3

Logistic Regression, Logit Models, and Logistic Discrimination

In logistic regression, there is a (binary) response of interest, and predictor variables are used to model the probability of that response. More generally, in a table of counts, primary interest is frequently centered on one factor that constitutes a response (dependent) variable. The other factors in the table are only of interest for their ability to help explain the response variable. Special kinds of models have been developed to handle these situations. In particular, rather than modeling log expected cell counts or log probabilities (as in log-linear models), when there is a response variable, various log *odds* related to the response variable are modeled.

The special case in which the response variable has only two categories is of particular interest and lends itself to an especially nice treatment. This is because, with only two categories, there is essentially only one way to define the odds. If p_1 is the probability in the first category and p_2 is the probability in the second category, then the odds of getting category one are p_1/p_2. The odds of getting category two are p_2/p_1. The important point is that *either* of these numbers, together with the fact that $p_1 + p_2 = 1$, completely determine both p_1 and p_2. So with two categories, the two choices for the odds lead to the same results. (In the last section of this chapter, we look at ratios p_1/p_2 but *without* $p_1 + p_2 = 1$; this causes complications.)

With a two-category response variable, we will examine models for $\log(p_1/p_2)$. When these models are regression type models, they are called *logistic regression models.* When these models are ANOVA type models, they are often referred to as *logit models.* The two terms "logit" and "logistic regression," as applied to models, are essentially two names for the

same idea. Technically, the terms logit and logistic are names for transformations. The logit transformation takes a number p between 0 and 1 and transforms it to $\log[p/(1-p)]$. The logistic transformation takes a number x on the real line and transforms it to $e^x/(1+e^x)$. Note that the logit transformation and the logistic transformation are inverses of each other. In other words, the logistic transformation applied to $\log[p/(1-p)]$ gives p and the logit transformation applied to $e^x/(1+e^x)$ gives x. Doing an analysis of data requires both of these transformations. It is largely a matter of personal preference as to which name is associated with the model.

The situation when there are more than two categories in the response variable is considerably more complicated because it is not clear which sets of odds to model. Several choices have been suggested; some of these are discussed in Section 6. As will be discussed in this chapter and shown in Chapter 11, the log odds models turn out to be equivalent to a log-linear model. *It is important to remember that log odds models are for use when relationships between the nonresponse factors (explanatory variables) are not of interest.* It is implicit in the definition of a logit model that no structure between the explanatory variables is taken into account. It *is* possible to use models for log odds that incorporate explanatory factor structure, but such models are not what are generally known as logistic regression or logit models.

Goodman (1973) proposes a multistep modeling procedure for response factors. His procedure involves collapsing on the response factor and fitting a log-linear model to the marginal table of the explanatory factors. This is followed by fitting a logit model for the response factor. Taken together, these give the probability of any cell in the table as the product of the marginal probability of its explanatory categories and the conditional probability of the cell given its explanatory categories. These recursive response models may or may not be log-linear models. Asmussen and Edwards (1983) give conditions for the equivalence of log-linear models and these multistep response models. Fienberg (1980, Chapter 7) gives a good brief discussion of response models and their limitations. More recently, graphical response models have been discussed by Asmussen and Edwards (1983), Edwards and Kreiner (1983), Kiiveri, Speed, and Carlin (1984), Kiiveri and Speed (1982), and Wermuth and Lauritzen (1983). Graphical models are discussed in the next chapter. Holland (1986) discusses statistics and causal inference, as do Glymour et al. (1987). The latter authors seem to have a substantially different perspective than that presented here. Goodman's basic procedure can be applied with more than one response factor and with responses involving more than two categories. In this chapter, attention is concentrated on models for one response factor that condition on all the explanatory factors. If more than one response factor is present, one simple approach just restricts the fitted models to log-linear models that condition on the explanatory factors.

Section 1 examines regression models for two category responses. Sec-

tions 2, 3, and 4 discuss measuring the fit of models, logistic regression diagnostics, and variable selection, respectively. Analysis of variance type models are examined in Section 5 for responses with two categories and in Section 6 for responses with more than two categories. Section 7 examines the analysis of retrospective studies via logistic discrimination; the distinction between retrospective and prospective studies is discussed in the next subsection.

Retrospective Versus Prospective Studies

It is important to distinguish between prospective and retrospective studies. It is important because aspects of the material in this chapter do not apply to retrospective studies. For our purposes, the distinction is based on the nature of the sampling scheme.

If the sampling is independent Poisson or if the categories of the response factor are in the same multinomial for every combination of the explanatory factors, the study is *prospective*. This is probably the most common way of thinking about data collection. For example, a prospective study of heart attack victims might take 250 people and examine each to determine whether they have had a heart attack and their levels of various explanatory factors. The explanatory factors may be such things as age, blood cholesterol, and blood pressure. In the prospective study, each combination of explanatory factors can be used to determine a population and an individual randomly falls into a response category. So, typically, there are many populations, and often each individual in the study is sampled from a different population. Prospective studies are (or can be thought of) as product-multinomials in which *the multinomial categories are the categories of the response factor.*

Obviously, in a random sample of 250 people from the general population, very few would have had heart attacks. An alternative sampling method is often used in the study of such rare events as heart attacks. One might sample 100 people who are known to have had heart attacks and 150 people who have not had heart attacks. Again, each individual is characterized by their levels of the explanatory factors. There are only two populations here: the heart attack victims and the subjects without heart attacks. The categories of the multinomials for the two populations are the different categories of explanatory factors. The analysis of such data involves describing the characteristics of the two groups in terms of the explanatory factors. The key fact in this second example is that *the response factor categories define different multinomial populations and the multinomial categories are the different combinations of the explanatory factors.* Generally, if, for all combinations of the explanatory factors, the various categories of the response factor occur in different multinomials, then the study is *retrospective.* In medical research, retrospective data collection corresponds to a *case-control* study. Clearly, for rare events, this

procedure has advantages over selecting a simple random sample. In a multinomial sample, so few of the rare events will occur as to give little power for determining their likelihood.

In the hypothetical retrospective study of heart attacks, let the index i denote a set of explanatory characteristics; let p_{1i} be the probability of that set for the heart attack population and p_{2i} be the probability for the population who have not had heart attacks. A parameter of interest is p_{1i}/p_{2i}, the ratio of the probabilities. (Note that, in general, $p_{1i} + p_{2i} \neq 1$, so these are *not* odds.) The parameter addresses the question of whether the explanatory characteristics i are more or less likely among heart attack victims than among subjects who have not had heart attacks. Unfortunately, this does not address the issue of cause and effect. It simply describes characteristics of the two populations. As will be seen in Section 7, inferences about $\log(\hat{p}_{1i}/\hat{p}_{2i})$ will be complicated by the fact that the probabilities apply to different multinomials. The asymptotic covariance of $\log(\hat{p}_{1i}/\hat{p}_{2i})$ does not simplify like it would if the probabilities were from the same multinomial. Moreover, one does not typically have the simplification that p_{1i}/p_{2i} is equal to m_{1i}/m_{2i}; thus, inferences about $\log(p_{1i}/p_{2i})$ cannot be made directly by examining $\log(\hat{m}_{1i}/\hat{m}_{2i})$.

Prospective studies do not directly address the issue of cause and effect either, but they come closer than retrospective studies. As discussed earlier, both multinomial sampling and some forms of product-multinomial sampling generate prospective studies. An example of a product-multinomial prospective study is to independently sample people for each set of explanatory characteristics and see how many have had heart attacks. In medical research, the data given by this sampling scheme are called *cohort* data, and *cross-sectional* data are used to indicate the results of simple multinomial sampling.

For cohort data, take p_{1i} to be the probability of a heart attack in the population defined by the ith set of explanatory variables. Obviously, $p_{2i} = 1 - p_{1i}$ is the probability of no heart attack in that population. The ratio p_{1i}/p_{2i} is the odds of having a heart attack for that population. (Recall that describing p_{1i}/p_{2i} as an odds is not appropriate in a retrospective study.) In a cohort study, one can argue that the ith population is a cause and that the ratio p_{1i}/p_{2i} is an effect. Unfortunately, populations usually involve more than the explanatory factors used to define them. Aspects of the ith population other than the values of the explanatory factors may be the true cause for p_{1i}/p_{2i}. We hope that random sampling within the population of people minimizes these effects.

For cross-sectional data, one can use the mental device of conditioning on the number of people who fall in each explanatory category to get an independent multinomial sample for each set of explanatory characteristics. Thus, we can treat cross-sectional prospective data as if they were cohort prospective data. In fact, any prospective study can be thought of as product-multinomial sampling with an independent sample for each set

of explanatory characteristics. This results from the fact that a prospective study is defined to be one in which all of the categories of the response factor are contained in the same multinomial.

Except for the last section, we restrict attention in this chapter to prospective studies. A convenience of dealing with prospective studies is that log odds defined for the response factor are the same using either probabilities or expected cell counts, i.e., $\log(p_{1i}/p_{2i}) = \log(m_{1i}/m_{2i})$.

3.1 Multiple Logistic Regression

This section is devoted to regression models for the log odds of a two-category response variable. The difference between this section and Section 2.6 is that here we consider the use of multiple predictor variables. The discussion will be centered around an example.

EXAMPLE 3.1.1. *Chapman Data.*
Dixon and Massey (1983) present data on 200 men taken from the Los Angeles Heart Study conducted under the supervision of John M. Chapman, UCLA. The data consist of seven variables:

Abbreviation	Variable	Units
Ag	Age:	in years
S	Systolic Blood Pressure:	millimeters of mercury
D	Diastolic Blood Pressure:	millimeters of mercury
Ch	Cholesterol:	milligrams per DL
H	Height:	inches
W	Weight:	pounds
CNT	Coronary incident:	1 if an incident had occurred in the previous ten years; 0 otherwise

Of the 200 cases, 26 had coronary incidents. The data are available electronically from STATLIB as well as through my web homepage:

`http://stat.unm.edu/~fletcher`

Additional information is given in the Preface.

As discussed in Section 2.6, such data can be viewed as a 200×2 contingency table in which the columns indicate presence or absence of a coronary incident and the rows indicate the 200 combinations of the explanatory variables Ag, S, D, Ch, H, and W associated with the men in the study. Each row is considered an independent binomial involving one trial. (If more than one person has the same combination of explanatory variables, it is irrelevant whether they are treated as binomials with one trial or grouped

together yielding a table with less than 200 rows.) The table has 200 counts and 400 cells, so the data are very sparse. As discussed in Section 2.6, when testing a logistic regression model against the saturated log-linear model (i.e., testing the logistic model for lack of fit), the asymptotic χ^2 approximation is notoriously bad. The test statistics are reasonable things to look at, but formal χ^2 tests are generally inappropriate because of the sparse data. Somewhat surprisingly, asymptotic χ^2 approximations do work for testing one logistic regression model (a full model) against another logistic regression (a reduced model).

Let p_i be the probability of a coronary incident for the ith man. We begin with the logistic regression model

$$\log[p_i/(1-p_i)] = \beta_0 + \beta_1 Ag_i + \beta_2 S_i + \beta_3 D_i + \beta_4 Ch_i + \beta_5 H_i + \beta_6 W_i, \quad (1)$$

$i = 1, \ldots, 200$. As discussed in Section 2.6 and Chapter 11, this is equivalent to a log-linear model for a two-way table in which the predictor variables are used to model the interaction, cf. Exercise 3.8.15. The model can be fitted using methods for log-linear models or the methods can be specialized for fitting logistic regression models, cf. Subsection 3.4.2 for SAS, BMDP, and GLIM commands. The actual methods for fitting logistic models will be examined in later chapters. The maximum likelihood fit of this model is given below.

Variable	Estimate	Std. Error	z
Intercept	−4.5173	7.451	−0.61
Ag	0.04590	0.02344	1.96
S	0.00686	0.02013	0.34
D	−0.00694	0.03821	−0.18
Ch	0.00631	0.00362	1.74
H	−0.07400	0.1058	−0.70
W	0.02014	0.00984	2.05

$$G^2 = 134.9, \qquad df = 193$$

The formula for G^2 is as in Section 2.6. The df is the number of cases, 200, minus the number of fitted parameters, 7. Based on the z values, none of the variables really stand out. There are suggestions of age, cholesterol, and weight effects. The G^2 would look good except that, as discussed earlier, there is no basis for comparing it to a standard.

Prediction follows much the same form as in Section 2.6,

$$\log[\hat{p}_i/(1-\hat{p}_i)] = \hat{\beta}_0 + \hat{\beta}_1 Ag_i + \hat{\beta}_2 S_i + \hat{\beta}_3 D_i + \hat{\beta}_4 Ch_i + \hat{\beta}_5 H_i + \hat{\beta}_6 W_i\,.$$

For a 60-year-old man with blood pressure of 140 over 90, a cholesterol reading of 200, who is 69 inches tall and weighs 200 pounds, the estimated log odds of a coronary incident are

$$\begin{aligned}\log[\hat{p}/(1-\hat{p})] = {} & -4.5173 + .04590(60) + .00686(140) - .00694(90) \\ & + .00631(200) - 0.07400(69) + 0.02014(200) = -1.2435.\end{aligned}$$

FIGURE 3.1. Coronary incident probabilities as a function of age. Solid curve–$Ch = 200$; dashed curve–$Ch = 300$.

The probability of a coronary incident is estimated as

$$\hat{p} = \frac{e^{-1.2435}}{1 + e^{-1.2435}} = .224\,.$$

Figure 3.1 gives plots of the estimated probability of a coronary incident as a function of age for people with $S = 140$, $D = 90$, $H = 69$, $W = 200$, and either $Ch = 200$ (solid line) or $Ch = 300$ (dashed line).

3.1.1 Informal Model Selection

We now consider fitting some reduced models. Simple linear logistic regressions were fitted for each of the variables with high z values, i.e., Ag, Ch, and W. Regressions with variables that seem naturally paired were also fitted, i.e., S,D and H,W. Listed below are the models, df, G^2, $A - q$, and A^*. The first two of these are the degrees of freedom and the likelihood ratio test statistic for testing against the saturated model. No *P values* are given because the asymptotic χ^2 approximation does not hold. Also given are two analogues of Mallow's C_p statistic, $A - q$ and A^*. $A - q$ was discussed in detail in Section 3.6. A^* is a modification of $A - q$ for logistic regression. $A - q \equiv G^2 - 2(df)$ and is the Akaike information criterion less the number of cells (200×2) in the table. A^* is a version of the Akaike information criterion defined for comparing submodels of model (1) to the full model. It is defined by

$$A^* = (G^2 - 134.9) - (7 - 2p)\,.$$

Here, 134.9 is G^2 for the full model (1), 7 comes from the degrees of freedom for the full model (6 explanatory variables plus an intercept), and p comes from the degrees of freedom for the submodel ($p = 1 +$ number of explanatory variables). The information in $A - q$ and A^* is identical: $A^* = 258.1 + (A - q)$. (The value 258.1 = number of cells $- G^2$[full model] $- p$[full model] $= 400 - 134.9 - 7$.) A^* is listed because it is a little easier to look at and takes values similar to C_p.

Model Variables	*df*	G^2	$A - q$	A^*
Ag,S,D,Ch,H,W	193	134.9	−251.1	7
Ag	198	142.7	−253.3	4.8
W	198	150.1	−245.9	12.2
H,W	197	146.8	−247.2	10.9
Ch	198	146.9	−249.1	9.0
S,D	197	147.9	−246.1	12.0
Intercept	199	154.6	−243.4	14.7

Of the models listed,

$$\log[p_i/(1 - p_i)] = \gamma_0 + \gamma_1 Ag_i \tag{2}$$

is the only model that is better than the full model based on the information criterion; i.e., A^* is 4.8 for this model, less than the 7 for model (1).

Asymptotically valid tests of submodels against model (1) are available. These are performed in the usual way; i.e., the differences in degrees of freedom and G^2's give the appropriate values for the tests. For example, the test of model (2) versus model (1) has $G^2 = 142.7 - 134.9 = 7.8$ with $df = 198 - 193 = 5$. Other tests are given below.

Tests against Model (1)

Model	*df*	G^2
Ag	5	7.8
W	5	15.2**
H,W	4	11.9*
Ch	5	12.0*
S,D	4	13.0*
Intercept	6	19.7**

All of the test statistics are significant at the .05 level, except for that associated with model (2). This indicates that none of the models other than (2) is an adequate substitute for the full model (1). In the table above, one asterisk indicates significance at the .05 level and two asterisks denotes significance at the .01 level.

Our next step is to investigate models that include Ag and some other variables. If we can find one or two variables that account for most of the

value $G^2 = 7.8$, we may have an improvement over model (2). If it takes three or more variables to explain the 7.8, model (2) will continue to be the best-looking model. [Note that $\chi^2(.95, 3) = 7.81$, so a model with three more variables than model (2) and the same fit as model (1) would still not demonstrate a significant lack of fit in model (2).]

Below are fits for all models that involve Ag and either one or two other explanatory variables.

Model Variables	*df*	G^2	A^*
Ag,S,D,Ch,H,W	193	134.9	7.0
Ag,S,D	196	141.4	7.5
Ag,S,Ch	196	139.3	5.4
Ag,S,H	196	141.9	8.0
Ag,S,W	196	138.4	4.5
Ag,D,Ch	196	139.0	5.1
Ag,D,H	196	141.4	7.5
Ag,D,W	196	138.5	4.6
Ag,Ch,H	196	139.9	6.0
Ag,Ch,W	196	135.5	1.6
Ag,H,W	196	138.1	4.2
Ag,S	197	141.9	6.0
Ag,D	197	141.4	5.5
Ag,Ch	197	139.9	4.0
Ag,H	197	142.7	6.8
Ag,W	197	138.8	2.9
Ag	198	142.7	4.8

Based on the A^* values, two models stand out:

$$\log[p_i/(1-p_i)] = \gamma_0 + \gamma_1 Ag_i + \gamma_2 W_i \tag{3}$$

with $A^* = 2.9$ and

$$\log[p_i/(1-p_i)] = \eta_0 + \eta_1 Ag_i + \eta_2 W_i + \eta_3 Ch_i \tag{4}$$

with $A^* = 1.6$.

The estimated parameters and standard errors for model (3) are

Variable	Parameter	Estimate	SE
Intercept	γ_0	-7.513	1.706
Ag	γ_1	0.06358	0.01963
W	γ_2	0.01600	0.00794

For model (4), these are

Variable	Parameter	Estimate	SE
Intercept	η_0	−9.255	2.061
Ag	η_1	0.05300	0.02074
W	η_2	0.01754	0.003575
Ch	η_3	0.006517	0.008243

The coefficients for Ag and W are quite stable in the two models. The coefficients of Ag, W, and Ch are all positive, so that a small increase in age, weight, or cholesterol is associated with a small increase in the odds of having a coronary incident. Note that we are establishing association, not causation.

As in regular regression, interpreting regression coefficients can be very tricky. The fact that the regression coefficients are all positive conforms with the conventional wisdom that high values for any of these factors increases one's chance of heart trouble. However, as in standard regression analysis, correlations between predictor variables can make interpretations of individual regression coefficients almost impossible.

It is interesting to note that from fitting model (1), the estimated regression coefficient for D, diastolic blood pressure, is negative. A naive interpretation would be that as diastolic blood pressure goes up, the probability of a coronary incident goes down. (If the log odds go down, the probability goes down.) This is contrary to common opinion about how these things work. Actually, this is really just an example of the fallacy of trying to interpret regression coefficients. The regression coefficients have been determined so that the fitted model explains these particular data as well as possible. As mentioned, correlations between the predictor variables can have a huge effect on the estimated regression coefficients. The sample correlation between S and D is .802, so estimated regression coefficients for these variables are unreliable. Moreover, it is not even enough just to check pairwise correlations between variables; any large partial correlations will also adversely affect interpretations. Fortunately, such correlations should not normally have an adverse affect on the predictive ability of the model; they only adversely affect attempts to interpret regression coefficients. In Chapter 13, we will see that the regression coefficients also depend on the precise form of the logit model. Other methods for modeling the probabilities that are both reasonable and very similar to logistic regression can have very different regression coefficients while giving very similar probabilities. Finally, in this particular example, another excuse for the D coefficient $\hat{\beta}_3$ being negative is that from the z value, β_3 is not significantly different from zero.

The estimated blood pressure coefficients from model (1) also suggest an interesting hypothesis. (The hypothesis would be much more interesting if the individual coefficients were significant, but we wish to demonstrate a modeling technique.) The coefficient for D is −.00694, which is approximately the negative of the coefficient for S, .00686. This suggests that

perhaps the difference $S - D$ would be just as valuable a predictor as the individual predictors S and D. We can evaluate this by fitting

$$\log[p_i/(1-p_i)] = \gamma_0 + \gamma_1 Ag_i + \gamma_2(S_i - D_i) + \gamma_3 Ch_i + \gamma_4 H_i + \gamma_5 W_i,$$

which gives $G^2 = 134.9$ on $df = 194$. This model is a special case of model (1), so a test of it against model (1) has

$$G^2 = 134.9 - 134.9 = 0$$

with $df = 194 - 193 = 1$. The G^2 is essentially zero, so the data are consistent with the reduced model. Of course, this reduced model was suggested by the fitted full model, so any formal test would be biased — but then one does not accept null hypotheses anyway, and the whole point of choosing this reduced model was that it seemed likely to give a G^2 close to that of model (1). We note that the new variable, $S - D$, is still not significant; it only has a z value of $.006834/.01877 = .36$.

Another way to view the procedure of the previous paragraph would be as a test of $H_0 : \beta_3 = -\beta_2$ in model (1). If we incorporate this hypothesis into model (1), we get

$$\begin{aligned} &\log[p_i/(1-p_i)] \\ &\quad = \beta_0 + \beta_1 Ag_i + \beta_2 S_i + (-\beta_2) D_i + \beta_4 Ch_i + \beta_5 H_i + \beta_6 W_i \\ &\quad = \beta_0 + \beta_1 Ag_i + \beta_2(S_i - D_i) + \beta_4 Ch_i + \beta_5 H_i + \beta_6 W_i \end{aligned}$$

as displayed above. (Whether we call the parameters β's or γ's is irrelevant.)

We learned earlier that, relative to model (1), either model (3) or (4) does an adequate job of explaining the data. This conclusion was based on looking at A^* values, but would also be obtained by doing formal tests of models. Thus, we know that age and weight are important variables in explaining coronary incidents. Moreover, cholesterol may also be an important variable. However, we have not explained most of the variability in coronary incidents.

Consider a measure analogous to R^2. The smallest interesting logistic regression model is $\log[p_i/(1-p_i)] = \gamma_0$. As seen earlier, this has $G^2 = 154.6$ on 199 degrees of freedom. The percent of variability explained by model (3) is

$$R^2(Ag, W) = \frac{154.6 - 138.8}{154.6} = .10\,,$$

which seems pretty pathetic. In fact, the R^2 from the full model is not much better

$$R^2(Ag, S, D, Ch, H, W) = \frac{154.6 - 134.9}{154.6} = .13\,.$$

We are a very long way from fitting the data as well as the saturated model fits. In fact, if we fit a 28-parameter model including all variables,

all squares of variables, and all two-factor cross-product terms, G^2 is 108.8, so R^2 is still only .30.

Granted, we have not explained most of the variation in the data, but it was probably not reasonable to think that we could. In standard regressions, a perfect fitting model can have a low R^2. This happens when there is substantial pure error in the model. The same thing happens in logistic regression. That fact will be illustrated in the next section.

3.2 Measuring Model Fit

In regression and ANOVA, R^2 is large when the pure error $\text{Var}(y_i) = \sigma^2$ is small. When σ^2 is unknown, there is always the hope that it will be small (if we can find the correct model). In logistic regression, $\text{Var}(y_i) = N_i p_i(1-p_i)$. There isn't any hope of making this universally small. You can only make it truly small by looking at uninteresting data — those with p_i near 0 or 1. Cases with a realistic chance of going either way make for large variability.

We begin our examination of how well models fit by looking at the likelihood ratio and Pearson test statistics as applied to logistic regression. As before, we have I independent binomials, each consisting of one trial. Thus, we have a logistic regression with I cases and the dependent variable y is either 0 or 1. Equivalently, we have an $I \times 2$ table with counts n_{ij}, where $n_{i1} + n_{i2} = 1$. Let the variable y correspond to counts in the first column of the table so that $y_i = n_{i1}$ and $1 - y_i = n_{i2}$. If a logistic regression model for $\log[p_i/(1-p_i)]$ is fitted, we obtain estimates $\hat{p}_i$ of the p_i's. For the corresponding log-linear model, $\hat{p}_i = \hat{m}_{i1}$ and $(1-\hat{p}_i) = \hat{m}_{i2}$. The likelihood ratio and Pearson test statistics against the saturated model were given in Section 2.6.

EXAMPLE 3.2.1. Suppose that in a true and very accurate model, there are 30 observations (i values) each with $p_i \doteq .1 \doteq \hat{p}_i$ and 30 with $p_i \doteq .9 \doteq \hat{p}_i$. From the first 30, we could then expect to get about three observations with $y_i = 1$. From (2.6.6), each of these observations has a crude standardized residual of about

$$\frac{(1-.1)}{\sqrt{.1(1-.1)}} = 3\,.$$

The other 27 observations will have residuals of

$$\frac{(0-.1)}{\sqrt{.1(1-.1)}} = -.333\,.$$

Similarly, from the second 30 observations, three would be about $(0-.9)/\sqrt{.9(1-.9)} = -3$ and 27 would be about $(1-.9)/\sqrt{.9(1-.9)} = .333$.

It is disturbing that this perfect model with a perfect fit has what usually would be considered large residuals.

Using the formula for X^2 from Section 2.6, the Pearson statistic will be about

$$X^2 \doteq 6(3^2) + 54(.333^2) = 60\,,$$

which is the number of cases. No matter how accurate the model is, the Pearson statistic for these observations will still be about 60. It will never get small. A similar phenomenon holds for the likelihood ratio test statistic. Under the same circumstances as discussed above,

$$G^2 \doteq 2[6\log(1/.1) + 54\log(1/(1-.1))] = 33.32\,.$$

(Asymptotic theory does not hold for these tests, so there is no reason to expect X^2 and G^2 to be about the same.) Note that a constant model for these data would have $\hat{p}_i \doteq .5$ and

$$G^2 = 2(60)\log(1/.5) = 83.18,$$

so for this essentially perfect model which is fit perfectly,

$$R^2 = \frac{83.18 - 33.32}{83.18} = .599;$$

not very high under the circumstances. In fact, since the probabilities in this example were chosen to be quite extreme (near 0 and 1), the observations have unusually low variability, which should actually inflate R^2. The moral is simply that one should not expect to see the very high R^2's that one sometimes gets in standard regression.

No matter how accurate the fitted model, the test statistics will not become arbitrarily small nor will R^2 approach 1. The likelihood ratio statistic G^2 will become small only as the intrinsic variability of the true model decreases, i.e., as all probabilities approach 0 or 1. The Pearson statistic evaluated at the true model will remain near I, the number of cases.

There are two morals to all of this. First, R^2 type measures can be used to measure relative goodness of fit but may be misleading if used to measure absolute goodness of fit. Models with low R^2's can fit great. Models with high R^2's can exhibit lack of fit. Second, residuals from logistic regression cannot be used without special consideration given to the 0-1 nature of the data (cf. Jennings, 1986).

EXAMPLE 3.2.2. Using model (4.1.4), one finds that of the 200 cases in the Chapman data, 26 cases had a crude standardized residual in excess of .97. The 26 cases were precisely the 26 cases that had coronary incidents. A method for identifying unusual cases that indicates that every case with a coronary event is unusual leaves something to be desired.

3.2.1 Checking Lack of Fit

Methods of checking for lack of fit in logistic regression have been discussed by Tsiatis (1980), Landwehr, Pregibon, and Shoemaker (1984), and Fienberg and Gong (1984). Their approaches are based on clustering near replicates of the regression variables so that something akin to pure error can be identified. We present here a method of evaluation inspired by standard residual analysis. Rather than clustering near replicates, it clusters cases with similar $\hat{p}_i$'s.

A standard method for identifying lack of fit in regression analysis is to plot the residuals against the predicted values. This plot should form a structureless horizontal band about zero (cf. Christensen, 1996b, Section 13.4). An equivalent plot would be the observations versus the predicted values. This plot should form a structureless band about the line with slope 1 and intercept 0. For logistic regression, a plot of observations versus predicted values should show predicted values near 0 having most observations equal to 0, and predicted values near 1 having most observations equal to 1; predicted values near .5 should have about equal numbers of observations that are 0s and 1s, etc. Such a plot could be difficult to interpret visually, so let's get cruder.

Break the predicted values into, say, 10 intervals: [0,.1), [.1,.2), [.2,.3), ..., [.9,1]. For each interval, find the number of cases that have $\hat{p}$ in the interval and the number of those cases that have $y = 1$. The midpoint of the interval multiplied by the number of cases should be close to the number of cases with $y = 1$. The fit within each interval can be summarized by looking at components of a Pearson-like statistic

$$\frac{[(\text{cases with } y = 1) - (\text{total interval cases})(\text{midpoint})]^2}{(\text{total interval cases})(\text{midpoint})}.$$

These case values can be added to obtain a summary measure of the goodness of fit.

Example 3.2.3. For the Chapman data using model (4.1.4), no $\hat{p}_i$ values are greater than .6. Intervals, total cases, coronary incidents, expected values (cases times midpoints), and components are listed below.

$\hat{p}$ Interval	Number of Cases	Coronary Incidents	Cases × Midpoint	Components
[0,.1)	99	5	4.95	0.0005
[.1,.2)	60	10	9	0.1111
[.2,.3)	22	2	5.5	2.2273
[.3,.4)	10	5	3.5	0.6429
[.4,.5)	7	2	3.15	0.4198
[.5,.6)	2	2	1.1	0.7364
				Total = 4.138

The component for the interval [.2,.3) is much larger than the others. If there is a lack of fit in evidence, it is most likely that people with estimated probabilities of a coronary incident between .2 and .3 actually have a considerably lower chance of having a coronary. On the other hand, if the components are at all analogous to χ^2's, even the interval [.2,.3) is not clear evidence of lack of fit. Based on this rather questionable comparison, neither the individual value 2.2273 nor the total 4.138 are unreasonable.

A possible improvement to this technique is, rather than taking the midpoint of the interval, to average the $\hat{p}$'s in the interval. For the interval [.2,.3), the average of the 22 $\hat{p}$'s is .24045 with a corresponding cell component of 2.0461. This indicates even less lack of fit.

3.3 Logistic Regression Diagnostics

We assume that the reader is familiar with diagnostics for standard regression. The diagnostics for logistic regression to be discussed are analogues of common methods used for standard regression. In standard regression, some of the usual diagnostic statistics are the residuals, standardized (studentized) residuals, standardized predicted (deleted) residuals (t residuals), Cook's distances, and the leverages, i.e., the diagonal elements of the projection operator ("hat matrix").

A primary use of residuals is in detecting outliers. However, as we have seen, for data consisting of 0s and 1s, the detection of outliers presents some unusual problems. When there are only two outcomes, it is difficult to claim that seeing either of them constitutes an outlier. If we have too many 0s or 1s in situations where we would not expect them (e.g., too many 1s in situations that we think have a small probability of yielding a 1), then we have a problem, but the problem is best thought of as a lack of fit. Moreover, we have seen that with 0-1 data, perfectly reasonable observations can have "unusually large" residuals.

Another use for residuals is in checking normality. For log-linear models, this can be thought of as checking how well the asymptotic theory holds. Unlike ANOVA type log-linear models, for 0-1 data the residuals are not asymptotically normal, so, again, the usual residual analysis is not appropriate.

All in all, the residuals (and modified residuals) do not seem very useful in and of themselves. We will concern ourselves with examining leverages and influential observations. In particular, we will examine the logistic regression analogue of Cook's distance that was discussed in Pregibon (1981) and Johnson (1985).

There are two questions frequently asked about influential observations. One is, "In what sense is this observation influential?" The question is crucial. Observations are not "influential" in a vacuum. They may be influential to the estimated regression parameters; they may be influential

to the fitted probabilities. They may be influential to just about anything. When examining influential observations, one first decides on the important aspects of the model and then examines influence measures appropriate to those aspects. The author agrees with Johnson (1985) that, typically, the primary concern should be about influence on the fitted probabilities. In logistic regression, Cook's distance is a direct influence measure relative to the fitted regression coefficients, but, as Johnson has shown, it can be viewed as an approximation to his Kullback-Leibler (K-L) divergence measure and Cook and Weisberg's (1982) likelihood distance measure. Although the author's inclination is toward the K-L divergence measure, the absence of readily available computer software dictates that the discussion be focused on Cook's distance.

The second frequently asked question about influential observations is, "Given some influential observations, what do you do about them?" My answer is that you should worry about them. Primarily, you should worry about whether it is more appropriate to ignore the fact that they are influential or eliminate their influence by deleting them from the data and then refitting the model. Of course, all of this is complicated by the fact that whether or not a case is influential depends on what model is being fitted. In the end, the answer to this question must depend on the data and the purpose of the analysis.

Many standard logistic regression programs provide diagnostics. For example, SAS PROC LOGISTIC and BMDP-LR both provide them and they can also be obtained from GLIM. Some sample commands are given in Subsection 3.4.2. In addition, many standard regression programs routinely provide diagnostics and these can also be used to obtain logistic regression diagnostics because the Newton-Raphson method of fitting logistic regression models amounts to doing a series of weighted regressions.

Standard diagnostic quantities for each case are the log odds

$$\log[\hat{p}_i/(1-\hat{p}_i)] = \hat{\beta}_0 + \hat{\beta}_1 Ag_i + \hat{\beta}_2 Ch_i + \hat{\beta}_3 W_i,$$

the predictive probability $\hat{p}_i$, the leverage $\hat{a}_{ii}$, the large sample standard error of $\hat{p}_i$, which is $\sqrt{\hat{p}_i(1-\hat{p}_i)(1-\hat{a}_{ii})}$, the standardized residual

$$r_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)(1-\hat{a}_{ii})}},$$

the Pearson residual

$$\tilde{r}_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}},$$

the square of which is the ith component of Pearson's chi-square, the deviance residual

$$\pm\sqrt{2[y_i \log(y_i/\hat{p}_i) + (1-y_i)\log((1-y_i)/(1-\hat{p}_i))]}$$

where the sign is taken to be the same as the sign of $y_i - \hat{p}_i$, and a version of Cook's distance for logistic regression. These formulae are for y_i either 0 or 1 and, as discussed earlier, residuals are not very interesting in this case. When y_i is a binomial count between 0 and N_i, the residuals can be useful for reasonably large N_i. The standardized residuals then become

$$r_i = \frac{y_i - N_i\hat{p}_i}{\sqrt{N_i\hat{p}_i(1-\hat{p}_i)(1-\hat{a}_{ii})}}$$

with similar adjustments to other quantities, cf. Subsection 3.4.1.

It should be mentioned that, properly defined, the logistic regression version of Cook's distance for the ith case requires computation of both the estimated logistic regression coefficients and the estimated coefficients when the ith case is deleted from the data. Both sets of estimates require iterative computations and it will be desired to investigate Cook's distance for every case in the data. This can be expensive. To reduce costs, it is common practice to use estimates when the ith case is deleted that are the result of one iteration of Newton-Raphson with starting values taken as the estimates from the full data set. In other words, this involves doing only one weighted regression. These one-step procedures will be the subject of our discussion.

We now present the procedure for obtaining diagnostic values in the context of fitting the model

$$\log[p_i/(1-p_i)] = \beta_0 + \beta_1 Ag_i + \beta_2 Ch_i + \beta_3 W_i$$

to the Chapman data. Specifically, we discuss how to use the diagnostic procedures in standard regression to obtain diagnostics for logistic regression.

Using a logistic regression program, we get the fit

Variable	Parameter	Estimate	SE
Intercept	β_0	−9.255	2.061
Ag	β_1	0.05300	0.02074
Ch	β_2	0.006517	0.003575
W	β_3	0.01754	0.008243

Having fit the model, create a data file that contains Ag, Ch, W, y, and $\hat{p}$, where y consists of the 0-1 counts y_i and $\hat{p}$ is a new variable that consists of the 200 values for $\hat{p}_i$. This data file is used as input into a regression program that allows (a) transformations of variables, (b) weighted regression, (c) computation of leverages, and (d) computation of Cook's distances. Using the transformation capability, define weights for the regression, say RWT, with $RWT_i = \hat{p}_i(1-\hat{p}_i)$. Also, define two variables Y_0 and Y with

$$Y_{0i} = \log[\hat{p}_i/(1-\hat{p}_i)]$$

and

$$Y_i = Y_{0i} + (y_i - \hat{p}_i)/RWT_i\,.$$

See Subsection 4.4.1 for computing methods when the data are not binary.

The variable Y_{0i} can be used to help verify that things are working as they should. Use the regression program to fit

$$Y_{0i} = \beta_0 + \beta_1 Ag_i + \beta_2 Ch_i + \beta_3 W_i + e_i$$

with the weights RWT_i. The regression coefficients from this fit should be identical to those obtained from the logistic regression program.

Now fit

$$Y_i = \beta_0 + \beta_1 Ag_i + \beta_2 Ch_i + \beta_3 W_i + e_i$$

with weights RWT_i. This gives estimated regression coefficients that are one additional step of the Newton-Raphson algorithm beyond those obtained by the logistic regression program. In this example, the regression gives

Variable	Parameter	Estimate	SE
Intercept	β_0	−9.256	2.066
Ag	β_1	0.05300	0.02077
Ch	β_2	0.006518	0.003578
W	β_3	0.017539	0.008248

$MSE = .9942$

The parameter estimates are very close to the original logistic regression estimates, but need not be identical to them. (They are close enough that the logistic regression program concluded that the results had converged.) The MSE is close to one, so the reported standard errors are also very close in the regression and the logistic regression. In logistic regression, there is no variance parameter to be estimated, as there is in standard regression, so for logistic regression, anything from a standard regression that involves the MSE must have that involvement eliminated. Appropriate standard errors are the reported standard errors divided by $\sqrt{MSE}$.

From this standard regression fit, we can also obtain leverages, standardized residuals, and Cook's distances. The leverages are precisely those suggested by Pregibon (1981). The standardized residuals and Cook's distances reported by a standard regression program also involve adjustments for the MSE. For logistic regression, those adjustments must be eliminated. The reported Cook's distances from the standard regression are essentially the one-step logistic regression Cook's distances. The difference in the Cook's distances is that the reported values are the true one-step estimates divided by the MSE. In the discussion below, the reported Cook's distances have been multiplied by MSE to give the appropriate values. In any case, because the values from the program are all being divided by the same

number, to make comparisons among cases the values could be used without modification. Another difference is that in standard regression, Cook's distance involves dividing by the number of regression parameters. Often in logistic regression programs, this division is not used. So in this example, to get the Cook's distances given by, say, BMDP-LR, the Cook's distances reported by the regression program have to be multiplied by $4\,MSE$. For what they are worth with 0-1 data, the standardized residuals reported by the regression program times $\sqrt{MSE}$ are the correct standardized residuals.

The nine cases with the highest leverages are

Case	19	38	41	84	108	111	116	153	157
Leverage	.104	.081	.149	.079	.147	.062	.067	.090	.093

The two that really stand out are Cases 41 and 108. Case 41 has $Ag = 40$, $W = 169$, and $Ch = 520$. This is an exceptionally high cholesterol value. Of the 200 cases, only 9 have Ch values over 400 and only 3 have Ch values over 428. These are Case 116 with $Ch = 453$, Case 38 with $Ch = 474$, and Case 41. A similar phenomenon occurs with Case 108. It has $Ag = 51$, $W = 262$, and $Ch = 269$. The weight of 262 pounds is extremely high within the data set. Of the nine cases with high leverage, only Cases 19, 41, and 111 correspond to men that had coronary incidents.

Denote the Cook's distances as C_i's. There are 32 cases with $C_i \geq .01$. These include all 26 of the individuals who had coronary incidents. Of the other six cases, four were also among the highest leverage cases and the remaining two also had reasonably high leverages.

Only four cases had $C_i \geq .05$. These are

Case	C_i	Leverage
41	.112	.149
86	.078	.008
126	.079	.042
192	.064	.022

All of these correspond to individuals with coronary incidents. If we compare the values $4\,C_i$ to a $\chi^2(4)$ distribution as suggested by Johnson (1985), we can get some global idea of the amount of influence each case is having. The conclusion is that none of the cases has much effect on the fitted model. The multiplier and *df* of 4 for calibrating C_i were the number of regression coefficients in the logistic regression. Of course, to compare these values to a chi-squared distribution, it is vital that the C_i's be computed properly; i.e., values from a standard regression program have to be multiplied by MSE.

Case 41 is easily the most influential, so it is of interest to examine what happens if this case is deleted from the data. For the most part, the fitted

p_i's are similar. The estimated coefficients with Case 41, without Case 41, and standard errors without Case 41 are given below.

Variable	Estimate with Case 41	Estimate without Case 41	SE without Case 41
Intercept	−9.255	−8.813	2.058
Ag	0.05300	0.05924	0.02138
Ch	0.006517	0.004208	0.003881
W	0.01754	0.01714	0.008216

The estimates have changed but not dramatically. Perhaps the most striking aspect is the change in the evidence for the effect of cholesterol. With Case 41 deleted, the estimate divided by the standard error is .004208/.003881 = 1.08. With Case 41 included, this value is .006517/.003575 = 1.82. The inclusion of cholesterol in the model was questionable before; without Case 41, there seems little need for it.

With Case 41 deleted, $G^2(Ag, Ch, W) = 132.8$ on 195 degrees of freedom. $G^2(Ag, W) = 134.0$ on 196 degrees of freedom. The difference in G^2's is $134.0 - 132.8 = 1.2$ with 1 degree of freedom. There is virtually no evidence for including cholesterol in the model. The estimated coefficients without Case 41 using only Ag and W are

Variable	Estimate	SE
GM	−7.756	1.745
Ag	0.06675	0.02013
W	0.01625	0.008042

Thus, we have found that almost all of the evidence for a cholesterol effect is based on the fact that one individual with a very high cholesterol level had a coronary incident. We can just drop that individual and state that for more moderate levels of cholesterol, the numerical level of cholesterol does not enhance our predictive ability. But this conclusion is already indicating that qualitatively different things happen at different cholesterol levels. Why not try to incorporate that idea into the models being considered? One could group cases based on cholesterol levels and fit different models to different groups. Rather than forming groups, perhaps cholesterol levels should be transformed before being used in the model. Whatever course the eventual analysis takes, Case 41 has directed our attention to the role of cholesterol. We now must question whether the current forms of our models are adequate for approximating the effect of cholesterol or whether the effect of cholesterol may be an oddity caused by one individual who just happened to have a coronary incident and very high cholesterol.

3.4 Model Selection Methods

Formal model selection methods can be based either on stepwise methods or finding best subsets of variables based on some criterion (e.g., Akaike's information). Fitting lots of models can be very expensive because each fit requires an iterative procedure. Stepwise methods are sequential, hence cheaper than best subset methods.

Standard computer programs are available for doing stepwise logistic regression, e.g., BMDP-LR and SAS PROC LOGISTIC. These operate in a fashion similar to standard regression (cf. Christensen, 1996a, 1996b; Draper and Smith, 1981; Weisberg, 1985). They are also very similar to the methods discussed in Sections 6.1 and 6.3. We will not give a detailed discussion.

To the best of the author's knowledge, the only standard computer program available for doing best subset logistic regression is SAS PROC LOGISTIC. This procedure is based on doing score tests, a subject that will be discussed below. In addition, programs for doing standard best subset selection can be used with one-step estimates of logistic regression parameters to identify good candidate models. To do this, the best subset program must allow weighted regression.

EXAMPLE 3.4.1. Model (3.1.1) was fitted to the Chapman data to obtain $\hat{p}_i$'s. We then defined two variables: a weight variable

$$RWT_i = \hat{p}_i(1-\hat{p}_i)$$

and a dependent variable

$$Y_i = \log[\hat{p}_i/(1-\hat{p}_i)] + (y_i - \hat{p}_i)/RWT_i\,.$$

The best subset regression program BMDP-9R was used employing the weights RWT to get best subsets of

$$Y_i = \beta_0 + \beta_1 Ag_i + \beta_2 S_i + \beta_3 D_i + \beta_4 Ch_i + \beta_5 H_i + \beta_6 W_i + e_i\,.$$

(Note the similarities to the procedure for getting diagnostic statistics.) The fits used in comparing various models are not fully iterated maximum likelihood fits. They involve one-step of the Newton-Raphson algorithm starting at the maximum likelihood fit for model (3.1.1). Determinations of best-fitting models are based on residual sums of squares rather than G^2's.

Based on the C_p statistic, the five best-fitting models are

Variables	C_p
Ag, Ch, W	1.66
Ag, W	2.88
Ag, Ch, H, W	3.13
Ag, S, Ch, W	3.49
Ag, D, Ch, W	3.59

The last three models are among the best because adding a worthless variable to the good model based on Ag, Ch, and W cannot do too much harm. The two most interesting models are precisely those identified earlier by less systematic means. In fact, in this example, the C_p statistics are very similar to the corresponding A^* values.

Of course, the C_p statistics are based on one-step fits. Below, we compare the MLEs for the model with Ag, Ch, and W to the one-step fit.

Variable	MLE	One-Step
Intercept	−9.2559	−9.21822
Ag	0.053004	0.0529624
Ch	0.0065179	0.00647380
W	0.017539	0.0174699

(Note that the MLEs differ slightly from the values given previously. The earlier values were obtained from the program GLIM. These values were obtained from BMDP-LR. It is normal for different [correct] programs to give *slightly* different answers.) The C_p's are not based on fully iterated fits, so it is probably a good idea to consider a larger number of models than one ordinarily would in standard regression. One hopes that the best fitting fully iterated models will be among the best fitting one-step models, but the relationship need not be exact.

This method of obtaining best subsets using one-step approximations is very natural, so it is not surprising that it has been discovered independently several times. The earliest references of which I am aware are Nordberg (1981, 1982).

As mentioned earlier, to the best of my knowledge, the only method for best subset regression that appears in a standard computer package is a method in SAS PROC LOGISTIC that gives the models with the best score tests. Score tests are arrived at in a similar fashion to the procedures discussed above. With regard to Example 3.4.1, to get the score test for dropping all of Ag, S, D, Ch, H, and W, fit the full regression model as indicated in the example but with one exception. The exception is that RWT and Y are defined as indicated using the $\hat{p}_i$'s, but in Example 3.4.1, the $\hat{p}_i$'s were obtained from a maximum likelihood fit of the full model, whereas for a score test, the $\hat{p}_i$'s are obtained from a maximum likelihood fit of the model that contains only an intercept. The score test statistic

for whether the six variables can be dropped is just the sum of squares for regression in the six-variable weighted regression model. The statistic is compared to a $\chi^2(6)$ distribution.

One nice thing about score tests is that the $\hat{p}_i$'s depend just on the intercept-only model, so getting the score statistics for testing any model against the intercept-only model is merely a matter of fitting a regression on that model. In other words, with the same definitions for RWT_i and Y_i, one can test the full model as well as models such as

$$Y_i = \beta_0 + \beta_1 Ag_i + \beta_4 Ch_i + e_i$$

and

$$Y_i = \beta_0 + \beta_1 Ag_i + \beta_4 Ch_i + \beta_6 W_i + e_i$$

simply by fitting the regressions and evaluating the sums of squares for regression.

Of course, the method presented in Example 3.4.1 is essentially the same except that the $\hat{p}_i$'s are taken from the (presumably more accurate) full model rather than the no-intercept model (which nobody takes seriously as a model). Also, some account of model size is being taken by looking at C_p statistics. If one specified best subset selection using the R^2 criterion in the regression program, *the only difference in the Nordbert and score procedures for choosing the best models would be in the choice of $\hat{p}_i$'s.*

3.4.1 Computations for Nonbinary Data

In this section and Section 3, we have considered only the case where the counts are either 0s or 1s. In Section 2.6 and in later chapters, we consider logistic models and/or theory for data involving counts that may be greater than 1. The computing methods discussed here are easily adapted to those situations. If the ith case has N_i trials (i.e., the possible values for y_i are $0, 1, \ldots, N_i$), then the appropriate weights are

$$RWT_i = N_i \hat{p}_i (1 - \hat{p}_i)$$

and the dependent variable in the regressions is

$$Y_i = \log[\hat{p}_i/(1 - \hat{p}_i)] + (y_i - N_i \hat{p}_i)/RWT_i \,.$$

Note that

$$\hat{p}_i/(1 - \hat{p}_i) = N_i \hat{p}_i/(N_i - N_i \hat{p}_i) \,.$$

Often logistic regression computer programs will provide the values $N_i \hat{p}_i$ rather than $\hat{p}_i$ as diagnostics. Finally, if all of the N_i's are large, then looking at the standardized residuals becomes reasonable. Also, when all the N_i's are large, tests against the saturated model can be validly compared to a chi-squared distribution.

3.4.2 Computer Commands

Below are SAS, BMDP, and GLIM commands for obtaining a logistic regression. The data are in a file 'chapman.dat' with eight columns: the case index, Ag, S, D, Ch, H, W, and Cnt. The file looks like this.

```
  1 44 124  80 254 70 190 0
  2 35 110  70 240 73 216 0
  3 41 114  80 279 68 178 0
  4 31 100  80 284 68 149 0
       data continue
199 50 128  92 264 70 176 0
200 31 105  68 193 67 141 0
```

We begin with SAS commands.

Perhaps the simplest way to fit the logistic regression model (4.1.4) in SAS is to use PROC GENMOD. The first line controls printing. The next four lines involve defining and reading the data and creating a variable "n" that gives the total number of possible successes for each case. The remaining lines specify the model and that a logistic regression is to be performed.

```
options ps=60 ls=72 nodate;
data chapman;
   infile 'chapman.dat';
   input ID Ag S D Ch H W Cnt;
   n = 1;
proc genmod data=chapman ;
   model Cnt/n = Ag Ch W / link=logit
                           dist=binomial;
run;
```

A more powerful program for logistic regression is PROC LOGISTIC.

```
options ps=60 ls=72 nodate;
data chapman;
   infile 'chapman.dat';
   input ID Ag S D Ch H W Cnt;
proc logistic data=chapman descending;
   model Cnt=Ag Ch W / waldcl waldrl plcl
                       influence iplots lackfit rsq;
   output out=chdiag predicted=phat;
run;
proc print data=chdiag;
run;
proc logistic data=chapman descending;
   model Cnt=Ag S D Ch H W / selection = score
                             best = 3 details;
```

```
run;
```

This program includes two calls of PROC LOGISTIC. The first is a standard procedure for obtaining a logistic regression. The second involves model selection. On the line with "proc logistic", one specifies the data being used and the command "descending". The command "descending" is used so that the program models the probabilities of events coded as 1 rather than events coded as 0. In other words, it makes the program model the probability of a coronary incident rather than the probability of no coronary incident. Standard output includes the estimated regression coefficients, standard errors, values of z^2, $P\ values$, and $e^{\hat{\beta}_k}$'s. The model statement is straightforward, specifying the dependent variable and the predictor variables. After the / on the model line, options are specified. "waldcl" causes the program to give the confidence intervals $\hat{\beta}_k \pm 1.96\,\mathrm{SE}(\hat{\beta}_k)$; call the interval (a, b). "waldrl" causes the program to give the values $e^{\hat{\beta}_k}$ and intervals (e^a, e^b). "plcl" gives alternative confidence intervals for the β_k's based on profile likelihoods. The command "influence" causes diagnostics to be presented, basically everything discussed by Pregibon (1981). This includes leverages, Cook's distance C_i (the same version as BMDP presents), and something called Cbar, which is $(1 - \hat{a}_{ii})C_i$. Index plots are given by specifying "iplots". For binary data, G^2 does not give a valid lack of fit test, "lackfit" gives a test similar in spirit to that discussed in Section 2. "rsq" gives values for R^2 and Adj. R^2, but R^2 is defined differently than it is here. The "output" command creates a SAS data set containing the diagnostics, so they can then be printed or manipulated in other ways.

The second proc logistic line was set up to do model selection. It uses the "selection" option. This can be set to "forward", "backwards", "stepwise", or "score". With the score option and "best = 3", the three one-variable models with the best score statistics, the three best two-variable models, the three best three-variable models, and so on, are all presented.

For BMDP, the commands are similar. You actually run the program BMDP-LR, so no statement of this being a logistic regression procedure is needed. Again the data are specified. The variables to be used are specified along with the dependent variable and the model. Interval and categorical variables must be specified prior to specifying the model.

```
/ INPUT     FILE = 'CHAPMAN.DAT'.
          FORMAT = FREE.
          VARIABLES = 8.
/ VARIABLE  NAMES = Index, Ag, S, D, Ch, H, W, Cnt.
          USE = Ag TO Cnt.
/ REGRESS   DEPENDENT = Cnt.
          INTERVAL = Ag, S, D, Ch, H, W.
             MODEL = Ag, Ch, W.
             MOVE =   0, 0, 0.
```

```
            METHOD = MLR.
/ PRINT     CELLS = MODEL.
/ END
```

The program is actually set up to do forward, backward, or stepwise regression. The "move" command was used to make the program fit only the model desired. Diagnostics are obtained by the "cells = model" specification.

Finally, for anyone who might want to use GLIM (still one of my favorites):

```
$units 200$
$data I Ag S D Ch H W Cnt $
$dinput 6$
$calc n = 1$
$yvar Cnt$
$error binomial n$
$fit Ag+Ch+W$
$display e$
$extract %vl$
$calc ahat=%vl*%wt/%sc$                  (leverages)
$calc r=(Cnt-%fv)/%sqrt(%sc*(1-ahat))$   (std. resids)
$calc C=(ahat/(1-ahat))*(r**2)$          (Cook's distances)
$look ahat r C$
```

"units" specifies the number of cases in the regression. After "dinput 6", DOS versions of GLIM prompt you for a file name, i.e., "chapman.dat". The Cook's distances are the same as those used in SAS and BMDP and 4 times those defined here. GLIM is similar in spirit to PROC GENMOD.

3.5 ANOVA Type Logit Models

In this section, analysis of variance type models for the log odds of a two-category response variable are discussed. We begin with a standard example.

EXAMPLE 3.5.1. Consider the muscle tension data of Example 3.7.1. Recall that the factors and levels are

Factor	Abbreviation	Levels
Change in muscle tension	T	High, Low
Weight of muscle	W	High, Low
Muscle type	M	Type 1, Type 2
Drug	D	Drug 1, Drug 2

and the data are

Tension (h)	Weight (i)	Muscle (j)	Drug (k) Drug 1	Drug 2
High	High	Type 1	3	21
		Type 2	23	11
	Low	Type 1	22	32
		Type 2	4	12
Low	High	Type 1	3	10
		Type 2	41	21
	Low	Type 1	45	23
		Type 2	6	22

Change in tension can be viewed as a response factor. Weight, muscle type, and drug are all explanatory variables. Thus, it is appropriate to model the log odds of having a high change in muscle tension. The three explanatory factors affect the log odds for high tension change. The most general model available is to use a model that includes all main effects and all interactions between the explanatory factors, i.e.,

$$\begin{aligned} \log(p_{1ijk}/p_{2ijk}) &= G + W_i + M_j + D_k \\ &\quad + (WM)_{ij} + (WD)_{ik} + (MD)_{jk} \\ &\quad + (WMD)_{ijk}\,. \end{aligned} \tag{1}$$

As usual, this is equivalent to a model with just the highest-order interactions; in this case,

$$\log(p_{1ijk}/p_{2ijk}) = (WMD)_{ijk}\,.$$

Model (1) can be fit by maximum likelihood. Reduced models can be tested. Estimates and asymptotic standard errors can be obtained. In other words, the analysis of model (1) is similar to that of an (unbalanced) standard ANOVA model or a log-linear model.

Of course, the analysis of model (1) should be similar to that of a log-linear model analysis because in a profound sense (alluded to in Section 2.6 and discussed in detail in Chapter 11), model (1) is precisely the same model as the saturated log-linear model, i.e.,

$$\log(m_{hijk}) = (\tau\omega\mu\delta)_{hijk}\,, \tag{2}$$

where we have used Greek equivalents of T, W, M, and D to emphasize that the parameters in (1) and (2) are different. Note that in both models (1) and (2), there is at least one parameter on the right-hand side for every term on the left-hand side. In both models, the data are fitted perfectly. In

examining the correspondence between logit models and log-linear models, it is crucial to keep in mind the fact that this is a prospective study, so

$$p_{1ijk}/p_{2ijk} = m_{1ijk}/m_{2ijk}\,.$$

Now consider a more interesting logit model than the saturated logit model (1). Consider, say,

$$\log(p_{1ijk}/p_{2ijk}) = W_i + (MD)_{jk} \tag{3}$$

where we have eliminated the redundant terms G, M_j, and D_k and assumed that the terms $(WM)_{ij}$, $(WD)_{ik}$, and $(WMD)_{ijk}$ add nothing to model (1). We wish to find the corresponding log-linear model. Model (3) is a model that explains tension change odds, so an effect, say W_i, alters the odds of high tension change. The odds cannot be altered without altering both the probability of high tension change and the probability of low tension change; thus, W_i affects both of these probabilities. In other words, the probabilities (and the expected cell counts) depend on both the tension change level T and the weight W. It follows that the logit effect W_i corresponds to a log-linear model interaction, say $(\tau\omega)_{hi}$. Similarly, the logit effect $(MD)_{jk}$ corresponds to the interaction $(\tau\mu\delta)_{hjk}$. As shown in Section 11.1, the log-linear model must contain $(\omega\mu\delta)_{ijk}$ terms, so model (3) is equivalent to

$$\log(m_{hijk}) = (\tau\omega)_{hi} + (\tau\mu\delta)_{hjk} + (\omega\mu\delta)_{ijk}\,. \tag{4}$$

Inclusion of the terms $(\omega\mu\delta)_{ijk}$ is required to deal with the sampling scheme when thinking of the sampling as product-multinomial (i.e., independent binomials) for every combination of the explanatory factors. This mental device was discussed in the subsection of the introduction on retrospective versus prospective studies.

In fact, these ideas extend to all logit and logistic regression models. Note that the shorthand notation used for ANOVA type log-linear models is easily adapted to ANOVA type logit models. Using this shorthand, the correspondence between logit models and log-linear models is illustrated in Table 3.1.

In each case, the effects in the logit model correspond to log-linear model effects that are the interaction between T and the logit model terms. In addition, the log-linear models always include the three-way interaction between the explanatory factors. Note that models (3) and (4) correspond to line 7 of Table 3.1. As another example, line 3 of the table indicates that the model

$$\log(p_{1ijk}/p_{2ijk}) = G + W_i + M_j + D_k + (WM)_{ij} + (MD)_{jk}$$

is equivalent to the model

$$\log(m_{hijk}) \quad = \quad \gamma + \omega_i + \mu_j + \delta_k + (\omega\mu)_{ij} + (\omega\delta)_{ik} + (\mu\delta)_{jk} + (\omega\mu\delta)_{ijk}$$

$$+ \tau_h + (\tau\omega)_{hi} + (\tau\mu)_{hj} + (\tau\delta)_{hk}$$
$$+ (\tau\omega\mu)_{hij} + (\tau\omega\delta)_{hik}.$$

Of course, the log-linear model can be written much more simply as

$$\log(m_{hijk}) = (\omega\mu\delta)_{ijk} + (\tau\omega\mu)_{hij} + (\tau\omega\delta)_{hik}\,.$$

TABLE 3.1. Correspondence Between Some Logit and Log-Linear Models

	Logit Model	Log-Linear Model
1)	{WM}{WD}{MD}	[WMD][TWM][TWD][TMD]
2)	{WM}{WD}	[WMD][TWM][TWD]
3)	{WM}{MD}	[WMD][TWM][TMD]
4)	{WD}{MD}	[WMD][TWD][TMD]
5)	{WM}{D}	[WMD][TWM][TD]
6)	{WD}{M}	[WMD][TWD][TM]
7)	{MD}{W}	[WMD][TMD][TW]
8)	{W}{M}{D}	[WMD][TW][TM][TD]
9)	{W}{M}	[WMD][TW][TM]
10)	{W}{D}	[WMD][TW][TD]
11)	{M}{D}	[WMD][TM][TD]

Given the log-linear models, the logit models can be obtained by subtraction. Using model (4), observe that

$$\begin{aligned}
\log(p_{1ijk}/p_{2ijk}) &= \log(m_{1ijk}/m_{2ijk}) \\
&= \log(m_{1ijk}) - \log(m_{2ijk}) \\
&= (\tau\omega)_{1i} + (\tau\mu\delta)_{1jk} + (\omega\mu\delta)_{ijk} \\
&\quad - (\tau\omega)_{2i} - (\tau\mu\delta)_{2jk} - (\omega\mu\delta)_{ijk} \\
&= [(\tau\omega)_{1i} - (\tau\omega)_{2i}] + [(\tau\mu\delta)_{1jk} - (\tau\mu\delta)_{2jk}] \\
&= W_i + (MD)_{jk}
\end{aligned}$$

where $W_i \equiv [(\tau\omega)_{1i} - (\tau\omega)_{2i}]$ and $(MD)_{jk} \equiv [(\tau\mu\delta)_{1jk} - (\tau\mu\delta)_{2jk}]$. Thus, model (4) implies model (3). Conversely, as will be seen in Chapter 11, the logit model (3) implies the log-linear model (4).

It is interesting to note that models other than (4) will also imply a logit structure of $\log(p_{1ijk}/p_{2ijk}) = W_i + (MD)_{jk}$. Any reduced model relative to (4) where the reduction involves only the $(\omega\mu\delta)_{ijk}$ terms also implies that $\log(p_{1ijk}/p_{2ijk}) = W_i + (MD)_{jk}$. For example, if $\log(m_{hijk}) = (\tau\omega)_{hi} + (\tau\omega\delta)_{hjk} + (\omega\mu)_{ij} + (\delta)_k$, then $\log(p_{1ijk}/p_{2ijk}) = \log(m_{1ijk}) - \log(m_{2ijk}) = W_i + (MD)_{jk}$. Such models imply that the logit structure holds *plus some additional conditions on the explanatory factors.* The logit structure of model (3) without any other conditions is equivalent to model (4), so if one is fitting logit models, the corresponding log-linear model must contain the

three-factor effects $(\omega\mu\delta)_{ijk}$. It will be recalled that our original argument for including the $(\omega\mu\delta)_{ijk}$ effects was based on the existence of product-binomial sampling. This condition is not necessary; a logit model implies the existence of the $(\omega\mu\delta)_{ijk}$ terms regardless of the sampling structure, cf. Section 11.1. Of course, in the absence of product-binomial sampling, it would seem to be difficult to interpret the terms $\log(p_{1ijk}/p_{2ijk})$ because we no longer have $p_{1ijk} + p_{2ijk} = 1$. Fortunately, if we condition on explanatory variables (i.e., if we condition on the marginal totals $n_{\cdot ijk}$), then for any of the standard prospective sampling schemes, the *conditional* sampling scheme is product-binomial and the standard interpretations can be used for the conditional distribution.

Before examining the actual analysis of the muscle tension data, we make one final comment about the logit model—log-linear model relationship. *A logit model can be thought of as a model fitted to a two-factor table, where one factor is tension and the other factor consists of all combinations of weight, muscle type, and drug.* The smallest interesting log-linear model is the model of independence:

$$\log(m_{hijk}) = \tau_h + (\omega\mu\delta)_{ijk}\,.$$

Looking at $\log(p_{1ijk}/p_{2ijk}) = \log(m_{1ijk}) - \log(m_{2ijk})$, we see that this model corresponds to a model $\log(p_{1ijk}/p_{2ijk}) = \tau_1 - \tau_2 \equiv G$, i.e., just fitting a grand mean. Intuitively, this would be the smallest interesting logit model. The saturated model for the two-factor table is the interaction model

$$\log(m_{hijk}) = \tau_h + (\omega\mu\delta)_{ijk} + (\tau\omega\mu\delta)_{hijk}\,,$$

which is the logit model $\log(p_{1ijk}/p_{2ijk}) = G + (WMD)_{ijk}$. The more interesting logit models correspond to modeling the interaction in this two-way table. They posit more interaction than complete independence, but less interaction than the saturated model. Note that thinking of this as a two-way table is also consistent with the idea of product-binomial sampling. The muscle tension data corresponds to a 8×2 table. Each row is a distinct set of explanatory variables, indexed by ijk. The columns are the two categories of the response, indexed by h. Each row is thought of as an independent binomial, so the row totals should be fixed by inclusion of a main effect for rows, i.e., the W, M, D three-way interaction.

We now return to the data analysis. Table 3.2 gives a list of logit models, *df*, G^2, *P values*, and $A-q$ values. The *df*'s, G^2's, P's, and $A-q$'s were actually obtained by fitting the corresponding log-linear models. Clearly, the best fitting logit models are the models {MD}{W} and {WM}{MD}. Both involve the muscle type—drug interaction and a main effect for weight. One of the models includes the muscle type—weight interaction.

We now take a closer look at the logit model {MD}{W}. As mentioned earlier, Table 3.2 was obtained by fitting the log-linear models corresponding to the logit model. The log-linear model corresponding to {MD}{W}

TABLE 3.2. Statistics for Logit Models

Logit Model	df	G^2	P	$A - q$
{WM}{WD}{MD}	1	0.111	.7389	−1.889
{WM}{WD}	2	2.810	.2440	−1.190
{WM}{MD}	2	0.1195	.9417	−3.8805
{WD}{MD}	2	1.059	.5948	−2.941
{WM}{D}	3	4.669	.1966	−1.331
{WD}{M}	3	3.726	.2919	−2.274
{MD}{W}	3	1.060	.7898	−4.940
{W}{M}{D}	4	5.311	.2559	−2.689
{W}{M}	5	11.35	.0443	1.35
{W}{D}	5	12.29	.0307	2.29
{M}{D}	5	7.698	.1727	−2.302

is [WMD][TMD][TW]. Each logit model term becomes an interaction with the response factor T and there is an interaction between all of the explanatory factors. The estimated expected cell counts for [WMD][TMD][TW] are given in Table 3.3.

TABLE 3.3. Estimated Expected Cell Counts for the Log-Linear Model [WMD][TMD][TW]

			Drug (k)	
Tension (h)	Weight (i)	Muscle (j)	Drug 1	Drug 2
High	High	Type 1	2.31	20.04
		Type 2	23.75	11.90
	Low	Type 1	22.68	32.96
		Type 2	3.26	11.10
Low	High	Type 1	3.69	10.97
		Type 2	40.24	20.10
	Low	Type 1	44.32	22.03
		Type 2	6.74	22.90

By taking the ratio of the high tension change estimates to the low tension change estimates, we obtain the estimated odds from the logit model. For example, as in Table 3.3, the high-tension, high-weight, type 1, drug 1 estimate is 2.308; the low-tension, high-weight, type 1, drug 1 estimate is 3.693. The ratio is $2.308/3.693 = .625$. This is the logit model estimate of the odds of a high-tension change for high-weight, type 1, drug 1. The estimated odds for all cells are given in Table 3.4.

The estimated odds of having a high tension change are 1.22 times greater for high-weight muscles than for low-weight muscles. For example, in Table 3.4, $.625/.512 = 1.22$ but also $1.22 = .590/.483 = 1.827/1.495 = .592/.485$. To put it another way, $\hat{m}_{11jk}\hat{m}_{22jk}/\hat{m}_{12jk}\hat{m}_{21jk} = 1.22$. This

TABLE 3.4. Estimated Odds of High Tension Change for the Logit Model {MD}{W}

		Drug	
Weight	Muscle	Drug 1	Drug 2
High	Type 1	.625	1.827
	Type 2	.590	.592
Low	Type 1	.512	1.496
	Type 2	.483	.485

corresponds to the main effect for weight in the logit model. The odds also involve a muscle type—drug interaction. The nature of this interaction is easily established. Consider the four estimated odds for high weights, $\hat{m}_{11jk}/\hat{m}_{21jk}$. These are the four values at the top of Table 3.4; e.g., for muscle type 1, drug 1, this is .625. In every muscle type—drug combination other than type 1, drug 2, the estimated odds of having a high tension change are about .6. The estimated probability of having a high tension change is about $.6/(1 + .6) = .375$. However, for type 1, drug 2, the estimated odds are 1.827 and the estimated probability of a high tension change is $1.827/(1 + 1.827) = .646$. The chance of having a high tension change is much greater for the combination muscle type 1, drug 2 than for any other muscle type—drug combination. A similar analysis holds for the low-weight odds $\hat{m}_{12jk}/\hat{m}_{22jk}$ but the actual values of the odds are smaller by a factor of 1.22 because of the main effect for weight.

The other logit model that fits quite well is {WM}{MD}. Tables 3.5 and 3.6 contain the estimated odds of high tension change for this model. The difference between Tables 3.5 and 3.6 is that the rows of Table 3.5 have been rearranged in Table 3.6. This sounds like a trivial change, but examination of the tables shows that Table 3.6 is easier to interpret.

TABLE 3.5. Estimated Odds for the Logit Model {WM}{MD}

		Drug	
Weight	Muscle	Drug 1	Drug 2
High	Type 1	.809	2.202
	Type 2	.569	.512
Low	Type 1	.499	1.358
	Type 2	.619	.557

Looking at the type 2 muscles, the high-weight odds are .919 times the low-weight odds. Also, the drug 1 odds are 1.111 times the drug 2 odds.

TABLE 3.6. Estimated Odds for the Logit Model {WM}{MD}

Muscle	Weight	Drug	
		Drug 1	Drug 2
Type 1	High	.809	2.202
	Low	.499	1.358
Type 2	High	.569	.512
	Low	.619	.557

Neither of these are really very striking differences. For muscle type 2, the odds of a high tension change are about the same regardless of weight and drug. Contrary to our previous model, they do not seem to depend much on weight, and to the extent that they do depend on weight, the odds go down rather than up for higher weights.

Looking at the type 1 muscles, we see the dominant features of the previous model reproduced. The odds of high tension change are 1.622 times greater for high weights than for low weights. The odds of high tension change are 2.722 times greater for drug 2 than for drug 1.

Both models indicate that for type 1 muscles, high weight increases the odds and drug 2 increases the odds. Both models indicate that for type 2 muscles, drug 2 does not substantially change the odds. The difference between the models {MD}{W} and {WM}{MD} is that {MD}{W} indicates that for type 2 muscles, high weight should increase the odds, but {WM}{MD} indicates little change for high weight and, in fact, what change there is indicates a decrease in the odds.

Incidentally, the reason for changing from Table 3.5 to Table 3.6 was the nature of the logit model. The model {WM}{MD} has M in both terms, so it is easiest to interpret when fixing the level of M. For a fixed level of M, the effects of W and D are additive, although the size of those effects change with the level of M.

This analysis of the data on change in muscle tension was intentionally performed at the lowest level of *technical* sophistication. The estimated expected cell counts were obtained by iterative proportional fitting. The entire analysis was based on these fitted values and the associated likelihood ratio test statistics. For example, conclusions about the importance of estimates were drawn without the benefit of standard errors for those estimates. Obtaining standard errors requires more computational sophistication. In particular, it requires fitting an auxiliary regression as discussed in Sections 6.7 and 10.2. However, it is interesting to see how much can be obtained from such a small computational investment.

There are two ways to fit an analysis of variance type logit (logistic) model. One way is to fit the corresponding log-linear model. The second way is to fit the logit model directly. This section has dealt exclusively

with fitting the corresponding log-linear models. Section 1 deals exclusively with fitting the logistic (logit) model directly. Although Section 1 deals specifically with regression models, the procedures for a direct fit of an ANOVA model are similar.

To reiterate, there are two principles that define the correspondence between logit models and log-linear models. Recall that effects in the logit model only involve the explanatory factors; e.g., logit effects for the tension change data never involve T, the response factor, only the explanatory factors. The first principle is that any effect in the logit model corresponds in the log-linear model to an interaction between the response factor and the logit effect. For example, a logit effect {MD} corresponds to a log-linear effect [TMD]. The second principle is that the log-linear model always includes the full interaction between the explanatory factors; e.g., all log-linear models include the [WMD] interaction. These principles also hold for logistic regression models. If g is an index or group of indexes that identify all levels of the predictor variables (i.e., explanatory factors), the log-linear model will have a term $u_{(g)}$ which is essentially the full interaction between the explanatory factors. Also, any linear logistic effect βx_g becomes a log-linear interaction $\eta_h x_g$ where h indexes the two levels of the response factor.

3.5.1 Computer Commands

The muscle tension data are listed in the file 'tenslr.dat' with one column for the number of high tension scores, one column for the low tension scores, and three columns of indices that specify the level of weight (high is 1), muscle type, and drug, respectively.

```
 3  3 1 1 1
21 10 1 1 2
23 41 1 2 1
11 21 1 2 2
22 45 2 1 1
32 23 2 1 2
 4  6 2 2 1
12 22 2 2 2
```

The following commands fit the model {WM}{WD}{MD} using SAS PROC GENMOD. This procedure works very much like GLIM. Note that the variable "n" is the total number of individuals with each level of weight, muscle type, and drug. As in Subsection 3.7.1, the "class" command is used to distinguish ANOVA type factors from regression predictors.

```
options ps=60 ls=72 nodate;
data tension;
   infile 'tenslr.dat';
```

```
   input H L W M D;
   n = H+L;
proc genmod data=tension;
   class W M D;
   model H/n = W*M W*D M*D / link=logit
                             dist=binomial;
run;
proc print data=chdiag;
run;
```

Alternatively, the log-linear model for [WMD][TWM][TWD][TMD] can be fitted as in Subsection 3.7.1. To fit other models such as {WM}{MD} or {WM}{D} using GENMOD, the model statement uses W*M M*D or W*M D, respectively.

3.6 Logit Models for a Multinomial Response

The basic method for dealing with a response variable (factor) with more than two levels is to arrange things so that only two things are compared at a time. One way of doing this is to identify pairs of levels to be compared. For example, if the response factor has R levels, comparing each level to the next level leads to modeling

$$\log(m_i/m_{i+1}), \qquad i = 1, \ldots, R-1\,, \tag{1}$$

or, equivalently,

$$\log(p_i/p_{i+1}), \qquad i = 1, \ldots, R-1\,.$$

These are the odds of getting level i relative to getting level $i+1$. They can be viewed as conditional odds given that either level i or $i+1$ occurs. To illustrate multinomial response models, consider the data of Exercise 3.7.2 presented in Table 3.1. The data involve factors of which we will treat abortion opinion as a response. The levels of abortion opinion are Yes, No, and Undecided. These indicate levels of support for legalized abortion. The model scheme indicated by equation (1) dictates looking at a series of odds: the odds of Yes to No and the odds of No to Undecided. In this case, the nominal levels of the response can be rearranged to suit us. For example, we could choose to look at the odds of No to Yes and of Yes to Undecided. These latter odds can be viewed as the conditional odds of No to Yes for people who have a clear opinion, and the odds of Yes to Undecided for people who are not opposed.

An alternative modeling scheme is for each level to be compared to a particular level; e.g., models can be formed for

$$\log(m_i/m_R), \qquad i = 1, \ldots, R-1\,. \tag{2}$$

With abortion opinions given in the order Yes, No, Undecided, these models involve the odds of Yes to Undecided and of No to Undecided. Again, one could (and in this case probably would) rearrange the order of the levels so that the level everything is compared to is a particularly interesting category.

If the same form model is used for each value of i, these methods are equivalent and both are equivalent to fitting a log-linear model. For example, the models

$$\log(m_{ijk}/m_{i+1\,jk}) = w_{2(j)} + w_{3(k)}, \qquad i = 1, \ldots, R-1,$$

and

$$\log(m_{ijk}/m_{Rjk}) = v_{2(j)} + v_{3(k)}, \qquad i = 1, \ldots, R-1,$$

are equivalent. (Note that the w and v parameters will also depend on i.) Both of these models are equivalent to

$$\log(m_{ijk}) = u_{23(jk)} + u_{12(ij)} + u_{13(ik)}\,.$$

Just as in two-category logit models, the interaction between all explanatory factors, $u_{23(jk)}$, is included in the model. The logit effects correspond in the log-linear model to interactions with the response factor. Given the log-linear model, the various logit models can be obtained by looking at differences. For example,

$$\log(m_{ijk}/m_{i+1\,jk}) = \log(m_{ijk}) - \log(m_{i+1\,jk})$$

and

$$\log(m_{ijk}/m_{Rik}) = \log(m_{ijk}) - \log(m_{Rjk})$$

lead to parametrizations such as

$$w_{2(j)} = u_{12(ij)} - u_{12(i+1\,j)}$$

and

$$v_{3(k)} = u_{13(ik)} - u_{13(Rk)}\,.$$

(Note that, as mentioned above, w_2 and v_3 depend on the category i that is being examined.) Fits for all of the models in (1) and (2) can be obtained by fitting one log-linear model.

Another way of reducing several response levels to binary comparisons is to pool response levels. One way to do this is to compare each level to the total of all other levels, e.g., model

$$\log\left(\frac{m_i}{\sum_{h\neq i} m_h}\right), \qquad i = 1, \ldots, R. \tag{3}$$

These are the odds of getting category i relative to not getting level i. Fitting these models requires fitting at least $R-1$ logit models. One log-linear model will not do. With the abortion opinion data, these are models

for the odds of Yes to not Yes, the odds of No to not No, and the odds of Undecided to not Undecided. These models focus on one category of response and ignore the structure of all other categories.

If the response levels have a natural ordering, say from smallest to largest, then it may be appropriate to look at *continuation ratios*

$$\log\left(\frac{m_i}{\sum_{h=i+1}^{R} m_h}\right), \qquad i = 1, \ldots, R-1\,. \tag{4}$$

These are the odds of getting level i relative to getting a category higher than level i. As always, we can rearrange the ordering of the response categories if it suits us. This method works very nicely for the abortion data, even though the response levels have no natural ordering. Think of the categories as being ordered as Undecided, Yes, No. Then the first model here has the odds of undecided to everything else, i.e., the odds of undecided to being decided. The second model has the odds of Yes to No, i.e., the odds of supporting legalized abortion relative to opposing it.

The odds in (4) are actually conditional odds. The probability of level i divided by the probability of a higher level is the odds of getting level i given that level i or higher is obtained. For example, the odds of Support to Oppose are actually conditional on being decided. As is seen in Exercise 3.8.14, fitting continuation ratio models for all i is equivalent to fitting a series of log-linear models.

Yet another possibility is to fit *cumulative logits*,

$$\log\left(\frac{\sum_{h=1}^{i} m_h}{\sum_{h=i+1}^{R} m_h}\right), \qquad i = 1, \ldots, R-1\,.$$

For abortion opinions ordered as Undecided, Yes, No, these models describe the odds of undecided to decided and the odds of not opposed to opposed.

EXAMPLE 3.6.1. We now examine fitting models to the data on race, sex, opinions on abortion, and age from Section 3.7. In a log-linear model, the variables are treated symmetrically. The analysis looks for relationships among any of the variables. Here, we consider opinions as a response variable. This changes the analysis in that [RSA] must be included in all models. Table 3.7 presents fits for all the models that include [RSA] and correspond to ANOVA type logit models.

The best fitting model is clearly [RSA][RSO][OA]. This model can be used directly to fit either the models in (1)

$$\begin{aligned} \log(m_{hi1k}/m_{hi2k}) &= w^1_{RS(hi)} + w^1_{A(R)}\,, \\ \log(m_{hi2k}/m_{hi3k}) &= w^2_{RS(hi)} + w^2_{A(k)}\,, \end{aligned}$$

the models in (2)

$$\log(m_{hi1k}/m_{hi3k}) = v^1_{RS(hi)} + v^1_{A(k)}$$

TABLE 3.7. Log-Linear Models for the Abortion Opinion Data

Model	df	G^2	$A-q$
[RSA][RSO][ROA][SOA]	10	6.12	−13.88
[RSA][RSO][ROA]	20	7.55	−32.45
[RSA][RSO][SOA]	20	13.29	−26.71
[RSA][ROA][SOA]	12	16.62	− 7.38
[RSA][RSO][OA]	30	14.43	−45.57
[RSA][ROA][SO]	22	17.79	−26.21
[RSA][SOA][RO]	22	23.09	−20.91
[RSA][RO][SO][OA]	32	24.39	−39.61
[RSA][RO][SO]	42	87.54	3.54
[RSA][RO][OA]	34	34.41	−33.59
[RSA][SO][OA]	34	39.63	−28.37
[RSA][RO]	44	97.06	9.06
[RSA][SO]	44	101.9	13.9
[RSA][OA]	36	49.37	−22.63
[RSA][O]	46	111.1	19.1

$$\log(m_{hi2k}/m_{hi3k}) \quad = \quad v^2_{RS(hi)} + v^2_{A(k)}\,,$$

or some variation of these. As discussed earlier, the first pair of models looks at the odds of supporting legalized abortion to opposing legalized abortion and the odds of opposing legalized abortion to being undecided. The second pair of models examines the odds of supporting legalized abortion to undecided and the odds of opposing to undecided. Of these, the only odds that seem particularly interesting to the author are the odds of supporting to opposing. In the second pair of models, choosing the category "undecided" as the standard level to which other levels are compared is particularly unintuitive. The fact that undecided is the last category is no reason for it to be chosen as the standard of comparison. Either of the other categories would be a better standard, so one of these should be used. Neither is obviously better than the other.

Neither of these pairs of models are particularly appealing, so we will only continue the analysis long enough to illustrate salient points and to allow comparisons with other models to be discussed later. The fitted values for [RSA][RSO][OA] are given in Table 3.8.

We consider only the odds of support relative to opposed. The odds can be obtained from the fitted values. For example, the odds for young white males are $100.1/39.73 = 2.52$. The full table of odds is given in Table 3.9.

Note that the values from age to age vary by a constant multiple depending on the ages involved. The odds of support decrease steadily with age. The model has no inherent structure among the four race-sex categories; however, the odds for white males and nonwhite males are surprisingly similar. Nonwhite females are most likely to support legalized abortion, white females are next, and males are least likely to support legalized abortion.

TABLE 3.8. Fitted Values for [RSA][RSO][OA]

			Age					
Race	Sex	Opinion	18-25	26-35	36-45	46-55	56-65	65+
White	Male	Support	100.1	137.2	117.5	75.62	70.58	80.10
		Oppose	39.73	64.23	56.17	47.33	50.99	62.55
		Undec.	1.21	2.59	5.36	5.05	5.43	8.35
	Female	Support	138.4	172.0	152.4	101.8	101.7	110.7
		Oppose	43.49	63.77	57.68	50.44	58.19	68.43
		Undec.	2.16	4.18	8.96	8.76	10.08	14.86
Nonwhite	Male	Support	21.19	16.57	15.20	11.20	8.04	7.80
		Oppose	8.54	7.88	7.38	7.11	5.90	6.18
		Undec.	1.27	1.54	3.42	3.69	3.06	4.02
	Female	Support	21.40	26.20	19.98	16.38	13.64	12.40
		Oppose	4.24	6.12	4.77	5.12	4.92	4.83
		Undec.	0.36	0.68	1.25	1.50	1.44	1.77

TABLE 3.9. Estimated Odds of Support versus Oppose

Legalized Abortion
(Based on the log-linear model [RSA][RSO][OA])

		Age					
Race	Sex	18-25	26-35	36-45	46-55	56-65	65+
White	Male	2.52	2.14	2.09	1.60	1.38	1.28
	Female	3.18	2.70	2.64	2.02	1.75	1.62
Nonwhite	Male	2.48	2.10	2.06	1.57	1.36	1.26
	Female	5.05	4.28	4.19	3.20	2.77	2.57

Confidence intervals for log odds or log odds ratios can be found using the methods of Section 10.2 or, alternatively, the methods of Section 11.1.

If we pool categories, we can look at the set of three models generated by (3) or the set of two models generated by (4). The set of three models consists of the odds of supporting, the odds of opposing, and the odds of undecided (in each case, the odds are defined relative to the union of the other categories). The two models from (4) are essentially continuation ratio models. The most interesting definition of these models is obtained by taking the odds of supporting to opposing and the odds of undecided to having an opinion. Fitting the models involves fitting log-linear models to two sets of data.

Eliminating all undecideds from the data, we fit [RSA][RSO][OA] to the $2 \times 2 \times 2 \times 6$ table with only the opinion categories "support" and "oppose." The estimated expected cell counts are given in Table 3.10. Note that the

estimated cell counts are very similar to those obtained when undecideds were included in the data. The odds of supporting relative to opposing are given below.

Odds of Support versus Opposed

		Age					
Race	Sex	18-25	26-35	36-45	46-55	56-65	65+
White	Male	2.52	2.14	2.09	1.60	1.38	1.28
	Female	3.18	2.70	2.64	2.01	1.75	1.62
Nonwhite	Male	2.48	2.11	2.06	1.57	1.36	1.26
	Female	5.08	4.31	4.22	3.22	2.79	2.58

Except for nonwhite females, the odds of support are essentially identical to those obtained with undecideds included. The G^2 for the fit without undecideds is 9.104 with 15 df. The G^2 for fitting [RSA][RO][SO][OA] is 11.77 on 16 df. The difference in G^2's is not large, so a logit model $\log(m_{hi1k}/m_{hi2k}) = R_{(h)} + S_{(i)} + A_{(k)}$ may fit adequately.

TABLE 3.10. Estimated Expected Cell Counts with Undecideds Eliminated

			Age					
Race	Sex	Opinion	18-25	26-35	36-45	46-55	56-65	65+
White	Male	Support	100.2	137.7	117.0	75.62	70.22	80.27
		Oppose	39.78	64.35	55.98	47.38	50.78	62.73
	Female	Support	139.2	172.2	152.3	101.6	101.7	109.9
		Oppose	43.78	63.77	57.71	50.41	58.28	68.05
Nonwhite	Male	Support	20.67	16.96	15.48	11.00	8.07	7.81
		Oppose	8.33	8.04	7.52	7.00	5.93	6.19
	Female	Support	20.84	25.17	20.21	16.78	13.98	12.97
		Oppose	4.11	5.84	4.79	5.22	5.02	5.03

We now pool the support and oppose categories to get a $2 \times 2 \times 2 \times 6$ table in which the opinions are "support or oppose" and "undecided." Again, the model [RSA][RSO][OA] is fitted to the data. For this model, we report only the estimated odds.

Odds of Being Decided on Abortion

		Age					
Race	Sex	18-25	26-35	36-45	46-55	56-65	65+
White	Male	116.79	78.52	32.67	24.34	22.26	16.95
	Female	83.43	56.08	23.34	17.38	15.90	12.11
Nonwhite	Male	23.76	15.97	6.65	4.95	4.53	3.45
	Female	68.82	46.26	19.25	14.34	13.12	9.99

Again, the estimated odds vary from age to age by a constant multiple. The odds decrease with age, so older people are less likely to take a position. White males are most likely to state a position. Nonwhite males are least likely to state a position. White and nonwhite females have odds of being decided that are somewhat similar.

The G^2 for [RSA][RSO][OA] is 5.176 on 15 *df*. The G^2 for the smaller model [RSA][RO][SO][OA] is 12.71 on 16 *df*. The difference is very large. Although a main-effects-only logit model fits the support-opposition table quite well, to deal with the undecided category requires a race-sex interaction.

We have pretty much exhausted what can be done easily by fitting ANOVA type log-linear models using iterative proportional fitting. However, computer programs are readily available for direct fitting of logit models. We illustrate some results for modeling the odds of support relative to opposition with undecideds eliminated from the data.

The model that we considered in detail was [RSA][RSO][OA]. This is equivalent to

$$\log(m_{hi1k}/m_{hi2k}) = (RS)_{hi} + A_k \tag{5}$$

which models the odds of supporting legalized abortion. To fit this model directly, we need to provide a computer program with the counts for all cells indicating support, i.e.,

$$n_{hi1k}, \qquad h = 1,2,\ i = 1,2,\ k = 1,\ldots,6\,,$$

and the total of support and opposition for all cells

$$N_{hik} = n_{hi\cdot k} = n_{hi1k} + n_{hi2k}\,.$$

For example, $n_{1111} = 96$, $n_{1211} = 140$, $n_{2111} = 24$, $n_{11\cdot 1} = 96 + 44 = 140$, $n_{12\cdot 1} = 140 + 43 = 183$, and $n_{21\cdot 1} = 24 + 5 = 29$. In addition, for each count and total, we need to provide the program with the corresponding indices h, i, and k. Fitting model (5) directly gives $G^2 = 9.104$ on 15 *df*, exactly the results from fitting the equivalent model [RSA][RSO][OA].

The table of odds has suggested two things: (1) odds decrease as age increases and (2) the odds for males are about the same. We want to fit models that incorporate these suggestions. Of course, because the data are suggesting the models, formal tests of significance will be even less appropriate than usual, but G^2's still give a reasonable measure of the quality of model fit.

We model the fact that odds are decreasing with age by incorporating a linear trend in ages. We do not have specific ages to associate with the age categories, so we simply use the codes $k = 1,2,\ldots,6$ to indicate ages. These scores lead to fitting the model

$$\log(m_{hi1k}/m_{hi2k}) = (RS)_{hi} + \gamma k\,. \tag{6}$$

The G^2 is 10.18 on 19 *df*, so the linear trend in coded ages fits very well. [Recall that model (5) has $G^2 = 9.104$ on 15 *df*, so a test of model (6) versus model (5) has $G^2 = 10.18 - 9.104 = 1.08$ on $19 - 15 = 4$ *df*.]

To incorporate the idea that males have the same odds of support, we recode the indices of the data. Recall that to fit model (5), we had to specify three index variables along with the numbers supporting and the totals. The indices for the $(RS)_{hi}$ terms are $(h, i) = (1, 1), (1, 2), (2, 1), (2, 2)$. We could recode the problem with an index, say $g = 1, 2, 3, 4$, and fit the model

$$\log(m_{g1k}/m_{g2k}) = (RS)_g + A_k$$

and get exactly the same fit. We can choose the recoding as

(h, i)	(1,1)	(1,2)	(2,1)	(2,2)
g	1	2	3	4

Note that, together, the subscripts g and k still distinguish all of the cases for which data are provided.

This recoding can now be modified, so models that treat males the same can be specified. If we want to treat males the same, then the codes for white males $g = 1$ and nonwhite males $g = 3$ must be made the same. On the other hand, we still have distinct data for white males and nonwhite males, so the fact that there are two replications on males must be accounted for. To treat males the same, recode g as (f, e) with

	wm	wf	nm	nf
g	1	2	3	4
f	1	2	1	3
e	1	1	2	1

where e is an index for replications and the codes wm, wf, nm, nf indicate white males, white females, nonwhite males, and nonwhite females, respectively. Now fit the model

$$\log(m_{fe1k}/m_{fe2k}) = (RS)_f + A_k\,. \tag{7}$$

The two male groups are only distinguished by the subscript e, and e does not appear on the right-hand side of the model, so the two male groups will be modeled identically. In fact, to use a logistic regression program, you typically do not even need to define the index e. But whether you define it or not, it exists implicitly in the model.

Model (7) is, of course, a reduced model relative to model (5). Model (7) has $G^2 = 9.110$ on 16 *df*, so the comparison between models has $G^2 = 9.110 - 9.104 = .006$ on $16 - 15 = 1$ *df*. We have lost almost nothing by going from model (5) to model (7).

Finally, we can write a model that incorporates both the trend in ages and the equality for males

$$\log(m_{fe1k}/m_{fe2k}) = (RS)_f + \gamma k\,. \tag{8}$$

This has $G^2 = 10.19$ on 20 *df*. Thus, relative to model (5), we have dropped 5 *df* from the model, yet only increased the G^2 by $10.19 - 9.10 = 1.09$.

For the alternative parametrization,

$$\log(m_{fe1k}/m_{fe2k}) = \mu + (RS)_f + \gamma k\,,$$

the estimates and standard errors using the side condition $(RS)_1 = 0$ are

Parameter	Estimate	SE	Est./SE
μ	1.071	.1126	9.51
$(RS)_1$	0	—	—
$(RS)_2$	.2344	.09265	2.53
$(RS)_3$	.6998	.2166	3.23
γ	−.1410	.02674	−5.27

All of the terms seem important. With this side condition, $(\widehat{RS})_2$ is actually an estimate of $(RS)_2 - (RS)_1$, so the z score 2.53 is an indication that white females have an effect on the odds of support that is different from males. Similarly, $(\widehat{RS})_3$ is an estimate of the difference in effect of nonwhite females and males. The estimated odds of support are

	Age					
Race-Sex	18-25	26-35	36-45	46-55	56-65	65+
Male	2.535	2.201	1.912	1.661	1.442	1.253
White female	3.204	2.783	2.417	2.099	1.823	1.583
Nonwhite female	5.103	4.432	3.850	3.343	2.904	2.522

These show the general characteristics discussed earlier. Also, they can be transformed into (conditional) probabilities of support. Probabilities are generally easier to interpret than odds. The estimated probability that a white female between 46 and 55 years of age supports legalized abortion is $2.099/(1 + 2.099) = .677$. The odds are about 2, so the probability is about twice as great that such a person will support legalized abortion rather than oppose it.

Similar ideas of modeling can be applied to the odds of having made a decision on legalized abortion.

Finally, a word about computing. The computations for models (6), (7), and (8) were executed using a computer program specifically designed for logit models. This was done because computer programs based on iterative proportional fitting cannot handle the corresponding log-linear models. Iterative proportional fitting only works for ANOVA type models. However, programs for fitting general log-linear models (e.g., GLIM) can handle the

log-linear models that correspond to (6), (7), and (8). The models are found in the usual way. Model (6) corresponds to

$$\log(m_{hijk}) = (RSA)_{hik} + (RSO)_{hij} + \gamma_j k$$

where we have added the highest-order interaction term not involving O and made the (RS) and γ terms depend on the opinion level j. Similarly, models (7) and (8) correspond to

$$\log(m_{fejk}) = (RSA)_{fek} + (RSO)_{fj} + (OA)_{jk}$$

and

$$\log(m_{fejk}) = (RSA)_{fek} + (RSO)_{fj} + \gamma_j k\,,$$

respectively.

In a somewhat different approach to treating response factors, Asmussen and Edwards (1983) allow the fitting of models that do not always include a term for the interactions among the explanatory factors. Instead, they argue that *log-linear models are appropriate for response factors as long as the model allows for collapsing over the response factors onto the explanatory factors*, cf. Section 5.3. These issues will also be discussed at the end of Section 6.8.

3.7 Logistic Discrimination and Allocation

How can you tell Swedes and Italians apart? How can you tell different species of irises apart? How can you identify people who are likely to have a heart attack or commit a crime? One approach is to collect data on individuals who are known to be in each of the populations of interest. The data can then be used to discriminate between the populations. To identify Swedes and Italians, one might collect data on height, hair color, eye color, and skin complexion. To identify irises, one might measure petal length and width and sepal length and width. Typically, data collected on several different variables are combined to identify the likelihood that someone belongs to a particular population. In a standard discrimination–allocation problem, independent samples are taken from each population. The use of these samples to characterize the populations is referred to as discrimination. Allocation involves identifying the population of an individual for whom only the variable values are known. The factor of interest in these problems is the population, but it is not a response factor in the sense used elsewhere in this chapter. In particular, *discrimination data arises from conducting a retrospective study.* The reader may want to review the subsection of the chapter introduction that discusses retrospective and prospective studies.

There has been extensive work done on the problems of discrimination and allocation. Introductions to the subject are contained in Anderson (1984), Christensen (1990), Hand (1981), Lachenbruch (1975), McLachlan (1992), Press (1984), and Rao (1973). The review article by Cheng and Titterington (1994) relates discriminant analysis to neural networks. Recent work on logistic discrimination includes Cox and Ferry (1991) and O'Neill (1994). Probably, the two most commonly used methods of discrimination are Fisher's linear discriminant function and logistic regression. Fisher's method is based on the idea that each case corresponds to a fixed population and that the variables for each case are observations from a multivariate normal distribution. The normal distributions for the populations are assumed to have different means but the same variances and covariances. The logistic regression approach (or as presented here, the log-linear model approach) treats the distribution for each population as a multinomial. Much of the theoretical work on discriminant analysis is done in a Bayesian setting and both methods lend themselves to the easy computation of posterior probabilities for a case to be in a particular population.

The weakness of Fisher's method is that the assumption of normality with equal covariances is often patently false. The case variables are often percentages, rates, or categorical variables. Even when the case variables are continuous on the entire real line, they are often obviously skewed. Frequently the variance-covariance matrices in the various populations are not even similar, much less identical. Fortunately, Fisher's method is somewhat insensitive (robust) to many of these difficulties, cf. Lachenbruch, Sneeringeer, and Revo (1973) and Press and Wilson (1978). Fisher's method is also easily generalized to handle unequal covariance matrices. The strength of Fisher's method is that for normal data it is more efficient than logistic discrimination, cf. Efron (1975).

EXAMPLE 3.7.1. Aitchison and Dunsmore (1975, p. 212) consider 21 individuals with 1 of 3 types of Cushing's syndrome. Cushing's syndrome is a medical problem associated with overproduction of cortisol by the adrenal cortex. The three types considered are related to specific problems with the adrenal gland, namely

A–adenoma
B–bilateral hyperplasia
C–carcinoma

The case variables considered are the rates at which two steroid metabolites are excreted in the urine. (These are measured in milligrams per day.) The two steroids are

TETRA – Tetrahydrocortisone

and

PREG – Pregnanetriol.

The data are listed in Table 3.11.

TABLE 3.11. Cushing's Syndrome Data

Case	Type	TETRA	PREG	Case	Type	TETRA	PREG
1	A	3.1	11.70	12	B	15.4	3.60
2	A	3.0	1.30	13	B	7.7	1.60
3	A	1.9	0.10	14	B	6.5	0.40
4	A	3.8	0.04	15	B	5.7	0.40
5	A	4.1	1.10	16	B	13.6	1.60
6	A	1.9	0.40	17	C	10.2	6.40
7	B	8.3	1.00	18	C	9.2	7.90
8	B	3.8	0.20	19	C	9.6	3.10
9	B	3.9	0.60	20	C	53.8	2.50
10	B	7.8	1.20	21	C	15.8	7.60
11	B	9.1	0.60				

The data determine the 3×21 table

	Case														
Type	1	2	3	4	5	6	7	8	$\cdots$	16	17	18	19	20	21
A	1	1	1	1	1	1	0	0	$\cdots$	0	0	0	0	0	0
B	0	0	0	0	0	0	1	1	$\cdots$	1	0	0	0	0	0
C	0	0	0	0	0	0	0	0	$\cdots$	0	1	1	1	1	1

The case variables TETRA and PREG are used to model the interaction in this table. The case variables are highly skewed, so, following Aitchison and Dunsmore, we analyze the transformed variables $TL \equiv \log(\text{TETRA})$ and $PL \equiv \log(\text{PREG})$. The transformed data are plotted in Figure 3.2.

Now consider the sampling scheme. For studies of this type, it is best modeled as involving independent samples from the three populations: A, B, and C. The sampling can be viewed as product-multinomial because all observations are intrinsically discrete. The categories for the product-multinomials consist of all the *observable* combinations of TL and PL. Although TL and PL are apparently continuous variables, the observations taken on TETRA and PREG are only known to be within $\pm.05$ mg and $\pm.005$ mg of their respective nominal values. Thus, the observations are discrete and product-multinomial sampling is appropriate. (Note that the PREG value for case 4 may be a typographical error.) The catch is that there are a huge number of possible categories. Most of these categories have no observations associated with them. If there are S observable combinations of the explanatory factors, we would like to perform a product-multinomial likelihood analysis of the $3 \times S$ table.

Unfortunately, the standard log-linear models for multinomial responses do not have maximum likelihood estimates because most of the column totals in the $3 \times S$ table are zero. To see that MLEs do not exist, observe

FIGURE 3.2. Cushing's Syndrome Data

that if they do exist, the fitted column totals must equal the observed column totals. Most of the S columns in the table will be unobserved, so most of the column totals will be zero. MLEs for log-linear models must be positive so that their logs can be taken; hence, the fitted column totals must all be positive. The fitted column totals cannot be both positive and zero.

In practice, the analysis is conducted as if the observed 3×21 table is obtained via product-multinomial sampling of the three populations. This works well in spite of the fact that the sampling scheme is palpably false. The 21 cases are included in the table precisely because they were observed. Thus, each column total *must* be at least one. If the sampling scheme were truly product-multinomial, there would be a positive probability of getting column totals equal to zero. Section 11.4 contains a more detailed discussion of these issues and a justification for treating the 3×21 table as product-multinomial. In the current section, we simply present the standard methodology.

One of the tricky things about this is that it *looks like* logistic regression, except that we have more than two possibilities for the response. But treating this as a logistic regression is wrong. In a logistic regression, there are cases with predictor variables associated with them, and each case randomly and independently falls into a response category; e.g., have a coronary incident or don't. In a logistic regression, when the responses are 0s and 1s, every case is a sample from a different population. But in this logistic discrimination, there are only three populations being sampled.

The sample sizes are larger, and the values of the predictor variables are actually the results of the sampling.

Because of the sampling scheme, when the samples from the various populations are of different sizes, the values m_{ij} are not directly useful in evaluating the relationship between populations and the predictor variables. For example, if we choose to sample 20,000 people from population A and only 10 from population B, the m_{1j}'s are not comparable to the m_{2j}'s. We must adjust for sample size before relating syndrome type to TL and PL. The evaluation of the relationship is based on the relative likelihoods of the three syndrome types. Thus, for any case j, our interest is in the relative sizes of p_{1j}, p_{2j}, and p_{3j}. Estimates of these quantities are easily obtained from the $\hat{m}_{ij}$'s. Simply take

$$\hat{p}_{ij} = \hat{m}_{ij}/n_{i\cdot}\,. \tag{1}$$

For a new patient of unknown syndrome type but whose values of TL and PL place him in category j, the most likely type of Cushing's syndrome is that which has the largest value among p_{1j}, p_{2j}, and p_{3j}. Clearly, we can estimate the most likely syndrome type. In practice, new patients are unlikely to fall into one of the 21 previously observed categories but the modeling procedure is flexible enough to allow allocation of individuals having any values of TL and PL. This will be discussed in detail in the subsection on allocation.

Discrimination

For each individual j, the variables $(TL)_j$ and $(PL)_j$ have been observed. We seek a model that can be used to classify observations into syndrome type. The main effects model is

$$\log(m_{ij}) = \alpha_i + \beta_j\,, \qquad i = 1,2,3\,, \qquad j = 1,\ldots,21\,.$$

We want to use TL and PL to help model the interaction, so fit

$$\log(m_{ij}) = \alpha_i + \beta_j + \gamma_{1i}(TL)_j + \gamma_{2i}(PL)_j\,, \tag{2}$$

$i = 1,2,3$, $j = 1,\ldots,21$.

This model is very similar to a log-linear version of the logit and logistic models discussed earlier. In particular, it has a separate term β_j for every combination of the explanatory variables. Taking differences gives, for example,

$$\log(m_{1j}/m_{2j}) = (\alpha_1 - \alpha_2) + (\gamma_{11} - \gamma_{12})(TL)_j + (\gamma_{21} - \gamma_{22})(PL)_j$$

which can be written as

$$\log(m_{1j}/m_{2j}) = \alpha + \delta_1(TL)_j + \delta_2(PL)_j\,.$$

Although this looks like a logistic regression model, it has a fundamentally different interpretation. Unlike logistic regression models, it is typically the case that

$$\log\left(\frac{m_{1j}}{m_{2j}}\right) \neq \log\left(\frac{p_{1j}}{p_{2j}}\right).$$

Moreover, the ratio p_{1j}/p_{2j} is not even an odds of type A relative to type B. Both numbers are probabilities, but they are probabilities from different populations. The correct interpretation of p_{1j}/p_{2j} is as a likelihood ratio, specifically the likelihood of type A relative to type B. A value p_{ij} is the likelihood within population i of observing category j. Having fitted model (2), the estimate of the log of the likelihood ratio is

$$\log\left(\frac{\hat{p}_{1j}}{\hat{p}_{2j}}\right) = \log\left(\frac{\hat{m}_{1j}/n_{1\cdot}}{\hat{m}_{2j}/n_{2\cdot}}\right) = \log\left(\frac{\hat{m}_{1j}}{\hat{m}_{2j}}\right) - \log\left(\frac{n_{1\cdot}}{n_{2\cdot}}\right).$$

It will be seen in Chapter 11 that, because interest is directed at comparing probabilities in *different* multinomials, asymptotic variances of estimates will be more complicated than for logistic regression.

Finally, it should be noted that *although odds depend on the sampling scheme, odds ratios do not.* Odds ratios are handled in exactly the same way regardless of whether the sampling scheme is prospective or retrospective.

The G^2 for model (2) is 12.30 on 36 degrees of freedom. As in Section 2.6, although G^2 is a valid measure of goodness of fit, G^2 cannot legitimately be compared to a χ^2 distribution. However, we can test reduced models. The model

$$\log(m_{ij}) = \alpha_i + \beta_j + \gamma_{1i}(TL)_j$$

has $G^2 = 21.34$ on 38 degrees of freedom and

$$\log(m_{ij}) = \alpha_i + \beta_j + \gamma_{2i}(PL)_j$$

has $G^2 = 37.23$ on 38 degrees of freedom. Neither of the reduced models provides an adequate fit. (Recall that χ^2 tests of model comparisons like these were valid.)

Table 3.12 contains estimated probabilities for the three populations. The probabilities are computed using equation (1) and model (2).

Table 3.13 illustrates a Bayesian analysis. For each case j, it gives the estimated posterior probability that the case belongs to each of the three syndrome types. The data consist of the observed TL and PL values in category j. Given that the syndrome type is i, the estimated probability of observing data in category j is $\hat{p}_{ij}$. Let $\pi(i)$ be the prior probability that the case is of syndrome type i. Bayes theorem gives

$$\hat{\pi}(i|Data) = \frac{\hat{p}_{ij}\pi(i)}{\sum_{i=1}^{3}\hat{p}_{ij}\pi(i)}.$$

Two choices of prior probabilities are used in Table 3.13: probabilities proportional to sample sizes, i.e., $\pi(i) = n_{i\cdot}/n_{\cdot\cdot}$ and equal probabilities

TABLE 3.12. Estimated Probabilities

	Group				Group		
Case	A	B	C	Case	A	B	C
1	.1485	.0012	.0195	12	.0000	.0295	.1411
2	.1644	.0014	.0000	13	.0000	.0966	.0068
3	.1667	.0000	.0000	14	.0001	.0999	.0000
4	.0842	.0495	.0000	15	.0009	.0995	.0000
5	.0722	.0565	.0003	16	.0000	.0907	.0185
6	.1667	.0000	.0000	17	.0000	.0102	.1797
7	.0000	.0993	.0015	18	.0000	.0060	.1879
8	.1003	.0398	.0000	19	.0000	.0634	.0733
9	.0960	.0424	.0000	20	.0000	.0131	.1738
10	.0000	.0987	.0025	21	.0000	.0026	.1948
11	.0000	.0999	.0003				

$\pi(i) = \frac{1}{3}$. Prior probabilities proportional to sample sizes are *rarely appropriate*, but they relate in simple ways to standard output, so we give them more prominence than they probably deserve. Both of the sets of posterior probabilities are easily obtained. The table of proportional probabilities is just the table of $\hat{m}_{ij}$ values. This follows from two facts: first, $\hat{m}_{ij} = n_{i\cdot}\hat{p}_{ij}$ and second, the model fixes the column totals, so $\hat{m}_{\cdot j} = 1 = n_{\cdot j}$. To obtain the equal probabilities values, simply divide the entries in Table 3.12 by the sum of the three probabilities for each case. Cases that are misclassified by either procedure are indicated with a double asterisk in Table 3.13.

Table 3.14 summarizes the classifications. With proportional prior probabilities, 16 of 21 cases are correctly allocated. With equal prior probabilities, 18 of 21 cases are correctly allocated. While Table 3.14 is useful, it ignores the clarity of the allocations. For example, case 4 with proportional probabilities is essentially a toss-up between types A and B. That information is lost in Table 3.14. (The probability of type A is slightly greater than one-half.) Another problem with Table 3.14 is that it tends to overestimate how well the discrimination would work on other data. The data were used to form a discrimination procedure and Table 3.14 evaluates how well it works by allocating the same data. This double dipping tends to make the discrimination procedure look better than it really is. Cross-validation can be used to reduce the bias introduced; for related work, see Geisser (1977) and Gong (1986). Finally, it is of interest to note that the difference in Table 3.14 between proportional probabilities and equal probabilities is that under proportional probabilities, one additional case in each of A and C is misclassified into B. That occurs because the prior probability for B is about twice as great as the values for A and C.

Readers who are familiar with normal theory discrimination may be interested in the analysis of these data contained in Christensen (1990). Taking the logs of tetrahydrocortisone and pregnanetriol is important in using Fisher's linear discrimination because the original data are clearly non-

TABLE 3.13. Probabilities of Classification

			Proportional Prior Probabilities			Equal Prior Probabilities		
Case		Group	*A*	*B*	*C*	*A*	*B*	*C*
1		*A*	.89	.01	.10	.88	.01	.12
2		*A*	.99	.01	.00	.99	.01	.00
3		*A*	1.00	.00	.00	1.00	.00	.00
4		*A*	.50	.50	.00	.63	.37	.00
5	**	*A*	.43	.57	.00	.56	.44	.00
6		*A*	1.00	.00	.00	1.00	.00	.00
7		*B*	.00	.99	.01	.00	.99	.01
8	**	*B*	.60	.40	.00	.72	.28	.00
9	**	*B*	.58	.42	.00	.69	.31	.00
10		*B*	.00	.99	.01	.00	.97	.03
11		*B*	.00	1.00	.00	.00	1.00	.00
12	**	*B*	.00	.29	.71	.00	.17	.83
13		*B*	.00	.97	.03	.00	.93	.07
14		*B*	.00	1.00	.00	.00	1.00	.00
15		*B*	.01	.99	.00	.01	.99	.00
16		*B*	.00	.91	.09	.00	.83	.17
17		*C*	.00	.10	.90	.00	.05	.95
18		*C*	.00	.06	.94	.00	.03	.97
19	**	*C*	.00	.63	.37	.00	.46	.54
20		*C*	.00	.13	.87	.00	.07	.93
21		*C*	.00	.03	.97	.00	.01	.99

TABLE 3.14. Summary of Classifications

	Proportional Prior Probabilities True Group			Equal Prior Probabilities True Group		
Allocated to Group	*A*	*B*	*C*	*A*	*B*	*C*
A	5	2	0	6	2	0
B	1	7	1	0	7	0
C	0	1	4	0	1	5

normal. Logistic discrimination imposes no such normality requirement. Without the log transform, Fisher's method misclassifies seven observations including five of the six in type *A*. Logistic discrimination on the untransformed data with proportional priors only misclassifies four observations and gets five of six correct in type *A*.

Allocation

If you stop and think about it, discrimination seems like a remarkably silly thing to do. Why take cases from known populations and reclassify them when the process of reclassification introduces errors? The reason discrim-

ination is interesting is because one can use a model that discriminates between cases from known populations to predict the population of an unknown case. In our example, a new patient can be measured for TL and PL, and then diagnosed as to type of Cushing's syndrome without direct examination of the adrenal cortex. (I have no idea if this is an accurate description of medical practice, but it illustrates the kind of thing that can be done.) We now consider the problem of allocating new cases to the populations.

Model (2) includes a separate term β_j for each case, so it is not clear how model (2) can be used to allocate future cases. We will begin with logit models and then work back to an allocation model. Model (2) has 30 parameters, only 9 of which are really of interest. Of these nine, only six are estimable. From (2), we can model the probability ratio of type A relative to type B

$$\begin{aligned} &\log(p_{1j}/p_{2j}) \\ &\quad = \log(m_{1j}/m_{2j}) - \log(n_{1\cdot}/n_{2\cdot}) \qquad (3) \\ &\quad = (\alpha_1 - \alpha_2) + (\gamma_{11} - \gamma_{12})(TL)_j + (\gamma_{21} - \gamma_{22})(PL)_j - \log(n_{1\cdot}/n_{2\cdot})\,. \end{aligned}$$

The log-likelihoods of A relative to C are

$$\begin{aligned} &\log(p_{1j}/p_{3j}) \\ &\quad = \log(m_{1j}/m_{3j}) - \log(n_{1\cdot}/n_{3\cdot}) \qquad (4) \\ &\quad = (\alpha_1 - \alpha_3) + (\gamma_{11} - \gamma_{13})(TL)_j + (\gamma_{21} - \gamma_{23})(PL)_j - \log(n_{1\cdot}/n_{3\cdot})\,. \end{aligned}$$

Fitting model (2) gives the estimated parameters.

Par.	Est.	Par.	Est.	Par.	Est.
α_1	0.0	γ_{11}	-16.29	γ_{21}	-3.359
α_2	-20.06	γ_{12}	-1.865	γ_{22}	-3.604
α_3	-28.91	γ_{13}	0.0	γ_{23}	0.0

where the estimates with values of 0 are really side conditions imposed on the collection of estimates to make it unique.

For a new case with values TL and PL, we plug estimates into equations (3) and (4) to get

$$\log(\hat{p}_1/\hat{p}_2) = 20.06 + (-16.29 + 1.865)TL + (-3.359 + 3.604)PL - \log(6/10)$$

and

$$\log(\hat{p}_1/\hat{p}_3) = 28.91 - 16.29(TL) - 3.359(PL) - \log(6/5)\,.$$

For example, if the new case has a tetrahydrocortisone reading of 4.1 and a pregnanetriol reading of 1.10, then $\log(\hat{p}_1/\hat{p}_2) = .24069$ and $\log(\hat{p}_1/\hat{p}_3) = 5.4226$. The likelihood ratios are

$$\hat{p}_1/\hat{p}_2 \quad = \quad 1.2721$$

$$\hat{p}_1/\hat{p}_3 = 226.45$$

and by division,

$$\hat{p}_2/\hat{p}_3 = 226.45/1.2721 = 178.01\,.$$

It follows that type A is a little more likely than type B and that both are much more likely than type C

One can also obtain estimated posterior probabilities for a new case. The posterior odds are

$$\frac{\hat{\pi}(1|Data)}{\hat{\pi}(2|Data)} = \frac{\hat{p}_1}{\hat{p}_2}\frac{\pi(1)}{\pi(2)} \equiv \hat{O}_2$$

and

$$\frac{\hat{\pi}(1|Data)}{\hat{\pi}(3|Data)} = \frac{\hat{p}_1}{\hat{p}_3}\frac{\pi(1)}{\pi(3)} \equiv \hat{O}_3\,.$$

Using the fact that $\hat{\pi}(1|Data) + \hat{\pi}(2|Data) + \hat{\pi}(3|Data) = 1$, we can solve for $\hat{\pi}(i|Data)$, $i = 1, 2, 3$. In particular,

$$\begin{aligned}
\hat{\pi}(1|Data) &= \left[1 + \frac{1}{\hat{O}_2} + \frac{1}{\hat{O}_3}\right]^{-1} = \frac{\hat{O}_2\hat{O}_3}{\hat{O}_2\hat{O}_3 + \hat{O}_3 + \hat{O}_2}\,,\\
\hat{\pi}(2|Data) &= \frac{1}{\hat{O}_2}\left[1 + \frac{1}{\hat{O}_2} + \frac{1}{\hat{O}_3}\right]^{-1} = \frac{\hat{O}_3}{\hat{O}_2\hat{O}_3 + \hat{O}_3 + \hat{O}_2}\,,\\
\hat{\pi}(3|Data) &= \frac{1}{\hat{O}_3}\left[1 + \frac{1}{\hat{O}_2} + \frac{1}{\hat{O}_3}\right]^{-1} = \frac{\hat{O}_2}{\hat{O}_2\hat{O}_3 + \hat{O}_3 + \hat{O}_2}\,.
\end{aligned}$$

Using TETRA = 4.10 and PREG = 1.10, the assumption $\pi(i) = n_{i\cdot}/n_{\cdot\cdot}$ and more numerical accuracy in the parameter estimates than was reported earlier,

$$\begin{aligned}
\hat{\pi}(1|Data) &= .433\\
\hat{\pi}(2|Data) &= .565\\
\hat{\pi}(3|Data) &= .002\,.
\end{aligned}$$

Assuming $\pi(i) = 1/3$ gives

$$\begin{aligned}
\hat{\pi}(1|Data) &= .560\\
\hat{\pi}(2|Data) &= .438\\
\hat{\pi}(3|Data) &= .002\,.
\end{aligned}$$

Note that the values of tetrahydrocortisone and pregnanetriol used are identical to those for case 5; thus, the $\hat{\pi}(i|Data)$'s are identical to those listed in Table 3.13 for case 5.

To use the log-linear model approach illustrated here, one needs to fit a 3×21 table. Typically, a data file of 63 entries is needed. Three rows of the data file are associated with each of the 21 cases. Each data entry has to be identified by case and by type. In addition, the case variables should be

included in the file in such a way that all three rows for a case include the corresponding case variables, TL and PL. Model (2) is easily fitted using GLIM.

It is easy to just fit log-linear or logistic models to data such as that in Table 3.11 and get $\hat{m}_{ij}$'s or $\hat{p}_{ij}$'s. If you treat these values as estimated probabilities for being in the various populations, you are doing a Bayesian analysis with prior probabilities proportional to sample sizes. This is rarely an appropriate methodology.

3.8 Exercises

EXERCISE 3.8.1. The auto accident data of Example 3.2.4 was actually a subset of a four-dimensional table. The complete data are given in Table 3.15. Analyze the data treating severity of injury as a response variable. What conclusions can you reach from examining the $\hat{m}_{hijk}$'s, the odds, and odds ratios?

TABLE 3.15. Automobile Accident Data

Small Cars ($h = 1$)

		Accident Type (k)			
		Collision		Rollover	
Injury (j)		Not Severe	Severe	Not Severe	Severe
Driver	No	350	150	60	112
Ejected (i)	Yes	26	23	19	80

Standard Cars ($h = 2$)

		Accident Type (k)			
		Collision		Rollover	
Injury (j)		Not Severe	Severe	Not Severe	Severe
Driver	No	1878	1022	148	404
Ejected (i)	Yes	111	161	22	265

EXERCISE 3.8.2. Breslow and Day (1980) present data on the occurrence of esophageal cancer in Frenchmen. Explanatory factors are age and alcohol consumption. High consumption was taken to be anything over the equivalent of one liter of wine per day. The data are given in Table 3.16. Analyze the data as a logit model. In your analysis, consider the information on ordered age categories.

EXERCISE 3.8.3. The data in the previous experiment is a series of 2×2 tables collected under five different age conditions. This is the same situation as the Mantel-Haenszel setup of Exercise 3.8.9. The Mantel-Haenszel

TABLE 3.16. Occurrence of Esophageal Cancer

Age	Alcohol Consumption	Cancer Yes	Cancer No
25-34	High	1	9
	Low	0	106
35-44	High	4	26
	Low	5	164
45-54	High	25	29
	Low	21	138
55-64	High	42	27
	Low	34	139
65-74	High	19	18
	Low	36	88
75+	High	5	0
	Low	8	31

test is one of conditional independence given age. It assumes that the model of no three-factor interaction holds. Respecify the test in terms of logit models.

EXERCISE 3.8.4. Haberman (1978) reports data from the National Opinion Research Center on attitudes toward abortion (cf. Table 3.17). The data were collected over 3 years. Analyze the abortion attitude data treating attitude as a response variable.

Respondents were identified by their years of education and their religious group. The groups used were Catholics, Southern Protestants, and other Protestants. Southern Protestants were taken as Protestants who live in or south of Texas, Oklahoma, Arkansas, Kentucky, West Virginia, Maryland, and Delaware. Attitudes toward abortion were determined by whether the respondent thought that legal abortions should be available under three sets of circumstances. The three circumstances are (a) a strong chance exists of a serious birth defect, (b) the woman's health is threatened, and (c) the pregnancy was the result of rape. A negative response in the table consists of negative responses to all circumstances. A positive response is three positives. A mixed response is any other pattern. Find an appropriate model for the data. Interpret the model and draw conclusions from the estimates. (Haberman also presents similar data based on three different circumstances: the child is not wanted, the family is poor, and the mother unmarried.)

EXERCISE 3.8.5. Feigl and Zelen (1965), Cook and Weisberg (1982), and Johnson (1985) give data on survival of 33 leukemia patients as a function of their white blood cell count and the existence of a certain morphological characteristic in the cells. The characteristic is referred to as either AG positive or AG negative. The binary response is survival of at least 52 weeks

TABLE 3.17. Abortion Attitudes among Caucasian Christians

Year	Religion	Years of Education	Attitude Negative	Mixed	Positive
1974	Prot.	0-8	7	16	49
	Prot.	9-12	10	26	219
	Prot.	12+	4	10	131
1974	Prot. S.	0-8	1	19	30
	Prot. S.	9-12	5	21	106
	Prot. S.	12+	2	11	87
1974	Cath.	0-8	3	9	29
	Cath.	9-12	15	30	149
	Cath.	12+	11	18	69
1973	Prot.	0-8	4	16	59
	Prot.	9-12	6	24	197
	Prot.	12+	4	11	124
1973	Prot. S.	0-8	4	16	34
	Prot. S.	9-12	6	29	118
	Prot. S.	12+	1	4	82
1973	Cath.	0-8	2	14	32
	Cath.	9-12	16	45	141
	Cath.	12+	7	20	72
1972	Prot.	0-8	9	12	48
	Prot.	9-12	13	43	197
	Prot.	12+	4	9	139
1972	Prot. S.	0-8	9	17	30
	Prot. S.	9-12	6	10	97
	Prot. S.	12+	1	8	68
1972	Cath.	0-8	14	12	32
	Cath.	9-12	18	50	131
	Cath.	12+	8	13	64

beyond the time of diagnosis. The data are given in Table 3.18. Fit a logistic regression model with separate slopes and intercepts for AG positives and negatives. Examine the data for influential observations. Consider whether a log transformation of the white blood cell count is useful. Evaluate models with (a) the same slope for both AG groups, (b) the same intercept for both AG groups, and (c) the same slope and intercept. Examine each model for influential observations.

EXERCISE 3.8.6. Finney (1941) and Pregibon (1981) present data on the occurrence of vasoconstriction in the skin of the fingers as a function of the rate and volume of air breathed. The data are reproduced in Table 3.19. A constriction value of 1 indicates that constriction occurred. Analyze the data.

EXERCISE 3.8.7. Mosteller and Tukey (1977) reported data on verbal test scores for sixth graders. They used a sample of 20 Mid-Atlantic and

TABLE 3.18. Data on Leukemia Survival

Survival	Cell Count	AG	Survival	Cell Count	AG
1	2,300	+	1	4,400	−
1	750	+	1	3,000	−
1	4,300	+	0	4,000	−
1	2,600	+	0	1,500	−
0	6,000	+	0	9,000	−
1	10,500	+	0	5,300	−
1	10,000	+	0	10,000	−
0	17,000	+	0	19,000	−
0	5,400	+	0	27,000	−
1	7,000	+	0	28,000	−
1	9,400	+	0	31,000	−
0	32,000	+	0	26,000	−
0	35,000	+	0	21,000	−
0	52,000	+	0	79,000	−
0	100,000	+	0	100,000	−
0	100,000	+	0	100,000	−
1	100,000	+			

TABLE 3.19. Data on Vasoconstriction

Constriction	Volume	Rate	Constriction	Volume	Rate
1	0.825	3.7	0	2.0	0.4
1	1.09	3.5	0	1.36	0.95
1	2.5	1.25	0	1.35	1.35
1	1.5	0.75	0	1.36	1.5
1	3.2	0.8	1	1.78	1.6
1	3.5	0.7	0	1.5	0.6
0	0.75	0.6	1	1.5	1.8
0	1.7	1.1	0	1.9	0.95
0	0.75	0.9	1	0.95	1.9
0	0.45	0.9	0	0.4	1.6
0	0.57	0.8	1	0.75	2.7
0	2.75	0.55	0	0.03	2.35
0	3.0	0.6	0	1.83	1.1
1	2.33	1.4	1	2.2	1.1
1	3.75	0.75	1	2.0	1.2
1	1.64	2.3	1	3.33	0.8
1	1.6	3.2	0	1.9	0.95
1	1.415	0.85	0	1.9	0.75
0	1.06	1.7	1	1.625	1.3
1	1.8	1.8			

New England schools taken from *The Coleman Report*. The dependent variable y was the mean verbal test score for each school. The predictor variables were x_1 — staff salaries per pupil, x_2 — percent of sixth grade fathers employed in white collar jobs, x_3 — a composite score measuring socioeconomic status, x_4 — the mean score on a verbal test administered to teachers, and x_5 — one-half of the sixth grade mothers' mean number of years of schooling. Schools meet a performance standard set for them (by me) if their average verbal test score is above 37. The data are given in Table 3.20.

(a) Using logistic regression on the 0-1 scores, find a good model for predicting whether schools meet the standard.

(b) Using standard regression on y, find several of the best predictive models. Compare these to your logistic regression model.

TABLE 3.20. Verbal Test Scores

Obs.	x_1	x_2	x_3	x_4	x_5	Score	y
1	3.83	28.87	7.20	26.60	6.19	1	37.01
2	2.89	20.10	−11.71	24.40	5.17	0	26.51
3	2.86	69.05	12.32	25.70	7.04	0	36.51
4	2.92	65.40	14.28	25.70	7.10	1	40.70
5	3.06	29.59	6.31	25.40	6.15	1	37.10
6	2.07	44.82	6.16	21.60	6.41	0	33.90
7	2.52	77.37	12.70	24.90	6.86	1	41.80
8	2.45	24.67	−0.17	25.01	5.78	0	33.40
9	3.13	65.01	9.85	26.60	6.51	1	41.01
10	2.44	9.99	−0.05	28.01	5.57	1	37.20
11	2.09	12.20	−12.86	23.51	5.62	0	23.30
12	2.52	22.55	0.92	23.60	5.34	0	35.20
13	2.22	14.30	4.77	24.51	5.80	0	34.90
14	2.67	31.79	−0.96	25.80	6.19	0	33.10
15	2.71	11.60	−16.04	25.20	5.62	0	22.70
16	3.14	68.47	10.62	25.01	6.94	1	39.70
17	3.54	42.64	2.66	25.01	6.33	0	31.80
18	2.52	16.70	−10.99	24.80	6.01	0	31.70
19	2.68	86.27	15.03	25.51	7.51	1	43.10
20	2.37	76.73	12.77	24.51	6.96	1	41.01

EXERCISE 3.8.8. *The Logistic Distribution.*
Show that $F(x) = e^x/(1 + e^x)$ satisfies the properties of a cumulative distribution function (cdf). Any random variable with this cdf is said to have a *logistic distribution*.

EXERCISE 3.8.9. *Stimulus–Response Studies.*
The effects of a drug or other stimulus are often studied by choosing r doses of the drug (levels of the stimulus), say $x_1, \ldots, x_r$, and giving the dose x_j to each of N_j subjects. The data consist of the number y_j who

exhibit some predetermined response. Often this response is the death of the subject, but it can be any measure of the effectiveness of the stimulus. Typically in *dose-response studies*, interest centers on the median effective dose, the $ED(50)$, or if the response is death, the median lethal dose, the $LD(50)$. The $LD(50)$ is that dose for which the probability is 0.5 that a subject will die. Frequently, a model of the form

$$\log\left[p_j/(1-p_j)\right] = \alpha + \beta \log(x_j)$$

is fitted to such data. Assume this model holds for any and all doses.

(a) Is the sampling scheme appropriate for a logistic regression?

(b) How could you estimate the $LD(50)$?

Exercises 11.8.2 and 11.8.3 give additional results on inference for the $LD(50)$.

Suppose that for each individual in the population there is a minimum dose x to which the individual will respond. (The individual is assumed to respond to all doses larger than x.) Let the random variable X be this minimum susceptibility for an individual chosen at random.

(c) Give an estimate of the median of X.

(d) Give an estimate of the 90th percentile of the distribution of X.

(e) What is the distribution of X?

EXERCISE 3.8.10. *Probit Analysis.*
An alternative to the logistic analysis of dose-response data is *probit analysis*. In probit analysis, the model is

$$\Phi^{-1}(p_j) = \alpha + \beta \log(x_j)$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution. Graph $\Phi(\cdot)$ and the logistic cdf and compare the general shapes of the distributions. In light of the two previous exercises, give a brief summary of the similarities and differences of logit and probit analysis. For more information on probit analysis, the interested reader should consult Finney (1971).

EXERCISE 3.8.11. Woodward et al. (1941) report several data sets, one of which examines the relationship between exposure to chloracetic acid and the death of mice. Ten mice were exposed at each dose level. The data are given in Table 3.21. Doses are measured in grams per kilogram of body weight. Fit the logistic regression model of Exercise 3.8.9 and estimate the $LD(50)$. Try to determine how well the model fits the data.

EXERCISE 3.8.12. Consider a sample of $j = 1, \ldots, r$ independent binomials $y_j \sim \text{Bin}(N_j, p_j)$, each with a covariate x_j. Suppose that for some cumulative distribution function $F(\cdot)$,

$$p_j = F(x_j).$$

TABLE 3.21. Lethality of Chloracetic Acid

Dose	Fatalities	Dose	Fatalities
.0794	1	.1778	4
.1000	2	.1995	6
.1259	1	.2239	4
.1413	0	.2512	5
.1500	1	.2818	5
.1588	2	.3162	8

Show that for some transformation $z_j = g(x_j)$ and parameters α and β, this logistic regression model holds:

$$\log\left(\frac{p_j}{1-p_j}\right) = \alpha + \beta z_j.$$

EXERCISE 3.8.13. The data of Exercise 3.8.2 are actually a retrospective study. A sample of cancer patients was compared to a sample of men drawn from the electoral lists of the department of Ille-et-Vilaine in Brittany. Reanalyze the data in light of this knowledge.

EXERCISE 3.8.14. Any multinomial response model can be viewed as the model for an $I \times J$ table. Assume product-multinomial sampling from J independent multinomials each with I categories. Define

$$\pi_{ij} = p_{ij} \Big/ \sum_{h=i}^{I} p_{hj}$$

so that the continuation ratios introduced in Section 4.2 are

$$\frac{\pi_{ij}}{1-\pi_{ij}} = \frac{p_{ij}}{\sum_{h=i+1}^{I} p_{hj}}.$$

Let $r_{ij} = n_{\cdot j} - \sum_{h=1}^{i-1} n_{hj}$; show that the product-multinomial likelihood

$$\prod_{j=1}^{J} \left\{ \frac{n_{\cdot j}!}{\prod_{i=1}^{I} n_{ij}!} \prod_{i=1}^{I} p_{ij}^{n_{ij}} \right\}$$

can be written as the product of binomial likelihoods, i.e.,

$$\prod_{j=1}^{J} \prod_{i=1}^{I-1} \binom{r_{ij}}{n_{ij}} \pi_{ij}^{n_{ij}} (1-\pi_{ij})^{r_{ij}-n_{ij}}.$$

Using this result and maximum likelihood estimation, show that a set of continuation ratio models can be fitted simultaneously to the entire table by

fitting each continuation ratio model separately. Note that the chi-square statistics for fitting each continuation ratio model can be added to get a chi-square statistic for the entire table.

EXERCISE 3.8.15. Give the log-linear model corresponding to (4.1.1).

EXERCISE 3.8.16. Analyze the trauma data that are described in Example 13.2.2.

4

Independence Relationships and Graphical Models

As mentioned in Section 3.7, all of the general principles of testing and estimation presented for three-factor tables also apply when there are additional classification factors. The main difference in working with higher-dimensional tables is that things become more complicated. First, there are many more ANOVA type models to consider. For example, in a four-factor table, there are 113 ANOVA models that include all of the main effects. In five-factor tables, there are several thousand models to consider. Second, a great many of the models require iterative methods for obtaining maximum likelihood estimates. Finally, interpretation of higher-dimensional models is more difficult.

In this chapter, we examine interpretations of models for four and higher-dimensional tables, graphical models, conditions that allow tables to be collapsed, and a variety of graphical models known as recursive causal models.

4.1 Model Interpretations

This section provides tools for interpreting log-linear models for higher-dimensional tables. The interpretations are based on independence and conditional independence. The tools are based on viewing higher-dimensional tables as three-dimensional tables. In Section 2, we present alternative methods based on exploiting the relationships between graph theory and conditional independence.

An example of a model with four factors is

$$\begin{aligned}\log(m_{hijk}) &= u + u_{1(h)} + u_{2(i)} + u_{3(j)} + u_{4(k)} \\ &\quad + u_{12(hi)} + u_{13(hj)} + u_{23(ij)} + u_{123(hij)} \\ &\quad + u_{14(hk)} + u_{24(ik)} \ .\end{aligned}$$

Eliminating redundant parameters gives

$$\log(m_{hijk}) = u_{123(hij)} + u_{14(hk)} + u_{24(ik)} \ .$$

The shorthand notation for this model is [123][14][24]. Our discussion of model interpretations will be based exclusively on the shorthand notation for models.

Consider the model [123][124]. If we think of all combinations of factors 1 and 2 as a single factor, then we get a three-factor table with factors (12), 3, and 4. The model [(12)3][(12)4] becomes a three-dimensional model of conditional independence. Given the levels of factors 1 and 2, factor 3 is independent of factor 4.

This trick of combining factors to reduce a four-factor model into a three-factor model is very useful. The model [123][14] can be considered as a three-factor model in which all combinations of factors 2 and 3 are a single factor. The model [123][14] can then be interpreted as saying that given factor 1, factor 4 is independent of factors 2 and 3. Note that the model puts no constraints on the relationship between factors 2 and 3, and that conditional probabilities involving factors 2, 3, and 4 can change with the level of factor 1, cf. Example 1.1.5.

Using the principle of combining factors, it is easy to see that [123][4] indicates that factor 4 is independent of factors 1, 2, and 3, but that the relationship between factors 1, 2, and 3 is unspecified. Also, [12][34] indicates that factors 1 and 2 may be related, factors 3 and 4 may be related, but 1 and 2 are independent of 3 and 4.

A second useful trick in interpreting models is looking at larger models. If a particular model is true, then any larger model is also true. If the larger model has an interpretation in terms of independence, then the smaller model admits the same interpretation.

For example, consider the three-factor models [12][3] and [12][23]. The smaller model [12][3] indicates that factors 1 and 2 are independent of factor 3. In particular, factors 1 and 3 are independent given the level of factor 2. This is the interpretation of the larger model [12][23]. The interpretation of the larger model is also valid for the smaller model. Note, however, that if two models are both valid and both are interpretable, then the smaller model gives the more powerful interpretation. Often, *we want to identify the smallest interpretable model that fits.*

Now consider the four-factor model [12][13][14]. One larger model is [123][14], so factors 2 and 3 are independent of factor 4 given factor 1. Similarly, [12][134] and [13][124] are also larger models, so we find that

given factor 1, the other three factors are all independent. Note that the structure of the model [12][13][14] makes this interpretation almost self-evident. Factor 1 is included in all three terms, so it is the variable that is fixed in the conditional probabilities. Factors 2, 3, and 4 are in separate terms, so they are independent given factor 1.

EXERCISE 4.1. By examining the probabilities p_{hijk}, show that the three larger models imply conditional independence for factors 2, 3, and 4 in [12][13][14].

A more complicated example is [12][13][24]. One larger model is [13][124]. Thus, given factor 1, factor 3 is independent of factors 2 and 4. Another larger model is [24][123]; thus, given factor 2, factor 4 is independent of factors 1 and 3.

Finally, consider the model [12][13][14][23]. The larger model [123][14] implies that 2 and 3 are independent of 4 given 1. In fact, this is the only simple interpretation associated with [12][13][14][23]. To see this, ignore factor 4. The three-factor model has the terms [12][13][23], which has no simple interpretation as a three-dimensional model. The next largest model is to replace [12][13][23] with [123]. If we do this in the four-factor model, we replace [12][13][14][23] with [123][14]. Any other model with a simple interpretation would have to be larger than [123][14]. However, because [123][14] already has a simple interpretation, we have the best explanation available. (Recall that the smaller the model, the more powerful the interpretation in terms of independence.)

Table 5.1 summarizes the discussion above and also includes some additional models. Note that it is the pattern of the models that determines interpretability. Just as [12][13][14] indicates that 2, 3, and 4 are independent given 1, the model [12][23][24] indicates that factors 1, 3, and 4 are independent given factor 2. Any relabeling of the factors in Table 5.1 gives another interpretable model. It is important to remember that while models imply certain interpretations, more than one model generates the same interpretation. For example, the model [123][14] gives the interpretation that given 1, factors 2 and 3 are independent of factor 4. Conversely, the condition that given 1, factors 2 and 3 are independent of factor 4 implies that [123][14] must hold. However, the independence condition is also consistent with the smaller model [12][13][23][14]. This smaller model is equivalent to the independence condition along with the additional condition that there is no u_{123} interaction.

Goodman (1970, 1971) and Haberman (1974a) introduced the concept of *decomposable* log-linear models. These models are also called *multiplicative. The class of decomposable models consists of all models that have closed form maximum likelihood estimates.* They also have simple interpretations in terms of independence or conditional independence. For example, all models for three factors other than [12][13][23] are decomposable. In Table

TABLE 4.1. Some Models and Their Conditional Independence Interpretations

Model	Interpretation
[123][124]	Given 1 and 2, factors 3 and 4 are independent.
[123][14][24]*	Given 1 and 2, factors 3 and 4 are independent.
[123][14]	Given 1, factor 4 is independent of factors 2 and 3.
[12][13][14][23]*	Given 1, factor 4 is independent of factors 2 and 3.
[123][4]	Factor 4 is independent of factors 1, 2, and 3.
[12][23][34][41]	Given 2 and 4, factors 1 and 3 are independent. Given 1 and 3, factors 2 and 4 are independent.
[12][13][14]	Given 1, factors 2, 3, and 4 are all independent.
[12][13][24]	Given 1, factor 3 is independent of factors 2 and 4. Given 2, factor 4 is independent of factors 1 and 3.
[12][34]	Factors 1 and 2 are independent of factors 3 and 4.
[12][13][4]	Factor 4 is independent of factors 1, 2, and 3. Given 1, factor 2 is independent of factor 3.
[12][3][4]	Factor 3 is independent of factors 1, 2, and 4. Factor 4 is independent of factors 1, 2, and 3.
[1][2][3][4]	All factors are independent of all other factors.

*These models imply their interpretations; however, the interpretations do not imply the models.

5.1, all models except the two with asterisks and [12][23][34][41] are decomposable. Note that [12][23][34][41] is not decomposable but is still characterized by its conditional independence relations. It is particularly easy to work with decomposable models because they have very simple structure. Often, results that are difficult or impossible to prove for arbitrary log-linear models can be shown for decomposable models, e.g., Bedrick (1983) and Koehler (1986). An exact characterization of decomposable models is given in the next section.

4.2 Graphical and Decomposable Models

Models that have interpretations in terms of conditional independence are known as *graphical* models. The terminology stems from the relationship of these models to graph theory. Berge (1973) gives a discussion of graph theory that is particularly germane but does not give statistical applications.

Edwards and Kreiner (1983) give an overview of the use of graphical log-linear models. More recently, Edwards (1995) provides an introduction to the uses of graphical models in statistics, including applications other than log-linear models. More advanced recent books include Whittaker (1990) and Lauritzen (1996).

Graphical models are determined by their two-factor interactions. The basic idea is that any graphical model containing all of the terms u_{12}, u_{13}, and u_{23} must also include u_{123}. To extend this, consider a graphical model that includes u_{12}, u_{13}, u_{23}, u_{24}, and u_{34}. The terms u_{12}, u_{13}, and u_{23} imply that u_{123} must be in the graphical model and u_{23}, u_{24}, and u_{34} imply that u_{234} must be in the model. A graphical model must contain u_{1234} if it includes all six of the two-factor terms that can be formed from the four factors, i.e., if it includes u_{12}, u_{13}, u_{14}, u_{23}, u_{24}, and u_{34}.

Definition 4.2.1. A model is *graphical* if, whenever the model contains all two-factor terms generated by a higher-order interaction, the model also contains the higher-order interaction. In graph theory, the corresponding idea is that of a *conformal* graph.

Obviously, we need to discuss both models and the two-factor effects generated by higher-order terms. The model [1234][345] is determined by the four-factor term u_{1234} and the three-factor term u_{345}. The four-factor term u_{1234} generates the two-factor terms u_{12}, u_{13}, u_{14}, u_{23}, u_{24}, and u_{34}. The three-factor u_{345} term subsumes the two-factor terms u_{34}, u_{35}, and u_{45}. We will refer to the four-factor term [1234] as generating [12], [13], [14], [23], [24], and [34]. Similarly, the three-factor term [345] generates the two-factor terms [34], [35], and [45]. Conversely, any graphical model that contains the two-factor effects [12], [13], [14], [23], [24], and [34] must include [1234] (or a larger term that subsumes [1234]) and a graphical model that includes [34], [35], and [45] also includes either [345] or a larger term.

EXAMPLE 4.2.2. *Three-Factor Models.*
The two-factor terms that are possible with three-factors are [12], [13], and [23]. The two-factor effects generate only one higher-order interaction, [123]. The only three-factor models that contain all of the two-factor terms are [123] and [12][13][23]. The model [123] contains all the two-factor effects and the higher-order term, so [123] is graphical. The model [12][13][23] contains all the two-factor effects generated by [123] but does not contain the higher-order term, so it is not graphical. None of the other three-factor models contain all of the two-factor interactions, so, by default, they are all graphical models.

EXAMPLE 4.2.3. *Four-Factor Models*
The model [123][24] is graphical because it includes the three-factor term [123] and *it does not contain all of the two-factor terms generated by any*

other higher-order terms. This follows because all higher-order terms other than [123] involve factor 4, so each generates at least 2 two-factor terms that involve factor 4. The model [123][24] includes only one such term, [24], so, by default, the model does not include all the two-factor terms for any higher-order interaction other than [123]. Similarly, the model [123][124] is graphical because the two-factor terms that are present only generate the three-factor interactions in the model.

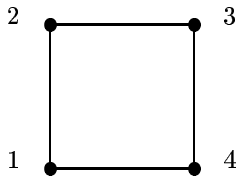
A model that is not graphical is [12][13][14][23]. It includes all of [12], [13], and [23], but it does not include [123]. The model [123][124][234] is not graphical because it contains all six of the possible two-factor effects but does not contain [1234]. Except for the models indicated by asterisks, all of the models in Table 5.1 are graphical models.

Any log-linear model can be embedded in a graphical model. This follows immediately from the fact that the saturated model is the graphical model having all possible two-factor effects. *To interpret a specific log-linear model, one seeks the smallest graphical models that contain the specific model.*

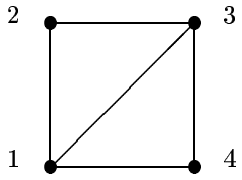
EXAMPLE 4.2.4. The nongraphical model [12][13][14][23] is a submodel of the graphical model [123][14]. The nongraphical model retains the conditional independence interpretation, 2 and 3 independent of 4 given 1 which is appropriate for the larger graphical model; however, the nongraphical model involves additional constraints.

Amazingly, graphical models can be displayed graphically. (Wonders never cease!) In this context, graphs are not directly related to Cartesian coordinates but rather to graph theory. A graph consists of *vertices* (nodes) and *edges. Vertices correspond to factors in log-linear models. Edges correspond to two-factor effects.* Note that graphs based on two-factor effects would be useless without a convention that dictates how two-factor effects determine a log-linear model. Thus, pictures of graphical models are worthless until after graphical models have been defined. A key feature of this subject is the one-to-one correspondence between graphical log-linear models and graphs. Every model determines a graph and every graph determines a model.

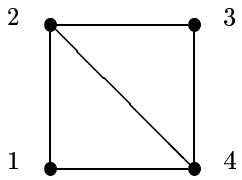
EXAMPLE 4.2.5. Consider the model [12][23][34][41]. Factors are points on a graph (*vertices*) and two-factor interactions are allowable paths (*edges*) between points. The graph is given below.



Now consider the model [123][134]. The two-factor terms generated by [123] are [12], [23], [13] and the terms [13], [34], [14] are generated by [134]. Note that [13] is common to both sets of two-factor terms. The corresponding graph is given below.

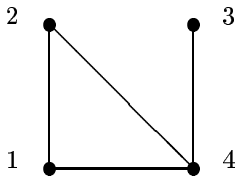


We can also read log-linear models directly from the corresponding graph. For example, the graph



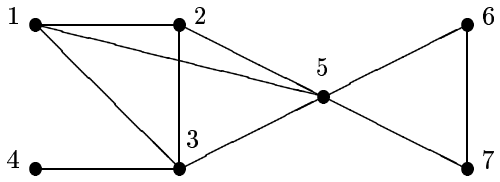
has the two-factor effects [12], [24], [14]; these generate the three-factor term [124]. The graph also contains the edges [23], [34], [24] that generate [234]. We have accounted for all of the edges in the graph, so the model is [124][234].

As another example, consider the following graph.



Again the edges [12], [24], [14] generate the three-factor term [124]. However, this graph does not have all of the edges that generate [234] because the graph does not contain [23]. The term [34] is not included in any larger term, so it must be included separately. The model is [124][34].

Finally, consider a seven-factor graph.

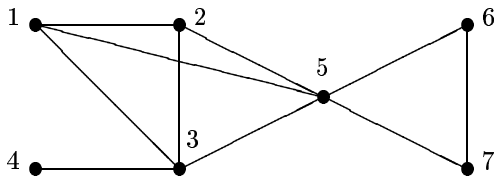


The graph contains all possible edges between the vertices 1, 2, 3, and 5; thus, the graphical log-linear model includes the term [1235]. Similarly, the graph contains all possible edges between the vertices 5, 6, and 7; therefore the log-linear model includes the term [567]. Finally, the graph contains the isolated edge [34], so this term must be in the model. These three terms account for all of the edges in the graph, so the graphical log-linear model is [1235][34][567].

Note that the graph also contains all possible edges between other sets of vertices; for example, all of the edges between 1, 2, and 5 are in the graph so the model includes [125]. However, [125] has already been forced into the model by the inclusion of [1235]. The graphical log-linear model is determined by the largest sets of vertices that include all possible edges between them. The set $\{1, 2, 5\}$ is unimportant because it is contained in the set $\{1, 2, 3, 5\}$. A set of vertices for which all the vertices are connected by edges is *complete*. Both the sets $\{1, 2, 5\}$ and $\{1, 2, 3, 5\}$ are complete. A maximal complete set (i.e., a complete set that is not contained in any other complete set), is called a *clique*. *The cliques of a graph determine the graphical log-linear model.* The set $\{1, 2, 3, 5\}$ is a clique, but the set $\{1, 2, 5\}$ is not maximal, so it is not a clique. In the graph of the model [1235][34][567], the cliques are $\{1, 2, 3, 5\}$, $\{3, 4\}$, and $\{5, 6, 7\}$. There is an obvious correspondence between the cliques and the $[\cdot]$ notation defining the

model. In the future, we will simply indicate the cliques as [1235], [134], and [567]. Because the cliques determine the model, the concept of a clique is of fundamental importance. Surprisingly, that importance need not be made explicit in the remainder of this discussion. However, in Wermuth's method of model selection, the role of cliques cannot be ignored, cf. Section 6.5.

Perhaps the most important reason for graphing log-linear models is that independence relations can be read directly from the graph. To do this, we need to introduce the concept of a chain. A *chain* is simply a sequence of edges that lead from one factor (vertex) to another factor. In the graph



there are a huge number of chains. For example, there is a chain from 1 to 2 to 5, a chain from 1 to 2 to 5 to 6, a chain from 1 to 2 to 5 to 7 to 6, a chain from 1 to 2 to 5 to 3 to 4, and many others. Note that a chain involves not only the end points but also all the intermediate points. In other words, there is a chain from 1 to 3 to 5 to 7, but the graph contains no chain from 1 to 3 to 7 because the graph does not include the edge [37]. We allow chains to begin and end at the same point, e.g., 3 to 1 to 5 to 3; in other words, *round-trips are allowed.* However, we do not allow the path to include a factor more than once. For example, we do not allow 3 to 5 to 6 to 7 to 5 to 1. Even though this path never uses the same edge twice, it does go through factor 5 twice. In a sense, there is no real loss in excluding such paths because we can still get from factor 3 to factor 1 by taking the path 3 to 5 to 1. We are just *not allowing ourselves to drive around in circles.* A formal definition of chains is given below.

Definition 4.2.6. Let h and j be factors and let $\{i_1, \ldots, i_k\}$ be a sequence of factors that are distinct from each other and from h and j. The sequence of edges $C_{hj} = \{[hi_1], [i_1 i_2], \ldots, [i_k j]\}$ is a *chain* between h and j. A graph contains the chain C_{hj} if the graph contains all of the edges included in the chain.

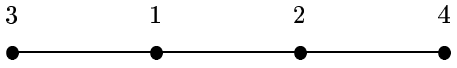
We get a degenerate chain if we start at h, go to another vertex i, and back to h. This is the only situation in which an edge could appear twice within the sequence of edges defining a chain. Nonetheless, this chain only contains one edge.

The key result on independence follows.

Theorem 4.2.7. Let the sets A, B, and C denote disjoint subsets of the factors in a graphical model. The factors in A are independent of the factors in B given C if and only if every chain between a factor in A and a factor in B involves at least one factor in C.

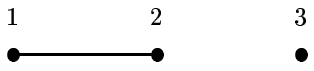
Proof. See Darroch, Lauritzen, and Speed (1980). □

EXAMPLE 4.2.8. The four-factor model [12][13][24] is illustrated below. It is graphical, so Theorem 4.2.7 applies. Rewrite the model as [31][12][24]. By the theorem, factor 3 is independent of 2 and 4 given 1, factors 3 and 1 are independent of 4 given 2, and factors 3 and 4 are independent given 1 and 2. There are three independence conditions here and all are necessary. For example, the model that only specifies 3 independent of 2 and 4 given 1 is [31][124], not [12][13][24].



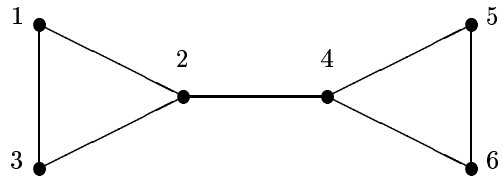
Theorem 4.2.7 also implies certain marginal independence relations. For example, factor 3 is independent of factor 2 given 1. This is a statement about the marginal distribution of factors 1, 2, and 3.

Consider the model [12][3]. There are no chains connecting factors 1 and 2 with 3, so every chain that connects them involves at least one member of the empty set. Thus, 1 and 2 are independent of 3 given the factors in the empty set; i.e., 1 and 2 are independent of 3.



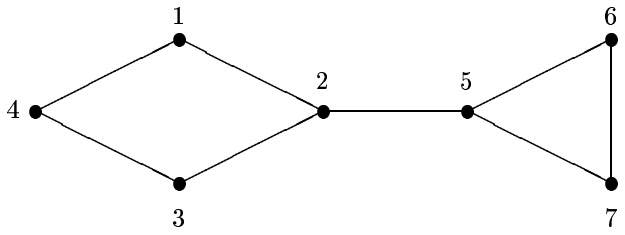
Marginally, we can conclude that 3 is independent of 2 and that 3 is independent of 1.

In the model [123][24][456], 1 and 3 are independent of 5 and 6 given 2 and 4. Similarly, 1 and 3 are independent of 4, 5, and 6 given 2. Also, 5 and 6 are independent of 1, 2, and 3 given 4.



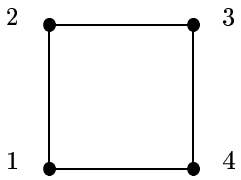
Many marginal independence relationships hold for this model. For example, 1 and 3 are independent of 4 and 5 given 2.

EXERCISE 4.2. (a) Graph the 10 graphical models in Table 5.1.
(b) List 10 of the independence relationships in [12][23][34][41][25][567].



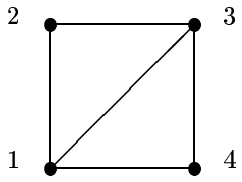
The decomposable models discussed at the end of the previous section form a subset of the graphical models. They are the graphical models that have the additional condition of being *chordal*. The terms given in parentheses in the example below are graph theory terms.

EXAMPLE 4.2.9. Suppose that a model contains the interactions [12], [23], [34], [41]. We can start at *any* of the points and travel in a cycle (*closed chain*) back to that point. For example, we can travel from 1 to 2, from 2 to 3, from 3 to 4, and from 4 back to 1.



The model (graph) is *chordal* if every such cycle among four or more vertices has a shortcut. A shortcut is called a *chord*. The cycle given above has two

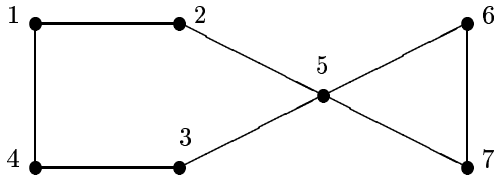
possible shortcuts: [31] and [24]; adding either or both of these would make the model chordal. For example, if the model also includes [31], a cycle from 1 back to 1 can be shortened by traveling [12], [23], [31].



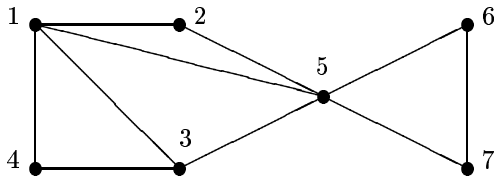
Similarly, a trip from 2 back to 2 can be shortened by traveling [23], [31], [12]. The graphical model generated by the terms [12], [23], [34], [41], and [31] is [123][134]. This decomposable model has the interpretation that given factors 1 and 3, factors 2 and 4 are independent. The maximum likelihood estimates are $\hat{m}_{hijk} = n_{hij\cdot}n_{h\cdot jk}/n_{h\cdot j\cdot}$.

The *length* of a chain is the number of edges in it. The closed chain [12], [23], [34], [41] has length four. The closed chain [12], [23], [31] has length three. A closed chain among four or more vertices is a closed chain of length four or more.

Consider the model [12][25][53][34][41][567] given below in graphical form.



This is not decomposable because the closed chain [12], [25], [53], [34], [41] involves five vertices and has no chords. The model requires the addition of at least two additional two-factor effects to convert it into a decomposable model. Adding one edge to the offending chain still leaves a cycle of length four without a chord. For example, adding [15] leaves the cycle [15], [53], [34], [41] without a chord. The following graph adds both [15] and [13].



This is now the graph of a decomposable log-linear model. All cycles of length four or more have a chord. The corresponding log-linear model is [143][135][125][567].

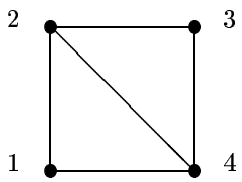
These ideas are formalized in the following definition.

Definition 4.2.10. Consider the chain $C_{hj} = \{[hi_1], [i_1 i_2], \ldots, [i_k j]\}$. The *length* of the chain is the number of edges (distinct elements) in C_{hj}. A chain C_0 is a *reduced chain* relative to C_{hj} if C_0 is a chain between h and j with length less than that of C_{hj} and if every factor involved in C_0 is also involved in C_{hj}. Note that it is the factors (vertices) that define reduced, not the effects (edges). A *closed chain* is a chain between a factor h and itself with length greater than 1. In particular, the chain from h to h, $C_{hj} = \{[hi], [ih]\}$, contains only one (distinct) edge, so it has length 1 and does not form a closed chain. Any two-factor term $[i_r i_s]$ is a *chord* of the closed chain $C_{hh} = \{[hi_1], \ldots, [i_{r-1} i_r], [i_r i_{r+1}], \ldots, [i_{s-1} i_s], [i_s i_{s+1}], \ldots, [i_k h]\}$ if the sequence $\{[hi_1], \ldots, [i_{r-1} i_r], [i_r i_s], [i_s i_{s+1}], \ldots, [i_k h]\}$ is a closed reduced chain relative to C_{hh}. It is allowable for either factor in $[i_r i_s]$ to be h. A model is *chordal* if every closed chain of length $k \geq 4$ generated by the model has a chord that is in the model. A model is *decomposable* if it is both graphical and chordal.

By definition, a closed chain must have a length of at least 2; it follows immediately that a closed chain must have a length of at least 3. Clearly, a closed chain of length three cannot have a chord, so it is natural that the definition of chordal models involves closed chains of length 4 or more. It is possible for a model to be chordal without being graphical. Chordal models have restrictions on the two-factor terms; they place no requirements on higher-order terms. In graph theory, decomposable models correspond to *acyclic hypergraphs.*

EXAMPLE 4.2.11. The effects [12], [23], [34] define a chain from 1 to 4. The effects [12], [24] define a reduced chain from 1 to 4. The effects in the model [12][23][34][41] define a closed chain from 1 back to 1 but also from 2 back to 2, from 3 back to 3, and from 4 back to 4. To see the last of these,

observe that the model contains the closed chain [41], [12], [23], [34]. The possible chords for these closed chains are [31] and [24]. To see that [24] is a chord, observe that [12], [24], [41] defines a closed reduced chain of [12], [23], [34], [41].

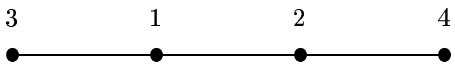


The nongraphical model [12][23][34][41][24] is chordal because any closed chain of length four has a chord that is in the model. This model has no closed chains of length greater than four because it involves only four factors. The model [12][23][34][41][24] is not decomposable because it is not graphical. It contains all of [12], [41], and [24] but not [124]. The corresponding decomposable model is [124][234]. This model generates precisely the two-factor terms [12], [23], [34], [41], and [24], so it is both graphical and chordal.

With four factors, the only graphical model that is not decomposable is [12][23][34][41].

Decomposable models have closed form estimates. We illustrate one simple case.

EXAMPLE 4.2.12. Consider the graph



The corresponding model is [12][13][24]. The probability structure can be read from the model,

$$p_{hijk} = \frac{p_{hi\cdot\cdot}p_{h\cdot j\cdot}p_{\cdot i\cdot k}}{p_{h\cdot\cdot\cdot}p_{\cdot i\cdot\cdot}},$$

where the terms in the numerator are determined by the terms in the model (the marginal probabilities in the numerator correspond to the margins fitted by the model) and the terms in the denominator correspond to the factors that appear in more than one term in the model. Marginal probabilities are estimated from marginal tables, e.g., $\hat{p}_{hi\cdot\cdot} = n_{hi\cdot\cdot}/n_{\cdot\cdot\cdot\cdot}$. The estimated expected counts are

$$\hat{m}_{hijk} = n_{\cdot\cdot\cdot\cdot}\hat{p}_{hijk} = \frac{n_{hi\cdot\cdot}n_{h\cdot j\cdot}n_{\cdot i\cdot k}}{n_{h\cdot\cdot\cdot}n_{\cdot i\cdot\cdot}}.$$

Decomposable models are closely related to *recursive causal models.* Recursive causal models use ideas from directed graphs to indicate causation. As mentioned in the introduction to Chapter 4, causation is not something that can be inferred from data. Any causation must be inferred from other sources. Recursive causal models are introduced in Section 4. The interested reader can also consult the relevant literature, e.g., Wermuth and Lauritzen (1983), Kiiveri, Speed, and Carlin (1984), and the fine expository paper by Kiiveri and Speed (1982).

The interplay between graph theory and statistics is a fascinating subject with implications for log-linear models, *covariance selection, factor analysis, structural equation models, artificial intelligence*, and *database management.* Reviews are given by Kiiveri and Speed (1982), Edwards (1995), Whittaker (1990), and Lauritzen (1996). For applications to log-linear models, see the references listed in the previous paragraph along with Darroch, Lauritzen, and Speed (1980). These articles cite a wealth of related work including the important contributions of Leo Goodman and Shelby Haberman.

4.3 Collapsing Tables

One important function of statistics is to summarize large batches of numbers. This is such a fundamental aspect of statistics that it is easily overlooked. For example, formal theories of statistics are generally based on the use of sufficient statistics, cf. Cox and Hinkley (1974). Intuitively, a sufficient statistic is simply a summary of the data that is sufficient for drawing valid conclusions about the data. The use of sufficient statistics is an enormous advantage both theoretically and practically.

An early step in analyzing many sets of data is the construction of a table. Tables organize data in a way that makes the data more understandable. Clearly, small tables are easier to understand than large tables. For example, a 3×5 table is typically easier to understand than a $3 \times 5 \times 4$ table. In this section, we establish conditions that, if satisfied, allow us to collapse a $3 \times 5 \times 4$ table of counts into a 3×5 table and still draw valid conclusions. Recall that collapsing is not always possible. Simpson's paradox is precisely the result of collapsing a table that cannot be validly collapsed. First, we discuss collapsing in three-factor tables and then extend the discussion to higher-order tables.

Collapsed tables are also used in analysis of variance. The three-factor ANOVA model

$$\begin{aligned} y_{ijk\ell} &= \mu + \alpha_i + \beta_j + \gamma_k \\ &\quad + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + e_{ijk\ell} \ , \end{aligned}$$

$i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$, $\ell = 1, \ldots, N$, is a model for analyzing the $I \times J \times K$ table of means $\bar{y}_{ijk\cdot}$. It is well known that with

no three-factor $(\alpha\beta\gamma)$ interaction, each of the two-factor interactions can be examined by looking at the corresponding two-factor marginal table. For example, the two-factor interaction $(\alpha\beta)$ can be investigated using the $I \times J$ table of $\bar{y}_{ij\cdot\cdot}$'s.

The situation with tables of counts is more complex. In general, *if a log-linear model has no three-factor interaction and if all two-factor interactions exist, it is not valid to draw conclusions about two-factor interactions from the two-factor marginal tables.*

As discussed earlier, two-factor interactions are closely related to odds ratios. A three-factor table is said to be collapsible over factor 1 if the odds ratios in the marginal table $p_{\cdot jk}$ are identical to the odds ratios for each row of the three-way table. In other words, we can collapse on rows (factor 1) if for all i, j, j', k, and k',

$$\frac{p_{\cdot jk}p_{\cdot j'k'}}{p_{\cdot j'k}p_{\cdot jk'}} = \frac{p_{ijk}p_{ij'k'}}{p_{ij'k}p_{ijk'}} \ . \tag{1}$$

If this is true, we can draw valid inferences about the relationship between factors 2 and 3 by looking only at the marginal table. (This is clearly an easier task than examining a separate two-way table for each level of factor 1.)

As shown in equation (3.2.3) in the subsection Odds Ratios and Independence Models, equation (1) holds if rows and columns are independent given layers. So, under this model, the relationship between columns and layers does not depend on rows. Similarly, the row-layer relationship can be investigated in the row-layer marginal table if rows and columns are independent given layers.

Note that in order to have equation (1) hold, it is not necessary that rows and columns be independent given layers. It is easily seen that if rows and layers are independent given columns, then (1) still holds. Moreover, if both models [13][23] and [12][23] hold, then we must have [1][23], so rows are independent of both columns and layers. Obviously, in this case, collapsing over rows is allowable.

The validity of collapsing over a factor is a property of the parameters p_{ijk}. Data analysis is based on the observed values n_{ijk}. It is important to realize that the MLE of the odds ratio $p_{\cdot jk}p_{\cdot j'k'}/p_{\cdot j'k}p_{\cdot jk'}$ under either model [13][23] or [12][23] is $n_{\cdot jk}n_{\cdot j'k'}/n_{\cdot j'k}n_{\cdot jk'}$ and that this is also the MLE from the marginal table of $n_{\cdot jk}$'s. Thus, data analysis is identical whether working with the three-dimensional table or with the collapsed table. Collapsed tables are such a useful and intuitive tool that we have already used them in data analysis. The reader should note that collapsed tables were an integral part of both Examples 3.2.2 and 3.2.3.

Our results on collapsing three-factor tables are summarized in the next theorem.

Theorem 4.3.1.

(a) If the model [13][23] holds, then the relationship between factors 2 and 3 can be examined in the marginal table $n_{\cdot jk}$ and the relationship between factors 1 and 3 can be examined in the marginal table $n_{i\cdot k}$.

(b) If either model [13][23] or [12][23] holds, then the relationship between factors 2 and 3 can be examined in the marginal table $n_{\cdot jk}$.

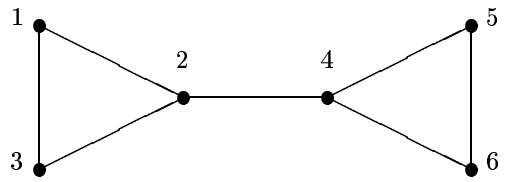
(c) If model [1][23] holds, then the relationship between factors 2 and 3 can be examined in the marginal table $n_{\cdot jk}$.

To extend collapsibility conditions to higher-order tables, use the same tricks as were used in Section 1 for interpreting higher-order models: reindexing and using larger models. Consider a table with five factors: 1, 2, 3, 4, and 5. Suppose our model is [1234][45][35]. This is contained in the model [1234][345]. By considering this as a three-factor table with factors 1-2, 3-4, and 5, we have that factor 1-2 and factor 5 are independent given factor 3-4. Thus, collapsing over factor 5 to examine the marginal table of factors 1, 2, 3, and 4 is valid. Also, collapsing over factors 1 and 2 gives a valid marginal table for examining factors 3, 4, and 5.

It is particularly easy to read off collapsibility from a graphical model.

Corollary 4.3.2. Let the sets A, B, and C denote a partition of the factors in a graphical model such that every chain between a factor in A and a factor in B involves at least one factor in C; then the relationships among the factors in A and C can be examined in the marginal table obtained by summing over the factors in B.

EXAMPLE 4.3.3. The model [123][24][456] can be graphed as below.



It follows that accurate conclusions can be drawn from the marginal tables $n_{123\cdots}$, $n_{1234\cdot\cdot}$, $n_{\cdots 456}$, and $n_{\cdot 2\cdot 456}$.

4.4 Recursive Causal Models

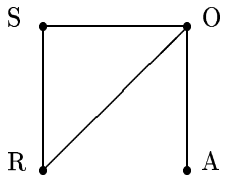
This section examines a class of graphical models that are useful for analyzing causal relationships. Causation is not something that can be established

by data analysis. Establishing causation requires logical arguments that go beyond the realm of numerical manipulation. For example, a well-designed randomized experiment can be the basis for conclusions of causality, but the analysis of an observational study yields information only on correlations. When observational studies are used as a basis for causal inference, the jump from correlation to causation must be made on nonstatistical grounds. In this section, we consider a class of graphical models that have causation built into them. The discussion focuses on appropriate graphs and their interpretations. Not all of the graphical models in this class correspond to log-linear models; thus, the new class is distinct from the graphical models considered in Section 2. For the models considered here, the numerical process of estimation is exceedingly simple.

The graphical models considered in this section are *recursive causal models.* Unlike most of the methods considered in Chapter 4, recursive causal models allow for multiple response factors. With multiple response factors, a given factor can serve as both a response, relative to some causal factors, and as a cause for other response factors. The term "recursive" indicates that response factors are not allowed to serve, even indirectly, as causes of themselves.

We begin with a discussion of models that involve only one response factor. In particular, we consider the abortion opinion data discussed in Sections 3.7 and 4.6.

EXAMPLE 4.4.1. *Abortion Opinion Data.*
The factors involved in the abortion opinion data are race R, sex S, opinion O, and age A. In Chapter 6, one of the better models found for these data is [RSO][OA]. This is a graphical model.



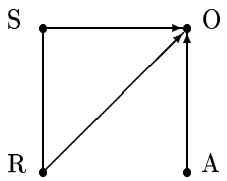
The model [RSO][OA] indicates that Race and Sex are independent of Age given Opinion. While this may explain the data well, it is difficult to imagine a social process that could cause independence between such tangible characteristics as Race, Sex, and Age given something as ephemeral as Opinion. In particular, it violates Asmussen and Edwards' (1983) criteria for response models (see the ends of Sections 4.6 and 6.8).

In Chapter 4, we have argued that when analyzing a response, one should condition on all explanatory factors. With Opinion taken as a response, any log-linear model should include the interaction term [RSA] for the

explanatory factors. In Section 4.6, we found that [RSA][RSO][OA] was a reasonable model. This model is not graphical. For example, the three-factor terms in the model, [RSA] and [RSO], imply the existence of [SA] and [SO] interactions. Taken together with the [OA] term, the model includes all of the two-factor terms included in [SAO]. By definition, if all these two-factor effects are included, a graphical model must also include [SAO]. Thus, [RSA][RSO][OA] is not graphical.

Note that based on the model [RSA][RSO][OA], any logit model for Opinion has an effect (RS) for Race-Sex interaction and a main effect A for age. In the discussion below, we present a recursive causal model that incorporates the same effects. The difference is that the recursive causal model is not a log-linear model but a *conjunction* of log-linear models.

While [RSO][OA] is a graphical model, it is not a recursive causal model for Opinion. Given below is the graph of a recursive causal model in which Opinion is the response, Race, Sex, and Age are direct causes of Opinion, and there is a joint effect for Race and Sex.



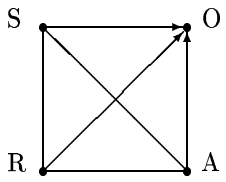
This is very similar to the graph for [RSO][OA]; however, some of the edges have been replaced by arrows. Arrows are called *directed edges*. Edges without arrowheads are *undirected edges*. Directed edges point to response factors; O is the only response factor. In this graph, the directed edges originate at the explanatory factors. The factors R, S, and A are each called a *direct cause* of O because there is a directed edge from each of R, S, and A to O. The undirected edge between R and S is unchanged from the graph of [RSO][OA]; the edge represents an interaction between R and S. Age involves no undirected edges.

With no loss of generality, the probability model corresponding to these four factors can be written as

$$\begin{aligned}&\Pr(R=h,S=i,O=j,A=k)\\&\qquad=\Pr(O=j|R=h,S=i,A=k)\Pr(R=h,S=i,A=k).\end{aligned}$$

Here, the probability is written as the product of a conditional probability of the response factor given its direct causes, $\Pr(O=j|R=h,S=i,A=k)$ and another term, $\Pr(R=h,S=i,A=k)$, that involves only the explanatory factors. Each of these terms is to be modeled with a log-linear model.

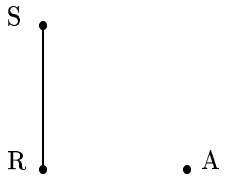
The term $\Pr(O = j|R = h, S = i, A = k)$ is a conditional probability; so, as discussed earlier, the corresponding log-linear model should include an [RSA] term. The log-linear model used is the graphical model that incorporates the explanatory factor edges for [RSA] and changes directed edges involving the four factors to undirected edges. In the following graph, the directed edges are retained to emphasize that there are two steps involved.



Changing directed edges to undirected edges, the graphical log-linear model is clearly the saturated model, cf. Section 2. The maximum likelihood estimate of $\Pr(R = h, S = i, O = j, A = k)$ is $n_{hijk}/n_{....}$ and, thus, the maximum likelihood estimate of $\Pr(O = j|R = h, S = i, A = k) \equiv p_{hijk}/p_{hi\cdot k}$ is

$$\frac{\hat{p}_{hijk}}{\hat{p}_{hi\cdot k}} = \frac{n_{hijk}}{n_{hi\cdot k}}.$$

The probability model for the explanatory factor term $\Pr(R = h, S = i, A = k)$ is also a log-linear model determined by the graph of the recursive causal model. The log-linear model is the graphical model obtained by dropping the response factor and the directed edges. It is given below.



The log-linear model for this graph is [RS][A]. It determines a marginal distribution for the explanatory factors. The model is that

$$\Pr(R = h, S = i, A = k) = \Pr(R = h, S = i)\Pr(A = k).$$

or, equivalently,

$$p_{hi\cdot k} = p_{hi\cdot\cdot}p_{\cdot\cdot\cdot k}.$$

The maximum likelihood estimates are

$$\hat{p}_{hi\cdot k} = \frac{n_{hi\cdot\cdot}}{n_{....}}\frac{n_{\cdot\cdot\cdot k}}{n_{....}}.$$

Combining the two sets of results gives

$$\begin{aligned}&\Pr(R=h,S=i,O=j,A=k)\\&\qquad=\Pr(O=j|R=h,S=i,A=k)\Pr(R=h,S=i)\Pr(A=k)\end{aligned}$$

or, equivalently,

$$p_{hijk}=\frac{p_{hijk}}{p_{hi\cdot k}}\,(p_{hi\cdot\cdot})\,p_{\cdot\cdot\cdot k}\,.$$

This probability model appears to be a saturated log-linear model but is not. Taking logs gives $\log(p_{hijk})$ as the sum of four additive terms, one of which involves all four indices. This would seem to be a saturated log-linear model. However, a two-stage modeling procedure was used, so the simple-minded approach is not appropriate. Using the maximum likelihood estimates from each stage gives

$$\hat{p}_{hijk}=\frac{n_{hijk}}{n_{hi\cdot k}}\frac{n_{hi\cdot\cdot}}{n_{\cdot\cdot\cdot\cdot}}\frac{n_{\cdot\cdot\cdot k}}{n_{\cdot\cdot\cdot\cdot}}$$

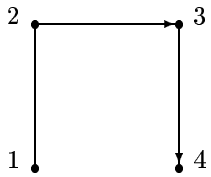
and estimated expected cell counts

$$\hat{m}_{hijk}=n_{\cdot\cdot\cdot\cdot}\frac{n_{hijk}}{n_{hi\cdot k}}\frac{n_{hi\cdot\cdot}}{n_{\cdot\cdot\cdot\cdot}}\frac{n_{\cdot\cdot\cdot k}}{n_{\cdot\cdot\cdot\cdot}}.$$

It is interesting to note that there is no data reduction involved in the estimation. An important, if often underemphasized, element of statistical modeling is that large amounts of data are reduced to manageable form by the use of sufficient statistics. This is certainly true for ANOVA type log-linear models where maximum likelihood estimates are completely determined by various *marginal* tables. The $\hat{m}_{hijk}$'s given above require knowledge of the n_{hijk}'s, so no data reduction has occurred. This is not always true; data reduction does occur for some recursive causal models.

We now consider recursive causal graphs with four factors, of which two are responses.

EXAMPLE 4.4.2. *Two Response Factors.*
Assume multinomial sampling for a table with four factors. Consider the recursive causal graph given below.



Factors 1 and 2 are purely explanatory and interact. Factor 3 has one direct cause, which is factor 2. Factor 4 has one direct cause, factor 3. In general, the probability model for four factors can be written in three terms:

$$\begin{aligned}\Pr(F_1 = h, F_2 = i, F_3 = j, F_4 = k) &= \Pr(F_1 = h, F_2 = i) \\ &\times \Pr(F_3 = j|F_1 = h, F_2 = i)\Pr(F_4 = k|F_1 = h, F_2 = i, F_3 = j).\end{aligned}$$

Based on the graph, we write the recursive causal probability model as

$$\begin{aligned}&\Pr(F_1 = h, F_2 = i, F_3 = j, F_4 = k) \\ &\qquad = \Pr(F_1 = h, F_2 = i)\Pr(F_3 = j|F_2 = i)\Pr(F_4 = k|F_3 = j).\end{aligned}$$

The first term, $\Pr(F_1 = h, F_2 = i)$, involves only the purely explanatory factors. The second term, $\Pr(F_3 = j|F_2 = i)$, involves the distribution for factor 3 given its direct cause, the purely explanatory factor 2. Note that factor 1 is an indirect cause of 3 because of its relationship with factor 2; however, only the direct cause F_2 is involved in the probability model. The third term, $\Pr(F_4 = k|F_3 = j)$, involves the distribution of F_4 given its direct cause F_3.

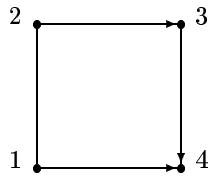
Again, estimation of probabilities is simple. The probability $\Pr(F_1 = h, F_2 = i)$ is estimated using the graphical log-linear model obtained by dropping the response factors 3 and 4 and all directed edges. In other words, it is estimated from the saturated model [12] for the marginal table involving only the first two factors. The term $\Pr(F_3 = j|F_2 = i)$ is estimated using the saturated model for the 2, 3 marginal table. The last term, $\Pr(F_4 = k|F_3 = j)$, is estimated from the saturated model for the 3, 4 marginal table. Thus,

$$\begin{aligned}\hat{p}_{hijk} &= \hat{p}_{hi\cdot\cdot}\frac{\hat{p}_{\cdot ij\cdot}}{\hat{p}_{\cdot i\cdot\cdot}}\frac{\hat{p}_{\cdot\cdot jk}}{\hat{p}_{\cdot\cdot j\cdot}} \\ &= \frac{n_{hi\cdot\cdot}}{n_{\cdot\cdot\cdot\cdot}}\frac{n_{\cdot ij\cdot}}{n_{\cdot i\cdot\cdot}}\frac{n_{\cdot\cdot jk}}{n_{\cdot\cdot j\cdot}}.\end{aligned}$$

Of course, the estimated expected cell counts are simply

$$\hat{m}_{hijk} = n_{\cdot\cdot\cdot\cdot}\hat{p}_{hijk}\,.$$

The graphical model can be made more interesting by inserting a direct cause between 1 and 4.



The recursive causal probability model is now

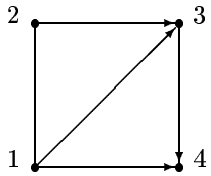
$$\Pr(F_1 = h, F_2 = i, F_3 = j, F_4 = k)$$
$$= \Pr(F_1 = h, F_2 = i)\Pr(F_3 = j|F_2 = i)\Pr(F_4 = k|F_1 = h, F_3 = j).$$

Again, the first term $\Pr(F_1 = h, F_2 = i)$, involves only the purely explanatory factors. The second term, $\Pr(F_3 = j|F_2 = i)$, involves the distribution for factor 3 given its direct cause. The difference between this model and the previous one is that the third term, $\Pr(F_4 = k|F_1 = h, F_3 = j)$, now involves the distribution of F_4 given both of its direct causes F_1 and F_3. Estimation is based on saturated models for appropriate marginal tables and yields

$$\begin{aligned}\hat{p}_{hijk} &= \hat{p}_{hi\cdot\cdot}\frac{\hat{p}_{\cdot ij\cdot}}{\hat{p}_{\cdot i\cdot\cdot}}\frac{\hat{p}_{h\cdot jk}}{\hat{p}_{h\cdot j\cdot}} \\ &= \frac{n_{hi\cdot\cdot}}{n_{\cdot\cdot\cdot\cdot}}\frac{n_{\cdot ij\cdot}}{n_{\cdot i\cdot\cdot}}\frac{n_{h\cdot jk}}{n_{h\cdot j\cdot}}.\end{aligned}$$

Note that a saturated model is used for the explanatory factors only because the graph indicates use of a saturated model. Unlike response factors, the model for explanatory factors is not required to be a saturated model for the appropriate marginal table.

Consider one final graph.



The probability model is

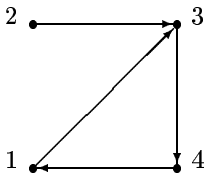
$$\Pr(F_1 = h, F_2 = i, F_3 = j, F_4 = k) = \Pr(F_1 = h, F_2 = i)$$
$$\times \Pr(F_3 = j|F_1 = h, F_2 = i)\Pr(F_4 = k|F_1 = h, F_3 = j).$$

Again, the first term, $\Pr(F_1 = h, F_2 = i)$, involves only the purely explanatory factors. The second term, $\Pr(F_3 = j|F_1 = h, F_2 = i)$, involves the distribution for factor 3 given both of its direct causes. The third term, $\Pr(F_4 = k|F_1 = h, F_3 = j)$ also conditions only on direct causes. Estimation is based on appropriate marginal tables,

$$\hat{p}_{hijk} = \frac{n_{hi\cdot\cdot}}{n_{\cdot\cdot\cdot\cdot}}\frac{n_{hij\cdot}}{n_{hi\cdot\cdot}}\frac{n_{h\cdot jk}}{n_{h\cdot j\cdot}}$$

with estimated expected cell counts $\hat{m}_{hijk} = n_{\cdot\cdot\cdot\cdot}\hat{p}_{hijk}$.

In general, a *causal graph* for a set of factors C includes a set M of purely explanatory factors, also called external or *exogenous* factors and a set $C - M$ of response factors, also called internal or *endogenous* factors. The exogenous factors have an undirected graph associated with them. Each endogenous factor is the end point for one or more directed edges. The directed edges can originate at either exogenous or endogenous factors. A causal graph is *recursive* if no endogenous factor is a cause of itself; in other words, if there are no directed pathways that lead from an endogenous factor back to itself. For example, the causal graph



is not recursive. Factor 2 is exogenous. The other three factors are endogenous. There is a directed path from 3 to 4 to 1 and back to 3, so 3 is a cause of itself and the graph is not recursive. In this example, all of the endogenous factors are causes of themselves.

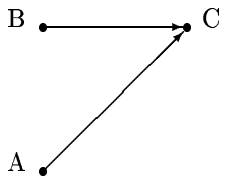
Generally, for a factor $F_i \in C - M$, its *direct causes* are the factors from which a directed edge goes to F_i. Let the set of all such factors be D_i. The probability for a recursive causal model is

$$\Pr(F_i = f_i \,:\, F_i \in C) = \Pr(F_i = f_i \,:\, F_i \in M) \prod_{F_i \in C-M} \Pr(F_i = f_i | D_i)\,.$$

The term $\Pr(F_i = f_i \,:\, F_i \in M)$ depends on the marginal graph for M, i.e., the graph that drops all endogenous factors and directed edges. Estimates of $\Pr(F_i = f_i \,:\, F_i \in M)$ are maximum likelihood estimates from the corresponding graphical log-linear model. Estimation of a term of the form $\Pr(F_i = f_i | D_i)$ is based on the saturated model for the marginal table with factors in $\{F_i\} \cup D_i$.

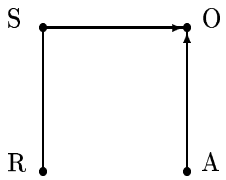
Putting arrowheads on edges and defining a new method for stating probability models has nothing fundamental to do with causation. These probability models can be used even when the causation suggested by the graph exists only in the head of the data analyst. Moreover, if enough models with nonsensical causations are fit, one that fits well may be found. Obviously, a well-fitting model does not establish that the causal patterns in the model are true. However, if the graph is a reasonable statement of a causal process, a well-fitting probability model adds credence to the hypothesized causal process. Evaluating how well recursive causal models fit is discussed later in this section.

A useful and interesting concept in recursive causal models is that of *configuration* $>$. It allows one to relate recursive causal models to decomposable log-linear models, cf. Section 2. Three factors A, B, and C are in configuration $>$ if C is caused by both A and B, but there is neither a directed nor an undirected edge between A and B. This is illustrated below.



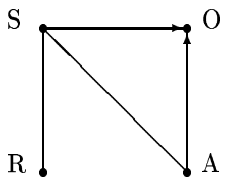
Recursive causal models are typically not log-linear, but there is a substantial intersection between the classes of models. The key result is that *a recursive causal graph contains no factors that are in configuration* $>$ *and the graph restricted to the exogenous variables is decomposable if and only if the recursive causal probability model is identical to the probability model determined by a decomposable log-linear model*, cf. Wermuth and Lauritzen (1983).

EXAMPLE 4.4.3. Consider the recursive causal graph given below. This is similar to one used in Example 4.4.1; however, the factor R has been eliminated as a direct cause for O.



The factors S, A, O are in configuration $>$; thus, the recursive causal graph is not equivalent to a decomposable log-linear model.

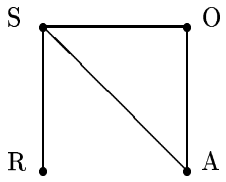
By connecting the nodes for S and A, the configuration $>$ can be eliminated.



The probabilities for this recursive causal graph are the probabilities for the model [RS][SA] in the exogenous marginal table, obtained by collapsing over the response factor O, times the conditional probabilities for $\Pr(O = j|S = i, A = k)$ from the saturated model collapsing over R. This gives

$$p_{hijk} = \left(\frac{p_{hi\cdot\cdot}p_{\cdot i\cdot k}}{p_{\cdot i\cdot\cdot}}\right)\left(\frac{p_{\cdot ijk}}{p_{\cdot i\cdot k}}\right) = \left(\frac{p_{hi\cdot\cdot}p_{\cdot ijk}}{p_{\cdot i\cdot\cdot}}\right).$$

Note that these are exactly the same *probabilities* as determined by the decomposable log-linear model [RS][SOA] defined by the *underlying undirected graph*



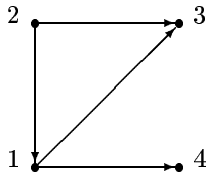
In the underlying undirected graph, directed edges are changed to undirected edges. The probability models are the same, so independence relationships are the same. For the decomposable model, the independence relationship is that R is independent of O and A given S. This holds for both the log-linear model and the recursive causal model.

The probability models associated with decomposable log-linear models are identical to probability models for recursive causal models that (1) have no configurations > and (2) have a decomposable exogenous factor graph. The exogenous factor graph is the graph with all response factors and directed edges eliminated. It follows that decomposable models can be thought of as a subset both of graphical log-linear models and of recursive causal models. A recursive causal model graph with a decomposable exogenous factor graph and no configurations > can be transformed into a decomposable log-linear model graph simply by changing directed edges to undirected edges. See Wermuth and Lauritzen (1983) and Kiiveri, Speed, and Carlin (1984) for the validity of these statements.

EXERCISE 4.3. In Example 4.4.1, it was stated that the model [RSO][OA] is not a recursive causal model for Opinion. However, [RSO][OA] is decomposable, so it corresponds to some recursive causal model. Explain why [RSO][OA] is not a recursive causal model for Opinion, and by changing the endogenous factor, give a recursive causal model that does correspond to [RSO][OA].

We now present a conditional independence result that holds for general recursive causal models. *Any endogenous (response) factor F_i is independent of the factors F_j for which it is not a direct or indirect cause given D_i, the direct causes of F_i.* Independence among exogenous factors is determined by the exogenous factor graph.

EXAMPLE 4.4.4. In the model associated with the following graph, factor 4 is independent of the factors for which it is not a cause, i.e., factors 2 and 3, given factor 1 which is the direct cause of 4.



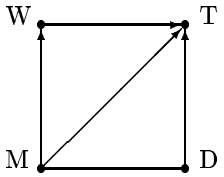
Another independence relation that can be read from the graph is that 3 is independent of 4 given 1 and 2, the direct causes of 3. Note that the graph contains no configurations >, so the independence relations are the same as in the underlying undirected graph. The decomposable model is [41][123] which has 4 independent of 2 and 3 given 1. The decomposable model also implies that 3 is independent of 4 given 1 and 2.

Wermuth and Lauritzen (1983) and Kiiveri, Speed, and Carlin (1984) examine the validity of conditional independence statements for recursive causal models. Another natural application of recursive causal graphs is in the analysis of structural equation models. Interpretations and conditional independence results hold as for the analysis of discrete data; see Kiiveri and Speed (1982).

Birch (1963), Goodman (1973), and Fienberg (1980) have examined methods of model selection that apply to recursive causal models. The methods are illustrated through an example.

EXAMPLE 4.4.5. Suppose muscle tension data similar to that of Examples 3.7.1 and 4.5.1 has been collected. The factors involved are T, the change in muscle tension, W, the weight of the muscle, M, the muscle type

and D, the drug administered. Each factor is at two levels. For ease of exposition, we will treat the sampling as multinomial. The graph below indicates a possible recursive causal scheme.



Change in muscle tension T is hypothesized to have all of W, M, and D as direct causes. Muscle weight W has muscle type M as a direct cause. The purely explanatory factors are M and D; they are allowed to interact with each other. Write $\Pr(T = h, W = i, M = j, D = k) = p_{hijk}$. Note that the indexing has changed from previous examples. The first factor is now at the upper right of the graph rather than the lower left and the order is counterclockwise rather than clockwise. *These changes are important for verifying the maximum likelihood estimates presented.* The expected cell counts for the recursive causal model are

$$\hat{m}_{hijk} = n_{....} \left(\frac{n_{..jk}}{n_{....}}\right) \left(\frac{n_{.ij.}}{n_{..j.}}\right) \left(\frac{n_{hijk}}{n_{.ijk}}\right).$$

The graph given above has one configuration > involving W, D, and T, so the probability model is not a decomposable log-linear model.

The lack of fit of the model can be tested in the usual way. Some algebra shows that

$$\begin{aligned} G^2 &= 2\sum_{hijk} n_{hijk} \log\left(\frac{n_{hijk}}{\hat{m}_{hijk}}\right) \\ &= 2\sum_{ijk} n_{.ijk} \log\left(\frac{n_{.ijk}}{n_{..jk} n_{.ij.}/n_{..j.}}\right). \end{aligned}$$

This is precisely the lack of fit statistic for testing the log-linear model [WM][MD] against the saturated model in the marginal table for W, M, and D. With each factor at two levels, it follows that the statistic has 2 degrees of freedom. In fact, the log-linear model [WM][MD] is consistent with the only conditional independence result available from the graph; the response factor W is independent of D given M, the direct cause of W.

The relationship of the lack of fit test with the log-linear model [WM][MD] can be seen through the probability modeling procedure. Recall that the basis of the modeling procedure is that one can always write

$$\Pr(T = h, W = i, M = j, D = k) = \Pr(M = j, D = k)$$

$$\times \Pr(W = i|M = j, D = k)\Pr(T = h|W = i, M = j, D = k).$$

and, based on the graph, the recursive causal probability model is

$$\begin{aligned}\Pr(T = h, W = i, M = j, D = k) &= \Pr(M = j, D = k) \\ &\times \Pr(W = i|M = j)\Pr(T = h|W = i, M = j, D = k).\end{aligned}$$

The only real modeling being done is replacing $\Pr(W = i|M = j, D = k)$ with $\Pr(W = i|M = j)$. The factor T is extraneous. The perfectly general statement

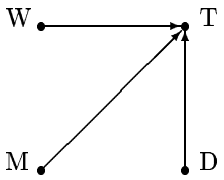
$$\Pr(W = i, M = j, D = k) = \Pr(M = j, D = k)\Pr(W = i|M = j, D = k)$$

has been replaced with

$$\begin{aligned}&\Pr(W = i, M = j, D = k) \\ &= \Pr(M = j, D = k)\Pr(W = i|M = j) \\ &= \Pr(M = j)\Pr(D = k|M = j)\Pr(W = i|M = j).\end{aligned}$$

This is just the model for conditional independence of W and D given M. It is not surprising that the test statistic involves only the aspect of the model that is not always true.

There is no particular reason to believe in a relationship between the type of drug used and the muscle type. It is also questionable whether muscle type really has an effect on weight. The graph that incorporates these ideas involves dropping one directed edge and the undirected edge. It is given below.



Note that by dropping M as a direct cause of W, W has been transformed into an exogenous factor. The estimated expected cell counts are

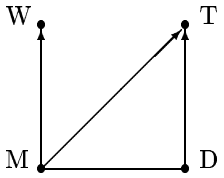
$$\hat{m}_{hijk} = n_{....}\left(\frac{n_{.i..}}{n_{....}}\frac{n_{..j.}}{n_{....}}\frac{n_{...k}}{n_{....}}\right)\left(\frac{n_{hijk}}{n_{.ijk}}\right).$$

This model can be checked for general lack of fit as above. It is not difficult to see that the test is identical to that for complete independence [W][M][D] in the marginal table collapsing over T. The likelihood ratio chi-squared has 4 degrees of freedom.

This new model was obtained from the previous one by dropping edges in the previous graph; thus, this model is a reduced model relative to the previous one.

It follows that a likelihood ratio test can be performed for comparing the two models. Not surprisingly, the test simplifies to that of [W][M][D] versus [WM][MD].

Another possible model eliminates the effect of muscle weight W on tension.

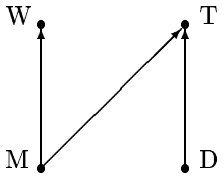


The maximum likelihood estimates are

$$\hat{m}_{hijk} = n_{....} \left(\frac{n_{..jk}}{n_{....}}\right) \left(\frac{n_{.ij.}}{n_{..j.}}\right) \left(\frac{n_{h.jk}}{n_{..jk}}\right).$$

The graph contains no configurations >, so the probability model and thus the maximum likelihood estimates are the same as for the decomposable log-linear model [WM][MTD] determined by the underlying undirected graph.

Consider one final model.



This incorporates independence of M and D from the exogenous factor graph and it has one conditional independence relation involving responses: W independent of T and D given M. These two relationships imply that the pair W and M are independent of D. The maximum likelihood estimates are

$$\hat{m}_{hijk} = n_{....} \left(\frac{n_{..j.}}{n_{....}}\right) \left(\frac{n_{...k}}{n_{....}}\right) \left(\frac{n_{.ij.}}{n_{..j.}}\right) \left(\frac{n_{h.jk}}{n_{..jk}}\right).$$

The likelihood ratio test statistic for this model can be separated into the sum of three terms. The first term is the statistic for testing [M][D]

versus [MD]. The second term is the statistic for testing [WM][MD] versus [WMD]. The last term is for testing [TMD][WMD] versus [TWMD]. The following series of equalities establishes the result.

$$
\begin{aligned}
G^2 &= 2\sum_{hijk} n_{hijk}\log\left(\frac{n_{hijk}}{\hat{m}_{hijk}}\right) \\
&= 2\sum_{hijk} n_{hijk}\left[\log(n_{hijk}) - \log(\hat{m}_{hijk})\right] \\
&= 2\sum_{hijk} n_{hijk}\log(n_{hijk}) - 2\sum_{hijk} n_{hijk}\log\left(\frac{n_{..j.}n_{...k}}{n_{....}}\right) \\
&\quad - 2\sum_{hijk} n_{hijk}\log\left(\frac{n_{.ij.}}{n_{..j.}}\right) - 2\sum_{hijk} n_{hijk}\log\left(\frac{n_{h.jk}}{n_{..jk}}\right) \\
&= 2\sum_{hijk} n_{hijk}\log\left(\frac{n_{hijk}\,n_{.ijk}}{n_{.ijk}\,n_{..jk}}n_{..jk}\right) \\
&\quad - 2\sum_{hijk} n_{hijk}\log\left(\frac{n_{..j.}n_{...k}}{n_{....}}\right) \\
&\quad - 2\sum_{hijk} n_{hijk}\log\left(\frac{n_{.ij.}}{n_{..j.}}\right) - 2\sum_{hijk} n_{hijk}\log\left(\frac{n_{h.jk}}{n_{..jk}}\right) \\
&= 2\sum_{hijk}\left[n_{hijk}\log(n_{..jk}) - n_{hijk}\log\left(\frac{n_{..j.}n_{...k}}{n_{....}}\right)\right] \\
&\quad + 2\sum_{hijk}\left[n_{hijk}\log\left(\frac{n_{.ijk}}{n_{..jk}}\right) - n_{hijk}\log\left(\frac{n_{.ij.}}{n_{..j.}}\right)\right] \\
&\quad + 2\sum_{hijk}\left[n_{hijk}\log\left(\frac{n_{hijk}}{n_{.ijk}}\right) - n_{hijk}\log\left(\frac{n_{h.jk}}{n_{..jk}}\right)\right] \\
&= 2\sum_{hijk} n_{hijk}\left[\log(n_{..jk}) - \log\left(\frac{n_{..j.}n_{...k}}{n_{....}}\right)\right] \\
&\quad + 2\sum_{hijk} n_{hijk}\left[\log\left(n_{.ijk}\right) - \log\left(\frac{n_{.ij.}n_{..jk}}{n_{..j.}}\right)\right] \\
&\quad + 2\sum_{hijk} n_{hijk}\left[\log(n_{hijk}) - \log\left(\frac{n_{h.jk}n_{.ijk}}{n_{..jk}}\right)\right] \\
&= 2\sum_{jk} n_{..jk}\left[\log\left(n_{..jk}\right) - \log\left(\frac{n_{..j.}n_{...k}}{n_{....}}\right)\right] \\
&\quad + 2\sum_{ijk} n_{.ijk}\left[\log\left(n_{.ijk}\right) - \log\left(\frac{n_{..jk}n_{.ij.}}{n_{..j.}}\right)\right]
\end{aligned}
$$

$$+ 2 \sum_{hijk} n_{hijk} \left[\log(n_{hijk}) - \log\left(\frac{n_{\cdot ijk} n_{h \cdot jk}}{n_{\cdot \cdot jk}} \right) \right].$$

The three terms in the last equality are precisely the three test statistics that were claimed. The existence of such breakdowns for G^2 is quite general.

4.5 Exercises

EXERCISE 4.5.1. Using the methods of Section 4.1, discuss the independence relationships for all of the models given below.

(a) [123][24][456]

(b) [12][13][23][24][456]

(c) [123][124][456]

(d) [123][24][456][15]

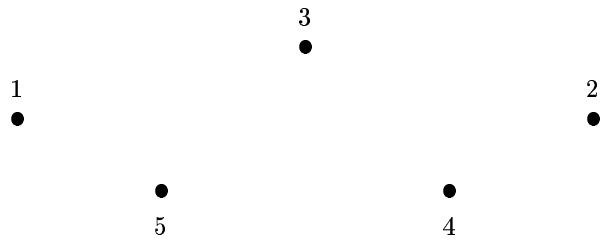
(e) [123][24][456][15][36]

EXERCISE 4.5.2. In the saturated log-linear model for a four-dimensional table, let $u_{34} = 0$ and let all of the corresponding higher-order terms also be zero, e.g., $u_{134} = 0$.

(a) Based on this model, find a formula for $\hat{m}_{hijk}$ without using graphical methods.

(b) Use graphical methods to find $\hat{m}_{hijk}$.

EXERCISE 4.5.3. The vertices for a five-factor model are given below. Connect the dots to give a graphical representation of the model [123][135][34][24]. Use the illustration to show that [123][135][34][24][25] is not a graphical model.



EXERCISE 4.5.4. Which of the models given below are graphical? Graph them. Which of these are decomposable? Discuss the independence relationships for all of the models. For each model, what marginal tables will provide valid inferences?

(a) [123][24][456]

(b) [12][13][23][24][456]

(c) [123][124][456]

(d) [123][24][456][15]

(e) [123][24][456][15][36]

EXERCISE 4.5.5. Consider all of the graphs in Example 4.4.2. Classify each as equivalent or not equivalent to a decomposable log-linear model. For those that are equivalent, prove the equivalence of the probability models.