Chapter 6

Model Selection Methods and Model Evaluation

This chapter examines methods of selecting models for high-dimensional tables. The model selection methods considered are the stepwise methods, e.g., forward selection and backward elimination, a modified backward elimination method from Aitkin (1978, 1979) that controls the experimentwise error rate, and a backward elimination method from Wermuth (1976) that is restricted to decomposable models. In addition, we discuss the use of the model selection criteria presented in Section 3.6. Of course, it would be foolish to choose a model simply because some model selection procedure presents it to you as a good model. Other considerations such as model interpretability and the consistency of the data with model assumptions may dictate choosing some other model. It is always wise to *use model selection methods to produce several apparently good models* that can be investigated further. In line with this approach, the analysis of residuals and influential observations is also discussed in this chapter.

Modeling is a useful process both for prediction of future observables and for describing the relationships between factors. Large models always reproduce the data on which they were fitted better than smaller models. The saturated model always provides a perfect fit of the data. However, smaller models have more powerful interpretations and are often better predictive tools than large models. *Often, our goal is to find the smallest model that fits the data.*

In model fitting, there are two approaches to specifying models: the descriptive approach and the causative approach. The descriptive approach simply describes the relationships that are observed. For example, in a three-way table, one might find that given a young child's educational sta-

tus, her father's educational status and her mother's educational status are independent. This can be used to describe the data, but it would be foolish to suggest that the child's status in any way determines her parents' status. In analyzing these factors, it makes sense to consider the parents' status as fixed and to determine its effect on the child's status. However, a statistical relationship between parents' status and child's status does not imply causation. Causation cannot be inferred on statistical grounds; it must be inferred from the subject matter. For these reasons, we will concentrate on descriptive modeling. Nonetheless, causation has important implications for statistical modeling. Causation is closely related to the existence of response factors; the analysis of response factors is treated in Chapter 4, Section 5.4, and is also discussed in Chapter 11.

6.1 Stepwise Procedures for Model Selection

Stepwise procedures assume an initial model and then use rules for adding or deleting terms to arrive at a final model. Stepwise procedures are categorized in three ways: *forward selection*, in which terms are added to an initial small model; *backward elimination*, in which terms are removed from an initial large model; and *composite methods*, in which terms can either be added to or removed from the initial model. Methods for choosing initial models and examples will be considered in the subsequent two sections.

Because of the huge number of terms available in high-dimensional models, more effort is expended in selecting an initial model than is commonly used in regression analysis. Moreover, because ANOVA type models use parameters that are not uniquely defined, stepwise procedures must be adjusted so that nonsensical models are not considered.

Often, at any given point, stepwise procedures are applied only to examine terms that involve the same number of factors. For example, sometimes in examining three-factor interactions, two-factor interactions are ignored and any three-factor interactions that are implied by the existence of higher-order interactions are forced into the model. (In this case, higher-order interactions are those that involve four or more factors.) We begin by considering this particular procedure. Later, an improved method (implemented in BMDP) will be discussed.

Stepwise procedures are sequential in that they assume a current model and look to add or delete terms one at a time to that model. When considering s-factor terms, the basic forward selection rule is

FS: (a) Add the s-factor term not already in the model that has the most significant test statistic.

(b) Continue adding terms until no term achieves a predetermined minimum level of significance.

The basic backward elimination rule is

BE: (a) Delete the s-factor term with the least significant test statistic among s-factor terms that are not forced into the model.

(b) Continue until all terms maintain a predetermined minimum level of significance.

The backward elimination procedure is based on comparing models and does not consider whether the reduced models fit relative to the saturated model. It is possible that a model may fit globally and that dropping an s-factor term may be acceptable but that the new smaller model may not fit globally. One might want to modify the procedure so that it stops before eliminating any effect that will cause the saturated model test to be rejected.

Note that a term can be forced into the model either by the sampling scheme or by the presence in the model of a higher-order term that implies the existence of the term in question. For example, in the model [1235][234], when considering three-factor terms for elimination, all of [123], [125], [135], and [235] are forced into the model by having [1235] in the model. The only three-factor term eligible for elimination is [234].

The composite method alternates between applying the forward selection rule and the backward elimination rule. For forward selection, the test statistics are the statistics for testing the current model against the larger models in which one additional term has been added. For backward elimination, the test statistics are the statistics for testing the current model against the reduced models in which one term has been eliminated. "Significance" of test statistics is measured by their P *values*. A test statistic fails to achieve a predetermined minimum level of significance, say α, if $P > \alpha$ and maintains that level of significance if $P < \alpha$. The level α is often taken as .10, .05, or .01.

As an alternative to considering only s-factor terms, one can consider adding or deleting either simple or multiple effects. *Adding a simple effect consists of adding an effect that does not imply the simultaneous addition of any other effects.* For example, if we have four factors, say R, S, O, and A, and the model is [RSO][SA], the only simple effects that could be added are [RA] and [OA]. To see this, note two things. First, all other two-factor terms are already in the model, so these are the only two-factor terms that can be added. Second, to add any three-factor terms, e.g., [RSA], also implies the addition of a new two-factor term. Therefore, adding any three-factor term implies the addition of more than one effect and thus is not the addition of a simple effect.

If we consider the deletion of simple effects from [RSO][SA], the only possible deletions are the [RSO] and [SA] terms. Deleting the [RSO] term leaves the model [RS][RO][SO][SA]. Deleting the SA effect leaves [RSO][A].

Addition of a multiple effect involves incorporating a new factor into some effect that already exists in the model. For the model [RSO][SA], possible multiple effects are constructed by adding A to [RSO] (giving [RSOA][SA]), by adding R to [SA] (giving [RSO][RSA]), and by adding O to [SA] (giving [RSO][OSA]). In addition, the term [A] is implicitly in the model, so the terms [RA] and [OA] can be added, giving [RSO][SA][RA] and [RSO][SA][OA], respectively. Also, the term [RO] is implicitly in the model, so the term [ROA] can be added, yielding the model [RSO][SA][ROA].

In forward selection, considering either addition of simple effects or addition of multiple effects is appropriate. Because addition of multiple effects involves consideration of a wider variety of additional effects, addition of multiple effects is generally preferred. In backward elimination, deletion of simple effects is the appropriate procedure.

As defined here, deleting multiple effects is the same procedure as deleting simple effects. For the model [RSO][SA], deletion of any factor from [RSO] leaves all of the implicit terms [RS], [SO], and [RO] unaffected. Thus, deletion of multiple effects allows consideration of only the models [RS][SO][RO][OA] and [RSO][A]. These are the same models considered in the deletion of simple effects. There is a key difference between adding and deleting multiple effects. Adding multiple effects to [RSO][SA] allows addition of interesting nonsimple effects like [ROA] because [RO] is implicitly in the model. Deleting a factor from an implicit term such as [RO] has absolutely no effect on the model [RSO][SA]. Other definitions of what it means to delete a multiple effect are possible, but the ones I am acquainted with can give stupid results. To be safe, when using computer software, one should always specify deletion of simple effects.

One reasonable approach to backward elimination in, say, a five-factor model is to eliminate first the five-factor effect if possible, then any unnecessary four-factor effects. When eliminating three-factor effects, restrict attention only to those three-factor effects not forced into the model by the included four-factor effects. Similarly, only consider for elimination two-factor effects that are not forced in by the included three- and four-factor effects. Also, *any effects forced into the model by the sampling scheme should never be considered for elimination.*

As is well known from regression analysis, the main virtue of stepwise methods is that they are fast and cheap. Their virtue is directly related to their fault. They are fast and cheap because the procedures put severe limits on the number of models that are considered. Because only a limited number of models are examined, the procedures can easily miss the best models. Stepwise methods do not give the best model based on any overall criteria of model fit (cf. Section 6); in fact, they can give models that contain none of the terms that are in the best models. Forward selection is a notoriously bad method of variable selection because it starts from an inadequate model and there is no guarantee that it will ever arrive at an adequate model. Backward elimination should give an adequate model

if the initial model is adequate, but the only way to ensure an adequate initial model is to use the saturated model. Combined methods improve on forward selection simply because they allow consideration of more models. However, combined methods do not ensure finding the best models either. A nontechnical problem with stepwise procedures is that they give a unique "best" answer. Typically, no uniquely correct model exists. *If stepwise methods are to be used, it is wise to use several variations and therefore arrive at several candidate models. These models should be evaluated on their interpretability and their consistency with model assumptions to arrive at one or more final models.*

6.2 Initial Models for Selection Methods

In this section, we discuss a variety of methods for arriving at an appropriate initial model from which to begin the search for a well-fitting model. *The examples in this section deal only with initial model selection.* An example incorporating various stepwise procedures is given in Section 3. This section examines three approaches to picking an initial model: all s-factor effects models, models based on tests of marginal and partial association, and models based on testing each term in the saturated model last.

6.2.1 All s-Factor Effects

The simplest way to choose an initial model is to take one that consists of all effects of a particular level s. For example, the initial model can be the model of all main effects, or all two-factor effects, or all three-factor effects, and so on. The initial model can be chosen as either the smallest of these models that fits the data or the largest of these models that does not fit the data. In particular, for a four-factor table, one can test the models

(a) [1][2][3][4]

(b) [12][13][14][23][24][34]

(c) [123][124][234][134]

(d) [1234]

against each other to determine the smallest model that fits the data. Suppose it is model (b). We can then consider eliminating terms from model (b). Another approach is to look at the largest model that does not fit the data. That would be model (a). We can then consider selecting terms to add to model (a). Elimination and selection can be performed either by ad hoc methods or by using the formal rules for backward elimination and forward selection.

Note that if we begin with model (b) [model (a)], it is very tempting to restrict attention to deleting (adding) only two-factor terms. Considering only two-factor terms is a substantial reduction in work as compared to investigating all levels of terms, but this simplification runs the risk of both missing some important terms and leaving in some unimportant terms.

Finally, it should be noted that if a combined stepwise procedure is to be used, either of the initial models is appropriate. However, different initial models may give different results.

EXAMPLE 6.2.1. Reconsider the data of Examples 3.7.1 and 4.5.1 on the relationship between two drugs and muscle tension. For each mouse, a muscle was identified and its tension was measured. A randomly chosen drug was given to the mouse and the muscle tension was measured again. The muscle was then tested to identify which type of muscle it was. The weight of the muscle was also measured. Factors and levels are tabulated below.

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

The sampling is product multinomial with the total count for each muscle type fixed. The data are

			Drug	
Tension	Weight	Muscle	Drug 1	Drug 2
	High	Type 1	3	21
		Type 2	23	11
High				
	Low	Type 1	22	32
		Type 2	4	12
	High	Type 1	3	10
		Type 2	41	21
Low				
	Low	Type 1	45	23
		Type 2	6	22

The test statistics are given below. Clearly, the only model that fits the data is the model of all three-factor interactions.

Model	*df*	G^2	*P*
[TWM][TWD][TMD][WMD]	1	0.11	.74
[TW][TM][WM][TD][WD][MD]	5	47.67	.00
[T][W][M][D]	11	127.4	.00

6.2.2 Examining Each Term Individually

An intuitively appealing method of selecting an initial model is to examine each term in the saturated model and include only those terms that are important. The question arises as to how one decides which terms are important. One reasonable approach is testing whether the terms are nonzero. One can then include the terms that are significantly different from zero and drop the rest. Unfortunately, the saturated model is overparametrized to the point that any terms except the u_{1234}'s can be dropped without affecting the model. The problem is in determining how to test whether the terms are zero. There are many possibilities. For example, to test whether $u_{123(hij)}$ is important in a four-dimensional table, we can test [12][13][234][134] versus [123][234][134] or we can test [234] versus [123][234] or any of a very large number of other model comparisons. The problem is to decide on which tests to examine.

Two methods have been proposed: testing each term last and the method suggested by Brown (1976) in which tests of marginal and partial association are performed for each term. These methods are examined in the next two subsections.

6.2.3 Tests of Marginal and Partial Association

Brown (1976) proposed looking at two tests for each term in the saturated model: a test of marginal association and a test of partial association. These tests can be used in a variety of ways to choose an initial model.

To test a particular term for *marginal association*, collapse over any factors not included in the term. The test of marginal association is based on the marginal table and consists of testing the model that involves only the term in question against the largest submodel that does not include the term. (Main effects are not tested for marginal association.)

Example 6.2.2. The test of marginal association for the $u_{1234(hijk)}$'s is the test of [123][124][134][234] versus [1234]. The test of marginal association for the $u_{123(hij)}$'s is the test of [12][23][13] versus [123]. The test of marginal association for the $u_{24(ik)}$'s is the test of [2][4] versus [24].

The test for *partial association* depends on the number of factors involved in the term. If the term involves s factors, the test of partial association is a test of the model with all s-factor (interaction) terms against the reduced model in which the term in question is dropped out. Thus, in a test of partial association, all other effects are fixed at a certain level of interaction.

Example 6.2.3. In a four-dimensional table, the test of partial association for the $u_{123(hij)}$'s is the test of [124][134][234] versus [123][124][134][234]. The test of partial association for the $u_{24(ik)}$'s is

the test of [12][13][14][23][34] versus [12][13][14][23][24][34]. In a four-dimensional table, the test of partial association for u_{1234} is identical to the test of marginal association.

Note that the degrees of freedom for the tests are the degrees of freedom for dropping the term in question; they are typically the same in the two tests.

There are a number of ways of choosing an initial model using Brown's tests: (a) include all terms with significant marginal tests, (b) include all terms with significant partial tests, (c) include all terms for which either the marginal or partial test is significant, (d) include all terms for which both the marginal and partial tests are significant.

Method (d) always gives the smallest model. Method (c) always gives the largest model. Method (d) can be used to determine an initial model for forward selection. Method (c) determines a model that might be used with backward elimination. Any of the four methods would give an appropriate initial model for combined stepwise selection.

An obvious ad hoc model selection approach is to restrict attention to models that are between the small model of method (d) and the large model of method (c). Perhaps the main fault with this method is that important terms could have been missed in model (c).

EXAMPLE 6.2.4. Brown's tests for the muscle tension data are presented in Table 6.1. The WMD term is clearly significant as is the WM term. In addition, several terms involving the change in muscle tension appear to be important, e.g., T, TM, TD, and possibly TMD.

Using significance levels of $\alpha = .01$ and $\alpha = .10$, the four initial models suggested by Brown's tests are

	$\alpha = .01$	$\alpha = .10$
Method (a):	[WMD][TMD]	[WMD][TMD][TWM]
Method (b):	[WMD][TD][TM]	[WMD][TMD]
Method (c):	[WMD][TMD]	[WMD][TMD][TWM]
Method (d):	[WMD][TD]	[WMD][TMD].

6.2.4 Testing Each Term Last

The basis of this method is testing whether each term can be dropped from the saturated model without a significant loss of explanatory power. The problem with this method is that it requires a reparametrization of the model. For example, the model $\log(m_{hijk}) = u_{24(ik)} + u_{1234(hijk)}$ is a saturated model, but if we drop the $u_{24(ik)}$'s, we get $\log(m_{hijk}) = u_{1234(hijk)}$ which is still a saturated model. Dropping the $u_{24(ik)}$'s does not change the

TABLE 6.1. Brown's Tests for the Muscle Tension Data

Effect	Partial Association G^2	P	Marginal Association G^2	P
T	6.04	.01	—	—
W	3.55	.06	—	—
M	1.18	.28	—	—
D	0.08	.78	—	—
TW	2.35	.13	0.06	.80
TM	6.81	.01	5.27	.02
WM	63.66	.00	62.25	.00
TD	6.02	.01	6.37	.01
WD	0.65	.42	1.12	.29
MD	0.17	.68	1.40	.24
TWM	1.00	.32	2.63	.10
TWD	0.01	.93	0.04	.85
TMD	2.86	.09	6.01	.01
WMD	35.65	.00	40.49	.00
TWMD	0.14	.70	0.14	.70

model. To test every term against the saturated model requires a regression parametrization in which dropping any term really reduces the model.

We begin with a simple example that assumes familiarity with estimation for analysis of variance under the "usual" constraints. After the example, we deal with the question of reparametrization. The discussion of reparametrization involves a more sophisticated use of linear model ideas than has been used thus far in the book.

EXAMPLE 6.2.5. Consider again the muscle-tension data of Example 6.2.1. This involves four factors each at two levels. We begin by examining a similar normal theory ANOVA model

$$\begin{aligned} y_{hijk} = \; & \mu + \alpha_h + \beta_i + \gamma_j + \eta_k \\ & + (\alpha\beta)_{hi} + (\alpha\gamma)_{hj} + (\alpha\eta)_{hk} \\ & + (\beta\gamma)_{ij} + (\beta\eta)_{ik} + (\gamma\eta)_{jk} \\ & + (\alpha\beta\gamma)_{hij} + (\alpha\beta\eta)_{hik} + (\alpha\gamma\eta)_{hjk} \\ & + (\beta\gamma\eta)_{ijk} + (\alpha\beta\gamma\eta)_{hijk} + e_{hijk} \; . \end{aligned}$$

With two levels in each factor, every interaction has one degree of freedom and corresponds to a contrast. For example, under the "usual" side conditions, the $(\alpha\beta)$ interaction contrast is

$$(\alpha\beta)_{11} - (\alpha\beta)_{12} - (\alpha\beta)_{21} + (\alpha\beta)_{22} \; .$$

The estimate of the contrast is

$$\bar{y}_{11\cdot\cdot} - \bar{y}_{12\cdot\cdot} - \bar{y}_{21\cdot\cdot} + \bar{y}_{22\cdot\cdot} \; .$$

Log-linear model estimation is analogous to the ANOVA procedure.

We are dealing with a saturated model

$$\log(m_{hijk}) = u + u_{1(h)} + \cdots + u_{234(ijk)} + u_{1234(hijk)},$$

so $\hat{m}_{hijk} = n_{hijk}$ for all h, i, j, and k. Define new parameters λ corresponding to each interaction contrast. For example, the u_{12} interaction corresponds to a contrast

$$4\lambda_{12} = u_{12(11)} - u_{12(12)} - u_{12(21)} + u_{12(22)}$$

or, equivalently,

$$16\lambda_{12} = 4[u_{12(11)} - u_{12(12)} - u_{12(21)} + u_{12(22)}].$$

This particular definition of the λ's relates to two things that will be examined later. One is ease of computation of the estimates; the other is a useful reparametrization of the saturated model. Let $w_{hijk} = \log(n_{hijk})$. The estimated contrast is

$$\begin{aligned} 16\hat{\lambda}_{12} &= 4[\bar{w}_{11..} - \bar{w}_{12..} - \bar{w}_{21..} + \bar{w}_{22..}] \\ &= [w_{11..} - w_{12..} - w_{21..} + w_{22..}] \, . \end{aligned}$$

Applying the usual side conditions, $u_{12(h\cdot)} = u_{12(\cdot i)} = 0$, leads to the parameter estimates

$$\frac{\hat{\lambda}_{12}}{4} = \hat{u}_{12(11)} = -\hat{u}_{12(12)} = -\hat{u}_{12(21)} = \hat{u}_{12(22)}.$$

Obviously, if you know one of the $\hat{u}_{12(hi)}$'s, you know them all. It is simpler to focus on $\hat{\lambda}_{12}$.

The estimates of all the $\hat{\lambda}$'s are 1/16th of the sums of 8 w_{hijk}'s minus the sum of the remaining 8 w_{hijk}'s, so all $\hat{\lambda}$'s have the same asymptotic standard error

$$\mathrm{SE}(\hat{\lambda}) = \left[\frac{1}{16}\sum_{hijk}\frac{1}{n_{hijk}}\right]^{1/2}.$$

The standard error depends on having a saturated model and is a generalization of the result for log odds ratios given earlier. Details are given in Section 10.2.

If we do an analysis of variance on the w_{hijk}'s, the sums of squares for various terms equal $16\hat{\lambda}^2$. Table 6.2 shows an analysis of variance. Table 6.3 gives values of $|16\hat{\lambda}|$ and $|z| = |\hat{\lambda}/\mathrm{SE}(\hat{\lambda})| = |16\hat{\lambda}/\mathrm{SE}(16\hat{\lambda})|$. The z values can be used to test $\lambda = 0$. The estimates in Table 6.3 were obtained from Table 6.2. For example, the source T has a sum of squares of .2208, so $|16\hat{\lambda}_T| = \sqrt{(16).2208}$. The standard error is $\mathrm{SE}(16\hat{\lambda}_T) = \sqrt{\sum(1/n_{hijk})}$.

TABLE 6.2. Analysis of Variance on $\log(n_{hijk})$ for the Muscle Tension Data

Source	*df*	SS
T	1	.2208
W	1	.3652
M	1	.0000
D	1	.9238
TW	1	.0522
TM	1	.4202
WM	1	5.631
TD	1	.1441
WD	1	.0080
MD	1	.2167
TWM	1	.1123
TWD	1	.0018
TMD	1	.2645
WMD	1	3.286
TWMD	1	.01188

TABLE 6.3. Muscle Tension Data: Estimates and Test Statistics for Model (2)

λ	$\lvert 16\hat{\lambda}\rvert$	$\lvert z\rvert$
T	1.880	1.44
W	2.417	1.85
M	0.002	0.00
D	3.846	2.94
TW	0.914	0.70
TM	2.593	1.98
WM	9.492	7.26
TD	1.518	1.16
WD	0.358	0.27
MD	1.862	1.42
TWM	1.340	1.03
TWD	0.172	0.13
TMD	2.057	1.57
WMD	7.251	5.55
TWMD	0.436	0.33

$\mathrm{SE}(16\hat{\lambda}) = 1.307$.

The main reason for using estimates of 16λ rather than λ is that the 16λ estimates are more comparable to another reparametrization of the saturated model that will be used later.

The important terms in Table 6.3 are λ_{WMD}, λ_{WM}, and perhaps λ_D. In other words, the main effect for D, the WM interaction, and the WMD interaction are the important terms in the model. By our rule for including lower-order terms, the inclusion of λ_{WMD} implies the model [WMD] which automatically includes both [WM] and [D]. It is interesting to note that the factor T does not appear in any important terms.

As mentioned at the beginning of the subsection, testing each term last requires that the saturated model be reparametrized into a regression model. A method is needed for relating the reparametrized results back to the original parametrization. This is most easily done when each factor is at only two levels. If each factor is at only two levels, there is one degree of freedom for each u term. Still, there are an infinite number of possible parametrizations. It is necessary to (arbitrarily) choose one.

EXAMPLE 6.2.6. Consider a $2 \times 2 \times 2 \times 2$ table. The model

$$\begin{aligned}\log(m_{hijk}) &= u + u_{1(h)} + u_{2(i)} + u_{3(j)} + u_{4(k)} \\ &\quad + u_{12(hi)} + u_{13(hj)} + u_{14(hk)} + u_{23(ij)} + u_{24(ik)} + u_{34(jk)} \\ &\quad + u_{123(hij)} + u_{124(hik)} + u_{134(hjk)} + u_{234(ijk)} \\ &\quad + u_{1234(hijk)}\end{aligned} \tag{1}$$

can be reparametrized as

$$\begin{aligned}\log(m_{hijk}) &= \lambda + (-1)^{h-1}\lambda_1 + (-1)^{i-1}\lambda_2 + (-1)^{j-1}\lambda_3 + (-1)^{k-1}\lambda_4 \\ &\quad + (-1)^{h+i-2}\lambda_{12} + (-1)^{h+j-2}\lambda_{13} + (-1)^{h+k-2}\lambda_{14} \\ &\quad + (-1)^{i+j-2}\lambda_{23} + (-1)^{i+k-2}\lambda_{24} + (-1)^{j+k-2}\lambda_{34} \\ &\quad + (-1)^{h+i+j-3}\lambda_{123} + (-1)^{h+i+k-3}\lambda_{124} \\ &\quad + (-1)^{h+j+k-3}\lambda_{134} + (-1)^{i+j+k-3}\lambda_{234} \\ &\quad + (-1)^{h+i+j+k-4}\lambda_{1234}\,.\end{aligned} \tag{2}$$

This parametrization gives the same estimates as using the "usual" side conditions, i.e., $0 = u_{1(\cdot)} = u_{2(\cdot)} = u_{3(\cdot)} = u_{4(\cdot)} = u_{12(\cdot i)} = u_{12(h\cdot)} = \cdots = u_{1234(\cdot ijk)} = u_{1234(h\cdot jk)} = u_{1234(hi\cdot k)} = u_{1234(hij\cdot)}$. Model (2) is given in matrix form in Example 10.4.1.

Other sets of side conditions correspond to other reparametrizations. For example, another frequently used set of side conditions are $0 = u_{1(1)} = u_{2(1)} = u_{3(1)} = u_{4(1)} = u_{12(1i)} = u_{12(h1)} = \cdots = u_{1234(1ijk)} = u_{1234(h1jk)} = u_{1234(hi1k)} = u_{1234(hij1)}$. Here, all u terms are set equal to zero for which any of h, i, j, or k is 1. If we let $\delta_{ab} = 1$ when $a = b$ and 0 otherwise

where a and b are any symbols, these side conditions correspond to the reparametrized model

$$\begin{aligned}\log(m_{hijk}) = {} & \gamma + \delta_{h2}\gamma_1 + \delta_{i2}\gamma_2 + \delta_{j2}\gamma_3 + \delta_{k2}\gamma_4 \\ & + \delta_{(h,i)(2,2)}\gamma_{12} + \delta_{(h,j)(2,2)}\gamma_{13} + \delta_{(h,k)(2,2)}\gamma_{14} \\ & + \delta_{(i,j)(2,2)}\gamma_{23} + \delta_{(i,k)(2,2)}\gamma_{24} + \delta_{(j,k)(2,2)}\gamma_{34} \\ & + \delta_{(h,i,j)(2,2,2)}\gamma_{123} + \delta_{(h,i,k)(2,2,2)}\gamma_{124} \\ & + \delta_{(h,j,k)(2,2,2)}\gamma_{134} + \delta_{(i,j,k)(2,2,2)}\gamma_{234} \\ & + \delta_{(h,i,j,k)(2,2,2,2)}\gamma_{1234}\,.\end{aligned} \tag{3}$$

Except for λ_{1234} and γ_{1234}, these parametrizations are *not* equivalent and can lead to different conclusions about which terms should be in a model.

As mentioned earlier, a primary difficulty in testing each term last is in relating the tests for the reparametrized model to tests for ANOVA type models. In the special case where each factor has two levels (categories), the relationship is simple, because each term in the ANOVA type models has one degree of freedom, just as each test in the reparametrized model has one degree of freedom. If a particular term has a large test statistic, the corresponding main effect or interaction is included in the model. For example, if we reject $H_0 : \lambda_{12} = 0$ (or $H_0 : \gamma_{12} = 0$), then our ANOVA model includes $u_{12(hi)}$. This implies that the ANOVA model will include (at least implicitly) $u_{1(h)}$ and $u_{2(i)}$ regardless of whether λ_1 (γ_1) and λ_2 (γ_2) are significantly different from zero. Note that because λ_{12} and γ_{12} are not equivalent, the results of this procedure depend on the parametrization chosen.

In fact, identifying important effects by testing all effects last does not provide a good end model. It provides an initial model from which some method of exploration (e.g., forward selection or combined stepwise) can be used to determine a final model.

EXAMPLE 6.2.7. Consider again the muscle tension data of Example 6.2.1. The λ values defined in Example 6.2.5 using the usual side conditions are exactly the same as the λ values defined in model (2). For example,

$$4\lambda_{12} = u_{12(11)} - u_{12(12)} - u_{12(21)} + u_{12(22)}.$$

We have already obtained estimates of the λ's, standard errors, and $|z|$ scores.

Similarly, if we use the parametrization of model (3) and the related side conditions, we find, for example, that

$$\gamma_{12} = u_{12(11)} - u_{12(12)} - u_{12(21)} + u_{12(22)};$$

however, $4\lambda_{12} \neq \gamma_{12}$. Some computer programs, e.g., GLIM, routinely provide estimates and standard errors for the parametrization of model (3).

The results along with $|z|$ values are reported in Table 6.4. There are now at least six interesting terms: γ_{WMD}, γ_{MD}, γ_{WM}, γ_D, γ_M, and γ_W. The $|z|$ value for γ_{WD} is also quite large. Again using the rule of including lower-order terms, the inclusion of γ_{WMD} implies the model [WMD], which, in turn, implies the inclusion of all of the other interesting terms. Once again, factor T does not appear.

The results from these two parametrizations are reasonably consistent for this data set, but it is not difficult to see how the analysis could go awry. Consider the terms for TWM and TMD. Using model (2), we get $|z(\hat{\lambda}_{TWM})| = 1.03$ and $|z(\hat{\lambda}_{TMD})| = 1.57$, so, although neither is significant, the TMD term seems considerably more important than the TWM term. However, in model (3), $|z(\hat{\gamma}_{TWM})| = 0.80$ and $|z(\hat{\gamma}_{TMD})| = 0.80$. In model (3), the TMD and TWM terms seem to be of equal importance. The problem is that the parameters are model dependent. Because they are from different models, $8\lambda_{TWM} \neq \gamma_{TWM}$. As a result, the test statistics are testing different things. The only exception to this is that $16\lambda_{TWMD} = \gamma_{TWMD}$.

The relationships between parameters in models (1), (2), and (3) are complex and, in the our view, often not worth pursuing. The simplest way to avoid the complexities of alternative parametrizations is to deal directly with model (1) and its submodels. In our view, the method of testing each term last can give some rough ideas about the analysis, but usually should not be considered to give anything more than *rough* ideas.

Again, we note a rather curious phenomenon in this example. The experiment was conducted to investigate changes in muscle tension. Neither parametrization shows the significance of any effect involving T, the change in tension. It is theoretically possible that none of the other factors relate to change in muscle tension, but in most studies of this type, the investigator conducts the experiment because he or she knows that there are relationships between the other factors and change in tension. As we saw from the tests of partial and marginal association, such relationships exist. The method employed has simply failed to find them.

Two final comments on the choice of an initial model. *Any effects that are forced into the model to deal with the sampling scheme should be included in any initial model and never deleted in any model selection method.* Also, one is rarely interested in models that are smaller than the model of complete independence. It is common practice to include at least the main effect for every factor in an initial model.

6.3 Example of Stepwise Methods

We now give detailed examples of forward selection and backward elimination. Reconsider the data of Example 3.7.2 in which there are four factors

TABLE 6.4. Muscle Tension Data — Model (3): Estimates, Standard Errors, and Test Statistics

γ	$\hat{\gamma}$	SE	$\lvert z\rvert$
T	−0.000	.8165	0.00
W	1.992	.6154	3.24
M	2.037	.6138	3.32
D	1.946	.6172	3.15
TW	0.716	.8569	0.84
TM	0.578	.8570	0.67
WM	−3.742	.8199	4.56
TD	−0.742	.9024	0.82
WD	−1.571	.6765	2.32
MD	−2.684	.7179	3.74
TWM	−0.888	1.104	0.80
TWD	−0.304	.9781	0.31
TMD	0.810	1.010	0.80
WMD	3.407	.9620	3.54
TWMD	0.436	1.307	0.33

defining a $2 \times 2 \times 3 \times 6$ table. Recall that the factors are

Factor	Abbreviation	Levels
Race	R	White, Nonwhite
Sex	S	Male, Female
Opinion	O	Yes = Supports Legalized Abortion
		No = Opposed to Legalized Abortion
		Und = Undecided
Age	A	18-25, 26-35, 36-45, 46-55, 56-65, 66+ years

The data are repeated in Table 6.5.

We begin by fitting the all three-factor model, the all two-factor model, and the complete independence (all one-factor) model.

Model	*df*	G^2
[RSO][RSA][ROA][SOA]	10	6.12
[RS][RO][RA][SO][SA][OA]	37	26.09
[R][S][O][A]	62	121.47

Clearly, both the all three-factor and the all two-factor models fit the data relative to the saturated model. Comparing the all three-factor and all two-factor models gives

$$\begin{aligned} G^2 &= 26.09 - 6.12 = 19.97, \\ df &= 37 - 10 = 27, \end{aligned}$$

TABLE 6.5. Abortion Opinion Data

Race	Sex	Opinion	Age					
			18-25	26-35	36-45	46-55	56-65	66+
White	Male	Yes	96	138	117	75	72	83
		No	44	64	56	48	49	60
		Und	1	2	6	5	6	8
	Female	Yes	140	171	152	101	102	111
		No	43	65	58	51	58	67
		Und	1	4	9	9	10	16
Nonwhite	Male	Yes	24	18	16	12	6	4
		No	5	7	7	6	8	10
		Und	2	1	3	4	3	4
	Female	Yes	21	25	20	17	14	13
		No	4	6	5	5	5	5
		Und	1	2	1	1	1	1

so there is no reason to reject the all two-factor model. The model of complete independence does not fit.

6.3.1 Forward Selection

First, we consider forward selection using the model of complete independence as our initial model. As a criterion for adding terms, we add the most significant term as long as the significance level is below .10. For those of us who are not wild about significance tests, some comments based directly on the likelihood ratio test statistics are also included. It should be remembered that the formal method of forward selection is based on the significance levels.

Three methods of forward selection have been discussed. These involve adding two-factor terms, adding simple effects, and adding multiple effects. Starting with the model of complete independence, the first step in all three methods is the same. We require the following fits:

Model	Added Term	*df*	G^2
[RS][O][A]	RS	61	119.45
[RO][S][A]	RO	60	107.48
[RA][S][O]	RA	57	115.46
[SO][R][A]	SO	60	112.28
[SA][R][O]	SA	57	120.75
[OA][R][S]	OA	52	59.78
[R][S][O][A]		62	121.47

All of the models with two-factor terms are compared to the model of complete independence; e.g., to test the RS term, $G^2 = G^2([R][S][O][A]) - G^2([RS][O][A]) = 121.47 - 119.45 = 2.02$. The degrees of freedom are $62 - 61 = 1$. The results of the tests are summarized below.

Term	*df*	G^2	*P*
RS	1	2.02	.1557
RO	2	13.99	.0009
RA	5	6.01	.3055
SO	2	9.19	.0101
SA	5	0.72	.9820
OA	10	61.69	.0000

The term OA has the smallest *P value*, so [OA] is added to the model of complete independence.

With the new model [R][S][OA], the second step of adding either two-factor effects or simple effects again remains the same. At this point, adding any three-factor term would imply adding more than one effect, so simple effects are only two-factor effects. Consideration of the addition of multiple effects leads to a different second step.

To add either a two-factor effect or a simple effect requires the following fits:

Model	Added Term	*df*	G^2
[RS][OA]	RS	51	57.76
[RO][OA][S]	RO	50	45.79
[RA][OA][S]	RA	47	53.77
[SO][OA][R]	SO	50	50.59
[SA][OA][R]	SA	47	59.06
[R][S][OA]		52	59.78

Again, all the models with an additional two-factor term are compared to the model [R][S][OA].

Term	*df*	G^2	*P*
RS	1	2.02	.1557
RO	2	13.99	.0009
RA	5	6.01	.3055
SO	2	9.19	.0101
SA	5	0.72	.9820

The term RO is added to the model, giving a base model of [RO][OA][S].

If addition of multiple effects to [R][S][OA] is considered, two more models must be evaluated. Adding an additional factor to the OA term leads to the models [R][SOA] and [S][ROA]. These models have the fits

Model	Added Term	*df*	G^2
[R][SOA]	SOA	35	47.91
[S][ROA]	ROA	35	32.31
[R][S][OA]		52	59.78

They are tested against [R][S][OA].

Term	*df*	G^2	*P*
SOA	17	11.87	.8082
ROA	17	27.47	.0516

These *P values* are larger than the *P value* for RO, so the term RO is still added to [R][S][OA], giving the new model [RO][OA][S].

In this example, addition of two-factor effects and addition of simple effects turn out to be identical procedures. We now follow this procedure to its conclusion. After establishing the end model, we will indicate how these procedures would have differed if our model selection procedure had been modified slightly. Finally, we will follow the method of addition of multiple effects to its conclusion.

After the second step of forward selection, the simple effects and two-factor effects procedures had arrived at a base model of [RO][OA][S]. The only simple effects that can be added are two-factor effects. There are four possible effects that can be added. The necessary model fits are

Model	Added Term	*df*	G^2
[RS][RO][OA]	RS	49	43.77
[RA][RO][OA][S]	RA	45	38.82
[SO][RO][OA]	SO	48	36.60
[SA][RO][OA]	SA	45	45.07
[RO][OA][S]		50	45.79

This leads to the following differences:

Term	*df*	G^2	*P*
RS	1	2.02	.1557
RA	5	6.97	.2228
SO	2	9.19	.0101
SA	5	0.72	.9820

Thus, [SO] is added to the model. This turns out to be the last step at which anything is added to the model. The next step examines

Model	Added Term	*df*	G^2
[RS][SO][RO][OA]	RS	47	34.25
[RA][SO][RO][OA]	RA	43	29.63
[SA][SO][RO][OA]	SA	43	35.33
[SO][RO][OA]		48	36.60

and

Term	*df*	G^2	*P*
RS	1	2.35	.1252
RA	5	6.97	.2228
SA	5	1.27	.9382

At this stage, all of the *P values* are in excess of .10, so no new term is added and the final model is [SO][RO][OA].

To see how adding two-factor effects can differ from adding simple effects, suppose that our criterion for stepping is having *P values* less than .15. With this criterion, the term RS would be added to the model, giving a new model of [RS][SO][RO][OA]. If we restrict ourselves to adding two-factor effects, the next step involves adding RA or SA. If we allow addition of simple effects, the three-factor term RSO could also be added. This is the first time that a simple effect is a three-factor effect because [RS][SO][RO][OA] is the first model that contains all three of the two-factor effects that correspond to a three-factor effect. In particular, with [RS], [SO], and [RO] in the model, adding [RSO] is adding a simple effect.

Recall that the rationale for starting our search with the model of complete independence was based in part on the fact that the all two-factor model gave an adequate fit. This provides a rationale for considering only the addition of two-factor terms. Unfortunately, the test of the all two-factor model against the all three-factor model can have very little power for identifying individual three-factor terms that are important. It is not safe to ignore all three-factor terms based on this one test. Thus, it is dangerous to consider adding only two-factor effects. By considering addition of simple effects, we at least admit the possibility of examining important three-factor effects.

We now return to examining forward selection with the addition of multiple effects. Recall that after the second step, we had arrived at a base model

of [RO][OA][S]. The multiple effects that can be added involve adding one new factor to a term already in the model, so these are the remaining two-factor effects (which are also simple effects) plus RSO, ROA, and SOA. Given below are statistics for fitting each model plus the differences in *df*'s and G^2's between the various models and [RO][OA][S]. Tests are based on these differences. We use a cutoff of $\alpha = .05$. (For this example, the first two steps do not change when $\alpha = .05$ is used instead of $\alpha = .10$.)

				Differences		
Model	Added Term	*df*	G^2	*df*	G^2	*P*
[OR][OA][SA]	SA	45	45.07	5	0.72	.9820
[RA][OR][OA][S]	RA	45	38.82	5	6.97	.2228
[RO][OA][OS]	SO	48	36.60	2	9.19	.0101
[RO][OA][SR]	RS	49	43.77	1	2.02	.1557
[RO][SOA]	SOA	33	33.92	17	11.87	.8082
[ROA][S]	ROA	35	32.31	15	13.48	.5654
[RSO][OA]	RSO	45	24.77	5	21.02	.0008
[RO][OA][S]	—	50	45.79	—	—	—

For testing against [RO][OA][S], the model with the smallest *P value* is [RSO][OA], with $P = .0008$. The *P value* is less than .05, so we take [RSO][OA] as our working model. Multiple effects that can be added to this are SA, RA, SOA, ROA, RSA, and RSOA. The statistics are given below.

				Differences		
Model	Added Term	*df*	G^2	*df*	G^2	*P*
[RSOA]	RSOA	0	0.00	45	24.77	.9938
[RSO][OA][SA]	SA	40	23.50	5	1.27	.9382
[RSO][OA][RA]	RA	40	17.79	5	6.97	.2228
[RSO][SOA]	SOA	30	22.09	15	2.68	.9998
[RSO][ROA]	ROA	30	11.29	15	13.48	.5655
[RSO][OA][RSA]	RSA	30	14.43	15	10.34	.7980
[RSO][OA]	—	45	24.77	—	—	—

For testing against the model [RSO][OA], every model has a *P value* greater than .05, so no new terms are added. The final model is [RSO][OA].

Because the addition of multiple effects leads to considering more models than the addition of simple effects, the author prefers the multiple effect option if you insist on doing forward selection.

6.3.2 Backward Elimination

We now consider applying backward elimination to the initial model containing all two-factor terms. We will use a cutoff value of $\alpha = .05$. Given

below are statistics for the model of all two-factor terms and the six models in which one of the two-factor terms has been dropped. (At this stage, the simple effects are precisely the two-factor effects.) The *df* and G^2 for testing each model against the saturated model are given. With this information, each reduced model can be tested against the all two-factor model by taking differences in the *df*'s and G^2's. For each of the reduced models, the differences are listed along with the *P value* for the test.

Model	Deleted Term	*df*	G^2	Differences *df*	Differences G^2	Differences *P*
[AS][AR][AO][OS][OR][SR]	—	37	26.09	—	—	—
[AS][AR][OS][OR][SR]	AO	47	89.24	10	63.15	.0000
[AR][AO][OS][OR][SR]	AS	42	27.28	5	1.19	.9461
[AS][AO][OS][OR][SR]	AR	42	32.98	5	6.89	.2289
[AS][AR][AO][OR][SR]	OS	39	36.12	2	10.03	.0067
[AS][AR][AO][OS][SR]	OR	39	41.33	2	15.24	.0005
[AS][AR][AO][OS][OR]	SR	38	28.36	1	2.27	.1319

Deleting the AS term gives the largest *P value*, so we choose the reduced model [AR][AO][OS][OR][SR].

Once again, the simple effects are the two-factor effects. The necessary statistics are

Model	Deleted Term	*df*	G^2	Differences *df*	Differences G^2	Differences *P*
[AR][AO][OS][OR][SR]	—	42	27.28	—	—	—
[AR][OS][OR][SR]	AO	52	89.93	10	62.65	.0000
[AO][OS][OR][SR]	AR	47	34.25	5	6.97	.2228
[AR][AO][OR][SR]	OS	44	36.80	2	9.52	.0086
[AR][AO][OS][SR]	OR	44	42.53	2	15.26	.0005
[AR][AO][OS][OR]	SR	43	29.63	1	2.35	.1252

Deleting the AR term gives the largest *P value*, so the new model is [AO][OS][OR][SR].

Simple effects are still two-factor effects, so the necessary statistics are

Model	Deleted Term	*df*	G^2	Differences *df*	Differences G^2	Differences *P*
[AO][OS][OR][SR]	—	47	34.25	—	—	—
[A][OS][OR][SR]	AO	57	95.94	10	61.69	.0000
[AO][OR][SR]	OS	49	43.77	2	9.52	.0085
[AO][OS][SR]	OR	49	48.57	2	14.32	.0008
[AO][OS][OR]	SR	48	36.60	1	2.35	.1252

We now delete SR and use the base model [AO][OS][OR]. The statistics are

Model	Deleted Term	df	G^2	Differences df	Differences G^2	Differences P
[AO][OS][OR]	—	48	36.60	—	—	—
[A][OS][OR]	AO	58	98.29	10	61.69	.0000
[AO][OR][S]	OS	50	45.79	2	9.19	.0101
[AO][OS][R]	OR	50	50.59	2	13.99	.0009

None of the *P values* is greater than .05, so we stop deleting terms and go with the model [AO][OS][OR]. Note that this model has the nice interpretation that given people's opinions; race, sex, and age are independent.

In this example, simple effects were always two-factor effects. If our cutoff level had been $\alpha = .01$, this would not have happened. With a cutoff of .01, the term OS can be dropped from [AO][OS][OR], yielding the model [AO][OR][S]. Deleting two-factor terms leads us to consider the reduced models [A][OR][S] (eliminating AO) and [AO][R][S] (eliminating OR). If we consider dropping simple effects, we would also consider the reduced model [AO][OR] (eliminating S). However, as mentioned earlier, it is typically not a good idea to drop main effects. If the initial model was the model of all three-factor effects, the difference between dropping three-factor effects and dropping simple effects could be substantial. Dropping simple effects is a more general procedure and seems more reasonable to the author.

6.3.3 Comparison of Stepwise Methods

Forward selection with $\alpha = .10$ and addition of simple effects lead to the model [RO][SO][OA]. It is easily seen that $\alpha = .05$ would lead to the same model. Forward selection with multiple effects and $\alpha = .05$ leads to [RSO][OA]. Backward elimination of simple effects from the all two-factor model with $\alpha = .05$ leads to [RO][SO][OA]. Frankly, we are lucky to have two methods give the same model. There is no reason that this needs to happen. Which is a better model? One is a special case of the other, so the test statistic is $36.60 - 24.77 = 11.83$ with $48 - 45 = 3$ degrees of freedom. This suggests quite strongly that [RSO][OA] is the better model. Note that it will not always be possible to test the results of different procedures because the models may not be comparable.

Stepwise methods are very sensitive to the cutoff values used. They are also very sensitive to the initial model. For backward elimination, we started with the model of all two-factor effects. The importance of the single three-factor effect RSO was washed out in testing the all two-factor model against the all three-factor model. Hence, it was decided to go with the two-factor model. If we had considered Brown's measures of partial and marginal association, we would have been better off (at least for these data). Brown's measures are given in Table 6.6. The term [RSO] stands out as a clearly important effect. In fact, even without using a stepwise procedure, Brown's

tests clearly suggest the model [RSO][OA], but that is a function of these particular data.

TABLE 6.6. Brown's Measures of Association

Effect	*df*	Partial G^2	*P*	Marginal G^2	*P*
R	1	1552.90	.00	—	—
S	1	25.21	.00	—	—
O	2	1532.82	.00	—	—
A	5	55.14	.00	—	—
RS	1	2.27	.13	2.02	.16
RO	2	15.24	.00	13.99	.00
SO	2	10.03	.01	9.19	.01
RA	5	6.89	.23	6.01	.31
SA	5	1.19	.95	0.72	.98
OA	10	63.15	.00	61.69	.00
RSO	2	10.51	.01	9.48	.01
RSA	5	2.55	.77	1.70	.89
ROA	10	7.17	.71	6.51	.77
SOA	10	1.43	1.00	1.41	1.00
RSOA	10	6.12	.81	6.12	.81

In this section, we have used several variations on stepwise regression. This is not just a pedagogical device. *If stepwise methods are to be used in spite of their well-known weaknesses, it is important to use several variations.* This allows the data to indicate several candidate models. *These models should be compared to see how well they fit the model assumptions. They should also be compared for interpretability.*

6.3.4 Computer Commands

BMDP-4F performs these stepwise procedures and gives initial models, including measures of partial and marginal association. Commands for backward elimination of simple effects are

```
/ INPUT     FILE = 'C:\LOGLIN\ABORT.DAT'.
          FORMAT = FREE.
          VARIABLES = 5.
/ VARIABLE  NAMES = R, S, A, O, N.
/ TABLE     INDEX = R, S, A, O.
            COUNT = N.
/ STAT      ALL.
/ FIT       MODEL = AS, AR, AO, OS, OR, SR.
            ASSOCIATION = 4.
            DELETE = SIMPLE.
            STEP = 10.
            PROB = .05.
```

```
/ PRINT     LINE = 79.
/ END
```

The "step" command specifies how many steps are allowed in the procedure. The "prob" command specifies the probability for stopping the procedure. With this program, *both* the *P value* for the individual term and the *P value* for testing the model against the saturated model must be less than the specified probability for the procedure to stop.

6.4 Aitkin's Method of Backward Selection

Aitkin (1978, 1979) suggests a model selection method that is closely related to the *all s-factor effects* method described in Section 2. After using backward selection to pick an all s-factor model, Aitkin's method provides for testing every model intermediate between the all s-factor model and the all $s-1$-factor model. The procedure also incorporates ideas on *simultaneous testing* that control the overall error rate for all tests performed.

Aitkin begins by testing the all $s-1$-factor model against the all s-factor model at a level, say, γ_s. (The choice of γ_s will be discussed later.) This is actually a test of whether the s-factor effects are needed in the model. Except for the choice of γ_s, this is exactly what was done in the subsection of Section 2 on All s-Factor Effects.

To describe the procedure precisely, we need some additional notation. Let G_s^2 be the likelihood ratio test statistic and let d_s be the degrees of freedom for testing the all s-factor model against the saturated model. To test the need for s-factor effects, we reject the null hypothesis of no s-factor effects if

$$G_{s-1}^2 - G_s^2 > \chi^2(1-\gamma_s, d_{s-1} - d_s) \; .$$

This is a test for the adequacy of the all $s-1$-factor model. Aitkin then identifies the smallest value of s for which the all s-factor effects model adequately fits the data. This model has s as the largest value such that $G_{s-1}^2 - G_s^2 > \chi^2(1-\gamma_s, d_{s-1} - d_s)$.

EXAMPLE 6.4.1. Consider the muscle tension data of Example 6.2.1. This is a four-factor table. As given in Example 6.2.1, the all s-factor models are

s	Model	d_s	G_s^2
4	[TWMD]	0	0
3	[TWM][TWD][TMD][WMD]	1	0.11
2	[TW][TM][WM][TD][WD][MD]	5	47.67
1	[T][W][M][D]	11	127.4

For reasons to be considered later, suppose $\gamma_4 = .05$, $\gamma_3 = .185$, and $\gamma_2 = .265$, then Aitkin's tests are

$s-1$ versus s	$G^2_{s-1} - G^2_s$	$\chi^2(1-\gamma_s, d_{s-1}-d_s)$
3 versus 4	$0.11 - 0 = .11$	$\chi^2(.95,1) = 3.841$
2 versus 3	$47.67 - 0.11 = 47.56$	$\chi^2(.815,4) = 6.178$
1 versus 2	$127.4 - 47.67 = 79.7$	$\chi^2(.735,6) = 7.638$

The largest value of s for which $G^2_{s-1} - G^2_s > \chi^2(1-\gamma_s, d_{s-1}-d_s)$ is $s = 3$. The model [TWM][TWD][TMD][WMD] fits the data according to Aitkin's criteria.

Before discussing Aitkin's method in general, we examine of its application to the race, sex, opinion, age data.

EXAMPLE 6.4.2. For the race, sex, opinion, age data of Section 3, take $\gamma_s = .10$ for all s. Recall from the previous section that the model of all two-factor effects fits well. Because of this, the procedure will never consider three-factor effects. In particular, it will never consider the model [RSO][OA] which fits very well.

The model of all two-factor effects has 37 degrees of freedom and $G^2 = 26.09$ for testing against the saturated model. The model of complete independence has 62 degrees of freedom for testing against the saturated model. In Aitkin's method, a model, say X, with all main effects and some two-factor effects is deemed inadequate if

$$G^2_X - 26.09 > \chi^2(.90, 62-37) = 34.39$$

or, equivalently, if

$$G^2_X > 34.38 + 26.09 = 60.47\,.$$

Any model that is not inadequate is adequate. A *minimal adequate model* is an adequate model that has no submodel that is deemed adequate. *Our primary interest is in identifying minimal adequate models.*

Among models with two-factor effects, the most informative models will be small models that fit and large models that do not fit. If a small model fits, any larger model also fits. If a large model does not fit, then any smaller model does not fit.

We begin by looking for small models that fit adequately. In particular, consider the first step of forward selection from the complete independence model.

Model	Added Effect	G^2
[R][S][OA]	AO	59.78
[R][SA][O]	SA	120.75
[RA][S][O]	RA	115.46
[R][SO][A]	SO	112.28
[RO][S][A]	RO	107.47
[RS][O][A]	RS	109.45

We are in luck! One of these models, [R][S][OA], has $G^2 < 60.47$, so it is deemed adequate. Thus, we have an extremely small model that is adequate and any larger model must also be adequate. The model [R][S][OA] is clearly a minimal adequate model. We have also established that any other minimal adequate models must have at least 2 two-factor effects (we have already checked all models with only 1 two-factor effect) and none of the two-factor effects can be OA (otherwise it will have [R][S][OA] as a submodel).

We now look for relatively large models that do not fit. Typically, a good approach is to look at the first step of backward elimination from the all two-factor effects model and hope to find some that do not fit adequately. For these data, however, we just established that any model larger than [R][S][OA] fits. In the first step of backward elimination, one two-factor effect is dropped out. Unless the two-factor effect dropped out is OA, we already know that the model will fit. The only model we need consider is [RS][RO][RA][SO][SA]; this model has a G^2 of 89.24. Once again, we are in luck. The value 89.24 is greater than the critical value 60.47, so [RS][RO][RA][SO][SA] is deemed inadequate. Thus, any model that does not include OA must be inadequate. Combining our two results, we have established that the only minimal adequate model is [R][S][OA].

We now consider what would occur if γ_2 was somewhat larger. Aitkin has suggested a method of choosing the γ_s's that will be discussed later. It begins with a level, say $\alpha = .05$, and for $t = 4$ factors with $s = 2$, Aitkin's choice is $\gamma_2 = 1 - (1-\alpha)^{\binom{4}{2}} = .265$, where $\binom{4}{2} = 6$ is the number of combinations of four things taken two at a time. Upon establishing that $\chi^2(.735, 25) = 28.97$ where $25 = 62 - 37$, a model X consisting of two-factor effects is deemed inadequate if

$$G_X^2 > 28.97 + 26.09 = 55.06\,.$$

The model [R][S][OA] has $G^2 = 59.78$, so it is no longer deemed adequate. However, it is very close to meeting the adequacy criterion. It makes sense to consider models that include [OA] and another two-factor effect. In particular, this is precisely the second step in forward selection starting with complete independence and adding simple effects. The models considered and their G^2's are

Model	G^2
[R][SA][OA]	59.06
[RA][S][OA]	53.77
[R][SO][OA]	50.59
[RO][S][OA]	45.79
[RS][OA]	57.76

The models [R][SA][OA] and [RS][OA] have $G^2 > 55.06$, so they are deemed inadequate. The models [RA][S][OA], [R][SO][OA], and

[RO][S][OA] are adequate and no smaller models are adequate, so the models [RA][S][OA], [R][SO][OA], and [RO][S][OA] are minimally adequate models. Working from the all two-factor model down, the model of all two-factor effects except [OA] is inadequate ($G^2 = 89.24$), so all adequate models contain [OA].

The inadequate models, [R][SA][OA] and [RS][OA] need to be considered as to the additional terms needed to make them adequate. If we add RA, SO, or RO, then we have made them larger than one of our minimally adequate models. The only two-factor effects that can generate additional minimally adequate models are SA and RS. If we add the appropriate effect to each model, we get [RS][SA][OA] in both cases. The G^2 for this model is 57.04. Because G^2 is greater than the critical value 55.06, [RS][SA][OA] is considered inadequate. Therefore, the only minimally adequate models are [RA][S][OA], [R][SO][OA], and [RO][S][OA]. All of these have simple interpretations in terms of conditional independence.

Aitkin's method applied to these data gives smaller models than any of the stepwise methods considered. Unfortunately, it missed the important [RSO] interaction.

General Discussion

We now present a general discussion of Aitkin's method. The method begins with a model of all s-factor effects that adequately fits the data, while the model with only the $s-1$-factor effects does not fit the data. The crux of the method is in identifying intermediate models that also give an adequate fit. If X is a *model that contains all $s-1$-factor effects and some but not all of the s-factor effects*, Aitkin tests the adequacy of X by rejecting adequacy of fit if

$$G_X^2 - G_s^2 > \chi^2(1-\gamma_s, d_{s-1} - d_s)$$

Here, G_X^2 is the likelihood ratio test statistic for testing model X against the saturated model. Note that the same criterion for rejection $\chi^2(1-\gamma_s, d_{s-1} - d_s)$ is used for any such model X. Also, if the all $s-1$-factor model is indeed an adequate fit, then because $G_X^2 - G_s^2 < G_{s-1}^2 - G_s^2$, the probability of a false rejection for any and all such models X is less than the probability of a false rejection of the all $s-1$-factor model.

The fact that the criterion of rejection does not depend on the particular model X leads to two important observations. If X is an adequate model, then any larger model, say W, must also be deemed adequate because $G_X^2 - G_s^2 \geq G_W^2 - G_s^2$. Similarly, if X is inadequate, then any smaller model W must also be deemed inadequate because $G_X^2 - G_s^2 \leq G_W^2 - G_s^2$. These facts together with Aitkin's restriction to only considering s-factor terms make it practical to examine all models that involve s-factor terms.

For the data of Example 6.2.1, Aitkin's method seeks to examine every model that is intermediate between the all three-

factor model [TWM][TWD][TMD][WMD] and the all two-factor model [TW][TM][WM][TD][WD][MD]. For example, in testing the intermediate, model [TWM][TWD][TMD], the value of $G^2_{[TWM][TWD][TMD]} - G^2_3 = 35.65$ is larger than $\chi^2(.815, 4) = 6.178$, so the model [TWM][TWD][TMD] is not considered adequate. (The test statistic was obtained from Table 6.1 and $.815 = 1 - \gamma_3$ from earlier in the section.) Note that the χ^2 value uses the 4 degrees of freedom associated with testing the all two-factor model against the all three-factor model. Similarly, any other model intermediate between the all two-factor and all three-factor models is tested against the all three-factor model using $\chi^2(.815, 4)$.

If the all two-factor model is adequate, the probability of a false rejection when testing the all two- and all three-factor models is $\gamma_3 = .185$. If the all two-factor model is adequate, then any intermediate model is adequate. Similarly, if an intermediate model is inadequate, then the all two-factor model must be inadequate. Because tests for intermediate models use the same χ^2 value but have smaller G^2 values than the all 2 versus all 3 test, a false rejection occurs for an intermediate model if and only if a false rejection occurs for the all two-factor model. (This is similar in spirit to Scheffé's method of multiple comparisons.)

Aitkin defines a subset of s-factor effects as a *minimal adequate subset* if no proper subset defines an adequate model. (The model is the model of all $s-1$-factor effects plus the subset of s-factor effects.) Typically, there will be several such models. Each of these models may be reduced further by testing any smaller-order (e.g., $s-1$) terms that are not forced into the model. These tests use the criterion of rejection appropriate for that order. [For an $s-1$-factor term, compare the test statistic to $\chi^2(1-\gamma_{s-1}, d_{s-2} - d_{s-1})$.] The fact that relatively few lower-order terms will not be forced into the model makes it practical to carry out this procedure.

The end result of Aitkin's model selection method is a collection of minimal adequate models. *These models should be compared for interpretability. The should also be compared to see how well they fit the model assumptions.* In fact, they can even be compared using the Adjusted R^2 or the Akaike information criteria discussed in Section 3.6.

Probably the main fault of Aitkin's method is the backward elimination in its first step. The first step is to decide on an adequate all s-factor model. For example, in a five-factor table, there are 10 three-factor terms. If 1 of the three-factor terms is substantial and the other 9 are not, then a test for all 10 terms will have little power to establish the need for this single three-factor term. We saw an example of this phenomenon earlier with the abortion opinion data. If it is important to pick up individual high-order interactions, Aitkin's method will be problematic. On the other hand, Aitkin's method may provide a good starting point to which an examination of higher-order interactions can be added.

Finally, we discuss the choice of γ_s's. Aitkin suggests choosing the γ_s's

so that there is a probability no greater than, say, γ of rejecting the main-effects-only model (i.e., complete independence) when main-effects-only is adequate. Suppose there are t factors in the table. When complete independence is true, the various tests for s-order interactions are asymptotically independent. Thus, asymptotically, the probabilities of not rejecting any test is the product of the probabilities for the individual tests and the γ_i's should be chosen to satisfy

$$1 - \gamma = \prod_{s=2}^{t} (1 - \gamma_s) \, .$$

(Note that we do not consider testing main effects, i.e., first-order "interactions.")

Particular values of the γ_s's can be chosen by analogy with a balanced t-factor analysis of variance. In a t-factor ANOVA, the number of s-factor effects is $\binom{t}{s}$. In a balanced ANOVA (with known variance), each of these tests might be conducted at some common level α.

Applying this idea to log-linear models, the probability of not rejecting any of the s-factor tests (assuming complete independence holds and tests are independent) is

$$1 - \gamma_s = (1 - \alpha)^{\binom{t}{s}} \, .$$

This determines a specific value for γ_s. The corresponding value of γ can be found using the binomial theorem. Because $2^t = \sum_{2=0}^{t} \binom{t}{s}$, we find that

$$\begin{aligned} \gamma &= 1 - \prod_{s=2}^{t} (1 - \gamma_s) \\ &= 1 - \prod_{s=2}^{t} (1 - \alpha)^{\binom{t}{s}} \\ &= 1 - (1 - \alpha)^{2^t - t - 1} \, . \end{aligned}$$

Aitkin (1979) suggests that it is reasonable to pick an α level that yields a γ between .25 and .5.

In Example 6.4.1, $t = 4$ and the γ_s values were chosen to satisfy

$$1 - \gamma_s = (1 - .05)^{\binom{4}{s}} \, ,$$

so $\gamma_4 = .05$, $\gamma_3 = .185$ and $\gamma_2 = .265$. These yield

$$\begin{aligned} \gamma &= 1 - \{(1 - .05)(1 - .185)(1 - .265)\}, \\ &= .431 \end{aligned}$$

which is in Aitkin's suggested range.

In the discussion of Aitkin's (1978) paper, D.R. Cox suggests that the emphasis on simultaneous testing in Aitkin's method is excessive. A compromise approach that seems intuitively appealing to the current author is, for some value α, to choose $\alpha = \gamma_2 = \cdots = \gamma_t$.

6.5 Model Selection Among Decomposable and Graphical Models

Wermuth (1976) has proposed a backward elimination technique that is restricted to the decomposable models discussed in Section 5.2. She focuses on identifying pairs of factors that can be viewed as conditionally independent. Wermuth's method is most easily understood in terms of graph theory and our general discussion of the method will center on that. However, we begin our discussion with an example that does not use graphical terms or motivations.

EXAMPLE 6.5.1. Consider again the race, sex, opinion, age data with a backward elimination cutoff of $\alpha = .05$. The initial model is the saturated model and we consider all factor pairs for possible conditional independence. This leads to the following lack of fit tests:

Factor Pair	Model	*df*	G^2	*P*
RS	[ROA][SOA]	18	20.45	.3080
RO	[RSA][SOA]	24	38.22	.0329
RA	[RSO][SOA]	30	22.09	.8506
SO	[RSA][ROA]	24	27.91	.2639
SA	[RSO][ROA]	30	11.29	.9989
OA	[RSO][RSA]	28	78.06	.0000

Note that for factor pair RS, the corresponding model [ROA][SOA] has R and S independent given O and A. Similar interpretations hold for the other pairs and models. The largest *P value* among the tests is for [RSO][ROA]. The *P value* is greater than .05, so we take as the model [RSO][ROA]. The conditional independence relation is that S and A are independent given R and O.

We have incorporated into the model the conditional independence of factors S and A. At the next step, we consider models that incorporate conditional independence between another pair of factors. In particular, we consider models that are reduced relative to [RSO][ROA] and incorporate an additional conditional independence. *The one exception is that the factor pair RO is in both terms of the model* [RSO][ROA], *so it cannot be considered.* As will be seen later in this section, incorporating a conditional independence between the pair RO would lead to a model that is not decomposable. For the pair OA, conditional independence is introduced by reducing the term [ROA] into [RO][RA]. The resulting model is [RSO][RO][RA]. The term [RO] is redundant because it is implied by [RSO]; thus, the reduced model is [RSO][RA]. A similar analysis holds for all other factor pairs except RO. The second step of the model selection method requires the following models and statistics:

				Differences		
Factor Pair	Model	*df*	G^2	*df*	G^2	*P*
—	[RSO][ROA]	30	11.29	—	—	—
OA	[RA][RSO]	38	80.45	8	69.16	.0000
RA	[OA][RSO]	45	24.77	15	13.48	.5657
OS	[RS][ROA]	34	30.29	4	19.00	.0011
RS	[SO][ROA]	33	23.12	3	11.81	.0083

The tests presented as differences are tests of the given models against the model [RSO][ROA]. The largest *P value* is again greater .05 and belongs to [OA][RSO], so this model is used for the next step. The model [OA][RSO] incorporates the conditional independence of R and A in addition to the conditional independence of S and A obtained from the first step of the selection procedure. The model [OA][RSO] has R and S independent of A given O.

In the next step, we begin with the model [OA][RSO]. The pairs SA and RA have already been identified for conditional independence. *There are no pairs that are contained in both terms of the model, so there are no pairs that cannot be considered for possible conditional independence.* All pairs other than SA and RA are considered for the possibility that they are conditionally independent. To incorporate conditional independence between O and A into the model [OA][RSO], break the term [OA] into [O][A] and use the model [O][A][RSO], which is equivalent to [A][RSO]. To incorporate conditional independence between O and S into [OA][RSO], break the term [RSO] into [RO][RS] and use the model [OA][RO][RS]. Similar models are used for the factor pairs RS and RO. The necessary tests are given below.

				Differences		
Factor Pair	Model	*df*	G^2	*df*	G^2	*P*
—	[OA][RSO]	45	24.77	—	—	—
OA	[A][RSO]	55	86.45	10	61.68	.0000
OS	[OA][RS][RO]	47	43.77	2	19.00	.0000
RS	[OA][SO][RO]	48	36.60	3	11.83	.0082
RO	[OA][SO][RS]	49	48.57	4	23.80	.0001

None of the *P values* is greater than .05, so we stop with the model [OA][RSO].

It is interesting to note that this happens to be the same model as achieved by forward selection of multiple effects. Note also that if $\alpha \leq .0082$, we would be led to consider R and S as conditionally independent and thus use the model [RO][SO][OA]. This is the other model achieved by stepwise regression.

In turn, [RO][SO][OA] would lead to taking S and O as conditionally independent and thus using the model [RO][S][OA]. These results can be seen from the following displays.

				Differences		
Factor Pair	Model	*df*	G^2	*df*	G^2	*P*
—	[OA][SO][RO]	48	36.60	—	—	—
SO	[RO][OA][S]	50	45.79	2	9.19	.0101
OA	[RO][SO][A]	58	98.29	10	61.69	.0000
RO	[OA][SO][R]	50	50.59	2	13.99	.0009

				Differences		
Factor Pair	Model	*df*	G^2	*df*	G^2	*P*
—	[RO][OA][S]	50	45.79	—	—	—
OA	[RO][S][A]	60	107.5	10	61.7	.0000
RO	[OA][R][S]	52	59.78	2	13.99	.0009

The model [RO][S][OA] is one of the minimally adequate models arrived at in our second application of Aitkin's method.

A key feature of Wermuth's method is that a pair of factors contained in more than one term in the model cannot be considered for conditional independence. Edwards and Havranek (1985) suggest dropping this requirement. Doing so changes the method from a search among decomposable models to a search among graphical models.

General Discussion

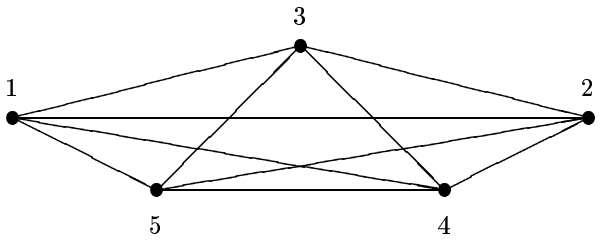
We now present a general discussion of Wermuth's method using graph-theoretic ideas. Recall that graphical models are completely determined by their two-factor effects. The procedure begins with the graphical model that includes all two-factor effects, i.e., the saturated model. Every two-factor effect is considered for deletion. The two-factor effect that generates the largest *P value* is deleted if the *P value* exceeds some cutoff point α.

Whichever effect is dropped, it determines two subsets of factors in which there are effects between every pair of factors. Recall that a subset with all possible two-factor effects is called *complete* and that if a complete subset is not strictly contained in any other complete subset, it is *maximal.* A maximal complete subset is a *clique.* Wermuth's method starts out with the clique based on all factors. Dropping one two-factor effect generates two cliques that each contain all but one factor.

Proceeding inductively, at any stage in Wermuth's method there are two or more cliques available. Two-factor effects that are part of more than one clique are not considered for elimination. It will be seen that the graphical model obtained by eliminating such effects is not decomposable. Among all other two-factor effects, the one with the largest *P value* is eliminated provided that the *P value* exceeds α.

Example 6.5.2. With five factors the initial model is [12345]. This can

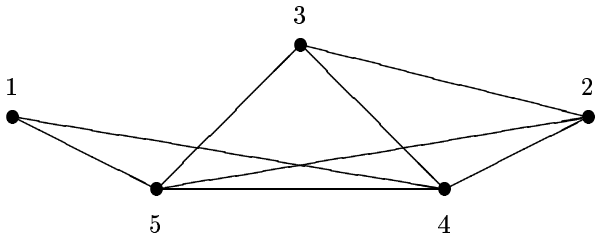
also be thought of as the initial clique.



All two-factor terms are considered for elimination. There are $\binom{5}{2} = 10$ of these. If, for example, [12] is dropped, it is easily seen from the graph that there are two new cliques, [1345] and [2345]. Thus, the graphical model is [1345][2345]. Another way of establishing the graphical model is to examine the nine remaining two-factor terms: [13], [14], [15], [34], [35], [45], [23], [24], [25]. Of these nine, the first six terms generate [1345] and the last six terms generate [2345]; thus, the model is [1345][2345]. The *P value* for dropping [12] is the *P value* for testing the model [1345][2345] versus [12345].

Having deleted [12], the second stage begins by considering the cliques of the model [1345][2345]. The cliques are [1345] and [2345]. The two-factor effects [34], [35], and [45] are contained in both cliques, so these are not considered for elimination. Among the other two-factor terms, the one with the largest *P value* is eliminated provided the *P value* exceeds α.

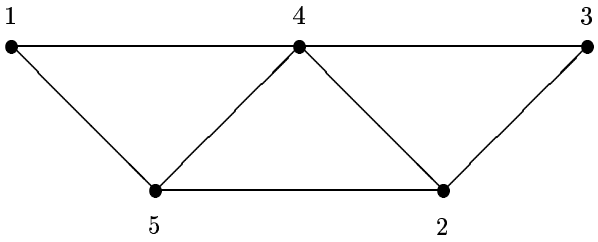
In considering whether to drop, say, [13], the test compares the graphical model with both [12] and [13] eliminated to the graphical model in which only [12] is eliminated. As discussed above, when [12] is eliminated, the model is [1345][2345]. If [13] is also eliminated, the clique [1345] breaks up into two complete subsets [145] and [345].



To see this, observe that [1345] is generated by [13], [14], [15], [34], [35], [45]. If [13] is dropped, the complete subsets are based on [14], [15], [45], and [34], [35], [45]. These generate [145] and [345], respectively. Note that [345] is not a clique because it is not maximal; it is contained in the complete subset [2345]. With both [12] and [13] eliminated, the cliques are [145] and [2345], so the model is [145][2345]. The test for eliminating [13] is the test of [145][2345] versus [1234][2345].

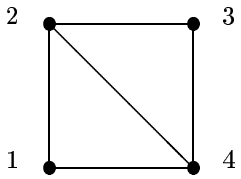
If [12] and [13] have been eliminated in the first two stages, the model is [145][2345]. The only two-factor effect contained in more than one clique is [45]. All other effects are considered for elimination. Suppose [35] is dropped. The resulting model is [145][234][245]. Now both [45] and [24] are in more than one clique, so neither can be eliminated. The process of testing and dropping two-factor terms continues until there are no tests with a *P value* greater than α.

To see that dropping an effect that is in more than one clique destroys decomposability, consider dropping [24] from the model [145][234][245].



From the two cliques involving [24], form the closed chain [34][45][52][23]. This chain of length 4 has one chord, [24]. If [24] is dropped, the model is no longer chordal, hence no longer decomposable. Whenever an effect contained in more than one clique is dropped, this construction of a closed chain of length four with no chords will work. Simply construct a closed chain out of the two vertices in the common effect and two other distinct vertices, one from each clique.

Conversely, if a model is decomposable and dropping a two-factor term makes it nondecomposable, then that two-factor term must be in more than one clique. For a model to become nondecomposable, the term eliminated must be the only chord in a closed chain of length four or more. If a closed chain of length four has only one chord, the chord must be in two complete subsets and thus in two cliques. In particular, if a decomposable model contains the closed chain [12][23][34][41] and contains *only* the one chord [24], then the model includes the complete sets [124] and [234], so [24] is contained in two different complete sets.



Each of the two complete sets must be contained in some clique and these cliques must be distinct. If the sets were in the same clique, then that clique would also have to include the chord [13]. By assumption, this cannot occur. In fact, this argument for closed chains of length four is sufficient for all cases because when a model becomes nondecomposable, the term eliminated must be the only chord in a closed chain of length four. In a decomposable model, any closed chain of length greater than four must have more than one chord because if the length is five or more and there is only one chord, there is a reduced closed chain of length at least four without a chord.

Clearly, Wermuth's method can be generalized to graphical models by removing the restriction that two-factor effects in more than one term are not considered for elimination. At each stage, all two-factor effects that have not been previously deleted can be considered for elimination. The corresponding models are the graphical models determined by the two-factor effects that have not been eliminated. This procedure was apparently first suggested by Edwards and Kreiner (1983). Model selection among graphical models is also discussed by Havranek (1984) and Edwards and Havranek (1985).

With four factors, the difference between Edwards and Kreiner's method and Wermuth's method occurs only at the second stage. This is due to the fact that there is only one graphical but nondecomposable model. As applied to the abortion opinion data, the term [SA] is dropped at the first stage, so at the second stage, Wermuth's method does not allow [RO] to be dropped. The method of Edwards and Kreiner has no such restriction.

For models with more than four factors, the difference between the graphical method and the decomposable method can be substantial. Restricting model search to graphical models seems like a very promising compromise between searching in the very large class of all ANOVA type models and searching in the very restrictive class of decomposable models. However, the difficulty of searching among graphical models should not be underestimated. Good (1975) has shown that for a 10-factor table there are almost 3.5 million graphical models.

With these methods as with all others, it is important to obtain several candidate models and to evaluate the models on grounds other than the values of their test statistics. Also, effects included because of the sampling design cannot be eliminated.

6.6 Use of Model Selection Criteria

The best approach to model selection for log-linear models would be to search through all models and choose, for closer examination, those with high values of Adj. R^2, low values of AIC, or extreme values of some other

model selection criterion, cf. Section 3.6. Such a procedure would require enormous amounts of computation. One possible way to reduce computations would be to base an initial search on models fitted by *weighted least squares* (cf. Sections 4.4 and 10.6) rather than models fitted by maximum likelihood.

At the moment, the author's best suggestion is to use a variety of model-fitting methods with a variety of critical (cutoff) values, and for stepwise methods, use a variety of initial models. By using several methods, we hope to find a wide range of possible models. These models can then be evaluated relative to each other with the help of the model selection criteria. Often, there is no need to decide on one particular model; a small number of alternative models may be more informative. If it is necessary to decide on only one model, the model with the lowest AIC (or highest Adj. R^2) may not be the best choice. Other considerations such as interpretability and consistency with assumptions may dictate choosing a model with a low AIC but not necessarily the lowest.

EXAMPLE 6.6.2. The evidence presented so far in the series of examples on the race, sex, opinion, age data strongly suggests to the author that the best model is [RSO][OA]. A formal comparison of all of the models arrived at by the various methods suggests the same conclusion.

Model	*df*	G^2	$A - q$	R^2	Adj. R^2
[RSO][OA]	45	24.77	−65.23	.80	.72
[RO][SO][OA]	48	36.60	−59.40	.70	.61
[R][S][OA]	52	59.78	−44.22	.51	.41
[RA][S][OA]	47	53.77	−40.23	.56	.42
[R][SO][OA]	50	50.59	−49.41	.58	.48
[RO][S][OA]	50	45.79	−54.21	.62	.53

In a less clear-cut situation, it would be wise to consider many more stepwise procedures than have been illustrated.

One worrisome aspect is that very little consideration has been given to three-factor effects other than RSO. The reader can check that none of the other three-factor effects substantially improves the model.

We have decided on one particularly good candidate model: [RSO][OA]. However, *the analysis of the data does not end with finding an appropriate ANOVA type model; that is just an important first step.* The model indicates that combinations of race and sex are independent of age given opinions about legalized abortions. Thus, we can collapse over some factors to study interrelationships in marginal tables. We can collapse over ages to study the relationships among race, sex, and opinion. We can collapse over race and sex to study the relationship between age and opinion. Cell counts for the collapsed tables need to be examined to study the nature of the relationships. These aspects of the analysis are discussed further in

Section 8. It is also necessary to evaluate whether the model really fits the data. To this end, Section 7 contains information on residual analysis and influential observations. Finally, the interpretability of the model should be examined. This model has a very nice interpretation with race and sex independent of age given opinion. Does the interpretation make any sense? As we will see in Section 8, interpretability is probably this model's weakest point. It can be argued that the appropriate analysis of these data involves explaining opinions on the basis of race, sex, and age. In that case, the methods of Chapter 4 should be used on these data.

6.7 Residuals and Influential Observations

In standard regression analysis, it is common practice to use residuals to check whether assumptions made in the model are valid and to detect the presence of observations that are unusually influential on the fit of the model. Three statistics that are commonly used for these purposes are the leverages, the standardized residuals, and Cook's distances. As seen in Chapter 4, residuals are not of much use when dealing with binary (0-1) data. However, in tables with reasonably large counts, residuals can be useful.

In a regression model $Y = X\beta + e$, the leverages are the diagonal elements of the projection matrix $X(X'X)^{-1}X'$. The matrix X is determined by the design of the data, i.e., the values of the predictor variables in the regression. The leverage measures how far the variables of a particular case are from the average of all of the cases. (The projection matrix is often called the "hat" matrix because it changes Y into $\hat{Y}$, i.e., $\hat{Y} = X(X'X)^{-1}X'Y$. Personally, I find this name distasteful. However, given the liberties I am about to take in naming residuals, I am probably not in a position to make a fuss.)

For log-linear models, there is an analogue of the leverage that depends both on the design and the probability of getting observations in a particular cell. Because the probabilities are unknown, they must be estimated; hence, we use estimated leverages. The estimated *leverage* of the ith case is denoted

$$\hat{a}_{ii} \ .$$

Leverages are discussed in detail in Chapter 10.

The log-linear analogue of a *standardized residual* is

$$r_i = \frac{n_i - \hat{m}_i}{\sqrt{\hat{m}_i(1 - \hat{a}_{ii})}} \ .$$

For a correct model, a large sample approximation for the distribution of r_i is $r_i \sim N(0, 1)$. Note that these are very similar to the *Pearson residuals* discussed earlier:

$$\tilde{r}_i = \frac{n_i - \hat{m}_i}{\sqrt{\hat{m}_i}} \ ;$$

they differ only in that the standardized residuals involve the leverages.

Actually, the residuals are the values $n_i - \hat{m}_i$, the difference between the observed and predicted values. As discussed in Section 2.1, these need to be standardized in some way. The Pearson residuals use a crude standardization, dividing by $\sqrt{\hat{m}_i}$. In Section 10.7, the Pearson residuals are referred to as the *crude standardized residuals* to distinguish them from the standardized residuals that involve a more sophisticated (and proper) standardization. This terminology was chosen by analogy with regression analysis. Unfortunately, it differs from the terminology used by many authors on log-linear models. Often, the Pearson (crude standardized) residuals are referred to as simply the standardized residuals, and the values defined here as the standardized residuals are referred to as the *adjusted* residuals.

Finally, *Cook's distance* for the ith case is a measure of the influence the ith case has on the fit of the model. In the context of fitting log-linear models, we drop each cell from the table, fit the remaining cells, and then estimate an expected cell count for the dropped cell. This is done without reference to any marginal totals that may be fixed by design; hence, it is most appropriate for Poisson sampling. If the model has p degrees of freedom, the analogue of Cook's distance can be written

$$C_i = \sum_{\text{all cells } r} \hat{m}_r \left[\log(\hat{m}_r/\hat{m}_{r(i)})\right]^2 \Big/ p$$

where $\hat{m}_{r(i)}$ is the estimate of the rth cell when the ith cell has been deleted (cf. Section 10.7). For Poisson sampling, these values can be calibrated by comparing them to a $\frac{1}{p}\chi^2(p)$ distribution. If $C_i > \chi^2(.5, p)/p$, cell i has a substantial influence. For multinomial or product-multinomial sampling, the degrees of freedom should be reduced by the number of independent multinomials. Because computation of all the $\hat{m}_{r(i)}$'s would require separate iterative procedures and be very expensive, it is suggested that a one-step estimate be used. Starting with the values $\hat{m}_r$, drop a cell and, rather than fully iterating, use just one step of the Newton-Raphson algorithm to obtain the $\hat{m}_{r(i)}$'s, cf. Section 10.7.

This definition of the Cook's distances has weaknesses; however, the situation is analogous to that of the Pearson residuals. The definition of the Pearson residuals as standardized residuals is weak, but the Pearson residuals are easy to compute and they contain valuable information. As will be seen below, using standard computer software, the standardized residuals are now easy to compute, so there is little reason to use the Pearson residuals. Similarly, using standard computer software, Cook's distances, as defined here, are easy to compute. Moreover, the author feels that they contain valuable information. Until something better becomes readily available, the author suggests examining these Cook's distances. Anderson (1992) discusses diagnostics for categorical data analysis and Thomas and

Cook (1989, 1990) discuss influence for generalized linear models (which include log-linear models, cf. Chapter 9).

6.7.1 Computations

We assume that the reader is capable of fitting an ANOVA model using a regression program. Our log-linear models are ANOVA type models. Good computer programs for doing regression generally provide leverages, standardized residuals, and Cook's distances. Also, they allow for computing weighted regressions.

After fitting the log-linear model, retain the counts for each cell, n_i, and the fitted values for each cell, $\hat{m}_i$. Now use the regression program to refit the ANOVA model, but use weighted regression with

$$\text{weight}_i = \hat{m}_i$$

and a dependent variable

$$Y_i = \log(\hat{m}_i) + (n_i - \hat{m}_i)/\hat{m}_i \, .$$

The leverages given by the program will be the $\hat{a}_{ii}$'s. The standardized residuals reported will be

$$r_i/\sqrt{\text{MSE}}$$

where $\sqrt{\text{MSE}}$ is the estimate of the standard deviation from the regression. Simply multiply the reported standardized residuals by $\sqrt{\text{MSE}}$ to obtain the correct values. The reported values of Cook's distances are

$$C_i/\text{MSE}$$

where C_i is computed using a one-step fit. The reported values C_i need to be multiplied by MSE. For the purpose of comparing the relative magnitudes of the r_i's or C_i's, the multiplication is irrelevant. For comparing standardized residuals to a $N(0,1)$ distribution or Cook's distances to a χ^2 distribution, the multipliers are important. It should also be noted that for large samples and a correct model, the MSE approaches a χ^2 distribution divided by its degrees of freedom. The large sample expected value for the MSE is 1.

Example 6.7.1. We now examine the leverages, standardized residuals, and Cook's distances for the abortion opinion data. In particular, we consider the fit of the model [RSO][OA]. The fitted values are given in Section 8 as Table 6.7.

The leverages are plotted against index values in Figure 6.1. The index values are just values $1, 2, \ldots, 72$ assigned to the cells. The G^2 for the model [RSO][OA] has 45 degrees of freedom. There are 72 cells, so there are

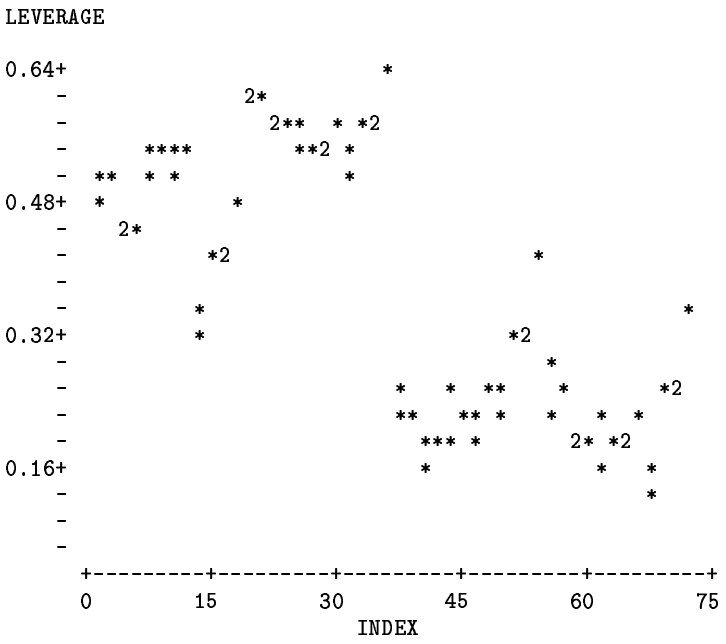


FIGURE 6.1. Leverage–Index Plot

$72 - 45 = 27$ degrees of freedom for the model. The sum of the 72 leverages must add up to 27. The average leverage is $\frac{27}{72} = .375$. The largest leverage is about .64, which is less than twice the average. None of the leverages seems excessively large. Leverages are rarely very large in balanced ANOVA type models.

Figures 6.2 and 6.3 contain a box plot and an index plot of the standardized residuals, respectively. The box plot identifies one very large residual and four other large residuals. In the index plot, only two residuals really stand out. There were 24 nonwhite males between 18 and 25 years of age who support legalized abortion; the estimated value from the model is only 14.52. This cell has a leverage of .222, a standardized residual of 2.82, and a Cook's distance of .085. The other large standardized residual is from the cell for nonwhite males above 65, who support legalized abortion. The observed value in this cell is 4, the fitted value is 10.90, the leverage is .181, the standardized residual is -2.31, and Cook's distance is .044. Considering that there are 72 cells, these values are not remarkably large. In fact, what is remarkable is that most of the standardized residuals are so tightly packed around zero.

Figure 6.4 contains a normal probability plot of the standardized residuals. If the asymptotic theory is valid, the plot should be approximately linear. It is not. Again, the problem seems to be that there are too many

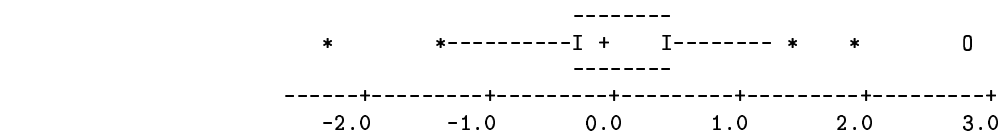


FIGURE 6.2. Standardized Residual–Box Plot

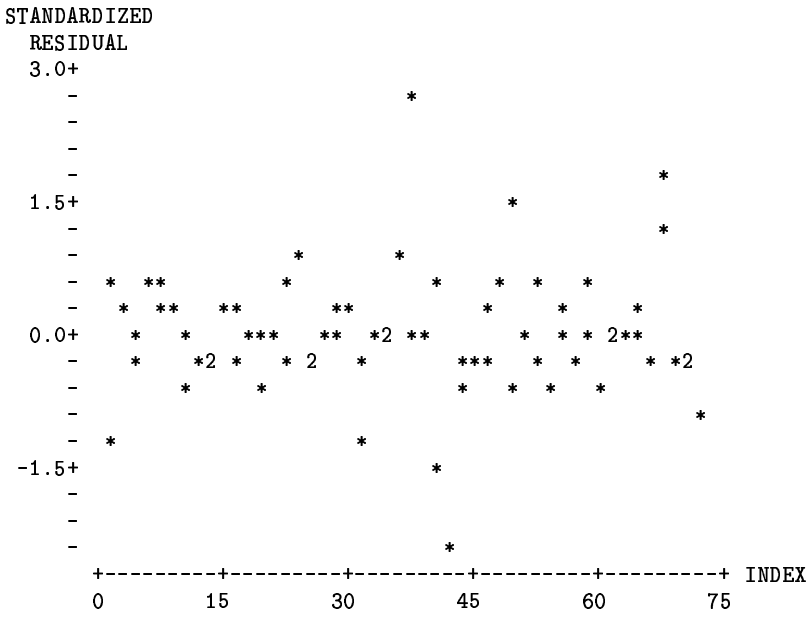


FIGURE 6.3. Standardized Residual–Index Plot

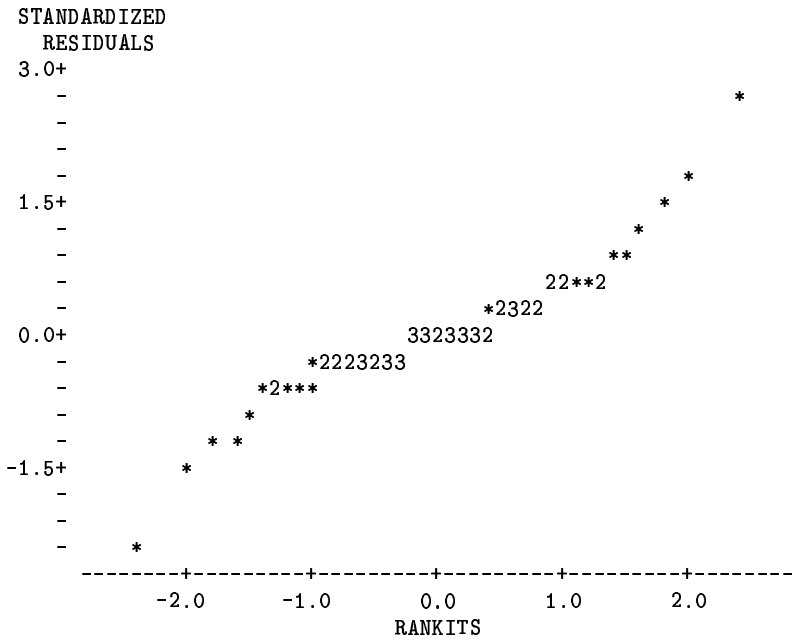


FIGURE 6.4. Standardized Residual–Rankit Plot

cells fitted too well by the model.

The MSE for the regression from which the residuals were obtained was .571. Under the sampling schemes considered, the MSE is a lack of fit test statistic. For large samples and a correct model, the MSE should be near 1. For incorrect models, it should be greater than 1. Again, the model fits better than we have any right to expect.

When examining leverages directly, we look for large individual leverages because they indicate that a cell is unlike the other cells. Large leverages are leverages near 1. Leverages also play a role in standardized residuals. Even if none of the leverages is large, they can be important. In particular, the range of leverages in this example is from .115367 to .632986. The factors $1/\sqrt{1-a_{ii}}$ are used to change Pearson residuals into standardized residuals. These factors range from 1.063 to 1.651, so a Pearson residual with the largest leverage would have a standardized residual more than 1.5 times larger than the same Pearson residual would have with the smallest leverage. This is not an inconsequential difference. Pearson residuals and standardized residuals can lead to very different conclusions.

Figure 6.5 contains a box plot of the Cook distances. Figure 6.6 contains an index plot of the Cook distances. Most cells have very little influence on the fit. The plots identify a number of relatively influential cases, but as compared to the calibrating values of $\frac{1}{27}\chi^2(27)$, none of the distances

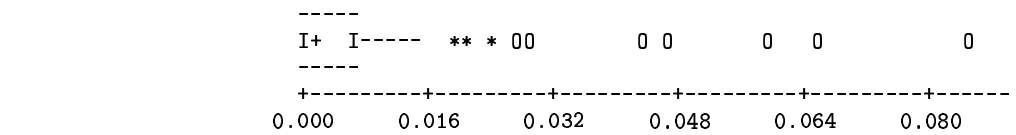


FIGURE 6.5. Cook's Distance–Box plot

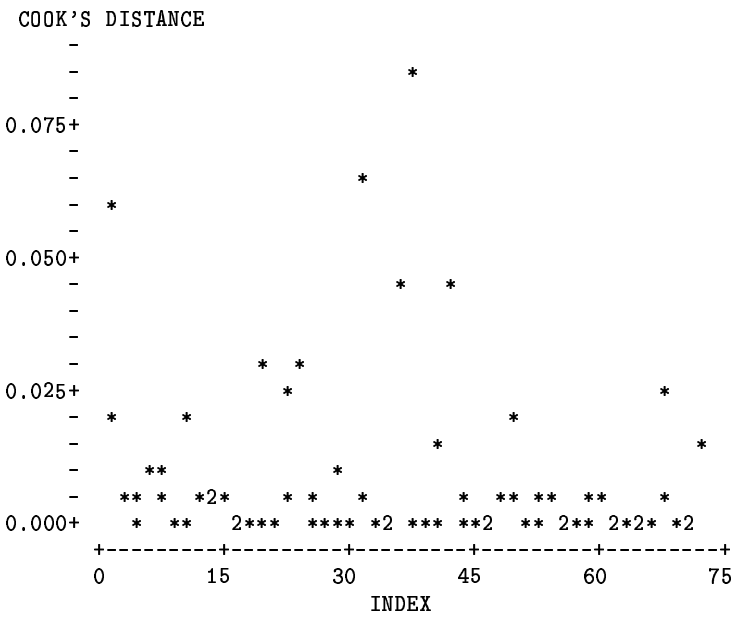


FIGURE 6.6. Cook's Distance–Index Plot

are substantial. For example, $\frac{1}{27}\chi^2(.5, 27) = 26.3/27 = .974$; none of the distances are anywhere near .974. The largest Cook's distance is the .085 for young nonwhite males who support legalized abortion.

6.7.2 Computing Commands

Below are commands that give the diagnostics provided by BMDP-4F.

```
/ INPUT     FILE = 'C:\LOGLIN\ABORT.DAT'.
          FORMAT = FREE.
          VARIABLES = 5.
/ VARIABLE  NAMES = R, S, A, O, N.
/ TABLE     INDEX = R, S, A, O.
            COUNT = N.
```

```
/ STAT      ALL.
/ PRINT     LINE = 79.
            LAMBDA.
            BETA.
            VAR.
            STAN.
            CHISQ.
            LRCHI.
/ FIT       MODEL = RSO, OA.
/ END
```

Diagnostics are also available from GLIM. The procedure for fitting log-linear models was illustrated in Subsection 3.7.1. The commands for diagnostics are exactly as in Subsection 4.4.2.

6.8 Drawing Conclusions

We have discussed model interpretation, model selection, and model validation. We have selected a model [RSO][OA] that is reasonably small, fits well, and has a simple interpretation. No cells seem to be unduly influential and no cells have outrageously bad fits. The model indicates that, given Opinion, Age is independent of Race and Sex. As discussed in the introduction to this chapter, the model is a description of the data and can be used to predict behavior in a similarly conducted study. It is not a statement about causation. In fact, it makes little sense to imagine that opinions about abortion cause the relative frequencies of Race and Sex to be independent of the frequencies of Age.

By itself, the model tells us nothing about the relationships among Race, Sex, and Opinion or about the relationship between Age and Opinion. Table 6.7 contains the estimated expected cell counts under the model [RSO][OA]. It is a complicated table, but much could be learned from studying it. Fortunately, there are easier ways to get at this information; we can collapse factors and study marginal tables.

As discussed in Section 5.3, the Race-Sex-Opinion relationships can be examined by collapsing over Age and looking at the Race-Sex-Opinion marginal table. This is given in Table 6.8. We see that whites are more likely to be in the survey than nonwhites. White females are a bit more likely to appear than white males. Among nonwhites, males and females are about equally likely. Ignoring the undecideds, the rate of support for legalized abortion is about the same for white and nonwhite males. It is higher for nonwhite females than white females. It is higher for white females than for white males. All of these things can be examined using odds ratios similar to Section 4.6. Formal inference requires standard errors as discussed in Section 10.2.

TABLE 6.7. Estimated Cell Counts under [RSO][OA]

			Age					
Race	Sex	Opinion	18-25	26-35	36-45	46-55	56-65	65+
White	Male	Support	105.5	132.1	114.5	76.94	72.81	79.19
		Oppose	41.87	61.93	54.95	47.98	52.34	61.93
		Undec.	1.39	2.50	5.27	5.27	5.54	8.04
	Female	Support	141.0	176.7	153.1	102.9	97.38	105.9
		Oppose	44.61	65.98	58.55	51.11	55.76	65.98
		Undec.	2.43	4.37	9.22	9.22	9.70	14.07
NonWhite	Male	Support	14.52	18.19	15.76	10.59	10.03	10.90
		Oppose	5.61	8.30	7.36	6.43	7.01	8.30
		Undec.	0.84	1.52	3.20	3.20	3.37	4.88
	Female	Support	19.97	25.01	21.67	14.57	13.79	14.99
		Oppose	3.91	5.79	5.14	4.48	4.89	5.79
		Undec.	0.35	0.62	1.32	1.32	1.39	2.01

TABLE 6.8. Race, Sex, Opinion Marginal Table

		Opinion			
Race	Sex	Support	Oppose	Undec.	Totals
White	Male	581	321	28	930
	Female	777	342	49	1168
Nonwhite	Male	80	43	17	140
	Female	110	30	7	147
	Totals	1548	736	101	2385

TABLE 6.9. Opinion, Age Marginal Table

	Age						
Opinion	18-25	26-35	36-45	46-55	56-65	65+	Totals
Support	281	352	305	205	194	211	1548
Oppose	96	142	126	110	120	142	736
Undec.	5	9	19	19	20	29	101
Totals	382	503	450	334	334	382	2385

To examine the relationship between Opinion and Age, we can collapse over Race and Sex. The marginal totals are given in Table 6.9. We see that some age groups are more likely to respond. There is more support than opposition in each age group. Undecideds increase with age. Also, the amount of support seems to decrease with age.

I find myself not really caring about the relative incidences of Race and Sex, but rather am interested in the relative support for legalized abortion among the different groups. This amounts to treating Opinion as a response variable and Race and Sex as explanatory variables; i.e., Race and Sex can be imagined to determine Opinion. In my experience, most contingency tables have one or more factors of particular interest that can be considered as response factors. (Of course, that is only in my experience.) Specific methods for analyzing tables with response factors were examined in Chapter 4.

If O were to be treated as a response and R, S, A as factors explaining that response, Asmussen and Edwards (1983) argue that, of the models listed in Example 6.6.2, only [RA][S][OA] is appropriate. Recall from Section 4.6 their contention that log-linear models are appropriate for response factors only if the model allows for collapsing over the response factors onto the explanatory factors, cf. Section 5.3. For example, the model [RSO][OA] can be collapsed over Race and Sex or collapsed over Age but not over the response factor opinion. Therefore, [RSO][OA] is not a reasonable log-linear model for the response factor O. It is illogical for Race and Sex to be independent of Age given the factor Opinion which is supposed to be a response. The response cannot generate independence between explanatory factors! On the other hand, models such as [RA][S][OA] are reasonable. Recall that [RA][S][OA] is one of the minimally adequate models found by Aitkin's method. However, you should also recall that Aitkin's method totally missed the important [RSO] interaction.

6.9 Exercises

EXERCISE 6.9.1. Reanalyze the auto accident data of Example 4.8.1 without treating any of the factors as a response factor.

EXERCISE 6.9.2. Reanalyze the abortion attitude data of Exercise 4.8.4 without treating any of the factors as a response.

EXERCISE 6.9.3. Using our discussion of graphical models and collapsibility in Chapter 5, argue that when the true model is [123][24][456], the test of marginal association for the $u_{123(ijk)}$'s does not ignore any vital information. Is the same conclusion appropriate when the true

model is [12][13][23][24][456]? What if the true model is [123][124][456], [123][24][456][15], or [123][24][456][15][36]?

Chapter 7

Models for Factors with Quantitative Levels

Just as in analysis of variance, if the levels of some factors are associated with quantitative values, these values can be used in the analysis. For example, in a two-factor ANOVA where factors are two kinds of fertilizer and levels are different quantitative amounts of fertilizers, an ANOVA would often examine linear and higher-order contrasts in the main effects and polynomial contrasts in the interaction (e.g., the linear-by-linear contrast).

In the analysis of categorical data, it is relatively rare that a factor has truly quantitative values associated with its levels. Often categories are ordered, but are not intrinsically quantitative. This is referred to as having *ordinal factor levels* or simply as having ordinal data. For example, socioeconomic status is often categorized as low, medium, and high. Surely, it would be advantageous to incorporate information about ordering into the analysis. The problem is in finding an appropriate method. The most commonly used method is to assign scores to the three levels of the factor. These scores can be any known numbers, say, x_1, x_2, and x_3. In fact, the most common method is to take $x_1 = 1$, $x_2 = 2$, and $x_3 = 3$. The analysis then proceeds as if the scores are true quantitative levels.

Alternatively, a factor might be income and the levels of the factor could be income intervals, say, less than \$20,000, \$20,000-\$40,000, and more than \$40,000. To have quantitative levels, we need one number associated with each level. Such numbers simply do not exist. As a practical matter, we could use the midpoints of the intervals as quantitative levels (scores). We can then develop models based on these approximate scores. Of course, if actual incomes are available for each individual, it would be more suitable to use the incomes in an appropriate regression analysis, cf. Section 4.1.

However, in practice it is not uncommon to encounter factors that have been created by categorizing continuous variables.

Continuous variables that have been categorized present some unique difficulties when assigning scores. With categories that are intervals, using the midpoints of the intervals is simple and appealing. In the income example above, the midpoint scores \$10,000 and \$30,000 may work reasonably well as quantitative levels for the first two income intervals. However there is no natural quantitative level to use for the category "more than \$40,000." Any analysis is dependent on the score that is chosen to represent the third category. Moreover, using midpoints may not be very efficient. Suppose it was known that most people in the less than \$20,000 interval had incomes near \$20,000. It would be better to use a score that was near \$20,000 rather than using the midpoint \$10,000. In practice, the method of determining a score will often depend on additional sources of information. If \$10,000 is not an appropriate score, there must be additional information leading to that conclusion. Use the same additional information to arrive at an alternative score.

In this chapter, we examine models that incorporate the quantitative nature of factor levels. When factor levels are not truly quantitative, the appropriateness of such models will be directly related to the appropriateness of the scores being used. We assume that the quantitative levels (i.e., scores) are known and consider linear models for the log of the expected cell counts (i.e., log-linear models). An alternative approach is to consider the scores as parameters and to estimate the scores. If the scores are parameters, then the models considered are no longer *linear* models for the log of the expected cell counts. Such models are discussed in Section 3.

7.1 Models for Two-Factor Tables

Consider a 3×4 table with quantitative levels x_1, x_2, x_3 and w_1, w_2, w_3, w_4. If the observations in the table were 12 normally distributed values y_{ij}, an analysis of variance would be appropriate. The model with no interaction is

$$y_{ij} = u + u_{1(i)} + u_{2(j)} + e_{ij} \ . \tag{1}$$

In this model, the 2 degrees of freedom for the main effect of Factor 1 can be broken into a linear contrast and a quadratic contrast. The three degrees of freedom for Factor 2 can be broken into a linear contrast, a quadratic contrast, and a cubic contrast. Equivalently, we can rewrite model (1) as a regression model

$$y_{ij} = u + \beta_1 x_i + \beta_2 x_i^2 + \eta_1 w_j + \eta_2 w_j^2 + \eta_3 w_j^3 + e_{ij} \, . \tag{2}$$

Now consider the full interaction model

$$y_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} + e_{ij} \, . \tag{3}$$

Note that with only one observation per cell, the terms $u_{12(ij)}$ are hopelessly confounded with the errors e_{ij}. Since our goal is only to draw analogies between analysis of variance and log-linear models, we need not concern ourselves with this confounding. To explore the interaction in model (3), we can consider contrasts in the interactions. Using the quantitative factor levels leads to considering things like the linear-by-linear, linear-by-quadratic, and quadratic-by-cubic interaction contrasts. In total, there are $(3-1)(4-1) = 6$ of these linearly independent interaction contrasts. Alternatively, we can rewrite model (3) using regression terms in place of the interactions,

$$y_{ij} = u + u_{1(i)} + u_{2(j)} + \gamma_{11}x_i w_j + \gamma_{12}x_i w_j^2 + \gamma_{13}x_i w_j^3 + \gamma_{21}x_i^2 w_j + \gamma_{22}x_i^2 w_j^2 + \gamma_{23}x_i^2 w_j^3 + e_{ij}.$$

This suggests a variety of partial interaction models that can be considered. The simplest of these models is

$$y_{ij} = u + u_{1(i)} + u_{2(j)} + \gamma_{11}x_i w_j + e_{ij}\,.$$

This is the model that provides for main effects in each factor, but models the interaction as consisting entirely of linear-by-linear interaction.

7.1.1 Log-Linear Models with Two Quantitative Factors

Exactly the same procedures are used when the data consist of counts. Consider an $I \times J$ table with quantitative levels $x_1, \ldots, x_I$ and $w_1, \ldots, w_J$. The model of independence is

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)}\,.$$

The model of full interaction (the saturated model) is

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}\,.$$

We can structure the interaction by considering a linear-by-linear association model

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + \gamma x_i w_j\,. \tag{4}$$

The maximum likelihood estimates $\hat{m}_{ij}$ must satisfy

$$\begin{aligned} \hat{m}_{i\cdot} &= n_{i\cdot}\,, \qquad i = 1, \ldots, I, \\ \hat{m}_{\cdot j} &= n_{\cdot j}\,, \qquad j = 1, \ldots, J, \end{aligned}$$

and

$$\sum_{ij} \hat{m}_{ij} x_i w_j = \sum_{ij} n_{ij} x_i w_j\,.$$

(This is easily seen from the results of Chapter 10.) Model (4) can be tested against the saturated model using either G^2 or X^2. The reduced model of independence can be tested against model (4) using either G^2, X^2, or $\hat{\gamma}/\mathrm{SE}(\hat{\gamma})$.

Often, model (4) is written in an equivalent form. If observations in all IJ cells are possible, model (4) is equivalent to

$$\log(m_{ij}) = u + u_{1(i)} + u_{2(j)} + \gamma(x_i - \bar{x}.)(w_j - \bar{w}.) \tag{5}$$

where $\bar{x}. = \frac{1}{I}\sum_{i=1}^{I} x_i$ and $\bar{w}. = \frac{1}{J}\sum_{j=1}^{J} w_j$. Frequently, the factor levels are equally spaced. This means that for some constants c and d, $x_{i+1} - x_i = c$, $i = 1, \ldots, I-1$, and $w_{j+1} - w_j = d$, $j = 1, \ldots, J-1$. This special case turns out to be equivalent to taking $x_i = i$, $i = 1, \ldots, I$ and $w_j = j$, $j = 1, \ldots, J$. Model (4) can be rewritten as

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + \gamma(i)(j)\,. \tag{6}$$

For equally spaced levels, model (6) is a reparametrization of model (4). The parameter γ in (6) is only identical to the parameter γ in (4) when the original scores are $x_i = i$ and $w_j = j$. In particular, γ in model (6) is equivalent to γdc in model (4). The scores $x_i = i$ and $w_j = j$ are used frequently when the factor levels are ordinal.

Model (4), when applied with scores that are equally spaced, is called the *uniform association* model. The name is apt because under this model, the odds ratios for consecutive table entries are identical. In particular,

$$\frac{m_{ij}m_{i+1\,j+1}}{m_{ij+1}m_{i+1\,j}} = e^{\gamma dc}\,,$$

$i = 1, \ldots, I-1$, $j = 1, \ldots, J-1$. To see this, note that

$$\begin{aligned}
&\log\left(\frac{m_{ij}m_{i+1\,j+1}}{m_{ij+1}m_{i+1\,j}}\right)\\
&\quad = \log m_{ij} - \log m_{ij+1} - \log m_{i+1\,j} + \log m_{i+1\,j+1}\\
&\quad = u + u_{1(i)} + u_{2(j)} + \gamma x_i w_j\\
&\qquad - u - u_{1(i)} - u_{2(j+1)} - \gamma x_i w_{j+1}\\
&\qquad - u - u_{1(i+1)} - u_{2(j)} - \gamma x_{i+1} w_j\\
&\qquad + u + u_{1(i+1)} + u_{2(j+1)} + \gamma x_{i+1} w_{j+1}\\
&\quad = \gamma[x_i w_j - x_i w_{j+1} - x_{i+1} w_j + x_{i+1} w_{j+1}]\\
&\quad = \gamma[x_i(w_j - w_{j+1}) - x_{i+1}(w_j - w_{j+1})]\\
&\quad = \gamma[x_i(-d) - x_{i+1}(-d)]\\
&\quad = \gamma d[x_{i+1} - x_i]\\
&\quad = \gamma dc\,.
\end{aligned}$$

If $x_i = i$ and $w_j = j$, then $d = 1$ and $c = 1,$, so consecutive log odds ratios equal γ.

The case of equal spacings arises very often, so we will always refer to model (4) and its equivalent, model (5), as the model of uniform association. If factor levels are not equally spaced, this terminology is meaningful in the sense that γ is a measure of association that applies uniformly, but any particular odds ratio must also be adjusted for differences in factor levels.

Finally, note that there is much more flexibility available than merely considering the independence model, the uniform association model, and the saturated model. The saturated model is equivalent to

$$\begin{aligned}\log m_{ij} \quad = \quad & u + u_{1(i)} + u_{2(j)} \\ & + \gamma_{1,1}x_i w_j + \gamma_{1,2}x_i w_j^2 + \cdots + \gamma_{1,J-1}x_i w_j^{J-1} \\ & + \gamma_{2,1}x_i^2 w_j + \gamma_{2,2}x_i^2 w_j^2 + \cdots + \gamma_{2,J-1}x_i^2 w_j^{J-1} \\ & \quad \vdots \\ & + \gamma_{I-1,1}x_i^{I-1} w_j + \gamma_{I-1,2}x_i^{I-1} w_j^2 + \cdots + \gamma_{I-1,J-1}x_i^{I-1} w_j^{J-1} .\end{aligned}$$

A wide variety of submodels can be fitted. For example, we could consider a second-order interaction model, say

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + \gamma_{1,1}x_i w_j + \gamma_{1,2}x_i w_j^2 + \gamma_{2,1}x_i^2 w_j + \gamma_{2,2}x_i^2 w_j^2 .$$

This model is larger than the uniform association model (5), but smaller than the saturated model. The maximum likelihood estimates for this model satisfy the equations

$$\begin{aligned}\hat{m}_{i\cdot} &= n_{i\cdot} , & i &= 1, \ldots, I, \\ \hat{m}_{\cdot j} &= n_{\cdot j} , & j &= 1, \ldots J, \\ \sum_{ij} x_i w_j \hat{m}_{ij} &= \sum_{ij} x_i w_j n_{ij} , \\ \sum_{ij} x_i w_j^2 \hat{m}_{ij} &= \sum_{ij} x_i w_j^2 n_{ij} , \\ \sum_{ij} x_i^2 w_j \hat{m}_{ij} &= \sum_{ij} x_i^2 w_j n_{ij} , \\ \sum_{ij} x_i^2 w_j^2 \hat{m}_{ij} &= \sum_{ij} x_i^2 w_j^2 n_{ij} .\end{aligned}$$

7.1.2 Models with One Quantitative Factor

Suppose that only the first factor in an $I \times J$ table has quantitative levels. Denote these levels as $x_1, \ldots, x_I$. We still want to consider models that are more general (larger) than the model of independence, but smaller than the saturated model. A frequently used model in this situation is the *column effects* model

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + \tau_j x_i . \tag{7}$$

This model implies that there is a linear effect on $\log m_{ij}$ from the rows of the table, but that the slope (τ) of this linear effect changes from column to column. In the case of I populations and $J = 2$ responses, model (7) is also a simple linear logistic regression model, cf. Section 2.6.

Although model (7) is appropriate when only one factor is quantitative, it can also be used when both factors are quantitative. Note that model (4) is a reduced model relative to model (7) in which the additional structure $\tau_j = \gamma w_j$ is imposed. Thus, the uniform association model assumes that the slopes change linearly with the columns. Model (7) is more general in that it allows arbitrary changes in the slopes. In particular, model (7) is equivalent to the model

$$\log(m_{ij}) = u + u_{1(i)} + u_{2(j)} + \gamma_{11} x_i w_j + \gamma_{12} x_i w_j^2 + \cdots + \gamma_{1,J-1} x_i w_j^{J-1} .$$

Maximum likelihood estimates for model (7) must satisfy

$$\begin{aligned} \hat{m}_{i\cdot} &= n_{i\cdot} , & i &= 1, \ldots, I, \\ \hat{m}_{\cdot j} &= n_{\cdot j} , & j &= 1, \ldots, J, \\ \sum_{i=1}^{I} \hat{m}_{ij} x_i &= \sum_{i=1}^{I} n_{ij} x_i , & j &= 1, \ldots, J . \end{aligned}$$

Testing is performed in the usual way.

More generally, we can consider any submodel of the saturated model. The saturated model can be reparametrized as

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + \tau_{1,j} x_i + \tau_{2,j} x_i^2 + \cdots + \tau_{I-1,j} x_i^{I-1} .$$

For $J = 2$, this is the $I - 1$-degree polynomial logistic regression model

$$\log \left(m_{i1} / m_{i2} \right) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_{I-1} x_i^{I-1} .$$

Of course, if the second factor in the table is quantitative rather than the first factor, we can write the saturated model as

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + \eta_{i,1} w_j + \cdots + \eta_{i,J-1} w_j^{J-1}$$

and consider reduced models. The *row effects* model

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + \eta_i w_j$$

is probably the most frequently used of these.

Although it seems to be done infrequently, there is no mathematical reason not to fit models using regression in place of main effects. For example, a reduced model relative to the uniform association model is

$$\log m_{ij} = u + \beta x_i + \eta w_j + \gamma x_i w_j .$$

This model also implies uniform association (in terms of the odds ratios when levels are equally spaced), but imposes additional constraints.

EXAMPLE 7.1.1. A sample of men between the ages of 40 and 59 was taken from the city of Framingham, Massachusetts. The men were cross-classified by their serum cholesterol and systolic blood pressure. We restrict attention to a subsample that did not develop coronary heart disease during a 6-year follow-up period. The data are given below.

Cholesterol	Blood Pressure (in mm Hg)				
(in mg/100 cc)	<127	127-146	147-166	167+	Totals
<200	117	121	47	22	307
200-219	85	98	43	20	246
220-259	119	209	68	43	439
≥260	67	99	46	33	245
Totals	388	527	204	118	1237

Consider four models:

Abbreviation	Model
[C][P][C_1]	$\log(m_{ij}) = u + u_{C(i)} + u_{P(j)} + C_{1i}(j)$
[C][P][P_1]	$\log(m_{ij}) = u + u_{C(i)} + u_{P(j)} + P_{1j}(i)$
[C][P][γ]	$\log(m_{ij}) = u + u_{C(i)} + u_{P(j)} + \gamma(i)(j)$
[C][P]	$\log(m_{ij}) = u + u_{C(i)} + u_{P(j)}$.

These are the row effects, column effects, uniform association, and independence models, respectively. The fits for the models relative to the saturated model are

Model	df	G^2	$A - q$
[C][P][C_1]	6	7.404	−4.596
[C][P][P_1]	6	5.534	−6.466
[C][P][γ]	8	7.429	−8.571
[C][P]	9	20.38	2.38

The best fitting model is

$$\log(m_{ij}) = u + u_{C(i)} + u_{P(j)} + \gamma(i)(j)\,.$$

Using the side conditions $u_{C(1)} = u_{P(1)} = 0$, the parameter estimates and standard errors are

Parameter	Estimate	Standard Error
u	4.614	.0699
$u_{C(1)}$	0	—
$u_{C(2)}$	−0.4253	.1015
$u_{C(3)}$	−0.0589	.1363
$u_{C(4)}$	−0.8645	.1985
$u_{P(1)}$	0	—
$u_{P(2)}$	0.0516	.0965
$u_{P(3)}$	−1.164	.1698
$u_{P(4)}$	−1.991	.2522
γ	0.1044	.0293

The estimated cell counts are

Estimated Cell Counts: Uniform Association

	Blood Pressure			
Cholesterol	<127	127-146	147-166	167+
<200	112.0	131.0	43.1	20.9
200-219	81.3	105.4	38.5	20.8
220-259	130.1	187.4	76.0	45.5
≥260	64.5	103.2	46.4	30.8

These are obtained from the uniform association model, so the odds ratios for consecutive table entries are identical. For example, the odds of blood pressure < 127 relative to blood pressure 127-146 for men with cholesterol < 200 are 1.11 times the similar odds for men with cholesterol of 200-219; up to roundoff error

$$\frac{112.0/131.0}{81.3/105.4} = \frac{112.0(105.4)}{81.3(131.0)} = e^{.1044} = 1.11$$

where $.1044 = \hat{\gamma}$. Similarly, the odds of blood pressure 127-146 relative to blood pressure 147-166 for men with cholesterol < 200 are 1.11 times the odds for men with cholesterol of 200-219:

$$\frac{131.0(38.5)}{105.4(43.1)} = e^{.1044} = 1.11\,.$$

Also, the odds of blood pressure < 127 relative to blood pressure 127-146 for men with cholesterol 200-219 are 1.11 times the odds for men with cholesterol of 220-259:

$$\frac{81.3(187.4)}{130.1(105.4)} = e^{.1044} = 1.11\,.$$

For consecutive categories, the odds of lower blood pressure are 1.11 times greater with lower blood cholesterol than with higher blood cholesterol.

The asymptotic 95% confidence interval for γ has end points $.1044 \pm 1.96(.0293)$. The interval is $(.047, .162)$. The corresponding interval for the odds ratio is $(e^{.047}, e^{.162})$ or $(1.05, 1.18)$. Thus, for consecutive categories, we are 95% confident that the odds of lower blood pressure are between 1.05 and 1.18 times greater with lower cholesterol than with higher cholesterol.

Of course, we can also compare nonconsecutive categories. For categories that are one step away from consecutive, the odds of lower blood pressure are $1.23 = e^{2(.1044)}$ times greater with lower cholesterol than with higher cholesterol. For example, the odds of having blood pressure < 127 compared to having blood pressure of $147 - 166$ with cholesterol < 200 are $1.23 = e^{2(.1044)}$ times those for cholesterol $200 - 219$. To check this, observe that

$$\frac{112.0(38.5)}{81.3(43.1)} = 1.23.$$

Similarly, the odds of having blood pressure < 127 compared to having blood pressure of 127-146 with cholesterol < 200 are 1.23 times those for cholesterol 220-259. Extending this leads to observing that the odds of having blood pressure < 127 compared to having blood pressure of 167+ with cholesterol < 200 are $2.559 = e^{9(.1044)}$ times those for cholesterol ≥ 260.

It is of interest to compare the estimated cell counts obtained under uniform association with the estimated cell counts under independence. The estimated cell counts under independence are

Estimated Cell Counts: Independence

	Blood Pressure			
Cholesterol	<127	127-146	147-166	167+
<200	96.3	130.8	50.6	29.3
200-219	77.2	104.8	40.6	23.7
220-259	137.7	187.0	72.4	41.9
≥260	76.85	104.4	40.4	23.4

With $\gamma > 0$, the uniform association model increases the estimated cell counts (relative to independence) for cells with (a) high cholesterol and high blood pressure and (b) low cholesterol and low blood pressure. Also, the uniform association model decreases the estimated cell counts for cells with (a) high cholesterol and low blood pressure and (b) low cholesterol and high blood pressure.

7.2 Higher-Dimensional Tables

The same basic methods used to incorporate quantitative levels into two-factor models can also be used in higher dimensions. For example, consider

an $I \times J \times K$ table with quantitative levels $x_1, \ldots, x_I$, $w_1, \ldots, w_J$, and $v_1, \ldots, v_K$. Assuming no three-factor interaction, there are three types of models that are particularly useful.

The *homogeneous uniform association* model is a model in which given a level for any factor, the remaining two factors display a uniform association. This model is

$$\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + \beta_1 x_i w_j + \beta_2 x_i v_k + \beta_3 w_j v_k \,.$$

Note that if $x_i = i$, $w_j = j$

$$\begin{aligned}
&\log\left(\frac{m_{ijk} m_{i+1\ j+1\ k}}{m_{i+1\ jk} m_{i\ j+1\ k}}\right) \\
&= u + u_{1(i)} + u_{2(j)} + u_{3(k)} + \beta_1 x_i w_j + \beta_2 x_i v_k + \beta_3 w_j v_k \\
&\quad + u + u_{1(i+1)} + u_{2(j+1)} + u_{3(k)} + \beta_1 x_{i+1} w_{j+1} \\
&\quad\quad + \beta_2 x_{i+1} v_k + \beta_3 w_{j+1} v_k \\
&\quad - u - u_{1(i+1)} - u_{2(j)} - u_{3(k)} - \beta_1 x_{i+1} w_j - \beta_2 x_{i+1} v_k - \beta_3 w_j v_k \\
&\quad - u - u_{1(i)} - u_{2(j+1)} - u_{3(k)} - \beta_1 x_i w_{j+1} - \beta_2 x_i v_k - \beta_3 w_{j+1} v_k \\
&= \beta_1 (x_i w_j - x_{i+1} w_j - x_i w_{j+1} + x_{i+1} w_{j+1}) \\
&= \beta_1 \,.
\end{aligned}$$

Thus, for any level of k, the log odds ratio for consecutive table entries equals β_1. Similarly, for j fixed, consecutive log odds ratios equals β_2; and for i fixed, consecutive log odds ratios equal β_3.

If Factor 3 does not have quantitative levels or if we merely wish to ignore the quantitative nature of the levels of Factor 3, we can write

$$\log(m_{ijk}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + \tau_{1k} x_i + \tau_{2k} w_j + \beta x_i w_j \,.$$

For each level of k, this model gives uniform associations. For a fixed level of i or a fixed level of j, odds ratios need not display uniform association.

If neither Factors 2 or 3 has quantitative levels or if we wish to ignore their quantitative nature, we can use the model

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{23(jk)} + \tau_{2j} x_i + \tau_{3k} x_i \,.$$

As before, models can be generalized by including powers of the x_i, w_j, and v_k scores. We can also model the three-factor interaction. The models

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + \beta x_i w_j v_k$$

and

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + \beta_1 w_j v_k + \beta_2 x_i v_k + \beta_3 x_i w_j + \gamma x_i w_j v_k$$

deal with the three-factor interaction while modeling two-factor interactions in alternative ways. Both of these models could be described as *heterogeneous uniform association* models.

EXAMPLE 7.2.1. In Chapter 6, we found that for the race, sex, opinion, age data, the model [RSO][OA] fits well. The ages are quantitative levels. We consider whether using the quantitative nature of this factor leads to a more succinct model. The age categories are 18-25, 26-35, 36-45, 46-55, 56-65, and 66+. For lack of a better idea, the category scores were taken as 1, 2, 3, 4, 5, and 6. Since the first and last categories are different from the other four, the use of the scores 1 and 6 are particularly open to question. Two models were considered:

Abbreviation	Model
[RSO][OA]	$\log(m_{hijk}) = u_{RSO(hij)} + u_{OA(jk)}$
[RSO][O_1]	$\log(m_{hijk}) = u_{RSO(hij)} + u_{A(k)} + O_{1j}k$
[RSO][O_1][O_2]	$\log(m_{hijk}) = u_{RSO(hij)} + u_{A(k)} + O_{1j}k + O_{2j}k^2$

Both of these are reduced models relative to [RSO][OA]. ([RSO][OA] is equivalent to $\log(m_{hijk}) = u_{RSO(hij)} + u_{A(k)} + O_{1j}k + O_{2j}k^2 + O_{3j}k^3 + O_{4j}k^4 + O_{5j}k^5$.) To compare models, we need the following statistics

Model	df	G^2
[RSO][OA]	45	24.77
[RSO][A][O_1][O_2]	51	26.99
[RSO][A][O_1]	53	29.33

Comparing [RSO][A][O_1] versus [RSO[OA] gives $G^2 = 29.33 - 24.77 = 4.56$ with degrees of freedom $53 - 45 = 8$. The G^2 value is not significant. Similarly, [RSO][A][O_1][O_2] is an adequate model relative to [RSO[OA]. The test for [O_2] has $G^2 = 29.33 - 27.99 = 1.34$ on 2 *df*, which is not significant. The model with only [O_1] fits the data well.

A primary difficulty with using quantitative factors is the necessity of assigning the factor scores. One way to avoid this problem is to estimate the factor scores. Methods for doing this are discussed in Section 3.

7.2.1 COMPUTING COMMANDS

Models with quantitative factors can be fit easily using several computer packages, e.g., SPLUS, GLIM, GENSTAT, and SAS PROC GENMOD. For example, the model $\log(m_{hijk}) = U_{RSO(hij)} + O_{1j}k + O_{2j}k^2$ can be fitted using SAS PROC GENMOD as given below.

```
options ps=60 ls=72 nodate;
data abort;
   infile 'abort.dat';
   input R S A O N;
   A2 = A * A;
```

```
proc genmod data=abort;
   class R S O;
   model N = R*S*O O*A O*A2 / link=log
                              dist=poisson;
run;
```

The key difference here from analysis of variance type models is that in the "class" command, age was not specified as a grouping (class) variable. The terms $O_{1j}k$ and $O_{2j}k^2$ are really interactions between the factor O and the predictors variables age and age squared. In the model statement, they are simply specified as interactions.

7.3 Unknown Factor Scores

The next step in generalizing the models of Sections 1 and 2 is to allow the factor scores to be unknown. In place of model (7.1.4), we assume the model

$$\log(m_{ij}) = \mu + \alpha_i + \beta_j + \gamma\nu_i\omega_j \tag{1}$$

where $\nu_i,\ i = 1, \ldots, I$, and $\omega_j,\ j = 1, \ldots, J$, are unknown parameters with

$$\sum_{i=1}^{I} \nu_i^2 = 1 = \sum_{j=1}^{J} \omega_j^2 \tag{2}$$

and

$$\alpha_\cdot = \beta_\cdot = \nu_\cdot = \omega_\cdot = 0.$$

The side conditions are imposed because the model is no longer log-linear. In general, for nonlinear models, the exact parametrization can be important in determining the properties of the model. The side conditions are necessary to have a well-defined parametrization. In particular, without condition (2), the parameter γ would not be well defined.

Model (1) is not log-linear, so the theoretical results used to justify fitting log-linear models do not apply. A separate theoretical development is required. Moreover, computer programs specifically developed for fitting log-linear models by maximum likelihood cannot be used to obtain the maximum likelihood fit of the model. Chuang (1983) used iteratively reweighted nonlinear least squares to obtain maximum likelihood estimates for models with unknown factor scores.

In addition to model (1), it is of interest to examine reduced models. In particular, the submodel of (1) with column main effects that are linear in the unknown factor scores is

$$\log(m_{ij}) = \mu + \alpha_i + \lambda\omega_j + \gamma\nu_i\omega_j \tag{3}$$

where

$$\alpha_\cdot = \nu_\cdot = \omega_\cdot = 0. \qquad \text{and} \qquad \sum_{i=1}^{I} \nu_i^2 = \sum_{j=1}^{J} \omega_j^2 = 1 \, .$$

Model (3) can also be written as

$$\log(m_{ij}) = \mu + \alpha_i + \nu_i \omega_j \tag{4}$$

with

$$\alpha_\cdot = \omega_\cdot = 0 \qquad \text{and} \qquad \sum_{j=1}^{J} \omega_j^2 = 1 \, .$$

Here, ν_i is equivalent to $\lambda + \gamma\nu_i$ in model (3), which is why the conditions $\nu_\cdot = 0$ and $\sum_{i=1}^{I} \nu_i^2 = 1$ are dropped. We can take model (4) one step further and write it as

$$\log(m_{ij}) = \alpha_{0i} + \alpha_{1i}\omega_j$$

where

$$\omega_\cdot = 0 \qquad \text{and} \qquad \sum_{j=1}^{J} \omega_j^2 = 1 \, .$$

The new parametrization is related to model (4) by $\alpha_{0i} \equiv \mu + \alpha_i$ and $\alpha_{1i} = \nu_i$. This version of the linear column effects model has the nice interpretation of *fitting separate lines in the unknown factor scores for each level of* i.

Similarly, if row effects are linear in the factor scores, the appropriate model is

$$\log(m_{ij}) = \mu + \lambda\nu_i + \beta_j + \gamma\nu_i\omega_j$$

with

$$\beta_\cdot = \nu_\cdot = \omega_\cdot = 0 \qquad \text{and} \qquad \sum_{i=1}^{I} \nu_i^2 = 1 = \sum_{j=1}^{J} \omega_j^2 \, .$$

This is equivalent to

$$\log(m_{ij}) = \mu + \beta_j + \nu_i\omega_j \tag{5}$$

and also to the separate lines model

$$\log(m_{ij}) = \beta_{0j} + \beta_{1j}\nu_i$$

where

$$\nu_\cdot = 0 \qquad \text{and} \qquad \sum_{i=1}^{I} \nu_i^2 = 1$$

in both models and $\beta_\cdot = 0$ in model (5).

If both main effects are linear, the model is

$$\log(m_{ij}) = \mu + \lambda_1 \nu_i + \lambda_2 \omega_j + \gamma \nu_i \omega_j$$

with

$$\nu_\cdot = \omega_\cdot = 0 \qquad \text{and} \qquad \sum_{i=1}^{I} \nu_i^2 = \sum_{j=1}^{J} \omega_j^2 = 1 \, .$$

An equivalent model is

$$\log(m_{ij}) = \mu + \nu_i + \omega_j + \gamma \nu_i \omega_j \tag{6}$$

where

$$\nu_\cdot = \omega_\cdot = 0 \, .$$

The relationship between the models is based on ν_i being equivalent to $\lambda_1 \nu_i$, ω_j being equivalent to $\lambda_2 \omega_j$, and γ being equivalent to $\gamma / \lambda_1 \lambda_2$.

When the factor categories are known to be ordered, a corresponding order can be imposed on the estimated factor scores. For example, models (1), (4), (5), and (6) can be fitted subject to the condition that $\nu_1 \leq \nu_2 \leq \cdots \leq \nu_I$. Similar orderings can be imposed on the column scores. Unfortunately, such constraints cause complications in the numerical procedures required to fit the models. In practice, it seems to be more common to fit the models without imposing order conditions on the scores. One can then verify whether the data are consistent with the a priori ordering.

Model (1) was first proposed by Fienberg (1968). Later, Goodman (1979, 1981), Anderson (1980), and Chuang (1983) extended the use of estimated factor scores. Johnson and Graybill (1972) proposed a model similar to (1) for standard analysis of variance. They built on earlier results in analysis of variance that are also of interest for log-linear models.

Tukey (1949) proposed a 1 degree of freedom test for nonadditivity in a two-way analysis of variance. Mandel (1961, 1971) extended Tukey's results by considering tests for more general models and presented a justification of the models based on unknown factor scores. Mandel's models differ from those considered by Johnson and Graybill in that they use the row and column effects in place of the unknown factor scores. These models and their justification apply equally well to log-linear models.

Mandel's models are

$$\log(m_{ij}) = \mu + \alpha_i + \beta_j + \delta_i \beta_j, \qquad \alpha_\cdot = \beta_\cdot = 0 \, , \tag{7}$$

and

$$\log(m_{ij}) = \mu + \alpha_i + \beta_j + \phi_j \alpha_i, \qquad \alpha_\cdot = \beta_\cdot = 0 \, . \tag{8}$$

The *Tukey model* is

$$\log(m_{ij}) = \mu + \alpha_i + \beta_j + \gamma \alpha_i \beta_j, \qquad \alpha_\cdot = \beta_\cdot = 0 \, . \tag{9}$$

It is easily seen that model (7) is equivalent to the linear row effects model (4). By equating

$$\beta_j = \omega_j$$

and

$$\delta_i = \nu_i - 1$$

and substituting into (4), we see that model (4) can be written as model (7). Conversely, if model (7) holds, write

$$\omega_j = \beta_j \Big/ \sqrt{\sum \beta_j^2}$$

and

$$\nu_i = (1 + \delta_i)\sqrt{\sum \beta_j^2}$$

to see that model (4) holds. Similarly, models (5) and (8) are equivalent and models (6) and (9) are identical. This establishes the justification for examining models (7), (8), and (9). They are equivalent to models with interesting interpretations in terms of underlying unknown factor scores.

If these models are appropriate and necessary, the maximum likelihood fits should be obtained. Maximum likelihood estimation and (generalized) likelihood ratio tests involving any of the models for unknown factor scores require specialized methods for fitting the log-nonlinear models. Standard programs for fitting log-linear models are not appropriate. However, by analogy with the two-stage fitting procedure commonly used in analysis of variance, a simple method can be derived for evaluating whether these models are necessary. All of the models considered contain the model of complete independence

$$\log(m_{ij}) = \mu + \alpha_i + \beta_j \tag{10}$$

as a submodel. This may not be obvious in models (4), (5), and (6), but it is in their equivalent versions (7), (8), and (9).

Models (7), (8), and (9) can be tested against model (10) in a very simple way. First, write

$$\tau_{ij} = \mu + \alpha_i + \beta_j$$

and note that if model (10) is true, (7) is equivalent to

$$\log(m_{ij}) = \mu + \alpha_i + \beta_j + \delta_i \tau_{ij}\,, \tag{11}$$

(8) is equivalent to

$$\log(m_{ij}) = \mu + \alpha_i + \beta_j + \phi_j \tau_{ij}\,, \tag{12}$$

and (9) is equivalent to

$$\log(m_{ij}) = \mu + \alpha_i + \beta_j + \gamma(\tau_{ij})^2\,. \tag{13}$$

Models (11), (12), and (13) would be log-linear if the τ_{ij}'s were known. Fit model (10) by maximum likelihood to obtain $\hat{\tau}_{ij}$. Substitute $\hat{\tau}_{ij}$ for τ_{ij} in models (11), (12), and (13) so that they are log-linear in the other parameters. Fit the models based on $\hat{\tau}_{ij}$ using standard log-linear methods and test them against model (10) in the usual way. If model (10) is true, these tests have asymptotic chi-squared distributions. For linear models, the validity of tests based on this two-stage fitting procedure was established by Milliken and Graybill (1970) and Rao (1965). For log-linear models, a corresponding result is given by Christensen and Utts (1992).

These models and methods can also be applied to higher-dimensional tables and to logit models. Example 7.4.1 gives the details for fitting a logit model.

EXAMPLE 7.3.1. Wing (1962), Haberman (1974b), and Fienberg (1980) have considered data on the relationship between length of hospitalization and frequency of visits for 132 long-term schizophrenic patients. Length of hospitalization was categorized as over 2 years but under 10, $(2, 10)$, over 10 years but under 20, $(10, 20)$, and over 20 years, 20+. Frequency of visits were regular, irregular (no home visits, hospital visits less than once a month), and never. The data are given in Table 7.1.

TABLE 7.1. Schizophrenic Data

Visitation Frequency (i)	Length of Hospitalization (j) in years $(2, 10)$	$(10, 20)$	20+
Regular	43	16	3
Irregular	6	11	10
Never	9	18	16

Fitting model (10) in the usual way and using the two-stage fitting procedure described above for the other models yields the lack of fit statistics given below.

Model	*df*	Two-Stage G^2
(10)	4	38.35
(11)	2	1.21
(12)	2	11.18
(13)	3	14.76

Except for model (10), these G^2's are not likelihood ratio lack of fit statistics for the models. However, the G^2 for model (10) can be subtracted from the other G^2's to obtain valid asymptotic χ^2 tests for model (10) versus the other models. The test statistics are as follows:

Model	df	Two-Stage G^2
(11)	2	37.14
(12)	2	27.17
(13)	1	23.59

All of the models fit better than (10), especially model (11).

One set of parameter estimates are given below for the two-stage fit of model (11). Recall that parameter estimates are not uniquely defined.

Parameter	Estimate
μ	-12.91
α_1	0.000
α_2	14.65
α_3	15.23
β_1	0.000
β_2	0.4727
β_3	0.4977
δ_1	5.034
δ_2	0.05197
δ_3	0.000

Observe that $\hat{\delta}_1 > \hat{\delta}_2 > \hat{\delta}_3 = 0$ with $\hat{\delta}_1$ much larger than the others. To draw conclusions about the meaning of the $\hat{\delta}$'s, we need to examine their multipliers in model (11), the $\hat{\tau}_{ij}$'s. The $\hat{\tau}_{ij}$'s are

	i		
j	1	2	3
1	3.305	3.051	2.612
2	2.473	2.220	1.780
3	2.939	2.685	2.246

In each row, the $\hat{\tau}_{ij}$'s are decreasing. With non-negative $\hat{\delta}$'s, as we move to the right in each row, the fitted counts decrease. For patients who are visited irregularly or never, the model indicates little decrease over time due to the interaction because the $\hat{\delta}$ values are near zero. However, $\hat{\delta}$ is large for row 1, so, according to the fitted model, the number of patients who have regular visits decreases dramatically over time.

EXERCISE 7.1. Show that models (7) and (11) are equivalent. Show that models (9) and (13) are equivalent.

7.4 Logit Models

Results analogous to Section 3 apply to models for the log odds. The example in this section involves a logit model with factors that are assumed

to have unknown quantitative category scores.

EXAMPLE 7.4.1. Rosenberg (1962) presents data on the relationships among Religion, Father's Educational Level, and Self-Esteem. The data are given in Table 7.2. Self-Esteem is considered the response.

TABLE 7.2. Data of Rosenberg (1962).

		Father's Educational Level					
Religion	Self-Esteem	8th or less	Some HS	HS Grad	Some Coll	Coll Grad	Post Coll
Catholic	High	245	330	388	100	77	51
	Low	115	152	153	40	37	19
Jewish	High	28	89	102	67	87	62
	Low	11	37	35	18	12	13
Protestant	High	125	234	233	109	197	90
	Low	68	91	173	47	82	32

Chuang (1983) presents a maximum likelihood analysis of Rosenberg's data that includes logit versions of Mandel's models (7.3.7) and (7.3.8) and the Tukey model (7.3.9). The expected cell counts are m_{ijk}, where i denotes religion, j denotes educational level, and k denotes self-esteem. A logit model imposes structure on

$$\tau_{ij} \equiv \log(m_{ij1}/m_{ij2})\,.$$

The logit versions of Mandel's models are

$$\tau_{ij} = \mu + \alpha_i + \beta_j + \delta_i\beta_j \tag{1}$$

and

$$\tau_{ij} = \mu + \alpha_i + \beta_j + \phi_j\alpha_i\,. \tag{2}$$

The logit version of the Tukey model is

$$\tau_{ij} = \mu + \alpha_i + \beta_j + \gamma\alpha_i\beta_j\,. \tag{3}$$

The additive model is

$$\tau_{Xij} = \mu + \alpha_i + \beta_j\,. \tag{4}$$

In Section 3, the models (7.3.11), (7.3.12), and (7.3.13) were used to simplify the two-stage fitting process. Their logit model analogues are

$$\tau_{ij} = \mu + \alpha_i + \beta_j + \delta_i\tau_{Xij}\,, \tag{5}$$

$$\tau_{ij} = \mu + \alpha_i + \beta_j + \phi_j\tau_{Xij}\,, \tag{6}$$

and

$$\tau_{ij} = \mu + \alpha_i + \beta_j + \gamma(\tau_{Xij})^2 . \tag{7}$$

As in Section 3, these models are equivalent to models (1), (2), and (3), respectively.

To fit (5), (6), and (7) using maximum likelihood requires the use of something other than a standard log-linear model or logistic model computer program. Chuang (1983) suggests modifying a nonlinear least squares program. The alternative two-stage procedure discussed in Section 3 is easily implemented with standard software. Begin by fitting model (4) to obtain the $\hat{\tau}_{Xij}$'s. The only nonlinear aspect to models (5), (6), and (7) is the presence of the τ_{Xij} parameters, so if these parameters were known, there would be no difficulty in obtaining fits for the models. In the two-stage procedure, the estimates $\hat{\tau}_{Xij}$ are substituted for the parameters. The resulting linearized models are fitted in the usual way with standard software. Test statistics for comparing the models to model (4) are also computed in the usual way.

Table 7.3 presents the results of fitting models (5), (6), and (7) by both maximum likelihood and the two-stage procedure. The G^2 values reported in Table 7.3 for the two-stage fits are not directly applicable because they are statistics for testing the models against the saturated model. The theoretical justification given in Christensen and Utts (1992) applies only to testing models (5), (6), and (7) against the additive model (4). Although the G^2's reported in Table 7.3 do not have a sound theoretical basis as test statistics, they are sometimes a valuable data analytic tool. This is not unreasonable because they are one-to-one functions of test statistics that have a sound basis.

TABLE 7.3. Model Fits

Model	*df*	MLE G^2	Two-Stage G^2
(5)	8	12.76	16.34
(6)	5	12.07	13.25
(7)	9	25.58	26.34
(4)	10	26.39	—

The results of testing the models with nonadditivity against the additive model are given in Table 7.4. In this example, the two-stage tests appear to be a little less powerful than the generalized likelihood ratio tests, but the qualitative conclusions about the best-fitting model are identical for the two methods. Model (5), i.e., model (1), appears to be the best-fitting model. Recall from Section 3 that this is the model that, for each religion, fits a line in the unknown factor scores associated with Father's Educational Level.

TABLE 7.4. Tests of Nonadditivity

Model	*df*	MLE G^2	Two-Stage G^2
(5)	2	13.63	10.05
(6)	5	14.32	13.14
(7)	1	0.81	0.05

If one of the log-nonlinear models is to be used in further work, an exact maximum likelihood fit of the model should be used. Nonetheless, the two-stage fitting procedure provides a simple yet valid diagnostic tool for checking whether Mandel's models and the Tukey model require more investigation.

7.5 Exercises

EXERCISE 7.5.1. Reanalyze the Intelligence versus Clothing table of Exercise 2.6.3 using the methods of Section 1. Note that this ignores the potentially complicating factor Standard and the complex sampling scheme.

EXERCISE 7.5.2. For the model

$$\log(m_{ij}) = u + u_{1(i)} + u_{2(j)} + \gamma_1 i\, j + \gamma_2 i^2 j\,,$$

find the log odds ratio

$$\log[m_{ij} m_{i+1\ j+1} / m_{i+1\ j} m_{i\ j+1}]$$

in terms of the model parameters.

EXERCISE 7.5.3. Duncan, Schuman, and Duncan (1973) and Duncan and McRae (1979) present data on evaluations made in 1959 and 1971 of the performance of radio and TV networks. The data are given in Table 7.5. Use the methods of Section 2 to analyze these data.

TABLE 7.5. Radio and Television Network Performance.

Year	Respondent's Race	Performance of Networks Poor	Fair	Good
1971	White	158	636	600
	Black	24	144	224
1959	White	54	253	325
	Black	4	23	81

EXERCISE 7.5.4. Assuming the use of consecutive integer scores, find the log odds ratios

$$\log[m_{ijk}m_{i+1\ j+1\ k}/m_{i+1\ jk}m_{i\ j+1\ k}]$$

and

$$\log[m_{ijk}m_{i+1\ j\ k+1}/m_{i+1\ jk}m_{i\ j\ k+1}]$$

in terms of the model parameters for the two heterogeneous uniform association models of Section 2.

EXERCISE 7.5.5. Use the methods of Section 3 to analyze the Intelligence – Clothing table of Exercise 2.6.3.