Chapter 8
Fixed and Random Zeros

Not infrequently, one encounters a table in which a number of the cell counts are 0s. These cells sometimes cause problems when fitting log-linear models. Recall that in our discussion of multiple logistic regression in Section 4.1, we had a 200×2 table that was riddled with zero counts. Except for the fact that some asymptotic results did not hold, the zeros caused no problems. Cells with zero counts merely have the potential to cause problems.

Cells with zero counts are classified in two ways: fixed and random. *Fixed zeros* are cells in which it is impossible to observe counts. Such cells must always be zero. *Random zeros* are cells that happen to have zero counts, but where it is possible to have positive counts. Tables with fixed zeros are called *incomplete tables*.

8.1 Fixed Zeros

EXAMPLE 8.1.1. Brunswick (1971) reports data on the health concerns of teenagers. These data have also been examined by Grizzle and Williams (1972) and Fienberg (1980). The data are given in Table 8.1. The two zeros in the table are fixed. It is physiologically impossible for males (of whatever age) to have menstrual difficulties. (Technically, I suppose the real issue here is whether males worry about menstrual difficulties. I suppose somewhere, sometime, some teenage male has worried about them, but one must admit that these are very nearly fixed zeros. We will treat them as such.)

TABLE 8.1. Health Concerns of Teenagers

		Health Concerns (k)			
Sex (i)	Age (j)	Sex, Reproduction	Menstrual Problems	How Healthy I Am	Nothing
Male	12-15	4	0	42	57
	16-17	2	0	7	20
Female	12-15	9	4	19	71
	16-17	7	8	10	31

The solution to dealing with fixed zeros is to throw them away. They are cells that do not really exist. One simply ignores those cells and fits a model to the cells that do exist. Thus, the model is fitted to an incomplete table.

In fact, it is impossible to include fixed zeros in a log-linear model. A log-linear model specifies a value of $\log(m_i)$ for every cell i. It is implicit that $\log(m_i)$ is defined. If a cell is a fixed zero, the probability of an observation occurring in that cell is zero, so the expectation, m_i, must also be zero. In such cases, $\log(m_i)$ is undefined. To fit log-linear models, one has to throw fixed zeros away.

If the Newton-Raphson algorithm of Chapter 10 is used to fit log-linear models, throwing out fixed zeros is no problem. The algorithm can easily handle the fact that not all combinations of the factor categories are considered in the model.

When using iterative proportional fitting, the situation is slightly more complex. The algorithm is based on having all combinations of the factor categories defined. Fortunately, there is a simple way around this problem. Recall that it is standard to start iterative proportional fitting with all initial cell estimates equal to 1 and that if the initial values satisfy the constraints of the model, the subsequent iterations also satisfy the constraints of the model. With all initial values of 1, the initial values satisfy the constraints of any interesting ANOVA model. To deal with fixed zeros, take the initial values of the corresponding cells to be 0. It is easily seen that if the initial value is 0, then all subsequent values will also be 0. (We use the definition that $0/0 = 0$.) Moreover, the constraints on the model with fixed zero cells eliminated are identical to the constraints on the complete model when the fixed zero cells are required to have fitted values of 0.

Although one can get the correct fitted values using iterative proportional fitting, the user often must provide the correct degrees of freedom. These are determined by the model with the fixed zero cells eliminated.

EXAMPLE 8.1.2. We illustrate the computation of degrees of freedom using the data of Example 8.1.1. Consider the model

$$\log(m_{ijk}) = M + S_i + A_j + H_k + (SA)_{ij} + (SH)_{ik} + (AH)_{jk}\,. \quad (1)$$

Clearly, the grand mean M can exist and we have data available for each sex, age, and health concern. Computing these degrees of freedom as usual, we have

Term	*df*
M	1
S	1
A	1
H	3

We also have data available for every combination of sex and age, and every combination of age and health concern. Again, computing degrees of freedom in the usual way, we get

Term	*df*
SA	1
AH	3

We do not have data available for every combination of sex and health concern. We have no data for males with menstrual difficulties. In the 2×4 table of sex and health concerns, we have only 7 cells instead of 8. The degrees of freedom for this table must be 7. To fit this table perfectly, we would use the parameters M, S_i, H_k, and $(SH)_{ik}$. The total number of degrees of freedom for these parameters must be 7. Because M has 1 degree of freedom, S has 1, and H has 3, that leaves only 2 degrees of freedom available for SH. Thus, model (1) has

$$1 + 1 + 1 + 3 + 1 + 2 + 3 = 12$$

degrees of freedom. The saturated model has 1 degree of freedom for each cell. There are nominally 16 cells, but 2 are fixed zeros, so the table has only 14 cells. For testing model (1) against the saturated model, the degrees of freedom are $14 - 12 = 2$.

These techniques are easily applied to all models for the health concern data. In testing [SA][H] against the saturated model, we have dropped the SH and AH terms. The degrees of freedom for these terms are put into the test degrees of freedom. Model (1) has 2 degrees of freedom for the test, SH has 2 degrees of freedom, and AH has 3 degrees of freedom, so the test of [SA][H] has $2 + 2 + 3 = 7$ degrees of freedom. Alternatively, we can do the calculation by noting that the saturated model has 14 *df*, while the [SA][H] model has terms and degrees of freedom $M(1)$, $S(1)$, $A(1)$, $SA(1)$, $H(3)$ for a total of 7 degrees of freedom. The test has $14 - 7 = 7$ degrees of freedom.

The fits for these data are summarized below.

Model	df	G^2
[SA][SH][AH]	2	2.03
[SH][AH]	3	4.86
[SA][AH]	4	13.45
[SA][SH]	5	9.43
[SA][H]	7	22.03
[SH][A]	6	15.64
[AH][S]	5	17.46
[S][A][H]	8	28.24

The best fitting model is either [SA][SH][AH] or [SH][AH], depending on whether health concerns are treated as a response variable or not.

One final note on fixed zeros. Because fixed zeros are really cells that do not exist, the existence of fixed zeros has no effect on the validity of large sample results. If the counts in the other cells are large, asymptotic results hold.

8.2 Partitioning Polytomous Variables

We now investigate a method that uses incomplete tables to examine category effects.

EXAMPLE 8.2.1. Duncan (1975) presents data on the earth-shattering question, "Who should shovel the snow from sidewalks?" The data are given in Table 8.2. Mothers were asked whether boys, girls, or both should do the shoveling. Mothers never responded that girls alone should do the shoveling, so the data are presented with only two categories. In addition, there are two explanatory factors, the mother's religion (R), Protestant, Catholic, Jewish, or other, and the year (Y) in which the question was asked. (Having lived 36 years in Minnesota and Montana, I am aware that a key factor has been left out. Father does the vast majority of the shoveling.)

We will treat mothers' opinions on shoveling (S) as a response variable and consider only log-linear models that correspond to logit models. The point of this example is to illustrate how to use tables with fixed zeros to answer questions about parameters in log-linear models.

Fitting the standard models gives

TABLE 8.2. Mothers' Opinions on Who Should Shovel Snow

Religion (i)	Year (j)	Shoveling (k) Boy	Both
Protestant	1953	104	42
	1971	165	142
Catholic	1953	65	44
	1971	100	130
Jewish	1953	4	3
	1971	5	6
Other	1953	13	6
	1971	32	23

Model	df	G^2
[RY][RS][YS]	3	0.4
[RY][RS]	4	21.5
[RY][YS]	6	11.2
[RY][S]	7	31.7

Clearly, the best fitting model is [RY][RS][YS]. We can write the model as

$$\log(m_{ijk}) = (RY)_{ij} + S_k + (RS)_{ik} + (YS)_{jk}\,, \tag{1}$$

$i = 1,2,3,4$, $j = 1,2$, $k = 1,2$. Because S is a response variable, the $(RY)_{ij}$'s must be in the model. The important terms are S_k, $(RS)_{ik}$, and $(YS)_{jk}$. In a logit model, S_k corresponds to the grand mean; not very interesting. The terms $(YS)_{jk}$ correspond to main effects in years with 1 degree of freedom. The terms $(RS)_{ik}$ correspond to main effects in religion with 3 degrees of freedom. Further analysis must examine the nature of the three degrees of freedom in (RS). We are really asking about relationships among the religion categories.

One way to proceed was illustrated in Section 4.6. We could incorporate constraints on the religions. For instance, we could treat all Protestants and Jews alike, while allowing Catholics and Others to have separate effects on the shoveling response. Such a procedure would involve recoding the indices. We could recode the index i into a new pair of indices g and h as follows:

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

where g indicates religion with no difference between Protestants and Jews, while h is simply used to tell Protestants and Jews apart. We can rewrite

model (1) as

$$\log(m_{ghjk}) = (RY)_{ghj} + S_k + (RS)_{ghk} + (YS)_{jk} \ . \tag{2}$$

To eliminate differences between Protestants and Jews in (RS), we drop the h, giving

$$\log(m_{ghjk}) = (RY)_{ghj} + S_k + (RS)_{gk} + (YS)_{jk} \ . \tag{3}$$

Both models (2) and (3) are actually models for incomplete tables. The possible values for g are 1, 2, 3. For h, they are 1, 2. For j and k, they are also 1 and 2. This suggests the existence of a $3 \times 2 \times 2 \times 2$ table. However, not all combinations of the indices are possible. Only when g is 1 can h be 2. Any cell $(g, 2, j, k)$ in the $3 \times 2 \times 2 \times 2$ table with $g = 2$ or 3 is a fixed zero.

It is obvious that quite a few questions can be addressed by creative reindexing. Duncan (1975) presented a particular pattern that is very flexible. Transform the index i into (r, s, t, u) where the correspondence is as follows:

i	1	2	3	4
(r, s, t, u)	(1,2,2,2)	(2,1,2,2)	(2,2,1,2)	(2,2,2,1)

Each 4-tuple has three 2s and one 1. If the 1 is in the third place, then the 4-tuple corresponds to $i = 3$, etc. The index r is 1 for Protestant and 2 otherwise. The index s is 1 for Catholic and 2 otherwise. Similarly, $t = 1$ indicates Jewish and $u = 1$ indicates Other. Including the year and shoveling factors, we now have a $2 \times 2 \times 2 \times 2 \times 2 \times 2$ table with many fixed zeros because mothers have only one religion. Using a natural identification, denote the factors corresponding to r, s, t, and u as P, C, J, and O. Model (1) can now be rewritten as

$$\log(m_{rstujk}) = (PCJOY)_{rstuj} + S_k + (PCJOS)_{rstuk} + (YS)_{jk} \ .$$

Moreover, we can denote this as [PCJOY][PCJOS][YS]. Note that the four factors P, C, J, O taken together are equivalent to the old factor R.

Now consider the model [PCJOY][YS][CS]. Recall that [PCJOY] is included because S is a response factor. The term [YS] seemed important in our original analysis, so it is retained. The new model replaces the terms $(RS)_{ik} = (PCJOS)_{rstuk}$ in model (1) with the terms $(CS)_{sk}$. In particular, the model is

$$\log(m_{rstujk}) = (PCJOY)_{rstuj} + S_k + (CS)_{sk} + (YS)_{jk} \ .$$

The logit model effect of the four different religions is being replaced with one effect that distinguishes Catholics from all other religions.

Earlier, we considered a model that treated Protestants and Jews the same, but allowed separate effects for Catholics and Others. In Duncan's setup, this is the model [PCJOY][YS][COS]. In the model

$$\log(m_{rstujk}) = (PCJOY)_{rstuj} + S_k + (COS)_{tuk} + (YS)_{jk} ,$$

the effect of religion on mothers' shoveling opinions is taken up by the $(COS)_{tuk}$ terms. Only three such terms exist: $(COS)_{12k}$, an effect for Catholics; $(COS)_{21k}$, an effect for others; and $(COS)_{22k}$. The $(COS)_{22k}$ term does not distinguish between Protestants and Jews.

Moreover, because the $2 \times 2 \times 2 \times 2 \times 2 \times 2$ table is so very incomplete, there is another way to do exactly the same thing. The model [PCJOY][YS][CS][OS] is equivalent to [PCJOY][YS][COS]. The model for [PCJOY][YS][CS][OS] is

$$\log(m_{rstujk}) = (PCJOY)_{rstuj} + S_k + (CS)_{sk} + (OS)_{uk} + (YS)_{jk} .$$

Catholics get the effects $(CS)_{1k} + (OS)_{2k}$. Others get the distinct effects $(CS)_{2k} + (OS)_{1k}$. Protestants and Jews both get $(CS)_{2k} + (OS)_{2k}$.

In a complete table, the 15 interactions between S and the religion factors P, C, J, and O would each have 1 degree of freedom. In fact, there are only 3 degrees of freedom for $(RS) = (PCJOS)$. Thus, there are a lot of redundant models. In fact, any term that includes three of P, C, J, and O is equivalent to any other term that includes three, and these are all equivalent to the four-factor term. This follows from the fact that any three of the indices r, s, t, and u completely determine the cell; e.g., if $r = 2$, $s = 2$, $t = 2$, then we must have $u = 1$. Because all terms with three of the factors are the same, models such as [PCJOY][YS][PCJS] and [PCJOY][YS][CJOS] are equivalent to each other and to [PCJOY][YS][PCJOS].

Although Duncan's method of reindexing is flexible, it is not a panacea. There are interesting questions that cannot be addressed using Duncan's method. For example, if we wanted a model that treated Protestant and Jews the same and also treated Catholics and Others the same, we could not arrive at such a model using Duncan's method.

Now let's see where Duncan's method gets us with the shoveling data. Some models and fits are given below:

Model	*df*	G^2
[PCJOY][PCJOS][YS] = [RY][RS][YS]	3	0.4
[PCJOY][PS][YS]	5	4.8
[PCJOY][CS][YS]	5	1.4
[PCJOY][JS][YS]	5	10.9
[PCJOY][OS][YS]	5	9.8
[PCJOY][YS] = [RY][YS]	6	11.2

The models that involve [JS] and [OS] do not fit very well. The model with [JS] only distinguishes between Jews and non-Jews. The relatively

poor fit indicates that Protestants, Catholics, and Others do not act the same. Similarly, the relatively poor fit of the model with [OS] indicates that Protestants, Catholics, and Jews do not act the same.

Both the models with [PS] and [CS] fit reasonably well. The model with [CS] fits especially well. This model suggests that Protestants, Jews, and Others act the same, but Catholic mothers have different opinions about who should shovel snow. The model with [PS] indicates that Catholics, Jews, and Others act the same, but Protestant mothers have different opinions. There are so few Jews and Others that it is not surprising that lumping them with either large category does not substantially hurt the fit. What we really know is that Protestants mothers have different attitudes than Catholic mothers and that if we want to lump Jewish and Other mothers with one of those categories, they seem to fit better with the Protestants than the Catholics. However, we know almost nothing about Jewish mothers and little about Other mothers. From looking at Table 8.2, it is clear that the Catholic mothers are much more egalitarian about shoveling than the Protestants or the Others.

We could go further with this analysis by considering models [RY][YS][CS][PS], [RY][YS][CS][JS], and [RY][YS][CS][OS], but considering what little difference there is between [RY][YS][CS] and [RY][YS][RS], there seems little point in pursuing the analysis further. Note that I have gotten lazy and started writing [RY] for [PCJOY].

8.3 Random Zeros

Random zeros are cells that have positive probability of occurring, but do not occur in the sample at hand. If a large enough sample was taken, these cells should eventually contain positive counts.

Random zeros present three problems. First, they suggest that asymptotic results are invalid. If the sample was large enough for asymptotic approximations, these cells ought not be zero. The discussions of small samples in Section 2.4 and conditional inference in Section 3.5 are relevant to this problem. Second, when a table includes random zeros, maximum likelihood estimates of the parameters may not exist. Finally, there is a practical problem in that some computer programs will give "MLEs" even when they do not exist. In this section, we concern ourselves primarily with problems related to the nonexistence of MLEs.

EXAMPLE 8.3.1. Consider a $4 \times 2 \times 3 \times 3$ table on the results of arthroscopic knee surgery. The four factors and their categories are listed as follows:

Factor Label	Factor Description	Categories
Type	Type of injury	Twist, Direct, Both, No injury
Sex	Sex of patient	Male, Female
Age	Age of patient	11-30, 31-50, 51-91
Result	Outcome of surgery	Excellent, Good, Fair-Poor

The data are given in Table 8.3. First, note that one cannot fit a saturated log-linear model. For a saturated model, $n_{hijk} = \hat{m}_{hijk}$. Because the model is log-linear, $\log(\hat{m}_{hijk})$ must be defined. However, for some cells, $n_{hijk} = 0$, so either $n_{hijk} \neq \hat{m}_{hijk}$ or $\log(\hat{m}_{hijk})$ is not defined.

TABLE 8.3. Data on Arthroscopic Knee Surgery

			Result (k)		
Type (h)	Sex (i)	Age (j)	Ex	Good	F-P
		11-30	21	11	4
	Male	31-50	32	20	5
		51-91	20	12	5
Twist					
		11-30	3	1	0
	Female	31-50	6	5	2
		51-91	6	3	1
		11-30	3	2	2
	Male	31-50	2	4	4
		51-91	0	0	0
Direct					
		11-30	0	1	1
	Female	31-50	0	0	0
		51-91	1	2	3
		11-30	7	1	1
	Male	31-50	11	6	2
		51-91	0	4	6
Both					
		11-30	1	0	0
	Female	31-50	1	1	1
		51-91	2	4	1
		11-30	0	0	0
	Male	31-50	1	2	1
		51-91	3	3	0
No Injury					
		11-30	1	0	0
	Female	31-50	1	2	0
		51-91	1	6	8

We can extend this argument to other models. If we consider models that include the $u_{TSA(hij)}$ terms (i.e., models that include [TSA]), the MLEs must satisfy the condition $n_{hij\cdot} = \hat{m}_{hij\cdot}$ for all h, i, and j. However, $n_{213\cdot} = n_{222\cdot} = n_{411\cdot} = 0$. If MLEs exist, then $\hat{m}_{213\cdot} = 0$, etc.

But, because we have a log-linear model, each term $\hat{m}_{213k}$ must be strictly

positive. Summing over k, $\hat{m}_{213\cdot}$ must also be strictly positive. Because $\hat{m}_{213\cdot} = 0$, we have a contradiction. The MLEs must not exist. Therefore, we cannot fit any log-linear models that include [TSA]. Similarly, $n_{4\cdot 13} = n_{4\cdot 12} = 0$, so MLEs do not exist for models that include [TAR]. All other marginal totals are positive, so models with [TSA] or [TAR] are our primary source of concern.

To some extent, these problems can be evaded. Suppose we wish to fit a model with [TSA]. If we think of TSA as one composite factor, we can think of the problem as being one of fitting a $(4 \times 2 \times 3) \times 3$ table, i.e., a 24×3 table. In this table, three of the "rows" have zero totals. If we drop these three rows from the table, we can fit the remaining 21×3 table. Now, suppose we fit the model [TSA][TR][AR]. We have 21 degrees of freedom in the model for fitting [TSA]. Normally, this would be 24 degrees of freedom for fitting a grand mean, main effects T, S, A; two-factor effects (TS), (TA), (SA); and the three-factor effect (TSA). However, because 3 rows are being dropped, we have only 21 degrees of freedom. For fitting [TR], we have an additional 8 degrees of freedom. Adding [TR] involves adding the main effect R with 2 degrees of freedom and the two-factor effect (TR) with 6 degrees of freedom. Finally, adding [AR] is equivalent to adding the two-factor effect (AR) with 4 degrees of freedom. The model has $21+8+4 = 33$ degrees of freedom. The table has $21 \times 3 = 63$ degrees of freedom, so the test of [TSA][TR][AR] has $63 - 33 = 30$ degrees of freedom. (Incidentally, $G^2 = 34.72$, so this is a very good model.)

We now consider the problem of determining the degrees of freedom for testing the more complex model, [TSA][TAR]. Recall that there are marginal totals of 0 associated with both of these terms. Of the 24 marginal totals associated with [TSA] (obtained by summing over R), three are 0. This leads us to fitting the $(24 - 3) \times 3$, TSA by R table considered above. Of the 36 marginal totals associated with [TAR] (obtained by summing over S), two are 0. This would normally lead us to fitting the $(36 - 2) \times 2$ TAR by S table. In fact, what we need to do is fit the intersection of these tables. The 21×3 table involves dropping the cells (h, i, j, k) with values (2,1,3,1), (2,1,3,2), (2,1,3,3), (2,2,2,1), (2,2,2,2), (2,2,2,3), (4,1,1,1), (4,1,1,2), and (4,1,1,3). The 34×2 table drops the cells (4,1,1,3), (4,2,1,3), (4,1,1,2), and (4,2,1,2). Note that both tables drop the cells (4,1,1,2) and (4,1,1,3), so a total of 11 cells are being dropped from the $4 \times 2 \times 3 \times 3$ table. We are left with a table that has $72 - 11 = 61$ cells.

We now compute the degrees of freedom for the model [TSA][TAR]. As before, fitting [TSA] alone accounts for 21 degrees of freedom. Determining the degrees of freedom for adding [TAR] is somewhat more complex. Normally, [TAR] alone would involve 36 degrees of freedom broken down as follows: (grand mean) : (1), T : (3), A : (2), R : (2), (TA) : (6), (TR) : (6), (AR) : (4), and (TAR) : (12). With the marginal zeros, there are only 34 degrees of freedom. The 2 degrees of freedom come out of the (TAR) interaction, so the correct degrees of freedom are (TAR) : (10). Adding [TAR] to

a model with [TSA] involves adding the effects R, (TR), (AR), and (TAR). The degrees of freedom for [TSA][TAR] are $21 + 2 + 6 + 4 + 10 = 43$. The lack of fit test for [TSA][TAR] has $61 - 43 = 18$ degrees of freedom; G^2 is 21.90.

The basic approach to dealing with models that have random zeros is to identify all cells that imply that MLEs do not exist. In other words, identify all cells for which the maximum likelihood constraints would imply that $\hat{m} = 0$. Such cells are dropped from the model and MLEs are found for the remaining cells. We are simply treating cells with "$\hat{m} = 0$" as fixed zeros. The degrees of freedom for the table are the number of cells in the full table minus the number of "fixed" zeros. The degrees of freedom for the model are the usual degrees of freedom for the model minus the number of degrees of freedom lost because there is "no information" available on some parameters. From Example 8.3.1, [TSA] usually involves 24 degrees of freedom related to the $4 \times 3 \times 3$ TSA marginal table. The table has $n_{213\cdot} = n_{222\cdot} = n_{411\cdot} = 0$. The nine cells involved in these three marginal totals are being treated like fixed zeros, so those nine cells "do not exist." There is no information available on 3 of the 24 cells of the TSA marginal table. Thus, 3 degrees of freedom are lost to the model because there is no information available.

We now consider one more example to set the ideas and illustrate some additional details.

EXAMPLE 8.3.2. The data in Table 8.4 were adapted from Lee (1980). The table involves four factors related to the survival of patients with stages 3 and 4 melanoma: Gender, Remission, Immunity, and Survival. Remission has three categories: still in remission, relapsed, never in remission. Immunity has three categories that were derived from results on six skin tests. One test score of at least 10 indicates good immunity. No test scores of at least 10 indicates no immunity. If more than half of the test scores are unknown and those that are known are less than 10, the immunity is taken as unknown. Finally, to get more interesting marginal zeros, Lee's count in cell (2,1,3,2) has been changed from 1 to 0.

Denote the factors Gender, Remission, Immunity, and Survival as G, R, I, and S, respectively. Consider fitting the model [GRI][GRS][GIS][RIS]. The likelihood equations are, for all h, i, j, and k,

$$\begin{aligned} n_{hij\cdot} &= \hat{m}_{hij\cdot}\,, \\ n_{hi\cdot k} &= \hat{m}_{hi\cdot k}\,, \\ n_{h\cdot jk} &= \hat{m}_{h\cdot jk}\,, \\ n_{\cdot ijk} &= \hat{m}_{\cdot ijk}\,. \end{aligned}$$

Many of these marginal tables have zero totals. The RIS marginal table has four zeros: $0 = n_{\cdot 131} = n_{\cdot 132} = n_{\cdot 112} = n_{\cdot 211}$. These involve

TABLE 8.4. Melanoma Data

Gender (h)	Remission (i)	Immunity (j)	Survival (k) Dead	Alive
		No Immunity	2	0
	Relapsed	Immunity	4	1
		Unknown	0	0
		No Immunity	0	1
Male	Remission	Immunity	1	10
		Unknown	3	3
		No Immunity	3	1
	None	Immunity	10	5
		Unknown	8	2
		No Immunity	2	0
	Relapsed	Immunity	3	4
		Unknown	0	0
		No Immunity	0	0
Female	Remission	Immunity	0	10
		Unknown	0	4
		No Immunity	2	0
	None	Immunity	6	3
		Unknown	3	8

the eight cells with counts of 0: (1,1,3,1), (2,1,3,1), (1,1,3,2), (2,1,3,2), (1,1,1,2), (2,1,1,2), (1,2,1,1), and (2,2,1,1). The GRI table has three zeros: $0 = n_{113\cdot} = n_{213\cdot} = n_{221\cdot}$. These involve six cases: (1,1,3,1), (1,1,3,2), (2,1,3,1), (2,1,3,2), (2,2,1,1), and (2,2,1,2). The GRS table has $n_{22\cdot 1} = 0$. This total involves the cases (2,2,1,1), (2,2,2,1), and (2,2,3,1). Finally, the GIS table has $n_{2\cdot 12} = 0$, so the cells (2,1,1,2), (2,2,1,2), and (2,3,1,2) are all 0.

Having listed all of the cells that would have to have $\hat{m} = 0$, we see that there are only 12 distinct cells. (These 12 happen to be all of the cells in the entire table with counts of 0, but that fact is irrelevant.) The full table is a $2 \times 3 \times 3 \times 2$ table having 36 cells. If we drop the 12 cells with $\hat{m} = 0$, we have 24 cells remaining in the table. Thus, the table has 24 degrees of freedom.

We now consider the degrees of freedom for the model. The RIS table has four zero totals. The corresponding cells are dropped from the table, so rather than being a 2×18 table, the G by RSI table is a 2×14 table. Thus, fitting [RIS] gives the model 14 degrees of freedom.

We can also compute the degrees of freedom for fitting [RIS] term by term. To compute the degrees of freedom term by term, we need to note

that the RI marginal table has $n_{\cdot 13\cdot} = 0$. Thus, this 3×3 table has one dropped cell and only 8 degrees of freedom. The degrees of freedom are allocated: (grand mean) : (1), R : (2), I : (2), and (RI) : (3) rather than the standard value 4. Moving up to the RIS $3 \times 3 \times 2$ table, we have 14 non-empty cells. Because the (RI) interaction has only 3 degrees of freedom, the 14 degrees of freedom correspond to: (grand mean) : (1), R : (2), I : (2), S : (1), (RI) : (3), (RS) : (2), (IS) : (2), which leaves (RIS) : (1) rather than the standard value of 4. Thus, with this pattern of random zeros, the four empty cells cause a reduction of 1 degree of freedom in the (RI) interaction and 3 degrees of freedom in the (RIS) interaction. Note that the only two-factor marginal table that contains a zero is the RI table. Thus, any other reductions in degrees of freedom must occur in three-factor interaction terms.

A similar analysis holds for the GRI marginal table. This $2 \times 3 \times 3$ table has three zeros, so three cells are dropped. There are $18 - 3 = 15$ degrees of freedom. The degrees of freedom are allocated: (grand mean) : (1), G : (1), R : (2), I : (2), (GR) : (2), (GI) : (2); once again (RI) : (3) (rather than 4) which leaves us with (GRI) : (2) (rather than 4).

The model [RIS][GRI] has 14 degrees of freedom for [RIS]. We then add the terms G, (GR), (GI), and (GRI) with $1 + 2 + 2 + 2 = 7$ degrees of freedom. Thus, [RIS][GRI] has $14 + 7 = 21$ degrees of freedom.

The $2 \times 3 \times 2$ GRS table has one zero, hence 11 degrees of freedom. They are allocated: (grand mean) : (1), G : (1), R : (2), S : (1), (GR) : (2), (GS) : (1), (RS) : (2), which leaves (GRS) : (1). Adding [GRS] to [RIS][GRI] involves adding the terms (GS) and (GRS) with $1 + 1 = 2$ degrees of freedom, so [RIS][GRI][GRS] has $21 + 2 = 23$ degrees of freedom.

Finally, the $2 \times 3 \times 2$ GIS table has one cell dropped for 11 degrees of freedom. The 1 degree of freedom lost is taken from the (GIS) interaction, so (GIS) has 1 degree of freedom. The model [RIS][GRI][GRS][GIS] involves adding (GIS) to the model [RIS][GRI][GRS]. The degrees of freedom are $23 + 1 = 24$.

Recall that the degrees of freedom for the table are 24. With 24 degrees of freedom for the model, we get a perfect fit. Having dropped out the cells that require $\hat{m} = 0$, the model [RIS][GRI][GRS][GIS] is a saturated model.

In order to implement this approach to dealing with random zeros, we must be able to identify all cells for which the likelihood equations imply that $\hat{m} = 0$. As in Examples 8.3.1 and 8.3.2, it is frequently easy to identify some cells that have $\hat{m} = 0$. Often, all of the cells with $\hat{m} = 0$ are easily identified. Cells are easy to identify if they correspond to marginal totals that are zero. Unfortunately, sometimes all of the marginal totals can be positive, but cells with $\hat{m} = 0$ still exist.

EXAMPLE 8.3.3. Consider a $2 \times 2 \times 2$ table with $n_{111} = n_{222} = 0$ and $n_{ijk} > 0$ for all other cells. If we fit the model [12][13][23], the likelihood

equations are

$$\begin{aligned} n_{ij\cdot} &= \hat{m}_{ij\cdot}\,, \\ n_{i\cdot k} &= \hat{m}_{i\cdot k}\,, \end{aligned}$$

and

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

Because n_{111} and n_{222} are the only 0s, all of the marginal totals listed above are positive. Looking at the marginal totals indicates no cause for concern. Nonetheless, these equations imply that $\hat{m}_{111} = \hat{m}_{222} = 0$.

To see this, first note that we must have $n_{\cdots} = \hat{m}_{\cdots}$. Writing out all of the likelihood equations involving n_{111} and n_{222} gives

$$\begin{aligned} n_{111} + n_{112} &= \hat{m}_{111} + \hat{m}_{112}\,, \\ n_{221} + n_{222} &= \hat{m}_{221} + \hat{m}_{222}\,, \\ n_{111} + n_{121} &= \hat{m}_{111} + \hat{m}_{121}\,, \\ n_{212} + n_{222} &= \hat{m}_{212} + \hat{m}_{222}\,, \\ n_{111} + n_{211} &= \hat{m}_{111} + \hat{m}_{211}\,, \\ n_{122} + n_{222} &= \hat{m}_{122} + \hat{m}_{222}\,. \end{aligned}$$

Adding these six equations together, we get

$$2(n_{111} + n_{222}) + n_{\cdots} = 2(\hat{m}_{111} + \hat{m}_{222}) + \hat{m}_{\cdots}\,.$$

Because $n_{\cdots} = \hat{m}_{\cdots}$, we have

$$n_{111} + n_{222} = \hat{m}_{111} + \hat{m}_{222}\,.$$

With $n_{111} = n_{222} = 0$, we need both $\hat{m}_{111} = 0$ and $\hat{m}_{222} = 0$. These two cells would be dropped from the $2 \times 2 \times 2$ table.

Although the author is unaware of any general method of identifying situations like that in Example 8.3.3, the process of fitting models can give hints as to whether such cells have been overlooked. In particular, if some estimated cell counts seem to be converging to zero or if the iterative estimated cell counts fail to converge, it would be wise to consider the possibility that this is due to cells that need to be dropped because of the pattern of random zeros.

Finally, it is interesting to look at the test of [TSA][TR][AR] versus [TSA][TAR] that can be obtained from Example 8.3.1. The test has $G^2 = 34.72 - 21.90 = 12.82$ with $df = 30 - 18 = 12$. In this example, the degrees of freedom for the test happen to correspond to the usual degrees of freedom for the (TAR) interaction with T at four levels, A at three levels, and R at three levels. However, the 12 degrees of freedom are actually arrived at quite differently. As established earlier, (TAR) has 10 degrees

of freedom. The other 2 degrees of freedom in the test come from the fact that [TSA][TR][AR] is fit to a 63-cell table rather than the 61-cell table of [TSA][TAR]. So the 12 degrees of freedom come from 10 degrees of freedom for (TAR) and 2 degrees of freedom for new cells.

This discussion also points out that there are some technical difficulties involved in testing [TSA][TR][AR] versus [TSA][TAR]. We are testing a 63-cell table against a 61-cell table. We have not discussed this sort of thing previously. Obviously, this requires a mathematical theory that embeds both of these within the 72-cell $4 \times 2 \times 3 \times 3$ table allowing for fixed zero cells, log-linear models on the nonzero cells, and reduced models in which fixed zeros are allowed to become unfixed. Moreover, asymptotic theory will be of limited value because all of these problems are being caused by small sample sizes.

8.4 Exercises

EXERCISE 8.4.1. Brown (1980) presents data that are reproduced in Table 8.5, on a cross-classification of 53 prostate cancer patients. The factors are acid phosphatase level in the blood serum, age, stage, grade, x-ray, and nodal involvement. Acid level and age are categorized as high or low. Stage is an indication of size and location of the tumor; a positive value is more serious. The grade and x-ray indicate whether biopsy and x-ray tests are positive for cancer. The final factor is the whether the lymph nodes are involved. Analyze the data using iterative proportional fitting and ANOVA type models.

EXERCISE 8.4.2. Extend your analysis of the Berkeley graduate admissions data (cf. Exercise 3.8.4) by incorporating the method of partitioning polytomous factors from Example 8.1.2.

EXERCISE 8.4.3. *Partitioning Two-Way Tables.*
Lancaster (1949) and Irwin (1949) present a method of partitioning tables that was used in Exercise 2.7.4. We now establish the validity of this method. Consider a two-dimensional $I \times (J + K - 1)$ table. The partitioning method tests for independence in two subtables. One table is a reduced $I \times K$ table consisting of the last K columns or the full table. The other table is an $I \times J$ table that uses the first $J - 1$ columns of the full table and also includes a column into which the last K columns of the full table have been collapsed. Write the data with three subscripts as n_{ijk}, $i = 1, \ldots, I$, $j = 1, \ldots, J$, and $k = 1, \ldots, L_j$, where

$$L_j = \begin{cases} 1 & \text{if } j \neq J \\ K & \text{if } j = J. \end{cases}$$

TABLE 8.5. Nodal Involvement in Prostate Cancer

		Low Acid							
Grade		−				+			
X−ray		−		+		−		+	
Involvement		No	Yes	No	Yes	No	Yes	No	Yes
Low	−	4	0	1	0	1	1	0	0
Low	+	0	0	0	1	1	0	0	1
High	−	3	0	2	0	1	0	0	0
High	+	2	1	0	0	4	0	0	0
Age	Stage								

		High Acid							
Grade		−				+			
X−ray		−		+		−		+	
Involvement		No	Yes	No	Yes	No	Yes	No	Yes
Low	−	5	1	1	0	0	1	0	0
Low	+	1	1	0	0	1	2	1	5
High	−	3	0	0	1	0	0	0	1
High	+	2	2	0	1	0	0	0	1
Age	Stage								

Consider the models

$$\log(m_{ijk}) = \alpha_i + \beta_{jk}, \tag{1}$$

$$\log(m_{ijk}) = \beta_{ij} + \beta_{jk}, \tag{2}$$

and

$$\log(m_{ijk}) = \gamma_{ijk}. \tag{3}$$

Model (3) is the saturated model so

$$\hat{m}^{(3)}_{ijk} = n_{ijk}.$$

Model (1) is the model of independence in the $I \times J + K - 1$ table, so

$$\hat{m}^{(1)}_{ijk} = n_{i\cdot\cdot} n_{\cdot jk} / n_{\cdots}.$$

(a) Show that the maximum likelihood estimates for model (2) are

$$\hat{m}^{(2)}_{ijk} = \begin{cases} n_{ijk} & \text{if } j \neq J \\ n_{iJ\cdot} n_{\cdot Jk} / n_{\cdot J\cdot} & \text{if } j = J. \end{cases}$$

(b) Show that both G^2 and X^2 are the same for testing model (2) against model (3) as for testing the reduced table for independence.

(c) Show that both G^2 and X^2 are the same for testing model (1) against model (2) as for testing the collapsed table for independence.

(d) Extend (b) and (c) by showing that all power divergence statistics are the same, cf. Exercise 2.7.8.

EXERCISE 8.4.4. *The Bradley-Terry Model.*
"Let's suppose, for a moment, that you have just been married and that you are given a choice of having, during your entire lifetime, either x or y children. Which would you choose?" Imrey, Johnson, and Koch (1976) report results from asking this question of 44 Caucasian women from North Carolina who were under 30 and married to their first husband. Women were asked to respond for pairs of numbers x and y between 0 and 6 with $x < y$. The data are summarized in Table 8.6. The most basic form for such experiments is to ask each woman to respond for all possible pairs of numbers. If this was done, there is a considerable amount of missing data. For example, in comparing 0 children with 1 child there are only $17+2 = 19$ responses rather than 44.

TABLE 8.6. Family Size Preference

Alternative Choice	Preferred Number of Children						
	0	1	2	3	4	5	6
0	—	17	22	22	15	26	25
1	2	—	19	13	10	9	11
2	1	0	—	11	11	6	6
3	3	1	7	—	6	2	6
4	1	10	12	13	—	4	0
5	1	11	18	15	17	—	11
6	2	13	20	22	14	12	—

This data collection technique is called the method of *paired comparisons*. It is often used for such things as taste tests. Subjects find it easier to distinguish a preference between two brands of cola than to rank their preferences among half a dozen. David (1988) provides a good survey of the literature on the analysis of preference data along with notes on the history of the subject. One particular model for preference data assumes that each item has a probability π_i of being preferred. Thus, in a paired comparison, the conditional probability that i is preferred to j is $\pi_i/(\pi_i + \pi_j)$. There are many ways to arrive at this model; the one given above is simple but restrictive. In other developments, the parameters π_i need not add up to one, but it is no loss of generality to impose that condition. Bradley and Terry (1952) rediscovered the model and popularized it. The Bradley-Terry model was put into a log-linear model framework by Fienberg and Larntz (1976). For I items being compared, their framework consists of fitting the incomplete $I(I-1)/2 \times I$ table in which the rows consist of all pairs of items and the columns consist of the preferred item. A test of the model $\log(m_{ij}) = u + u_{1(i)} + u_{2(j)}$ is a test of whether the Bradley-Terry model holds.

(a) Rewrite Table 8.6 in the Fienberg-Larntz form.
(b) Test whether the Bradley-Terry model fits.

(c) Show that under the log-linear main effects model the odds of preferring item j to item j' is the ratio of a non-negative number depending on j and a non-negative number depending on j'. Show that this is equivalent to the Bradley-Terry model.

(d) Estimate the probabilities π_i.

Chapter 9
Generalized Linear Models

Generalized linear models are a class of models that generalize the linear models used for regression and analysis of variance. They allow for more general mean structures and more general distributions than regression and analysis of variance. Generalized linear models were first suggested by Nelder and Wedderburn (1972). An extensive treatment is given by McCullagh and Nelder (1989). Generalized linear models include logistic regression as a special case. Another special case, Poisson regression, provides the same analysis for count data as log-linear models. The discussion here involves more distribution theory than has been required elsewhere in this book; in particular, it makes extensive use of the exponential family of distributions and the gamma distribution. Information on these distributions can be obtained from many sources, e.g., Cox and Hinkley (1974). Section 1 presents the family of distributions used in generalized linear model theory. Estimation of the linear parameters is dealt with in Section 2. Model fitting and estimation of dispersion are examined in Section 3; both of these topics involve a version of the likelihood ratio test statistic called the *deviance*. Section 4 contains a summary and discussion. We begin with a brief review of some ideas from regression and analysis of variance and three examples of generalized linear models.

Regression and analysis of variance are fundamental tools in statistics. A multiple regression model is

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + e_i,$$

$i = 1, \ldots, n$, where $\mathrm{E}(e_i) = 0$ and, typically, $x_{i1} = 1$, $i = 1, \ldots, n$, so that β_1 is an intercept. The x_{ij}'s are all assumed to be known predictor

variables; the β_j's are fixed unknown parameters. A one-way analysis of variance can be written as

$$\begin{aligned} y_{ij} &= \mu + \alpha_i + e_{ij} \\ &= \mu \cdot (1) + \alpha_1 \delta_{1i} + \cdots + \alpha_t \delta_{ti} + e_{ij} \end{aligned}$$

where $i = 1, \ldots, t$, $j = 1, \ldots, n_i$, $\mathrm{E}(e_{ij}) = 0$, and δ_{hi} is 1 if $h = i$ and 0 otherwise. In the analysis of variance, μ and the α_i's are fixed unknown parameters, while the multiplier 1 for μ and the δ_{hi}'s play roles analogous to the x_{ij}'s in regression.

The key aspect of both regression and ANOVA is that they involve observations whose expected value is a linear combination of known predictor variables. In regression,

$$\mathrm{E}(y_i) = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

and in analysis of variance

$$\mathrm{E}(y_{ij}) = \mu \cdot (1) + \alpha_1 \delta_{1i} + \cdots + \alpha_t \delta_{ti} \,.$$

If the observations y in regression and ANOVA have the same variance and are uncorrelated, regression and ANOVA provide the best estimates of (estimable) parameters among all estimators that are unbiased linear functions of the observations. If we go a step further and assume that the observations have independent normal distributions with the same variance, then the usual estimates are the best among all unbiased functions of the observations. In both of these statements, "best" means that the estimates have minimum variance. Under the assumption of independent normal distributions with the same variance, the usual estimates are also maximum likelihood estimates. The current chapter is concerned with finding maximum likelihood estimates for a more general set of models. The models are less restrictive in that they allow more general forms of linearity and distributions other than the normal.

We now set some matrix notation. Both regression and analysis of variance are *linear models*, cf. Christensen (1996b). Let y_i be an observation. It is of no significance how we subscript the observations. Any convenient method of subscripting is acceptable whether it be one subscript as in regression, two subscripts as in one-way analysis of variance, or three subscripts as in two-way ANOVA with replications. Let $x_i' = (x_{i1}, \ldots, x_{ip})$ be a $1 \times p$ row vector of predictor variables. Let $\beta = (\beta_1, \ldots, \beta_p)'$ be a $p \times 1$ column vector of unknown parameters. A typical normal theory linear model assumes

$$y_i \sim N(x_i'\beta, \sigma^2)$$

where $i = 1, \ldots, n$ and the y_i's are independent. For pedagogical reasons, it is advantageous to write

$$y_i \sim N(m_i, \sigma^2), \qquad m_i = x_i'\beta.$$

We begin our discussion of generalized linear models by considering three examples. *In each case, we take* $y_1, \ldots, y_n$ *independent.* The models are specified by their distributions and mean structure.

For normal data,

$$y_i \sim N(m_i, \sigma^2), \qquad \mathrm{E}(y_i) = m_i, \qquad m_i = x_i'\beta.$$

This is the model for analysis of variance and regression.

For Poisson data,

$$y_i \sim \mathrm{Pois}(m_i), \qquad \mathrm{E}(y_i) = m_i, \qquad \log(m_i) = x_i'\beta.$$

This is just a log-linear model for a table containing n cells where the count in each cell has a Poisson distribution with parameter m_i. It is important to note that n is the number of cells and *not* the observation vector as it is elsewhere in this book. As we have mentioned before and as is shown in Chapter 12, under very weak conditions the analysis of a contingency table under Poisson sampling is the same as the analysis under multinomial sampling. It is interesting to note that the framework for generalized linear models assumes independent observations, so it does not apply directly to multinomial sampling or to general product-multinomial sampling. It is the equivalence of the maximum likelihood analyses under the Poisson, multinomial, and product-multinomial sampling schemes that makes generalized linear models a useful tool for contingency tables.

For binomial sampling, we take y_i to be the *proportion* of successes, so

$$N_i y_i \sim \mathrm{Bin}(N_i, p_i), \qquad \mathrm{E}(y_i) = p_i \equiv m_i,$$

$$\log\left(\frac{m_i}{1-m_i}\right) = \log\left(\frac{p_i}{1-p_i}\right) = x_i'\beta.$$

Note that N_i is a known quantity and not a parameter. The model is simply a logistic regression or logit model. The data consist of proportions obtained from independent binomial random variables and the mean structure is a linear model in the log odds. Alternative models for binomial regression will be mentioned later. Except in this chapter, y_i for binomial regression is always taken to mean the number of successes, rather than the proportion of successes. The change is made to fit the binomial distribution into the family of generalized linear models.

9.1 Distributions for Generalized Linear Models

The normal, Poisson, and binomial distributions considered above are members of the exponential family of distributions. A random variable y has

a distribution in the exponential family if it has a probability density or mass function that can be written as

$$f(y|\theta) = R(\theta)\exp\left[\sum_{j=1}^{v} q_j(\theta)\, t_j(y)\right] h(y)$$

where $\theta = (\theta_1, \ldots, \theta_s)'$ is a vector of parameters. If $v = 1$, the family is referred to as a one-parameter exponential family; the one parameter can be taken as $q_1(\theta)$.

The theory of generalized linear models requires the distribution of y to be in a subclass of the one-parameter exponential family. The density or mass function must have the form

$$f(y|\theta, \phi; w) = \exp\left\{\frac{w}{\phi}[\theta y - r(\theta)]\right\} h(\phi, y, w) \tag{1}$$

where θ, ϕ, and w are scalars. By assumption, w is a fixed known number. The role of ϕ in this function is curious; it is treated as an unknown constant but not as a parameter. With ϕ constant, the distribution (1) is in the one-parameter exponential family; just take $R(\theta) = \exp[-wr(\theta)/\phi]$, $q(\theta) = w\theta/\phi$, $t(y) = y$, and $h(y) = h(\phi, y, w)$. In practice, ϕ is often an unknown parameter. As such, the distribution need not be in the exponential family relative to the two parameters θ and ϕ because the function $h(\phi, y, w)$ need not satisfy the conditions of a two-parameter exponential family. The value ϕ is simply a *positive* number that is convenient for defining various special cases. The particular form of $f(y|\theta, \phi; w)$ in (1) is chosen so that the maximum likelihood estimate of θ does not depend on ϕ. This will be discussed in more detail in Section 2.

For the family of distributions (1), the expected value of y depends on θ but not on ϕ. For any distribution,

$$1 = \int f(y|\theta, \phi; w)\, dy$$

where it is understood that integration is always replaced by summation when y has a discrete distribution. Taking the derivative with respect to θ on both sides gives

$$0 = \int \dot{f}(y|\theta, \phi; w)\, dy \tag{2}$$

where $\dot{f}$ is the derivative of f with respect to θ and f satisfies conditions so that the derivative can be taken under the integral sign. From the exact form of $f(y|\theta, \phi; w)$ in (1), it is easily seen that (2) is

$$0 = \frac{w}{\phi} \int (y - \dot{r}(\theta))\, f(y|\theta, \phi; w)\, dy$$

where $\dot{r}(\theta)$ is the derivative $dr(\theta)/d\theta$. It follows that

$$\mathrm{E}(y) \equiv m = \dot{r}(\theta).$$

Typically, $\dot{r}$ is an invertible function, so θ is also a function of the mean, say

$$\theta = \dot{r}^{-1}(m).$$

Linear structure for distributions of the form (1) is most naturally specified by

$$\theta = x'\beta \tag{3}$$

where, as usual, β is a vector of unknown parameters and x is fixed and known. Note that with $\theta = \dot{r}^{-1}(m)$, the linear structure $x'\beta$ in equation (3) is also a function of the mean. In fact, the analysis of generalized linear models can be carried through when the linear structure is a more general function of the mean,

$$g(m) = x'\beta,$$

as long as it is possible to write $\theta = g_*(x'\beta)$ for some function $g_*(\cdot)$.

A *generalized linear model* consists of independent observations y_i, $i = 1, \ldots, n$, with

$$y_i \sim f(y_i|\theta_i, \phi, w_i), \qquad \mathrm{E}(y_i) \equiv m_i, \qquad g(m_i) = x_i'\beta.$$

If $g(m_i) = \theta_i$, the model is a *canonical* generalized linear model. In other words, a canonical model has $g(\cdot) \equiv \dot{r}^{-1}(\cdot)$.

Names have been given to the various components of generalized linear models. The linear structure $x'\beta$ is called the *linear predictor*. The function $g(\cdot)$ that specifies the relationship $g(m) = x'\beta$ between the mean and the linear predictor is called the *link function*. If $g(m) = \theta$, the function $g(\cdot)$ is called the *canonical link function*. The density $f(y|\theta, \phi; w)$ is often called the *error function* and the parameter ϕ is often called the *dispersion parameter*.

In Section 3, we will need to know the variance of y when y has a density of the form (1). Taking the second derivative with respect to θ on both sides of

$$1 = \int f(y|\theta, \phi; w)\, dy$$

and assuming that derivatives can be taken under the integral gives

$$\begin{aligned} 0 &= \int \ddot{f}(y|\theta, \phi; w)\, dy \\ &= \frac{w}{\phi} \int \frac{d\left[(y - \dot{r}(\theta))\, f(y|\theta, \phi; w)\right]}{d\theta} dy \\ &= \frac{w^2}{\phi^2} \int (y - \dot{r}(\theta))^2 f(y|\theta, \phi; w)\, dy \\ &\quad + \frac{w}{\phi} \int -\ddot{r}(\theta) f(y|\theta, \phi; w)\, dy \end{aligned}$$

where two dots indicate a second derivative. It follows immediately that $0 = \left[\mathrm{Var}(y)w^2/\phi^2\right] - [\ddot{r}(\theta)w/\phi]$ and, thus,

$$\mathrm{Var}(y) = \ddot{r}(\theta)\phi/w.$$

The function $\ddot{r}(\theta)$ is often written as a function of m,

$$V(m) \equiv \ddot{r}\big(\dot{r}^{-1}(m)\big) = \ddot{r}(\theta).$$

$V(m)$ is generally referred to as the *variance function.*

We now review how the general distribution theory applies to the three examples given earlier: normal, Poisson, and binomial sampling.

If $y \sim N(m, \sigma^2)$, the density for y real is

$$\begin{aligned} f(y|m;\sigma^2) &= \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\frac{(y-m)^2}{2\sigma^2}\right] \\ &= \exp\left(\frac{-m^2}{2\sigma^2}\right)\exp\left(\frac{my}{\sigma^2}\right)\left(\frac{1}{\sqrt{2\pi}\,\sigma}e^{-y^2/2\sigma^2}\right). \end{aligned}$$

To see that the density has the form of equation (1), identify $\theta = m$, $w = 1$, $\phi = \sigma^2$, $r(\theta) = m^2/2$, and $h(\phi, y, w) = \left(1/\sqrt{2\pi\sigma}\right)e^{-y^2/2\sigma^2}$. The expected value of y is m, so the canonical linear structure is $\theta = m = x'\beta$. The canonical link leads to a standard linear model.

For $y \sim \mathrm{Pois}(m)$, the probability mass function on $y = 0, 1, 2, \ldots$ is

$$\begin{aligned} f(y|m) &= \frac{m^y e^{-m}}{y!} \\ &= \exp(-m)\exp(y\,\log(m))\,(1/y!)\,. \end{aligned}$$

Identify $\theta = \log(m)$, $w = 1$, $\phi \equiv 1$, $r(\theta) = m$, and $h(\phi, y, w) = (1/y!)$. It is well known that for a $\mathrm{Pois}(m)$ distribution, the expected value and variance are both m. To see this from the general distribution theory, observe that the mean is $\dot{r}(\theta)$ and with $w = 1$ and $\phi \equiv 1$, the variance is $\ddot{r}(\theta)$. From $\theta = \log(m)$ and $r(\theta) = m$, it follows that $r(\theta) = e^\theta$ and thus $\dot{r}(\theta) = e^\theta$ and $\ddot{r}(\theta) = e^\theta$. Again, using $\theta = \log(m)$ gives $m = \dot{r}(\theta) = \ddot{r}(\theta)$. The expected value of y is m, so the canonical linear structure is $\theta = \log(m) = x'\beta$. The canonical link leads to a standard log-linear model for Poisson data.

For $Ny \sim \mathrm{Bin}(N, p)$ with N known, the mass function on $Ny = 0, \ldots, N$ is

$$\begin{aligned} f(y|p) &= \binom{N}{Ny}p^{Ny}(1-p)^{N-Ny} \\ &= \binom{N}{Ny}(1-p)^N\left(\frac{p}{1-p}\right)^{Ny} \\ &= (1-p)^N\exp\left[Ny\log\left(\frac{p}{1-p}\right)\right]\binom{N}{Ny}. \end{aligned}$$

Identify $\theta = \log\left(\frac{p}{1-p}\right)$, $w = N$, $\phi \equiv 1$, $r(\theta) = -\log(1-p)$, and $h(\phi, y, w) = \binom{N}{Ny}$. The expected value of y is $m \equiv p$, so the canonical linear structure is $\theta = \log\left(\frac{p}{1-p}\right) = \log\left(\frac{m}{1-m}\right) = x'\beta$. The canonical link leads to a standard logistic (logit) model. (See Exercise 9.5.1.)

In general, the inverse of any cumulative distribution function (cdf) $F(\cdot)$ makes a reasonable link function for binomial data, i.e., $g(p) = F^{-1}(p)$. $F(u) = e^u/(1+e^u)$ is the cdf of the logistic distribution and defines logistic regression. Probit regression is the procedure based on taking $F(u) = \Phi(u)$ where $\Phi(u)$ is the cdf of a standard normal distribution. A third example is complementary log-log regression which uses $F(u) = 1 - \exp[1 - \exp(e^u)]$.

As a last example, consider the gamma distribution. The gamma distribution is defined by the probability density function

$$f(y|\alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} e^{-\lambda y} y^{\alpha-1}$$

for $y > 0$. The density depends on two parameters, α and λ. The expected value of a gamma distribution is

$$\mathrm{E}(y) \equiv m = \frac{\alpha}{\lambda}$$

and the variance is

$$\mathrm{Var}(y) = \frac{\alpha}{\lambda^2}.$$

To indicate that y has a gamma distribution, write

$$y \sim \mathrm{Gamma}(\alpha, \lambda).$$

Special cases of the gamma distribution include exponential distributions with mean $1/\lambda$, i.e., $\mathrm{Gamma}(1, \lambda)$ and $\chi^2(n)$ distributions, $\mathrm{Gamma}(\frac{n}{2}, \frac{1}{2})$'s.

The gamma density can be rewritten as

$$f(y|\alpha, \lambda) = \left(\frac{\lambda}{\alpha}\right)^\alpha \exp\left[\alpha\left(\frac{-\lambda}{\alpha}\right) y\right] \left(\frac{\alpha^\alpha}{\Gamma(\alpha)} y^{\alpha-1}\right).$$

To see the gamma density in the form of equation (1), identify $\theta = -\lambda/\alpha$, $w = 1$, $\phi = 1/\alpha$, $r(\theta) = -\log(\lambda/\alpha)$, and $h(\phi, y, w) = \alpha^\alpha y^{\alpha-1}/\Gamma(\alpha)$. The expected value of y is $m = \alpha/\lambda = -1/\theta$, so the canonical linear structure is $\theta = -1/m = x'\beta$. Note that the distribution is only defined for $y > 0$; thus, for any gamma distribution, $m > 0$. It follows that when using the canonical link, restrictions must be placed on the parameter vector β to ensure that the expected value is positive. (See Exercise 9.5.2.)

The canonical generalized linear model for n independent observations with gamma distributions is

$$y_i \sim \mathrm{Gamma}(\alpha, \lambda_i), \qquad \mathrm{E}(y_i) = m_i = \frac{\alpha}{\lambda_i}, \qquad \frac{-1}{m_i} = x_i'\beta$$

where β is restricted so that $-x_i'\beta > 0$ for all i. Gamma distribution regression is useful for modeling situations in which the coefficient of variation is constant. The coefficient of variation is

$$\frac{\sqrt{\text{Var}(y_i)}}{\text{E}(y_i)} = \frac{\sqrt{\alpha/\lambda_i^2}}{\alpha/\lambda_i} = \frac{1}{\sqrt{\alpha}}.$$

When the data appear to have a constant coefficient of variation, using the gamma distribution is an alternative to doing a standard linear model analysis on the logs of the data, cf. Christensen (1996b, Section 13.7). Note that the constant coefficient of variation does not depend on the choice of link function. Often, noncanonical links such as the identity, $m = x'\beta$, and the log, $\log(m) = x'\beta$, are used with gamma distributed data, cf. McCullagh and Nelder (1989). The identity link requires restrictions on β; the log link does not. When the coefficient of variation is small, the log link analysis is very similar to the linear model analysis on the logs of the data. *The log link is probably the most commonly used for gamma regression. Exponential regression* is the special case with $\phi = 1$ and (usually) a log link.

9.2 Estimation of Linear Parameters

In their most natural form, generalized linear models assume n independent observations with

$$y_i \sim f(y_i|\theta_i, \phi; w_i)$$

and

$$\theta_i = x_i'\beta$$

for some vector of parameters $\beta = (\beta_1, \ldots, \beta_p)'$. The density $f(y_i|\theta_i, \phi; w_i)$ is defined by equation (9.1.1). The linear structure given above uses the canonical link function. More generally, the linear structure can be defined by

$$g(m_i) = x_i'\beta.$$

For this extension, assume $\theta_i = \dot{r}^{-1}(m_i)$ and

$$\theta_i = \dot{r}^{-1}\big(g^{-1}(x_i'\beta)\big) \equiv g_*(x_i'\beta).$$

The likelihood function for the generalized linear model with canonical link function is

$$\begin{aligned} L(\beta;\phi) &= \prod_{i=1}^{n} f(y_i|\theta_i, \phi; w_i) \\ &= \prod_{i=1}^{n} f(y_i|x_i'\beta, \phi; w_i). \end{aligned}$$

Using equation (9.1.1), the log-likelihood is

$$\begin{aligned} \ell(\beta;\phi) &\equiv \log[L(\beta;\phi)] \\ &= \sum_{i=1}^{n} \log[f(y_i|x_i'\beta,\phi;w_i)] \qquad (1) \\ &= \sum_{i=1}^{n} \frac{w_i}{\phi}\left[x_i'\beta\, y_i - r(x_i'\beta)\right] + \sum_{i=1}^{n} \log[h(\phi,y_i,w_i)]\,. \end{aligned}$$

With ϕ fixed, the maximum likelihood estimate of β is obtained by solving for β in the likelihood equations

$$\frac{\partial \ell(\beta;\phi)}{\partial \beta_j} = 0, \qquad (2)$$

$j = 1,\ldots,p$. It is a simple matter to see that taking the partial derivatives $\partial\ell(\beta;\phi)/\partial\beta_j$ leads to likelihood equations of the form

$$\frac{Q_j(\beta)}{\phi} = 0$$

for some functions $Q_j(\cdot)$, $j = 1,\ldots,p$. Obviously, the solution $\hat{\beta}$ to such likelihood equations does not depend on the value of ϕ. Thus, $\hat{\beta}$ is the maximum likelihood estimate for any value of ϕ; i.e., it is the maximum likelihood estimate regardless of the true value of ϕ.

Essentially, the same analysis holds when a linear structure $g(m_i) = x_i'\beta$ is assumed. With $\theta_i = g_*(x_i'\beta)$, simply use $g_*(x_i'\beta)$ in place of $x_i'\beta$ in equation (1). The only problem is that the partial derivatives become slightly more difficult to find.

Maximum likelihood estimates are invariant under transformations of the parameters. In other words, given a maximum likelihood estimate for a parameter, any function of the maximum likelihood estimate is the maximum likelihood estimate for the corresponding function of the parameter. For a discussion of this property see Cox and Hinkley (1974, p. 287). Given a maximum likelihood estimate for β, say $\hat{\beta}$, we immediately obtain an estimate of the expected value m_i, namely

$$\hat{m}_i = g^{-1}(x_i'\hat{\beta}),$$

an estimate of the linear predictor $g(m_i)$, namely

$$g(\hat{m}_i) = x_i'\hat{\beta},$$

and an estimate of θ_i, namely

$$\hat{\theta}_i = g_*(x_i'\hat{\beta}).$$

Solving the likelihood equations (2) is typically accomplished by using the Newton-Raphson algorithm. For generalized linear models, this reduces to performing a series of weighted least squares regressions and is known as *iteratively reweighted least squares.* Sections 10.5 and 11.3 give details for the special cases of log-linear modeling and logistic regression.

Under suitable conditions, the estimate $\hat{\beta}$ and smooth functions of $\hat{\beta}$, e.g., $\hat{m}_i = g^{-1}(x_i'\hat{\beta})$, have asymptotic multivariate normal distributions. Moreover, an estimate of the asymptotic covariance matrix of $\hat{\beta}$ is easily obtained from the iteratively reweighted least squares algorithm. Under suitable conditions, this estimate is consistent and also yields estimated asymptotic covariance matrices for smooth functions of $\hat{\beta}$. Given the estimates and the estimated asymptotic covariance matrices, standard normal theory methods for tests and confidence regions can be applied to yield asymptotic statistical inferences.

This brief discussion of estimation has not addressed several important points. To perform the differentiations, the $x_i'\beta$'s need to define a regression so that the β_j's are well defined. The partial derivatives need to be derived and shown to be of the form $Q_j(\beta)/\phi$, cf. Exercise 9.5.3. A solution to the likelihood equations must be shown to give the maximum of the log-likelihood. Exact conditions for the asymptotic results need to be stated; the necessary conditions may differ for different generalized linear models. For example, in regression analysis, one typically thinks about having the number of observations n go to infinity; however, for contingency table data, the number of cells in the table is n and is typically considered fixed, while the number of counts within the cells is assumed to get large. For more information on many of these issues see McCullagh and Nelder (1989).

9.3 Estimation of Dispersion and Model Fitting

Generalized linear model theory focuses on the estimation of linear parameters. The general theory seems to be less well developed for the purposes of model fitting and dispersion estimation. The basic statistics used in model fitting and estimating functions of the dispersion parameter ϕ are the *deviance* and a generalization of the Pearson test statistic. There are patterns common to the use of these statistics, but specifics vary from case to case. Our discussion focuses on two general asymptotic approaches. In one approach, the number of observations n is allowed to go to infinity. This approach is appropriate for many linear model and logistic regression problems. The second approach fixes n and uses asymptotics based on other aspects of the model. This approach is appropriate for many contingency table problems. We begin by defining the statistics.

The *standardized deviance* is simply the asymptotic form of the likelihood ratio statistic for testing a generalized linear model against the cor-

responding saturated model. Remember that in the likelihood analysis of a generalized linear model, the dispersion parameter ϕ is treated as fixed. A saturated model is simply one in which the number of parameters is so large that the data are fit perfectly. In particular, the model

$$y_i \sim f(y_i|\theta_i, \phi; w_i),$$

with no restrictions on the θ_i's, is saturated because there are as many parameters θ_i as there are observations. The maximum likelihood estimates have $\hat{m}_i = y_i$. This is easily established from the likelihood equations for the θ_i's. Substituting θ_i for $x_i'\beta$ in (9.2.1) and taking partial derivatives with respect to the θ_i's gives $y_i = \dot{r}(\hat{\theta}_i) = \hat{m}_i$ as a solution to the equations. The estimates of the θ_i's for the saturated model are determined by the estimates of the m_i's.

The parameters β and $m = (m_1, \ldots, m_n)'$ are assumed to be interchangeable, so write

$$\ell(m; \phi) \equiv \ell(\beta; \phi).$$

Also, write $y = (y_1, \ldots, y_n)'$. The *standardized deviance* is two times the difference between the maximum of the log-likelihood under the saturated model and the maximum of the log-likelihood under the specified generalized linear model, i.e.,

$$D^*(\hat{m}; \phi) = 2\left[\ell(y; \phi) - \ell(\hat{m}; \phi)\right].$$

Here, y is used in $\ell(y; \phi)$ because y is the maximum likelihood estimate of m for the saturated model. From inspection of (9.2.1), it is easily seen that the standardized deviance can be written as

$$D^*(\hat{m}; \phi) = \frac{D(\hat{m})}{\phi}$$

for a function $D(\hat{m})$ that does not depend on ϕ. Define the function $D(\hat{m})$ to be the *deviance* of the generalized linear model. Recall that in many important special cases, $\phi = 1$. For normal theory linear models, $D(\hat{m})$ is the sum of squares error.

As mentioned before, the likelihood analysis treats ϕ as fixed and ignores the fact that the dispersion ϕ is a parameter. The standardized deviance $D^*(\hat{m}; \phi)$ is only the likelihood ratio test statistic when ϕ is known. When ϕ is unknown, $D^*(\hat{m}; \phi)$ is not even a statistic because it depends on an unknown parameter. $D(\hat{m})$, on the other hand, does not depend on ϕ, so it is a statistic.

Another statistic used to evaluate models and estimate dispersion is the *generalized Pearson statistic*. The Pearson statistic is defined as

$$X^2 = \sum_{i=1}^{n} \frac{w_i\,(y_i - \hat{m}_i)^2}{V(\hat{m}_i)}$$

where $V(\cdot)$ is the variance function defined in Section 1. See Exercise 9.5.6.

The Pearson statistic can be used for consistent estimation of ϕ for large n. The variance of y_i is $V(m_i)\phi/w_i$ so, clearly,

$$\mathrm{E}\left(\frac{w_i\,(y_i - m_i)^2}{V(m_i)}\right) = \phi.$$

By Chebyshev's Weak Law of Large Numbers (cf. Rao, 1973, p. 112), if

$$\frac{1}{n^2}\sum_{i=1}^{n}\frac{w_i^2\mathrm{E}(y_i - m_i)^4}{V(m_i)^2} \to 0$$

as $n \to \infty$, then

$$\frac{1}{n}\sum_{i=1}^{n}\frac{w_i\,(y_i - m_i)^2}{V(m_i)} \xrightarrow{P} \phi.$$

It follows that if $V(\cdot)$ is a continuous function and $\hat{m}_i \xrightarrow{P} m_i$ for all i,

$$\frac{X^2}{n-p} \xrightarrow{P} \phi$$

where p is the number of parameters in β and we are assuming that the $n \times p$ model matrix

$$X = \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix}$$

has $\mathrm{rank}(X) = p$ not depending on n. Typically, we take

$$\hat{\phi} = \frac{X^2}{n-p}.$$

Continuous functions of the dispersion, say $d(\phi)$, are estimated with $d(\hat{\phi})$. If the Pearson estimate is consistent, continuous functions of it are also consistent estimates.

Under some large sample conditions with fixed n, for all values of ϕ the standardized Pearson statistic has the asymptotic distribution

$$\frac{X^2}{\phi} \sim \chi^2(n-p).$$

In these cases, standard methods of variance estimation for normal data can be applied to give asymptotic confidence intervals and tests for ϕ, cf. Christensen (1996a, Sec. 2.6). Using properties of the χ^2 distribution, the asymptotic distribution also leads to the approximation

$$\mathrm{E}(X^2) \doteq \phi \cdot (n-p).$$

Once again, an obvious estimate of ϕ is

$$\hat{\phi}_P = \frac{X^2}{n-p}.$$

Similarly, under certain conditions with n fixed, the standardized deviance $D^*(\hat{m};\phi)$ has the asymptotic distribution

$$D^*(\hat{m};\phi) \sim \chi^2(n-p)$$

for all values of ϕ. By definition,

$$D(\hat{m}) = \phi D^*(\hat{m};\phi),$$

so, asymptotically,

$$D(\hat{m}) \sim \phi\chi^2(n-p).$$

Again, when the asymptotic distribution is valid, standard methods of variance estimation for normal data can be applied to give asymptotic confidence intervals and tests for ϕ. An obvious point estimate of ϕ is

$$\hat{\phi}_D = \frac{D(\hat{m})}{n-p}.$$

Unfortunately, the deviance-based estimate is frequently inconsistent when $n \to \infty$. Even in the simplest binomial case, $y_i \sim \text{Bin}(1,p)$, this estimate is not consistent. For this case, $\phi \equiv 1$, but for large samples, it is easily seen that

$$\frac{D}{n-1} \xrightarrow{P} -2\left[p\log(p) + (1-p)\log(1-p)\right]$$

which is not typically equal to 1.

Deviances and Pearson statistics can also be used to evaluate the adequacy of generalized linear models. If null distributions are available for the statistics, tests of the adequacy of models can be performed. These can be unconditional, exact conditional, approximate conditional, or asymptotic distributions. If null distributions are not available, the statistics can be used in an exploratory fashion to give rough ideas of model adequacy.

If the data follow a true one-parameter exponential family distribution, the dispersion parameter ϕ is identically constant and, without loss of generality, we can take $\phi \equiv 1$. If the model is correct and the Pearson statistic gives a consistent estimate of ϕ,

$$\frac{X^2}{n-p} \xrightarrow{P} 1.$$

If $X^2/(n-p)$ is substantially larger than 1, it is an indication that the model is incorrect. With $\phi \equiv 1$, the standardized deviance equals the deviance.

If the deviance estimate of ϕ is consistent, the deviance can also be used to evaluate lack of fit. When n is fixed, if the deviance has an asymptotic χ^2 distribution, the deviance is a lack of fit test statistic that can be used in a formal asymptotic test of the generalized linear model against the saturated model. Similarly, if X^2 has an asymptotic χ^2 distribution, the Pearson statistic can be used in a lack of fit test.

To test a model $g(m_i) = x_i'\beta$ against a reduced model, say $g(m_i) = x_{0i}'\gamma$ in a one-parameter family, simply compare the difference in the deviances to an appropriate χ^2 distribution. In particular, the asymptotic generalized likelihood ratio test rejects the adequacy of the reduced model at the α level if

$$D(\hat{m}_0) - D(\hat{m}) > \chi^2(1-\alpha, p-p_0).$$

Here, $\hat{m}_0$ and $D(\hat{m}_0)$ are the maximum likelihood estimate of m and the deviance under the reduced model. The model is a reduced model in the sense that $X_0 = XB$ for some matrix B where

$$X_0 = \begin{bmatrix} x_{01}' \\ \vdots \\ x_{0n}' \end{bmatrix}$$

and

$$\operatorname{rank}(X_0) = p_0.$$

Such reduced model tests tend to be asymptotically valid under weaker conditions than general lack of fit tests. In particular, the tests are often valid under both asymptotic approaches discussed here. Less formally, if $(D(\hat{m}_0) - D(\hat{m}))/(p-p_0)$ is a credible estimate of $\phi \equiv 1$ under the reduced model, it makes sense to reject the reduced model whenever the estimate is much larger than 1.

Both Poisson regression and logistic regression fit into this one-parameter framework. For Poisson regression, the deviance is G^2 and often can be used for lack of fit tests. In both Poisson and logistic regression, the asymptotic χ^2 approximation for the test of a model against a reduced model is often valid. However, we have seen that the lack of fit statistic for logistic regression is typically *not* asymptotically χ^2. Care must be used in applying the asymptotic results given above. The specifics of each situation must be considered.

For generalized linear models with a nontrivial dispersion parameter, we can only test reduced models against larger models. An appealing asymptotic test is to reject the adequacy of the reduced model at the α level if

$$\frac{\big(D(\hat{m}_0) - D(\hat{m})\big)/(p-p_0)}{D(\hat{m})/(n-p)} > F(1-\alpha, p-p_0, n-p).$$

This relies not only on $\big(D(\hat{m}_0) - D(\hat{m})\big)/\phi$ and $D(\hat{m})/\phi$ being asymptotically χ^2 but also on them being asymptotically independent. As discussed

earlier, the χ^2 approximation to the distribution of $D(\hat{m})/\phi$ frequently requires asymptotics based on fixed n. For normal theory models, this is the usual F test.

If (a) n is large, (b) $D(\hat{m})/(n-p)$ is a consistent estimate of ϕ, and (c) $\big(D(\hat{m}_0) - D(\hat{m})\big)/\phi$ is asymptotically χ^2, we get the asymptotic null distribution

$$\frac{\big(D(\hat{m}_0) - D(\hat{m})\big)/(p-p_0)}{D(\hat{m})/(n-p)} \sim \frac{\chi^2(p-p_0)}{p-p_0}$$

with a corresponding test. If $X^2/(n-p)$ is a consistent estimate of ϕ, the Pearson estimate can be used in the denominator of the asymptotic test. This is of particular importance when $D(\hat{m})/(n-p)$ is not consistent but $X^2/(n-p)$ is. Again, a less formal evaluation can be made if $\big(D(\hat{m}_0) - D(\hat{m})\big)/(p-p_0)$ is a plausible estimate of ϕ under the reduced model. The reduced model is called in question when the test statistic is much larger than 1.

Note that as $n-p$ approaches infinity, the $F(p-p_0, n-p)$ distribution approaches a $\chi^2(p-p_0)/(p-p_0)$. Even though the appropriate asymptotic distribution for large n is a rescaled χ^2, for data analytic purposes it may not be unreasonable to use F tables instead.

Normal linear models and gamma distribution regression both fit into the nontrivial dispersion parameter framework. As always, appropriate conditions must be met for the asymptotic results to be valid. For normal theory linear models, the deviance and Pearson statistic both equal the error sum of squares and the F distribution is exact. It does not rely on any asymptotic arguments.

9.4 Summary and Discussion

Without a doubt, iteratively reweighted least squares for generalized linear models is a remarkably useful computing device. A review of the use of iterative generalized least squares in statistical estimation is given by del Pino (1989). Iteratively reweighted least squares is a special case of iterative generalized least squares.

Generalized linear models are designed to treat independent observations that have a distribution in the one-parameter exponential family. They also provide maximum likelihood estimates of appropriate functions of the β parameters when each observation has a distribution in a particular family of two-parameter distributions. This two-parameter family is chosen so that a trick commonly used in estimation for normal theory linear models works for the entire family. The trick is that maximum likelihood estimates of β can be found easily for any value of ϕ. These estimates do not depend on ϕ; therefore, the estimates must be maximum likelihood even when ϕ is an unknown parameter. Given the maximum likelihood estimates of β,

finding the maximum likelihood estimate of ϕ requires solving one equation: $d\ell(\hat{\beta};\phi)/d\phi = 0$. This method is used by Christensen (1996b, Section 2.4) to find maximum likelihood estimates for normal theory linear models. Maximum likelihood estimation of ϕ seems preferable to the essentially ad hoc methods that are illustrated above.

Of the examples that we have considered, Poisson regression and logistic regression are generalized linear models for one-parameter exponential families. Poisson sampling seems to be relatively uncommon for contingency tables; the standard sampling schemes are multinomial and product-multinomial. However, under very mild conditions, maximum likelihood estimates for Poisson sampling are also maximum likelihood estimates for the other sampling schemes. Of course, logistic regression can also be viewed as a special case of product-multinomial log-linear modeling. In regard to Poisson sampling, Santner and Duffy (1989, Problem 3.3) present an interesting data set. The data, originally given in Quine (1975), are on the number of absences of 113 Australian school children. The data are categorized using four factors: age at three levels, sex, cultural background (aboriginal, white), and learning ability (slow, average). The number of absences for different children might be considered as observations on independent Poisson random variables. Note that the number of cross-classifications from the four factors is $3 \times 2 \times 2 \times 2 = 24$, but there are 113 Poisson observations. The analysis of such data would be analogous to a four-factor analysis of variance with unequal numbers of replications on the various treatments. Moreover, Santner and Duffy (1989, p. 135) suggest that the data suffer from *overdispersion*, i.e., $\phi > 1$. It is interesting to note that an observed value of $X^2/(n-p)$ much larger than 1 can indicate *either* lack of fit *or* overdispersion. See McCullagh and Nelder (1989, Sections 4.5, 5.5) for discussion of overdispersion.

The other two examples considered in this chapter, normal theory linear models and gamma distribution regression, involve the two-parameter family of distributions that was used in the basic theory. Generalized linear model methods can be used to analyze other useful models; see McCullagh and Nelder (1989) for a broad range of applications.

While generalized linear models are a useful idea and provide an excellent computing device, care must be taken in their application. For log-linear models, Poisson sampling does not always lead to the same analysis as multinomial and product-multinomial sampling. The distinctions as well as the similarities must be kept in mind. The validity of asymptotic distributions must also be examined carefully. As seen in Section 11.2, a careful analysis of asymptotic issues can be quite complicated. Some extensions of the basic theory such as overdispersion, e.g., allowing ϕ to be a nondegenerate parameter in binomial and Poisson sampling, and *quasi-likelihood* methods have been proposed. Such extensions are widely accepted as providing valuable data analytic tools; however, many people have difficulty in understanding the theoretical basis for them.

9.5 Exercises

EXERCISE 9.5.1. Show that for the binomial model of Section 1, $r(\theta) = \log(1 + e^{\theta})$ and that $\dot{r}(\theta) = p$ and $\ddot{r}(\theta) = p(1 - p)$.

EXERCISE 9.5.2. For the gamma model of Section 1, use the definition of θ and $r(\theta)$ to show that the mean and variance are as given.

EXERCISE 9.5.3. Show that the likelihood equations (9.2.2) have the form $Q_j(\beta)/\phi = 0$, $j = 1, \ldots, p$.

EXERCISE 9.5.4. Show that if $f(y_i|\theta, \phi; w)$ from (9.1.1) is the common density of independent observations y_i, $i = 1, \ldots, n$, then $\sum_{i=1}^{n} y_i$ has a density $f(y_i|\theta_*, \phi_*, w_*)$ for some θ_*, ϕ_*, and w_*.

EXERCISE 9.5.5. Let $Y = (y_1, \ldots y_n)'$. Show that a generalized linear model with canonical link has $X'Y$ as a sufficient statistic.

EXERCISE 9.5.6. Using the definitions of this chapter, find the Pearson statistic (defined in Section 3) for Poisson and binomial regression in terms of the y_i's and m_i's. Show that these are identical to the Pearson statistics defined in Chapter 2.