Chapter 5

Log-linear models II: three-way contingency tables

Log-linear models can be used to analyse contingency tables of all orders. Here we will consider three-way tables; the extension to higher order tables follows in a reasonably obvious way. A major difference between two-way and higher order tables is the way in which the concept of independence generalises. In two-way tables there is only one concept of independence: the two factors are either independent or dependent. With three factors there are four types of independence, and another, related concept.

5.1 Types of independence

For three factors $A$, $B$ and $C$:

1. $A$, $B$ and $C$ mutually independent;
2. $A$ independent of $B$ and $C$, together;
3. $A$ independent of $B$ for each level of $C$ (i.e. conditional independence of $A$ and $B$, given the level of $C$);
4. $A$ independent of $B$, marginally over $C$;
5. no three-factor interaction.

Note that the roles of $A$, $B$ and $C$ can be interchanged in each of (2), (3) and (4).

5.2 Log-linear models for three-way tables

Let $y_{ijk}$ denote the observed frequencies, with marginal totals $y_{ij.}, y_{i..}$, etc and let $\lambda_{ijk}, \lambda_{ij.}, \lambda_{i..}$, etc, denote the corresponding expected frequencies.

Then the (saturated) log-linear model is:

$$\log(\lambda_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk}.$$ 

A saturated model is one that has as many (independent) parameters as there are observations. Here, this implies that $\hat{\lambda}_{ijk} = y_{ijk}$. 

40
5.3 Hierarchical models

A hierarchical model is one for which inclusion of a term implies inclusion of all lower-order ‘relatives’. E.g. inclusion of the two-factor interaction, $\alpha\beta_{ij}$, implies inclusion of the corresponding main effects, $\alpha_i$ and $\beta_j$, and the mean, $\mu$.

Notation for hierarchical models

\[
\begin{align*}
[ABC] & \quad \log(\lambda_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk} \\
[AB][AC][BC] & \quad \log(\lambda_{ijk}) = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} \\
[AB][AC] & \quad \log(\lambda_{ijk}) = \mu + \alpha_i + \beta_j + \alpha\gamma_{ik} \\
[AB][C] & \quad \log(\lambda_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} \\
[A][B][C] & \quad \log(\lambda_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k
\end{align*}
\]

5.4 Log-linear models and independence

In most cases it is possible to express the various forms of independence in terms of a log-linear model as indicated below. If, for example, the model $[AC][BC]$ provides a good fit to the data then we would conclude that, for each level of factor $C$, there is no significant association between factors $A$ and $B$.

(a) $A$, $B$ and $C$ mutually independent: $[A][B][C]$

(b) $A$ independent of $B$ and $C$ together: $[A][BC]$

(c) $A$ independent of $B$ for each level of $C$: $[AC][BC]$

(d) $A$ independent of $B$, marginally over $C$: no 3-factor model

(e) no three-factor interaction: $[AB][AC][BC]$

5.4.1 Interpretation of the three-factor interaction

No three-factor interaction implies that the association between factors $A$ and $B$, say, is the same for each level of factor $C$. For example, for a table with factors gender, smoking and lung-cancer, no three-way interaction would imply that the odds ratio in favour of a smoker dying of lung cancer is the same for males and females, even though the incidence of lung cancer might be substantially different for male and female smokers.
Example: Car preferences (as a three-way contingency table)

GLIM analysis

```
[i]  ? $units 8
[i]  ? $data y$read 168 32 68 12 84 164 16 24
[i]  ? $calc s=gl(2,4):r=gl(2,1):p=gl(2,2)
[i]  ? $fact s 2 r 2 p 2$yvar y$error p$
[i]  ? $fit r+s+p$
   scaled deviance = 172.03 at cycle 4
   d.f. = 4
[i]  ? $fit s+r*p$
   scaled deviance = 164.41 at cycle 4
   d.f. = 3
[i]  ? $fit r+s*p$
   scaled deviance = 153.40 at cycle 4
   d.f. = 3
[i]  ? $fit p+s*r$
   scaled deviance = 19.236 at cycle 3
   d.f. = 3
[i]  ? $fit s*r+r*p$
   scaled deviance = 11.620 at cycle 3
   d.f. = 2
[i]  ? $fit s*p+r*r$p
   scaled deviance = 145.78 at cycle 4
   d.f. = 2
```

The only models that provide an adequate fit to the data are \([SR][SP]\) and \([SR][SP][RP]\). Further, \([SR][SP][RP]\) is not significantly better than \([SR][SP]\) hence we conclude that \([SR][SP]\) is the most appropriate model. The interpretation of this model is that, given gender (S), preference is independent of residence.
Example: Occupation, education and aptitude

The following data were obtained from 4353 individuals classified according to occupation (4 groups: 1 = self-employed, business; 2 = teacher; 3 = self-employed, professional; 4 = salary-employed), education (4 levels) and aptitude (5 levels as measured by a Scholastic Aptitude Test).

<table>
<thead>
<tr>
<th>E</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| o | 1 1 | 42.000 | 55.000 | 22.000 | 3.000 |
| o | 2 | 72.000 | 82.000 | 60.000 | 12.000 |
| o | 3 | 90.000 | 106.000 | 85.000 | 25.000 |
| o | 4 | 27.000 | 48.000 | 47.000 | 8.000 |
| o | 5 | 8.000 | 18.000 | 19.000 | 5.000 |

| o | 2 1 | 0.000 | 0.000 | 1.000 | 19.000 |
| o | 2 | 0.000 | 3.000 | 3.000 | 60.000 |
| o | 3 | 1.000 | 4.000 | 5.000 | 86.000 |
| o | 4 | 0.000 | 0.000 | 2.000 | 36.000 |
| o | 5 | 0.000 | 0.000 | 1.000 | 14.000 |

| o | 3 1 | 1.000 | 2.000 | 8.000 | 19.000 |
| o | 2 | 1.000 | 2.000 | 15.000 | 33.000 |
| o | 3 | 2.000 | 5.000 | 25.000 | 83.000 |
| o | 4 | 2.000 | 2.000 | 10.000 | 45.000 |
| o | 5 | 0.000 | 0.000 | 12.000 | 19.000 |

| o | 4 1 | 172.000 | 151.000 | 107.000 | 42.000 |
| o | 2 | 208.000 | 198.000 | 206.000 | 92.000 |
| o | 3 | 279.000 | 271.000 | 331.000 | 191.000 |
| o | 4 | 99.000 | 126.000 | 179.000 | 97.000 |
| o | 5 | 36.000 | 35.000 | 99.000 | 79.000 |

Scaled deviance = 1357.0 at cycle 5
D.f. = 69

Scaled deviance = 1179.6 at cycle 5
D.f. = 57

Scaled deviance = 1319.6 at cycle 5
D.f. = 57

Scaled deviance = 228.22 at cycle 4
D.f. = 60

Scaled deviance = 190.81 at cycle 4
D.f. = 48

Scaled deviance = 50.892 at cycle 4
D.f. = 48
The model \([OE][AE]\) (given education, occupation is independent of aptitude) provides a good fit to the data, but the model \([OE][OA][EA]\) provides a significantly better fit. Try looking at the residuals for the model \([OE][AE]\).

Most of the large residuals come from the teachers, almost all of whom have education level 4. Investigate what happens when teachers are omitted.

Omit teachers
ACD CHAPTER 5: Three-way contingency tables

We conclude that, given education level, occupation is independent of aptitude, especially for occupational groups other than teachers. For teachers there are too few with lower educational levels to be able to draw meaningful conclusions.

5.5 Collapsibility

When dealing with higher dimensional contingency tables it is often desirable to collapse the table over one or more factors in order to simplify the analysis, and subsequent interpretation. In particular, people often consider only the two-factor marginal tables obtained by collapsing, or summing, over all other factors. For some cases this is a reasonable thing to do; for others it is totally inappropriate. Here we consider the case of collapsing over one factor in a three-way table, though the results can readily be extended to cover tables of higher dimension.

Consider a three-way table with factors $A$, $B$ and $C$. If we collapse over factor $C$ (say), to produce a two-way table with factors $A$ and $B$, then the only association we can investigate is the association between factors $A$ and $B$. It can be shown that the association between factors $A$ and $B$ in the collapsed table will be the same as in the three-way table if, and only if, in the three-way table:

(i) there is no three factor interaction, and

(ii) at least one of the two factor interactions involving $C$ is zero.

In practice we interpret no interaction as meaning negligible interaction.

Example: Car preferences (continued)

The table below gives estimates, with standard errors in parentheses, of the parameters associated with the various interactions for various models. These parameters measure the association between the relevant factors, and a value of zero indicates no association.

For the three-way table it is possible, using the model with all three two-factor interactions, to assess the association between each pair of factors; for the collapsed tables it is only possible to assess the association between the remaining two factors.

From the parameter estimates for the model $[SR][SP][RP]$ we find that there is significant association between factors $S$ and $R$ and between factors $S$ and $P$, but not between factors $R$ and $P$. Omitting the $R.P$ interaction we find that the estimates of the other interaction parameters are different, but only slightly.

Looking at the estimates of the interaction parameters obtained from the collapsed tables we find that we get exactly the same values for the estimates of $S.R$ and $S.P$ as we did for $[SR][SP]$, but
that we get a substantially different value for $R.P$. This implies that it is OK to collapse over $R$ or $P$ ($R.P$ is not significant), but that it is not OK to collapse over $S$.

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter estimates (standard errors)</th>
<th>Scaled deviance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S.R</td>
<td>S.P</td>
</tr>
<tr>
<td>[SR][SP][RP]</td>
<td>2.29 (0.21)</td>
<td>-0.82 (0.23)</td>
</tr>
<tr>
<td></td>
<td>145.64</td>
<td>11.48</td>
</tr>
<tr>
<td>[SR][SP]</td>
<td>2.31 (0.21)</td>
<td>-0.91 (0.22)</td>
</tr>
<tr>
<td></td>
<td>152.79</td>
<td>18.63</td>
</tr>
</tbody>
</table>

**Collapsed tables**

<table>
<thead>
<tr>
<th></th>
<th>S.R</th>
<th>S.P</th>
<th>R.P</th>
</tr>
</thead>
<tbody>
<tr>
<td>over $S$ [$RP$]</td>
<td>-0.60 (0.22)</td>
<td></td>
<td>7.62</td>
</tr>
<tr>
<td>over $R$ [SP]</td>
<td></td>
<td>-0.91 (0.22)</td>
<td>18.63</td>
</tr>
<tr>
<td>over $P$ [SR]</td>
<td>2.31 (0.21)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>152.79</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

It is worth noting that the parameter estimates, and their standard errors, and the scaled deviances that are obtained from the collapsed 2-way tables, can be obtained by fitting appropriate models to the (full) 3-way table.
Chapter 6

Log-linear models III: model selection

6.1 Model selection

Selecting the ‘best’ model involves a trade-off between adequacy of fit and simplicity. We usually want to find the simplest model, or models, which provide a good fit to the data.

For even moderate sized problems, the number of possible models is very large; often tens of thousands.

There will often be a number of models that it is not possible to choose between on statistical grounds. Only pairs of models for which one is a special case of the other can be formally compared.

There is no all-purpose, best method of model selection. It is possible, and often desirable, to use stepwise procedures, similar to those used in (standard) regression.

Depending on the nature of the factors being considered (response or explanatory) and the way in which the study was carried out, it may be appropriate to consider only a restricted range of models. For example, if there is a single response factor and two or more explanatory factors, then it doesn’t make much sense to conclude that, given the response, the explanatory factors are independent. Sometimes the design of the study implies that certain terms should be in any model considered; the model that consists of just those terms is referred to as the minimal model.

There are a couple of approaches that get mentioned in a number of texts on the subject namely:

1. If there is one response factor (R) and a set of explanatory factors (E1, E2, ..., Ek) then start with the model \([E1\ E2\ \ldots\ Ek][R]\) and add interactions between R and the Ej’s until you find an adequate fit (or fits). This approach can be extended in an “obvious” way if there is more than one response factor.

2. Fit the sequence of models:
   main effects only
   all two-factor interactions
   all three-factor interactions
   etc
   and determine the simplest such model that provides a good fit to the data. Then consider dropping interaction terms from that model until the smallest (hierarchical) model that provides an adequate fit is obtained.
Another approach, advocated by some authors, is to use the “*Akaike Information Criterion*” (AIC).

\[
\text{AIC} = \text{(scaled) deviance} + 2k
\]

where \(k\) is the number of (linearly independent) parameters for the model (so that the scaled deviance has \(n - k\) degrees of freedom).

The model with the smallest AIC is preferred.

### 6.2 Interpretation and uses of hierarchical models

Two cases:

(i) At least one two-factor interaction not significant ⇒ (conditional) independence, the nature of which can readily be inferred from the graph of the smallest graphical model that contains the specific model.

**Graphical models:** A model is graphical if, whenever the model contains all two-factor terms generated by a higher-order interaction, the model also contains the higher-order interaction.

**Chain:** A sequence of “edges” that links two factors.

**(Conditional) independence:** Let the sets \(A\), \(B\) and \(C\) denote disjoint subsets of the factors in a graphical model. Then the factors in \(A\) are independent of the factors in \(B\) given \(C\) if and only if every chain between a factor in \(A\) and a factor in \(B\) involves at least one factor in \(C\).

(ii) All two-factor interactions significant — used as a smoothing device. The following is an example of this type of use.
Example: Compliance with a Screening Test for Cancer (males only)

<table>
<thead>
<tr>
<th></th>
<th>Primary</th>
<th>Secondary</th>
<th>Tertiary</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
<td>N</td>
<td>R</td>
</tr>
<tr>
<td>Smokers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>3</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>E</td>
<td>1.4</td>
<td>12.7</td>
<td>3.9</td>
</tr>
<tr>
<td>%</td>
<td>8</td>
<td>71</td>
<td>21</td>
</tr>
<tr>
<td>Non-smokers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>13</td>
<td>26</td>
<td>4</td>
</tr>
<tr>
<td>E</td>
<td>14.1</td>
<td>25.2</td>
<td>3.7</td>
</tr>
<tr>
<td>%</td>
<td>33</td>
<td>58</td>
<td>9</td>
</tr>
</tbody>
</table>

Legend:

- C = compliance; N = non-compliance; R = refuse to participate
- O = observed frequency; E = expected frequency;
- % = expected frequency as a percentage for that group of subjects (e.g. primary educated smokers)

Factors:  
- S = smoking; EL = education level; RES = response (C, N or R)

Variable:  
- $x = EL$

<table>
<thead>
<tr>
<th>Model</th>
<th>Scaled Deviance</th>
<th>df</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S + EL + RES + S.EL$</td>
<td>40.36</td>
<td>10</td>
<td>56.36</td>
</tr>
<tr>
<td>$S + EL + RES + S.EL + EL.RES$</td>
<td>25.76</td>
<td>6</td>
<td>49.76</td>
</tr>
<tr>
<td>$S + EL + RES + S.EL + S.RES$</td>
<td>16.71</td>
<td>8</td>
<td>36.71</td>
</tr>
<tr>
<td>$S + EL + RES + S.EL + S.RES + EL.RES$</td>
<td>2.61</td>
<td>4</td>
<td>30.61</td>
</tr>
<tr>
<td>$S + EL + RES + S.EL + S.RES + RES.x^*$</td>
<td>3.04</td>
<td>6</td>
<td>27.04</td>
</tr>
</tbody>
</table>

* model used to calculate expected frequencies