Chapter 3

Categorical Data

3.1 Introduction

A particularly important application of generalised linear models is the analysis of categorical data. Here, the data are observations of one or more categorical variables on each of a number of units (often individuals). Therefore, each of the units are *cross-classified* by the categorical variables. For example, the dataset `job` represents the cross-classification of 901 individuals according to their income (4 levels) and job satisfaction (4 levels).

	Job Statisfaction			
	Very	Little	Moderately	Very
Income (US$)	Dissatisfied	Dissatisfied	Satisfied	Satisfied
<6000	20	24	80	82
6000–15000	22	38	104	125
15000–25000	13	28	81	113
>25000	7	18	54	92

Source: 1984 General Social Survey.

A cross-classification table, like the one above is called a *contingency table*. This is a *two-way table*, as there are two classifying variables. We also describe the table above as a 4×4 contingency table.

Cell Type	Sex	Remission No	Remission Yes
Nodular	Male	1	4
	Female	2	6
Diffuse	Male	12	1
	Female	3	1

The above `lymphoma` dataset, representing 30 lymphoma patients classified by sex, cell type of lymphoma and response to treatment, is an example of a three-way contingency table. It is a $2 \times 2 \times 2$ or 2^3 table.

Each position in a contingency table is called a *cell* and the number of individuals in a particular cell is the *cell count*. Partial classifications derived from the table are called *margins*. For a two-way table these are often displayed in the margins of the table as below.

	Job Statisfaction				
Income (US$)	Very Dissatisfied	Little Dissatisfied	Moderately Satisfied	Very Satisfied	Total
<6000	20	24	80	82	**206**
6000–15000	22	38	104	125	**289**
15000–25000	13	28	81	113	**235**
>25000	7	18	54	92	**171**
Total	**62**	**108**	**319**	**412**	**901**

These are one-way margins as they represent the classification of items by a single variable; Histological type and Response respectively.

For *multiway* tables, higher margins may be calculated. For example, for `lymphoma`, the two-way Cell type/Sex margin is

Sex	Cell Type Nodular	Cell Type Diffuse
Male	5	13
Female	8	4

We can model contingency table data using generalised linear models. To do this, we assume that the cell counts are observations of independent Poisson random variables. This is intuitively sensible as the cell counts are non-negative integers (the sample space for the Poisson distribution). Therefore, if the table has n cells, which we label $1, \ldots, n$, then the observed cell counts $y_1, \ldots, y_n$ are assumed to be observations of independent Poisson random variables $Y_1, \ldots, Y_n$. We denote the means of these Poisson random variables by $\mu_1, \ldots, \mu_n$. The canonical link function for the Poisson distribution is the log function, and we assume this link function throughout. A generalised linear model for Poisson data using the log link function is called a *log-linear model*.

The explanatory variables in a log-linear model for contingency table data are the cross-classifying variables. As these variables are categorical, they are *factors*. As usual with factors, we can include interactions in the model as well as just main effects. Such a model will describe how the expected count in each cell depends on the classifying variables, and any interactions between them. Interpretation of these models will be discussed further in Section 3.4.

Log-linear data structure for the `job` dataset.

Cell(i)	income	satis	count (y_i)
1	<6000	Dissatis	20
2	<6000	L.Dissatis	24
3	<6000	MSatis	80
4	<6000	VSatis	82
5	6-15	Dissatis	22
6	6-15	L.Dissatis	38
7	6-15	MSatis	104
8	6-15	VSatis	125
9	15-25	Dissatis	13
10	15-25	L.Dissatis	28
11	15-25	MSatis	81
12	15-25	VSatis	113
13	>25000	Dissatis	7
14	>25000	L.Dissatis	18
15	>25000	MSatis	54
16	>25000	VSatis	92

Log-linear data structure for the `lymphoma` dataset.

Cell (i)	Response (y_i)	Explanatory variables		
		Cell Type	Sex	Remission
1	1	Nodular	Male	No
2	2	Nodular	Female	No
3	12	Diffuse	Male	No
4	3	Diffuse	Female	No
5	4	Nodular	Male	Yes
6	6	Nodular	Female	Yes
7	1	Diffuse	Male	Yes
8	1	Diffuse	Female	Yes

3.2 Multinomial sampling

Although the assumption of Poisson distributed observations is convenient for the purposes of modelling, it is often untenable, as a result of the way in which the data have been collected. Frequently, when contingency table data are obtained, the total number of observations (the *grand total*, the sum of all the cell counts) is fixed in advance. In this case, no individual cell count can exceed the prespecified fixed total, so the assumption of Poisson sampling is invalid as the sample space is bounded. Furthermore, with a fixed total, the observations can no longer be observations of independent random variables.

For example, consider the `lymphoma` dataset.

Cell Type	Sex	Remission	
		No	Yes
Nodular	Male	1	4
	Female	2	6
Diffuse	Male	12	1
	Female	3	1

It may be that these data were collected over a fixed period of time, and that in that time there happened to be 30 patients. In this case, the Poisson assumption is perfectly valid. However, it

may have been decided at the outset to collect data on 30 patients, in which case the grand total is fixed at 30, and the Poisson assumption is not valid.

When the grand total is fixed, a more appropriate distribution for the cell counts is the *multinomial* distribution. The multinomial distribution is the distribution of cell counts arising when a prespecified total of N items are each independently assigned to one of n cells, where the probability of being classified into cell i is p_i, $i = 1, \ldots, n$, so $\sum_{i=1}^{n} p_i = 1$. The probability function for the multinomial distribution is

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}; \mathbf{p}) &= P(Y_1 = y_1, \ldots, Y_n = y_n) \\ &= \begin{cases} N! \prod_{i=1}^{n} \frac{p_i^{y_i}}{y_i!} & \text{if } \sum_{i=1}^{n} y_i = N \\ 0 & \text{otherwise} \end{cases} \quad (1) \end{aligned}$$

It is straightforward to see that the binomial is the special case of the multinomial with two cells ($n = 2$).

The expected value of a vector of multinomial cell counts $\mathbf{Y}$ is $N\mathbf{p}$, that is

$$\mu_i = E(Y_i) = Np_i \qquad i = 1, \ldots, n.$$

We can still use a log-linear model for contingency table data when the data have been obtained by multinomial sampling. We model $\log \mu_i = \log(Np_i)$, $i = 1, \ldots, n$ as a linear function of explanatory variables. However, such a model must preserve $\sum_{i=1}^{n} \mu_i = N$, the grand total which is fixed in advance, by design.

From (1), the log likelihood for a multinomial log-linear model is

$$\log f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\mu}) = \sum_{i=1}^{n} y_i \log \mu_i - N \log N - \sum_{i=1}^{n} \log y_i! + \log N!.$$

Therefore, the maximum likelihood estimates $\hat{\boldsymbol{\mu}}$ maximise $\sum_{i=1}^{n} y_i \log \mu_i$ subject to the constraints $\sum_{i=1}^{n} \mu_i = N = \sum_{i=1}^{n} y_i$ (multinomial sampling) and $\log \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ (imposed by the model).

For a Poisson log-linear model,

$$f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\mu}) = \prod_{i=1}^{n} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}$$

Therefore,

$$\begin{aligned} \log f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\mu}) &= -\textstyle\sum_{i=1}^{n} \mu_i + \sum_{i=1}^{n} y_i \log \mu_i - \sum_{i=1}^{n} \log y_i! \qquad (2) \\ &= -\textstyle\sum_{i=1}^{n} \exp(\log \mu_i) + \sum_{i=1}^{n} y_i \log \mu_i - \sum_{i=1}^{n} \log y_i!. \end{aligned}$$

Now any Poisson log-linear model with an intercept can be expressed as

$$\log \mu_i = \alpha + \text{ other terms depending on } i \qquad i = 1, \ldots, n$$

Now

$$\begin{aligned} & \frac{\partial}{\partial \alpha} \log f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\mu}) = -\sum_{i=1}^{n} \exp(\log \mu_i) + \sum_{i=1}^{n} y_i \\ \Rightarrow \quad & \sum_{i=1}^{n} \hat{\mu}_i = \sum_{i=1}^{n} y_i. \end{aligned} \tag{3}$$

From (2), we notice that, at $\alpha = \hat{\alpha}$ the log likelihood takes the form

$$\log f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\mu}) = -\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} y_i \log \mu_i - \sum_{i=1}^{n} \log y_i!.$$

Hence, when we maximise the log-likelihood, for a Poisson log-linear model with intercept, with respect to the other log-linear parameters, we are maximising $\sum_{i=1}^{n} y_i \log \mu_i$ subject to the constraints $\sum_{i=1}^{n} \mu_i = \sum_{i=1}^{n} y_i$ from (3) and $\log \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ (imposed by the model).

Therefore, the maximum likelihood estimates for multinomial log-linear parameters are identical to those for Poisson log-linear parameters. Furthermore, the maximised log-likelihoods for both Poisson and multinomial models take the form $\sum_{i=1}^{n} y_i \log \hat{\mu}_i$ as functions of the log-linear parameter estimates. Therefore any inferences based on maximised log-likelihoods (such as likelihood ratio tests) will be the same.

Therefore, we can analyse contingency table data using Poisson log-linear models, even when the data has been obtained by multinomial sampling. All that is required is that we ensure that the Poisson model contains an intercept term.

3.3 Product multinomial sampling

Sometimes, the grand total is fixed in advance, as a result of certain margins being prespecified. For example, consider the `lymphoma` dataset:

Cell Type	Sex	Remission No	Remission Yes
Nodular	Male	1	4
	Female	2	6
Diffuse	Male	12	1
	Female	3	1

It may have been decided at the outset to collect data on 18 male patients and 12 female patients. Alternatively, perhaps the distribution of both the Sex and Cell type of the patients was fixed in

advance to be

Sex	Cell Type Nodular	Diffuse
Male	5	13
Female	8	4

In cases where a margin is fixed by design, the data consist of a number of fixed total subgroups, defined by the fixed margin. Neither Poisson nor multinomial sampling assumptions are valid. The appropriate distribution combines a separate, independent multinomial for each subgroup. For example, if just the Sex margin is fixed above, then the appropriate distribution for modelling the data is two independent multinomials, one for males with $N = 18$ and one for females with $N = 12$. Each of these multinomials has four cells, representing the cross-classification of the relevant patients by Cell Type and Remission. Alternatively, if it is the Cell type/Sex margin which has been fixed, then the appropriate distribution is four independent two-cell multinomials (binomials) representing the remission classification for each of the four fixed-total patient subgroups.

When the data are modelled using independent multinomials, then the joint distribution of the cell counts $Y_1, \ldots, Y_n$ is the product of terms of the same form as (1), one for each fixed-total subgroup. We call this a distribution a *product multinomial*. Each subgroup has its own fixed total. The full joint density is a product of n terms, as before, with each cell count appearing exactly once.

For example, if the Sex margin is fixed for `lymphoma`, then the product multinomial distribution has the form

$$f_{\mathbf{Y}}(\mathbf{y};\mathbf{p}) = \begin{cases} N_m! \prod_{i=1}^4 \frac{p_{mi}^{y_{mi}}}{y_{mi}!} N_f! \prod_{i=1}^4 \frac{p_{fi}^{y_{fi}}}{y_{fi}!} \\ \qquad \text{if } \sum_{i=1}^4 y_{mi} = N_m \text{ and } \sum_{i=1}^4 y_{fi} = N_f \\ 0 \qquad \text{otherwise} \end{cases}$$

where N_m and N_f are the two fixed marginal totals (18 and 12 respectively), $y_{m1}, \ldots, y_{m4}$ are the cell counts for the Cell type/Remission cross-classification for males and $y_{f1}, \ldots, y_{f4}$ are the corresponding cell counts for females. Here $\sum_{i=1}^4 p_{mi} = \sum_{i=1}^4 p_{fi} = 1$, $E(Y_{mi}) = N_m p_{mi}$, $i = 1, \ldots, 4$, and $E(Y_{fi}) = N_f p_{fi}$, $i = 1, \ldots, 4$.

Using similar results to those used in §3.2 (but not proved here), we can analyse contingency table data using Poisson log-linear models, even when the data has been obtained by product multinomial sampling. However, we must ensure that the Poisson model contains a term corresponding to the fixed margin (and all marginal terms). Then, the estimated means for the specified margin are equal to the corresponding fixed totals.

For example, for the `lymphoma` dataset, for inferences obtained using a Poisson model to be valid when the Sex margin is fixed in advance, the Poisson model must contain the Sex main effect (and the intercept). For inferences obtained using a Poisson model to be valid when the Cell type/Sex margin is fixed in advance, the Poisson model must contain the Cell type/Sex interaction, and all marginal terms (the Cell type main effect, the Sex main effect and the intercept).

Therefore, when analysing product multinomial data using a Poisson log-linear model, certain terms must be present in any model we fit. Their removal is prohibited. Otherwise the inferences do not remain valid.

A consequence of this result is that we can sometimes think of logistic regression models for binomial data as log-linear models. For example, consider the data analysed in S-Plus worksheet 6 (presented on the next page). We previously analysed this table as binomial data with four explanatory variables S, T, B and R. However, as each combination of the four explanatory factors is present exactly once, we can think of these data as a five-way $3 \times 2 \times 2 \times 2 \times 2$ contingency table, with Survival (U) as the additional classifying variable. Then we can fit log-linear models to the data.

S	T	B	R	Survived? Yes	No
Anterior	≤ 12 hours	Yes	Active	53	6
			Placebo	42	7
		No	Active	207	20
			Placebo	220	42
	> 12 hours	Yes	Active	50	8
			Placebo	44	12
		No	Active	241	29
			Placebo	257	36
Inferior	≤ 12 hours	Yes	Active	41	7
			Placebo	32	5
		No	Active	223	22
			Placebo	210	20
	> 12 hours	Yes	Active	40	4
			Placebo	50	4
		No	Active	226	11
			Placebo	226	13
Other	≤ 12 hours	Yes	Active	12	2
			Placebo	20	8
		No	Active	73	9
			Placebo	83	13
	> 12 hours	Yes	Active	18	2
			Placebo	17	5
		No	Active	90	13
			Placebo	102	18

How do these kinds of model differ?

The logistic regression implicitly assumes that the $STBR$ margin is fixed in advance. The

binomial denominators are input as fixed constants and no model is specified for them. The likelihood is then a product of binomials (a special case of product multinomial).

We know that we can model data of this form using a Poisson log-linear model. However, if we are assuming that the $STBR$ margin is fixed in advance, then $S*T*B*R$ (the $STBR$ interaction and all marginal terms) must be present in any log-linear model we fit. Now suppose that i and j are two cells in the same row of this table, *i.e.* they take identical values of S, T, B and R, but different values of U. Furthermore, i and j are the only cells contributing to a particular fixed marginal total, so y_i and y_j have a fixed sum (call this N). Now

$$\log \mu_i - \log \mu_j = \log \frac{\mu_i}{\mu_j} = \log \frac{Np_i}{Np_j} = \log \frac{p_i}{1-p_i}$$

as $p_i + p_j = 1$ in this product binomial scheme. So we can express logit p_i as the difference between two log cell means. Furthermore, as a log-linear model expresses $\log \mu_i - \log \mu_j$ as a linear function of explanatory variables, it is therefore equivalent to a logistic regression model for the product binomial cell probabilities.

Any term which appears as a function of S, T, B or R (and not of U) in the log-linear model, disappears when we consider $\log \mu_i - \log \mu_j$, as i and j take identical values of S, T, B and R. Any term which depends on U remains. For example, the U main effect appears as $\beta_U(\text{yes}) - \beta_U(\text{no})$ in every logit. The US interaction appears as $\beta_{US}(\text{yes}, s_i) - \beta_U(\text{no}, s_i)$, as $s_i = s_j$. Therefore, this term depends on the value of S for the row. It describes how logit p_i depends on S and hence is the main effect of S in the logistic model.

In general, if we have a $2 \times m_1 \times m_2 \times \cdots \times m_r$ contingency table representing the cross-classification of binary variable U and variables $X_1, \ldots, X_r$, then the log-linear model for the full table which includes $X_1 * X_2 * \cdots * X_r$, is equivalent to the logistic regression model for $P(U = 1)$, with $X_1, \ldots, X_r$ as potential explanatory variables. The intercept for the logistic regression model is derived from the U main effect in the log-linear model; the main effect for X_1 in the logistic regression model is derived from the UX_1 interaction in the log-linear model; the X_1X_2 interaction in the logistic regression model is derived from the UX_1X_2 interaction in the log-linear model, *etc.*

Both of these equivalent models assume (implicitly) that the $X_1X_2 \cdots X_r$ margin is fixed in advance, by design. Even if this is not the case, the inferences are still valid, and we can still use these models to learn about the way in which U depends on the explanatory variables. However, if this margin is not fixed in advance, allowing log-linear models which do not include $X_1 * X_2 * \cdots * X_r$ may provide interesting information about relationships between the other variables. For example, for the `lymphoma` dataset, we can assume that the Cell-type/Sex margin is fixed and still learn about how Remission depends on these variables. But if this margin is not fixed in advance, we can also use log-linear models to learn about how Cell-type and Sex are associated.

If margins are fixed in advance, then this must be respected. If they are not, then more can be learn from the data by not imposing unnecessary conditions on the models.

3.4 Interpreting log linear models

Log linear models for contingency tables enable us to determine important properties concerning the joint distribution of the classifying variables. In particular, the form of our preferred log linear model for a table will have implications for how the variables are associated.

Each combination of the classifying variables occurs exactly once in a contingency table. Therefore, the model with the highest order interaction (between all the variables) and all marginal terms (all other interactions) is the saturated model. The implication of this model is that every combination of levels of the variables has its own mean (probability) and that there are no relationships between these means (no structure). The variables are highly dependent.

To consider the implications of simpler models, we first consider a two-way $r \times c$ table where the two classifying variables R and C have r and c levels respectively. The saturated model $R * C$ implies that the two variables are associated. If we remove the RC interaction, we have the model $R + C$,

$$\log \mu_i = \alpha + \beta_R(r_i) + \beta_C(c_i) \qquad i = 1, \ldots, n$$

where $n = rc$ is the total number of cells in the table. Because of the equivalence of Poisson and multinomial sampling, we can think of each cell mean μ_i as equal to Np_i where N is the total number of observations in the table, and p_i is a cell probability. As each combination of levels of R and C is represented in exactly one cell, it is also convenient to replace the cell label i by the pair of labels j and k representing the corresponding levels of R and C respectively. Hence

$$\log p_{jk} = \alpha + \beta_R(j) + \beta_C(k) - \log N \quad j = 1, \ldots, r, \quad k = 1, \ldots, c.$$

Therefore

$$\begin{aligned} P(R = j, C = k) &= \exp[\alpha + \beta_R(j) + \beta_C(k) - \log N] \\ &\qquad j = 1, \ldots, r, \quad k = 1, \ldots, c. \end{aligned}$$

so

$$\begin{aligned} 1 &= \textstyle\sum_{j=1}^{r} \sum_{k=1}^{c} \exp[\alpha + \beta_R(j) + \beta_C(k) - \log N] \\ &= \tfrac{1}{N} \exp[\alpha] \textstyle\sum_{j=1}^{r} \exp[\beta_R(j)] \sum_{k=1}^{c} \exp[\beta_C(k)] \end{aligned}$$

Furthermore

$$\begin{aligned} P(R = j) &= \textstyle\sum_{k=1}^{c} \exp[\alpha + \beta_R(j) + \beta_C(k) - \log N] \\ &= \tfrac{1}{N} \exp[\alpha] \exp[\beta_R(j)] \textstyle\sum_{k=1}^{c} \exp[\beta_C(k)] \quad j = 1, \ldots, r, \end{aligned}$$

and

$$\begin{aligned} P(C=k) &= \sum_{j=1}^{r} \exp[\alpha + \beta_R(j) + \beta_C(k) - \log N] \\ &= \tfrac{1}{N} \exp[\alpha] \exp[\beta_C(k)] \sum_{j=1}^{r} \exp[\beta_R(j)] \quad k = 1, \ldots, c. \end{aligned}$$

Therefore

$$\begin{aligned} P(R=j)P(C=k) &= \tfrac{1}{N} \exp[\alpha] \exp[\beta_C(k)] \exp[\beta_R(j)] \times 1 \\ &= P(R=j, C=k) \\ & \qquad\qquad j = 1, \ldots, r, \quad k = 1, \ldots, c. \end{aligned}$$

Absence of the interaction RC in a log linear model implies that R and C are independent variables. Absence of main effects is generally less interesting (implying uniformity of a particular margin). Generally, main effects are not removed from a log linear model.

In multiway tables, absence of a two-factor interaction does not necessarily mean that the two variables are independent. For example, consider the `lymphoma` dataset, with 3 binary classifying variables Sex (S), Cell type (C) and Remission (R). A reasonable log linear model for these data is $R * C + C * S$. Hence the RS interaction is absent. The estimated cell means, converted to probabilities, for this model are

Cell Type	Sex	Remission	
		No	Yes
Nodular	Male	0.0385	0.1282
	Female	0.0615	0.2051
Diffuse	Male	0.3824	0.0510
	Female	0.1176	0.0157

Hence the estimated probabilities for the two-way Sex/Remission margin (together with the corresponding one-way margins) are

Sex	Remission		
	No	Yes	
Male	0.4208	0.1792	0.6
Female	0.1792	0.2208	0.4
	0.6	0.4	1

It can immediately be seen that this model does not imply independence of R and S, as $\hat{P}(R,S) \neq \hat{P}(R)\hat{P}(S)$. What the model $R * C + C * S$ implies is that R is independent of S *conditional on* C, that is

$$P(R, S|C) = P(R|C)P(S|C)$$

Another way of expressing this is

$$P(R|S,C) = P(R|C),$$

that is, the probability of each level of R given a particular combination of S and C, does not depend on which level C takes. [Equivalently, we can write $P(S|R,C) = P(S|C)$]. This can be observed by calculating the estimated odds in favour of $R =$ yes over $R =$ no for the `lymphoma` dataset.

Cell Type	Sex	Remission		Odds
		No	Yes	
Nodular	Male	0.0385	0.1282	3.33
	Female	0.0615	0.2051	3.33
Diffuse	Male	0.3824	0.0510	0.13
	Female	0.1176	0.0157	0.13

Therefore, the odds depend only on a patient's Cell type, and not on their Sex.

In general, if we have an R-way contingency table with classifying variables $X_1, \ldots, X_r$, then a log linear model which does not contain the X_1X_2 interaction (and therefore by the principle of marginality contains no interaction involving both X_1 and X_2) implies that X_1 and X_2 are *conditionally independent* given $X_3, \ldots, X_r$, that is

$$P(X_1, X_2|X_3, \ldots, X_r) = P(X_1|X_3, \ldots, X_r)P(X_2|X_3, \ldots, X_r).$$

The proof of this is similar to the proof in the two-way case. Again, an alternative way of expressing conditional independence is

$$P(X_1|X_2, X_3, \ldots, X_r) = P(X_1|X_3, \ldots, X_r)$$

or

$$P(X_2|X_1, X_3, \ldots, X_r) = P(X_2|X_3, \ldots, X_r).$$

Although, for the `lymphoma` dataset, R and S are conditionally independent given C, we have already seen that they are not marginally independent.

Sex	Remission		Odds
	No	Yes	
Male	0.4208	0.1792	0.43
Female	0.1792	0.2208	1.23

Male patients have a much lower probability of remission. The reason for this is that, although R and S are not directly associated, they are both associated with C. Observing the estimated values

it is clear that patients with $C =$ nodular have a greater probability of remission, and furthermore, that female patients are more likely to have this cell type than males. Hence females are more likely to have $R =$ yes than males.

Suppose the factors for a three-way tables are X_1, X_2 and X_3. We can list all possible dependence structures using the graphs (drawn in class) and the following.

1. Model 1 containing the terms X_1, X_2, X_3. All factors are mutually independent.
2. Model 2 containing the terms $X_1 : X_2, X_3$. The factor X_3 is jointly independent of X_1 and X_2.
3. Model 3 containing the terms $X_1 : X_2, X_2 : X_3$. The factors X_1 and X_3 are conditionally independent given X_2.
4. Model 4 containing the terms $X_1 : X_2, X_2 : X_3, X_1 : X_3$. There is no conditional independence structure. This is the model without the highest order interaction term.
5. Model 5 containing $X_1 : X_2 : X_3$. This is the saturated model. No more simplification of dependence structure is possible.

Conditional and marginal association of two variables can therefore often appear somewhat different. Sometimes, the association can be 'reversed' so that what looks like a positive association marginally, becomes a negative association conditionally. This is known as *Simpson's paradox.*

In 1972-74, a survey of women was carried out in an area of Newcastle. A follow-up survey was carried out 20 years later. Among the variables observed in the initial survey was whether or not the individual was a smoker and among those in the follow-up survey was whether the individual was still alive, or had died in the intervening period.

Smoker	Dead	Alive	Odds(Dead)
Yes	139	443	0.31
No	230	502	0.46

Therefore, looking at this table, it appears that the non-smokers had a greater probability of dying. However, there is an important extra variable to be considered, related to both smoking habit and mortality – age (at the time of the initial survey). When we consider this variable, we get the table on the following page. Conditional on every age at outset, it is now the smokers who have a higher probability of dying. The marginal association is reversed in the table conditional on

age, because mortality (obviously) and smoking are associated with age. There are proportionally many fewer smokers in the older agegroups (where the probability of death is greater).

Age	Smoker	Dead	Alive	Odds(Dead)	Odds ratio
18–34	Yes	5	174	0.029	1.02
	No	6	213	0.028	
35–44	Yes	14	95	0.147	2.40
	No	7	114	0.061	
45–54	Yes	27	103	0.262	1.44
	No	12	66	0.182	
55–64	Yes	51	64	0.797	1.61
	No	40	81	0.494	
65–74	Yes	29	7	4.143	1.15
	No	101	28	3.607	
75–	Yes	13	0	—	—
	No	64	0	—	

When making inferences about associations between variables, it is important that all other variables which are relevant are considered. Marginal inferences may lead to misleading conclusions.