHg)	Sample Size	Observed Heart Disease	Fitted		Standardized Residual	
			Independence Model	Linear Logit	Independence Model	Linear Logit
<117	156	3	10.8	5.2	−2.62	−1.11
117–126	252	17	17.4	10.6	−0.12	2.37
127–136	284	12	19.7	15.1	−2.02	−0.95
137–146	271	16	18.8	18.1	−0.74	−0.57
147–156	139	12	9.6	11.6	0.84	0.13
157–166	85	8	5.9	8.9	0.93	−0.33
167–186	99	16	6.9	14.2	3.76	0.65
>186	43	8	3.0	8.4	3.07	−0.18

Source: Data from Cornfield (1962).

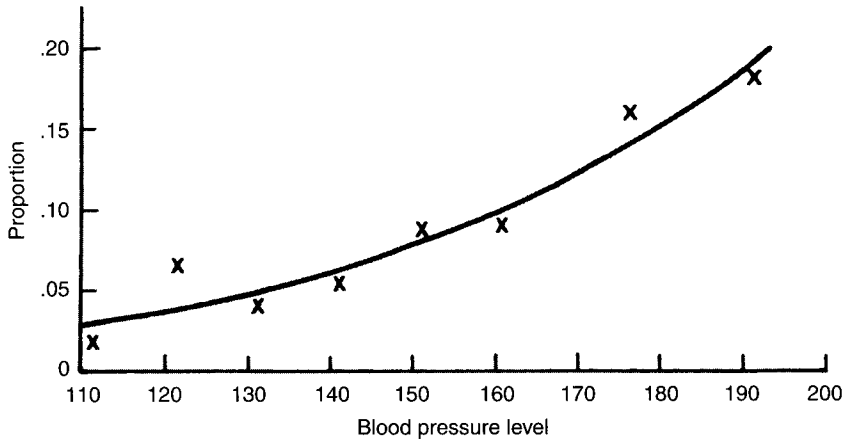

Figure 6.3 Sample proportions and estimated probabilities of heart disease for linear logit model.

with scores $\{x_i\}$ for systolic blood pressure level. We used scores (111.5, 121.5, 131.5, 141.5, 151.5, 161.5, 176.5, 191.5). The nonextreme scores are midpoints for the intervals of blood pressure. The trend in standardized residuals disappears for this model, and only the second category shows some evidence of lack of fit. A single relatively large residual is not surprising, however. With many residuals, a few may be large merely by chance. Here the overall fit statistics ($G^2 = 5.91$, $X^2 = 6.29$, with df = 6) do not indicate problems. In analyzing residual patterns, we should be cautious about attributing patterns to what might be chance variation from a model.

A useful graphical display for showing lack of fit compares sample and fitted proportions by plotting them against each other or by plotting both of them against explanatory variables. For the linear logit model, Figure 6.3 plots both the sample proportions and the estimated probabilities of heart disease against blood pressure. The fit seems decent.

Studying residuals helps us understand either why a model fits poorly or where there is lack of fit in a generally good-fitting model. The next example illustrates the second case.

6.2.3 Example: Admissions to Graduate School at Florida

Table 6.6 refers to graduate school applications for the 23 departments in the College of Liberal Arts and Sciences at the University of Florida during the 1997–1998 academic year. It cross-classifies the applicant's gender, department to which he or she applied, and whether he or she was admitted, which we treat as the response variable. For gender i in department k, let y_{ik} denote the number admitted and let π_{ik} denote the probability of admission. We treat $\{Y_{ik}\}$ as independent bin(n_{ik}, π_{ik}). Other things being equal, we would hope the admissions decision is independent of gender. The model with no gender effect, given the department, is

$$\text{logit}(\pi_{ik}) = \alpha + \beta_k^D.$$

However, this model fits rather poorly ($G^2 = 44.74$, $X^2 = 40.85$, df = 23).

The software output in Table 6.6 reports standardized residuals $\{r_i\}$ for the number of females who were admitted. For instance, the Astronomy department admitted 6 females,

Table 6.6 Graduate School Admissions by Gender and Department, with Standardized Residuals for Model of No Gender Effect

	Females		Males		Std. Res		Females		Males		Std. Res
Dept	Yes	No	Yes	No	(Fem, Yes)	Dept	Yes	No	Yes	No	(Fem, Yes)
anth	32	81	21	41	−0.76	ling	21	10	7	8	1.37
astr	6	0	3	8	2.87	math	25	18	31	37	1.29
chem	12	43	34	110	−0.27	phil	3	0	9	6	1.34
clas	3	1	4	0	−1.07	phys	10	11	25	53	1.32
comm	52	149	5	10	−0.63	poli	25	34	39	49	−0.23
comp	8	7	6	12	1.16	psyc	2	123	4	41	−2.27
engl	35	100	30	112	0.94	reli	3	3	0	2	1.26
geog	9	1	11	11	2.17	roma	29	13	6	3	0.14
geol	6	3	15	6	−0.26	soci	16	33	7	17	0.30
germ	17	0	4	1	1.89	stat	23	9	36	14	−0.01
hist	9	9	21	19	−0.18	zool	4	62	10	54	−1.76
lati	26	7	25	16	1.65						

Source: Data courtesy of Prof. James Booth.

which was 2.87 estimated standard deviations higher than the model predicted. Each department has only a single nonredundant standardized residual, because of marginal constraints for the model. The model has fit $\hat{\pi}_{ik} = (y_{1k} + y_{2k})/n_{+k}$, corresponding to an independence fit ($\hat{\pi}_{1k} = \hat{\pi}_{2k}$) in each partial table. Now,

$$y_{1k} - n_{1k}\hat{\pi}_{1k} = y_{1k} - n_{1k}\frac{(y_{1k} + y_{2k})}{n_{+k}} = \frac{n_{2k}}{n_{+k}}y_{1k} - \frac{n_{1k}}{n_{+k}}y_{2k} = -(y_{2k} - n_{2k}\hat{\pi}_{2k}).$$

Thus, standard errors of $(y_{1k} - n_{1k}\hat{\pi}_{1k})$ and $(y_{2k} - n_{2k}\hat{\pi}_{2k})$ are identical. The standardized residuals are identical in absolute value for males and females but of different sign. Astronomy admitted 3 males, and their standardized residual was −2.87; the number admitted was 2.87 estimated standard deviations lower than predicted.

Having a single nonredundant value r_i for each df is an advantage of standardized residuals over Pearson (or deviance) residuals. The model of conditional independence has df = 1 for each partial table. Only one bit of information exists about how the data depart from the model, yet the Pearson residual for males need not equal the Pearson residual for females in absolute value. The $\{r_i\}$ for females who were admitted in each department satisfy $\sum_{i=1}^{23} r_i^2 = X^2$, their squares giving 23 df = 1 components for the Pearson statistic. The 46 squared Pearson residuals would have the same sum, but each has null distribution smaller than χ_1^2.

Departments with large standardized residuals reveal the reason for the lack of fit. Significantly more females were admitted than the model predicts in the Astronomy and Geography departments, and fewer in the Psychology department. Without these three departments, the model fits reasonably well ($G^2 = 24.37$, $X^2 = 22.75$, df = 20).

For the complete data, adding a gender effect to the model does not provide an improved fit ($G^2 = 42.36$, $X^2 = 38.99$, df = 22), because the departments just mentioned have associations in different directions and of greater magnitude than other departments. This model has an ML estimate of 1.19 for the gender conditional odds ratio, the odds of admission being 19% higher for females than males, given department. By contrast, the marginal table collapsed over department has a sample odds ratio of 0.94, the overall odds of admission being 6% lower for females. This illustrates Simpson's paradox (Section 2.3.2),

the estimated conditional association having different direction than the estimated marginal association.

6.2.4 Influence Diagnostics for Logistic Regression

Other regression diagnostic tools are also helpful in assessing fit. These include plots of ordered standardized residuals against normal percentiles (Haberman 1973a) and analyses that describe an observation's influence on parameter estimates and fit statistics. Whenever a residual indicates that a model fits an observation poorly, it can be informative to delete the observation and refit the model to remaining ones. This is equivalent to adding a parameter to the model for that observation, forcing a perfect fit for it.

For ungrouped binary data, the notion of an outlier is not as clear as in ordinary regression. Copas (1988) used a probabilistic definition whereby, if the fitted model were true, the observation would be very unlikely to occur. But then, if $\hat{\pi}_i$ is close to 1 or close to 0 over certain regions of explanatory variable values, it is not at all surprising to observe some outliers. Copas studied how various models differ in their sensitivity to outliers.

As in ordinary regression, a single observation can be quite influential in determining parameter estimates. The greater an observation's leverage, the greater its potential influence. The fit could be quite different if an observation that appears to be an outlier on y and has large leverage is deleted. However, a single observation can have a much more exorbitant influence in ordinary least-squares regression than in logistic regression, since ordinary regression has no bound on the distance of y_i from its expected value. In Section 4.5.6 we observed that the GLM estimated hat matrix

$$\hat{\boldsymbol{H}}_{at} = \hat{\boldsymbol{W}}^{1/2}\boldsymbol{X}(\boldsymbol{X}'\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{W}}^{1/2}$$

depends on the fit as well as the model matrix $\boldsymbol{X}$. For logistic regression, recall (from Section 5.5.2) that the weight matrix $\hat{\boldsymbol{W}}$ is diagonal with element $\hat{w}_i = n_i\hat{\pi}_i(1-\hat{\pi}_i)$ for the n_i observations at setting i of predictors. Points that have extreme predictor values need not have high leverage. In fact, the leverage can be relatively small if $\hat{\pi}_i$ is close to 0 or 1.

Several measures describe the effect of removing an observation from the data set. They are related algebraically to the observation's leverage (Pregibon 1981, Williams 1987). In logistic regression, the observation could be a single binary response or a binomial response for a set of subjects all having the same predictor values (i.e., *ungrouped* or *grouped* data). For each observation, influence measures of deleting the observation include:

1. For each model parameter, the change in its estimate. This change, divided by its standard error, is called *Dfbeta*.
2. A measure of the change in a joint confidence interval for the parameters. This confidence interval displacement diagnostic is denoted by c.
3. The change in X^2 or G^2 goodness-of-fit statistics. Pregibon (1982) showed that the change in X^2 approximates the squared standardized residual for that observation.

For each measure, the larger the value, the greater the influence. With continuous or multiple predictors, it can be informative to plot these diagnostics, for instance, against the estimated probabilities.

Table 6.7 Diagnostic Measures for Logistic Regression Models Fitted to Heart Disease Data

Blood Pressure	*Dfbeta*	c	Pearson X^2 Diff.	Likelihood-Ratio G^2 Diff.	Pearson X^2 Diff.[a]	Likelihood-Ratio G^2 Diff.[a]
111.5	0.49	0.34	1.22	1.39	6.86	9.13
121.5	−1.14	2.26	5.64	5.04	0.02	0.02
131.5	0.33	0.31	0.89	0.94	4.08	4.56
141.5	0.08	0.09	0.33	0.34	0.55	0.57
151.5	0.01	0.00	0.02	0.02	0.70	0.66
161.5	−0.07	0.02	0.11	0.11	0.87	0.80
176.5	0.40	0.26	0.42	0.42	14.17	10.83
191.5	−0.12	0.02	0.03	0.03	9.41	6.73

[a]Independence model; other values refer to linear logit model with blood pressure predictor.

We illustrate the diagnostics using the linear logit model for Table 6.5, which has blood pressure as a predictor for heart disease. Table 6.7 contains simple approximations (due to Pregibon 1981) for the *Dfbeta* measure for the coefficient of blood pressure, the confidence interval diagnostic c, the change in G^2, and the change in X^2 (which is the square of the standardized residual, r_i^2). All their values show that deleting the second observation has the greatest effect. This is not surprising, as that observation has the only relatively large residual. By contrast, Table 6.7 also contains the changes in X^2 and G^2 for deleting observations in fitting the independence model. At the low and high ends of the blood pressure values, several changes are very large. However, these all relate to removing an entire binomial sample at a blood pressure level instead of removing a single subject's binary observation. Such subject-level (ungrouped data) deletions have little effect even for this model.

6.3 SUMMARIZING THE PREDICTIVE POWER OF A MODEL

In ordinary regression, R^2 describes the reduction in the conditional variation of the response compared with the marginal variation. It and the multiple correlation R describe how well the explanatory variables can predict the response, with $R = 1$ for perfect prediction. Despite various attempts to define analogs for categorical response models, no proposed measure is as widely useful as R and R^2. In this section we present a few ways proposed for summarizing predictive power.

6.3.1 Summarizing Predictive Power: *R* and *R*-Squared Measures

For any GLM, the correlation $R(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ between the observed responses $\{y_i\}$ and the model's fitted values $\{\hat{\mu}_i\}$ measures predictive power. For least-squares regression, R is the multiple correlation between Y and the predictors. An advantage of the correlation, relative to its square, is the appeal of working on the original scale and its approximate proportionality to effect size: For a small effect with a single predictor, doubling the slope corresponds approximately to doubling R.

In logistic regression with ungrouped data, $\hat{\mu}_i$ for a particular model is the estimated probability $\hat{\pi}_i$ for binary observation i. So, $R(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ is then the correlation between the n binary $\{y_i\}$ observations (1 or 0 for each) and the estimated probabilities. The highly discrete nature of y can suppress the range of possible R values. Nevertheless, R is useful for

comparing fits of different models for the same data. A caveat is that with many predictors the R estimates can become highly biased upwards in estimating the true correlation, $R(\boldsymbol{Y}, E(\boldsymbol{Y}|\boldsymbol{X}))$, so it can be misleading to compare sample R values for models with greatly different df values. A jackknife adjustment can reduce this bias (Zheng and Agresti 2000).

Another way to measure the association between the binary responses $\{y_i\}$ and their fitted values $\{\hat{\pi}_i\}$ uses the proportional reduction in squared error

$$1 - \frac{\sum_i (y_i - \hat{\pi}_i)^2}{\sum_i (y_i - \bar{y})^2},$$

obtained by using $\hat{\pi}_i$ instead of $\bar{y} = \sum_j y_j/n$ as a predictor of y_i (Efron 1978). Amemiya (1981) suggested a related measure that weights squared deviations by inverse predicted variances. For logistic regression, unlike normal GLMs, these and $R(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ need not be nondecreasing as the model gets more complex. Like any correlation-type measure, they can depend strongly on the range of observed values of explanatory variables, and as computed for sample data are biased upward as estimates of corresponding population measures. Bias corrections are possible (e.g., Liao and McGee 2003).

6.3.2 Summarizing Predictive Power: Likelihood and Deviance Measures

Other measures of predictive power directly use the likelihood function. Denote the maximized log likelihood by L_M for a given model, L_S for the saturated model, and L_0 for the null model containing only an intercept term. Probabilities are no greater than 1.0, so log likelihoods are nonpositive. As the model complexity increases, the parameter space expands, so the maximized log likelihood increases. Thus, $L_0 \leq L_M \leq L_S \leq 0$. The measure

$$\frac{L_M - L_0}{L_S - L_0} \tag{6.3}$$

falls between 0 and 1. It equals 0 when the model provides no improvement in fit over the null model, and it equals 1 when the model fits as well as the saturated model. A weakness is that the log likelihood is not an easily interpretable scale. Interpreting the numerical value is difficult, other than in a comparative sense for different models.

For N independent Bernoulli observations, the maximized log likelihood is

$$\log \prod_{i=1}^{N} [\hat{\pi}_i^{y_i}(1 - \hat{\pi}_i)^{1-y_i}] = \sum_{i=1}^{N} [y_i \log \hat{\pi}_i + (1 - y_i)\log(1 - \hat{\pi}_i)].$$

The null model gives $\hat{\pi}_i = (\sum_i y_i)/N = \bar{y}$, so that

$$L_0 = N[\bar{y}(\log\ \bar{y}) + (1 - \bar{y})\log(1 - \bar{y})].$$

The saturated model has a parameter for each subject and implies that $\hat{\pi}_i = y_i$ for all i. Thus, $L_S = 0$ and (6.3) simplifies to

$$D = \frac{L_0 - L_M}{L_0}.$$

McFadden (1974) proposed this measure.

Suppose there are multiple observations at each setting of explanatory variables. Then, the data file can take the grouped-data form of N binomial counts with binomial indices $\{n_i\}$, rather than the ungrouped form of N Bernoulli indicators each with $n_i = 1$. The saturated model then has a parameter for each count. It gives N fitted proportions equal to the N sample proportions of success. Then L_S is nonzero and (6.3) takes a different value than when calculated using individual subjects. For N binomial counts, the maximized likelihoods are related to the G^2 goodness-of-fit statistic by $G^2(M) = -2(L_M - L_S)$, so (6.3) becomes

$$D^* = \frac{G^2(0) - G^2(M)}{G^2(0)}.$$

Goodman (1971a) and Theil (1970) discussed this and related partial association measures.

With grouped data D^* can be large even when predictive power is weak at the subject level. For instance, a model can fit much better than the null model even though fitted probabilities are close to 0.50 for the entire sample. In particular, $D^* = 1$ when it fits perfectly, regardless of how well one can predict individual subjects' responses on Y with that model. Also, suppose that the population satisfies the given model, but not the null model. As the sample size $n = \sum_i n_i$ increases with number of settings N fixed, $G^2(M)$ behaves like a chi-squared random variable but $G^2(0)$ eventually grows unboundedly. Thus, $D^* \to 1$ (in probability) as $n \to \infty$, and its magnitude tends to depend on n. This measure confounds model goodness of fit with predictive power. Similar behavior occurs for R^2 in regression analyses when calculated using *means* of y values (rather than individual y values) at N different x settings. It is more sensible to use D for binary, ungrouped data.

6.3.3 Summarizing Predictive Power: Classification Tables

A *classification table* cross-classifies the binary response with a prediction of whether $y = 0$ or 1. The prediction for observation i is $\hat{y} = 1$ when $\hat{\pi}_i > \pi_0$ and $\hat{y} = 0$ when $\hat{\pi}_i \leq \pi_0$, for some cutoff π_0. One possibility is $\pi_0 = 0.50$. Another is the sample proportion of 1 outcomes, which is $\hat{\pi}_i$ for the model containing only an intercept term. Rather than using $\hat{\pi}_i$ from the model fitted to the data set of which y_i was one element, it is better to make the prediction with the "leave-one-out" cross-validation approach by which $\hat{\pi}_i$ is based on the model fitted to the other $n - 1$ observations.

Using a classification table, we can summarize the predictive power by

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

(Recall Section 2.1.3.) An overall summary of predictor power is the proportion of correct classifications. This estimates

$$\begin{aligned} P(\text{correct classification}) &= P(y = 1 \text{ and } \hat{y} = 1) + P(y = 0 \text{ and } \hat{y} = 0) \\ &= P(\hat{y} = 1 \mid y = 1)P(y = 1) + P(\hat{y} = 0 \mid y = 0)P(y = 0), \end{aligned}$$

which is a weighted average of sensitivity and specificity.

A classification table has limitations: It collapses continuous predictive values $\hat{\pi}$ into binary ones. The choice of π_0 is arbitrary. Results are sensitive to the relative numbers of

times that $y = 1$ and $y = 0$. For example, if a low proportion of observations have $y = 1$, the model fit may never have $\hat{\pi}_i > 0.50$, in which case one never predicts $\hat{y} = 1$. Again, the main use is for comparing different models with the same data.

6.3.4 Summarizing Predictive Power: ROC Curves

The classification table summaries depend on the cutoff π_0 for making classifications. A *receiver operating characteristic* (ROC) curve is a plot of sensitivity as a function of (1 − specificity) for the possible π_0. A ROC curve is more informative than a classification table, because it summarizes predictive power for all possible π_0. When π_0 is near 0, almost all predictions are $\hat{y} = 1$; then, sensitivity is near 1, specificity is near 0, and the point (1 − specificity, sensitivity) $\approx$ (1, 1). When π_0 is near 1, almost all predictions are $\hat{y} = 0$; then, sensitivity is near 0, specificity is near 1, and (1 − specificity, sensitivity) $\approx$ (0, 0). A ROC curve usually has a concave shape connecting the points (0, 0) and (1, 1).

For a given specificity, better predictive power corresponds to higher sensitivity. So, the better the predictive power, the higher the ROC curve. In a summary sense, the greater the area under the ROC curve, the better the predictions. In fact, the area under a ROC curve is identical to the value of another measure of predictive power, the *concordance index* (Hanley and McNeil 1982). Consider all pairs of observations (i, j) for which $y_i = 1$ and $y_j = 0$. The concordance index c is the proportion of such pairs for which $\hat{\pi}_i > \hat{\pi}_j$; that is, it is the relative frequency of the pairwise predictions and the outcomes being concordant, the observation with the larger y also having the larger $\hat{\pi}$. A value $c = 0.50$ means predictions are no better than random guessing. This corresponds to a model having only an intercept term and an ROC curve that is a straight line connecting points (0, 0) and (1, 1).

6.3.5 Example: Evaluating Predictive Power for Horseshoe Crab Data

Table 6.2 shows the correlation $R(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ for some models fitted to the horseshoe crab data for predicting whether a female crab had at least one satellite. Color alone (C) has $R = 0.285$, width alone (W) has $R = 0.402$, and using both ($C + W$) increases R to 0.452. The simpler model ($C = \text{dark} + W$) that uses color as binary merely to indicate whether a crab is dark does nearly as well, with $R = 0.447$. These models fit essentially as well as more complex models not shown in the table. For example, the model that adds an interaction term to the model ($C = \text{dark} + W$) has $R = 0.452$.

Other measures of predictive power have different magnitudes but similar results in comparing various models. For example, the concordance index $c = 0.639$ with model (C) (in factor form), 0.742 with model (W), 0.771 with model ($C + W$), 0.772 with model ($C = \text{dark} + W$), and 0.772 for the model that adds an interaction term to this model.

Next, we illustrate a classification table, for the model ($C + W$). Of the 173 crabs, 111 had a satellite, for a sample proportion of 0.642. Table 6.8 shows classification tables using $\pi_0 = 0.50$ and $\pi_0 = 0.642$ with cross-validated predictions. When $\pi_0 = 0.642$, from Table 6.8 the estimated sensitivity $= 74/111 = 0.667$ and specificity $= 42/62 = 0.677$. The proportion of correct classifications is $(74 + 42)/173 = 0.671$.

Figure 6.4 shows how PROC LOGISTIC in SAS reports the ROC curve for the model ($C + W$). When $\pi_0 = 0.642$, specificity $= 0.68$, sensitivity $= 0.67$, and the point plotted has coordinates (0.32, 0.67). The area under the curve is $c = 0.771$.

Table 6.8 Classification Tables for Horseshoe Crab Mating Data

	Prediction, $\pi_0 = 0.642$		Prediction, $\pi_0 = 0.500$		
Actual	$\hat{y} = 1$	$\hat{y} = 0$	$\hat{y} = 1$	$\hat{y} = 0$	Total
$y = 1$	74	37	94	17	111
$y = 0$	20	42	34	28	62

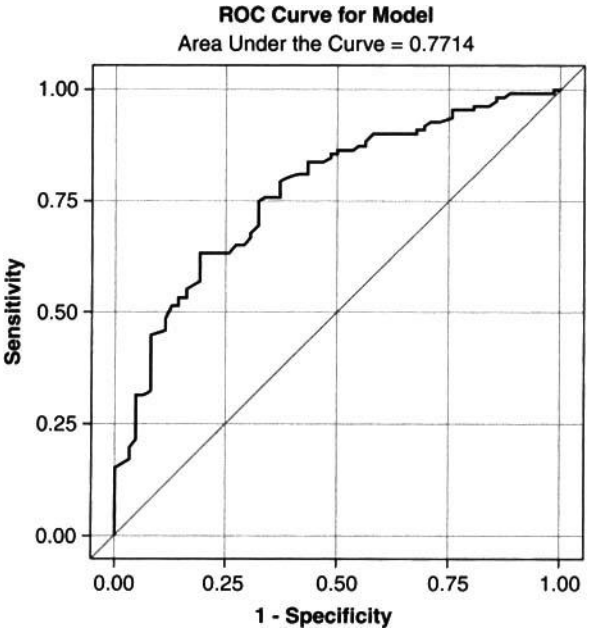

Figure 6.4 ROC curve (from SAS PROC LOGISTIC) for logistic regression model estimating the probability a crab has satellites, using width and color predictors.

6.4 MANTEL–HAENSZEL AND RELATED METHODS FOR MULTIPLE 2 × 2 TABLES

The analysis of the graduate admissions data in Section 6.2.3 used the model of conditional independence. This model is an important one in biomedical studies that investigate whether an association exists between a treatment variable and a disease outcome after adjusting for a possibly confounding variable that might influence that association. We next present the test of conditional independence as a logistic model analysis for a $2 \times 2 \times K$ contingency table. We also present a test and a related estimation method, due to Mantel and Haenszel (1959), that seem non-model-based but relate to the same logistic model.

We illustrate using Table 6.9, showing results of a clinical trial with eight centers. The study compared two cream preparations, an active drug and a control, on their success in curing an infection. This table illustrates a common pharmaceutical application, comparing two treatments on a binary response with observations from several strata. The strata are

Table 6.9 Clinical Trial Relating Treatment to Response for Eight Centers, with Expected Value and Variance (of Success Count for Drug) Under Conditional Independence

		Response				
Center	Treatment	Success	Failure	Odds Ratio	μ_{11k}	var(n_{11k})
1	Drug	11	25	1.19	10.36	3.79
	Control	10	27			
2	Drug	16	4	1.82	14.62	2.47
	Control	22	10			
3	Drug	14	5	4.80	10.50	2.41
	Control	7	12			
4	Drug	2	14	2.29	1.45	0.70
	Control	1	16			
5	Drug	6	11	∞	3.52	1.20
	Control	0	12			
6	Drug	1	10	∞	0.52	0.25
	Control	0	10			
7	Drug	1	4	2.0	0.71	0.42
	Control	1	8			
8	Drug	4	2	0.33	4.62	0.62
	Control	6	1			

Source: Beitler and Landis (1985).

often medical centers or clinics; or, they may be levels of age or severity of the condition being treated; or, they may be combinations of levels of several control variables; or, they may be different studies of the same sort summarized in a meta-analysis.

6.4.1 Using Logistic Models to Test Conditional Independence

For a binary response Y, we analyze the effect of a binary predictor X, conditional on the category of a qualitative covariate Z. Let $\pi_{ik} = P(Y = 1|X = i, Z = k)$. Consider the model

$$\text{logit}(\pi_{ik}) = \alpha + \beta x_i + \beta_k^Z, \quad i = 1, 2, \quad k = 1, \ldots, K, \tag{6.4}$$

where $x_1 = 1$ and $x_2 = 0$. This model assumes that the XY conditional odds ratio is the same at each category of Z, namely, $\exp(\beta)$. The null hypothesis of XY conditional independence is H_0: $\beta = 0$. The Wald statistic is $(\hat{\beta}/SE)^2$. The likelihood-ratio statistic is the difference between deviance statistics for the reduced model

$$\text{logit}(\pi_{ik}) = \alpha + \beta_k^Z \tag{6.5}$$

and the full model. These tests are sensible when X has a similar effect at each category of Z. They have df $= 1$.

Alternatively, since the reduced model (6.5) is equivalent to conditional independence of X and Y, we can test conditional independence using a goodness-of-fit test of that model. Such a test has df $= K$ when X is binary. This corresponds to comparing model (6.5) and the saturated model, which permits $\beta \neq 0$ in (6.4) and also contains $(K - 1)$ XZ

interaction parameters. The likelihood-ratio test statistic partitions into two components, the likelihood-ratio statistic with df = 1 for testing H_0: $\beta = 0$ in model (6.4) and the likelihood-ratio statistic with df = $(K - 1)$ for testing the fit of model (6.4) and thus equality of the K odds ratios (Goodman 1969, Cheng et al. 2010).

When no interaction exists or when the conditional XY association has relatively little variation among the levels of Z, it follows from results in Section 5.3.7 that the approach using df = K of testing conditional independence is less powerful, especially when K is large. When model (6.4) holds, both tests have the same noncentrality. Thus, the test of $\beta = 0$ in model (6.4) is more powerful, since it has fewer degrees of freedom. However, when the direction of the conditional XY association varies among categories of Z, it can be less powerful.

6.4.2 Cochran–Mantel–Haenszel Test of Conditional Independence

Mantel and Haenszel (1959) proposed a non-model-based test of H_0: conditional independence in $2 \times 2 \times K$ tables. Focusing on retrospective studies of disease, they treated response (column) marginal totals as fixed. Thus, in each partial table k of cell counts $\{n_{ijk}\}$, their analysis conditioned on both the treatment (e.g., group) totals $\{n_{1+k}, n_{2+k}\}$ and the response outcome totals $\{n_{+1k}, n_{+2k}\}$. The usual sampling schemes then yield a hypergeometric distribution (3.17) for the first cell count n_{11k} in each partial table. That count determines $\{n_{12k}, n_{21k}, n_{22k}\}$, given the marginal totals.

Under H_0, the hypergeometric mean and variance of n_{11k} are

$$\mu_{11k} = E(n_{11k}) = n_{1+k}n_{+1k}/n_{++k},$$
$$\text{var}(n_{11k}) = n_{1+k}n_{2+k}n_{+1k}n_{+2k}/[n_{++k}^2(n_{++k} - 1)].$$

Cell counts from different partial tables are independent. The test statistic combines information from the K tables by comparing $\sum_k n_{11k}$ to its null expected value. It equals

$$\text{CMH} = \frac{[\sum_k (n_{11k} - \mu_{11k})]^2}{\sum_k \text{var}(n_{11k})}. \tag{6.6}$$

This statistic has a large-sample chi-squared null distribution with df = 1.

When the odds ratio $\theta_{XY(k)} > 1$ in partial table k, we expect that $(n_{11k} - \mu_{11k}) > 0$. When $\theta_{XY(k)} > 1$ in every partial table or $\theta_{XY(k)} < 1$ in each table, $\sum_k (n_{11k} - \mu_{11k})$ tends to be relatively large in absolute value. This test works best when the conditional XY association is similar in each partial table. In this sense it is similar to the tests of H_0: $\beta = 0$ in logistic model (6.4). When the sample sizes in the strata are moderately large, this test usually gives similar results. In fact, it is a score test of H_0: $\beta = 0$ in that model (Birch 1964b, 1965, Darroch 1981, Day and Byar 1979).

Cochran (1954) proposed a similar test statistic. He treated the rows in each 2×2 table as two independent binomials rather than a hypergeometric. Cochran's statistic is (6.6) with $\text{var}(n_{11k})$ replaced by

$$\text{var}(n_{11k}) = n_{1+k}n_{2+k}n_{+1k}n_{+2k}/n_{++k}^3.$$

Because of the similarity in their approaches, we call (6.6) the *Cochran–Mantel–Haenszel* (CMH) *statistic*. The Mantel and Haenszel approach using the hypergeometric is more

general in that it also applies to some cases in which the rows are not independent binomial samples from two populations. Examples are (1) retrospective studies and (2) randomized clinical trials with the available subjects (usually volunteers) randomly allocated to two treatments. In the first case the column totals are naturally fixed. In the second, under the null hypothesis the column margins are the same regardless of how subjects are assigned to treatments, and randomization arguments lead to the hypergeometric for each 2×2 table.

Mantel and Haenszel (1959) proposed (6.6) but with a continuity correction. The P-value from the test then better approximates an exact conditional test, based directly on the convolution of the hypergeometric distributions rather than the chi-squared approximation (Section 7.3.5). However, that test tends to be conservative. Mantel and Fleiss (1980) stated that the asymptotic approximation for this test is adequate if the potential values for $\sum(n_{11k} - \mu_{11k})$, for the fixed margins in each 2×2 table, can exceed ± 5. The CMH statistic generalizes for $I \times J \times K$ tables (Section 8.4.3).

6.4.3 Example: Multicenter Clinical Trial Revisited

For the multicenter clinical trial introduced at the beginning of Section 6.4, Table 6.9 reports the sample odds ratio for each table and the expected value and variance of the number of successes for the drug treatment (n_{11k}) under H_0: conditional independence. In each table except the last, the sample odds ratio shows a positive association. Thus, it makes sense to combine results using CMH $= 6.38$, with df $= 1$. There is considerable evidence against H_0 ($P = 0.012$).

Similar results occur in testing H_0: $\beta = 0$ in logistic model (6.4). The model fit has $\hat{\beta} = 0.777$ with $SE = 0.307$. The Wald statistic is $(0.777/0.307)^2 = 6.42$ ($P = 0.011$). The likelihood-ratio statistic equals 6.67 ($P = 0.010$).

6.4.4 CMH Test Is Advantageous for Sparse Data

In summary, for the main-effects logistic model (6.4), the CMH statistic is the score statistic alternative to the likelihood-ratio or Wald test of H_0: $\beta = 0$. As $n \to \infty$ with fixed K, all three tests have the same asymptotic chi-squared behavior under H_0. An advantage of the CMH statistic is that its chi-squared limit also applies with an alternative asymptotic scheme in which $K \to \infty$ as $n \to \infty$. The asymptotic theory for likelihood-ratio and Wald tests requires the number of parameters (and hence K) to be fixed, so it does not apply to this scheme.

Here is an application of this type: Suppose each stratum has a single matched pair of subjects, one in each group. Then, $n_{1+k} = n_{2+k} = 1$ for each k and $n = 2K$, so $K \to \infty$ as $n \to \infty$. Table 6.10 shows the data layout for this situation. When both subjects in stratum k make the same response, as in the first case in Table 6.10, $n_{+1k} = 0$ or $n_{+2k} = 0$. Given the marginal counts, the internal counts are then completely determined, and $\mu_{11k} = n_{11k}$ and $\text{var}(n_{11k}) = 0$. When the subjects make differing responses, as in the second

Table 6.10 Two Examples of a Stratum Containing a Matched Pair

	Response		Response	
Element of Pair	Success	Failure	Success	Failure
First	1	0	1	0
Second	1	0	0	1

case, $n_{+1k} = n_{+2k} = 1$, so that $\mu_{11k} = 0.50$ and $\text{var}(n_{11k}) = 0.25$. Thus, a matched pair contributes to the CMH statistic only when the two subjects' responses differ. Let K^* denote the number of the K tables that satisfy this. Although each n_{11k} can take only two values, the central limit theorem implies that $\sum_k n_{11k}$ is approximately normal for large K^*. Then, the distribution of CMH is approximately chi-squared.

Usually, when K grows with n, each stratum has few observations, so the full table is sparse. There may be more than two observations, such as case–control studies that match several controls with each case. The nonstandard setting in which $K \to \infty$ as $n \to \infty$ is called *sparse-data asymptotics*. Ordinary ML estimation then breaks down because the number of parameters is not fixed, instead having the same order as the sample size. In particular, the chi-squared approximation is good for the likelihood-ratio and Wald statistics for testing conditional independence when K is fixed and small relative to n and the strata marginal totals mostly exceed about 5 to 10.

6.4.5 Estimation of Common Odds Ratio

It is more informative to estimate the strength of association than to test hypotheses about it. When the association seems stable among partial tables, we can combine the K sample odds ratios into a summary measure of conditional association. The logistic model (6.4) implies homogeneous association, $\theta_{XY(1)} = \cdots = \theta_{XY(K)} = \exp(\beta)$. The ML estimate of the common odds ratio is $\exp(\hat{\beta})$.

Other estimators of a common odds ratio are not model-based. Woolf (1955) proposed an exponentiated weighted average of the K sample log odds ratios. Let $p_{ij|k} = n_{ijk}/n_{++k}$. Mantel and Haenszel (1959) proposed

$$\hat{\theta}_{\text{MH}} = \frac{\sum_k (n_{11k} n_{22k}/n_{++k})}{\sum_k (n_{12k} n_{21k}/n_{++k})} = \frac{\sum_k n_{++k} p_{11|k} p_{22|k}}{\sum_k n_{++k} p_{12|k} p_{21|k}}. \tag{6.7}$$

This gives more weight to strata with larger sample sizes. With fixed K, $\log(\hat{\theta}_{\text{MH}})$ is slightly less efficient than the ML estimator $\hat{\beta}$ unless $\beta = 0$ (Tarone et al. 1983). However, it is preferred over the ML estimator when K is large and the data are very sparse. The ML estimator $\hat{\beta}$ of the log odds ratio then tends to be too large in absolute value. For sparse-data asymptotics with only a single matched pair in each stratum, for instance, $\hat{\beta} \xrightarrow{p} 2\beta$. (see Exercise 11.29.)

Robins et al. (1986) derived an estimated variance for $\log(\hat{\theta}_{\text{MH}})$ that applies both for standard asymptotics with large n and fixed K and for sparse-data asymptotics in which K is also large. Expressing $\hat{\theta}_{\text{MH}} = R/S = \left(\sum_k R_k\right) / \left(\sum_k S_k\right)$ with $R_k = n_{11k} n_{22k}/n_{++k}$, their derivation showed that $(\log \hat{\theta}_{\text{MH}} - \log \theta)$ is approximately proportional to $(R - \theta S)$. They also showed that $E(R - \theta S) = 0$ and derived the variance of $(R - \theta S)$. Their result is

$$\begin{aligned}\hat{\sigma}^2[\log \hat{\theta}_{\text{MH}}] = {} & \frac{1}{2R^2} \sum_k n_{++k}^{-1} (n_{11k} + n_{22k}) R_k \\ & + \frac{1}{2S^2} \sum_k n_{++k}^{-1} (n_{12k} + n_{21k}) S_k \\ & + \frac{1}{2RS} \sum_k n_{++k}^{-1} [(n_{11k} + n_{22k}) S_k + (n_{12k} + n_{21k}) R_k].\end{aligned}$$

For the eight-center clinical trial summarized by Table 6.9,

$$\hat{\theta}_{MH} = \frac{\sum_k (n_{11k}n_{22k}/n_{++k})}{\sum_k (n_{12k}n_{21k}/n_{++k})} = \frac{(11 \times 27)/73 + \cdots + (4 \times 1)/13}{(25 \times 10)/73 + \cdots + (2 \times 6)/13} = 2.13.$$

For $\log \hat{\theta}_{MH} = 0.758$, $\hat{\sigma}[\log \hat{\theta}_{MH}] = 0.303$. A 95% confidence interval for the common odds ratio is $\exp(0.758 \pm 1.96 \times 0.303)$ or (1.18, 3.87). Similar results occur using model (6.4). The 95% confidence interval for $\exp(\beta)$ is $\exp(0.777 \pm 1.96 \times 0.307)$, or (1.19, 3.97), using the Wald interval, and (1.20, 4.02) using the likelihood-ratio interval. Although the evidence of an effect is considerable, inference about its size is rather imprecise with such a small sample. The odds of success may be as little as 20% higher with the drug, or they may be as much as four times as high.

If the true odds ratios are not identical but do not vary much, $\hat{\theta}_{MH}$ still is a useful summary of the conditional associations. Similarly, the CMH test is a powerful summary of evidence against H_0: conditional independence, as long as the sample associations fall primarily in a single direction. It is not necessary to assume equality of odds ratios to use the CMH test or $\hat{\theta}_{MH}$.

6.4.6 Meta-analyses for Summarizing Multiple 2 × 2 Tables

A *meta-analysis* is a statistical analysis that combines information from several studies. For comparing two treatments on a binary response, the analysis refers to a $2 \times 2 \times K$ table, one 2×2 table for each study. For a particular effect measure, such as the odds ratio or a difference of proportions, here we consider the simplifying assumption that the population values of the measure are identical in each study. This is usually unrealistic, but is often adequate for providing a simple summary of the effect when the true effect does not vary much among studies. Sections 6.4.10 and 13.3.6 generalize to allow for heterogeneity among the effects.

Consider first the significance test of the null hypothesis of no effect, that is, conditional independence between the treatment and the response for each study. The logistic model (6.4) is a natural one for such an analysis. We test H_0: $\beta = 0$ using the likelihood-ratio test or the Cochran–Mantel–Haenszel (CMH) test (6.6). As mentioned in Section 6.4.4, the CMH test is advantageous for highly sparse data. When asymptotics are unsuitable even for that test, we can use a small-sample generalization of Fisher's exact test to multiple 2×2 tables, as presented in Section 7.3.5. For the CMH test or for the small-sample test, tables for which there are either no successes or no failures provide no information about whether there is truly an association and make no contribution to the test. (Recall that Section 6.4.4 discussed this for matched pairs.) There is no reason to use some device such as adding a small constant to cells of the table so those tables enter the analysis, because they are uninformative about the odds ratio (Agresti and Hartzel 2000).

Consider next summarizing the size of the effect. For the logistic model (6.4), we can use the ML estimate of the odds ratio $\exp(\beta)$ and a corresponding confidence interval. For highly sparse data, we can instead use the Mantel–Haenszel estimate $\hat{\theta}_{MH}$ and its corresponding interval. A small-sample interval can guarantee a lower bound for the coverage probability (Section 16.6.6). For all such frequentist analyses, tables for which there are either no successes or no failures provide no information about the size of the common odds ratio and do not contribute to the estimate.

6.4.7 Meta-analyses for Multiple 2 x 2 Tables: Difference of Proportions

The difference of proportions and the relative risk are alternative effect measures that are simpler to interpret than the odds ratio. A common difference of proportions for each study is the parameter δ in a model

$$\pi_{ik} = \alpha + \delta x_i + \beta_k^Z, \quad i = 1, 2, \quad k = 1, \ldots, K,$$

that replaces the logit link in model (6.4) by the identity link.

Mantel–Haenszel-type estimates are also available for such measures. In stratum k, denote the binomial "success" counts by $s_k = n_{11k}$ and $t_k = n_{21k}$ based on sample sizes $m_k = n_{1+k}$ and $n_k = n_{2+k}$, and let $N_k = m_k + n_k$. With $w_k = m_k n_k / N_k$, the estimator of a common difference of proportions is the weighted average of the stratum-specific estimates $\hat{\delta}_k = [(s_k/m_k) - (t_k/n_k)]$,

$$\hat{\delta}_{\text{MH}} = \left(\sum_k w_k \hat{\delta}_k\right) \Big/ \left(\sum_k w_k\right) \tag{6.8}$$

(Greenland and Robins 1985). An estimated variance,

$$\hat{\sigma}^2(\hat{\delta}_{\text{MH}}) = \left[\hat{\delta}_{\text{MH}}\left(\sum_k P_k\right) + \left(\sum_k Q_k\right)\right] \Big/ \left(\sum_k w_k\right)^2,$$

with

$$P_k = [m_k^2 t_k - n_k^2 s_k + m_k n_k (n_k - m_k)/2]/N_k^2,$$
$$Q_k = [s_k(n_k - t_k) + t_k(m_k - s_k)]/2N_k,$$

applies under both standard and sparse-data asymptotics (Sato 1989).

Under standard asymptotics, the ML model-based estimator is preferred because it is more efficient. However, ML fitting difficulties often arise when both probabilities are near 0 or near 1, and the $\{\pi_{ik}\}$ must be constrained to fall between 0 and 1. Here is an alternative approach that is then asymptotically efficient and does not have boundary problems: Express the score or profile likelihood $100(1 - \alpha)\%$ confidence interval for the difference of proportions (see Section 3.2.5) for study k alone as $d_k \pm z_{\alpha/2} s_k$, where d_k is the midpoint of that interval (i.e., *not* the sample difference of proportions $\hat{\delta}_k$) and s_k is a "pseudo standard error" that is taken to be the width of the interval divided by $2z_{\alpha/2}$. Then, taking weight $w_k = [1/(s_k^2)]/[\sum_i 1/(s_i^2)]$, we form $\hat{\delta} = \sum_k w_k d_k$, $SE = [\sum_k 1/(s_k^2)]^{-1/2}$, and the summary interval $\hat{\delta} \pm z_{\alpha/2}(SE)$. Unlike Wald methods, this does not require using unreliable sample standard errors from each study but merely uses a midpoint and width based on information obtained from the likelihood function.

To illustrate, the eight-center clinical trial data of Table 6.9 was analyzed in Sections 6.4.3 and 6.4.5 with CMH methods and with logistic model (6.4). For summarizing the effect by a common difference of success proportions between drug and control, the Mantel–Haenszel-type estimate (6.8) is $\hat{\delta}_{\text{MH}} = 0.130$ with $SE = 0.050$. Using the alternative method just mentioned that combines information from the eight center-specific score confidence intervals, we get $\hat{\delta} = 0.128$, $SE = 0.049$, and a 95% confidence interval for a common difference of proportions of (0.032, 0.224).

For the difference of proportions, tables for which there are either no successes or no failures provide no information about whether the true common value δ is nonzero (i.e., the significance testing problem) but they do give information about the magnitude of the effect. If each treatment, for example, has a very large number of failures and no successes, then we have evidence that both population proportions are close to 0 and that the difference is small. Thus, such data do have an impact on practical significance. (See Exercise 6.33 for an illustration.)

Agresti and Hartzel (2000) discussed ways of summarizing information from multiple tables and gave many additional references. Tian et al. (2009) proposed an alternative approach designed for small-sample cases in which some centers may have no outcomes of a particular type.

6.4.8 Collapsibility and Logistic Models for Contingency Tables

We have seen that conditional associations in partial tables usually differ from marginal associations. Under certain *collapsibility conditions* given in Section 2.3.6, however, they are the same. For odds ratios, recall that for three-way tables, XY marginal and conditional odds ratios are identical if either Z and X are conditionally independent or if Z and Y are conditionally independent.

For instance, suppose that a clinical trial studies the association between a binary treatment variable X ($x_1 = 1, x_2 = 0$) and a binary response Y, using data from K centers (Z). The logistic model (6.4), namely,

$$\text{logit}(\pi_{ik}) = \alpha + \beta x_i + \beta_k^Z, \quad i = 1, 2, \quad k = 1, \ldots, K, \tag{6.9}$$

has the same treatment effect β for each center. Since the model has no restriction on the conditional association of Z with X or with Y, this effect may differ after collapsing the $2 \times 2 \times K$ table over centers. The estimated XY conditional odds ratio, $\exp(\hat{\beta})$, typically differs from the sample odds ratio in the marginal 2×2 table.

Next, consider the simpler model that lacks center effects, $\text{logit}(\pi_{ik}) = \alpha + \beta x_i$. This states that, for each treatment, the success probability is identical for each center. The model satisfies a collapsibility condition for the XY association, because it states that Z is conditionally independent of Y, given X. So, when center effects are negligible and the simpler model fits nearly as well, the estimated treatment effect is approximately the marginal XY odds ratio.

6.4.9 Testing Homogeneity of Odds Ratios

The homogeneous association condition $\theta_{XY(1)} = \cdots = \theta_{XY(K)}$ for $2 \times 2 \times K$ tables is equivalent to logistic model (6.9). A test of homogeneous association is implicitly a goodness-of-fit test of this model. The usual G^2 and X^2 test statistics provide this, with $\text{df} = K - 1$. They test that the $K - 1$ parameters in the saturated model that are the coefficients of interaction terms [cross-products of the indicator variable for X with $(K - 1)$ indicator variables for categories of Z] all equal 0.

For the eight-center clinical trial data in Table 6.9, $G^2 = 9.75$ and $X^2 = 8.03$ ($\text{df} = 7$) do not contradict the hypothesis of equal odds ratios. It is reasonable to summarize the conditional association by a single odds ratio (e.g., $\hat{\theta}_{\text{MH}} = 2.13$ or $e^{\hat{\beta}} = 2.17$) for all eight partial tables.

6.4.10 Summarizing Heterogeneity in Odds Ratios

In practice, the effect of interest is often similar from stratum to stratum. In multicenter clinical trials comparing a new drug to a standard, for example, if the new drug is truly more beneficial, the population effect is usually positive in each stratum.

In strict terms, however, a model with homogeneous effects is unrealistic. Consider the odds ratio, to illustrate. First, we rarely expect the true odds ratio to be *exactly* the same in each stratum, because of unmeasured covariates that affect it. Breslow (1976) discussed modeling of the log odds ratio using a set of explanatory variables. Second, the model regards the strata effects $\{\beta_k^Z\}$ as fixed effects, treating them as the only strata of interest. Often the strata are merely a sampling of the possible ones. Multicenter clinical trials have data for certain centers but many other centers could have been chosen. Scientists would like their conclusions to apply to all such centers, not only those in the study.

A somewhat different logistic model treats the true log odds ratios in the partial tables as a random sample from a $N(\mu, \sigma^2)$ distribution. Fitting the model yields an estimated mean log odds ratio and an estimated variability about that mean. The inference applies to the population of strata rather than only those sampled. This type of model uses *random effects* in the linear predictor to induce this extra type of variability. In Chapter 13, we discuss GLMs with random effects, and in Section 13.3.5 we fit such a model to Table 6.9.

6.4.11 Propensity Scores in Observational Studies

We finish this section by mentioning a more challenging setting for analyzing conditional associations – observational studies in which we want to compare two groups while controlling for possibly confounding variables $\boldsymbol{x}$. Rosenbaum and Rubin (1983) proposed methods of adjusting for bias in making such comparisons. They defined the *propensity* as the probability of being in one group, for a given setting of the explanatory variables $\boldsymbol{x}$. They used logistic regression to estimate how propensity depends on $\boldsymbol{x}$. In comparing the groups on the response variable, they showed how to control for differing distributions of the groups on $\boldsymbol{x}$ by adjusting for the estimated propensity. This is done by using the propensity to match samples from the groups or to subclassify subjects into several strata consisting of intervals of propensity scores or to adjust directly by entering the propensity in the model.

For any study that is observational rather than randomized, there is still the limitation that propensity score methods adjust only for observed confounding covariates and not for unobserved ones. Also, the methods work better in larger samples, so observed covariates tend to be more truly balanced in the subclassifications. In various writings, Rubin has pointed out that confidence in causal conclusions based on such methods must rely on how consistent the results are with other evidence and how sensitive the conclusions are to reasonable deviations such as in the effects of unobserved covariates.

6.5 DETECTING AND DEALING WITH INFINITE ESTIMATES

The log-likelihood function for logistic regression models is strictly concave. ML estimates exist and are unique except in certain boundary cases. Estimates do not exist or may be infinite when there is no overlap in the sets of explanatory variable values having $y = 0$ and having $y = 1$ (Albert and Anderson 1984).

6.5.1 Complete or Quasi-complete Separation

The space of explanatory variable values is said to have *complete separation* when a hyperplane can pass through that space such that on one side of that hyperplane $y_i = 0$ for all observations, whereas on the other side, $y_i = 1$ always. This means that there exists a vector $\boldsymbol{b}$ such that

$$\boldsymbol{b}^T \boldsymbol{x}_i > 0 \text{ whenever } y_i = 1,$$
$$\boldsymbol{b}^T \boldsymbol{x}_i < 0 \text{ whenever } y_i = 0.$$

There is then *perfect discrimination*, as we can predict the sample outcomes perfectly by knowing the predictor values.

Figure 6.5 illustrates for a single explanatory variable. Here, $y = 0$ at $x = 10, 20, 30, 40$, and $y = 1$ at $x = 60, 70, 80, 90$. For $\boldsymbol{x}_i = (1, x_i)^T$, the predictor $\boldsymbol{b}^T \boldsymbol{x}_i = -50 + x_i$ [i.e., $\boldsymbol{b}^T = (-50, 1)$] gives perfect predictions. An ideal fit has $\hat{\pi} = 0$ for $x < 50$ and $\hat{\pi} = 1$ for $x > 50$. By letting $\hat{\beta} \to \infty$ and, for fixed $\hat{\beta}$, letting $\hat{\alpha} = -\hat{\beta}(50)$ so that $\hat{\pi} = 0.50$ at $x = 50$, we can generate a sequence with ever-increasing value of the likelihood function that comes successively closer to a perfect fit.

In practice, most software fails to recognize when some ML estimates are actually infinite. After a few cycles of iterative fitting, the log likelihood looks flat at the working estimate, and convergence criteria are satisfied. Because the log likelihood is so flat and because the variance of $\hat{\beta}_j$ comes from the negative inverse of the matrix of second derivatives, software typically reports huge standard errors. For the data in Figure 6.5, for instance, PROC GENMOD in SAS reports $\text{logit}(\hat{\pi}) = -192.2 + 3.8x$ with standard errors of 8.0×10^8 and 1.5×10^7.

In practice, an indication of complete separation is when the fitted prediction equation perfectly predicts the response outcome for the entire data, giving $\hat{\pi} = 1.0$ (to many decimal places) whenever $y = 1$ and $\hat{\pi} = 0.0$ whenever $y = 0$. A related indication is that the reported maximized log-likelihood value is 0 to many decimal places. Another warning signal is when standard errors seem unnaturally large. When there is indication of complete separation for a model containing several predictors, using the forward selection algorithm can reveal a subset of them for which complete separation occurs once they are all used.

A weaker condition that causes at least one estimate to be infinite, called *quasi-complete separation*, occurs when a hyperplane separates explanatory variable values with $y = 1$ and with $y = 0$, but cases exist with both outcomes on that hyperplane. For example, this

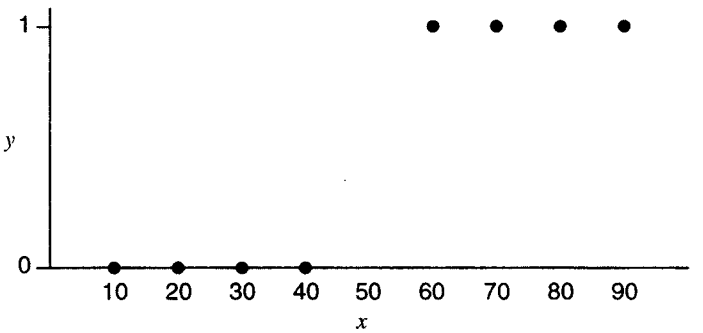

Figure 6.5 Perfect discrimination resulting in an infinite logistic regression parameter estimate.

happens if we add to Figure 6.5 two observations at $x = 50$, one with $y = 1$ and one with $y = 0$. With quasi-complete separation, there is not perfect discrimination for all observations. The maximized log likelihood is then strictly less than 0. An indication of quasi-complete separation is that some observations have $\hat{\pi} = 1.0$ or 0.0. Again, a warning signal is when reported standard errors seem unnaturally large.

When complete or quasi-complete separation do not occur, all ML estimates are finite and unique. Quasi-complete separation is more common than complete separation. It is more liable to happen with qualitative predictors than quantitative predictors. If any category of a qualitative predictor has either no cases with $y = 0$ or no cases with $y = 1$, there is quasi-complete separation when that variable is entered as a factor in the model (i.e., using an indicator variable for that category). With many predictors, it's a good idea to cross-classify each qualitative predictor with y to check for an empty cell, which is a sufficient condition for quasi-complete separation.

With an infinite estimate, Wald inference is worthless. By contrast, we can still compute likelihood-ratio and score tests and invert them to get a confidence interval. For example, the likelihood still has a maximized value at the infinite estimate for a parameter, so we can compare its value to the value when the parameter is equated to some fixed value such as zero. For the data in Figure 6.5, the likelihood-ratio test statistic for H_0: $\beta = 0$ is 11.09 ($df = 1, P = 0.001$), and the 95% confidence interval for β is $(0.06, \infty)$, so we can conclude that the effect is positive in the population.

6.5.2 Example: Multicenter Clinical Trial with Few Successes

Table 6.11 shows results of a clinical trial conducted at five centers. The purpose was to compare an active drug to placebo for treating fungal infections, with a binary (success, failure) response. For these data, let Y = response, X = treatment (1 = active drug, 0 = placebo), and Z = center.

Table 6.11 Clinical Trial Relating Treatment to Response, Showing also *XY* and *YZ* Marginal Tables

		Response (Y)		YZ Marginal	
Center (Z)	Treatment (X)	Success	Failure	Success	Failure
1	Active drug	0	5		
	Placebo	0	9	0	14
2	Active drug	1	12		
	Placebo	0	10	1	22
3	Active drug	0	7		
	Placebo	0	5	0	12
4	Active drug	6	3		
	Placebo	2	6	8	9
5	Active drug	5	9		
	Placebo	2	12	7	21
XY	Active drug	12	36		
marginal	Placebo	4	42		

Source: Data courtesy of Diane Connell, Sandoz Pharmaceuticals Corporation.

Centers 1 and 3 had no successes. Thus, the 5×2 YZ marginal table relating response to center collapsed over treatment, shown on the right side of Table 6.11, contains zero counts. Infinite ML estimates occur for terms in logistic models relating to the YZ association. An example is the model

$$\text{logit}(\pi_{ik}) = \beta x_i + \beta_k^Z.$$

[We take out the intercept from (6.9), so the $\{\beta_k^Z\}$ need no constraint; then, these refer to each center's effect rather than contrasts between each center and a baseline center.] The likelihood function increases continually as β_1^Z and β_3^Z decrease toward $-\infty$; that is, as the logit decreases toward $-\infty$, so the fitted probability of success decreases toward the ML estimate of 0 for those centers.

Because of the infinite estimates, we cannot conduct a Wald test of the center effects in Table 6.11. However, SAS (PROC GENMOD) reports a maximized log-likelihood value of -28.87 for this model and -40.58 when the center term is removed from the model, so the likelihood-ratio statistic for this effect equals 23.42 (df = 4).

The counts in the 2×2 marginal table relating response to treatment, shown in the bottom panel of Table 6.11, are all positive. The empty cells affect the center estimates, but not the treatment estimate, for this model. In the limit as the log likelihood increases, the fitted values have a log odds ratio $\hat{\beta} = 1.55$ ($SE = 0.70$). Most software reports this but, instead of $\hat{\beta}_1^Z = \hat{\beta}_3^Z = -\infty$, reports large numbers with extremely large standard errors. For instance, PROC GENMOD in SAS reports values of about -26 for $\hat{\beta}_1^Z$ and $\hat{\beta}_3^Z$, with standard errors of about 200,000.

The treatment estimate $\hat{\beta} = 1.55$ also results when we delete centers 1 and 3 from the analysis. When a center contains responses of only one type, it provides no information about this odds ratio. (It does provide information about the size of some other measures, such as the difference of proportions, as discussed above in Section 6.4.6.) Such tables also make no contribution to standard tests of conditional independence, such as the Cochran–Mantel–Haenszel test.

An alternative strategy in multicenter analyses combines centers of a similar type. Then, if each resulting partial table has responses with both outcomes, the inferences use all data. For Table 6.11, perhaps centers 1 and 3 are similar to center 2, since the success rate is very low for that center. Combining these three centers and refitting the model to this table and the tables for the other two centers yields $\hat{\beta} = 1.56$ ($SE = 0.70$). Usually, this strategy produces results essentially the same as from deleting tables with no outcomes of a particular type.

6.5.3 Remedies When at Least One ML Estimate Is Infinite

What can you do if there is complete or quasi-complete separation and thus at least one ML estimate does not exist? As just mentioned, you can still usually do inference about that effect. For example, you can conduct a likelihood-ratio test. If $\hat{\beta} = \infty$, a profile likelihood confidence interval will have the form (L, ∞). With quasi-complete separation, some parameter estimates may be unaffected, and their inference will resemble the usual. With small samples and categorical predictors, you can use the specialized exact conditional methods to be presented in Section 7.3.

Alternatively, you can make some adjustment so all estimates are finite. For example, if a category of a qualitative predictor has no cases with $y = 1$, perhaps combine that category

with a similar one such that outcomes of both type then occur. Some approaches smooth the data, thus producing finite estimates. The Bayesian approach (Section 7.2) is the best known way of doing that. The amount of smoothing for the resulting estimates depend strongly on the variability in the Bayes prior distribution.

A related way maximizes a *penalized likelihood* function. This adds a term to the ordinary log-likelihood function such that maximizing the amended function smooths the estimates by shrinking them toward 0 (Firth 1993a). Section 7.4.5 introduces this approach, which corresponds to using the Bayesian posterior mode induced by the Jeffreys prior distribution. For the data in Figure 6.5, this method replaces the infinite estimate of β by $\hat{\beta} = 0.067$ ($SE = 0.042$). The corresponding 95% penalized profile likelihood confidence interval is (0.013, 0.334). Its highly asymmetric form about $\hat{\beta}$ reflects the highly nonsymmetric appearance of the log-likelihood function for such data.

6.6 SAMPLE SIZE AND POWER CONSIDERATIONS

In any statistical procedure, the sample size n influences the results. Strong effects are likely to be detected even when n is small. By contrast, detection of weak effects requires large n. A study design should reflect the sample size needed to provide good power for detecting the effect.

6.6.1 Sample Size and Power for Comparing Two Proportions

For test statistics having large-sample normal distributions, power calculations can use ordinary methods. To illustrate, consider a test comparing binomial parameters π_1 and π_2 for two medical treatments. An experiment plans independent samples of size $n_i = n/2$ receiving each treatment. The researchers expect $\pi_i \approx 0.60$ for each, and a difference of at least 0.10 is important. In testing H_0: $\pi_1 = \pi_2$, the variance of $\hat{\pi}_1 - \hat{\pi}_2$ is $\pi_1(1 - \pi_1)/(n/2) + \pi_2(1 - \pi_2)/(n/2) \approx 0.60 \times 0.40 \times (4/n) = 0.96/n$. In particular,

$$z = \frac{(\hat{\pi}_1 - \hat{\pi}_2) - (\pi_1 - \pi_2)}{\sqrt{0.96/n}}$$

has approximately a standard normal distribution for π_1 and π_2 near 0.60.

The power of an α-level test of H_0 is approximately

$$P\left[\frac{|\hat{\pi}_1 - \hat{\pi}_2|}{\sqrt{0.96/n}} \geq z_{\alpha/2}\right].$$

When $\pi_1 - \pi_2 = 0.10$, for $\alpha = 0.05$, this equals

$$\begin{aligned} &P\left[\frac{(\hat{\pi}_1 - \hat{\pi}_2) - 0.10}{\sqrt{0.96/n}} > 1.96 - 0.10\sqrt{n/0.96}\right] \\ &\quad + P\left[\frac{(\hat{\pi}_1 - \hat{\pi}_2) - 0.10}{\sqrt{0.96/n}} < -1.96 - 0.10\sqrt{n/0.96}\right] \\ &= P[z > 1.96 - 0.10\sqrt{n/0.96}] + P[z < -1.96 - 0.10\sqrt{n/0.96}] \\ &= 1 - \Phi[1.96 - 0.10\sqrt{n/0.96}] + \Phi[-1.96 - 0.10\sqrt{n/0.96}], \end{aligned}$$

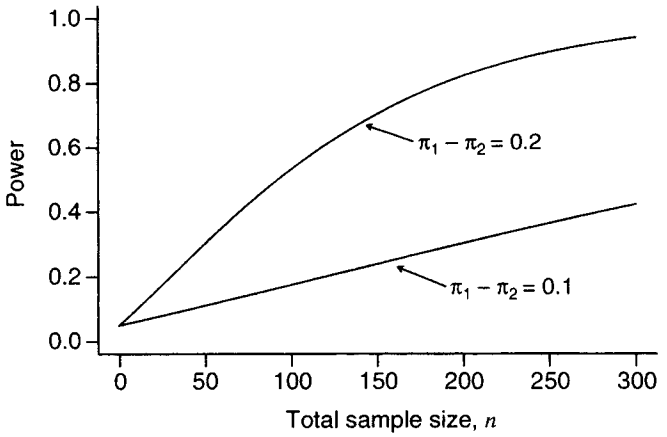

Figure 6.6 Approximate power for testing equality of proportions, with true values near middle of range and $\alpha = 0.05$.

where Φ is the standard normal cdf. The power is approximately 0.11 when $n = 50$ and 0.30 when $n = 200$. It is not easy to attain significance when effects are small and the sample size is not very large. Figure 6.6 shows how the power increases in n when $\pi_1 - \pi_2 = 0.10$. By contrast, it also shows how the power improves when $\pi_1 - \pi_2 = 0.20$.

For specified $P(\text{type I error}) = \alpha$ and $P(\text{type II error}) = \beta$ (and hence power $= 1 - \beta$), we can determine the sample size needed to attain those values. A study using $n_1 = n_2$ requires approximately

$$n_1 = n_2 = (z_{\alpha/2} + z_\beta)^2[\pi_1(1 - \pi_1) + \pi_2(1 - \pi_2)]/(\pi_1 - \pi_2)^2.$$

For a test with $\alpha = 0.05$ and $\beta = 0.10$ when π_1 and π_2 are truly about 0.60 and 0.70, $n_1 = n_2 = 473$. Similarly, with about 473 subjects in each group, a 95% confidence interval has only a 0.10 chance of containing 0 when actually, $\pi_1 = 0.60$ and $\pi_2 = 0.70$.

This sample-size formula is approximate and may underestimate slightly the actual values required. It is adequate for most practical work, though, in which only rough conjectures are available for π_1 and π_2. Farrington and Manning (1990) and Fleiss et al. (2003, Chap. 4) showed more precise formulas.

6.6.2 Sample Size Determination in Logistic Regression

Consider now the model $\text{logit}[\pi(x_i)] = \alpha + \gamma x_i$, $i = 1, \ldots, n$, in which x is quantitative. [We use γ so as not to confuse with $\beta = P(\text{type II error})$.] The sample size needed to achieve a certain power for testing H_0: $\gamma = 0$ depends on the variance of $\hat{\gamma}$. This depends on $\{\pi(x_i)\}$, and formulas for n use a guess for $\hat{\pi} = \pi(\bar{x})$ and the distribution of X. The effect size is the log odds ratio τ comparing $\pi(\bar{x})$ to $\pi(\bar{x} + s_x)$, the probability at a standard deviation above the mean of x. For a one-sided test when X is approximately normal, Hsieh (1989) derived

$$n = [z_\alpha + z_\beta \exp(-\tau^2/4)]^2(1 + 2\hat{\pi}\delta)/(\hat{\pi}\tau^2), \tag{6.10}$$

where

$$\delta = [1 + (1 + \tau^2)\exp(5\tau^2/4)]/[1 + \exp(-\tau^2/4)].$$

The value n decreases as $\hat{\pi} \to 0.50$ and as $|\tau|$ increases.

We illustrate for modeling the effect of x = cholesterol level on the probability of severe heart disease for a population for which that probability at an average level of cholesterol is about 0.08. Researchers want the test to be sensitive to a 50% increase in this probability, for a standard deviation increase in cholesterol. The odds of severe heart disease at the mean cholesterol level equal $0.08/0.92 = 0.087$, and the odds one standard deviation above the mean equal $0.12/0.88 = 0.136$. The odds ratio equals $0.136/0.087 = 1.57$, and $\tau = \log(1.57) = 0.450$. For $\alpha = 0.05$ and $\beta = 0.10$, $\delta = 1.306$ and $n = 612$.

6.6.3 Sample Size in Multiple Logistic Regression

A multiple logistic regression model requires larger n to detect effects. Let R denote the multiple correlation between the predictor X of interest and the others in the model. The formula (6.10) for n divides by $(1 - R^2)$. In that formula, $\hat{\pi}$ is evaluated at the mean of all the explanatory variables, and the odds ratio refers to the effect of X at the mean level of the other predictors.

Consider the example in Section 6.2.2 when blood pressure is also a predictor. If the correlation between cholesterol and blood pressure is 0.40, we need $n \approx 612/[1 - (0.40)^2] = 729$.

These formulas provide, at best, very approximate indications of sample size. Most applications have only a crude guess for $\hat{\pi}$ and R, and X may be far from normally distributed.

6.6.4 Power for Chi-Squared Tests in Contingency Tables

When hypotheses are false, squared normal and X^2 and G^2 statistics have large-sample noncentral chi-squared distributions (Section 5.3.8). Suppose that H_0 is equivalent to model M for a contingency table. Let π_i denote the true probability in cell i, and let $\pi_i(M)$ denote the value to which the ML estimate $\hat{\pi}_i$ for model M converges, where $\sum_i \pi_i = \sum_i \pi_i(M) = 1$. For a multinomial sample of size n, the noncentrality parameter for X^2 equals

$$\lambda = n \sum_i \frac{[\pi_i - \pi_i(M)]^2}{\pi_i(M)}. \tag{6.11}$$

This has the same form as X^2, with π_i in place of the sample proportion p_i and $\pi_i(M)$ in place of $\hat{\pi}_i$. The noncentrality parameter for G^2 equals

$$\lambda = 2n \sum_i \pi_i \log \frac{\pi_i}{\pi_i(M)}. \tag{6.12}$$

When H_0 is true, all $\pi_i = \pi_i(M)$. Then, for either statistic, $\lambda = 0$ and the ordinary (central) chi-squared distribution applies.

To determine the approximate power for a chi-squared test with df = ν, (1) choose a hypothetical set of true values $\{\pi_i\}$, (2) calculate $\{\pi_i(M)\}$ by fitting to $\{\pi_i\}$ the model M

Table 6.12 Power of Chi-Squared Test for $\alpha = 0.05$

	Noncentrality													
df	0.0	0.2	0.4	0.6	0.8	1.0	2.0	3.0	4.0	5.0	7.0	10.0	15.0	25.0
1	.050	.073	.097	.121	.146	.170	.293	.410	.516	.609	.754	.885	.972	.998
2	.050	.065	.081	.098	.115	.133	.226	.322	.415	.504	.655	.815	.944	.996
3	.050	.062	.075	.088	.102	.116	.192	.275	.358	.440	.590	.761	.917	.993
4	.050	.060	.071	.082	.093	.106	.172	.244	.320	.396	.540	.716	.891	.989
6	.050	.058	.066	.075	.084	.094	.146	.206	.270	.336	.468	.644	.843	.980
8	.050	.057	.064	.071	.079	.087	.131	.182	.238	.296	.417	.588	.799	.968
10	.050	.056	.062	.068	.075	.082	.121	.166	.215	.268	.379	.542	.760	.956
20	.050	.053	.056	.060	.063	.066	.096	.125	.158	.193	.273	.402	.611	.883
50	.050	.052	.054	.056	.059	.061	.076	.092	.110	.129	.173	.250	.398	.687

Source: Reprinted with permission from G. E. Haynam, Z. Govindarajulu, and F. C. Leone, in *Selected Tables in Mathematical Statistics*, eds. H. L. Harter and D. B. Owen. Chicago: Markham, 1970.

for H_0, (3) calculate the noncentrality parameter λ, and (4) calculate $P[X^2_{\nu,\lambda} > \chi^2_\nu(\alpha)]$. Table 6.12 shows an excerpt from a table of noncentral chi-squared probabilities for step 4 with $\alpha = 0.05$.

6.6.5 Power for Testing Conditional Independence

We use an example based on one in O'Brien (1986). A standard fetal heart rate monitoring test predicts whether a fetus will require nonroutine care following delivery. The standard test has categories (worrisome, reassuring). The response Y is whether the newborn required some nonroutine medical care during the first week after birth (1 = yes, 0 = no). A new fetal heart rate monitoring test is developed, having categories (very worrisome, somewhat worrisome, reassuring). A physician plans to study whether this new test can help make predictions about the outcome; that is, given the result of the standard test, is there an association between the response and the result of the new test? A relevant statistic tests the effect of the new monitoring test in the logistic model having the new test (N) and the standard test (S) as qualitative predictors.

To help select n, a statistician asks the physician to conjecture about the joint distribution of the explanatory variables, with questions such as "What proportion of the cases do you think will be scored 'reassuring' by both tests?" For each NS combination, the physician also guessed $P(Y = 1)$. Table 6.13 shows one scenario for marginal and conditional

Table 6.13 Scenario for Power Computation

Standard Test	New Test	Joint Probability	P(nonroutine care)
Worrisome	Very worrisome	0.04	0.40
	Somewhat worrisome	0.08	0.32
	Reassuring	0.04	0.27
Reassuring	Very worrisome	0.02	0.30
	Somewhat worrisome	0.18	0.22
	Reassuring	0.64	0.15

Source: Reprinted with permission from O'Brien (1986).

probabilities. These yield a joint distribution $\{\pi_{ijk}\}$ from their product, such as $0.04 \times 0.40 = 0.016$ for the proportion of cases judged worrisome by the standard test and very worrisome by the new test and requiring nonroutine medical care. These joint probabilities yield fitted probabilities $\boldsymbol{\pi}(M_0)$ and $\boldsymbol{\pi}(M_1)$ for the null and alternative logit models. (We can get these by entering $\{\pi_{ijk}\}$ in percentage form as counts in software for logistic regression, fitting the relevant model, and dividing the fitted counts by 100 to get the fitted joint probabilities.) The likelihood-ratio test comparing these models has noncentrality (6.12) with $\boldsymbol{\pi}(M_1)$ playing the role of $\boldsymbol{\pi}$ and $\boldsymbol{\pi}(M_0)$ playing the role of $\boldsymbol{\pi}(M)$.

For the scenario in Table 6.13, the noncentrality equals $0.00816n$, with $df = 2$. For $n = 400$, 600, and 1000, the approximate powers when $\alpha = 0.05$ are 0.35, 0.49, and 0.73. This scenario predicts 64% of the observations to occur at only one combination of the factors. The lack of dispersion for the factors weakens the power.

6.6.6 Effects of Sample Size on Model Selection and Inference

The effects of sample size suggest some cautions for model selection. For small n, the most parsimonious model accepted in a goodness-of-fit test may be quite simple. By contrast, larger samples usually require more complex models to pass goodness-of-fit tests. Then, some effects that are statistically significant may be weak and substantively unimportant. With large n it may be adequate to use a model that is simpler than models that pass goodness-of-fit tests. An analysis that focuses solely on goodness-of-fit tests is incomplete. It is also necessary to estimate model parameters and describe strengths of effects.

These remarks merely reflect limitations of significance testing. In many areas of application, null hypotheses are rarely true. With large enough n, they will be rejected. A more relevant concern is whether the difference between true parameter values and null hypothesis values is sufficient to be important. Many methodologists overemphasize testing and underutilize estimation methods such as confidence intervals. When the P-value is small, a confidence interval specifies the extent to which H_0 may be false, thus helping us determine whether rejecting it has practical importance. When the P-value is not small, the confidence interval indicates whether some plausible parameter values are far from H_0. A wide confidence interval containing the H_0 value indicates that the test had weak power at important alternatives.

NOTES

Section 6.1: Strategies in Model Selection

6.1 **AIC, BIC**: For cogent arguments supporting the use of AIC, see Burnham and Anderson (2010). A modified version is recommended if the number of parameters is large. Some statisticians believe that BIC can select an overly simple model. For this and other critiques, see articles by Gelman and Rubin, Firth and Kuha, Raftery, Weakliem, and Xie, in the February 1999 issue of *Sociological Methods and Research*.

Section 6.2: Logistic Regression Diagnostics

6.2 **Diagnostics**: Olive and Hawkins (2005) presented graphics that are useful for variable selection. As an alternative to the residual methods discussed, smoothing the residuals before plotting them (e.g., using methods to be presented in Section 7.4) can be helpful (Fowlkes 1987,

Lloyd 1999, Sec. 5.4). Cook and Weisberg (1999, Chap. 22) and Landwehr et al. (1984) showed other examples of useful diagnostic plots. For other logistic regression diagnostics, see Copas (1988), who also considered resistant fitting methods (e.g., to take misclassification into account), Hosmer and Lemeshow (2000, Chap. 5), Johnson (1985), and Pregibon (1981).

Section 6.3: Summarizing the Predictive Power of a Model

6.3 **R^2 measures**: Amemiya (1981), Efron (1978), Hu et al. (2005), Liao and McGee (2003), Maddala (1983), Schemper (2003), and Zheng and Agresti (2000) and references therein reviewed R^2 measures for binary regression. Hosmer and Lemeshow (2000, Sec. 5.2.3) discussed classification tables and their limitations. Pepe (2004) and references therein surveyed ROC methodology.

Section 6.4: Mantel–Haenszel and Related Methods for Multiple 2 × 2 Tables

6.4 **DIF**: One application of CMH methods is *differential item functioning*: comparing groups in terms of how different they are in responding to items on a questionnaire, after adjusting for overall abilities or scores. See Holland and Wainer (1993).

6.5 **Breslow–Day test**: An analog of $\hat{\theta}_{MH}$ and $\hat{\delta}_{MH}$ summarizes relative risks from several strata (Greenland and Robins 1985). Breslow and Day (1980, p. 142) proposed an alternative large-sample test of homogeneity of odds ratios. In each partial table let $\{\hat{\mu}_{ijk}\}$ have the same marginals as the data observed, yet have odds ratio equal to $\hat{\theta}_{MH}$. Their test statistic has the Pearson form comparing $\{n_{ijk}\}$ to $\{\hat{\mu}_{ijk}\}$. Tarone (1985) showed that, because of the inefficiency of $\hat{\theta}_{MH}$, the Breslow–Day statistic must be adjusted for it to have exactly a limiting chi-squared null distribution with df $= K - 1$. This adjustment is usually minor. Other work on comparing odds ratios and estimating a common value includes Breslow and Day (1980, Sec. 4.4), Donner and Hauck (1986), Gart (1970), Jones et al. (1989), and Liang and Self (1985). For modeling the odds ratio, see Breslow (1976), Breslow and Day (1980, Sec. 7.5), and Prentice (1976a). Breslow emphasized retrospective studies, in which the conditional approach is natural since the outcome totals are fixed.

Section 6.5: Detecting and Dealing with Infinite Estimates

6.6 **Infinite ML**: For discussion of this topic, including other link functions and GLMs, see Albert and Anderson (1984), Haberman (1974a), Santner and Duffy (1986), Silvapulle (1981), and Wedderburn (1976).

6.7 **High imbalance**: King and Zeng (2001) and Owen (2007) discussed applications in which one outcome category is much more common than the other. Examples include rare diseases, fraudulent use of a credit card, and non-spam email messages in spam folders. King and Zeng proposed a sampling design of sampling all possible cases of the rare outcome and a much smaller fraction of the other outcome. Owen showed that under a sampling scheme for which $n \to \infty$ while the number of outcomes in one category remains finite, a limit exists for the estimated parameter vector that depends on the distribution of the $\boldsymbol{x}$ values.

Section 6.6: Sample Size and Power Considerations

6.8 **Noncentral chi-squared**: Gail and Gart (1973) and Suissa and Shuster (1985) studied sample size for obtaining fixed power in Fisher's test. Farrington and Manning (1990) considered sample size for nonnull effects for the difference of proportions and relative risk using score-type tests. For sample size determination in logistic regression, see Hsieh et al. (1998), Lyles et al. (2006), Schoenfeld and Borenstein (2005), Vaeth and Skovlund (2004), and Whittemore

(1981). Lachin (1977) considered $I \times J$ tables. Drost et al. (1989), Haberman (1974a, pp. 109–112), Meng and Chapman (1966), Mitra (1958), and Patnaik (1949) derived theory for asymptotic nonnull behavior of chi-squared statistics; see also Section 16.3.5. O'Brien's (1986) simulation results suggested that the noncentral chi-squared approximation for G^2 holds well for a wide range of powers. Read and Cressie (1988, pp. 147–148) listed other articles that studied the nonnull behavior of X^2 and G^2.

EXERCISES

Applications

6.1 For the horseshoe crab mating data, the maximized log-likelihood value is -112.88 for the model with only an intercept, -97.87 for the model with weight as a predictor, -97.23 for the model with width as a predictor, and -96.45 for the model using both as predictors. Conduct (**a**) a test of H_0: $\beta_1 = \beta_2 = 0$ for the joint effects, and (**b**) separate tests for the partial effects. Why does neither test in part (b) show evidence of an effect when the test in part (a) shows strong evidence?

6.2 For the horseshoe crab mating data, Table 6.14 shows ML estimates for two models using weight and color (with dark color as the baseline) as predictors of satellite presence. Compare the models using a likelihood-ratio test and using AIC. Select a model, and interpret its estimates.

Table 6.14 Effects for Two Models with Predictors of Crab Satellites, for Exercise 6.2

	Model 1		Model 2	
Term	Estimate	*SE*	Estimate	*SE*
Intercept	−4.53	1.00	−1.19	2.30
Weight	1.69	0.39	0.19	1.03
Color 1	1.27	0.85	−0.43	5.40
Color 2	1.41	0.54	−1.27	2.58
Color 3	1.08	0.59	−6.73	3.44
Weight×Color 1			0.85	2.16
Weight×Color 2			1.21	1.14
Weight×Color 3			3.56	1.56
Log-likelihood	−94.27		−90.83	
AIC	198.54		197.66	

6.3 The book's website (`www.stat.ufl.edu/~aa/cda/cda.html`) has a $2 \times 3 \times 2 \times 2$ table relating responses on frequency of attending religious services, political views, opinion on making birth control available to teenagers, and opinion about whether premarital sex before marriage is wrong. Treating opinion about premarital sex as the response variable, use backward elimination to select a model. Interpret.

6.4 For Table 10.1, treating marijuana use as the response variable, build a model with alcohol use, cigarette use, gender, and race as potential explanatory variables.

Summarize your strategy for selecting a model, and interpret your final choice of model.

6.5 For Table 6.4, fit the stage 3 model denoted there by $(E * P + G)$. Use parameter estimates to interpret the G effect and the dependence of the E effect on P.

6.6 According to the *Independent* newspaper (London, Mar. 8, 1994), the Metropolitan Police in London reported 30,475 people as missing in the year ending March 1993. For those of age 13 or less, 33 of 3271 missing males and 38 of 2486 missing females were still missing a year later. For ages 14 to 18, the values were 63 of 7256 males and 108 of 8877 females; for ages 19 and above, the values were 157 of 5065 males and 159 of 3520 females. Analyze by building a model, and interpret. (Thanks to Pat Altham for showing me these data.)

6.7 Fowlkes et al. (1988) reported results of a survey of employees of a large national corporation to determine how satisfaction depends on race, gender, age, and regional location. The data are at the book's website. Build a logistic model for these data and carefully interpret the parameter estimates.

6.8 Table 6.15 shows the results of a study about Y = whether a patient having surgery with general anesthesia experienced a sore throat on waking (0 = no, 1 = yes) as a function of the D = duration of the surgery (in minutes) and the T = type of device used to secure the airway (0 = laryngeal mask airway, 1 = tracheal tube). Use a model-building strategy to select a logistic model for these predictors. For your model, interpret parameter estimates, and conduct inference about the effects.

Table 6.15 Data for Exercise 6.8 on Surgery and Sore Throats

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

Source: Data from "Binary Data" by D. Collett, in *Encyclopedia of Biostatistics*, 2nd ed. Hoboken, NJ: Wiley, 2005, pp. 439–446.

6.9 Refer to the previous exercise. Use a measure of predictive power to compare the fits of various models to these data.

6.10 Refer to the previous two exercises. For your preferred model:

a. Summarize predictive power using classification tables with $\pi_0 = 0.50$ and $\pi_0 = \bar{y}$. In each case, report and interpret the sensitivity and specificity.

b. Summarize predictive power using an ROC curve. Report and interpret the concordance index.

6.11 Discern the reasons that Simpson's paradox occurs for the graduate admissions data of Table 6.6.

6.12 Refer to Exercise 2.15 on graduate school admissions and gender. Fit the model of no G effect, given the department. Use X^2 to test the fit. Obtain standardized residuals, explain how they relate to X^2, and interpret the lack of fit.

6.13 Conduct a residual analysis for the independence model with Table 5.5 on treating leprosy. What type of lack of fit is indicated?

6.14 For the horseshoe crab data, use methods such as Section 6.3 shows to evaluate predictive power for logistic models that include weight and color as explanatory variables.

6.15 Table 6.16 refers to the effectiveness of immediately injected or $1\frac{1}{2}$-hour-delayed penicillin in protecting rabbits against lethal injection with β-hemolytic streptococci.

a. Let $X =$ delay, $Y =$ whether cured, and $Z =$ penicillin level. Fit the logistic model (6.4). Argue that the pattern of 0 cell counts suggests that (with no intercept) $\hat{\beta}_1^Z = -\infty$ and $\hat{\beta}_5^Z = \infty$. What does your software report?

b. Using the logistic model, conduct the likelihood-ratio test of XY conditional independence. Interpret.

Table 6.16 Data for Exercise 6.15 on Penicillin Treatment for Streptococcus

Penicillin Level	Delay	Response	
		Cured	Died
$\frac{1}{8}$	None	0	6
	$1\frac{1}{2}$ h	0	5
$\frac{1}{4}$	None	3	3
	$1\frac{1}{2}$ h	0	6
$\frac{1}{2}$	None	6	0
	$1\frac{1}{2}$ h	2	4
1	None	5	1
	$1\frac{1}{2}$ h	6	0
4	None	2	0
	$1\frac{1}{2}$ h	5	0

Source: Reprinted with permission from Mantel (1963).

c. Test XY conditional independence using the Cochran–Mantel–Haenszel test. Interpret.

d. Estimate the XY conditional odds ratio using (i) ML with the logistic model, and (ii) the Mantel–Haenszel estimate. Interpret.

6.16 Refer to Table 2.6. Use the CMH statistic to test independence of death penalty verdict and victim's race, controlling for defendant's race. Conduct another test of this hypothesis, and compare results.

6.17 Treatments A and B were compared on a binary response for 40 pairs of subjects matched on relevant covariates. For each pair, treatments were assigned to the subjects randomly. Twenty pairs of subjects made the same response for each treatment. Six pairs had a success for the subject receiving A and a failure for the subject receiving B, whereas the other 14 pairs had a success for B and a failure for A. Use the Cochran–Mantel–Haenszel procedure to test independence of response and treatment. (In Section 11.1 we present an equivalent test, McNemar's test.)

6.18 For the data summarized in Figure 1 of the 2011 *Lancet* article by Rothwell et al. (**377**: 31–41) from eight studies on the effect of daily aspirin on cancer deaths, conduct a meta-analysis that combines a significance test with a confidence interval to summarize the size of effect. Interpret.

6.19 For the data summarized in Figure 1 of the 2010 *American Statistician* article by Kulinskaya et al. (**64**: 350–356), conduct a meta-analysis that combines a significance test with a confidence interval to summarize the size of effect. Interpret.

6.20 A data set at the text website from a 2005 article by D. Potter (*Statist. Med.* **24**: 693–708) describes results from a study in which subjects received a drug and the outcome measures whether the subject became incontinent ($y = 1$, yes; $y = 0$, no). The three explanatory variables are lower urinary tract variables that represent drug-induced physiological changes.

a. Find the prediction equations when each predictor is used separately in logistic regressions.

b. Try to fit a main-effects logistic model containing all three predictors. What does your software report for the effects and their standard errors? (The ML estimates are actually $-\infty$ for x_1 and x_2 and ∞ for x_3.) Can you see a pattern in the data that is responsible for this behavior?

6.21 Refer to the example of complete separation in Section 6.5.1. For the 8 observations, randomly generate values for a second predictor from the $N(0, 1)$ distribution. Taking both explanatory variables in your model, is there still complete separation? Is there quasi-complete separation? What does your software report for the model parameter estimates and SE values?

6.22 Refer to the multicenter clinical trial of Table 6.11.

a. Fit the main effects model considered in the text with your favorite software (omitting the intercept), and summarize results.

b. For Center 1, add ε successes for the active treatment, and report the impact (if any) on β_1^Z and $\hat{\beta}$. Do this for $\varepsilon = 10^{-6}$, $\varepsilon = 10^{-3}$, $\varepsilon = 0.50$. Do such centers give any information about the treatment log odds ratio effect, as described by $\hat{\beta}$ and its SE?

6.23 Apply the logistic regression model to the 2×2 table consisting of the data for Center 5 in Table 6.9, where $x = 1$ for drug and $x = 0$ for control.

a. Report the ML estimate $\hat{\beta}$.

b. What does your software report when you try to fit this model? Explain why.

c. Can you construct a 95% confidence interval for β? Show how.

6.24 For the example in Section 6.6.1, suppose $\pi_1 = 0.70$ and $\pi_2 = 0.60$. What sample size is needed for the test to have approximate power 0.80, when $\alpha = 0.05$, for (**a**) H_a: $\pi_1 \neq \pi_2$ and (**b**) H_a: $\pi_1 > \pi_2$?

6.25 For the example in Section 6.6.1 with equal treatment sample sizes, suppose $\pi_1 = 0.63$ and $\pi_2 = 0.57$. Explain why the joint probabilities in the 2×2 table are 0.315 and 0.185 for treatment A and 0.285 and 0.215 for treatment B. For the model of independence, explain why the fitted joint probabilities are 0.30 for success and 0.20 for failure, in each row. Show that X^2 has noncentrality parameter $0.00375n$ and df = 1. For $n = 200$ and $\alpha = 0.05$, find the power.

6.26 An experiment is designed to compare two treatments on a three-category response. The researcher expects the conditional distributions to be approximately (0.2, 0.2, 0.6) and (0.3, 0.3, 0.4).

a. With 100 observations for each treatment and $\alpha = 0.05$, find the approximate power to compare the distributions using (i) X^2 and (ii) G^2. Compare results.

b. What sample size is needed for each treatment for the tests in (a) to have approximate power 0.90?

6.27 The horseshoe crab width values in Table 4.3 have $\bar{x} = 26.3$ and $s_x = 2.1$. If the true relationship were similar to the fitted equation in Section 5.1.3, about how large a sample yields P(type II error) = 0.10, with $\alpha = 0.05$, for testing H_0: $\beta = 0$ against H_a: $\beta > 0$?

6.28 This book's website (`www.stat.ufl.edu/~aa/cda/cda.html`) contains a five-way table relating occupational aspirations (high, low) to gender, residence, IQ, and socioeconomic status. Analyze these data.

6.29 In recent years there has been controversy about the effects of rosiglitazone (an antidiabetic drug) on myocardial infarction (MI) and cardiovascular mortality. Review the 2010 meta-analysis by S. Nissen and K. Wolski in *Archives of Internal Medicine* (**14**: 1191–1201). Conduct your own analysis of the effects of rosiglitazone on MI.

Theory and Methods

6.30 For a sequence of s nested models $M_1, \ldots, M_s$, model M_s is the most complex. Let v denote the difference in residual df between M_1 and M_s.

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

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

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

6.31 Prove that the Pearson residuals for the linear logit model applied to a $I \times 2$ contingency table satisfy $X^2 = \sum_{i=1}^{I} e_i^2$. [*Hint*: Start with the X^2 sum over the $2I$ cells and combine the two terms from the same row.] Note that this holds for a binomial GLM with a linear trend for *any* link function.

6.32 For ungrouped binary data, explain why when $\hat{\pi}_i$ is near 1, residuals are necessarily either small and positive or large and negative. What happens when $\hat{\pi}_i$ is near 0?

6.33 For a $2 \times 2 \times K$ table from a multicenter clinical trial, one center has entries $(0, n)$ in row 1 and $(0, 2n)$ in row 2 (i.e., no successes for either treatment).

a. Explain why there is *no* information in this table about whether there is an association, regardless of the value of n. [*Hint*: Show that $\hat{\pi}_1 - \hat{\pi}_2 = 0$ has estimated null $SE = 0$, and the P-value is 1.0 for Fisher's exact test or for an unconditional exact test.]

b. Explain why there *is* information in the table about the size of association, in terms of the difference of proportions, and the precision of information increases as n increases. Illustrate by finding the 95% score confidence intervals for π_1, π_2, and $\pi_1 - \pi_2$, when $n = 10$ and when $n = 100$. (See `www.stat.ufl.edu/~aa/cda/R` for R functions. Note that Wald intervals are useless for such data.)

6.34 Refer to logit model (6.4) for a $2 \times 2 \times K$ contingency table $\{n_{ijk}\}$. Using a basic result for testing in exponential families, explain why uniformly most powerful unbiased tests of conditional XY independence are based on $\sum_k n_{11k}$ (Birch 1964b; Lehmann and Romano 2005, Sec. 4.8).

6.35 Suppose that $\{\pi_{ijk}\}$ in a $2 \times 2 \times 2$ table are, by row, (0.15, 0.10 / 0.10, 0.15) when $Z = 1$ and (0.10, 0.15 / 0.15, 0.10) when $Z = 2$. For testing conditional XY independence with logistic models having Y as a response, explain why the likelihood-ratio test comparing models $X + Z$ and Z is not consistent but the likelihood-ratio test of fit of the XY conditional independence model is.

6.36 For 2×2 tables with all marginal totals positive, explain what patterns of 0 cell counts correspond to (**a**) complete separation and (**b**) quasi-complete separation.

6.37 For k explanatory variables, suppose logistic regression has finite parameter estimates when used with each predictor alone. Explain why infinite estimates could occur when the predictors are all used in a main-effects model. Sketch a graph with $k = 2$ to illustrate this.

6.38 In Table 6.11, suppose the outcome of 0 successes for the active drug in Centers 1 and 3 was instead a positive count, but there were still no successes for placebo in those centers. Explain why all estimates would be finite for the main-effects model fitted in Section 6.5.2, but infinite estimates would occur for the more general model permitting center-by-treatment interaction.

6.39 Explain why complete or quasi-complete separation would not cause ML estimates to be infinite if you were using the identity link function but might cause other problems with the iterative fitting process.

6.40 For a GLM, let $\hat{\boldsymbol{\mu}}^{(-)} = (\hat{\mu}^{(-1)}, \ldots, \hat{\mu}^{(-n)})$, where $\hat{\mu}^{(-i)}$ denotes the estimate of $E(Y_i)$ for observation i after fitting the model without that observation. The *leave-one-out cross-validation* adjustment to the predictive measure $R(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ is $\text{corr}(\boldsymbol{y}, \hat{\boldsymbol{\mu}}^{(-)})$. For binary data, consider the model, $\text{logit}(\pi_i) = \alpha$ for all i. Show that $\hat{\pi}_i = \bar{y}$, $\hat{\pi}^{(-i)} = [n/(n-1)][\bar{y} - (1/n)y_i]$, and hence $\text{corr}(\boldsymbol{y}, \hat{\boldsymbol{\pi}}^{(-)}) = -1$. This suggests that leave-one-out cross-validation can be misleading for estimating the correlation with model $\text{logit}(\pi_i) = \alpha + \beta x$ when the true effect is very weak (Zheng and Agresti 2000).

6.41 Using graphs or tables, explain what is meant by *no interaction* in modeling response variable Y and explanatory variables X and Z when:

a. All variables are continuous (multiple regression).

b. Y and X are continuous, Z is categorical (analysis of covariance).

c. Y is continuous, X and Z are categorical (two-way ANOVA).

d. Y is binary, X and Z are categorical (logistic regression).