4 Log-linear Models for Contingency Tables

4.1 Introduction

Data often arise in the form of counts of events, or are presented as the number of data units that are categorized by certain combinations of attributes, or characteristics. Such data are generally presented in the form of contingency tables.

Categorical variables, or factors, are very common. Some may be thought of as ordinal (e.g. age-group, level of credit risk, level of satisfaction, degree of disability, attitude to some political issue), while others are inherently unordered (drug prescribed, fertilizer applied, crop variety planted, sex, ethnic origin, chemical catalyst used). The analysis of contingency tables often appears early on in statistics courses, with associated χ^2 tests of association, and it is not generally appreciated how complex the analysis of such data can be. There are a number of problems.

4.1.1 Curse of Dimensionality

If a data sample of size n is presented as a multi-way table of counts, categorized by p categorical variables (factors), then the number of cells is an exponential function of the number of factors. For example, if all the factors have just 3 levels then the number of cells is 3^p. For $p = 6$ (not very many variables), the number of cells is 729. Recall that many contingency table methods require that, under any suggested model, each cell should have ≥ 5 observations. This would mean that, under the null hypothesis that there is no effect of any factor, each cell would have the same expected count (required to be ≥ 5), and we would need $n = 5 \times 3^6 = 3,645$ observations. Under a more complex model (allowing main effects and interactions up to some proposed degree), n would have to be very much greater. In general, the more complex the model, the larger the required data set.

In general, high-dimensional tables tend to have very many cells with counts < 5, or counts of zero. There are various approaches.

The complexity of proposed models can be reduced by using expert knowledge to suggest graphical models of causality and interaction, in essence models of conditional independence between factors. See, for example, Whittaker J (1990), Cowell R J *et al* (1999), Bishop C M (2006) Chapter 8.

Another approach might be aggregate over some of the factors and produce tables of reduced dimensionality, thus ignoring the possibility of higher degree interactions. This can introduce serious problems if thoughtlessly applied, and in the next subsection we consider a simple example of what is more generally called Simpson's paradox.

4.1.2 Dangers of Aggregation: Simpson's Paradox

Suppose that a study was carried out using, as the data sample, the first 100 men and the first 100 women convicted of shoplifting or theft in an English city in one year. Thus the numbers of convicted men and women were constrained to be equal. Sentences were classified as lenient or severe according to the average sentence given for each particular crime. The resulting data are summarized in the following contingency table:

	Sentence		
Sex	Lenient	Severe	**Total**
Male	40	60	100
Female	60	40	100
Total	100	100	200

Thus 60% of males were given a severe sentence, as compared with 40% of females. However there is another factor, not controlled for: number of previous convictions. Let us look again.

		Sentence		
No. Previous Convictions	*Sex*	Lenient	Severe	**Total**
0	Male	10 (100%)	0 (0 %)	10 (100%)
	Female	50 (71%)	20 (29%)	70 (100%)
1-2	Male	12 (60%)	8 (40%)	20 (100%)
	Female	10 (50%)	10 (50%)	20 (100%)
≥ 3	Male	18 (26%)	52 (74%)	70 (100%)
	Female	0 (0%)	10 (100%)	10 (100%)
		100	100	200

It is clear that collapsing these data into two dimensions has a misleading effect. The two associations, number of previous convictions with sex, number of previous convictions with sentence, combine to produce a misleading two-way table of sex versus sentence. The three-way table makes it apparent that women are treated more severely than men, while the two-way table (the marginal table summed over previous convictions) suggests that they are treated more leniently.

4.1.3 Examples/Types of Contingency Tables

Contingency tables are of different types. Consider the following four examples.

TABLE 1: Evans (January 2007). Data on occurrences of fatal train collisions, derailments and overruns from 1967 to 2006 The accidents are categorized into 5-year time periods, and into whether they would have been prevented if Automatic Train Protection had been installed.

	ATP-preventable		
Period	Yes	No	**Total**
1967-1971	9	16	25
1972-1976	7	7	14
1977-1981	5	5	10
1982-1986	3	8	11
1987-1991	4	6	10
1992-1996	2	4	6
1997-2001	2	1	3
2002-2006	0	1	1
Total	32	48	80

TABLE 2: Yule (1900). Classifies 205 married couples according to the height of each partner.

	Wife			
Husband	Tall	Medium	Short	**Total**
Tall	18	28	14	60
Medium	20	51	28	99
Short	12	25	9	46
Total	50	104	51	205

TABLE 3: Gillet. A randomized clinical trial of an influenza vaccine. Patients were randomly allocated to two groups, one group given the new vaccine and the other a placebo. The responses were the level of antibody found in the blood six weeks after vaccination.

	Response			
Treatment	Small	Moderate	Large	**Total**
Placebo	25	8	5	38
Vaccine	6	18	11	35
Total	31	26	16	73

TABLE 4: Bishop (1969). Classifies 715 babies according to their 'survival' (survived or died), clinic attended (clinic A or B), and level of ante-natal care received (low or high).

		Survival		
Clinic	*Antenatal care*	Survived	Died	**Total**
A	Low	176	3	179
	High	293	4	297
B	Low	197	17	214
	High	23	2	25
	Total	689	26	715

How should these different tables be modelled?

4.2 Sampling Schemes

A further modelling difficulty relates to the sampling scheme that has produced the data. We need to carefully consider how the data in these tables were collected as this will determine the appropriate probability models to associate with the data, as well as the types of hypotheses that can be tested for.

Case (a): Nothing Fixed
Here, we simply record the number of occurrences or individuals arising in particular categories over, say, a time period of interest. The overall total, as well as the cell counts, are all realizations of random variables. This is the case in Table 1, and we expect to model the cell counts as independent Poisson random variables. It follows that the overall total is also Poisson.

Case (b): Total Sample Size Fixed
We record information on a random sample of individuals, according to the classifications, up to a certain (possibly pre-specified) sample size. This is the case in Table 2 in which the overall sample size of 205 is fixed and cannot be regarded as a random variable.

Case (c): One or more margins Fixed
In Table 3, the row totals corresponding to the 'treatment' margin were constrained by the study design to be 38 and 35 respectively. By implication the overall sample size is also fixed.

Also consider, Table 4. Suppose that a sample of 179+297=476 babies is taken from clinic A, and a sample of 214+25=239 babies from clinic B, and then information is recorded on the level of antenatal care, and survival of the babies born. Then the 'clinic' margin would be considered fixed (476 in Clinic A and 239 in Clinic B) by the sampling design.

Alternatively, let us suppose that the 'clinic'×'care' two-way margin is fixed. Then the following quantities (sample sizes) are fixed by the design:
179 for clinic A, low antenatal care
297 for clinic A, high antenatal care
214 for clinic B, low antenatal care
25 for clinic B, high antenatal care.

In general, if we have a contingency table with m classifying factors, then up to $m-1$ margins can be fixed by the sampling design.

In this chapter (Chapter 4) we will focus our discussion on the analysis of two-way tables, such as Tables 1-3. Three-way (and higher-way) tables, such as Table 4, will be discussed in the next chapter.

4.3 Probability Distributions

Consider the two-way table below with two factors, A and B, where the former occurs at J levels, and the latter at K levels. Let Y_{jk} be the frequency for the (j,k)-th cell of the table.

	B_1	B_2	...	B_K	**Total**
A_1	Y_{11}	Y_{12}	...	Y_{1K}	$Y_{1.}$
A_2	Y_{21}	Y_{22}	...	Y_{2K}	$Y_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_J	Y_{J1}	Y_{J2}	...	Y_{JK}	$Y_{J.}$
Total	$Y_{.1}$	$Y_{.2}$	...	$Y_{.K}$	$Y_{..} = n$

Then,

$$Y_{j.} = \sum_{k=1}^{K} Y_{jk}, \quad j = 1, \ldots, J,$$

and

$$Y_{.k} = \sum_{j=1}^{J} Y_{jk}, \quad k = 1, \ldots, K,$$

are the row and column totals respectively. Also,

$$Y_{..} = \sum_{j=1}^{J}\sum_{k=1}^{K} Y_{jk} = \sum_{j=1}^{J} Y_{j.} = \sum_{k=1}^{K} Y_{.k} = n \quad j = 1, \ldots, J; k = 1, \ldots, K,$$

is the total sample size n.

Case (a):
Here, it is supposed that the Y_{jk} are mutually independent and follow a Poisson distribution, each with parameter λ_{jk}. Label the realizations of these cell counts by the y_{jk}. Then the joint distribution of the cell counts is:

$$f(\mathbf{y}; \boldsymbol{\lambda}) = \prod_{j=1}^{J}\prod_{k=1}^{K} \frac{\lambda_{jk}^{y_{jk}} e^{-\lambda_{jk}}}{y_{jk}!} \tag{1}$$

Also note that if we set $\lambda_{..} = \sum_{j=1}^{J}\sum_{k=1}^{K} \lambda_{jk}$, then the distribution of the overall total ('sample size' n) is

$$f(n) = \frac{\lambda_{..}^{n} e^{-\lambda_{..}}}{n!} \tag{2}$$

Case (b):
Using (1) and (2), the conditional distribution of the cell counts, given that the total sample size, $\sum_{j=1}^{J}\sum_{k=1}^{K} y_{jk}$, is equal to n, is:

$$f(\mathbf{y}|n) = f(\mathbf{y}, n)/f(n) \tag{3}$$

$$= \frac{\prod_{j=1}^{J}\prod_{k=1}^{K} \frac{\lambda_{jk}^{y_{jk}} e^{-\lambda_{jk}}}{y_{jk}!}}{\frac{\lambda_{..}^{n} e^{-\lambda_{..}}}{n!}} = n! \prod_{j=1}^{J}\prod_{k=1}^{K} \frac{\theta_{jk}^{y_{jk}}}{y_{jk}!} \tag{4}$$

where $\theta_{jk} = \lambda_{jk}/\lambda_{..}$, is the cell probability (of landing in the jth row and kth column of the table).

Thus, when the sample size is fixed by design, the cell counts follow a *Multinomial distribution* with parameters n and $\{\theta_{jk}\}$, where $\sum_{j=1}^{J}\sum_{k=1}^{K}\theta_{jk}=1$.

Case (c):
Suppose that we fix the margin corresponding to factor A. This corresponds to the row totals, $\{y_{j.}\}$, being fixed by design. We say that A is an *explanatory* variable or factor, whereas B is a *response* factor.
By analogy with Case (b), the (joint) distribution of the cell counts at the j-th row of the table, is again Multinomial, with parameters $y_{j.}$, and $\{\theta_{jk}\}$, where $\sum_{k=1}^{K}\theta_{jk}=1$ ($\theta_{jk}=\lambda_{jk}/\lambda_{j.}$):

$$f(y_{j1},\ldots,y_{jK}|y_{j.})=y_{j.}!\prod_{k=1}^{K}\frac{\theta_{jk}^{y_{jk}}}{y_{jk}!}$$

Assuming that the rows are mutually independent, the joint distribution of the cell counts for the whole table is:

$$f(\mathbf{y}|y_{j.},j=1,\ldots,J)=\prod_{j=1}^{J}y_{j.}!\prod_{k=1}^{K}\frac{\theta_{jk}^{y_{jk}}}{y_{jk}!} \tag{5}$$

where $\sum_{k=1}^{K}\theta_{jk}=1$ for $j=1,\ldots,J$. (In this case, we say the cell counts follow a *Product Multinomial distribution.*

4.4 Log-linear Models

Recall the following result regarding the mean of the Multinomial distribution.

Lemma 4.1
Suppose that the joint distribution of $Y_1,\ldots,Y_N$ is Multinomial with parameters n and $\{\theta_i\}$, thus

$$f(\mathbf{y})=n!\prod_{i=1}^{N}\frac{\theta_i^{y_i}}{y_i!}$$

where the $\{\theta_i\}$ are non-negative, $\sum_{i=1}^{N}\theta_i=1$, and $\sum_{i=1}^{N}y_i=n$.

Then

$$E[Y_i]=n\theta_i$$

We shall use this general result to compute the expectations of the cell frequencies in cases (b) and (c). In so doing, we also consider the forms of these expressions under certain hypotheses of interest, and then transform these into 'log-linear form'.

Case (a) nothing fixed
Here, we are simply dealing with the expectations of Poisson random variables, and so

$$E[Y_{jk}]=\lambda_{jk} \tag{6}$$

Now under the hypothesis of *independence* between factors A and B, the probability, θ_{jk}, that, given an occurrence, it falls into the (j,k)-th cell of the table, can be written as

$$\theta_{jk} = \theta_{j.} \times \theta_{.k}$$

where $\sum_{k=1}^{K}\sum_{j=1}^{J}\theta_{jk} = 1$, and $\theta_{j.}$ and $\theta_{.k}$ are the marginal probabilities of landing in the j-th row and k-th column of the table, respectively. But $\theta_{jk} = \lambda_{jk}/\lambda_{..}$, $\theta_{j.} = \lambda_{j.}/\lambda_{..}$, and $\theta_{.k} = \lambda_{.k}/\lambda_{..}$. Therefore

$$\lambda_{jk} = \frac{\lambda_{j.} \times \lambda_{.k}}{\lambda_{..}}$$

and, under independence,

$$E[Y_{jk}] = \lambda_{jk} = \frac{\lambda_{j.} \times \lambda_{.k}}{\lambda_{..}} \tag{7}$$

Taking the logarithm of (6), and with an appropriate re-definition of the resulting terms on the RHS, we have

$$\eta_{jk} = \log E[Y_{jk}] = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} \tag{8}$$

Applying the same procedure to the 'independence' expression (7) yields

$$\eta_{jk} = \log E[Y_{jk}] = \mu + \alpha_j + \beta_k \tag{9}$$

By analogy with analysis of variance, the hypothesis of independence between factors A and B is equivalent to $(\alpha\beta)_{jk} = 0$ for $j=1,\ldots,J$, $k=1,\ldots,K$, i.e. the 'interaction' terms are zero.

Case (b):total sample size n fixed
Under this (multinomial) distribution

$$E[Y_{jk}] = n\theta_{jk} \tag{10}$$

Now under the hypothesis of *independence* between factors A and B, the probability for the (j,k)-th cell of the table, θ_{jk}, can be written as

$$\theta_{jk} = \theta_{j.} \times \theta_{.k}$$

where $\theta_{j.}$ and $\theta_{.k}$ are the marginal probabilities of landing in the j-row and k-th column of the table, respectively. Hence, under independence,

$$E[Y_{jk}] = n \times \theta_{j.} \times \theta_{.k} \tag{11}$$

Again, taking the logarithms of (10) yields

$$\eta_{jk} = \log n + \log \theta_{jk} \tag{12}$$

Applying the same procedure to the 'independence' expression (11) yields

$$\eta_{jk} = \log n + \log \theta_{j.} + \log \theta_{.k} \tag{13}$$

For this case, it again turns out that the full model (12), and the independence model (13) can be represented by (8) and (9) respectively. [Note one difference however. In Case (a) the constant term is unknown and is a parameter to be estimated. In Case (b) the constant term is a function of the overall total count and was fixed by the sample design.]

Case (c): one margin fixed (e.g. row totals)
Under this (product multinomial) distribution

$$E[Y_{jk}] = y_{j.}\theta_{jk} \tag{14}$$

where $\sum_k \theta_{jk} = 1$ for $j = 1, \ldots, J$.

Now consider the hypothesis of 'homogeneity', in which the conditional probability for being in the k-th cell given the j-th row, does not depend on the row j, i.e.

$$\theta_{jk} = \theta_k$$

Hence, under homogeneity,

$$E[Y_{jk}] = y_{j.}\theta_k \tag{15}$$

Taking the log's of (14) and (15) yields

$$\eta_{jk} = \log E[Y_{jk}] = \log y_{j.} + \log \theta_{jk}$$

and

$$\eta_{jk} = \log E[Y_{jk}] = \log y_{j.} + \log \theta_k$$

which can, once again, be related to (8) and (9) respectively, by reparameterizing.

So this time, homogeneity (rather than independence, which has no meaning in this case anyway), requires that all of the 'interaction' terms are equal to zero.

The 'log-linear models' that have arisen from the above discussion actually amount to saying that

$$\eta_i = \log E[Y_i] = \mathbf{x}_i'\boldsymbol{\beta}, \quad i = 1, \ldots, N \tag{16}$$

for appropriately chosen $\mathbf{x}_i$ and $\boldsymbol{\beta}$. Thus, we relate the mean of the cell counts, to the linear predictor η_i through the *log link function*. Furthermore, in (a), we have a Poisson error structure for these counts; thus, in this case, we can test hypotheses using the corresponding GLM. The vector $\boldsymbol{\beta}$ is estimated using maximum likelihood estimation, which, by the invariance property, yields the fitted means, $\hat{\mu}_i = e^{\hat{\eta}_i}$.

But what about cases (b) and (c)?

(Product-) Multinomial data may be analyzed as though they were independent Poisson data, with log-linear predictor, PROVIDED TERMS CORRESPONDING TO THE FIXED MARGINS ARE INCLUDED IN THE MODEL. The deviance will be correct, as well as the estimates.

Why is this? Briefly, we have shown that a multinomial distribution can be obtained from independent Poisson observations, conditioned on the overall total. If observations from a multinomial distribution are treated as independent Poisson observations and we fit a log-linear model that includes a constant term, then the corresponding maximum likelihood estimate of the intercept turns out to be precisely that estimate that constrains the overall fitted total to

equal the overall observed total. The required conditioning is immediately imposed.

Suppose, for example, that the row totals are fixed by design, i.e. the $\{y_{j\cdot}\}$ are fixed. We always fit the terms

$$\mu + \alpha_j$$

so that the required conditioning is imposed. The other terms in the linear predictor of the full model are

$$\beta_k + (\alpha\beta)_{jk}$$

corresponding to θ_{jk}: terms in this part of the predictor can be removed to test for the various hypotheses.

As is the case with ANOVA models, our statistical model is over parameterized. Thus, in order to obtain (unique) estimates, we need to impose appropriate constraints. We use the *corner point* constraints, now the default R. Thus we set:

$$\alpha_1 = 0, \quad \beta_1 = 0, \quad (\alpha\beta)_{1k} = 0,\ k{=}1, \ldots, K, \quad (\alpha\beta)_{j1} = 0,\ j{=}1, \ldots, J.$$

Imposing any suitable set of constraints on the 'full' model, the free parameters are: 1 from μ, $(J-1)$ from the $\{\alpha_j\}$, $(K-1)$ from the $\{\beta_k\}$, and $(J-1)(K-1)$ from the $\{(\alpha\beta)_{jk}\}$, which adds up to JK, the total number of cells in the table! Thus, our full model (8) corresponds to the 'saturated' model.

Example 4.2 (Heights)
Suppose that the data in TABLE 2 were collected in such a way that only the total sample size is fixed by the design. Determine whether there is any association between the height of the husband and that of the wife.

Solution:

```
> count <- c(18, 28, 14, 20, 51, 28, 12, 25, 9)
> w <- rep(c("T", "M", "S"), 3)
> h <- c(rep("T", 3), rep("M", 3), rep("S", 3))

> wife <- factor(w)
> husband <- factor(h)

> height <- data.frame(count, husband, wife)
```

```
> height
  count husband wife
1    18       T    T
2    28       T    M
3    14       T    S
4    20       M    T
5    51       M    M
6    28       M    S
7    12       S    T
8    25       S    M
9     9       S    S

> height.glm <- glm(count ~ husband + wife, family = poisson)

> summary(height.glm)

Call:
glm(formula = count ~ husband + wife, family = poisson)

Deviance Residuals:
      1        2        3        4        5        6        7        8
 0.8490  -0.4482  -0.2424  -0.8699   0.1092   0.6645   0.2304   0.3404
      9
-0.7508

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.9165     0.1218  32.152  < 2e-16 ***
husbandS     -0.7665     0.1784  -4.295 1.74e-05 ***
husbandT     -0.5008     0.1636  -3.061  0.00221 **
wifeS        -0.7126     0.1709  -4.168 3.07e-05 ***
wifeT        -0.7324     0.1721  -4.256 2.08e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 50.5890  on 8  degrees of freedom
Residual deviance:  2.9232  on 4  degrees of freedom
AIC: 56.57

Number of Fisher Scoring iterations: 4
```

The (scaled) deviance of 2.9232 on the χ^2_4 distribution is not significant.

```
> dev <- height.glm$deviance
> pchisq(dev, 4, lower.tail=F)
[1] 0.570763
```

Thus, there is no evidence to reject the hypothesis of independence.

The fitted means can be obtained under this 'independence' model:

```
> round(fitted(height.glm), 3)
     1      2      3      4      5      6      7      8      9
14.634 30.439 14.927 24.146 50.224 24.629 11.220 23.337 11.444

> sum(fitted(height.glm))
[1] 205
```

Note that the fitted total is precisely the fixed total as required.

The estimates of the parameters in our linear predictor can be extracted using the `coef` function, or the `dummy.coef` function if the zero parameters are to be included.

```
> coef(height.glm)
(Intercept)    husbandS    husbandT       wifeS       wifeT
  3.9165008  -0.7664785  -0.5007753  -0.7125653  -0.7323679

> dummy.coef(height.glm)
Full coefficients are

(Intercept):      3.916501
husband:                M          S          T
                0.0000000 -0.7664785 -0.5007753
wife:                   M          S          T
                0.0000000 -0.7125653 -0.7323679
```

Thus

$$\widehat{\mu} = 3.9165, \quad \widehat{\alpha}_M = 0, \quad \widehat{\alpha}_S = -0.76648, \ldots, \ldots, \widehat{\beta}_T = -0.73237$$