5 Log-linear Models for 3-way Contingency Tables

5.1 Introduction

We extend our techniques from the previous chapter in order to deal with tables of count data classified according to 3 factors. You will see that there are many different sorts of independence between the variables that are possible. Extensions to more than 3 factors can be deduced by analogy from the 2 and 3 factor cases, and is left to you as an exercise!

There are 4 possibilities to consider: (i) Nothing is fixed by the sampling design, (ii) Only the total sample size fixed, (iii) One factor (or variable) is fixed by the design, so this must be considered as an explanatory variable, while the other two may be regarded as response variables (or we may regard one as a response variable and condition on the other), and finally (iv) two factors are fixed so that we have one response variable and two explanatory variables. Case (i) is not very common in practice; and since we would want to fit a constant term in the linear predictor anyway, the log-linear models for case (ii) should suffice, allowing for a different interpretation of the constant term.

Cases (i)-(iv) reflect the possible models when the sampling design dictates the inclusion of certain terms in the null model. You should also be aware that there are models that allow a factor to be regarded as an explanatory variable even if the data are purely observational, with no margins fixed in advance.

We consider cases (ii)-(iv) in turn, stating the relevant probability models for the realizations of the table, the expressions for the expectations of the cells counts, along with the corresponding log-linear models. We shall also indicate the terms that need to be included in the linear predictor in order to justifiably fit a GLM with Poisson error structure and log link.

5.2 Notation

The data are classified according to 3 factors, A, B, and C, each having J,K, and L levels, respectively.

Let Y_{jkl} be the count for the cell at the j-th row, k-th column, and l-th layer of the table, and let y_{jkl} be its realization.

Let θ_{jkl} be the probability for the cell at the j-th row, k-th column, and l-th layer of the table.

5.3 Three Response Variables

(Case (ii).) With just the sample size fixed, the appropriate distribution is multinomial with parameters n, and $\{\theta_{jkl}\}$

$$f(\mathbf{y}, \boldsymbol{\theta}) = n! \prod_{j=1}^{J} \prod_{k=1}^{K} \prod_{l=1}^{L} \frac{\theta_{jkl}^{y_{jkl}}}{y_{jkl}!}$$

where $\sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{l=1}^{L} \theta_{jkl} = 1$ and $\sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{l=1}^{L} y_{jkl} = n$.

Thus

$$E[Y_{jkl}] = n\theta_{jkl} \tag{1}$$

Taking the logarithm of (1) yields

$$\log E[Y_{jkl}] = \log n + \log \theta_{jkl}$$

which corresponds to the (saturated) log-linear model

$$\eta_{jkl} = \underline{\mu} + \alpha_j + \beta_k + \gamma_l + (\alpha\beta)_{jk} + (\alpha\gamma)_{jl} + (\beta\gamma)_{kl} + (\alpha\beta\gamma)_{jkl} \tag{2}$$

The underlined term 'corresponds' to the $\log n$ term; since n is specified by design, then we must also always include the constant term when considering smaller, more parsimonious models.

The number of independent parameters is equal to

$$1+(J-1)+(K-1)+(L-1)+(J-1)(K-1)+(J-1)(L-1)+(K-1)(L-1)+(J-1)(K-1)(L-1) = JKL.$$

There are eight different log-linear hypotheses corresponding to types of independence. Each of these hypotheses can be determined by assessing the fit of the relevant model. These are outlined below.

H_0: Complete Independence Model
This model is also known as the *mutual independence model*, since it indicates that there is no association between the variables. That is, A is independent of B, is independent of C, is independent of A.

Clearly, under this model, $\theta_{jkl} = \tilde{\theta}_{j\diamond\diamond} \times \tilde{\theta}_{\diamond k\diamond} \times \tilde{\theta}_{\diamond\diamond l}$ and so

$$E[Y_{jkl}] = n\, \tilde{\theta}_{j\diamond\diamond}\, \tilde{\theta}_{\diamond k\diamond}\, \tilde{\theta}_{\diamond\diamond l} \tag{3}$$

This implies that

$$\log E[Y_{jkl}] = \log n + \log \tilde{\theta}_{j\diamond\diamond} + \log \tilde{\theta}_{\diamond k\diamond} + \log \tilde{\theta}_{\diamond\diamond l}$$

and so the corresponding log-linear model is

$$\eta_{jkl} = \underline{\mu} + \alpha_j + \beta_k + \gamma_l \tag{4}$$

with

$$1 + (J-1) + (K-1) + (L-1) = J + K + L - 2$$

independent parameters.

H_1: Joint Independence Model
Suppose, for definiteness, that the first variable is independent of the other two. That is, A is *jointly independent* of B and C, (where is a partial association between B and C).

Clearly, under this model, $\theta_{jkl} = \tilde{\theta}_{j\diamond\diamond} \times \tilde{\theta}_{\diamond kl}$ which implies that

$$E[Y_{jkl}] = n\,\tilde{\theta}_{j\diamond\diamond}\,\tilde{\theta}_{\diamond kl} \tag{5}$$

The logarithm of this expression is given by

$$\log E[Y_{jkl}] = \log n + \log \tilde{\theta}_{j\diamond\diamond} + \log \tilde{\theta}_{\diamond kl}$$

The corresponding log-linear model is

$$\eta_{jkl} = \underline{\mu} + \alpha_j + \beta_k + \gamma_l + (\beta\gamma)_{kl} \tag{6}$$

This has

$$1 + (J-1) + (K-1) + (L-1) + (K-1)(L-1) = J + KL - 1$$

independent parameters.

There are three such joint independence models as the variables A, B and C are permuted.

H_2: Conditional Independence Model
Suppose, for definiteness, that given the level of the first variable/factor, the other two are independent. We say that$B$ and C are *conditionally independent* of A. That is B and C are independent, given A. (There are partial associations between A and B, and A and C).
Under this model, we have the factorization $\theta_{jkl} = \tilde{\theta}_{j\diamond\diamond} \times \tilde{\theta}_{jk\diamond} \times \theta_{j\diamond l}$. In this case, the expectation of the cell counts is given by

$$E[Y_{jkl}] = n\,\tilde{\theta}_{j\diamond\diamond}\,\tilde{\theta}_{jk\diamond}\,\tilde{\theta}_{j\diamond l} \tag{7}$$

Upon taking the logarithm of (7), we obtain

$$\log E[Y_{jkl}] = \log n + \log \tilde{\theta}_{j\diamond\diamond} + \log \tilde{\theta}_{jk\diamond} + \log \theta_{j\diamond l}$$

This corresponds to the log-linear model

$$\eta_{jkl} = \underline{\mu} + \alpha_j + \beta_k + \gamma_l + (\alpha\beta)_{jk} + (\alpha\gamma)_{jl} \tag{8}$$

This has

$$1 + (J-1) + (K-1) + (L-1) + (J-1)(K-1) + (J-1)(L-1) = J(K+L-1)$$

independent parameters.

There are three such conditional independence models as the variables A, B and C are permuted.

H_3:Partial Association Model
This model is also called the *homogeneous association model*. Here, each pair of factors is unaffected by the level of the third. It is the only hypothesis that cannot be given an interpretation in terms of conditional probability.

Under this model, $n\theta_{jkl}$ factorizes into the form

$$E[Y_{jkl}] = n\,\tilde{\theta}_{jk\diamond}\,\tilde{\theta}_{j\diamond l}\,\tilde{\theta}_{\diamond kl} \tag{9}$$

Taking the logarithm of this yields

$$\log E[Y_{jkl}] = \log n + \log \tilde{\theta}_{jk\diamond} + \log \tilde{\theta}_{j\diamond l} + \log \tilde{\theta}_{\diamond kl}$$

which corresponds to

$$\eta_{jkl} = \underline{\mu} + \alpha_j + \beta_k + \gamma_l + (\alpha\beta)_{jk} + (\alpha\gamma)_{jl} + (\beta\gamma)_{kl} \tag{10}$$

This has

$$JKL - (J-1)(K-1)(L-1)$$

independent parameters.

There are three such conditional independence models as the variables A, B and C are permuted.

It is easily seen that these hypotheses (models) are related. For example,

$$H_0 \subset H_1 \subset H_3, \quad H_0 \subset H_2 \subset H_3 \quad \text{and} \quad H_1 \cap H_2 = H_0$$

Note that **non-comprehensive** models, which do not contain all main effects are also possible. One such example is where,

$$E[Y_{jkl}] = n\tilde{\theta}_{jk\diamond}$$

This corresponds to

$$\eta_{jkl} = \underline{\mu} + \alpha_j + \beta_k + (\alpha\beta)_{jk}$$

with

$$1 + (J-1) + (K-1) + (J-1)(K-1) = JK$$

independent parameters

In this model, the level of factor C does not affect anything. It may be omitted and the $J \times K$ margin should be analysed instead.

5.4 Two Response Variables+One Explanatory Variable

(Case (iii).) Let us suppose, for definiteness, that the margin corresponding to C is fixed, i.e. the layer totals, $\{y_{..l}\}$, are fixed by design. Then, A and B are considered response variables, while the sample design requires the factor C to be considered as an explanatory variable. The appropriate distribution for the realizations of the table is the product multinomial distribution:

$$f(\mathbf{y};\boldsymbol{\theta}) = \prod_{l=1}^{L} y_{..l}! \prod_{j=1}^{J}\prod_{k=1}^{K} \frac{\theta_{jkl}^{y_{jkl}}}{y_{jkl!}}$$

where $\sum_{j=1}^{J}\sum_{k=1}^{K}\theta_{jkl} = 1$, for $l = 1,\ldots,L$.

The expected values of the cell counts take the form

$$E[Y_{jkl}] = y_{..l}\theta_{jkl} \tag{11}$$

Taking the logarithm of this expression yields

$$\log E[Y_{jkl}] = \log y_{..l} + \log \theta_{jkl}$$

Thus the corresponding log-linear model, which is also the saturated model, is:

$$\eta_{jkl} = \underline{\mu} + \alpha_j + \beta_k + \underline{\gamma_l} + (\alpha\beta)_{jk} + (\alpha\gamma)_{jl} + (\beta\gamma)_{kl} + (\alpha\beta\gamma)_{jkl} \tag{12}$$

with JKL independent parameters. The term $\log y_{..l}$ necessitates the inclusion of the terms $\mu + \gamma_l$.

Possible hypotheses are indicated below.

Independence of Responses at each level of the explanatory variable
in which the response variables are independent, given the level of the explanatory variable.

Expectation:

$$E[Y_{jkl}] = y_{..l}\,\tilde{\theta}_{j\diamond l}\,\tilde{\theta}_{\diamond kl} \tag{13}$$

log-linear model:

$$\eta_{jkl} = \underline{\mu} + \alpha_j + \beta_k + \underline{\gamma_l} + (\alpha\gamma)_{jl} + (\beta\gamma)_{kl} \tag{14}$$

No. of Independent Parameters:

$$L(J + K - 1)$$

Homogeneity Model
in which the association between the response variables is the same at each level of the explanatory variable.

EXPECTATION:

$$E[Y_{jkl}] = y_{\cdot\cdot l}\,\tilde{\theta}_{jk\diamond} \tag{15}$$

LOG-LINEAR MODEL:

$$\eta_{jkl} = \underline{\mu} + \alpha_j + \beta_k + \underline{\gamma_l} + (\alpha\beta)_{jk} \tag{16}$$

NO. OF INDEPENDENT PARAMETERS:

$$JK + L - 1$$

5.5 One Response Variable+Two Explanatory Variables

(Case (iv).) Suppose, for definiteness, that the first variable, A, is a response, while the sampling design requires the other two factors to be explanatory variables. This means that the totals $y_{\cdot kl}$ are fixed by the sampling design. Then the appropriate distribution is

$$f(\mathbf{y};\boldsymbol{\theta}) = \prod_{k=1}^{K}\prod_{l=1}^{L} y_{\cdot kl}! \prod_{j=1}^{J} \frac{\theta_{jkl}^{y_{jkl}}}{y_{jkl}!},$$

where $\theta_{\cdot kl} = 1$ for all $k = 1,\ldots,K$, $l = 1,\ldots,L$. This allows for a different, independent multinomial distribution for levels of A for every combination of the levels of B and C, and the expectations of the cell counts take the form

$$E[Y_{jkl}] = y_{\cdot kl}\theta_{jkl}$$

Taking the logarithm of this equation yields

$$\log E[Y_{jkl}] = \log y_{\cdot kl} + \log \theta_{jkl}$$

This corresponds to the (saturated) log-linear model

$$\eta_{jkl} = \underline{\mu} + \alpha_j + \underline{\beta_k + \gamma_l} + (\alpha\beta)_{jk} + (\alpha\gamma)_{jl} + \underline{(\beta\gamma)_{kl}} + (\alpha\beta\gamma)_{jkl} \tag{17}$$

Relevant hypotheses include,

Same distribution for A ($j = 1,\ldots,J$) for all levels of factor B
in which the pattern of responses for A is the same at each level of the explanatory variable B, regardless of the level of C.

EXPECTATION:

$$E[Y_{jkl}] = y_{\cdot kl}\tilde{\theta}_{j\diamond l} \tag{18}$$

LOG-LINEAR MODEL:

$$\eta_{jkl} = \underline{\mu} + \alpha_j + \underline{\beta_k + \gamma_l} + (\alpha\gamma)_{jl} + \underline{(\beta\gamma)_{kl}} \tag{19}$$

NO. OF INDEPENDENT PARAMETERS:

$$L(J + K - 1)$$

Same distribution for A $(j = 1, \ldots, J)$ for all combinations of levels of B and C

EXPECTATION:

$$E[Y_{jkl}] = y_{.kl}\tilde{\theta}_{j\diamond\diamond} \tag{20}$$

LOG-LINEAR MODEL:

$$\eta_{jkl} = \underline{\mu} + \alpha_j + \underline{\beta_k + \gamma_l} + \underline{(\beta\gamma)_{kl}} \tag{21}$$

NO. OF INDEPENDENT PARAMETERS:

$$KL + J - 1$$

Notes

• We have selected particular variables to be either response or explanatory. We can often permute these selections, yielding a corresponding change in the log-linear model. Once you understand the basic principle of deriving these expressions, then such changes and extensions should not pose any problem.

• In deciding which model to endorse, out of a number of valid possibilities which fit the data adequately, we should try to adhere to the principle or *law of parsimony*. This basically says that we try to go for the simplest model possible: it is considered preferable to endorse a simpler model which describes the data adequately, rather than a more complicated one which leaves very little of the variability unexplained.

5.6 Log-Linear Models and Logistic Regression

A brief introduction. [See the example that follows].

Often, in the context of log-linear modelling, we are primarily interested in a particular response variable, and aim to see how it 'depends' on a number of explanatory variables. If such a response is binary, a natural question is *what is the relationship between logistic regression and log-linear modelling for data sets to which both methods can be applied?*

It is the case that there is an exact relationship between the two approaches, subject to the following conditions.

(a) All possible combinations of levels of the explanatory variables are represented in the data, and

(b) the log-linear models include the main effect of the response and a saturated model for the explanatory variables.

To say that the results match exactly here means that the residual deviances will be the same for the corresponding models, and that the corresponding parameter estimates will be the same as well. That is, the main effects of explanatory variables in the logistic regression are identified with interactions between the response variable and explanatory variables in the log-linear model, (and so on). This is because, subject to conditions (a) and (b), the two models amount to the same thing - they both involve logs of probabilities. (You are spared the details)!

5.7 An Example

Consider the data of Bishop (1969), TABLE 4 from Chapter 8, which is repeated below.

Clinic	Antenatal Care	Survival		
		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

Let us assume that only the sample size, 715, has been fixed by design. We shall try to identify whether there is any association between any of the three response variables, which shall be labelled clinic, care, and surv. The data can be read in as follows:

```
> count <- c(176, 3, 293, 4, 197, 17, 23, 2)
> cl <- c(rep("A", 4), rep("B", 4))
> ca <- rep(c("L", "L", "H", "H"), 2)
> su <- rep(c("S", "D"), 4)
> clinic <- factor(cl)
> care <- factor(ca, levels = c("L", "H"))
> surv <- factor(su, levels = c("S", "D"))
> health <- data.frame(clinic, care, surv, count)
> health
  clinic care surv count
1      A    L    S   176
2      A    L    D     3
3      A    H    S   293
4      A    H    D     4
5      B    L    S   197
6      B    L    D    17
7      B    H    S    23
8      B    H    D     2
```

The code (along with output) for testing for (complete) independence between the variables is given below.

```
> health.glm <- glm(count ~ clinic + care + surv, family = poisson, data = health)
> summary(health.glm)

Call:
glm(formula = count ~ clinic + care + surv, family = poisson,
    data = health)

Deviance Residuals:
     1       2       3       4       5       6       7       8
-5.072  -2.470   5.654  -1.501   5.782   4.326  -9.600  -1.069

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.52990    0.05744  96.270  < 2e-16 ***
clinicB     -0.68895    0.07928  -8.690  < 2e-16 ***
careH       -0.19926    0.07517  -2.651  0.00803 **
survD       -3.27714    0.19978 -16.404  < 2e-16 ***
```

```
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1066.43  on 7  degrees of freedom
Residual deviance:  211.48  on 4  degrees of freedom
AIC: 259.66

Number of Fisher Scoring iterations: 5
```

The (scaled) deviance is 211.48 on 4 degrees of freedom; this is highly significant, and so we are forced to reject the hypothesis of independence between all 3 response variables.

Now let us try the model with all three main effects and all three of the two-factor interactions.

```
> summary(update(health.glm,  ~ clinic * care * surv - clinic:care:surv))

Call:
glm(formula = count ~ clinic + care + surv + clinic:care + clinic:surv +
    care:surv, family = poisson, data = health)

...

    Null deviance: 1.0664e+03  on 7  degrees of freedom
Residual deviance: 4.3256e-02  on 1  degrees of freedom
AIC: 54.226

Number of Fisher Scoring iterations: 3
```

The scaled deviance is 0.04326 on 1 df. Clearly this (partial association) model fits. We now need to see if a more parsimonious model would fit.

We try dropping two-factor interaction terms from the previous model until we find a model that adequately fits the data. This leads us to the following model:

```
> health.glm <- glm(count ~ clinic + care + surv + clinic:care + clinic:surv,
                            family = poisson, data = health)
> summary(health.glm)

Call:
glm(formula = count ~ clinic + care + surv + clinic:care + clinic:surv,
    family = poisson, data = health)

Deviance Residuals:
       1          2          3          4          5          6          7
-0.027693   0.221610   0.021487  -0.178476   0.000894  -0.003044  -0.002617
       8
 0.008894

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   5.17257    0.07495  69.011  < 2e-16 ***
clinicB       0.11057    0.10321   1.071    0.284
careH         0.50635    0.09462   5.351 8.74e-08 ***
survD        -4.20469    0.38077 -11.042  < 2e-16 ***
```

```
clinicB:careH -2.65345    0.23157 -11.458  < 2e-16 ***
clinicB:survD  1.75550    0.44963   3.904 9.45e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1.0664e+03  on 7  degrees of freedom
Residual deviance: 8.2289e-02  on 2  degrees of freedom
AIC: 52.265

Number of Fisher Scoring iterations: 4
```

The deviance of 0.08229 on 2 degrees of freedom is not significant and so we endorse this model (other models with all main effects and 2 out of the 3 two-factor interactions were not found to fit adequately, as is the case for the smaller models). This corresponds to the *conditional independence* model of (8). This says that *given the clinic, survival prospects, and standard of antenatal care are independent.*

The estimated parameters and fitted responses are shown below.

```
> dummy.coef(health.glm)
Full coefficients are

(Intercept):     5.172571
clinic:                 A         B
               0.0000000 0.1105693
care:                   L         H
               0.0000000 0.5063463
surv:                   S         D
                0.000000 -4.204693
clinic:care:          A:L       B:L       A:H       B:H
                0.000000  0.000000  0.000000 -2.653447
clinic:surv:          A:S       B:S       A:D       B:D
                0.000000  0.000000  0.000000  1.755504

> coef(health.glm)
  (Intercept)       clinicB         careH         survD clinicB:careH
    5.1725707     0.1105693     0.5063463    -4.2046926    -2.6534465
clinicB:survD
    1.7555041

> fitted(health.glm)
176.367647   2.632353 292.632353   4.367647 196.987448  17.012552  23.012552
         8
  1.987448
```

Suppose our interest is actually in how the (binary) response variable SURV, depends on CLINIC and CARE. It may have occurred to us to alternatively fit a **logistic regression model** for SURV with explanatory variables CLINIC and CARE.

This can be done in R as follows. Note that this requires the data to be entered differently.

```
> surv <- c(176, 293, 197, 23)
> died <- c(3, 4, 17, 2)
> cl <- c("A", "A", "B", "B")
> ca <- rep(c("L", "H"), 2)
> total <- surv + died
> clinic <- factor(cl)
> care <- factor(ca, levels = c("L", "H"))
> prop <- surv/total
> HLR <- data.frame(surv, died, prop, total, clinic, care)
> HLR
   surv died      prop total clinic care
1  176    3 0.9832402   179      A    L
2  293    4 0.9865320   297      A    H
3  197   17 0.9205607   214      B    L
4   23    2 0.9200000    25      B    H
```

Hence, a logistic regression can be fitted for the observed proportions of surviving babies, as follows.

```
> health.glm <- glm(prop ~ clinic + care, family = binomial, weights = total,
                            data = HLR)
```

```
> summary(health.glm)

Call:
glm(formula = prop ~ clinic + care, family = binomial, data = HLR,
    weights = total)

Deviance Residuals:
       1         2         3         4
-0.11107   0.09263   0.04706  -0.14186

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   4.1372     0.5077   8.149 3.66e-16 ***
clinicB      -1.6991     0.5307  -3.202  0.00137 **
careH         0.1104     0.5610   0.197  0.84403
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 17.828399  on 3  degrees of freedom
Residual deviance:  0.043256  on 1  degrees of freedom
AIC: 19.4

Number of Fisher Scoring iterations: 4
```

This model clearly fits,since the residual deviance of 0.0433 is not significant when compared with a χ^2_1 distribution. We now need to see if a more parsimonious model would fit.

```
> anova(health.glm, test = "Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: prop

Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                       3    17.8284
clinic  1   17.746         2     0.0823 2.524e-05 ***
care    1    0.039         1     0.0433    0.8434
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

shows that there is no need for the CARE term in a model which already contains CLINIC. Checking the other way, it is seen that CLINIC is necessary in a model containing CARE, so that CLINIC can not be dropped. Hence, an appropriate fitting model for the proportions (log odds) of surviving babies has only the main effect of CLINIC. The estimated parameters and fitted responses from this model are shown below.

```
> health.glm <- update(health.glm,  ~ . - care)
> dummy.coef(health.glm)
Full coefficients are

(Intercept):     4.204693
clinic:                 A         B
```

```
                0.000000 -1.755504

> fitted(health.glm)
        1         2         3         4
0.9852941 0.9852941 0.9205021 0.9205021

> fitted(health.glm) * total
        1         2         3         4
176.36765 292.63235 196.98745  23.01255
```

Note that this corresponds exactly to the results of the log linear approach above.